Revisão sistematica das teorias norteadoras do Visual Transformers

Joelson S. Junior[1], Giancarlo Lucca[2], Diego Bottero[1],
Graçaliz P. Dimuro[1], Helida Santos[1]

[1]Centro de Ciências Computacionais – Universidade Federal do Rio Grande(FURG),
Rio Grande-RS, Brasil
[2]PGEEC – Universidade Católica de Pelotas (UCPEL), Pelotas-RS, Brasil

Abstract. *Transformers have emerged as a powerful architecture in Artificial Intelligence, revolutionizing various Natural Language Processing and image processing tasks. This paper presents a comprehensive analysis of the historical evolution of Transformers, emphasizing their self-attention mechanism and culminating with the introduction of the novel Visual Transformers model. We explore the main contributions made by key works in the field leading up to the development of Visual Transformers. Conducting a systematic analysis helps in gaining a deeper understanding of the functioning of this model and identifying the key topics for each approach.*

Resumo. *Os Transformers emergiram como uma poderosa arquitetura em Inteligência Artificial, revolucionando várias tarefas de processamento de linguagem natural e processamento de imagem. Este artigo apresenta uma análise abrangente sobre a evolução dos Transformers historicamente, destacando seu mecanismo de autoatenção e finalizando com o novo modelo de Transformers Visual. Exploramos as principais contribuições dos trabalhos-chave na área até o desenvolvimento do Visual Transformers. Realizar uma análise sistemática nos ajuda a entender melhor o funcionamento desse modelo e quais são os tópicos chaves para cada abordagem.*

1. Introdução

A inteligência artificial (IA) [Russell 2010] é uma área da ciência da computação que está em grande evolução conforme avançamos em algoritmos e em poder computacional, esses avanços nos possibilitaram revolucionar diversas áreas distintas da sociedade, tais como a indústria [Zhang and Man 1998], a saúde [Liu et al. 2021] e a educação [Ouyang and Jiao 2021]. Uma das abordagens que tem se mostrado cada vez mais relevantes em IA são as Redes Neurais [Abdi et al. 1999]. A sua aprendizagem, que tem como inspiração os neurônios do cérebro. Dessa forma, são realizadas diversas atividades complexas, que envolvem o reconhecimento de fala [Hinton et al. 2012], tradução de idiomas [Di Gangi et al. 2019] e toda uma vasta área envolvendo visão computacional [Jing et al. 2019].

Os modelos inspirados na estrutura do cérebro humano consistem em camadas de neurônios interconectados que processam informações em paralelo, resultando em predições precisas e eficientes. Nos últimos anos, o uso de redes neurais se tornou

cada vez mais comum em diversas áreas da computação, impulsionando avanços significativos na capacidade de processamento de dados e no desenvolvimento de algoritmos mais inteligentes e sofisticados [Valenzuela et al. 2023]. Além das redes neurais com múltiplos neurônios, outros modelos foram desenvolvidos visando resolver problemas de tarefas mais específicas de IA, as que mais se relacionam com este trabalho são as Redes Neurais Convolucionais (CNN, do inglês Convolutional Neural Network) [LeCun et al. 1998] e as Redes Neurais Recorrente (RNN, do inglês Recurrent Neural Network) [Rumelhart et al. 1986].

As RNNs foram apresentadas no artigo *Learning representations by back-propagating errors* [Rumelhart et al. 1986], onde os autores propõem uma arquitetura de rede neural que lida com dados sequenciais, como series temporais, escritas e sons. Essa arquitetura tem como principal ideia permitir que informações anteriores influenciem no processamento atual dos dados, sendo uma ótima ferramenta para o processamento de linguagem natural, reconhecimento de fala e previsões de séries temporais. Porém a possibilidade do treinamento de dados sequenciais com propagação de informação tem uma desvantagem que é a impossibilidade de realizar esse processamento o em paralelo, já que o próximo o dado a ser processado tem dependência do dado anterior.

Já as CNNs foram introduzidas pelo artigo *Gradient-Based Learning Applied to Document Recognition* [LeCun et al. 1998]. O trabalho utilizou essa nova arquitetura para a tarefa de reconhecimento de caracteres manuscritos, demonstrando um grande avanço em relação aos métodos tradicionais da época. A arquitetura proposta foi chamada de LeNet-5 e consistia em camadas convolucionais, camadas de subamostragem e camadas totalmente conectadas.

Sabendo que modelos de RNNs possuem uma limitação referente à paralelização de processos, uma nova arquitetura foi apresentada pelo artigo: *Attention is all you need* [Vaswani et al. 2017], esse estudo apresentou uma nova arquitetura chamada Transformers que, ao contrário das redes neurais convolucionais e recorrentes, foi projetada para processar sequências de dados em paralelo, sem a necessidade de manter uma memória interna de longo prazo.

Os Transformers são especialmente úteis em tarefas de processamento de linguagem natural, como sumarização de texto e resposta a perguntas, que exigem a compreensão de relacionamentos complexos entre diferentes partes da sequência. Porém percebeu-se também que esse novo modelo poderia ser utilizado para outros tipos de problemas como, por exemplo, em uma variação recente desse modelo, chamada de visual Transformer apresentada no trabalho *An image is worth 16x16 words: Transformers for image recognition at scale* [Dosovitskiy et al. 2020], que aplica com sucesso Transformers em tarefas de visão computacional, como classificação e segmentação de imagens.

Compreendendo toda a evolução até o momento para o surgimento de Transformers, esse trabalho apresenta uma construção histórica dos principais trabalhos relacionados aos Transformers que ajudaram a construir esse novo modelo hoje chamado de Visual Transformers.

Este trabalho está organizado da seguinte forma. Na Seção 2 são apresentados os principais elementos do Transformers no processo de linguagem natural. Na Seção 3, realizamos um estudo sobre o *BERT*, uma poderosa técnica de pré-treinamento que

impulsionou o avanço em tarefas de compreensão de linguagem natural. Na Seção 4, apresentamos uma abordagem que amplia a ideia de pré-treinamento autoregressivo para a compreensão de linguagem, permitindo uma maior generalização. Na Seção 5 exploramos a aplicação de Transformers para o reconhecimento de imagens em larga escala, demonstrando um marco significativo na extensão dos Transformers além do processamento de texto. Por fim, na Seção 6 apresentamos a conclusão do trabalho onde realizamos um breve comentário do impacto do desenvolvimento dos Transformers.

2. Transformers: Da linguagem natural à visão computacional

Para a realização desta análise histórica nesta seção, elencamos os principais artigos relacionados aos Transformers que contribuíram para a construção do *Visual Transformers* [Dosovitskiy et al. 2020]. Listamos nesta seção os artigos em ordem cronológica e, em seguida, os analisamos individualmente, destacando suas principais contribuições para o tema de pesquisa. Com isso, criamos a Figura 1, que representa a linha do tempo histórica do desenvolvimento dos artigos relacionados ao Visual Transformers.

Figure 1. Histórico até o Visual Transformers.

Conforme a Figura 1 apresentada, é evidente que após o lançamento do artigo pioneiro que introduziu a arquitetura dos Transformers em 2017, foram publicados outros dois artigos que desempenharam um papel crucial no desenvolvimento teórico subjacente à construção dos Transformers visuais, como descrito no artigo intitulado *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* [Dosovitskiy et al. 2020]. Esses trabalhos foram fundamentais para o avanço e aprimoramento dos Transformers aplicados à tarefa de reconhecimento de imagens em grande escala.

2.1. *Attention is All You Need*

Nos últimos anos, houve um notável avanço na área de redes neurais, especialmente em 2017, com a introdução do artigo *Attention is All You Need* [Vaswani et al. 2017]. Esse artigo trouxe à tona um novo modelo de rede neural chamado Transformer, que foi inicialmente desenvolvido para lidar com problemas relacionados ao processamento de linguagem natural. No entanto, à medida que esse modelo era explorado com mais profundidade, percebeu-se que ele poderia ser adaptado e utilizado em diversas outras aplicações no campo da inteligência artificial.

O modelo Transformer propõe uma arquitetura inovadora, que se baseia exclusivamente no mecanismo de atenção, dispensando os conceitos clássicos das redes neurais convolucionais e recorrentes. Essa abordagem revolucionária trouxe uma série de benefícios e melhorias em relação aos modelos anteriores, permitindo o processamento eficiente de abundância de dados e a captura de relações de longo alcance. O mecanismo de atenção, fundamental no modelo Transformer, pode ser representado pela seguinte equação:

$$\textit{Self-Attention}(Q, K, V) = \textit{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{1}$$

Na Equação 1 a matriz de *queries* Q é obtida multiplicando-se a representação da entrada pelos pesos da matriz W_q. De maneira similar, a matriz de *keys* K é obtida pela multiplicação da representação da entrada pelos pesos da matriz W_k, enquanto a matriz de *values* V é calculada a partir da representação da entrada e dos pesos da matriz W_v, a dimensão d_k é referente a dimensão da matriz K.

A operação de similaridade utilizada entre as matrizes de consulta Q e chave K é calculada por meio de uma operação de multiplicação de matrizes. Essa operação permite medir o grau de relacionamento entre os elementos das duas matrizes, indicando quais partes da entrada são mais relevantes para a tarefa em questão.

Após a multiplicação das matrizes Q e K, os resultados são normalizados para garantir que os pesos de atenção estejam na mesma escala. A normalização é crucial para a atenção ser distribuída adequadamente e para nenhum elemento dominar a contribuição da atenção. Uma técnica comum para a normalização é dividir os valores resultantes pela raiz quadrada da dimensão da matriz de consulta d_k.

Em seguida, os valores normalizados passam por uma função de ativação, geralmente a função Softmax [Rumelhart et al. 1986]. A função Softmax é aplicada para converter os valores de similaridade em pesos de atenção, que variam de 0 a 1, refletindo a importância relativa de cada elemento da entrada para a tarefa em mãos.

Dessa forma, a multiplicação de matrizes, normalização e função Softmax permitem que o modelo atribua pesos de atenção a cada elemento da entrada, concentrando-se nos elementos mais relevantes para a realização da tarefa em questão. Essa abordagem de atenção permite que o modelo Transformer foque em partes relevantes dos dados de entrada, capturando relações complexas e aprendendo representações mais robustas. Além disso, o mecanismo de atenção também permite o processamento paralelo dos dados, contribuindo para a eficiência computacional do modelo.

Ao contrário das abordagens anteriores baseadas em redes neurais recorrentes, essa nova arquitetura não requer recorrência para realizar a captura de dependências temporais. Em vez disso, ela utiliza o mecanismo de atenção para obter informações contextuais de todas as posições sequenciais da entrada, permitindo assim a paralelização eficientemente durante o treinamento. Essa característica dos Transformers possibilita o uso do treinamento em lote, no qual várias sequências de entrada são processadas simultaneamente. Isso acelera o processo de treinamento e, ao mesmo tempo, melhora a generalização do modelo em relação aos dados.

O treinamento em lote é uma técnica importante no contexto dos Transformers, ao permitir aproveitar a capacidade de processamento paralelo dos mecanismos de atenção. Com isso, é possível processar um conjunto de exemplos de treinamento ao mesmo tempo, reduzindo o tempo necessário para treinar o modelo em comparação com abordagens sequenciais.

Essa capacidade de paralelização eficiente durante o treinamento contribui para a escalabilidade dos modelos baseados em Transformers, possibilitando o processamento de grandes volumes de dados de forma mais rápida e eficaz. Além disso, o treinamento em lote também ajuda a evitar o *overfitting*, melhorando a capacidade do modelo de generalizar para novos exemplos.

Essas características dos Transformers e do treinamento em lote têm impulsionado avanços significativos no campo da inteligência artificial, especialmente em tarefas relacionadas ao processamento de linguagem natural, como tradução automática, análise de sentimento e geração de texto. A capacidade de capturar relações de longo alcance e processar eficientemente grandes quantidades de dados possibilita a criação de modelos mais poderosos e precisos.

3. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*

Com o surgimento dos modelos baseados em Transformers, uma série de abordagens foram desenvolvidas com o propósito de abordar questões relacionadas ao *Natural Language Process* (NLP) [Otter et al. 2020]. Entre esses modelos, o *BERT (Bidirectional Encoder Representations from Transformers)* se destacou, conforme apresentado no artigo intitulado *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* [Devlin et al. 2018].

O modelo *BERT* utiliza a arquitetura dos Transformers para aprender representações bidirecionais das palavras, permitindo assim a compreensão do contexto em que uma palavra está inserida, considerando tanto as palavras anteriores quanto as posteriores no texto. Essa abordagem de compreensão contextual é crucial em muitos cenários de *NLP*, uma vez que é fundamental compreender o contexto e o significado das palavras.

O *BERT* foi desenvolvido para realizar tarefas de previsão da próxima palavra em uma sequência, auxiliando no processo de aprendizagem de linguagem relevante para o contexto da sequência. Por meio desse modelo, torna-se possível utilizar conjuntos de dados de menor tamanho e adaptá-los para tarefas específicas, como classificação e tradução de textos. Essa flexibilidade proporciona um modelo altamente eficaz, que pode ser adotado por diversos pesquisadores, demandando menor poder computacional para ser ajustado conforme suas próprias necessidades.

4. *XLNet: Generalized Autoregressive Pretraining for Language Understanding*

O advento desses modelos trouxe à tona o *XLNet* [Yang et al. 2019], que se destaca como um modelo pré-treinado autoregressivo generalizado (ARPT). Similar ao *BERT*, o *XLNet* também é fundamentado na arquitetura *Transformer*, porém emprega uma abordagem

inovadora de codificação denominada *Transformer-XL* [Dai et al. 2019], que permite capturar informações com um alcance mais extenso. Essa capacidade de memorização de longo alcance confere ao *XLNet* uma maior robustez para lidar com tarefas de compreensão de linguagem natural, como tradução automática, sumarização de texto e questionamento e resposta.

Ao adotar uma abordagem autorregressiva, o *XLNet* consegue gerar predições de palavras sequencialmente, considerando o contexto de toda a sentença anterior, representando uma melhoria significativa em relação aos modelos prévios que trabalhavam bidirecionalmente. Essa característica autorregressiva possibilita ao *XLNet* uma melhor compreensão das relações entre as palavras e, consequentemente, uma maior precisão na geração de traduções, resumos e respostas em tarefas de *NLP*.

O *XLNet* foi treinado com um enorme conjunto de dados de texto e código, possibilitando gerar texto de alta qualidade, traduzir idiomas com precisão e responder perguntas de forma informativa.

Um dos fatores mais importantes foi disponibilizar a abordagem de código aberto, que impulsionou a colaboração e o compartilhamento de conhecimento na área de Inteligência Artificial, permitindo que especialistas e entusiastas trabalhassem em conjunto para impulsionar ainda mais o desenvolvimento e avanço da tecnologia. Com o acesso ao código-fonte do *XLNet*, a comunidade pode adaptar o modelo às suas necessidades específicas e também contribuir para aprimorá-lo, promovendo um ciclo de aprendizado contínuo e acelerado na área de *NLP* e IA em geral. Essa colaboração aberta tem sido essencial para o progresso da pesquisa e aplicação de modelos de linguagem e inteligência artificial em diversos campos e setores.

5. *An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale*

A percepção da melhoria dos modelos de *NLP* utilizando a arquitetura *Transformer* motivou pesquisadores da Google AI a desenvolverem um novo modelo de aprendizado de máquina para reconhecimento de imagens. Esse novo modelo apresenta a capacidade de aprender as relações entre diferentes partes de uma imagem, possibilitando a identificação de objetos em cenas com maior precisão.

Essa nova abordagem na área de visão computacional representa um avanço significativo, pois o modelo consegue compreender as interações entre os elementos presentes em uma imagem, considerando a disposição espacial e as conexões entre eles. Essa capacidade de aprender as relações entre as partes da imagem permite uma análise mais profunda e contextualizada, resultando em uma identificação mais precisa dos objetos presentes nas cenas.

Nesse novo modelo, cada imagem é dividida em *patches* de 16x16 pixels, e cada *patch* é convertido em uma sequência de *tokens*. Cada *token* é uma pequena unidade que carrega um significado, semelhante a palavras ou números nos modelos anteriores.

Essa abordagem permite representar a imagem como uma sequência de *tokens*, possibilitando utilizar a arquitetura *Transformer* para aprender as dependências espaciais entre os patches da imagem. Ao passar todos os tokens para o decodificador do modelo *Transformers*, o sistema consegue capturar informações mais complexas e abrangentes

sobre a imagem.

A utilização do decodificador do modelo *Transformers* para processar a sequência de tokens possibilita ao modelo aprender as relações e interações entre os patches da imagem. Isso implica em uma melhor compreensão das características visuais da imagem, o que, por sua vez, resulta em uma classificação mais precisa.

Ao classificar as imagens, o modelo utiliza os tokens passados pelo decodificador. Esses tokens representam as informações extraídas dos patches da imagem, permitindo ao modelo realizar uma análise mais profunda e abrangente.

Em resumo, esse novo modelo combina o poder da arquitetura Transformer com a representação em forma de sequência de tokens para aprender as relações espaciais entre os patches de uma imagem. Essa abordagem proporciona uma melhor compreensão das características visuais e uma classificação mais precisa das imagens, tornando-se uma técnica promissora para tarefas de reconhecimento de imagens e visão computacional, em geral.

6. Conclusão

O desenvolvimento dos Transformers tem sido uma verdadeira revolução na área de Inteligência Artificial. Inicialmente, esses avanços foram mais notáveis no campo do Processamento de Linguagem Natural (NLP), onde modelos como o BERT e o XLNet trouxeram melhorias significativas em tarefas de compreensão de linguagem e processamento de texto.

Os Transformers, por sua capacidade de compreender as relações contextuais entre as palavras e sequências de texto, permitiram um salto qualitativo no desempenho de modelos de linguagem, permitindo a resolução de problemas mais complexos de NLP, tais como a tradução automática, a geração de texto e o desenvolvimento de chatbots mais avançados e interativos.

Esses avanços pioneiros no NLP inspiraram pesquisadores a explorar o potencial dos Transformers em outras áreas da inteligência artificial, incluindo a Visão Computacional. Como mencionado anteriormente, o uso de Transformers em modelos de aprendizado de máquina para reconhecimento de imagens trouxe resultados promissores, permitindo a compreensão das relações espaciais entre partes das imagens e melhorando a precisão na classificação e identificação de objetos.

O desenvolvimento e a evolução dos Transformers representam um marco significativo na história da Inteligência Artificial, e os progressos alcançados até o presente momento são apenas o início do potencial revolucionário dessa abordagem em diferentes campos e aplicações. À medida que novas pesquisas são realizadas e novos conhecimentos são compartilhados, é possível esperar que a Inteligência Artificial continue a se desenvolver, trazendo benefícios cada vez mais amplos para a sociedade em geral.

References

Abdi, H., Valentin, D., and Edelman, B. (1999). *Neural networks*. Number 124. Sage.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., and Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Di Gangi, M. A., Negri, M., and Turchi, M. (2019). Adapting transformer to end-to-end spoken language translation. In *Proceedings of INTERSPEECH 2019*, pages 1133–1137. International Speech Communication Association (ISCA).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97.

Jing, Y., Yang, Y., Feng, Z., Ye, J., Yu, Y., and Song, M. (2019). Neural style transfer: A review. *IEEE transactions on visualization and computer graphics*, 26(11):3365–3385.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Liu, P.-r., Lu, L., Zhang, J.-y., Huo, T.-t., Liu, S.-x., and Ye, Z.-w. (2021). Application of artificial intelligence in medicine: an overview. *Current Medical Science*, 41(6):1105–1115.

Otter, D. W., Medina, J. R., and Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, 32(2):604–624.

Ouyang, F. and Jiao, P. (2021). Artificial intelligence in education: The three paradigms. *Computers and Education: Artificial Intelligence*, 2:100020.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

Russell, S. J. (2010). *Artificial intelligence a modern approach*. Pearson Education, Inc.

Valenzuela, O., Catala, A., Anguita, D., and Rojas, I. (2023). New advances in artificial neural networks and machine learning techniques. *Neural Processing Letters*, pages 1–4.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Zhang, J. and Man, K.-F. (1998). Time series prediction using rnn in multi-dimension embedding phase space. In *SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 98CH36218)*, volume 2, pages 1868–1873. IEEE.