数理・情報のフロンティア
2021 年度採択研究者

2021 年度
年次報告書

柴 康太

東京大学 大学院工学系研究科
大学院生（博士課程）

積層型 AI チップの低電力高効率アーキテクチャ

§1. 研究成果の概要

2021 年度はプルーニングされた AI モデルを高効率に圧縮する技術の研究をおこなった。広く使われているランダムプルーニングではニューラルネットワークで不要なニューロンやシナプスを刈り取ることで高い推論精度を維持しながらパラメータを大幅に削減することができる。しかし、生成される疎行列の構造が不規則になるためデータの圧縮が非効率であり記憶容量が増大する。この課題を解決するため 2021 年度は 2 つの研究を行った。

1 つ目は疎行列の圧縮手法である。データを高効率に圧縮する手法として N:M プルーニングが提案されており、重み行列の M 個の要素から構成されるブロックから N 個のパラメータだけを残して他のパラメータをプルーニングすることによって規則的な構造生み出し圧縮効率を向上させている。しかし、厳しいプルーニング制約のために推論精度が悪化してしまう課題が存在する。本研究ではランダムプルーニングで生成された推論精度は高いが不規則な疎行列を並び替えることによって、N:M プルーニングのような規則的な行列を生成し高い推論精度と高い圧縮効率を同時に達成する手法を提案した。本手法を活用することで推論精度を維持したまま記憶容量の従来比 63%削減を達成した。

2 つ目は推論処理ハードウェアである。並び替え自体は学習後にクラウドでおこなわれるためエッジデバイスのハードウェアリソースを消費しないが、並び替えられた行列の処理はエッジで高効率におこなう必要がある。しかし、並び替えられた行列を処理する際に外部メモリへのアクセスがランダムになる傾向があるため、ランダムアクセスが不得意な DRAM はストールを引き起こす。そこで本研究ではランダムアクセス可能な積層 SRAM を活用することによってオンチップバッファの 20%削減と演算性能の 1.4 倍改善が同時に達成されることをサイクルシミュレーションで確認した。

【代表的な原著論文情報】

1) K. Shiba, M. Okada, A. Kosuge, M. Hamada, and T. Kuroda,
"Polyomino: A 3D-SRAM-Centric Architecture for Randomly Pruned Matrix Multiplication with Simple Rearrangement Algorithm and x0.37 Compression Format,"
IEEE International New Circuits and Systems Conference (NEWCAS), June 2022, submitted.

2) 柴康太, 小菅敦丈, 濱田基嗣, 黒田忠広,
"[招待講演] 三次元積層 SRAM と近接場無線接続技術,"
電子情報通信学会(IEICE) 集積回路研究会(ICD) メモリ技術と集積回路技術一般, Apr. 2022.

3) K. Shiba, T. Omori, K. Ueyoshi, S. Takamaeda-Yamazaki, M. Motomura, M. Hamada, and T. Kuroda,
"A 96-MB 3D-Stacked SRAM Using Inductive Coupling with 0.4-V Transmitter, Termination Scheme and 12:1 SerDes in 40-nm CMOS,"
IEEE International Symposium on Circuits and Systems (ISCAS), May 2022, accepted.

4) T. Omori, K. Shiba, A. Kosuge, M. Hamada, and T. Kuroda,

"A Physical Verification Methodology for 3D-ICs Using Inductive Coupling,"
IEEE Electrical Design of Advanced Packaging and Systems (EDAPS), pp. 72-74, Dec. 2021.