「革新的コンピューティング技術の開拓」
2018年度採択研究者

2018 年度
実績報告書

高前田 伸也

北海道大学大学院情報科学研究科
准教授

アーキテクチャとアルゴリズムの協調による高効率深層学習システムの創出

§1. 研究成果の概要

本研究は、利用可能なエネルギーが制限されているエッジコンピューティング環境において、高速かつリアルタイムに安心・安全な認識・判断処理を行うことができる機械学習システム、特に、深層学習システムを実現することを目的に、ハードウェア・アーキテクチャとアルゴリズムの協調による高精度化、高速化、高電力効率化、高信頼化技術の実現を目指すものである。

2018 年 10 月～2019 年 3 月までの半年間では主に、軽量なハードウェア上で高い認識精度を達成するための、低ビット精度演算に基づくハードウェア指向ニューラルネットワークモデルとそのハードウェアアーキテクチャに関する研究、および、深層ニューラルネットワークを高速処理するハードウェアの自動設計を可能にする高位合成コンパイラの開発に取り組んだ。

低ビット精度演算に基づくハードウェア指向ニューラルネットワークについては、ディザ拡散を用いた高精度な二値化ニューラルネットワークモデル Dither NN と、対応する軽量ハードウェアアーキテクチャを提案した（図 1）。二値化ニューラルネットワークは重みと活性を 1 ビットで表現することにより、演算の軽量化とデータ量の削減ができるという利点がある一方で、認識精度が低いという致命的な課題が存在する。本研究では、二値化ニューラルネットワークの演算の

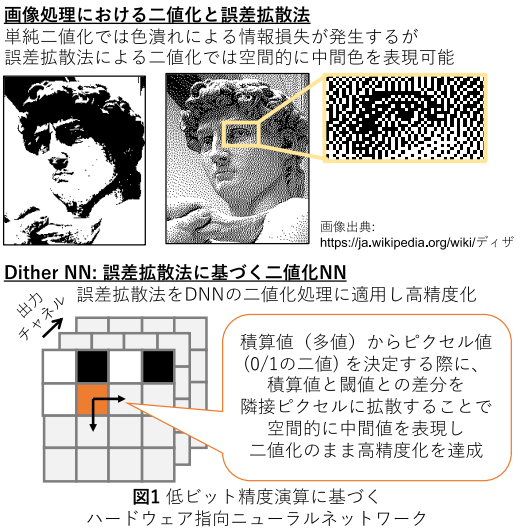

図1 低ビット精度演算に基づく
ハードウェア指向ニューラルネットワーク

NNgen:
DNNハードウェアの自動合成を可能にする高位合成コンパイラ
Pythonによる演算グラフから
DNN専用HWを自動的に合成し
FPGA上にDNN-HWを短期間で実現

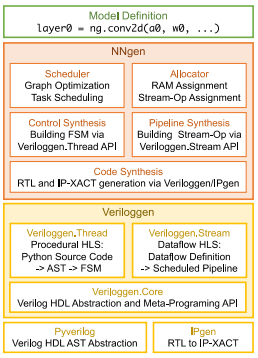

図2 DNNハードウェア
自動合成コンパイラ

内部に存在する多値の情報を積極的に活用することで、従来の二値化ニューラルネットワーク同等の回路規模と計算速度で、より高い認識精度を達成することを可能にした。本成果は国際会議 FPT'18 にて発表し、Best Paper Award を受賞した。

アーキテクチャとアルゴリズムの研究に加えて、ニューラルネットワークを高速に処理可能な専用ハードウェアの自動設計を可能にする、高位合成コンパイラ NNgen の開発を行った(図 2)。演算グラフのレベルで処理を記述するだけで、専用ハードウェアを合成可能なコンパイラであり、現在オープンソース化に向けて準備を進めている。

§2. 研究実施体制

① 研究者: 高前田 伸也 (北海道大学大学院情報科学研究科 准教授)

② 研究項目

・二値化ニューラルネットワークに基づく、高精度モデルの開発

・深層ニューラルネットワーク処理ハードウェア向け高位合成コンパイラの開発