Gradually Global Pruning for Object Detection Network

Haoze Sun
Department of Weapon and Control
Army Academy of Armored Force
Beijing, china
sunhz1989@163.com

Tianqing Chang
Department of Weapon and Control
Army Academy of Armored Force
Beijing, china
changtianqing@263.com

Lei Zhang
Department of Weapon and Control
Army Academy of Armored Force
Beijing, China
zhangl_1974@163.com

Guozhen Yang
Department of Weapon and Control
Army Academy of Armored Force
Beijing, china
diegorevil@163.com

Bin Han
Department of Weapon and Control
Army Academy of Armored Force
Beijing, china
Han_b08@163.com

***Abstract*—State-of-the-art object detection methods employ the deep convolutional neural Network (CNN) to achieve excellent results on several public datasets. However, most of these methods depend on complex backbone networks which limit the application in many real-world scenarios. In this paper, a novel gradual global pruning is presented to remove the redundant filters for object detection models, which can largely reduce the model size of the network while maintain the network accuracy. Firstly, the importance of individual filter is evaluated globally across all the network layers by calculating the sum of its absolute weights. Then, a progressive sparsity control function is proposed to gradually reduce the proportion of pruning in the implementation process. Finally, this paper finetunes the pruned network to restore its accuracy. The experimental results on Pascal VOC2007 dataset show that the proposed method can drastically reduce the model size of the network with negligible loss in detection accuracy.**

Keywords—gradual global pruning, object detection, deep convolutional neural Network

I. INTRODUCTION

Deep Convolutional Neural Network (CNN) [1] based methods currently achieve great success on most of the computer vision tasks. State-of-the-art object detection method, such as Faster-RCNN [2], YOLO [3], SSD [4], also employ the powerful tool to achieved excellent results on several public datasets (Pascal VOC [5], KITTI [6], etc). However, most of these methods depend on very complex backbone networks (VGGNet [7], ResNet [8]), which require high storage and operation capabilities of hardware devices, thus limiting the application in many real world scenarios such as robotics, self-driving car and augmented reality. For example, VGG-16 network contains about 138M floating-point parameters and the model size is more than 500M, which greatly exceeds the capacity of most embedded platforms. As such, more efficient and lightweight detection models are highly desired for the application in many resource-limited platforms.

Some attempts have been made to reduce the storage and computation costs by model pruning. LeCun et al. [9] used the idea of information-theoretic to remove some unimportant weights from a trained network with negligible loss in accuracy. However, this method is not applicable for deep CNN model due to the expensive memory and computation costs. Han et al. [10] proposed a simple pruning approach by removing weights with values below a threshold and fine-tuning the pruning model to recover its accuracy. This iterative procedure is performed several times, generating a very sparse model. However, such a non-structured sparse model should be supported by specialized hardwares and softwares for efficient inference, which is difficult and expensive for the application in real-world scenarios. Li et al. [11] proposed a structured pruning strategy which aims at directly removing filters as a whole. The filter level pruning could reduce the memory footprint dramatically and is far more efficient and independent to specialized hardware/software platforms. However, [11] used a layer-by-layer fixed pruning manner, which is less efficient and computational intensive in the process of implementation.

In this paper, a novel gradual global pruning is presented to prune redundant filters in detection models, which can largely reducing the model size of the network while maintain the network accuracy. Firstly, we evaluate the importance of individual filters globally across all the network layers by calculating the sum of its absolute weights. Then, we propose a progressive sparsity control function to gradually reduce the proportion of pruning in the implementation process. Finally, we finetune the pruned network to restore its accuracy. The proposed gradual global pruning is evaluated on the Pascal VOC2007 dataset and implemented on the widely-used SSD detection framework which is constructed on VGG-16 backbone network. The experimental results show that the proposed method can drastically reduce the model size of the network with negligible loss in detection accuracy. Comparing to layer-size pruning and equal proportion pruning, the proposed gradual global pruning achieves better performance in detection accuracy and pruning efficiency.

978-1-7281-1223-7/19/$31.00 ©2019 IEEE

Dalian, China
March 29-31, 2019

II. Overview of the Proposed Pruning Method

A. General process of network pruning

Network pruning usually follows the three-step principle of "training-pruning-finetune". Firstly, the initialized network is trained to provide the basis for subsequent network pruning. Secondly, according to the specific rules, the importance scores of the neurons is sorted and the neurons whose importance degree is lower than the threshold are removed and shown in Figure 1. The last step is to fine-tune the pruned network. In general, pruning will cause the network parameters to get rid of the local optimum, resulting in degradation of network performance. Therefore, it is necessary to retrain the network to recover its accuracy. Network pruning is actually an iterative process in which the repetition of pruning and finetuning alternates until the desired model is achieved. In addition, in the process of pruning, it is necessary to properly control the proportion of each step of pruning. If too many neurons are removed at one time, the model may be seriously damaged, so that the original accuracy of the model cannot be restored by fine-tuning.

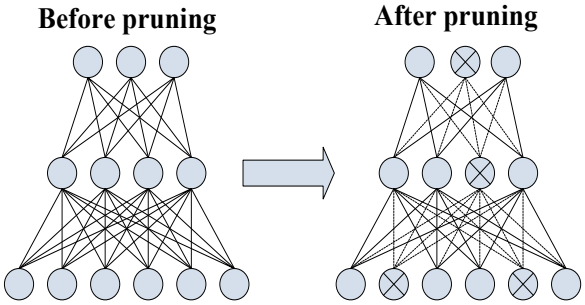

Figure 1. An illustration of the network pruning

B. Importance evaluation for neurons

The importance evaluation is the core problem for network pruning, which score each neuron in the network to reflect the influence of the current neurons on the network performance. Generally, neurons with small importance scores will be removed. At present, there are two main ways to evaluate the importance of neurons, namely correlation-based evaluation criteria and weight-based evaluation criteria. The correlation-based criteria judges the importance of neurons by calculating the correlation between them. Strong correlation indicates that the activation of the neurons in the previous layer has a greater decisive effect on the activation of current neurons. The weight-based criterion judges the importance by calculating the weight of the neuron, and the weight is considered to be proportional to the influence of the neuron on the network. Generally, the weight-based criterion do not need to calculate the response of each layer in each step of the reduction process, and the calculation amount is relatively small, which is more suitable for the transfer learning task. Therefore, we uses the weight-based criterion to judge the importance of neurons.

The weight-based criterion usually adopts two regularization methods of L1-norm and L2-norm, which present no noticeable difference in our experiments. Therefore, the L1-norm with less computation cost is adopted in this paper. Thus, for a convolution filter, its importance score is defined as Equation (1).

$$Score(l,i) = \frac{1}{n_c \times n_m \times n_n} \sum \left| F_i^l \right| \tag{1}$$

Where $n_c \times n_m \times n_n$ indicates total number of convolution kernels, F_i^l indicates the *I-th* convolution kernel in the *L-th* layer.

C. Gradually global network pruning

At present, most of the network pruning methods adopt the layer-by-layer fixed manner, in which only one layer of neurons is pruned in each round. This layer-wise manner can adapt the network to the new input distribution gradually through multiple rounds of fine-tuning, and has little influence on the network performance. However, it also has many shortcomings: firstly, for a deep convolution network, the finetuning after each pruning round will bring huge computation and time consumption; secondly, each convolution layer in the network has different redundancy, which makes it difficult to determine the pruning proportion for each layer in order to achieve a good performance. In order to solve this problem, a gradually global pruning scheme is proposed. In each round of pruning, the importance of all neurons is uniformly evaluated, and then the neurons that have the least impact on network performance are removed and fine-tune the pruned network iteratively. Compared with the layer-wise manner, the global pruning scheme has stronger adaptability and the implementation process is more efficient.

For the pruning ratio of each round, a progressive sparsity control function is designed as Equation (2).

$$s_t = s_f + (s_i - s_f)(1 - \frac{t}{n})^3 \tag{2}$$

where s_t indicates the sparsity of the network after the *T-round* pruning, $s_i = 0$ indicates the initial sparsity of the network, s_f indicates the predetermined target sparsity, and n indicates the total number of pruning round. Then the pruning ratio corresponding to the *T-th* round is expressed as Equation (3).

$$r_t = s_t - s_{t-1} \tag{3}$$

With the sparsity function, we can flexibly control the pruning proportion of each round of the network. In the initial stage of pruning, the network also contains a large number of redundant parameters, so we can prune the network with a large proportion. As the number of neurons in the model decreases gradually, the proportion of pruning also needs to be reduced accordingly. In the experiment, it is found that this pruning scheme is beneficial to maintain the accuracy of the network after fine-tuning.

In addition, it should be noted that the weight-based evaluation criteria mentioned above cannot be directly applied to global pruning, which is mainly due to the systematic deviation of contributions of different layers.

Therefore, it is necessary to eliminate the importance bias of different layers. In this paper, a simple balance method is adopted, which divides the scores of neurons in each layer directly by the mean of that layer to carry out the contribution depolarization. The adjusted importance evaluation method is defined as Equation (4).

$$Score_{modified}(l,i) = \frac{Score(l,i)}{\frac{1}{N_l}\sum_{i=0}^{N_l} Score(l,i)} \tag{4}$$

D. Implementation process

Table I shows the implementation process of the proposed pruning method. Compared to the layer-wise manner, which only performs on a certain layer at a time, the global pruning manner proposed can remove the redundant neurons in the whole network, so it can greatly reduce the number of fine-tuning needed for network pruning, and also help to recover the performance degraded by network pruning. In addition, the global pruning manner does not need to specify the proportion of each layer that needs to be removed in a certain round. Under a given network performance indicator, this scheme can automatically approach the optimal structure of the network through repeated global filtering, pruning, and fine-tuning.

TABLE I. PROCEDURE OF THE GRADUALLY GLOBAL PRUNING SCHEME

1: Evaluate the current model on test set
2: Calculate the importance score for all the neurons
3: Sort the filters across the whole network by their importance scores
4: According to the sort, remove filters with scores below the thresholds and their corresponding feature maps.
5: Retrain the pruned network to recover its performance degraded by network pruning.
6: Go to step 1 and continue the pruning procedure until the target performance is obtained.

III. EXPERIMENTS

In this section, the classical one-stage detection model SSD is used as the pruned network to verify the effectiveness of the proposed pruning method. We set three different target pruning proportion, 25%, 50%, 75% and compare the gradually global pruning scheme with layer-wise and equal-proportion pruning scheme. We train the network on the union of Pascal VOC 2007 trainval and VOC 2012 trainval datasets, and evaluate on the Pascal VOC2007 test set. Throughout training we use a batch size of 32, a momentum of 0.9 and a decay of 0.0005. What's more, we initialize the learning rate 10^{-3} and train for 10 epoch, and then decay it to 10^{-4} for the next 10 epoch training. All experiments were implemented on a PC with Intel core E5-2650 CPU, a NVIDIA TITIAN X GPU. The PC operating system was Ubuntu 16.04. All the experiments are conducted within Caffe [12] platform.

Table II shows the performance of the proposed gradually global pruning method corresponding to three different pruning proportions with 10 rounds pruning. It can be observed that the accuracy of the model is well maintained. Even when the target pruning proportion is 75%, the mAP loss of the model is only 3.1%, while the model parameters decrease 86.1%. Compared to the layer-wise scheme, the global pruning scheme achieves better accuracy for each target pruning proportion, and the pruning process is also more concise.

TABLE II. PRUNING RESULTS OF SSD ON VOC2007

Method	Target pruning proportion	Pruning round	mAP	Model Size (Million Parameters)
SSD	0	0	74.3%	33.1
Layer-wise	25%	13	73.5%	18.6
	50%	13	72.4%	10.3
	75%	13	70.9%	4.8
Gradually global pruning	25%	10	73.9%	18.2
	50%	10	72.8%	10.7
	75%	10	71.2%	4.6

Fig. 2 shows the change of detection accuracy when the target pruning ratio is 75%, using the proposed pruning method and using the the equal-proportion pruning method respectively. Equal proportion pruning, as the name implies, uses the same pruning proportion in each round. As shown in Fig. 2, the equal proportion pruning achieves a high detection accuracy at the beginning of pruning. However, as the model size continues to decrease, the proportion of the removed neurons in the current model becomes larger and larger, resulting in a rapid decline in mAP. Finally, the proposed pruning method achieves a detection accuracy of 71.2%, outperforming equal-proportion pruning by 3.6%. These results demonstrate the effectiveness and superiority of the proposed pruning method.

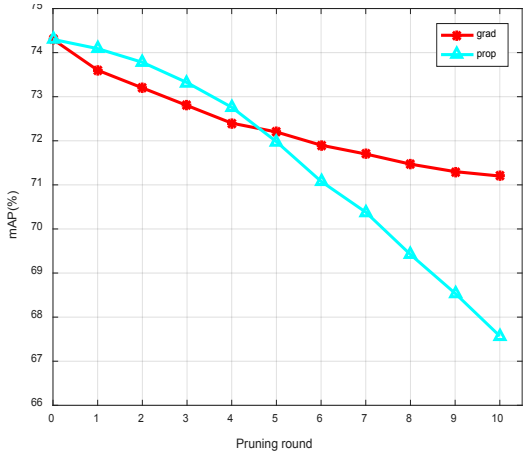

Figure 2. mAP with dirrerent pruning ratio

IV. CONLCUSION

In this paper, a novel gradual global pruning is presented to remove the redundant filters for object detection models, which can largely reducing the model size the network while maintain the network accuracy. Firstly, we evaluate the importance of individual filter globally across all network layers by calculating the sum of its absolute weights. Then, we propose a progressive sparsity control function to gradually reduce the proportion of pruning in the implementation process. Finally, we finetune the pruned network to restore its accuracy. The experimental results on Pascal VOC2007 dataset show that the proposed method can drastically reduce the model parameters of the network with negligible loss in detection accuracy. Comparing to the layer-by-layer fixed pruning and equal proportion pruning, the proposed gradual global pruning achieves better performance in detection accuracy and implementation efficiency.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, G. Hinton. ImageNet classification with deep convolutional neural networks// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.

[2] S. Ren, K. He, R. Girshick. Faster R-CNN: Towards real-time object detection with region proposal networks// Proceedings of the 2015 Advances in Neural Information Processing Systems (NIPS2015). CANADA: MIT Press, 2015: 91-99.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[4] W. Liu, D. Anguelov, D. Erhan. SSD: single shot multibox detector// Proceedings of 2016 European Conference on Computer Vision. Berlin: Springer press, 2016:21-37.

[5] M. Everingham, G. Van, C. K. Williams, et al. The pascal visual object classes (VOC) challenge. International Journal of Computer Vision, 2010, 88(2): 303-338.

[6] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? the kitti vision benchmark suite, CVPR(2017), pp. 3354-3361.

[7] K. Simonyan, A. Zisserman. Very deep convolutional networks for large-scale image recognition[J]. Computer Science, 2014.

[8] K. He, X. Zhang, S. Ren. Deep residual learning for image recognition// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2016:770-778.

[9] Y. LeCun, J. S. Denker, and S. A. Solla. Optimal braindamage. In NIPS, pages 598–605, 1990.

[10] S. Han, J. Pool, J. Tran. Learning both weights and connections for efficient neural networks. //Advances in Neural Information Processing Systems, Massachusetts: MIT Press, 2015: 1135-1143.

[11] H. Li, A. Kadav, I. Durdanovic. Pruning filters for efficient convnets. arXiv preprint arXiv:1608.08710, 2016.

[12] Y. Jia, E. Shelhamer, J. Donahue. Caffe: Convolutional Architecture for Fast Feature Embedding, in Proceedings of the 22nd ACM International Conference on Multimedia, pp. 675–678.