Parameter Efficient Fine Tuning (PEFT)

A screenshot of a computer

Description automatically generated

* In contrast to full fine-tuning where every model weight is updated during supervised learning, parameter efficient fine tuning methods only update a small subset of parameters.
* Some path techniques freeze most of the model weights and focus on fine tuning a subset of existing model parameters, for example, particular layers or components.
* Other techniques don't touch the original model weights at all, and instead add a small number of new parameters or layers and fine-tune only the new components.
* PEFT Fine tuning saves space and is flexible.

PEFT Methods

A screenshot of a computer

Description automatically generated

PEFT Techniques 1 : LoRA

A diagram of a algorithm

Description automatically generated

A screenshot of a math problem

Description automatically generated

If instead, you want to carry out a different task, say Task B, you simply take the LoRA matrices you trained for this task, calculate their product, and then add this matrix to the original weights and update the model again. The memory required to store these LoRA matrices is very small.

A diagram of a task

Description automatically generated

Switch out the weights when you need to use them, and avoid having to store multiple full-size versions of the LLM.