氏名（本籍）	菅間　幸司（東京都）
学位の種類	博　士（工学）
学位授与番号	甲 第１０４号
学位授与日付	令和３年３月２５日
専攻	システム工学専攻
学位論文題目	Behavior-based DNN Compression:Pruning and Facilitation Methods
学位論文審査委員	（主査）教　授　和田　俊和 （副査）教　授　風間　一洋 講　師　八谷　大岳 栗田　多喜夫（学外委員）

論文内容の要旨

This thesis presents a set of methods for compressing Deep Neural Network (DNN) models. In various Computer Vision tasks, DNN models show excellent performance and are expected to be used in a lot of industrial applications. However, due to their heavy computational requirements, it is difficult to implement them on the edge devices with limited computational resources. Therefore, it is desired to develop a method to compress large DNN models without accuracy degradation.

An effective approach for this purpose is pruning. Pruning is a technique to reduce the computational cost of pretrained DNN models by removing redundant neurons. The most important feature required for the pruning methods is to preserve the accuracy of DNN models as well as possible, while making them smaller. If the model is severely damaged by pruning, its accuracy may not be recovered by retraining. Therefore, there is a demand for a pruning method that can preserve the model accuracy well.

In this thesis, we present two pruning methods, Neuro-Unification (NU) and Reconstruction Error Aware Pruning (REAP), and two methods for facilitating pruning, Pruning Ratio Optimizer (PRO) and Serialized Residual Network (SRN). The outlines of the proposed methods are as follows.

Neuro-Unification (Related publications: [1, 2])

Neuro-Unification (NU) is a neuron behavior-based pruning method. The neuron behavior is defined by a vector composed of the neuron's outputs corresponding to several input data. In NU, we do not just prune but unify a pair of neurons having similar behaviors. Unifying a pair of neurons is, in other words, 1) pruning one of them, and 2) updating the weights connected to the other one so that the behavior of the pruned one is reconstructed. Therefore, NU suffers smaller error and preserves the model accuracy better than the methods that only conduct pruning.

Reconstruction Error Aware Pruning (Related publications: [3, 4, 5])

Reconstruction Error Aware Pruning (REAP) is an extended version of NU and is the best pruning method in terms of reconstructing the layer-wise error.

In REAP, the behavior of the pruned neuron is reconstructed from the behaviors of all remaining neurons in the same layer by using least squares method. Thus, the error caused by pruning can be compensated even better.

The problem of REAP is the large computational cost for selecting the neurons to be pruned. In a naïve way, we once try to prune each one of them, reconstruct its behavior, and check which neuron has the smallest reconstruction error. This is computationally expensive, especially when we have lots of neurons (e.g. 4,096 neurons). Therefore, we developed a biorthogonal system-based algorithm with which we can compute the reconstruction errors of all neurons in one-shot.

Pruning Ratio Optimizer

As the DNN models usually have several layers, we need to tune the pruning ratio (the ratio of the neurons to be pruned) in each layer properly, in order to preserve the accuracy well. REAP is the best method in terms of minimizing the layer-wise error. However, we do not know how much this layer-wise error affects the model performance. Thus, we cannot tune the pruning ratios based on only the layer-wise errors.

Therefore, we propose Pruning Ratio Optimizer (PRO), a method for tuning the pruning ratios based on the error in the final layer of the model (while neuron selection in each layer is based on the layer-wise error). The idea of PRO is to select the layer where pruning will have the smallest impact on the final layer, and to prune some neurons in the selected layer, repeatedly. PRO is an efficient and easy-to-use method, as it takes a greedy approach and has few hyper-parameters to be controlled.

Serialized Residual Network (Related publications: [6])

A limitation of the layer-wise pruning methods including REAP is that ResNet is difficult to prune. The ResNet architecture is composed of the stacked blocks with branched paths. In each block, the inputs are propagated as they are in one path, the linear transformations (convolutions) are performed in the other path, and both are eventually added. At this addition, two inputs must have the same dimensions, which means the layers having the branched paths cannot be pruned. This limitation is significant for us, because ResNet architecture is used for various DNN models.

Therefore, we propose a technique to transform a trained ResNet model into an equivalent model having a serial architecture whose weights are partially fixed for conducting identity mapping. We call this serialized model a Serialized Residual Network (SRN) model. Although the SRN model has more computational complexity than the original ResNet model, we can reduce its complexity drastically by pruning with smaller degradation, because the SRN model has a serial architecture and we can perform pruning in any layer.

Apart from the purpose of facilitating pruning, the proposed method can also be used for improving the accuracy of the pretrained ResNet model. This way of using SRNs is effective if the advantage of improved accuracy outweighs the disadvantage of increased complexity. The problem is that when training the SRN model, it suffers the *side effect* of L2 regularization. L2 regularization strongly penalizes the weights that are far from zero. The identity mapping portion of the weight matrices/tensors in the SRN model contains the weights that are equal to 1, and 1 is a large value for a DNN's weight. These weights are penalized too strongly by L2 regularization, and updated significantly during training, which results in accuracy degradation.

In order to solve this problem, we suggest Elastic Weight Regularization (EWR). While L2 regularization strongly penalizes the weights far from zero, EWR penalizes the weights that are far from their initial values. With EWR, the weights that are initially equal to 1 can avoid being penalized too strongly, and the training of SRN models can be stabilized.

The highlight of this thesis is REAP. REAP is theoretically the best pruning method among the methods that conduct pruning based on layer-wise error. In addition, the biorthogonal system-based algorithm for neuron selection is a novel way of using biorthogonal system.

References

[1] Koji Kamma, Yuki Isoda, Sarimu Inoue, and Toshikazu Wada. Behavior-based compression for convolutional neural networks. In *Proceedings of International Conference on Image Analysis and Recognition (ICIAR)*, Vol. 11662, pp. 427–439, 2019.

[2] Koji Kamma, Yuki Isoda, Sarimu Inoue, and Toshikazu Wada. Neural behavior-based approach for neural network pruning. *IEICE Transactions on Information and Systems*, Vol. E103-D, pp. 1135–1143, 2020.

[3] Koji Kamma and Toshikazu Wada. Reconstruction error aware pruning for accelerating neural networks. In *Proceedings of International Symposium on Visual Computing (ISVC)*, pp. 59–72, 2019.

[4] Koji Kamma and Toshikazu Wada. Reap: A method for pruning convolutional neural networks with performance preservation. *IEICE Transactions on Information and Systems*, Vol. E104-D, pp. 194–202, 2021.

[5] Koji Kamma and Toshikazu Wada. Accelerating the convolutional neural networks by smart channel pruning. In *Proceedings of the 22nd Meeting on Image Recognition and Understanding (MIRU)*, 2019.

[6] Koji Kamma and Toshikazu Wada. Serialized residual network. In *Proceedings of the 23rd Meeting on Image Recognition and Understanding (MIRU)*, 2020.

論文審査の結果の要旨

菅間幸司君の博士論文審査会は，2021 年 1 月 27 日に開催した．候補者の論文は，Deep Neural Network (DNN)の圧縮技術に関するものであり，層単位の圧縮法である NU と REAP，DNN 全体の圧縮を行う際に適切に各層の圧縮率を決める PRO，及び，分岐を含む Residual Net を直列化して圧縮を行う SRN に関するものである．このうち，REAP による圧縮法において，最小二乗法を何度も解くことなく，双直交基底を用いて One shot で再構成後の誤差が最小になる削除ニューロンを求める手法は，独創性があり，特に高く評価できるものである．また，層単位の圧縮法を DNN 全体に適切に適用するための圧縮率を求める手法である PRO や SRN も，理論と技術の両面において優れている．これらの技術を用いて圧縮した DNN は，すでに一部社会実装もされており，博士論文として十分な内容を含んでいると判断する．

最終試験の結果の要旨

審査委員より，DNN 及びその学習法や圧縮法，さらにそれらの実験方法などについて，複数の専門的質問が出され，候補者は，専門的な知識と客観的なデータに基づき，分かりやすく適切な回答を行っている．このことから，数学， Deep Neural Network，機械学習などに関する博士の学位を持つ者として，標準以上の専門的知識を有していると判断できた．また，このような専門的な知識に基づく回答を，可能な限り平易な言葉で説明しており，専門的な知識をわかりやすく伝えるという点でも，博士の学位を与えるにふさわしいと判断する．