Integrating knowledge distillation of multiple strategies

Jindong Min[a], Mingxia Wang[*b]
[a]School of Computer Engineering and Science, Shanghai University, Shanghai, China; [b]Qingdao Etsong Technology Co., Ltd, Qingdao, Shandong, China

ABSTRACT

With the increase of the complexity of the real visual target detection task and the improvement of the recognition accuracy, all aspects of the target detection network model have also become more complex and huge, causing difficult deployment and poor inference timelines. In this paper, knowledge distillation is used to compress the huge and complex deep neural network model to another lightweight network model. Different from traditional knowledge distillation methods, we propose a novel knowledge distillation that incorporates multi-faceted features, called M-KD. In this paper, when training and optimizing the deep neural network model for target detection, the diverse knowledge of the teacher network is transferred to the student network. Moreover, we also introduce an intermediate an intermediate guidance layer between teacher network and the student network to make up for the huge difference between them. Finally, this paper adds an exploration module to the traditional knowledge distillation teacher-student network model. Lastly, comprehensive experiments in this paper using different distillation parameter configurations demonstrate that our proposed new network model achieves substantial improvements in speed and accuracy performance.

Keywords: Object detection, knowledge distillation, convolutional network, model compression

1. INTRODUCTION

The knowledge distillation model utilizes a large amount of knowledge contained in the large and complicated teacher network model to guide the optimization of the simple student network model with fewer parameters, and improve the accuracy of the student network model. However, many research methods on knowledge distillation are based on the output of the teacher network directly, causing knowledge hidden in the teacher network not fully utilized. Moreover, some studies output a student network by integrating many teacher models, which disregard differences between them.

In order to resolve the problems above, we propose a multi-step knowledge distillation model. We add multiple intermediate guidance layer networks between the teacher network and the student network, reducing the gap between them and preventing error iteration considerably. Additionally, inspired by the deep neural network feature attention map and the relationship between the layers of the network channel[1,2], we allow students to learn the feature attention map and layers' relation of teacher model to acquire more knowledge.

In this paper, our new knowledge distillation model, M-KD, integrates the knowledge of the soft probability soft target output of the teacher network, the relationship between the layers of the deep neural network and the feature attention map of the hidden layer of the network. To learn and discover some new knowledge features, we divide the student network model into two parts: inheritance and exploration. Experiments show that these two parts together greatly improve the diversification and multi-feature of student network model learning.

2. RELATED WORKS

This paper studies adopting knowledge distillation to the deep convolutional neural network model in object detection. The core idea of the knowledge distillation model is to use a large deep network with better performance as the teacher network to guide the learning of a pure student network, so that the performance of the student network is close to or even better than that of the teacher network.

Knowledge distillation is a relatively new research topic in the field of computer vision. Many achievements have been made in model compression. By mining the knowledge in the teacher network, we distill this rich knowledge into the student network, so that the student network and the teacher network are consistent. Hinton et al.[3] first formally put

* sharon7326@sina.com

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125064M
 · 0277-786X · doi: 10.1117/12.2662513

forward the concept of knowledge distillation in the computer vision neighborhood, taking the soft target of the output result of the teacher network as the target of the student network, training the student network and improving the performance of the student network. In order to learn the feedback information in the student network in traditional distillation process, Zhang et al.[4] proposed a deep mutual learning model. Multiple student networks are trained at the same time, and they learn from each other through the ground truth and the output results of multiple networks to make progress together. Zagoruyko et al.[5] noticed that the feature map in the convolutional network plays a great role in the final image classification, and transferred the feature map of the teacher network to make the feature map of the student network and the feature map of the teacher network as similar as possible. Yim et al.[6] used the FSP matrix to measure the relationship between the features of two layers, obtained by multiplying and summing the feature map of the previous layer with the feature map of the next layer. Ji et al.[7] proposed a new feature self-distillation method. This method does not establish distillation between different depth feature maps, but guides the establishment of feature maps of the same level through the integration and refinement of each depth feature layer.

Based on the existing knowledge distillation methods, our proposed knowledge distillation method improves the difficult problems in the current knowledge distillation methods. The proposed method bridges the gap between teacher model and student model greatly and the proposed student network exploration module enables the student network to break through the limitation of teachers' network knowledge and bring better performance.

3. METHOD

3.1. Multi-step Intensive Guided Knowledge Distillation

In order to reduce the huge gap between the teacher network model and the student network model, as shown in Figure 1, we add a series of transition layers[8] between the teacher network model and the student network model. These transition layers are called teaching assistant layers. A method of training teaching assistants for intensive guided knowledge distillation. We teach student networks using knowledge drawn from medium-sized teaching assistants and teachers. In addition, this special form of structurally connected distillation also greatly facilitates teaching assistantships.

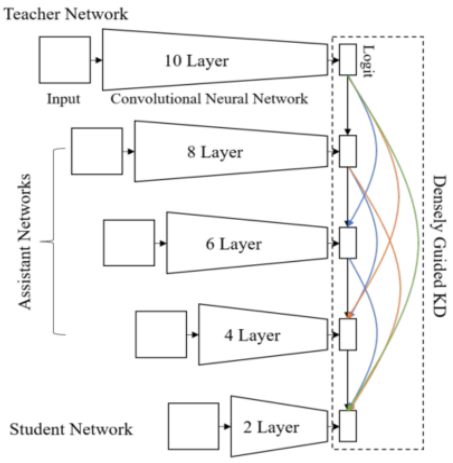

Figure 1. Multi-step intensive guided knowledge distillation.

We use several intermediate distillation losses between the original teacher network model and the student network model. Specifically, for example, there are two teacher assistant (TA) layers and with a teacher network model *T*, so that the loss can be expressed as:

$$L_{A_1} = L_{T \to A_1} \tag{1}$$

$$L_{A_2} = L_{T \to A_2} + L_{A_1 \to A_2} \tag{2}$$

The loss function of the small and lightweight student network model can be expressed as follows:

$$L_S = L_{T\to S} + L_{A_1\to S} + L_{A_2\to S} \tag{3}$$

If there are *N* intermediate teacher assistant models, the student loss function is expressed as:

$$L_S = L_{T\to S} + \sum_{i=1}^{N} L_{A_i\to S} \tag{4}$$

Our multi-step knowledge distillation uses multiple intermediate transition layers, namely teacher assistant networks for knowledge distillation, which can enable the large teacher network to train the small student network well and accurately through the effective guidance and transition of multiple teaching assistant networks in the middle layers.

3.2 Inherit and explore distillation network models

The student network in our proposed knowledge distillation model is divided into two new and different modules: inheritance and exploration, which is different from the traditional student network model that only consists of one part and its structure is shown in Figure 2.

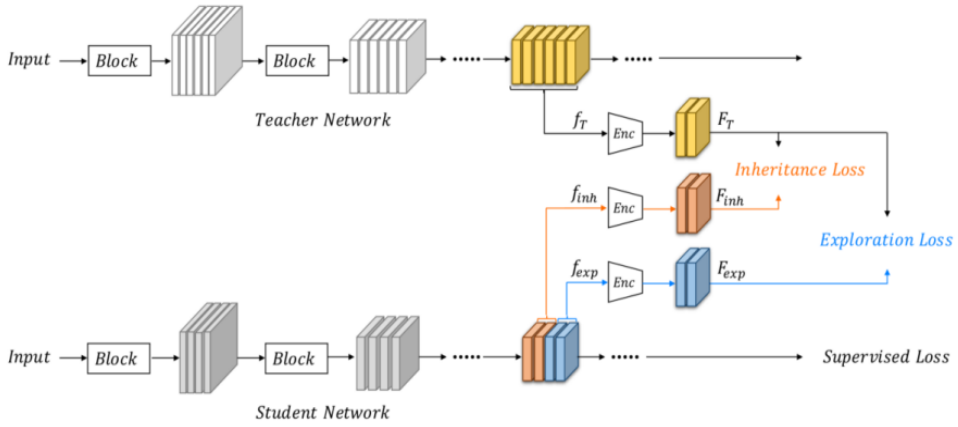

Figure 2. Inherit and explore frameworks.

In the above knowledge distillation model, we use f_T to represent the feature knowledges of the teacher network, and represent the features of the inherited part and the exploration part of the student network as f_{inh} and F_{exp}, respectively. Since the information such as the size and shape of these features may be different, it is difficult for us to measure the similarity and difference between these features. In order to solve the inconsistent feature size of teacher-student network models, we try to use a feature encoder to integrate the features of teacher-student network into a shared latent feature space of the same dimension to achieve the same feature shape and size of teacher and student networks. The integrated features are denoted by F_T, F_{inh} and F_{exp}, respectively. Through the feature encoder method, we extract the corresponding feature knowledge from the convolutional blocks of specific channels of the teacher neural network model. The loss function of the auto-encoders used during training can be express as follows:

$$L_{rec} = ||f_T - R(f_T)||^2 \tag{5}$$

where the feature map of the teacher network is denoted by f_T and the output of the auto-encoder is denoted by $R(f_T)$. The inheritance loss function L_{inh} inherits the existing feature knowledge hidden in the teacher model by minimizing the difference between it and the teacher feature F_T, and is expressed as:

$$L_{inh} = || \frac{F_{inh}}{\|F_{inh}\|_2} - \frac{F_T}{\|F_T\|_2} ||_1 \tag{6}$$

L_{exp} exploration function is designed to be just the opposite of the L_{inh} inheritance function, and L_{exp} learns a representation that is different from inheritance. The representation of the exploration function is similar to the inherited loss function above, an obvious approach is to minimize the negative difference between and F_T:

$$L_{exp} = -|| \frac{F_{exp}}{\|F_{exp}\|_2} - \frac{F_T}{\|F_T\|_2} ||_1 \tag{7}$$

We clearly find that the sign change of the exploration loss L_{exp} is not simply determined by pushing F_{exp} to learn a

negative teacher factor $-F_T$, which obviously correlates with F_T. The purpose of the L_{exp} loss function is to make the exploration part focus on the different regions of the image.

4. EXPERIMENTS

4.1 Experiment details

In various experiments on multiple datasets in this paper, we use commonly used image detection performance evaluation metrics such as AP, AP_{60}, AP_S and FPS as evaluation metrics. The main backbone network models for image object detection in this paper include YOLO, Fast RCNN, SSD, and RefineDet[9,10]. Our proposed new knowledge distillation model is compared and tested on the improved RCNN network Mask RCNN and Cascade Mask RCNN. In the experiment, we specified the backbone network of each target detection network as ResNet32 and ResNet64. In the experiment, we first pre-trained the backbone network model of the new knowledge distillation architecture on the large deep learning dataset ImageNet, and then the data A series of fine-tuning is performed on the results of pre-trained neural network models or various parameters on CIFAR-100. We use MBGD and set the initial learning rate to 0.2, momentum to 0.9 and minibatch size to 34. The distillation temperature τ decrease from 5 to 1. All experiments in this paper are based on the deep learning platform TensorFlow on a GeForce RTX 3090.

4.2 Experiment results

Compared with the latest three current knowledge distillation model methods[11-13], the new knowledge distillation method M-KD proposed in this paper, our proposed target detection knowledge distillation architecture model has been significantly improved in the performance of AP indicators, and the real-time performance of the model. In order to guarantee, it is easier to deploy edge devices. In various ablation experiments based on the CIFAR-100 dataset, the target detection models of YOLO, Fast RCNN, SSD and RefineDet compared the latest knowledge distillation models, and the AP indicators were improved by 1.6, 1.2, 1.2 and 1.4 respectively. On the classic instance segmentation model Mask RCNN, the AP metric is improved by 1.9. In various comparative experiments, we clearly observe AP improvements of 1.2 and 1.0 on the base ResNet32 and ResNet64 backbone object detection models, respectively. In addition to the improvement in the AP target detection index, the FSP, the number of pictures that can be processed per second, has also been improved a lot, and the detection speed of the model is faster. This is a good illustration of applying our proposed knowledge distillation model M-KD to these deeper and more complex deep learning target detection network models, and the performance improvement is even more improved.

5. CONCLUSIONS

In this paper, our proposed model integrates various knowledge, and the dense guidance layer in the middle can effectively avoid wrong iterations and improve the student network's performance. Compared with other state-of-the-art knowledge distillation methods, this novel knowledge distillation architecture model proposed in this paper is greatly improved and advanced. Furthermore, it shows that the model exhibits continuous optimization in various comparative experiments and various deep learning dataset settings, improving object detection models faster and yielding higher levels of overall performance. The comprehensive knowledge distillation method proposed in this paper can also be applied to deep learning fields such as small sample learning and natural language processing, and has good results in other fields.

REFERENCES

[1] He, K., Zhang, X., Ren, S., et al., "Deep residual learning for image recognition," IEEE Conf. on Computer Vision and Pattern Recognition, 770-778 (2016).

[2] Cai, Z. and Vasconcelos, N., "Cascade R-CNN: High quality object detection and instance segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence 43(5), 1483-1498 (2019).

[3] Hinton, G., Vinyals, O. and Dean, J., "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, (2015).

[4] Zhang, Y., Xiang, T., Hospedales, T. M., et al., "Deep mutual learning," IEEE Conf. on Computer Vision and Pattern Recognition, 4320-4328 (2018).

[5] Zagoruyko, S. and Komodakis, N., "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," arXiv preprint arXiv:1612.03928, (2016).

[6] Yim, J., Joo, D., Bae, J., et al., "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," IEEE Conf. on Computer Vision and Pattern Recognition, 4133-4141 (2017).
[7] Ji, M., Shin, S., Hwang, S., et al., "Refine myself by teaching myself: Feature refinement via self-knowledge distillation," IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 10664-10673 (2021).
[8] Son, W., Na, J., Choi, J., et al., "Densely guided knowledge distillation using multiple teacher assistants," IEEE/CVF Inter. Conf. on Computer Vision, 9395-9404 (2021).
[9] Huang, Z., Shen, X., Xing, J., et al., "Revisiting knowledge distillation: An inheritance and exploration framework," IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 3579-3588 (2021).
[10] He, K., Gkioxari, G., Dollár, P., et al., "Mask R-CNN," IEEE Inter. Conf. on Computer Vision, 2961-2969 (2017).
[11] Zhang, L. and Ma, K., "Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors," Inter. Conf. on Learning Representations, (2020).
[12] Wang, T., Yuan, L., Zhang, X., et al., "Distilling object detectors with fine-grained feature imitation," IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 4933-4942 (2019).
[13] Heo, B., Kim, J., Yun, S., et al., "A comprehensive overhaul of feature distillation," IEEE/CVF Inter. Conf. on Computer Vision, 1921-1930 (2019).