Elton Cardoso do Nascimento 233840

IA024 - Redes Neurais Profundas para Processamento de Linguagem Natural - 1s2024

Leitura do Artigo "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"

O artigo apresenta o desenvolvimento de um novo modelo de NLP, o BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers), um modelo encoder-only Transformer. A principal inovação neste modelo é o uso de informações tanto de tokens tanto lendo da direita para a esquerda quanto da esquerda para a direita em todas as camadas de atenção, ao contrário dos modelos anteriores causais, que realizam a operação apenas da esquerda para a direita. Isso permite ao modelo obter performance melhor que o estado da arte em diversas tarefas.

Para obter essa nova característica, o treino do modelo é feito mascarando aleatoriamente tokens da entrada para serem posteriormente preditos, processo chamado de "masked language model" (MLM), que se contrapõe a tarefa de predizer apenas o token no final da sentença. O modelo também é capaz de obter informações sobre relações entre duas sentenças, realizando para isso um segundo treino em que deve predizer de uma segunda sentença é continuação da primeira. É apresentado um teste de ablação mostrando que ambos os treinamentos são importantes para a performance do modelo . Uma diferente forma de realizar o encoding da entrada é apresentada onde, além de somar o enconding posicional ao embedding do token, é adicionado o embedding relacionado a primeira ou a segunda sentença. Além disso, tokens especiais são utilizandos no tokenizador, como início de sentença (CLS) e separador de sentenças (SEP).

A avaliação do modelo se foca na realização de fine-tuning para tarefas específicas com modificações mínimas adicionando camadas finais do modelo, obtendo performance melhor que estado da arte nas tarefas com pouco custo. Porém também são apresentadas avaliações sobre o uso de transferência de aprendizado feature-based, obtendo performance similar a primeira técnica.

Por fim, o trabalho consegue demonstrar que aumentar o tamanho do modelo ao extremo pode permitir melhorias em tarefas com datasets pequenos.