2023 年度年次報告書
社会変革に向けた ICT 基盤強化
2023 年度採択研究代表者

李 鵬

会津大学 コンピュータ理工学部
上級准教授

環境適応型エッジ AI による巨大モデル利活用基盤

研究成果の概要

本研究の目的は、スマートフォンなどの汎用 CPU や GPU を備えたエッジデバイス上で巨大 AI モデルの推論を効率的に実行可能とする手法の開発することである。エッジデバイスに搭載されるモデルは、そのデバイスが存在する環境に適した対話や画像処理ができれば良い。そのローカル性の特徴を活かし、元の巨大 AI モデルを多数の小型エキスパートモデルに再構築し、環境の変化や通信・計算資源の変動に応じてエッジデバイス上での必要充分なエキスパートを動的に切り替える技術を確立する。本研究は 2023 年 10 月から開始され、以下の研究タスクを順調に完了しました。

A. エッジ環境に適用する巨大 AI モデルの再構築
本研究タスクでは、Mixture-of-Experts（MoE）技術を活用し、元の巨大 AI モデルを複数のエキスパートモデルに分割して再構築する。現在の MoE は、主にデータセンター環境を対象としており、複数の同質の専門家ネットワークを構築し、入力データを適切なエキスパート（専門家ネットワーク）とマッチングするゲーティングネットワークを設計する。しかし、エッジ環境では、計算および通信リソースは異質性と多様性であるため、専門家モデルを最適化することを行う。具体的には、デバイス親和性モデルの構築と異種専門家ネットワークの構築を提案しました。

B. 微調整（fine-tuning）技術による再訓練コストの削減
高精度を達成するために、MoE 化したモデルを再訓練する必要がある。しかし、複数のエキスパートを含むモデル全体は大きいため再訓練のコストが高くなる。そこで、微調整技術を導入し、再訓練のコストを大幅に削減する。具代的には、アダプタ技術を用いた微調整を提案し、強化学習に基づく同期方法を設計した。

【代表的な原著論文情報】

1) Fahao Chen, Peng Li, Shengli Pan, Lei Zhong, and Jing Deng, “Giant Could be Tiny: Efficient Inference of Giant Models on Resource-Constrained UAVs”, IEEE Internet-of-Things Journal, accepted.
2) Tianyu Qi, Yufeng Zhan, Peng Li, and Yuanqing Xia, “Tomtit: Hierarchical Federated Fine-Tuning of Giant Models based on Autonomous Synchronization”, IEEE International Conference on Computer Communications (INFOCOM) 2024, accepted.