Acceleration of Target Detection Based on Forced Knowledge Distillation

Jing Wen*, Chenggui Gong
School of Electronic Information Engineering, Beijing Jiao tong, Beijing, China

*Corresponding author e-mail: 16120203@bjtu.edu.cn

Abstract. In recent years, deep learning has achieved outstanding results on many problems such as computer vision, natural language processing and so on. The research of network model compression and acceleration can make the network model run efficiently on resource-constrained devices by greatly reducing the amount of computation of the network when the performance of the network decreases slightly. At present, knowledge distillation has achieved good results in classification tasks, but it shows strong limitations in the face of more complex tasks like detection. In this work, we propose a forced knowledge distillation framework, which focuses on improving the ability of feature extraction in detection. The whole model can be compact and fast without losing much accuracy. We use PASCAL VOC, KITTI and MSCOCO data sets to make a comprehensive evaluation. The results show that the forced knowledge distillation framework can fully learn the knowledge of teacher networks and achieve better detection results in smaller student networks.

1. Introduction
As a popular sub-field of machine learning in recent years, deep learning [1] has been applied successfully in many fields. Although the deep network model performs well in many experiments, it is still constrained by time and space in practical application. This situation poses a new challenge to us: how can we make the network model run efficiently on resource-constrained devices by reducing the computational load of the network greatly while maintaining the performance of the existing neural network basically unchanged.

Network compression and acceleration technology includes low-Rank [2], Pruning [3], Quantization [4], and Knowledge Distillation [5]. At present, these methods have achieved good results in classification tasks, and even the compressed model in a specific scene outperforms the original model. However, when faced with more complex tasks such as detection and semantics segmentation, they have not achieved good results. To solve this problem, this paper proposes an end-to-end framework, which applies the core idea of knowledge distillation to detection tasks; makes comparative experiments on the effects of various knowledge distillation methods in detection, and gets the best knowledge distillation method applied to the framework in detection tasks; and uses two-stage training method to optimize student network’s effect; conducts detailed experiments on multiple open datasets, and make the analysis of the experimental results. Results show that the framework proposed in this paper can greatly improve the training effect of student network, even in some data sets, the effect is basically the
same as that of teacher network, which achieves the original intention of network compression. Through comparative experiments, the best way of knowledge distillation for detection task is obtained.

2. Related work

2.1. Faster RCNN
In recent years, with the development of deep learning, convolutional neural network (CNN) has been widely used in various tasks of computer vision. R-CNN [7] proposed by Girshick et al. is widely used in detection. Fast R-CNN [8] proposed by Microsoft Asia Research Institute has greatly accelerated the recognition of interest regions by introducing RoI pooling layer. Faster R-CNN [9] goes further on the basis of Fast R-CNN. By introducing region proposal network (RPN), it makes the extraction of RoI more efficient and reliable.

2.2. Knowledge Distillation
Knowledge distillation refers to making small models fit large models by transferring large amounts of data, so that small models can learn function mapping similar to large models. Knowledge distillation was first proposed by Hinton in 2015. And then Romero proposed FitNets [6]. By adding loss into the middle layer, the network is divided into two parts while training, so that the output of the middle layer of the teacher network and the student network is as close as possible. Then the FSP method proposed by Yim et al. [11] transfers the relativity between feature maps as knowledge to student networks. Zagoruyko et al. [12] transfers the attention of feature maps as knowledge to student networks.

3. Forced knowledge distillation
Faster RCNN is used as the basic target detection framework. Faster RCNN consists of three modules: 1) feature extraction part of shared convolution layers; 2) region proposal network (RPN); 3) classification and regression network (RCN). RCN and RPN both use the output of 1) as their input, and RCN also uses the result of RPN as its input. The first part feature extraction is the common input of RPN and RCN. The quality of feature extraction has an important impact on the test results. Therefore, the guidance of teacher network for feature extraction of student network is very important.

3.1. Overall framework of forced knowledge distillation
The overall framework of forced knowledge distillation is shown in Figure 1. Firstly, we use hint-based learning [6], FSP method [11] and Attention method [12] to make the feature representation of student network and teacher network as similar as possible. We use L2 loss and Normalized L2 loss to learn the feature representation of teacher network. Secondly, by using forced distillation, students network can acquire teachers network’s knowledge, at the same time, it greatly promotes the learning of feature representation and accelerates the convergence of the overall model.

Figure 1. Overall framework of forced knowledge distillation
The training loss of the overall framework of forced distillation is defined as follows:

\[
L_{FKD} = \gamma L_{feature} + L_{cls} + L_{reg} \\
L_{cls} = L_{rpn.cls} + L_{rcn.cls} \\
L_{reg} = L_{rpn.reg} + L_{rcn.reg}
\] (1)

The definitions of $L_{cls}$ and $L_{reg}$ are the same as Faster RCNN, Softmax loss and smooth L1 loss are used respectively. $L_{cls}$ includes the classification loss of RPN and RCN, and $L_{reg}$ includes the bbox regression loss of RPN and RCN. $\gamma$ is a super parameter used to adjust the balance between different losses. $L_{feature}$ is the loss of feature extraction. This paper uses L2 loss and Normalized L2 loss experiments respectively. The definition of L2 loss is as follows:

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

Among them, $\mathcal{L}$ denotes all the teacher-student knowledge pairs that we want to transfer knowledge; $K_T$ and $K_S$ denote teachers' knowledge and students' knowledge respectively when knowledge is transferred. Normalized L2 loss is defined as follows:

\[
L_{feature} = \sum_{j \in \mathcal{L}} \left( \frac{K_S}{\|K_S\|_2} - \frac{K_T}{\|K_T\|_2} \right)^2
\] (3)

On the basis of L2 loss, the knowledge of both teachers and students is normalized, which has an important impact on the training of students network.

3.2. Forced knowledge distillation

Because the feature extraction part is the basis of the follow-up steps of Faster RCNN framework. In order to make the student network get better test results, it is necessary to make the output of the feature extraction part of the student network approach the teacher network as far as possible. Therefore, in order to promote the learning of feature extraction, we use forced knowledge distillation, that is, the student network directly inherits the weight of RPN and RCN. As a result, only the weight of feature extraction is different between student network and teacher network, and the approximation of feature representation is realized by ground truth approximation.

3.3. Learning procedure

The training process of the whole student network includes two stages. Firstly, we minimize the loss function $L_{FKD}$, and inherit the weight of FPN and RCN to promote the reduction of $L_{feature}$. Then, the $L_{feature}$ of feature extraction is cancelled and the routine loss function $L_{ori}$ is used to fine-tune network.

\[
L_{ori} = L_{cls} + L_{reg}
\] (4)

The following algorithm 1 explains the learning process.

| Algorithm 1 Transfer the distilled knowledge in detection |
|----------------------------------------------------------|
| Stage 1: Learning Feature Representation, Inheriting FPN and RCN Weights |
| Weights of the student and teacher networks: $W_s, W_t$ |
| 1: $W_s = \arg\min_{W_s} L_{FKD}(W_s, W_t)$ |
| Stage 2: Training for the original task |
| 1: $W_s = \arg\min_{W_s} L_{ori}(W_s)$ |
4. Experiment and analysis
We choose Resnet of different layers as our teacher/student model and use two different settings for students and teachers. We evaluated several common detection datasets, including KITTI, PASCAL VOC 2007 and MS COCO. For all data sets, the model is evaluated by means average precision (mAP) of IoU = 0.5. For COCO datasets, in addition to generic metrics, we also evaluated the average mAP of IoU [{0.5:0.05:0.95}] (expressed as mAP [0.5, 95]).

4.1. Experimental of compulsory knowledge distillation Algorithm
We experimented with three different knowledge transfer methods and two loss functions applied to COCO datasets. Teacher and student network are Resnet101 and Resnet18, respectively. Table 1 is the experimental results for this part.

| Table 1. Different knowledge transfer methods and loss function in Detection |
|-----------------------------|-----------------|-----------------|-----------------|
|                             | Hint Learning   | FSP             | Attention       |
| L2 loss                     | 32.9 (+1.7)     | 33.1 (+1.9)     | 33.9 (+2.7)     |
| Normalized L2 loss          | 33.2 (+2.0)     | 33.5 (+2.3)     | 34.3 (+3.1)     |

From Table 1, we can see that in this task, the knowledge transfer method based on Attention can get the best results in our forced knowledge distillation detection framework. At the same time, Normalized L2 loss is more stable than L2 loss, which shows that normalization plays an important role in the success of knowledge transfer.

Secondly, we use Attention method to do knowledge transfer, Normalized L2 loss to do loss function, and make full experiments on different data sets and different teacher-student networks to comprehensively investigate the effect of our forced knowledge distillation framework. Table 2 is the experimental results for this part.

| Table 2. Results of different data sets and different teacher-student network pairs |
|-------------------------------|-----------------|-----------------|-----------------|-----------------|
| Student | Teacher | PASCAL07 | COCO@.5 | COCO@[.5,.95] | KITTI |
|--------|---------|---------|---------|-------------|-------|
| Res18  | Res50   | 72.8 (+1.2) | 52.1 (+0.8) | 32.6 (+1.4) | 64.3 (+1.0) |
|        | Res101  | 74.3 (+2.8) | 53.3 (+2.0) | 34.3 (+3.1) | 65.6 (+2.3) |
| Res50  | -       | 73.4     | 54.7     | 33.6     | 65.7     |
|        | Res101  | 74.9 (+1.5) | 57.0 (+2.3) | 36.9 (+3.3) | 66.8 (+1.1) |
| Res101 | -       | 75.1     | 59.7     | 38.1     | 66.7     |

Table 2 shows the mAP of Res18 and Res50 on three different data sets under the guidance of different capacity Resnet. It is easy to see from the table that the deeper the model is, the better the model performance will be, but we know that the smaller the model is, the faster the model runs. It can be seen that the effect of each model has been improved to a certain extent after being instructed, and getting help from a better teacher will make the experiment more effective. Secondly, we can see that on the PASCAL and KITTI datasets, the effect of the student model after guidance is much closer to that of the teacher model. Even under the guidance of Res101, Res50 surpasses the teacher network slightly on the KITTI dataset. On the contrary, when the data set is much larger, it is very difficult for the student model to surpass the more complex teacher model. This shows that for large-scale data sets, models with higher capacity are more worthwhile.

5. Conclusion
We propose an end-to-end framework, which uses forced knowledge distillation to accelerate the CNN feature extraction part of the detection framework. We can guide students to learn more complex teacher models, and we explore a variety of knowledge transfer methods and loss functions, and find the most...
suitable combination for feature extraction part of detection framework to improve its performance. Full experiments on multi-data sets have resulted in consistent improvements in student networks, even slightly surpassing teacher networks on KITTI data sets. But as we all know, compared with the teacher's network, the student's network has obviously improved in speed and reduced in space occupancy.

Acknowledgments

Wen Jing (1993.7-), female, Han nationality, Taiyuan City, Shanxi Province, Master of Control Science and Engineering, Grade 2016, School of Electronic Information Engineering, Beijing Jiaotong University. Postal Address: Wenjing (Received) 15652599523, Beijing Jiaotong University, Haidian District, and Beijing.

Gong Chenggui (1993.7-), male, Han nationality, Bijie City, Guizhou Province, Master of Traffic Information Engineering and Control, Grade 2017, School of Electronic Information Engineering, Beijing Jiaotong University. Postal Address: Gong Chenggui (Received) 18810255373, Beijing Jiaotong University, Haidian District, and Beijing.

References

[1] Lecun Y, Bengio Y, Hinton G. Deep learning. [J]. Nature, 2015, 521 (7553): 436.
[2] 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.
[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] Leng C, Li H, Zhu S, et al. Extremely Low Bit Neural Network: Squeeze the Last Bit Out with ADMM [J]. 2017.
[5] Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network [J]. arXiv preprint arXiv:1503.02531.
[6] Romero A, Ballas N, Kahou S E, et al. FitNets: Hints for Thin Deep Nets [J]. Computer Science, 2015.
[7] 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.
[8] Girshick R. Fast R-CNN [J]. Computer Science, 2015.
[9] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [C]// International Conference on Neural Information Processing Systems. 2015.
[10] Pan S J, Yang Q. A Survey on Transfer Learning [J]. IEEE Transactions on Knowledge & Data Engineering, 2010, 22 (10): 1345 - 1359.
[11] 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 & Pattern Recognition. 2017.
[12] Zagoruyko S, Komodakis N. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer [J]. 2016.
[13] Chen G, Choi W, et al. Learning Efficient Object Detection Models with Knowledge Distillation [J]. NIPS, 2017.