DeepID-Net: Deformable Deep Convolutional Neural Networks for Object Detection

Wanli Ouyang, Xiaogang Wang, Xingyu Zeng, Shi Qiu, Ping Luo, Yonglong Tian,
Hongsheng Li, Shuo Yang, Zhe Wang, Chen-Change Loy, Xiaoou Tang
The Chinese University of Hong Kong

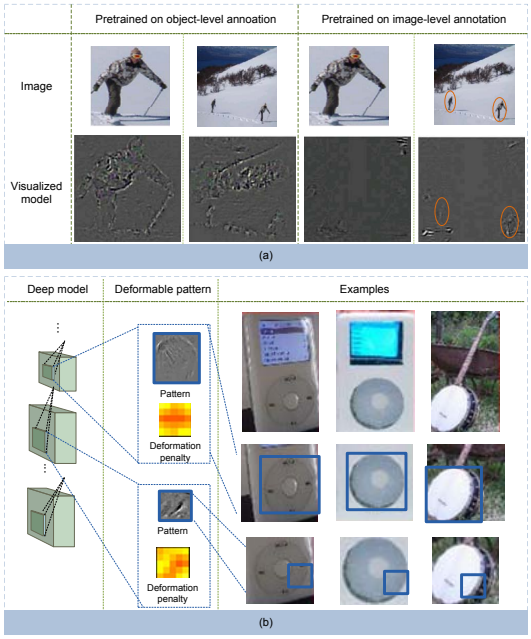

Figure 1: The motivation in new pretraining scheme (a) and jointly learning feature representation and deformable object parts shared by multiple object classes at different semantic levels (b). In (a), Model pretrained on image-level annotation is more robust to size and location change while model pretrained on object-level annotation is better in representing object with tight bounding box. In (b), when ipod rotates, its circular pattern moves horizontally at the bottom of the bounding box. Therefore, the circular patterns have smaller penalty moving horizontally but higher penalty moving vertically. The curvature part of the circular pattern are often at the bottom right positions of the circular pattern. *Best viewed in color.*

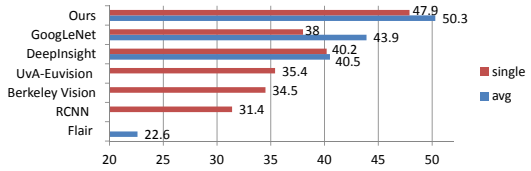

Figure 2: Mean AP on ILSVRC 2014 test data.

Object detection is one of the fundamental challenges in computer vision. It has attracted a great deal of research interest. In this paper, we propose deformable deep generIc object Detection convolutional neural NETwork (DeepID-Net). In DeepID-Net, we jointly learn the feature representation and part deformation for a large number of object categories. We also investigate many aspects in effectively and efficiently training and aggregating the deep models, including bounding box rejection, training schemes, context modeling, and model averaging. The proposed new framework significantly advances the state-of-the-art for deep learning based generic object detection, such as the well known RCNN [3] framework. Table 1 provides detailed component-wise experimental results on how our approach can improve the mean Averaged Precision (AP) obtained by RCNN [3] from 31.0% to mean AP 50.3% step-by-step on the ImageNet Large Scale Visual Recognition Challenge 2014 (ILSVRC2014) object detection task. Comparison with state-of-the-art is shown in Fig. 2.

There are three contributions of this paper:

1. A new scheme for pretraining the deep CNN model. We propose to pretrain the deep model on the ImageNet image classification and localization dataset with 1000-class object-level annotations instead of with image-level annotations, which are commonly used in existing deep learning object detection [3, 5]. Then the deep model is fine-tuned on the ImageNet/PASCAL-VOC object detection dataset with 200/20 classes, which are the targeting object classes in the two datasets. The motivation of this pretraining scheme is shown in Fig. 1(a).
2. A new deformation constrained pooling (def-pooling) layer, which enriches the deep model by learning the deformation of object parts at any information abstraction levels. The def-pooling layer can be used for replacing the max-pooling layer and learning the deformation properties of parts. The motivation of this new layer is shown in Fig. 1(b). The def-pooling layer is a more general representation of the deformation constraint in the deformable part based model (DPM) [2] and the deformation layer in [4].
3. A new deep learning pipeline for object detection as shown in Fig. 3. Detailed component-wise analysis is also provided through extensive experimental evaluation, as shown in Table. 1. It provide a global view for people to understand the deep learning object detection pipeline. Table 1 summarizes how performance gets improved by adding each component step-by-step into our pipeline. RCNN has mAP 29.9%. With bounding box rejection, changing A-net to G-net, replacing image-level annotation by object-level annotation in pretraining, combining candidates from selective search and edgeboxes [6], the def-pooling layer, pretraining the object-level annotation with multiple scales [1], and adding the contextual information from image classification scores, mAP is increased to 48.2%. With model averaging, the final result is 50.7%.

Table 1: Experimental results on the pipeline for single model tested on ILSVRC2014 val2.

pipeline	RCNN	+reject bbox	A-net to G-net	bbox pretrain	+edge box	+Def pooling	multi-scale pretrain	+ctx	+bbox
mAP (%)	29.9	30.9	37.8	40.4	42.7	44.9	47.3	47.8	48.2

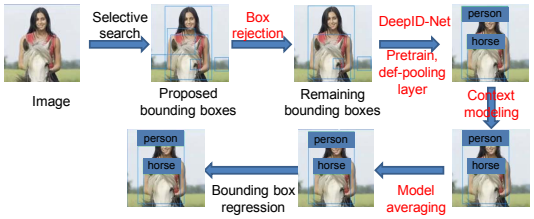

Figure 3: Overview of our approach. Find detailed description in the paper. Texts in red highlight the steps that are not present in RCNN [3].

[1] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014.

[2] P. Felzenszwalb, R. B. Grishick, D.McAllister, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans. PAMI*, 32: 1627–1645, 2010.

[3] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[4] W. Ouyang and X. Wang. Joint deep learning for pedestrian detection. In *ICCV*, 2013.

[5] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014.

[6] C. Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.