Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 12, n. 1, 2026.

Trabalho apresentado no XLIV CNMAC, Fundação Getulio Vargas - Rio de Janeiro - RJ, 2025.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Combinando Autoencoder com Mascaramento e Transformadas Digitais para Classificação de Imagens

Amanda P. de O. Ornelas[1], João B. Florindo[2]
IMECC-UNICAMP, Campinas, SP

O avanço do aprendizado profundo permitiu a diversas áreas a aplicação desta tecnologia. Na área da medicina, por exemplo, surgiu a possibilidade de utilizar essas técnicas para segmentação e classificação de imagens. O grande problema dessa aplicação na medicina consiste no fato de que existe muitas vezes uma limitação no conjunto de imagens rotuladas (necessário para o treinamento de um algoritmo supervisionado). Mas, apesar da escassez de imagens anotadas, há um grande número de imagens não rotuladas. Dessa maneira, surgiu a necessidade de aplicar métodos para o treinamento de redes neurais que fizessem algum uso também dessas imagens não rotuladas. Assim, o aprendizado autossupervisionado tem se apresentado como uma solução promissora para este desafio, uma vez que ele é capaz de produzir representações úteis utilizando um conjunto de dados não rotulados [4]. Um modelo recentemente desenvolvido de aprendizado autossupervisionado é o Masked Autoencoder (MAE).

O MAE, proposto por [3], divide as imagens em *patches* (partes) e realiza o mascaramento de uma porcentagem desses *pacthes*. O objetivo é que o modelo aprenda características relevantes do dado ao tentar recuperar a imagem original. Para isso, uma estretégia seria mascarar uma grande quantidade de *patches* aleatoriamente para, assim, eliminar a redundância e criar uma tarefa desafiadora. É importante que os *patches* mascarados sigam uma distribuição uniforme, impedindo que tenham mais *patches* mascarados próximos ao centro da imagem (ou seja, previnindo um viés central). Assim, o codificador irá operar apenas nos *patches* mascarados e retornar uma representação latente desses *patches*. Em seguida, o decofidicador tem como entrada o conjunto das representações latentes obtidas pelo codificador e os *patches* mascarados com suas respectivas posições para finalmente obter a imagem original. O decodificador e o codificador são construídos de maneira independente, pois o decodificador é utilizado apenas durante a etapa de pré-treinamento, sendo retirado do modelo após a conclusão dessa etapa, já que o mesmo não será utilizado nas tarefas de classificação ou segmentação de imagem. O codificador e o decodificador do MAE consistem em um *Vision Transformer* (ViT).

Os Transformers foram propostos originalmente para tradução automática de textos [5]. Recentemente, [1] aplicou uma ideia semelhante para processamento de imagens, construindo o *Vision Transformer*. A ideia consiste em particionar as imagens e transformá-las em uma sequência de dados para aplicar o modelo Transformer.

Este projeto visa utilizar o MAE em conjunto com uma operação de processamento de imagem: a Transformada de Fourier.

Conforme [2], a Transformada de Fourier, proposta originalmente pelo matemático francês Jean Baptiste Joseph Fourier, é amplamente usada em processamento de imagens. Ela consiste em uma outra maneira de representar a informação visual da imagem, ou seja, fora do "domínio espacial". Dizemos que a imagem pela Transformada de Fourier está representada no "domínio das frequências". Este domínio contém informações sobre as frequências da imagem. Além disso, é possível retornar a imagem ao "domínio espacial" sem perda de informação.

[1] a224208@dac.unicamp.br
[2] florindo@unicamp.br

Este projeto propõe a construção de um novo modelo para processamento de imagens médicas utilizando imagens não rotuladas:

- O pré-treinamento terá como entrada a imagem mascarada da Transformada de Fourier. E como saída a reconstrução da Transformada de Fourier da imagem. Ou seja, a imagem que deve ser reconstruída é a Transformada de Fourier da imagem original.

A ideia consiste em utilizar o codificador resultante das operações de pré-processamento, descartar o decodificador, e acrescentar um novo decodificador responsável pela tarefa de classsificação das imagens médicas.

O modelo encontra-se atualmente em fase de testes. Espera-se que a aplicação da Transformada de Fourier contribua para a melhoria dos resultados do MAE, uma vez que representará um desafio adicional para o codificador.

Referências

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit e N. Houlsby. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". Em: **Proceedings of the International Conference on Learning Representations (ICLR)**. Google Research, Brain Team. 2021.

[2] R. C. Gonzalez e R. E. Woods. **Processamento Digital de Imagens**. Trad. por Cristina Yamagami e Leonardo Piamonte. 3ª ed. Tradução da obra original publicada em 2008 pela Pearson Education, Inc. São Paulo: Pearson Education do Brasil, 2010.

[3] K. He, X. Chen, S. Xie, Y. Li, P. Dollár e R. Girshick. "Masked Autoencoders Are Scalable Vision Learners". Em: **IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. Facebook AI Research (FAIR). 2022, pp. 15979–15988. DOI: 10.1109/CVPR52688.2022.01553.

[4] S. C. Huang, A. Pareek, M. Jensen, M. P. Lungren, S. Yeung e A. S. Chaudhari. "Self-supervised learning for medical image classification: a systematic review and implementation guidelines". Em: **NPJ Digital Medicine** 6.1 (2023), p. 74. DOI: 10.1038/s41746-023-00811-0.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser e I. Polosukhin. "Attention Is All You Need". Em: **31st Conference on Neural Information Processing Systems (NIPS 2017)**. 2017.