Destaques da leitura de “**LoRA: Low-Rank Adaptation of Large Language Models**”

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

Os destaques da leitura do artigo foram os seguintes:

- Achei impressionante que os autores propuseram uma técnica de treinamento que é capaz de reduzir, em diversas ordens de magnitude, o número de parâmetros treináveis, com uma performance similar senão melhor ao *fine-tuning* completo, e com a possibilidade de mesclar os parâmetros ΔW ao modelo final, resultando em latência adicional zero para inferência;
- Achei muito relevante a comparação que os autores fazem com técnicas já existentes de adaptação dos modelos, considerando Adapter Layers, Prefix-embedding e Prefix-layer tuning;
- A leitura do artigo me estimulou a estudar mais sobre assuntos de Álgebra Linear que não tinha muita familiaridade com, como:
    - Single Value Decomposition (SVD);
    - Rank (ou posto matricial em português);
    - Rank Decomposition.
- A análise sobre a dimensionalidade de modelos, de Deep Learning, pré-treinados e, similarmente, como os ajustes em um *fine-tuning* ocorrem em um espaço de baixa dimensionalidade (ou rank):
    - Isso me fez questionar por que modelos grandes, como o GPT-3, têm uma performance muito melhor, dado que grande parte dos problemas de Machine Learning possuem uma estrutura intrínseca de “low-rank”;
    - O artigo “Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning” (https://arxiv.org/abs/2012.13255) que os autores do LoRA utilizam como motivação teórica me instigou. Pretendo ler este trabalho para compreender um pouco mais a fim de responder os questionamentos que levantei no item anterior.
- A avaliação extensa do LoRA aplicado a diferentes arquiteturas de redes para NLU (Natural Language Understanding) NLG (Natural Language Generation), considerando as redes RoBERTa, DeBERTa, GPT-2 e GPT-3;
- Achei bacana que os autores listam, quase que detalhadamente, o regime de treinamento e os hiperparâmetros utilizados (otimizador, learning rate, warmup ratio, LR scheduler, batch size, número de épocas, parâmetro alpha e rank do LoRA, …);
- Achei muito interessante as análise feitas sobre os subespaços de similaridade entre diferentes ranks r utilizando o SVD. Essa abordagem me fez refletir sobre novas possibilidades de visualização e de interpretação do que ocorre dentro de modelos de Deep Learning;
- Para compreender mais a intuição por trás do LoRA, assisti um vídeo discutindo o artigo: https://www.youtube.com/watch?v=dA-NhCtrrVE&ab_channel=ChrisAlexiuk.