Destaques da leitura de **“BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”**

Nome: Caio Petrucci dos Santos Rosa **RA:** 248245

Os destaques da leitura do artigo foram os seguintes:

- A proposição do uso de redes neurais pré-treinadas em grandes quantidades de dados por diversas épocas (128B de *tokens* em aproximadamente 40 épocas, conforme descrito na Apêndice A.2) combinado com um pequeno *fine-tuning* de até 4 épocas, em vez de treinar redes “do zero” como *task-specific*;
- Também achei interessante que, prevendo uma tendência que posteriormente seria seguida pela OpenAI no artigo “Language Models are Unsupervised Multitask Learners” do GPT-2, os autores avaliaram a performance do modelo em diferentes *benchmarks* em função de sua quantidade de parâmetros, considerando modelos de 110M até 340M parâmetros;
- Para o pré-treinamento, foram utilizadas duas tarefas: uma tarefa *sentence-level* - dado duas sentenças A e B, determinar se a sentença B ocorre depois da sentença A - e outra *token-level* - prever um *token* mascarado no meio de uma sentença -.
- O modelo foi treinado (*fine-tuning*) e avaliado em diversos *datasets*, contendo diferentes tarefas, dentre eles foram:
    - GLUE *benchmark*: é uma coleção de diversas tarefas de *natural language understanding* (descritas no Apêndice B.1 do artigo);
    - SQuAD v1.1 e v2.0: é um conjunto de pares de perguntas e documentos contendo respostas, em que o modelo deve predizer qual a resposta contida dentro do documento para a pergunta;
    - SWAG: é uma coleção de grupos de sentenças onde, dado uma sentença inicial, o modelo deve escolher a melhor continuação entre outras 4 sentenças.
- As comparações entre as abordagens *feature-based* e *fine-tuning-based* foram relevantes e mostraram que o ajuste fino dos parâmetros é sempre beneficial pois, de acordo com os autores, pode ajudar o modelo a ajustar representações que não são tão facilmente transferíveis para a nova tarefa;
- As discussões relacionadas às diferentes estratégias de *masking* tanto sobre a avaliação de diferentes formas de mascarar os *inputs* e que, eventualmente substituir os *tokens* a serem mascarados por um *token* aleatório ou ele mesmo em vez do *token* `[MASK]`, ajuda o modelo a lidar com certa incompatibilidade entre as tarefas de pré-treinamento e as tarefas de *fine-tuning* considerando que o *token* `[MASK]` muitas vezes não irá aparecer na *downstream-task*;
- Ainda ficou um pouco confuso para mim parte da representação do *input* e do *output*, principalmente como os *tokens* `[CLS]` e `[SEP]` funcionam mais profundamente;
- Por fim, achei interessante que os autores decidiram realizar uma comparação mais minuciosa, principalmente no Apêndice, entre o trabalhos do GPT da OpenAI e do próprio BERT, o que trouxe alguns *insights* sobre a diferenças entre as arquiteturas de Transformers Encoder-only e Decoder-only, como a velocidade de convergência durante o treinamento.