第1章 マイクロソフトとMetaが提供してくれるオープンソースのLLM

ローカル環境でGPTベースの大規模言語モデルを使う

佐々木 峻

● GPTベースのLLMをローカルで試せる

2022年11月にChatGPTが発表されてから，多くの注目を浴び，日々新たな使い方や連携したサービスが発表されています．一方，GPTベースのLLM（大規模言語モデル）も数多く開発され，中には商用利用可能なライセンスでモデルを配布しているところもあり，環境さえあれば，GPTベースのLLMをローカルで試すことができるようになりました．

直近では，2023年7月にMetaがマイクロソフトと協力してLLaMA2というモデルを発表し話題になりました．このモデルは個人が申請すれば，モデルそのものをダウンロードして，要求仕様さえ満たせば個人所有のPC上で動作可能，さらに商用利用可能ということで，例えばOpenAIやAzureといったクラウド・サービスを利用できない環境で動かすシステムでも使用できる選択肢として考えられています．

本稿では，モデルが配布されているダウンロード可能なLLMに焦点を当て，これまでの流れを簡単に振り返ります．その後，ChatGPTとの違いや，チューニングのための手法について解説します．

ChatGPTとダウンロード可能なLLMとの違い

LLaMAなどのLLMモデルはChatGPTとは何が違うのかを説明します．ここではMetaが発表したダウンロード可能なLLMであるLLaMA2を例にして，ChatGPTとの違いを見ていきます．

● サービスとモデル

ChatGPTはOpenAIが提供するサービスであり，一方でLLaMAは機械学習モデルというのが根本的な違いです．

ChatGPTはサービスのため，ウェブ画面，またはOpenAIのAPIを通じてのみ使えます．

LLaMA2はMetaがモデル・ファイルをリポジトリから公開しており，各個人のローカル環境にインストール可能です．従って，モデルの出力をそのまま利用できるのと，さらに独自のデータセットと学習方法を使ってファイン・チューニングして自分の目的に特化したモデルを作成するということも可能です．ただし，LLMのモデル・サイズ次第で求められるGPUが変わり，大きいサイズのモデルには強力なGPUが必要になります．

● モデルのサイズ

オープンなLLaMAは各ローカルで実行されることを想定されているので，ChatGPTで利用されているモデルとはモデル・サイズが異なります．

モデル・サイズを表すパラメータ数で見てみると，ChatGPTで利用されているGPT-3.5は3550億個とされている一方，LLaMA2は幾つか種類があり，70億，130億，330億，650億個のモデルが用意されています．最も大きなサイズで650億個なので，GPT-3.5と比較すると1/6程度のパラメータ数で動いており，トレーニングや推論がより低い性能のPCでも動かすことができるようになっています．

● 対応言語

LLMはどの言語の学習データを使用するかによって，出力できる言語が変わってきます．例えばChatGPTで使用されているGPT-3.5は，日本語の出力にも対応しています．その反面，LLaM2Aは日本語の出力には対応していません．このようにLLMによって対応している言語は異なるので，使用する際には確認しておくことが必要です．

幾つかの軸でChatGPTとLLaMA2の違いを紹介しました．LLaMA2などのオープンソースのモデルはローカルで動かせる分，自分でファイン・チューニング・モデルが作成できるなど自由度が高い利用方法が可能ですが，対応言語などの制約はあるので，利用シーンなどを考慮して適切なモデル・サービスを選ぶことが重要になります．まとめると**表1**のような違いになります．