



Processamento de Linguagem Natural com Transformers

Processamento de Linguagem Natural com Transformers

Componentes da Arquitetura BERT

A arquitetura do BERT é baseada na arquitetura dos Transformers, que foram introduzidos no artigo "Attention is All You Need". O Transformer usa uma arquitetura de encoder-decoder, mas BERT usa apenas a parte do encoder. Aqui estão os componentes chave da arquitetura BERT:

Input Embeddings: A entrada para o BERT é uma sequência de tokens, que são primeiros convertidos em vetores, ou "embeddings". Estes são construídos a partir de três tipos diferentes de embeddings: Token embeddings (para converter palavras em vetores), Segment embeddings (para distinguir diferentes sentenças) e Position embeddings (para incorporar a posição das palavras na sequência). Estes três tipos de embeddings são somados para produzir a entrada final.

Transformers Encoders: A parte central da arquitetura do BERT é uma série de camadas idênticas, cada uma das quais é um bloco de Transformer Encoder. Cada bloco possui duas subcamadas: a primeira é uma atenção multi-cabeça que permite que o modelo ponde diferentes palavras em uma sentença ao gerar a representação de uma palavra específica e a segunda é uma rede neural feed-forward simples.

Self-Attention Mechanism: Dentro da subcamada de atenção multi-cabeça nos blocos de Transformer, o BERT utiliza um mecanismo chamado "self-attention". Esse mecanismo permite que o modelo considere o contexto das palavras, olhando para outras palavras na mesma sentença.

Fully Connected Layer: Na extremidade da rede, BERT tem uma única camada totalmente conectada que é usada para tarefas de classificação. Para tarefas de previsão de token (por exemplo, preencher um espaço em branco), a saída desta camada é alimentada através de uma camada softmax para produzir probabilidades para cada token no vocabulário.

Tokens Especiais: A arquitetura do BERT também faz uso de tokens especiais, como [CLS] (usado como a primeira entrada para tarefas de classificação) e [SEP] (usado para separar duas sentenças).

Esses são os componentes básicos da arquitetura do BERT. As versões específicas do BERT, como o BERT-base e o BERT-large, diferem no número de camadas de Transformers, no tamanho dos vetores de embedding e no número de cabeças de atenção.

Criaremos o modelo BERT a partir do zero no Estudo de Caso a seguir.