Processamento de Linguagem Natural com Transformers

Arquitetura Vision Transformer

O Vision Transformer (ViT) é um modelo de aprendizado profundo desenvolvido pelos pesquisadores do Google Brain. Foi introduzido em um paper intitulado "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" e representa uma abordagem radicalmente diferente para a visão computacional.

https://arxiv.org/abs/2010.11929

Tradicionalmente, as Redes Neurais Convolucionais (CNNs) têm sido o padrão para tarefas de visão computacional, como reconhecimento de imagem. CNNs se destacam ao processar dados no formato de grade localmente e hierarquicamente, aproveitando a estrutura 2D das imagens.

O Vision Transformer, no entanto, aplica a arquitetura Transformer diretamente às imagens. Para fazer isso, a imagem é dividida em um conjunto de patches fixos, que são então “achatados” e processados sequencialmente pelo Transformer.

Essencialmente, cada patch de uma imagem é tratado como um "token" de uma sequência, similar a como as palavras em uma frase são tratadas em tarefas de processamento de linguagem natural. Este é o significado do termo "An Image is Worth 16x16 Words" no título do paper - cada patch de 16x16 pixels é tratado como uma "palavra".

Um token de classificação é adicionado à sequência de entrada, e este token é usado para a predição de classe na saída do modelo. Além disso, como nos Transformers de linguagem, são adicionadas posições relativas aos patches para manter um senso de localização espacial na imagem.

O ViT mostrou resultados competitivos quando treinado em grandes quantidades de dados e seu desempenho aumenta à medida que a quantidade de dados de treinamento aumenta, ao contrário das CNNs que tendem a saturar.

O modelo ViT é estudado na prática no curso de Análise de Imagens com IA, aqui na DSA.

Não demorou muito para que os pesquisadores tentassem unir o modelo BERT (um dos modelos Transformer mais poderosos em PLN) com o modelo ViT. Por que não criar um modelo capaz de capturar a relação entre texto e imagem? E assim nasceu a arquitetura VilBERT, tema da aula a seguir.