ディープラーニングのモデル軽量化技術

山本　康平　　橘　素子
前野　蔵人

近年、AIの基幹技術であるディープラーニングの適用事例が急速に拡大している。これまでは、大規模なGPUを搭載したオンプレミスのワークステーションやクラウドの利用が主流であったが、2016年ごろからエッジデバイスへの組込み実装や専用チップが登場し始めた。現在では、車載やスマートフォン、組込みIoTデバイスなどの多様なエッジデバイスに広がりつつある。しかし、一般に高精度なディープラーニングのモデルは、動作のために大容量のメモリーを必要とすることやその消費電力の高さから、エッジデバイスへの搭載が困難であった。

そこで、OKIでは元の精度（画像認識や音声認識などの推定精度を指す）を維持しつつモデルを軽量化し、演算リソースを大きく削減する技術を研究開発をしている。本稿では、そのモデル軽量化技術の現状と課題、及びOKI独自の技術を紹介する。

モデル軽量化技術とは

モデル軽量化技術とは、モデルの精度を維持しつつパラメーター数や演算回数を低減する手法の総称である。近年のディープラーニングは、実行に膨大なメモリーや演算能力を必要とすることから、モデル軽量化技術の必要性が高まっている。

ディープラーニングのモデルとは、狭義には4層以上に多層化したニューラルネットワークのことを指し、層間結合やバイアスなどの係数として、多数のパラメーターを持つ。通常、それらのパラメーターは16～32ビットの浮動小数点で表現される。ディープラーニングには、「学習」と「推論」の2つのフェーズがある。「学習」は、大量のデータを利用してパラメーターを最適化する処理であり、「推論」は学習によって最適化されたパラメーターを用いて未知のデータに対する答えを求める処理である。

エッジデバイスなどの処理能力の限られる実行環境では、学習よりも演算リソースが少なく済む推論機能だけを実装するのが一般的であるが、それでも高精度なモデルをエッジデバイス上で動作させることは難しい。その理由は、高精度なモデルほど膨大なパラメーター数や演算量を必要とするためである。そこで、モデル軽量化技術を適用することにより、それらの制約を軽減し高精度なモデルの推論機能をエッジデバイス上で高速に動作させることができる。

モデル軽量化技術の現状と課題

(1) モデル軽量化技術の分類

モデル軽量化技術には多様なアプローチが提案されているが、概ね6種類に分類できる。**表1**に分類とともにメモリー量・演算量（積和演算の回数）・併用の容易さ・精度への影響度の4つの観点での比較（△→○→◎の順で優位）を示す。「メモリー量」及び「演算量」はそれぞれの削減が期待で

表1 モデル軽量化技術の分類

分類	概要	メモリー量削減効果	演算量削減効果	併用の容易さ	精度への影響度
低ランク近似	重み行列を低ランク行列に分解・近似	○	△	○	○
量子化	演算のビット精度を低ビットに削減	◎	△	○	△
蒸留	大規模な学習済モデルを用いて小規模なモデルを学習	△	△	○	△
重み共有	重み係数を複数の結合で共有	○	△	△	○
高効率構造	畳込演算を複数の軽負荷な畳込演算の組合わせで代替	○	○	○	△
枝刈り	学習後のモデルから重要性の低いニューロンを削減	○	○	◎	◎

きる度合いを示す。「併用の容易さ」はその他の軽量化技術との組合せの容易さを示し、「精度への影響度」は軽量化技術を適用した際に生じる精度劣化の低減度合いを示している。それぞれの手法の詳細は以下のとおりである。

■**低ランク近似**:ディープラーニングにおける大部分の演算が大規模な行列演算で表現できることを利用して、その大規模な行列を小さな行列に数学的に分解・近似することで軽量化する。この手法は主にメモリー使用量の削減に向く。

■**量子化**:パラメーターを8ビット以下の固定小数点や整数に置き換えることで軽量化するが、丸め誤差や数値表現範囲の狭まりの影響で精度が劣化する。特に、4ビット未満の場合に精度が大きく劣化することが知られている。

■**蒸留**:大規模な学習済みの「教師」モデルと、小規模かつ未学習の「生徒」モデルを用意し、生徒モデルの出力と教師モデルの出力の差を最小化するように生徒モデルを学習する手法である。ただし、生徒モデルの選択に任意性が残り、最適な選択が難しいため、その他手法に比べて**表1**に記載の観点で劣る傾向にある。

■**重み共有**:モデルの重み係数を異なるニューロン間の接続で共有した上で学習する手法である。一つの係数を複数共有利用するため、メモリー使用量を削減できる。一方で、演算量の削減効果は少ない。

■**高効率構造**:ディープラーニングで最も多用されるネットワーク構造である畳み込みニューラルネットワーク（CNN）の畳込演算を、複数の軽負荷な畳込演算の組合わせで代替させた構造である。例えば、それぞれ同じデータを入力し独立に畳込演算させた後に結果を統合する並列的な組合わせ方法や、多次元の畳込演算を複数の低次元な畳込演算で代替し直列的に組み合わせる方法がある。効率的な構造であるが、大規模モデルほどの精度を持たないことが知られている。

■**枝刈り**:大規模なモデルの学習後、重要度の低いニューロンを削減する手法である。この考え方は、人の脳細胞が認知能力を確立するとともに減少していくことや、細胞が多少死滅しても、認知能力に影響が出ないことに似ているため、それを工学的に積極的に活用しようというアプローチである。この方法は、モデル構造を大きく変更しないため、その他の軽量化技術との組合わせの相性が良い。

これらの方式の中で、併用の容易さと精度への影響度のバランスに優れる技術が「枝刈り」である。ただし、精度への影響度が優位となるのは、次項記載の課題に対する適切な工夫を施した場合である。枝刈りの手法には、大別してニューロン単位とチャネル単位の2種類のアプローチがある。ニューロン単位とは、ニューラルネットワークの基本要素であるニューロンごとの重要度に基づき削減するものであり、チャネル単位とは、CNNに用いられる重み係数のグループであるフィルター単位やその演算結果の集合であるチャネル単位での重要度に基づき削減するものである。ニューロン単位の削減では、モデル全体に散在する重要度の低いニューロンをきめ細やかに削減でき、精度を維持しつつ高い削減率を達成しやすい。しかし、CNNではフィルターが複数のニューロンから成る構造を持つため、その一部を削減しても構造自体をそのまま保持する必要がある。それがメモリーアクセスの頻出などの問題に繋がり、演算効率を上げにくいといった実装面での課題となる。一方でチャネル単位の削減では、チャネルのデータを生成するフィルター単位での削減となることから、メモリー使用量と処理速度の両面で大きなメリットがある。

(2) チャネル単位の枝刈り手法の課題

チャネル単位枝刈りの従来技術には、「チャネル重要度の指標」と「チャネル削減率の設定」の二つの課題があった。

一つ目の課題は、チャネルの重要度を測る指標が各層に対して独立に計算される方式となっている点である。このような指標を用いると、例えば、ある層では重要でないと判断したチャネルが、別の層にとっては必要であった可能性が残る。すなわち、精度に貢献する重要なチャネルの喪失により、モデル軽量化後の精度劣化度が大きくなることが予想できる。従来技術を参照すると、各フィルターを構成する値の絶対和が大きいほど重要なチャネルと見なす指標[1)]や、推論時に削除しても計算結果の変化が小さいチャネルを重要でないチャネルと見なす指標[2),3)]がある。しかし、それらの指標はいずれも層ごとに独立な計算方式によって算出される。そのため、層内では良好な比較が行えるが、層間を考慮すると必ずしも最適なチャネルが選ばれているわけではなく、非効率な選択となりがちであった。従って、全ての層との関係を考慮できるような指標が望まれている。

二つ目の課題は、チャネルの削減率を層単位で個別に設定しなければならない点である。各層に割り当てる削減率はユーザーに委ねられるが、適切に設定しなければ精度を大きく損なってしまう。その理由は、CNNを構成する複数の畳込層のそれぞれが、枝刈りに対して異なる感度を持つためである[1)]。感度とは、チャネル削減率の精度への影響度合いである。例えば、ある層は削減率を高く設定しても精度への影響は少ないが、他のある層に対してそれと同等の削減率を設定すると著しい精度劣化を招く。そのため、ユーザーは感度を考慮しながら適切に削減率を選択しなければなら

ない。しかし、その感度の分析作業は試行錯誤と専門知識が必要であり、かつ最適な選択が難しい。さらに、より層数の多い大規模モデルに適用する場合には、削減率の必要設定数が多くなり、難度が飛躍的に高まってしまう。すなわち、チャネル削減率を層ごとに設定する作業を不要とし、モデル全体で一つのチャネル削減率を設定でき、その上で最適な層間の削減率の配分がなされる手法が望まれている。

PCAS 技術

OKIは前節に記載した二つの課題に対応した独自のモデル軽量化技術として、CNNモデルを対象としたチャネル単位の枝刈りを最適に行うPCAS（Pruning Channels with Attention Statistics）技術を保有している。本技術は、従来手法の抱える課題を解決し、高い精度を維持しながらモデルをメモリー使用量と演算量の両面で軽量化できることを特徴としている。

(1) 技術概要

PCAS技術の概要を図1に示す。軽量化の対象となるCNNモデルの層間に、新たなニューラルネットワークモデル（アテンションモジュールと呼ぶ）を挿入し、そのモジュールだけを対象とした学習を実行する。モジュールの出力層のニューロン数は、オリジナルネットワークの挿入前段の層の出力チャネル数と1:1に対応し、各ニューロンの値は対応するチャネルの値を次の層に伝播する量を制御するゲートとして機能する。この構成での学習とは、精度に寄与するチャネルに対応するニューロンほど高い値を出力することを目的とした、モジュールを構成するパラメーターの最適化となる。こうしてモジュール出力層の各ニューロンは、対応するチャネルの重要度を示すようになる。学習の完了したモジュールは、推論を実行すると個々のデータについてそれぞれ最適な重要度を出力できるようになるが、その平均値をチャネルの重要度として用いることができる。

PCAS技術のチャネル重要度指標は、従来技術[1),2),3)]の指標とは異なり、層間の関係を考慮できる。これはすなわち、前節の一つ目の課題を解決している。チャネルの重要度は、各モジュールがオリジナルの畳込層を挟みつつ全て接続された状態で学習が実行されるため、ゲートの重みとして相互に影響しながら最適化が進む。従って、各層のチャネルの重要度は、層全体に渡って最適化された値となる。この場合、ある層で重要でないと判断されたチャネルは、別の層でも重要でない可能性が高くなり、各重要度は互いに影響を受けにくい性質を持つことを意味する。すなわち、モデル全体として重要でないチャネルの削減が容易となり、結果として精度劣化を軽減する効果が得られる。

PCAS技術がモデル軽量化のために必要とするチャネル削減率は、層単位ではなくモデル全体で1つである。これは、前節の二つ目の課題の解決を意味している。具体的には、PCAS技術のチャネル重要度指標が、異なる層間でも同じ基準で評価できることを利用し、モデル全体の削減率を達成するまで全ての層の全てのチャネルの重要度に基づき、重要度の低いチャネルを削減していくことで、層ごとに異なる量のチャネル削減ができる。その後、削減したモデルのネットワークで再学習（ファインチューニング）することで軽量化が完了する。なお、挿入したモジュールは、チャネルの重要度を推定した後は取り外すため、これによる演算規模の増大は推論時に影響しない。

以上から、PCAS技術は層ごとのチャネル削減率の設定を不要としつつ、人手を介在しないことからチャネル削減率の配分が最適となり、精度を維持しつつメモリー使用量と演算量の大幅な削減の実現が期待できる。

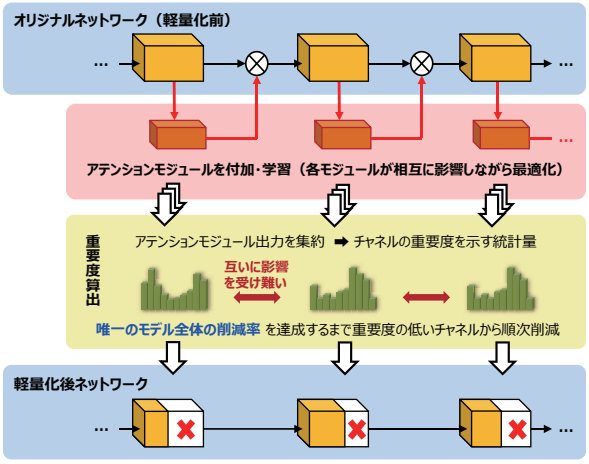

図 1　PCAS 技術適用の概念図

(2) 評価結果

一般的にディープラーニングのベンチマークで使われているデータセットと50層のモデルを用い、PCAS技術の有効性を確認した。その結果を図2に示す。左側の軸がパラメーター数及び演算回数であり、チャネル単位の枝刈り手法の適用前の状態をそれぞれ100%とした割合で表現している。また、右側の軸は精度を表している。

PCAS技術による結果は、軽量化前からの精度劣化が無い状態で、パラメーター数も演算回数も半分以下に削減できている。さらに、最近のトップクラスの国際学会で発表されている従来技術[2),3),4)]のベンチマークでも、同じ条件で、演算回数削減率を12ポイント、パラメーター数削減率を13ポイント程度改善でき、演算量とメモリー使用量の両面で効

率の良いモデルを実現できることを確認している。

この結果はモデル全体で1つのチャネル削減率を元に得られたものである。従って、層ごとの削減率の決定や感度の分析を一切行っていないにもかかわらず、優れた結果が得られることも示している。

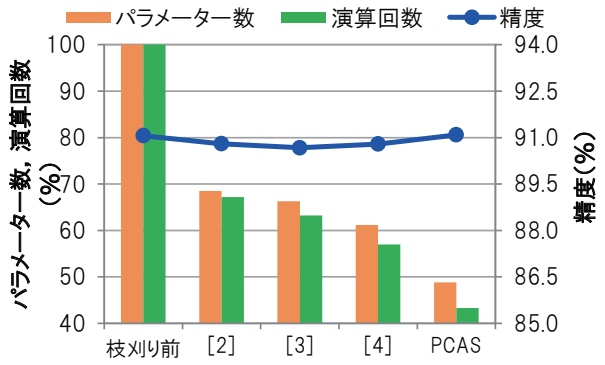

図２ 評価結果

今後の展望

本稿では、OKI独自のモデル軽量化技術であるPCAS技術を紹介した。現在は、更にモデル軽量化効果とハードウェア実装への親和性を高めるため、量子化との併用への対応を進めている。こうして実現する省リソースで高精度なディープラーニングモデルは、エッジ領域でのAI実装の普及を大きく加速することが期待され、OKIの多様なAIエッジソリューションに本技術を適用するために、開発を進めていく予定である。

謝辞

この成果の一部は、国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の委託業務の結果得られたものです。

◆◆

■参考文献

1）Hao Li, Asim Kadav, Igor Durdanovic and Hanan Samet, Hans Peter Graf: Pruning Filters for Efficient ConvNets, International Conference on Learning Representations (ICLR), 2017.

2）Yihui He, Xiangyu Zhang and Jian Sun: Channel Pruning for Accelerating Very Deep Neural Networks, International Conference on Computer Vision (ICCV), 2017.

3）Jian-Hao Luo, Jianxin Wu and Weiyao Lin: ThiNet: A Filter Level Pruning Method for Deep Neural Network Compression, International Conference on Computer Vision (ICCV), 2017.

4）Z. Huang and N. Wang:Data-Driven Sparse Structure Selection for Deep Neural Networks, European Conference on Computer Vision (ECCV), 2018.

●筆者紹介

山本康平:Kohei Yamamoto. 経営基盤本部 研究開発センター イノベーション推進室

橘素子:Motoko Tachibana. 経営基盤本部 研究開発センター イノベーション推進室

前野蔵人:Kurato Maeno. 経営基盤本部 研究開発センター イノベーション推進室

TIPS【基本用語解説】

ニューロン

ニューラルネットワークを構成する基本的な要素。多数の入力をもち、それらと重みとの線形結合に活性化関数などの演算を行い出力する構造を持つ。