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.

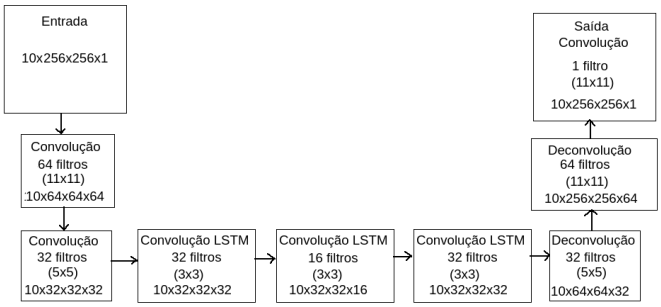

Figura 1. Ilustração da arquitura de *autoencoder* Espaço-Temporal montada.

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.

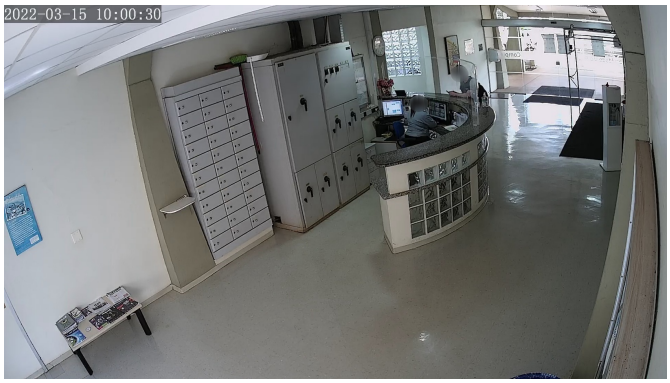

Figura 2. Exemplo de frame considerado cotidiano na entrada da FEEC-Unicamp.

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.

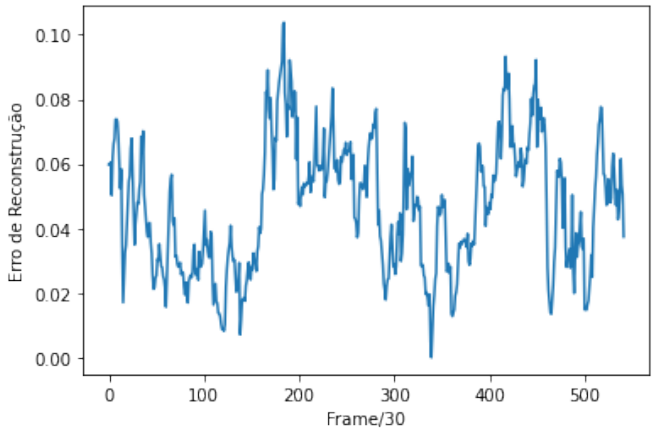

Figura 5. Curva do erro de reconstrução de um vídeo teste de conteúdo considerado cotidiano. Nota-se o pico de erro em 0.10, ou 10%.

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

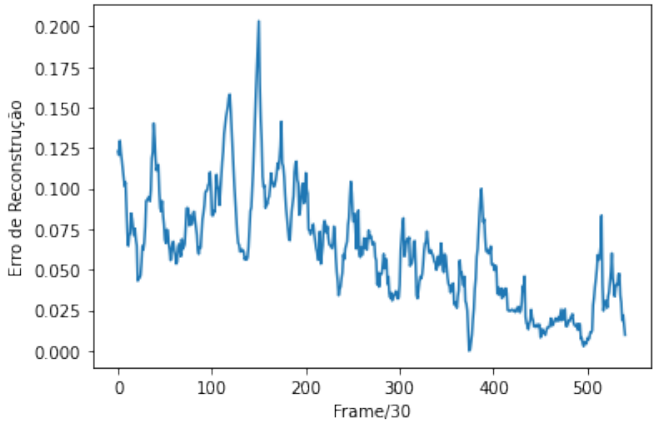

Figura 6. Curva do erro de reconstrução das sequências de teste por frame normalizado.

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.