Assessment of Object Detection Using Deep Convolutional Neural Networks

Ajeet Ram Pathak, Manjusha Pandey, Siddharth Rautaray and Karishma Pawar

Abstract Detecting the objects from images and videos has always been the point of active research area for the applications of computer vision and artificial intelligence namely robotics, self-driving cars, automated video surveillance, crowd management, home automation and manufacturing industries, activity recognition systems, medical imaging, and biometrics. The recent years witnessed the boom of deep learning technology for its effective performance on image classification and detection challenges in visual recognition competitions like PASCAL VOC, Microsoft COCO, and ImageNet. Deep convolutional neural networks have provided promising results for object detection by alleviating the need for human expertise for manually handcrafting the features for extraction. It allows the model to learn automatically by letting the neural network to be trained on large-scale image data using powerful and robust GPUs in a parallel way, thus, reducing training time. This paper aims to highlight the state-of-the-art approaches based on the deep convolutional neural networks especially designed for object detection from images.

Keywords Computer vision · Deep convolutional neural networks
Deep learning · Object detection

A. R. Pathak (✉) · M. Pandey · S. Rautaray
Data Science Center of Excellence, School of Computer Engineering,
Kalinga Institute of Industrial Technology (KIIT) University, Bhubaneswar, India
e-mail: ajeet.pathak44@gmail.com

M. Pandey
e-mail: manjushapandey82@gmail.com

S. Rautaray
e-mail: sr.rgpv@gmail.com

K. Pawar
Department of Computer Engineering & IT, College of Engineering Pune (COEP),
Pune, India
e-mail: kvppawar@gmail.com

S. Bhalla et al. (eds.), *Intelligent Computing and Information and Communication*,
Advances in Intelligent Systems and Computing 673,
https://doi.org/10.1007/978-981-10-7245-1_45

1 Introduction

Computer vision technology has been extensively used in different segments like industry, automation, consumer markets, medical organizations, entertainment sectors, defense, and surveillance, to mention a few. The ubiquitous and wide applications like scene understanding, video surveillance, robotics and self-driving cars triggered vast research in the domain of computer vision during the most recent decade. Visual recognition systems encompassing image classification, localization, and detection have achieved great research momentum due to significant development in neural networks especially deep learning, and attained remarkable performance [1]. The last 4 years witnessed a great improvement in performance of computer vision tasks especially using deep convolution neural networks (DCNNs) [2].

Several factors are responsible for proliferation for DCNNs viz. (i) Availability of large training datasets and fully annotated datasets (ii) Robust GPU to train large-scale neural network models in a parallel way (iii) State-of-the-art training strategies and regularization methods. Object detection is one of the crucial challenges in computer vision and it is efficiently handled by DCNN [3], Restricted Boltzmaan Machine (RBM) [4], autoencoders [5], and sparse coding representation [6]. This paper aims to highlight state-of-the-art approaches for object detection based on the DCNNs.

The contents of the paper are portrayed as follows. Section 2 introduces object detection. The fundamental building blocks of DCNNs from the perspective of object detection are enunciated in Sect. 3. The state-of-the-art DCNN-based approaches for object detection are discussed in Sect. 4. The paper is concluded in Sect. 5.

2 Object Detection

An image or video contains single or more than one classes of real-world objects and abstract things like human, faces, building, scene, etc. The aim of object detection is to determine whether the given instance of the class is present in the image, estimate the location of the instance/instances of the all the classes by outputting the bounding box overlapping the object instance along with obtained accuracy of detection irrespective of partial occlusions, pose, scale, lightening conditions, location, and camera position. It is generally carried out using feature extraction and learning algorithms. Object detection is a preliminary step for various computer vision tasks like object recognition, scene understanding from images and activity recognition, anomalous behavior detection from videos. Detecting instance or instances of the single class of object from image or video is termed as *single object class detection*. *Multi-class object detection* deals with detecting instances of more than one class of objects from the image or video. Following challenges need to be handled while detecting objects from the images.

- Image-based challenges

Many computer vision applications require multiple objects to be detected from the image. Object occlusions (partial/full occlusion), noise, and illumination changes make detection challenging task. Camouflage is a challenge in which object of interest is somewhat similar to the background scene. This challenge needs to be handled in surveillance applications. It is also necessary to detect objects under conditions of multiple views (lateral, front), poses, and resolutions. The object detection should be invariant to scale, lighting conditions, color, viewpoint, and occlusions.

- Processing challenges

Detecting objects at large scale without losing accuracy is a primary requirement of object detection tasks. Some applications require robust and efficient detection approaches, whereas others require real-time object detection. Thanks to specialized hardware like GPUs and deep learning techniques which allow to train multiple neural networks in parallel and distributed way helping to detect objects at real-time.

3 Building Blocks of Convolutional Neural Network

DCNNs was first used for image classification. After achieving state-of-the-art performance in image classification, DCNN has been used for more complex tasks like object detection from images and videos.

3.1 CNN Architecture

Convolutional neural network (CNN) is a kind of feedforward neural network in which the neurons are connected in the same way as the neurons present in the brain of animal's visual cortex area. Figure 1 shows the architecture of CNN. Being hierarchical in nature, CNN encompasses convolution layer with activation function like Rectified Linear Unit (ReLU), followed by pooling layer and eventually fully connected layers. The pattern of CONV-ReLU-POOL is repeated in such a manner that image reduces spatially. The neurons are arranged in the form of three dimensions—width, height, and depth. Depth corresponds to color channels in the image. The image to be recognized is fed as input in terms of [$width \times height$ $depth$]. For the desired number of filters, the image is convolved with the filter function in order to get the specific feature. This process is repeated for the desired number of filters and accordingly feature map is created. This is done by applying the dot product of the weight assigned to the neuron and the specified region in the image. In this way, the output of a neuron is computed by convolution layer.

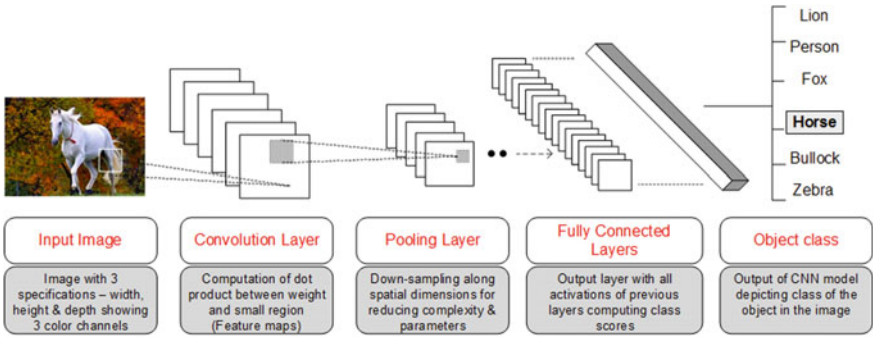

Fig. 1 Architecture of convolutional neural network

Rectified Linear Unit (ReLU) applies activation function at elemental level making it nonlinear. Generally, convolution layer is followed by pooling layer to reduce the complexity of the network and the number of parameters in learning by down-sampling the feature map along spatial dimensions. The last layer in the CNN is a fully connected layer which gives the output of the image recognition task in the form of scores representing the object classes. Highest score represents the presence of a corresponding class of the object in the image.

3.2 Pooling Layers

Addition of pooling layer amidst the consecutive layers of convolutional layer reduces the number of parameters and complexity of the network, and thus, control overfitting. Pooling layer is translation-invariant and it takes activation maps as input and operates on every patch of the selected map. There are various kinds of pooling layers.

- Max pooling: In this pooling, each depth slice of input is operated using pooling. Figure 2 shows working of max pooling where a filter of size 2×2 is applied over a patch of an activation map with a stride of 2. The max value among each entry of 4 numbers is chosen and stored into the matrix, getting a spatially resized map. Another pooling approach is average pooling which takes an average of neighborhood pixels. Max pooling gives better results compared to average pooling [7].
- Deformation constrained pooling (Def-pooling): In order to apply deformation of object parts along with geometric constraints and associated penalty, def-pooling is applied [8]. It has the ability to learn deformable properties of object parts and shares visual patterns at any level of information abstraction and composition.

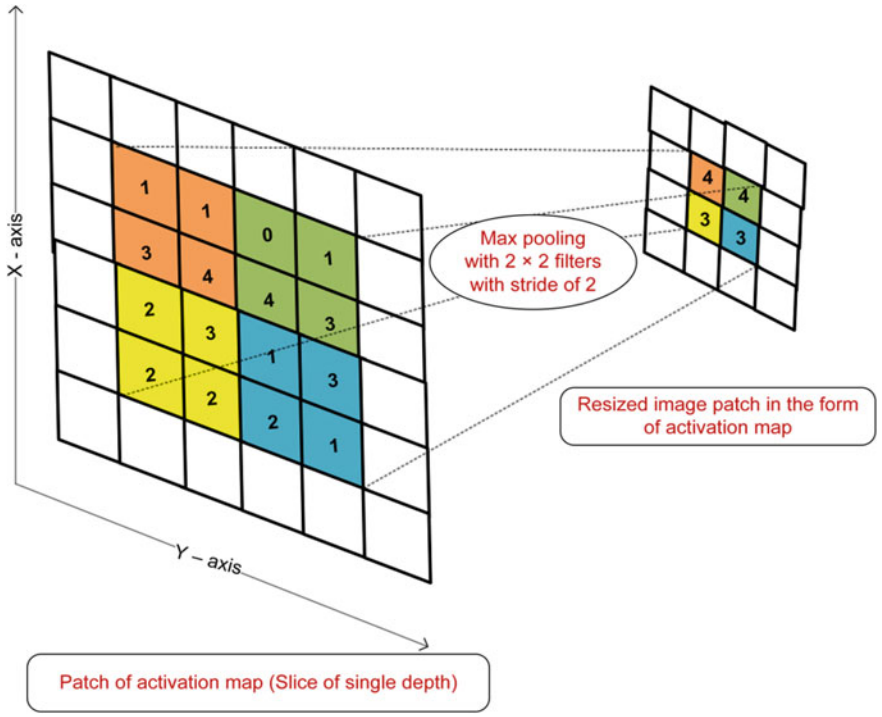

Fig. 2 Max pooling over the slice of single depth

- Fully connected layers: These layers perform high-level reasoning in CNN. They exhibit a full connection to all the decision functions in the previous layer and convert 2D features into one-dimensional feature vector. Fully connected layers possess a large amount of parameters and so require powerful computational resources.

3.3 Regularization

"*Regularization is defined as any modification we make to a learning algorithm that is intended to reduce its generalization error but not its training error* [9]." Due to a large amount of parameters for training and ability of architectures to learn more abstractions using deep learning, there are chances of obtaining negative performance on test data from the model, i.e., model learns too well such that it poorly generalizes in case of new data. This is known as overfitting. Regularization strategies are required in order to stop overfitting.

4 State-of-the-Art Object Detection Approaches Using DCNN

Table 1 compares the state-of-the-art discussion of DCNN-based approaches for object detection. DCNNs have been extensively used for image classification and achieved state-of-the-art results [1].

For the very first time, DCNNs have been used for object detection by Szegedy et al. [10]. The authors have formulated object detection as a regression problem for object bounding box masks and defined object detection as estimation and localization of class and object from the image, respectively. Their approach is known as DetectorNet in which the last layer of AlexNet [1] architecture is replaced with regression layer in order to localize the objects using DNN-based object mask regression. To precisely detect the multiple instances of the same object, DetectorNet applies multi-scale box inference with refinement procedure. But, this approach lacks multiple classes of objects for detection since it uses only single mask regression.

To demystify the working of features extracted in CNN model and diagnose the errors associated with the model, Zeiler and Fergus (ZFNet) [3] put forth a novel visualization technique based on multilayered deconvolution network (deconvnet). This model used deconvnet in order to project features back into pixel space of the image.

Deformable DCNN (DeepID-Net) encompasses feature representation learning, part deformation learning, context modeling, model averaging, and bounding box location refinement [11] and uses cascaded CNN for object detection. It works on deformable part objects. Regions with CNN features (R-CNN) [12] take an input image and evaluate '*n*' number of bottom-up region proposals using segmentation. Once region proposals are obtained, it classifies proposals using class-relevant SVMs to get classified regions. This method acts as a baseline model for a large number of approaches put forth for object detection. Fast R-CNN [13] is the extended version of R-CNN to improve the speed of training and testing phase and improve the detection accuracy. Fast R-CNN suffers from the drawback of calculating the proposals for each region in the image, thus, incurring the large cost of computation, this drawback has been removed in Faster R-CNN by Ren et al. [14]. The authors put forth region proposal network (RPN) which is a fully convolution network in which input image is shared with the detector network, this network simultaneously calculates object bounds and object features at each point, thus freeing cost of region proposals.

This method merges RPN with fast R-CNN and creates a unified network. It works on the principle of "attention mechanism" in which RPN guides network where to search for the object bounds. In the paper, by Markus et al. [15], multiple CNN models are used to detect objects at multiple scales. The papers by Lee et al. [16] and Cheng et al. [17] are based on the region-based proposals. Lee et al. [16] handled the issue of intra-class and interclass variability among objects using multi-scale templates of CNN and non-maximum suppression method. On the other

Table 1 Comparative study of State-of-the-approaches for object detection

Paper	Approach	Issues	Features	CNN configuration	Pooling	Classifier	Regularization technique	Deep network training
ImageNet [1]	CNN	Reducing detection errors	Image classification for high-resolution images	5 CNN layers + 3 Fully conn. Layers	Max pooling	Softmax	DropOut	Supervised learning mode, Local response normalization and stochastic gradient descent
ZFNet [3]	CNN and deconv. Neural network	Demystifying the working of intermediate CNN layers and classifier	Visualization of intermediate features in the network	Fully supervised CNN models	Max pooling	Softmax	DropOut, ReLU as activation function	Back-propagating derivative of cross entropy function for training and SGD
DetectorNet [10]	CNN	Object detection and localization	Multi-object detection, DNN-based regression for localization	5 CNN layers + 2 Fully conn. Layers	Max pooling	Softmax	Use of negative classes to regularize the network	Adaptive gradient algorithm based on Stochastic optimization
DeepID-Net [11]	Cascaded CNN	Handling of deformable part objects	Generic object detection	13 CNN layers + 3 Fully conn. Layers (O-Net and T-Net)	Def-pooling	SVM	Pre-training scheme & jointly learning feature representation	Multi-stage deep training model
Faster R-CNN [14]	Region-based proposal network	Computation cost of region-Based proposal handling translation	Cost-free region proposals, Real-time object	VGG-Net (13 CNN & 3 fully conn. layers) + ZF Net	RoI pooling layer	R-CNN acts as a classifier	"Image-centric" sampling strategy based on back-propagation and SGD	"Attention" mechanism of Neural network, back-propagation

(continued)

Table 1 (continued)

Paper	Approach	Issues	Features	CNN configuration	Pooling	Classifier	Regularization technique	Deep network training
Markus et al. [15]	Multi-scale model	Visual Pedestrian detection with high accuracy	Multi-scale person detection	3 stages of CNNs processing image patches	Max pooling	Non-maximum suppression	DropOut	SGD with mini-batch training
RIFD-CNN [17]	Region-based proposal (Fisher discriminative)	Object rotation, intra-class and interclass similarity of objects	Rotation invariant, interclass & intra- class object detection	AlexNet + VGG-Net	Max pooling	Softmax, Linear Support Vector Machine (SVM)	Discrimination regularization	CNN-based pre-training and fine-tuning using SGD
MSS-CNN [18]	CNN working on image pyramids	Handling of scale variation	Multi-scale, context-aware object detection and localization	Image pyramid-based CNN	Max pooling	structured SVM, bounding box regression	Training based on the data augmentation	Multi-scale training, non-max suppression to resolve multiple detection issue

hand, in RIFD-CNN by Cheng et al. [17], the issue of object rotation and intra-class and interclass variability is handled by introducing rotation-based layer and Fisher discriminative layer in the network, respectively. For detecting small objects and localizing them, contextual information based on the multi-scale model of CNN is used in [18], handling the issue of variation in scaling of objects. The aforementioned approaches mainly focus on the specific challenge in object detection viz. multi-scale model, fast and real-time detection, detection accuracy, and localization, interclass and intra-class variation of objects.

It is worth important to amalgamate challenges and address them using unified object detection framework applicable to detect objects in different complex scenarios and thereby enhance the usability of such object detection systems.

5 Conclusion

This paper compares some of the noteworthy approaches to object detection based on DCNNs. DCNN-based approaches are found to be suitable for images and can also be applicable to detect moving objects from the video [19]. The need of the hour is to develop object detection model which can be generalized to work in different application scenarios like face recognition, emotion detection, abandoned object detection (Suspicious object detection), etc. The role of "transfer learning" method for training deep networks would help to cope with the issue [20].

The efficacy of object detection frameworks mainly depends on the learning mode, method of processing the images (parallel programming) and also the platform (CPU, GPU). Continuous change in the scene implies a change in the behavior of objects to be detected, therefore, it is mandatory for such systems to continuously learn the multitude features of objects and detect them despite of a change in their orientation, views, and forms. In addition to this, real-time detection of objects [21] helps to take proactive measures or acts as alarming conditions for effectively monitoring and controlling the public and private places requiring utmost security.

Object detection is a very promising area which can be applied in computer vision and robotics systems, surveillance based on the drone cameras, etc. It is extremely useful in places like deep mines, expeditions to exploring deep ocean floor where human presence is not feasible.

References

1. Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. (2012).
2. He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In: European Conference on Computer Vision, pp. 346–361. Springer International Publishing. (2014).

3. M.D. Zeiler, R. Fergus.: Visualizing and understanding convolutional neural networks. In: ECCV. (2014).
4. R. Salakhutdinov, G.E. Hinton.: Deep boltzmann machines. In: AISTATS, (2009).
5. S. Rifai, P. Vincent, X. Muller, et al.: Contractive auto-encoders: explicit invariance during feature extraction. In: ICML (2011).
6. Yang, Jianchao, Kai Yu, Yihong Gong, and Thomas Huang.: Linear spatial pyramid matching using sparse coding for image classification. In: Computer Vision and Pattern Recognition, CVPR. pp. 1794–1801. IEEE. (2009).
7. D. Scherer, A. Müller, S. Behnke.: Evaluation of pooling operations in convolutional architectures for object recognition. In: ICANN. (2010).
8. W. Ouyang, P. Luo, X. Zeng, et al.: Deepid-net: Deformable deep convolutional neural networks for object detection. In: Computer Vision and Pattern Recognition, pp. 2403–2412. IEEE. (2015).
9. I. Goodfellow, Y. Bengio, and A. Courville.: Deep Learning. MIT Press. (2016).
10. C. Szegedy, A. Toshev, D. Erhan.: Deep neural networks for object detection. In: Proceedings of the NIPS. (2013).
11. W. Ouyang, P. Luo, X. Zeng, et al.: DeepID-Net: multi-stage and deformable deep convolutional neural networks for object detection. In: Proceedings of the CVPR. (2015).
12. R. Girshick, J. Donahue, T. Darrell, et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the CVPR. (2014).
13. R. Girshick.: Fast R-CNN. In: ICCV. (2015).
14. S. Ren, K. He, R. Girshick, and J. Sun.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: TPAMI, pp. 91–99. IEEE. (2016).
15. Eisenbach, Markus, Daniel Seichter, Tim Wengefeld, and Horst-Michael Gross.: Cooperative multi-scale Convolutional Neural Networks for person detection. In: International Joint Conference on Neural Networks (IJCNN), pp. 267–276. IEEE. (2016).
16. B. Lee, E. Erdenee, S. Jin, and P. K. Rhee. Efficient object detection using convolutional neural network-based hierarchical feature modeling. In: Signal, Image Video Process. vol. 10, no. 8, pp. 1503–1510, (2016).
17. Cheng, Gong, Peicheng Zhou, and Junwei Han.: RIFD-CNN: Rotation-invariant and fisher discriminative convolutional neural networks for object detection. In: Computer Vision and Pattern Recognition, pp. 2884–2893. IEEE. (2016).
18. E. Ohn-Bar and M. M. Trivedi.: Multi-scale volumes for deep object detection and localization. In: Pattern Recognition, vol. 61, pp. 557–572. Elsevier (2017).
19. S. H. Shaikh, K. Saeed, and N. Chaki.: Moving Object Detection Approaches, Challenges and Object Tracking. In: Moving Object Detection Using Background Subtraction, pp. 5–14. Springer International Publishing (2014).
20. Dauphin, G.M. Yann, X. Glorot, S. Rifai, Y Bengio, I. Goodfellow, E. Lavoie, X. Muller et al.: Unsupervised and transfer learning challenge: a deep learning approach. In: ICML Workshop on Unsupervised and Transfer Learning, pp. 97–110. (2012).
21. P. Viola, M. Jones.: Rapid object detection using a boosted cascade of simple features. In: Computer Vision and Pattern Recognition. pp. I-511-I-518. IEEE. (2001).