Artificial Intelligence (Machine Learning & Deep Learning) [Course]

Week 10 – Hugging Face -Fine-Tuning LLMs (PEFT, QLoRA)

[See examples / code in GitHub code repository]

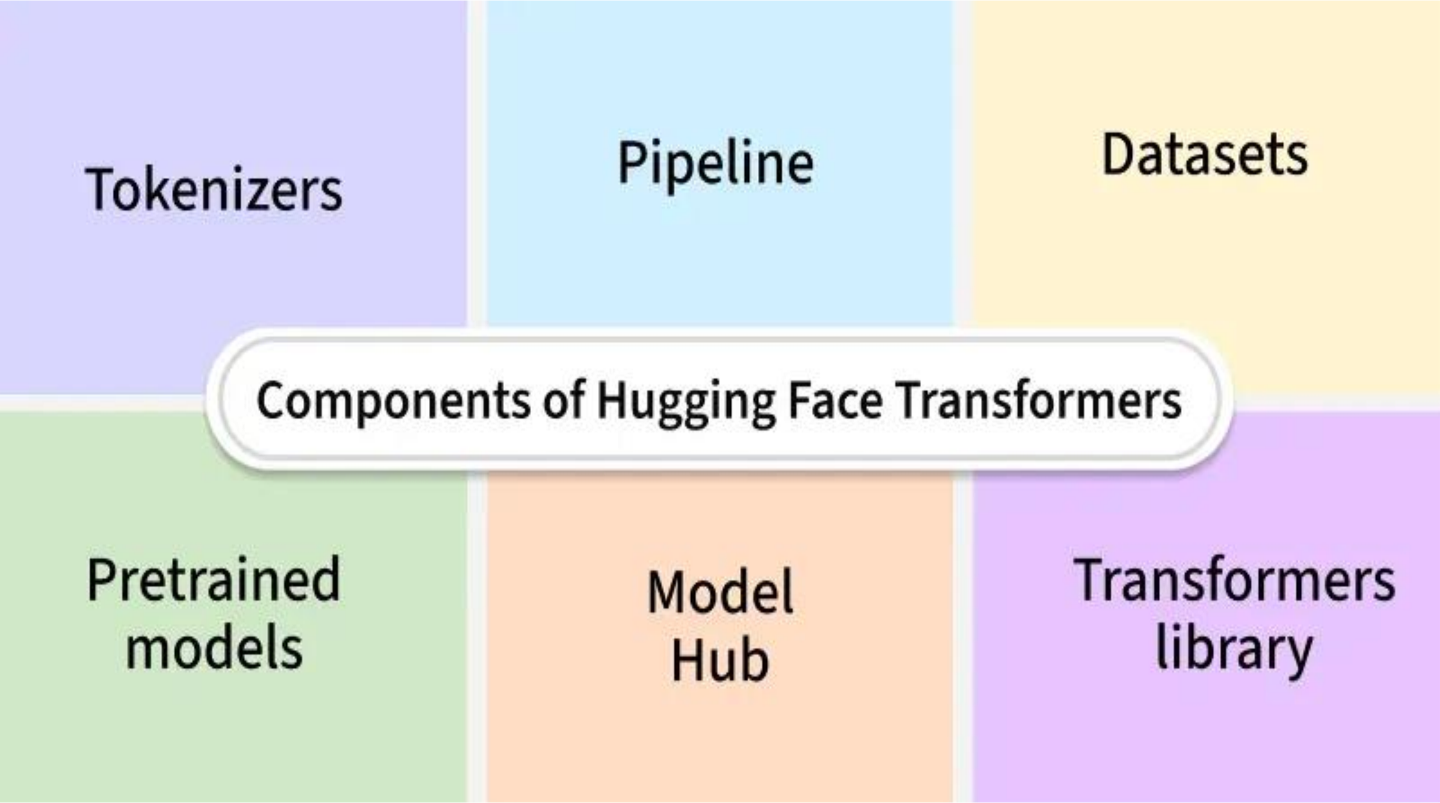
It is not about Theory, it is 20% Theory and 80% Practical – Technical/Development/Programming [Mostly Python based]

Hugging Face & Transformers

Hugging Face Transformers is an open source library that provides easy access to thousands of machine learning models for natural language processing, computer vision and audio tasks. Built on top of frameworks like PyTorch and TensorFlow it offers a unified API to load, train and deploy models such as BERT, GPT and T5. Its versatility and large model hub make it a go-to tool for both beginners and researchers to build AI applications with minimal effort.

Core Components

Lets see core components of Hugging Face Transformers:



Components of Hugging Face Transformers

Tokenizers

Pipeline

Datasets

Pretrained models

Model Hub

Transformers library

Reference:
https://www.geeksforgeeks.org/artificial-intelligence/Introduction-to-hugging-face-transformers/

Hugging Face & Core Elements

Tokenizers: This is responsible for efficiently converting raw text into tokens that transformer models can understand. It ensures text is appropriately tokenized, padded and truncated to match the model's input requirements.

Pipeline: Pipeline abstraction provides a simple interface for running pre trained models on a variety of tasks. It allows users to easily interact with models without writing custom code making it accessible for beginners or for rapid prototyping.

Datasets: This provides access to a wide range of datasets for training and evaluating models. It simplifies the data pipeline, supporting large scale datasets and making it easy to load, filter and preprocess data for use with transformer models.

Transformers Library: It supports PyTorch, TensorFlow and JAX enabling users to train, fine tune and use pre trained models across different frameworks. It removes much of the complexity, allowing users to focus on model development and experimentation.

Model Hub: This is a central repository that hosts thousands of pre trained models from Hugging Face and the community. Users can easily download models, fine tune them and share them with others.

Pre trained Models: Hugging Face provides a vast collection of pre trained models for NLP tasks including text classification, translation, question answering, text generation and more. These models are built on transformer architectures like BERT, GPT-2, T5, RoBERTa, DistilBERT and others.

Reference:

https://www.geeksforgeeks.org/artificial-intelligence/Introduction-to-hugging-face-transformers/

Fine-Tuning LLMs (PEFT, QLoRA)

Fine-tuning large language models (LLMs) is used for adapting LLM's to specific tasks, improving their accuracy and making them more efficient. However full fine-tuning of LLMs can be computationally expensive and memory-intensive. QLoRA (Quantized Low-Rank Adapters) is a technique used to significantly reduces the computational cost while maintaining model quality.

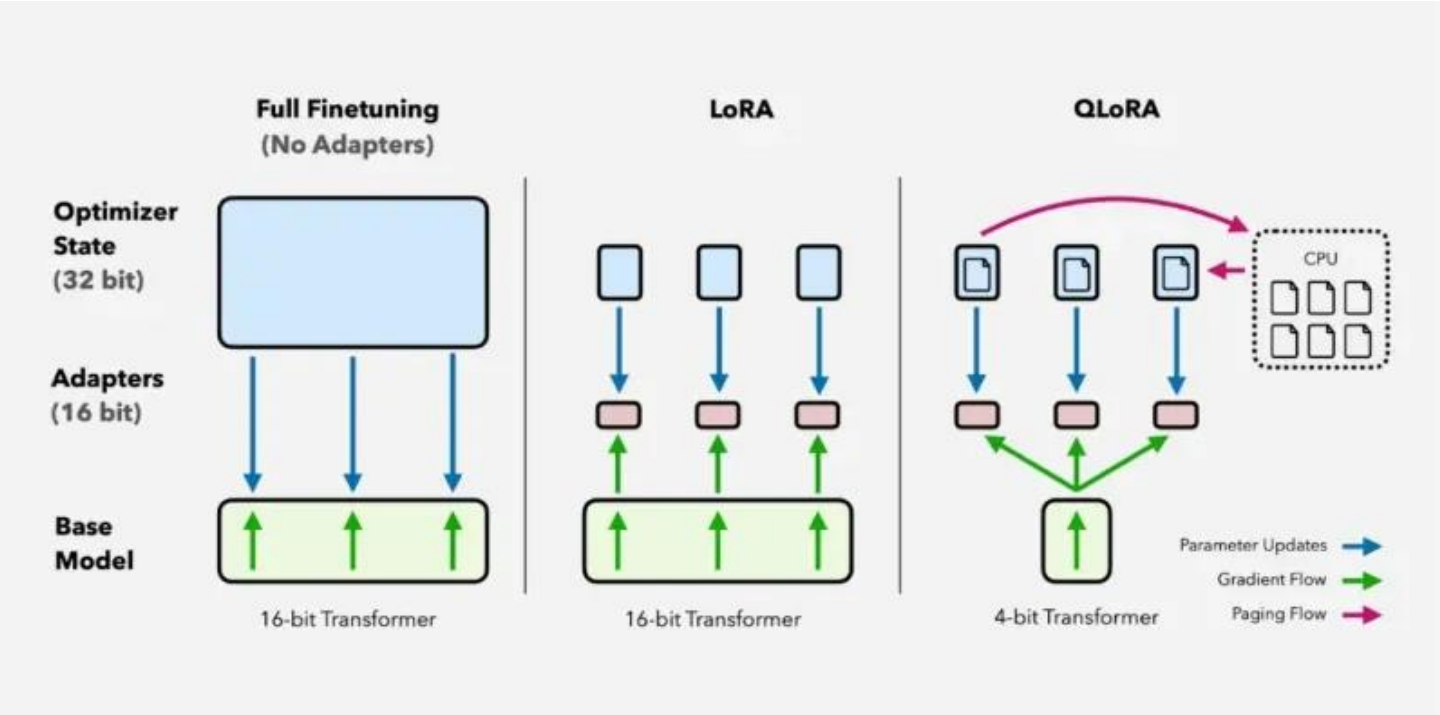
What is QLoRA?

QLoRA is a advanced fine-tuning method that quantizes LLMs to reduce memory usage and applies Low-Rank Adaptation (LoRA) to train a subset of model parameters. This allows:

Lower GPU memory requirements : Fine-tuning large models on consumer GPUs.

Faster training : Using fewer parameters speeds up the process.

Preserved model quality : Achieves similar performance to full fine-tuning.



Reference:

https://www.geeksforgeeks.org/nlp/fine-tuning-large-language-models-llms-using-qlora/

Reference Code: https://github.com/ShahzadSarwar10/FULLSTACK-WITH-AI-BOOTCAMP-B1-MonToFri-2.5Month-Explorer/blob/main/Week10/Case-10-2-Fine-TuningLargeLanguageModels-LLMs-Using-QLoRA.py

Thank you - for listening and participating

- □ **Questions / Queries**
- □ **Suggestions/Recommendation**
- □ **Ideas…..?**

Shahzad Sarwar
Cognitive Convergence
https://cognitiveconvergence.com
shahzad@cognitiveconvergence.com
voice: +1 4242530744 (USA) +92-3004762901 (Pak)