Course Conclusion

Congratulations! You've successfully completed the course and are now equipped with pretrained transformers, parameter-efficient fine-tuning (PEFT) using low-rank adaptation (LoRA) and quantized low-rank adaptation (QLoRA), and fine-tuning using Hugging Face and PyTorch.

At this point, you know that:

- Hugging Face is a platform and community dedicated to machine learning (ML) and data science, which assists users in developing, deploying, and training ML models.
- PyTorch is a software-based open-source deep learning framework for building neural networks and supporting their architectures.
- Fine-tuning LLMs adapt pretrained models for specific tasks or domains using domain-specific data.
- Upon fine-tuning a model completely, it provides 90% accuracy to the model. However, by fine-tuning only the final layer of the model, training becomes much faster but decreases the model's performance.
- Benefits of fine-tuning:
    - Enhances efficiency and saves time
    - Transfers learning, time, and resource efficiency
    - Tailors responses and task-specific adaptation
    - Addresses issues like overfitting, underfitting, catastrophic forgetting, and data leakage
- Approaches of fine-tuning language models:
    - Self-supervised fine-tuning
    - Supervised fine-tuning
    - Reinforcement learning from human feedback
    - Direct preference optimization
- Hugging Face's built-in data sets can be loaded using the load_dataset function. The tokenizer function extracts the text from the data set example and applies the tokenizer. The evaluation function evaluates the model's performance after fine-tuning it.

- SFT Trainer (or supervised fine-tuning trainer) simplifies and automates many training tasks, making the process more efficient and less error-prone compared to training with PyTorch directly.
- Parameter-efficient fine-tuning (PEFT) methods reduce the number of trainable parameters that should be updated to adapt a large pretrained model to specific downstream applications effectively.
- Methods of PEFT are selective, additive, and reparameterization fine-tuning.
- Soft prompts are learnable tensors concatenated with the input embedding that can be optimized to a data set; however, ranks minimize the number of vectors for space spanning.
- LoRA helps complex ML for specific uses by adding lightweight plug-in components to the original model. It reduces the number of trainable parameters using pretrained models and matrix algebra to decompose weight updates into low-rank matrices.
- In LoRA with PyTorch, the model uses the internet movie database (IMDB) data set and the class to create iterators for training and testing data sets; however, using the IMDB dataset and LoRA with Hugging Face simplifies the model training process.
- QLoRA is a fine-tuning technique in ML for optimizing performance; however, quantization reduces the precision of numerical values to a finite set of discrete levels by defining the quantization range and levels.
- Model quantization reduces the precision of model parameters by reducing the model size and improving inference speed by maintaining the model's accuracy.
- Some of the model quantization techniques are:
    - Uniform quantization
    - Non-uniform quantization
    - Weight clustering
    - Pruning