Optimization of CNN Structure based on Principal Component Analysis and Sparsification

Arashi Hirose, Yoshihiro Maeda, and Takayuki Hamamoto

***Abstract*—In this paper, we propose an optimization method for CNN models that combines the search for the optimal structure using principal component analysis and sparsity-based pruning. Conventional pruning methods search for the optimal structure of CNN by using principal component analysis, and the pruned model is retrained. However, this method requires a lot of retraining time because the learned weights are not used. The proposed method determines the optimal number of channels using principal component analysis, and model weight is pruned based on sparsity while reutilizing the learned weights. The proposed method can make pruned models that maintain accuracy with minimal retraining.**

***Index Terms*—CNN, pruning, PCA, sparsification**

I. Introduction

Convolutional neural network (CNN) performs highly in various tasks such as object recognition. However, larger CNN models are required for high performance, leading to increased memory usage and computation. Therefore, we require CNN pruning methods, which reduce the weight of the original model while maintaining its accuracy and significantly reducing computational resources. Method [1] uses principal component analysis (PCA) to search for the optimal structure of a CNN to construct the pruned model. However, when pruning the weights, the weights are initialized without utilizing the learned weights, which requires enormous retraining time. Li *et al.* [2] proposed a pruning method that utilizes the fact that filters with small magnitude weights tend to be less important. However, this method determines the optimal model structure by iterative re-training; thus, it causes a time-consuming process. This paper proposes a pruning method for CNN models that combines PCA-based search for the optimal structure and sparsity-based pruning that maintain accuracy with little retraining time.

II. PROPOSED METHOD

The proposed method first determines the optimal number of channels for each layer of the learned model by using PCA as in [1]. We obtain N feature maps resulting from the forward propagation of each layer, where N is the number of input images. Therefore, the number of samples is $N \times H \times W$ for each channel, where H and W denote the vertical and horizontal sizes of the feature map, respectively. PCA is then performed on the feature maps obtained for each layer. Based on the contribution ratio obtained by PCA, the number of eigenvalues required to reach a specified cumulative contribution ratio α is determined as the optimal number of channels for that layer.

TABLE I
Top-1 accuracy of object detection task.

compression rate	method [1] 200epoch	method [1] 20epoch	the proposed method 20epoch
43%	93.13	89.37	**93.25**
47%	93.04	88.75	**93.30**
56%	92.99	88.88	**93.17**
65%	**93.14**	89.12	92.83

After determining the optimal structure of the model by PCA, the proposed method removes unimportant channels by pruning based on L1 norm [3]. The proposed method calculates the L1 norm of the weights for each filter, which represents its importance [2]. Then, the filters corresponding to output channels with small values of the L1 norm are removed from the original model until the number of channels determined by PCA is reached. The remaining channels are given the learned parameters of the original model.

III. FINDINGS

We experimented to show the significance of the proposed method for the object recognition task. We used the vgg16_bn model and the CIFAR10 dataset. We used Garg *et al.* [1] as a comparison method. The cumulative contribution rate α was set to 99.0%. TABLE I shows the top-1 accuracies of the object detection task. Comparing the proposed method learned in 20 epochs with [1] learned in 200 epochs, the accuracy of the proposed method is almost the same as that of [1]. Regarding the accuracy of 20 epochs, the proposed method achieves a higher accuracy than [1]. This indicates that [1] requires a large number of epochs to recover the accuracy, while the proposed method can recover the accuracy in smaller epochs.

IV. CONCLUSION

We proposed a pruning method based on the search for the optimal structure of a CNN model using principal component analysis and sparsity. The proposed method reduces the retraining time by about ten times with almost the same accuracy as the conventional method.

References

[1] I. Garg *et al.*, “A Low Effort Approach to Structured CNN Design Using PCA,” IEEE Access, vol. 8, pp.1347-1360, 2020.

[2] H. Li *et al.*, “Pruning filters for efficient convnets,” in ICLR, 2016.

[3] W. Wen *et al.*, “Learning structured sparsity in deep neural networks,” in NIPS, 2016.

*Research supported by JSPS KAKENHI(21K17768)

Arashi Hirose, Yoshihiro Maeda, and Takayuki Hamamoto are with the Tokyo University of Science, 6-3-1 Niijuku, Katsushika-ku, Tokyo 125-8585, Japan.