Review of Deep Learning Techniques for Object Detection and Classification

Mohd Ali Ansari(✉) and Dushyant Kumar Singh

Motilal Nehru National Institute of Technology Allahabad, Allahabad, India
mohdaliucer@gmail.com, dushyant@mnnit.ac.in

Abstract. Object detection and classification is a very important integrant of computer vision domain. It has its role in various sectors of life as security, safety, fun, heath & comfort etc. Under safety and security, surveillance is one critical application area where, Object detection has gained the growing importance. Object in such case could be human being and other suspicious and sensitive objects. Correct detection and classification on accuracy measures is always a challenge in these problems. Now days, deep learning techniques are getting utilized as an effective and efficient tool for different classification problems. Looking over these facts, a review of available deep learning architectures has been presented in this paper, for the problem of object detection and classification. The classification models considered for review are AlexNet, VGG Net, GoogLeNet, ResNet. The dataset used for experimentation is Caltech-101 dataset and the standard performance measures utilized for evaluation are True Positive Rate (TPR), False Positive Rate (FPR) and Accuracy.

Keywords: Object detection · Classification · Deep learning
Convolutional Neural Network (CNN) · AlexNet · VGG net · ResNet

1 Introduction

Computer vision provides the ability to the machine to see and gather information from the environment. This field contains methods for acquiring, processing and analyzing the images, to be able to extract important information from them. Recently in computer vision, a lot of research has been seen for classification and recognition of objects in images and videos. Many applications are using object classification and recognition technique to solve the real world problem.

Frame-differencing and Background Subtraction are the two major techniques for object detection in an image or video. Noises are the biggest reason due to which the efficiency of these approaches is affected most. Due to the noise and motion, in frame differencing it creates a lot of data; there is an added difficulty in differencing images, as the noise has similar properties in different images or videos. In case of Background subtraction, due to motion in the background, it's become difficult to identify which part of an image is background, which makes the efficiency lower. Other approaches work on object features and a classifier. In this approach firstly extract some feature from the object after that using some classifier technique to classify the objects on the basis of extracted feature [10].

S. Verma et al. (Eds.): CNC 2018, CCIS 839, pp. 422–431, 2019.
https://doi.org/10.1007/978-981-13-2372-0_37

In object detection technique the toughest part is to detect and identify the features in the raw input data and on the basis of that feature it detects objects. While in deep learning there is no manual step for finding the feature of an object. In deep learning, at the time of training, it discovers the most useful. In deep learning, there is no need to select any special feature to classify and for the detection of the object. In comparison to other classification and detection technique, deep learning has better accuracy if using sufficient amount of depth in the classification model.

2 Related Work

There are several approaches proposed by the researcher using different techniques of classification and recognition.

Krizhevsky et al. [1] proposed the technique for object classification. They perform classification task on 1.28 million images that belong to 1000 classes. In this technique, they use CNN for object classification. They use 5 convolutional layers and 3 fully-connected layers. They use different filter size at different convolutional layer with the different stride. AlexNet obtains 57.0% accuracy for top-1 while for top-5 it obtains 80.3% accuracy. Simonyan and Zisserman [2] perform classification task on 1.3 million images that belong to 1000 classes. In this technique, they use CNN for object classification. They make the network that contains 19 layers out of which 16 are the convolutional layer and 3 are the fullyconnected layer. They use very small filter size to all convolutional layer with one stride. VGG obtains 70.5% accuracy for top-1 while for top-5 it obtains 90.0% accuracy.

Szegedy et al. [3] proposed the technique for object classification. In this technique, they use inception module for object classification. They make the network that contains 22 layers. They use 1×1, 3×3, 5×5 filters to convolutional layer. GoogLeNet obtains 68.7% accuracy for top-1 while for top-5 it obtains 88.9% accuracy He et al. [4] make deeper neural network for more accurate object classification. They present a residual network to training that are substantially deeper than those used previously ResNet can get more accuracy as we increase depth. ResNet trained on imagenet dataset that contain approx 1.2 million images with approximately 2000 classes. Resnet-152 obtains 80.62% accuracy for top-1 while for top-5 it obtains 95.51% accuracy.

3 Deep Learning Models

CNN is composed of multiple layers; each layer has specific work to do. To extract useful information pass the input through the layers [7]. CNN contains multiple layers each layer have some parameters that are trained on the data set, CNN automatically extracts most useful information or feature. CNN is better to work with images.

3.1 AlexNet

This model is trained on a subset of the ImageNet database [1], which is used in ImageNet Large-Scale Visual Recognition Challenge (ILSVRC). The model is trained on more than a million images and can classify images into 1000 object categories. As the winner of ILSVRC 2012, the AlexNet architecture has about 650 thousand neurons and 60 million parameters. AlexNet includes five convolutional layers, two normalization layers, three maxpooling layers, three fullyconnected layers, and a linear layer with softmax activation function in the output. Moreover, it uses the dropout regularization method to reduce overfitting in the fullyconnected layers and applies Rectified Linear Units (ReLUs) for the activation of those and the convolutional layers (Fig. 1).

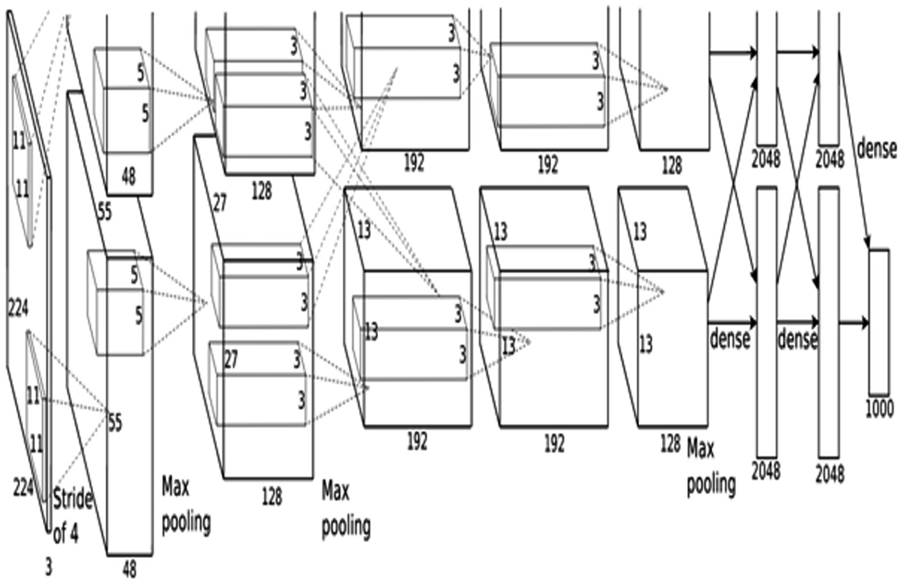

Fig. 1. AlexNet CNN architecture [1].

3.2 GoogLeNet

The GoogLeNet architecture was first introduced by Szegedy et al. in their 2014 [3]. GoogLeNet is an inception architecture that enables one to increase the width and depth of the network for an improved generalization capacity per a constant computational complexity. GoogLeNet architecture involves 6.8 million parameters with nine inception modules, two convolutional layers, one convolutional layer for dimension reduction, two normalization layers, four max-pooling layers, one average pooling, one fullyconnected layer, and a linear layer with softmax activation function in the output. Each inception module in turn contains two convolutional layers, four convolutional layers for dimension reduction, and one maxpooling layer. GoogLeNet also uses dropout regularization in the fullyconnected layer and applies the ReLU activation function in all of the convolutional layers (Fig. 2).

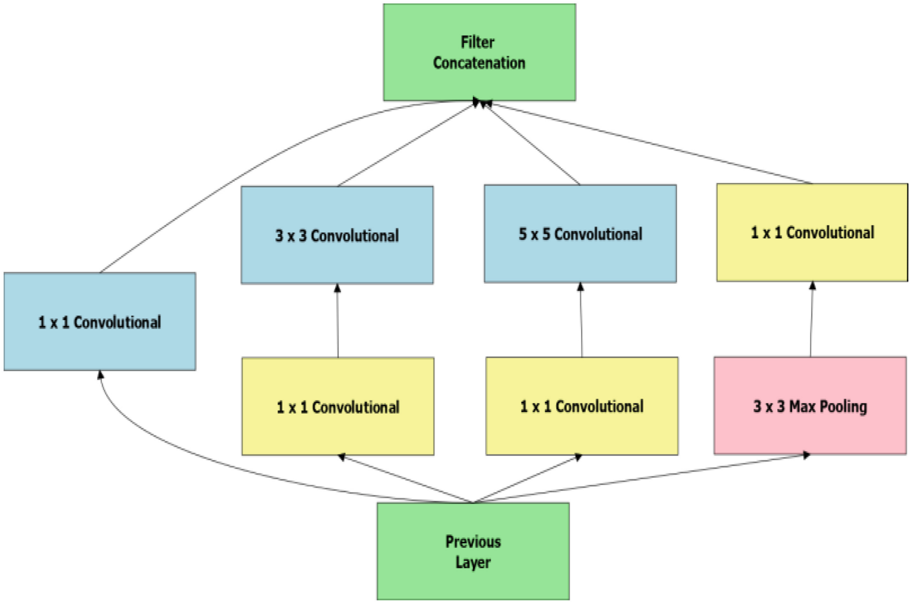

Fig. 2. GoogLeNet inception model [3].

3.3 VGG

The VGG network architecture was introduced by Simonyan and Zisserman [2]. The largest VGGNet architecture involves 144 million parameters from 16 convolutional layers with very small filter size of 3×3, five max-pooling layers of size 2×2, three fullyconnected layers, and a linear layer with Softmax activation function in the output. This model also uses dropout regularization in the fullyconnected layer and applies ReLU activation to all the convolutional layers. In Table 1 FS stands for Filter Size while CL stands for convolution layer.

3.4 ResNet

The ResNet architecture was first introduced by He et al. in their 2015 [5]. ResNet is a classification model that is totally different from our previous models. In ResNet author use very deep network to train model. When they use very deep neural network then they expected high accuracy but in reality the training error increased. To overcome the training error problem author uses the residual model. In Table 2 FS stands for Filter Size while CL stands for convolution layer.

Table 1. VGG CNN architecture [2].

VGG16	VGG19
16 weight layer	19 weight layer
Input (224 × 224 RGB image)	
3 × 3 FS-64 CL 3 × 3 FS-64 CL	3 × 3 FS-64 CL 3 × 3 FS-64 CL
Maxpool	
3 × 3 FS-128 CL 3 × 3 FS-128 CL	3 × 3 FS-128 CL 3 × 3 FS-128 CL
Maxpool	
3 × 3 FS-256 CL 3 × 3 FS-256 CL 3 × 3 FS-256 CL	3 × 3 FS-256 CL 3 × 3 FS-256 CL 3 × 3 FS-256 CL 3 × 3 FS-256 CL
Maxpool	
3 × 3 FS-512 CL 3 × 3 FS-512 CL 3 × 3 FS-512 CL	3 × 3 FS-512 CL 3 × 3 FS-512 CL 3 × 3 FS-512 CL 3 × 3 FS-512 CL
Maxpool	
3 × 3 FS-512 CL 3 × 3 FS-512 CL 3 × 3 FS-512 CL	3 × 3 FS-512 CL 3 × 3 FS-512 CL 3 × 3 FS-512 CL 3 × 3 FS-512 CL
Maxpool	
FC(4096)	
FC(4096)	
FC(1000)	
Softmax	

Table 2. First column a plain network with 34 parameter layers. Second column is a residual network with 34 parameter layers. The blue color shortcuts increase dimensions.

34 Layer Plain	34 Layer Residual
Input image	
7 x 7 FS-64 CL,/2	7 x 7 FS-64 CL,/2
3 x3 FS-64 CL	3 x3 FS-64 CL
3 x3 FS-64 CL	3 x3 FS-64 CL
3 x3 FS-64 CL	3 x3 FS-64 CL
3 x3 FS-64 CL	3 x3 FS-64 CL
3 x3 FS-64 CL	3 x3 FS-64 CL
3 x3 FS-64 CL	3 x3 FS-64 CL
3 x3 FS-128 CL,/2	3 x3 FS-128 CL,/2
3 x3 FS-128 CL	3 x3 FS-128 CL
3 x3 FS-128 CL	3 x3 FS-128 CL
3 x3 FS-128 CL	3 x3 FS-128 CL
3 x3 FS-128 CL	3 x3 FS-128 CL
3 x3 FS-128 CL	3 x3 FS-128 CL
3 x3 FS-128 CL	3 x3 FS-128 CL
3 x3 FS-128 CL	3 x3 FS-128
3 x3 FS-256 CL,/2	3 x3 FS-256 CL,/2
3 x3 FS-256 CL	3 x3 FS-256 CL
3 x3 FS-256 CL	3 x3 FS-256 CL
3 x3 FS-256 CL	3 x3 FS-256 CL
3 x3 FS-256 CL	3 x3 FS-256 CL
3 x3 FS-256 CL	3 x3 FS-256 CL
3 x3 FS-256 CL	3 x3 FS-256 CL
3 x3 FS-256 CL	3 x3 FS-256 CL
3 x3 FS-256 CL	3 x3 FS-256 CL
3 x3 FS-256 CL	3 x3 FS-256 CL
3 x3 FS-256 CL	3 x3 FS-256 CL
3 x3 FS-256 CL	3 x3 FS-256 CL
3 x3 FS-512 CL,/2	3 x3 FS-512 CL,/2
3 x3 FS- 512 CL	3 x3 FS- 512 CL
3 x3 FS- 512 CL	3 x3 FS- 512 CL
3 x3 FS- 512 CL	3 x3 FS- 512 CL
3 x3 FS- 512 CL	3 x3 FS- 512 CL
3 x3 FS- 512 CL	3 x3 FS- 512 CL
Avg Pooling	
FC 1000	

4 Experimental Results

There are four classification model AlexNet, VGG-16, ResNet-50 and Inception-v3 [8, 9] used in this paper. To check the performance of above mentioned models on other datasets. In this paper we used Caltech-101 dataset, which contains 101 classes and approximately 10k images. This dataset contains large number of images, so we reduced the number of images down to 1400. Then we apply testing on this reduced dataset to all four classification models. To check the performance of classification models, we have used True Positive Rate (TPR), False Positive Rate (FPR), Precision and Accuracy [5, 6], which are described below (Tables 3, 4, 5 and 6).

Table 3. Confusion matrix for AlexNet model.

Total input (1420)	Ant	Beaver	Cougar	Electric guitar	Flamingo	Grand piano	Other
Ant	8	0	0	0	0	0	12
Beaver	0	8	0	0	0	0	12
Cougar	0	0	24	0	0	0	16
Electric guitar	0	0	0	8	0	0	12
Flamingo	0	0	0	0	17	0	23
Grand piano	0	0	0	0	0	14	6
Other	1	2	4	2	0	0	771

Table 4. Confusion matrix for VGG model.

Total input (1420)	Ant	Beaver	Cougar	Electric guitar	Flamingo	Grand piano	Other
Ant	14	0	0	0	0	0	6
Beaver	0	11	0	0	0	0	9
Cougar	0	0	33	0	0	0	7
Electric guitar	0	0	0	13	0	0	7
Flamingo	0	0	0	0	21	0	19
Grand piano	0	0	0	0	0	15	5
Other	2	1	1	2	0	0	873

Table 5. Confusion matrix for ResNet model.

Total input (1420)	Ant	Beaver	Cougar	Electric guitar	Flamingo	Grand piano	Other
Ant	15	0	0	0	0	0	5
Beaver	0	15	0	0	0	0	5
Cougar	0	0	36	0	0	0	4
Electric guitar	0	0	0	17	0	0	3
Flamingo	0	0	0	0	24	0	16
Grand piano	0	0	0	0	0	16	4
Other	0	2	0	1	0	0	965

Table 6. Confusion matrix for inception model.

Total input (1420)	Ant	Beaver	Cougar	Electric guitar	Flamingo	Grand piano	Other
Ant	14	0	0	0	0	0	6
Beaver	0	14	0	0	0	0	6
Cougar	0	0	37	0	0	0	3
Electric guitar	0	0	0	19	0	0	1
Flamingo	0	0	0	0	30	0	10
Grand piano	0	0	0	0	0	19	1
Other	1	1	4	1	0	0	1027

True Positive Rate (TPR): It is ratio of correctly classified elements [5, 6].

$$\mathrm{TPR} = \frac{TP}{\mathrm{TP} + \mathrm{FN}} \tag{1}$$

Precision: It is ratio of correctly classified elements with total correct classification.

$$\mathrm{Precision} = \frac{TP}{\mathrm{TP} + \mathrm{FP}} \tag{2}$$

False Positive Rate (FPR): It is ratio of incorrect elements that classified correct.

$$\mathrm{FPR} = 1 - \mathrm{TNR} \tag{3}$$

Accuracy: It is ratio of correctly classified element with total number of prediction.

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{4}$$

Table 7. TPR, precision, FPR, and accuracy.

	Inception	ResNet	VGG	AlexNet
TPR	0.815	0.765	0.693	0.612
Precision	0.974	0.963	0.943	0.905
FPR	0.169	0.231	0.331	0.506
Accuracy	0.817	0.766	0.69	0.598

Figure 3 shows the accuracy of AlexNet is minimum among all, Precision is approx same in all model and FPR is maximum is AlexNet and minimum in Inception model. Table 7 shows that inception model having the best accuracy among these models. It also shows that inception model is best in precision among them.

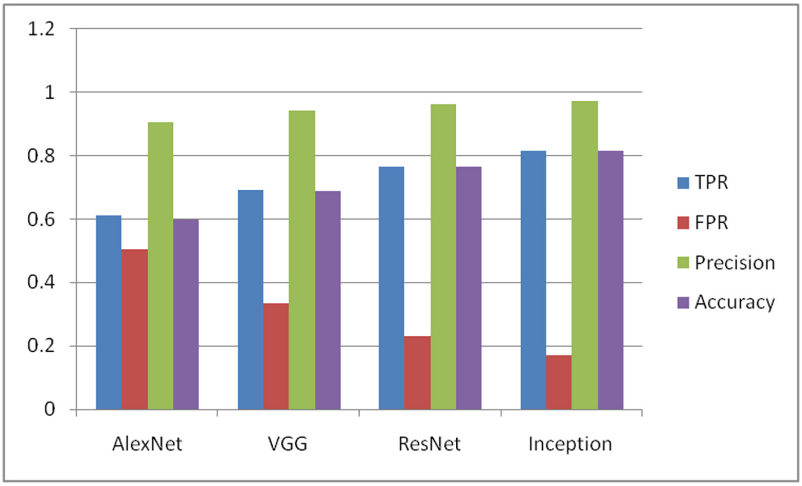

Fig. 3. TPR, FPR, precision and accuracy graph.

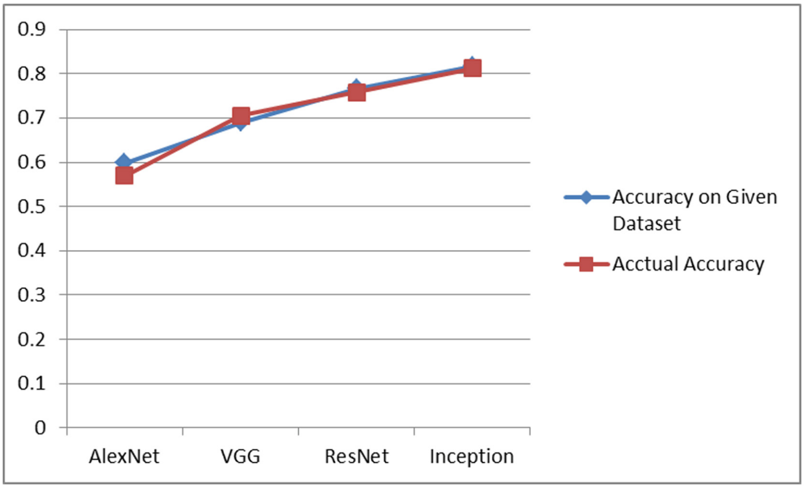

Fig. 4. FPR, precision and accuracy graph. (Color figure online)

In Fig. 4 there are two lines, red line represents the accuracy on given dataset Caltech-101 and the blue one represent the accuracy according to claimed accuracy [1–4]. Figure 8 shows, there is no difference between accuracies. In above graph accuracy is calculated on the basis of classification of objects correctly. But if we calculate the probability of the object in top-5 predicted objects by models then we get following accuracy improvement AlexNet obtains 57.0% accuracy for top-1 while for top-5 it obtains 80.3% accuracy, VGG obtains 70.5% accuracy for top-1 while for top-5 it obtains 90.0% accuracy, Resnet-152 obtains 75.8% accuracy for top-1 while for top-5 it obtains 92.9% accuracy while Inception obtains 81.2% accuracy for top-1 while for top-5 it obtains 95.8% accuracy.

5 Conclusion

There are four different classification and recognition approaches is presented in this paper and performed comparison on these classification models. For comparison of classification algorithm we used four parameters true positive rate, precession, false positive rate and accuracy. These derivatives shows which comparison model is better with comparison to other. Inception classification model having the highest accuracy and lowest false positive rate among all, while AlexNet classification model have the lowest accuracy and highest false positive rate among all.

References

1. 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)
2. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
3. Szegedy, C., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
5. Agarwal, A., Gupta, S., Singh, D.K.: Review of optical flow technique for moving object detection. In: 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I), pp. 409–413. IEEE, December 2016
6. https://en.wikipedia.org/wiki/Confusion_matrix
7. https://www.analyticsvidhya.com/blog/2017/06/architecture-of-convolutional-neural-networks-simplified-demystified/
8. Shin, H.C., et al.: Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging **35**(5), 1285–1298 (2016)
9. Xia, X., Xu, C., Nan, B.: Inception-v3 for flower classification. In: 2017 2nd International Conference on Image, Vision and Computing (ICIVC), pp. 783–787. IEEE, June 2017
10. Singh, D.K.: Gaussian elliptical fitting based skin color modeling for human detection. In: 2017 IEEE 8th Control and System Graduate Research Colloquium (ICSGRC), pp. 197–201. IEEE, August 2017