Proceedings of the 36th Chinese Control Conference
July 26-28, 2017, Dalian, China

A Review of Object Detection Based on Convolutional Neural Network

Wang Zhiqiang[1], Liu Jun[1]

1. Fundamental Science on Communication Information Transmission and Fusion Technology Laboratory, Hangzhou 310018, China
E-mail: 1374113850@qq.com

Abstract: With the development of intelligent device and social media, the data bulk on Internet has grown with high speed. As an important aspect of image processing, object detection has become one of the international popular research fields. In recent years, the powerful ability with feature learning and transfer learning of Convolutional Neural Network (CNN) has received growing interest within the computer vision community, thus making a series of important breakthroughs in object detection. So it is a significant survey that how to apply CNN to object detection for better performance. First the paper introduced the basic concept and architecture of CNN. Secondly the methods that how to solve the existing problems of conventional object detection are surveyed, mainly analyzing the detection algorithm based on region proposal and based on regression. Thirdly it mentioned some means which improve the performance of object detection. Then the paper introduced some public datasets of object detection and the concept of evaluation criterion. Finally, it combed the current research achievements and thoughts of object detection, summarizing the important progress and discussing the future directions.

Key Words: Convolutional Neural Network, object detection, region proposal, regression, datasets

1. Introduction

With the development of mobile internet and the popularization of various social media, the amount of image data on Internet has increased rapidly, but human beings cannot process efficiently so many image data. So it is expected to carry out these data processing automatically with the aid of computer to solve large-scale visual problems. With a deeper understanding of image processing technology, comprehensive understanding of the image and accurate identification of the target object of the image becomes more and more important [1]. The people not only concern about the classification of images simply, but also want to accurately obtain the semantic category of object and the location in the image [2], so the object detection technology had received wide attention [2, 3] . Object detection technology aims to detect the target objects with the theories and methods of image processing and pattern recognition, determine the semantic categories of these objects, and mark the specific position of the target object in the image [4, 5].

In the actual application, it is a very challenging task to use the computer technology to automatically detect objects. Complex background, noise disturbance, occlusion, low-resolution, scale and attitude changes, and other factors all will seriously affect the object detection performance. The conventional object detection method was based on the hand-crafted feature, it is not robust to illumination change, lacking good generalization ability. And it is acknowledged that progress of object detection has been very slow during 2010-2012 in PASCAL VOC challenge [6], with small gains obtained by building ensemble systems and employing minor variants of traditional methods [7]. For this reason, a variety of methods are proposed to improve the performance of object detection. Convolutional neural network (CNN) [8] as a successful model of deep learning [9], has the ability of hierarchical learning features, and the research [10, 11] shows that the feature extracted by CNN has a stronger ability of discrimination and generalization than hand-crafted feature.[1]

The CNN has achieved a great success in the many areas of computer vision. In 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [12], Hinton and his student Krizhevsky [13] had applied CNN to image classification and achieved top5 error 15.3% vs 26.2% of the conventional method.

After this turning point, CNN became dominant in the later computer vision tasks because of obvious success. Then in 2013, Ross Girshick proposed the R-CNN (Regions with CNN features) [7] method and applied it in object detection successfully. According to current related academic and technical progress, the deep learning method can achieve higher precision and make test-time shorter than previous method.

2. Convolutional Neural Network

In the field of computer vision, feature extraction and classification has always been a very important research direction. In the conventional image processing tasks, the extracted features are often pre-designed features [14] based on statistical regularities or prior knowledge. So it cannot represent the original image information comprehensively and accurately. CNN provides an end-to-end learning model in which parameters can be trained by the gradient descent method. The well-trained CNN can learn the features of the image more fully, and we can regard it as a better black Box to extract feature.

As an important branch of neural network, CNN increases the concept of receptive field and shared weights [8] , which not only greatly reduces the parameters of training, but also reduces the complexity of network model. The features of each layer are generated from previous layer's local area (receptive field) by sharing the weight of

the convolutional kernel. These characteristics make the CNN more suitable for the learning and represent of image features than other neural networks, and it can also keep the translation and scale invariance to a certain extent.

In a typical CNN, the first few layers are usually alternating layers of convolution and pooling, and the final layers of the network near the output layer are usually full-connected network [15] . The training of CNN mainly uses the forward propagation and BP (Back Propagation) algorithm to learn the layer-connection weights, bias and other parameters. The training is a supervised learning process that requires the image data as input and the corresponding labels to optimize the network parameters, finally it will obtain an optimized-weight model.

CNN is composed of different functional layer structure. Typical CNN has convolutional layer, pooling layer and fully-connected layer. However, CNN adds many new layers in the process of evolution and improvement, such as SPP-layer which existed in the SPP-net [16], the ROI(Region of Interest)-pooling layer of Fast R-CNN [17], and the Region Proposal Network (RPN) layer of Faster R-CNN. According to the specific problems, improving the traditional CNN structure can achieve the better performance. In this section, we introduce the basic network structure of a typical CNN and BP algorithm.

2.1 The fundamental structure of CNN

As shown in Figure 1，typical CNN structure mainly consists of the input layer, the convolutional layer, pooling layer, full connect layer and output layer.

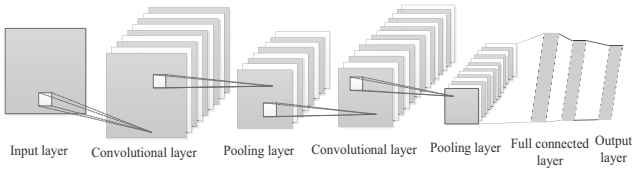

Fig 1: typical CNN structure

The input to the convolutional neural network is usually the original image X_{In} . X_j^l denotes the $l-th$ layer's $j-th$ feature map. We assume X_j^l is a feature map in the convolutional layer，and X_j^l is generated by formula (1):

$$X_j^l = f(\sum_{i \in M_j} X_i^{l-1} * W_{ij}^l + b_j^l) \quad (1)$$

Here, W^l is the l layer feature map corresponding weight matrix，* is symbol of convolution which specific filters for feature maps of $l-1$ layer, the calculated results are added to the bias b_j^l, finally X_j^l is obtained by operations of nonlinear activation function f(x) like Rectified Linear Units (ReLU).

The pooling layer, also called the down-sampling layer, usually follows the convolutional layer, and down sampling the previous feature map according to the fixed rule [18]. The specific rules are: max-pooling, average-pooling, stochastic pooling, overlapped pooling etc. The pooling layer function mainly has two aspects: 1) reducing the dimensionality of feature map; 2) keep scale invariance. Assumed X_j^l is feature map in pooling layer, and pooling operation is described by formula (2):

$$X_j^l = f(\beta_j^i pooling(X_j^{l-1}) + b_j^l) \quad (2)$$

$pooling(x)$ denotes the rule of down-sampling function, β_j^i is the weight of pooling. In general, β is a fixed value，bias b and activation function f(x) is not used .So the general form of pooling operations is formula (3):

$$X_j^l = \beta_j^i pooling(X_j^{l-1}) \quad (3)$$

In a fully-connected network, the feature maps of images are concatenated into a one-dimensional feature vector as input to a fully-connected network. The output of the fully-connected layer can be obtained by making weighted summation to the input and responsed by the activation function [19], shown as formula (4):

$$X^l = f(w^l X^{l-1} + b^l) \quad (4)$$

2.2 Back Propagation

The BP (back propagation) algorithm is used to adjust the weight parameters of neural network. For CNN, the main optimization parameters are convolution kernel parameters, pooling-layer weights, full-connected layer weights and bias parameters. The essence of BP is to compute the partial derivative of the residuals for each layer parameter, and learn an association rule between the residuals and the network weights, then adjust the weight of the network to make the network output closer to the given expected value.

The training goal of the CNN is to minimize the loss function $E(w,b)$ of the network. Common loss functions are MSE (Mean Squared Error) function, NLL (Negative Log Likelihood) function and so on. MSE is defined by formula (5), and NLL is defined by formula (6).

$$MSE(W,b) = \frac{1}{|Y|}\sum_{i=1}^{|Y|}(Y(i) - \bar{Y}(i))^2 \quad (5)$$

$$NLL(W,b) = -\sum_{i=1}^{|Y|}\log Y(i) \quad (6)$$

In the training phase, the residuals are propagated backwards through gradient descent, the trainable parameters of each layer are updated layer by layer in CNN. The learning rate η is used to control the strength of BP, the weight W_i is updated by formula (7),and the bias b_i is updated by formula (8).

$$W_i = W_i - \eta\frac{\partial E(W,b)}{\partial W_i} \quad (7)$$

$$b_i = b_i - \eta\frac{\partial E(W,b)}{\partial b_i} \quad (8)$$

3. Object Detection Based On CNN

In 2013, the organizer of the ImageNet competition added the object detection task with 200 objects in 40,000 images. However, the winner of the competition uses a hand-crafted feature with an only 22.581% mean Average Precision (mAP). Benefitting from deep learning and region proposal algorithm, R-CNN achieve a large improvement which is up to 43.933% mAP in ILSVRC

2014. R-CNN [7] first comp up with the widely used method based on CNN in object detection.

This section analyzes and summarizes the object detection methods based on CNN, and it can be divided into four part: Part 1, introduce the traditional object detection pipeline. Part 2, introduce the CNN object detection framework with region proposal. Part 3, the method that transforms the detection problem into regression problem. Part 4, there are some methods to improve the performance of object detection.

3.1 Traditional Object Detection Method

As shown in Fig 2, the traditional object detection framework is generally divided into four stages: Firstly, generating candidate regions on the given image by sliding window, extracting the relevant features from these regions, using the trained classifier to classify and identify the regions, finally, revising and optimizing the results of detection by NMS (Non Maximum Suppression).

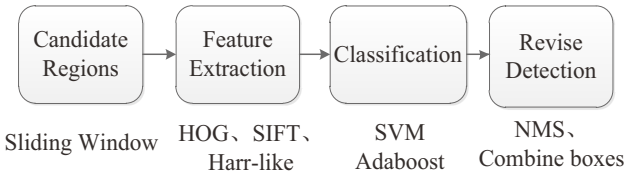

Fig 2: traditional object detection pipeline

(1) Generating Candidate Regions

This stage is to obtain the location of the object. However, the object may appear anywhere in the image. The size and aspect ratio of object is also uncertain, so using sliding window strategy to traverse the entire image with a series of scale and aspect ratio sliding-window. This exhaustive strategy includes most of the possible locations of the object， but there are also some obvious disadvantages: the large complexity of time，too many redundancy windows. That would severely impact the speed and performance of feature extraction.

(2) Feature Extraction

Feature extraction directly influences the design and performance of the classifier. However, it is hard to design a robust feature because of variety of external factors, such as movements of objects, illumination change, complex and changeable environment. Some hand-crafted features are widely used in this period, such as Scale Invariant Feature Transform (SIFT), Histograms of Oriented Gradients (HOG) and Local Binary Patterns (LBP).

(3) Classification

The SVM or AdaBoost classifiers are widely used to classify the extracted features in this stage.

(4) Revise Detection Results

There are still many redundant windows after classification, so it is necessary to remove redundant windows and optimize the detection results by Non-Maximum Suppression (NMS) and combining overlapped Bounding box.

Above all, traditional object detection has two main problems: first, the method based on sliding window is not purposeful enough, high time-complexity, and it has too many redundant windows; secondly, the hand-crafted features are not robust enough for variety of changes.

3.2 Object Detection Based On Regions with CNN features(R-CNN)

In order to solve the problems of redundant windows, region proposal [20] provides a good solutions, which is figure out in advance the possible location of objects in image. Region proposal make good use of the information of image, such as color, edge and texture，which can ensure that the system has a relatively high recall under the conditions of less windows (thousands or even hundreds). It greatly decreases time-complexity of the next actions and quality of region proposals are higher than sliding windows. The common region proposals algorithm primarily adopts selective Search [20] , edge Boxes [21], etc.

With the candidate regions, the rest work is actually the image classification task for the candidate regions (feature extraction plus classification). The success of AlexNet [13] in image classification shows the CNN has strong ability in feature extraction. So in 2014, Ross B. Girshick proposed the region proposal with CNN instead of traditional sliding windows with hand-crafted features, designed R-CNN detection framework. It resulted in great success in object detection and open the door of object detection based on CNN.

In R-CNN [7], there are three modules: first, generate category-independent region proposals. The second module is a large CNN that extracts a fixed-length feature vector from each region. In the last module, the region would be separated into the object and background by specific linear SVMs. In order to improve accuracy for localization, the author trained a linear regression model[7] to modify the coordinates of detection boxes inspired by the bounding box regression in DPM [22].These improvements are obviously successful, in VOC2012 datasets， R-CNN achieve 53.7% mAP than 35.1% mAP of DPM HSC [23].

However, the R-CNN had generated two thousand candidate windows per image while testing. Each region would be extracting features via CNN, resulting in the feature computation time-consuming (50s per image). The another problem is that prevalent CNNs require a fixed input size, so R-CNN fit the input image to the fixed size via cropping[13, 24] or warping[7, 25]. But it would cause a loss of image information, such as aspect ratio and the scale.

He Kaiming et al. [16] proposed the SPP-Net to overcome the existing defects in R-CNN, SPP-net runs the convolutional layer only once on entire image to get the feature maps. Because of sharing computation, it makes the test time shorter than the R-CNN 24-64 times. The proposal regions have different size, but the full-connected layer required a fixed input size. So He Kaiming proposed the Spatial Pyramid Pooling (SPP)-layer. SPP-layer is putted behind to the last convolutional layer, the position of the candidate windows obtained by the selective search relative to the original image is mapped to the last convolutional layer features maps. Then the patch features would be pooling by a multi-level spatial pyramid and generates a fixed-length feature vector representation for each window. These representations are provided to the fully-connected network. SPP-Net can preserve both the

global and local information of image, so it had a higher mAP than R-CNN.

Fast R-CNN proposed the Region of Interest (RoI) pooling layer based on the SPP-layer and it can be regard as a single-level SPP-layer, it also can process the image of various scale. The Fast R-CNN method has the following advantages relative to SPP-net:

- Expect region proposal, Fast R-CNN is an end-to-end training object detection framework, instead of SPP training in a multi-stage pipeline (feature extraction, SVM classifier and bounding box regression).
- Fast R-CNN uses multi-task loss on each labeled RoI to jointly train for classification and bounding-box regression.
- SPP-net cannot update the convolutional layers while training, however, the Fast R-CNN can update all network layers. And for an object detection task, fine-tuning the parameters of convolutional layers is also necessary and important, because the preceding convolutional layers can preserve more location information.

Region proposal computation is still a bottleneck in the Fast R-CNN, high-quality region proposals directly impact the speed and accuracy of object detection. Ren et al. [17] proposed a Region Proposal Network (RPN) that can make object detection nearly be real-time. RPN is a Fully Convolutional Network (FCN) that can simultaneously predict object bounds and objectness scores at each position. The core idea of RPN is using CNN to generate region proposal with a high recall directly, it not only accelerated detecting speed (sharing computation) and but also can be combined with CNN for an end-to-end training. Faster R-CNN has a 73.2% mAP on VOC2007 and can reach a detection speed of about 5 images per second during testing.

Faster R-CNN [17] had merged region proposal and CNN classification to an end-to-end detection network，having a good improvement in both speed and accuracy. However, Faster R-CNN still cannot meet the requirements of real-time object detection (about 20 images per second) with high accuracy. Because the computation of CNN feature extraction is still time-consuming, so this is a huge potential improvement to find and research.

In general, R-CNN, SPP-Net, Fast R-CNN and Faster R-CNN is proposed and developed one by one, the object detection that based on CNN had a simpler pipeline, a higher precision and less time. The R-CNN series of methods have been a very important branch of object detection at present.

3.3 Object Detection Based On Regression

Faster R-CNN method is currently the major of object detection, but the speed cannot meet the real-time requirements. So the importance of You Only Look Once (YOLO) [26] methods has been rising up. These approaches adopt the idea of regression, that is dividing the input image into several cells and each cell predicts boxes and class probability.

YOLO transforms the detection problem to regression problem and accelerate the speed of detection greatly, it can process 45 images per second. And YOLO makes predictions by global information of image and make good use of the context, so the percentage of false positive sharply declines. But YOLO also has the problem: the average precision is only 63.4% mAP on VOC2007 without the region proposal.

SSD [27] used a combination of the regression idea of YOLO and the anchor [17] method of Faster R-CNN, making predictions from feature maps of different scales. It not only achieved high detection accuracy like Faster R-CNN, but also guaranteed the real-time in speed like YOLO.SSD have a mAP 72.1% on VOC2007 test with 58 FPS in GPU, it makes possible to take a real-time object detection with high accuracy for practical applications.

3.4 The Methods to Improve Performance

The methods based on region proposal and based on regression are two important research directions of object detection. But beyond that, the researchers also proposed some excellent methods to make better performance, the following are the details.

(1) **Redesign feature extraction structure.** Faster R-CNN or SSD generally adopted the ZF, VGG [28] network that stacked by convolutional layer and pooling layer to extract feature. With the increased depth of network, the time spent would be increased due to more feature computation. It can even cause performance degradation of network, because the convolutional feature extraction of each layer is accompanied by the loss of the information of previous layer. So PVANET [29] adopted advanced module (C-ReLU [30]、Inception [31]) instead of general convolutional and pooling layer, and using Residual connection[32] to avoid the performance degradation with deeper network. C-ReLU is used in first several layers (to conv3_4) of CNN to reduce computation by half without losing accuracy. Inception building blocks is putted in the rest of feature extraction network. Inception module can extract abundant features because of wider structure and reduce the loss of feature information. And the PVANET can achieve a higher mAP (84.9%) with a pretty good speed.

(2) **Online Hard Example Mining (OHEM).** OHEM [33] applied hard example mining in the SGD (stochastic gradient descent) algorithm, Fast R-CNN automatically select the appropriate region proposals to train as positive and negative examples in the training process. The experimental results suggest that using OHTM can raise 4% mAP of Fast R-CNN on VOC2007 and VOC2012 test, and OHEM is also adopted by R-FCN for good performance.

(3) **Concatenation of multi-scale features**. Fast R-CNN and Faster R-CNN both use the feature of last layer to take detection. But the high-layer features lose much detail information because of pooling and it would affect the accuracy of lactation. HyperNet [34], PVANET et al. make use of multi-scale features, not only exploiting high-layer semantic information to classify, but also considering the low-level textural features to more accurate location.

(4) **Use of context information.** The researches [35, 36] suggest that it would achieve better detecting performance with using the context of region proposal with extracting features of region proposal. In ILSVRC2016, GBD-Net [37]

expanded the window on candidate boxes to use the contextual visual information, it raised 2.2% mAP with basic network of ResNet-269 on ImageNet DET dataset.

4. Datasets and Evaluation

Deep Neural Network needs large amount of labeled data for training model, nowadays the most common used datasets of object detection are ImageNet, PASCAL VOC and MS COCO. PASCAL VOC provides a standard image labeling and evaluation system. PASCAL VOC image dataset include 20 categories, the dataset has a high-quality and labeled completely image what is very suitable for examining the algorithm performance. For object detection, ImageNet provides an important source of data for object detection because of a specific bounding labeling training set. COCO is sponsored by Microsoft, the annotation includes categories, location information and semantic text description of image. The open-source of COCO dataset also make contributions to the improvement of object detection.

The results of the object detection can be expressed as follow: Input image $\boldsymbol{I}$ is tested by the detector, obtain the bounding box $\boldsymbol{B}$ of each object, corresponding category label $\boldsymbol{c}$ and confidence level $\boldsymbol{f}$. The evaluation of Multi-Objects detection in the same image is considered as separate detection results, the ground truth boxes $\boldsymbol{B_g}$. If the predicted bounding box satisfies the following formula (9), it is regarded as correct.

$$a = \frac{area(B \bigcap B_g)}{area(B \bigcup B_g)} \geq a_0 \tag{9}$$

Here, $\boldsymbol{a}$ is an evaluation parameter Intersection Over Union (IOU)，representing the overlap rate of the ground truth box and object window predicted by the detector. a_0 is a previously set threshold value, the general value is 0.5.

To calculate average precision (AP), we need to know some concepts, shown in Table 1 below.

Table 1: Concepts about binary classification

Concepts	Explanation
True Positive (TP)	The number of samples which the true is predict as positive
True Negative (TN)	The number of samples which the true is predict as negative
False Positive (FP)	The number of samples which the false is predict as positive
False Negative (FN)	The number of samples which the false is predict as negative

Precision is defined as the following formula (10):

$$P = \frac{TP}{TP + FP} \tag{10}$$

Recall is defined as the following formula (11):

$$R = \frac{TP}{TP + FN} \tag{11}$$

Accuracy is defined as the following formula (12):

$$AC = \frac{TP}{TP + FN + TN + FP} \tag{12}$$

The Precision Recall (PR) curve was drawn by the recall against corresponding precision, as shown Figure 3 [38], and the value of average precision (AP) is the area value under of the curve, it can be calculate by integral of function of precision and recall.

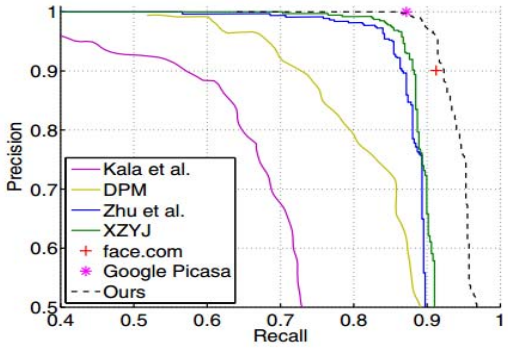

Fig 3: The PR curve

For assessing overall system performance, it is commonly using mean Average Precision (mAP). mAP is calculated by using the following formula (13).

$$mAP = AP/N \tag{13}$$

Here, N is the numbers of object category. The formula represents that calculating the AP for each category and averaging the AP of all categories. In addition, the training time of model, storage space, execution efficiency and transfer ability of different scenarios et al. are also important criteria to evaluate the performance of object detection algorithm.

5. Conclusion

The paper focused on the object detection based on CNN, the structure of CNN, the framework of object detection based on CNN and the methods of improving detection performance are introduced. CNN has strong ability in feature extraction, it can compensate for the drawback existing in hand-crafted features. CNN also has better advantage than conventional methods on real-time, accuracy, adaptability, but it still has lots of room for improvement. Improving the structure of CNN can reduce the loss of feature information, fully utilizing the relations of object and context and building the fuzzy inference can make the computer better deal with the problems like occlusion and low-resolution. Enhance of intelligence and the practicability of object detection based on CNN are the key point in future research.

References:

[1] C. Szegedy, A. Toshev and D. Erhan, Deep Neural Networks for object detection, *Advances in Neural Information Processing Systems*, 2013, 26: 2553-2561.

[2] K.Q. Huang, W.Q. Ren and T.N. Tan, A review on image object classification and detection, *Chinese Journal of Computers*, 2014.

[3] X. Zhang, Y.H. Yang, Z. Han, H. Wang, and C. Gao, Object class detection: A survey, *ACM Computing Surveys*, 46(1): 28-36, 2013.

[4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, ImageNet Large Scale Visual Recognition Challenge, *International Journal of Computer Vision*, 115(3): 211-252, 2015.

[5] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick, Microsoft COCO: Common Objects in Context, 8693: 740-755, 2014.
[6] D. Hoiem, S.K. Divvala and J.H. Hays, Pascal VOC 2008 Challenge, *WORLD LITERATURE TODAY*, 2009.
[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, *Computer Science*: 580-587, 2014.
[8] N. Chumerin, convolutional neural network, 2015.
[9] Y. Lecun, Y. Bengio and G. Hinton, Deep learning., *Nature*, 521(7553): 436-44, 2015.
[10] A.S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, CNN Features Off-the-Shelf: An Astounding Baseline for Recognition, 512-519, 2014.
[11] H. Kataoka, K. Iwata and Y. Satoh, DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition, *Computer Science*, 50(1): 815-830, 2013.
[12] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and F.F. Li. ImageNet: A large-scale hierarchical image database, in Proc. Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. 2009, 248-255.
[13] A. Krizhevsky, I. Sutskever and G.E. Hinton. ImageNet classification with deep convolutional neural networks, in Proc. International Conference on Neural Information Processing Systems. 2012, 1097-1105.
[14] Nixon and Mark. *Feature Extraction & Image Processing for Computer Vision (Third Edition)*: Publishing House of Elec, 2013.
[15] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, 86(11): 2278-2324, 1998.
[16] K. He, X. Zhang, S. Ren, and J. Sun, Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition, *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 37(9): 1904-16, 2015.
[17] 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 & Machine Intelligence*: 1-1, 2016.
[18] H. Wu and X. Gu. *Max-Pooling Dropout for Regularization of Convolutional Neural Networks*: Springer International Publishing, 2015.
[19] J. Bouvrie, Notes on Convolutional Neural Networks, *Neural Nets*, 2006.
[20] J.R. Uijlings, K.E. Sande, T. Gevers, and A.W. Smeulders, Selective Search for Object Recognition, *International Journal of Computer Vision*, 104(2): 154-171, 2013.
[21] C.L. Zitnick and P. Dollár. *Edge Boxes: Locating Object Proposals from Edges*, 2014.
[22] P.F. Felzenszwalb, R.B. Girshick, D. Mcallester, and D. Ramanan, Object detection with discriminatively trained part-based models., *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 32(9): 1627, 2014.
[23] X. Ren and D. Ramanan. Histograms of Sparse Codes for Object Detection, in Proc. Computer Vision and Pattern Recognition. 2013, 3246-3253.
[24] M.D. Zeiler and R. Fergus, Visualizing and Understanding Convolutional Networks, 8689: 818-833, 2013.
[25] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition, *Computer Science*, 50(1): 815-830, 2013.
[26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, You Only Look Once: Unified, Real-Time Object Detection, *Computer Science*, 2015.
[27] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, and A.C. Berg, SSD: Single Shot MultiBox Detector, 2015.
[28] K. Simonyan and A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, *Computer Science*, 2014.
[29] K.H. Kim, S. Hong, B. Roh, Y. Cheon, and M. Park, PVANET: Deep but Lightweight Neural Networks for Real-time Object Detection. 2016.
[30] W. Shang, K. Sohn, D. Almeida, and H. Lee, Understanding and Improving Convolutional Neural Networks via Concatenated Rectified Linear Units. 2016.
[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions, in Proc. Computer Vision and Pattern Recognition. 2014, 1-9.
[32] K. He, X. Zhang, S. Ren, and J. Sun, Deep Residual Learning for Image Recognition, 770-778, 2015.
[33] A. Shrivastava, A. Gupta and R. Girshick, Training Region-Based Object Detectors with Online Hard Example Mining, 2016.
[34] T. Kong, A. Yao, Y. Chen, and F. Sun, HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection, 845-853, 2016.
[35] S. Gidaris and N. Komodakis, Object Detection via a Multi-region and Semantic Segmentation-Aware CNN Model, *Computer Science*: 1134-1142, 2015.
[36] S. Bell, C.L. Zitnick, K. Bala, and R. Girshick. Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks, in Proc. IEEE Conference on Computer Vision and Pattern Recognition. 2016, 2874-2883.
[37] X. Zeng, W. Ouyang, B. Yang, J. Yan, and X. Wang. *Gated Bi-directional CNN for Object Detection*: Springer International Publishing, 2016.
[38] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, Joint Cascade Face Detection and Alignment, 8694: 109-122, 2014.