TSUBAME 共同利用　令和6年度　産業利用　成果報告書

利用課題名　自動運転 AI の研究開発
英文：R&D for automotive AI

利用課題責任者　関川雄介

所属　デンソーアイティーラボラトリ
URL https://www.d-itlab.co.jp/

邦文抄録（300 字程度）
車載 AI プロセッサは電力・演算性能が限られることから軽量な DNN モデルが求められる。一方で，近年の高精度な DNN モデルはパラメータ数が増大し演算量が増加する傾向にある。そこで車載 AI に適用できうる DNN 圧縮技術を確立する必要がある。本研究課題では特にニューラルネットワークの極低ビット（2bit~4bit）への量子化技術に着目し、その有効性に関する検討を進めた。これまでにも様々な量子化手法が提案されているが、それらの有効性は共通の実験条件下で十分に比較・評価は行われてこなかった。そこで画像分類問題において、有望な既存の量子化手法を公平な実験条件下でベンチマーク評価を行うとともに、精度劣化を抑えるために重要な要素を明らかにした。

英文抄録（100 words 程度）
Lightweight DNN models are essential for automotive AI processors due to constraints in power and computational resources. However, recent high-accuracy DNNs tend to be larger and more computationally intensive. This highlights the need for effective DNN compression techniques suitable for automotive AI applications.

This study focuses on extremely low bit quantization (2 to 4bits) that reduces memory size and computational complexity. Although various quantization methods have been proposed, their performance has not been thoroughly compared under the identical experimental conditions. To address this, we conducted a benchmark evaluation of promising existing quantization techniques for image classification under fair settings, while also clarifying the essential components to suppress accuracy degradation.

Keywords: 5つ程度
車載 A I、DNN、モデル圧縮、量子化、低ビット

背景と目的

車載 AI プロセッサなどのエッジ端末では電力・演算性能が限られることから軽量な DNN モデルが求められる。一方、近年の高精度な DNN モデルはパラメータ数が増大し演算量が増加する傾向にある。そこで、エッジ端末に適用できる DNN 圧縮技術を確立を目指す。本プロジェクトでは、極低ビットでの様々な量子化手法のベンチマークを通して、精度を担保するために必要な技術を炙り出し、さらに精度劣化を抑えるための検討材料を揃える。

概要

ニューラルネットワークの低ビット化は、浮動小数点の学習済み模型に対して重みと活性化関数に量子化関数を差し込むことで実現できる(図 1)。

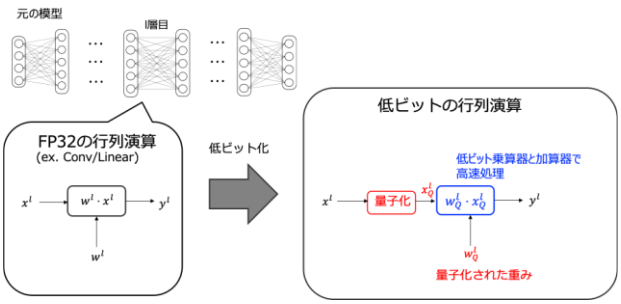

図 1:ニューラルネットワークの低ビット化

これにより、ニューラルネットワークの演算の高負荷箇所である行列演算は整数演算で実行でき、高速な処理が可能になる。

一方、極低ビット(2~4 ビット)では量子化誤差が大きくなるため、単純に量子化関数を差し込むだけでは著しく

精度劣化を起こす。

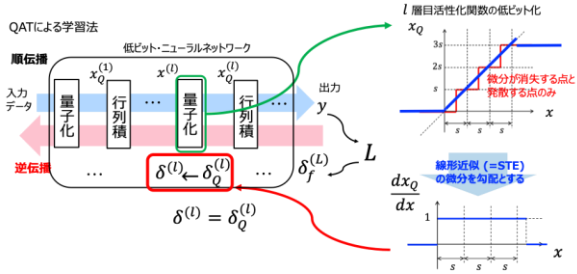

図 2:低ビットニューラルネットワークの学習方法

そこで、精度劣化を防ぐために極低ビットでは、元の全ての学習用のデータセットを用いて End-to-End で再学習する手法である Quantization-Aware-Training(QAT)が発展してきた(図 2)。QAT では順伝播時は重みと活性化関数を低ビット化したネットワークを通して出力するのに対して、逆伝播時はその低ビット化の離散的な振る舞いのために有限値を持つ勾配を通すことができないという課題がある。そこで逆伝播で勾配を通すために、量子化関数の clip の内側はそのまま微分を通す Straight-through-estimator(STE)と呼ばれる近似的な手法がしばしば用いられてきた。

STE の DNN への応用は 2012 年に Hinton による講義[1]でその原型となるアイデアが提唱されて以来、幅広く発展してきた。とりわけ成功を収めている手法としては、重みだけではなく、量子化関数の間隔幅(ステップ幅)自体も STE を用いて学習することでより高い表現力を獲得するアプローチが挙げられる(PACT[2],LSQ[3])。さらに、それらの手法の不等間隔幅への拡張したアプローチ(LCQ[4],nuLSQ[5])などがあげられる。ステップ幅を学習する手法の中でも STE

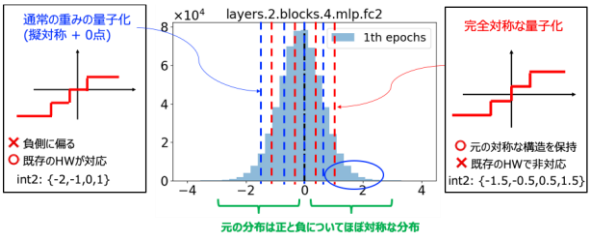

図 3:重みの分布の構造と量子化関数の違い

を関数全体に使う(PACT)か、あるいは部分的に使う(LSQ)かによって勾配の近似方法が異なり、学習後の精度に影響を及ぼす。加えて、重みの量子化手法についても大きく分けて3種類の方式が存在する：0 値をレベル数に含めつつ、負側の量子化レベルを正側より一つ多くとるほぼ対称な方式(LSQ[3])で実用的に用いられている方式，元の浮動小数点の分布の対称性を尊重して、0 値をレベル数に持たない完全に対称な量子化(UniQ[6],N2UQ[7],LQ-Nets[8])の方式や 0 値をレベル数にもったまま完全に対称な構造を保つために標準的な量子化から負側のレベル数を一つ削る量子化ノ方式(LSQ-SignMag/APoT[9]/LCQ[4])がある。それらに加えて、各手法ごとに学習を安定化させる独自の勾配を調整する技術や量子化関数の初期化法も併せて開発されてきており、精度劣化を抑えるためにどの要素が支配しているのが判別できない状況となっている。

そこで本研究では精度に影響を及ぼす可能性が高い構成要素(1. 重みの学習率と減衰の設定 2. ステップ幅の勾配調整法 3. ステップ幅の初期化法 4. ステップ幅に対する学習率と減衰の調整)ごとに精度を評価した上で、高精度を実現する構成要素の組み合わせのもとで最後に代表的な量子化手法のベンチマーク評価を画像分類問題で行った。データセットしては、Imagenet を用いて、模型としては Swin-T で評価を行なった。また、オプティマイザーについては学習済み模型で用いた AdamW を採用した。各構成要素の精度評価は Imagenet の訓練データの 9 割を 15epoch 学習して残りの 1 割をヴァリデーション・データとして行った。そして最終的なベンチマーク評価では全ての訓練データを用いて 90epoch を学習して公式の validation 用のデータセットを用いてテスト精度評価を行った。代表的な量子化手法として等間隔量子化手法である LSQ, UniQ,LSQ-SignMag, N2UQ と不等間隔量子化手法である APoT, LQ-Nets, LCQ, nuLSQ を評価した。

結果および考察

ここでは 2~4bit の量子化のうち、各手法の差異が最も顕著になる 2bit の結果について報告する。まず、各構成要素の精度評価について報告する。

1. 重みの学習率と減衰の設定

各手法による評価の偏りを減らして評価の公平性を保つために、最終評価で用いる量子化手法を用いずに、各層の入力・重みごとに MSE でステップ幅を決める単純な量子化手法(MSEQ)を用いてグリッド探索で重み

の学習率と減衰を決定した(表 1)。

表 1:MSEQ での重みの学習率(lr)と減衰(wd)のグリッド探索

lr \ wd	0.00625	0.0125	0.025	0.05
0.001	69.241	69.269	69.091	68.881
0.0001	72.353	72.224	**72.498**	72.321
0.00001	64.726	64.813	64.744	64.721

2. ステップ幅の勾配調整法

ステップ幅は各層ごとに共通のものを用いるため、その層の出力数と大きく乖離がある。そのため、連鎖律に従って逆伝播で STE 勾配を計算するとステップ幅の勾配量は大きくなり、また層ごとに不均一な大きさになる[3]。それを解消するためにこれまで提案されてきた技術としては、層のパラメータとビット数に依存する経験的に決定した定数でスケールさせる勾配スケール法(GS)や重みの分布を正規化したのち量子化する方法がある。さらに後者にも、重みの分布の標準偏差で正規化する方法(LWN[4])やエントロピーが大きくなるように正規化する方法(EPWR[7])がある。また今回は AdamW を optimizer として用いるため、そもそも勾配の大きさがリスケールされて、自然と勾配の大きさが調整される側面がある。そのため、上記の勾配調整法の評価に加えて、勾配調整法を用いない場合の評価も行なった。その結果、精度の差異は小さいものの、表 2 に示すように標準偏差で分布を正規化する LWN が最も良い精度を示した。

表 2: LSQ でのステップ幅の勾配調整法の比較

勾配調整法	GS+AdamW	LWN+AdamW	EPWR+AdamW	AdamW
val精度	73.18	**73.38**	73.20	73.24

3. ステップ幅の初期化法

ステップ幅の初期化も各手法ごとに様々なものが用いられてきた。定数値で初期化する手法や、分散を用いて初期化する方法(LSQ 初期化)、量子化誤差の平均二乗誤差を最小にするように初期化する方法(MSE 初期化)もある。また、MSE の初期化をさらに簡単化した NMSE 初期化も提案されている[6]。これらのステップ幅の初期化法について LSQ で評価比較を行なった。表 3 に示すように、定数初期化は他の初期化法に比べて著しい精度劣化を示し、MSE 初期化が最も精度が良い結果となった。

表 3: LSQ でのステップ幅の初期化法の比較

初期化法	定数初期化	LSQ初期化	NMSE初期化	MSE初期化
val精度	41.47	72.57	72.75	**73.38**

4.各量子化手法の量子化パラメータに対する学習率と減衰の調整

1~3 で選ばれた精度の最も良い組み合わせを用いて、各量子化パラメータについての学習率と減衰をグリッド探索した。量子化パラメータとは例えば LSQ や UniQ ではステップ幅、PACT では閾値、nuLSQ では複数のレベル数といった手法ごとに存在する量子化関数を形成するための学習パラメータを意味する。

1~4 で選ばれた量子化関数の勾配調整法(LWN)と初期化法(MSE)、ならびに重みと量子化パラメータの学習率と減衰を用いて、各量子化手法の精度評価を行なった(表4)。その結果、等間隔量子化手法の中では LSQ

表 4: 量子化手法のベンチマーク評価

	quantizer	SwinT-tiny[蒸留] 80.916/95.422	SwinT-small[蒸留なし] 83.050/96.864
等間隔	1. LSQ	76.172/92.908	77.874/94.180
	2. UniQ	**76.378**/92.984	**78.154**/94.168
	3.LSQ-SignMag	75.468/92.444	77.376/93.836
	4. PACT	75.318/92.284	76.836/93.682
	5. N2UQ	74.728/92.256	77.812/94.076
不等間隔	6. APoT	74.832/92.171	76.856/93.556
	7. LQ-Nets	75.282/92.508	77.590/93.942
	8. LCQ	75.899/92.805	77.906/94.034
	9. nuLSQ	**76.698**/93.142	**78.310**/94.288

よ UniQ が高精度を示した。特筆すべき点として、比較手法の中で最も新しい手法である N2UQ は、比較対象の中で最も高精度と原論文中で報告されていたが、本研究による統一的な実験環境下の再実験により、むしろ従来手法である LSQ や UniQ の方が高い精度を達成していたことが明らかになった。さらに LSQ と UniQ を比較すると、僅差ながら UniQ が高精度な結果を示している。これは先述の通り、学習済みの重みの分布の構造を反映した完全に対称な量子化構造の方が、量子化誤差を下げることができ、最終的な精度も良い結果となることを意味する(図 3)。一方、対称性を保ちつつ 0 値を量子化レベルを含める LSQ-SignMag は精度劣化を引き起こしている。これは負側のレベル数を一つ削減することによって、表現力が下がり、最終的な精度に悪影響を及ぼした結果と考えられる。

さらに不等間隔量子化手法でも同様に各手法の比較を

行った。その結果、LSQ をベースとして不等間隔に拡張した nuLSQ が、他の不等間隔手法と比較して最も高い精度を達成した。

まとめ、今後の課題

本研究では、既存の量子化手法を共通の実験条件したでベンチマーク評価を実施した。その結果、量子化関数を STE で学習する研究の中で初期の手法である LSQ とそれの重みの完全対称版の UniQ、そして不等間隔版の nuLSQ が高精度な結果が得られた。

今後は、本研究成果を受けてさらに精度劣化を抑える手法の検討をする。現時点では以下の3つの方向性が有望であると考えている。I. nuLSQ の完全対称版への拡張を行う、II. 層ごとに LSQ と UniQ を適応的に選択する、III.層ごとに用いている共通のステップ幅をチャンネルごと/トークンごと/ヘッドごとなど、高速な整数演算の処理の要請を満たす範囲で、量子化粒度を変える。

参考文献

[1] Hinton, G. (2012). Neural networks for machine learning. Coursera, video lectures

[2] Choi, J., Wang, Z., Venkataramani, S., Chuang, P.I.J., Srinivasan, V., Gopalakrishnan, K.: Pact: Parameterized clipping activation for quantized neural networks. International Conference on Learning Representation (2018)

[3] Esser, S.K., McKinstry, J.L., Bablani, D., Appuswamy, R., Modha, D.S.: Learned step size quantization. In: International Conference on Learning Representations (2019)

[4] Yamamoto, K.: Learnable companding quantization for accurate low-bit neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2021)

[5] Gongyo, S., Liang, J., Ambai, M., Kawakami, R., & Sato, I. (2024). Learning Non-uniform Step Sizes for Neural Network Quantization. In Proceedings of the Asian Conference on Computer Vision (2024)

[6] Pham, P., Abraham, J.A., Chung, J.: Training multi-bit quantized and binarized networks with a learnable symmetric quantizer. IEEE Access 9 (2021)

[7] Liu, Z., Cheng, K.T., Huang, D., Xing, E.P., Shen, Z.: Nonuniform-to-uniform quantization: Towards accurate quantization via generalized straight-through estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2022)

[8] Zhang, D., Yang, J., Ye, D., Hua, G.: Lq-nets: Learned quantization for highly accurate and compact deep neural networks. In: Proceedings of the European conference on computer vision (2018)

[9] Li, Y., Dong, X., Wang, W.: Additive powers-of-two quantization: An efficient nonuniform discretization for neural networks. In: International Conference on Learning Representations (2020)