パラメータ効率の良いチューニング方法

- LLMは膨大なパラメータを持つため、推論に必要なメモリ量が莫大
  - ➡ モデルを圧縮することでメモリ使用を削減できる
- 圧縮の方法として一般的には「量子化」が行われる
  - ➡ 量子化とは、浮動小数店からビット数の小さい整数にモデルのパラメータやアクティベーション (中間層) の値を変換して圧縮すること
  - ➡ 量子化を考慮した学習 (QAT) と学習後の量子化 (PTQ) があるが、後者の方が計算コストがはるかに低い
  - ➡ LLMは小さい言語モデルよりも量子化の影響を受けにくく、同じメモリコストであれば小さなモデルよりも量子化した大きなモデルを使用する方が良い

量子化に関する実証的分析と知見

- アクティベーションの量子化は、モデルが大きくなると外れ値が発生して困難になる
- 量子化されたLLMの性能を向上するには、効率的なファインチューニングを行うのがよい
- 8ビットの量子化ではまず性能は低下しないが、4ビットや3ビットの量子化については、性能が下がらないような戦略が必要