進捗報告

佐藤　力

富山県立大学
u220029@st.pu-toyama.ac.jp

July 15, 2025

研究テーマとその説明

研究テーマ

大規模言語モデルに組み込む動的適応プルーニングの提案手法

大規模言語モデル

大規模言語モデル (LLM) とは大量のデータとディープラーニング（深層学習）技術によって構築された言語モデル。テキストの生成や質疑応答などができる

大規模言語モデルの課題点

大規模言語モデル (LLM) は莫大なパラメータを保有する。
モデルのサイズが大きくなり推論に時間がかかる。
計算リソースが多く消費電力も大きい。

研究テーマとその説明

プルーニング

モデル削減の手法の一つ。影響の少ないパラメータを削減する手法であり、パラメータを減らすことで計算量も減る。
結果としてモデルのサイズ削減のみならず推論時間、モデル完成までの時間短縮も見込める。

図 1: プルーニング

研究テーマとその説明

静的プルーニング

静的プルーニングとは学習終了後や決められた周期で行われるプルーニング。

静的プルーニングの課題点

・学習終了時に行うプルーニングは、学習中不要な重みも計算が実行されるため、訓練時間や電力消費の増大に直結する。
・周期的に行われるプルーニングでは、一時的に重要度が低く見えただけで実際には必要なパラメータを削除してしまう可能性がある。

研究テーマとその説明

動的プルーニング

動的プルーニングとは学習中に不要な重みを探知してプルーニングする手法。

動的プルーニングのメリット

・学習中に不要な重みを特定して、プルーニングをするため計算リソースを減らすことができる。
・常にパラメータの動きを監視できる状態であるため、一時的に重要度が低いパラメータを削除するリスクが減らせる。

研究テーマとその説明

プルーニングの課題

LLM のような膨大なパラメータを扱うモデルでは、正則化をかけても過学習の抑制と損失関数の最小点のバランスを取ろうとするため、不要な重みが漸近的な 0 になるだけで、完全な 0 にならないことがある。

研究の着眼点

・漸近的な 0 の重みを 0 にしたい。
・一時的に重要度が低い重みではなく冗長に下回る重みの削除を目指したい。

提案手法

提案手法

重みの更新式にターミナルアトラクタを組み込む。

ターミナルアトラクタ

指定した時間内に指定した値に到着する。
そのため、今回問題視されている漸近的な 0 な問題も解決できると考えられる。

ファインチューニング

事前学習がされているモデルに行う追加学習。
今回はこのファインチューニングを中心とした学習に着眼し、研究を進めようと考えている。

学習の流れ

学習の流れ

1. 学習開始
2. 不要な重みの特定
3. 検証・評価
4. ターミナルアトラクタの組み込み
5. 特定の重みを 0 にする

引き続きステップ 1 に戻る
学習終了後、モデルの最終的な評価を行う。

現在の進捗

現在の進捗

前回、windows で作った環境が Linux に移行できなかったため、モデルを動かす環境の構築を完成させた。