Destaques da leitura de **“Language Models are Unsupervised Multitask Learners”**

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

Os destaques da leitura do artigo foram os seguintes:

- Os autores tiveram como principal objetivo treinar um modelo capaz de aprender a executar diversas tarefas e lidar com uma generalização OOD (*out-of-distribution*). Para isso, tentaram uma abordagem similar ao *multi-task training*, realizando um aprendizado não-supervisionado implícito para diversas tarefas através do aprendizado auto-supervisionado para a tarefa de predição do próximo *token*;
- Para o treinamento de um modelo capaz de realizar diversas tarefas, os autores construíram um *dataset* chamado WebText, que contempla diversas formas de “documentos de alta qualidade”. Achei interessante a heurística que utilizaram para encontrar e considerar os documentos de “alta qualidade”, que foi um Karma positivo de 3 na plataforma Reddit;
- Também achei muito interessante a abordagem de representação do *input* para o modelo, que se baseou principalmente no Byte-Pair Encoding (BPE) mas considerando *tokens* a nível de byte. Não ficou nítido para mim, apenas pela leitura do artigo, como a implementação desse tipo de tokenização ocorre, já que a implementação em si não é descrita em detalhes;
- As pequenas alterações no modelo - dado que apenas moveram as camadas de *layer normalization* para a entrada e para a saída de cada bloco *transformer* e adicionaram um *scaling factor* para *residual connections* - mostram a importância que o conjunto de dados e as diferentes avaliações, tanto dos dados como do modelo, são muito importantes para um bom resultado, além dos detalhes minuciosos de implementação de uma arquitetura de rede;
- Achei importante o fato de que os autores avaliaram o modelo em diversos *datasets* e tipos de tarefas - *question-answering*, tradução, sumarização, compreensão de leitura e de dependências de longo alcance, e nomeação de entidades -, o que mostrou com clareza a hipótese de que “*language models are unsupervised multitask learners*”. Além disso, consideraram quais tipos de tarefa que o modelo performou melhor e quais sofreram maior influência a fatores como o tamanho e a capacidade da rede, por exemplo a tarefa de *question-answering*;
- A análise da performance do modelo em função de seu tamanho foi extremamente relevante e trouxe a ideia de que a escala em modelos de linguagem trazem um ganho absurdo de desempenho. Os autores deixam claro que o modelo aparentou estar longe de *overfitting* e que, na verdade, ainda estava em *underfitting*. Porém, apesar dessas observações, tive a impressão de que ainda não tinham dimensão das capacidades emergentes que surgiriam com o aumento da escala;
- Achei interessantíssima a avaliação do quanto o modelo poderia estar memorizando dos dados e o levantamento quantitativo do *overlap* dos dados de treinamento com os dados de teste tanto no WebText quanto nos outros *datasets* utilizados no trabalho.