Summary and Highlights

Congratulations! You have completed this lesson. At this point in the course, you know that:

- Parameter-efficient fine-tuning (PEFT) methods reduce the number of trainable parameters.
- PEFT includes various fine-tuning methods, such as selective fine-tuning, additive fine-tuning, and reparameterization fine-tuning.
- Low-rank adaptation (LoRA) reduces the trainable parameter by leveraging pre-trained models with high-dimensional matrices. The simplest way to explain LoRA is by using an algebra matrix.
- The original weight matrix remains frozen while fine-tuning LoRA; however, the LinearWithLoRA class copies the original linear model and creates a LoRALayer object.
- LoRA with PyTorch uses an Internet Movie Database (IMDB) dataset to review movies. LoRALayer class implements the LoRA module with two low-rank matrices.
- LoRA with HuggingFace helps to train models easily using a tokenizer to create input IDs and loads bidirectional representation for transformers (BERT) like models from the HuggingFace transformer library.
- Quantized low-rank adaptation (QLoRA) is a fine-tuning technique in machine learning designed to optimize the performance and efficiency of large language models (LLMs).
- Quantization reduces the precision of the numerical values to a finite set of discrete levels, decreasing memory usage and enabling efficient computation on hardware with limited precision.
- The quantization range lies between -1 to 1 and uses 4-bit NormaFloat (NF4) and double quantization methods.
- QLoRA reduces the memory footprints using model parameters, gradients, two-state optimizers, and activations.
- Model quantization reduces the precision of model parameters by reducing the model size and improving inference speed by maintaining the model's accuracy. The useful tools and libraries for model quantization are TensorFlow Lite and PyTorch.
- Some of the model quantization techniques are:
    - Uniform quantization
    - Non-uniform quantization
    - Weight clustering
    - Pruning and quantization