Nome: Fabio Grassiotto RA: 890441
IA024 - Aula 5_6
Leitura do Artigo “LORA Low-Rank Adaptation of Large Language Models”

Resumo do Artigo:

Este artigo apresenta uma nova técnica, chamada de **LoRA** (de *Low-Rank Adaptation,* ou Adaptação de Matriz de Baixo Ranking), para o *fine-tuning* de modelos de linguagem.
As principais contribuições que posso ressaltar deste artigo são:

- A hipótese formulada pelos autores que a alteração dos pesos que ocorre no processo de adaptação de modelos para novas tarefas *downstream* em um modelo de rede neural tem um rank de matriz (ou seja, em álgebra linear, a quantidade de linhas linearmente independentes da matriz) intrinsicamente baixo, o que favorece a proposta da técnica apresentada no artigo.
- A análise apresentada das técnicas já existentes para adaptação de modelos para tornar o processo mais eficiente. Nessa análise, os autores argumentam que as técnicas apresentadas apresentam limitações que a proposta do LoRA busca resolver.
- Proposta do LoRA como nova abordagem para substituir o processo de *fine-tuning* de modelos de linguagem com a geração de matrizes de menor rank que representam as matrizes de pesos de um modelo base. No LoRA, as matrizes de pesos do modelo base são congeladas e duas matrizes de baixo rank são injetadas no processo de treinamento do modelo.
- A apresentação dos detalhes da aplicação da técnica à arquitetura de Transformers em especial a aplicação da decomposição para as matrizes utilizadas no processo de auto atenção W_q, W_k, W_v e W_o. Os autores apresentam os ganhos relativos à quantidade de memória de vídeo utilizada para o treinamento (de 1.2TB para 350GB) assim como a redução do tamanho dos *checkpoints* do modelo em cerca de 10.000 vezes.
- A apresentação de experimentos e métricas detalhadas de avaliação dos modelos treinados com LoRA em comparação com fine-tuning tradicional e outras técnicas de adaptação de modelos mostrando as vantagens do LoRA.
- No final do artigo os autores identificam novas direções para o trabalho futuro, ressalto duas principais: (1) a combinação do LoRA com outros métodos de adaptação (o que podemos verificar que ocorreu nos dias de hoje com o QLora, por exemplo) e (2) a necessidade de foco na pesquisa para aumentar a compreensão dos mecanismos que permitem o processo de *fine-tuning* e LoRA, dado que o processo não é totalmente claro.

Artigos Relevantes:

A Note on LoRA, por Vlad Fomenko, Han Yu, Jongho Lee, Stanley Hsieh, Weizhu Chen ([2404.05086] A Note on LoRA (arxiv.org))– Uma análise recente da contribuição do artigo original por alguns dos autores, com explicações práticas do uso da técnica.

The Expressive Power of Low-Rank Adaptation, por Yuchen Zeng, Kangwook Lee https://arxiv.org/abs/2310.17513 - Um artigo recente, que busca explicar a base teórica do processo utilizado pelo LoRA.

QLoRA: Efficient Finetuning of Quantized LLMs por Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, Luke Zettlemoyer ([2305.14314] QLoRA: Efficient Finetuning of Quantized LLMs (arxiv.org)) – Proposta do uso de quantização juntamente com a técnica de LoRA.

What is Low-Rank Adaptation (LoRA) | explained by the inventor Na verdade um vídeo no Youtube: https://www.youtube.com/watch?v=DhRoTONcyZE&t=133s o Autor do LoRA, Edward Hu, explica qual o entendimento dele da técnica do LoRA.