C-002

FPGA 実装に向けた部分 2 値化 CNN の蒸留に基づく精度向上
On Accuracy Improvement Based on Distillation of Partially-Binarized CNNs for FPGA Implementation

笠島　華乃† 黒木　修隆† 沼　昌宏†
Kano Kasashima Nobutaka Kuroki Masahiro Numa

1. はじめに

近年，ニューラルネットワークを用いることにより，高精度な画像認識や音声認識が実現されている。なかでも，画像認識分野で有効な畳み込みニューラルネットワーク（CNN：Convolutional Neural Networks）に注目が集まっている。これまで CNN を実装するため一般的に用いられてきた GPU（Graphics Processing Unit）は，エッジ・コンピューティングに適用するためには消費電力が大きいという問題がある [1]。そこで，CNN を書き換え可能な FPGA（Field Programmable Gate Array）上で実装することで，低消費電力化を実現する手法が注目されている。FPGA で実現可能な回路規模には制限があるため，実現する機能を推論に限定するとともに，FPGA 実装に適した重みや入力値の 2 値化に関する研究が行われている [2]。その一方で，2 値化の適用によって認識精度が低下する問題点がある。

そこで本稿では，画像認識 CNN の FPGA 実装に向けた 2 値化 CNN による精度低下を補填することを目的として，蒸留を用いた精度向上手法を提案する。

2. 提案手法

2.1 2 値化 CNN によるリソース削減と認識精度

一般に，画像認識 CNN の重みや入力には 32 bit 浮動小数点数が用いられる。重みと入力を 1 と –1 とで表現する 2 値化 CNN [2] では，畳み込み演算における乗算を XNOR で，加算を bit-counting 演算で実現できる。bit- counting 演算とは，入力に含まれる“1”の数をカウントする演算である。これによって，32 bit 浮動小数点数の非バイナリ CNN と比較すると，58 倍の演算高速化が行える [2]。さらに，重みの保持に関して 32 倍の省メモリ効果が見込まれるほか，XNOR による乗算と入力値の 2 値化の効果により，全体として 32 倍を超えるメモリ削減効果と乗算器の削減効果が見込まれる。

ここで，FPGA による実装の対象とする 2 値化 CNN の実用性について評価する。表 1 に，一般的な CNN である AlexNet と，2 値化 CNN である XNOR-Net の演算精度を評価した結果を示す [2]。この結果は，2 値化によって精度が 11～12.4 pt 低下することを示しているが，この大幅な精度低下が 2 値化適用における問題となっている。

2.2 蒸留による精度向上手法

2 値化による精度低下を抑制するため，精度向上手法が必要である。蒸留 [3] は，学習方法によって精度を向上させる手法であり，2 値化 CNN のリソース削減効果を維持しつつ，精度の向上が期待できる点を特徴とする。蒸留によって，

表 1 精度評価

手法	Top-1 Accuracy	Top-5 Accuracy
AlexNet（浮動小数点数）	56.6%	80.2%
XNOR-Net（2 値化 CNN）	44.2%	69.2%

大規模モデルの知識を小規模モデルに継承させることで，小規模モデルの精度を向上させる。大規模モデルを教師モデル，小規模モデルを生徒モデルとしたとき，教師モデルの出力を正解ラベルとして与え，生徒モデルの学習を行う。これによって，ラベル間の類似度に関する情報を反映した学習が行える。蒸留の概要を図 1 に示す。まず，入力画像と正解ラベルをもとに，初めに教師モデルの学習を行い，次に生徒モデルの学習を行う。生徒モデルの学習には，画像に対する正解ラベル：hard target と，教師モデルの推論結果：soft target の 2 つを用いる。

2.3 アンサンブル学習と蒸留

アンサンブル学習 [4] とは，別々に学習させた学習器の結果を統合して考えることで，予測能力を向上させる手法である。図 2 に示すバギングは，アンサンブル学習の分野において主流のアルゴリズムであり，並列に学習器を処理させる

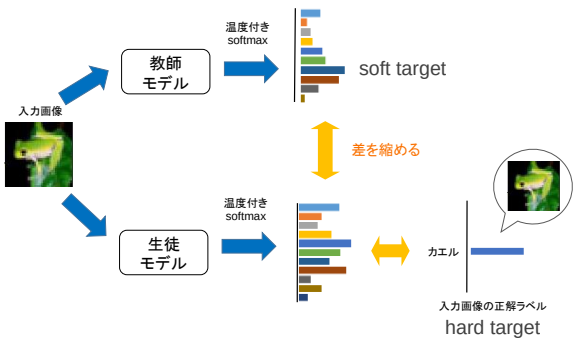

図 1 蒸留の概要

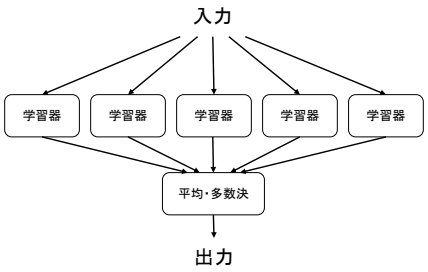

図 2 バギング

† 神戸大学，Kobe University

ことができる。図 2 に示されるように，訓練用データの部分集合を作成してそれぞれに学習させることで，複数の学習器を作成する [4]。推論の際には，学習器それぞれの出力の多数決，または平均によって最終的な出力を決定する。

蒸留では，高い認識精度をもつ教師モデルの出力を生徒モデルに反映させることによって，正解クラス以外の値も反映させ，精度を向上させる。そのため，教師モデルの高精度化が精度向上に影響する。

そこで提案手法として，蒸留の教師モデルにアンサンブル学習を適用する。本稿では，蒸留によく用いられる平均を用いてアンサンブル学習を行う [3]。教師モデル 5 個，10 個に対して，アンサンブル学習を行い，その出力を蒸留の soft target とする。アンサンブル学習は複数のモデルを用いて精度を向上させる手法であり，FPGA で実装しようとする場合は必要なリソースが増加する。しかし，学習段階での蒸留と組み合わせることにより，FPGA 上に実装する CNN 回路のリソースを増やすことなく，精度向上効果が見込まれる。

3. 実験と評価

3.1 認識精度の評価

提案手法の認識精度に関する評価を行うため，16 層 CNN を教師モデル，5 層部分 2 値化 CNN を生徒モデルとして用いた。表 2 に示すように，生徒モデルについては 2 値化を行わないモデル A のほか，畳み込み層と全結合層までの各層について上から順に 2 値化の範囲を広げたモデル B～E を用意して，各モデルに対して認識精度評価を行った。CIFAR-10 [5] を用いて，ソフトウェア上で以下の各手法による学習と推論を行い，認識精度の評価を行った。

i) 従来手法　：部分 2 値化 CNN
ii) 提案手法 1：蒸留を用いた部分 2 値化 CNN
iii) 提案手法 2：教師モデル 5 個をアンサンブル学習，蒸留を用いた部分 2 値化 CNN
iv) 提案手法 3：教師モデル 10 個をアンサンブル学習，蒸留を用いた部分 2 値化 CNN

評価対象である 5 層部分 2 値化 CNN については，蒸留を用いて学習を行った。さらに，提案手法 2，3 では教師モデルのアンサンブル学習を行った。

認識精度に関する評価結果を表 3 に示す。生徒モデル A~D において，提案手法が高い認識精度を示しており，なかでも提案手法 2，3 が最も高い精度を示した。この結果から，蒸留によって 2 値化 CNN の導入による精度低下が抑制でき，アンサンブル学習と組み合わせることで，さらにその効果が高まることが確認できた。一方で，モデル E については 0.2～0.5 pt の範囲で認識精度が低下した。これは，全結合層にまで 2 値化の範囲を広げた結果，表現できる値の範囲が狭まり，蒸留の学習結果をうまく反映できなかったためと考えられる。

3.2 ハードウェア・リソースの評価

画像認識 CNN の FPGA 実装に向け，2 値化を行った際のリソース削減効果を評価する。本稿では，表 2 で示した各層のうち，最も多くのリソースを必要とする畳み込み層 3 に関する必要リソース数を比較する。すなわち，モデル A，B，C の畳み込み層 3 と，モデル D，E の 2 値化畳み込み層 3 との間で比較評価する。

表 4 に FPGA へのマッピング結果を示す。2 値化の導入によって，畳み込み回路の LUT を 95% 以上，FF を 99% 以上，BRAM を 87% 以上削減する効果が確認できた。

表 2　生徒モデルの部分 2 値化 CNN

層構造	生徒モデル				
	A	B	C	D	E
畳み込み層 1	通常	2 値化	2 値化	2 値化	2 値化
畳み込み層 2	通常	通常	2 値化	2 値化	2 値化
畳み込み層 3	通常	通常	通常	2 値化	2 値化
全結合層 1	通常	通常	通常	通常	2 値化
全結合層 2	通常	通常	通常	通常	通常

表 3　認識精度結果（%）

手法	生徒モデル				
	A	B	C	D	E
従来手法	79.3	77.4	74.0	69.8	65.4
提案手法 1	82.4	79.0	75.8	70.7	65.2
提案手法 2	82.8	80.2	76.1	70.9	65.2
提案手法 3	83.0	80.4	76.0	71.8	64.9

表 4　マッピング結果

リソース種別	利用数（率）			
	LUT	FF	BRAM	MUX
畳み込み回路	42,653 (14.1%)	2,375 (0.4%)	512 (50.0%)	372 (0.2%)
2 値化畳み込み回路	1,953 (0.6%)	16 (< 0.1%)	64 (6.2%)	0 (0%)

4. まとめ

本稿では，画像認識 CNN の実装に必要な FPGA リソース削減のために導入される 2 値化によって生じる精度低下を補償することを目的として，アンサンブル学習に基づく蒸留を適用することで，2 値化によるリソース削減効果を維持しつつ，認識精度を向上させる手法を提案した。ソフトウェア上でのシミュレーションの結果，蒸留のみを用いた提案手法によって，認識精度が 0.9 pt ～ 3.1 pt 向上する効果を確認した。さらに，教師モデルにアンサンブル学習を適用し，認識精度を比較した結果，認識精度が 2 pt ～ 3.7 pt 向上する効果が確認され，アンサンブル学習との組合せで，さらなる精度低下抑制効果が期待できることを確認した。

また，FPGA へのマッピングの結果，2 値化によって畳み込み回路の LUT を 95% 以上，FF を 99% 以上，BRAM を 87% 以上削減可能となった。よって，提案手法に基づいて 2 値化 CNN を FPGA 上に実装することで，リソースを削減しつつ，精度低下を抑制できることを確認した。

参考文献

[1] H. Terada, H. Shouno, “B-DCGAN: Evaluation of binarized DCGAN for FPGA”, CVPR, 2018.
[2] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, “XNOR-Net: ImageNet classification using binary convolutional neural networks,” CVPR, 2016.
[3] G. Hinton, O. Vinyals, J. Dean, “Distilling the knowledge in a Neural Network,” Machine Learning, 2015.
[4] 上田修功，“アンサンブル学習”，情報処理学会論文誌，vol. 46，no. SIG15，pp. 11-20，2005 年 10 月．
[5] A. Krizhevsky, “The CIFAR-10 dataset”, http://www.cs.toronto.edu/~kriz/cifar.html