量子化による認識精度低下を抑制する深層学習向け重み正規化手法

1210111 土居 拓矢 (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

深層学習は積和演算の量が膨大であるため、演算速度の向上と省電力化のためにハードウェア化することが有効である。回路規模削減と省メモリ化のためにパラメータを量子化する研究が行われているが、認識精度の低下が問題となっている[1]。先行研究[2]では学習中のパラメータにビットマスクを施すことでこの問題の解決を試みたが、学習が効果的に行えないケースがあった[2]。そこで本研究では、学習中ではなく学習後のパラメータにビットマスクを行うとともに、畳み込み層ごとにビットマスクのパターンを適用することで、量子化による認識精度低下を抑制する手法を考案する。量子化のみを行った場合との認識精度を比較、評価する。

2. 提案手法

提案手法の概要を図 1 に示す。提案手法は学習後の畳み込み層の重み係数に対して量子化を行う。まず、想定する量子化ビット数を決め、それより少し大きいビット数に量子化し、その差分をビットマスクする。これにより、ハードウェア実装時には単純量子化時のビット数と同じ回路規模であるとみなせる。しかし実際には、単純量子化時のビット幅よりビット幅が大きくなっており、認識精度低下の抑制が期待される。丸め方法は、最近接丸めを採用している。

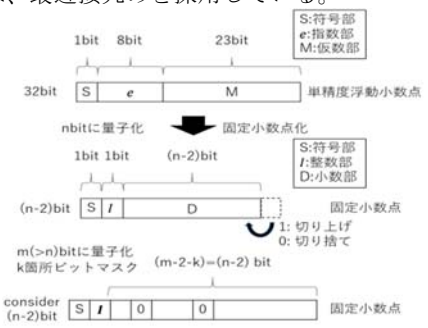

図 1　提案手法の概要

3. 先行研究からの課題とその対策

先行研究では 2 つの課題が残されていた。1 つは学習から推論までに膨大な時間がかかることである。先行研究ではビットマスクの位置を変えるたびに学習をやり直す必要があり、全てを試すとなると膨大な時間がかかる。そこで本研究では学習済みモデルに対して量子化とビットマスクを行うことで、一番時間のかかる学習は最初の一度のみとした。

もう 1 つは層ごとの重み係数の精度を考慮せず、すべて同じパターンでビットマスクしていることである。これでは、1 層目の重み係数の範囲をカバーできたとしても、2 層目の重み係数の範囲をカバーできるとは限らない。そこで本研究では層ごとに適切なビットマスクのパターンを割り当てるようにした。これにより、層ごとの重み係数を広くカバーできると考えた。

4. 評価環境

フレームワークには caffe を用いた。ネットワークには LeNet_caffe(LeNet)、Cifar10_quick、Cifar100_normal を用いた。画像データセットは LeNet には Mnist を、Cifar10_quick には Cifar10 を、Cifar100_normal には Cifar100 を用いて、提案手法の有効性を評価した。

5. 評価結果

LeNet_caffe に量子化のみを行った場合とすべての畳み込み層に同じビットマスクパターンを適用した場合の結果を表 2、表 3 に示す。3bit も 4bit も量子化のみでは精度が得られないが、ビットマスクをして見かけのビット幅を広げることで高い精度が得られていることがわかる。また 3bit 精度で認識精度が最も高いパターンは X0.0XX で、逆に最も低いパターンは X0.X0000X であった。この 2 つのパターンによる違いを調べるため、3bit 精度においてそれぞれが取り得る値を重み係数のヒストグラムに合わせた結果を図 2 と図 3 に示す。図 2 と図 3 にそれぞれ 1 層目と 2 層目の畳み込み層の重み係数の値の分布を示す。ヒストグラムの下にある線はヒストグラムの横軸と同じであり、点は左にあるビットマスクパターンの取り得る値を示している。図 2 では、認識精度の高いパターンは重み係数の値の範囲を広くカバーできており、取り得る値も均等に分布していることがわかる。それに比べ、認識精度の低いパターンは取り得る値が極端に偏っており、重み係数の範囲を全体的にカバーできていない。一方、図 3 ではどちらも取り得る値は少なく、また極端に偏っていることがわかる。

表 2. 3bit 精度

量子化ビット数とビットマスク位置		認識精度	比較
3_0_0	XX.X	0.114	0
4_1_0_0	X0.XX	0.232	0.117
5_2_0_1_0	X0.0XX	0.964	0.850
5_2_0_2_0	X0.X0X	0.952	0.838
6_3_0_1_2_0	X0.00XX	0.935	0.821
6_3_0_1_3_0	X0.0X0X	0.902	0.788
8_5_0_2_3_4_5_0	X0.X0000X	0.114	0.000

表 3. 4bit 精度

量子化ビット数とビットマスク位置		認識精度	比較
4_0_0	XX.XX	0.232	0
5_1_0_0	X0.XXX	0.972	0.740
6_2_0_1_0	X0.0XXX	0.976	0.744
6_2_0_2_0	X0.X0XX	0.955	0.723
6_2_0_3_0	X0.XX0X	0.899	0.667
7_3_0_1_2_0	X0.00XXX	0.971	0.739
7_3_0_1_3_0	X0.0X0XX	0.968	0.736

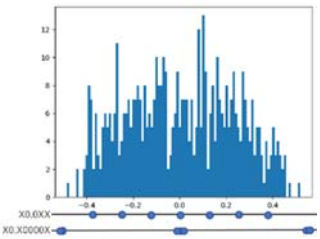

図 2. 1 層目の重み係数の分布

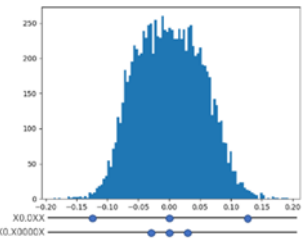

図 3. 2 層目の重み係数の分布

そこで、これらの結果を参考に、層ごとに異なるビットマスクを行った結果を表 4、表 5 に示す。いずれの場合も層ごとに異なるビットマスクとすることで認識精度が向上した。特に認識精度が高い組み合わせは 1 層目も 2 層目もビットマスクの位置が歯抜け状態ではなく、上位ビットから順番にマスクされている。上位ビットから順番にマスクをすることで取り得る値が均等になり、重み係数の範囲を広くカバーできたことで認識精度が向上したと考えられる。

表 4. 3bit 異なる場合の精度

畳み込み層1	畳み込み層2	認識精度	比較
3_0_0	3_0_0	0.114	0
5_2_0_1_0	6_3_0_1_2_0	0.962	0.848
5_2_0_1_0	7_4_0_1_2_4_0	0.847	0.733

表 5. 4bit 異なる場合の精度

畳み込み層 1	畳み込み層2	認識精度	比較
4_0_0	4_0_0	0.232	0
6_2_0_1_0	7_3_0_1_2_0	0.976	0.744
6_2_0_1_0	8_4_0_1_2_3_0	0.964	0.733
6_2_0_1_0	8_4_0_1_2_4_0	0.948	0.716

6. まとめ

提案手法によって、単純な量子化のみでは認識精度が低いビット数の場合でも、認識精度を向上させることができた。また層ごとに適切なビットマスクをすることでさらに認識精度が向上することが明らかになった。特に上位ビットからマスクをして組み合わせたパターンは、他のパターンよりも高い認識精度が得られた。

参考文献

[1] S. Zhou, Z. Ni, X. Zhou, H. Wen, Y. Wu, and Y. Zou, “DoReFa-Net: Training low bitwidth convolutional neural networks with low bitwidth gradients”, arXiv preprint arXiv: 1606.06160, Jun. 2016.
[2] 井上裕太, “重み量子化時の識別精度低下を抑制する深層学習向け重み正規化手法の検討”, 高知工科大学 , 2019 年 3 月.