C-005

画像分類 CNN の FPGA 実装に向けた インクリメンタル量子化手法によるリソース削減

Resource Reduction by Incremental Quantization Method for FPAG Implementation of Image Classification CNN

山本 晃暉† 黒木 修隆† 沼 昌宏†
Koki Yamamoto Nobutaka Kuroki Masahiro Numa

1. はじめに

近年，畳み込みニューラルネットワーク（CNN: Convolutional Neural Network）を用いることで，高精度な画像認識が実現されている。CNN 実装のために一般に用いられる GPU（Graphics Processing Unit）は，消費電力が大きいという問題がある。そこで，FPGA（Field Programmable Gate Array）上にハードウェアとして実装することで，低消費電力化を実現する研究に注目が集まっている。しかし，FPGA で実装可能な回路規模には制限があるため，FPGA 実装を考慮した演算手法や回路構成が必要となる。そこで，一般に 32 bit 浮動小数点数で表現される CNN のパラメータを，少ないビット幅で表現し，軽量化する量子化手法に注目が集まっている。しかし，量子化の適用には，認識精度の低下を招くという欠点がある。

そこで本稿では，CNN 実装に必要な FPGA のリソースを削減しつつ，CNN の量子化によって発生する精度低下を抑えるために，量子化誤差に基づくインクリメンタル量子化手法を提案する。

2. 提案手法

2.1 量子化誤差を考慮したインクリメンタル量子化

CNN の軽量化のために量子化を適用した場合に発生する精度低下を抑えるために，量子化誤差の大きな重みを優先して段階的に量子化する，インクリメンタル量子化手法を提案する。

本手法では，量子化の適用範囲を量子化適用率 r_{q} で制限する。また，量子化適用率を量子化適用系列 $R_{\mathrm{q}} = (r_1, r_2, \cdots, r_{n-1}, 1)$ の要素として格納する。また，R_{q} の要素は順に大きくなる $(r_1 < r_2 < \cdots < r_{n-1} < 1)$ ように設定する。これには，精度に影響を与える量子化誤差が大きい値を，学習初期に優先的に処理する狙いがある。

図 1 に量子化の対象範囲について示す。ここで，i 番目の量子化後の値を q_i とし，量子化を適用する上下の閾値をそれぞれ $t_{i,+}$, $t_{i,-}$ とする。量子化 1 回目，すなわち $i = 1$ の時，R_{q} より r_1 を取り出す。このとき，量子化範囲は図 1 の緑の範囲に相当する。この範囲の値を量子化した結果，q_i として固定する。その後，固定していない重みについて学習し，その結果をもとに更新する。この処理を R_{q} の各要素について行い，すべての重みについて適用する。

2.2 焼きなまし法を応用した精度向上手法

焼きなまし法 [1] は，金属の熱処理に着想を得た最適化問題のアルゴリズムである。図 2 に示すように，近傍と比較して極小のエネルギーを示す局所解を，ある確率であえてエネルギーが増える方向に更新することで，真の解である大域解へ到達させることを目指す手法である。

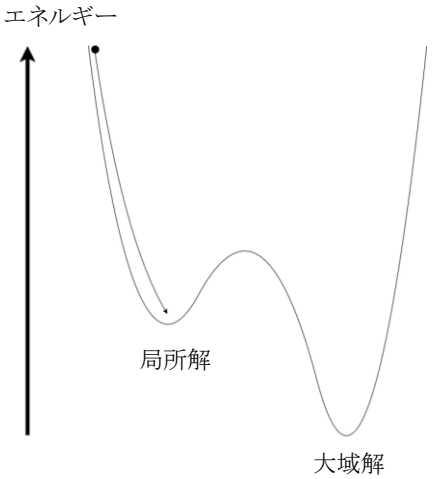

図 2 焼きなまし法のイメージ

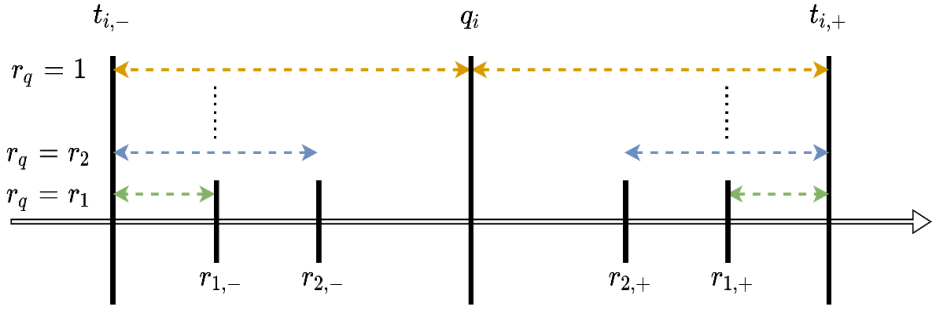

図 1 量子化対象範囲の設定

† 神戸大学 Kobe University

Copyright © 2022 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

表 1　実験結果(VGG16)

手法	ビット幅 [bit]	認識精度 [%]	パラメータ容量 [bit]	容量比 [%]
VGG16	32	49.64	1,088,222,336	100.00
従来手法 (INQ)	5	51.01	170,372,672	15.66
	4	50.85	136,378,240	12.53
	3	49.99	102,383,808	9.41
提案手法 (+焼きなまし)	5	51.95	170,372,672	15.66
	4	51.05	136,378,240	12.53
	3	51.14	102,383,808	9.41

表 2　実験結果(MobileNetV2)

手法	ビット幅 [bit]	認識精度 [%]	パラメータ容量 [bit]	容量比 [%]
MobileNetV2	32	58.44	75,263,104	100.00
従来手法 (INQ)	5	53.65	16,139,584	21.44
	4	51.60	13,949,824	18.53
	3	21.69	11,760,064	15.63
提案手法 (+焼きなまし)	5	57.37	16,139,584	21.44
	4	57.60	13,949,824	18.53
	3	46.09	11,760,064	15.63

表 3　マッピング結果

対象回路	リソース利用数(利用率 [%])			
	LUT	FF	DSP	BRAM
VGG16 (32 bit)	664,400 (218.8%)	206,592 (97.7%)	6,096 (217.7%)	768 (74.6%)
VGG16 (5 bit)	204,096 (67.2%)	8,584 (4.1%)	1,664 (59.4%)	96 (9.3%)

ここで, 2.1 節の量子化の過程で固定した重みの一部を局所解とみなし, 再度学習することで, CNN の精度向上を図る。

3. 実験と評価

3.1 演算精度の評価

提案手法の認識精度に対する評価を行う。100 クラス分類のデータセットである CIFAR-100 を用いて, 事前学習済みの VGG16 [2] と MobileNetV2 [3] を対象として, ソフトウェア上で従来手法と提案手法で量子化し, 推論を行った。実験環境を以下に示す。

i) 従来手法: INQ [4] にて再学習
(5, 4, 3 bit, 学習回数 2 回)

ii) 提案手法: 提案手法にて再学習
(5, 4, 3 bit, 学習回数 2 回, $R_q = (0.2, 0.5, 0.7. 1)$, 焼きなまし法あり)

表 1 に VGG16 に関する実験結果を, 表 2 に MobileNetV2 に関する実験結果をそれぞれ示す。

VGG16 では, 従来手法より 2.31 pt 精度が向上し, パラメータ容量が最大 90% 削減されている。

MobileNetV2 では, 従来手法よりも精度低下幅を抑えることに成功し, 4 bit に圧縮した場合, 手法適用前からパラメータ容量を約 81% 削減し, 精度低下を 0.84 pt に抑えた。

3.2 ハードウェア・アーキテクチャの評価

提案手法を用いたパラメータ容量圧縮の効果を評価するため, VGG16 の畳み込み層の 1 層目に対して提案手法を適用することを想定して, リソース利用率の比較を行った。

表 3 にマッピング結果を示す。提案手法を適用することで, 畳み込み演算回路のリソース数に関して, LUT を 69.3%, FF を 95.8%, DSP を 72.7%, パラメータを保存するリソースである BRAM を 87.5%, それぞれ削減できることを確認した。

4. まとめ

本稿では, 量子化による精度低下を抑えることを目的として, 量子化誤差を考慮したインクリメンタル量子化手法を提案した。

ソフトウェア上でのシミュレーションの結果, VGG16 では, 提案手法によって認識精度が最大 2.31 pt 向上され, パラメータ容量は最大 90% 削減された。MobileNetV2 では, パラメータ容量を 80% 程度削減しても, 認識精度の低下を 1 pt 以内に抑えられた。

また, FPGA へのマッピングの結果, LUT を 69.3%, FF を 95.8%, DSP は 72.7%, BRAM を 87.5% 削減する効果が確認できた。

よって, 提案手法を用いた量子化によって, CNN の精度低下を抑制しつつ, パラメータ容量を圧縮し, リソース利用数を削減できることを確認した。

今後の課題としては, MobileNetV2 における精度低下が挙げられる。圧縮時の状態を調べることで, より効率よく圧縮できると考えている。また, 焼きなましの適用方法についても試行回数や探索方法について改良することで, 精度面での改善が期待される。

参考文献

[1] S. Kirkpatrick, C. D. Gelatt JR and M. P. Vecchi, "Optimization by simulated annealing," Science. 220 (4598): 671–680., 1983.

[2] K. Simonyan, "Very deep convolutional networks for large-scale image recognition," https://arxiv.org/abs/1409.1566, 2015.

[3] S. Mark, H. Andrew, Z. Menglong, Z. Andrey, and C. Liang-Chieh, “MobileNetV2: Inverted residuals and linear bottlenecks,” 2018.

[4] A. Zhou, "Incremental network quantization: towards lossless CNNs with low-precision weights," https://arxiv.org/abs/1702.03044, 2017.

Copyright © 2022 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.