Elton Cardoso do Nascimento - 233840

IA024 - Redes Neurais Profundas para Processamento de Linguagem Natural - 1s2024

Leitura do Artigo "QLoRA: Efficient Finetuning of Quantized LLMs" (Dettmers et al.)

O trabalha apresenta uma nova técnica para realizar o finetuning de modelos com alto custo de memória. QLoRA permite realizar o finetuning do maior modelo disponível publicamente na época utilizando apenas uma GPU comercial, diminuindo a barreira que existe entre instituições de pesquisa e grandes empresas. Ele realiza isso se baseando no LoRA e adicionando três aspectos: quantização em **4-bit NormalFloat** (NF4), que permite eficientemente representar dados distribuídos normalmente utilizando 4-bits; **dupla quantização**, que quantiza as constantes de quantização; e **otimização de paginação**, utilizando a funcionalidade de memória unificada do CUDA para evitar problemas com picos de uso de memória. O processo de treino utilizando o QLoRA consiste então em desquantizar a constante de quantização, que é utilizada para desquantização os pesos, que são então utilizandos para realizar as operações. Com isso, a técnica utiliza ao mesmo tempo 2 precisões diferentes, uma em 4 bits para armazenamento e outra em 16 bits para realizar as operações.

A técnica apresentada se mostra eficiente no uso de memória e eficaz quantitativamente, se igualando ao finetuning com LoRA sem a técnica. Além das avaliações quantitativas, o trabalho também realiza uma avaliação qualitativa, desenvolvendo protocolos específicos para isso. A avaliação realizada considera tanto avaliações realizadas por humanos quanto utilizando outras LLMs (GPT-4) para realizar a avaliação, gerando rankings e pontuações Elo. É apresentada uma lista com exemplos de problemas qualitativos do modelo treinado, como se recusar a responder ou apresentar informações incorretas.