# Object Detection Based on Multiscale Merged Feature Map

Zhaohui Luo, Hong Zhang, Zeyu Zhang, Yifan Yang(✉), and Jin Li

Image Processing Center, Beihang University, Beijing 100191, China
lzhmolly@163.com, Stephenyoung@163.com

**Abstract.** In object detection, high quality feature map is of great importance for both object location and classification. This paper presents a new network architecture to get higher quality feature map, which combines the feature map from shallow convolution layers with deep convolution layers by up–sampling and concatenating. It adopts a one-stage network, which does not rely on region proposal, to directly predict the location and classification of objects using the high quality feature map. With the input images of size 300 * 300, this network can be trained efficiently to achieve solid results on well-known object detection benchmarks: 77.7% on VOC2007, outperforming a comparable state of the art SSD [1], YOLO [5] and Faster R-CNN [4] model.

**Keywords:** Object detection · Up-sampling · Concatenating · Feature map

## 1 Introduction

Convolutional neural networks (CNNs) have made impressive improvements in many computer vision tasks for several years, such as image classification, object detection and semantic segmentation and other tasks. As for object detection, most of current state-of-the-art object detection methods adopt two kind of architecture. One use a Region Proposal Network (RPN) [4] to propose potential bounding boxes that are widely called region proposals, then use a high-quality classification network to detect real objects from those region proposals, as is exemplified in R-CNN [2], Fast R-CNN [3] and Faster R-CNN [4]. The other use a single network that do not rely on region proposal to directly predict the bounding boxes and classification of objects, such as YOLO [5] and SSD [1]. The first kind of method can achieve better results, while it is computationally intensive. The second kind of method is relatively fast and can be applied to real-time applications, but there is still a lot of room for improvement in detection accuracy.

For the second method, it skips the process that generating region proposals to directly detect object from the feature layer. Thereby its performance greatly depend on the quality of feature layer. However there is an imbalance in the depth of feature layer between object location predication and object classification predication. Previous works have shown that feature map from shallow convolution layer capture more fine details of the input objects, which is important for object location predication. However object classification predication usually adopt feature maps from deep convolution layers to have a more accurate judgement in image classification. So the feature map


Y. Wang et al. (Eds.): IGTA 2018, CCIS 875, pp. 80–87, 2018.
https://doi.org/10.1007/978-981-13-1702-6_8

from shallow convolution layer lacks for classification information from higher layer, which may yield false positives. This can explain why SSD [1] perform not well in small target detection.

To solve this problem, we present a new network architecture for object detection in this paper, which generate object location predication and classification predication from the new feature maps that combines the feature map from shallow convolution layers with it from deep convolution layers by up–sampling and concatenating. By this way, the new feature map not only preserve fine details of the input objects but also have better object classification information, which is a good solution to the imbalance problem mentioned above.

## 2 Related Work

### 2.1 Early Object Detection Networks

With the development of deep convolution neural networks, object detection has made great progress in accuracy. There are two types of object detectors based on deep learning, one way adopt the popular two-stage object detection strategy, which firstly take a region proposal network (RPN [4]) to generate potential bounding box and then use a classification network to recognize the real objects from those region proposal. Another way directly predict object location and classification with a one-stage network that do not relay on region proposals. Methods based on region proposals are firstly used by R-CNN [2], it generates region proposals by a traditional way from Selective Search [6], after that those region proposals are resized into the same size and put into a classical classification network to recognize its categories. With the spatial pyramid pooling layer (SPP net [7]) put forward, Fast R-CNN [3] allow region proposals with different size and scale share the base convolution layer to get features for the later classification process, which largely reduce the convolution computation in feature extraction of region proposals. Afterwards Faster R-CNN [4] no longer uses traditional method Selective Search [6] to extract region proposals, it also use a convolution neural network named RPN [4] to generated region proposals.

In another way, there are also a lot of object detectors that adopt one-stage network and skep the process of extracting region proposals. For example, OverFeat [8] predicts a bounding box directly from each location of the topmost feature map after knowing the confidences of the underlying object categories by a classification network. Lately YOLO [5] begin to predict bounding boxes and the confidence of containing objects at the same time from the topmost feature map, and it share the bounding box for multiple categories. Afterwards SSD [1] do not just take use of the topmost feature map to predict the final result, it begin to adopt multiscale feature maps from different depth convolution layer, which make great improvements to the accuracy of object detectors with one-stage network. Our method try to concatenate those multiscale feature map by up-sampling process to gain higher quality feature map for bounding box prediction and categories prediction.

### 2.2 Base Network

Almost all image processing and pattern recognition approaches need to extract features from raw image data. In early research, hand-engineered features like Harr [9] features and HOG [10] features are well-designed and widely used. With the development of deep convolution neural networks, object classification network are widely transformed to extract high quality features for different image processing task. Those object classification network are called base network, such as AlexNet [11], VGG [12], and ResNet [13] and GoogLeNet [15–18]. Those base network have different model size and can achieve different precision, we should take the demand of different task into consideration to choose a proper base network to extract features from a large number of raw image data.

### 2.3 Up-Sampling and Concatenating

Semantic segmentation networks usually combine semantic information with appearance information by the way of up-sampling and concatenating. AS is exemplified in FCN [14], it begins to merge semantic information from a deep, coarse layer with appearance information from a shallow, fine layer to produce accurate and detailed segmentations. It shows that feature map from deeper layer provides classification information while the feature map from shallower layer provides location information. Although object detection task is not as strict as semantic segmentation task to the accurate boundary of objects, it just needs a bounding box to display object location. This architecture can also be applied to object detection to improve the accuracy by merging location information and classification information.

## 3 Our Approach

### 3.1 Feature Extraction Network

Figure 1 shows the structure of our detection model, we basically follow the method of SSD 1. Several feature layers are added to the end of the base network VGG16 12, we use the conv4 3, conv7 (fc7), conv8_2, conv9_2, and conv10_2 layer to build the up-sampling and merging architecture. It is processed as follows: the conv10_2 layer is 2x up-sampled and its size changed from $3 \times 3$ into $5 \times 5$, then concatenated with the conv9_2 layer by a concatenating layer to obtain the new $5 \times 5$ feature map with increased channel dimensions. In this illustration the 2x up-sampling layer is initialized by bilinear interpolation. At the same time, the new $5 \times 5$ feature map is 2x up-sampled and concatenated with conv8_2 layer to get the new $10 \times 10$ feature map. The same up-sampling and merging process is applied to conv7 layer and conv4_3 layer to get the new $19 \times 19$ feature map and the new $38 \times 38$ feature map. We adopt this up-sampling and merging method to get multiscale high quality feature map, then we predict both location and classifications from those new $5 \times 5$, $10 \times 10$, $19 \times 19$, $38 \times 38$ feature map, and the original $1 \times 1$, $3 \times 3$ feature map. Those new feature maps have fine details of objects appearance and good classification information to accurately detect objects in the image.

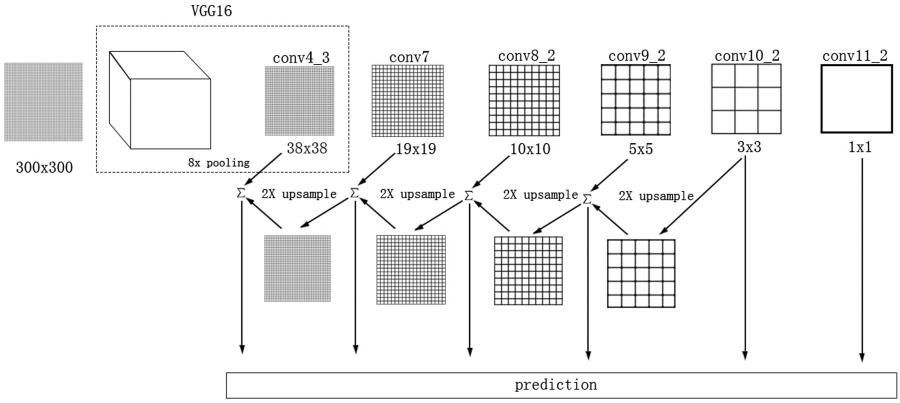


**Fig. 1.** Key idea of our object detection network.

### 3.2 Prediction Layer

Since object classification requires larger receptive field, generally it uses the deepest feature map followed by full connection layers to predict object categories. However, when we adopt the method mentioned above that uses different levels of feature maps to predict object classification and object location at the same time, the receptive field of feature map from shallower convolution layer is small, which will lead to a misclassification. To solve this problem, we use deeper convolution network for object classification behind those feature maps to expand its receptive filed. As is shown in Fig. 2, behind that feature map, a deeper convolution network (two $3 \times 3$ convolution layers) is adopted for object classification, while the location prediction use only one $3 \times 3$ convolution layer.

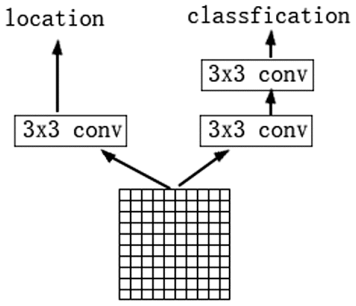


**Fig. 2.** Key idea of our object detection network.

### 3.3 Towards a More Efficient Detection Network

In our design, original feature map after the base network is concatenated with the upsampling feature map from deeper convolution layer to generate new feature map.

Thereby the new feature map will have a larger number of channels due to the concatenating strategy. However, the dimensions of feature map have great effects on the complexity and computation of the classification prediction layer and location prediction layer. So we add a $3 \times 3$ convolution layer after the concatenation layer to fix the number of channels to 256 for all different scale feature maps. By this sample design, we reduce the extra computation cost brought from the up-sampling and concatenating strategy.

## 4 Experiments on Object Detection

### 4.1 Implementations

The input images are resized to $300 \times 300$. We train our models end to end, fine-tune the model based on VGG16 [10] and pre-train it on the ImageNet ILSVRC CLS-LOC 20 dataset. We use SGD with initial learning rate $10 - 3$, 0.9 momentum, 0.0005 weight decay, and batch size 24.

### 4.2 VOC 2007 Detection

On PASCAL VOC 19 dataset, we compare against Fast R-CNN [3] and Faster R-CNN [4], SSD [1] on PASCAL VOC2007 test (4952 images), and trained our model on the union of 2007 trainval and 2012 trainval. Training images are resized into $300 \times 300$.

Table 1 shows that our method is already more accurate than Faster R-CNN 4 and SSD 1. We can clearly see that our method perform well on majority of categories, especially it improves a lot on the detection of aero plane, bird, person and other categories. However it has much worse performance on the detection of bottle, in large part it is because that bottle usually is not only small but also has a bigger length-width ratio than other categories, while feature map in our design responsible for small objects only has 2 ratio default box so that it is hard to detect objects that has a bigger length-width ratio like bottles. But overall, our method gains a significant improvements on the detection of most categories on PASCAL VOC2007. In Fig. 3 we show some detection results on PASCAL VOC2007.

**Table 1.** Test detection results in PASCAL VOC2007. Both fast and faster R-CNN resize input images' minimum dimension to 600. SSD and our method take use of $300 \times 300$ input images.

| Method | mAP | Areo | Bike | Bird | Boat | Bottle | Bus | Car | Cat | Chair | Cow |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Fast [3] | 70.0 | 77.0 | 78.1 | 69.3 | 59.4 | 38.3 | 81.6 | 78.6 | 86.7 | 42.8 | 78.8 |
| Faster [4] | 73.2 | 76.5 | 79.0 | 70.9 | 65.5 | **52.1** | 83.1 | 84.7 | 86.4 | 52.0 | **81.9** |
| SSD300[1] | 74.1 | 74.6 | 80.2 | 72.2 | 66.2 | 47.1 | 82.9 | 83.4 | 86.1 | 54.4 | 78.5 |
| Ours | **77.7** | **87.8** | **84.1** | **80.4** | **75.3** | 46.0 | **89.0** | **88.6** | **87.3** | **55.3** | 77.9 |
| Method | mAP | Table | Dog | Horse | Mbike | Person | Plant | Sheep | Sofa | Train | tv |
| Fast [3] | 70.0 | 68.9 | 84.7 | 82.0 | 76.6 | 69.0 | 31.8 | 70.1 | 74.8 | 80.4 | 70.4 |
| Faster [4] | 73.2 | 65.7 | **84.8** | 84.6 | 77.5 | 76.7 | 38.8 | 73.6 | 73.9 | 83.0 | 72.6 |
| SSD300[1] | 74.1 | 73.9 | 84.4 | 84.5 | 82.4 | 76.1 | 48.6 | 74.3 | **75.0** | 84.3 | 74.0 |
| Ours | **77.7** | **76.5** | 83.7 | **87.1** | **87.5** | **79.5** | **55.4** | **76.7** | 70.8 | **89.1** | **76.5** |

**Fig. 3.** Samples of detection results on VOC2007. Detections with scores higher than 0.6 are showed and each color corresponds to an object category.

Table 2 shows the speed and accuracy between Faster R-CNN [4], SSD [1] and our method. Our method outperform Faster R-CNN [4] in both speed and accuracy. Although it is slow than SSD300 [1], its accuracy is 3.6% higher than SSD300 [1]. Generally speaking, our object detector can achieve quite high performance, at the same time it runs very fast in the NVIDIA 1080 Ti with 44.8 FPS.

**Table 2.** The efficiency on VOC2007-test. We test those models on NVIDIA 1080Ti.

| Model | mAP | Input size | Times/ms | FPS |
|---|---|---|---|---|
| Faster [4] | 73.2 | about $1000 \times 600$ | 61.8 | 16.2 |
| SSD300 [1] | 74.1 | $300 \times 300$ | 11.5 | 87.0 |
| Our method | 77.7 | $300 \times 300$ | 22.3 | 44.8 |

### 4.3 Experiments on VOC2007

To get a detection model with high performance and high efficiency, we test our models in different configurations.

Table 3 shows the accuracy of our models in different configurations. If we only adopt up-sampling and concatenating strategy, the model can reach 76.9% mean AP with the VGG16 [12] base network. When different depth of classification and location predication layer is applied, the performance increase by 0.8% mean AP. this strategy use two $3 \times 3$ convolution layers for object classification, while one $3 \times 3$ convolution layer for location predication.

**Table 3.** Effects of various design choices. Up-sampling means model with up-sampling and concatenating feature map. conf&loc means model with different depth of location prediction layer and classification prediction layer.

| Model | mAP |
|---|---|
| Up-sampling | 76.9 |
| Up-sampling + conf&loc | 77.7 |

## 5 Conclusion

We present an object detection network based on multiscale merged feature map. Our system naturally take use of the state-of-the-art image classification models: VGG16 [12] as a base network, it achieves accuracy competitive with the Faster R-CNN [4] and SSD [1]. Our method aims to obtain high quality feature map by up-sampling and concatenating different scale feature map. Moreover high quality feature map is of great importance in other computer vision task, such as object classification and semantic segmentation. We believe that our design principle is not only applicable to object detection but also widely applicable to other computer vision tasks in future work.

## References

1. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
2. Girshick, R., Donahue, J., Darrell, T., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
3. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
4. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
5. Redmon, J., Divvala, S., Girshick, R., et al.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
6. Uijlings, J.R.R., Van De Sande, K.E.A., Gevers, T., et al.: Selective search for object recognition. Int. J. Comput. Vis. **104**(2), 154–171 (2013)
7. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8691, pp. 346–361. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10578-9_23
8. Sermanet, P., Eigen, D., Zhang, X., et al.: OverFeat: integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229 (2013)
9. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recogn. **29**(1), 51–59 (1996)
10. Papageorgiou, C.P., Oren, M., Poggio, T.: A general framework for object detection. In: Sixth International Conference on Computer Vision 1998, pp. 555–562. IEEE (1998)

11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
13. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
14. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
15. Szegedy, C., Liu, W., Jia, Y., et al.: Going deeper with convolutions. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
16. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456 (2015)
17. Szegedy, C., Vanhoucke, V., Ioffe, S., et al.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
18. Szegedy, C., Ioffe, S., Vanhoucke, V., et al.: Inception-v4, inception-ResNet and the impact of residual connections on learning. In: AAAI, pp. 4278–4284 (2017)
19. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. IJCV **88**, 303–338 (2010)
20. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. IJCV **115**, 211–252 (2015)