A Critical Review of Object Detection using Convolution Neural Network

Sehar Un Nisa[1], Muhammad Imran[1],
[1]Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST) Islamabad
sehar.rahes@gmail.com, malikimran110@gmail.com,

***Abstract*— To recognize an object, for a human, is an easy task but for machines, to perform the same task with same efficiency is a complex task. For computer systems, images are sets of numeric values that have no meaning in itself. To make these numbers useful, diverse techniques have been proposed. Comparative to others, deep learning approaches achieved state-of-the-art performance in many computer vision applications, such as object detection, image classification, image retrieval, human pose estimation. To detect object of interest, Convolutional Neural Network (CNN) has been observed widely successful method. Few factors are there to get better accuracy and performance for instance efficient model, larger datasets and hardware support. This study aims to review CNN methods for object detection by highlighting the contribution and challenges from few recent research papers. Also how well to use these CNN techniques in combination to other methods for best suited results with other. Better performance such as increased accuracy, fast processing reduce error rates also introduced few new concerns and issues in parallel regarding the discussed methods such as time consumption, anonymous behavior of Neural Network. To address these issues a conceptual model is presented using CNN and Lease Square Support Vector Machine (LS-SVM).**

Keywords*— *computer vision, Convolutional Neural Network, , detection, image, Lease Square Support Vector Machine, object

I. INTRODUCTION

Human recognizes objects in an image with little effort but for computer aided systems the task of object recognition is still a challenge for many decades. Premise of object detection lies human visual attention. Visual attention of human has been studied in many domains such as psychology and human neural system. Researchers are working to imitate neural system so that computers can visualize objects as we human do. The task of computer vision applications is to detect concentrated parts for classification, recognition or detection in images. State-of-the-art performance has been observed using certain techniques of machine learning among them Convolution Neural Network (CNN) models are at top in object detection.

Apart from the fact that CNN performance has been prominently increasing but still it is also induces certain limitations to it. According to Alex et al. [1], CNN can perform well on natural image classification, and proved a base model for later researchers [4, 5, 6, 7, 8]. Limitation of AlexNet model was that there were no clear implementation understanding of this model discussed in their study. Recently a lot of work has been done to understand the architecture of CNN [2, 4, 7]. CNN models required fixed size images which is a challenge when dealing with images with scalable sizes. To address this, new approaches have been introduced in [4, 8]. Training with limited data rises overfitting problem [1, 6, 9]. It arises whenever network has learned the training data, but cannot implement those learned features on real time or unknown images. Reason of this overfitting issue is due to less training dataset or less training time. Two approaches have been introduced in [1] to reduce overfitting Data augmentation and Dropout. Data augmentation [1, 8, 9] allows to transform images from original image with very little computation and these images do not need to be stored on disk, which saves memory. Dropout approach randomly drop neurons from network during training to prevent it from co-adapting too much. A lot of computational resources are needed [2, 3] which takes time to train network. Powerful GPU (Graphical Processing Units) are needed to speed up training process [1]. A state-of-the-art CNN can produce images that is unrecognizable to humans with 99.99% confidence which is still an open problem on generality of CNN performance.

In this work state-of-the-art models of object detection task will be reviewed. A new model will be proposed in this work. ImageNet dataset will be used in proposed model.

II. LITERATURE REVIEW

Krizhevsky, et al. [1] proposed a technique using Convolution Neural Network (CNN) to detect classification from natural images. Object detection and classification tasks are different but someway related. In object classification task object instances in an image are estimated for class label. Conversely in object detection task not only class label but location of instances are also estimated. Therefore proposed framework has been used in several models of object detection. The proposed network architecture used five convolution layers with max-pooling layers and three fully connected layers. Their major contribution is reducing overfitting problem; which occurs due to less training dataset or less training time. Two approaches have been introduced to reduce Overfitting: Data augmentation and Dropout. Data augmentation allows to transform images from original image with little computation. Augmented images need not to be stored on memory. By doing so memory can be saved. In Dropout approach neuron are dropped randomly from network at the time of training so that it would not accustomed too much. The discussed framework is proved a base model for later computer vision researchers. The proposed model won ILSVRC-2012 and achieved top-1 error rates of 37.5% and top-5 error rates of 17.0%.

978-1-5386-9609-5/19/$31.00 ©2019 IEEE

Szegedy, et al. [2] introduced a framework to address the problem of classifying an object with location estimation that is still a challenging task in object detection domain. The proposed model detects multiple instances of objects in a single image. Two approaches are introduced for object detection in presented model. Firstly authors deal this problem as regression problem. Regression is not only useful in object classification but also in localization of objects as well. For classification authors have used AlexNet and replaced last layer with regression layer. Network architecture of presented model used Deep Neural Network (DNN) based regression consists of seven layers, in which two layers are fully connected layers and five of them are convolution layers. Each layer have used max pooling with Rectified Linear Unit (ReLU) additionally. Secondly authors aim to detect multiple instances of different sizes with low computational cost. The discussed model is named as DetectorNet and achieved Mean Average Precision (mAP) of 30.41%.

Handling deformation is still a challenge in object detection task. Deformation can be refer as part of a moving object, for instance the body parts of a pedestrian detector. Ouyang, et al. [3] have introduced deformation convolutional neural networks model to detect objects in an image. To handle deformation, authors have presented a deformation pooling layer. The proposed work scarce in learning the deformation constraint and geometric model of object parts. So def-pooling can be proved as a replacement for max and average pooling strategies which are commonly used. To increase the learning process it is quite popular in CNN models to use pre-trained network. Number of approaches use AlexNet as their baseline network model, by doing so they accelerate the process of learning to result better performance. Multiple contributions of discussed model include integration of object parts deformation, learning of feature representation, localization refinement of bounding box. Authors have improved the mAP from 31% to 50.3% obtained by R-CNN

Although AlexNet won the ILSVRC2012 competition but its major drawback is that it requires fixed resolution of input images, which effects the precision for the images of variance size or scale. He, et al. [4] have discussed a SPP-Net to address the requirement of fixed resolution of images. In this model, authors have introduced the Spatial Pyramid Pooling to solve the artificial problem of the need of fixed resolution images for CNN model. SPP-Net can be used to improve the accuracy of prevailing architectures as well. SPP-Net utilizes the general CNN architecture with replacement of a spatial pyramid pooling layer with last pooling layer. Advantage of using SPP is that fixed-length representation are extracted from arbitrary images. These extracted features are useful for handling different size, scale and aspect ratio, and can be applied in any CNN architecture to enhance its performance. PASCAL VOC 2007-2012 and Caltech101 are used for training and testing of this model. Proposed model achieved 60% Mean Average Precision (mAP).

Girshick, et al. [5] proposed an algorithm to detect scalable objects and achieved better performance on PASCAL VOC dataset as compared to previous models. Indroduced model consists of two aspects. Firstly, CNN can be applied to learning feature representation, localization refinement of bounding box and object segmentation. Secondly, as many models underwent overfitting due to insufficiant labbeled data, supervised pretraining with generalized fine-tuning can be utilized to improve performance. Authors have suggested to pretrain network model on ImageNet dataset with object level annotations. Existing models of CNN have been using image level annotation for the task of object detection.

Girshick, et al. [6] discovered the learning pattern of CNN. But in recent years there is no prominent result recorded on PASCAL VOC. Authors claimed that they have achieved better performance on PASCAL VOC as compared to previous best recorded results. Their approach consists of two aspects. Firstly to localize and segment object, to obtain that CNN can be applied to learning feature representation, localization refinement of bounding box. Authors have suggested to pretrain network model on ImageNet dataset with object level annotations rather than image level annotations. Their stage-by-stage training scheme solves overfitting problem by adding regularization constraint to parameter. Size of a network plays a vital role to increase the performance of CNN, so to achieve this, authors have adopted approach of combining different structures together to increase the size of a network by using the output of former network in the previous ones. The proposed approach use selective search for candidate bounding boxes, a detector rejects bounding boxes which are similar to background. The proposed approach have been ranked 2 in ILSVRC 2014. Authors have improved mAP by 30% as compared to previous best result on Pascal VOC.

Sermanet, et al. [7] proposed a model that integrate multiple tasks of object detection on a single network simultaneously. Before this model all these tasks were completed on separate networks. Authors have used same network architecture for classification used in AlexNet. Drawback of AlexNet architecture was that there were no clear description of how that network worked. In this work authors have vividly describe the working of CNN model that is a major contribution. Proposed model is winner of object detection and localization in ILSVRC 2013 and 4th in object classification. Proposed model has achieved mAP of 29.9% on ImageNet and winner of ILSVRC competition of 2013.ImageNet and Pascal VOC 2012 datasets are used for training and testing of model.

Szegedy, et al. [8] proposed a large network architecture to achieve high quality performance for object detection. One of the challenges CNN models are facing today is to learn the ability to generalize. Many research models have been introduced to solve generalization by using large networks with maximum number of parameters. Authors introduced a novel approach with deeper structure of using large network with 22 layers to predict multiple instance of objects in a single image. The proposed model is divided into two tasks. At first stage bounding boxes are estimated that might have potential existence of object instances. At second stage CNN classifier is used to predict and localized these estimated

bounding boxes. As large number of parameters requires massive computational efforts to train network. Authors have solved this problem by a constant cost for computational budget. To avoid overfitting problem authors have adopted data augmentation approach. The proposed model has achieved mAP of 43.9% on ImageNet dataset. MNIST and CIFER datasets are used for training and testing of model.

Erhan, D., C. Szegedy, et al. [9] have used deep CNN to detect objects from a single image that can envisages multiple bounding boxes at a time for objects of interest. Authors have used a detector called 'DeepMultiBox'. This approach is scalable and can generalize across multiple datasets to predict location of interest. To detect multiple instance from a single image is a challenging task. Important feature of this algorithm is that it is able to capture multiple instances of objects of the same class. Proposed model has generalization ability not only for the categories on which it was trained on but for the unknown categories as well. For each image multiple bounding boxes are estimated and confidence score of how likely that bounding box contain the object of interest are predicted. The proposed model introduced loss, that uses and train bounding box predictor with network training. The proposed model resulted 40% top-5 on ILSVRC 2012.Their pipeline work includes to use localization and recognition paths into a single network to extract both location and class label information in a single feed-forward pass through network.

Simonyan, K. and A. Zisserman et al. in [10] have used a model to increase the size of the network by adding multiple convolutional layers in comparison to the AlexNet model. Use of very small convolutional filters in all layers also improves its efficiency. Increase in size results the increase of generalize ability. Larger networks improves performance with the risk of overfitting and it needs a lot computational resources. This model achieved a top-5 error rate of 7.3% and ranked second in ILSVRC 2014.

Barat, C. and C. Ducottet et al. [11] have used structural representations on top of pre-trained CNN feature to improve image classification. Authors claimed that performance of the CNN model can be increase by two ways by using of a pre-trained CNN and secondly single feature extraction way. These methods can be useful for training and testing and achieve good performance on standard CPU. Authors demonstrated that the powerful edit distance can be very effective for image classification, if we carefully choose the edit operations and their associated costs. The model combined CNN features and Spatial Pyramid Pooling. Their major contribution is that they have avoided computing cost by replacing GPU with standard 16-core CPU.

He et al [12] Identified problem of CNN thirst for large amount of data to train its network. This problem is inevitable. To solve this issue earlier models used Rectified Linear Unit (ReLU). ReLU has revolutionized CNN performance for object detection. It is first introduced in AlexNet when first time CNN performed state-of-the-art on natural images. Previous work was done for digits not for natural images. Later models used AlexNet to achieve better performance. Proposed model enhanced ReLU and introduced Parametric Rectified Linear Unit (PReLU). Authors have train their model on ImageNet and used architecture of AlexNet. Doing so gives edge to proposed model to work well on task of image classification and recognition. This model achieved 4.94% top-5 test error on ImageNet 2012.

Razavian, et al. [13] achieved high performance in object recognition. Authors have identified that using already trained model can give better results and can be useful for multiple vision tasks. Authors have trained their model on OverFeat that used AlexNet as a base model. Reason of its good performance is that they have used feature extracted from OverFeat network. Proposed model approach for using trained OverFeat can be useful for not only object recognition task but also for classification and localization task as well. It gives variety for different vision tasks. SVM classifier was applied on features extracted from proposed model. Jittering was applied for image augmentation. The proposed model is developed with CNN and SVM classifiers that gives the edge for better performance. Authors have achieved mAP of 73.0% with diverse datasets of Oxford5K, Paris6K, Sculptures6K, Holidays and UKbench.

He et al. [14] have proposed a deep model that gives good result with low computational cost. Behavior of deep model is discussed in this study. Authors have implemented a deep network of 152 layers with low complexity and achieved top 5 error of 3.57. CIFER-10 dataset is used for training and testing purpose. Authors have used RCNN as base model for the task of object detection. RCNN is customized and VGG is replaced with ResNet .The proposed model achieved 73.8% mAP.

Oquab, et al. [15] proposed model that can recognize with limited amount of training data. There are factors of images that effect detection task, for instance dimension of object, size of images, multiple classes present in image and background .The proposed model is trained with large amount of images of ImageNet dataset that is almost 15 million images and 22000 classes. Later the proposed model is tested on PASCAL VOC. AlexNet architecture has been used as a base model. Model is pretrained on ImageNet. The proposed model achieved 82.8% mAP.

III. PROPOSED MODEL

First step of proposed model is to remove noise from images. Secondly features are extracted with a method combination of HoG (Histogram of Gradient) and PCA (Principle component Analysis). For dimensionality reduction LDA (Linear Discriminant Analysis) has been used. Then the refined image is used as an input to CNN model. CNN model obtain deep feature as well as apply morphological segmentation on image. In next step all the features computed by ConvNet after segmentation will forward for object extraction. Later Least Square Support Vector Machine (LS SVM) is used to estimate predicted score of object present in image where we will get

desired object. To train and test the model ImageNet data set will be used.

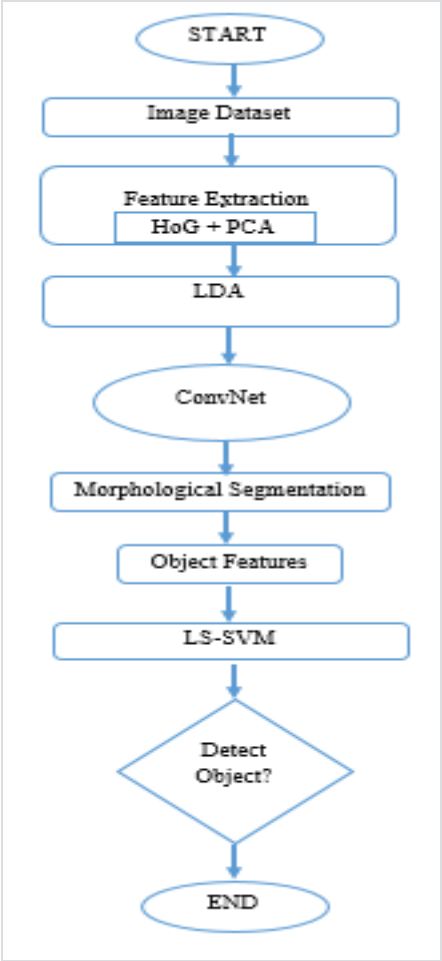

Figure 1: PROPOSED MODEL named VECTOR

IV. Conclusion

In this paper, a review of different approaches is presented to enlighten the work done for saliency detection using CNN. Later on a critical evaluation is done on the basis of techniques used, limitations and challenges. It is very clear from study that some issues are still open problems and needs intense research. To address these issues a conceptual model is proposed that ideally solves problems of overfitting, computational cost, multiple instance detection and deformation. Some of the issues discussed in critical evaluation section are artificial and avoidable. To address these issues a conceptual model is proposed that ideally solves problems of overfitting, computational cost, multiple instance detection and deformation. Some of the issues discussed in critical evaluation section are artificial and avoidable. Few solution of these problems are suggested in this review.

REFERENCE

[1] Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097-1105.

[2] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in Advances in Neural Information Processing Systems, 2013, pp. 2553-2561.

[3] W. Ouyang, X. Zeng, X. Wang, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, C. C. Loy, K. Wang, J. Yan, and X. Tang, "DeepID-Net: Deformable Deep Convolutional Neural Networks for Object Detection," IEEE Transactions on Pattern analysis and machine intelligence, 2016, pp. 1-1

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," IEEE Transactions on Pattern analysis and machine intelligence, vol. 37, pp. 1904-1916, 2015.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-Based Convolutional Networks for Accurate Object Detection and Segmentation," IEEE Transactions on Pattern analysis and machine intelligence, vol. 38, pp. 142-158, 2016.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580-587.

[7] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in International Conference on Learning Representations, 2013, pp. 1312-6229.

[8] C. Szegedy, L. Wei, J. Yangqing, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1-9.

[9] Erhan, D., C. Szegedy, et al. Scalable object detection using deep neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp 1011-1030.

[10] Simonyan, K. and A. Zisserman, "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv: , 2014, pp 1409.1556.

[11] Barat, C. and C. Ducottet, "String representations and distances in deep convolutional neural networks for image classification." Pattern Recognition., 2016, pp 1304-1320.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 'Delving Deep into Rectifiers: Surpassing Human-Level Performance on Imagenet Classification', in Proceedings of the IEEE international conference on computer vision, 2015), pp. 1026-34.

[13] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson, 'Cnn Features Off-the-Shelf: An Astounding Baseline for Recognition', in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014), pp. 806-13.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 'Deep Residual Learning for Image Recognition', in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016), pp. 770-78.

[15] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic, 'Learning and Transferring Mid-Level Image Representations Using Convolutional Neural Networks', in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014), pp. 1717-24.

[16] Guo, Yanming, et al. "Deep learning for visual understanding: A review."Neurocomputing 2015.

TABLE 1: CRITICAL EVALUATION

Sr	Authors	Tech.	Performance parameters			Limitation	AlexNet Architecture	Bounding Boxes	ImageNet Pretraining	Fixed-size images	Datasets		
			Error rate – 1 (%)	Error rate -5 (%)	Mean Average Precision (mAP)						Name	Classes	Total Images
1	**Krizhevsky, et al. 2012**	CNN	37.50	17.0.	None	□ Required fixed resolution images. □ Depth of the network is strictly considered, removing any layer resulted in lower performance	Yes	No	Yes	Yes	ImageNet[1]	20000	Used 1.2 million images from total 15 million
2	**Szegedy, et al. 2013**	CNN	NA	NA	30.41%.	□ Computational cost at training time. □ Detect limited number of classes.	Yes	Yes	No	No	PASCAL[2] VOC 2007-2012	20	Used VOC2007 (5000 images-for testing) and VOC2012 (11 thousand images-for training)
3	**Ouyang, et al. 2016**	CNN	NA	NA	69%		Yes	Yes	Yes	No	ImageNet	20000	15 Million
											MS-COCO[6]		
											PASCAL VOC	20	16 thousands
4	**He, et al. 2015**	CNN+SVM	27.86	8.06	59.20%	• Proposed model has used small dataset that rise question on its accuracy.	No	Yes	Yes	No	PASCAL VOC	20	9963 (50.29% - train, 49.7% - test)
											Caltech101[3]	102	9144
5	**Girshick, et al. 2016**	CNN+SVM	NA	NA	62.4%.		Yes	Yes	Yes	No	ImageNet	20000	456191(train - 86.7%, Val - 4.4%, Test - 8.8%)
											PASCAL VOC	20	16 thousands
6	**Girshick, et al. 2014**	CNN	NA	NA	58.5%.	Feature computation is time consuming.	Yes	Yes	Yes	No	ImageNet	1000	15 Million
											PASCAL VOC	20	16 thousands
7	**Sermanet, et al. 2013**	CNN	35.74	14.20	29.90%	□ Require labeled training data □ Slow performance of localization task □ Slow GPU performance	No	Yes	Yes	No	ImageNet	20000	15 Million

											PASCAL VOC	20	16 thousands
8	**Szegedy, et al. 2015**	CNN	NA	6.67	43.90%	□ Max Memory usage □ Computational cost is maximum	Yes	Yes	No	No	ImageNet	20000	15 Million(8%–train, 0.3% - val, 0.1% - test)
											MNIST[4]	NA	70,000
											CIFAR[5]	10	60,000
9	**Erhan, et al (2014)**	CNN +SVM			59.4%	Do not use Location and class label information in a single pass	Yes	Yes	Yes	No	15Scene, Caltech101,Pascal VOC 2007, MIT indoor		
10	**Simonyan, et al (2014)**	CNN			53.3%		Yes	Yes	Yes	No	ImageNet	20000	15 Million
											PASCAL VOC	20	16 thousands
11	**Barat, et al (2016)**	CNN				• Integrate edit operations into graph representations for relationship between neighboring strings. • Improvement required for slightly different visual content	No	Yes	Yes	Yes	ImageNet	20000	15 Million
											PASCAL VOC	20	16 thousands
12	**Kaiming et al (2015)**	CNN		4.94%		• Good result for classification task, but lack in object detection task	Yes	Yes	Yes	No	ImageNet	20000	15 Million
											PASCAL VOC	20	16 thousands
13	**Razavian et al (2014)**	CNN + SVM			73.0%		No	Yes	Yes	Yes	Oxford5K, Paris6K, Sculptures6K, Holidays and UKbench		
14	**Kaiming et al (2016)**	CNN		3.57	73.8%		Yes	Yes	Yes	No	ImageNet	20000	15 Million
											PASCAL VOC	20	16 thousands
15	**Oquab et al (2014)**	CNN			82.8%		Yes	Yes	Yes	No	ImageNet	20000	15 Million
											PASCAL VOC	20	16 thousands