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:

* Proposta do **LoRA** como nova abordagem para o *fine-tuning* de modelos de linguagem.
* Detalhes da aplicação da técnica à arquitetura de Transformers.
* Apresentação de métricas detalhadas de avaliação dos modelos treinados com LoRA em comparação com fine-tuning tradicional.

**Artigos Relevantes:**A Note on LoRA ([[2404.05086] A Note on LoRA (arxiv.org)](https://arxiv.org/abs/2404.05086) )– Uma análise recente da contribuição do artigo original dos autores, com explicações práticas.

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

QLoRA: Efficient Finetuning of Quantized LLMs ([[2305.14314] QLoRA: Efficient Finetuning of Quantized LLMs (arxiv.org)](https://arxiv.org/abs/2305.14314)) -