**Knowledge Transfer for Object Detection**

Chenggui Gong	extsuperscript{a}, Xiao Zhang

Beijing Jiaotong University, Beijing, China

	extsuperscript{a}17120224@bjtu.edu.cn

**Abstract.** Object detection is one of the basic tasks of computer vision. Although deep neural networks have greatly promoted the development of target detection, the training of deep neural networks often requires certain computing resources, especially when dealing with more complex detection tasks. It will become more complicated, and the increase in model complexity will increase its demand for computing resources. This is not desirable for platforms with limited computing resources. Without sacrificing excessive detection accuracy, how to make the model run efficiently on a platform with limited resources has become a big challenge. In order to solve this problem, we have introduced the idea of knowledge distillation in this work. We use the knowledge learned by complex models as prior knowledge, and pass the prior knowledge to the small-scale neural network model, so that the small-scale neural network model can acquire the similar functions of large-scale networks. Based on this, we proposed a framework for target detection based on knowledge distillation and conducted experiments on PASCAL VOC [1] and KITTI [2]. The experimental results show that the proposed framework can make the smaller student network show similar performance to the larger teacher network. This provides a possibility for the deep learning system to run efficiently on a resource-limited platform.

1. Introduction

For human beings, vision is an important way of perception. Most of the external information is perceived by vision. Humans and animals perceive the external environment through vision to obtain various information that is important to the survival of the body. It is of great research significance to let computers perceive and understand the environment like human beings. The improvement of hardware performance, the continuous optimization of intelligent algorithms and the rise of big data have greatly promoted the development of computer vision. Target detection is one of the important components of computer vision. It is widely used in pedestrian detection, robotics and intelligent transportation.

In recent years, with the development of deep learning, deep neural networks have been widely used in target detection, semantic segmentation and other fields with their powerful feature extraction capabilities. The powerful ability of deep neural networks comes from the huge parameters in their networks. When the network becomes complicated, the parameters in the network will increase sharply. The huge network training often means the consumption of a large amount of computing resources and the cost of time. With the popularity of smart wearable devices in social life, how to effectively deploy deep learning systems on portable devices with limited resources has become a new challenge.

The acceleration and compression of the network is very beneficial for deploying the deep neural network model on a platform with limited resources. At present, the main network acceleration and compression technologies are divided into parameter pruning and sharing [3], low rank decomposition...
[4] and Knowledge distillation [5]. Parameter pruning and sharing is to remove the redundant parameters in the network and reduce the scale of the model by cutting off the secondary connections in the network or sharing parameters. Stanford University took the lead in proposing a deep compression parameter pruning method, but because this pruning method is often not effective in the hardware implementation process, structural pruning and gradient pruning are optimized. Low rank decomposition is to reduce the amount of calculation by matrix analysis of convolution calculation to compress and accelerate the network. However, the scalability of low rank decomposition is not strong. The implementation of low rank method is not easy because it involves Calculate costly decomposition operations, and the current method is to perform low-rank approximation layer by layer, unable to perform global parameter compression. In addition, decomposition also requires a lot of retraining to achieve convergence. The main idea of knowledge distillation is to make the output of its model close to the output of a larger teacher network by training a more compact student network. Although these network acceleration and compression methods have achieved good results in image classification, the effects in target detection, semantic segmentation and other complex tasks are not very prominent.

In view of the above problems, in this paper, we propose a knowledge forced distillation framework for target detection, and do a detailed experiment on multiple data sets. The experimental results prove that the framework proposed in this paper can greatly improve the training effect of student networks.
2. Related work

2.1. Object Detection
Target detection is one of the important computer vision tasks, which is the basis of computer vision tasks such as instance segmentation and target tracking. 2012 is a watershed in image feature extraction. The re-emergence of the convolutional neural network quickly squeezes the way of hand-crafted features with excellent feature extraction capabilities, and quickly sweeps across the major computer vision areas. In the field of deep learning, target detection is mainly divided into two-stage detection and one-stage detection. In the two-stage detection, RCNN [6] first proposed the area with convolutional neural network function for target detection. It extracts a set of region proposals through selective search algorithm, and rescales these region proposals to fixed size, then, the proposal is finally sent to the convolutional neural network to extract features. Although the RCNN effect is amazing, the redundant feature calculation of a large number of overlapping region proposal greatly reduces the speed of detection. SPPNet [7] proposed a spatial pyramid pooling module to generate fixed-length feature representations, enabling CNN to accept image input of any size, avoiding repeated feature calculations and improving detection speed. Fast RCNN [8] combines the advantages of RCNN and SPPNet, it greatly promotes the improvement of target detection accuracy and speed performance. Faster R-CNN [9] proposed the RPN network, which is the first end-to-end target detection framework. FPN [10] proposed a top-down pyramid structure based on Faster R-CNN to capture semantic information of various scales. FPN has shown great advantages in object detection at various scales.

Compared to the two-stage object detection which regards the process from coarse to fine, the one-stage detection is completed in one step. YOLO [11] is the first one-stage detector in the field of deep learning. It divides the image into regions, and then performs bounding box prediction and confidence calculation for each region at the same time. Although its detection speed is fast, its detection accuracy lower than two-stage detectors, especially for small target detection. SSD [12] proposed using multi-resolution detection technology to improve the detection accuracy of the one-stage detector. RetinaNet [13] explored the reasons for the decline in detection accuracy of one-stage detectors and used Focal Loss to solve the category imbalance in network training, making the network more focused on those samples that are difficult to train. In this way, RetinaNet not only has fast detection speed, but also the detection accuracy is comparable to the two-stage detector.

2.2. Knowledge Distillation
In the past few years, deep neural networks have made great progress in many visual tasks, but most models require higher computing resources, which limits their application on resource-constrained platforms. Knowledge distillation is a more effective method of network compression and acceleration. It uses a small model to fit the large model, allowing the small model to learn a function map similar to the large model. In this way, the small-scale neural network model can acquire similar functions of large-scale networks. Then the small-scale network model is transplanted on a limited resource platform. This not only saves resources but also completes tasks. [5] proposed a framework for knowledge distillation, which uses the student teacher network paradigm to guide the output of student networks to accelerate network training. [14] proposes to learn the knowledge in neural networks to solve Network compression. In order to fully learn the teacher network, FitNets uses the student network to imitate the full feature map of the teacher network. P. Luo [15] use more compact high-level neurons to represent knowledge. The work of T. Chen [16] speeds up the experimental process by transferring knowledge from the previous network to each new deeper or wider network. The FSP knowledge transfer algorithm proposed by Yim. [17] transfers the correlation between feature maps in the original network as knowledge to the student network.

3. Methods
In this section we mainly introduce the methods used in this paper and the loss function.
3.1. Attention knowledge transfer
Attention is a key aspect of our visual experience, and we need to stay focused to create a visual
representation with detail and consistency. The core idea of the Attention transfer algorithm is to improve
the performance of the convolutional neural network by shifting the Attention information, so that the
network is more focused on the information that is useful for the current task. The knowledge distillation
framework of this paper is shown in Fig 1. In this framework, we use the teacher network’s attention
information to guide the student network so that the student network can get better performance.

3.2. Loss Function
The knowledge forced distillation framework needs to learn the feature extraction part of the teacher
network and strengthen the knowledge transfer of the feature extraction part. The feature knowledge
transfer function is defined as follows:

\[
L_{feature} = \sum_{j \in L} \frac{K_S}{\|K_S\|_2} - \frac{K_T}{\|K_T\|_2}^2
\] (1)

Among them, L represents the knowledge pairs of all teachers and students we want to realize
knowledge transfer. \(K_T\) stands for teacher knowledge and \(K_S\) stands for student knowledge.

4. Experimental evaluation
In this section we mainly introduce the basic components of the network, the data sets used and the
evaluation metrics.

4.1. Basic components of the network
The teacher and student models we defined in the experiment are standard CNN architectures, including
the convolutional layer, the fully connected layer, the Relu, the dropout layer, and the softmax layer. We
choose different layers of Resnet [18] as our teacher-student model, using two different settings for
students and teachers. In the experiment, we used a smaller network as the student network and a larger
network as the teacher network.

4.2. Datasets and Evaluation metrics
We evaluated our approach on several commonly used public datasets, including KITTI, PASCAL VOC
2007. KITTI and PASCAL are relatively small datasets with fewer object categories and tag images. For
all datasets, we follow the PASCAL VOC convention and evaluate the model by mean average precision
(mAP) of IoU=0.5. We did all the experiments on the 1080Ti server.

4.3. Experimental Results
We used a network of teachers to guide the student network and experimented on different data sets. The
experimental results are shown in Table 1. Among them, S represents the student network, and T
represents the teacher network. From the experimental results, we can see that the large-scale teacher
network detection accuracy is high, and the student network detection accuracy is much lower than that
of the teacher network. However, the student network has improved its performance after being guided
by the teacher network, and its detection accuracy is almost the same as the teacher network. We all know
that small-scale networks are faster than large-scale networks, and the computing resources are not
demanding, but the detection accuracy is low. Through knowledge distillation, we can enable small-scale
networks to learn about the performance of large-scale models, and achieve the effect of model
compression acceleration.
Table 1. Detection Performance of the Proposed Method

|                | PASCAL07 | KITTI  |
|----------------|----------|--------|
| S: ResNet-18   | 71.5     | 63.3   |
| T: ResNet-101  | 75.1     | 66.7   |
| S+T: ResNet-18 + ResNet-101 | 74.6 | 65.8 |

S denotes student network, T denotes teacher network, S+T denotes student network guided by the teacher network.

5. Conclusion

We all know that deep learning models generally improve the accuracy of the network by deepening or widening the number of layers of the network. But the larger the model, the more computing resources it consumes. With the popularity of wearable devices in life. Deploying a deep learning system on a resource-constrained platform has become an urgent problem to be solved. The quality of the features is related to the accuracy of the model. In this paper, we use a large network of teachers to guide the learning of small-scale student network. Experiments show that after the student network is guided by the teacher network, the network detection accuracy has been improved. In this way, we have achieved compression and acceleration of the model.

References

[1] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]// IEEE Conference on Computer Vision & Pattern Recognition. 2012.
[2] Everingham M, Winn J. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Development Kit[J]. 2006, 111(1):98-136.
[3] Han S, Mao H, Dally W J. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding[J]. Fiber, 2015, 56(4):3-7.
[4] Zhang X, Zou J, He K, et al. Accelerating Very Deep Convolutional Networks for Classification and Detection[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, 38(10):1943-1955.
[5] Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network[J]. Computer Science, 2015, 14(7):38-39.
[6] Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, 2014.
[7] He K, Zhang X, Ren S, et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 37(9):1904-16.
[8] Girshick, Ross. Fast R-CNN[J]. Computer Science, 2015.
[9] Ren, Shaoqing, He, Kaiming, Girshick, Ross. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J].
[10] Lin T Y, Dollár, Piotr, Girshick R, et al. Feature Pyramid Networks for Object Detection[J]. 2016.
[11] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection[J]. 2015.
[12] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[J]. 2015.
[13] Lin T Y, Goyal P, Girshick R, et al. Focal Loss for Dense Object Detection[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, PP(99):2999-3007.
[14] Romero A, Ballas N, Kahou S E, et al. FitNets: Hints for Thin Deep Nets[J]. Computer Science, 2014.
[15] Luo P, Zhu Z, Liu Z, et al. Face Model Compression by Distilling Knowledge from Neurons[C].
[16] Chen T, Goodfellow I, Shlens J. Net2Net: Accelerating Learning via Knowledge Transfer[C]. //ICLR, San Juan, Puerto Rico, 2016.

[17] Yim J, Joo D, Bae J, et al. A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning[C]. //IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. IEEE, 2017.4133-4141.

[18] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[J]. 2015.