畳込みニューラルネットワーク向け重み量子化に関する研究

Training Convolutional Neural Networks with Weights Quantization

1205065 氏原 収悟 （集積システム研究室）
（指導教員 密山 幸男 准教授）

1. はじめに

画像識別に用いられる深層学習の畳込みニューラルネットワーク（CNN）は，膨大な演算量という問題点がある．そのためソフトウェアによる実装では処理速度や消費電力の面で限界がある．そこで高速化や低消費電力化を目指したハードウェア実装に関する研究が盛んに行われている．ハードウェア化において，回路規模やメモリ使用量を削減するために深層学習に用いられる重み係数などのビット数を削減することが有効であり，様々な量子化手法が報告されている[1]-[4]．そこで本研究では，フレームワークの中でも自由度が高いといわれている Chainer を用いて，量子化手法を評価できる環境を構築する．さらに，重み係数量子化の 1 手法を提案し，既存の量子化手法による認識精度と比較評価を行う．

2. 量子化の既存手法

重み係数の量子化は，まず単精度浮動小数点による学習処理によって得られた重み係数について，量子化を施すことである．量子化を施す手順はいくつかの方法がある，学習によって得られた全ての重み係数を一斉に量子化する単純量子化や，部分的段階的に重み係数の量子化を進めるインクリメンタル量子化などがある．

二値化モデルでは，BNN[1]が単純量子化を採用している．BNN は，重み係数と入力値を二値化して識別を行う．BNN では活性化関数などに Sign 関数を用いて二値化を行う．三値化モデルとしては，TWN[2][3]が単純量子化を採用している．TWN は，重み係数を三値化して識別を行うものであり，各値の割合を決めるスケーリング係数と閾値を用いて三値化を行う．TWN[2]（以後 TWN1）ではスケーリング係数は正負で同じ値を用いる．TWN[3]（以後 TWN2）では正負で異なる値を用いる.インクリメンタル量子化は INQ[4]に用いられている．INQ では特別な演算式によって，量子化を行う．

3. 提案手法

提案手法では，出力層に近い層の重み係数が入力層に近い重み係数より重要であるという考えに基づいたものである．入力層を量子化しても出力層に近い層で学習による調整が利くように，入力のほうから順に量子化を行っていく．

4. 評価環境

BNN，TWN1，TWN2，INQ を提案手法の比較対象とするため，それぞれで用いられている VGG-8，VGG-9，ReNet-18 の三種類のネットワーク構成を用いた．画像データセットとしては，Cifar-10 と，ImageNet を用いる．

評価実験において，ネットワーク構成を VGG-8，VGG-9 とするときのパラメータ構成は表 1 にした．ネットワーク構成を ResNet-18 としたとき，パラメータ構成は表 2 にした．

既存の量子化手法による認識精度は，文献で報告されている値を採用する．構築した評価環境を用いて求めた量子化前の認識精度は，各文献に記載されている値と異なるため，評価指標として量子化前の認識精度と量子化後の認識精度の差を用いる．

表 1 VGG-8, 9 のパラメータ

パラメータ	学習時	量子化時
epoch数	160	20
初期学習率	0.1	0.5
学習率減衰のepoch数	80及び120	5毎
学習率の減衰率	10%	50%
バッチサイズ	100	100
重み減衰	0.0001	0.0001
モーメンタム	0.9	0.9

表 2 ResNet-18 のパラメータ

パラメータ	学習時	量子化時
epoch数	35	2
初期学習率	0.1	0.25
学習率減衰のepoch数	30	1毎
学習率の減衰率	10%	25%
バッチサイズ	64	64
重み減衰	0.0001	0.0001
モーメンタム	0.9	0.9

5. 評価結果

VGG-9 のネットワーク構成を用いて BNN と提案手法の量子化前との認識精度の差を表 3 に示す．BNN に比べて提案手法の結果が悪い．提案手法では ReLU 関数を用いたため，負の値が消され，取りうる値が少なくなってしまった可能性が考えられる．一方、VGG-8 のネットワーク構成を用いて TWN1 と提案手法の量子化前との認識精度の差を表 4 に示す．TWN1 と比較して提案手法の結果が悪い．

表 3 BNN との認識精度比較（VGG-9，Cifar-10）

手法	重み係数	量子化前との認識精度の差
BNN	1bit	0.08%
提案手法	5bit	-0.93%
	4bit	-3.08%

表 4 TWN1 との認識精度比較（VGG-8，Cifar-10）

手法	重み係数	量子化前との認識精度の差
TWN1	2bit	-0.32%
提案手法	5bit	-2.03%
	4bit	-3.63%

ResNet-18 のネットワーク構成を用いて TWN1，2 と提案手法の量子化前との認識精度の差を表 5 に示す．TWN2 と比較して提案手法の結果がかなり悪い．提案手法では，小さい値の場合は単純な切捨てになってしまうため，ある一定以下の値が 0 になるケースが多かった可能性がある．

表 5 TWN1，2 との認識精度の比較（ResNet-18, ImageNet）

手法	重み係数	量子化前との認識精度の差
TWN1	2bit	-3.60%
TWN2	2bit	0.30%
提案手法	5bit	-9.52%

同様に、ResNet-18 を用いて INQ と提案手法の量子化前との認識精度の差を表 6 に示す．INQ に比べて提案手法の結果がかなり悪い．提案手法では量子化は単純な切り上げのため悪かったと考えられる．

表 6 INQ との認識精度の比較（ResNet-18, ImageNet）

手法	重み係数	量子化前との認識精度の差
INQ	5bit	0.72%
	4bit	0.63%
	3bit	-0.19%
	2bit	-2.25%
提案手法	5bit	-9.52%

6. まとめ

畳込みニューラルネットワーク向け重み量子化のための評価環境を Chainer で構築し，量子化手法について既存手法との比較評価を行った．量子化後認識精度が悪化した原因として量子化を行う手順よりも量子化の演算方法や，活性化関数も含めた学習パラメータ設定の違いが考えられる．現在は Chainer 以外にも，より自由度が高く扱いやすいフレームワークもあると考えられるので，フレームワーク選択の見直しも行う必要がある．

参考文献

[1] R. Zhao, W. Song, W. Zhang, T. Xing, J. Lin, M. Srivastava, R. Gupta and Z. Zhang, “Accelerating binarized convolutional neural networks with software-programmable FPGAs,” in Proc. Int’l Symposium on Field-Programmable Gate Arrays (ISFPGA), Feb. 2017.
[2] F. Li, B. Zhang and B. Liu, “Ternary weight networks,” in Proc. Neural Information Processing Systems (NIPS), Dec. 2016.
[3] C. Zhu, S. Han, H. Mao and W. J. Dally, “Trained ternary quantization,” in Proc. Int’l Conference on Learning Representations (ICLR), Apr. 2017.
[4] A. Zhou, A. Yao, Y. Guo, L. Xu and Y. Chen, “Incremental networks quantization: towards lossless CNNs with low-precision weights,” in Proc. Int’ l Conference on Learning Representations (ICLR), Apr. 2017.