Destaques da leitura de “**QLoRA: Efficient Finetuning of Quantized LLMs**”

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

Os destaques da leitura do artigo foram os seguintes:

- Os autores apresentaram uma abordagem inovadora para a fine-tuning, baseado em LoRA, em combinação com quantização de LLMs, que foi capaz de manter a eficácia dos modelos quase sempre, se comparado a um fine-tuning completo em um modelo de 16-bit;
- O que mais achei interessante no artigo foi o tipo de parâmetro NF4, que, na minha visão, foi a técnica central no processo de quantização proposto. Porém, ainda fiquei um pouco em dúvida de como de fato é feita a conversão bit a bit de um dado do tipo FP16 para um dado NF4;
- Também achei bastante interessante as técnicas de quantização dupla (*double quantization*) e de otimizadores paginados (*paged optimizers*), porém, principalmente para a quantização dupla, ainda ficou um pouco confuso como seria uma implementação “from scratch” desse tipo de técnica, além do qual foi o real impacto positivo que esse método teve no treinamento do LLM proposto;
- Além disso, no apêndice, os autores tentam demonstrar uma validação empírica de que os parâmetros das redes neurais seguem uma distribuição normal, dado que a aplicação de NF4 tem como pressuposto isso;
- Achei que, no entanto, o artigo tem uma certa confusão estrutural e de objetivo. Por um lado, os autores propõem uma nova técnica de fine-tuning de LLMs utilizando quantização com poucos bits. Por outro lado, os autores também se prolongam em diversos aspectos, não relacionados ao fine-tuning com quantização, do processo de avaliação do modelo Guanaco, baseado no Llama, em comparação a diversos outros modelos da literatura, que não foram treinados com QLoRA. Por conta disso, não ficou tão claro para mim se o principal objetivo é a melhoria da eficiência através da quantização ou a introdução de um novo LLM destinado a aplicações como chatbot;
- Mesmo assim, analisando a conteúdo apresentado em si, achei muito interessante as técnicas de avaliação tanto quantitativas quanto qualitativas dos resultados do modelo Guanaco e a competição estilo campeonato que foi feita com outros LLMs ou sistemas baseados em LLMs, como o ChatGPT, que realizaram para ranquear a qualidade do modelo.