A Deep Neural Network Compression Algorithm Based on Knowledge Transfer for Edge Device

Chao Li
Institute of Computing Technology
Chinese Academy of Sciences
BeiJing, China
lichao@ict.ac.cn

Xiaolong Ma
Institute of Computing Technology
Chinese Academy of Sciences
BeiJing, China
maxiaolong@ict.ac.cn

Zhulin An
Institute of Computing Technology
Chinese Academy of Sciences
BeiJing, China
anzhulin@ict.ac.cn

Yongjun Xu
Institute of Computing Technology
Chinese Academy of Sciences
BeiJing, China
xyj@ict.ac.cn

***Abstract*—Poster: The computation and storage capacity of the edge device are limited, which seriously restrict the application of deep neural network in the device. Toward to the intelligent application of the edge device, we introduce the deep neural network compression algorithm based on knowledge transfer, a three-stage pipeline: lightweight, multi-level knowledge transfer and pruning that reduce the network depth, parameter and operation complexity of the deep learning neural networks. We lighten the neural networks by using a global average pooling layer instead of a fully connected layer and replacing a standard convolution with separable convolutions. Next, the multi-level knowledge transfer minimizes the difference between the output of the "student network" and the "teacher network" in the middle and logits layer, increasing the supervised information when training the "student network". Lastly, we prune the network by cuts off the unimportant convolution kernels with a global iterative pruning strategy.**

Keywords—edge device, deep learning, neural network compression, knowledge transfer

I. INTRODUCTION

With the advancement of Edge Computing and IoT (Internet of Things) technologies, there are more and more edge devices, and they are becoming more and more intelligent[1][2]. Deep Learning is a state-of-the-art technology in the field of artificial intelligence in recent years[3]. It has achieved breakthrough achievements in many fields such as computer vision and natural language processing. In recent years, more and more deep neural network models have been proposed, such as AlexNet[4], VGGNet[5], GoogLeNet[6] and ResNet[7], which are becoming more and more accurate in target recognition and getting deeper and deeper. The deep neural network has the characteristics of deep layers and large parameters, which will bring huge storage cost and computational overhead. However, the computing power and storage capacity of the devices are limited, which seriously restricts the application of deep neural networks to the edge devices. TABLE I shows the depth, parameters and computation amount of some classical neural networks. We can see that the computation and storage cost of deep neural networks are very expensive. In addition, Han et al[8] show that deep neural networks running on edge devices will increase memory access and consume a lot of battery power. Although cloud computing can run deep neural networks at the cloud and provide cloud services to the edge devices, some intelligent applications may encounter high network latency, high power consumption, and low security.

TABLE I. DEEP NERUAL NETWORK PARAMETERS

Deep neural network	Depth	Size (MB)	Computation times (Millions)	Parameters (Millions)
AlexNet	8	200	720	60
VGG-16	16	~550	15300	138
GoogLeNet	22	~50	1550	6.8
ResNet	101	~170	11300	42

Since the deep neural network cannot run directly on the edge devices and the cloud computing service is not apply for edge devices, then we consider to compress the deep neural network and run it at the edge devices. Deep neural network compression can effectively reduce the parameter amount and reduce its computational and storage overhead while the performance of the network model is slightly reduced[9][10]. The knowledge transfer method is a mainstream approach compressing the deep neural network and it can realize neural network compression without additional runtime library or special hardware support 错误! 未找到引用源。. As shown in the Fig.1, we propose the deep neural network compression algorithm based on

knowledge transfer, a three stage pipeline: lightweight, multi-level knowledge transfer and pruning.

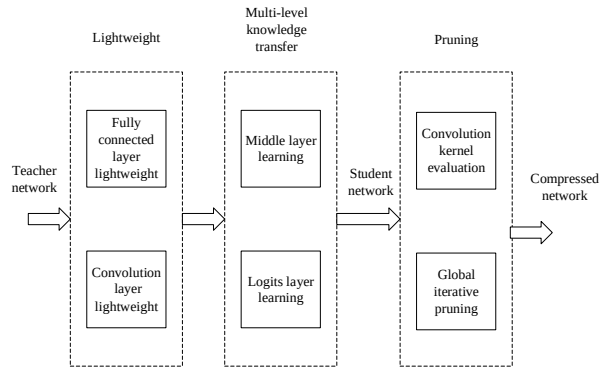

Fig. 1. The three stage compression pipeline of the deep neural network compression algorithm based on knowledge transfer

Lightweight is designed from the perspective of both the fully connected layer and the convolution layer. They can greatly reduce the storage and computational cost of deep neural network without reducing the classification accuracy. Reducing storage cost requires reducing the amount of parameters in fully connected layer, while reducing computational cost requires reducing the amount of computation of convolution layer. Lightweight designs a compact “student network” structure that uses a global average pooling layer instead of a fully connected layer and replaces a standard convolution with separable convolutions.

In order to solve the problem of the lack of supervised information in the knowledge transfer method, multi-level knowledge transfer method increases the effective supervised information during the training process, so that the classification accuracy of “student network” can be improved. The method of multi-layer knowledge transfer uses a combination of the middle layer learning and logits layer learning. The whole framework is shown in Fig.3. The middle layer refers to the hidden layer before the logits layer of the deep neural network. The output of these layers is the feature map learned by the neural network. The logits layer is the output of the layer before softmax activation. Multi-level knowledge transfer learns network changes in the middle layers and adds the soft target with the random factors to the logits layer to increase the supervised information in the process of training and improve the classification accuracy of the “student network” model trained, reduce the loss of classification performance due to compression of neural networks.

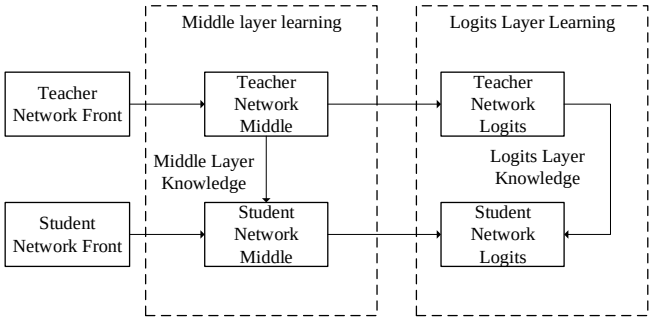

Fig. 2. Knowledge transfer learning framework

In the pruning stage, the ”student network” model compressed by the lightweight and multi-level knowledge transfer method is used as a pruning model, and the data-independent convolution kernel evaluation method is used to evaluate the importance of each convolution kernel in the pruning model, using a channel-based pruning. The pruning method cuts off the unimportant convolution kernels in the pruning model and uses a global iterative pruning strategy. Deep neural network pruning usually involves three steps: evaluating convolution kernels, cutting convolution kernels and fine-tuning.

REFERENCES

[1] Shi W, Cao J, Zhang Q, et al. Edge Computing: Vision and Challenges[J]. IEEE Internet of Things Journal, 2016, 3(5):637-646.

[2] Shi W, Dustdar S. The Promise of Edge Computing[J]. Computer, 2016, 49(5):78-81.

[3] Lecun Y, Bengio Y, Hinton G. Deep learning.[J]. Nature, 2015, 521(7553):436.

[4] A.Krizhevsky, I.Sutskever, G.E.Hinton. Imagenet classification with deep convolutional neural networks[J]. , 2012, 1097-1105

[5] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. international conference on learning representations, 2015.

[6] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[J]. computer vision and pattern recognition, 2015: 1-9.

[7] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[J]. computer vision and pattern recognition, 2016: 770-778.

[8] Han S, Mao H, Dally W J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding[J]. arXiv preprint arXiv:1510.00149, 2015.

[9] Anwar S，Sung WY．Coarse pruning of convolutional neural networks with random masks[C]．IEEE Conference on Learning and Representation(ICLR)，2017，134-145.

[10] Denil M, Shakibi B, Dinh L, et al. Predicting Parameters in Deep Learning[J]. neural information processing systems, 2013: 2148-2156.