ISBN 978-981-14-1684-2
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
(WCSE 2019 SUMMER)
Hong Kong, 15-17 June, 2019, pp. 136-142
doi: 10.18178/wcse.2019.06.020

Traffic Sign Recognition Based on Up-sampling Convolution

Yitian Lu [1,2], Ping Jiang [1], Shun Nishide [2], Xin Kang [2] and Fuji Ren [2 +]

[1] School of Electrical Engineering, Nantong University, No.9 Seyuan Road, Chongchuan District, Nantong, Jiangsu, China

[2] Faculty of Engineering, Tokushima University 2-1 Minami Josanjima, Tokushima, 770-8506, Japan

Abstract. This paper presented a method makes traffic sign recognition faster and more accurate. Traditional faster detectors are limited by their accuracy and are not sensitive to small objects, in the area of self-driving, it has some inconspicuous but important object of concern, such as traffic sign. We noticed that most traffic signs in dataset is small and easily to confuse with complex backgrounds. In this situation, after a series of convolutional layers, some of these traffic signs can't be detected or classified correctly, and the problem of neglect happens a lot. In order to settle this problem and optimize the result, we simplified the SSD structure and introduced an up-sampling structure to make the geometric details of small objects distinctly. This method significantly improved the result of recognition, we got 97.6% mAP on The German Traffic Sign Benchmark with 96 × 96 input and SSD300 has 79.7% mAP on same dataset.

Keywords: traffic sign recognition, up-sampling, small objects.

1. Introduction

Neural networks involving image learning generally use convolutional neural networks as the basis, but after years of practice and development, the convolutional neural network has inherent problems in structure: the high-level layer has a large receptive field and strong semantic information representation, but low resolution brings weak ability of the geometric detail information representation. In this respect, the low-level layer's receptive field is relatively small, and the geometric detail information representation ability is formidable, although it has high resolution, each feature map contains scant contextual semantic information.

On this basis, the SSD strategy is to use multi-scale feature maps to predict objects, high-level feature information with larger receptive fields to predict large objects, and low-level feature information with smaller receptive fields to predict small objects. This brings up a problem: when using the feature information of the low-level network to predict small objects, due to the lack of high-level semantic features, SSD has a poor detection effect on small objects.

In order to tackle the lack, we simplified the model and utilized multi-scale features, followed by an up-sampling structure. For small objects, a shallow convolutional network can better preserve its geometric features, and multi-scale features contain rich semantic information [10][12]. In addition, up-sampling can magnify small features[21]. Theoretically, the model can significantly improve the detection ability of small objects and accuracy while maintaining detection speed. In our experience, we got 97.6% mAP on dataset.

This paper first introduces some state-of-the-art works about object detection and briefs problems they encountered at the time, then compared with recently proposed detection systems, we have summarized some superior methods for detecting small objects and combined them on our model, proved to be superior to the original one which trained on traffic sign dataset.

+ Corresponding author. Tel.: + 088 - 656 - 9684; fax: + 81886566575.
E-mail address: ren@is.tokushima-u.ac.jp.

2. Related Work

In the field of neural networks and deep learning, many models can be utilized to identify and detect objects, but due to different subjects, the performance of the model is also uneven.

In the task of object detection, the classical machine learning algorithm used the Histogram of Oriented Gradient (HOG), support vector machine (SVM) [1][2][23], etc. to extracted features and classified them. In recent years, deep learning has become popular, current state-of-the-art object detection models can be roughly divided into two methods. One method, called as two-stage object detector. The previous object detection models used selective search to find the area of the image where objects may exist, such as R-CNN [3], it needed to perform a forward feature extraction for each proposal extracted by selective search (about 2000 proposals per image), which the amount of calculation is large and cannot meet the requirements of real-time detection. Faster R-CNN [4] is a typical example that used Region Proposal Network (RPN) to extract proposal via sharing convolutional layers, it can achieve 83.8% accuracy on the voc2012 dataset. He K solved the problem that RoI Pooling will deform the RoI region[24] during the pooling process with the location information is not accurately extracted by Mask R-CNN [5], accomplished the segmentation task by improving the structure of Faster R-CNN. These two-stage object detection models can achieve the most advanced requirements in terms of accuracy at the time, but due to the features must be extracted first, then predict and identify the object's position, the detection speed is inevitably constrained.

To solve the shortcomings of the slow operation speed of the first one, the second method proposed. For example, Yolo (you only look once) and the subsequently proposed upgraded version Yolo9000, Yolov3 [6][7][8], as its name suggests, returned the position of the bounding box and the category directly at the output layer that achieve one-stage detection. Although the speed has a great improvement, the accuracy is lower than the first method. Single shot multibox detector (SSD) [9] inherited the idea of regression from Yolo, based on VGG network, completed objects location and classification once, that met the requirements of detection speed, although it is not as accurate as the two-stage detector, but it has higher precision than Yolo.

Traffic sign is a common typical small object, and in the actual observation, it is easy to be affected by occlusion and its own deformation, but to summarize these two methods above, most of them are lack in small object detection. In each image, such as SSD, the detector had to evaluate abundant candidate regions, but most of these regions does not contain the object that we need. In response to these problems, researchers had also proposed many effective suggestions. Reference [11] proposes that the network can effectively identify deformed objects by changing the shape of the convolution kernel; PSPNet [12] used a way to add contextual semantic information based on FCN [13], which can not only found small objects in the picture, but also distinguished targets in complex backgrounds; Online hard example mining (OHEM) [14] and Focal Loss (FL) [10] compensated the weight of small positive samples with algorithms.

For solve the problem that the extracted objects are not obvious due to insufficient training data, a feature extraction method based on up-sampling is proposed. After extracting the effective information in images, the features of the small objects are amplified and try to fuse features of different scales. After that, the non-maximum value suppression method is used to enhance the effect, then combined with the existing model to improve the recognition ability of the network and train a more rapid and effective model.

3. Up-sampling SSD

Our goal is to make the neural network has a more remarkable ability in detecting small objects such as traffic sign. The original SSD model has been excellent in detection speed, therefore all we have to do is improve its accuracy and the ability to detect small objects.

3.1. Issues in preparation

In our early research, we found that some of the images in the dataset were too small in resolution, there were also cases where the image was too bright or faint. After a deep convolutional processing, the geometric features of these images will be eliminated, which is not conducive to network extraction and learning features.

(a) Overly bright samples (b) Dim samples

Fig. 1: Some samples of 'speed limit 70km/h'. Images above illustrate the difference in size and contrast between the light and dark in the training samples. From human perspective, it is arduous to distinguish the classes of image.

In response to such problems, our work is to resolve the contradiction between low resolution and superior detection results. We first improved from model structure.

3.2. Framework

The deeper the convolutional network, the stronger the semantic information. But the deeper network structure will cause the feature map of the last layer to be too small. For example, the size of input image will from 32×32 becomes 2×2 after VGG, which is not conducive to subsequent detection and regression.

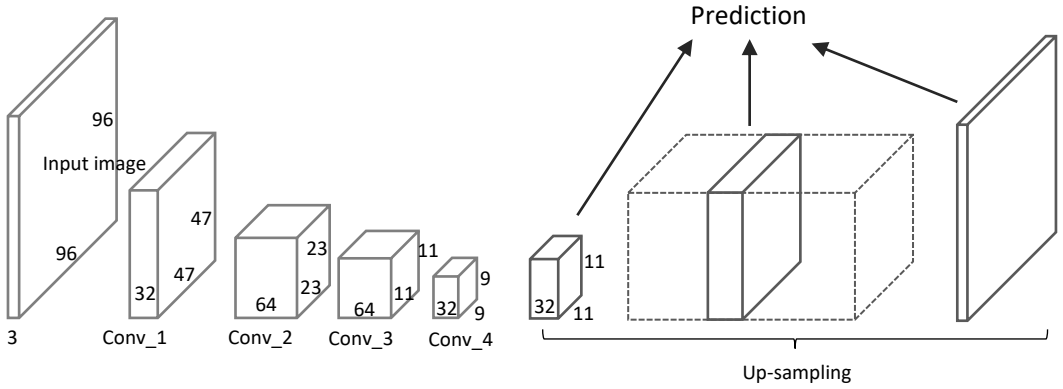


Fig. 2: Overview of our proposed model. Given an input image, first use a simple CNN to get the feature map, then send to up-sampling process, fuse deconvolution feature maps to predict location and confidences. The prediction part uses same measures and parameters with SSD basically.

For the task of improving the semantic information, not only [15], but also [16], [17], [18] and [19] are show a multi-scale approach to extract semantic information. In order to enhance the character expression capability of small objects, we use an additional up-sampling layer to put the geometric features of the object in size, that we can get the final feature maps where have the same size as input images, and then combined with the characteristic diagram of the previous high semantic information to forecast the object class. In this experiment, we use a series of successive deconvolutional layers to fulfill the up-sampling operation. As Fig. 2 demonstrates, different from the SSD, because of the simplified levels, we use all feature maps after up-sampling layers fusing and predicting. What surprised us is that this method can not only make up the deficiency of small object detection, but also speed up the training speed to a certain extent.

3.3. Training

In a network, feature maps from different levels are considered to have different receptive field sizes, some state-of-the-art object detection frameworks use similar pyramid structure prediction methods or enhance data size to improve the effectiveness of detection results. However, contrary to training large and

complex datasets that in most object detectors, we use less data augmentation methods because the dataset we used is not very large, so that it will encounter less trouble of overfitting, and the following experiments had also improved the accuracy of the model trained by the dataset after complex data augmentation is not as high as the model not adopted.

4. Experiment

In this section, we evaluate our model on the task of traffic sign recognition, then compared with initial SSD model. Try to control variables, use different optimizers and set different learning rates to force the model more outstanding.

4.1. Dataset

In this paper, we used The German Traffic Sign Benchmark (GTSRB)[20] as the main training dataset to detect and classify traffic signs, in addition, the pre-trained VGG network weights on ImageNet are loaded in SSD to extract feature maps when comparing. This dataset has 43 classes, over 39k train and 12k val images cover almost all signs we can see daily, and there are comparisons between glare and low light. Meanwhile, for the purpose of simulating the practical observation of traffic signs in the field of view, the problem of deformation and occlusion may encounter, the appropriate data augmentation processing is done.

Fig. 3: The German Traffic Sign Benchmark

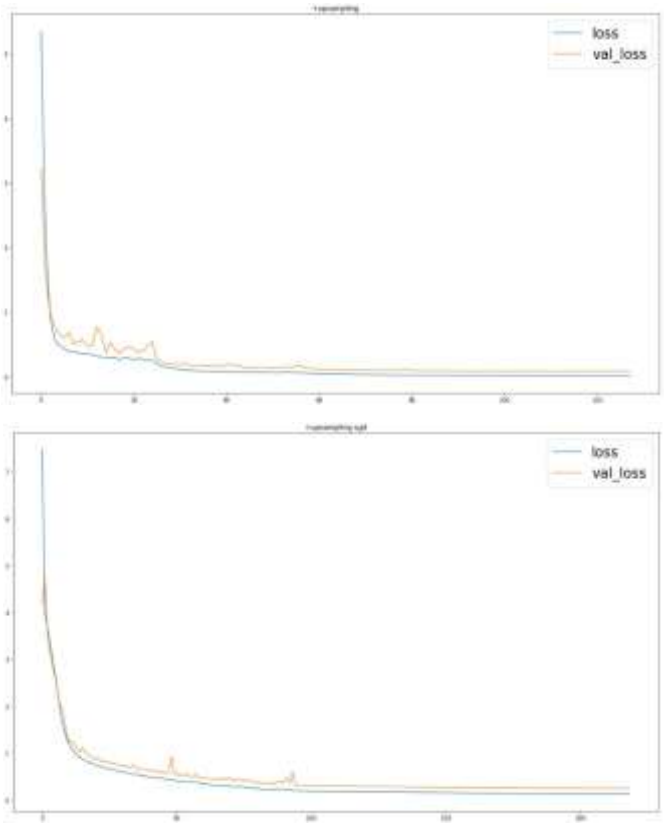
4.2. Results

On this dataset, we compare against SSD300 and SSD512 on GTSRB respectively, our model used deconv5, deconv6, deconv7, deconv8 to predict location and confidences. First fine-tune the model use SGD and Adam, find a better optimizer. We adopted a learning rate reduction strategy to ensure the orderly decline of gradient. Initial learning rate 0.001 and reduction factor 0.2, if the loss is not reduced after a certain epoch, next learning rate will become initial learning rate multiply factor, until the minimum learning rate 10-5. The batch size set as 16.

Table 1: Effect of different optimizers on model performance. We use both SGD and Adam to calculate the optimal gradient, we did this work to certificate which is more effective in our work

optimizer	loss	val_loss	mAP
SGD	0.0914	0.1640	94.9
Adam	0.0181	0.0831	97.6

Table 1 shows that our test result. Adam falls more quickly and reaches stable value faster. SGD can be more stable, but it requires more iterations. It is difficult for SGD to choose an appropriate initialization and learning rate, which Adam is a better selection.



(a) Adam (b) SGD

Fig. 4: SGD and Adam downward trend

We used the same parameters as those used above to train original SSD, compared with our model. Table 2 shows the results. As we can see, SSD512 is deeper than SSD300, which also indicates that deeper convolutional network is not conducive to identify small objects. In fact, when we only simplified the feature extraction structure, we can yield better results than the original SSD.

After we applied the up-sampled structure, the accuracy of each class increased significantly, bring about the final mAP boost. In terms of experimental results, the multi-scale prediction combined with the high-resolution feature maps can better recognize the traffic signs, and the high geometric features after up-sampling are more advantageous.

Table 2: GTSRB test results. The horizontal axis represents the classes of signs. SSD300 and SSD512 parameters include data augmentation as well as the original. 'ours' refers to the simplified feature extraction network, 'up' is the up-sampling process, and only the brightness is processed, while 'DA' is completely based on the SSD data augmentation

Method	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
SSD300	91.4	86.6	87.2	81.1	90.6	77.5	79.5	87.4	79.3	90.3	89.9	89.4	83.2	90.7	90.6
SSD512	55.4	76.5	84.4	78.8	90	67.5	74.7	74.2	69.3	89.7	90.5	89.3	90.7	90.9	90
Ours+up	100	99.8	97	91.9	97.3	98	94.2	92.9	93.6	100	97.2	99.7	98.5	99.5	100
Ours+up+DA	100	90.9	90.9	91.4	90.9	95.1	95.8	88.5	91.1	99.3	94.5	90.5	90.9	92	98.8

Method	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
SSD300	90.9	93.5	78.7	76.7	53.7	38.3	79	88.6	77.1	87.2	87	85	88.8	83.3	85.5
SSD512	91.7	88.1	85.8	80.6	45.8	39.5	41.2	85.3	67.9	53.6	77.5	62.6	69.6	88.3	77.5
Ours+up	99.7	100	100	95.1	100	100	97.6	92.1	97.1	100	97.7	93.7	80.4	99.8	99.8
Ours+up+DA	95	97.5	88	89.8	54.8	66.7	95.3	79.9	91.2	98.6	90.6	92.8	95.3	99.3	92.7

Method	31	32	33	34	35	36	37	38	39	40	41	42	43	mAP
SSD300	68.2	90.3	87.4	79.4	82	90.1	61.4	38.1	84.5	14.8	84.3	85.8	71.1	79.7
SSD512	52.7	92.9	96.4	71.4	73	86.2	76.1	24.4	77.7	36.9	73.2	58.7	71.6	73.4
Ours+up	95.9	100	100	100	100	99.9	100	97.3	99.6	99.8	92.2	99.3	99.2	97.6
Ours+up+DA	87.2	98.8	100	88	83.1	97.1	76.3	48	85.1	51.7	90.3	91.2	85.3	88.6

In addition, we also tried to utilize data augmentation, but the experimental results display that only a few classes prediction can meet or exceed the original results. The probable cause is that it does not work well for this dataset, which has the opposite effect. We tried the data augmentation built into Keras and redesigned several different sets of parameters as well, still didn't meet expectations.

Since as the last layer of convolution, conv4 also has higher semantic information. We try to use conv4, deconv6, deconv7, deconv8 as fusion prediction as well, made a comparative experiment with the same parameters and data processing medium, but the result showed that using all deconvolutional layers was slightly better. Table 3 and Figure 5 illustrate training result. It is means that in our case, the higher resolution of deconvolution layers can detect small objects better when the semantic information contained is relatively similar.

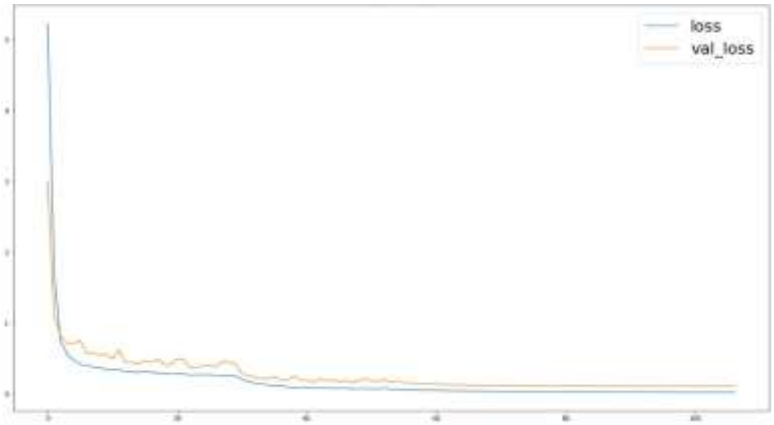


Fig. 5: Conv4, deconv6, deconv7, deconv8 training result.

Table 3: A comparison between different layers

Layers	loss	val_loss	mAP
Deconv5,6,7,8	0.0181	0.0831	97.6
Conv4, deconv6,7,8	0.0224	0.0992	97.2

5. Conclusion

This paper proposes a model that uses a simple convolutional neural network combined with SSD regression prediction function, and attaches an up-sampling structure, which surpasses the original model's detection ability in traffic signs recognition without sacrificing detection speed. Traffic signs are a symbolic symbol, each symbol has a specific shape meaning. Subsequent work will be applied to traffic signs around the recognition of human poses [22].

The experiment results suggest that this structure can be applied to many small object detection fields. For larger objects, it may require some special functions to reach the state-of-the-art level.

6. Acknowledgment

This research has been partially supported by JSPS KAKENHI Grant Number 15H01712.

7. References

[1] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005, 1: 886-893.

[2] Sánchez AVD. Advanced support vector machines and kernel methods[J]. Neurocomputing, 2003, 55(1-2): 5-20.

[3] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[4] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal

networks[C]//Advances in neural information processing systems. 2015: 91-99.

[5] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, 2017: 2980-2988.

[6] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[7] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[J]. arXiv preprint, 2017.

[8] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv: 1804.02767, 2018.

[9] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.

[10] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[J]. IEEE transactions on pattern analysis and machine intelligence, 2018.

[11] Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks[J]. CoRR, abs/1703.06211, 2017, 1(2): 3.

[12] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[C]//IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2017: 2881-2890.

[13] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.

[14] Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 761-769.

[15] Zhang S, Wen L, Bian X, et al. Single-shot refinement neural network for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4203-4212.

[16] Zhang Z, Qiao S, Xie C, et al. Single-Shot Object Detection with Enriched Semantics[R]. Center for Brains, Minds and Machines (CBMM), 2018.

[17] Chen L C, Collins M, Zhu Y, et al. Searching for efficient multi-scale architectures for dense image prediction[C]//Advances in Neural Information Processing Systems. 2018: 8713-8724.

[18] Lin G, Milan A, Shen C, et al. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 5168-5177.

[19] Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(4): 834-848.

[20] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. Man vs. computer:Benchmarking machine learning algorithms for traffic sign recognition. Neural networks, vol. 32, pp. 323–332, 2012.

[21] Shou Z, Chan J, Zareian A, et al. Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos[C]//Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. IEEE, 2017: 1417-1426.

[22] Yuming Xu, Fuji Ren, XIN KANG, Shun Nishide and Ping Jiang, Human Pose Recognition in Robots Based on Angle of Joint Vector, Proceedings of The 12th International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE17), Chengdu, 2017, pp. 304-315.

[23] Wang X, Jin C, Liu W, et al. Feature fusion of hog and wld for facial expression recognition[C]//Proceedings of the 2013 IEEE/SICE International Symposium on System Integration. IEEE, 2013: 227-232.

[24] Sun, X., Lv, M., Quan, C., & Ren, F. (2017, October). Improved facial expression recognition method based on ROI deep convolutional neutral network. In 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII) (pp. 256-261). IEEE.