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

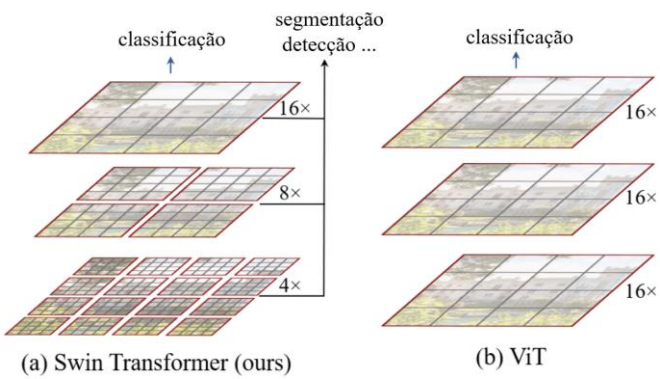

Figura 1. Comparação entre os mapeamentos (a) hierárquico do Swin Transformer e (b) tradicional do Vision Transformer. Adaptada de [7].

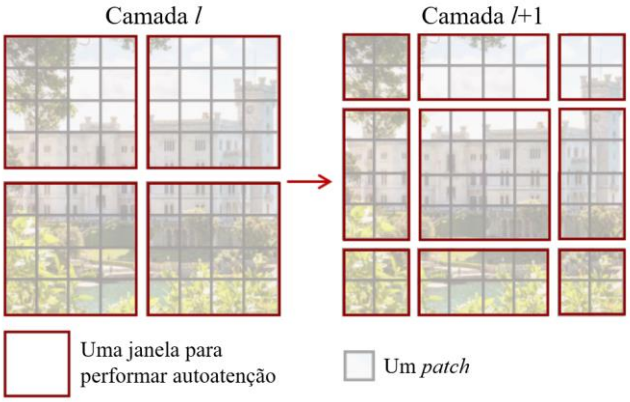

Figura 2. Uma ilustração da abordagem de janela deslocada. Adaptada de [7].

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.