SCIENCE CHINA
Information Sciences

• LETTER •

December 2022, Vol. 65 229101:1–229101:2
https://doi.org/10.1007/s11432-020-3162-4

Knowledge transferred adaptive filter pruning for CNN compression and acceleration

Lihua GUO*, Dawu CHEN & Kui JIA

School of Electronic and information Engineering, South China University of Technology, Guangzhou 510641*, China*

Received 16 April 2020/Revised 14 June 2020/Accepted 29 September 2020/Published online 14 March 2022

Citation Guo L H, Chen D W, Jia K. Knowledge transferred adaptive filter pruning for CNN compression and acceleration. Sci China Inf Sci, 2022, 65(12): 229101, https://doi.org/10.1007/s11432-020-3162-4

Dear editor,

It is necessary to compress the convolutional neural network (CNN) for resource-constrained devices deployment, resulting in accelerated model training and inference. "Soft" filter pruning (SFP) [1] conducts dynamic pruning operations during training time, allowing updates of the pruned filters to compensate for any possible misoperation. However, the pruning exhibits less efficiency owing to the lack of supervision provided regarding the filters' importance for specific learning tasks. To provide this supervision, we have made use of the knowledge transfer based on network compression techniques. Filter importance concerning specific learning tasks could be gained from a high-performance teacher network. We have noted that previous methods merge the strategy of distilling the task-specific knowledge of the teacher network directly into a lightweight student network. In this study, we employ the normalization statistics of feature maps (NSFM) as the transferable knowledge representation for soft pruning of network filters, and this method is termed as transferred adaptive filter pruning (TransFP), where the adaptivity of filter pruning results from the selection of the adaptive filter according to the importance of the mask of each filter.

Model and methodology. In TransFP, a deep CNN should be first trained as a teacher network, and then the knowledge information is gained from the teacher network. Simultaneously, a small and compact CNN is designed as a student network. The student network will be further compressed and accelerated by a soft filter pruning under the guidance of the knowledge information obtained from the teacher network. During soft pruning, a weight mask is designed for each filter channel, which indicates the importance of each filter channel based on a global pruning rate. Mask weight is specially designed to implement the structure filter pruning and it indicates the importance of filter channel during filter pruning. The smaller is the value of the weight mask, the less important are the filters, which will be pruned.

For deep CNN networks, the convolutional operation layer can be expressed by $Y = F(X, W)$, where F represents the nonlinear mapping function between the input X and output Y. In CNN, the common non-linear mapping function is RELU. For defining the importance of output channels, a weight mask is assigned for each filter channel, which is multiplied by the nonlinear mapping function F. Besides, M is defined as the vector of a weight mask. The operation becomes $Y = M \bigotimes F(X, W)$, where $\bigotimes$ refers to the element-wise multiplication operation between the weight mask of each channel and the output of the nonlinear mapping function. In addition, the weight masks in the network are learnable parameters, which can be trained jointly with the network weights. We can manipulate the four-step pruning strategy with the weight mask as follows.

(1) Filter selection: after a certain period (e.g., one training epoch), the importance of each channel is evaluated using the absolute value of their corresponding weight mask, and those channels are identified using a small weight mask under a given pruning rate.

(2) Filter pruning: after the selection process, the values of selected filters are set to zero to achieve the same effect as removing the filters from the network.

(3) Reconstruction: to recover the capacity and further improve the performance of the model, the pruned model is reconstructed by allowing the filters with zero weight mask to be updated in the next training period.

(4) Compact model creation: the model converges after iterating over the filter selection, i.e., filter pruning and reconstruction steps, the filters with a zero weight mask and the corresponding feature maps from the model are removed to obtain a real compact model.

If the knowledge can be distilled from the teacher network and can be transferred to the student network, it would be very helpful to train the student network and improve its performance. The knowledge is usually gained from a high-performance teacher network to help the channel selection of the student network. In this method, the teacher network should distill the knowledge to preserve the discriminative representation while the student network can retain powerful recognition performance similar to the teacher network. From our observation, the normalization statistics from each instance can be distilled as the knowledge because of their discriminative information. Furthermore, the

* Corresponding author (email: guolihua@scut.edu.cn)

first and second-order statistics could be extracted and these normalization statistics can be formulated as follows:

$$\begin{aligned}\mu_{li} &= \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} a_{lihw},\\ \sigma_{li}^2 &= \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}(a_{lihw}-\mu_{li})^2,\end{aligned} \tag{1}$$

where a_{lihw} is the ith feature map, μ_{li} is the mean and σ_{li} is the standard deviation of the ith feature map. H and W are the height and width of feature map. These statistical values of all maps are respectively concatenated to form feature vectors as follows:

$$\begin{aligned}\mu_l &= (\mu_{l1}, \mu_{l2}, \ldots, \mu_{lC}),\\ \sigma_l &= (\sigma_{l1}, \sigma_{l2}, \ldots, \sigma_{lC}).\end{aligned} \tag{2}$$

These two feature vectors are regarded as knowledge. Without loss of generality, transfer losses are positioned between the student and teacher layers with same number of channels. Let s and t denote student and teacher networks with weights of w_s and w_t, respectively. The new total loss combines both the cross-entropy loss and knowledge difference loss, which is defined as follows:

$$\begin{aligned}L_{\text{total}} = {} & L_{\text{CE}}(f(x, w^s, M), y)\\ & + \frac{\lambda}{2}\frac{1}{L}\sum_{l\in L}\left(\left\|M_l u_l^s - u_l^t\right\|_2 - \left\|M_l \sigma_l^s - \sigma_l^t\right\|_2\right),\end{aligned} \tag{3}$$

where $L_{\text{CE}}(f(x, w^s, M), y)$ denotes the cross entropy loss (where M refers to the mask weight), L is the number of layers, and (u_l^s, u_l^t), (σ_l^s, σ_l^t) are the pairs of the mean vector and the standard deviation vector from the student and the teacher, respectively. In the final loss function, the weight mask is modeled not only into the crossing entropy loss but into the knowledge difference loss as well, which is jointly trained with the network parameters.

Conclusion and future work. In this study, a knowledge-transferred adaptive soft-pruning method exploiting the normalization statistics is proposed to compress and accelerate deep CNNs. Simultaneously, a teacher-student learning mechanism is applied to guide the soft pruning process. According to this mechanism, the knowledge is distilled from the teacher network based on the normalization statistics and then transferred to supervise the pruning of the student network filters. Experimental results have revealed that this method can achieve competitive performance compared to other existing methods. The main advantage of this method is that it can achieve smaller drops in classification accuracies at the same rates of compression. Furthermore, this method could be combined with other acceleration algorithms, e.g., matrix decomposition and low precision weights, to further improve the performance.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 61771201) and Guangzhou Science and Technology Projects (Grant No. 201707010141).

References

1 He Y, Kang G L, Dong X Y, et al. Soft filter pruning for accelerating deep convolutional neural networks. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018. 2234–2240