



*Processamento de Linguagem Natural com Transformers*

# Processamento de Linguagem Natural com Transformers

Quantized Low-Ranking Adaptation (QLoRA)

QLoRA é uma extensão LoRA para aumentar a eficiência quantizando valores de peso da rede original, desde tipos de dados de alta resolução, como Float32, até tipos de dados de resolução mais baixa, como int4. Isso leva a demandas reduzidas de memória e cálculos mais rápidos.

Existem três otimizações principais que o QLoRA traz além do LoRA, o que torna o QLoRA um dos melhores métodos PEFT: 4-bit NF4 Quantization, Double Quantization e Unified Memory Paging. Vejamos uma descrição de cada um dos métodos.

#### 4-bit NF4 Quantization

A quantização NormalFloat4 de 4 bits é um processo de 3 etapas: Normalização, Quantização e Desquantização.

Como parte das etapas de normalização e quantização, os pesos são ajustados para uma média zero e uma variância unitária constante. Um tipo de dados de 4 bits pode armazenar apenas 16 números. Como parte da normalização, os pesos são mapeados para esses 16 números, distribuídos com centro em zero e, em vez de armazenar os pesos, a posição mais próxima é armazenada.

Para Desquantizar os valores, fazemos exatamente o inverso.

Obviamente, há uma perda de dados quando normalizamos e quantizamos, à medida que passamos do FP32, que é um tipo de dados de alta resolução, para um tipo de dados de baixa resolução. A perda não é enorme, desde que não haja valores discrepantes no tensor de entrada, o que pode afetar o `absmax()` e eventualmente perturbar a distribuição. Para evitar esse problema, geralmente quantizamos os pesos de forma independente por blocos menores, o que normalizará os valores discrepantes.

A quantização NormalFloat de 4 bits é aplicada aos pesos do modelo original, os pesos do adaptador LoRA serão FP32, pois todo o treinamento acontecerá nesses pesos. Feito todo o treinamento, os pesos originais serão desquantizados.

#### Double Quantization

A quantização dupla reduz ainda mais o consumo de memória, quantizando constantes de quantização. Na etapa anterior de quantização FP4 de 4 bits, calculamos a constante de quantização. Mesmo isso pode ser quantizado para melhor eficiência e é isso que fazemos na Dupla Quantização.

Como a quantização é feita em blocos, para evitar outliers, normalmente 64 pesos em 1 bloco, teremos 1 constante de quantização. Essas constantes de quantização podem ser quantizadas ainda mais, para reduzir o consumo de memória.

## Unified Memory Paging

Juntamente com as técnicas acima, o QLoRA também utiliza o recurso de memória unificada da nVidia, que permite transferências contínuas de páginas GPU-> CPU, quando a GPU fica sem memória, gerenciando assim os picos repentinos de memória na GPU e ajudando em problemas de estouro/sobrecarga de memória.

LoRA e QLoRA são duas das técnicas mais amplamente utilizadas para ajuste fino com eficiência de parâmetros.