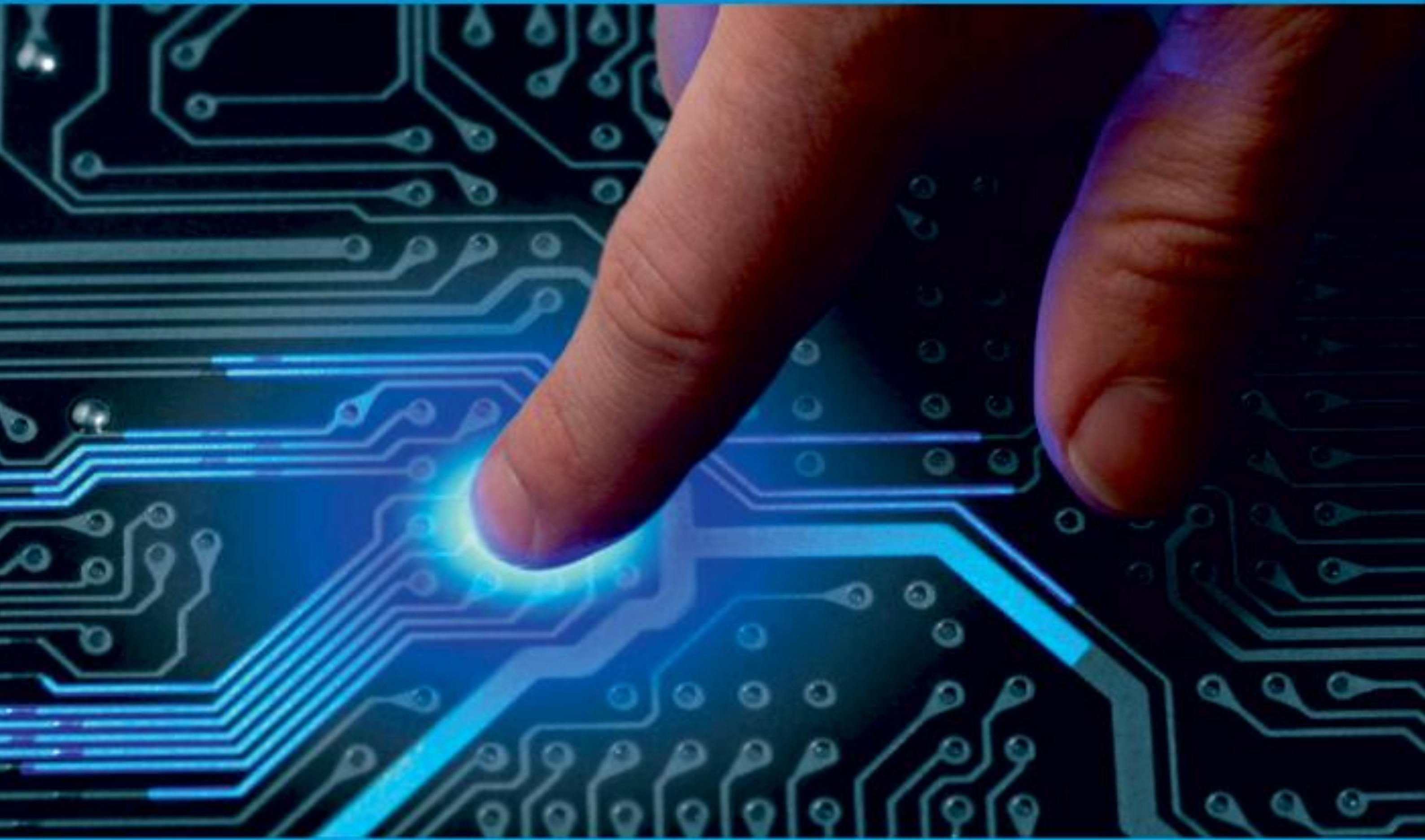
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 5, May 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

| DOI: 10.15680/IJIRCCE.2022.1005076|

Object Detection from Scratch with Deep Supervision Techniques

S Anil Kumar[1], M. Lakshmi Sowjanya[2], K.Tejaswini[3], K.Sai Prathima[4], K. Revanth[5]

Associate Professor, Department of CSE, Tirumala Engineering College, NRT, Andhra Pradesh, India

UG Student, Department of CSE, Tirumala Engineering College, NRT, Andhra Pradesh, India

UG Student, Department of CSE, Tirumala Engineering College, NRT, Andhra Pradesh, India

UG Student, Department of CSE, Tirumala Engineering College, NRT, Andhra Pradesh, India

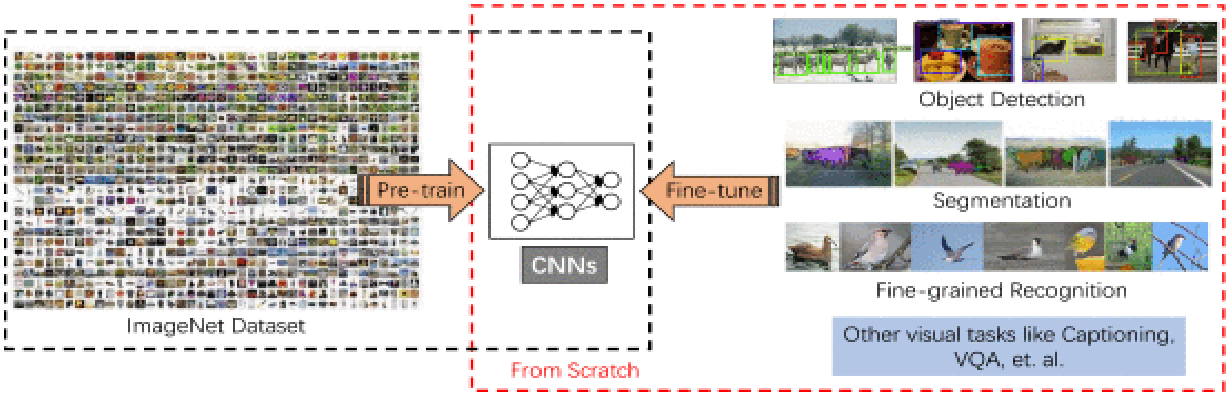
UG Student, Department of CSE, Tirumala Engineering College, NRT, Andhra Pradesh, India

ABSTRACT: In this paper, we propose Deeply Supervised Object Detectors (DSOD), an object detection framework that can be trained from scratch. Recent advances in object detection heavily depend on the off-the-shelf models pre-trained on large-scale classification datasets like ImageNet and OpenImage. However, one problem is that adopting pre-trained models from classification to detection task may incur learning bias due to the different objective function and diverse distributions of object categories. Techniques like fine-tuning on detection task could alleviate this issue to some extent but are still not fundamental. Furthermore, transferring these pre-trained models across discrepant domains will be more difficult (e.g., from RGB to depth images). Thus, a better solution to handle t hese critical problems is to train object detectors from scratch, which motivates our proposed method. Previous efforts on this direction mainly failed by reasons of the limited training data and naive backbone network structures for object detection. In DSOD, we contribute a set of design principles for learning object detectors from scratch. One of the key principles is the *deep supervision*, enabled by layer-wise dense connections in both backbone networks and prediction layers, plays a critical role in learning good detectors from scratch. After involving several other principles, we build our DSOD based on the single-shot detection framework (SSD). We evaluate our method on PASCAL VOC 2007, 2012 and COCO datasets. DSOD achieves consistently better results than the state-of-the-art methods with much more compact models. Specifically, DSOD outperforms baseline method SSD on all three benchmarks, while requiring only 1/2 parameters. We also observe that DSOD can achieve comparable/slightly better results than Mask RCNN [1] + FPN [2] (under similar input size) with only 1/3 parameters, using no extra data or pre-trained models.

KEYWORDS: Object detection, deeply supervised networks, learning from scratch, densely connected layers

I.INTRODUCTION

GENERIC object detection is the task that we aim to local- ize various objects in a natural image automatically. This task has been heavily studied due to its wide applica- tions in surveillance, autonomous driving, intelligent secu- rity, etc. In the recent years, with the progress of more and more innovative and powerful Convolutional Neural Net - works (CNNs) based object detection systems have been proposed, the object detection problem has been one of the fastest moving areas in computer vision. To achieve desired performance, the common practice in advanced object detection systems is to fine-tune models pre-trained on ImageNet [3]. This fine-tuning process can be viewed as transfer learning [4]. Specifically, as is shown in Fig. 1, researchers usually train CNN models on large - scale classification datasets like ImageNet [3] first, then fine- tune the models on target tasks, such as object detection Learning from scratch means we directly train models on these target tasks without involving any other additional data or extra fine-tuning processes. Empirically, fine-tuning from pre-trained models has at least two advantages. First, there are numerous state-of-the- art pre-trained CNN models publicly available. It is conve- nient for researchers to reuse the learned parameters in their own domain-specific tasks. Second, fine-tuning on pre- trained models can quickly convergence to a final state and requires less instance-level annotated training data than basic classification task. However, the critical imitations are also obvious when adopting the pre-trained models for object detection: (I) Lim- ited design space on network structures. Existing object detectors directly adopt the pre-trained networks, and as a conse- quence, there is little flexibility to control/adjust the detailed network structures, even for small changes of network design. Furthermore, the pre-trained models are mostly from large-scale classification task, which are usually very heavy (con- taining a huge number of parameters) and are not suitable for some specific scenarios. The heavy network structures will bound the requirement of computing resources. (II) Learning/ optimization bias. Since there are some differences in both objective functions and the category distributions between classification and detection tasks, these differences may lead to different searching/optimization spaces. Therefore, learn- ing may be biased towards a local minimum when all param- eters are initialized from classification pre-trained models, which is not the best for target detection task. (III) Domain mis- match. As is well-known, fine-tuning can mitigate the gap between different target category distribution. However, it is still a severe problem when the source domain (e.g., Image- Net) has a huge mismatch to the target domain such as depth images, medical images, etc [36].



II. RELATED WORK

Object Detection. Modern CNN-based object detectors can mainly be divided into two groups: (i) proposal-based/two-stage methods; and (ii) proposal-free/one-stage methods.Proposal-based family includes R-CNN [5], Fast R-CNN [6], Faster R-CNN [7], R-FCN [8] and Mask RCNN [1]. R-CNN uses selective search [43] to first generate potential object regions in an image and then perform classification on the proposed regions. R-CNN requires high computational costs since each region is processed by the CNN network separately. Fast R-CNN improves the efficiency by sharing computation of backbone networks and Faster R-CNN uses neural networks (i.e., RPN) to generate the region proposals. R-FCN further improves speed and accuracy by removing fully-connected layers and adopting position-sensitive score maps for final detection.

Recently, in order to realize real-time object detection, the proposal-free methods like YOLO [10] and SSD [9] have been proposed. YOLO uses a single feed-forward convolutional network to predict object classes and locations directly, which no longer requires a second per-region classification operation so that it is extremely fast. SSD further improves YOLO in several aspects, including (1) use small convolutional filters to predict categories and anchor offsets for bounding box locations; (2) use pyramid features for prediction at different feature scales; and (3) use default boxes and aspect ratios for adjusting varying object shapes. Some other proposal-free detectors also be proposed recently, e.g., RetinaNet [11], Scale-Transferrable [44], Single-shot Refinement [45], RFB Net [46], CornetNet [47], ExtremeNet [48], etc. Our proposed DSOD is built upon SSD framework and thus it inherits the speed and accuracy advantages of SSD, while produces more compact and flexible models.

Network Architectures for Detection. Since there are significant efforts that have been devoted to design network architectures for image classification, many diverse and powerful networks are emerged, such as AlexNet [49], VGGNet [50], GoogLeNet [51], ResNet [52], DenseNet [39], etc. Meanwhile, several advanced regularization techniques [53], [54] also have been proposed to further enhance the model capabilities. In practice, most of the detection methods [5], [6], [7], [9] directly utilize these structures pre-trained on ImageNet as the backbone network for detection task.

Some other works try to design specific backbone network structures for object detection, but still require to pre-train on ImageNet classification dataset in advance. Specifically, YOLO [10] defines a network with 24 convolutional layers followed by 2 fully-connected layers. YOLO9000 [55] improves YOLO by proposing a new network named Darknet-19, which is a simplified version of VGGNet [50]. YOLOv3 [56] further improve the performance through involving residual connection on Darknet-19 and other techniques. Kim et al. [57] proposes PVANet for fast object detection, which consists of the simplified “Inception” block from GoogleNet. Huang et al. [58] investigated various combination of network structures and detection frameworks, and found that Faster R-CNN [7] with Inception-ResNet-v2 [59] achieved very promising performance. In this paper, we also consider designing a suitable backbone structure for generic object detection. However, the pre-training operation on ImageNet is no longer required by the proposed DSOD.

Learning Deep Models from Scratch. To the best of our knowledge, there are no previous works that train deep CNN-based object detectors from scratch. Thus, our proposed approach has very appealing advantages over existing solutions. We will elaborate and validate the method in the following sections. In semantic segmentation, Jégou et al. [60] demonstrated that a well-designed network structure can outperform state-of-the-art solutions without using the pre-trained models. It extends DenseNets to fully-convolutional networks by adding an upsampling path to recover the full input resolution.

III.DSOD

In this section, we first introduce the whole framework of our DSOD architecture, following by several important design principles. Then we describe the objective function and training settings in detail.

3.1 Network Architecture

Similar to SSD [9], our proposed DSOD method is a multi-scale and proposal-free detection framework. The network structure of DSOD can be divided into two parts: the backbone sub-network for feature extraction and the front-end sub-network for prediction over multi-resolution feature maps. The backbone sub-network is a variant of the deeply supervised DenseNets [39] structure, which is composed of a *stem block*, four *dense blocks*, two *transition layers* and two *transition w/o pooling layers*. The front-end subnetwork (or named *DSOD prediction layers*) fuses multi-scale prediction responses with an elaborated *dense structure*. Fig. 2 illustrates the proposed DSOD prediction layers along with the plain structure used in SSD [9]. The full DSOD network architecture[1] is detailed in Table 1. Now we elaborate each component and the corresponding design principle in the following.

3.2 Design Principles

Principle 1: Proposal-Free. In order to reveal the potential influences in learning object detection from scratch, we investigated all the state-of-the-art CNN-based object detectors under the default settings. As aforementioned, R-CNN and Fast R-CNN require external object proposal generators like selective search. Faster R-CNN and R-FCN require integrated region-proposal-network (RPN) to generate relatively fewer region proposals. YOLO and SSD are single-shot and proposal-free methods (one-stage), which handle object location and bounding box coordinates as a regression problem. We observe that only proposal-free methods (one-stage detectors) can converge successfully without the pre-trained models if we follow the original settings without involving some significantly modifications (e.g., replacing RoI pooling with RoI align [1], adopting Sync BN [61] or Group Norm [62] to mitigate small batch-size issue, etc.). We conjecture this is due to the RoI pooling (Regions of Interest) in the other two categories of methods — RoI pooling uses quantization to generate features for each region proposals, which causes misalignments that hinders/reduces the gradients being smoothly back-propagated from region-level to convolutional feature maps. The proposal-based methods work well with pre-trained network models because the parameter initialization is good for those layers before RoI pooling, while this is not true for training from scratch.

Hence, we arrive at the first principle: training detection network from scratch requires a proposal-free framework, even if there is no BN layer [54] included in the network structures (In contrast, norm layer is critical for both Sync BN [61] and Group Norm [62] methods to train region-based/two-stage detectors from scratch). In practice, we derive a

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005076|

multi-scale proposal-free framework from the SSD framework [9], as it could reach state-of-the-art accuracy while offering fast processing speed.

$$Pi=\phi i[P1/4(xL),P1/2(xM),xH],$$

3.3 Training Objective

Our whole training objective loss is derived from SSD [9] and Fast RCNN [6], which is a weighted sum of the classification loss (cls) and the localization loss (reg)

$$L(p,p*,r,g)=1N(Lcls(p,p*)+\alpha p*Lreg(r,g)),(2)$$

where p denotes a discrete probability distribution that is computed by a softmax over the K+1 outputs. p∗ is the ground-truth class. r is the bounding-box regression offsets and g is the ground-truth bounding-box regression target. α is the coefficient to balance the two losses. Following Fast RCNN [6], we also adopt the L1 loss for bounding-box regression

$$Lreg(r,g)=\sum i\in\{x,y,w,h\}smoothL1(ri-gi).(3)$$

Experiments

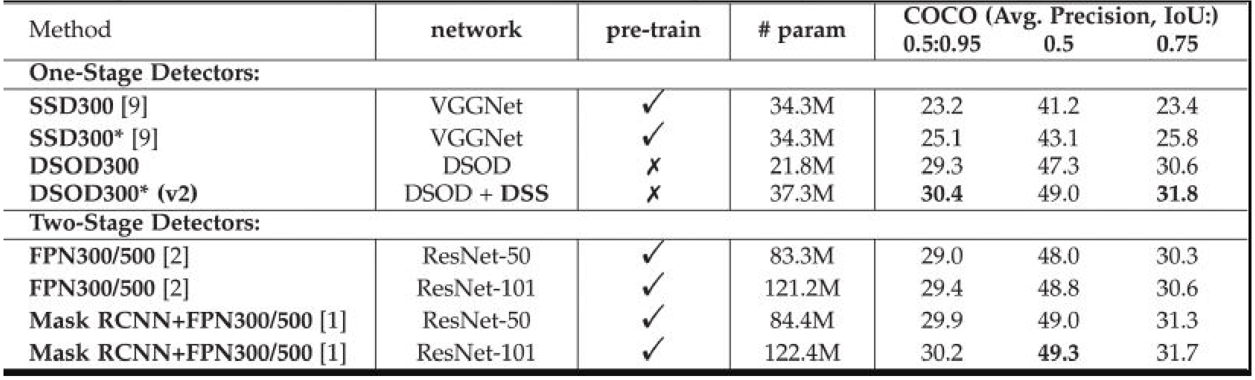
Our experiments are conducted on the widely used PASCAL VOC 2007, 2012 and MS COCO datasets that have 20, 20, 80 object categories respectively. We adopt the standard mean Average Precision (mAP) to measure the object detection performance.

4.1 Ablation Study on PASCAL VOC2007

We first investigate each component and design principle of our DSOD framework. The results are mainly summarized in Tables 6 and 3. We design several controlled experiments on PASCAL VOC 2007 with our DSOD300 (with 300 × 300 inputs) for this ablation study. A consistent setting is imposed on all the experiments, unless when some components or structures are examined. In this study, we train the models with the combined training set from VOC 2007 trainval and 2012 trainval (“07+12”), and test on the VOC 2007 test set.

4.1.1 Configurations in Dense Blocks

In this section, we first investigate the impact of different configurations in dense blocks of the backbone sub-network. The results are mainly summarized in Table 2 and Table 3.*Compression Factor in Transition Layers.* We compare two compression factor values (θ = 0.5, 1) in the transition layers of DenseNets. Results are shown in Table 3 (rows 2 and 3). Compression factor θ = 1 means that there is no feature map reduction in the transition layer, while θ = 0.5 means half of the feature maps are reduced. We can observe that θ = 1 obtains 2.9 percent higher mAP than θ = 0.5.



Method	network	pre-train	# param	COCO (Avg. Precision, IoU:)		
				0.5:0.95	0.5	0.75
One-Stage Detectors:						
SSD300 [9]	VGGNet	✓	34.3M	23.2	41.2	23.4
SSD300* [9]	VGGNet	✓	34.3M	25.1	43.1	25.8
DSOD300	DSOD	✗	21.8M	29.3	47.3	30.6
DSOD300* (v2)	DSOD + **DSS**	✗	37.3M	**30.4**	49.0	**31.8**
Two-Stage Detectors:						
FPN300/500 [2]	ResNet-50	✓	83.3M	29.0	48.0	30.3
FPN300/500 [2]	ResNet-101	✓	121.2M	29.4	48.8	30.6
Mask RCNN+FPN300/500 [1]	ResNet-50	✓	84.4M	29.9	49.0	31.3
Mask RCNN+FPN300/500 [1]	ResNet-101	✓	122.4M	30.2	**49.3**	31.7

# Channels in Bottleneck Layers. As shown in Table 3 (rows 3 and 4), we observe that wider bottleneck layers (with more channels of response maps) improve the performance greatly (4.1 percent mAP).

# Channels in the 1st Conv-Layer. We observe that a large number of channels in the first conv-layers is beneficial, which brings 1.1 percent mAP improvement (in Table 3 rows 4 and 5).

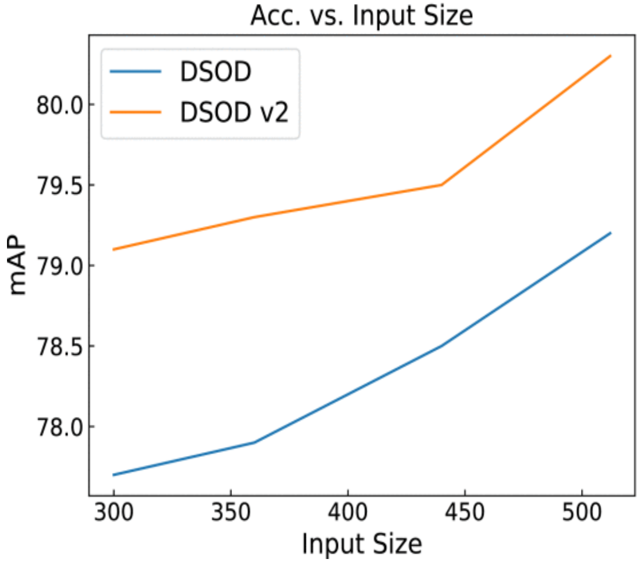
Growth Rate. A large growth rate k is found to be much better. We observe 4.8 percent mAP improvement in Table 3 (rows 5 and 6) when increase k from 16 to 48 with 4k bottleneck channels.

4.7 From DSOD to DSOD (v2)

Compared with DSOD, DSOD v2 includes the extra DSS module to further enhance the supervision signal under the training from scratch scenario. The comparison results of DSOD and DSOD v2 are shown in Table 9. We can see that DSOD v2 improves the performance consistently on both PASCAL VOC and COCO datasets under different training sets. In DSOD v2, we also replace the pre-activation of BN [69] in DSOD with post-activation (replacing BN-ReLU-Conv with the Conv-BN-ReLU manner), as shown in Fig. 5. We observe that this operation can improve the detection performance with about 0.6 percent mAP.

V.DISCUSSION

Better Model Structure versus More Training Data. An emerging idea in the computer vision community is that object detection or other vision tasks might be solved with deeper and larger neural networks backed with massive training data like ImageNet [3]. Thus more and more large-scale datasets have been collected and released recently, such as the Open Images dataset [71], which is 7.5x larger in the number of images and 6x larger of categories than that of ImageNet. We definitely agree that, under modest assumptions that given boundless training data and unlimited computational power, deep neural networks should perform extremely well. However, our proposed approach and experimental results imply an alternative view to handle this problem: a better model structure might enable similar or better performance compared with complex models trained from large data. Particularly, our DSOD is only trained with 16,551 images on VOC 2007, but achieves competitive or even better performance than those models trained with 1.2 million + 16,551 images.



In this premise, it is worthwhile rehashing the intuition that as datasets grow larger, training deep neural networks becomes more and more expensive. Thus a simple yet efficient approach becomes increasingly important. Despite its conceptual simplicity, our approach shows great potential under this setting.

Why Training from Scratch? There are many successful cases that fine-tuning works well and achieves consistent improvement, especially in object detection areas. So why do we still need to train object detectors from scratch? As aforementioned briefly, the critical importance of training from scratch has at least two aspects. First, there may have big domain differences between the pre-trained and the target one. For instance, most pre-trained models are learned on large-scale RGB dataset like ImageNet. It is fairly difficult to transfer RGB models to depth images, multi-spectrum images, medical images, etc. Some advanced domain adaptation techniques have been proposed and could mitigate this problem. But what an amazing thing if we have a technique that can train object detector from scratch. Second, fine-tuning restricts

| DOI: 10.15680/IJIRCCE.2022.1005076|

the design space of network structures for object detection. This is very critical for the deployment of applying deep neural networks to some resource-limited Internet-of-Things (IoT) scenario.

VI. CONCLUSION

We have presented Deeply Supervised Object Detector (DSOD), a simple yet efficient framework for learning object detectors from scratch. Without using pre-trained models from ImageNet, DSOD demonstrates competitive performance to state-of-the-art detectors such as SSD, Faster R-CNN, R-FCN, FPN, Mask RCNN, etc. on the popular PASCAL VOC 2007, 2012 and MS COCO datasets, meanwhile, with only 1/2, 1/4 and 1/10 parameters compared to SSD, R-FCN and Faster R-CNN, respectively. Due to the learning from scratch property, DSOD has great potential on domain-different scenarios, such ad depth, medical, multi-spectral images, etc. Our future work will consider learning object detectors directly in these diverse domains, as well as learning ultra efficient DSOD models to support resource-bounded devices.

REFERENCES

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” inProc. IEEE Int. Conf. Comput. Vis., 2017, pp. 2980–2988.

[2] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 936–944.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, et al., “ImageNet: A large- scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 248–255.

[4] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Learning and trans- ferring mid-level image representations using convolutional neu- ral networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 1717–1724.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hier- archies for accurate object detection and semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 580–587.

[6] R. Girshick, “Fast R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis.,2015, pp. 1440–1448.

[7] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real- time object detection with region proposal networks,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2015, pp. 91–99.

[8] Y. Li, K. He, J. Sun, et al., “R-FCN: Object detection via region- based fully convolutional networks,” in Proc. Int. Conf. Neural Inf.Process. Syst., 2016, pp. 379–387.

[9] W. Liu, D. Anguelov, D. Erhan, et al., “SSD: Single shot multibox detector,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 21–37.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE Conf. Com- put. Vis. Pattern Recognit., 2016, pp. 779–788.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

Scan to save the contact details

www.ijircce.com