第2章 小さいコストで大規模モデルをカスタマイズできる「LoRA」を試す

ローカルLLMを自分用にファイン・チューニング

佐々木 峻

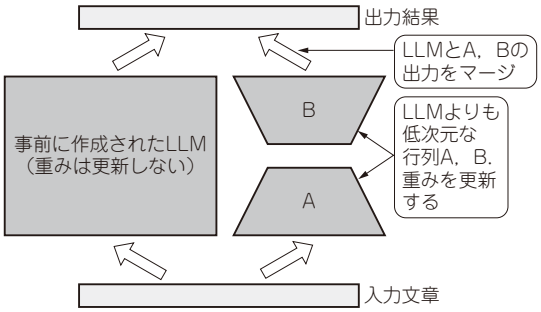

図1[1] **学習時のLLMのベース・モデルとLoRAの関係**
文献(1)の図に筆者が日本語のコメントを追記

前章では，頒布されているローカルLLM（Large Language Models）を使って，幾つかの文章生成タスクを試しました．頒布されているモデルをそのまま利用するだけでもさまざまな用途に対応できました．

しかし，ローカルLLMの良いところはベースのモデル（ファイン・チューニングされていないもの）から，自分用にファイン・チューニング済みモデルを作ることができる，つまり自分用にカスタマイズできることです．

本稿ではLLMのファイン・チューニング手法のLoRA（Low-Rank Adaptation of Large language Models）を使ってローカルLLMをカスタマイズします．

（編集部）

● 小さいコストで大規模モデルを学習できる「LoRA」

LoRAはマイクロソフトの研究者が2021年に提唱した学習手法[1]です．より小さいコストで大規模なモデルを学習できるということで注目を集めました．この手法により，サイズの大きいLLMをローカル・マシン上でファイン・チューニングすることが可能になりました．

● LoRAが小コストで学習できる仕組み

通常，ファイン・チューニングというのは，モデルの各パラメータを更新するため，学習元のモデル・サイズが大きいほどメモリも大きいサイズが必要です．しかし，LLMは，モデル・サイズ自体が非常に大きく，ファイン・チューニングするにも高性能なGPUマシンを長時間動かす必要がありました．

LoRAを使った学習の構成を**図1**に示します．左側が学習元のLLMのベース・モデルで，右側がLoRAで学習するための軽量モデルです．LoRAは，モデルのパラメータを更新するのではなく，モデルの出力を調整するための軽量なモデルを学習することで，出力だけを効率的に調整します．LoRAでは左側のベース・モデルのパラメータは更新せず，右側の軽量なモデルを学習することで，より少ないリソースで学習します．

LoRAを，GPT-3（パラメータ数1750億）に適用すると，学習時に必要なGPUメモリが1.2Tバイトから350Gバイトと，約1/3になったとのことです．この手法によって，ローカルLLMを個人のマシンでも，性能が低いGPUでも，より短い時間でカスタマイズすることが可能になりました．

実践！Chat形式に特化させたLLMへカスタマイズ

日本語に強いjapanese-stablelm-base-alpha-7b[2]をベースのモデルとして，チャット形式の出力を学習したモデルを作成します．

ベースのモデルをそのまま使用すると，意図しない形式で文章が生成されてしまいます．そこで，欲しい形式で文章を出力できるようにLoRAを使います．

● 使用するデータセット

質問と回答がセットでそろっているデータセットを使用する必要があります．今回はdatabricks-dolly-15k-ja[3]というデータセットを使用します．このデータセットのうち，今回はシンプルな質問と回答のみを使用して学習元データとします．