Automated Hyperparameter Optimization and Novel Techniques for Reduced Computational Cost of Deep Neural Networks

坂井，靖文

(Degree)
博士（工学）

(Date of Degree)
2025-03-25

(Resource Type)
doctoral thesis

(Report Number)
甲第9219号

(URL)
https://hdl.handle.net/20.500.14094/0100496500

（別紙様式 3）

論文内容の要旨

氏　　名　坂井　靖文

専　　攻　情報科学専攻

論文題目（外国語の場合は，その和訳を併記すること。）

Automated Hyperparameter Optimization and Novel Techniques for Reduced Computational Cost of Deep Neural Networks

深層ニューラルネットワークの計算コスト削減に向けた自動ハイパーパラメータ最適化および新規手法

指導教員　天能　精一郎

（注）2，000 字～4，000 字でまとめること。

This dissertation presents computational cost reduction techniques of deep neural networks (DNNs), including automated hyperparameter search, not only for established fields such as image processing but also for emerging fields such as materials discovery.

Existing DNN computational cost reduction techniques often require complex manual hyperparameter tuning, limiting their efficiency and broad applicability. Furthermore, methods effective in established fields like image processing often prove inefficient when applied to emerging domains such as materials discovery. To address these limitations, we propose five novel techniques: (1) an 8-bit shifted dynamic fixed-point (S-DFP) quantization method minimizing accuracy loss; (2) a structured pruning method using zero-padding layers to improve compression efficiency, especially for networks with shortcut connections; (3) a gradient-aware automatic pruning rate search (GAPRS) algorithm automating optimal pruning rate selection; (4) a novel self-supervised learning method employing atom replacement for efficient learning with unlabeled materials data; and (5) a novel active learning method leveraging expanded features of material structural data to reduce the cost of generating labeled training data. We evaluated these techniques extensively. For image processing, experiments on ImageNet and CIFAR-10 using various ResNet architectures demonstrated significant reductions in model size and computational cost. For example, a 60.1% parameter reduction in ResNet-50 on ImageNet with 76.17% accuracy retained. For materials discovery, experiments on the Open Catalyst 2022 benchmark dataset using various graph neural networks (GNNs) such as PaiNN for material energy prediction showed that our method achieved the lowest prediction error with the fewest training data compared to existing methods.

Chapter 1 shows research background of the dissertation and basic characteristics for existing DNN computational cost reduction techniques such as quantization, pruning, self-supervised learning, and active learning.

In Chapter 2, challenges in DNN computational cost reduction are summarized: (1) trade-off between prediction accuracy, computation speed and model size of DNNs, (2) manual hyperparameter tuning for accelerating methods of DNN, and (3) DNNs computational cost for emerging fields such as materials discovery.

In Chapter 3, a quantization method using shifted dynamic fixed-point format for DNNs training is proposed. To prevent accuracy degradation due to quantized DNN training when using conventional DFP format, S-DFP shifts the data representable range of DFP from a large value area to a small value area by adding bias to the exponent of DFP. Using the proposed S-DFP format for quantized DNNs training, the quantized model can be trained using 8-bit fixed-point precision without significant accuracy degradation on the ImageNet task with ResNet-32, ResNet-50, ResNet-101, and ResNet-152. Since the validation accuracy of the quantized model with proposed S-DFP was achieved equivalent accuracy with FP32 regardless of the depth in ResNet, the proposed S-DFP is expected to significantly improve the accuracy of the quantized model regardless of the DNN model type, such as transformer other than ResNet. In addition, the proposed method achieved the equivalent accuracy with conventional 16-bit DFP instead of using 8-bit S-DFP.

In Chapter 4, to reduce the workload of inefficient manual pruning rate assignment, we describe our proposed automatic pruning rate search method named GAPRS for structured pruning. In the proposed method, by selecting a pruning rate in which the pruning error is smaller than a pre-defined threshold, the accuracy degradation of the pruned model is suppressed. Furthermore, we describe a structured pruning method without degrading the prediction accuracy of complex DNNs with shortcut connection. To improve the inference speed and the compression ratio for the DNN model with a shortcut connection such as ResNet, zero-padding layers are inserted into all input paths to the addition operator. The experimental results demonstrate that the pruned model accuracy can be effectively controlled by the proposed method. In addition, the pruning rate derived by the proposed method can be converged. We experimentally showed the superiority of GAPRS against various state-of-the-art methods on CIFAR-10 and ImageNet using various ResNet architectures. For ResNet-50, our method achieved a 60.1% reduction in parameters and a 59.8% reduction in FLOPs while maintaining 76.17% accuracy on ImageNet. These results demonstrate the effectiveness of GAPRS and our proposed method of inserting zero-padding layers. in achieving high compression ratios without model accuracy degradation.

Chapter 5 presents novel self-supervised learning (SSL) and active learning methods designed to reduce the computational cost of materials discovery using DNNs. Our proposed mask-less SSL method employs atom replacement with unlabeled data to improve the accuracy of catalyst energy prediction using GNNs. Unlike existing SSL methods that utilize fictitious "mask" atoms, our approach leverages only real elements, thereby promoting more efficient learning. We evaluated this method on three benchmark datasets (OC22, Poisoned Catalyst, and Expanded Poisoned Catalyst) using CGCNN and PaiNN GNNs, achieving the lowest prediction error across all experiments. Furthermore, we introduce a novel active learning method to enhance the accuracy of DNN-based catalyst energy prediction with limited training data. This method leverages all available material properties (structure and energy) to minimize the required training data. Evaluations on three benchmark datasets (OC22, NRR, and ORR) using PaiNN and EquiformerV2 GNNs demonstrate that our method achieves the lowest mean absolute error (MAE) among all evaluated sampling methods even with the fewest data.

Chapter 6 summarizes the conclusions of this study. This dissertation presents methods for achieving low-cost DNN computation and automated hyperparameter optimization to accelerate DNNs, addressing both established applications such as image processing and emerging fields such as materials discovery. Our work contributes to bridging the DNN computational demand-supply gap in diverse fields, including image processing and materials discovery, by avoiding inefficient manual hyperparameter tuning.

Keywords: Neural networks, Computational cost reduction, Model compression, Quantization, Pruning, Self-supervised learning, Active learning, Dynamic fixed-point, Structured pruning, Pruning rate search, mask less SSL, Feature sampling.

（別紙１）　　論文審査の結果の要旨

氏名	坂井　靖文		
論文題目	Automated Hyperparameter Optimization and Novel Techniques for Reduced Computational Cost of Deep Neural Networks (深層ニューラルネットワークの計算コスト削減に向けた自動ハイパーパラメータ最適化および新規手法)		
審査委員	区　分	職　名	氏　　名
	主　査	教授	天能　精一郎
	副　査	教授	太田　能
	副　査	教授	滝口　哲也
	副　査	教授	川口　博
	副　査	准教授	和泉　慎太郎

要　　旨

本学位論文では、画像処理などの確立された分野だけでなく、材料探索などの新興分野においても適用可能な、自動ハイパーパラメータ探索を含む深層ニューラルネットワーク（DNN）の計算コスト削減技術が提案されている。

既存の DNN 計算コスト削減技術は、複雑な手動によるハイパーパラメータ調整を必要とする場合が多く、その効率性と広範な適用性を制限している。さらに、画像処理などの DNN の有効性確立された分野で有効性が実証されている計算コスト削減手法も、材料探索などの新興分野に適用すると非効率となることがある。これらの課題に対処するため、本研究では 5 つの新規な手法が提案されている。(1) 量子化時の精度低下を最小限に抑える 8 ビットシフト動的固定小数点(S-DFP)量子化手法、(2) ショートカット接続を持つアーキテクチャ構成の DNN において圧縮効率を向上させるゼロパディング層を用いた構造化プルーニング手法、(3) 最適なプルーニング率の選択を自動化し、煩雑な手動調整を不要にする勾配を活用した自動プルーニング率探索アルゴリズム、(4) ラベルなし材料データを用いた効率的な学習のための原子置換を用いた、マスクを使用しない自己教師あり学習手法、そして(5) 材料構造データの拡張特徴量を活用して材料探索分野におけるラベル付き訓練データ生成のコストを削減する能動学習手法である。これらの提案手法について、本論文では広範なベンチマークデータセットと DNN モデルにより大規模に評価されている。画像処理については、様々な ResNet アーキテクチャを用いて ImageNet と CIFAR-10 のベンチマークデータセットで実験を行い、モデルサイズと計算コストの大幅な削減が実証されている（ImageNet 上の ResNet-50 でパラメータを 60.1%削減しながら、76.17%の精度を維持）。材料探索については、PaiNN などの様々なグラフニューラルネットワーク（GNN）を用いて Open Catalyst 2022 ベンチマークデータセットで実験を行い、既存の手法と比較して、最も少ない訓練データで最低の予測誤差が達成されている。

本学位論文は 7 章で構成されている。

第 1 章では、論文の研究背景と、量子化、プルーニング、自己教師あり学習、能動学習といった既存の DNN の計算コスト削減手法の基本的な特徴について述べられている。

第 2 章では、DNN 計算コスト削減手法の３つ課題について述べられている。つまり、1: 計算コスト削減手法を適用 DNN における予測精度と、計算速度およびモデルサイズのトレードオフ、2: DNN の計算コスト削減手法における、手動による非効率なハイパーパラメータ調整と、3: 材料探索といった新たな DNN の適用領域における DNN の計算コストについてである。

第 3 章は(1)に対応する。DNN の計算高速化、特に DNN の訓練高速化のための、シフトされた動的固定小数点形式(S-DFP: Shifted Dynamic Fixed Point format)を使用した量子化方法が提案されている。従来の動的固定小数点形式(DFP: Dynamic Fixed Point format)を使用した場合の量子化された DNN モデルの訓練による精度低下を防ぐために、S-DFP は DFP の指数(exponent)にバイアスを加えることで、DFP の表現可能な範囲を大きな値の領域から小さな値の領域にシフトさせる。ResNet-32、ResNet-50、ResNet-101、ResNet-152 を使用した ImageNet タスクで、提案された S-DFP 形式を使用して量子化された DNN モデルを訓練したところ、予測精度の大幅な劣化がなく、8 ビット固定小数点精度で量子化されたモデルを訓練できることが示されている。また、提案された S-DFP を使用した量子化モデルの予測精度は、ResNet の層の深さに関係なく FP32 と同等の精度を達成したことが示されている。さらに、提案された方法は、8 ビット S-DFP を使用する代わりに、従来の 16 ビット DFP と同等の精度が達成されている。この研究は以下の学術論文として出版された。Sakai, Y., & Tamiya, Y. (2021), S-DFP: shifted dynamic fixed point for quantized deep neural network training, *Neural Computing and Applications*, 1-8.

氏名	坂井　靖文

第 4 章は(2)と(3)に対応する。ショートカット接続を持つ複雑な DNN モデルの予測精度を低下させることなく、構造化プルーニングによるモデル圧縮化率を改善する手法が提案されている。ResNet のようなショートカット接続を持つ DNN モデルでは、ショートカット接続の結合点において、加算演算子による行列の連結操作(concatenate)が実行される。ここで、ショートカット接続を持つ DNN モデルに対して構造化プルーニングを適用すると、結合点において連結操作を行う複数の行列の行列サイズの不一致が発生するため、連結操作が実行できなくなる。そこで、ショートカット接続を持つ複雑な DNN モデルの計算速度、特に推論の計算速度と圧縮率を向上させるために、加算演算子への全ての入力経路にゼロパディング層を挿入する。ベンチマークタスクである CIFAR-10 と ImageNet と、さまざまな ResNet アーキテクチャを使用して提案手法が評価され、ResNet-50 では、DNN モデルのパラメータを 60.1%削減し、FLOPs を 59.8%削減しながら、ImageNet で 76.17%の精度を維持した。また非効率的な手動によるプルーニング率調整の作業負荷を軽減するために、GAPRS と名付けられた構造化プルーニングのための自動プルーニング率探索手法が提案されている。提案された方法では、プルーニングされる重みの大きさで定義されるプルーニング誤差が、勾配を用いて事前に定義されたしきい値よりも小さくなるようにプルーニング率を選択することで、プルーニングされたモデルの精度低下を抑制する。実験結果は、提案された方法によってプルーニングされたモデルの精度を効果的に制御できることを示している。提案された方法によって導出されたプルーニング率は収束可能である。さらに、提案手法はプルーニング前後の重みの大きさの差分のみを利用していることから、量子化時に発生する量子化誤差を用いることで、量子化にも適用可能である。ベンチマークタスクである CIFAR-10 と ImageNet と、さまざまな ResNet アーキテクチャを使用して、多様な最先端手法に対する GAPRS の優位性が実験的に示されている。例えば、ResNet-50 では、パラメータを 60.1%削減しながら、ImageNet で 76.17%の精度を実現している。この結果は、DNN モデルの予測精度の大幅な劣化を引き起こすことなく、高いモデル圧縮率を達成する GAPRS の有効性を示している。これらの研究は以下の学術論文と国際会議として出版、発表された。Sakai, Y., Eto, Y., & Teranishi, Y. (2022), Structured pruning for deep neural networks with adaptive pruning rate derivation based on connection sensitivity and loss function, *Journal of Advances in Information Technology*, *1*. Sakai, Y., Iwakawa, A., Tabaru, T., Inoue, A., & Kawaguchi, H. (2022), Automatic pruning rate derivation for structured pruning of deep neural networks, In *IEEE international conference on pattern recognition (ICPR),* 2561-2567.

第 5 章は(4)と(5)に対応する。また材料探索の計算コスト削減を目的とした、自己教師あり学習と能動学習手法が提案されている。提案される自己教師あり学習手法は、既存手法が架空の「マスク」原子を使用するのに対し、本手法は現実の元素のみを使用することで効率的な学習を実現する。CGCNN と PaiNN GNN を用いて、OC22 を含む 3 つのベンチマークデータセットで提案手法は評価され、全てのケースで最低の予測誤差が達成されている。さらに、限られた訓練データで DNN による触媒エネルギー予測精度を向上させるための新規な能動学習手法が提案されている。提案手法は、構造やエネルギーなどの利用可能な全ての材料特性を訓練データの選択に活用することで、必要な訓練データ数を最小限に抑えている。PaiNN と EquiformerV2 を用いて、OC22 を含む 3 つのベンチマークデータセットで本手法の有効性が検証され、評価された全てのサンプリング手法の中で、提案手法はデータ数が最も少ない場合でも最低の平均絶対誤差を達成している。これらの研究は以下の 2 件の国際会議で発表された。Sakai, Y., Matsumura, N., Inoue, A., Kawaguchi, H., Thang, D., Ishikawa, A., Höskuldsson, Á. B. & Skúlason, E. (2024), Active Learning for Graph Neural Networks Training in Catalyst Energy Prediction, In *IEEE International Joint Conference on Neural Networks (IJCNN),* 1-8. Sakai, Y., Dang, T., Fukuta, S., Shirahata, K., Ishikawa, A., Inoue, A., H. Kawaguchi, Höskuldsson, Á. B. & Skúlason, E. (2023), Self-Supervised Learning with Atom Replacement for Catalyst Energy Prediction by Graph Neural Networks, *Procedia Computer Science*, *222*, 458-467.

第 6 章では、本学位論文の結論について述べられている。

本論文では、画像処理などの確立された応用分野だけでなく、材料探索などの新興分野について、低コストな DNN 計算と DNN 高速化のための自動ハイパーパラメータ最適化手法を研究したものであり、画像処理や材料探索を含む様々な分野における DNN 計算の需要と供給のギャップの解消に大きく貢献した。本論文はディープニューラルネットワークの性能最適化のための自動ハイパーパラメータ探索について重要な知見を得たものとして価値ある集積である。提出された論文はシステム情報学研究科学位論文評価基準を満たしており，学位申請者の坂井靖文は，博士（工学）の学位を得る資格があると認める。