An Ensemble Method of CNN Models for Object Detection

Jinsu Lee
School of ICT
University of Science and Technology (UST)
Daejeon, Republic of Korea
jinsulee@ust.ac.kr

Sang-Kwang Lee
SW·Contents Research Laboratory
Electronics and Telecommunication
Research Institute
Daejeon, Republic of Korea
sklee@etri.re.kr

Seong-Il Yang
SW·Contents Research Laboratory
Electronics and Telecommunication
Research Institute
Daejeon, Republic of Korea
siyang@etri.re.kr

***Abstract*—Object detection is a research field that deals with detecting objects of a certain class in digital images and videos. Traditional methods of object detection were based on pre-structured features and had limitation on accuracy and computational efficiency. As deep learning had been proved to be a breakthrough, researches about object detection method based on deep learning, especially CNN, started. CNN-based object detection methods can be divided into two types. One is two-stage detector that once region proposals are generated, then they are classified. The other is one-stage detector that detects and classifies the object without generating region proposals. In two-stage detector case, combining CNN models is one of the ways to improve the accuracy in detection, which is called ensemble. In ensemble method, the region proposals generated from each CNN models are combined, classified, and finally voted. When selecting CNN models to be used in ensemble method, various properties of them should be considered in order to enhance complementary strength. In this paper, we propose advanced ensemble method in object detection with novel methods of model selecting and box voting. It is proved with experiment that the accuracy in object detection increased with our proposed methods. Also, combining the original method and our proposed method is expected to further increase the accuracy in detection and make ensemble model more robust.**

***Index Terms*—Object detection, Ensemble method, CNN**

I. Introduction

Object detection is a research field that deals with detecting instances of semantic objects of a certain class in digital images and videos [1]. In order to detect an object, first of all, the features should be extracted from given image, and these features largely affects the accuracy in detection. At early research stage, the objects are detected using the feature points which are pre-structured with image processing techniques such as SIFT [2], SURF [3] and HOG [4]. However, in ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012, deep convolutional neural networks (CNN) showed outstanding performance in image classification, beating other image processing-based classification methods. Since then, there have been intense studies about image classification and object detection method using deep neural networks, especially CNN.

The performance of CNN has been improved as having deeper layers. However, CNN with deep structure can cause over-fitting or vanishing gradient problem. In recent studies, sparse structure of CNN [5] or residual learning method [6] is considered as a solution of these problems. Not only the structure of CNN, but also the methodology of object detection based on CNN has been studied. R Girshick et al. [7] first applied CNN to object detection method, called R-CNN. The detection method of R-CNN consists of three parts. Firstly, region proposals, which mean the parts of image where any objects can be, are extracted using Selective search algorithm [8]. Then, the feature maps are extracted from each region proposal. At last, each region proposal is classified, using the feature maps. This two-stage detector shows high accuracy in detection but requires a lot of time cost because of the algorithm that generates region proposals. One-stage detector such as YOLO [9] and SSD [10] is designed to relieve the time cost, but shows relatively lower accuracy in detection.

The accuracy of two-stage detector can be further increased by combining CNN models, which is called "ensemble". In ensemble method, the region proposals generated from the final feature map are shared between CNN models. Then, each CNN model classifies the shared region proposals. The results of classification from each model are unified and voted at last. The sharing and voting process in ensemble method requires additional time cost, so it is usually considered in time-free application.

In this paper, we propose advanced ensemble method in object detection, especially focusing on Faster R-CNN [11], with novel model selecting and box voting methods. With proposed model selecting method, the CNN models are selected based on not only overall mean accuracy precision (mAP) but also AP according to the class and the size of objects so that the complementary strength between each model can be leveraged. And with our proposed box voting method, the classes predicted by multiple classifiers are voted, based on their per-class AP, which makes the voting process more reasonable.

In Section II, we have a survey on CNN-based object detection method and CNN model used as feature extractor. In Section III, proposed box voting and model selecting methods are explained. The result of experiments can be checked in Section IV, and in Section V, our conclusion and future works are mentioned.

978-1-5386-5041-7/18/$31.00 ©2018 IEEE

II. Related Works

A. CNN-based object detection method

R-CNN (Regions with CNN features) [7] method is a object detection method that first applied deep neural network. The detection process of R-CNN method consists of three parts. First, region proposals which mean some parts of an image where an object can be are extracted, using Selective search algorithm. Then, the region proposals are transformed into a fixed size, and the features of them are extracted using CNN. Finally, based on the extracted features, the region proposals are classified with Support Vector Machine (SVM) classifier. For more accurate prediction of box coordinates, a calibration process, called bounding-box regression, is applied to the final prediction result. As an experiment with PASCAL VOC 2010 dataset, R-CNN method showed mAP of 53.7%.

However, R-CNN method requires a lot of time in training and inference mainly for two reasons. One is that the training and inference process is separated into CNN, classifier, and bounding-box regression. The other reason is that CNN should be computed for every region proposals. In Fast R-CNN [12] method, to overcome these shortcomings, classifier and bounding-box regression is trained at once as computing the summation of the losses. Also, CNN is computed for an input image, and RoI (Region of Interest) pooling is applied to the extracted feature map. With this structural modification, Fast R-CNN showed higher accuracy and speed in object detection.

The algorithm that generates the region proposals in R-CNN and Fast R-CNN is separated from CNN. This algorithm is not trainable and requires a lot of computation. In Faster R-CNN [11] method, an additional CNN is used to generate the region proposals instead of Selective search algorithm. The input to the additional CNN, called RPN (Region Proposal Networks), is the final feature map extracted from CNN model, and the output is the generated region proposals. As integrating the region proposal generating process into the main CNN model, speed and accuracy in detection was increased.

The object detection methods mentioned above are regarded as two-stage detector in respect that the detection process is composed of region proposal generating part and classification part. These methods are high in accuracy but relatively low in speed to be applied to real-time application. This led the research in object detection method to one-stage detector. YOLO (You Only Look Once) [9] method is one of the one-stage detection methods, which considers the whole detection process as single regression problem. Instead of generating region proposals, the input image is divided into a number of grids, and a specific number of bounding boxes are assigned to each grid. With YOLO method, 45 images can be processed in a second, which implies that the object detection with YOLO can be applied to real-time system. However, the accuracy in detection was 10% lower than Faster R-CNN method. SSD (Single Shot multibox Detector) [10] method, another one-stage detector, exploits multiple feature maps with different sizes which were used in CNN computation, unlike other detection methods that exploit only the final feature map. The feature maps in early stage are large in size, which means they are used to detect small size of objects. In contrast, the feature maps in late stage are small in size, which means they are used to detect large size of objects. As getting rid of region proposal generating process, SSD method showed higher performance than Faster R-CNN in respect of speed and even in accuracy on some occasions.

B. CNN for object detection

In object detection, CNN model that extracts feature from given image affects the performance of detection model. The first CNN architecture with deep layers was AlexNet [13] that won ILSVRC 2012 with outstanding performance. AlexNet consists of 5 convolutional layers and 3 fully connected layers, and utilized 2 GPUs (Graphic Processing Unit) in parallel for quick computation. As an activation function, ReLU (Rectified Linear Unit) is applied instead of hyperbolic tangent or logistic regression, because of its property of non-saturating non-linearity. Also, in order to prevent over-fitting problem, data augmentation and Dropout [14] are applied.

The size of convolutional filters in AlexNet is various from 3×3 to 11×11. However, K Simonyan et al. [15] noticed that large size of convolutional filter is inefficient in respect of accuracy and speed. In their proposed CNN architecture, VGG Net, convolutional filters with 3×3 are adopted, with the fact that multiple layers of 3×3 convolutional filters have the effect of single larger convolutional filter such as 5×5 or 11×11. The advantage of convolutional filter with 3×3 size has proved in many researches, and it has been exploited in many CNN architectures developed afterward, even though VGG Net won second place in ILSVRC 2014.

The first place in ILSVRC 2014 was won by GoogLeNet which applied Inception module that consists of pooling layer and convolutional layers with various sizes. The filters with various size can extract various features, however, which requires huge amount of computation. In order to relieve this computational load, 1×1 convolutional layers are added to Inception module. With 1×1 convolutional filter, the dimension of feature maps can be decreased so that the computational load is also decreased. Although GoogLeNet outperformed VGG Net in ILSVRC 2014, its structural complexity made it difficult to be used in real applications.

CNN architectures are developed in the way of increasing the depth of layers. However, deep neural network can lead to problems such as over-fitting, vanishing gradient, or degradation of accuracy with some reasons. K He et al. [6] developed a CNN architecture, named ResNet, adopting the concept of residual learning. Residual learning is a learning method that a set of layers learns not the output but the difference between the input and the output, which enables the model to learn subtle change sensitively. ResNet is a CNN architecture that has same structure with VGG Net but where residual learning is applied. With residual learning, the accuracy was consistently increased as the network gets deeper to 152 layers.

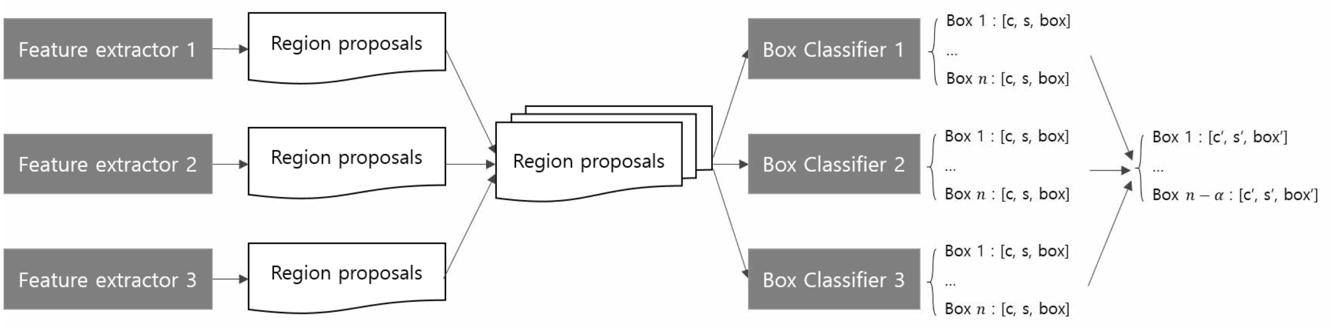

Fig. 1: Ensemble method in two-stage detector.

III. Ensemble of CNN Models

One of the ways to enhance the accuracy in object detection is to combine CNN models, which is called "ensemble". J Huang et al. [16] proposed an ensemble method of Faster R-CNN, and this achieved the state of the art performance on the 2016 COCO object detection challenge. In order to construct ensemble detector, first of all, CNN models to be used must be selected. They proposed model selecting method based on category-wise AP (Average Precision) vectors in order to acquire complementary advantages of each model. The diversity between models is computed as cosine distance between the category-wise AP vectors. If cosine distance between two models is higher than some threshold, the model with higher mAP is selected, and the other is discarded. The region proposals generated from each selected model are combined into a set of region proposals, shown as in Fig. 1. The set is shared between each model so that each model classifies it and calibrates the box location. The results of box location from each model are voted in the way of assigning more weight to the result with high confidence. The coordinate of a voted box $Loc_{(x,y,w,h)}$ is calculated as:

$$Loc_{(x,y,w,h)} = \frac{\sum_{i=1}^{N} c_i \cdot Loc_{(x,y,w,h)_i}}{\sum_{i=1}^{N} c_i} \tag{1}$$

where c_i and $Loc_{(x,y,w,h)_i}$ are the confidence and the coordinate from i-th model.

We propose novel model selecting and class voting methods in order to further increase the accuracy in detection of ensemble method. In model selecting, we consider AP according to not only the class but also the size of object. Some models may show high performance in detecting large size of objects, while others in detecting small size of objects. The complementary strength between them can be acquired, avoiding biased selection of models. The diversity between the models is computed as cosine distance between the object size-wise AP vectors. And in box voting, the confidence of class is voted by assigning the per-class AP as weight. The voted confidence c is computed as:

$$c = \frac{\sum_{i=1}^{N} AP_i \cdot c_i}{\sum_{i=1}^{N} AP_i} \tag{2}$$

TABLE I: List of feature extractors used in experiment. Stride refers to the output stride of extracted RPN feature map, and Loss ratio refers to the ratio of the location loss and objectness loss.

Feature extractor	Stride	Loss ratio	mAP
Inception V2	8	2:1	0.5470
		1:1	0.5359
		1:2	0.5390
	16	2:1	0.6940
		1:1	0.6597
		1:2	0.6858
ResNet-50	8	2:1	0.6662
		1:1	0.6685
		1:2	0.6682
	16	2:1	0.6744
		1:1	0.6553
		1:2	0.6558
ResNet-101	8	2:1	0.5859
		1:1	0.6267
		1:2	0.2995
	16	2:1	0.7007
		1:1	0.6267
		1:2	0.2995
ResNet-152	8	1:2	0.6793
	16	2:1	0.6932
		1:1	0.6803
		1:2	0.6607

where AP_i and c_i are the AP and the confidence from i-th model. The final class is predicted as the class with the maximum value of the voted confidences. This improvement in class voting method can increase the accuracy in detection, considering that most mis-predicted cases in example results are made because of the class, rather than the box coordinates.

IV. Experiments and Results

The proposed ensemble method is evaluated with experiment on the comparison with the original ensemble method with PASCAL VOC 2012 dataset [17]. PASCAL VOC 2012 dataset consists of 17,125 images labeled with 20 kinds of classes, such as aeroplane, person and chair. The feature extractors used in experiment are Inception V2 [18], ResNet-50, ResNet-101, and ResNet-152 with different hyperparameters.

TABLE II: Effect of proposed model selecting method. The selecting method based on AP according to class refers to the method proposed by J Huang et al. [16], and the selecting method based on AP according the object size refers to the proposed method in this paper.

Selecting method	Selected feature extractors	mAP
AP according to class	ResNet-101 w/ stride 16, loss ratio 2:1	0.8241
	ResNet-101 w/ stride 8, loss ratio 2:1	
	InceptionV2 w/ stride 8, loss ratio 2:1	
AP according to object size	ResNet-101 w/ stride 16, loss ratio 2:1	0.8519
	InceptionV2 w/ stride 16, loss ratio 2:1	
	ResNet-101 w/ stride 8, loss ratio 1:1	

The detailed information of the feature extractors is described in Table I. Other hyperparameters which are not described in Table I, such as learning rate, anchor scale, and aspect ratio, are fixed as same value among every feature extractor for unbiased comparison. It means that each model may not show their respectively best performance because they are not under optimal configuration.

Table II shows the feature extractors which are selected according to different model selecting methods and their performance. Different feature extractors are selected from the list of CNN models described in Table I according to the selecting methods, which implies that the feature extractors have their own strength in different respects. Given this, in ensemble method, it is suggested that the feature extractors be selected based on the criteria that is considered to be important in the situation. The result shows that mAP increased by about 3% with our proposed method. Combining two selecting method is expected to further increase the accuracy in detection.

Proposed box voting method was evaluated with the ensemble model which is shown in the last row in Table II. When applying our proposed box voting method, mAP has increased to 0.8704, showing 2% increase compared with the original method.

V. Conclusion

In this paper, we proposed model selecting and box voting methods in ensemble method of two-stage detectors for the purpose of improvement in the accuracy in object detection. In proposed model selecting method, object size is also considered as criteria to select feature extractors. And in box voting method, we improved class voting process by exploiting per-class AP as weight. As the result of experiment with PASCAL VOC 2012 dataset, the effect of our proposed methods are proved with increased mAP. Combining the model selecting methods based on class and object size is expected to further increase the accuracy in detection and make the ensemble model more robust.

In future work, some experiments on model selecting method based on both category-wise AP and object size-wise AP will be conducted. Other features such as shape of image or major color in image also can be used in model selecting. In addition, further evaluation of the proposed model selecting method is required to prove the robustness on various dataset.

The ensemble method in object detection is employed in the situation that accuracy is more important than time cost. However, it requires a lot of additional time cost for the reason that it is consisted of several feature extractors in parallel. Thus, there needs study about time-efficient ensemble method to enhance accuracy with less time cost.

Acknowledgment

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (R0118-16-1005, Digital Content In-house R&D)

References

[1] Wikipedia, ‘Object detection’, 2018. [Online]. Available: https://en.wikipedia.org/wiki/Object_detection. [Accessed: 27-Jul-2018].

[2] D.G. Lowe, “Object Recognition from Local Scale-Invariant Features”, Proceedings of the Seventh IEEE International Conference on Computer Vision, pp. 2:1150-1157, 1999.

[3] H. Bay, A. Ess, T. Tuytelaars, and L.V. Gool, “Speeded-Up Robust Features (SURF)”, Computer Vision and Image Understanding, Vol. 110, No. 3, pp. 346-359, 2008.

[4] N. Nadal and B. Triggs, “Histograms of oriented gradients for human detection”, 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), 2005.

[5] C. Szegedy et al., “Going deeper with convolutions”, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-9, 2015.

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition”, 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”, Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 580-587, 2014.

[8] J.R.R. Uijlings, K.E.A. van de Sande, T. Gevers, and A.W.M. Smeulders, “Selective search for object recognition”, International Journal of Computer Vision, Vol. 104, No. 2, pp. 154-171, 2013.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection”, Computer Vision and Pattern Recognition 2016, pp.779-788, 2016.

[10] W. Liu et al, “SSD: Single Shot MultiBox Detector”, European Conference on Computer Vision 2016, pp. 21-37, 2016.

[11] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, No. 6, pp. 1137-1149, 2017.

[12] R. Girshick, “Fast R-CNN”, IEEE International Conference on Computer Vision 2015, pp. 1440-1448, 2015.

[13] A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks”, Conference on Neural Information Processing Systems 2012, pp. 1097-1105, 2012.

[14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting”, Journal of Machine Learning Research, Vol. 15, pp. 1929-1958, 2014.

[15] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition”, International Conference on Learning Representations 2015, 2015.

[16] J. Huang et al, “Speed/accuracy trade-offs for modern convolutional object detectors”, Computer Vision and Pattern Recognition 2017, pp. 7310-7319, 2017.

[17] M. Everingham, L. Gool, C. Williams, J. Winn, and A. Zisserman, “The Pascal Visual Object Classes (VOC) Challenge”, International Journal of Computer Vision, Vol. 88, No. 2, pp. 303-338, 2010.

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision”, Computer Vision and Pattern Recognition 2016, pp. 2818-2826, 2016.