Nome: Fernando Buligon Antunes **Data:**24/03/2025

O que são Transformers

Se tratam de um tipo de rede neural, modelos capazes de entender, traduzir ou escrever textos. Alguns dos modelos famosos que são baseados em transformers são BERT, GPT-3 e T5. Até um tempo atrás o modelo usado para processamento de linguagem natural era o RNN (Recurrent Neural Network), que havia sérios problemas, principalmente pois caso estivesse analisando um texto minimamente grande quando ele chegava no final ele já se “esquecia” do início, também não era capaz de compreender que a ordem das palavras importa, então por exemplo, se você tentasse traduzir um texto a saída provavelmente seria algo totalmente sem sentido e fora de contexto. Os transformers entraram para resolver esse exato problema, permitindo que os modelos fossem treinados com base de dados muito maiores do que o RNN, por exemplo, o GPT-3 foi treinado com praticamente 45 terabytes de dados textuais, então a apresentadora define transformers como uma combinação entre rápido desenvolvimento e uma base de dados realmente grande, gerando ótimos resultados.

Há três inovações que permitem esse modelo funcionar tão bem:

- Positional Encodings: É a ideia de tokenizar cada palavra presente em uma sentença antes de enviar ela para a rede neural;
- Attention: Permite que o modelo analise todas as palavras individualmente em uma sentença antes de fazer uma decisão sobre como traduzir na saída;
- Self-Attention: Permite a rede neural analisar o contexto da palavra, fazendo com que não aconteçam erros na presença de duas palavras iguais com sentidos diferentes.

Ao final do vídeo é feito uma breve apresentação do TensorFlow Hub, um repositório repleto de modelos pré treinados.

Seção 8 - Transformers, Bert, GPT e mais

Primeiro é feita uma introdução sobre os transformers, definindo como os mais refinados modelos em processamento de linguagem natural atualmente, depois ele faz um resuminho contendo quase tudo que foi falado no vídeo anterior, com o diferencial de que mostra a arquitetura de um transformer, a ENDEC, que seriam o codificador e o decodificador, o codificador basicamente transforma os dados para uma maneira na qual o computador consiga processar, e o decodificador volta esses dados para uma forma na qual nós consigamos visualizar e entender. Também é apresentado o Multi-Head Attention, que ao invés de um único processo de Attention ser processado, são processados múltiplos por vez e o resultado de cada um é somado em um total, isso faz com que a precisão do sistema seja maior.

Depois é falado sobre o Bidirectional Encoder Representation from Transformers (BERT), e como o próprio nome já sugere, ele é baseado em transformers, possui apenas a parte do encoder e é bidirecional, sendo capaz de interpretar texto nas duas direções. No total existem 24 variações do BERT, que se diferenciam pelo número de camadas de

encoders e unidades na camada oculta, todas essas variações são pré treinadas, o que é bem importante considerando que o modelo padrão do BERT em inglês demorou quatro dias para ser treinado usando 16 TPUs. Esse treino é feito em duas etapas, a primeira é a masked language modeling, em que 15% dos tokens recebem a tag MASK e o modelo tenta prever qual seria a tag que melhor se encaixa naquele token através de ajuste de pesos,, se baseando nas palavras não mascaradas, quando ele termina é retornado uma lista com as probabilidades, a outra técnica é a next sentence pedicition, em que as sentença são divididas em pares, com o objetivo de prever se a sentença é a continuação de seu par, é um problema de classificação binária, tendo como únicas saídas 0 ou 1. O BERT funciona da seguinte maneira: primeiro o texto é convertido em três camadas de embedding, passando pela de token, em que faz a tokenização padrão, colocando a tag CLS no começo da primeira sentença e a SEP no final de todas as sentenças, depois vem a camada de segment, que separa todas as sentenças em segmentos diferentes, por exemplo, a primeira sentença vai ser a A e a segunda a B, e por último, a camada de position, que adiciona um número a cada token, representando sua posição. Caso alguma palavra não esteja presente no vocabulário, ela é dividida até que seja encontrada e depois é sinalizada com "##". Caso queira usar o BERT em português existem o BERTimbau que é próprio da língua portuguesa e o Multilingual Bert, que foi treinado em mais de 100 idiomas usando a wikipedia.

Depois é mostrado as cinco variações principais do BERT:

- ALBERT: Versão mais leve, disponível em diferentes versões baseadas pelo tamanho, usa menos parâmetros mas mesmo assim mantém uma performance superior a outros modelos;
- RoBERTa: É implementado com Pytorch e tem o diferencial que não possui a etapa de next sentence prediction, e é treinado em diferentes gêneros textuais;
- ELECTRA: Faz uso de uma técnica de substituição de token, então ao invés de receber máscaras eles são substituídos, e assim como o ALBERT possui variação de acordo com o tamanho desejado;
- XLNET: É um modelo bem potente baseado em "Large Bidirectional transformer", e faz uso da permutação, prevendo os tokens de forma aleatória;
- DistilBert: É uma versão mais compacta e rápida do Bert, então é baseado em knowledge distillation, em que um modelo menor é treinado a partir do maior.

Hugging Face, empresa especializada em NLP, possuindo uma comunidade gigante onde temos modelos, dataset e serviços disponíveis. O Hugging Face apresenta uma lista de tarefas que são muito simples de se fazer usando seu pipeline, por exemplo, se quiser fazer uma análise de sentimento, você só precisa instanciar o objeto, passar o texto e ele te devolve com o resultado.

OpenAI, empresa dona da famosa coleção de modelos GPT, cada um contendo performances diferentes para casos diferentes, o diferencial dela é que existe método e modelo genérico para qualquer tarefa, por exemplo, com o Hugging Face se você quer um método para análise de sentimento, você busca o modelo de análise de sentimento e adapta seu método, com a OpenAI você só passa o prompt e ele já vai entender o que você busca, outra diferença é que os modelos são todos em nuvem, então não há a necessidade de baixar. Apesar de todos os seus benefícios, os modelos GPT não são gratuitos a nível de

produção, é necessário uma autenticação por chave, mas para o curso, foi usado uma versão gratuita com dezoito dólares livres para gastar.

Depois foi feito um tour pelo repositório de modelos da Hugging face, foi nele que encontramos os modelos usados para a parte prática. No total foram feitas quatro funções, question answering, fill mask, summarization e text generation.

Na parte prática da OpenAI, foi preciso gerar uma chave de acesso para conseguir se conectar a API que passa o que foi pedido para o gpt-3.5-turbo, que foi o modelo escolhido. Consegui gerar a chave mas diferentemente do que foi mostrado no curso, não recebi o limite para uso, tentei criar uma conta nova e mesmo assim meu crédito para uso era zero, dei uma pesquisada e outras pessoas com o mesmo objetivo que eu também tiveram problemas sem soluções
https://community.openai.com/t/how-can-i-get-free-trial-credits/26742/11.

Na imagem abaixo é possível ver que não houve gasto e que o limite só vai resetar dia 01/04, excedendo a data limite de entrega do card atual, por isso, segui acompanhando a parte prática sem testar.