パラメータ効率の良いチューニング方法

- アダプター・チューニング
  - 小さなニューラルネットワークのモジュールをTransformerに組み込む
  - アダプターは特定のタスクに最適化されるが、元の言語モデルのパラメータはそのままのため、学習するパラメータを減らすことができる
- Prefix Tuning
  - Transformer層に学習可能なベクトルの集合を付加する
  - ベクトルの値だけが学習されるため、パラメータ効率が良い

パラメータ効率の良いチューニング方法

- プロンプト・チューニング
  - 入力層に学習可能なプロンプトのベクトルを埋め込む
  - 入力プロンプトを補強して、特定のタスクに特化させられる
- Low-Rank Adaptation (LoRA)
  - モデルのパラメータの一部をより小さな行列に近似し、この行列に対して更新を行うことで次元数を減らして効率的に学習する
  - 出力にかかる時間が増加したり入力テキストの長さが制限されない