AI コンピューティング アーキテクチャの研究

准教授　劉　載勲
研究分野：コンピュータビジョン・DNN プロセッサ
ホームページ：http://www.artic.iir.titech.ac.jp

研究目的・内容

深層ニューラルネット(DNN)技術の勃興とともに、人工知能(AI)コンピューティングの分野が大きく進展しています。従来型のコンピューティングが「手続き型」であるのに対し、AI コンピューティングの分野は「構造型」であることを大きな特徴としています。その違いをアーキテクチャ(=処理方式)の革新に活かすことで、これまでよりも大幅にエネルギー効率や処理速度が高いコンピューティングシステムの実現が可能となります(図 1)。このような観点から世界中でアーキテクチャ変革の大規模競争が始まっています。

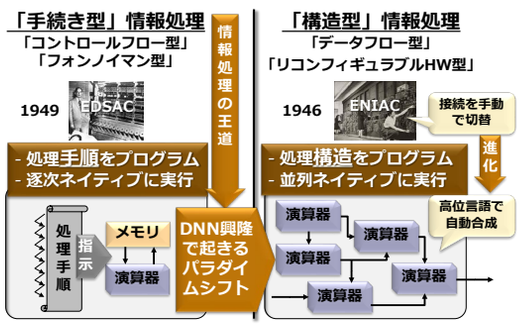

図 1.　AI コンピューティング時代の情報処理パラダイムの変革

劉研究室では、その流れを先導するために本村研究室と一体となって、科学技術創成研究院 AI コンピューティング研究ユニットを構成しています。アルゴリズム、アーキテクチャ、回路の協調性に注目した機械学習のためのソフトウェア・ハードウェア協調システムの実現を目指します。

研究テーマ

１．DNN アクセラレータの研究

DNN は高い学習能力と推論精度のため、多くの分野における応用が期待されていますが、その代償として多くの計算量を要求します。そのため、現在実用化されている DNN はクラウドサーバーがその処理を担う場合がほとんどと言えます。DNN の膨大な計算量は結果としてネットワーク環境の安定性が保証できないドローンやロボット、車などの自立型の組込みシステムにおいてその実用化を妨げる大きな原因となっています。

我々は、組込みシステムにおける処理時間、消費電力、計算資源の制約下で DNN の実用化を達成するべく、量子化、枝刈り、蒸留などの様々な近似コンピューティングを用いた DNN の効率的な実装アルゴリズムについて研究を行うと同時に、それに基づく低電力かつ高速な DNN 処理を支援するハードウェアアーキテクチャの研究を行います。アルゴリズムとハードウェアの設計段階からシステム全体の効率的な構成を考えることによって既存研究では実現できなかった DNN に対する高い処理性能と低電力化の実現を目指します。

２．DNN の知見を活かした既存機械学習の性能向上と実用化の研究

現在 DNN を支える深層学習はその学習対象をニューラルネットワークとする場合がほとんどです。しかし、深層学習そのものは非線形変換の繰り返しによって機械学習の表現力を向上させるための学習概

念であり、その応用先をニューラルネットワークに限ったものではありません。機械学習にはニューラルネットワーク以外にも、サポートベクターマシン（SVM）、ランダムフォーレスト（RF）、ブースティング決定木（BDT）など様々な手法が存在し、少ない計算量で高い表現力を実現することが可能です。これらの機械学習手法が持つ主な問題は、その表現能力を活かすために必要な学習能力の不在と言えます。

我々は DNN のテクニックの中で小規模ネットワークの学習能力不足を補う蒸留と呼ばれる手法を SVM、RF、BDT などの非ニューラルネットワークの機械学習に適用することで、DNN では実現が困難な低計算量化を実現します。また今までの研究で蓄積した既存機械学習手法の FPGA 実装方式の知見を活用（図2、3）し、機械学習アクセラレータにおける新たなブレークスルーを目指します。

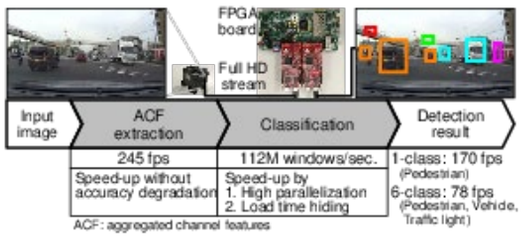

図 2. BDT による物体検出 HW

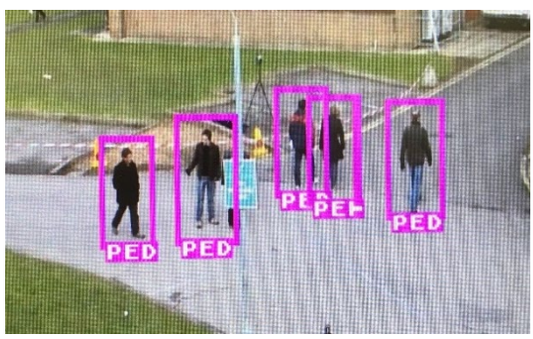

図 3. BDT による歩行者検出 HW デモ

３．ニューロモルフィックコンピューティングによる学習・推論の研究

人間の脳は情報を処理するニューロンとニューロン間を接続するシナプスで構成されています。各ニューロン間ではディジタルの 2 値信号に似た電気的スパイクの有無を用いて情報伝達が行われ、伝達されたスパイクが各ニューロンの膜電位の変化と発火を引き起こすことで情報処理が行われます。DNN で代表される人工ニューラルネットワークは人間の脳を模倣していると言われますが、実数値の積和演算を用いる点や信号の時間的ズレを用いない点で生体脳とは大きく異なり、エネルギー効率の面で人間の脳に比べて大きく劣ります。我々は生体脳の特徴をより正確に再現したニューロモルフィックコンピューティングに基づく新たな学習・推論アルゴリズムの研究とその実装方式の研究に取り組みます。

●教員からのメッセージ

2019 年度 9 月まで大阪大学 情報科学研究科にて主にコンピュータービジョンを対象に機械学習とパターン認識、そしてそのためシステムアーキテクチャ設計の研究を行ってきました。2019 年度 10 月から科学技術創成研究院（すずかけ台）の AI コンピューティング研究ユニットに参加し、日本の中核研究拠点を作ることを目指して活動しています。本村研究室と劉研究室に配属される学生の指導は AI コンピューティング研究ユニットとして、本村研究室と劉研究室の区別なく、一体となって行います。AI コンピューティング研究ユニットの HP を参照してください。

●関連する業績、プロジェクトなど

1. 基盤（B） 近似コンピューティングを活用した深層ニューラルネットワークアクセラレータの開発
2. 8 件の特許出願（https://www.j-platpat.inpit.go.jp/で「劉 載勲」「ユ ジェフン」を検索）

受賞　18 年 T-SLDM Best Paper Award、16 年 SISA Student Best Paper Award（指導学生）、15 年 IEEE 関西支部学生研究奨励賞、14 年 画像電子学会優秀論文賞、13 年 ITC-CSCC Best Paper Award、12 年 スマートインフォメディアシステム研究会若手研究優秀賞