Lab: QLoRA with Hugging Face

In this lab, you will learn to perform parameter efficient fine-tuning (PEFT) using Quantized Low-Rank Adaptation (QLoRA) with Hugging Face.

If you have access to a CUDA (Compute Unified Device Architecture)-enabled GPU, this optional lab will walk you through the process of loading a model, enabling it for quantized training, and adapting its layers using QLoRA.

This lab teaches you how to fine-tune using Hugging Face and covers the pros and cons of QLoRA fine-tuning and model quantization in general.

NOTE:To successfully complete this optional lab, you'll need to run it on a computer equipped with a CUDA-enabled GPU. You can either download the notebook and run it locally or use a cloud account with GPU access.

To download the notebook, Right-click and open the given link in a new tab:
Lab: QLoRA with Hugging Face

Skills Network