Nome: Fabio Grassiotto RA: 890441
IA024 - Aula 6_7
Leitura do Artigo “QLORA Efficient Finetuning of Quantized LLMs”

Resumo do Artigo:

Este artigo apresenta QLORA, um método de *fine*-tuning eficiente que se propõe a reduzir a memória necessária para o processo de refinamento no treinamento de modelos grandes de linguagem.

As principais contribuições que posso ressaltar deste artigo são:

- A apresentação de inovações relacionadas ao processo de treinamento que definem o método em si:

    - A introdução do tipo de dado **4-bit NormalFloat**, a ser utilizado no processo de *fine-tuning* no lugar de inteiros e números de ponto flutuante de 4 bits.
    - **Quantização Dupla**, um método para quantizar as próprias constantes de quantização, economizando cerca de 3 GB para um modelo de 65B.
    - **Otimizadores Paginados,** uma otimização para utilizar memória unificada de GPUs Nvidia para evitar uso excessivo de memória no processamento de sequências longas em um batch.
    - **Adaptadores Low-Ranking (LoRA),** o método apresentado do QLoRA faz a retropropagação de gradientes através de um modelo de linguagem pré-treinado quantizado e congelado de 4 bits em adaptadores LoRA. Esses adaptadores permitem um ajuste fino eficiente, preservando o desempenho.

- As análises quantitativa e qualitativa do processo de *fine-tuning* de uma quantidade grande de modelos de linguagem disponíveis utilizando a técnica de QLORA e a apresentação dos resultados obtidos incluindo a comparação de métricas.
- A discussão das limitações do método apresentado, incluindo a falta de avaliação de modelos grandes (33B e 65B) devido aos custos envolvidos, assim com a avaliação de métricas para modelos do tipo instrucional.