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

2019 年度
実績報告書

高前田　伸也

東京大学大学院情報理工学系研究科
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

2019 年 4 月～2020 年 3 月までの 1 年間は、軽量ハードウェアで高い認識精度を達成する二値化ニューラルネットワークの新たな活性関数とそのハードウェアアーキテクチャに関する研究、ベイジアンニューラルネットワークに適した FPGA 向けオーバーレイアーキテクチャ、動的無効ニューロン予測による計算高速化技術、および、学習済みモデルからニューラルネットワークアクセラレータのハードウェア記述を自動合成する高位合成コンパイラ NNgen の開発の研究に取り組んだ。

二値化ニューラルネットワークの活性化関数およびハードウェアアーキテクチャに関する研究では、隣接ニューロン間の差分を二値化する差分二値化という新しい活性化関数を開発した（図 1）。隣接ニューロンの大小関係を二値化することで、情報の強弱を次のレイヤーに伝播しやすくなり、認識精度の向上を達成した。また、本手法に適したハードウェアアーキテクチャを提案し、回路面積の増加はわずかであることを確認した。

ベイジアンニューラルネットワークに適した FPGA 向けオーバーレイアーキテクチャの研究では、重みの確率分布から実際の処理に用いる値をサンプリングする回路について、逆関数法とルックアップテーブルを用いる軽量アーキテクチャを開発した。

動的無効ニューロン予測による計算高速化技術については、活性化関数 ReLU により多くのニューロンの出力値が 0 になることに着目して、出力が 0 になるニューロンを予測するモデルとそのためのアーキテクチャ支援により、出力値 0 のニューロンの計算を動的に省略する方式を提案した。

ニューラルネットワークハードウェアの高位合成コンパイラ NNgen については、処理速度および回路面積の最適化の他、汎用ニューラルネットワーク表現モデル ONNX からのインポート機能、学習済みモデルの量子化機構などを新規開発し、オープンソースソフトウェアとして GitHub 上（https://github.com/NNgen/nngen）で公開した（図 2）。

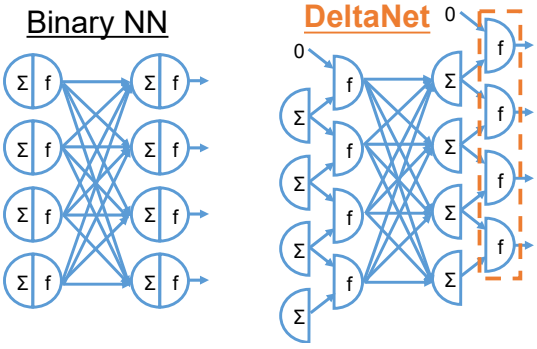

図 2 差分二値化活性化関数 Delta

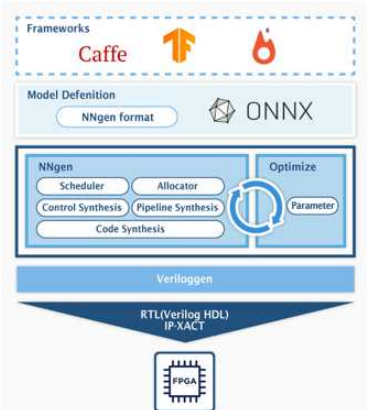

図 1 NNgen