Detection of Objects Using Deep Supervised Approach

Mr. G. Pavan Kumar[1], N. Sai pragna[2], N. Anusha[3], S. Bala Sushma[4], P. Anuhya[5]
[1]Asst. Prof, Dept of CSE, Tirumala Engineering College, Narasaraopet, Guntur, A.P., India
[2,3,4,5]B. Tech Students, Dept of CSE, Tirumala Engineering College, Narasaraopet, Guntur, A.P., India

ABSTRACT - We introduce the Modified Deeply Supervised Object Detector (MDSOD), a system for learning object detectors from the ground up. Modern object depends heavily on the off-shelf networks pre-trained in large-scale classification datasets such as ImageNet, since both loss functionality and category distribution between classification and detection tasks are different. Model adjustment for the detection task will somewhat but not fundamentally mitigate this bias. Furthermore, it is much more difficult to migrate previously trained models to different domains from classification to detection. A safer way to address these two important issues is to train object detectors from scratch, which is what motivates our proposed MDSOD. Previous attempts in this direction have largely failed due to much more complex failure functions and a lack of training data in target detection. We contribute a series of design guidelines for training object detectors from scratch to MDSOD.

Keywords: Object Detection, Supervised Networks, Deep Learning

I. INTRODUCTION

Convolutional Neural Networks (CNNs) have significantly improved performance in a wide range of computer vision tasks, including image recognition, object identification, image segmentation, and so on. Many novel CNN network architectures have been proposed in recent years.

[1] propose a "Inception" module that concatenates feature maps created by different sized filters.

[2] DenseNets with dense layer-wise relations are proposed. The accuracy of certain vision operations has significantly increased as a result of these excellent network architectures. Among them, object detection is one of the most rapidly evolving areas due to its numerous uses in surveillance, autonomous vehicles, and other fields.

However, there are several significant drawbacks to using pre-trained networks for target detection: (1) Limited design area for the structure. The pre-trained network models are primarily from ImageNet-based classification tasks, which are usually very heavy and have a large number of parameters. Existing object detectors use pre-trained networks directly, so there is no ability to control/adjust the network architectures. The heavy network architectures often limit the amount of computational capital required. (2) Bias of learning We claim that since the loss functions and category distributions for classification and detection tasks vary, this can result in separate searching/optimization spaces. As a result, learning could be skewed against a local minimum that is not optimal for the detection mission. (3) Mismatch of domains. As is well established, fine-tuning will help to close the distance caused by different target group distribution. However, it remains a serious issue where the source domain (ImageNet) is vastly different from the target domain, such as depth images, medical images, and so on.

II. RELATED WORK

[1] R-CNN, Fast R-CNN, Faster R-CNN, and R-FCN are all proposal-based methods. R-CNN employs selective search to create possible object regions in an image before classifying the suggested regions. R-CNN has high computing costs since the CNN network processes each region separately. Fast R-CNN and Faster R-CNN increase throughput by exchanging computation and generating area ideas with neural networks. R-FCN increases speed and precision much more by eliminating fully-connected layers by using position-sensitive score maps for final detection.

[3] The creation of network architectures for image classification has received a lot of attention. Many new networks have appeared, including AlexNet, VGGNet, GoogLeNet, ResNet, and DenseNet. In the meantime, many regularisation approaches have been proposed to further improve the model's capabilities. The majority of detection methods use pre-trained ImageNet models as the backbone network.

[4] Other works build specialised backbone network architectures for object detection, but they also enable the network to be pre-trained on the ImageNet classification dataset first. YOLO, for example, specifies a network of 24 convolutional layers followed by two completely connected layers.

[7] proposes PVANet for object identification, which is made up of a condensed version of GoogleNet's "Inception" block. [11] researched different network structure and identification system combinations and discovered that Faster R-CNN with Inception-ResNet-v2 obtained the best results. We also look at network architectures for generic object

detection in this article. The proposed DSOD, however, does not require pre-training on ImageNet.

[11] demonstrated that without the use of pre-trained models, a well-designed network structure will outperform state-of-the-art solutions DenseNets is extended to complete convolutional networks by including an upsampling path to retrieve the full input resolution.

III. PROPOSED ARCHITECTURE

The proposed MDSOD approach is a multi-scale proposal-free detection framework, analogous to SSD. MDSOD's network configuration is split into two parts: the backbone sub-network for feature extraction and the front-end sub-network for prediction over multi-scale response maps. The backbone sub-network is a version of the strongly supervised DenseNets system, consisting of a stem block, four thick blocks, two transformation layers, and two transition w/o pooling layers. The front-end subnetwork combines multi-scale prediction responses with a dense framework that has been elaborated. Figure 1 depicts the proposed MDSOD prediction layers as well as the basic framework of SSD's multi-scale forecasting charts.

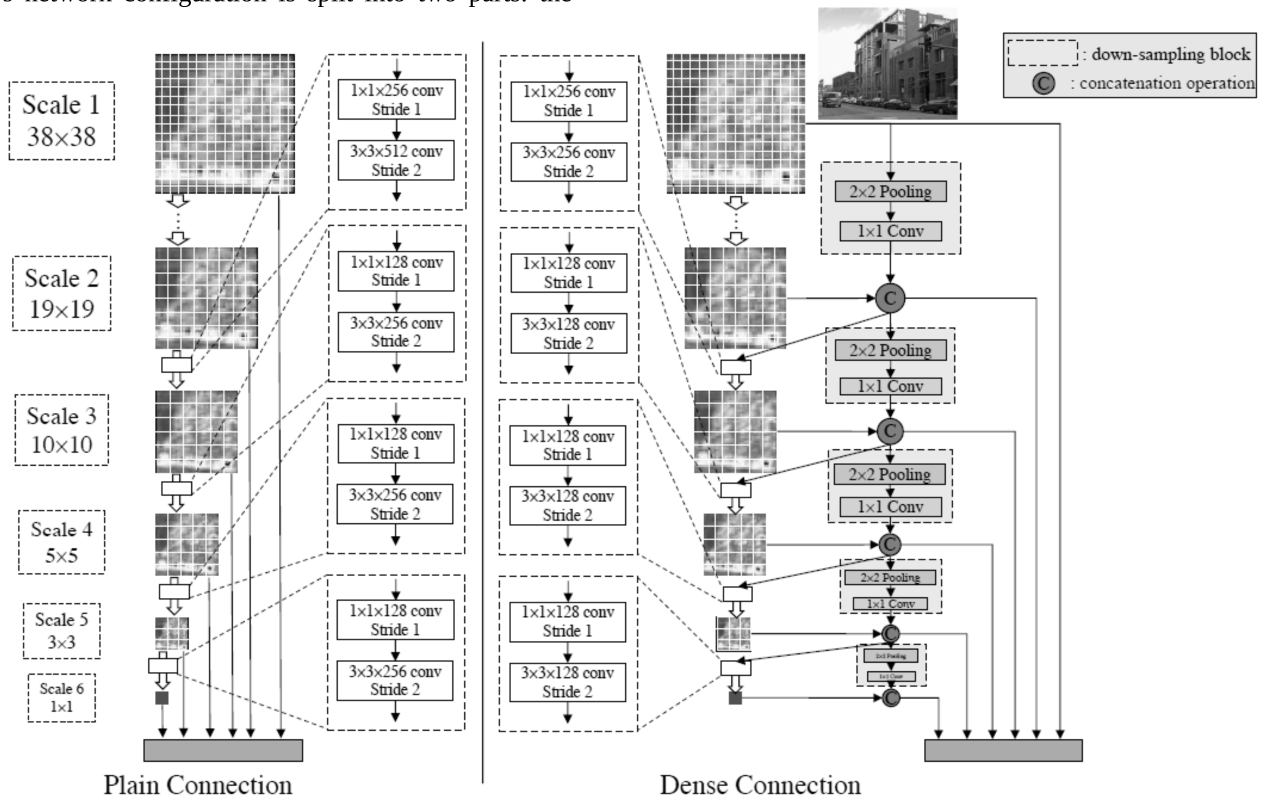
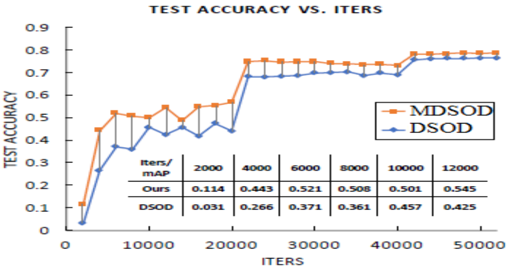


Figure 1. Proposed Modified DSOD architecture

IV. RESULTS AND DISCUSSION

In MDSOD, we find that models still achieve the highest precision with 62k iterations. MDSOD, on the other hand, needs about 50k iterations to reach final convergence of the same batch size. As a result, MDSOD has a 38 percent higher convergence speed than DSOD. Figure 1 depicts a comparison of preparation and research accuracy using the DSOD system.

On the VOC 2007 test range, DSOD and MDSOD are compared. We present MDSOD and DSOD data from six separate iterations. MDSOD achieves much higher precision with the same number of training iterations and also achieves quicker convergence than the baseline DSOD.



Iters/ mAP	2000	4000	6000	8000	10000	12000
Ours	0.114	0.443	0.521	0.508	0.501	0.545
DSOD	0.031	0.266	0.371	0.361	0.457	0.425

Figure 2. Comparison of Test results

V. CONCLUSION

Modified Deeply Supervised Object Detector (MDSOD) is a basic but powerful system for training object detectors from scratch. DSOD outperforms state-of-the-art detectors such as SSD, Faster R-CNN, and R-FCN on common datasets without using pre-trained models on ImageNet, with parameters compared to SSD, R-FCN, and Faster R-CNN. DSOD has a lot of promise in a variety of scenarios such as depth, medical, multi-spectral videos, and so on. Our future work will take into account these domains, as well as learning ultra-efficient DSOD models to support resource-constrained devices.

VI. REFERENCES

[1]. S. Bell, C. Lawrence Zitnick, et al. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In CVPR, 2016.

[2]. L.-C. Chen, G. Papandreou, I. Kokkinos, et al. Semantic image segmentation with deep convolutional nets and fully connected crfs. In ICLR, 2015. 1

[3]. J. Deng, W. Dong, R. Socher, L.-J. Li, et al. Imagenet: A large-scale hierarchical image database. In CVPR, 2009. 1,[

[4]. R. Girshick. Fast r-cnn. In ICCV, 2015.

[5]. R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.

[6]. X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In AISTATS, 2010.

[7]. S. Gupta, J. Hoffman, and J. Malik. Cross modal distillation for supervision transfer. In CVPR, 2016.

[8]. B. Hariharan, P. Arbel´aez, R. Girshick, and J. Malik. Hyper-columns for object segmentation and fine-grained localization. In CVPR, 2015.

[9]. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016.

[10]. G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. In CVPR, 2017.

[11]. J. Huang, V. Rathod, C. Sun, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In CVPR, 2017.

[12]. S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167, 2015.