Knowledge Distillation based Compact Model Learning Method for Object Detection

Jong Gook Ko, Wonyoung Yoo
Content Research Division,
Communications & Media Research Laboratory
ETRI (Eletronics Telecommunications Research Instutue)
Daejeon, Korea
jgko@etri.re.kr, zero2@etri.re.kr

***Abstract*—Recently, video analysis technology through deep learning has been developing at a very rapid pace, and most of the technology related to improving recognition performance in server environment is being developed. However, in addition to video analysis technology in the existing server environment, the demand of object detection in visual image analysis have been increasing recently in embedded boards of low specification and mobile environments such as smartphones, drones, and industrial boards. Despite the significant improvement in the accuracy of existing object detectors, image processing for real-time applications often requires a lot of runtime. Therefore, many studies are being conducted on lightweight object detection technology, and knowledge distillation is one of the solutions. Efforts such as model compression use fewer parameters, but there is a problem that accuracy is significantly reduced. In this paper, we propose method to improve the performance of lightweight mobilenet-SSD models in object detection by using knowledge transfer methods. We conduct evaluation with PASCAL VOC dataset. Our results show detection accuracy improvement in object detection.**

Keywords—object detection, knowledge distillation, lightweight deep learning model

I. Introduction

Object detection technology has developed rapidly through deep learning, but most of the technologies are based on high-performance GPU server environments. Recently, as video analysis demands in mobile and embedded environments are increasing for application utilization and service expansion, it is necessary to develop image analysis technologies in low-specification embedded and mobile environments rather than high specification server environments. Deep learning has gone beyond simply the level of "good models" If a smaller model can achieve as much performance as a larger model, it can be said to be more efficient in terms of computing resources, energy, and memory. For example, if you want to use an application using deep learning on your phone, and you want to use a model that requires hundreds of megabytes of memory, you need to connect to an online cloud server and use resources such as GPU, but if you make this model small enough, you can calculate it with just your phone's CPU, it would be better in many ways. Various compression methods such as pruning and knowledge distillation are being studied for a lighter deep learning model

The Neural Network pruning method is to find and remove the less important parameters of the Neural Network. In many cases, the Neural Network is configured to allow a sufficient number of parameters, because it is not known how many parameters are optimal to use when starting learning. After completing the learning of this structure, we can check the status of the parameters and see that there are many parameters that do not affect the results. The Neural Network pruning method removes the parameter from the Neural Network by selecting a parameter that has less impact on the result. Removal of parameters with less impact can result in loss of accuracy, and additional learning can be done while running to compensate for this loss of accuracy.

Knowledge distillations also referred to as the Teacher-Student Network model for learning a student network from the Teacher Network. When there is a Dataset you want to learn, the Teacher Network learns Dataset first. After that, Student Network, which is smaller than Teacher Network, uses Teacher Network to learn Dataset. This process is expressed in the term Distillation, which is to condense the knowledge of Dataset to the Student Network. Several papers have confirmed that the Student Network model, which has been learned in this way, performs better than the model without a Teacher Network. This is known as Knowledge-distillations because the knowledge of Teacher Network, a larger Neural Network, has been transferred into Student Network, a smaller Neural Network. Various studies are under way, including 'Distilling the Knowledge in a Neural Network'[1] published by Professor Geoffrey Hinton in 2015

Models produced through various model compression methods reduce parameters or channel numbers of models to result in significant speed-ups and reduce model size, but reduce the accuracy of compressed models. To solve this problem of accuracy reduction, we propose compact model learning method for object detection using Knowledge distillation in this paper.

II. Related works

Hinton et al. [1] argues that the success of the KD may be due to an incorrect class of output distribution, which provides information about class relationships. Hinton et al. [1] propose knowledge multiplication as a more common example than [2], which applies the prediction of the teacher model as a 'soft label' to suggest more loss of temperature cross entropy instead of L2 loss. Vapnik & Izmailov [3] studied the effects of knowledge transfer using secondary information called Private Information from a learning theory perspective. Lopez-Paz et al[4] established a connection between the KD and privileged information. Ba [5] compress the deep network into a shallow but wider network in which the compressed model mimicked 'logits'. Romero et al. [6] Introduce a two-step strategy for training deep networks. The middle layer of

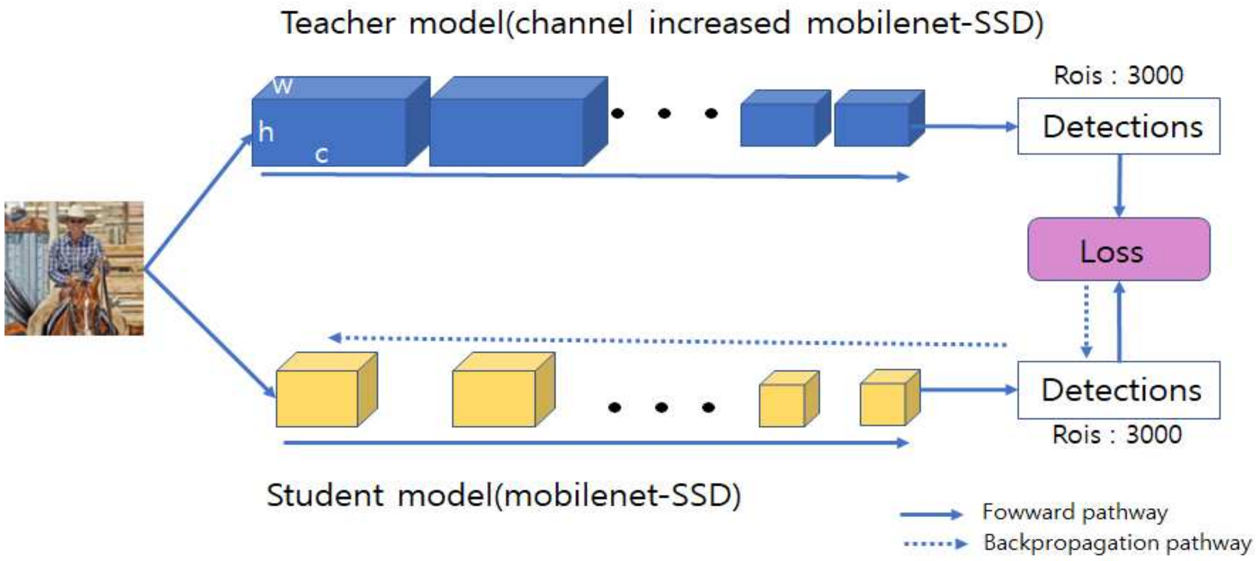

Figure 1. The proposed compact model learning method of object detection using Knowledge distillation

the teacher provides 'hint' to guide the training of the student model. Furlanello et al.[7] proposes KD techniques to boost quality, they also provide intuitions for the effectiveness of KD. Knowledge distillation is another way to maintain accuracy with model compression. [9] proposes an algorithm to train a single neural network by imitating the output of a model ensemble.

Recently, Phuong[8]showed a faster convergence rate in distillation. Furthermore Anil et al. [10] and Furlanello et al. [11] analyzed that distilling the student model on its own, either through mutual or self- distillation, improves quality, even if it does not involve a strong teacher.

III. Knowledge Distillation Based Compact Model Learning Method for Object Detection

In this work, we adopt the Mobilenet-SSD as the object detection framework. The SSD architecture is one stage detection which learns to predict bounding box locations and classify the locations in one pass. SSD can be trained end to end while Faster-RCNN cannot. The Mobilenet-SSD architecture consists of a base network(mobilenet) followed by SSD detection layers(several convolutional layers)

A. Overall Structure

Our overall compact model learning framework is illustrated in Figure 1. The compact object detection model for accuracy enhancement used mobilenet-ssd, and the teacher model used for knowledge transfer used the mobilenet-ssd model with increased channel size. The Teacher mobilenet-ssd has the same structure as the student model except for the increased channel of each conv. In other words, it performs object detection learning by extracting the same structure and the same number of candidate areas (Rois: 3,000)

For the transfer of knowledge, the Teacher model and the student model are learned as follows:

- Teacher model:

  Perform forward propagation on the input image ➔ output confidence values for each RoIs.

- Student model:

  Perform forward propagation on the input image ➔ output confidence values for each RoIs ➔ calculation of loss based on teacher and student loss ➔ perform back propagation based on the loss

B. Knowledge Distillation for Object Detection

Two loss functions are used to train the model for object detection: classification loss and regression loss.

In this paper, knowledge transfer was performed by passing only classification loss information of the teacher model to the student model except for regression loss information. The final loss of the student model is calculated as in Equation 1 below.

$$\text{Loss} = \alpha(\text{CE\_loss}(S_{conf}, GT)) + \beta\,(\text{KLDiv}\left(\frac{T_{conf}}{Temp}, \frac{S_{conf}}{Temp}\right))$$

$$\alpha + \beta = 1 \qquad (1)$$

where S_{conf} and T_{conf} are confidence values of Student and Teacher network. GT and Temp value are Ground Truth label and the temperature value to make the output soft probability. CE_loss means cross entropy function and KLDiv means Kullback–Leibler divergence function.

IV. Experiments

In this section, we show experimental results of our proposed method for object detection on server with Titan XP GPU. We used PASCAL VOC dataset for object detection accuracy experiment. VOC2012+VOC2007 was used for learning and VOC2007 test set was used for testing.

Table 1 shows the performance comparison of the mobilenet-ssd of baseline for PASCAL VOC test dataset, teacher mobilenet-ssd for knowledge transfer, and mobilenet-ssd model based on knowledge transfer. Teacher model of channel increased mobilenet-ssd is a high accuracy but heavy model with slow processing speed. The knowledge distillation based Mobilenet-SSD proposed in this paper has the same processing speed as the existing Mobilenet-SSD, but the accuracy is increased by 0.5 mAP.

TABLE I. COMPARISON OF OBJECT DETECTION ACCURACY EVALUATED ON PASCAL VOC DATASET

Models	**Accuracy**	**Inference Speed**
Mobilenet-SSD (baseline)	67.5%	5ms
Teacher Mobilenet-SSD (channel increased model)	69.7%	9ms
KD-Mobilenet-SSD (proposed KD based compact model)	**68% (+0.5%)**	5ms

V. CONCLUSION

We propose compact model learning method for object detection using knowledge distillation. Heavy Mobilenet-SSD that has increased channels are used as a teacher model to guide the learning of student Mobilenet-SSD model. We show the performance comparison of baseline student model, teacher model, and KD-student model. Notably, the KD-Mobilenet-SSD models trained with our learning method execute at the same speed as baseline model with accuracy improvement at PASCAL VOC dataset. In this paper, we do not use regression loss but only classification loss for Knowledge distillation. We are planning to improve object detection accuracy by using both of them in next time.

ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIP) (No. 2018-0-00198), Object information extraction and real-to-virtual mapping based AR technology)

REFERENCES

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

[2] J. Ba and R. Caruana. Do deep nets really need to be deep? In Advances in neural information processing systems, pages 2654–2662, 2014

[3] Vapnik, V. and Izmailov, R. Learning Using Privileged Information: Similarity Control and Knowledge Transfer. Journal of machine learning research, 16(2023-2049):2, 2015

[4] Lopez-Paz, D., Bottou, L., Scholkopf, B., and Vapnik, V. ¨ Unifying Distillation and Privileged Information. arXiv preprint arXiv:1511.03643, 2015

[5] J. Ba and R. Caruana. Do deep nets really need to be deep? In Advances in neural information processing systems, pages 2654–2662, 2014

[6] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio. Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550, 2014.

[7] Furlanello, T., Lipton, Z. C., Tschannen, M., Itti, L., and Anandkumar, A. Born again neural networks. arXiv preprint arXiv:1805.04770, 2018

[8] Phuong, M. and Lampert, C. Towards understanding knowledge distillation. In International Conference on Machine Learning, pp. 5142–5151, 2019

[9] C. Bucilua, R. Caruana, and A. Niculescu-Mizil. Model compression. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 535–541. ACM, 2006

[10] Anil, R., Pereyra, G., Passos, A., Ormandi, R., E. Dahl, ´ G., and E. Hinton, G. Large scale distributed neural network training through online distillation. arXiv preprint arXiv:1804.03235, 2018.

[11] Furlanello, T., Lipton, Z. C., Tschannen, M., Itti, L., and Anandkumar, A. Born again neural networks. arXiv preprint arXiv:1805.04770, 2018.