Prática: Introdução a Modelos Transformers (I)

Eduardo Prasniewski

1 Video

Explica o que são Transformers. Criados em 2017 por pesquisadores da Google, pelo famoso artigo "Attention is all you need". Positional encoding é a ideia de ao invés de olhar para as palavras sequencialmente (como nas RNNs) é atribuir um número a palavra antes de ser enviada para o modelo. Attention é a estratégia usada para que o modelo utilize informações ja processadas para a atual decisão (olhando palavras anteriores). Já self-attention, o grande triunfo do artigo, diz respeito ao modelo entender o contexto em que a palavra está inserida.

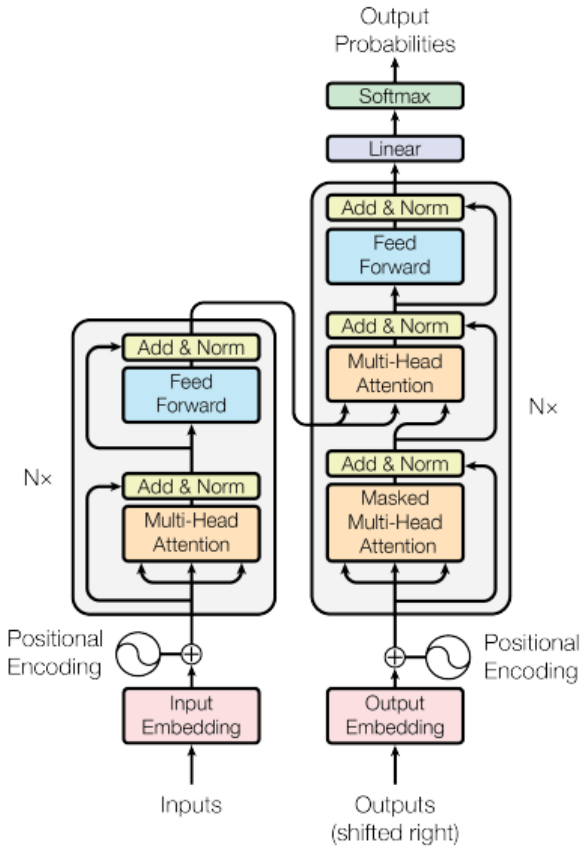

Figura 1: Arquitetura Transformer

2 Minicurso Udemy

2.1 Seção 8

Explica superficialmente com uma abordagem matemática como funciona o mecanismo da arquietetura transformer, primeiramente explica o positional enconding e depois aprofunda

no self-attention, que é utilizada no decoder dos transformers, por exemplo, msotra como é calculada as matrizes query (Q), key (K) e value (V). Logo em seguida aborda os conceitos de decoder, os quais são muito semelhantes as camadas de encoder, sendo a principal diferença a mask-multihead-attention.

Posteriormente introduz a arquitetura BERT (Bidirectional Encoder Representations from Transformers), o qual utiliza apenas a perate de encoder dos Trasnformers. É um modelo pré-treinado open-source, o qual utilizou dados de Toront BookCorpus e a Wikipedia. Porém também é disponível multilingual e possui um modelo BERTimbau, um fine-tuning para lingua portuguesa realizado em cima do BERT. Ainda possui outras variantes, como alBERT, RoBERTa, XLNet e DistilBERT.

O professor apresenta duas empresas principais de modelos de LLM, a Hugging Face e a OpenAI, resaltando as vantagens e desvantagens principalmente relacionadas a open-source e praticidade de uso. Dentro da plataforma do Hugging Face há variados modelos, principalmente provindos de fine-tuning, um desses foi utilizado para demonstrar como é feito o uso do *pipeline* com a biblioteca *transformers*, para a aplicação de Perguntas e Respostas, Preenchimento de Lacunas, Resumo e Geração de Texto.

Por fim, faz uma demonstração do uso do ChatGPT usando a API da OpenAI, porém, como é necessário uma conta na OpenAI com cartão para o uso do trial, acabei não seguindo com a prática, porém acompanhei a aula. ChatGPT é um gerador de text com viés genérico, se adaptando a conversa para melhorar suas respostas. Para auxiliar a especialização, é possível determinar agentes, como *system* e o *assistant* que determinam o papel e contexto da conversa.