Huggingface

HuggingFace是一個LLM模型的雲端平台(Hub)，擁有全世界的LLM開源模型，資料集。

註冊Huggingface帳號

執行程式前先去Huggingface註冊帳號，並獲取存取Token

Transformers 套件

* Hugging Face結合平台內所有的LLM與資料集，推出Transformers套件，讓LLM的引用、操作、微調，與資料集的使用變得更方便
* Pipeline子套件：單純使用LLM(inference)
* Datasets：LLM資料集與資料操作
* LLM訓練套件(AutoModel、AutoTokenizer)
* PEFT(Parameter-Efficient Fine Tuning)：高效率微調

Pipeline

* Pipeline函數是Transformer用來呼叫大型語言模型(LLM)最簡單的方式
* 指定『任務』，pipeline會自動幫你搭配最適合的LLM，自動幫你完成任務
* 如沒有指定，Pipeline會自動挑選適合的LLM

Datasets套件

* Datasets套件可以呼叫LLM常用資料集
* 也可以針對資料集進行內容操作，以符合LLM的格式

LLM訓練套件

* LLM進行訓練是很複雜且困難，必須特定的模型、斷詞工具(Tokenizer)
* Transformer除了提供各種LLM與斷詞工具，還協助使用者自動挑選，例如AutoModel、AutoTokenizer
* 提供簡化訓練過程，例如Trainer、TrainingArguments

PEFT效率微調

* LLM已經大到無法進行微調
* GPU記憶體不足(Nvidia 4090 GPU: 24G ram)
* PEFT(Parameter-Efficient Fine Tuning)
* 透過降低權重精準度(32bit->4bit)
* 只微調外加Adapter層
* Lora、Qlora、unsloth