No. A-12

会津大学

３次元アレイ内データ循環による畳み込みニューラルネットワークの高効率ニアデータ処理

上級准教授　富岡　洋一 / 名誉教授　Stanislav Sedukhin

省エネかつ高性能なニューラルネットワーク用プロセッサの開発

関連発明：データ処理装置、及びこれにおけるデータ処理方法（特願2017-117686【特許第7014393号】）、データ処理装置及びデータ処理方法（特願2020-043230【特許第7467786号】）、テンソルメモリ及びテンソルプロセッサ（特願2022-056069）

概要

〇近年，Convolutional Neural Network (CNN)等の深層学習技術の発展により画像認識の品質が急速に向上しています。このような最先端の画像認識技術を監視、見守り、自動運転システム等多くのシステムで活用していくことが期待されており、データ処理速度、エネルギー効率に対する要求も、日々厳しくなっています。
〇今日のコンピューティングでは、計算よりも、メモリアクセスやデータ伝送に要する電力の方が支配的となっており、データの近くにProcessing Element (PE)を配置し、メモリアクセスや長距離のデータ転送を避けることでエネルギー効率を向上するニアデータ処理方式のアーキテクチャが注目されています。
〇我々はニアデータ処理方式でCNNのマルチチャネルの畳込み計算を実行する3次元アレイプロセッサとアルゴリズムを提案しました。本アレイプロセッサは、アレイプロセッサ内部でデータを循環してデータを最大限再利用しつつ最小ステップ数で畳込み演算を実行します。メモリアクセスも削減し、大規模並列計算を実現できることから、電力効率と性能効率の良い計算を実現します。

実用化の可能性

〇監視カメラや自動運転、ディープラーニングサーバ等、幅広い画像認識アプリケーションを応用先として想定しています。
〇提案方式では、テンソルデータの要素数と同じ並列度で計算を行いますが、大規模なCNNを少ない回路資源で実現するブロッキング方式についても検討しています。

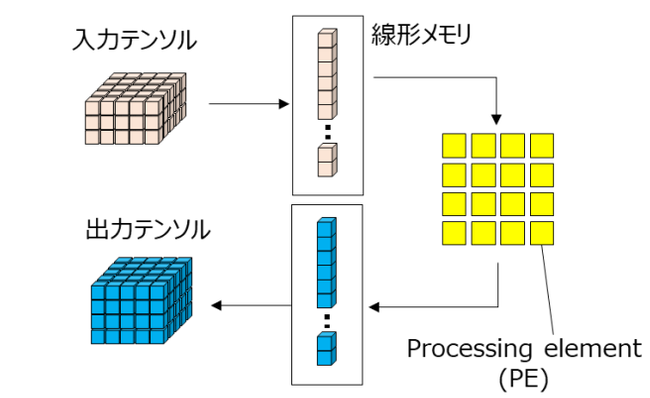
UBICからのメッセージ

ここ数年の深層学習の進展と普及により、さまざまな分野において実用面での適用が加速しています。人間の脳を模擬したニューラルネットワークですが、現状の計算機アーキテクチャでは、データ伝送部分が消費電力と処理性能の向上を妨げる一因となっています。本技術は、3次元アレイプロセッサ内でデータを循環させることにより、上記課題を解決し、省エネと高性能化を目指します。深層学習は、ソフトウェア・ハードウェアの両面から、ますますの発展が期待されています。

研究概要図

従来方式

テンソルデータを1次元データとして線形メモリに保存
そこそこの並列度、高い動作周波数でリアルタイム処理を実現



提案方式

テンソルデータを形状を維持したまま保持
高並列度、低動作周波数、ニアデータ処理で高電力効率化