H-028

プルーニング，量子化，行列分解による全結合ネットワークのパラメータ削減

西野 駿佑 †　数藤 恭子 †
Shunsuke Nishino　Kyoko Sudo

† 東邦大学

1. はじめに

近年，画像や音声，自然言語処理など，様々なメディア処理を目的とした高精度な深層学習モデルが提案されている．こうしたモデルを環境や人間のセンシングといった実課題に応用するため，スマートフォンや Raspberry Pi のようなエッジデバイスに深層学習モデルを搭載するニーズは高い．しかし，膨大な数のパラメータを持つ複雑なモデルは，容量が大きく推論時の計算量も多いため，メモリやディスク容量，電力等が制限された小型端末への搭載は困難である．学習済みモデルのアーキテクチャを近似し軽量化することで，こうした端末でも動作させることが可能になり，応用範囲が拡大すると考えられる．そこで本研究では，認識精度を維持しつつ，モデルを圧縮することを目的とする．従来から知られているパラメータ削減手法である，プルーニングと量子化に加えて，線形演算の行列分解による計算削減を組み合わせた効果的な圧縮方法を提案する．

具体的な検討対象のモデルとして，近年様々な画像認識タスクにおいて有効性が注目されており，エッジデバイスへの適用ニーズも高いと考えられる Vision Transformer[2] を用いる．CIFAR10 の学習済みモデルを元の 6 割程度のパラメータ数に削減しても同等の識別精度を保つことが可能であることを示す．

2. 関連研究

従来から提案されている深層学習のモデル圧縮の代表的な方法の 1 つに，パラメータの冗長性を削減するプルーニングがある．LeCun ら [8] は，各パラメータによる誤差関数の 2 次微分に基づき，影響が少ないパラメータを削減することでネットワークが圧縮できることを示した．Han ら [4] は，しきい値以下の重みへの結合を除き再学習することで，LeNet や AlexNet などの畳み込みネットワークにおいて，精度低下を数 % から数 10% に抑えつつ 9 割程度のパラメータ削減を行った．Sajjad ら [10] は Transformer のモデルの一種 BERT の層ごとの dropping を提案し，精度の低下を数 % に抑えつつ 4 割程度のパラメータを削減可能なことを示した．プルーニングはモデル容量の大幅な削減が可能であるが，パラメータ削減後に精度を回復するためのモデルの再学習を行うことを前提としている．

深層学習モデルのもう 1 つの代表的な圧縮方法にパラメータの量子化がある．Courbariaux ら [5, 6] は，パラメータの重みと出力を 1 または −1 に量子化しながら学習する BinaryConnect を提案し，中間層が 3 層の Multi Layer Perceptron のネットワークで，精度低下を 10% 以下に抑え，計算量を 3 分の 2 程度削減した．Han ら [3] はこれらのプルーニングと量子化を組み合わせることでより大幅な容量削減に成功している．しかし量子化は，多くの場合に再学習を必要とする [3, 5, 6].

これらの方法の他に，全結合層の重み行列に対し行列分解による低ランク近似を行う手法がある．Sainath ら [7] は，全結合層のみから成る音声認識モデルに対し，最終層以外のパラメータを行列分解することでモデル容量を削減した．畳み込みニューラルネットワークなどの多くの深層学習モデルでは，全結合層は主に出力層付近のみで用いられているため，行列分解を深層学習モデルに適用した例は多くない．しかし，低ランク近似の精度が良ければ，モデルの精度を低下させることなく重みの数と計算量を減らすことが可能である．パラメータ数の削減による推論時における計算速度の向上と消費電力の減少も期待される．一般的にプルーニングや量子化ほどの圧縮率は得られないが，学習を必要としないことが利点である．そのため，他の手法との組み合わせでの利用も容易であり，全体としてより多くの容量削減が可能になると考えられる．

そこで本研究では，主に全結合層からなる Transformer のモデルを対象とし，プルーニングと量子化と行列分解を組み合わせて圧縮する実験を行う．Sajjad ら [10] の Transformer モデルのプルーニングによる圧縮の研究において，精度の低下率はプルーニングを施す層に依存することが示されていることから，適用する層を適切に選択し，精度の低下を最小限に抑えて圧縮可能な適用方法を考察する．

3. 提案手法

3.1. 概要

本研究では，プルーニングと量子化によって大幅なモデル容量削減を可能にした Han ら [3] の手法をベースラインとし，学習済みの深層学習モデルの全結合パラメータの削減手法として，プルーニングと量子化に行列分解を組み合わせる方法を提案する．パラメータの削減を行う層を自動的に決定する手法として，出力ベクトルにおける要素の負値の個数の割合を指標とする手法を提案する．

3.2. プルーニング

プルーニングは重み行列の要素のうち，値の絶対値が小さい要素を削除することでパラメータ数を削減する手法である．要素数を N とする重み行列 W を考える．W の各要素を $w_i(i = 0, \ldots, N-1)$ とし，N 個のうち $p\%$ 削減するとき，$|w_i|$ が小さい要素から順に $0.01pN$ 個削減することを割合 p でプルーニングを行うと表現するものとする．

学習済みモデルをプルーニングすると，精度が低下してしまうため，Han らはプルーニング後にファインチューニングを行うことで精度を回復させている [4]．本研究では，Han らが提案したプルーニング手法を採用し，割合 p を決めてプルーニングを行い，その後ファインチューニングを行う．

3.3. 量子化

深層学習のパラメータは一般的に 32 ビットの浮動小数点型であり，量子化は浮動小数点型から整数型へと変換することによってモデル容量を削減する手法である．量子化は再学習を必要としない Post Training Quantization[11] を用いた．

$x \in \mathbb{R}$ を量子化の対象となるパラメータとし，そのデータ範囲を $[-\alpha, \alpha]$ とする．x を b ビットへと量子化するとき，量子化後のパラメータ $x_q \in \mathbb{Z}$ が取り得る範囲は $x_q \in [-2^{b-1}+1, 2^{b-1}-1]$ となる．

浮動小数点型から整数型へと変換する場合，あるスケールファクター $s \in \mathbb{R}$ との積をとった結果を整数型へと丸める:

$$x_q = clip(round(s \cdot x), -2^{b-1}+1, 2^{b-1}-1) \quad (1)$$

$$\text{ただし，} s = \frac{2^{b-1}-1}{\alpha} \quad (2)$$

$$clip(x, l, u) = \begin{cases} l, & x < l \\ x, & l \le x \le u \\ u, & x > u \end{cases} \quad (3)$$

スケールファクターの定め方には様々な手法があるが，本研究では動的量子化という手法を用いる．動的量子化はモデルの推論時のデータ範囲によってスケールファクターを決定する．

3.4. 行列分解

本研究では，行列分解のアルゴリズムとして特異値分解を用いる．全結合層の重み行列 $W(m$ 行 n 列) を対象として，特異値の個数 (ランク) を r とすると，W は行列 $U(m$ 行 r 列)，$D(r$ 行 r 列)，$V(r$ 行 n 列) の 3 つの行列の積で $W = UDV$ のように分解される．

行列 U，D，V はそれぞれ mr，r^2，nr 個のパラメータを持つ．元のパラメータ数よりも特異値分解後の 3 つの行列のパラメータ数の合計が小さくなるには，r が式 (4) の条件を満たす必要がある．

$$mr + r^2 + nr < mn. \quad (4)$$

特異値分解後，全結合層の重み行列 W は 3 つの行列積で表現される．1 つの全結合層で表すことは不可能であるため，Sainath らの方法 [7] に倣い，重み行列 W を持つ全結合層を，それぞれが重み行列 U，D，V を持つ 3 つの全結合層へと変換する．

3.5. プルーニング，量子化，特異値分解の組み合わせ

本研究では，既存手法のプルーニングと量子化，提案手法の全結合層への特異値分解を組み合わせる．特異値分解後の行列は実数値を取るため，量子化は特異値分解後に施す必要がある．そこで，プルーニング, 特異値分解，量子化の順序で処理を行う．

3.6. パラメータ削減を行う層の選択

全結合層の出力部分に用いられている ReLU や GELU などの活性化関数は，入力が負値の場合は 0 に近い値を出力する．そのため，各層の出力ベクトルの要素の負値の値が相対的に多い層は，圧縮してもそれ以降の層の重みの値が全体の出力に与える影響は少ないと考えられる．Sajjad ら [10] の研究においても，プルーニングを適用する層により精度の低下に数 % 程の違いがあると示されている．そこで，精度の低下を最小限に抑えるには，削減可能な層を選択する必要があると考えられる．本研究では，プルーニングを適用する層を選択する方法として，学習済みモデルに学習データの一部を再入力し，各層の出力ベクトルにおける要素の負値の割合を求めて指標とすることを提案する．

4. 実験

4.1. 実験条件

Vision Transformer のモデル (ViT-Base[2]) に含まれる MLP (全結合層) を対象として，ViT-Base を CIFAR10 で学習したモデルの容量削減を行う．オリジナルのモデルの精度は 97.5%，容量は 335MB である．本研究では，A. 容量削減手法の組み合わせパターンと，B. ViT-Base モデル中の全結合層 12 個のどの部分に

適用するかのパターンについて実験を行い，元のモデルに対する容量の削減率と，認識精度の維持の観点から最適な方法を考察する．

A. 手法の組み合わせパターン
4 種類の組み合わせ方を次のように定義する．
P: プルーニングのみ
P+Q: プルーニング · 量子化
P+SVD: プルーニング · 特異値分解
P+SVD+Q: プルーニング · 特異値分解 · 量子化
ただし，プルーニングは重みを削減する割合を 10% から 90% の間を 10% 刻みで実験を行い，特異値分解の特異値の個数は 300 とした．

B. 全結合層の選択パターン
ViT-Base モデル中の全結合層 12 個のどの部分に適用するかのパターンは，全 12 層を前半（第 1 層から第 6 層）と後半（第 7 層から第 12 層）に分け，前半のみに適用する場合・後半のみに適用する場合・全てに適用する場合，の 3 通りとする．これは，表 1 に示すように各層における出力ベクトルの要素の符号の割合を調べ，前半（第 1 層から第 6 層）よりも後半（第 7 層から第 12 層）の方が負値の割合が多いことから，前半と後半でパラメータの削減率や削減による精度低下の傾向に違いがあるのではないかと予測した事による．

4.2. 結果

図 1 に前半，後半，全ての層で適用する層を変えたときの実験結果を示す．ただし，図 1 の各点は左から順にプルーニングの割合が 10%, 20%, . . . , 90% の場合を表している．本実験では Wu[11] の実験に倣い，元モデルの精度 (97.5%) から精度の低下が 1% 以下を許容できる範囲と考える．図 1a，1b より，出力ベクトルの負値の割合が低い前半の層に適用した場合は精度が低下が大きく，負値の割合が大きい後半の層に適用した場合には精度の低下が起こらなかった．また図 1c より，全ての層に対して実験を行った場合，前半の層だけで実験したときと同様に，プルーニングの割合を上げていくと精度が低下した．プルーニングの割合で，それに続く特異値分解と量子化による削減率が変わる理由は，プルーニングの段階で重み行列のサイズが小さくなるほど特異値分解と量子化で削減することのできる要素は減るからである．図 1a より，前半の層へのプルーニングの適用が精度の低下を引き起こしている．前半の層は負値の割合が低いことが確認できていることから，プルーニングを適用する層は，出力ベクトルの負値の割合を指標とし，割合が高い層のみに適用することで精度を低下させることなくパラメータ数を削減できると考えられる．

表 1: 全 12 層の全結合層の出力に対する負値の個数の割合

層	1	2	3	4	5	6
負値の割合	88%	89%	89%	89%	90%	90 %
層	7	8	9	10	11	12
負値の割合	91%	94%	94%	97%	98%	98%

容量の観点からは，最初に適用したプルーニング単体の場合のグラフから 1% より大きく精度を落とすことなく容量の削減に成功した．精度の低下が 1% 以下であった後半における結果 (図 1b) より，プルーニング単体の場合とプルーニング・特異値分解・量子化を組み合わせた場合を比較すると，最大でプルーニング単体の場合から追加で約 20% の削減に成功し (プルーニング割合が 10% の場合．図 1b の最左の点)，最小でプルーニング単体の場合から追加で約 8% の削減に成功した (プルーニング割合が 90% の場合．図 1b の最右の点)．この結果より，複数の手法を組み合わせることで容量削減の効果を高めることができると考えることができる．

5. まとめ

深層学習モデルの圧縮について，従来から知られているプルーニングと量子化に行列分解を組み合わせた適用方法を精度維持とモデルの容量の観点から考察した．各層の出力ベクトルに負値の割合が大きい層を圧縮することで，精度の低下なく容量削減が可能であると考えられる．実験では，プルーニングと量子化と行列分解を出力ベクトルにおける要素の負値の割合が多い後半の層のみに適用する場合が最も精度の低下がなく圧縮可能であった．プルーニング後の圧縮された重み行列に対して行列分解を適用した場合でも，精度の低下なく更に容量削減可能であった．

参考文献

[1] A. Vaswani, N. Hazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaise, and I. Polosukhin, ”Attention is all you need,” In Neural Information Processing Systems (NIPS), 2017.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D.Weissenborn, X. Zhai, T. Unterhiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, ”An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” In IEEE International Conference on Learning Representations (ICLR), 2021.

[3] S. Han, H. Mao, and W.J. Dally, “Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding,”In-

ternational Conference on Learning Representations (ICLR), 2016.

[4] S. Han, J. Pool, J. Tran, and W. Dally, “Learning both weights and connections for efficient neural network,” Neural Information Processing Systems(NIPS), 2015.

[5] M.Courbariaux, Y.Bengio, J.P.David, ”BinaryConnect: Training Deep Neural Networks with binary weights during propagations, ” Advances in Neural Information Processing Systems, 3123-3131, 2015.

[6] I.Hubara, M.Courbariaux, D.Soudry, R.E.Yaniv, Y.Bengio, ”Binarized neural networks, ” Advances in Neural Information Processing Systems, 4107-4115, 2016.

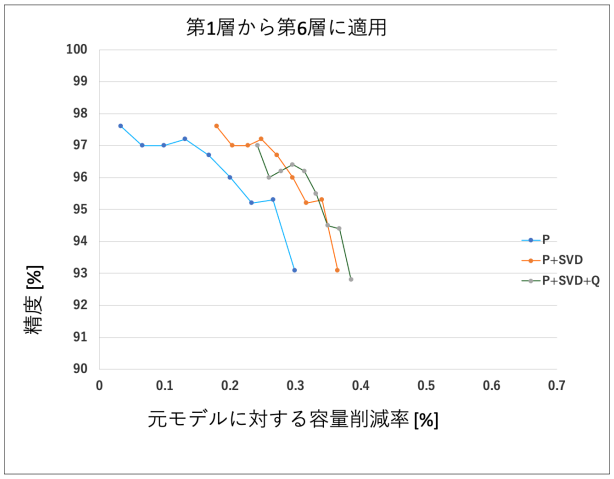
[7] T. N. Sainath, B. Kingsbury, V. Sindhwani, E.Arisoy, B. Ramabhadran, “Low-rank matrix factorization for Deep Neural Network training with high-dimensional output targets,” International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013.

[8] LeCun, Yann, Denker, John S, Solla, Sara A, Howard, Richard E, and Jackel, Lawrence D. Optimal brain damage. In NIPs, volume 89, 1989.

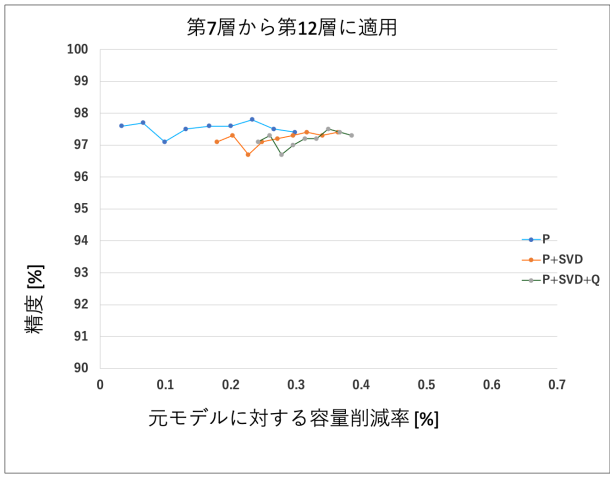
[9] I.Hubara, M.Courbariaux, D.Soudry, R.E.Yaniv, Y.Bengio, ”Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations, ” The Journal of Machine Learning Research, 18-1, 6869-6898, 2017.

[10] H.Sajjad, F.Dalvi, N.Durrani, P.Nakov, ”Poor Man’s BERT: Smaller and Faster Transformer Models, ” arXiv preprint arXiv:2004.03844v1, 2020.

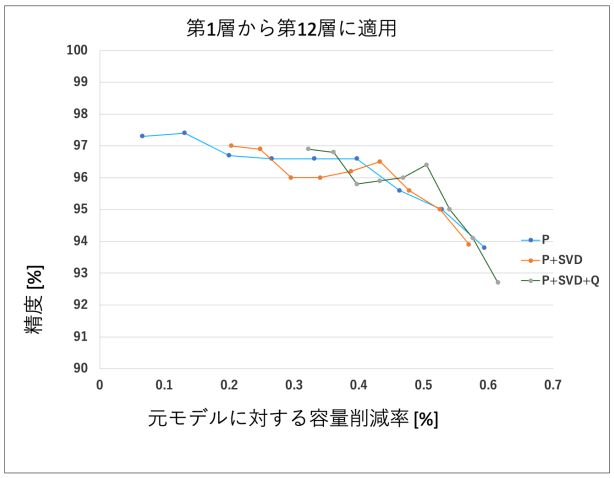
[11] H.Wu, P.Judd, X.Zhang, M.Isaev, P.Micikevicius, ”Integer Quantization for Deep Learning Inference: Principles and Empirical Evaluation, ” arXiv preprint arXiv:2004.09602v1, 2020.



(a) 第 1 層から第 6 層までを対象とした際の実験結果



(b) 第 7 層から第 12 層までを対象とした際の実験結果



(c) 第 1 層から第 12 層までを対象とした際の実験結果

図 1: 層を変えたときの精度と容量削減率