進捗報告

佐藤　力

富山県立大学
u220029@st.pu-toyama.ac.jp

June 17, 2025

TOYAMA Prefectural University

研究タイトル

（仮）大規模言語モデルの動的適用プルーニング.

研究内容

・背景：LLM では学習するために必要なパラメータが膨大である
・目的：LLM で問題視されているモデルの大きさを簡略化するためのもの
・概要：現在提案されている手法には量子化や静的プルーニングがあるが、動的プルーニングに変換することでモデルサイズの削減につながるのではないか

仮決定したこと

プルーニングする対象

重みが定番らしいので重みを選定

重要度評価指標

影響力を単純に図るため絶対値の大きさ

動的適応のロジック

正解とモデルが予想した値との差を関数化した損失関数の定めた規定のラインを指定

TOYAMA Prefectural University

進捗

モデルの選定

Llama3

環境の構築

Anaconda 仮想環境下での CUDA 環境の構築