Anais do
Décimo e Quarto Encontro dos Alunos e Docentes do Departamento de Engenharia de Computação e Automação

Campinas, São Paulo
25 e 26 de agosto de 2022

Faculdade de Engenharia Elétrica e de Computação
Universidade Estadual de Campinas

Diretoria da Faculdade de Engenharia Elétrica e de Computação

Prof. José Alexandre Diniz - Diretor

Prof. Paulo Cardieri – Diretor Associado

Conselho Departamental do Departamento de Engenharia de Computação e Automação Industrial

Alim Pedro de Castro Gonçalves - Chefe

Christian Rodolfo Esteve Rothenberg

Denis Gustvo Fantinato

Eduardo Alves do Valle Junior

Eleri Cardozo

Eric Rohmer

Fernando José Von Zuben

José Mario De Martino

Léo Pini Magalhães

Letícia Rittner

Levy Boccato – Vice-chefe

Marco Aurélio Amaral Henriques

Matheus Souza

Paula Dornhofer Paro Costa

Rafael Ferrari

Ricardo Ribeiro Gudwin

Romis Ribeiro De Faissol Attux

Wu Shin –Ting

Logotipo do evento: Rodolfo Luis Tonoli
Diagramação dos anais: Wu Shin Ting
Editores: Léo Pini Magalhães, Rodolfo Toloni e Wu Shin Ting

Apresentação

É com muita satisfação que apresentamos esta coletânea de contribuições para a Décima Quarta Edição do Encontro de Alunos e Docentes do Departamento de Engenharia de Computação e Automação - EADCA - da Faculdade de Engenharia Elétrica e de Computação da Unicamp, realizado nos dias 25 e 26 de agosto de 2022. Esta edição retorna ao formato presencial, visto que a edição anterior, em 2020, foi realizada de forma remota mas nem por isso menos exitosa.

Para esta edição recebemos 24 contribuições mostrando o esforço de alunos e professores na retomada das atividades presenciais neste primeiro ano acadêmico presencial após dois anos de atividades eminentemente remotas e de isolamento social impostos pela pandemia. Observamos que foi mantido o formato de quatro páginas para os trabalhos de modo a que os autores apresentassem as idéias do seu trabalho de forma abrangente.

Esta publicação apresenta adicionalmente os resumos das quatro palestras proferidas. Uma delas no âmbito de atividades de nossa Agência de Inovação e as outras três por professores recentemente aposentados e assim homenageando todos os docentes que já atuaram em nosso departamento.

Somos gratos a todos os membros do DCA, especialmente à Sra. Clarice Fincatti da Silva e à chefia do departamento, pelo apoio demonstrado desde a primeira hora; à Diretoria da FEEC pelo incentivo e pelo decisivo suporte; aos coordenadores das sessões, que tão gentilmente aceitaram os convites recebidos; aos autores e orientadores, que cuidadosamente conduziram o trabalho de elaboração dos resumos, assegurando a qualidade desta coletânea; e aos alunos do DCA por terem comparecido a este encontro que lhes pertence.

Por fim, esperamos que o evento deste ano tenha também mantido a tradição de contribuir para a formação de pesquisadores no seio de um profícuo ambiente científico.

Cordialmente,

Léo Pini Magalhães Rodolfo Luis Tonoli Wu Shin-Ting

Comissão Organizadora

agosto de 2022

FEEC | UNICAMP
EADCA22
XIV ENCONTRO DE ALUNOS E DOCENTES DO DCA

Palestras Convidadas

Exemplos de Aplicações de IA no Brasil: Trinta anos de IA na FEEC

Prof. Fernando A. C. Gomide

A inteligência artificial é um assunto que vem sendo estudado na FEEC desde a década de 80. As atividades de ensino e pesquisa em IA na FEEC tiveram, desde suas origens, uma visão de engenharia e suas aplicações na indústria, empresas e áreas correlatas. A apresentação, após explicitar a visão da IA da FEEC na época, mostra exemplos de aplicações onde a FEEC e seus alunos estiveram envolvidos nas últimas três décadas. Exemplos de aplicação de interesse contemporâneo também serão brevemente discutidos.

Navegar é Preciso

Prof. Rafael Santos Mendes

O objetivo desta palestra é mostrar, através das diversas técnicas de navegação no mar, como conceitos geométricos e trigonométricos, assim como a teoria de filtragem estocástica se articulam para permitir a localização de um objeto em movimento. São inicialmente abordadas as técnicas de pilotagem e navegação por referências costeiras e em seguida, após a introdução de conceitos elementares de astronomia e de trigonometria esférica, são apresentados alguns aspectos da navegação astronômica. Finalmente, discute-se o funcionamento básico do Sistema de Posicionamento Global (GPS) relacionando-o com as técnicas de filtragem estocástica, em particular com o filtro de Kalman Estendido (EKF).

CPS, Sistemas Ciberfísicos: Ensino e Desenvolvimento

Prof. José Raimundo de Oliveira

Esta palestra procura apresentar os conceitos inerentes ao projeto de sistemas ciberfísicos (CPS, do inglês *Cyber-Physical System*) e como isto pode interferir na formação de novos engenheiros em nossa escola. Os sistemas ciberfísicos (CPS) são sistemas que são construídos e dependem da sinergia de componentes físicos e computacionais. Os projetistas de CPS enfrentam o desafio de gerenciar restrições em vários domínios de conhecimento como, por exemplo: software, transdutores, processamento de sinal, comunicação e rede, arquitetura de computador, controle, simulação e modelagem de processos físicos. A formação de especialistas em algum destes domínios pode até ajudar o desenvolvimento de componentes individuais de projeto. Mas na medida que que a sofisticação dos sistemas ciberfísicos aumenta, aumenta também a necessidade de desenvolvedores com conhecimento interdisciplinar de como esses componentes interagem na composição com outros.

Propriedade Intelectual e Inovação na Unicamp

Elisama Campelo Santos

Na palestra abordamos os conceitos básicos de propriedade intelectual (PI); programa de computador e sua proteção; patentes e questões entre "Proteger ou Publicar?"; e as orientações de como a Inova Unicamp auxilia nesses processos e casos de exemplo.

FEEC | UNICAMP
EADCA22
XIV ENCONTRO DE ALUNOS E DOCENTES DO DCA

Sessão Técnica 1

In-band Inter Packet Gap Telemetry (IPGNET): Unlocking novel network monitoring methods

Francisco Germano Vogt , Christian Esteve Rothenberg

{f234632@dac.unicamp.br, chesteve@dca.fee.unicamp.br}

Departamento de Engenharia de Computação e Automação (DCA)

Faculdade de Engenharia Elétrica e de Computação (FEEC)

Universidade Estadual de Campinas (Unicamp)

Campinas, SP, Brasil

Abstract – Network monitoring is a fundamental task to provide good network management and performance. Since the SDN emergence, the In-band Network Telemetry (INT) has been demonstrated as an efficient network monitoring framework. Using INT, we can collect network information hop-by-hop directly from the data plane by including this information in the network production traffic. However, this information collection is limited by available packet size and processing overhead, making it critical to choose what information to collect and when to collect it. So, in this work, we propose the Inter-Packet Gap (IPG) per-hop monitoring using INT. We argue that by monitoring the IPG hop-by-hop, it is possible to correlate the data and identify network problems like (1) Network congestions, (2) Network delay, and (3) Microbursts and their contributing flows. Our preliminary results show that IPGNET can efficiently detect the microbursts on multiple queues and report all the contributing flows.

Keywords – Programmable Data Planes, SDN, P4, In-band Network Telemetry, Network Monitoring

(Texto removido pela cláusula de sigilo dos convênios com as empresas.)

Estudo de Abordagem de Detecção de Anomalias no Contexto do Monitoramento Inteligente com Câmeras

Guilherme Magalhães Soares , José Mario De Martino
{g217241@dac.unicamp.br, martino@unicamp.br}

Departamento de Engenharia de Computação e Automação (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Campinas, SP, Brasil

Resumo – Câmeras de segurança têm se tornado cada vez mais presentes no espaço público para o monitoramento de ocorrências e ações ilícitas. Entretanto, a análise de vídeos é um trabalho manual exaustivo, muitas vezes ineficaz para oferecimento de auxílio em tempo real. Buscando mitigar esse problema, este projeto descreve o estudo da utilização de redes neurais da arquitetura *autoencoder* na detecção de situações consideradas anômalas. Nesse estudo foram utilizadas imagens de câmera de monitoramento localizada na entrada principal da Faculdade de Engenharia Elétrica e de Computação da Unicamp. O estudo focou a detecção de aglomeração no local como evento anômalo. Os resultados dos experimentos indicam que o *autoencoder* implementado é adequado para a detecção da anomalia escolhida.

Palavras-chave – Visão computacional, Deep Learning, Reconhecimento de Anomalias.

1. Introdução

Atualmente, os sistemas de segurança por monitoramento são altamente dependentes de verificação manual e, em diversos casos, infecazez a depender da razão entre número de funcionários alocados para o serviço e de câmeras conectadas ao sistema. Entretanto, devido ao aumento da capacidade de processamento de dados dos sistemas computacionais atuais e refinamento de técnicas de Inteligência Artificial (IA), a aplicação de sistemas autônomos para reconhecimento de atividades anômalas, como aglomerações, roubos e assaltos, por exemplo, tem se tornado cada vez mais acessível, possibilitando auxílio ágil.

O Monitoramento Inteligente com Câmeras propõe, então, uma abordagem utilizando Visão Computacional e técnicas de Reconhecimento de Atividades Humanas (RAH) para detectar e reconhecer esses eventos em contexto de câmeras de segurança. Para tanto, o estudo e a construção de *datasets* são extremamente necessários, podendo estabelecer a viabilidade e a qualidade do modelo de aprendizado profundo construído.

Neste contexto, busca-se a construção de um *dataset* de vídeos de câmera de segurança e análise sobre capacidade de reconhecimento de atividades anômalas por arquiteturas de inteligência artificial no conjunto obtido.

2. Proposta

Para entendimento sobre o problema proposto, foi necessário o estudo de conceitos associados às áreas de Aprendizado de Máquina e Reconhecimento de Atividades Humanas. Para a implementação da solução proposta, também foi necessário o estudo da linguagem de programação *Python* e do *framework* Keras.

Este trabalho utiliza como base a arquitetura Autoencoder Espaço-temporal [1] , aplicando-a no contexto de câmeras de segurança usando como dados obtidos das câmeras da Faculdade de Engenharia Elétrica e de Computação (FEEC). O treinamento e testes foram realizados no ambiente *Google Colabolatory*.

2.1. Aprendizado de Máquina

As redes convolucionais [4] são referências clássicas na área de Visão Computacional e, ultimamente, diversos modelos baseados em Vision Transformer (ViT) [2] as tem superado em áreas como segmentação e classificação de imagens, detecção de objetos e síntese de imagens. Entretanto, essa arquitetura necessita de grandes quantias de dados e anotações extensivas, impraticáveis para análise de acontecimentos anômalos.

A escolha por um método de aprendizado não supervisionado é, portanto, o ideal para o problema, evitando processos de anotação. Além disso, como algumas ações humanas dependem de movimentação e, assim, são definidas em mais de um frame, levando à escolha por um modelo que possua memória, como LSTM [3] .

Com isso, escolheu-se como estudo de caso a arquitetura *autoencoder* [6] com camadas LSTM Convolucionais [7] , denominada *autoencoder* Espaço-Temporal , ilustrada na Figura 1, capaz de aprender exatamente o que é dado como entrada e, assim, distinguir *outliers* (pontos fora do padrão). Neste caso, o *autoencoder* recebe uma

imagem, realiza a passagem pela rede neural e tenta reconstruir a imagem original na saída, retornando o quanto ele errou no processo. Ou seja, espera-se que treinando o modelo com vídeos considerados normais, ou dentro do cotidiano, a arquitetura retorne erros altos nos vídeos que contenham alguma anomalia.

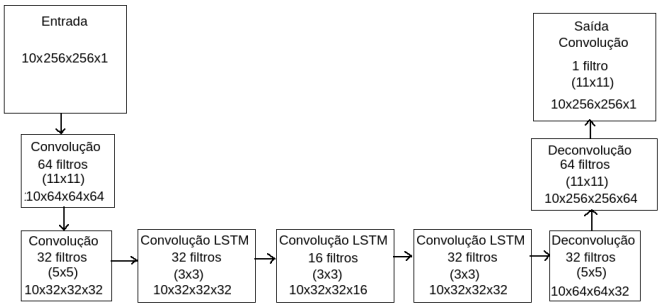

Figura 1. Ilustração da arquitura de *autoencoder* Espaço-Temporal montada.

Nesse contexto, cada vídeo foi separado em *frames* de 256x256 pixels compondo sequências de dez *frames* a partir da técnica de *Sliding Window* - ou Janela Deslizante. Cada sequência é composta por frames intercalados de trinta em trinta, com configuração para, além da sequência original como, por exemplo, *frame* 0, 30, 60, etc., obter também variações dela, aumentando a quantidade de sequências total e, assim, possibilitando treinamento maior.

A métrica para reconhecimento de anomalia é o erro da reconstrução da sequência devolvida pelo *autoencoder*. Neste caso, será montado um gráfico de erro normalizado ao longo dos *frames* do vídeo de teste, computado a partir da norma da subtração entre as sequências de entrada e de saída do *autoencoder*. Para consideração de *threshold*, ou limiar de erro, foi feito análise a partir de vídeos de teste considerados cotidianos e, a partir do pico gerado, definir este valor.

2.2. Conjunto de dados e Reconhecimento de Anomalias

Analisando a literatura sobre RAH, é possível identificar duas abordagens para construção de *datasets* na área: unimodal e multimodal. A primeira se refere à ingestão de apenas um tipo de dado, como um conjunto de imagens, enquanto a segunda, a mais de um, como imagens e sensor de profundidade, por exemplo. Como o escopo do projeto é a análise em câmeras de segurança, optou-se pela organização de dados unimodais: vídeos sem áudio.

Nesse contexto, foram utilizados vídeos gerados pelas câmeras de segurança da FEEC com perspectiva estática para a entrada principal do prédio. Os vídeos obtidos tem duração de uma hora, percorrendo dois dias inteiros. Seu conteúdo apresenta a recepção, armários, a porta automática de entrada e o dispensor de álcool como estruturas fixas, enquanto a quantidade de pessoas pelo local é a única característica variável.

Para análise, foi considerado cotidiano uma baixa movimentação de pessoas pelo local, sem obstrução visual das estruturas fixas do ambiente, como ilustrado na Figura 2. Por outro lado, sequências de *frames* que contenham aglomerações de pessoas, como na Figura 3, são consideradas anômalias.

Para validação da qualidade de vídeo e estruturação dos arquivos, utilizou-se como embasamento *datasets* bem reconhecidos na área, como o UCF101 [9] , Kinetics [8] e Moments in Time [5] . Nesse aspecto, apenas vídeos que não possuem quedas abruptas na quantidade de *frames* por segundo (FPS) ou com imagens apresentando deterioração, como pontos pretos ou regiões sem nitidez, foram selecionados. A resolução original das imagens - 1920x1080 - se mostrou satisfatória, comparada aos trabalhos anteriormente mencionados.

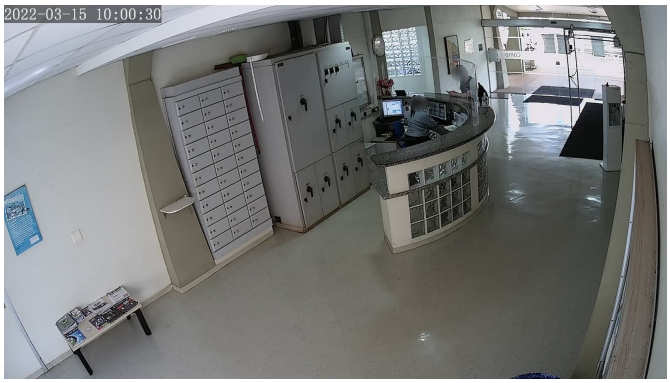

Figura 2. Exemplo de frame considerado cotidiano na entrada da FEEC-Unicamp.

Figura 3. Exemplo de frame considerado anomalia devido à uma aglomeração na entrada da FEEC-Unicamp.

3. Resultados

Para definição do limiar de erro do sistema, foram utilizados vídeos de teste cujo conteúdo é considerado cotidiano. Um exemplo é exposto na Figura 4 e o gráfico de erro de reconstrução gerado a partir deste é ilustrado na Figura 5.

Figura 4. Frame considerado cotidiano utilizado em vídeos de teste do sistema.

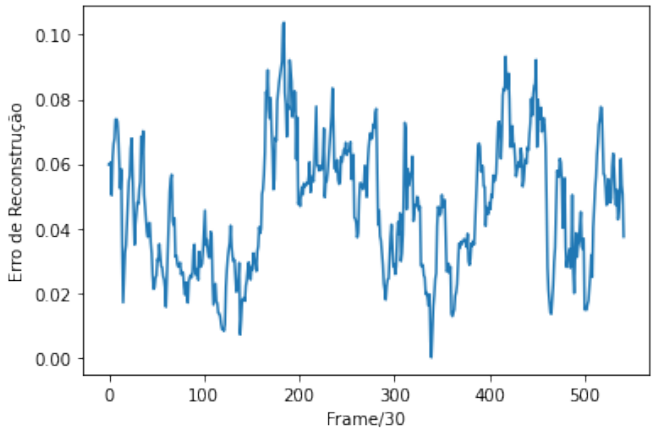

Figura 5. Curva do erro de reconstrução de um vídeo teste de conteúdo considerado cotidiano. Nota-se o pico de erro em 0.10, ou 10%.

Os testes com outros vídeos de conteúdo semelhante apresentaram este mesmo comportamento, permitindo definir o *threshold* como 0.10 dentro da escala de 0 a 1 estabelecida.

Com isso, o gráfico ilustrado na Figura 6 foi gerado utilizando o vídeo de teste de conteúdo anômalo.

O *autoencoder* retornou um pico de erro de reconstrução acima de vinte porcento no *frame* 3570, que é exposto na Figura 7. Como pode ser visto, ocorre uma grande aglomeração de pessoas, o que é considerado anormal pelo padrão adotado, demonstrando a capacidade de reconhecimento da arquitetura escolhida.

Além disso, nota-se a diminuição progressiva do erro ao longo dos *frames*, com eventuais picos locais. Isto está

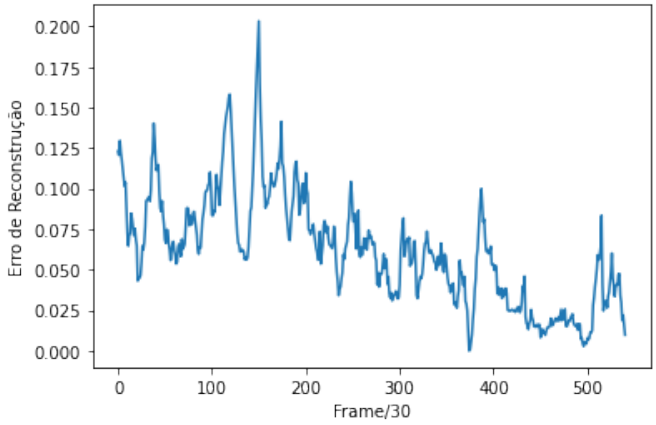

Figura 6. Curva do erro de reconstrução das sequências de teste por frame normalizado.

Figura 7. Frame de maior erro de reconstrução no vídeo de teste. Nota-se aglomeração expressiva, apontada como anomalia pelo *autoencoder*.

atrelado à uma redução no número de pessoas na entrada, com movimentações menores ao longo do tempo, o que é esperado para o local analisado.

4. Conclusões

Neste trabalho foi construído um conjunto de vídeos da entrada principal do prédio da FEEC-Unicamp para análise da capacidade de reconhecimento de atividades humanas consideradas anômalas para o contexto da faculdade por redes neurais.

Foi considerado como normal, ou cotidiano, vídeos contendo baixa movimentação de pessoas, e, como teste, um vídeo contendo aglomeração de pessoas durante o primeiro dia de aula na universidade no ano de 2022.

Como estudo de caso, utilizou-se a arquitetura *autoencoder* com camadas convolucionais e LSTM, a qual demonstrou eficiência na detecção dos momentos considerados anômalos. A métrica utilizada se baseou no erro de reconstrução da sequência de *frames*, com um pico no momento de maior movimentação.

Em trabalhos futuros pretende-se analisar a capacidade de generalização do *autoencoder*, explorando a possibilidade de utilização de mais de uma câmera, mudando o ambiente e, ainda assim, conseguir distinguir momentos considerados anômalos.

Agradecimentos

Este projeto se insere no âmbito do Programa DAI (Doutorado Acadêmico para Inovação) do CNPq que busca fortalecer a pesquisa, o empreendedorismo e a inovação em Instituições Científica, Tecnológica e de Inovação, por meio do envolvimento de estudantes de graduação e pós-graduação em projetos de interesse do setor empresarial.

Referências

[1] Yong Shean Chong and Yong Haur Tay. Abnormal event detection in videos using spatiotemporal autoencoder. *CoRR*, abs/1701.01546, 2017.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.

[3] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[4] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

[5] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, and Aude Oliva. Moments in time dataset: one million videos for event understanding, 2019.

[6] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

[7] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *CoRR*, abs/1506.04214, 2015.

[8] Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. A short note on the kinetics-700-2020 human action dataset, 2020.

[9] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012.

Revisão bibliográfica comparativa entre Swin Tranformers e ConvNeXt e proposta de aplicação em Monitoramento Inteligente com Câmeras

Rômullo Randell Macedo Carvalho, José Mario De Martino
{r217905@dac.unicamp.br, martino@unicamp.br}

Departamento de Engenharia de Computação e Automação (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Campinas, SP, Brasil

Resumo – Até a década passada, Redes Convolucionais (ConvNets) representavam, em geral, a melhor solução para Reconhecimento de Atividades Humanas (HAR) utilizando imagens como objeto de estudo. No entanto, outros métodos em Aprendizado de Máquina foram implementados e têm evoluído para obtenção de melhores resultados, com abordagens semelhantes ou não às ConvNets. Entre os algoritmos que se propuseram a isso estão Swin Transformer e ConvNeXt. Esse artigo pretende estudar esses dois modelos e compará-los sucintamente, além de propor um estudo futuro em monitoramento inteligente com câmeras em que podem ser úteis.

Palavras-chave – ConvNeXt; Swin Transformer; Reconhecimento de Atividade Humana.

1. Introdução

Apesar de seu treinamento por *backpropagation* remontar à década de 1980 [5], apenas em 2012 as Redes Convolucionais (ConvNets) foram utilizadas de forma a mostrar seu real potencial em recursos visuais, com a introdução da AlexNet [4]. Desde então, a área evoluiu rapidamente, enfatizando diversos aspectos (precisão, eficiência e escalabilidade), popularizando-se a ponto de se tornar dominante sobre outras arquiteturas em Visão Computacional.

Em concomitância, as Redes Neurais Recorrentes apresentavam os resultados mais satisfatórios em Processamento de Linguagem Natural (PNL). Esse cenário mudou, em 2017, com a proposta da arquitetura dos Transformers [12], que, a princípio, dispensava completamente recorrências e convoluções.

Embora as tarefas de Visão Computacional e de PNL apresentem diferenças estruturais significativas, a evolução dos Transformers para linguagem trouxe o questionamento de o quão satisfatórios seriam os resultados de alguma adaptação visual dessa arquitetura em relação às ConvNets. A introdução de Vision Transformers (ViT) [2] gerou impacto, ainda mais com a ajuda de tamanhos maiores de modelos e conjunto de dados, superando ConvNets padrão por uma margem significativa em classificação de imagens.

Apesar desses resultados positivos em classificação de imagens, o estudo em Visão Computacional é bem mais amplo, reunindo diversas outras tarefas específicas. A complexidade da arquitetura do ViT é quadrática em relação ao tamanho da entrada e, mesmo isso não sendo um problema inaceitável para classificação de imagens como proposto em [2], usando os dados do ImageNet [1], torna-se intratável para entradas de maior resolução.

Nesse ponto, destaca-se que o ViT é bem semelhante aos Transformadores originalmente empregados em PNL, sem mesmo introduzir um viés (bias) indutivo específico da imagem. Em geral, as soluções em boa parte das tarefas de Visão Computacional utilizam um modelo de "janela deslizante" e totalmente convolucional, como é o caso das ConvNets.

Para solucionar tal problema, os Transformers Hierárquicos empregam uma abordagem híbrida. Swin Transformer (Swin-T) [7], nessa direção, adota a estratégia similar a "janela deslizante", aproximando-se da característica convolucional das ConvNets e obtendo bons resultados em uma série de tarefas de visão computacional.

Em contrapartida, o sucesso do Swin-T atraiu atenção novamente para modelos ConvNets, uma vez que seu desempenho, em parte, deve-se a uma característica comum dessas redes. A arquitetura ConvNeXt [9], baseada inteiramente em ConvNets, apresenta resultados competitivos com os do Swin-T.

Este artigo apresenta considerações sobre as duas arquiteturas ConvNeXt e Swin Transformer, comparando seus resultados quando pertinente. Por fim, projeta-se em que ambas podem contribuir para a proposta de trabalho futuro em monitoramento inteligente de câmeras.

2. Revisão Bibliográfica

Diferentemente das ConvNets, que evoluíram progressivamente na última década, a adoção do Vision Transformers foi uma mudança de passo. Nesta seção, apresentamos as características que permitem que duas arquiteturas de ponta, uma de cada paradigma, tenham mais recentemente se destacado das demais.

2.1. Swin Transformers

Em oposição aos tokens de palavras (unidades fundamentais de processamento em Transformers linguísticos), os elementos visuais podem variar substancialmente em escala: a segmentação semântica, por exemplo, requer uma previsão densa no nível do pixel. Nos modelos predecessores [12] [2] do Swin-T (Swin Transformer), os tokens são em escala fixa e, portanto, inadequada a muitas tarefas de visão computacional.

Na Figura 1 [7], é apresentada uma comparação do mapeamento da imagem entre um Swin-T e um ViT. Enquanto este mantém uma janela fixa com ao longo das camadas, aquele utiliza uma representação hierárquica que inicia com janelas (delineadas em vermelho) com poucos *patches* (delineados em cinza) e, gradualmente, mescla *patches* vizinhos em camadas mais profundas da arquitetura. A complexidade linear é alcançada computando a autoatenção localmente dentro de janelas não sobrepostas.

Uma característica importante de projeto do Swin Transformer é a estratégia de "janela deslocada" entre camadas consecutivas de autoatenção. No exemplo da Figura 2 [7], na camada *l*, apresenta-se um esquema comum de particionamento de janela; enquanto isso, na camada *l*+1, o particionamento é deslocado, resultando em novas janelas. A computação de autoatenção nas novas janelas cruza os limites das

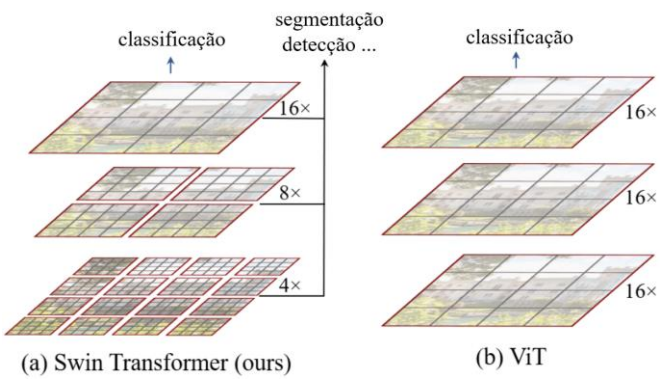

Figura 1. Comparação entre os mapeamentos (a) hierárquico do Swin Transformer e (b) tradicional do Vision Transformer. Adaptada de [7].

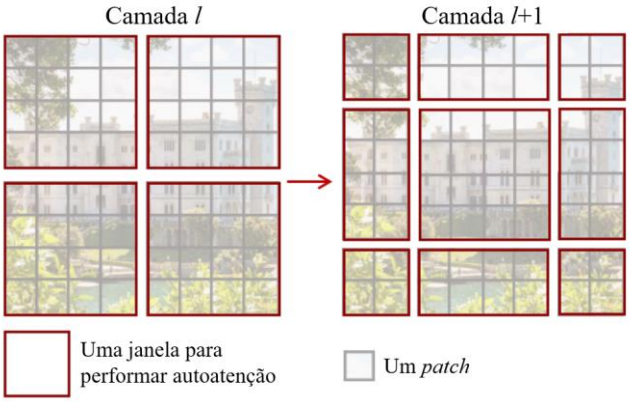

Figura 2. Uma ilustração da abordagem de janela deslocada. Adaptada de [7].

janelas anteriores na camada *l*, proporcionando conexões entre elas, o que aumenta significativamente o poder de modelagem.

Segundo os resultados obtidos em [7], o Swin Transformer apresenta desempenho superior nas tarefas de reconhecimento de classificação de imagem, detecção de objetos e segmentação semântica em relação aos modelos ViT [2] [11] e ResNe(X)t [3] [13] com latência semelhante nas três tarefas. Uma comparação com a ConvNeXt é realizada na Seção 3.

2.2. ConvNeXt

O sucesso dos Transformers em muitas tarefas de visão computacional se deve, em parte, a introduzir convoluções em suas arquiteturas. Essas tentativas, no entanto, têm impactos: a implementação da autoatenção da janela deslizante pode ser cara [10] ou necessitar de abordagens avançadas [8]. Em vez disso, uma alternativa seria inserir características dos Transformers em ConvNets em busca de desempenhos melhores.

A ConvNeXt nasce dessa busca e é construída inteiramente a partir de módulos de ConvNets puras. Inicialmente, foi baseada em uma Rede Neural Residual (ResNet) com treinamento melhorado (utilizando otimizador AdamW e distribuindo melhor a computação em cada estágio, por exemplo) e, gradualmente, sua arquitetura foi aproximada de uma visão hierárquica, semelhante a um Swin-T [9].

Em [9], ConvNeXts foram avaliadas em algumas tarefas de visão, como classificação de imagens com ImageNet [1], detecção/segmentação de objetos com COCO [6] e segmentação semântica com ADE20K [8]. Segundo os resultados obtidos, as ConvNeXts competem favoravelmente com Transformers em termos de precisão, escalabilidade e robustez. Esses resultados são apresentados parcialmente e discutidos na seção seguinte.

3. Resultados Analisados

Apresenta-se na Tabela 1 a comparação entre os resultados do Swin Transformer [7] e da ConvNeXt [9] para a tarefa de classificação de imagens utilizando o banco de dados ImageNet (ImageNet-1K para teste e ImageNet-1K e ImageNet-22K para treino). O treinamento, para ambas as arquiteturas, foi realizado com uma GPU A100, com imagens de tamanho 224×224 ou 384×384 e com mais operações de ponto flutuante por segundo (FLOPs) para as versões maiores das redes (descritas na legenda da Tabela 1).

No caso específico dos resultados da Tabela 1, as diferentes versões da ConvNeXt tiveram, durante o teste, taxas de transferência (imagens por segundo) e acurácia superiores do que as respectivas versões do Swin Transformer. Esse comparativo indica maior eficiência e taxa de acertos da arquitetura puramente convolucional para essa tarefa específica.

Vale ressaltar que estes resultados apresentam superioridade da ConvNeXt em uma tarefa específica. Os próprios autores proponentes dessa arquitetura [9] avaliam que os Transformers devem desempenhar melhor em outras tarefas, sobretudo, as que exijam saídas discretas, esparsas ou estruturadas.

Modelo [1]	Tamanho da imagem	FLOPs	Taxa [2]	Acurácia [3,4] (Teste em IN-1K)	
				Treino IN-1K	Treino IN-22-K
Swin-T	224^2	4,5G	1325,6	81,3	– [5]
ConvNeXt-T	224^2	4,5G	**1943,5**	**82,1**	–
Swin-S	224^2	8,7G	857,3	83,0	–
ConvNeXt-S	224^2	8,7G	**1275,3**	**83,1**	–
Swin-B	224^2	15,4G	662,8	83,5	85,2
ConvNeXt-B	224^2	15,4G	**969,0**	**83,8**	**85,8**
Swin-B	384^2	47,1G	242,5	84,5	86,4
ConvNeXt-B	384^2	45,0G	**336,6**	**85,1**	**86,8**
Swin-L	224^2	34,5G	435,9	–	86,3
ConvNeXt-L	224^2	34,4G	**611,5**	84,3	**86,6**
Swin-L	384^2	103,9G	157,9	–	87,3
ConvNeXt-L	384^2	101,0G	**211,4**	85,5	**87,5**

Tabela 1. Comparação de resultados entre Swin Transformer e ConvNeXt. Adaptada de [9].

[1] Termos -T, -S, -B e -L utilizados de sufixo para representar respectivamente as versões muito pequena (do inglês, *tiny*), pequena (*small*), base (*based*) e grande (*large*).
[2] Taxa de transferência, medida em imagens por segundo.
[3] IN-1K: ImageNet-1K, com 1000 classes e aproximadamente 1,2 milhões de imagens.
[4] IN-22K: ImageNet-1K, com 21841 classes e aproximadamente 14 milhões de imagens.
[5] –: resultados não obtidos.

4. Proposta de Trabalho Futuro

A sucinta revisão bibliográfica descrita ao longo deste artigo tem por objetivo servir de arcabouço teórico para uso de ambas as arquiteturas descritas (ConvNeXts e ViT, mais especificamente Swin Transformers) e de possíveis variações (inclusive, de proposição própria) para o monitoramento inteligente de câmeras.

Pretende-se utilizar vídeos gravados em áreas externas para detecção de atividades anômalas, ou seja, incomuns para o ambiente analisado. Para tal, o banco de dados deve conter imagens de câmeras externas da Universidade Estadual de Campinas (Unicamp) e da cidade de Campinas, com a devida permissão dos órgãos responsáveis e seguindo as orientações usuais de direito de imagens.

A princípio, o objetivo de aplicabilidade prevê uma detecção semiautomática, em que o algoritmo sinaliza uma possível anomalia que deve ser, ainda, avaliada por uma pessoa responsável pelo acionamento de qualquer medida posterior necessária. Inicialmente, as anomalias (aglomerações, assaltos, acidentes etc.) não devem ser identificadas pelo próprio algoritmo, porém, existe planejamento de expansão da proposta para tal.

5. Conclusões

A ConvNeXt, um modelo ConvNet puro, pode ter um desempenho tão bom quanto um Transformer (com visão hierárquica, como o Swin Transformer) em tarefas de classificação de imagem, detecção de objetos, instância e segmentação semântica. Somado a isso, a natureza puramente convolucional para treinamento e teste (mais “simples de implementar”) torna atrativo o uso dessa arquitetura em tarefas de visão computacional.

Entretanto, está muito longe de terem se esgotados os vieses de estudo tanto para Vision Transformers quanto para as ConvNets: o uso da arquitetura dos Transformers para tarefas de visão computacional é recente e ainda mais o contraponto realizado com a ConvNext. Por enquanto, é possível conjecturar que algumas tarefas podem ser mais adequadas a uma ou a outra.

Nesse cenário, ambas as arquiteturas devem ser consideradas e testadas para a tarefa proposta de Monitoramento Inteligente com Câmeras. Próximos trabalhos devem direcionar qual das duas ou mesmo outra arquitetura apresenta melhores resultados para a tarefa.

Agradecimentos

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e da Empresa SiDi.

Referências

[1] Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In: **CVPR**, 2009.

[2] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. **arXiv preprint arXiv:2010.11929.** https://arxiv.org/abs/2010.11929.

[3] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In: **Proceedings of the IEEE conference on computer vision and pattern recognition.** 2016. p. 770–778.

[4] Krizhevsky, A.; Sutskever, I.; Hinton, G. Imagenet classification with deep convolutional neural networks. In: **NeurIPS**, 2012

[5] LeCun, Y.; Boser, B.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W.; Jackel, L. D. Backpropagation applied to handwritten zip code recognition. **Neural computation**, 1989.

[6] Lin, T.-Y; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick. C. L. Microsoft COCO: Common objects in context. In: **ECCV**. 2014.

[7] Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. 2021.

[8] Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. **Swin transformer:** Hierarchical vision transformer using shifted windows. 2021.

[9] Liu, Z.; Mao, H.; Wu, C. Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. 2022. p. 11976-11986. https://arxiv.org/abs/2201.03545.

[10] Ramachandran, P.; Parmar, N.; Vaswani, A.; Bello, I.; Levskaya, A.; Shlens, J. Stand-alone self-attention in vision models. **NeurIPS**, 2019.

[11] Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. **arXiv preprint arXiv:2012.12877**. 2020.

[12] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L; Polosukhin, I. Attention is all you need. In: **Advances in neural information processing systems**, 30. 2017. https://arxiv.org/abs/1706.03762?context=cs.

[13] Xie, S.; Girshick, R.; Dollar, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. p. 1492– 1500. 2017.

[14] Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; Torralba, A. Semantic understanding of scenes through the ADE20K dataset. In: **IJCV**, 2019.

Species distribution modeling applied to MILPA agroecological consortium in Brazil

Matheus Gustavo Alves Sasso, matheus.sasso17@gmail.com
Paula Dornhofer Paro Costa, paulad@unicamp.br

Departamento de Engenharia de Computação e Automação (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Campinas, SP, Brasil

Abstract – Despite playing an essential role in the Brazilian economy, traditional agriculture usually leads to several environmental issues. As an alternative to traditional agriculture, agroecology studies how poly-culture creates beneficial interactions with the environment that can reduce the ecological impact of agriculture. In this paper we propose the Species Distribution Modeling (SDM) algorithms to identify regions in which species from the MILPA agroecological consortium can grow in Brazil. We use a pipeline composed by a One Class Support Vector Machine (OCSVM) combined with multiple ensemble classifiers and neural networks to generate adaptability maps. We also propose the creation of a common database of environmental variables regarding the Brazilian territory which we call *br-env*.

Keywords– Species distribution modeling, Agroecology, MILPA, br-env, Adaptability maps.

1. Introdução

Technological advances aimed at the monocultures cultivation have fulfilled the social role of meeting the food sufficiency of the world population through the development of more efficient irrigation techniques, the mechanization of processes, the use of fertilizers, the application of pesticides, and the creation of transgenics. (more productive hybrid varieties) [1]. However, the high density of a single species (low genetic variability) promotes unstable systems to climate changes, requires constant management due to the depletion of soil nutrients (generating the need to use fertilizers), and provides a high density of pests (requiring the use of pesticides) [2] . In other words, monocultures can be held responsible for important environmental impacts such as contamination of soils, rivers, and air [2].

In the Brazilian context, other problems directly related to monoculture are also identified, such as the deforestation of Brazilian biomes (Amazon, Cerrado, Pantanal), often resulting from the land grabbing process. Also, it promotes the rural exodus and the increase in social inequalities since small farmers do not have the power to compete in the market with large farmers [3].

To face problems related to monoculture, sustainable strategies gained strength, such as agroecological crops, polycultures, and Agroforestry Systems (SAF). Those techniques use the understanding of ecological processes in favor of agricultural production, intending to mitigate the socio-environmental impacts related to monocultures and also increase output by raising the Land Equivalent Ratio Land Equivalent Ratio (LER) [4]–[6]. In these alternatives, each plant plays one or more roles within the system, such as, for example, the functions of retaining nitrogen, producing litter, protecting the system as a live fence for animals or as a barrier against winds, and acting as a ground cover, among others [2]. Past experiences of farmers and indigenous peoples discovered species that can grow together since they play complementary roles, which is defined by the theory of agroecological consortiums [7], [8].

However, agroecology faces the challenge of scale. Replacing the intensive use of agricultural techniques with agroecological consortiums required the understating of how well they can suit to environmental conditions of a region as granular as possible. In this context, the ecological niche theory establishes the relevance of studying relationships between species and the specific requirements of an area as crucial mechanisms for the survival of species and the maintenance of biodiversity [9], [10].

Therefore, the present work aims to explore a machine learning approach called Species Distribution Modeling (SDM) as a tool for identifying ecological niches of agroecological consortium. SDM techniques seek to model their suitability to a proposed region based on environmental data in coordinates of species occurrences. The model predicts how the same species could develop without major environmental interventions in inference time [11]. In this way, SDM techniques can help identify territory regions where an agroecological consortium shows potential for agricultural production.

In particular, this work focuses on the Brazilian territory. It will have as the object of study thw

species of the MILPA consortium, an agroecological consortium initially cultivated by the Mayans, which integrates variations of species of corn, beans, squash, and pepper [12]. We also highlight that MILPA species are relevant to the food security objective from Food and Agriculture Organization of the United Nations [13].

2. Proposal

The general objective of this work is to study, apply and evaluate computational techniques for modeling the distribution of species from the MILPA consortium to identify regions from Brazil suitable for the establishment of agroecological crops and understand, in more detail, the potentials and limitations of SDMs. We describe the criteria to decided which species from MILPA consortium we decided to study, resulting in the following species: Zea mays, Cucurbita pepo, Cajanus cajan and Capsicum annuum.

The first step to applying SDMs focus in Brazil was having data representing the country's environment. For this reason we propose *br-env*, a standardized ensemble of environmental information from multiple sources framed in the Brazilian territory. Next, we explore SDMs as a binary classification problem from a two steps algorithm. First, we generate pseudo-absence species using a OCSVM once this data is not naturally available. Next, we evaluated different classifiers that calculate the probability of a coordinate being present or absent. In addition, once *br-env* is composed of 99 environmental variables, we assessed the dimension reduction algorithm VIF to verify if a reduced environment space could be sufficient to classify the species well.

The output of our experiments consists of 32 distribution maps for each considered species, classifier, and environment size. We evaluated experiments according to the performance metrics AUC and TSS and qualitative aspects of the generated maps. The results of this project are accessible in https://github.com/AI-Unicamp/easy-sdm

2.1 *br-env*

A crucial part of this work consisted of preparing the data before applying machine learning models. The data engineering phase started by downloading the raw environmental databases and species occurrence to create a 3D array with 99 environment variables composed from Bioclim, Envirem, and Soilgrids databases for the Brazilian territory coordinate limits. This step is species independent and resulted in br-env, a database of aggregated environmental data for Brazil that is used in this study and can be used for further studies on any specie. From a practical point of view, the br-env consists of a Numpy array with metadata information to reference each column to an environment variable.

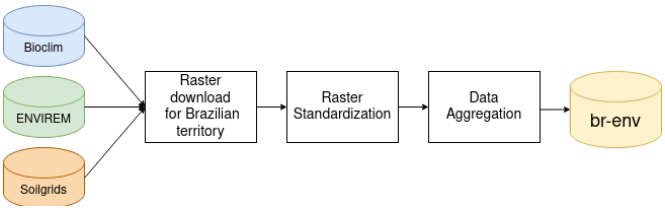

Figure 1. The figure represents the br-env creation process. First we download data from Bioclim, Envirem and Soilgrids. Next we standardize the data from each source to common specifications. Finally we describe the aggregation process to a unique array, which we call br-env.

2.2 Pseudo-absences generation

We built dataset rows corresponding to presences according to the species GBIF occurrence (OCC) coordinates. However, absent species data are not available. For this reason, we generated pseudo-absence (PSA) coordinates representing regions where the plant species could not grow.

Among the techniques mapped from the literature in we applied Random Selection with Environmental Profiling (RSEP) considering the best trade-off between simplicity and performance. To extract environmental profiling, we applied OCSVM, once we can use a semi-supervised approach, in which the algorithm is trained only with presence data and the resulting model, at inference time, labels coordinates as presence or absence [14].

We randomly generated the same number of pseudo-absences as the number of available occurrences to keep a balanced dataset and to attenuate the effects of prevalence in the classification models performance [15].

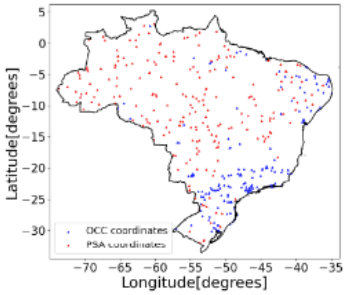

Figure 2. Example result of pseudo species generation for Zea mays.

2.3 The Variance Inflation Factor (VIF)

In the dataset construction, 99 features were present. Many of the environment variables have colinear dependencies with each other. Therefore, to evaluate behave of a reduced number of components and keep the performance metrics, we applied the VIF algorithm in parallel for the whole data before the cross-validation.

VIF is calculated by the $VIF = 1/(1 - R2)$. This formula measures if one variable can be predicted based on the features. If R2 is small, a specific variable can not be predicted from the information from other variables,

showing a small colinearity with them, offering evidence that an algorithm could use this variable as an information source to determine the target label.

We decided to keep a group of features with the constraint VIF<10 for each part compared to the others as explained in [16] when VIF>10, two variables are highly correlated, and can one of them can be discarded. Hence, we developed an iterative process that calculates the VIF factor for each feature and removed the biggest one until all of them respect the constrain.

2.4 The classifiers

In this section, we describe the proposed classifiers. We focused on comparing the performance of popular ensemble techniques (GB, XGB and RF) with a MLP. The tree ensemble algorithms are built with multiple decision trees, but they differ on how to integrate them. Decision trees learn data through an architecture represented by branch nodes and leaf nodes, the former contains a condition to split the data, and the latter helps to decide a class to a new data point [17].

GB and XGB represent the boosting ensemble technique that combines many weak decision tree classifiers to build a robust classifier. The observation weights are adjusted based on the previous classification, which replaces the approach of creating only one predictive solid model. Like ANN, boosting techniques are nonparametric machine learning, so the models can be adjusted according to the observed data, being adequate for the domain-specific tasks, as SDMs. The main difference between GB and XGB is that the last one used advanced regularization techniques like L1 and L2, which are expected to improve model generalization capabilities [18].

RF is also an ensemble technique that produces each decision tree independently and combines the results at the end of the process by the majority votes of each tree. Algorithm that follow this procedure are classified as a bagging algorithms. The literature shows that this approach has good performances in problems which the number of variables are much larger than the number of observations, which is also true for SDMs [19].

MLP is a conventional neural network approach in which neurons with nonlinear activation functions are structured as a network with multiple layers. The first one has the number of neurons equal to the number of features. The last one has the number of neurons dependent on the loss function and the problem objective. ANN learns through the back-propagation process in which the neuron's weights are updated with the error between a predicted target data and the labeled one [20], [21]. For our SDM approach, the number of neurons in the first layer is the number of environment variables. As we modeled it as a binary classification problem, we used a sigmoid function in the last neuron that calculates the prediction errors through a log-loss function [16].

3. Results

The experiments results we present in this chapter correspond to distribution models that are obtained varying three different aspects of the modeling process:

- **species modeled:** Zea mays, Cucurbita pepo, Cajanus cajan or Capsicum annuum;
- **training data:** with complete set of environmental variables available or VIF-reduced number of variables;
- **binary classifier:** Random Forest, Gradient Boosting, XGBoost or Multilayer Perceptron).

In summary, we conducted 32 experiments (4 species × 2 dataset configurations × 4 binary classifiers). For each experiment, we evaluated the measured performance of the classifier (AUC and TSS metrics), and the resulting distribution map characteristics.

In Figure 3 we show an example distribution map. In Tables 1-4 we expose the results of the proposed experiments for each of the studies species.

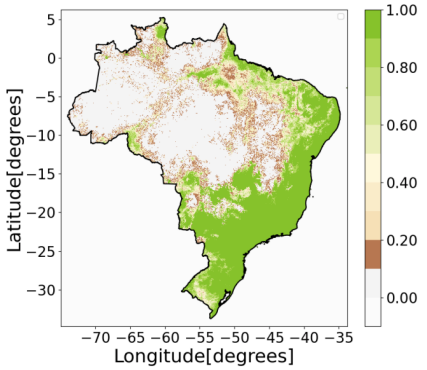

Figure 3. Example distribution for Zea mays using XGB as classifier.

Experiment Setup		KFold Metrics			
		AUC		TSS	
Estimator	Is VIF applied?	Mean	Std	Mean	Std
XGB	Yes	0,932	0,007	0,690	0,037
XGB	No	0,942	0,010	0,726	0,051
RF	Yes	0,943	0,014	0,735	0,050
RF	No	**0,949**	0,004	**0,749**	0,037
GB	Yes	0,924	0,018	0,672	0,038
GB	No	0,938	0,005	0,717	0,064
MLP	Yes	0,918	0,013	0,650	0,019
MLP	No	0,927	0,021	0,700	0,076

Table1. Zea mays experiments results

Experiment Setup		KFold Metrics			
		AUC		TSS	
Estimator	Is VIF applied?	Mean	Std	Mean	Std
XGB	Yes	0,835	0,077	0,565	0,118
XGB	No	0,908	0,041	**0,614**	0,1290
RF	Yes	0,881	0,060	0,565	0,057
RF	No	**0,929**	0,032	0,551	0,125
GB	Yes	0,860	0,121	0,488	0,176
GB	No	0,905	0,054	**0,614**	0,129
MLP	Yes	0,828	0,058	0,483	0,085
MLP	No	0,873	0,049	0,597	0,085

Table 2. Cucurbita pepo experiments results

Experiment Setup		KFold Metrics			
		AUC		TSS	
Estimator	Is VIF applied?	Mean	Std	Mean	Std
XGB	Yes	0,906	0,047	0,659	0,087
XGB	No	0,932	0,035	**0,736**	0,131
RF	Yes	0,914	0,039	**0,736**	0,139
RF	No	**0,946**	0,035	**0,736**	0,093
GB	Yes	0,919	0,041	**0,736**	0,108
GB	No	0,927	0,042	0,722	0,098
MLP	Yes	0,905	0,046	0,651	0,110
MLP	No	0,902	0,029	0,659	0,060

Table 3. Cajanus cajan experiments results

Experiment Setup		KFold Metrics			
		AUC		TSS	
Estimator	Is VIF applied?	Mean	Std	Mean	Std
XGB	Yes	0,967	0,004	0,8069	0,016
XGB	No	**0,980**	0,002	**0,853**	0,004
RF	Yes	0,965	0,004	0,801	0,015
RF	No	0,974	0,003	0,824	0,011
GB	Yes	0,967	0,004	0,809	0,017
GB	No	0,977	0,003	0,840	0,018
MLP	Yes	0,957	0,002	0,763	0,014
MLP	No	0,969	0,005	0,789	0,015

Table 4. Capsicum annuum experiments results

4. Conclusions

Despite being the Brazilian economy flagship, traditional agriculture has several problems regarding environmental degradation and efficiency in production by area. Understanding the regions in Brazil where producers can grow agroecological consortiums is a starting point to increase the adoption of sustainable crops in the place of monoculture and reduces it negative impact on natural resources. For this reason, this work aimed to explore how Species Distribution Modeling (SDM) can suggest regions for the MILPA consortium to grow in Brazil. According to ecological concepts, this consortium considers species from big groups (corn, squash, bean, and pepper) to create positive interactions between them and the environment to amplify food production without requiring pesticides and fertilizers.

We presented br-env, a standardized three-dimensional array that uses raster image data from Bioclim, Envirem and Soilgrids to represent the Brazilian environment conditions. Br env could be used in further research to create SDM independent of species and modeling choices. Besides, we detail how we used a pipeline of a OCSVM plus an ensemble classifier to generate distribution maps based on species occurrences and pseudo-absences. According to several experiments, we compare the proposed ensemble classifiers (XGB, GB, RF, and MLP) applied to the MILPA species we selected (Zea mays, Cucurbita pepo, Cajanus cajan and Capsicum annuum) to evaluate the classification algorithms according to the AUC and TSS metrics and the VIF algorithm.

Based on the performed experiments, we concluded this work by answering the research questions. The algorithms had similar performances according to the performance metrics for each species. They show that using and OCSVM to generate pseudo absences had a more significant impact independent of the classification algorithm in terms of metrics. Still, those classifiers were relevant in the style of distribution, with RF being granular in its decisions and XGB, GB and MLP being more decisive. AUC and TSS metrics should be analyzed as a pair once they evaluate the predicted habitats according to different perspectives. Also, we can notice that apply VIF variables. Besides, we identify species with fewer occurrences as Cucurbita pepo, which tend to perform worse in the metrics, and species with concentrated occurrences as Capsicum annuum tend to perform well. Finally, we can conclude that species to compose the MILPA consortium could grow together in a coastal part of the Northeast region of Brazil. We could join part of the Southeast region if we remove the optional species Capsicum annuum.

Finally, we highlight the importance of considering statistical, computational, and ecological knowledge to evaluate the generated habitat distributions. Understanding specific properties of a target species, is substantial to make a crop succeed when using the distribution's manual.

Acknowledgements

I would like to thank Paula Dornhofer Paro Costa for guiding me on the journey of being a researcher and for enabling me to turn the initial ideas I had from this project into solid work. I would also like to thank my parents, Maria Angela Alves and Antonio Carlos Sasso, for their financial, loving support and for making me believe in education since I was a kid. Finally, I would like to thank all my friends who have substantially influenced the person and professional I am.

References

[1] M. Duru, O. Therond, and M. Fares, 'Designing agroecological transitions; A review', *Agron. Sustain. Dev.*, vol. 35, no. 4, pp. 1237–1257, Oct. 2015, doi: 10.1007/s13593-015-0318-x.

[2] S. R. Gliessman, *Agroecology: The Ecology of Sustainable Food Systems, Third Edition*, 0 ed. CRC Press, 2014. doi: 10.1201/b17881.

[3] L. A. Martinelli, R. Naylor, P. M. Vitousek, and P. Moutinho, 'Agriculture in Brazil: impacts, costs, and opportunities for a sustainable future', *Curr. Opin. Environ. Sustain.*, vol. 2, no. 5–6, pp. 431–438, Dec. 2010, doi: 10.1016/j.cosust.2010.09.008.

[4] E. Barrios *et al.*, 'The 10 Elements of Agroecology: enabling transitions towards sustainable agriculture and food systems through visual narratives', *Ecosyst. People*, vol. 16, no. 1, pp. 230–247, Jan. 2020, doi: 10.1080/26395916.2020.1808705.

[5] T. Juniper, *What has nature ever done for us? how money really does grow on trees*. Santa Fe, NM: Synergetic Press, 2013.

[6] Y. Yu, T.-J. Stomph, D. Makowski, and W. van der Werf, 'Temporal niche differentiation increases the land equivalent ratio of annual intercrops: A meta-analysis', *Field Crops Res.*, vol. 184, pp. 133–144, Dec. 2015, doi: 10.1016/j.fcr.2015.09.010.

[7] C. Kremen, A. Iles, and C. Bacon, 'Diversified farming systems: an agroecological, systems-based alternative to modern industrial agriculture', *Ecol. Soc.*, vol. 17, no. 4, 2012.

[8] E. A. Frison and I. P. of E. on S. F. Systems, 'From uniformity to diversity: a paradigm shift from industrial agriculture to diversified agroecological systems', IPES, Report, 2016. Accessed: Oct. 30, 2021. [Online]. Available: https://cgspace.cgiar.org/handle/10568/75659

[9] T. Václavík and R. K. Meentemeyer, 'Equilibrium or not? Modelling potential distribution of invasive species in different stages of invasion: Equilibrium and invasive species distribution models', *Divers. Distrib.*, vol. 18, no. 1, pp. 73–83, Jan. 2012, doi: 10.1111/j.1472-4642.2011.00854.x.

[10] J. L. Brown, 'SDMtoolbox: a python-based GIS toolkit for landscape genetic, biogeographic and species distribution model analyses', *Methods Ecol. Evol.*, vol. 5, no. 7, pp. 694–700, Jul. 2014, doi: 10.1111/2041-210X.12200.

[11] J. Elith and J. R. Leathwick, 'Species Distribution Models: Ecological Explanation and Prediction Across Space and Time', *Annu. Rev. Ecol. Evol. Syst.*, vol. 40, no. 1, pp. 677–697, Dec. 2009, doi: 10.1146/annurev.ecolsys.110308.120159.

[12] S. Teran and R. Heilskov, 'Las plantas de la milpa entre los mayas: etnobotanica de las plantas cultivadas por campesinos mayas en las milpas del noreste de Yucatan', 1998.

[13] FAO, 'FAO Strategic framework 2022-31'. FAO, 2021. Accessed: Jun. 30, 2022. [Online]. Available: http://www.fao.org/3/ne577en/ne577en.pdf

[14] M. Goldstein and S. Uchida, 'A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data', *PLOS ONE*, vol. 11, no. 4, p. e0152173, Apr. 2016, doi: 10.1371/journal.pone.0152173.

[15] M. Barbet-Massin and W. Jetz, 'A 40-year, continent-wide, multispecies assessment of relevant climate predictors for species distribution modelling', *Divers. Distrib.*, vol. 20, no. 11, pp. 1285–1295, Nov. 2014, doi: 10.1111/ddi.12229.

[16] J. Rew, Y. Cho, and E. Hwang, 'A Robust Prediction Model for Species Distribution Using Bagging Ensembles with Deep Neural Networks', *Remote Sens.*, vol. 13, no. 8, p. 1495, Apr. 2021, doi: 10.3390/rs13081495.

[17] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, 'An introduction to decision tree modeling', *J. Chemom.*, vol. 18, no. 6, pp. 275–285, Jun. 2004, doi: 10.1002/cem.873.

[18] T. Chen and C. Guestrin, 'XGBoost: A Scalable Tree Boosting System', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[19] G. Biau and E. Scornet, 'A random forest guided tour', *TEST*, vol. 25, no. 2, pp. 197–227, Jun. 2016, doi: 10.1007/s11749-016-0481-7.

[20] A. K. Jain, Jianchang Mao, and K. M. Mohiuddin, 'Artificial neural networks: a tutorial', *Computer*, vol. 29, no. 3, pp. 31–44, Mar. 1996, doi: 10.1109/2.485891.

[21] L. Noriega, 'Multilayer perceptron tutorial', *Sch. Comput. Staffs. Univ.*, 2005.

FEEC | UNICAMP
EADCA22
XIV ENCONTRO DE ALUNOS E DOCENTES DO DCA

Sessão Técnica 2

Single Neural Network Sensor Fusion for Automotive Application

Marcelo Eduardo Pederiva , José Mario De Martino and Alessandro Zimmer
{m122580@dac.unicamp.br, martino@unicamp.br, Alessandro.Zimmer@thi.de}

Departamento de Engenharia de Computação e Automação (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Campinas, SP, Brasil

Abstract – Autonomous vehicles are becoming a reality in ground transportation. Computational advancement has enabled powerful methods to sense, map, locate, and process large amounts of data required to drive on urban streets safely. The fusion of multiple sensors allows for building accurate world models to improve autonomous vehicles' navigation and behavior. Among the current techniques, the fusion of LIDAR, RADAR, and Camera data have shown significant improvement in perception tasks. Current methods use parallel networks to explore each sensor separately. Despite its significant accuracy, its response time and high demand for computational resources are still limitations for a real-world self-driving application. Fusing these sensors using a single network is still an open question and a promising candidate to avoid these problems. The paper presents a Ph.D. project under development. It presents a preliminary approach for early sensor fusion and its performance with one sensor to detect 3D objects.

Keywords – Object Detection, Sensor Fusion, Machine Learning, Multi-task Learning, Autonomous Vehicles

1. Introduction

Autonomous vehicles have been the target of great interest in universities, research centers, and industry. With the advance of computer technology and computational techniques, autonomous cars' implementation became increasingly viable. However, implementing autonomous vehicles on urban streets requires a thorough perception of the environment, including detecting objects and their movements. These tasks require exteroceptive sensors to measure the car's surroundings. Sensors in this category include Cameras, Radio Detection and Ranging sensors (RADAR), and Light Detection and Ranging sensors (LIDAR).

Currently, LIDARs are widely used to detect objects around the vehicle. LIDAR data can be used to precisely estimate an object's geometry [11]. However, this type of sensor presents significant limitations in estimating motion. In contrast, RADARs allow robust motion estimation, providing accurate object velocity and direction measurements. Additionally, RADARs support reliable detection despite adverse weather conditions [12]. LIDARs and RADARs have significant limitations concerning recognizing objects despite their reliable performance in the situations mentioned. Object recognition using Cameras is an improving research area [1]. The state-of-art object recognition approaches using cameras deliver accurate real-time recognition (under 0.04 seconds) of several objects simultaneously observed in an image [2].

In general, the environment perception has been performed using a combination of two sensors: LIDAR-Camera, LIDAR-RADAR, or RADAR-Camera. Methods based on LIDAR-Camera have shown convincing results concerning visual detection, distance, and geometry estimation of objects [3]. However, the methods are not adequate to estimate objects' velocities [13]. Velocity estimation is better tackled by LIDAR-RADAR fusion methods [9]. However, the absence of cameras in LIDAR-RADAR approaches precludes objects' visual identification, impacting autonomous decision-making. Finally, the RADAR-Camera detection shows gains in performance for detecting objects in low light and rainy/cloudy weather [4]. Although RADARs are reliable all-weather sensors, they can not provide a dense environment sampling as LIDARs. Consequently, the combination RADAR-Camera does not support a high-quality geometry estimation of the detected objects.

At the end of 2020, the first LIDAR-RADAR-Camera fusion Deep Learning-based for 3D object detection in a real-world scene was proposed. Based on Frustum PointNet (F-PointNet) [8], the method uses the Fast R-CNN object detector to estimate a Region of Interest (RoI) of the camera view. This output is combined with the LIDAR measurements to estimate and classify the 3D object model. Simultaneously, the RADAR sensor provides the detections' velocity estimation with a different neural network. As a result, the model achieved high accuracy and small velocity error compared to the current methods that used only a two sensors fusion [10].

The mentioned approaches and main 3D object detection models use a Late Fusion of the sensors. Based

on parallel networks, each sensor passes through a different neural network, and in the end or in the middle, the results are combined. This method provides a precise result. However, it requires a high demand for computing resources, including footprint and energy consumption. On the other hand, the Early Fusion fuses the sensors' information before the network, implementing a single architecture for the prediction. This process has low computation requirements and a low memory budget. Nevertheless, the learning process of mixed features of multiple different sensors increases the prediction challenge.

Although Early Fusion approaches for 3D object detection are still challenging, their low computational demand is of great relevance for application in an autonomous vehicle with limited resources. Therefore, our project aims to reduce computational demand with a new competitive fusion approach based on a single Neural Network. This network will combine and analyze the information acquired by LIDAR, RADAR, and Camera sensors to detect vehicles on the street.

The remaining paper is organized as follows. We first review the related work in section 2. Then in section 3, we present some experiments based on a known dataset. Next, we show some preliminary results and analysis in section 4. Finally, in section 5, we conclude the paper and present the remaining approaches in the project.

2. Related work

The main challenge for implementing an Early Fusion is effectively merging data from different types of sensors. For this, it is necessary to represent all the data in a single reference.

Image-based Object Detection research has been fast developed in recent years. Furthermore, current methods present a high accuracy and fast response in detecting multiple objects presented in the same image. In this way, this project aims to use the Camera sensor as a reference for other sensors and build our model based on the state-of-art 2D Object Detection models.

In the 2D Object Detection field, among the state-of-art models, the "You Only Look Once", namely YOLO, has been standing out in 2D object detection models in the last years. Its ability to recognize objects' bounding boxes and classify them quickly makes it an excellent candidate for many tasks. Consequently, variations of YOLO have surged in different fields, such as Object Detection, Object Tracking, Image Segmentation, and Landmark Detection. Due to this versatility and its remarkable performance in different fields, the YOLO approach is a promising initial candidate to use as inspiration to develop our 3D detection model.

Our approach converts the labels of the cars' position into a three-dimensional grid, a spatial representation of the environment. The grid is divided into a Sx×Sy×Sz grid, where Sx, Sy, and Sz represents the division in each X, Y, and Z axis, respectively. As a result, our model uses $Sx = 13$, $Sy = 5$, and $Sz = 13$. The Figure 1 shows the grid representation, where each grid cell stores an 9 length array prediction: $class, P, x, y, z, w, h, l, \theta$. The first value represents the object class, followed by a confidence of the grid cell estimation, then the six values representing the 3D bounding box (center of the object and its dimensions (width, height, length)), and the last, the rotation angle in Y axis of the object.

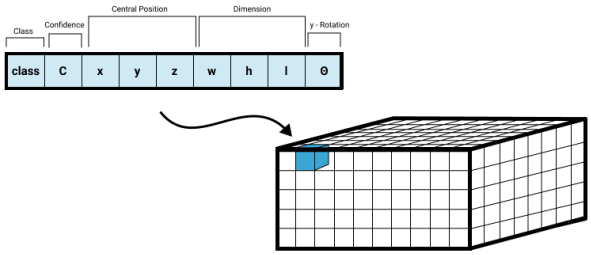

Figure 1. Three-dimensional grid of features.

Based on Supervised Learning, the network was trained to return a similar three-dimensional grid as provided by the label. In other words, the network output a tensor of predictions accommodated in a $13 \times 5 \times 13 \times 9$ tensor.

For the input of our network, the point cloud sensors (LIDAR and RADAR) are converted to the camera reference and are concatenated into an input tensor $width \times height \times 8$, where 8 represents the channels of each sensor: R, G, B, X, Y, Z, Vx, Vz. Furthermore, each channel is normalized individually.

2.1. Network

The proposed network is based on a Multi-Task Learning approach with hard parameter sharing. In other words, the architecture presents a sequence of convolutional layers and in the end, the last layer is branched into five output tasks.

The network combines two types of convolutional layers, 2D and 3D. First, it is used a known backbone as a 2D feature extractor. We choose the ResNet50, aiming for a combination of good performance in a fast response. It receives the input tensor (W×H×channels) and outputs a 15×15×2048 shape. This output pass trough one convolutional layer with 625 filters $kernel\_size =$

1, $strides = 1$, and leakyReLU activation function. Next, to introduce one dimension to the architecture, it is passed by a reshape layer (25×17×25×17). This new shape passes by a sequence of 3D convolutional and residual layers, representing a ResNet approach with 3D convolutions. Finally, the output is branched by five 3D convolutional layers. Each branch will predict each characteristic (class, confidence, position, dimension, and rotation angle) separately and be concatenated into a Sx×Sy×Sz×9 tensor of predictions.

2.2. Loss

$$loss = loss_{conf} + loss_{class} + loss_{box} \quad (1)$$

To calculate the error, we consider the L2 loss of all prediction features. The loss equation (Equation 1) is divided by 3 losses: Confidence loss ($loss_{conf}$), Class loss ($loss_{class}$), Box loss ($loss_{box}$).

The Confidence loss is represented by the reliance on each model prediction. It is represented by correct cell prediction ($loss_{obj}$) and incorrect cell prediction ($loss_{noobj}$). The Classification loss is represented by the error in the prediction of the object's classification ($loss_{class}$). Finally, the Box loss is defined by the errors of the Intersection over Union score ($loss_{IoU}$) and the y-axis rotational estimation ($loss_{\theta}$). Furthermore, each error has a weight parameter (λ) to balance the relevance of each feature in the total loss value. Currently, the weight values is defined as follows: $\lambda_{obj} = 20$; $\lambda_{noobj} = 1$; $\lambda_{class} = 1$; $\lambda_{IoU} = 10$; $\lambda_{\theta} = 10$.

3. Experiments

Our project starts exploring its performance using only the Camera as input. With a good candidate for 3D object detections, the model will be tested with LIDAR data. Finally, the RADAR sensor will be linked to the tensor to predict objects' velocity at a later phase.

Our model performance has been tested in the KITTI dataset. However, as the KITTI test dataset is closed and has a limited number of submissions, for experimental studies, we divided the open content (7581 images) into train and test sets for our model. The training step was done with 80% to train and 10% for validation. The last 10% was used for the testing step. The model was trained in a GPU: RTX2080ti and CPU: i7 9700KF with 200 epochs and used the Adam optimizer with a learning rate = 0.003.

4. Preliminary Results

In the Object Detection field, the Intersection over Union (IoU) score is a particular evaluation metric used to represent the matching percent between the predicted bounding box and the ground truth. Then, to evaluate the mean Average Precision (mAP), we considered a correct detection if the prediction presents an $IoU > 0.7$ and an object recognized if it presents an $IoU > 0.1$.

Table 1. Model's Performance.

	mAP	Mean IoU	Max IoU	Average Recognition
3D	25.34 %	0.3636	0.9119	79.81 %
2D	25.38 %	0.3667	0.9537	79.02 %

As a result, our proposed model using only the camera sensor achieve great precision and recognized most of the vehicles presented in the scene. In Table 1, the results considering the 3D and 2D (Bird Eye View) IoU scores are shown. The *mAP*, *mean IoU*, and *Max IoU* represent the mean Average Precision, the mean Intersection over Union, and the maximum Intersection over Union, respectively.

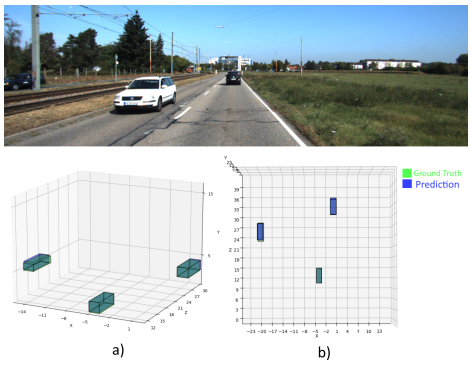

Figure 2. Example of model detection. The top image is the camera input. On a), we have an overview and on b), a bird-eye-view perspective.

The Figure 2 shows the model performance detecting 3 vehicles correctly. There are two vehicles fully visible and one partially occluded in this view. The ground truth is represented by the green box and the prediction with a blue box.

Table 2. Monocular Detection on KITTI dataset. Methods with * trained/tested with a small dataset.

Methods	mAP_{3D}	mAP_{2D}
Ground-Aware [6]	14.94%	20.29%
MonoEF [15]	15.62%	22.00%
MonoFlex [14]	15.30%	21.62%
GUPNet [7]	16.80%	23.23%
MonoCon [5]	17.64%	24.07%
Ours*	**25.34%**	**25.38%**

The state-of-art models presented in Table 2 were trained using all training set (open content) and were

tested in the closed test set. The KITTI test set presents similar images, in the same environments, from the open content. Although our model was trained and tested in a small dataset, its results showed a good candidate for 3D object detection.

5. Conclusions and Remaining Work

This study presents a beginning approach to fuse sensors' data. It shows the use of a single network to predict the 3D bounding box of objects for autonomous vehicles. The initial results, using a single sensor, showed itself a good candidate for 3D object detection. Using a small dataset to train and test, the model showed a competitive result against state-of-art models.

For the remaining work, LIDAR data will be implemented in the input of the network aiming to enhance the prediction precision. Next, the RADAR data will be implemented to estimate the velocity of the objects.

The detection model proposed by this research has applications that are not limited to autonomous cars. With technical improvement, the model can be used in different areas that need an autonomous perception of the environment.

Acknowledgment

The authors acknowledge the funding received from The National Council for Scientific and Technological Development (CNPq 141061/2021-9).

References

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal Speed and Accuracy of Object Detection. http://arxiv.org/abs/2004.10934, 2020.

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection, 2020.

[3] Luca Caltagirone, Mauro Bellone, Lennart Svensson, and Mattias Wahde. LIDAR–camera fusion for road detection using fully convolutional neural networks. *Robotics and Autonomous Systems*, 111:125–131, 2019.

[4] Kamil Kowol, Matthias Rottmann, Stefan Bracke, and Hanno Gottschalk. YOdar: Uncertainty-based Sensor Fusion for Vehicle Detection with Camera and Radar Sensors. http://arxiv.org/abs/2010.03320, 2020.

[5] Xianpeng Liu, Nan Xue, and Tianfu Wu. Learning auxiliary monocular contexts helps monocular 3d object detection, 2021.

[6] Yuxuan Liu, Yuan Yixuan, and Ming Liu. Ground-aware monocular 3d object detection for autonomous driving. *IEEE Robotics and Automation Letters*, 6(2):919–926, 2021.

[7] Yan Lu, Xinzhu Ma, Lei Yang, Tianzhu Zhang, Yating Liu, Qi Chu, Junjie Yan, and Wanli Ouyang. Geometry uncertainty projection network for monocular 3d object detection, 2021.

[8] Charles R. Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J. Guibas. Frustum PointNets for 3D Object Detection from RGB-D Data. https://arxiv.org/abs/1711.08488, 2018.

[9] Meet Shah, Zhiling Huang, Ankit Laddha, Matthew Langford, Blake Barber, Sidney Zhang, Carlos Vallespi-Gonzalez, and Raquel Urtasun. LiRaNet: End-to-End Trajectory Prediction using Spatio-Temporal Radar Fusion. https://arxiv.org/abs/2010.00731, 2020.

[10] L. Wang, T. Chen, C. Anklam, and B. Goldluecke. High dimensional frustum pointnet for 3d object detection from camera, lidar, and radar. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1621–1628, 2020.

[11] Ruisheng Wang. 3d building modeling using images and lidar: a review. *International Journal of Image and Data Fusion*, 4(4):273–292, 2013.

[12] Bin Yang, Runsheng Guo, Ming Liang, Sergio Casas, and Raquel Urtasun. RadarNet: Exploiting Radar for Robust Perception of Dynamic Objects. https://arxiv.org/abs/2007.14366, 2020.

[13] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3d object detection and tracking. https://arxiv.org/abs/2006.11275, 2021.

[14] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3289–3298, June 2021.

[15] Yunsong Zhou, Yuan He, Hongzi Zhu, Cheng Wang, Hongyang Li, and Qinhong Jiang. Monocular 3d object detection: An extrinsic parameter free approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7556–7566, June 2021.

Deep learning for Tractography

Gilbert Pla Martínez, Wu Shin-TIng

{g234612@dac.unicamp.br, ting@dca.fee.unicamp.br}

Departamento de Engenharia de Computação e Automação (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Campinas, SP, Brasil

Abstract – Tractography is one of the most valuable tools for neurosurgeons in preoperative planning since it allows the visualization of both white matter structures and the fibers' distribution in a patient's brain. The best-known classical techniques are either deterministic or probabilistic in providing fiber orientations in voxel resolution. With advances in machine learning, particularly deep learning, a series of elements and new ways of obtaining the fiber structures have been introduced, making tractography a promising, handy, and reliable tool for doctors in their daily diagnostic process. A typical consistent tractography pipeline comprises two stages: diffusion directionality modeling and diffusion-oriented tracking. This paper reviews deep learning-based estimations of local fiber orientations representable by either diffusion tensors or diffusion orientation distribution functions.

Keywords – Tractography, diffusion tensor imaging, fiber orientation distribution functions

1. Introduction

Neural fibers are responsible for highly complex brain connections. Diffusion Magnetic Resonance Images have been used in clinical and research applications to infer white matter structures and brain fiber interconnections, far beyond the resolutions of current images [1] [2]. The process is called tractography. Magnetic resonance imaging has also seen the development of artificial intelligence algorithms. More recently, deep learning-based techniques have been perfecting the levels of reliable neural pathways for making diagnoses and providing a new perspective and interpretation of the information and the results obtained [3]. Algorithms based on deep learning are data driven, so the selection of datasets, input data, pre-processing stages and the associated labels for training imply the elaboration of a training paradigm that guarantees the robustness of the model developed. Despite the progress made so far, there is still a long way to walk to predict tracking directions and stopping within diffusion magnetic resonance imaging because of the complexity of neural pathways [12].

2. Proposal

As deep learning techniques depend on the input data and the manipulation and pre-processing of these data, we proposed a bibliographic survey of the latest deep learning based estimations of fiber orientations algorithms, emphasizing the training datasets capable of deriving the mapping rules from the raw diffusion data to the neural pathways and the applied deep learning algorithms for conducting such derivations.

Diffusion-Weighted magnetic resonance Imaging (DWI) and High Angular Resolution Diffusion Imaging (HARDI) are the most known techniques for sampling the raw diffusion signals of water molecules within a voxel of brain tissue. Most applied ways to synthesize from these raw datasets diffusion directionality in the brain are Diffusion Tensor Imaging (DTI) and Orientation Distribution Function (ODF). From the estimated local diffusion directionality and start seed points, plausible white matter streamlines are expected to be tracked. Although we are looking for a learning algorithm that can derive mapping rules between DWI scans and neural fibers directly from DWI scans and reference streamlines, we chose to first deal with the problem in two stages like Benou and Raviv [11].

There are four major groups of deep learning algorithms: (1) supervised (Convolutional Neural Networks, Long Short Term Memory Networks, Recurrent Neural Networks), (2) semi-supervised (Generative Adversarial Networks), (3) unsupervised (Autoencoders-Autodecoders, Support Vector Machine), and (4) reinforcement (Deep Reinforcement Learning).

2.1 Article Review

In this section, we briefly describe the articles analyzed so far, highlighting the training data used and the deep learning algorithms applied for estimating the diffusion orientations in each brain voxel.

DeepDTI [4] is an algorithm that minimizes the data requirements for its operation. It uses as input an image with b=0 (non-diffusion weighted scan) and 6 DWIs together with the synthesized DTI and T1 and T2-weighted volumes from the WU-Minn Human

Connectome Project (HCP) databases. Data from 70 unrelated subjects were used, with 40 subjects for training, 10 for validation, and 20 for evaluation or testing. The learning algorithm is a deep convolutional network or CNN with 10 layers of 3 dimensions each. The output of this CNN also corresponds to a high-quality volume with b = 0 and 6 volumes of DWIs optimized by the diffusion encoded directions. The authors quantify the performance of the learning algorithm using the quality of the output images, DTI metrics, DTI-based tractography of the reconstruction, and analysis of specific tracts.

In the same direction, SuperDTI [5] is a method also based on CNN, aiming to take advantage of the elements of DTI-based methods. The data are from the databases of the international HCP, consisting of DWIs of 50 subjects divided into 40 for training and validation and 10 for tests. Although the authors also use non-diffusion-weighted and 6 DWI volume, SuperDTI differs from DeepDTI in training CNN parameters. They are trained separately by the FA and MD maps and the eigenvectors, pairwisely. This method eliminates the noise-sensitive tensor fitting process and has quantification errors close to 5% in the regions of interest that contain target white matter and fibers.

Karimi and Gholipour [6] also estimated a diffusion tensor image using six diffusion-weighted scans. They further proposed exploiting the relationships between diffusion signals and tensors in neighboring voxels to improve the tensor estimation accuracy. They applied two-stage transformer neural networks as a learning algorithm. The first estimates the diffusion tensors according to the diffusion signals in a neighborhood of voxels The second refines the estimation of the tensors by learning the relationships between the diffusion signals and the tensors estimated by the first network. They evaluated the proposed method with HCP, scans from the Pediatric Imaging, Neurocognition, and Genetics (PING) dataset, and Vein of Galen Malformation (VOGM) scans.

The diffusion orientation distribution functions head another way to determine the fiber spatial orientations. As they require higher angular resolution diffusion signals, it has driven work related to training algorithms for increasing angular resolution while keeping acquisition time low. Jha, R. R. et al. proposed a machine based on generative adversarial networks (GAN) [7] to obtain more gradient directions for under-sampled DWI volumes with reduced number of directions in q-space.

In [8], Jha, R. R. et al. designed a GAN-based model for synthesizing multi-shell multi-tissue fiber orientation distribution function from the spherical harmonic coefficients of a single-shell HARDI volume at a b-value of 1000. The HARDI signals are transformed into the spherical harmonic coefficients to train the neural network. The performance of the learning capability was evaluated with the HARDI multi-shell dataset from the HCP: 100 randomly selected subjects having volumes acquired with different gradients: b = 0, b = 1000, b = 2000, b = 3000.

On the one hand, the fiber orientation distribution function is better estimated from high angular resolution diffusion imaging. On the other hand, signal acquisition is much more time-consuming. Rui Zeng et al. [9] devised a 3D convolutional neural network to enhance the angular resolution of low-quality single-shell low angular resolution diffusion image (LARDI) data, making them equivalent to those derived from high-quality multi-shell HARDI acquisition. The machine also learns to remove false fibers and recover some fibers present in the original volume, thus allowing a more reliable tract reconstruction in practical clinical situations. The authors randomly selected 110 subjects from HCP, 50 for training, 50 for validation, and 10 for testing.

Lyon, M et al. [10] also investigated a way to overcome a long time in scanning diffusion signals if high angular resolution. They proposed a recurrent CNN autoencoder architecture to infer higher angular resolution diffusion signals without spherical harmonic coefficients. A 3D convolutional long short term memory (ConvLSTM) cell is applied to model the relationship of q-space cells. The authors used HCP datasets for training and evaluating the performance of the proposed model by measuring the deviation of estimated diffusion signals from the ground truth across multiple diffusion directions.

2.2 Training Datasets and Deep Learning Algorithms for Diffusion Orientation Estimation

Table 1 summarizes the learning algorithms used in reviewed articles and the type of data used for training. Supervised learning is the most used, followed by the semi-supervised and one unsupervised learning.

None of the proposals use the raw data directly from the diffusion signals (see Table 1). Instead, the authors preprocessed the data to make them fittable to the learning architecture. Among the papers studied, the combination of DTI model and deep convolutional neural networks seems the one that presents fewer pre-processing stages.

Method	Deep learning algorithm	Input Data	Output Data
DeepDTI	Supervised	T1-weighted T2-weighted + 1 DWI (b = 0) + 6 weighted DWIS	High quality (b = 0) DWI and 6 weighted DWIs
SuperDTI	Supervised	1 DWI (b = 0) + 6 weighted DWIS	Maps(DWIs – FA, DWIs – MD and DWIs – eigenvectors)
Karimi, D Gholipour, A	Semi-supervised	Volumetric 3D Image with 6 Channels (1 channel = 1 of 6 Normalized DWI)	Volumetric 3D Image with 6 Channels (1 channel = 1 of 6 tensor elements)
Jha, . R. R. et.al.	Semi-supervised	Input DWI slice I and Ground truth slice G	Full sample HARDI close to G
Jha, R. R et. al. (VRfRNet)	Supervised	Spherical Harmonics Coefficients (Applied Q-ball imaging on 1K HARDI data)	Predicted Multi-Shell Multi-Tissue fODF
Zeng, R et. al	Supervised	Single-Shell LARDI-FOD image	Super-resolved FOD image
Lyon, M et. al.	Unsupervised	3D dMRI patches + corresponding b-vectors	3D dMRI patches along the given diffusion direction

Table 1. Summary of learning algorithms and data type used.

3. Discussions

All the works analyzed implemented a methodology to obtain the fiber reconstruction with the best possible performance using deep learning and the facilities this technology provides. In general, the data used in each algorithm are from healthy subjects, and the results are compared with the known ground truth, lacking extensive tests in DWI volumes for patients with anatomical malformation. .

As machines learn with data from healthy people, we expect their performance with dMRI of non-healthy people to be inferior to that obtained with healthy people. It would be interesting to train these machines with data of non-healthy people and fine-tuning the training parameters to try to make them appropriate for clinical reality, gain insight into the problems, and look for novel solutions.

The datasets of the Human Connectome Project are used in all the works studied. In some cases, the authors specify the dataset used within the Human Connectome Project: in SuperDTI the HCP Young Adult dataset, in DeepDTI the Human Connectome Project (HCP) WU-Minn-Ox Consortium, Jha, R.R. et al. used the multi-shell HARDI from the WU-Minn Human Connectome Project (HCP) dataset.

The input data for each model varies depending on the type of application to be used, although in general the pre-processing stage largely determines the dimensions and the type of data to be used and introduced into the deep neural networks as inputs. The models based on diffusion tensors require fewer data pre-processing steps than the ones based on orientation distribution functions. Nevertheless, diffusion orientation distribution functions are much more information concerning neural pathways.

4. Conclusions

Deep learning techniques are promising for developing the estimation of local fiber orientation. However, the correct selection of the data to obtain the desired model with good performance is still challenging due to the difficulty involved in making a good selection. In addition, both the choice of the data and its pre-processing stages will affect the degree of complexity of the model developed, an element that affects the further practical implementation of the model obtained. This review improved our understanding of the potential challenges in elaborating a learning algorithm that maps the DWI scans in diffusion direction models (DTI or ODF), their training, validation, and testing. To achieve our goal, we must study alternatives for fiber tracking and analyze a better paradigm to map DWI in tracts. Directly from DWI scans to tractography or in two steps?

Acknowledgments

The authors thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for providing funding for this work.

References

[1] Poulin, P., Jörgens, D., Jodoin, P.-M. and Descoteaux, M., «Tractography and machine learning: Current state and open challenges,» *Magnetic Resonance Imaging,* vol. 64, pp. 37-48, 2019.

[2] Tax, C. M., Bastiani, M., Veraart, J., Garyfallidis, E. and Okan Irfanoglu, M., «What's new and what's next in diffusion MRI preprocessing,» *NeuroImage,* vol. 249, nº 118830, 2022.

[3] Chaudhari, A., Sandino, C. , Cole, E.,Larson, D., Gold, D., Vasanawala, S., Lungren, M., Hargreaves, B. and Langlotz, C., «Prospective Deployment of Deep Learning in MRI: A Framework for Important Considerations, Challenges, and Recommendations for Best Practices,» *J Magn Reson Imaging,* vol. 54, nº 2, pp. 357-371, Aug 2021.

[4] Tian, Q., Bilgic, B., Fan, Q., Liao, C., Ngamsombat, C., Hu, Y., Witzel, T., Setsompop,

K., Polimeni, J. R. and Huang, S. Y., «DeepDTI: High-fidelity six-direction diffusion tensor imaging using deep learning,» *NeuroImage,* vol. 219, nº 117017, 2020.

[5] Li, H., Liang, Z., Zhang, C., Liu, R., Li, J., Zhang, W., Liang, D., Shen, B., Zhang, X., Ge, Y., Zhang, J. and Ying, L., «SuperDTI: Ultrafast DTI and fiber tractography with deep learning,» *Magnetic Resonance in Medicine,* vol. 86, nº Issue 6, pp. 3334-3347, 2021.

[6] Karimi, D. and Gholipour, A., «Diffusion tensor estimation with transformer neural networks,» *Artificial Intelligence in Medicine,* vol. 130, nº 102330, 2022.

[7] Jha, . R. R. , Gupta, H. , Pathak, S. K., Schneider, W., Kumar, B. R. and Nigam, A., «Enhancing hardi reconstruction from undersampled data via multi-context and feature inter-dependency gan,» *IEEE 18th International Symposium on Biomedical Imaging (ISBI),* pp. 1103-1106, 2021.

[8] Jha, R. R., Pathak, S. K., Nath, V., Schneider, W., Rathish Kumar, B., Bhavsar, A. and Nigam, A. «VRfRNet: Volumetric ROI fODF reconstruction network for estimation of multi-tissue constrained spherical deconvolution with only single shell dMRI,» *Magnetic Resonance Imaging,* vol. 90, pp. 1-16, 2022.

[9] Zeng, R. , Lv, J., Wang, H., Zhou, L., Barnett, M., Calamante, F. and Wang, C., «FOD-Net: A deep learning method for fiber orientation distribution angular super resolution,» *Medical Image Analysis,* vol. 79, nº 102431, 2022.

[10] Lyon, M., Armitage, P. and Álvarez, M. A., «Angular Super-Resolution in Diffusion MRI with a 3D Recurrent Convolutional Autoencoder,» *arXiv preprint arXiv:2203.15598,* 2022.

[11] Benou, I. and Raviv, T. R., «DeepTract: A Probabilistic Deep Learning Framework for White Matter Fiber Tractography,» Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Lecture Notes in Computer Science. , Shenzhen, China, 2019, doi:10.1007/978-3-030-32248-9.

[12] Poulin, P., Jörgens, D., Jodoina, P.-M. and Descoteauxa, M., «Tractography and machine learning: Current state and open challenges,» Magnetic Resonance Imaging, vol. 64, pp. 37-48, December 2019.

Redução do consumo de memória no algoritmo de criptografia pós-quântica SABER

George Gigilas Junior, Marco Aurélio Amaral Henriques

{georgejuniorg@yahoo.com.br, marco@dca.fee.unicamp.br}

Departamento de Engenharia de Computação e Automação (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Campinas, SP, Brasil

Abstract – SABER é um algoritmo pós-quântico IND-CCA2 seguro de troca de chaves, finalista da terceira rodada do processo de padronização de algoritmos pós-quânticos do NIST. Visando tornar sua execução eficiente em ambientes computacionais restritos, este trabalho implementa, em um microcontrolador ARM Cortex-M0+, otimizações da literatura de outras arquiteturas para reduzir o consumo de memória de pilha desse algoritmo e de suas variações LightSABER e FireSABER. Considerando a versão de segurança média, foi possível reduzir o consumo de memória de pilha de suas funções entre 18% e 39%, comparado com resultados para ARM Cortex-M0 disponíveis na literatura. Ademais, descobriu-se que a flag de compilação -O1 produziu os melhores resultados para a versão de menor segurança, e que a flag -O2 foi a melhor para as demais versões.

Keywords – Criptografia pós-quântica, ARM Cortex M0+, SABER, reticulados.

1. Introdução

Com os avanços no desenvolvimento de computadores quânticos, um poder computacional ainda maior se aproxima. Apesar de promissor para diversas pesquisas, ele também se mostra uma ameaça para a segurança de diversas aplicações na internet. Já existem algoritmos quânticos capazes de quebrar os esquemas de criptografia assimétrica atuais, baseados em fatoração de números primos ou em logaritmos discretos.

Visando algoritmos que permaneçam seguros após o advento dessa tecnologia, o órgão americano National Institute of Standards and Technology (NIST) promoveu um processo de padronização de algoritmos pós-quânticos, iniciado em 2016 [9]. Ele é composto por algumas rodadas em que os melhores algoritmos são selecionados e outros descartados, tanto para algoritmos de troca de chaves quanto para algoritmos de assinatura digital. Em 2021, eram quatro os candidatos de troca de chaves na terceira rodada, sendo três deles baseados em problemas com reticulados, o que tornava muito provável que algum algoritmo baseado em reticulados fosse escolhido.

Porém, os problemas com reticulados são pesados computacionalmente. As chaves desses esquemas são bastante grandes, o que faz com que as operações fiquem mais caras. Diante disso, foram propostas várias otimizações na literatura para viabilizar os algoritmos em ambientes computacionais restritos, que podem apresentar dificuldades para executá-los pela quantidade limitada de memória disponível.

Uma vez que dispositivos IoT (Internet of Things), que possuem recursos limitados, estão cada vez mais presentes no dia-a-dia, é importante que eles também sejam protegidos. Com base nisso, decidimos buscar otimizar o consumo de memória do algoritmo SABER [4], que ainda era finalista quando a pesquisa foi iniciada. Também foram avaliados os requisitos de memória para as variações LightSABER e FireSABER. A primeira é uma versão mais leve e menos segura, enquanto a segunda é mais robusta e mais segura. Por fim, fizemos diversos testes para determinar a melhor flag de otimização para cada versão do algoritmo.

Vale notar que, apesar de não ter sido escolhido no processo do NIST, o SABER é bastante similar ao algoritmo escolhido (CRYSTALS-KYBER [2]), o que possivelmente permite um intercâmbio de técnicas de otimização entre os dois algoritmos. Além disso, ele se mostrou bastante eficiente em ambientes restritos, o que pode ser bastante importante em alguns contextos.

1.1 SABER

O SABER é um mecanismo de encapsulamento de chaves (KEM) pós-quântico que, como todo KEM, faz a geração de chaves, o encapsulamento e o desencapsulamento de mensagens (de 32 bytes, que são usadas como chaves de sessão em algoritmos de criptografia simétrica). Ele é IND-CCA2 seguro [1], o que significa que nenhum adversário possui vantagem significativa para conseguir distinguir pares de texto cifrado baseados nas mensagens cifradas por ele.

Como citado anteriormente, a segurança do algoritmo se baseia em problemas com reticulados, mais especificamente o Module Learning with Rounding (Module-LWR). Tal problema corresponde a uma variação do Learning with Errors (LWE) em que,

ao invés de inteiros, tem-se polinômios nas entradas das matrizes e os erros (ou ruídos) são calculados de forma determinística. Já o LWE, se trata do problema de resolver equações matriciais (correspondentes a combinações lineares de vetores) acrescidas de ruídos, como ilustrado na Figura 1. A matriz **A** e o vetor **b** formam a chave pública, o vetor **s** é a chave privada, o vetor **e** é composto pelos ruídos e o inteiro *p* define o conjunto em que as operações são realizadas [10].

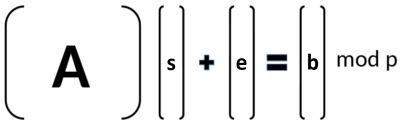

Figura 1. Esquematização do LWE.

1.2 LightSABER e FireSABER

O LightSABER e o FireSABER, como citado anteriormente, são variações oficiais do SABER, com mudança em alguns parâmetros que possibilitam níveis de segurança e de complexidade distintos. Entre os parâmetros que se diferem, o principal é o número de polinômios no vetor secreto e na matriz pública, sendo o segundo correspondente ao quadrado do primeiro. Dessa forma, o LightSABER possui 2 polinômios no vetor **s** (4 na matriz **A**), enquanto o vetor **s** no SABER conta com 3 polinômios (9 na matriz **A**) e o FireSABER possui 4 polinômios no vetor **s** (16 na matriz **A**). Por consequência, as três versões possuem tamanho de chaves e nível de segurança distintos: a versão mais leve possui chave pública de 672 bytes, chave privada de 1.568 bytes e nível de segurança quântica de 115; a versão original possui chave pública de 992 bytes, chave privada de 2.304 bytes e nível de segurança quântica de 180; a versão mais robusta possui chave pública de 1.312 bytes, chave privada de 3.040 bytes e nível de segurança quântica de 245.

2. Proposta

Diante do contexto de criptografia pós-quântica, a pesquisa tem como proposta reduzir os requisitos de memória de pelo menos um dos algoritmos submetidos ao NIST, para serem executados de forma eficiente em ambientes computacionais restritos. Ao longo dos estudos, optou-se por estudar o SABER e suas variações e executá-los em um microcontrolador ARM Cortex-M0+, dispositivo bastante limitado.

Para tanto, estudamos otimizações propostas na literatura para implementações em ARM Cortex-M0 e Cortex-M4. Vale ressaltar que as otimizações são voltadas para reduzir o consumo de memória de pilha, memória extra alocada ao longo da execução do programa. A memória de programa, por sua vez, é bastante abundante e é mais do que suficiente, portanto ela não é alvo de otimizações. Quanto às variáveis globais, elas armazenam variáveis importantes que não podem ser reduzidas.

Com relação às diferenças entre as arquiteturas ARM Cortex-M0 e Cortex-M0+, a mudança do pipeline de três para dois estágios pode impactar a contagem de ciclos. Apesar de a frequência de *clock* ter aumentado, ela não impacta os resultados pois foi medido o número de ciclos de *clock*. De forma geral, ambas são bastante parecidas e o consumo de memória deve ser bem próximo para as duas arquiteturas.

2.1 Otimizações propostas na literatura

Como já foi visto, a segurança e a complexidade do algoritmo é bastante relacionada às operações com polinômios, que são bastante custosas, especialmente porque todos eles têm 256 coeficientes de 13 bits. Por esse motivo, grande parte das otimizações propostas na literatura são feitas acerca disso. Em todo o esquema, a multiplicação de polinômios é a que mais consome memória de pilha e, como o módulo da aritmética modular do SABER não é uma potência de dois (é um número primo), ele não utiliza a Number Theoretic Transform (NTT) [5], que é a forma mais eficiente de fazer tal operação. Alternativamente, utiliza-se uma combinação dos algoritmos de Toom-Cook [3] e Karatsuba [3], que é a forma compatível mais eficiente. O problema é que esses algoritmos consomem memória de pilha da ordem de O(n), já que são recursivos e requerem memória adicional (não são *in-place*).

Em troca de velocidade, Karmakar et. al [7] propuseram uma versão *in-place* do algoritmo de Karatsuba que consome O(log n) de memória de pilha para as multiplicações, sendo vantajoso para ambientes restritos em memória. Além disso, Karmakar et. al propõem uma estratégia *just-in-time* para gerar a matriz **A** e o vetor secreto **s**. A partir dela, os polinômios são gerados um por vez, reaproveitando seu espaço na memória, de acordo com a demanda nas operações. Dessa forma, apesar de as operações ficarem mais lentas, pela necessidade de gerar os polinômios várias vezes, o consumo de memória de pilha reduz bastante. Foi necessário adaptar essas otimizações para o LightSABER e para o FireSABER, já que Karmakar et. al implementaram-nas apenas para o SABER.

Além dessas otimizações, estudamos propostas já desenvolvidas para dispositivos ARM Cortex-M4, buscando trazer as que não dependem de instruções específicas para o ARM Cortex-M0+. Dentre as propostas da literatura [8], foram implementadas duas delas. A primeira corresponde à codificar os coeficientes dos polinômios do vetor secreto **s** com apenas 4 bits, ao invés de 13 bits. A literatura afirma que essa mudança não impacta a segurança do algoritmo, por não utilizar o NTT. Adicionalmente, essa

mudança torna as funções de empacotamento (que compactam os coeficientes dos polinômios) mais simples, aumentando o desempenho do algoritmo. A segunda proposta consiste em utilizar uma versão *in-place* da verificação do texto cifrado (durante a decifração), reduzindo o uso de memória de pilha.

3. Resultados

Aplicando as otimizações descritas, o código resultante foi executado na placa de desenvolvimento FRDM-KL25Z, que possui MCU KL25Z128 (processador ARM Cortex-M0+), 128 kB de memória flash, 16 kB de memória SRAM e 48 MHz de frequência de clock. A IDE MCUXpresso, desenvolvida pela NXP, foi utilizada para a realização dos testes, já que ela possui uma ferramenta integrada para medição de uso de memória de pilha. Essa IDE possui um compilador embutido GNU Arm Embedded Toolchain 2021.07. Para contar o número de ciclos de *clock*, utilizou-se a interface CMSIS (Cortex Micro-Controller Software Interface and Standard).

Algoritmo		Geração de chaves (variação %)	Encapsulamento (variação %)	Desencapsulamento (variação %)
SABER (-O2)	Memória (kB)	4,13	3,75	3,77
	Ciclos	4.495.576	5.940.149	6.930.342
LightSABER (-O1)	Memória (kB)	3,36 (-19%)	3,46 (-8%)	3,49 (-7%)
	Ciclos	2.172.163 (-52%)	3.157.820 (-47%)	3.815.365 (-45%)
FireSABER (-O2)	Memória (kB)	4,88 (+18%)	4,00 (+7%)	4,02 (+7%)
	Ciclos	7.743.499 (+72%)	9.630.721 (+62%)	10.973.725 (+58%)

Tabela 1. Consumo de memória de pilha e ciclos de clock dos algoritmos na melhor flag de compilação.

Função		SABER em M0 [7]	SABER em M4 *speed/memory* [1] [6]	SABER em M4 *speed* [2] [6]	Este trabalho
Geração de chaves	Memória (kB)	5,03	3,79 (-25%)	6,64 (+32%)	4,13 (-18%)
	Ciclos (mil)	4.786	820 (-83%)	645 (-87%)	4.495 (-6%)
Encapsulamento	Memória (kB)	5,12	3,18 (-38%)	7,32 (+43%)	3,75 (-27%)
	Ciclos (mil)	6.328	1.059 (-83%)	851 (-87%)	5.940 (-6%)
Desencapsulamento	Memória (kB)	6,22	3,19 (-49%)	7,32 (+18%)	3,77 (-39%)
	Ciclos (mil)	7.509	1.038 (-86%)	774 (-90%)	6.930 (-8%)

[1] Versão dedicada a otimização de memória de pilha, mas também levando em conta a velocidade.
[2] Versão dedicada a otimização de velocidade de execução.

Tabela 2. Resultados dos melhores testes do SABER comparados com os valores de referência.

Quanto à compilação, fizemos testes para diversas flags de otimização, para as três versões do algoritmo, comparando ciclos de clock e memória de pilha. Para o LightSABER, a flag -O1 produziu resultados melhores. Já para as outras versões, a flag com melhor custo-benefício foi a -O2. Ao contrário do esperado, a flag -O3 não se comportou tão bem, o que sugere que ela deve ser utilizada com cautela ou até evitada. Na Tabela 1, estão dispostos os melhores valores de ciclos de clock e de memória de pilha encontrados. Como esperado pela quantidade de operações, os ciclos de clock variaram bastante conforme a versão do algoritmo. Já o consumo de memória de pilha se manteve baixo nas três versões.

Para comparação, utilizamos os valores disponibilizados no oficial do algoritmo e pelo PQM4 (*framework* responsável por bibliotecas, avaliação e

testes com algoritmos de criptografia pós-quântica [6]). Porém, não foram disponibilizados dados para o LightSABER e para o FireSABER em Cortex-M0, impossibilitando tal comparação. Os dados comparativos se encontram na Tabela 2 e indicam uma melhora considerável em relação à implementação em Cortex-M0, provavelmente por conta das otimizações trazidas da implementação para Cortex-M4. Ainda sobre a Tabela 2, nota-se que, como esperado, a implementação em Cortex-M4 apresentou melhores resultados, pelo seu poder computacional e por ter mais otimizações feitas para essa arquitetura. Mesmo assim, os resultados deste trabalho são importantes pois apresentaram uma grande redução de memória de pilha e ainda uma redução de ciclos de clock. Destacamos o menor gasto de memória de pilha, comparado com a versão focada em velocidade, e o consumo de memória de pilha similar ao da versão focada em memória.

4. Conclusões

Com a finalidade de diminuir os requisitos de memória do algoritmo SABER (e suas variações) e torná-lo eficiente em ambientes computacionais restritos, estudamos e implementamos diversas otimizações da literatura para serem executadas em um dispositivo ARM Cortex-M0+. Ademais, fizemos testes com várias flags de otimização na compilação para determinar a melhor delas para cada algoritmo. Concluímos que a flag -O1 mostrou os melhores resultados para o LightSABER e, a flag -O2, para as demais versões.

Comparando os resultados obtidos com os resultados para Cortex-M0 disponíveis na literatura [7], conseguimos reduzir entre 6% e 8% o número de ciclos de clock e entre 18% e 39% o uso de memória. Embora não tenha resultados nessa plataforma para as variações do SABER, espera-se que os valores sejam similares e proporcionais, considerando as diferenças entre as versões. Também comparamos com resultados de memória obtidos para Cortex-M4 e chegamos próximo dos resultados da versão otimizada em memória e até superar os resultados da versão que foca em velocidade.

A partir deste estudo, pode-se continuar a trazer otimizações desenvolvidas para outras plataformas, visando principalmente otimizar a velocidade (sem comprometer o consumo de memória), uma vez que o consumo de memória já foi bastante reduzido. Para trabalhos futuros, propomos uma otimização mais detalhada a nível de linguagem assembly, podendo melhorar ainda mais o desempenho do algoritmo.

Agradecimentos

Agradecemos aos membros do grupo ReGrAS (Research Group on Applied Security), que participaram e contribuíram para discussões acerca dos estudos desenvolvidos nesta pesquisa, e ao Programa Institucional de Bolsas de Iniciação Científica (PIBIC) do CNPq e à Pró-Reitoria de Pesquisa da Unicamp, pela oportunidade e suporte da pesquisa.

Referências

[1] Bogdanov, D. (2005), “IND-CCA2 secure cryptosystems”, University of Tartu.

[2] Bos, J., Ducas, et. al (2018), “CRYSTALS – Kyber: a CCA-secure module-lattice-based KEM”, 2018 IEEE European Symposium on Security and Privacy, EuroS&P.

[3] Crandall, R., Pomerance, C. (2005), “Prime Numbers – A Computational Perspective”, Second Edition, Springer, Section 9.5.1: Karatsuba and Toom–Cook methods, p.473.

[4] D’Anvers, JP., et. al (2018), “Saber: Module-LWR Based Key Exchange, CPA-Secure Encryption and CCA-Secure KEM”, In: Joux A., Nitaj A., Rachidi T. (eds) Progress in Cryptology - AFRICACRYPT 2018 - 10th International Conference on Cryptology in Africa, Marrakesh, Morocco, May 7-9, 2018, Proceedings. Lecture Notes in Computer Science 10831, Springer 2018, ISBN 978-3-319-89338-9.

[5] Hedge, S., Nagapadma, R. (2019), “Number Theoretic Transform for Fast Digital Computation”, Department of Electronic and Communication Engineering, NIE Institute of Technology.

[6] Kannwischer, M. et. al (2019), “PQM4: Post-quantum crypto library for the ARM Cortex-M4”, https://github.com/mupq/pqm4. (acessado em 27/06/2022)

[7] Karmakar, A., et. al (2018), “Saber on ARM. CCA-secure module lattice-based key encapsulation on ARM”, In Transactions in Cryptographic Hardware and Embedded Systems.

[8] Mera, J., et. al (2020), “Time-memory trade-off in Toom-Cook multiplication: an application to module-lattice based cryptography”, In Transactions in Cryptographic Hardware and Embedded Systems.

[9] National Institute of Standard and Technology – NIST (2016), “Request for Comments on Post-Quantum Cryptography Requirements and Evaluation Criteria”, Notice 81 FR 50686, p. 50686-50687.
https://csrc.nist.gov/projects/post-quantum-cryptography (acessado em 09/08/2022)

[10] Regev, O. (2005), “The Learning with Errors Problem”, Courant Institute of Mathematical Sciences, New York University.

Análise da aplicação da tecnologia FIDO (Fast Identity Online) como segundo fator de autenticação de usuários

Leone Vinícius de Oliveira , Marco A. A. Henriques
{l1178685@dca.unicamp.br, maah@unicamp.br}

Departamento de Engenharia de Computação e Automação (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Campinas, SP, Brasil

Resumo – A necessidade de corrigir as falhas ao autenticar usuários é uma urgência cada vez maior conforme o uso da internet aumenta. Métodos como login e senha, biometria e envio de senhas únicas já são amplamente utilizadas, mas possuem falhas na segurança em determinados contextos. Com a intenção de resolver esse problema, a *FIDO Alliance* desenvolveu o sistema de autenticação FIDO (Fast Identity Online). Esse método utiliza um par de chaves pública e privada, onde a chave pública é distribuída aos clientes e a privada é armazenada no autenticador, que pode ser o smartphone (com NFC ou Bluetooth) ou um token USB. A principal diferença da FIDO com os outros métodos é que a autenticação acontece apenas na presença de um dispositivo que guarda chaves criptográficas de forma segura, impedindo ataques como *phishing* (obtenção de senhas por meio da indução do usuário ao erro) e *man in the middle* (interceptação de comunicação e alteração de mensagens), já que o atacante precisaria ter acesso físico ao dispositivo para poder se passar por um usuário legítimo. Neste trabalho estamos avaliando as diferentes formas de emprego dessa tecnologia e os custos, vantagens e desvantagens de cada uma delas, de maneira a identificar possíveis aplicações no contexto de autenticação em sistemas sensíveis na universidade.

Palavras-chave – chaves criptográficas, autenticação, *Fast Identity Online*, autenticação de múltiplos fatores

1. Introdução

A constante evolução da internet demanda a agilização e a garanta de segurança ao autenticar usuários. O método mais utilizado para isso ainda é com a verificação através de um *login* e de uma senha. Este método tem sérios problemas de segurança e de usabilidade, uma vez que o usuário pode esquecer da sua senha, criar várias contas com a mesma senha, sofrer ataques como *phishing* (obtenção de senhas por meio da indução do usuário ao erro) e *man in the middle* (interceptação de comunicação e alteração de mensagens), entre outras inconveniências.

Outros métodos de autenticação surgiram com a intuição de solucionar essas falhas. A utilização da biometria para autenticar os usuários soluciona alguns dos problemas do método anterior, mas o custo para sua implementação é alto e tem um menor alcance aos usuários. Outro método de autenticação foi através do uso de senhas únicas, que poderiam ser enviadas por SMS, email ou aplicativos específicos com essa funcionalidade, como *Google Authenticator*. Estes, por sua vez, podem ter problemas ao transmitir a senha única e podem gerar dificuldades ao usuário para a compreensão de como utilizar corretamente. Considerando todos esses pontos, a *FIDO Alliance* desenvolveu protocolos para o método de autenticação FIDO(*Fast Identity Online*).

2. Chaves criptográficas

As chaves criptográficas são um conjunto de informações (bits, números, letras, etc.) que representam a identidade de um usuário. Com elas, é possível codificar e decodificar mensagens. Existem as chaves simétricas, onde a chave para codificar é a mesma para decodificar, e as assimétricas, que são chaves diferentes para codificar e decodificar uma mesma mensagem.

O método de criptografia que utiliza chaves assimétricas também é chamado de criptografia por chave pública. Nele a chave privada deve ser armazenada com sigilo e somente o usuário dono deve ter acesso à ela. Já a pública pode ser distribuída aos serviços que a requisitarem para autenticar o usuário que possui seu par privado. Uma mensagem criptografada por uma chave privada só pode ser decodificada pela chave pública que compõe o seu par, e vice-versa.

Utiliza-se o método de criptografia por chave pública para prover sigilo à mensagem que deve ser enviada e para garantir sua autoria. Ao criptografar a mensagem utilizando a chave pública, apenas o usuário que possuir sua chave privada par tem a capacidade de decodificá-la, garantindo assim o sigilo. Já quando a mensagem é criptografada pela chave privada, qualquer um que possuir a chave pública par pode ler a mensagem, mas o destinatário saberá que somente o dono da chave privada par foi capaz de escrever aquela mensagem codificada.

3. O protocolo SSH

O protocolo SSH é um conjunto de regras que determina diretrizes de como conectar dois computadores de forma segura utilizando a criptografia com um par de chaves pública-privada. Para configurar a criptografia como SSH, três requisitos devem ser alcançados: sigilo, autenticação da origem e integridade da mensagem. Os dois primeiros requisitos já caracterizam uma comunicação por par de chaves assimétricas. O terceiro requisito define que a mensagem deve ser entregue sem sofrer nenhuma alteração.

Neste protocolo, o computador local requisita uma conexão com a máquina remota desejada, que já deve ter um *software* com o protocolo SSH. Então esta máquina, que é o servidor SSH, envia um desafio para o computador local, que deve resolvê-lo e confirmar sua identidade criptografando a mensagem com a chave privada. O servidor SSH verifica então a resposta utilizando a chave pública par para descriptografar a mensagem e então permitir o acesso do usuário. Toda essa conversa entre os computadores acontece de forma que o usuário não necessite realizar nenhuma ação [1].

4. *FIDO Alliance*

A *FIDO Alliance*, consórcio de empresas que pretende mudar a forma como a autenticação online é feita, desenvolveu três conjuntos de especificações e regras: FIDO Universal Second Factor (FIDO U2F), FIDO Universal Authentication Framework (FIDO UAF) e FIDO2. Juntos, estes protocolos têm o objetivo de mostrar como alcançar uma comunicação mais simples, mais segura e mais robusta para o usuário.

4.1. FIDO U2F

Esse conjunto de protocolo tem como objetivo assegurar a segurança ao utilizar o protocolo FIDO como segundo fator de autenticação. Para isso, um dispositivo que seja compatível com os protocolos FIDO deve ser utilizado, que chamaremos de dispositivo U2F. Ele pode ser um token USB ou um smartphone que tenha como se comunicar com as tecnologias Bluetooth ou NFC. Quando o usuário se cadastra em um serviço, ele deve conectar o dispositivo U2F, que gera o par de chaves público-privado, armazenando localmente a chave privada e registrando a chave pública par no servidor do serviço requisitado. Quando o usuário se identificar novamente, o serviço o identifica através do dispositivo U2F. Esse conjunto de protocolos também definem qual a criptografia utilizada e como a comunicação com o dispositivo será realizada [2]. Ao utilizar um dispositivo U2F, é garantido que o usuário está fisicamente em posse do dispositivo, impedindo que alguém que possua seus *login* e senha possa se passar por ele.

4.2. FIDO UAF

Esse conjunto de protocolos tem como objetivo fornecer um mecanismo de autenticação unificado que substituam as senhas. Os protocolos FIDO UAF determinam como devem ser feitos o registro, a autenticação, a confirmação da transação e a exclusão de cadastro de usuários [3].

4.3. FIDO2

FIDO2 é a junção de dois conjuntos de protocolos: *Client-to-Authenticator Protocols* (CTAP) e *W3C Web Authentication* (WebAuthn). Elses se complementam para garantir a segurança na autenticação em aplicativos web. O CTAP foi desenvolvido pela *FIDO Alliance* e tem como objetivo determinar como a interface autenticadora deve receber e passar os parâmetros e desafios necessários para a autenticação, como as mensagens devem ser criptografadas e como o meio de transporte assegura a integridade das mensagens enviadas [4].

5. O protocolo FIDO

O protocolo FIDO determina como devem ser feitos o registro e a autenticação do usuário utilizando a criptografia por um par de chaves pública-privada. Tanto no registro quanto na autenticação, são determinadas as especificações entre três sujeitos: o Autenticador FIDO, que pode ser um token USB ou um smartphone com NFC ou Bluetooth; o cliente FIDO, que é o navegador que requisita o cadastro; e o servidor FIDO, que verifica se a autenticação é bem sucedida ou não.

5.1. Registro

Quando o cliente pede para se cadastrar no servidor FIDO, ele retorna ao cliente informações da conta, um desafio e informações opcionais de segurança que ajudam a evitar ataques. No pacote de autenticação, o cliente determina qual vai ser o método de autenticação e adiciona também informações de segurança necessárias para o autenticador. O cliente então envia esse pacote ao autenticador para ele verificar o usuário, criar um par de chaves pública e privada e a credencial, e retornar ao cliente um pacote com a sua credencial para aquele serviço específico, a chave pública e informações sobre o sucesso ou a falha na autenticação do usuário. O cliente envia esse pacote ao servidor, que armazena a chave pública e a credencial do cliente. Vale notar que o usuário apenas teve que se autenticar localmente com o autenticador, sem a necessidade dele enviar nenhuma informação online [5].

5.2. Autenticação

Ao desejar acessar o serviço novamente, o usuário se identifica e então é realizada a autenticação. Nesse processo, o servidor FIDO envia uma mensagem ao cliente com um desafio e dados opcionais de segurança. O cliente então cria um pacote com os dados do cliente e com o desafio, e o envia para o autenticador. Este então verifica o usuário e depois busca a chave privada para resolver o desafio, verificando também se apenas um acesso foi realizado, aumentando assim a segurança na comunicação. Após isso, o autenticador retorna ao cliente um pacote onde indica se a autenticação do usuário foi bem sucedida e, caso positivo, envia a resposta do desafio. O cliente repassa o pacote ao servidor, que busca a chave pública da credencial do cliente no banco de dados e verifica a resposta do desafio. Caso a resposta esteja correta, o usuário estará autenticado. Nota-se que mais uma vez o usuário precisou se autenticar apenas localmente com o autenticador [5].

6. Conclusões e próximos passos

Através de um exemplo básico de funcionamento, foi possível perceber a praticidade da autenticação. Após se identificar, o usuário deve realizar uma pequena ação requisitada pelo dispositivo autenticador, que pode ser um toque no sensor ou apertar um botão. Como todo o processo de autenticação acontece apenas localmente, com o cliente enviando ao servidor FIDO informações de sucesso ou de falha, foi concluído que a segurança é maior utilizando os protocolos FIDO, quando comparado com outros métodos de autenticação. Apesar disso, esse método de implementação ainda não está amplamente utilizado por três principais fatores: (i) nem todos os dispositivos são compatíveis com a tecnologia FIDO; (ii) uma dependência aos dispositivos autenticadores, que leva ao investimento para a aquisição de tokens USB ou às inconveniências de garantir que o usuário esteja com um smartphone compatível sempre com a bateria carregada; (iii) e explicar aos usuários o passo adicional para a autenticação, que, mesmo sendo mais prático que verificar uma senha única em um aplicativo próprio ou em outros canais, ainda é existente e menos robusto que a simples identificação por *login* e senha.

A continuação deste trabalho tem como objetivo mostrar como a tecnologia FIDO pode substituir os métodos tradicionais de segundo fator de autenticação, utilizando como exemplo a comunidade existente na Unicamp, onde o público alvo é diversificado. Deve ocorrer primeiramente uma identificação do método já existente e determinar suas dificuldades. Após isso deve-se determinar como realizar a substituição e verificar se ela é plausível ou não. Por fim deve-se justificar os resultados obtidos.

7. Agradecimentos

Eu tenho uma enorme gratidão ao meu professor orientador Marco Amaral Henrique, que me passou todas as principais orientações e direcionamentos para que a pesquisa pudesse ser concluída. Além disso, agradeço sua paciência e didática para me passar seus conhecimentos fundamentais para a melhor compreensão dos resultados obtidos. Agradeço também meus pais e amigos, Lilian Margareth da Silva Oliveira, Jânio José de Oliveira e João Vitor Tobias da Silva, que sempre se colocaram a disposição para apoiar minha pesquisa e não deixar faltar motivação para concluí-la.

Referências

[1] D. J. Barrett and R. E. Silverman, *SSH, the Secure Shell: The definitive guide*. O'Reilly and Associates, 2001.

[2] D. Balfanz, S. Srinivas, and E. Tiffany, "Universal 2nd factor (u2f) overview," *FIDO Alliance*, 2014.

[3] R. Lindemann and E. Tiffany, "Fido uaf protocol specification," *FIDO Alliance*, 2020.

[4] J. Bradley *et al.*, "Client to authenticator protocol (ctap)," *FIDO Alliance*, 2022.

[5] G. Prado, M. Landi, and A. Shikiar, "Introdução à autenticação fido," 2019.

FEEC | UNICAMP
E A D C A 22
XIV ENCONTRO DE ALUNOS E DOCENTES DO DCA

Sessão Técnica 3

Arquitetura baseada em internet das coisas robóticas para aplicação em casas inteligentes de pessoas com dificuldade de locomoção com foco em idosos

Felipe Augusto Oliveira Mota , Suzana Viana Mota , César Bastos da Silva , Vinicius Emanoel Ares , Victor Fermán , Arnaud de Jarcy , Eric Rohmer
{felipeaomota@gmail.com, rohmer@unicamp.br}

Departamento de Engenharia de Computação e Automação (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Campinas, SP, Brasil

Resumo – Diante do crescimento, proporcional, da população idosa, faz-se necessário desenvolver tecnologias assistivas que beneficiem a qualidade de vida deles. Assim, é proposto o desenvolvimento de uma arquitetura baseada em internet das coisas robóticas para aplicação em casas inteligentes com foco no público citado. Para realização geral do projeto, primeiro investiga-se o público-alvo, define-se recomendações para trabalhar com a tecnologia e por fim propõe uma arquitetura inicial. A solução deverá especificar um sistema remoto com uma estrutura local composta por módulos de atuação que gerenciem todas as coisas robóticas. Além disso, um dispositivo, mediador para padronizar a relação do usuário com toda estrutura.

Palavras-chave – casa inteligente, dificuldade de locomoção, idoso, internet das coisas robóticas.

1. Introdução

A proporção da população idosa tem aumentado. O relatório das Nações Unidas sobre envelhecimento da população Global (apud [11]) apontava que 25% da população europeia já era idosa. O crescimento projetado para os próximos 15 anos na América Latina e Caribe é de 71%, na Ásia de 66% e na África é de 64%. A qualidade de vida dos idosos é influênciada por: estilo de vida, sexo, idade, escolaridade, etnia, capacidade física, doenças e renda. Além da qualidade do sono, capacidade funcional e características sociodemográficas [5]. A qualidade de vida dos idosos também pode ser conquistada pelo estado psicológico e mental do idoso [8].

A atenção a essas necessidades perpassa pela Tecnologia Assistiva (TA). A TA fornece uma interface para incapacidade das pessoas em colaborar com a tecnologia disponível para realizar as atividades esperadas. Funcionando como uma interface entre pessoas e sistema para colaborar tanto para operar a tarefa dada, atribuídas por pessoas e executadas pelo sistema [4]. Diante disso, surge o seguinte questionamento: Como gerar qualidade de vida para os idosos, através da tecnologia assistiva?

Uma solução pode ser a implantação de casas inteligentes para idosos. Uma casa inteligente é composta por sensores, aparelhos e dispositivos de conexão que podem ser operados remotamente. Normalmente, esses sistemas são conectados em rede de comunicação sem fio e utilizam um padrão para sua comunicação. Pode-se também fazer o gerenciamento de energia, segurança e experiência de conforto doméstico [14].

Para implementar casas inteligentes, [11] e [10] utilizam a Internet das Coisas Robóticas (*Internet of Robotic Things*, IoRT). [11] justificam que o financiamento para apoiar os projetos de robótica para enfrentar os desafios do envelhecimento da população tem aumentado. Já [10] abordam que o cenário mundial caminha para tecnologias sem contato, onde os robôs estão no centro.

A IoRT pode ser definida como uma infraestrutura global para a sociedade da informação que permite serviços robóticos avançados, interconectando coisas robóticas com base em tecnologia de informação e comunicação interoperáveis, existentes e em execução [13].

Ademais, os idosos, além das questões inerentes a idade, eles podem ser classificados dentro do filão de Pessoas com Dificuldade de Locomoção (PDL). Que inclui também os dependentes de cadeira de rodas, de muletas, as gestantes, os obesos, pessoas com deficiências temporárias, entre outros [3]. Ao desenvolver uma proposta de casa inteligente para o idoso pode-se alcançar outros indivíduos considerados PDL.

Assim, o objetivo do presente trabalho consiste em discutir e apresentar a proposta inicial, do desenvolvimento de uma arquitetura baseada em IoRT a fim de propiciar casas inteligentes para idosos e, por conseguinte, para outros tipos de PDLs.

2. Proposta

A arquitetura deverá ser aplicada para gerir, ao mesmo tempo, dispositivos inteligentes comuns e robôs que assistam o indivíduo em sua distinta rotina doméstica e não apenas automatizem serviços. Espera-se disponibilizar uma plataforma que permita a união de monitoramentos, serviços e assistências. Sendo possivel alinhar monitoramento de saúde, serviços domésticos, cuidados médicos, manipulação, mobilidade, telepresença, entreternimento, monitoramento e telemetria. A Figura 1 representa a proposta.

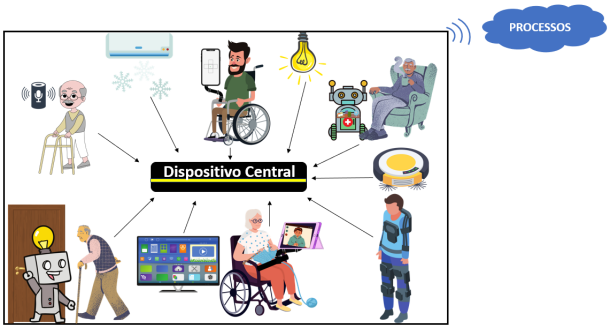

Figura 1. Visão geral da proposta.

O projeto precisar ser modular, onde seja possível conectar e utilizar diferentes equipamentos, de acordo com as necessidades do usuário. A arquitetura proposta deverá ser cognitiva e desenvolver todos os seus processos robustos em nuvem. Dentre outros processos, deverão ser realizados em nuvem toda a parte de inteligência computacional, banco de dados e *digital twin* das coisas robóticas inseridas nas residências. Com isso, há a possibilidade de menor manutenção local, atualização ágil da parte lógica e melhor gerencimento dos processos.

2.1. Metodologia Adotada

O método para execução do trabalho é baseado na metodologia trifásica (TPM) [6], que tem como foco o processo de design IoT, abordando questões inerentes ao desenvolvimento de uma solução IoT independente do fabricante. Os autores definem as 3 fases da estrutura como:

- Compreendendo o negócio - tem como objetivo apresentar o problema a ser resolvido, são levantadas as questões sobre o mesmo para levar a compreensão das demandas do projeto;
- Definição de requisitos - busca discutir os problemas levantados na primeira fase a fim de levantar requisitos necessários para o desenvolvimento do projeto;
- Implementação - serão realizadas as escolhas tecnológicas do projeto, avaliação e comparação de tecnologias disponíveis no mercado e também é realizado o monitoramento e o controle dos dispositivos (coisas) utilizadas no projeto.

Este atual trabalho, por estar em fase inicial, abordará a realização e resultados preliminares da primeira fase metodológica. Possibilitando prospecções para as fases posteriores. Foi desenvolvida uma pesquisa bibliográfica para compreender os requisitos do público em foco e também as recomendações para o trabalho com IoRT. Isso possibilitou a geração das exigências inciais para a arquitetura da proposta.

3. Resultados

3.1. Compreensão do Público

Pode-se afirmar que o sucesso na proposição de uma solução depende, necessariamente, do conhecimento das diversas demandas características da pessoa idosa no contexto de suas atividades diárias [9].

Uma casa segura para o idoso é um ambiente planejado com a finalidade de minimizar ao máximo o risco de acidentes, oferecendo segurança, conforto, independência e qualidade de vida [7]. Já, os PDLs de forma geral, apresentam dificuldades para se deslocarem de um ponto a outro do espaço urbano podem representar tanto um desafio a ser superado, quanto um cansaço desencorajante em seus movimentos reduzido [3].

Ainda, ao refletir sobre os idosos, uma casa inteligente deve ser dividida em física e psicológica. Com isso, a residência inteligente deve propiciar quatro princípios: Fortalecer a capacidade do idoso de estabelecer conectividade; Reduzir as dificuldades e barreiras dos aspectos de produtos e serviços para estabelecer conectividade; Aumentar as oportunidades de estabelecer conectividade; Melhorar a qualidade da conectividade [8].

Em estudo realizado, [10] comprovaram que a introdução de um mediador, que oriente todas as ações com um único padrão de atuação, ajuda numa casa IoRT a difundir os serviços e utilitários fornecidos. A IoRT com dispositivo mediador manteve todos os membros das famílias agindo com autonomia e mantendo seus postos sociais de costume na organização familiar. Sem o mediador, os mais velhos dependiam dos mais jovens para utilização da tecnologia. Com o mediador, os mais velhos não tinham necessidade de aprender a operar todos os dispositivos existentes, atribuindo a esse público uma maior independência, dessa forma, a solução proposta adota a relação usuário-mediador-sistema.

Assim, é possível que o indíviduo controle e interaja com o dispositivo mediador. A partir do mediador

há um contato com o dipositivo central que ajuda a controlar as atividades locais. Tudo apoiado pela gestão dos processos baseados em nuvem.

3.2. Recomendações para Aplicações IoRT

Os sistemas IoRT devem ser capazes e lidar com eficiência em todas as situações e desafios complexos dentro do ambiente empregado [12]. Os autores, ainda, afirmam que todas as aplicações IoRT requerem uma abordagem de arquitetura multidimensional e multicamada para realizar todas as diferentes tarefas de detecção, atuação, rede e interação. Eles definem que as características de um sistema IoRT são: sensoriamento (detecção), atuação, controle, planejamento, percepção e cognição.

De formar resumida, pode-se definir uma arquitetura IoRT em 4 níveis [1]: 1-Elementos Físicos; 2-Redes e Infraestrutura; 3-Aplicativos e Serviços; 4-Pessoas e Processos de Negócio.

Outras propostas de camadas mais extensas e especificadas sãos propostas em [12] e [13]. Este último, cita, dentre outras condições, o que é relatado abaixo para justificar sua proposta:

- Novo conceito de atuação como serviço, para garantir a adaptabilidade e iteração do usuário para dispositivos IoRT. Aqui, o conceito de máquina-a-máquina (M2M) é estendido para máquina-maquina-atuador (M2M2A);
- A percepção em robótica é considerada como uma combinação de informação de sensores com modelagem de conhecimento para permitir que a robótica realize interação máquina-humano;
- Quanto a cognição, a IoRT pode aproveitar a inteligência local e distribuída.
- As duas funções mais importantes executadas pela IoRT são Serviços Robóticos de Interação (RoIS) e Interação Humano-Robô (HRI). Para realizar interações e reações, vários equipamentos podem ser necessários: câmera; microfone, radar, Lidar, dentre outros senores.

Em complemento ao que já foi exposto pode-se acrescentar os conceitos de *edge computing* e *fog compuntig*. As duas processam dados sem necessidade da nuvem, o que auxilia como uma filtragem, ajuda a diminuir o tráfego de dados e podem decorrer premissas para contribuir na segurança e agilidade da informação. O que diferencia o trabalho das duas é o local onde o processamento ocorrerá. A decisão de onde o sistema deve atuar depende da eficiência e adaptabilidade diante do serviço que se deseja executar [2].

A *fog* define uma névoa para processar dados abaixo da nuvem, criando uma espécie de nuvem específica (identificada) para o projeto. Ela processa dados mais robusto que a *edge*. Esta, por sua vez processa os dados nos chamados dispositivos de borda, que podem ser desde os dispositivos que estão sensoriando o ambiente ou até num *gateway* local. Assim, tendo um processamento muito mais próximo das coisas.

3.3. Arquitetura Inicial

Para as proposições iniciais, considera-se todos os dispositivos como robôs e descreve-se a relação usuário, sistema IoRT e ambiente, sem diferenciar o que é conectividade física e psicológica. Ao pensar nessa relação, determina-se os seguinte passos:

1. O usuário solicita o serviço ao mediador;
2. O mediador comunica com a central (gateway);
3. Central atua de acordo com a regra imposta pela nuvem;
4. Central envia comando para dispositivo inteligente;
5. Dispositivo inteligente atua sobre o ambiente;
6. Ação Concluída? Envia resultado para central;
7. Central salva em nuvem que atividade foi realizada;
8. Nuvem salva no “perfil” do usuário que a atividade foi realizada;
9. Nuvem repassa para central a alteração de status da atividade;
10. Central comunica mediador;
11. Mediador comunica usuário.

A arquitetura inicial proposta é apresentada na Figura 2. Nela é possível contemplar toda organização dividida em 4 camadas. A computação de borda (*edge*) auxilia numa filtragem inicial de dados, podendo dedicar aos dispositivos um processamento local. Mais próxima a rede local, tem-se a névoa (*fog*), garantindo identificação dos servidores remotos que auxiliam em ações mais específicas do sistema. Por fim, em nuvem (que pode ser contrada de terceiros) existe o apoio a serviços mais robustos. Ainda precisa ser definido melhor o que será processado na névoa e o que será processado na nuvem. Toda ação iniciando através de um mediador e com requisitos de segurança em todos os níveis.

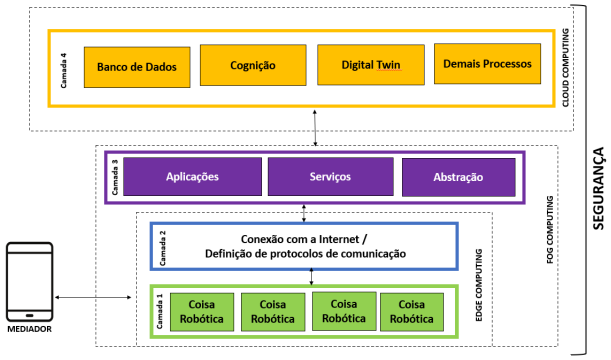

Figura 2. Arquitetura Inicial Proposta

4. Considerações Finais

Propôs-se uma arquitetura inicial IoRT para casas inteligentes de PDLs, com foco em idosos. Percebeu-se a necessidade de um mediador para padronizar o controle do usuário. Foi possível descrever as ações da relação do usuário com o sistema e também foi determinada a arquitetura inicial, divida em 4 camadas. Essas definições permitiram compreender, inicialmente, o comportamento e disposição do sistema. Porém, ainda há necessidade de precisar os locais onde cada processo será executado. Para continuidade das fases metodológicas, serão pesquisadas a rotinas dos idosos e outros PDLs brasileiros, além de um estudo mais aprofundado acerca de IoRT, desenvolvendo simulações e testes. Ao fim, espera-se uma estrutura cognitiva que consiga se adaptar à realidade do usuário, que monitore a posição do indivíduo dentro da residência para fomentar a tomada de decisão robótica e que gerencie todos os processos sistêmicos e de inteligência computacional em nuvem.

Agradecimentos

Ao IFNMG pela bolsa concedida através do Programa de Bolsas para Qualificação de Servidores (PBQS).

Referências

[1] Michel Albonico, Adair José Rohling, Paulo Jr. Varela, and Juliano Soares dos Santos. Mining evidences of internet of robotic things (iort) softwarefrom open source projects. *Simpósio Brasileiro de Componentes, Arquiteturas E Reutilização de software (SBCARS)*, 2021.

[2] Rajkumar Buyya and Satish Narayana Srirama. *Fog and Edge Computing: Principles and Paradigms*. Wiley, 2019.

[3] Cristiane Rose Duarte and Regina Cohen. Afeto e lugar: Pessoas com dificuldades de locomoção e espaço urbano. `http://www.bengalalegal.com/afetoelugar`, 2005.

[4] Anil Kumar Dubey, Shankar Hari Mewara, Khushbu Gulabani, and Prakriti Trivedi. Challenges in design deployment of assistive technology. *International Conference on Signal Propagation and Computer Technology (ICSPCT)*, 2014.

[5] Luana Karoline Ferreira, Juliana Fernandes Filgueiras Meireles, and Maria Elisa Caputo Ferreira. Avaliação do estilo e qualidade de vida em idosos: Uma revisão de literatura. *Revista Brasileira de Geriatria e Gerontologia*, 21(5):639–651, 2018.

[6] Luiz Carlos Branquinho Caixeta Ferreira, Omar Carvalho Branquinho, Pedro Rinaldo Chaves, Paulo Cardieri, Fabiano Fruett, and Michel Daoud Yacoub. A pbl-based methodology for iot teaching. *IEEE Communications Magazine*, 57(11):20–26, 2019.

[7] Márcia Maria Vieira Hazin. Os espaços residenciais na percepção dos idosos ativos. Master's thesis, Universidade Federal de Pernambuco, Recife, Março 2012.

[8] Yuqi Liu and Ryoichi Tamura. How can smart home help "new elders" aging in place and building connectivity. *16th International Conference on Intelligent Environments (IE)*, 2020.

[9] Fausto Orsi Medola. Design de produtos assistivos para idosos. *Estudo Interdiciplinar Envelhecimento*, 25(Edição Especial):14–23, 2020.

[10] Byeong June Moon, Sonya S. Kwak, Dahyn Kang, Hanbyeol Lee, and Jongsuk Choi. The effects of internet of robotic things on in-home social family relationships. *29th IEEE International Conference onRobot and Human Interactive Communication (RO-MAN*, 2020.

[11] Tamanna Motahar, Fahim Md. Farden, Dibya Prokash Sarkar, Atiqul Md. Islam, Maria E. Cabrera, and Maya Cakmak. Sheba: A low-cost assistive robot for older adults in the developing world. *28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2019.

[12] Anand Nayyar, Ranbir Singh Batth, and Amandeep Nagpal. Internet of robotic things: Driving intelligent robotics of future- concept, architecture, applications and technologies. *4th International Conference on Computing Sciences*, 2018.

[13] Partha Pratim Ray. Internet of robotic things: Concept,technologies, and challen. *IEEE Access*, 4:9489–9500, 2016.

[14] Safdar Rizvi, Izaan Sohail, Mehreen M Saleem, Areeba Irtaza, Maria Zafar, and Mehak Syed. A smart home appliances power management system for handicapped, elder and blind people. *4th International Conference on Computer and Information Sciences (ICCOINS)*, 2018.

Evaluation of open source network operating systems for 5G disaggregated cell site gateway solutions

Alan Teixeira da Silva and Christian Esteve Rothenberg

{alan.teixeirads@gmail.com, chesteve@dca.fee.unicamp.br}

Departamento de Engenharia de Computação e Automação (DCA)

Faculdade de Engenharia Elétrica e de Computação (FEEC)

Universidade Estadual de Campinas (Unicamp)

Campinas, SP, Brasil

Abstract – The deployment of fifth generation mobile networks (5G) has been steadily increasing worldwide since the past year as more countries adhered to the movement (e.g., passing legislation and auctioning radio spectrum). Thus, the level of infrastructure development required to smoothly apply the technology has escalated substantially as well and the convergence of parallel fields - such as Network Operating Systems, the development of data plane programmability brought by Programming Protocol-independent Packet Processors (P4) language and white cell site gateways - is imperative to further improve performance. In this context, this work evaluates the suitability of Software for Open Networking in the Cloud (SONiC) and P4 Integrated Network Stack (PINS) in a disaggregated architecture for 5G mobile infrastructure scenario involving Disaggregated Cell Ste Gateway (DCSG) cell site router family.

Keywords – Network Operating System, Disaggregated Cell Site Gateway, SONiC, PINS, P4

(Texto removido pela cláusula de sigilo dos convênios com as empresas.)

Transferência de Estilo para Síntese de Fala Expressiva

Leonardo B. de M. M. Marques , Lucas H. Ueda , Paula D. P. Costa
{l218479@dac.unicamp.br, l156368@dac.unicamp.br, paulad@unicamp.br}

Departamento de Engenharia de Computação e Automação (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Campinas, SP, Brasil

Resumo – Embora pesquisas recentes em sistemas de conversão de texto em fala tenham mostrado melhorias significativas na naturalidade e inteligibilidade, transmitir aspectos expressivos da fala por meio da síntese ainda é um problema em aberto. Esta é uma característica crucial para permitir que agentes artificiais socialmente interativos exibam comportamentos típicos da comunicação humana. A maneira mais comum de abordar a expressividade é considerar os estilos de fala, uma descrição de alto nível das maneiras de falar, como "narrativa", "amigável" ou "sussurrando". Neste contexto, este trabalho aborda a seguinte questão: como modelar estilos de fala de maneira realista? Para resolver esse problema, exploramos o uso de modelos generativos de difusão, e visamos usar recursos prosódicos de baixo nível da fala, frequência fundamental, duração e energia, a fim de obter representações de estilo que melhor condicionam o modelo texto-fala.

Palavras-chave – síntese de fala expressiva, estilos de fala, transferência neural

1. Introdução

Devido ao rápido desenvolvimento das técnicas neurais de modelagem acústica e geração de formas de onda, as tecnologias de texto-fala estão reduzindo progressivamente a lacuna entre a fala natural e a sintética [9]. No entanto, um problema ainda em aberto é a síntese de fala expressiva realista [8].

A síntese de fala expressiva pode ser caracterizada como um problema de mapeamento do tipo *one-to-many*, uma vez que um mesmo fonema pode apresentar diferentes produções acústicas e prosódicas, observáveis, por exemplo, em diferentes entonações, sotaques, ritmos e velocidade de produção [12].

Os estilos de fala são definidos como os atributos globais que descrevem a emoção, afeto e/ou atitude social transmitida através da fala por um falante em um domínio particular. Leitura, locutor, conversação ou emoção (feliz, zangado, triste, etc...) são alguns exemplos [9].

Nas conversas cotidianas face-a-face, as interações contém sinais sociais como humor, empatia e compaixão por meio tanto do conteúdo linguístico quanto do estilo da fala. Assim, para alcançar um meio de comunicação mais afetivo e humano, é de grande importância que os sistemas de conversão de texto em fala sintetizem falas com estilos de fala adequados que estejam de acordo com o contexto conversacional [7].

As principais abordagens que visam introduzir expressividade tentam modelar o estilo usando uma rede neural para gerar um vetor latente único e global através do aprendizado não supervisionado que represente o

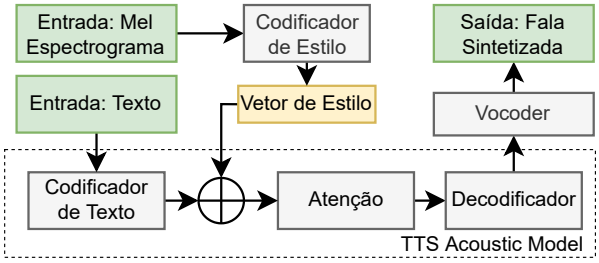

Figura 1. Arquitetura de um TTS neural incorporando estilo.

estilo [14]. A grande maioria desses trabalhos usa como entrada o mel-espectrograma (uma representação do sinal de áudio no domínio do tempo e da frequência) e, através de uma rede recorrente e convolucional (chamada de codificador de referência [10]), constrói-se um vetor de referência, que é utilizado por outros módulos na obtenção do vetor de estilo. À composição desses módulos junto com o codificador de referência, dá-se o nome de codificador de estilo. A arquitetura típica de um sistema desse tipo é apresentada na Figura 1, na qual os blocos verdes são as entradas e saída, os cinzas são redes neurais, e em amarelo destaca-se o vetor de estilo.

Existem várias tentativas de melhorar a capacidade de modelagem de estilo do vetor de referência, como a partir desse realizar o aprendizado de um banco de vetores chamado "Global Style Tokens" (GST) [11], esperando que cada um capture um aspecto global aleatório da distribuição de áudio, como por exemplo velocidade de fala, ruído de fundo, timbre, etc. Outra abordagem consiste no uso de modelos generativos, como o auto-encoder variacional (VAE) [15] e fluxos normalizadores

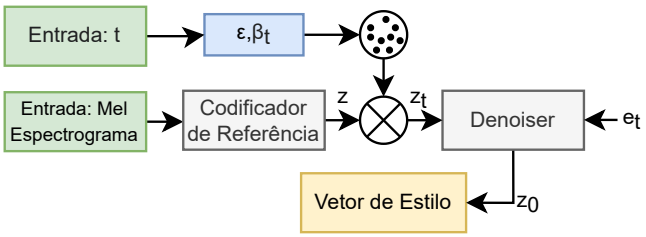

Figura 2. Codificador de estilo baseado em modelos de difusão.

(*flows*) [1]. Um problema recorrente desses modelos é a questão do vazamento de informações: não se sabe se o vetor de estilo está modelando apenas o estilo isoladamente: não há interpretabilidade nos módulos citados acima, e informações do falante, ruído de fundo, ambiência, ou quais outras características podem estar sendo modeladas. Alguns trabalhos já propuseram métodos para desembaraçar os atributos do falante e do estilo obtidas através do mel-espectrograma de referência [2].

Com base no exposto, recentemente, o desafio de incorporar estilo em sistemas TTS expressivos foi subdividido em duas questões principais [13]: como obter um vetor de estilo significativo, dado o rótulo de estilo e como injetar adequadamente o vetor de estilo em um modelo acústico texto-fala. Nesse contexto, o presente trabalho se concentra no primeiro problema: como modelar o estilo de fala de maneira realista.

2. Proposta

Dado o recente sucesso de modelos de difusão [3] na tarefa de síntese de imagem a partir do texto, obtendo melhor desempenho que as GANs, inicialmente exploramos a modelagem de estilo através desses. Modelos de difusão [5] são um tipo de modelo generativo consistindo em uma cadeia de Markov que gradativamente remove a informação presente nos dados através da adição de ruído sequencial. Dessa maneira, leva-se a distribuição original dos dados à uma distribuição gaussiana. Após esse procedimento, o processo reverso gradativo de reconstrução é aprendido através de redes neurais, criando-se assim a capacidade de sintetizar um dado partindo de uma amostra de ruído gaussiano.

O processo de geração do vetor de estilo, mostrado na Figura 2, consiste então na entrada com o mel-espectrograma de referência, o qual se deseja capturar o estilo, que é transformado num vetor de referência, denominado z, através do codificador de referência. Entra-se também com o número t de passos de ruído gaussiano que serão adicionados ao vetor de referência z, levando esse o vetor de referência para o nível de ruído z_t, a partir da seguinte equação:

$$\boldsymbol{z}_t(\boldsymbol{z}, \boldsymbol{\epsilon}) = \sqrt{\bar{\alpha}_t}\boldsymbol{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad (1)$$

na qual $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ e $\bar{\alpha}_t$ é um hiper parâmetro do processo de difusão.

Esse vetor ruidoso de referência é então utilizado como conhecimento a priori para iniciar o processo reverso de difusão. Em cada passo, uma única rede neural, o "denoiser", recebe o vetor de referencia ruidoso no nível de ruído t, z_t, juntamente com o vetor que sinaliza o passo que está sendo executado, e_t e retorna o vetor reconstruído z_{t-1}. A modelagem do processo reversa também é feita através de distribuições gaussianas, tendo a média e variância modelada por redes neurais a partir de z_t, de acordo com a equação a seguir:

$$p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t) := \mathcal{N}(\boldsymbol{z}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{z}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{z}_t, t)), \quad (2)$$

na qual $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ representam a média e a matriz de covariância respectivamente.

Esse processo é então repetido t vezes para reconstruir o vetor o qual foi aplicado t passos de ruído, sendo o resultado final, z_0, o denominado vetor de estilo, que será usado para condicionar o modelo acústico de texto-fala a gerar fala no estilo da referência dada. O vetor de estilo é então concatenado à saída do codificador de texto e ambos servem de entrada para os módulos de atenção e o decodificador. Construiu-se a hipótese de que, a partir processo de remoção de ruído iterativo, o modelo de difusão era capaz de selecionar os atributos mais importantes do vetor ruidoso e reconstruir o vetor de estilo da maneira que melhor condicionasse o modelo acústico.

Seguindo a tendência de modelos TTS de começarem a usar atributos prosódicos de baixo nível para realizar a síntese de fala [6], e também com o intuito de controlar as informações que o codificar de estilo recebe/processa, a abordagem de utilizar esses atributos ao invés do mel-espectrograma será investigada. Especificamente, ao entrarmos com um mel-espectrograma para referência de estilo, não há garantia de que a rede explicitamente esteja modelando o estilo, podendo capturar outros atributos indesejados presentes, como o timbre do falante, ruído de fundo, ambiência, entre outros.

Hipotetizamos que um número finito de características prosódicas de um sinal de áudio seja suficiente para conseguir capturar seu estilo. Dessa maneira, ao introduzir essas características no codificador de estilo, pretende-se mitigar problemas de vazamento de informações não desejadas (como timbre do falante, ruído de fundo), que ocorrem atualmente ao se entrar com o mel-espectrograma, avançando num caminho da interpretabilidade para obter representações de estilo mais significativas.

Assim, considera-se utilizar um subconjunto de parâmetros do GeMAPS [4], um grupo de parâmetros acústicos selecionados com base no potencial de indicar características afetivas fisiológicas, utilização em trabalhos passados e significância teórica; a fim de ser um padrão para pesquisas futuras. Visa-se realizar um estudo baseado na importância de atributos para avaliar quais são suficientes para capturar bem o estilo em diferentes bancos de dados contendo diferentes estilos.

Adicionalmente, um objetivo também é o de mudar do aprendizado não-supervisionado para o supervisionado, a fim de obter uma melhora no desempenho. Com a introdução dos rótulos de estilo, é possível adicionar módulos classificadores de emoção após o codificador de estilo para fazer com que os gradientes do classificador torne os vetores de estilo mais discriminativos, tendo esses a informação que distingue os estilos. Também, a fim de fazer com que o codificador de estilo não aprenda informações de falante, é possível introduzir um classificador de locutor com uma camada de reversão do gradiente, com o objetivo de se afastar dos mínimos.

3. Resultados

Um banco de dados de uma única falante em Português do Brasil foi usado. Ele contém 15 horas de fala, sendo 6 de conteúdo expressivo, falado por uma atriz de voz profissional. Os estilos presentes no banco de dados são categorizados como “animado”, “acolhedor” e “ríspido”, e foram projetados para aplicações baseadas em serviços com foco em consumidores. Existe um total de 12400 enunciados neutros, 1307 animados, 1308 acolhedores, e 1256 ríspidos. Pra cada categoria, 90% das sentenças foram usadas para treinamento e 10% para validação e teste.

Para avaliar o desempenho do modelo, foi realizado um experimento perceptual no qual 30 participantes foram solicitados para ouvir e atribuir valores de naturalidade e expressividade de cada síntese. Especificamente, comparou-se o modelo proposto baseado em difusão, com aqueles que eram estado-da-arte na literatura: o VAE e o VAE+Flow. Para avaliar a naturalidade, uma frase do conjunto de teste foi sintetizada por cada modelo, que recebiam o mel-espectrograma correspondente como entrada de estilo, e então solicitou-se o julgamento de 0 a 100 o quão natural cada áudio soava. As médias do resultado são mostrado na Figura 3. Nele, observa-se que, enquanto que nos estilos acolhedor e neutro os desempenho são bastante similares, o modelo de difusão obtém melhor desempenho nos estilos ríspido e animado.

Para avaliar a expressividade, um experimento de preferência ABX foi conduzido, no qual cada um dos modelos recebia uma frase e uma referência de estilo também do conjunto de teste não correspondentes com intuito de realizar a transferência de estilos. Os participantes então escolhiam qual síntese tinha o estilo mais parecido com o da referência, para avaliar o quão bem o estilo foi transferido para o conteúdo textual. Nessa, comparou-se o de difusão com o VAE+Flow, com a opção de “ambos são igualmente parecidas” também inclusa. Os resultados são mostrado na Figura 4. No caso de transferência paralela, na qual o mel-espectrograma de estilo de referência tem o mesmo conteúdo que o texto de entrada, o modelo de difusão obteve melhor desempenho, enquanto no caso não paralelo (texto da referência de estilo diferente do texto de entrada) foi um pouco pior. Considerando o caso dos dois agregados, o modelo de difusão obteve 14.41% de preferência a mais que o VAE+Flow.

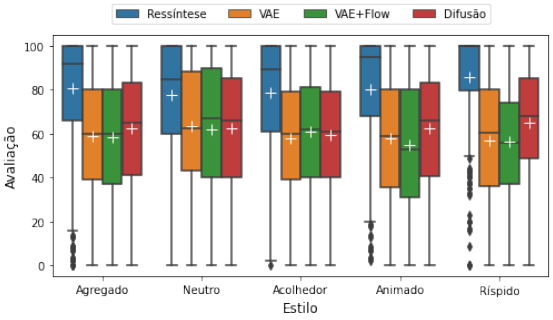

Figura 3. Experimento de naturalidade

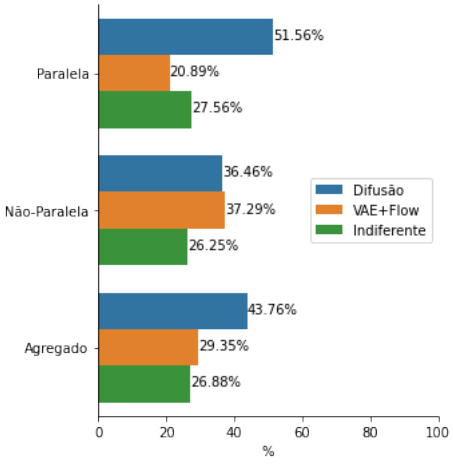

Figura 4. Experimento ABX de preferência de estilo

4. Conclusões

Foram investigados as principais técnicas para a modelagem de estilo baseado em modelos generativos não-supervisionados. Experimentando os modelos de difusão para gerar os vetores de estilos, observou-se uma melhoria na expressividade e naturalidades nos estilos mais energéticos: ríspido e animado. Nos estilos neutro e aco-

lhedor, os modelos obtiveram desempenhos similares. Para trabalhos futuros, busca-se fazer uso de técnicas supervisionadas e utilizar atributos prosódicos a fim de obter representações de estilo mais significativas para melhor condicionar o modelo acústico de texto-fala.

Agradecimentos

Os autores agradecem ao Centro de Pesquisa e Desenvolvimento (CPQD), em especial ao Flávio O. Simões, Mário Uliani Neto, Edson J. Nagle, Fernando O. Runstein, e Bianca Dal Bó, pelo apoio, disponibilização dos recursos e banco de dados; e ao Ministério da Ciência, Tecnologia e Inovações pelo apoio e financiamento deste projeto. Este trabalho é apoiado pelo BI0S - Instituto Brasileiro de Ciência de Dados, bolsa #2020/09838-0, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

Referências

[1] Vatsal Aggarwal, Marius Cotescu, Nishant Prateek, Jaime Lorenzo-Trueba, and Roberto Barra-Chicote. Using vaes and normalizing flows for one-shot text-to-speech synthesis of expressive speech. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6179–6183. IEEE, 2020.

[2] Xiaochun An, Frank K. Soong, and Lei Xie. Disentangling style and speaker attributes for tts style transfer. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 30:646–658, jan 2022.

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[4] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2015.

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[6] Adrian Łańcucki. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6588–6592. IEEE, 2021.

[7] Jingbei Li, Yi Meng, Chenyi Li, Zhiyong Wu, Helen Meng, Chao Weng, and Dan Su. Enhancing speaking styles in conversational text-to-speech synthesis with graph-based multi-modal context modeling. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7917–7921. IEEE, 2022.

[8] Yi Ren, Ming Lei, Zhiying Huang, Shiliang Zhang, Qian Chen, Zhijie Yan, and Zhou Zhao. Prosospeech: Enhancing prosody with quantized vector pre-training in text-to-speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7577–7581. IEEE, 2022.

[9] Manuel Sam Ribeiro, Julian Roth, Giulia Comini, Goeric Huybrechts, Adam Gabryś, and Jaime Lorenzo-Trueba. Cross-speaker style transfer for text-to-speech using data augmentation. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6797–6801, 2022.

[10] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J. Weiss, Rob Clark, and Rif A. Saurous. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron, 2018.

[11] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A. Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis, 2018.

[12] Ning-Qian Wu, Zhao-Ci Liu, and Zhen-Hua Ling. Discourse-level prosody modeling with a variational autoencoder for non-autoregressive expressive speech synthesis. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7592–7596, 2022.

[13] Fengyu Yang, Jian Luan, and Yujun Wang. Improving emotional speech synthesis by using sus-constrained vae and text encoder aggregation. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8302–8306, 2022.

[14] Yuanhao Yi, Lei He, Shifeng Pan, Xi Wang, and Yujia Xiao. Prosodyspeech: Towards advanced prosody model for neural text-to-speech. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7582–7586, 2022.

[15] Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling. Learning latent representations for style control and transfer in end-to-end speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6945–6949. IEEE, 2019.

Separando atributos de fala: conversão texto-fala expressiva entre locutores cruzados baseada na aprendizagem de representações

Lucas H. Ueda, Leonardo B.M.M. Marques , Paula Dornhofer Paro Costa
{l1156368@dac.unicamp.br, l218479@dac.unicamp.br, paulad@.unicamp.br}

Departamento de Engenharia de Computação e Automação (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Campinas, SP, Brasil

Resumo – A conversão texto-fala expressiva entre locutores cruzados consiste na transferência de um estilo de fala de um locutor para outro que nunca gravou falas com tal estilo. A efetividade nessa tarefa permite que seja possível transferir uma fala expressiva para locutores que só temos falas neutras em posse, amenizando assim a necessidade de gravação de dados expressivos de um novo locutor. O aprendizado de representações consiste em construir espaços onde os atributos de interesse são modelados, em particular, se os atributos são modelados independentemente torna-se possível condicioná-los de forma independente. O presente trabalho busca, através do aprendizado de representações, gerar espaços onde os atributos expressivos da fala (prosódia) e o timbre da voz sejam independentes, permitindo que um locutor neutro fale de forma expressiva sem nunca tê-las gravado.

Palavras-chave – síntese de fala, fala expressiva, modelagem de sequência, aprendizado de representações, prosódia, transferência de estilo.

1. Introdução

A fala sintetizada está presente cada vez mais no nosso cotidiano, das vozes gravadas na secretária eletrônica alguns anos atrás até os recentes assistentes virtuais controlados por voz. Os recentes avanços na área de aprendizado de máquina, em particular as redes neurais artificiais, possibilitaram que esses sistemas gerem voz artificial com qualidade próxima à fala humana [7, 8, 13].

No entanto, a comunicação oral humana não se baseia somente no conteúdo da mensagem a ser transmitida, mas também na forma como essa mensagem é realizada. Uma mesma frase pode ser lida de diferentes formas, alterando-se a entonação, o ritmo ou mesmo a emoção da fala. Essas diferentes formas de se enunciar uma mesma frase é o que chamamos de prosódia [1]. A prosódia está relacionada ao "como se fala" e não ao "o que se fala", e ela é responsável por não somente tornar uma frase mais interessante, como também auxilia na compreensão de seu conteúdo [2].

Diversos trabalhos propuseram formas de se incorporar a expressividade em sistemas de conversão texto-fala. Em [9], é proposto o *Reference Encoder*, um codificador de representação de estilos onde a prosódia de uma fala de referência é transferida para a fala sintética. Já em [10], uma base com diferentes estilos é utilizada, e o espaço gerado pelo *Reference Encoder* é analisado e utilizado para condicionar o estilo desejado na fala sintética. Mais recentemente, abordagens que modelam explicitamente componentes prosódicos vem sendo propostos, como o *FastPitch* [13], que propõe a modelagem da duração dos fonemas e da curva de frequência fundamental como parte da incorporação de expressividade no modelo.

Neste trabalho apresentamos resultados iniciais com foco na modelagem de representações de estilos. Utilizamos a arquitetura *FastPitch* como base do nosso modelo proposto, além disso, incorporamos o codificador de estilo *Reference Encoder* para gerar o espaço de estilos e assim condicionar o estilo alvo à fala sintetizada final. Adicionamos também componentes que buscam isolar os atributos modelados por cada componente do modelo final.

2. Base de dados

A base de dados utilizada consiste em cerca de 15 horas de áudios gravados por uma locutora brasileira, disponibilizada pela Fundação CPQD. Quatro estilos de fala foram gravados: animado, acolhedor, ríspido e neutro. O estilo animado se caracteriza por uma fala alegre, que transmita energia positiva. O acolhedor como uma fala calma, tranquila e compreensiva. O ríspido como alguém irritado e que cobra o interlocutor por algo. E por fim o neutro, somente com a leitura do texto desejado. A volumetria da base disponível é apresentada na Tabela 1.

As falas gravadas consistem em amostras de áudio em arquivos *wav* à 22KhZ, conjuntamente com a transcrição fonética falada na amostra.

Estilo	Horas aproximadas
Neutro	11
Animado	2
Acolhedor	2
Ríspido	2

Tabela 1. Volumetria aproximada da base de dados por estilos (em horas).

3. Modelo

Modelos de conversão texto-fala frequentemente são divididos em dois módulos principais, o modelo acústico e o vocoder. O modelo acústico é responsável por mapear o texto de entrada (ou sua transcrição fonética) no mel-espectrograma da fala sintetizada. Mel é uma escala perceptiva que divide o espectro em bandas para representar os tons como se fossem iguais em distância um do outro, levando em consideração que o ouvido humano não percebe frequências em uma escala linear [11]. Essa escala mapeia o espectro para que as variações tonais sejam percebidas linearmente pelos humanos. Para gerar de fato o sinal de fala é necessário mapear este mel-espectrograma em amostras de áudio e o vocoder é responsável por isso.

Nosso modelo acústico proposto (Figura 1) tem como base a arquitetura *FastPitch* proposto por [13].

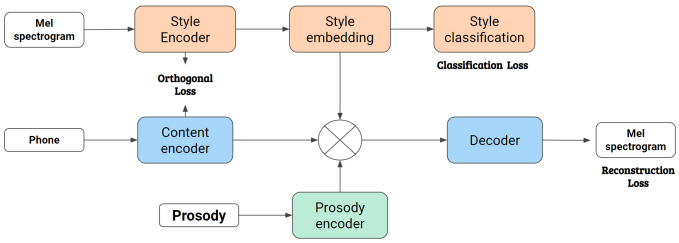

Figura 1. Arquitetura do modelo proposto.

Essa arquitetura consiste em uma camada de *encoder* que codifica os fonemas de entrada em representações densas de dimensão 384. Essas representações são então utilizadas por 3 módulos distintos, pelo preditor de duração, responsável por predizer a duração que cada fonema terá na fala sintetizada, pelo preditor de pitch, responsável por prever a curva de frequência fundamental da fala sintetizada, e por fim pelo *decoder*, que recebe a soma residual dessas representações e as predições dos demais módulos para predizer o mel-espectrograma da fala sintetizada. Adicionalmente, utilizamos o *Reference Encoder* proposto por [9], para gerar representações de estilos dos mel-espectrogramas de referência, onde tais representações são somadas a saída do *encoder* e utilizadas para condicionar os estilos na fala sintetizada. Como forma de gerar representações separáveis entre os diferentes estilos, uma camada de classificação é adicionada após o codificador de estilo, responsável por classificar os estilos a partir das representações geradas [12]. Além disso, como uma forma de induzir o modelo a não codificar o conteúdo da frase de entrada conjuntamente com o estilo, utilizamos uma função de perda que mede a ortogonalidade entre as representações originadas do codificador de estilo e do codificador de fonemas, similar ao proposto por [5].

Para condicionar um estilo específico na fala sintetizada condicionamos o *decoder* ao centroide das regiões de cada estilo no espaço gerado pelas representações, similar ao proposto por [4]. Por fim, para gerar as amostras de fala, utilizamos um vocoder capaz de mapear o mel-espectrograma em amostras de áudio, baseado na arquitetura *HiFi-GAN* [3].

3.1. Experimentos

Realizamos ao todo cinco experimentos (Tabela 2). As duas primeiras baseadas em técnicas mais simples, em uma criamos modelos especialistas para cada estilo baseado no ajuste fino das arquiteturas nos dados expressivos (Vanilla 1), e em outra utilizamos um mapeamento simples, conhecido como *look-up table*, para condicionar cada estilo a partir de camadas de embedding simples (Vanilla 2).

Nome	Descrição
Vanilla 1	Ajuste fino
Vanilla 2	Mapeamento de estilo por tabela
Baseline	Não supervisionado
Experimento 1	Proposto não balanceado
Experimento 2	Proposto balanceado

Tabela 2. Descrição dos experimentos realizados.

Para os demais experimentos, inicializamos as arquiteturas com pesos de um *FastPitch* pré-treinado, e realizamos os ajustes nas camadas adicionais que o modelo proposto possui por 200k iterações. Para uma primeira abordagem, adicionamos o *Reference Encoder* de forma simples, e sem supervisão no treinamento (Baseline). Já os Experimentos 1 e 2, consistem na arquitetura proposta apresentada, com a função de perda de ortogonalidade e classificação, treinadas em duas partições de dados diferentes, uma com os dados totais disponíveis, e outra com uma partição balanceada (mesma quantidade de horas para cada estilo).

4. Resultados

Para analisar as representações de estilos utilizamos a técnica *UMAP* [6] e projetamos a representação 2D delas. Para todos os experimentos a fala sintetizada final é inteligível e sem presença de ruídos.

O espaço gerado pela Baseline pode ser observado na Figura 2. Nota-se que para os diferentes estilos as representações se concentram em diferentes regiões do espaço. No entanto, é possível observar alguns pontos sobrepostos entre os estilos. É interessante observar que houve uma separação mais clara entre o estilo neutro (em verde) e os demais.

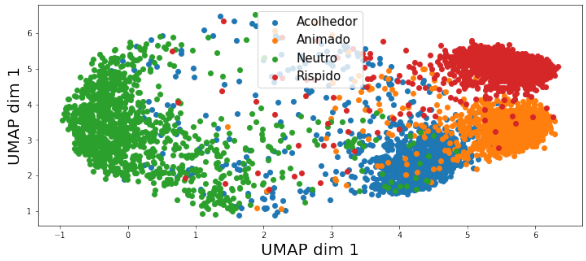

Figura 2. Espaço de estilos da Baseline.

Já para o Experimento 1 (Figura 3), nota-se que os diferentes estilos apresentam clusteres bem separados, com exceção de poucos pontos sobrepostos ao estilo neutro na região central da figura.

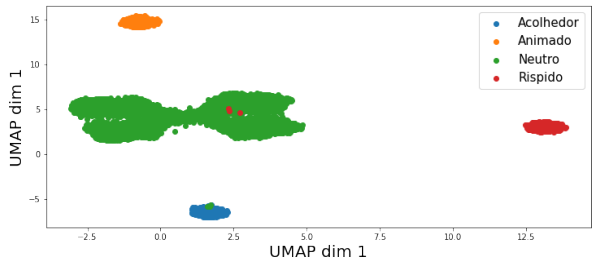

Figura 3. Espaço de estilos do Experimento 1.

Por fim, o uso de dados balanceados no Experimento 2 (Figura 4) modelou claramente quatro clusteres distintos para cada um dos estilos presentes na base.

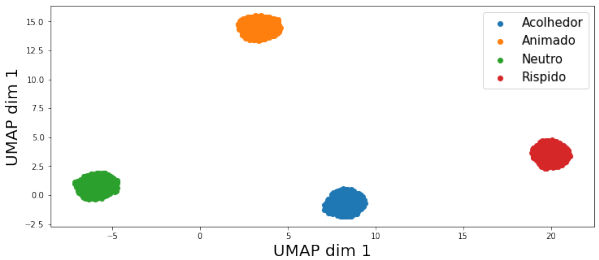

Figura 4. Espaço de estilos do Experimento 2.

Para todos os experimentos, a fala sintetizada final continua inteligível e com qualidade próxima a das falas gravadas. As abordagens baseadas em representações (Baseline, Experimentos 1 e 2), conseguem gerar modulações diferentes na fala final ao se condicionar a diferentes regiões do espaço de estilos, enquanto que nos dois experimentos Vanilla o condicionamento é único para cada estilo.

5. Conclusões

O problema de se modelar fala expressiva em arquiteturas de conversão texto-fala é complexo e possui diferentes abordagens na literatura. Neste trabalho propomos o uso de duas delas, a modelagem explícita de componentes prosódicos utilizando a arquitetura *FastPitch*, e o uso de um codificador de estilos (*Reference Encoder*). Além disso, propomos o uso de duas funções de perda para auxiliar as representações geradas pelo *Reference Encoder*, a de classificação e a de ortogonalidade entre o conteúdo e o estilo. Os resultados mostraram que de fato, o uso desses dois componentes adicionais auxiliam a arquitetura em gerar representações mais separadas entre os diferentes estilos. No entanto, para avaliar a capacidade de cada modelo gerar os estilos alvos desejados ainda é necessário uma avaliação perceptual.

5.1. Próximos passos

A importância das representações geradas serem desembaraçadas de outras informações é, particularmente, importante quando se deseja transferir as representações para outros locutores (transferência de estilo), dito isso, um próximo passo é realizar experimentos com uma base multi-locutor e testar se é possível transferir as representações de um estilo para um locutor que nunca gravou dados naquele estilo.

Agradecimentos

Os autores agradecem ao Centro de Pesquisa e Desenvolvimento (CPQD), em especial ao Flávio O. Simões, Mário Uliani Neto, Edson J. Nagle, Fernando O. Runstein, e Bianca Dal Bó, pelo apoio, disponibilização dos recursos e banco de dados; e ao Ministério da Ciência, Tecnologia e Inovações pelo apoio e financiamento deste projeto. Este trabalho é apoiado pelo BI0S - Instituto Brasileiro de Ciência de Dados, bolsa #2020/09838-0, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

Referências

[1] Plínio A. Barbosa. *Prosódia*. Parábola, May 2019.
[2] F. Eyben, S. Buchholz, N. Braunschweiler, J. Latorre, V. Wan, M. J. F. Gales, and K. Knill. Unsu-

pervised clustering of emotion and voice styles for expressive TTS. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4009–4012, March 2012. ISSN: 2379-190X.

[3] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17022–17033. Curran Associates, Inc., 2020.

[4] O. Kwon, I. Jang, C. Ahn, and H. Kang. An Effective Style Token Weight Control Technique for End-to-End Emotional Speech Synthesis. *IEEE Signal Processing Letters*, 26(9):1383–1387, September 2019. Conference Name: IEEE Signal Processing Letters.

[5] Tao Li, Xinsheng Wang, Qicong Xie, Zhichao Wang, and Lei Xie. Cross-speaker emotion disentangling and transfer for end-to-end speech synthesis, 2021.

[6] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [cs, stat]*, September 2020. arXiv: 1802.03426.

[7] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning. *arXiv:1710.07654 [cs, eess]*, February 2018. arXiv: 1710.07654.

[8] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783, 2018.

[9] R. J. Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J. Weiss, Rob Clark, and Rif A. Saurous. Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron. *arXiv:1803.09047 [cs, eess]*, March 2018. arXiv: 1803.09047.

[10] Alexander Sorin, Slava Shechtman, and Ron Hoory. Principal Style Components: Expressive Style Control and Cross-Speaker Transfer in Neural TTS. In *Interspeech 2020*, pages 3411–3415. ISCA, October 2020.

[11] S. S. Stevens, J. Volkmann, and E. B. Newman. A Scale for the Measurement of the Psychological Magnitude Pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190, January 1937. Publisher: Acoustical Society of America.

[12] Lucas H. Ueda, Paula D. P. Costa, Flavio O. Simoes, and Mário U. Neto. Are we truly modeling expressiveness? A study on expressive TTS in Brazilian Portuguese for real-life application styles. In *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, pages 84–89, 2021.

[13] Adrian Łańcucki. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6588–6592, 2021.

FEEC | UNICAMP
E A D C A 22
XIV ENCONTRO DE ALUNOS E DOCENTES DO DCA

Sessão Técnica 4

Extração Automática de Metadados de Imagens de Acervos Digitais de Museus Brasileiros

Vagner Inácio de Oliveira , Paula D. Paro Costa , Dalton Martins
{vagner.inol@gmail.com,paulad@unicamp.br,daltonmartins@unb.br}

Departamento de Engenharia de Computação e Automação (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Campinas, SP, Brasil

Abstract – A anotação de metadados de acervos digitais é realizada, tipicamente, por diferentes profissionais especializados, configurando uma atividade complexa, trabalhosa, que demanda grande quantidade de tempo, frequentemente sujeita a falhas humanas, altos custos e problemas na recuperação das informações de acordo com o desejado. Avanços recentes em inteligência artificial, particularmente técnicas de *Deep Learning*, têm mostrado seu potencial na realização de reconhecimentos visuais e na interpretação de objetos em imagens. Nesse contexto, o presente trabalho apresenta o EMA, um conjunto de dados de imagens oriundas do patrimônio cultural brasileiro com mais de 11.000 imagens rotuladas de objetos pertencentes a dezessete museus brasileiros, disponibilizadas pelo Projeto Tainacan. O conjunto de dados EMA é uma contribuição para o desenvolvimento de ferramentas de anotação de metadados automatizadas. Este projeto também apresenta resultados parciais da rede neural residual ResNet50 como *baseline* para o conjunto de dados, resultando em uma taxa de reconhecimento superior a 86%.

Keywords – Patrimônios Culturais Digitais, Tesauro, Anotação Automática, Deep Learning, Computer Vision

1. Introdução

As coleções digitais são uma maneira eficaz de possibilitar ao público a exploração do patrimônio cultural dos museus. Elas são particularmente relevantes em um país como o Brasil, onde os museus que preservam a história do país estão a milhares de quilômetros de distância, tornando-os inacessíveis à maioria das pessoas e difíceis de serem estudados por historiadores e pesquisadores em geral. Adiciona-se a este fator o risco de desastres, vide os ocorridos em menos de uma década com três museus brasileiros: Museu da Língua Portuguesa em São Paulo em 2015, o Museu Histórico Nacional no Rio de Janeiro em 2018 e, mais recentemente, o Museu de História Natural em Minas Gerais em 2020.

Apesar de todas as dificuldades enfrentadas pelos museus brasileiros, o país possui uma quantidade expressiva de acervos digitalizados. O Instituto Brasileiro de Museus (IBRAM) dá acesso pela internet a mais de 15.000 itens, de dezessete museus, em conjunto com seus respectivos metadados anotados com contexto histórico. A principal tecnologia da informação por trás disso é o Tainacan [3], uma plataforma de código aberto para criação de acervos digitais no WordPress, que também permite acesso programável ao banco de dados [1].

A anotação de metadados completa e confiável é fundamental para agregar significado às imagens do acervo digital de um museu. A imagem de um garfo, por exemplo, torna-se uma imagem irrelevante de um objeto se não for indicado que foi utilizado por algum personagem histórico durante um jantar onde foram tomadas grandes decisões ou que seu material representa todo um período histórico. Essa anotação de metadados é normalmente conduzida por vários profissionais especializados e é uma atividade complexa, trabalhosa e demorada, frequentemente levando a altos custos, falhas humanas e mal-entendidos. Com isso, inúmeros acervos digitalizados no Brasil e no mundo sofrem com a falta de informações de metadados, tornando os bens culturais pouco atrativos e seu potencial completo inexplorado.

Para enfrentar o problema, este trabalho propõe o uso de algoritmos de aprendizado de máquina, especificamente, modelos de *computer vision*, como ferramentas de auxílio para que profissionais especializados conduzam processos de anotação de metadados mais eficientes, confiáveis e potencialmente menos dispendiosos e descreve a construção de um conjunto de dados de imagens como um passo necessário para o desenvolvimento de ferramentas de anotação de metadados baseadas em IA para bens de patrimônios culturais.

2. Proposta

O principal objetivo deste projeto será a criação de uma base de dados confiável de imagens rotuladas para uma posterior aplicação no desenvolvimento de um modelo de extração automática de metadados. A validação do conjunto de dados será efetuada através de um modelo que

servirá como *baseline*.

3. Metodologia

O primeiro passo em nossa metodologia envolveu um estudo do acervo digital gerenciado pelo IBRAM integrado pelo Tainacan [3]. Coletamos os metadados de todos os objetos da coleção, que contém 15.651 objetos de dezessete museus ("JSON Metadata" na Figura 1).

Cada objeto do acervo é categorizado de acordo com seu tesauro. Um tesauro é definido como um conjunto de conceitos, denominados termos ou descritores, determinados de acordo com sua função ou estrutura, ordenados de forma clara e inequívoca, com base no estabelecimento de relações entre eles [2].

Como primeira abordagem do problema, focamos no tesauro mais frequente na coleção, "interior", correspondendo a 18,6% do total de itens. No contexto de patrimônio cultural, o termo refere-se a objetos da vida cotidiana usados no interior das casas, tal como um ferro a carvão usado para passar roupas quando não havia eletricidade. Também realizamos uma entrevista com uma museóloga do IBRAM que confirmou que muitos museus no Brasil se dedicam a mostrar como as pessoas viviam no passado, mostrando, por exemplo, como brancos e negros viviam na época da escravidão. Ela também enfatizou a relevância de desenvolver ferramentas automáticas ou semiautomáticas para ajudar os museólogos a gerar metadados para itens digitalizados.

Como segundo passo, analisamos qual campo de metadados poderia ser utilizado para rotular suas imagens correspondentes. Identificamos que os campos de metadados "título", "denominação", "tipo de material" e "técnica" são os que fornecem uma descrição geral de cada item. No entanto, verificamos que os campos "tipo de material" e "técnica" nem sempre eram preenchidos e que o campo "título" às vezes substituía uma descrição precisa por um alias que não descreve o objeto adequadamente. Por esse motivo, adotamos o campo "denominação" como o campo de destino para extrair os rótulos do nosso conjunto de dados.

Mais uma vez, nos deparamos com uma vasta quantidade de termos usados para descrever os objetos "interiores" da coleção, e decidimos analisar as palavras utilizadas com maior frequência para descrever os objetos. Como resultado dessa análise, decidimos manter apenas as 31 palavras mais frequentes como rótulos das imagens, se tornando estes, o nome das pastas no banco de dados de imagens, pois também facilita a sua utilização no modelo. Alguns exemplos de objetos "interiores" obtidos da coleção e seus respectivos rótulos podem ser vistos na Figura 2.

Figura 2. Exemplos de objetos pertencentes ao tesauro "interior". Seus rótulos são castiçal *(candlestick)*, panela (*pan*) e lampião (*kerosene lamp*).

4. Resultados

O principal resultado do presente trabalho é o conjunto de dados de imagens EMA, com 11.996 imagens, correspondentes a 2.922 objetos de dezessete museus brasileiros, rotulados de acordo com 31 classes. Como prova de conceito de uso do conjunto de dados EMA para treinar um modelo de *Deep Learning* (DL) para reconhecimento de objetos de "interiores" de patrimônios culturais, construímos um classificador de imagens que conta com a rede pré-treinada ResNet50. Adotando o método de aprendizagem por transferência, treinamos a camada final usando as imagens originais sem aumento de dados ou quaisquer transformações. Usamos 80% das imagens para treinar o modelo e as imagens restantes foram usadas para validação e testes. O modelo foi aplicado com fastai, uma biblioteca de DL de código aberto construída em cima do PyTorch.

A precisão de treinamento e validação ao final de 6 épocas foi de 86,7%. As classificações mais confusas estão resumidas na Tabela 3 e mostram as limitações de nossa metodologia. Por exemplo, os resultados dos quatro rótulos para identificar talheres: garfo, faca de mesa, colher e também talheres. Esses quatro rótulos resultaram em muitos erros de classificação, já que o rótulo dos talheres engloba garfo, faca e colher. Notamos também, por exemplo, a confusão entre as classes luminária e arandela. Uma luminária pode ter partes de uma arandela, por isso não é fácil resolver esse tipo de classificação.

5. Conclusões

A anotação de metadados em coleções digitais é uma tarefa desafiadora. Os problemas típicos incluem falta de informação e classificações erradas, principalmente devido a diferenças significativas entre objetos modernos e seus equivalentes no passado. Esses problemas podem causar problemas de recuperação de dados ou associar um item ao contexto errado, dificultando o acesso ao conhecimento que o objeto pode oferecer.

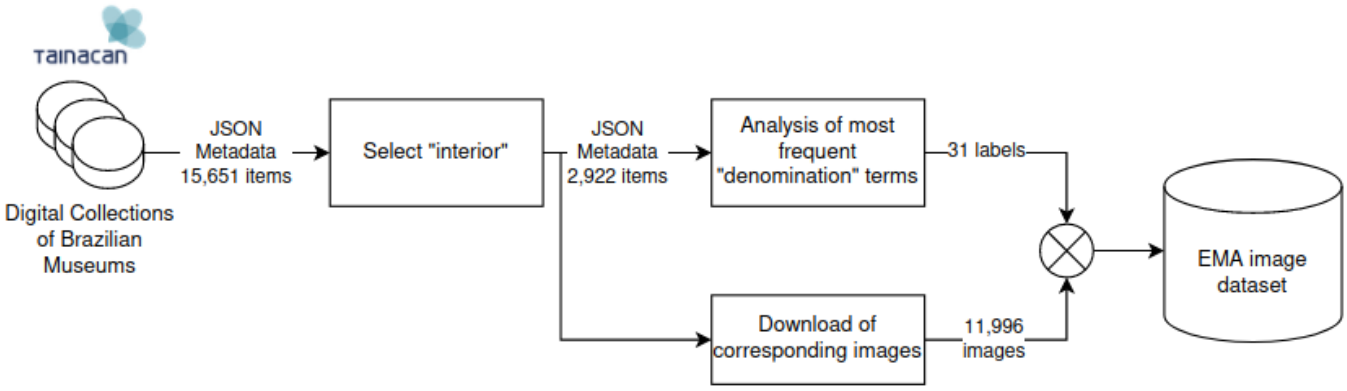

Figura 1. Passos para a construção do conjunto de dados EMA.

Actual	Predicted	occurrences
spoon	cutlery	58
table knife	cutlery	44
cutlery	spoon	34
cutlery	table knife	32
sideboard	curtain	31
luminaire	sconce	26
sconce	luminaire	17
cutlery	fork	16
curtain	sideboard	15
fork	cutlery	11
bed	jug	7
fork	spoon	5
dish	spoon	4
cup of tea	spoon	3
mirror	chest	3
saucer	cup of tea	3
sideboard	table	3
chest	mirror	2
cup	spoon	2
cup of tea	saucer	2
luminaire	table	2
table	luminaire	2
table	sideboard	2
chest of drawers	chest	1
chest of drawers	fork	1
glass	spoon	1
lamp	jug	1
luminaire	table knife	1
table knife	spoon	1

Figura 3. Confusões com maior frequência

Neste projeto, apresentamos nossos primeiros passos para o desenvolvimento de ferramentas de anotação de metadados baseadas em IA para ajudar os museólogos a melhorar a qualidade geral da anotação de coleções digitais. Em particular, apresentamos o EMA, um conjunto de dados de imagens rotuladas com mais de 11.000 imagens de objetos históricos encontrados em dezessete museus brasileiros. O código implementado para executar todas as etapas de processamento e classificação descritas neste trabalho e as instruções para solicitação do conjunto de dados estão disponíveis no repositório do projeto [4].

Também apresentamos resultados de um modelo *baseline* para este conjunto de dados por meio de um treinamento do modelo ResNet50. Nosso modelo conseguiu obter 86,7% de precisão no reconhecimento de objetos, mostrando a consistência do conjunto de dados e o potencial dessa abordagem.

Trabalhos futuros incluem explorar o desempenho de outras arquiteturas de DL e aumentar o conjunto de dados com outras coleções de patrimônios culturais para uma generalização do modelo. Também planejamos desenvolver um aplicativo que indicará rótulos durante os processos de anotação.

Agradecimentos

Agradecemos a Amanda Oliveira, do IBRAM, pelas valiosas informações que forneceu sobre sua experiência nos museus brasileiros. Agradecemos também à comunidade Tainacan pelo apoio.

Referências

[1] Governo Federal. Instituto brasileiro de museus - ibram, 2022.

[2] Helena Dodd Ferrez. Tesauro de objetos do patrimônio cultural nos museus brasileiros. *Rio de Janeiro: Fazer Arte. Gerência de Museus da Secretaria Municipal de Cultura*, 2016.

[3] University of Brasília. Tainacan, 2021.

[4] P. D. P. Costa e D. L. Martins V. De Oliveira. "ema's project repository", 2022.

Controle de qualidade automatizado em espectroscopia por ressonância magnética através da rede neural ART Fuzzy e algoritmo k-means

Gabriel Dias , Thays Abreu , Simone Appenzeller , Sergio Dertkigil , Letícia Rittner
{g172441@dac.unicamp.br, thays@unicamp.br, appenzel@unicamp.br, sergiosj@unicamp.br, lrittner@unicamp.br}

Departamento de Engenharia de Computação e Automação (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Campinas, SP, Brasil

Resumo – A Espectroscopia por Ressonância Magnética (ERM), frequentemente, produz espectros com ruído e artefatos que podem induzir erros de quantificação ou interpretação que impactam o uso clínico dos dados espectroscópicos. Neste trabalho foi proposto um modelo de clusterização para controle de qualidade em ERM usando a rede neural ART Fuzzy e o algoritmo k-Means. Utilizaram-se 22048 espectros de aquisições multivoxel provenientes do Hospital de Clínicas da Unicamp. O método proposto utiliza uma abordagem hierárquica: primeiro, a rede neural ART Fuzzy busca clusterizar espectros que apresentam grande variação para filtrar sinais válidos. A partir deste cluster, faz-se uma nova clusterização usando k-means de forma a obter padrões de boa qualidade. Foram realizadas comparações quantitativas dos clusters obtidos com métricas tradicionais usadas para avaliar a qualidade dos espectros mostraram que, aproximadamente, 93% atendem ao critério métrico SNR, 85% ao critério FWHM, 100% ao CRLB e 81% considerando os três critérios simultaneamente.

Palavras-chave – Espectroscopia por ressonância magnética, Controle de qualidade, Aprendizado de máquina, Rede neural ART fuzzy, k-means

1. Introdução

A Espectroscopia por Ressonância Magnética (ERM) viabiliza a identificação e quantificação de neurometabólitos para monitorar alterações metabólicas cerebrais de forma não invasiva. Essas informações podem dar suporte ao diagnóstico e tratamento de condições neurológicas tais como epilepsia, Alzheimer e tumores no cérebro [9]. No entanto, os espectros, sinais no domínio da frequência produzidos pela ERM, costumam ser gerados com artefatos que podem induzir erros de quantificação ou interpretação, reduzindo o valor clínico da técnica. Alguns dos fatores que comprometem a qualidade do espectro são correntes parasitas, baixa supressão da influência da água, contaminação lípidica e volume parcial [3]. Quando não cientes desses fenômenos, médicos podem facilmente tirar conclusões incorretas de mapas metabólicos de baixa qualidade, pois os softwares tradicionais de quantificação tais como LCModel [13], TARQUIN [14] e Osprey [10] não verificam a qualidade dos espectros a serem analisados de forma automatizada. Ao mesmo tempo, a inspeção manual é afetada pela subjetividade humana e muito demorada para o método de aquisição multivoxel [11].

Existem métricas (medidas de qualidade), tais como Relação Sinal Ruído (SNR - do inglês *Signal to Noise Ratio*), Largura a meia altura (FWHM - do inglês *Full Width at Half Maximum*) e limite inferior de Cramér-Rao (CRBL - do inglês *Cramér Rao Lower Bound*) que são utilizadas para controle de qualidade de espectro [12]. No entanto, essas métricas precisam de valores limiares para definir se um espectro é ruidoso ou não, sendo que não há um consenso entre os especialistas sobre esses valores e eles podem variar de acordo com a aplicação [6].

Abordagens que usam aprendizado de máquina devem desempenhar um papel cada vez mais importante na criação de métodos automatizados rápidos e confiáveis para ERM [6]. Até o momento, existem poucos trabalhos para controle de qualidade em ERM [11, 2, 4, 7].

Dentre os trabalhos mais promissores na identificação de espectros de boa qualidade, Pedrosa de Barros et al. [11] utilizou um algoritmo supervisionado baseado em floresta aleatória que obteve uma área abaixo da curva (AUC - do inglês *Area Under the Curve*) de 0.976. Utilizou-se um conjunto de 9756 dados multivoxel rotulados por especialistas que foram submetidos a um processo de engenharia de atributos. Extraíram-se atributos do sinal no domínio do tempo e da frequência. Alguns deles foram derivados diretamente dos dados brutos, como a magnitude de um dado ponto ou de métricas estatísticas como média, desvio padrão, curtose e distorção. O modelo foi implementado em um módulo de extensão do software tradicional JMRUI [8], mas é necessário que seja investigada sua aplicação para scanners e formas de aquisição que não foram utilizados pelos autores no estudo.

Gurbani et al. [2] propôs um modelo supervisionado

através de uma arquitetura de rede neural convolucional profunda. Foram utilizados um conjunto de 8894 espectros adquiridos pela metodologia multivoxel que foram rotulados por especialistas. A rede pôde ser treinada sem engenharia de atributos, apenas com os pontos da curva espectral. Atingiu-se uma AUC de 0.951 com a modelagem, que também precisa ser avaliada em outros contextos de aquisição, afim de que seja analisada sua capacidade de generalização.

2. Proposta

O trabalho aqui apresentado propôs, pela primeira vez, um método hierárquico baseado em dois modelos não-supervisionados para a tratativa do controle de qualidade de ERM. Diferentemente de outras propostas que utilizaram engenharia de atributos para modelagem [7, 4, 11], este trabalho empregou apenas os pontos da curva espectral. Também, a modelagem aqui apresentada destaca-se pela quantidade de 22048 espectros disponíveis para estudo, uma quantidade consideravelmente maior do que em outros trabalhos [2, 11].

O método foi baseado na rede neural artificial ART Fuzzy [1] e o algoritmo k-means [5]. A rede neural ART Fuzzy pertence à familia de redes ART (do inglês *Adaptative Ressonance Theory*), possui treinamento não supervisionado e engloba em sua arquitetura cálculos baseados na lógica nebulosa, tendo como principais caracteristicas a estabilidade e a platicidade. O algoritmo k-means é um algoritmo numérico não supervisionado, não determinístico e iterativo de clusterização, sendo amplamente utilizado em diversos problemas devido a sua eficácia, simplicidade e rapidez.

2.1. Conjunto de Dados

O conjunto de dados foi composto por 106 aquisições multivoxel na região do corpo caloso de voluntários saudáveis e pacientes do Hospital de Clínicas (HC - UNICAMP). O protocolo da aquisição incluiu: imagens ponderadas em T1; Sequência de pulso 2D (PRESS); ângulo de excitação de 90º; TE longo: 144ms e TR: 2000ms e as varreduras foram realizadas com um scanner Philips 3T. Cada aquisição foi constituída por uma malha de espectros de 16 linhas por 13 colunas, totalizando 22048 espectros utilizados nesse estudo. Todos os participantes do estudo assinaram o Termo de Consentimento Livre e Esclarecido (TCLE).

2.2. Pré-Processamento

Os espectros advindos do scanner constituem uma curva com 1024 pontos. Para a modelagem, foi selecionada a região de frequência com os principais neurometabólitos que são quantificados [15], resultando em espectros com 330 pontos (Fig. 1). Utilizou-se a normalização Min-Max [16] destes dados.

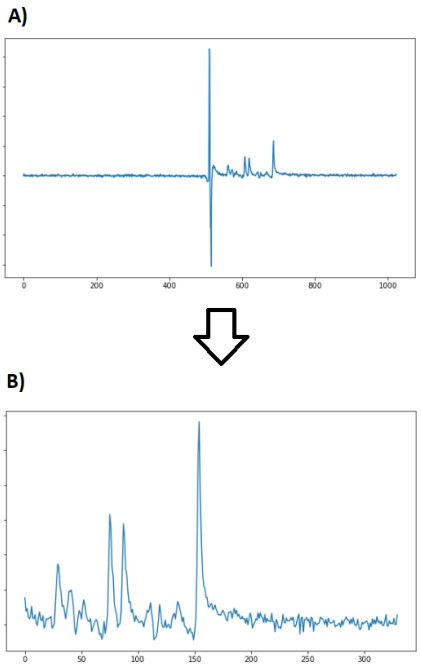

Figura 1. Em (A) espectro bruto com 1024 pontos. (B) região metabólica selecionada com 330 pontos.

2.3. Treinamento

O método hierárquico proposto é composto de 2 etapas: a clusterização pela rede neural ART Fuzzy, para descartar espectros de baixo sinal; e a clusterização dos espectros selecionados na primeira etapa, buscando encontrar o cluster com espectros de boa qualidade (Fig.2).

Para a calibração do modelo ART Fuzzy, o parâmetro de escolha (α) foi sempre fixado no valor de 0.0001. Variou-se a quantidade de épocas, parâmetro de vigilância (ρ) e parâmetro de treinamento (β) para ser feita a validação do modelo em diversos casos. Aplicou-se uma varredura baseada em *Grid Search* para os parâmetros de vigilância e de treinamento, de 0.3 à 0.9 com um passo de 0.5, e para a quantidade de épocas, de 1 até 10. Para uma quantidade de épocas igual a 6, α = 0.0001, ρ = 0.7 e β = 0.7, obteve-se a clusterização desejada (Fig. 3).

Em seguida, a partir do grupo separado pela rede neural ART Fuzzy (Fig. 3A), fez-se uma nova clusterização usando o algoritmo k-means de forma a obter um grupo de espectros que apresentasse padrões de boa qualidade. Apenas altos valores do número de classes k geraram grupos com distinções significativas. Atingiu-se o resultado almejado para um valor de k = 15 (Fig. 4).

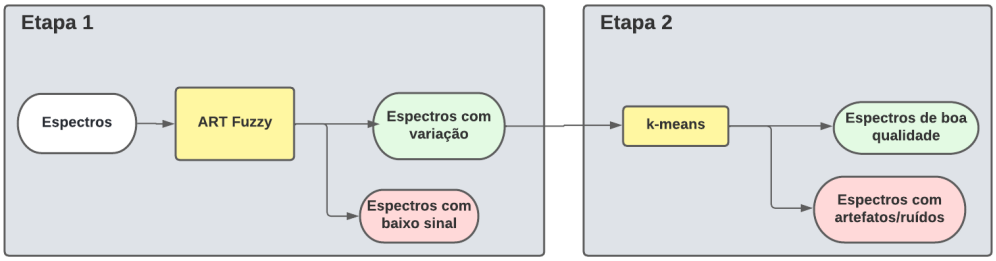

Figura 2. Fluxograma da abordagem hierárquica proposta. Na primeira Etapa, a rede neural ART Fuzzy busca clusterizar espectros que apresentam variação para que os espectros de baixo sinal sejam desconsiderados.Em seguida, encaminha-se para a segunda Etapa com uma nova clusterização usando k-means de forma a obter um grupo com padrões de boa qualidade.

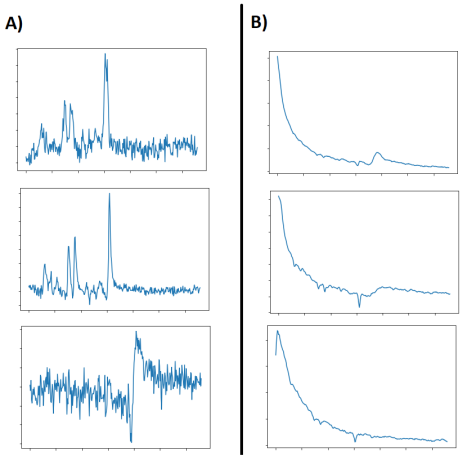

Figura 3. Resultado da clusterização por ART Fuzzy: (A) exemplos de espectros que pertencem ao grupo considerado de interesse para a próxima etapa do método; (B) exemplos com pouca variação tidos como espectros de baixo sinal.

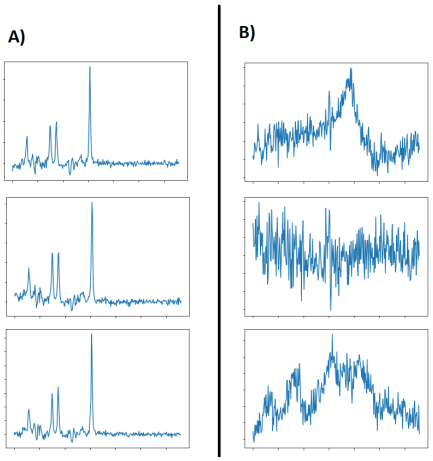

Figura 4. Clusterização obtida após k-means: (A) exemplos de espectros que pertencem ao grupo de boa qualidade desejado; (B) exemplos de espectros ruidosos e com artefatos que foram clusterizados nos demais grupos.

3. Resultados

Os agrupamentos finais obtidos pelo método foram avaliados qualitativamente (Fig. 5). Comparações quantitativas com métricas tradicionais usadas para avaliar a qualidade dos espectros de ERM também foram conduzidas. Para essa análise, utilizaram-se os espectros classificados como de boa qualidade pelo modelo. Os valores típicos de limiar utilizado para as métricas tradicionais são: SNR>10; FWHM<10; CRLB<50 [12]. Aproximadamente, 93% dos espectros contemplaram os critérios métricos SNR, 85% para o FWHM, 100% para o CRLB e 81% considerando os três critérios simultaneamente.

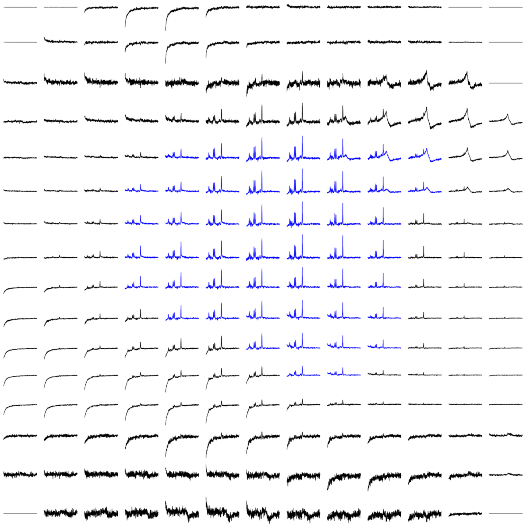

Figura 5. Clusterização resultante em uma grade espectral multivoxel de um determinado paciente: espectros tidos como de boa qualidade pelo modelo final (em azul); espectros considerados ruidosos ou com artefato (em preto)

4. Conclusões

O modelo proposto foi capaz de encontrar padrões de boa qualidade nos espectros obtidos a partir de aquisições multivoxel. Trata-se de um modelo rápido, facil-

mente implementável e consistente com as métricas tradicionais de qualidade usadas. Em trabalhos futuros, serão investigados outros protocolos de aquisição multivoxel para os dados espectroscópicos. Também, outras técnicas de aprendizado de máquina serão exploradas com análises sobre os espectros considerados corrompidos cujos artefatos poderiam ser corrigidos por meio de técnicas existentes. Dessa forma, o método proposto por esse projeto de pesquisa instiga a obtenção de uma ferramenta automatizada e consistente para atuar no controle de qualidade de ERM, garantindo maior valor clínico para a técnica.

Agradecimentos

Gabriel Dias agradece ao Scholarship Program da DeepMind (`deepmind.com/scholarships`) pela bolsa de estudos concedida que possibilita dedicação integral ao programa de pós-graduação. Thays Abreu agradece à bolsa CAPES.

Referências

[1] Gail A Carpenter, Stephen Grossberg, and David B Rosen. Fuzzy art: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, 4(6):759–771, 1991.

[2] Saumya S Gurbani, Eduard Schreibmann, Andrew A Maudsley, James Scott Cordova, Brian J Soher, Harish Poptani, Gaurav Verma, Peter B Barker, Hyunsuk Shim, and Lee AD Cooper. A convolutional neural network to filter artifacts in spectroscopic mri. *Magnetic Resonance in Medicine*, 80(5):1765–1775, 2018.

[3] Roland Kreis. Issues of spectral quality in clinical 1h-magnetic resonance spectroscopy and a gallery of artifacts. *NMR in Biomedicine*, 17(6):361–381, 2004.

[4] Sreenath P Kyathanahally, Victor Mocioiu, Nuno Pedrosa de Barros, Johannes Slotboom, Alan J Wright, Margarida Julià-Sapé, Carles Arús, and Roland Kreis. Quality of clinical brain tumor mr spectra judged by humans and machine learning tools. *Magnetic Resonance in Medicine*, 79(5):2500–2510, 2018.

[5] J MacQueen. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297, 1967.

[6] Andrew A Maudsley, Ovidiu C Andronesi, Peter B Barker, Alberto Bizzi, Wolfgang Bogner, Anke Henning, Sarah J Nelson, Stefan Posse, Dikoma C Shungu, and Brian J Soher. Advanced magnetic resonance spectroscopic neuroimaging: Experts' consensus recommendations. *NMR in Biomedicine*, 34(5):e4309, 2021.

[7] Bjoern H Menze, B Michael Kelm, Marc-André Weber, Peter Bachert, and Fred A Hamprecht. Mimicking the human expert: pattern recognition for an automated assessment of data quality in mr spectroscopic images. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 59(6):1457–1466, 2008.

[8] A Naressi, Chantal Couturier, I Castang, R De Beer, and Danielle Graveron-Demilly. Java-based graphical user interface for mrui, a software package for quantitation of in vivo/medical magnetic resonance spectroscopy signals. *Computers in Biology and Medicine*, 31(4):269–286, 2001.

[9] Sarah J Nelson. Multivoxel magnetic resonance spectroscopy of brain tumors. *Molecular Cancer Therapeutics*, 2(5):497–507, 2003.

[10] Georg Oeltzschner, Helge J Zöllner, Steve CN Hui, Mark Mikkelsen, Muhammad G Saleh, Sofie Tapper, and Richard AE Edden. Osprey: Open-source processing, reconstruction & estimation of magnetic resonance spectroscopy data. *Journal of Neuroscience Methods*, 343:108827, 2020.

[11] Nuno Pedrosa de Barros, Richard McKinley, Urspeter Knecht, Roland Wiest, and Johannes Slotboom. Automatic quality control in clinical 1h mrsi of brain cancer. *NMR in Biomedicine*, 29(5):563–575, 2016.

[12] Danilo Pereira, Larissa Ganaha, Simone Appenzeller, and Leticia Rittner. Open-source toolbox for analysis and spectra quality control of magnetic resonance spectroscopic imaging. In *Medical Imaging 2021: Biomedical Applications in Molecular, Structural, and Functional Imaging*, volume 11600, pages 64–71. SPIE, 2021.

[13] Stephen W Provencher. Estimation of metabolite concentrations from localized in vivo proton nmr spectra. *Magnetic Resonance in Medicine*, 30(6):672–679, 1993.

[14] Greg Reynolds, Martin Wilson, Andrew Peet, and Theodoros N Arvanitis. An algorithm for the automated quantitation of metabolites in in vitro nmr signals. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 56(6):1211–1219, 2006.

[15] Mohamed Saber and Frank Gaillard. MR spectroscopy, May 2008.

[16] Wikipedia. Feature scaling. `http://en.wikipedia.org/wiki/Feature_scaling`, 2013. Online; accessed 17 August 2022.

Segmentação automática dos lobos pulmonares em imagens de CT utilizando U-Net: comparando abordagens 2D e 2D-estendida

Jean Antonio Ribeiro , Leticia Rittner
{j265739@dac.unicamp.br, lrittner@unicamp.br}

Departamento de Engenharia de Computação e Automação (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Campinas, SP, Brasil

Resumo – O desenvolvimento de algoritmos eficientes e robustos em ferramentas automatizadas para segmentação pulmonar e seus lobos é fundamental para o diagnóstico e acompanhamento de doenças pulmonares. Este trabalho apresenta uma abordagem para segmentação automática de lobos pulmonares utilizando redes neurais profundas em imagens de CT. Para isso, foram utilizadas duas abordagens em conjunto com a rede U-Net: 2D e 2D-estendida. Na abordagem 2D, a rede é treinada fatia por fatia utilizando convolução 2D e, na predição, os resultados são empilhados para formar a segmentação volumétrica. Na abordagem 2D-estendida, utiliza-se também convolução 2D, mas são colocadas na entrada da rede 3 fatias (como canais), para reter parte da informação volumétrica. Testes utilizando um dataset público, que inclui 50 imagens com anotações, para cada um dos lobos pulmonares, mostraram que a abordagem 2D-estendida pode alcançar uma média de *Dice* superior a 91%, sendo ligeiramente superior à abordagem 2D.

Palavras-chave – segmentação automática, lobos pulmonares, fissuras, rede neural convolucional, U-Net, tomografia computadorizada

1. Introdução

O pulmão humano é dividido em cinco lobos que são separados pelas fissuras lobares. O pulmão direito tem três lobos: inferior, superior e médio. O pulmão esquerdo tem somente dois lobos: superior e inferior. Os sistemas brônquico e vascular são isolados com conexões mínimas entre os lobos, e os lobos são considerados unidades com funções relativamente independentes. Assim, as doenças pulmonares podem ser limitadas a um lobo individual.

Geralmente, imagens de tomografia computadorizada (CT) são utilizadas para identificar anormalidades ou pequenas lesões pulmonares, dificilmente vistas em imagens radiográficas. No entanto, a análise dessas imagens não é simples, principalmente quando o radiologista precisa de uma solução rápida. A quantidade de fatias geradas durante o exame por paciente é grande, podendo conter centenas de fatias, cada uma das quais precisa ser analisada pelo radiologista. Este processo é demorado, difícil de ser realizado e envolve algum grau de variabilidade entre observadores.

Além disso, os limites lobares, definidos pelas fissuras pulmonares, muitas vezes são parcialmente invisíveis nas imagens de CT [3]. Em pulmões saudáveis, as bordas dos lobos são definidas por fissuras visíveis, que muitas das vezes podem estar incompletas, dificultando a identificação correta dos limites lobares [13].

Alterações morfológicas em lobos específicos também podem ocorrer durante o progresso patológico de doenças pulmonares [3]. Doenças como fibrose ou enfisema podem obscurecer a forma e a aparência das fissuras lobares. Em pacientes com COVID-19, as imagens de CT apresentam sombras em vidro fosco e os lobos são preenchidos com líquido pleural, tornando invisíveis partes das fissuras lobares [16].

Na segmentação de lobos pulmonares, George et al. [4] desenvolveram um algoritmo utilizando redes neurais profundas, para identificar os limites lobares, e *random walker* (RW), para gerar as segmentações finais. O treinamento e a inferência foram realizados em fatias axiais 2D. Em relação a U-Net 3D, a U-Net 2D possui a vantagem que não utilizar, nas imagens de entradas, a profundidade existente nos volumes de CT. Por sua vez, Imran et al. (2020) [8] combinaram características da rede V-Net [5] e redes progressivas aninhadas holisticamente [6]. Para evitar o *overfitting*, foram utilizadas somente fatias axiais, em que pelo menos um lobo do pulmão está presente.

Visto que a localização e distribuição da doença pulmonar é um fator significativo na determinação de um tratamento adequado, muitos trabalhos tem sido desenvolvidos com o objetivo de alcançar uma segmentação adequada dos lobos pulmonares. Do ponto de vista técnico, a segmentação precisa do lobo pulmonar pode melhorar as tarefas clínicas subsequentes, incluindo a previsão de malignidade do nódulo, avaliação e quantificação de doenças pulmonares, reduzindo o espaço de busca para lobos mais propensos a serem afetados [3].

2. Proposta

A seguir, são descritos os datasets que foram utilizados nos experimentos. Também é descrita a implementação das etapas de pré-processamento, treinamento e predição nas abordagens 2D e 2D-estendida.

2.1. Dataset

Foram utilizados quatro datasets: Medical Segmentation [10] (9 volumes), VESSEL12 [14] (23 volumes), MOSMED [11] (172 volumes) e Tang, Zhang and Xie [15] (50 volumes). Nos 3 primeiros datasets, as anotações foram geradas de forma automática pelo *framework lungmask* [7], consideradas neste trabalho como padrão prata (*silver standard*). O único dataset com padrão ouro (*gold standard*) disponível é o de Tang, Zhang and Xie [15], que segundo os autores, foram geradas por radiologistas.

2.2. Método

Primeiramente, todas as imagens de CT foram normalizadas para valores da Unidade Hounsfield (HU) no intervalo de [-1024, 600] e depois normalizadas para um intervalo entre 0 e 1. O otimizador utilizado foi Adam [9], com *learning rate* de 5×10^{-4} e batches de 6 imagens. O modelo foi implementado utilizando a biblioteca de código aberto PyTorch Lightning. Todos os experimentos foram executados com 60 épocas. O modelo foi treinado e avaliado nas vistas axiais, cada uma com tamanho de 512×512. Com o objetivo de reduzir a quantidade de memória utilizada na GPU e ajudar na generalização do modelo, um *RandomCrop* de $64 \times 256 \times 256$ foi realizado em cada uma das fatias, em ambas as abordagens 2D e 2D-estendida. Para otimizar os pesos do modelo, durante o treinamento, foi utilizada a função de custo *Dice*.

Foram desenvolvidas duas abordagens para a segmentação dos lobos pulmonares em imagens de CT: 2D e 2D-estendida. Para isso, foi utilizada a rede U-Net, que é uma rede amplamente utilizada na segmentação de imagens médicas [3]. A rede é treinada fatia por fatia utilizando convolução 2D, contudo, na abordagem 2D-estendida, são colocadas na entrada da rede 3 fatias, como se fossem três canais, obedecendo a uma determinada regra de espaçamento. A ideia de utilizar 3 fatias consecutivas, como entrada da rede, foi proposta inicialmente por Pereira et al. [12], com o intuito de trazer a informação volumétrica para a U-Net 2D.

Na abordagem 2D-estendida, três variações de espaçamento foram implementadas: sem espaçamento (*consec.*), a cada 2 fatias (*salta 1*) e a cada 3 fatias (*salta 2*). Na variação do tipo *consec.*, as imagens são formadas pelas fatias $n-1$, n e $n+1$; na variação do tipo *salta 1*, as imagens são formadas pelas fatias $n-2$, n e $n+2$; e, finalmente, na variação *salta 2*, as imagens são formadas pelas fatias $n-3$, n e $n+3$. Na abordagem 2D, cada imagem é gerada por cada uma das fatias que formam o volume 3D (Fig. 1).

Na etapa de predição, as fatias são empilhadas para formar a segmentação volumétrica. Em seguida, o desempenho do volume é por meio do Coeficiente *Dice* [2]. A saída da rede é composta de seis canais: fundo, lobo direito superior (RUL), lobo direito do meio (RML), lobo direito inferior (RLL), lobo esquerdo superior (LUL) e lobo esquerdo inferior (LLL). O conjunto de treinamento é composto por todos os volumes dos datasets com padrão-prata (Medical Segmentation, VESSEL12 e MOSMED). O conjunto de teste é composto de 40 volumes do dataset de Tang, Zhang and Xie [15]. Os 10 volumes restantes, deste último dataset, foram separados para utilização nos experimentos com *finetuning*.

3. Experimentos e Resultados

A segmentação do lobo pulmonar foi avaliada por meio da métrica *Dice Similarity Coefficient* (*Dice*). O *Dice* é amplamente utilizado para avaliar a similaridade entre duas amostras [2].

Na avaliação dos resultados, foram realizados 2 conjuntos de experimentos. No primeiro conjunto, o modelo foi treinado somente com máscaras padrão-prata. A predição do modelo foi comparada às máscaras padrão-ouro, no conjunto de testes. No segundo conjunto, o modelo também foi treinado com as máscaras padrão-prata e foi realizado o *finetuning* dos pesos, utilizando-se os 10 volumes com padrão-ouro, separados para este fim. O objetivo aqui foi verificar se, na falta de um conjunto grande com anotações manuais, o *finetuning* de um modelo pré-treinado com máscaras geradas automaticamente é capaz de melhorar a performance do modelo. Dentro de cada conjunto de experimentos (com e sem *finetuning*), foram comparadas as abordagens 2D e 2D-estendida, esta última variando-se o espaçamento das fatias de entrada.

Na abordagem utilizando a rede U-Net 2D-estendida sem *finetuning*, os valores médio de *Dice* foram superiores a 90%, sendo 95,3% o maior deles, correspondente a abordagem *consec.*. Na abordagem *consec.* com *finetuning*, os resultados foram melhores do que nas abordagens com espaçamento *salta 1* e *salta 2*. Em geral, nas abordagens utilizando 2D-estendida, os resultados foram ligeiramente melhores do que na abordagem 2D. Por causa da anatomia do pulmão e da pouca visibilidade apresentada pela fissura horizontal direita em imagens de CT, segmentar o lobo superior direito é mais difícil do que os outros lobos [1]. Consequentemente os valores de *Dice*, desse lobo, foram inferiores aos demais (Tabela 1).

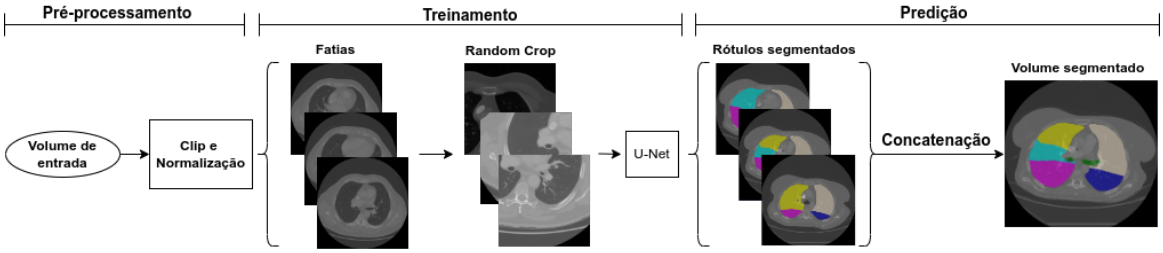

Figura 1. Fases do algoritmo implementado utilizando a rede U-Net 2D. No caso da abordagem 2D-estendida, cada fatia representa uma imagem com três fatias.

Finetuning	Espaçamento	*Dice*	LUL	LLL	RUL	RML	RLL
Não	*consec.*	0.909±0,045	0.958±0,031	0.955±0,026	0.903±0,071	0.778±0,128	0.950±0,024
	salta 1	0.920±0,042	0.961±0,032	0.955±0,026	0.914±0,068	0.817±0,125	0.951±0,023
	salta 2	0.915±0,044	0.961±0,028	0.955±0,023	0.908±0,074	0.804±0,124	0.945±0,025
	2D	0.908±0,043	0.957±0,028	0.954±0,024	0.899±0,072	0.794±0,120	0.935±0,027
Sim	*consec.*	0.896±0,051	0.955±0,038	0.953±0,031	0.895±0,085	0.739±0,136	0.938±0,037
	salta 1	0.873±0,068	0.946±0,042	0.940±0,035	0.863±0,097	0.695±0,197	0.922±0,049
	salta 2	0.863±0,065	0.943±0,044	0.938±0,036	0.844±0,098	0.676±0,171	0.915±0,055
	2D	0.891±0,061	0.949±0,038	0.945±0,036	0.884±0,083	0.741±0,176	0.937±0,036

Tabela 1. Média dos valores de *Dice* obtidos no dataset de teste de Tang, Zhang and Xie [15]. O treinamento do modelo foi realizado com e sem *finetuning*. Para cada lobo, é mostrado o valor médio de *Dice* correspondente a variação de espaçamento entre fatias que é utilizado na abordagem 2D-espandida.

Na Figura 2, é mostrada uma imagem de CT, a anotação *Ground Truth* (GT), correspondente para cada um dos lobos pulmonares, e uma imagem de predição. Na imagem de predição, é possível observar que alguns voxels foram segmentados de forma incorreta, quando comparados com a anotação GT. Cada tom de cor, na anotação GT e na imagem de predição, representa um lobo do pulmão. Geralmente, muitas das segmentações incorretas são causadas por fissuras incompletas que estão presentes nas imagens de CT, por exemplo, originadas na ocorrência de lesões patológicas próximas às bordas dos lobos.

4. Conclusões

Este trabalho apresentou uma abordagem para segmentação automática de lobos pulmonares utilizando redes neurais profundas em imagens de CT. Para isso, foram utilizadas duas abordagens em conjunto com a rede U-Net: 2D e 2D-estendida. A qualidade das segmentações produzidas foi avaliada por meio do Coeficiente *Dice*. Os valores de *Dice* alcançados pelo modelo ainda são baixos, quando comparados com o estado da arte. Especificamente, o baixo valor de *Dice*, apresentado pela segmentação do lobo superior direito, indica que a presente abordagem precisa ser melhorada. Nas segmentações produzidas pela abordagem 2D-estendida *consec.*, o valor médio de *Dice* foi superior a 91% sem *finetuning* e 89% utilizando *finetuning*. Na abordagem 2D, o valor médio de *Dice* foi de 90,8% e 89.1%, nas abordagens com e sem *finetuning*, respectivamente. Neste caso, a utilização de *finetuning* ajudou na generalização da rede. Em comparação com a abordagem 2D, a abordagem 2D-estendida foi ligeiramente superior.

Baseado nos resultados experimentais, percebe-se que esta abordagem é uma boa base para projetos futuros, envolvendo a segmentação dos lobos pulmonares. Entretanto, as segmentações finais apresentam ruídos e precisam de ajustes. Uma solução, para este problema, seria a utilização de outras redes, juntamente com técnicas de regularização e argumentação. Trabalhos futuros incluem treinar o modelo em um dataset maior e avaliar outros recursos, tanto computacionais quanto metodológicos, com o objetivo de melhorar a precisão da segmentação em diferentes datasets. Esse dataset deverá conter imagens sintéticas e de CT que representem uma ampla variedade de doenças pulmonares, como COVID-19 e câncer de pulmão. Devido à diferença nos tamanhos dos lobos, uma função de custo que utilize uma média ponderada por lobos deverá ser desenvolvida, com o objetivo de melhorar a qualidade das segmentações produzidas.

Agradecimentos

Este projeto foi parcialmente apoiado pelas bolsas da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) 88887.513444/2020-00.

Figura 2. Demonstração de uma imagem de CT, anotação GT e imagem de predição.

Referências

[1] P. Cronin, B. Gross, A. Kelly, S. Patel, E. Kazerooni, and R. Carlos. Normal and accessory fissures of the lung: Evaluation with contiguous volumetric thin-section multidetector CT. *Eur. J. Radiol.*, 75:e1–8, 11 2009.

[2] L. R. Dice. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302, 1945.

[3] T. Doel, D. J. Gavaghan, and V. Grau. Review of automatic pulmonary lobe segmentation methods from CT. *Comput. Med. Imaging Graphics*, 40:13–29, 2015.

[4] K. George, A. Harrison, D. Jin, Z. Xu, and D. Mollura. Pathological Pulmonary Lobe Segmentation from CT Images Using Progressive Holistically Nested Neural Networks and Random Walker. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 195–203, Cham, 09 2017. Springer.

[5] E. Gibson, F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. Davidson, S. P. Pereira, M. J. Clarkson, and Dean C. Barratt. Automatic Multi-Organ Segmentation on Abdominal CT With Dense V-Networks. *IEEE Trans. Med. Imaging.*, 37(8):1822–1834, 2018.

[6] A. P. Harrison, Z. Xu, K. George, Le Lu, R. M. Summers, and D. J. Mollura. Progressive and Multi-path Holistically Nested Neural Networks for Pathological Lung Segmentation from CT Images. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, pages 621–629, Cham, 2017. Springer.

[7] J. Hofmanninger, F. Prayer, J. Pan, S. Röhrich, H. Prosch, and G. Langs. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *Eur. Radiol. Exp.*, 4:50, 08 2020.

[8] A. Imran, A. Hatamizadeh, S. P. Ananth, X. Ding, N. Tajbakhsh, and D. Terzopoulos. Fast and automatic segmentation of pulmonary lobes from chest CT using a progressive dense V-network. *Comput. Methods Biomech. Biomed. Eng.: Imaging Visualization*, 8(5):509–518, 2020.

[9] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA, USA, 05 2015.

[10] MedSeg. COVID-19 CT segmentation dataset. urlhttp://medicalsegmentation.com/covid19, 2020. acessado em 10/08/2022.

[11] S. P. Morozov, A. E. Andreychenko, N. A. Pavlov, A. V. Vladzymyrskyy, N. V. Ledikhova, V. A. Gombolevskiy, I. A. Blokhin, P. B. Gelezhe, A. V. Gonchar, and V. Yu. Chernina. MosMedData: Chest CT Scans With COVID-19 Related Findings Dataset, 2020.

[12] M. Pereira, I. Fantini, R. Lotufo, and L. Rittner. An extended-2d cnn for multiclass alzheimer's disease diagnosis through structural mri. In *Medical Imaging 2020: Computer-Aided Diagnosis*, volume 11314, pages 438–444. SPIE, 2020.

[13] B. N. Raasch, E. W. Carsky, E. J. Lane, J. P. O'Callaghan, and E. R. Heitzman. Radiographic anatomy of the interlobar fissures: a study of 100 specimens. *AJR.*, 138:1043–1049, 06 1982.

[14] R. D. Rudyanto, S. Kerkstra, and E. M. [van Rikxoort] et al. Comparing algorithms for automated vessel segmentation in computed tomography scans of the lung: the VESSEL12 study. *Med. Image Anal.*, 18(7):1217–1232, 2014.

[15] H. Tang, C. Zhang, and X. Xie. Automatic Pulmonary Lobe Segmentation Using Deep Learning. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1225–1228, Venice, Italy, 03 2019. IEEE.

[16] S. Zheng, W. Nie, L. Pan, B. Zheng, Z. Shen, L. Huang, C. Pei, Y. She, and L. Chen. A dual-attention V-network for pulmonary lobe segmentation in CT scans. *IET Image Proc.*, 15(8):1644–1654, 2021.

Análise Comparativa de Métodos de Aprendizado de Máquina para Equalização

Luan Lopes Fontes, Romis Attux
l239682@dac.unicamp.br, attux@dca.fee.unicamp.br

Departamento de Engenharia de Computação e Automação (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Campinas, SP, Brasil

Resumo – Em sistemas de comunicação digital, os efeitos do meio de transmissão são fatores limitantes para o desempenho atingível. Uma estratégia para mitigar tais efeitos é lançar mão de um filtro que recebe o nome de equalizador. Embora a teoria clássica de equalização seja construída sobre estruturas lineares, filtros não-lineares têm recebido crescente atenção nas últimas décadas tanto pela necessidade de expandir os sistemas de comunicação quanto pela maior disponibilidade de hardware de alto desempenho. Neste trabalho, buscaremos realizar uma análise comparativa abrangendo estruturas lineares e não-lineares de equalização no contexto de um leque diversificado de modelos de canal.

Palavras-chave – Aprendizado de máquina, equalização de canais.

1. Introdução

A equalização não-linear de canais de comunicação baseada em algoritmos de aprendizado de máquina [1] é um assunto de grande interesse devido ao crescimento exponencial na demanda por taxas de transmissão e robustez que caracteriza os sistemas de comunicação modernos [2].

A área de aprendizado de máquina tem experimentado, ao longo da última década, um enorme desenvolvimento, o que se deve ao crescimento explosivo na produção de dados, ao maior acesso a hardware paralelo e aos aportes teóricos trazidos pelos estudos na área de *deep learning*. Nesse contexto, deve-se analisar qual pode ser o real impacto dessa "revolução" nas tarefas de processamento de sinais ligadas à comunicação. O problema de equalização, por sua generalidade [1] e importância histórica, é uma opção natural e importante, e abordá-lo é a tarefa central deste projeto.

Neste trabalho, faremos uma análise comparativa envolvendo equalizadores lineares e não-lineares no contexto de modelos lineares e não-lineares de canal. Na seção 2, faremos uma breve apresentação do problema a ser resolvido; na seção 3, discutimos os testes e resultados, e, na seção 4, expomos nossas conclusões.

2. Comunicação e Equalização

Um sistema de comunicação digital (SCD) é, essencialmente, composto por um transmissor, um canal, e um receptor. As informações a serem transmitidas são codificadas em um alfabeto, isto é, um conjunto finito de símbolos. Neste trabalho, considera-se um modelo discreto equivalente em banda base do SCD, o qual facilita a análise sem perda de generalidade [3].

O sinal gerado $s(k)$ é um processo estocástico com amostras discretas, independentes e identicamente distribuídas (i.i.d.) pertencentes a um alfabeto finito. O tipo de modulação empregada no SCD é o que define o alfabeto - por exemplo, uma modulação 2-PAM (ou 2-PSK, em banda base) dá origem ao alfabeto {-1, +1}, que é o empregado neste trabalho.

2.1. Modelo de Canal

Ao passar por um canal, o sinal transmitido sofre distorções, o que ocasiona perda de informação. A modelagem de um canal, isto é, sua representação matemática, leva em conta, via de regra, a interferência intersimbólica (IIS), o ruído e eventuais distorções não-lineares [1].

No intuito de reverter os efeitos introduzidos pelo canal de comunicação, favorecendo a recuperação do sinal original, utiliza-se um filtro denominado equalizador. A Fig. 1 é um esquema simplificado de SCD com equalização:

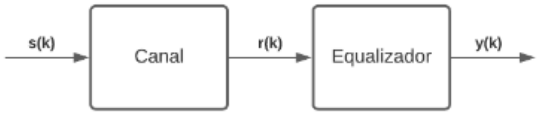

Figura 1 – Esquema Simplificado de SCD

A recuperação da informação transmitida pelo equalizador pode ocorrer de dois modos: estimação de sequência ou símbolo a símbolo. Na estimação de sequência, um conjunto de amostras recebidas é usado para se obter, através do critério de máxima verossimilhança, o sinal original [4]. Este tipo de equalizador pode ser implementado por meio do algoritmo de Viterbi [5].

Por outro lado, a metodologia de estimação de símbolo a símbolo usa um número fixo de amostras recebidas para estimar um único símbolo a cada período de amostragem. Equalizadores que utilizam essa técnica permitem que seus parâmetros sejam ajustados ao longo do tempo para se adequarem a variações no canal, merecendo o epíteto de *adaptativos*. O equalizador de memória finita símbolo a símbolo ótimo é dado pelo critério de Máxima Probabilidade A Posteriori (MAP - Máximo A Posteriori), e é chamado de equalizador bayesiano.

Os parâmetros dos equalizadores adaptativos são determinados através de algoritmos de treinamento, de modo a atender um certo critério de desempenho. Nesse contexto, o treinamento pode ser supervisionado ou não-supervisionado. No treinamento supervisionado, um *sinal de treinamento* ou *sinal piloto* é enviado periodicamente durante a transmissão das informações. O receptor possui uma cópia desse sinal piloto, que é utilizada para adaptar os parâmetros do equalizador durante o período de treinamento - desse modo, durante esse período, não existe transmissão efetiva de informação.

O treinamento *não-supervisionado*, também conhecido como *cego*, é caracterizado pela não existência de um sinal piloto, sendo a adaptação do equalizador feita com base exclusivamente no sinal recebido e em estatísticas do sinal transmitido [3]. Neste trabalho, enfocaremos o problema supervisionado apenas.

Os equalizadores adaptativos podem ser implementados por meio de filtros i.e. estruturas lineares ou não-lineares. Existe uma vasta bibliografia sobre a aplicação de filtros lineares na equalização de canais, com destaque para os filtros FIR (do inglês *Finite Impulse Response*). Este modelo é atraente pela sua baixa complexidade computacional e seu amplo arcabouço matemático, edificado sobre a teoria de sistemas lineares.

Por outro lado, os recentes avanços tecnológicos têm viabilizado e incentivado a utilização e estudo de estruturas não-lineares e, assim, computacionalmente mais complexas, como os filtros polinomiais [1] e as redes neurais artificiais [2,3], sendo essas o alvo de estudo neste trabalho. Dessa forma, aplicaremos tanto filtros lineares quanto não-lineares, de modo a explicitar a diferença de desempenho entre os dois métodos e os casos em que cada um é mais adequado.

3. Resultados

Considerou-se um modelo de canal com IIS e ruído. A SNR (*Signal-to-Noise Ratio*), que mede nível de ruído do canal, variou de 6 a 20 dB. As simulações contaram com 100.000 amostras e foram realizadas para o canal $H_1(z) = 1 + 0{,}6z^{-1}$, bem como para uma extensão não-linear do mesmo, da forma $x_p(k) = x(k) + 0.3x^3(k)$.

Foram utilizados um filtro linear com resposta ao impulso finita e uma rede neural do tipo perceptron de múltiplas camadas (MLP, do inglês multilayer perceptron). Em todos os testes, o equalizador bayesiano foi implementado como uma referência de desempenho.

As amostras geradas foram divididas em conjuntos de treinamento e teste, na proporção 80%-20%. O filtro linear foi treinado num esquema de regressão logística com entropia cruzada, e a MLP com esse mesmo critério e com o critério de erro quadrático médio.

3.1 - Canais Lineares

Primeiramente, consideramos o canal $H_1(z) = 1 + 0{,}6z^{-1}$ com dois atrasos de equalização: d = 0 e d = 2. Consideramos um equalizador com duas entradas em todos os casos.

Para o caso de atraso nulo, a distribuição dos dados, *x(k) x x(k-1)*, ocorre segundo o que mostra a Fig. 2.

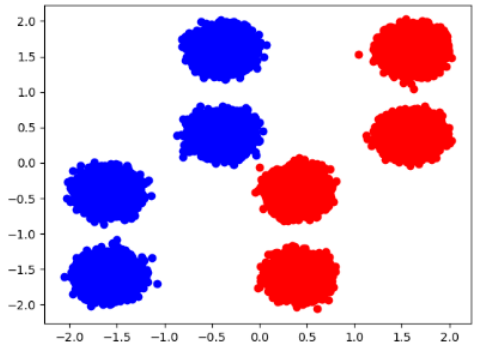

Figura 2 – Distribuição dos Dados para Atraso Nulo (Azul – Classe -1 / Vermelho – Classe +1)

Percebe-se que as classes são, em essência, linearmente separáveis. Por esse motivo, espera-se que o equalizador linear tenha bom desempenho, apenas ligeiramente inferior ao de equalizadores não-lineares. Na Fig. 3, as curvas de BER do equalizador linear, de duas implementações de uma MLP e do equalizador de Bayes mostram isso.

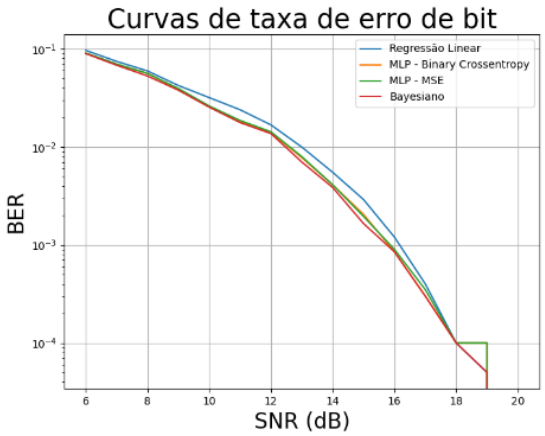

Figura 3 – Curvas de Taxa de Erro de Bit (BER) – Atraso Nulo

Por outro lado, quando o atraso é igual a dois, os estados não são linearmente separáveis, como mostra a Fig. 4.

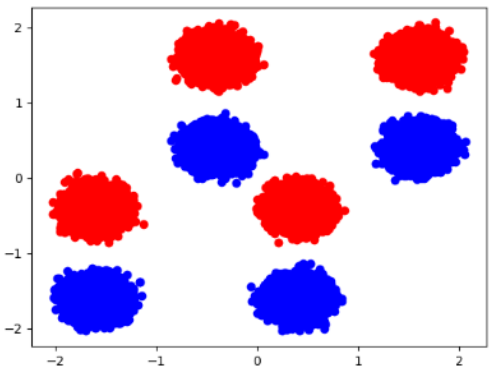

Figura 4 - Distribuição dos Dados para Atraso Dois (Azul – Classe -1 / Vermelho – Classe +1)

Nesse caso, a estrutura linear não será capaz de separar os estados, o que leva às curvas de BER da Fig. 5.

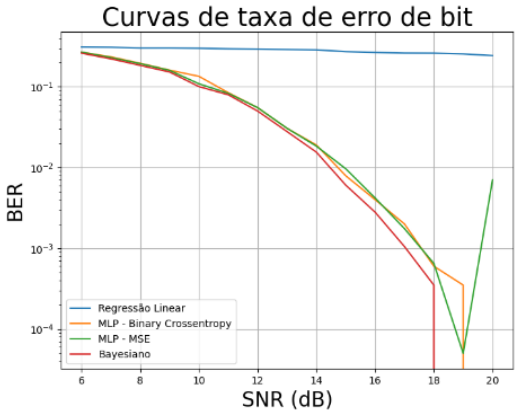

Figura 5 – Curvas de Taxa de Erro de Bit (BER) – Atraso Dois

Para as SNRs mais altas, na Fig. 3 e, especialmente, na Fig. 5, nota-se uma distorção que é causada, provavelmente, pela dificuldade de estimar a probabilidade de erro ou por algum problema de convergência da rede numa simulação.

3.2 - Canal Não-Linear

Por fim, consideramos um canal não-linear simples. Será considerada a saída do canal linear com função de transferência H1(z), $y_l(k)$, e um termo relativo a seu cubo, levando à seguinte expressão para a saída:

$$y(k) = y_l(k) + 0{,}3\, y_l^3(k) \quad (1)$$

A distribuição dos dados para esse canal e atraso igual a dois é mostrado na Fig. 6.

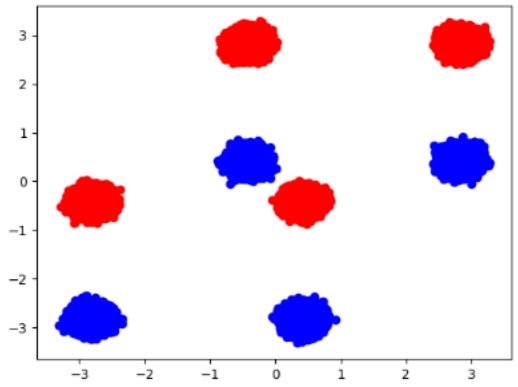

Figura 6 - Distribuição dos Dados para Atraso Dois (Azul – Classe -1 / Vermelho – Classe +1) – Canal Não-Linear

Neste caso, a distorção não-linear não é particularmente marcante, o que leva a curvas de BER similares às mostradas na Fig. 5. É isso que nos mostra a Fig. 7.

4 – Conclusões

Neste trabalho, que contempla parte dos resultados obtidos numa iniciação científica financiada pela FAPESP (proc. 2021/01684-7), fez-se uma análise comparativa inicial de estruturas lineares e não-lineares de equalização. Os resultados obtidos foram condizentes com as expectativas associadas a equalizadores lineares e não-lineares para cada espécie de canal. O resultado do equalizador de Bayes serviu sempre como referência.

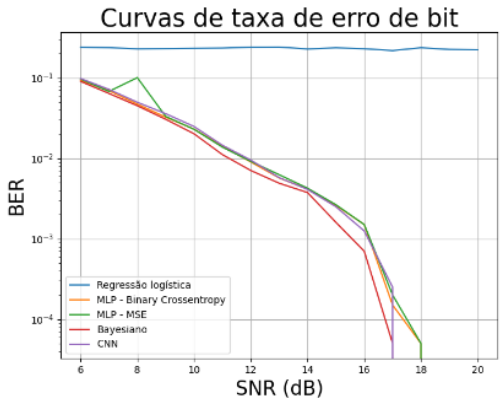

Figura 7 – Curvas de Taxa de Erro de Bit (BER) – Atraso Dois – Canal Não-Linear

Agradecimentos

Os autores agradecem à FAPESP (proc. 2021/01684-7) e ao CNPq (proc. 308811/2019-4) o apoio financeiro.

Referências

[1] J. M. T. Romano, R. R. F. Attux, C. C. Cavalcante, R. Suyama, *Unsupervised Signal Processing: Channel Equalization and Source Separation*, CRC Press, 2010.

[2] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press, 2016.

[3] S. Haykin, Adaptive Filter Theory, Prentice Hall, 1996.

[4] G. Forney, "Maximum-Likelihood Sequence Estimation of Digital Sequences in the Presence of Intersymbol Interference", IEEE Transactions on Information Theory, Vol. 18, No. 3, pp. 363 – 378, 1972.

[5] G. Forney, "The Viterbi Algorithm", Proceedings of the IEEE, Vol. 61, No. 3, pp. 268 – 278, 1973.

FEEC | UNICAMP
EADCA22
XIV ENCONTRO DE ALUNOS E DOCENTES DO DCA

Sessão Técnica 5

FastSLAM 1.0 aplicado em um cadeira de rodas inteligente.

César Bastos da Silva, Vinicius Emanoel Ares, Felipe Augusto Oliveira Mota,
Victor Fermán, Arnaud de Jarcy, Eric Rohmer
{cesar.silva2612@gmail.com, rohmer@unicamp.br}

Departamento de Engenharia de Computação e Automação (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Campinas, SP, Brasil

Resumo – Nos dias de hoje, é evidente o crescimento na quantidade de indivíduos que apresentam algum grau de dificuldade em habilidades como, enxergar e caminhar. Visando minimizar o impacto imposto por tais deficiências, há um aumento na necessidade de novas tecnologias assistivas. Dentre elas, é possível destacar a cadeira de rodas inteligente, a qual visa aumentar a mobilidade funcional de pessoas com deficiência tanto nos membros inferiores, como também superiores. Para gerar um equipamento que proporcione cada vez mais autonomia ao indivíduo, a transformação da cadeira em um sistema robótico autônomo que consiga mapear se localizar é essencial, tornando-a similar a um robô de serviço. Visto isso, este trabalho propõe aplicar o FastSlam 1.0, utilizando recursos naturais detectados por um Lidar 2D e sua odometria. O sistema será testado no ambiente de simulação CoppeliaSim, com integração via ROS, visando apresentar os resultados parciais do algoritmo implementado.

Palavras-chave – Coloque aqui palavras-chave que as pessoas que procuram seu artigo podem usar ou que podem ajudar os comitês de revisão ou editores a atribuí-lo aos revisores apropriados.

1. Introdução

Nos dias atuais, aproximadamente 200 milhões de pessoas experimentam dificuldades funcionais consideráveis e há uma expectativa de crescimento para os próximos anos, o que mostra há necessidade em criar tecnologias facilitadoras para pessoas que possuam deficiências que atrapalham seu conforto e qualidade de vida [8]. Esta área da pesquisa é conhecida como tecnologia assistiva, responsável por estudar novos meios de aumentar a mobilidade e condições para pessoas com deficiência [12].

Um exemplo que vem sendo estudado é a cadeira de rodas motorizada, que quando utilizada em conjunto com técnicas de robótica se torna inteligente, tendo seu funcionamento bastante similar a robôs de serviço[9]. Há duas categorias que relacionam o grau de autonomia da cadeira de rodas inteligente, alterando o tipo de interação e controle provindo do usuário, a semiautônoma e a autônoma. Um dos principais pilares para o desenvolvimento de um sistema autônomo consiste no desenvolvimento de um sistema de mapeamento e localização (SLAM), sendo utilizado desde aos mais simples robôs, como robô aspirador, aos mais complexos como carros autônomos [2].

A área de SLAM já vem sendo estudada a algumas décadas, sendo possível encontrar diferentes técnicas que resolvem este problema, se diferindo inicialmente aos sensores utilizados para detecção do seu entorno e do próprio veículo. No caso, para alcançar uma autonomia completa do veículo, apenas sensores *on board* devem ser utilizados[4]. O método mais simples é a partir da aquisição da velocidade das rodas do veículo, aliadas ao modelo cinemático é possível prever um valor para a pose do veículo, entretanto, este método apresenta um alto erro, principalmente para longas trajetórias e com o escorregamento das rodas [1].

Por outro lado, diferentes sensores exteroceptivos são utilizados, como câmera e Lidar, são os mais utilizados para localização e robôs móveis. Esses sensores junto de algoritmos específicos, buscam características e pontos de referência (*landmark*) que possam ser utilizados para ajudar a localização do robô no mundo. Dentre os algoritmos encontrados na literatura é possível destacar o *Split–and–merge* e Incremental para lidar [7], ORB e SURF *features* para câmera [10].

Dentre os métodos mais clássicos de localização, é possível destacar o *Extended Kalman Filter* (EKF) SLAM, Monte-Carlo, FastSLAM 1.0 e o *Extended Information Filter* (EIF) SLAM. O EKF–SLAM se baseia na utilização de uma equação diferencial estocástica e linear, sendo necessário linearizar a média e covariâncias das relações não lineares no sistema, seu principal limitação é relacionada ao sistema ter um número alto de *landmarks*, podendo deixar o tempo de execução elevado. Já o método de Monte–Carlo utiliza do filtro de partículas para determinar a pose do robô sem a necessidade de conhecimento da pose inicial, realizando etapas de reamostragem a cada iteração [11].

O FastSLAM 1.0 propõe a utilização de um filtro de partículas junto com filtros de Kalman individuais para cada *landmark*, ganhando em velocidade de processamento. Por fim, apresenta–se o *Extended Information Filter* (EIF) SLAM, que processa a informação ao longo de toda a trajetória para realizar a estimação da pose [5].

Este trabalho propõe apresentar os resultados ainda preliminares de uma simulação de localização e mapeamento baseado em um lidar com 180°, utilizando o algoritmo *split-and-merge* junto do FastSLAM 1.0, visto que historicamente estes se mostra mais eficiente dentre os apresentado. Um robô genérico móvel representa a cadeira de rodas em um ambiente simulado no CoppeliaSim.

2. Metodologia

Para realizar a implementação do presente trabalho serão necessários os seguintes itens:

- Ambiente Simulado;
- Passar as informações via ROS;
- Modelo cinemático do veículo;
- Detecção de *landmarks*, *Split-and-Merge*;
- Implementação do FastSLAM.

2.1. Ambiente Simulado:

Para realizar a simulação será utilizado o simulador de robôs CoppeliaSim, um *framework* versátil, muito utilizado nas área de robóticas, possuindo uma API que permite conexão direta via ROS. O robô utilizado será um Pioneer 3–DX(Figura 1, um veículo a ser utilizado de forma genérica.

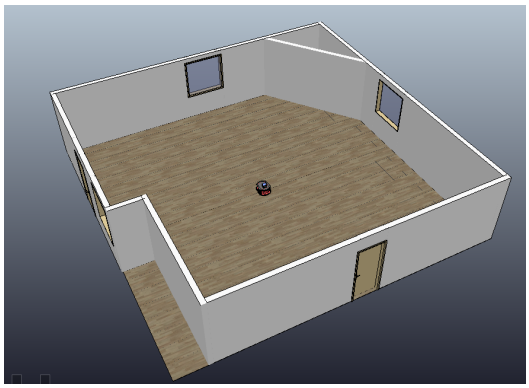

Figura 1. Pioneer 3-DX na simulação.

2.2. Modelo cinemático:

Esta etapa é essencial para a etapa de predição e para gerar um sinal de odometria para o sistema. A seguir serão apresentadas as Eq. 1 e 2, explicitando a relação da velocidade das rodas do veículo para estimar a posição do robô [11].

$$V = \frac{V_R + V_L}{2} \tag{1}$$

$$\omega = \frac{V_R - V_L}{2b} \tag{2}$$

Onde:

V_R : Velocidade linear da roda direita;

V_L : Velocidade linear da roda esquerda;

V : Velocidade linear do robô;

b : Distância do centro de rotação até a roda.

Dividindo então as velocidades linear e angular pelo tempo de amostragem (discretização pelo método de Euler), é possível definir o quanto o veículo se movimentou (Δs) e qual a sua variação angular ($\Delta\theta$). Com isso encontra-se a função apresentada na Eq. 3, que descreve o movimento do robô aproximado de $[x \;\; y \;\; \theta]^T$ para $[x' \;\; y' \;\; \theta']^T$.

$$\begin{bmatrix} x' \\ y' \\ \theta' \end{bmatrix} = \begin{bmatrix} x \\ y \\ \theta \end{bmatrix} + \begin{bmatrix} \Delta s cos(\theta + \frac{\Delta\theta}{2}) \\ \Delta s sen(\theta + \frac{\Delta\theta}{2}) \\ \Delta\theta \end{bmatrix} \tag{3}$$

2.3. *Split-and-Merge*

Um dos métodos bastante consolidados para extração de linhas retas a partir de um conjunto de pontos bidimensionais, com base apenas na geometria básica. Estes dados nesta aplicação, são as distâncias lidas pelo lidar.

Seu algoritmo começa com a fase de *SPLIT*, onde é formada uma reta entre o primeiro e último ponto do grupo de dados, procurando o ponto mais distante da reta, maior que um determinado ϵ, que será utilizado na próxima fase e então repete-se o cálculo anterior, até que todos pontos do subgrupo criado possuam distância menor que ϵ, chegando então a fase de *MERGE*, onde se encontra a reta desejada, o último ponto dessa reta será utilizado para uma nova fase de *SPLIT*. Isso é realizado, até que todos pontos estejam em algum subgrupo.

2.4. FastSLAM

A proposta apresentada em [5], conhecido por ser computacionalmente eficiente, lidando bem com diferentes dados, devido a amostragem do tipo Monte–Carlo e com não linearidades de modelos, sendo necessária a linearização. Busca decompor o problema do SLAM em

dois fatores, a localização do robô e a estimação dos *landmarks* que são condicionados pela pose do robô. Cada *landmark* é modelado independente, descrito por uma série de distribuições, aplicando um Filtro de Kalman 2x2 para cada *landmark* presente, reduzindo o tempo de processamento, sendo significantemente mais rápido que o EKF–SLAM.

Seu algoritmo consiste em quatro etapas, primeiramente a posição de cada partícula é atualizada, apenas com a odometria. Vale ressaltar que na versão 2.0, a posição dos *landmarks* influencia diretamente na pose. A próxima etapa realiza a atualização dos *landmarks* de cada partícula com seu EKF correspondente e então calcula-se os pesos de cada partícula [5]. Por fim, uma etapa essencial, a reamostragem das partículas, responsável por eliminar as partículas que possuem menor peso, logo menos informação útil [3].

A etapa de reamostragem é realizada através do método de amostragem estocástica universal, pois esta apresenta uma baixa variância, possuindo uma implementação eficiente. De uma maneira bem simples, é uma roleta com o número de setas correspondente ao número total de partículas e cada partícula pode ser escolhida proporcionalmente ao seu peso

3. Resultados

Um ambiente interno é simulado no *CoppeliaSim* e então são implementados os métodos apresentados na Seção anterior, resultando em um FastSLAM *online*. O robô é liberado no ambiente para se movimentar livremente, sem uma trajetória pré-definida e então, todos os seus dados vão sendo recebidos.

Split-and-Merge

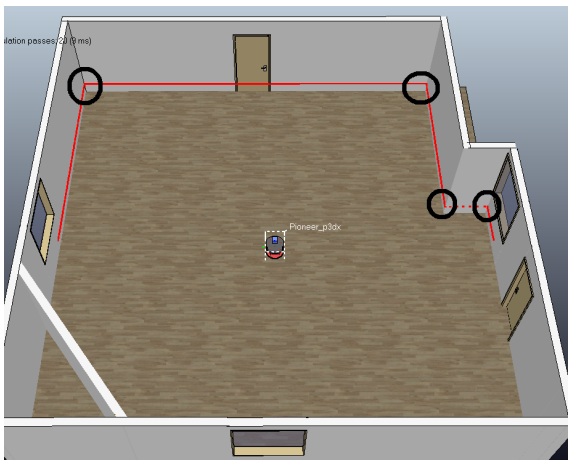

Figura 2. Landmarks a serem detectados pelo *Split-and-Merge*.

Para demonstrar o funcionamento do algoritmo, o veículo é colocado parado de frente para uma parede e então são retiradas os *landmarks* detectados. Como apresentado na Figura 2, há quatro pontos de interesse, os cantos das paredes, encontrado os pontos (4.93693 m, -1.38314 rad), (4.08377 m, -1.29278 rad), (6.06039 m, -0.667244 rad) e (6.69507 m, 0.799297 rad), sendo distância e ângulo em relação ao robô, respectivamente.

FastSLAM 1.0

Após a implementação da detecção de *landmarks* e de cada etapa, basta realizar a isimulação do algoritmo FastSlam 1.0. Como mencionado, o veículo foi liberado para andar livremente pelo mapa e seus valores foram sendo apresentados graficamente no Rviz. Nas Figuras 3 e 4 é apresentado o avanço da lozalição dos *landmarks* e na localização do veículo. Em verde são os *landmarks* salvos em uma das partículas, em vermelho o valor real da posição do veículo e em azul a localização dada pela etapa de predição do FastSLAM 1.0.

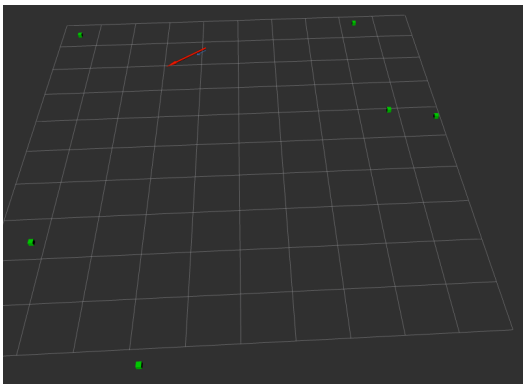

Figura 3. Momento que o erro da odometria não comprometeu o algoritmo.

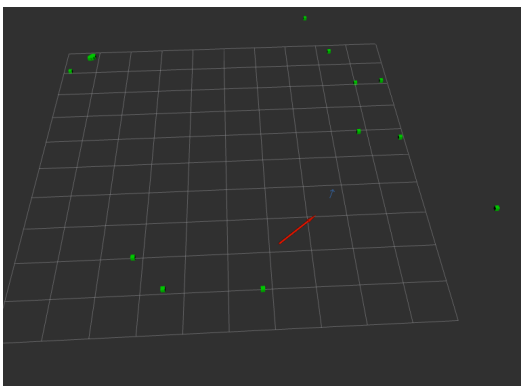

Figura 4. ALgoritmo começa a encontrar falsos negativos.

4. Conclusões

O trabalho consistia apresentar os resultados preliminares de um sistema de localização de um robô imerso no ambiente simulado CoppeliaSim, interconectando os algoritmos implementados a simulação via ROS. Primeiramente, percebe-se a boa performance do método de detecção de *landmarks*, o *Split-and-merge*, por mais que tenha uma simplicidade, se mostrou eficiente. Então, é apresentado o FastSLAM 1.0, que inicialmente há uma boa resposta de determinação da posição real dos *landmarks*, entretanto, como a atualização da posição de cada partícula depende unicamente do sinal de controle, velocidade das rodas, o sistema vai encontrando falsos positivos, resultando em um mapa degradado. Uma boa solução que será implementada no futuro, é o FastSLAM 2.0, que adiciona na atualização da posição do robô influência dos *landmarks*, o que melhora o sistema como um todo [6].

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Referências

[1] Tiago Giacomelli Alves. Sistema de controle de pose para uma cadeira de rodas inteligente. Master's thesis, Universidade Federal do Rio Grande do Sul, Brasil, 2018.

[2] Luca Calabrese. Robust odometry, localization and autonomous navigation on a robotic wheelchair. Master's thesis, Politecnico Di Milano, Italy, 2013.

[3] Márcio Nunes de Miranda. Algoritmos genéticos: fundamentos e aplicações, 2007.

[4] Julio Fajardo, Victor Ferman, Jabes Guerra, Antonio Ribas Neto, and Eric Rohmer. Lmi methods for extended filters for landmark-based mobile robot localization. In *2021 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, pages 511–517. IEEE, 2021.

[5] Michael Montemerlo, Sebastian Thrun, Daphne Koller, Ben Wegbreit, et al. Fastslam: A factored solution to the simultaneous localization and mapping problem. *Aaai/iaai*, 593598, 2002.

[6] Michael Montemerlo, Sebastian Thrun, Daphne Koller, Ben Wegbreit, et al. Fastslam 2.0: An improved particle filtering algorithm for simultaneous localization and mapping that provably converges. In *IJCAI*, volume 3, pages 1151–1156, 2003.

[7] V. Nguyen, A. Martinelli, N. Tomatis, and R. Siegwart. A comparison of line extraction algorithms using 2d laser rangefinder for indoor mobile robotics. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1929–1934, 2005.

[8] World Health Organization. *World report on disability 2011*. World Health Organization, 2011.

[9] Richard C Simpson. Smart wheelchairs: A literature review. *Journal of rehabilitation research and development*, 42(4):423, 2005.

[10] Shaharyar Ahmed Khan Tareen and Zahra Saleem. A comparative analysis of sift, surf, kaze, akaze, orb, and brisk. In *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, pages 1–10, 2018.

[11] Sebastian Thrun. Probabilistic robotics. *Communications of the ACM*, 45(3):52–57, 2002.

[12] Salifu Yusif, Jeffrey Soar, and Abdul Hafeez-Baig. Older people, assistive technologies, and the barriers to adoption: A systematic review. *International journal of medical informatics*, 94:112–116, 2016.

Path planning and terramechanics coupling for rovers in rough lunar terrain

Vinicius Emanoel Ares , Felipe Augusto Oliveira Mota , César Bastos da Silva , Eric Rohmer
{vearesfem@gmail.com, rohmer@unicamp.br}

Departamento de Engenharia de Computação e Automação (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Campinas, SP, Brasil

Abstract – It is proposed the research of soil-wheel contact models, known as terramechanics, and integrating those models into path planning algorithms. The coupling of these two domains will enable the development of path planning systems for space explorations rovers in rough terrains. The proposed approach has three main applications. First, terramechanics simulations enable the trafficability evaluation of the terrain, so the paths will be generated to minimize the risk of the rover getting stuck. Second, the energy necessary to drive depends on how much the rover sinks or slips. The proposed simulation will estimate the energy consumption beforehand, and choose paths of minimum energy consumption. And third, it will enable the development of a control system based on marks left by wheels in the soil. The simulation would be used to train such a system. It is planned to validate the developed simulations in a soil-wheel test bed of Tohoku University. Afterwards, tests will be performed with a lunar rover prototype at University of Luxembourg. This research is part of an ongoing project partially founded by a CAPES-PrInt scholarship and will be in cooperation with institutions of Luxembourg, USA and Japan. The simulation platform proposed will help the Japan partners who are developing rovers to search for water on the Moon.

Keywords – terramechanics, rover, path planning, lunar water.

1. Introduction

It is proposed the research of soil-wheel contact models, known as terramechanics, and integrating those models into path planning algorithms. The coupling of these two domains will enable the development of path planning systems for space explorations rovers in rough terrains. The research is a part of the ongoing project partially founded by a CAPES-PrInt scholarship, ”Simulation platform as testbed for high and low level control of the Sorato lunar water prospecting robot under realistic conditions”. The proposed approach has three main applications: 1 – safe paths, 2 – minimum energy consumption, and 3 – control by camera. First, terramechanics simulations enable the trafficability evaluation of the terrain, so the generated paths will avoid very rugged regions, minimizing the risk of the rover slipping or getting stuck. Second, the energy necessary to drive depends on how much the rover sinks or slips, which is dependent on the slopes and level of aggregation of the terrain. The proposed simulation will estimate the energy consumption beforehand, and choose paths of minimum energy consumption. Third, in a simulation that features terramechanics, the rover wheels will leave marks on the soil. That will enable the future development of a system that uses the rover camera and sensors to control a straight path based on soil marks.

Space exploration missions involving astronauts are more expensive and dangerous than using rovers. On the other hand, there are interactions that occur in the wheels of rovers that must be studied in other to achieve a safe ride. Some of the phenomena that can occur between the rover and the lunar regolith is side slippery, when the rover is traversing a slope, or accumulation of soil around a wheel that can make it stuck. Getting stuck is specially critical for space missions, by not having means to unstuck a vehicle, this can jeopardize a multi-million dollar program. By means of experiments and simulation, the conditions that could hinder a rover stuck can be predicted and prevented.

The technology of space exploration rovers can also be applied to drive in rough terrains found in disaster sites. Dr. Eric Rohmer, while working in the Space Robotics Laboratory has helped the development of the hardware and software of the Quince rescue robot. This robot has faced a real challenge in 2011 when a tsunami hit the East cost of Japan, resulting in the Fukushima nuclear accident. The environment inside the plant was too dangerous for anyone to walk inside. Three Quince robots were retrofitted to carry sensors and enter the accident environment to perform several mapping, inspection and radiation sensing missions inside the Fukushima power plant [1].

Simulation software play an important role in the development of robot hardware and software. They help reducing the cost and time that would otherwise be spent

in numerous prototypes and tests. Conversely, simulation helps to define which experiments are more important, hence resources can be focused. Today, the two most popular simulation software for robotics are CoppeliaSim and Gazebo, according to [4], [7] and [10].

CoppeliaSim and Gazebo are able to address multibody dynamics simulations, where several rigid body elements connect and move relative to each other, these are used to simulate the moving parts of a robot. What those simulators lack though, is the ability of performing simulations considering a deformable medium, like tires and soil. Terramechanics is the field that deals with the dynamics of the contact between a wheel and the soil. When the focus is in robot that use legs rather then wheels, the term Terradynamics is preferred. [6] Terramechanics allows to determine the effect on the wheels, i.e loads, as well the effects on the soil, i.e. the wheel impressions. Moreover, a simulator with a realistic terramechanics model can be used as a tool to train machine learning algorithms for many kind of situations, like path planing, slippery detection, path following, optimal exploration or swarm control strategies.

The developed terramechanics models will be integrated as APIs in the robot simulation software CoppeliaSim (previously called V-REP). CoppeliaSim is developed by Coppelia Robotics, a company based in Zurich, Switzerland and managed by Dr. Marc Freese. The supervisor of this research, Dr. Eric Rohmer, has collaborated in the development of CoppeliaSim, notably in the development of APIs [9]. An advantage of using CoppeliaSim in this research is having easy access to the main developer, Dr. Marc Freese, that will be helpful if issues arise during the integration of the terramechanics API into the simulator.

2. Proposal

It is proposed the development and coupling of terramechanics and path-planning models. The models will be integrated as plugins into the robot simulator CoppeliaSim (former V-REP). Then, CoppeliaSim with the plugins will be used to simulate the motion of a rover in lunar regolith. This would allow obtaining paths that are optimum in terms of safety against the risk of getting stuck and energy saving. It will also make possible to represent marks left in the soil by wheels, see figure 1 for an example. Future research can use these wheel marks to develop control based on marks captured by the rover camera. Besides simulation, tests in a single-wheel test bed will be used to calibrate the parameters of the terramechanics model. Then, a rover prototype, Explorer 1 (EX1) will be used to validate the developed models. The rover EX1 is being developed in the Space Robotics Lab at Tohoku University, Japan. The supervisor of this research, Dr. Eric Rohmer, participated in the development of rovers for space exploration [8] as a post doc and researcher at Tohoku University. Nowadays Dr. Eric Rohmer is in active collaboration with the Space Robotics Lab. The simulator used in this research, CoppeliaSim, was developed by the Swiss company Coppelia Robotics. Dr. Eric Rohmer also participated in the development of CoppeliaSim and collaborated with plugins.

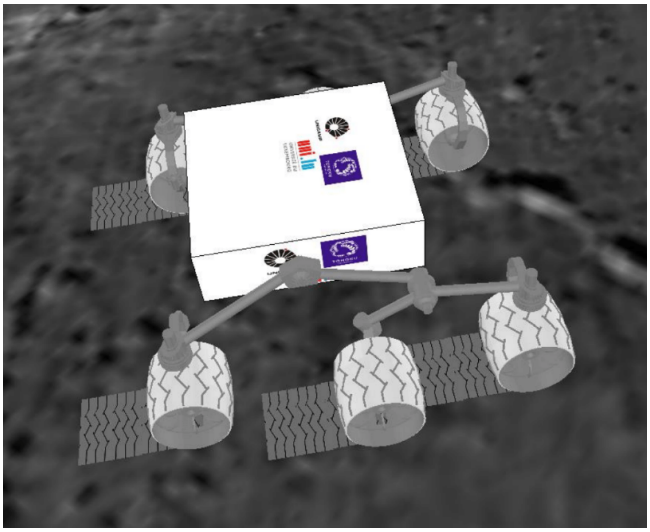

Figure 1. Wheel marks.

The locomotion dynamics can exhibit longitudinal and lateral slip. Longitudinal slip is measured by the slip ratio, which is the ratio between the circumference velocity of the wheel and the longitudinal traveling velocity of the vehicle [3], expressed by the equation:

$$s = \left\{ \begin{array}{ll} (r\omega - v_x)/r\omega & (r\omega > v_x : driving) \\ (r\omega - v_x)/v_x & (r\omega < v_x : braking) \end{array} \right\}$$

The lateral slip is measured by the slip angle, which is the arc-tangent between the lateral and longitudinal traveling velocities [3], given by the expression:

$$\beta = arctan(v_y/v_x)$$

To understand the phenomena that takes place between a wheel and granular media, three approaches can be used: experimental, analytical and numerical simulation. For the later option, a useful tool is the Discrete Element Method. Figure 2 shows a contact model of two particles, the basis for the Discrete Element Method.

A terramechanics model will be implemented in the C++ programming language. It will then be integrated

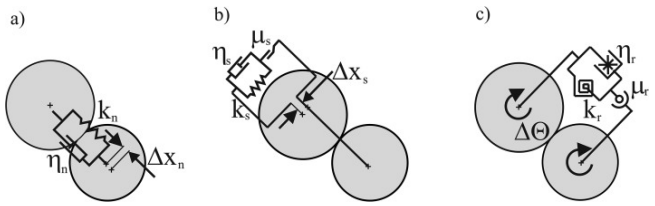

Figure 2. Contact model for Discrete Element Method [2]. Reactions forces of the type normal (a), tangential (b) and rolling (c).

as a plugin in the CoppeliaSim robot simulator. Experimental validation of the terramechanics model will be carried out in the test bed of Space Robotics Lab. The experimental setup is shown in figure 3. In this setup, the conveyance and the wheel can be drive independently by motors, this allows to vary the slip ratio as desired. The steering part enables to choose a steering angle. As the conveyance moves in a single direction, the steering angle is equivalent to the slip ratio β. With a force and torque sensor it is possible to measure forces in 3 directions and torques in 3 directions.

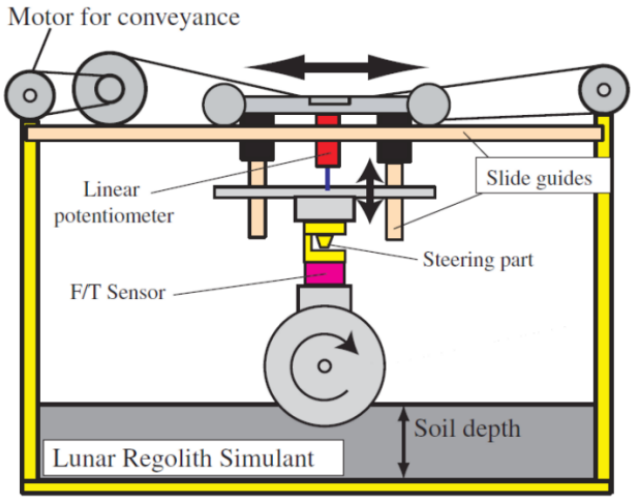

Figure 3. Terramechanics experimental setup to measure forces in a single wheel. [3].

2.1. Objectives

The proposed research intends to meet the following objectives:

1. Study existing terramechanics models, implement and compare them regarding accuracy and computational cost. Calibrate and validate models in a single-wheel test bed;
2. Implement models in C++ for fast computation and integrate in CoppeliaSim. This will help Tohoku University researchers in rover development;
3. Develop new path-planning models, that consider trafficability evaluation and terramechanics simulations. Run simulations to verify capabilities of safe and energy efficient path generation;
4. Run simulations of a real rover, EX1, in CoppeliaSim, aiming for accurate simulation of lunar conditions. Perform laboratory tests with rover EX1 to validate the simulations.

Figure 4 shows the Lunar Testbed of Luxembourg University, where it is intended to accomplish objective 4.

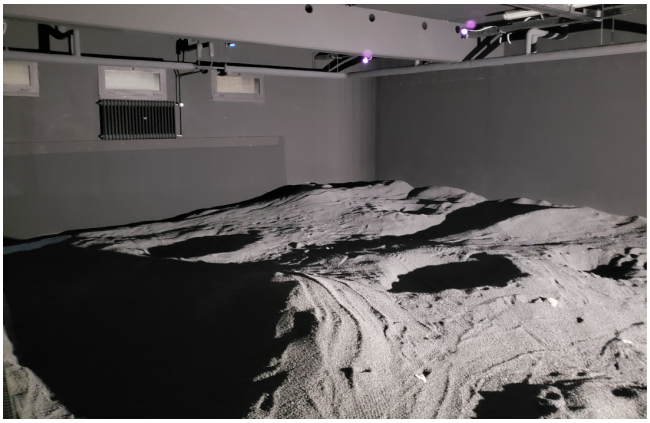

Figure 4. Luxembourg lunar test bed.

3. Expected Results

There are two aspects that will be analysed and validated: the rover simulation platform and the terramechanics path planner. The lunar test bed experimental setups of universities of Luxembourg and Tohoku are important to validate the rover simulation platform. Once it is validated, it can be used to evaluate the performance of the terramechanics path planner.

3.1. Validation of the rover simulation platform

In the lunar test bed, it will be prepared a terrain topology set up. Then that terrain topology will be scanned to create a digital elevation map. Wherefore we will have two versions of the same terrain topology, one real and other virtual. In that topology we propose two paths, one straight line path connecting two spots, and one path with curves. Those two paths are then run in the rover simulation platform and in the lunar test bed with the rover prototype. The energy consumed during the runs are recorded and compared, simulation versus experimental test. Depending on the difference found, the rover simulation platform will be considered validated, otherwise further refinements will be implemented.

3.2. Performance analysis of the terramechanics path planner

The evaluation of the proposed path planner will have two goals:

1. Check whether the terramechanics path planner is able to generate safer paths (less risk of getting stuck) compared to reference path planners.
2. Check whether the terramechanics path planner is able to generate paths with lower energy consumption compared to reference path planners.

The path planners used as reference for comparisons will be A*, RRT [5] and RRT*. For goals 1 and 2, candidate paths are generated in reference path planners and in the terramechanics path planner. Then, those candidate paths are run in the rover simulation platform and in the lunar test bed. After those runs, it is compared the number of times the rover got stuck and how much energy was consumed in the path.

4. Conclusion

With the scientific literature perused it was possible to find relevant gaps in the research area and thus define a research subject. This field is particularly interesting for research because the technologies of terramechanics and path planning has broad application like self-driving cars, off-road vehicles, agricultural robotics, and socially assistive robots.

Acknowledgments

The authors would like to thank the SEW-Eurodrive and Professor Eleri Cardozo for the financial support. Also, we are gratefull to Victor Ferman and Jorge Rufino Fernández Herrera for helping in the adaptation to LCA laboratories.

References

[1] Eric Rohmer webpage. `https://www.dca.fee.unicamp.br/~eric/`.

[2] Józef Horabik and Marek Molenda. Parameters and contact models for dem simulations of agricultural granular materials: A review. *Biosystems Engineering*, 147:206–225, 2016.

[3] G. Ishigami and K. Yoshida. Steering characteristics of an exploration rover on loose soil based on all-wheel dynamics model. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3099–3104, 2005.

[4] Serena Ivaldi, Jan Peters, Vincent Padois, and Francesco Nori. Tools for simulating humanoid robot dynamics: A survey based on user feedback. In *2014 IEEE-RAS International Conference on Humanoid Robots*, pages 842–849, 2014.

[5] Steven M. LaValle. Rapidly-exploring random trees : a new tool for path planning. *The annual research report*, 1998.

[6] Chen Li, Tingnan Zhang, and Daniel I. Goldman. A terradynamics of legged locomotion on granular media. *Science*, 339(6126):1408–1412, mar 2013.

[7] Lenka Pitonakova, Manuel Giuliani, Anthony Pipe, and Alan Winfield. Feature and performance comparison of the V-REP, Gazebo and ARGoS robot simulators. 02 2018.

[8] Eric Rohmer, Giulio Reina, and Kazuya Yoshida. Dynamic simulation-based action planner for a reconfigurable hybrid leg–wheel planetary exploration rover. *Advanced Robotics*, 24(8-9):1219–1238, 2010.

[9] Eric Rohmer, Surya P. N. Singh, and Marc Freese. V-REP: A versatile and scalable robot simulation framework. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1321–1326, 2013.

[10] R R Shamshiri, I A Hameed, L Pitonakova, C Weltzien, S K Balasundram, and I J Yule. Simulation software and virtual environments for acceleration of agricultural robotics: Features highlights and performance comparison. In *2018 Int J Agric & Biol Eng*, page 15–31, 2018.

Otimizações em operações de criptografia pós-quântica baseada em reticulados para ambientes restritos

Felipe J. A. Rampazzo , Marco A. A. Henriques
{f233261@dca.unicamp.br, maah@unicamp.br}

Departamento de Engenharia de Computação e Automação (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Campinas, SP, Brasil

Resumo – Os esquemas de criptografia de chave pública estão fadados à obsolescência com a possibilidade do surgimento de computadores quânticos, visto que a segurança desses sistemas é sustentada em problemas matemáticos que estes computadores conseguiriam resolver em tempo polinomial. Devido aos efeitos catastróficos que o surgimento desse computador traria para a computação de um modo geral, novos esforços têm sido feitos na criação de algoritmos pós-quânticos que sejam resistentes aos futuros ataques dos computadores quânticos. O National Institute for Standards and Technology - NIST - EUA tem coordenado os estudos neste tema e já selecionou alguns algoritmos baseados em reticulados como padrões de criptografia pós-quântica. Nesse contexto, este trabalho busca avaliar as possibilidades de melhoria em relação ao desempenho e custos de implementação de algoritmos criptográficos pós-quânticos baseados em reticulados em ambientes restritos. A justificativa para tal abordagem se deve às importantes limitações desses ambientes, que geralmente carecem do poder computacional e funcionalidades necessárias às operações criptográficas pós-quânticas.

Palavras-chave – criptografia pós-quântica, reticulados, ambientes computacionais restritos, processadores ARM

1. Introdução

Nos últimos anos é visto uma crescente de projetos para a construção de um computador quântico funcional. Todavia, seu surgimento colocaria em risco a segurança dos criptossistemas de chave pública atuais, mais especificamente os de troca de chaves e assinaturas.

Com a publicação do trabalho de [1], verificou-se que um computador quântico conseguiria quebrar esses criptossistemas em tempo polinomial, tornando-os fadados à obsolescência. Desde então, novos esforços têm sido feitos na criação de algoritmos pós-quânticos que sejam resistentes aos futuros ataques desses computadores.

Na busca desses novos esquemas criptográficos, o NIST iniciou em 2016 o Processo de Padronização de Criptografia Pós-Quântica (PQC Standardization Process), um concurso com o objetivo de selecionar os novos padrões de criptografia de chave pública que sejam resistentes aos computadores quânticos.

Após três rodadas, os primeiros escolhidos foram CRYSTALS-Kyber na categoria de mecanismos de encapsulamento de chave (public-key encapsulation mechanism - KEM) e os de assinatura selecionados foram CRYSTALS–Dilithium, FALCON e SPHINCS, sendo o primeiro deles o recomendado para uma implementação inicial [2]. Uma quarta rodada está em andamento para a definição de outros algoritmos a serem padronizados.

Os métodos matemáticos que garantem a segurança vistos nesses algoritmos incluem os baseados em códigos, reticulados, multivariados, hashes e isogenias. Contudo, ao analisar os algoritmos já padronizados, percebe-se que os construídos sobre problemas de reticulados são predominantes visto que, dos quatro definidos, três sustentam sua segurança nessas estruturas.

Com o intuito de explorar esses problemas, este trabalho procura avaliar as possibilidades de otimização das principais operações presentes em algoritmos criptográficos pós-quânticos baseados em problemas de reticulados em ambientes restritos de software e hardware.

2. Criptografia baseada em reticulados

Os reticulados são conjuntos de pontos discretos no espaço n-dimensional Euclidiano R^n , descritos como todas as combinações lineares inteiras de vetores independentes [3]. Na Figura 1, é mostrado um exemplo de reticulado formado por todas as combinações lineares de $\mathbf{b_1}$ e $\mathbf{b_2}$, como por exemplo o vetor $\mathbf{s}$ formado pela combinação $-2b_1 + b_2$. **Definição 1:** seja $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_m$ vetores linearmente independentes em R^n. Um reticulado Λ com bases $\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_m\}$, em que $m \leq n$ [4, p. 50], é definido por:

$$\Lambda = \{u_1\mathbf{b}_1 + \ldots + u_m\mathbf{b}_m : u_1, \ldots, u_m \in Z\}$$

A criptografia baseada em reticulados usa conjecturas de difícil solução em reticulados de R^n como prova

de segurança para a construção de sistemas criptográficos [3]. O primeiro trabalho desta área é datado em 1996 com a proposta de [5], definida como Short Integer Solution (SIS), que consiste em encontrar o vetor mais curto de um reticulado inteiro dado uma função de mão única baseada neste reticulado. Ainda neste trabalho, foi definido que a complexidade de resolver este problema no caso médio é tão difícil quanto no pior caso.

Contudo, é no ano de 2005 com o trabalho de [6] que a ligação entre sistemas criptográficos e reticulados ganham mais atenção. Regev propõe um método chamado Learning With Errors (LWE) que se torna a base dos criptossistemas que se respaldam em reticulados. No LWE, há duas instâncias do problema consideradas NP-difícil, ou seja, que são tão difíceis de resolver quanto os problemas mais difíceis em NP. Esses problemas são conhecidos como Shortest Vector Problem (SVP) e o Closest Vector Problem (CVP).

No SVP, dado dois vetores bases $\mathbf{b_1}$ e $\mathbf{b_2}$ de um reticulado, deve-se encontrar o vetor mais curto que pertença ao reticulado a partir dessas bases. Na Figura 1, esse vetor é representado por $\mathbf{s}$.

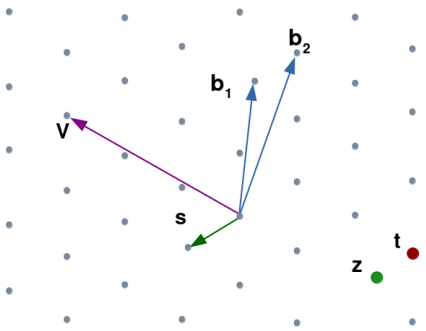

Figura 1. Shortest Vector Problem (SVP) e Closest Vector Problem (CVP). Fonte: `https://blog.cloudflare.com/towards-post-quantum-cryptography-in-tls/`

Já no CVP, a dificuldade está em encontrar um vetor $\mathbf{z}$ que não pertença necessariamente ao reticulado e que seja o mais próximo de um vetor t que é pertencente ao reticulado. Ambos esses problemas se mostraram seguros tanto em ataques de computadores clássicos quanto, na teoria, dos quânticos. Vemos um reticulado de apenas duas dimensões na figura 1. Porém, em modelos reais, as dimensões dos reticulados são muito grandes e, somada a escolha de bases pouco ortogonais entre si, dificultam muito a solução do problema.

2.1. Problema Learning With Errors

O LWE é um problema baseado em reticulados voltado a aplicações criptográficas, em que é necessário resolver um sistema de equações lineares sobre um módulo primo inteiro q. Para montar o sistema criptográfico utilizando o LWE deve-se:

1. Escolher um vetor secreto s, tal que $\mathbf{s} \in Z_q^n$.
2. Vetores públicos (matriz) $\mathbf{a}^i = (a_1^i, \ldots, a_n^i)$, tal que $1 \leq i \leq n$, obtidos por meio de uma distribuição uniforme em Z_q^n.
3. Por fim, deve-se escolher um vetor de erro com coeficientes pequenos, tal que $\mathbf{e} = (e^1, \ldots e^n) \in Z_q^n$ via uma distribuição χ.
4. Assim, a distribuição LWE pode ser dada como

$$b^i = \sum_{j=1}^{n} a_j^i s_j + e^i q$$

A dificuldade em encontrar o vetor secreto $\mathbf{s}$ é aumentada com a adição dos erros ao sistema. Sem esses ruídos, esse sistema poderia ser resolvido a partir da eliminação de Gauss. Assim, para todo $\mathbf{a} \cdot \mathbf{s} \approx b$, uma adição suscetiva de erros sempre resulta em uma aproximação cumulativa que cresce tanto a ponto de não sobra nenhuma informação na "aproximação" obtida, algo que poderia favorecer a descoberta de $\mathbf{s}$. Essa afirmação vale mesmo quando o erro é pequeno.

Para cifrar um bit x é necessário gerar um vetor $\mathbf{v} \in 0, 1^m$ e gerar o par que representa o texto cifrado

$$(\mathbf{c}_1, c_2) = (A\mathbf{v}, \mathbf{b} \cdot \mathbf{v} + x \cdot \left\lfloor \frac{q}{2} \right\rfloor)$$

. Já para decifrar, usamos a chave secreta $\mathbf{s}$ calculada anteriormente e computamos
$c_2 - \mathbf{c}_1 \cdot \mathbf{s} = (A\mathbf{s} + \mathbf{e}) \cdot \mathbf{v} + x \left\lfloor \frac{q}{2} \right\rfloor - (A\mathbf{v}) \cdot \mathbf{s}$
$= (A\mathbf{s}) \cdot \mathbf{v} - (A\mathbf{v}) \cdot \mathbf{s} + \mathbf{e} \cdot \mathbf{v} + x \left\lfloor \frac{q}{2} \right\rfloor \approx x \left\lfloor \frac{q}{2} \right\rfloor$.

Se x for 1, o valor estaria mais próximo de $\left\lfloor \frac{q}{2} \right\rfloor$ do que de zero, visto que os vetores de erro adicionados são pequenos. Do contrário, o valor estaria mais próximo de zero.

Novas variantes do LWE surgiram com o tempo, como o Ring-LWE proposto por [7], em que as operações são sobre anéis polinomiais, e não vetores n-dimensionais do LWE, o que reduz o tamanho das chaves, uma vez que somente a primeira linha da matriz precisa ser armazenada e o restante é calculado partindo da linha anterior, tornando-as próximas do tamanho das chaves do RSA e melhorando a eficiência das construções baseadas em reticulados.

Outra variante do LWE pode ser encontrada em [8] e [9] chamada Module-LWE, visando resolver fragilidades tanto do LWE quanto do RLWE. A diferença principal para o RLWE é a substituição dos elementos do anel único por elementos modulares de um mesmo anel. Este é geralmente um anel ciclotômico de potência de

dois, ou seja, um anel em $Z[\mathbf{X}]/\langle \mathbf{X}^n + 1\rangle$ com $n = 2^k$ [10]. O Module-LWE se tornou modelo mais utilizado desde então nos algoritmos criptográficos baseados em reticulados, implementado nos padrões já definidos de algoritmos pós quânticos, como o CRYSTALS-Kyber, pois algumas operações são facilitadas neste tipo de anel, tornando-os mais performáticos.

2.2. CRYSTALS-Kyber

O CRYSTALS-Kyber é um algoritmo para realizar troca de chaves de forma segura a ataques de computadores pós-quânticos, por meio de um mecanismo de encapsulamento de chaves. O algoritmo é do tipo IND-CCA2 e baseado em um variante do LWE em reticulados modulares, o Module-LWE. A chave encapsulada é formada por 256 bits, permitindo seu uso por algoritmos de chave simétrica como o AES, o que facilita o desenvolvimento de protocolos híbridos, que unem a criptografia pós-quânticas e clássica. Todo o processo de KEM do CRYSTALS-Kyber pode ser dividido em três operações principais: geração de chaves, encapsulamento (ou encriptação) e desencapsulamento (ou decriptação).

Na etapa de de geração de chaves, um par de chaves pública e privada é gerado pela parte que deseja inciar a comunicação segura. Os parâmetros iniciais a serem definidos são um módulo q para os coeficientes e um módulo polinomial $x^n + 1$ (para reduzir o grau do polinômio). A chave pública é formada por duas partes: uma matriz e um vetor. As entradas da matriz $\mathbf{A}$ são polinômios em que os coeficientes são randômicos sobre o módulo q.

Então, é gerado a chave secreta $\mathbf{s}$, porém, com os coeficientes do polinômios "pequenos", da mesma forma que o vetor de erros $\mathbf{e}$. Por fim, para obter a segunda parte da chave pública, o vetor $\mathbf{t}$, realizamos uma multiplicação e adição de matrizes, tal que $\mathbf{t} = \mathbf{As} + \mathbf{e}$.

A seguir, é realizado o encapsulamento da chave secreta compartilhada em que são necessários dois vetores de polinômios $\mathbf{s}$ e $\mathbf{e_1}$ e um polinômio e_2. Para cifrar a mensagem, primeiro, é preciso transformá-la em um polinômio. No caso do CRYSTALS-Kyber, cada bit representa um coeficiente do polinômio. Antes de cifrar a mensagem, precisamos "expandir"o polinômio, para evitar os erros adicionados, multiplicando-o por $\lceil q/2 \rfloor$.

Após esse processo, cifra-se *m* com a chave pública $(\mathbf{A}, \mathbf{t})$, tal que $\mathbf{u} = \mathbf{A}^T\mathbf{r} + \mathbf{e_1}$ e $v = \mathbf{t}^T\mathbf{r} + e_2 + m$. Tanto o vetor $\mathbf{u}$ quanto o polinômio *v* são enviados e consistem no texto cifrado. A dificuldade de se obter m se baseia em quão complexo é resolver o problema do SVP.

A última etapa consiste no desencapsulamento da chave secreta compartilhada. Em posse da chave secreta $\mathbf{s}$, podemos decifrar a mensagem da seguinte forma: $m = v - \mathbf{s}^T\mathbf{u}$. Porém, o resultado ainda não é o que se espera, visto que $m = \mathbf{e}^T\mathbf{r} + e_2 + m + \mathbf{s}^T\mathbf{e_1}$. É necessário desfazer a "expansão", verificando se os valores são mais próximos de $\lceil q/2 \rfloor$ ou de (0 ou q) e assim definir se os bits são 0 ou 1.

Todas essas etapas passam por um processo de conversão que aumenta o nível de segurança do algoritmo de IND-CPA para IND-CCA2, por meio da Transformação Fujisaki-Okamoto.

O CRYSTALS-Kyber conta ainda com funções hash padronizadas pelo NIST na FIPS 202, que inclui as pertencentes à família SHA3 e SHAKE. Além disso, a multiplicação de vetores e matrizes são realizadas a partir de uma implementação da transformada rápida de Fourier para operar sobre anéis e corpos finitos, chamada Number Theoretic Transform (NTT).

3. Resultados

Neste trabalho, dois ambientes restritos são estudados para verificar as possibilidades de melhoria e implementação do algoritmo CRYSTALS-Kyber. O primeiro ambiente é em hardware, com mudanças nas operações de hash em placas ARM M0+. Já o segundo ambiente restrito é baseado em software, em que levantou-se problemas iniciais da implementação para switches programáveis em P4. Ainda não foi definido qual dos ambientes serão alvo da pesquisa, apenas estudos iniciais acerca das características e desafios de cada um deles. Os resultados já obtidos desses ambientes podem ser vistos a seguir.

3.1. Ambiente restrito em hardware - ARM M0+

Os testes iniciais foram realizados em uma placa Freedom FRDM-KL25Z, equipada com MCU KL25Z128, ARM Cortex-M0+, 128 KB de memória FLASH, clock de 48 MHz e 16 KB de memória SRAM, sendo que desses, 3 KB são dedicados ao uso do ambiente de desenvolvimento.

Como pode ser visto em [11], foi possível executar as três funções principais do algoritmo. Contudo, algumas modificações foram necessárias, como a substituição da biblioteca OpenSSL para a geração de números aleatórios pela função *srand* da linguagem C. Embora haja impactos na segurança, essa ação foi precisa dada as limitações do ambiente.

Outra modificação foi aplicada nas funções hash do CRYSTALS-Kyber como a de saída extensível (XOF), as H e G, uma função pseudo-aleatória (PRF) e uma de derivação de chave (KDF). O intuito dessas modificações foi medir o impacto de cada uma na eficiência do algoritmo

em ambientes que carecem de aceleradores em hardware.

3.2. Ambiente restrito em software - P4

O P4 é uma linguagem utilizada para definir como os pacotes serão processados no plano de dados por dispositivos de rede programáveis. Esses dispositivos também precisam prover serviços de segurança para diversos contextos e é possível aplicá-las para a detecção de intrusão, mitigação de DDoS, implantação de ACLs, firewalls, entre outras [12].

Todavia, há uma dificuldade de se aplicar sistemas criptográficos nestes dispositivos, uma vez que apenas operações de aritmética simples [13] são permitidas, além da ausência de funções hashes seguras criptograficamente. Há trabalhos que implementam hashes seguros em P4, como em [14] e [15], porém estes não são os padrões definidos pelo NIST no concurso dos novos algoritmos PQC.

Outra barreira do P4 é a ausência de *loops*, ainda que a recirculação de pacotes possa ser usada para simular essa propriedade. Contudo, o número de vezes que um pacote pode recircular é limitado em alguns modelos de switches. A funcionalidade *#define*, nativa do P4 e presente também na linguagem C, ameniza essa limitação ao criar um *loop* "desenrolado"e evitar a redundância de linhas de código, que naturalmente ocorre devido as características intrínsecas da linguagem P4.

4. Conclusões

Com o iminente surgimento de um computador quântico e os riscos que este trará para as comunicações seguras, novos esforços estão sendo empregados na busca por sistemas criptográficos que sejam seguros a ataques desses dispositivos.

Entretanto, a construção matemática desses criptossistemas são pesados do ponto de vista computacional, o que exige implementações mais otimizadas, principalmente se forem aplicados em ambientes restritos, tanto em software como em hardware.

Dadas essas motivações, este trabalho busca efetuar otimizações nas operações mais pesadas dos algoritmos PQC, como as funções hash e NTT. Já em um ambiente restrito em software como visto na linguagem P4, a contribuição reside na implementação do algoritmo PQC em si, visto que este cenário possui grandes limitantes para o desenvolvimento.

Referências

[1] P. W. Shor, "Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer," *SIAM Journal on Computing*, vol. 26, no. 5, pp. 1484–1509, 1997.

[2] D. Moody, "Status report on the third round of the NIST post-quantum cryptography standardization process," tech. rep., National Institute of Standards and Technology, jul. 2022.

[3] S. I. Costa, F. Oggier, A. Campello, J.-C. Belfiore, and E. Viterbo, *Lattices Applied to Coding for Reliable and Secure Communications*. Springer International Publishing, 2017.

[4] M. F. Bollauf, "Códigos, reticulados e aplicações em criptografia," Master's thesis, Universidade Estadual de Campinas, Campinas-SP, Brasil, 2015.

[5] M. Ajtai, "Generating hard instances of lattice problems (extended abstract)," in *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '96, (New York, NY, USA), p. 99–108, Association for Computing Machinery, 1996.

[6] O. Regev, "On lattices, learning with errors, random linear codes, and cryptography," *Journal of the ACM (JACM)*, vol. 56, no. 6, 2009.

[7] V. Lyubashevsky, C. Peikert, and O. Regev, "On ideal lattices and learning with errors over rings," *Journal of the ACM (JACM)*, vol. 60, nov 2013.

[8] A. Langlois and D. Stehlé, "Worst-case to average-case reductions for module lattices," *Designs, Codes and Cryptography*, vol. 75, 06 2014.

[9] Z. Brakerski, C. Gentry, and V. Vaikuntanathan, "Fully homomorphic encryption without bootstrapping," *Electronic Colloquium on Computational Complexity (ECCC)*, vol. 18, p. 111, jan. 2011.

[10] M. R. Albrecht and A. Deo, "Large modulus ring-lwe $\geq$ module-lwe," in *Advances in Cryptology – ASIACRYPT 2017* (T. Takagi and T. Peyrin, eds.), (Cham), pp. 267–296, Springer International Publishing, 2017.

[11] L. F. C. Ferro, F. J. A. Rampazzo, and M. A. A. Henriques, "Estudos de otimização do algoritmo de criptografia pós-quântica CRYSTALS-KYBER," in *Anais Estendidos do XXI Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais (SBSeg Estendido 2021)*, Sociedade Brasileira de Computação - SBC, 2021.

[12] E. F. Kfoury, J. Crichigno, and E. Bou-Harb, "An exhaustive survey on p4 programmable data plane switches: Taxonomy, applications, challenges, and future trends," *IEEE Access*, vol. 9, pp. 87094–87155, 2021.

[13] Y. Gao and Z. Wang, "A review of p4 programmable data planes for network security," *Mobile Information Systems*, vol. 2021, pp. 1–24, 11 2021.

[14] D. Scholz, A. Oeldemann, F. Geyer, S. Gallenmüller, H. Stubbe, T. Wild, A. Herkersdorf, and G. Carle, "Cryptographic hashing in p4 data planes," pp. 1–6, 09 2019.

[15] S. Yoo and X. Chen, "Secure keyed hashing on programmable switches," SPIN '21, (New York, NY, USA), p. 16–22, Association for Computing Machinery, 2021.

Avaliação de esquemas pós-quânticos de assinatura baseada em atributos

Érico Rolim , Marco Henriques
{e170610@dac.unicamp.br, maah@unicamp.br}

Departamento de Engenharia de Computação e Automação (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Campinas, SP, Brasil

Resumo – Esquemas de assinaturas baseadas em atributos (attribute-based signatures — ABS) utilizando curvas elípticas já foram muito estudados, com múltiplas implementações validadas. Entretanto, não se pode dizer o mesmo sobre esquemas baseados em reticulados, apesar das principais propostas de esquemas pós-quânticos se basearem neles. Isso torna necessário o estudo mais profundo do uso de reticulados nesse tipo de esquema. Esse trabalho busca implementá-los de forma eficiente, ao mesmo tempo que avalia o nivel de segurança que oferecem.

Palavras-chave – post-quantum lattice attribute signature

1. Introdução

1.1. Assinaturas baseadas em atributos

Assinaturas baeadas em atributos são uma classe de construção criptográfica com propriedades interessantes de proteção de privacidade. Permitem a criação de assinaturas onde o foco é nos atributos do indivíduo que gerou a assinatura, sem reveler a identidade desse indivíduo [4]. Atributos podem ser de vários tipos, como o fato do indivíduo ser aluno de um determinado curso ou instituição, ou pertencer a uma certa classe trabalhadora em uma empresa. Normalmente, esses atributos são fornecidos e certificados por uma entidade central.

Há, ainda, aplicações menos óbvias da tecnologia, como em frotas de veículos que se comunicam para transmitir informações sobre movimento e condições de vias [1]: uma entidade central distribui identidades para todos os veículos, e eles podem se comunicar sem revelar sua identidade por completo para outros veículos.

O mecanismo básico de operação da assinatura baseada em atributos é o mesmo, independentemente dos detalhes do esquema criptográfico de cada proposta. Há quatro etapas distintas [3]:

- Setup (configurar): a entidade gera sua chave secreta e define quantos (e talvez quais) atributos serão suportados
- Extract (extrair): a entidade cria uma chave para um certo indíviduo, correspondente aos atributos possuídos pelo indivíduo
- Sign (assinar): o indíviduo cria uma mensagem, e a assina com sua chave; dependendo do esquema criptográfico utilizado, pode ser possível não assinar a mensagem com todos os atributos possuídos pelo indivíduo
- Verify (verificar): um terceiro verifica a mensagem e a assinatura, confirmando que foi assinada por alguém com certos atributos, mas sem saber a identidade desse indivíduo

1.2. Criptografia pós-quântica

Ocorreram muitos avanços em computação quântica nos últimos anos, levando a uma preocupação no mundo da criptografia: os esquemas criptográficos modernos (baseados em grandes números primos e curvas elípticas), se atacados por um computador quântico suficientemente poderoso, podem ser quebrados. Isso levou ao estudo de algoritmos e construções que não tenham as mesmas fraquezas [2].

Entretanto, esquemas de assinaturas baseadas em atributos, em sua maioria, utilizam emparelhamento de curvas elípticas [5]. Dos poucos esquemas utilizando reticulados, e, portanto, possivelmente resistentes a ataques de um computador quântico, não encontramos implementações concretas ou avaliações claras de suas propriedades de segurança ou eficiência.

2. Proposta

A proposta desse trabalho de conclusão de curso, portanto, é avaliar múltiplos aspectos dos esquemas de assinaturas baseadas em atributos presentes em artigos publicados nos últimos anos. As características que serão avaliadas são explicadas abaixo:

- Eficiência: uso de recursos computacionais e tamanho dos artefatos gerados (chaves e assinaturas)
- Segurança: qual o nível equivalente de segurança garantido pela prova de segurança do esquema

Para isso, será necessário implementar os esquemas descritos nos artigos encontrados, uma vez que nenhum conta com implementações públicas dos esquemas criptográficos.

O trabalho começará com uma avaliação mais profunda do esquema explicado em [3], que já foi implementado, com algumas mudanças, no repositório `https://github.com/regras/labs`, por um ex-aluno da FEEC.

Agradecimentos

Agradeço o grupo de estudo ReGrAS, liderado pelo professor Marco Henriques, pelas discussões e apresentações sobre criptografia pós-quântica.

Agradeço minha família por todo apoio e incentivo.

Referências

[1] Hui Cui, Robert H Deng, and Guilin Wang. An attribute-based framework for secure communications in vehicular ad hoc networks. *IEEE/ACM Transactions on Networking*, 27(2):721–733, 2019.

[2] Taniya Hasija, KR Ramkumar, Amanpreet Kaur, Sudesh Mittal, and Bhupendra Singh. A survey on nist selected third round candidates for post quantum cryptography. In *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, pages 737–743. IEEE, 2022.

[3] Xie Jia, Hu Yupu, Gao Juntao, Gao Wen, and Li Xuelian. Attribute-based signatures on lattices. *The Journal of China Universities of Posts and Telecommunications*, 23(4):83–90, 2016.

[4] Hemanta K Maji, Manoj Prabhakaran, and Mike Rosulek. Attribute-based signatures. In *Cryptographers' track at the RSA conference*, pages 376–392. Springer, 2011.

[5] Prince Silas Kwesi Oberko, Victor-Hillary Kofi Setornyo Obeng, Hu Xiong, and Saru Kumari. A survey on attribute-based signatures. *Journal of Systems Architecture*, page 102396, 2022.

FEEC | UNICAMP
EADCA22
XIV ENCONTRO DE ALUNOS E DOCENTES DO DCA

Sessão Técnica 6

Como a plataforma ION garante a segurança das identidades digitais descentralizadas

Rodrigo S. P. Hirao, Marco A. Amaral Henriques
r186837@dac.unicamp.br, maah@unicamp.br

Departamento de Engenharia de Computação e Automação (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Campinas, SP, Brasil

Resumo – Sistemas de gestão de identidades digitais são usualmente feitos utilizando um sistema central para o armazenamento e prova dos dados, o que pode causar desconfiança sobre a segurança e privacidade envolvida durante o processo. Foi feito o esforço pela W3C de especificar uma arquitetura padrão para identidades digitais descentralizadas, que se caracterizam por colocar sob a guarda e controle dos seus usuários seus dados de identificação e demais metadados relacionados, gerando assim o conceito de identidade autossoberana. Essa especificação foi utilizada para a definição do protocolo Sidetree. A plataforma ION é a implementação do protocolo Sidetree utilizando a blockchain da criptomoeda Bitcoin (como âncora de confiança), o sistema de arquivos distribuído IPFS (como sistema de armazenamento endereçado por conteúdo) e a base de dados MongoDB (como cache da blockchain). Utilizando as ferramentas e especificações disponibilizadas pela ION foi estudado, por diversos métodos, o fluxo de uma identidade desde sua criação até sua resolução, para, assim, poder verificar como o sistema garante a segurança envolvida em cada etapa.

Palavras-chave – identidade descentralizada, identidade autossoberana, plataforma ION, blockchain

1. Introdução

Em diversos contextos é necessária a identificação do usuário para o uso de um serviço e para isso é usado uma identidade acompanhada de uma credencial. A identidade possibilita a identificação única do usuário dado o contexto e a credencial atesta a autoridade do usuário sobre essa identidade [4].

As identidades podem ser classificadas em 3 categorias [7]. A primeira é a identidade física, que está sob o controle do usuário [8]. Um exemplo é o RG, que pode ser guardado na carteira e usado quando o usuário achar preciso, tendo como credenciais a foto do rosto e a assinatura. A segunda categoria é a do documento de identidade digital, que nada mais é que uma representação digital do documento físico. A terceira é a do documento de identidade eletrônica, a identidade que foi criada digitalmente e não possui uma representação física. Esta identidade precisa de um provedor de identidades (*IdP*) relacionada a um provedor de serviço (*SP*).

Uma identidade centralizada contém um *SP* que também é o *IdP*, assim o SP tem o controle total dos dados do usuário. Isso pode criar problemas como o usuário ter que gerenciar diversas identidades para diversos *SP*s.

Uma outra solução então é uma identidade federada, onde um *IdP* é feito por um serviço terceiro de confiança e o *SP* se comunica com o *IdP* para provar as credenciais. Embora esse método evite o problema anteriormente citado ele mantém o controle sobre a identidade por parte de um provedor, comprometendo a privacidade e segurança [10].

Uma outra solução, mais recente, é a adoção do conceito de identidades autossoberanas (*SSI*) [4], onde o usuário tem um maior controle de suas informações privadas armazenadas consigo e não depende totalmente de um terceiro para se autenticar [8]. Porém tal solução pode ser mais complexa na implantação e uso cotidiano, pois usuários leigos podem ter maior dificuldade para provar o controle de sua identidade.

Para obter uma *SSI* digital foi proposto o modelo de Identidades Digitais Descentralizadas (*DID*), com os objetivos de ser descentralizado, persistente, verificável criptograficamente e resolvível [5]. Dispondo de uma solução para o modelo de *SSI* que consiga provar sua identidade sem a necessidade de uma IdP, um usuário poderá fazer a autenticação diretamente com o provedor de serviços (*SP*).

Foi feito um esforço pela *W3C* para definir padrões na implementação de *DIDs*[9], de modo a deixar as tecnologias usadas acima do protocolo como uma escolha da implementação do protocolo. Assim foi criado o protocolo *Sidetree*[2] que segue as especificações da *W3C* possibilitando a escolha de uma tecnologia para sistema de ancoragem de dados sensíveis, uma para armazenamento endereçado por conteúdo (*CAS*) e uma para banco de dados (*DB*) de cache, o que resultou na implementação

da plataforma *ION*, que usa a *Sidetree* com a *blockchain Bitcoin* como sistema de ancoragem, o sistema de arquivos distribuído *IPFS* (*InterPlanetary File System*) como *CAS* e a base de dados *MongoDB* como cache.

2. Proposta

Após o entendimento da especificação de DID pela W3C e do protocolo da Sidetree, a plataforma ION foi estudada detalhadamente por meio de seu código e documentações, com o fim de executar o protocolo diretamente acima da bitcoin e ipfs, sem utilizar as ferramentas fornecidas pela Microsoft, fortalecendo sua descentralização pela independência de ferramentas centralizadas.

Ao conseguir usar a DID em um nível mais baixo de implementação foi possível analisar as âncoras de confiança atreladas às tecnologias envolvidas, como a bitcoin e ipfs. Assim conseguimos verificar as assinaturas digitais e a gestão das chaves que as garantem e avaliar a segurança dessa implementação do protocolo.

Após a avaliação das âncoras de confiança, avaliaremos se há possíveis melhorias no protocolo a que possam ser feitas sem comprometer a segurança já estabelecida.

3. Especificações de DID pela W3C

A *W3C* define uma *DID* como uma *URI* (*Uniform Resource Identifier*) **did:{método}:{identificador}** Figura 1, sendo o **{método}** a definição de como a *DID* foi implementada e o **{identificador}** o identificador único da *DID* relativo ao método utilizado. Tal *URI* referencia um documento *DID* formatado como *JSON* (*Javascript Object Notation*) que possui provas criptográficas para garantir as propriedades da *DID*, como chaves públicas relacionadas à chave privada que apenas o controlador possui. Além disso, o documento também possui informações sobre todos os serviços referenciados pela *DID*, que podem também usar alguma chave do documento como prova criptográfica de que o serviço existiu e criou tal documento.

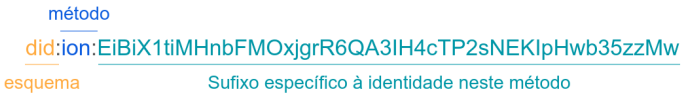

Figura 1. URI de uma DID de exemplo

4. O protocolo Sidetree

A *Sidetree* é o protocolo que implementa as especificações fornecidas pela *W3C* disponibilizando uma biblioteca em *NodeJS*, que pode ser implementada com qualquer interface para seu sistema de ancoragem, *CAS* e base de dados de cache. Para tanto a implementação da *Sidetree* precisa definir esses módulos de sistema de ancoragem, *CAS* e *DB* de cache. Na biblioteca disponibilizada já existem módulos de exemplo que são implementados com a *blockchain* da *Bitcoin* como sistema de ancoragem, o *IPFS* como *CAS* e o *MongoDB* como base de dados de cache.

4.1. Criação de uma DID

Durante o fluxo de criação de uma *DID*, é criada uma lista de objetos de atualização do documento e essa atualização pode ser uma criação de uma nova *DID*. Cada documento é derivado de um *"delta"* que define a diferença entre o estado anterior da *DID* e o atual, onde o estado anterior pode ser vazio no caso da inexistência da *DID*.

Cada operação do delta poderá adicionar, remover ou atualizar chaves públicas no formato de *JSON Web Key* (*JWK*) [6] e serviços, bem como definir seus papéis e objetivos. Após a criação do *delta*, este deverá passar por um processo de *hash sha256* e codificação em *base64url* resultando no que é chamado de *"deltaHash"*. Tendo o *delta* e o *deltaHash*, podemos enviar os dados usando *HTTP* a um servidor executando uma implementação da *Sidetree* com o método escolhido, que por sua vez irá inserir estes dados em uma fila onde o *deltaHash* é a chave de busca e o delta é o resultado (Figura 2).

Um programa chamado Escritor em Lote é executado no nó em paralelo (Figura 3), a fim de consumir a fila para inserção e criar o registro no *CAS* desejado (*IPFS* no caso do ION), o qual é chamado de *Chunk File*. Em seguida é criado um arquivo de provas para operações de atualização chamado de *Provisional Proof File* e um terceiro arquivo apontando para os endereços tanto do *Chunk File* quanto do *Provisional Proof File*, chamado de *Provisional Index File*. Por fim é criado um arquivo para provas de linhagem da *DID* chamado de *Core Proof File* e um arquivo final chamado de *Core Index File* que terá os endereços do *Core Proof File* e do *Provisional Index File*, criando uma árvore com o *Core Index File* como raiz.

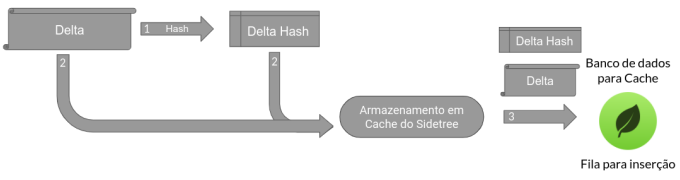

Figura 2. Criação de uma DID na fila de cache.

Feitas as operações desejadas no *CAS*, o *Core Index File* será adicionado, concatenado com o número de

operações existentes na forma **{número de operações}.{endereço do core index file}** no sistema de ancoragem, como pode ser visto na Figura 3.

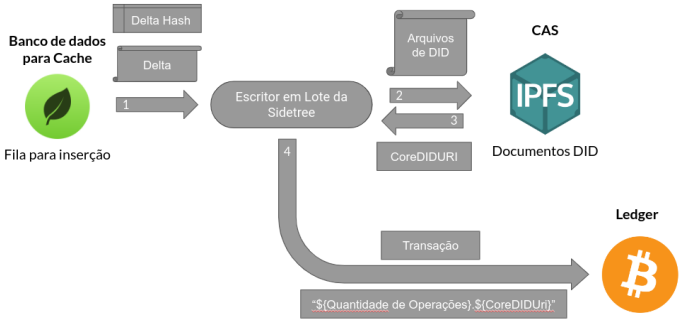

Figura 3. Criação de uma DID no CAS e no sistema de ancoragem a partir da fila.

4.2. Resolução de uma DID

Para resolver uma *DID* pela *URI* podemos seguir 2 caminhos, dependendo de sua formatação: caso ela esteja no formado de *URI* longa (**did:{método}:{deltaHash}:{base64(delta)}**) basta decodificar a {base64(delta)}. Porém, para uma *URI* curta (Figura 1) não conseguimos obter de imediato o conteúdo da *DID*.

É preciso percorrer o sistema de ancoragem procurando pelas operações *DID* do método indicado pela *URI* e, um por um, devemos pegar o *Core Index File* e percorrer a árvore no *CAS* até chegarmos nos *Chunk Files*, que irão descrever os deltas dessa transação no sistema de ancoragem. Em seguida, pegamos os deltas obtidos e guardamos em um banco de dados intermediário de cache usando o *hash* do delta como chave, como pode ser visto na Figura 4.

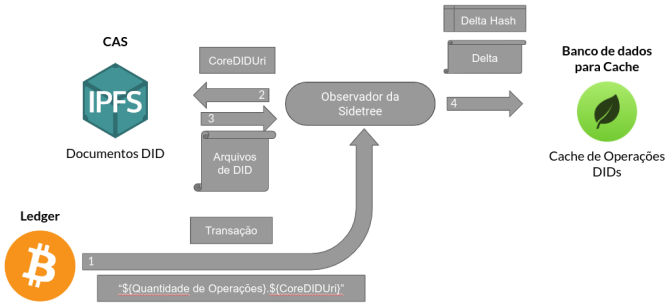

Figura 4. Resolução de uma DID para o cache

Após ter adicionado todos os deltas encontrados no sistema de ancoragem no banco de dados intermediário é possível encontrar rapidamente o documento desejado (por meio do hash do delta presente na *URI* da *DID*) no cache criado usando o banco de dados intermediário (Figura 5).

Figura 5. Resolução de uma DID a partir do cache.

5. ION: uma implementação da Sidetree

A Sidetree possibilita uma implementação com escolha de *CAS* e Sistema de ancoragem, assim surgindo implementações diversificadas, como a *Element* que usa a *Ethereum* com a *IPFS*, ou a photon que usa a *Amazon QLDB* com a *Amazon S3*. A *ION* utiliza a *blockchain* da *Bitcoin*, afirmando que essa é a *blockchain* aberta mais segura [3], em conjunto com o *IPFS* (como *CAS*).

A *ION* oferece uma ferramenta para implementação de um servidor *ION* (chamado de nó), usando a *API* da *Sidetree*, que já possui a blockchain de *Bitcoin* e o sistema de arquivos distribuído *IPFS* implementados como padrão, criando assim pontos de conexão (*endpoints*) *http* do tipo *rest* para gerenciamento básico de uma *DID*, como criação e resolução, por exemplo [1].

6. Resolução de URI em ION

Para avaliar se o entendimento do protocolo está correto e se o estabelecimento das âncoras de confiança foi feito corretamente para a *DID* da Figura 1, utilizamos as ferramentas disponibilizadas pela *ION* e acompanhamos passo a passo a resolução de uma *DID*, conforme a Figura 6.

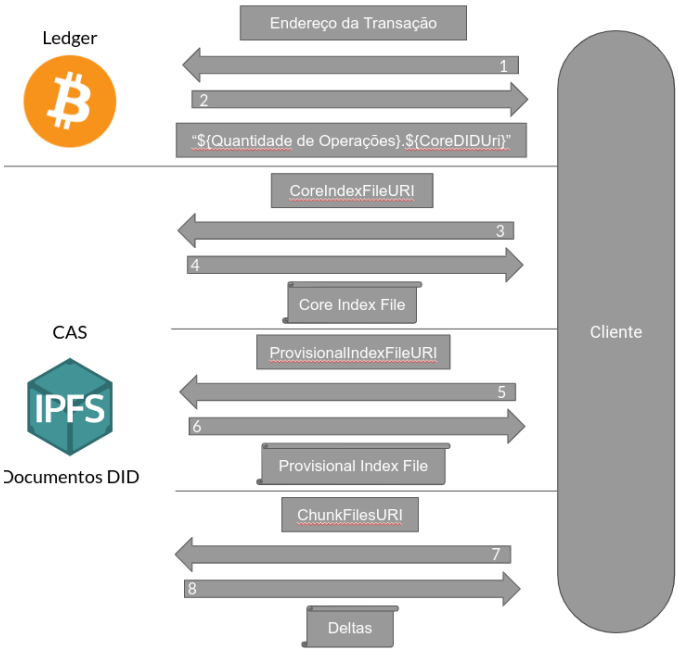

Figura 6. Fluxo para resolução manual da DID.

O problema inicial encontra-se ao tentar descobrir o endereço de transação, para tanto, foi necessário observar a bitcoin durante a ancoragem da *DID*, o que resultou em uma única transação com o padrão *"ion:1.{código}"* durante um intervalo de vários minutos, indicando similaridade com o que foi ancorado recentemente.

Usamos o *hash {código}* para buscar o Core Index File em um domínio disponível para busca na IPFS, o que retorna um arquivo compactado com o *Core Index File*. Este possui o endereço para o *Provisional Index File*.

Assim, repetindo o passo anterior e buscando pelo arquivo na *IPFS* temos o *Provisional Index File* compactado, que, após descompactado, resulta no *JSON* com o endereço de um *Chunk File*.

Com o endereço do *Chunk File* podemos repetir novamente o processo e buscar o arquivo compactado na *IPFS*, para assim conseguir o delta.

Com esses 3 arquivos podemos derivar o documento *DID* referente à Figura 1, por estes possuírem todas as informações necessárias para a composição do documento. Assim, constatamos que foi possível percorrer todos os passos necessários para identificar e encontrar um recurso na blockchain e no IPFS a fim de verificar como estão asseguradas as garantias de confiança no documento *DID*.

7. Conclusões e trabalhos futuros

Devido à busca por soluções autossoberanas de identidades a solução de *DIDs* se mostrou interessante, tendo um padrão criado pela *W3C* para sua implementação e um protocolo acima desse padrão (*Sidetree*) fornecendo uma fácil implementação de diferentes métodos de *DID*, como a *ION*, que funciona usando a *Bitcoin* como sistema de ancoragem e depende da *IPFS* para armazenar os documentos *DID* gerados.

O armazenamento *IPFS*, ao indexar seus arquivos usando um *hash* do conteúdo do arquivo, impediu o arquivo de ser alterado sem mudar seu índice. Porém evidenciou a necessidade de um sistema de ancoragem imutável para provar a ordem e atualidade de seus documentos *DID*, uma vez que a *DID* pode ter sido removida ou atualizada em uma transação futura, mas o documento antigo continua na rede *IPFS*.

Observando o funcionamento passo a passo da resolução de uma *DID ION* foi possível perceber sua dependência da imutabilidade e persistência da *Bitcoin*, uma vez que é possível adquirir uma *DID* desatualizada, mesmo se for encontrado o bloco correto na blockchain da *Bitcoin*. Assim, é necessário fazer uma leitura completa da *blockchain* pelo menos uma vez.

Este fato torna difícil a resolução de DIDs sem a utilização de um servidor executando um serviço *ION* com um banco de dados de cache.

Não foi contemplado neste artigo o funcionamento das operações de atualização, remoção e recuperação da *DID*. Outro trabalho futuro seria interessante analisar as diferentes possibilidades de sistemas de ancoragem, no lugar da *Bitcoin*, que possui um elevado tempo entre suas transações e pode ter um custo muito elevado para o contexto de *DIDs*.

Referências

[1] Daniel Buchner. Did method implementation using the sidetree protocol on top of bitcoin, 2020. https://github.com/decentralized-identity/ion Acessado em 26 Junho 2022.

[2] Daniel Buchner, Orie Steele, and Troy Ronda. Sidetree v1.0.0, 2021. identity.foundation/sidetree/spec/ . Acessado em 26 Junho 2022.

[3] Daniel Buchner and Henry Tsai. Q&a about ion, 09 2021. https://github.com/decentralized-identity/ion/blob/66813123cf81ace05cea2039e93ef263952d6283/docs/Q-and-A.md Acessado em 26 Junho 2022.

[4] Md Sadek Ferdous, Farida Chowdhury, and Madini O. Alassafi. In search of self-sovereign identity leveraging blockchain technology. *IEEE Access*, 7:103059–103079, 07 2019.

[5] Kim Hamilton-Duffy, Ryan Grant, and Adrian Gropper. Use cases and requirements for decentralized identifiers, 2021. https://www.w3.org/TR/did-use-cases/ . Acessado em 26 Junho 2022.

[6] M. Jones. Json web key (jwk), 05 2015. https://datatracker.ietf.org/doc/html/rfc7517 Acessado em 26 Junho 2022.

[7] Frederico Schardong and Ricardo Custódio. Self-sovereign identity: A systematic map and review. *ACM Comput. Surv. 1*, Article 1, 08 2021.

[8] Johannes Sedlmeir, Reilly Smethurst, Alexander Rieger, and Gilbert Fridgen. Digital identities and verifiable credentials. *Bus Inf Syst Eng*, 63:603–613, 12 2020.

[9] Manu Sporny, Dave Longley, Markus Sabadello, Drummond Reedand, Orie Steele, and Christopher Allen. Decentralized identifiers (dids) v1.0, 2021. https://w3c.github.io/did-core/ . Acessado em 26 Junho 2022.

[10] Andrei Volkov. Addressing the challenges facing decentralized identity systems. *iSChannel*, pages 10–15, 09 2020.

Aplicação de esquemas de assinatura digital pós-quântica baseada em reticulados para hardwares restritos

Rodrigo Duarte de Meneses , Marco Aurélio Amaral Henriques
{r197962@dac.unicamp.br, maah@dca.fee.unicamp.br}

Departamento de Engenharia de Computação e Automação (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Campinas, SP, Brasil

Abstract – Neste artigo, apresentamos o esquema de assinatura digital pós-quântica Dilithium, parte da suíte CRYSTALS (Cryptographic Suite for Algebraic Lattices), composta por algoritmos criptográficos baseados em reticulados. O Dilithium é um dos candidatos ao processo de padronização de algoritmos pós-quânticos do NIST (National Institute of Standards and Technology), e possui grande relevância em aplicações no contexto de ambientes computacionais restritos. As tentativas de otimização dos recursos computacionais do Dilithium para viabilizar sua implementação em hardwares restritos comumente buscam aprimorar os cálculos associados à NTT. Entretanto, a implementação do Dilithium com funções de hash criptográfico alternativas constitui uma opção válida para potenciais otimizações do algoritmo. Este trabalho busca analisar o impacto das funções de hash nos requisitos de processamento do Dilithium para sua posterior aplicação em ambientes computacionais restritos.

Keywords – criptografia pós-quântica, criptografia baseada em reticulados, assinaturas digitais

1. Introdução

1.1. Motivações e relevância

De acordo com os padrões vigentes de criptografia de chave simétrica, os criptossistemas mais populares são baseados no RSA ou em curvas elípticas. A segurança desses esquemas é associada à dificuldade do problema do logaritmo discreto e de fatoração de primos.

Com o advento da computação quântica, a viabilização de algoritmos como o algoritmo de Shor [12] mostra que esses problemas podem ser resolvidos em tempo polinomial [11]. Com isso, a segurança atribuída aos criptossistemas mais amplamente utilizados foi comprometida. O desafio de desenvolvimento de esquemas de encriptação resistentes aos ataques de um computador quântico configura a área da criptografia pós-quântica.

Em 2016, o NIST anunciou um processo de padronização de algoritmos pós-quânticos [7]. Dentre os algoritmos finalistas do terceiro round, destacam-se os algoritmos de assinatura digital pós-quântica baseados em reticulados – FALCON e CRYSTALS-Dilithium.

O FALCON utiliza-se do criptossistema NTRU e apresenta a maior velocidade de verificação e também os menores tamanhos de chave necessários [8]. Entretanto, a utilização do criptossistema NTRU constitui uma dificuldade para sua implementação em ambientes computacionais restritos. O Dilithium, por sua vez, não depende do criptossistema NTRU para o seu funcionamento e pode ser melhor aplicado no contexto de hardwares restritos.

1.2. Princípios de funcionamento

O problema básico sob o qual o Dilithium adquire sua segurança é o problema MLWE (Module Learning With Errors) [10], esquematizado na Figura 1.

O MLWE consiste em uma matriz $\mathbf{A}$ e um conjunto de vetores $\mathbf{t}$, $\mathbf{s_1}$ e $\mathbf{s_2}$ com seus elementos em um anel polinomial R. Esse problema pode ser descrito como a dificuldade de distinguir um vetor qualquer $\mathbf{t} \in R^k$ de um vetor da forma $\mathbf{t} = \mathbf{A} \cdot \mathbf{s_1} + \mathbf{s_2}$. Esse problema é uma generalização do problema LWE (Learning With Errors), onde tomamos $R = \mathbb{Z}_q$. Os problemas MLWE e LWE são considerados difíceis [9] [6].

$$\mathbf{t} = \begin{pmatrix} t_1 \\ \vdots \\ t_k \end{pmatrix} = \underbrace{\begin{pmatrix} a_{1,1} & \dots & a_{1,l} \\ \vdots & \ddots & \vdots \\ a_{k,1} & \dots & a_{k,l} \end{pmatrix}}_{\text{uniform, public}} \underbrace{\begin{pmatrix} s_{1,1} \\ \vdots \\ s_{1,l} \end{pmatrix}}_{\text{short}} + \underbrace{\begin{pmatrix} s_{2,1} \\ \vdots \\ s_{2,k} \end{pmatrix}}_{\text{short}}$$

Figura 1. Esquema básico do problema MLWE.

No Dilithium, a matriz $\mathbf{A}$ e o vetor $\mathbf{t}$ são tomados como parâmetros públicos e os vetores $\mathbf{s_1}$ e $\mathbf{s_2}$ constituem a chave privada. O vetor $\mathbf{s_2}$ é também chamado de vetor de erros, por ser responsável por incluir os ruídos no cálculo de $\mathbf{t}$. O anel polinomial utilizado no esquema é da forma $R = \mathbb{Z}_q[X]/(X^n + 1)$, com $q = 8.380.417$ e $n = 256$.

A maior dificuldade, portanto, advém do armazenamento dos parâmetros $\mathbf{A}$, $\mathbf{t}$, $\mathbf{s_1}$ e $\mathbf{s_2}$. Por se tratarem de matrizes e vetores de polinômios, o armazenamento desses parâmetros ocupa muito espaço de memória. A solução utilizada é gerar os parâmetros a partir de uma seed ρ através de uma função de hash criptográfico. Com isso, é necessário que armazenemos apenas a seed ρ, e os parâmetros são gerados conforme a demanda.

Para esse propósito são empregadas as rotinas ExpandA, ExpandS e ExpandMask, chamadas de funções de expansão. Essas funções são responsáveis por gerar os parâmetros necessários a partir de uma seed aleatória ρ.

1.3. Trabalhos relacionados

As tentativas de otimização do Dilithium para implementação em ambientes computacionais restritos [5] [3] costumam ter como objetivo aumentar sua eficiência através de otimizações no cálculo da NTT (Number Theoretic Transform), uma forma particular da transformada discreta de Fourier para corpos finitos. Essas otimizações são importantes no sentido de agilizar as multiplicações entre polinômios, um dos processos de maior custo computacional no Dilithium [4]. No entanto, as otimizações propostas não aumentam a eficiência das funções de expansão (ExpandA, ExpandS e ExpandMask), utilizadas em todo algoritmo.

Diante disso, esta pesquisa busca analisar a possibilidade implementação do Dilithium com diferentes funções de hash criptográfico e seu desempenho em hardwares restritos.

2. Proposta

Tendo em vista que a utilização das funções de hash criptográfico é um dos aspectos mais computacionalmente custosos no Dilithium [4], é válido analisar como essas funções são utilizadas.

As funções de hash são empregadas majoritariamente nas funções de expansão. Como visto, as funções de expansão são responsáveis por gerar parâmetros do algoritmo a partir de seeds amostradas de uma distribuição uniforme. Na implementação original do Dilithium, são utilizadas as funções SHAKE-128 e SHAKE-256 para o cálculo dos hashes.

Na documentação submetida para o round 3 do NIST [4] é apresentada uma implementação alternativa utilizando o AES-ctr (modo counter) como função de hash. Essa comparação visa explicitar a eficiência do SHAKE em relação ao AES-ctr quando implementados em software.

No entanto, os resultados disponíveis na literatura [1] apontam para um overhead computacional associado a essas funções para o cálculo de hash. Com o intuito de implementação em ambientes restritos, a utilização de funções de hash com menores requisitos de memória e processamento consitui uma possível complementação às otimizações nos cálculos da NTT.

3. Resultados

A partir da proposta de otimização através das funções de hash criptográfico, buscamos determinar se o impacto dessas funções é de fato significativo nos requisitos de memória e processamento do Dilithium. Fazemos esta análise determinando a contribuição percentual das funções de expansão (que correspondem a maior parte da utilização das funções de hash) nos ciclos de processador utilizados para as principais rotinas do Dilithium.

A Tabela 1 mostra a média e mediana dos ciclos de processador para cada uma das três funções de expansão utilizadas, calculadas para 100.000 iterações do algoritmo. Para cada iteração é feita a geração de chaves, assinatura de uma mensagem aleatória de 59 bytes e verificação dessa assinatura. A decisão de apresentar também a mediana foi feita devido à presença de alguns valores atípicos (outliers), fenômeno esperado por conta da utilização de rejection sampling no Dilithium.

Adotamos o set de parâmetros 3 indicado na documentação [4], e todas as medidas foram feitas por um laptop com processador Intel Core i5-8265U CPU 1.6 GHz, SO Ubuntu 20.04.4 LTS e memória RAM de 8 Gb.

-	ExpandA	ExpandS	ExpandMask
Média	161.088	2.718	4.670
Mediana	145.069	2.398	4.566

Tabela 1. Medidas dos ciclos de processador para as funções de expansão.

Na Tabela 2 são apresentados os valores de média e mediana das três funções principais do Dilithium: KeyGen (geração de chaves pública e privada), Sign (procedimento de assinatura) e Verify (verificação de assinatura).

-	KeyGen	Sign	Verify
Média	305.068	1.236.360	304.456
Mediana	299.730	992.632	296.411

Tabela 2. Medidas dos ciclos de processador para as funções principais.

A partir da frequência com que cada uma das funções de expansão é utilizada nas rotinas principais do Di-

lithium, pode-se analisar a contribuição de cada uma das funções de expansão para as rotinas principais, da perspectiva da média de ciclos de processador. Isso é feito calculando a razão entre os ciclos de processador de cada uma das funções de expansão em relação às rotinas principais. Por fim, somamos as contribuições das funções de expansão de forma a obter a sua contribuição geral para cada uma das rotinas. Os resultados obtidos são expressos na Tabela 3.

-	KeyGen	Sign	Verify
ExpandA	52.8%	13.0%	52.9%
ExpandS	0.8%	0%	0%
ExpandMask	0%	0.3%	0%
Σ	53.6%	13.3%	52.9%

Tabela 3. Percentual de contribuição de cada uma das funções de expansão para a média de ciclos de processador das funções principais.

Percebemos, da Tabela 3, que as funções de expansão correspondem, em média, a mais de metade dos ciclos de processador utilizados para as funções de geração de chaves e verificação de assinatura. Esse resultado respalda a hipótese de que as funções de hash possuem um impacto substancial nos requisitos de processamento do algoritmo.

4. Conclusões

A fim de viabilizar a aplicação do Dilithium em ambientes computacionais restritos, buscou-se determinar quais rotinas do algoritmo são responsáveis pelo maior percentual de processamento realizado. Conforme indicado pelos resultados, as funções de expansão correspondem à grande parte dos ciclos de processador. Assim, indica-se a otimização das funções de hash como uma potencial abordagem no sentido de possibilitar a aplicação do Dilithium em hardwares restritos.

Com a possibilidade de otimização do Dilithium a partir de funções de hash criptográfico alternativas, pode-se apontar como posteriores etapas desta pesquisa a implementação do Dilithium com as funções de hash da família BLAKE [2], ou o KangarooTwelve, que aparentam ser opções com menor custo computacional para o cálculo de hashes. A partir dessas possíveis otimizações, podemos buscar a aplicação do Dilithium em ambientes computacionais restritos, tais quais os microcontroladores ARM Cortex M0+ e Cortex M4.

Agradecimentos

Agradecemos ao suporte do Research Group of Applied Security (ReGrAS), da Faculdade de Engenharia Elétrica e Computação (FEEC), para elaboração da pesquisa, bem como para as discussões que se provaram imprescindíveis em sua produção.

Referências

[1] Jean-Philippe Aumasson. Too much crypto. *Real-World Crypto*, May 2020.

[2] Jean-Philippe Aumasson et al. Sha-3 proposal blake. `https://decred.org/research/aumasson2010.pdf`, 2010. (acessado em 08/08/2022).

[3] Luke Beckwith et al. High-performance hardware implementation of crystals-dilithium. *2021 International Conference on Field-Programmable Technology (ICFPT)*, November 2021.

[4] Vadim Lyubashevsky et al. Crystals-dilithium: Algorithm specifications and supporting documentation. `https://pq-crystals.org/dilithium/`, 2021. (acessado em 08/08/2022).

[5] Youngbeom Kim et al. Crystals-dilithium on armv8. *Hindawi Security and Communication Networks*, 2022.

[6] Daniele Micciancio. *Complexity of Lattice Problems*. Kluwer Academic, 2002.

[7] National Institute of Standards and Technology. Request for comments on post-quantum cryptography requirements and evaluation criteria. `https://csrc.nist.gov/projects/post-quantum-cryptography`, 2016. (acessado em 08/08/2022).

[8] Thomas Prest. 3rd round update on the falcon candidate algorithm. `https://csrc.nist.gov/Presentations/2021/falcon-round-3-presentation`, 2021. (acessado em 08/08/2022).

[9] Oded Regev. On lattices, learning with errors, random linear codes, and cryptography. *Journal of the ACM*, 2009.

[10] Gregor Seiler. Rystals-dilithium: A lattice-based digital signature scheme. *Conference on Cryptographic Hardware and Embedded Systems*, September 2018.

[11] Peter W. Shor. Algorithms for quantum computation: discrete logarithms and factoring. *Proceedings of the 35th Annual Symposium on Foundations of Computer Science*, pages 124–134, November 1994.

[12] Peter W. Shor. Polynomial time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Journal on Computing*, 1997.

Hierarchical control-plane and advanced telemetry in the context of programmable packet-optical networks

Rossano P. Pinto and Christian E. Rothenberg

{rossano@gmail.com, chesteve@dca.fee.unicamp.br}

Department of Computer Engineering and Industrial Automation (DCA)

School of Electrical and Computer Engineering (FEEC)

State University of Campinas (Unicamp)

Campinas, SP, Brasil

Abstract – In SDN, the control plane is usually delegated to a centralized off-premise controller. In some situations it's desired to have some on-premise or even embedded control. Hierarchical control plane split actions between local agents and a global controller. In order to make a local and a global actor work in a synchronized way, they must coordinate they're actions. With the use of advanced telemetry it's possible to monitor network changes and act accordingly in a fast and timely fashion.

Keywords – Hierarchical control plane, ONOS, P4, gNMI, streaming telemetry, packet-optical networks.

(Texto removido pela cláusula de sigilo dos convênios com as empresas.)

Guided assembly using deep learning-based object recognition and augmented reality instructions

Mariana Zaninelo Reis, Prof. Dr. José Mario De Martino

{m223752@g.unicamp.br, martino@dca.fee.unicamp.br}

Departamento de Engenharia de Computação e Automação (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Campinas, SP, Brasil

Abstract – Efficient assembly is vital to guarantee quality and celerity in manufacturing processes. Usually, the assembly tasks in a non-robotic process are expressed by assembly instructions that provide text and drawing information, both in 2D and 3D, about the form and order of the execution of tasks. However, these instructions are rarely interactive and, depending on their complexity, hard to follow. The main objective of this master's work is to develop an approach capable of providing the user with interactive step-by-step guidance for assembly tasks. Deep learning techniques will be applied to recognize, in real-time, equipment parts, indicate the assembly order, and the coupling to other parts. The assembly instructions will be provided exploring augmented reality. We plan to implement the solution on a mobile device and assess and compare our solution with more traditional ones, considering assembly speed-up and perceptual feedback from users.

Keywords – Augmented Reality, Deep Learning, Guided Assembly, Object Recognition.

1. Introduction

Industrial production lines are usually automated for most of their activities. However, some specific assembly processes are still predominantly performed manually. For these tasks, execution guides are used that can be printed or made available on support computers.

These assembly guides have technical drawings, assembly instructions, as well as related processes that must be executed, such as measurements and verification of aspects that guarantee the quality of the product.

Depending on the level of complexity of the assembly, these guides are not always enough for the assemblers to perform their tasks in a simple way. The main reason for this is the fact that they are not interactive and do not allow any kind of verification of the activities they are carrying out.

New ways of providing these assembly instructions have been proposed recently. Among the main motivations are the intention to improve the user experience and the performance indices in the manufacturing lines, such as task execution time and assertiveness level, in addition to guaranteeing the quality of the operations being performed.

The use of augmented reality combined with real-time object recognition is an area of study that can help improve assembly instructions.

Augmented reality offers an interactive experience of a real-world environment enhanced by computer-generated images and information. In an augmented reality environment, the user does not lose his natural perceptions, which is essential in its use in an industrial setting. Augmented reality applications are already widely applied or studied in some areas, such as games, medical applications, industrial applications, product design and development, assembly lines and navigation.

The application of augmented reality to promote guided assembly can be beneficial in terms of enabling user interactions. Superimposing synthetic images on real objects and providing assembly instructions that promote greater interaction has proven to be favorable in carrying out complex tasks.

There are already studies showing gains over conventional methods of assembly instructions [2]. Benefits such as reduced assembly time and reduced level of difficulty in performing tasks can be observed. However, some aspects need to be analyzed in the use of augmented reality, such as the level of familiarization and acceptance of the technology by users, the type of device that will be used and its

implications, in addition to the costs and resources for implementation.

There are two ways to launch augmented reality experiences, with or without the use of markers. Applications that use markers, usually QR codes or some image to help position the synthetic imagery, are more common. However, their use is not always viable, depending on the scenario where the montage is carried out and the availability of space to adequately place the markers.

Markerless augmented reality experiences have been more discussed lately. However, depending on the application, it is necessary to reference or superimpose computer-generated images on real objects [1]. The combination of object recognition techniques has been explored in conjunction with augmented reality applications without the use of markers so that computer-generated images and information can be referenced.

Object recognition can improve the positioning of synthetic images [4], making the augmented reality experience more realistic. In guided assembly, object recognition can improve the interactive experience and help to guide the activities performed.

The evolution of algorithms based on deep learning for object recognition tasks made it possible to apply them in more complex activities and to obtain greater levels of assertiveness [5, 6]. Currently, there are models based on deep learning that can be used on mobile devices, covering their application area.

The use of neural networks, specifically deep learning, for object recognition allows state-of-art recognition of complex assembly parts.

2. Proposal

The main objective of this work is to develop a guided assembly solution that provides the user with instructions to perform tasks in an interactive way. Algorithms based on deep learning will be used to recognize equipment parts to be assembled, providing the correct assembly sequence in real-time. Augmented reality is the technology that will be responsible for relaying assembly instructions to the user.

A mini bench vise will be used as a case study. The bench vise, printed in a 3D printer, is shown in Figure 1. The reason for choosing this object is that it is composed of some parts, which will allow the study of the use of the proposed solution in tasks that require the execution of some steps.

To use algorithms based on deep learning for object recognition, it is necessary to create a dataset that will be used for training the neural network. The dataset will consist of photos of each piece, varying their dispositions and the background in which they are found.

Building a dataset with many samples, specifically in this case, where it is necessary to have photos of the parts, is not an easy task. Neural networks, in general, require a significant amount of data for their training. Therefore, we intend to use image augmentation techniques to improve the performance of the neural network.

The creation of new architectures is not foreseen in this work, the intention is to use existing models that are already known to be efficient for this type of task. Some architectures will be analyzed to define the one that best meets the needs of the project.

The use of augmented reality associated with the work proposal occurs because it allows interaction with the user and facilitates complex work instructions [7, 8].

A mobile device will be used to carry out the proof of concept, a cell phone or tablet. The intention is to place it on a support in front of the activity, so that the user can view the information available and have their hands free to carry out the assembly.

Figure 1: mini bench vise used as study object

As a way of analyzing the results, we intend to compare the assembly performed using conventional tasks guides and the proposed solution. The parameters of comparison are the execution time of the activity and the level of difficulty to perform the assembly. For the

analysis of the level of difficulty of the task, perceptive evaluations of the users will be necessary.

If the hypothesis initially raised is true, it is expected that there will be a decrease in the execution time of the activity, as well as in the level of difficulty for execution.

3. Results

Regarding the dataset, the photos that will initially compose it have already been taken. The images were divided into ten classes, nine of which correspond to individual pieces and one that includes multiple pieces in the same images. Each class has three hundred images, totaling a dataset with three thousand samples.

The images have already been separated and stored in folders related to their respective classes, which depending on the type of neural network to be used, is sufficient to be used as a form of labeling. In addition, the entire dataset was labeled using bounding boxes, aiming to test neural networks that use this type of annotation, such as, for example, the YOLO [3, 9, 10] network. An example of samples that are present in each of the ten classes in the dataset is illustrated in Figure 2.

Figure 2: dataset samples

4. Conclusions

No conclusions have yet been reached about the proposed work, however, it is expected that there will be a decrease in the time of the mapped assembly task, in relation to traditional methods of instructions, as well as an improvement in the user experience during execution.

It is also expected that the work will contribute to the development of augmented reality, as it will explore a scenario of interest in several areas, which is the association of real-time object recognition with markerless applications.

References

[1] AKGUL, O.; PENEKLI, H. I.; GENC, Y. **Applying Deep Learning in Augmented Reality Tracking**. 2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS). **Anais**... Em: 2016 12TH INTERNATIONAL CONFERENCE ON SIGNAL-IMAGE TECHNOLOGY & INTERNET-BASED SYSTEMS (SITIS). nov. 2016.

[2] ALVES, J. et al. **Comparing Spatial and Mobile Augmented Reality for Guiding Assembling Procedures with Task Validation**. 2019 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC). **Anais**... Em: 2019 IEEE INTERNATIONAL CONFERENCE ON AUTONOMOUS ROBOT SYSTEMS AND COMPETITIONS (ICARSC). abr. 2019.

[3] BAHRI, H.; KRČMAŘÍK, D.; KOČÍ, J. **Accurate Object Detection System on HoloLens Using YOLO Algorithm**. . Em: 2019 INTERNATIONAL CONFERENCE ON CONTROL, ARTIFICIAL INTELLIGENCE, ROBOTICS & OPTIMIZATION (ICCAIRO). IEEE Computer Society, 1 maio 2019. Disponível em: <https://www.computer.org/csdl/proceedings-article/iccairo/2019/357200a219/1iQ32yocCDm>. Acesso em: 9 ago. 2022

[4] GARON, M.; LALONDE, J.-F. Deep 6-DOF Tracking. **IEEE Transactions on Visualization and Computer Graphics**, v. 23, n. 11, p. 2410–2418, nov. 2017.

[5] LALONDE, J.-F. **Deep Learning for Augmented Reality**. 2018 17th Workshop on Information Optics (WIO). **Anais**... Em: 2018 17TH WORKSHOP ON INFORMATION OPTICS (WIO). jul. 2018.

[6] LI, W. et al. Integrated Registration and Occlusion Handling Based on Deep Learning for Augmented Reality Assisted Assembly Instruction. **IEEE Transactions on Industrial Informatics**, p. 1–11, 2022.

[7] NISHIHARA, A.; OKAMOTO, J. **Object recognition in assembly assisted by augmented reality system**. 2015 SAI Intelligent Systems Conference (IntelliSys). **Anais**... Em: 2015 SAI INTELLIGENT SYSTEMS CONFERENCE (INTELLISYS). nov. 2015.

[8] PARK, K.-B. et al. Deep learning-based smart task assistance in wearable augmented reality. **Robotics and Computer-Integrated Manufacturing**, v. 63, p. 101887, 1 jun. 2020.

[9] REDMON, J. et al. **You Only Look Once: Unified, Realtime Object Detection**. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). **Anais**... Em: 2016 IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR). jun. 2016.

[10] REDMON, J.; FARHADI, A. **YOLOv3: An Incremental Improvement**. arXiv, , 8 abr. 2018. Disponível em: <http://arxiv.org/abs/1804.02767>. Acesso em: 9 ago. 2022