第6章 LoRA，量子化，QLoRA，Key-Valueキャッシング

ローカルLLMを「賢く，小さく，高速に」動かすワザ

川原田 将之

● ローカルLLMの課題「賢く，小さく，速く」をいかに実現させるかが問題

手元でLLMを動かすローカルLLMには，主に4つのメリットがあります．

- 低レイテンシ（高速な応答）…クラウドとの通信が不要なため，応答速度が格段に向上します．
- オフライン動作…インターネットがつながっていない環境でも，LLMを利用できます．
- プライバシ保護…自分や会社の重要なデータをクラウドに送信しないため，情報漏洩のリスクを低減できます．
- コスト削減…クラウド・サービスのAPI利用料を節約できます．

ローカルLLMは上記のメリットがある一方で，クラウドと比べて計算能力やメモリ容量に制約があります．その制約の中で，LLMをいかにして賢く，速く動かすかが大きな課題です．そこで本稿では，この課題を解決するため，**表1**の3つの技術を解説します．

まずはLLMの基本的な仕組みをサッと見てみる

LLMが文章を理解したり生成したりする基本的な仕組みを説明します．現在のLLMのほとんどは，トランスフォーマというアーキテクチャをベースにしています．

表1　本稿で扱うLLMを賢く，小さく，速くする技術

課　題	使用する技術	技術の概要
賢く	LoRAによる効率的なファインチューニング	特定のタスクにLLMを特化させるための，パラメータ効率が高い学習手法
小さく	量子化によるモデルの軽量化	モデルのメモリ使用量を抑え，限られたリソースでも動かせるようにする
速く	Key-Valueキャッシングによる推論の高速化	LLMの文章生成（推論）処理を効率化する

● Attention（アテンション）機構

トランスフォーマの最大の特徴は，Attention（アテンション）機構です．一言でいうと，文章中の単語の関連性に重み付けをする仕組みです．例えば，「昔 ある ところ に」という入力から次に続く単語「おじいさん」を予測する場合（**図1**），アテンション機構は「ところ」や「に」といった単語が，次に来るであろう人物を表す単語と強い関連性を持つと判断し，それらの単語に高い重みを置きます．このように，文脈に応じて単語間の関連性の強弱を動的に学習することで，LLMは文脈を深く理解し，自然な文章を生成できます．

● LLMのアーキテクチャの特徴

トランスフォーマは元々，機械翻訳のために考案さ

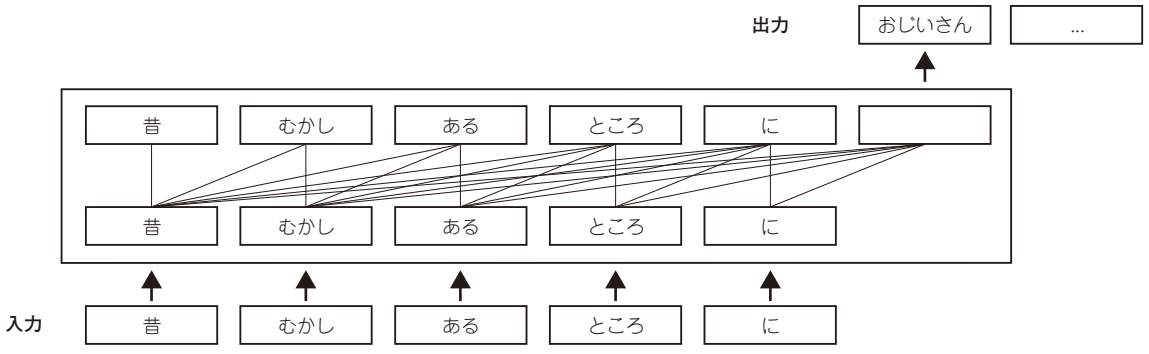

図1　アテンション（注意）機構の仕組み