Co-occurrence Matrix Analysis Based Semi-Supervised Training for Object Detection

Min-Kook Choi[1], Jaehyeong Park[1], Jihun Jung[1], Heechul Jung[2], Jin-Hee Lee[1], Woong-Jae Won[1], Woo Young Jung[1], Jincheol Kim[3], and Soon Kwon*[1]

DGIST[1], KAIST[2], SK Telecom[3], Republic of Korea

{MKCHOI, STILLRUNNING, JIHUN.JUNG, JHLEE07, WWJ, WYJUNG, SOONYK}@DGIST.AC.KR[1], HEECHUL@KAIST.AC.KR[2], JINCHEOL.B.KIM@SK.COM[3]

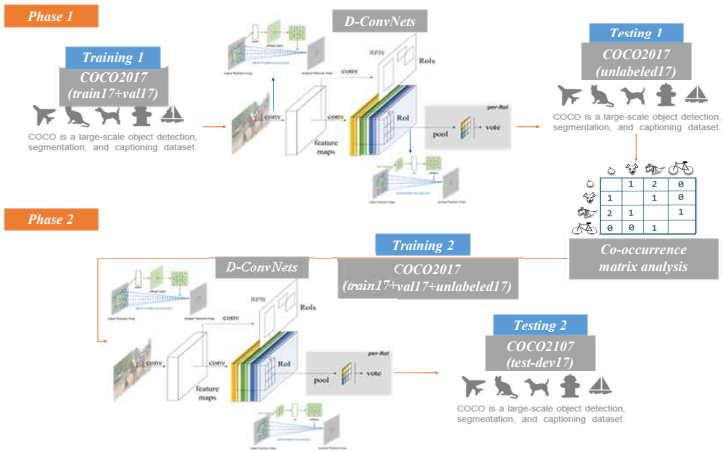


Figure 1. **Overview of the proposed SSL pipeline**. The pro-posed technique consists of two steps. In the first step, the detector for labeling the unlabeled data is trained with the existing annotated data (Training 1), and then the inferring process for the unlabeled data is performed (Testing 1). After performing pseudo-labeling through one-hot-vector encoding and co-occurrence matrix analysis, a new network is trained (Training 2) and the data for evaluation are inferred (Testing 2).

Abstract

One of the most important factors in training object recognition networks using convolutional neural networks (CNN) is the provision of annotated data accompanying human judgment. Particularly, in object detection or semantic segmentation, the annotation process requires considerable human effort. In this paper, we propose a semi-supervised learning (SSL)-based training methodology for object detection, which makes use of automatic labeling of un-annotated data by applying a network previously trained from an annotated dataset. Because an inferred label by the trained network is dependent on the learned parameters, it is often meaningless for re-training the network. To transfer a valuable inferred la- bel to the unlabeled data, we propose a re-alignment method based on co-occurrence matrix analysis that takes into account one-hot-vector encoding of the estimated label and the correlation between the objects in the image. We used an MS-COCO detection dataset to verify the performance of the proposed SSL method and deformable neural networks (D-ConvNets) [1] as an object detector for basic training. The performance of the existing state-of-the-art detectors (D-ConvNets, YOLO v2 [2], and single shot multi-box detector (SSD) [3]) can be improved by the proposed SSL method without using the additional model parameter or modifying the network architecture.

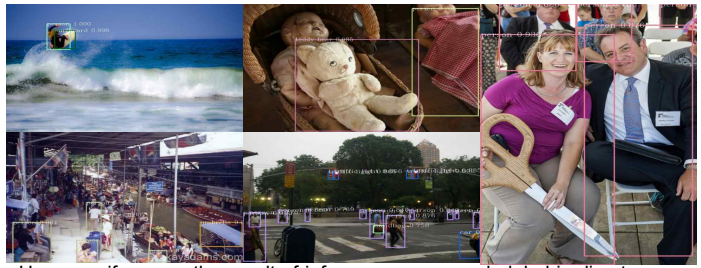
1. Introduction

- There are two main types of end-to-end training object detectors (one and two stage detector) that utilize CNN as a backbone architecture. Both types of networks have played a significant role in improving the dramatic performance of CNN and the decoder network for multi-tasking.
- Despite the dramatic improvement in performance of state-of-the-art detectors, object detectors trained by machine learning techniques have the disadvantage of having a large capacity for the refined datasets for training. Figure 2 shows examples of annotation tools guided by human's efforts.
- In this paper, we propose a simple but powerful one-hot-vector encoding based on the SSL idea and a semi-supervised training method through co-occurrence matrix analysis (see Fig. 1).
- As a result of testing the SSL scheme with the MS-COCO detection dataset, we confirmed the performance improvement in the state-of-the-art detectors such as deformable neural networks (D-ConvNets) [1], YOLO v2 [2], and single shot multi-box detector (SSD) [3] in terms of accuracy using mean average precision (mAP) without any additional parameter or architecture modification.

Figure 2. **Examples of annotation tools for the object detection.** Annotation process that rely on human labor impose a heavy burden on learning-based algorithms, especially annotation costs to investigate well-trained network for objection detection.

2. Method

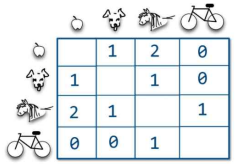
- The latest performance networks deduce a bounding box of the correct form that can be used as training data in a specific object or visual environment. Below figure shows the inferenced output from D-ConvNets trained with MS-COCO dataset.



- However, if we use the result of inference as a pseudo label in direct way, we cannot maximize the training efficiency by dependency of parameter and data.
- In order to compensate for the effect of pseudo-labeling during training, 1) the inference result is encoded as a one-hot-vector and 2) the co-occurrence matrix obtained from the prior knowledge is used to recalculate whether the inference result is suitable for training.

$$p(x_1,\dots,x_n)=\prod_{i=1}^{n}p(x_i|x_1,\dots,x_{i-1})\approx\max p(x|x_{i-1}),\forall i=1,\dots,n \quad \text{Conditional Independence}$$

$$p(\text{apple}\ |\text{dog},\text{horse},\text{bike})\approx p(\text{apple}\ |\text{horse})$$



Conditional Marginalization

0	0.298	0.413	0.289
0.877	0	0.089	0.035
0.893	0.065	0	0.042
0.903	0.037	0.06	0

Max-normalization

0	0.722	1.0	0.7
1.0	0	0.101	0.04
1.0	0.073	0	0.5
1.0	0.41	0.067	0

- To reflect the extracted correction probabilities, we need to re-scale the inferred softmax probability for pseudo-labeling with the co- occurrence matrix values.

$$LB(g_i)=\begin{cases}[\hat{x},\hat{y},\hat{w},\hat{h},c], & \dfrac{\exp(q_j)}{\sum_{j=1}^{n}\exp(q_j)}\sigma>\rho_{co}\\ [\,], & \text{otherwise}\end{cases}$$

3. Experimental Results

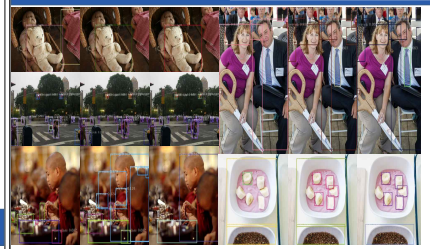


Figure 3. **Examples of pseudo-labeling results.** There is a large difference in the result of pseudo labeling according to the set threshold value. For the first row, we could remove the bounding box for the mis-inferred object, and for the second and third rows, we detected additional objects in the complex scene. The fourth row detected a small tie object, which is difficult to deduce in a complex scene, based on the relation between objects. The final row detected additional bounding boxes of undetected objects.

Table 1: MS-COCO detection dataset evaluations for [0.5:0.05:0.95] using D-ConvNets with different parameters.

Model (backbone, SSL parameter(s), training dataset)	mAP
D-ConvNets (ResNet-101, none, train17 + val17)	36.3
D-ConvNets (ResNet-101, $\rho = 0.5$, train17 + val17 + unlabeled17)	37.0
D-ConvNets (ResNet-101, $\rho = 0.7$, train17 + val17 + unlabeled17)	36.7
D-ConvNets (ResNet-101, $\rho = 0.5$, $\rho_{co} = 0.1$, train17 + val17 + unlabeled17)	37.6
D-ConvNets (ResNet-101, $\rho = 0.5$, $\rho_{co} = 0.2$, train17 + val17 + unlabeled17)	37.3
D-ConvNets (ResNet-101, $\rho = 0.5$, $\rho_{co} = 0.3$, train17 + val17 + unlabeled17)	**37.8**
D-ConvNets (ResNet-101, $\rho = 0.5$, $\rho_{co} = 0.4$, train17 + val17 + unlabeled17)	37.5

Table 2: MS-COCO detection dataset evaluations for [0.5:0.05:0.95] using different architectures with or without the proposed SSL ($\rho = 0.5$, $\rho_{co} = 0.3$).

Model (backbone CNN)	mAP	mAP with SSL
SSD [3] (ResNet-101)	24.1	**25.3**
YOLO v2 [2] (Darknet-19)	24.0	**25.1**
D-ConvNets [1] (ResNet-101)	36.3	**37.8**

Reference

[1] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in International Conference on Computer Vision (ICCV), 2017.

[2] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in Computer Vision and Pattern Recognition (CVPR), 2017.

[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in European Conference on Computer Vision (ECCV), 2016.