Analytics Vidhya

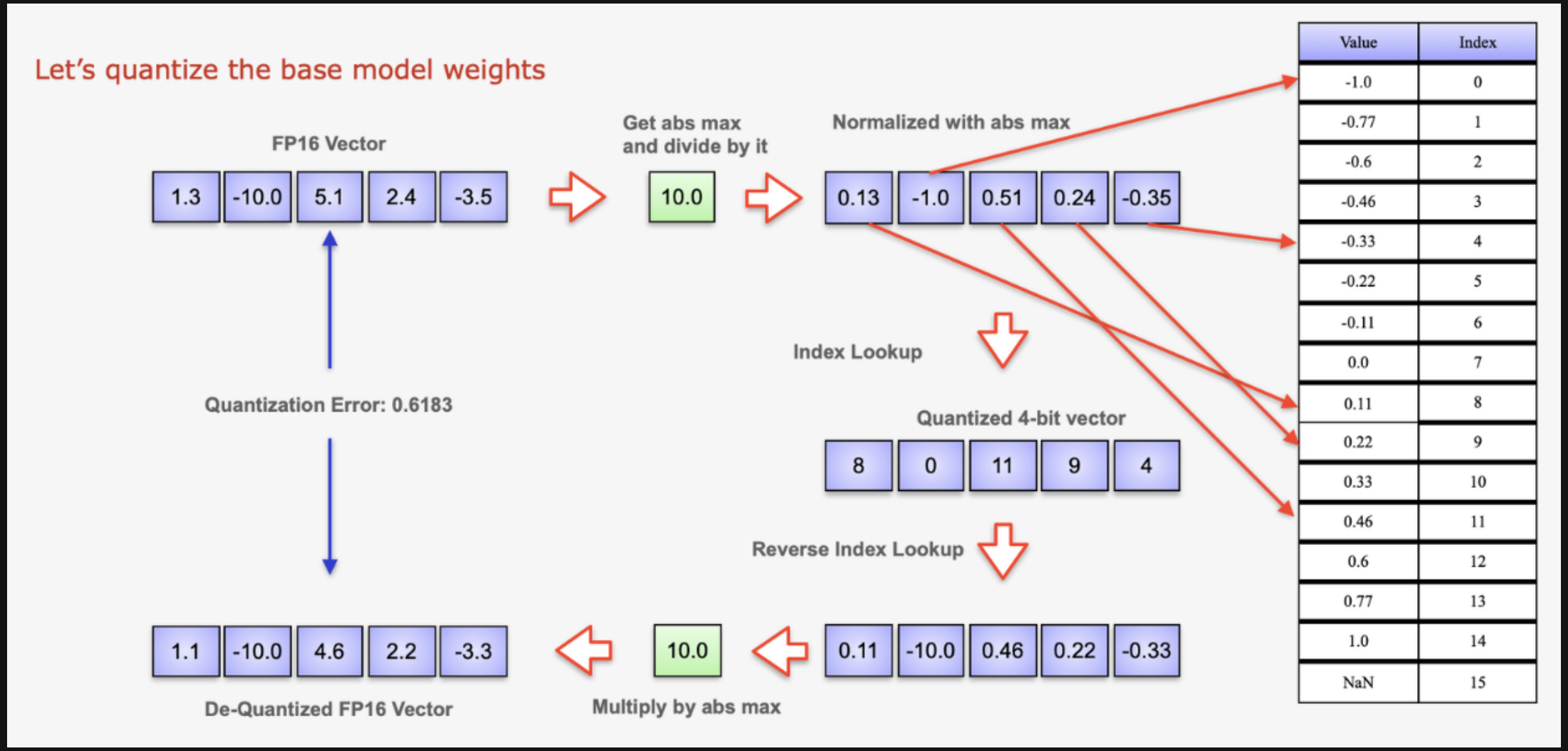
Quantization and QLoRA

Instructor

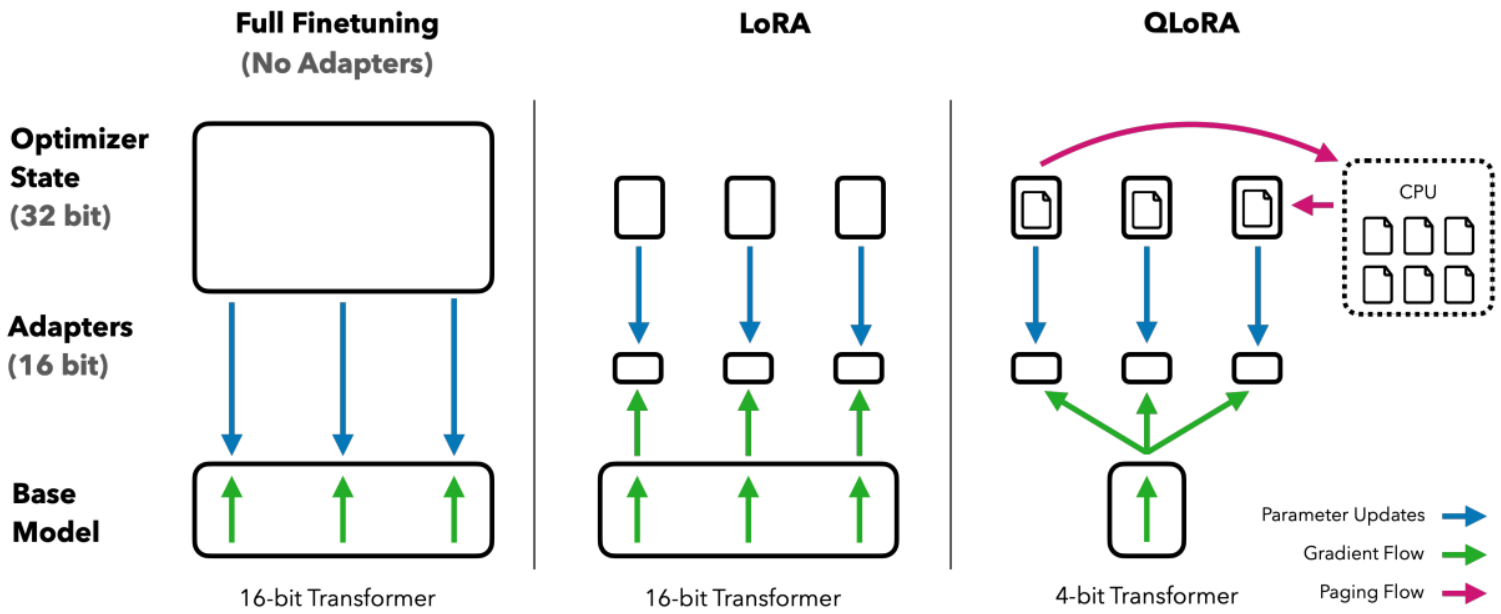
Sourab Mangrulkar

Machine Learning Engineer at 🤗,
Creator of 🤗 PEFT

Compressing base model using Quantization



QLoRA



Overview

- Asymmetric NF4 DType: [-1.0, -0.7, -0.53, -0.39, -0.28, -0.18, -0.09, 0.0, 0.08, 0.16, 0.25, 0.34, 0.44, 0.56, 0.72, 1.0]
- Double QuantizationPaged Optimizers
- QLoRA has one storage data type (NF4) and a computation data type (16-bit Bfloat). Dequantize the storage data type to the computation data type to perform the forward and backward pass, but only compute weight gradients for the LoRA parameters which use 16-bit Bfloat.
- LoftQ is a method to initialize LoRA weights such that the quantization error is minimized. Improves performance of QLoRA

Analytics Vidhya

QLoRA Finetuning cost

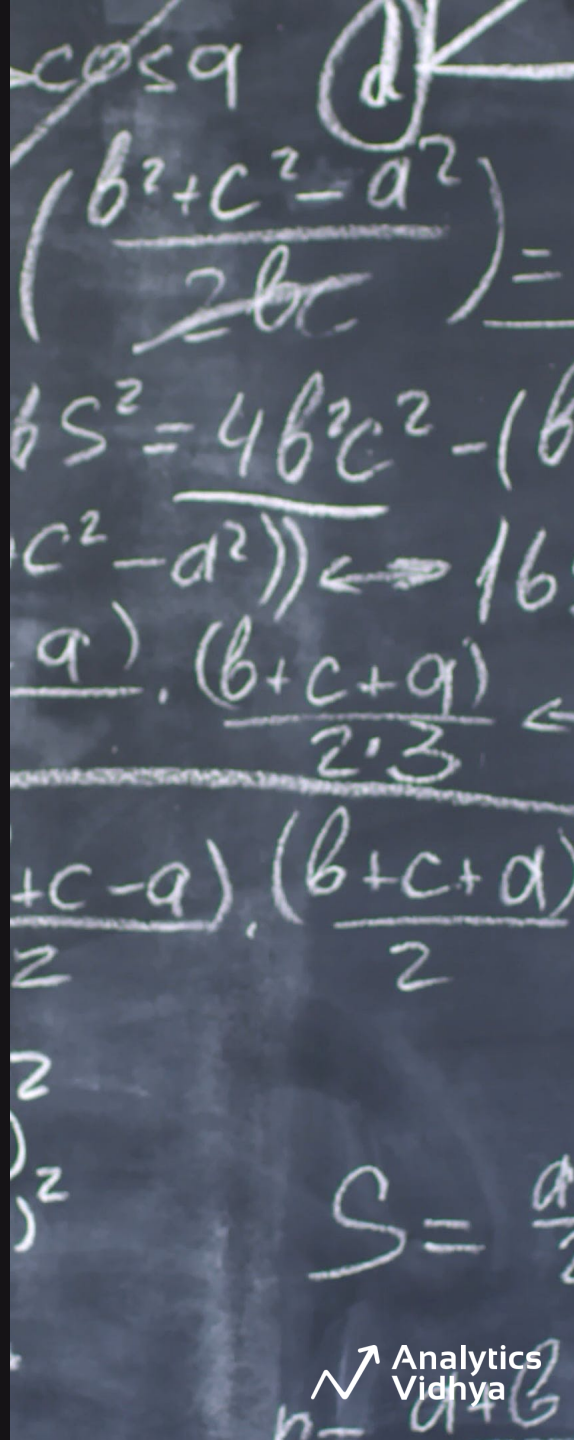
Finetuning Mistral-7B in mixed-precision using Adam Optimizer.
trainable: 21,549,136 || all params: 7,263,322,192 || trainable%: 0.296

Weights - 0.5 bytes / parameter

Gradients - 2 bytes / parameter

Optimizer state - 4 bytes / parameter (FP32 copy) + 8 bytes / parameter (momentum & variance estimates)

Total training cost: 16 bytes/parameter * 7 billion parameters * 0.0029 + 14 = 112 * 0.00296 + 4 GB ~ 4.5 GB



Analytics Vidhya