채널 프루닝과 전이 학습을 이용한 경량 DNN 모델 개발

Development of a Lightweight DNN Model using channel Pruning with Transfer Learning

사라 수알리힌, 김덕환 [1]

Sara Sualiheen, Deok-Hwan Kim[1]
Department of Electrical and Computer Engineering, Inha University, Incheon, South Korea
sarasualiheen@inha.edu, deokhwan@inha.ac.kr

Deep neural networks (DNNs) have been widely used in various applications, however the computational complexity and memory requirements of DNNs are becoming challenging, especially in resource-constrained devices, such as mobile phones and embedded systems. In this paper, we propose a lightweight DNN model using channel pruning to address the computational complexity and memory re resource-constrained devices. Our approach combines channel prunir to maintain accuracy. Evaluation on the CIFAR10 dataset, ResNet50, x 78% test accuracy, 89% train accuracy, and 73% validation accu unpruned model. The pruned model is suitable for applications with

1. Introduction

AI and DL have achieved remarkable success in various fields such as image recognition [1], natural language processing [2], and speech recognition [3]. However, the high computational requirements, large storage, and longer training times associated with DL models have created a need for lightweight models that can run on resource-constrained devices. Channel pruning is an effective approach for developing lightweight DNN models by removing redundant connections from pre-trained models, resulting in a more compact and efficient model [4].

The main contribution of the study includes:

- Applying transfer learning approach to the pre-trained ResNet50 and fine-tune it on CIFAR10 dataset.
- Dense layer added after Global Average Pooling for flexibility and accuracy enhancement.
- Fine-tuning the pruned model on CIFAR-10 dataset to compensate the loss accuracy due to channel pruning.

2. Related Works

Channel pruning is a hot topic of discussion for many years since 1990's [5, 6]. [7] introduced a novel channel pruning technique that balances computational complexity and accuracy in neural networks. Similarly, [8] presents a heuristic-based filter pruning method for deep neural networks, enabling deployment on resource-constrained devices. The method is validated on diverse architectures and datasets such as AlexNet, VGG16, ResNet34, CIFAR10, CIFAR100, and ImageNet.

In [9], authors introduce a structured pruning method for deep neural networks that dynamically determines the pruning rate for each layer based on the gradient and loss function, eliminating manual assignment. The method is evaluated on CIFAR-10 with VGG-16 and ResNet using

[1] Corresponding Author

iterative pruning techniques.

3. Proposed Methodology

3.1. Proposed Framework

The proposed framework employs channel pruning to reduce computational complexity while preserving accuracy in deep neural networks. The study introduces a framework (Fig1) that incorporates transfer learning on the pre-trained ResNet50 model. By utilizing transfer learning, the classification layer of ResNet50 is modified from 1000 to 10 classes to align with the CIFAR10 dataset. The model is then fine-tuned on the CIFAR-10 dataset, followed by weight pruning to reduce the computational complexity of the model. The advantage of transfer learning includes reduced training time and small dataset implication. The channel pruning technique removes unnecessary weights from the model, resulting in a more compact and efficient network. The performance of both the pruned-ResNet50 and pre-trained ResNet50 models have been evaluated to compare their accuracy.

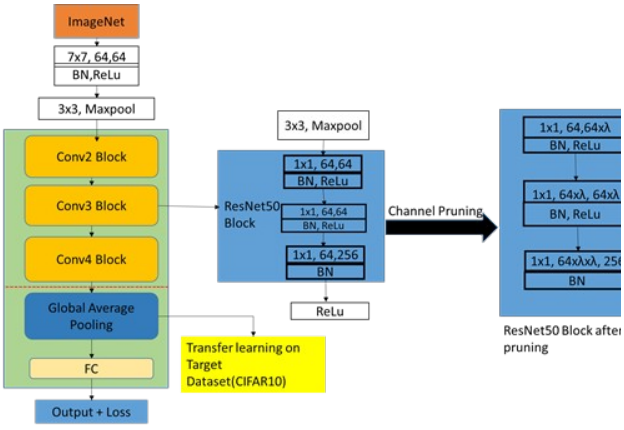

Figure 1. Proposed Modification in the ResNet50

The proposed objectives were achieved by applying the following algorithm.

Algorithm: Achieving Channel Pruning

1. **INPUT:** CIFAR-10
2. **INPUT:** Pre-Trained ResNet50
3. **OUTPUT:** Channel Pruned ResNet50
4. for all layers
5. For n=1 to N, do
6. Compute weight value
7. Sort in descending order
8. Remove 10% of channels from each block
9. End for
10. Retrain the pruned model
11. Return the pruned model

The initial step involves taking the pre-trained ResNet50 model trained on ImageNet with 1000 classes and applying transfer learning to adapt it to the CIFAR10 dataset with 10 classes. Channel pruning is then applied to all blocks of ResNet50 by sorting weights in descending order based on their magnitude and removing weights below a 10% threshold. This process converts the DNN model into a more compact form.

4. Experiments

4.1. Experimental setup

Python (Tensor Flow) experiments were conducted on a computer system powered by a 12th Generation Intel® Core™ i9-12900K 3.20GHz processor.

4.2. Dataset

The CIFAR-10 dataset [10], is a popular resource for training machine learning and computer vision models. It consists of 60,000 color images that are 32x32 pixels in size, with 6,000 images for each of the 10 classes. These classes are airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks.

4.3. Experimental result

To validate our approach, we utilized the CIFAR-10 dataset to assess the performance of our pruned network. The application of CIFAR-10 allowed us to retrain ResNet50 on the target dataset, resulting in reduced computational complexity and improved accuracy. This fine-tuning process enhances the model's ability to adapt and generalize, potentially reducing parameters and computational complexity.

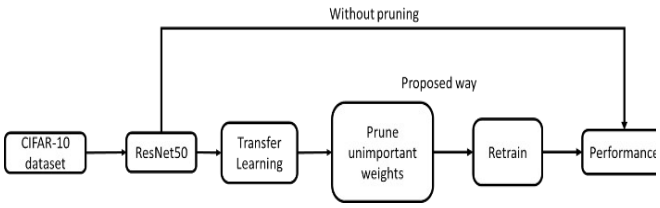

Figure 2. Proposed Methodology

<Table 1> Number of parameters in ResNet50 and Pruned ResNet50

ResNet50 and pruned ResNet50 Parameters		
Block #	**ResNet50**	**Pruned ResNet50**
1	183506	165155
2	107336	96602
3	197436	177692
4 to N	233,867,308	210,480,578
Total	234,355,586	210,920,027

The pruned network achieved a train accuracy of 89% and a test accuracy of 78%.

By applying pruning, the ResNet50's parameters were reduced from 234,355,586 to 210,920,027.

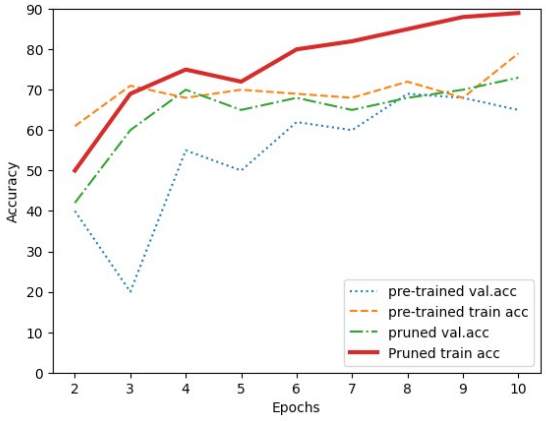

Figure 3. Pruning on RESNET-50

In Figure3, the bold line represents the higher training accuracy of the pruned model (89%) compared to the pre-trained ResNet50 (79%). The dashed line shows the pruned model's accuracy on the validation dataset (73%), while the dotted line represents the pre-trained ResNet50's accuracy (65%). The test accuracy of the pruned ResNet50 is 78%, while the pre-trained model achieves 73%.

<Table 2>comparing the performance of pruned ResNet50 and TF-Lite model.

Model	Accuracy	Loss	Processing time on dataset	Number of parameter
Pruned ResNet50	78%	0.45	1.5 sec/image	21 million
TF-Lite	77%	0.48	2.5 sec/image	25 million

Our results in Table2 demonstrate superior performance compared to the TF-Lite model. Our analysis demonstrates that channel pruning with transfer learning is an effective approach to create lightweight DNN models with minimal compromise on accuracy.

5. Conclusions

We propose a method to create a lightweight DNN model using channel pruning and transfer learning. By combining these techniques, we effectively reduce the computational complexity while preserving accuracy. Our findings indicate that the pruned network is well-suited for resource-constrained devices like those in blockchain networks and edge computing environments.

Acknowledgment

This research was supported in part by the BK21 Four Program funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea(NRF)and in part by Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE)(P0017124, The Competency Development Program for Industry Specialist) in part by the National Research Foundation of Korea(NRF) grant funded by the Korean government(MSIT) (No. NRF-

2021R1F1A1050750).

References

[1] Khan, M., Jan, B., Farman, H., Ahmad, J., Farman, H., & Jan, Z. (2019). Deep learning methods and applications. Deep learning: convergence to big data analytics, 31-42.

[2] Chen, P. F., Chen, L., Lin, Y. K., Li, G. H., Lai, F., Lu, C. W., ... & Lin, T. Y. (2022). Predicting Postoperative Mortality with Deep Neural Networks and Natural Language Processing: Model Development and Validation. JMIR Medical Informatics, 10(5), e38241.

[3] Liang, S., & Yan, W. (2022). Multilingual speech recognition based on the end-to-end framework. Multimedia Tools and Applications.

[4] Luo, J. H., & Wu, J. (2020). Neural network pruning with residual-connections and limited-data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1458-1467).

[5] Hanson, S., & Pratt, L. (1988). Comparing biases for minimal network construction with back-propagation. Advances in neural information processing systems, 1.

[6] LeCun, Y., Denker, J., & Solla, S. (1989). Optimal brain damage. Advances in neural information processing systems, 2.

[7] Yu, X., Serra, T., Ramalingam, S., & Zhe, S. (2022, June). The combinatorial brain surgeon: Pruning weights that cancel one another in neural networks. In International Conference on Machine Learning (pp. 25668-25683). PMLR.

[8] Choudhary, T., Mishra, V., Goswami, A., & Sarangapani, J. (2022). Heuristic-based automatic pruning of deep neural networks. Neural Computing and Applications, 34(6), 4889-4903.

[9] 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.

[10] Yang, C., & Liu, H. (2022). Channel pruning based on convolutional neural network
sensitivity. Neurocomputing, 507, 97-106.