

組込みシステム向け DNN 開発支援システムの開発

松本 斗貴[†] 中本 幸一[†] 山本 椋太[‡] 本田 晋也[‡] 若林一敏[§]
[†]兵庫県立大学大学院 [‡]名古屋大学大学院 [§]日本電気（株）

1 はじめに

近年、深層学習が注目を集め、様々な分野で利用され始めている。様々な学習方法が提案され、GPU を用いて 100 万枚を超える大規模な学習サンプルを数日から数週間かけて学習し、高い認識精度を達成した事例も存在する。

しかし、組込みシステムにおいては厳しいリソース制約が存在するため、低消費電力、省メモリかつリアルタイム性を考慮した実装が求められる [1]。その中で、GPU を利用した実装は消費電力が過大となり、CPU だけを利用した実装は性能が不足し、リアルタイム性を満足できない。

そこで、我々は、FPGA を用いた組込みシステム向け深層学習フレームワークの構想 [2] を提案した。

2 組込みシステム向け深層学習フレームワークの構想

提案したフレームワークの概要を図 1 に示す。提案したフレームワークでは組込みシステムの厳しいリソース制約を満たすために FPGA を使用する。その理由は、FPGA は特定の演算に対する回路を柔軟に構築可能であり、CPU と比べて高速かつ GPU と比べて電力消費を抑えて演算することが可能だと言われているためである。

さらに、提案したフレームワークでは、深層学習による学習と推論を別の環境で行う。はじめに、既存の汎用マシン向け深層学習フレームワークによる学習をデスクトップ PC 等のリソースの豊富な環境を行う。学習の結果、深層学習の推論において必要なパラメータが計算される。これらのパラメータおよび学習に使用したニューラルネットワークの構成情報から、組込みシステム向けに深層学習の推論部分の C 言語ソースコードを生成し、推論処理だけを組込みシステムで行うことで厳しいリソース制約を満たすことを目標としている。

提案したフレームワークでは、深層学習の学習と推

論を分けて行うことに加えて、FPGA を使用することで厳しいリソース制約を満たし、組込みシステムで深層学習を利用することを目指している。

しかし、深層学習に推論処理だけを行う場合においてもリソースを多く要する。その理由は、DNN の特性である膨大な量の浮動小数点数パラメータを使用した積和演算にある。推論に必要なパラメータの数は小規模なネットワークで 120 万個あり、これらのパラメータは浮動小数点数であるため 32bit で表現される。これらの膨大な量の浮動小数点数パラメータを用いた大量の積和演算は、組込みシステムの厳しいリソース制約を満足できない。

そこで、小規模なネットワーク、データセットを用いて学習した結果において、浮動小数点数パラメータを、整数に近似する方法 [3] を提案した。この手法により、32bit 表現だったパラメータの値を 1bit や 2bit の整数で表現することができる。その結果、メモリ使用量を大幅削減し、さらに積和演算をシフト演算と論理演算に置き換えることが可能になり、深層学習による推論処理を組込みシステムの厳しいリソース制約を満たしつつ行える。

しかし、何百万とパラメータの値を手作業で近似する作業は非常に非効率である。さらに近似によって浮動小数点数パラメータを近似する方法は複数存在する。

そこで、ユーザが任意の近似方法で選択し、パラメータを近似し、精度の測定できるシステムの開発を行った。

3 DNN 開発支援システムの開発

提案システムの概要を図 2 に示す。提案システムでは最初に既存のオープンソースニューラルネットワークライブラリにて学習を行う。現在、対応しているライブラリは Keras である。学習の結果、推論に必要なパラメータと学習に用いたネットワークの構成情報が出力される。それらのファイルをパラメータ近似ツールに入力し、ユーザが任意の近似方法でパラメータを近似し、近似した結果、どの程度の精度が得られたのかを容易に確認できるシステムである。そして、このシステムは組込みシステム向け深層学習フレームワークの一部として利用される。

Development of DNN development support system for embedded system

[†] Toki Matsumoto, Yukikazu Nakamoto : Graduate School of Applied Informatics University of Hyogo

[‡] Ryota Yamamoto, Shinya Honda : Graduate School of Infomatics, Nagoya University

[§] Kazutoshi Wakabayashi : NEC Corporation Inc.

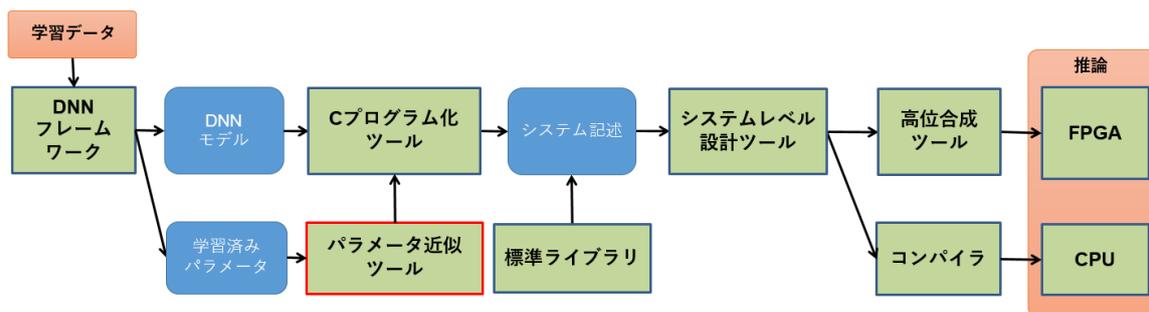


図 1: 組込みシステム向け深層学習フレームワーク

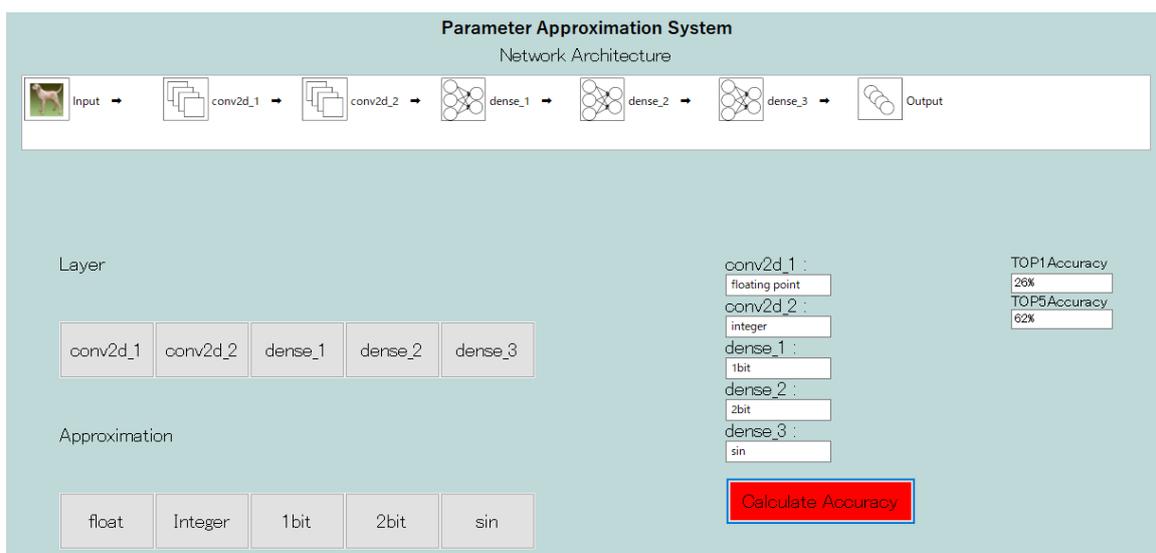


図 2: 提案システム

4 おわりに

我々が提案した既存のオープンソースニューラルネットワークライブラリによって学習した結果のパラメータを整数に近似し、推論する方法は、小規模なネットワーク、データセットにおいて有効な手法である。本来、浮動小数点数パラメータを整数に近似することで情報が欠損し、推論精度は低下する可能性が考えられる。

つまり、小規模なネットワーク、データセットにとっては有効な手法ではあるが、どのようなネットワークにおいても有効な手法であるとは限らない。

そのため、今後、提案システムを用いて様々なネットワークやデータセットで学習した結果に対して、パラメータを近似する事がどのような影響を及ぼすのかを検証する必要があると考えられる。

5 謝辞

この成果は、国立研究開発法人新エネルギー・産業技術総合開発機構 (NEDO) の委託業務の結果得られ

たものである。また、一部 JSPS 科研費 17H00762 の助成を受けている。

参考文献

- [1] 高田広章: 組込みシステム開発技術の現状と展望, 情報処理学会論文誌, Vol. 42, No. 4, pp. 930-938 (2001).
- [2] 山本 椋太, 岡本 卓也, 本田 晋也, ほか: FPGA を用いた組込みシステム向け深層学習フレームワークの構想, 情報処理学会組込みシステムワークショップ (2017)
- [3] Qian Zhao, Toki Matsumoto, Yukikazu Nakamoto, Shinya Honda, Kazutoshi Wakabayashi "A Compact and Efficient Inference Technique for Deep Neural Networks on FPGAs"