Single-Shot Object Detection with Enriched Semantics

Zhang, Liqiang

August 10, 2018

Abstract

In this paper, the author propose a novel single shot object detection network named Detection with Enriched Semantics (DES). This motivation is to enrich the semantics of object detection features within a typical deep detector, by a semantic segmentation branch and a global activation module. The segmentation branch is supervised by weak segmentation ground-truth, i.e., no extra annotation is required. Comprehensive experimental results on both PASCAL VOC and MS COCO detection datasets demonstrate the effectiveness of the proposed method. And They achieve an mAP of 81.7 on VOC2007 test and an mAP of 32.8 on COCO test-dev with an inference speed of 31.5 milliseconds per image on a Titan Xp GPU. With a lower resolution version, we achieve an mAP of 79.7 on VOC2007 with an inference speed of 13.0 milliseconds per image.

1. Introduction

With the emergence of deep neural networks, computer vision has been improved significantly in many aspects such as image classification [2], object detection [1], and segmentation. Among them, object detection is a fundamental task which has already been extensively studied. Currently there are mainly two series of object detection frameworks: the two-stage frameworks such as Faster-RCNN and R-FCN which extract proposals, followed by per-proposal classification and regression; and the one-stage frameworks such as YOLO [20] and SSD, which apply object classifiers and regressors in a dense manner without objectnessbased pruning. Both of them do classification and regres- sion on a set of pre-computed anchors.

In this paper, The author aim to address the problem discussed above, by designing a novel single shot detection network, named Detection with Enriched Semantics (DES), which consists of two branches, a detection branch and a segmentation branch. The detection branch is a typical single shot detector, which takes VGG16 as its backbone, and detect objects with multiple object detection feature maps in different layers. This is shown in the upper part of Fig. 2.

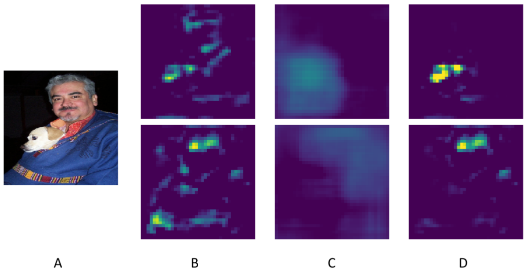

Figure 1. Low level features augmented with semantic meaningful features from the segmentation branch. A: original image fed into our detection network. B: original low level detection features (X) for the input image. C: semantic meaningful features (Z) from the segmentation branch. D: augmented low level features which is then used in the later stages for our detection network.

Fig. 1 gives an illustration of this semantic augmentation process. After the original low level features (B) are activated by segmentation features (C), the augmented low level features (D) can capture both the basic visual pattern as well as the semantic information of the object. This can be considered as an attention process, where each channel of the original low level feature map is activated by a semantically meaningful attention map, to combine both basic visual pattern and semantically meaningful knowledge.

2. Proposed method

Detection with Enriched Semantics (DES) is a single-shot object detection network with three parts: a single shot detection branch, a segmentation branch to enrich semantics at low level detection layer, and a global activation module to enrich semantics at higher level detection layers. As shown in the left lower part in Fig. 1, their segmentation branch takes *conv4_3* as input, rep- resented by the black arrow pointed from *conv4_3* to segmentation branch.

Mathematically, let $X \in \mathbb{R}^{C\times H\times W}$ be the low level detection feature map from the detection branch, $G \in 0, 1, 2, \ldots, N^{H\times W}$ be the segmentation ground-truth where

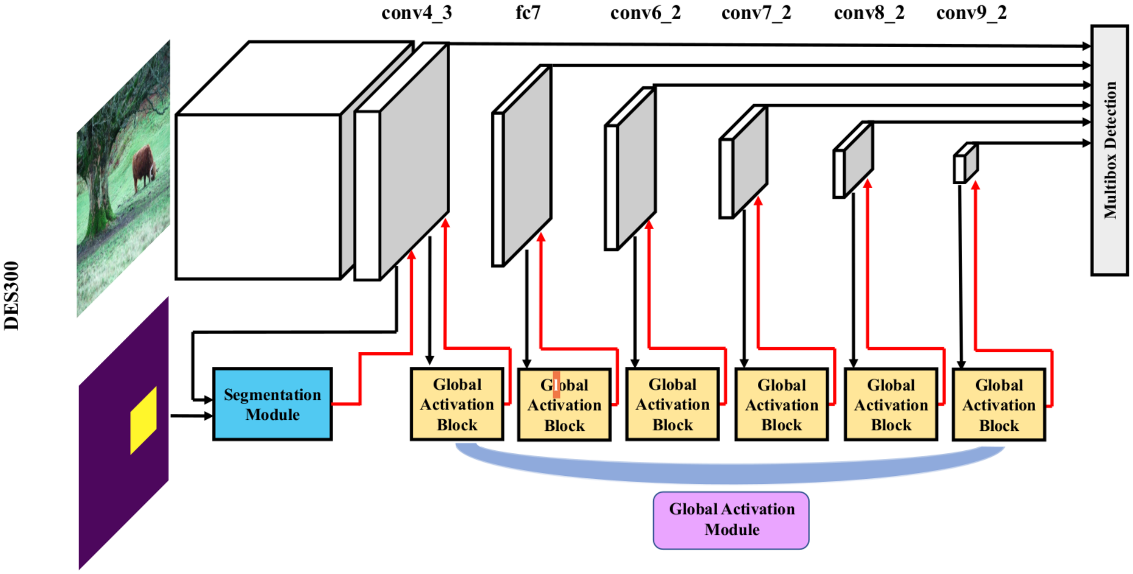

Figure 2. Pipeline for DES: the upper half is the object detection branch for DES which has six prediction source layers from *conv4_3* up to *conv9_2*; the lower half is the segmentation branch and the global activation module. The segmentation branch is added at the first prediction source layer *conv4_3*. The global activation module consists of six global activation blocks. Those global activation blocks are added at each prediction source layer. The black arrows pointed to those modules are the input flow, and the red arrows pointed out from those modules are the output flow to replace the original feature map.

N is the number of classes (20 for VOC and 80for COCO). The segmentation branch computes $Y \in \mathbb{R}^{(N+1)\times H\times W}$ as the prediction of per-pixel segmentation where

$$Y = \mathcal{F}(\mathcal{G}(X)) \tag{1}$$

satisfying

$$Y \in [0,1]^{(N+1)\times H\times W}, \sum_{c=0}^{N} Y_{c,h,w} = 1. \tag{2}$$

$\mathcal{G}(X) \in \mathbb{R}^{C'\times H\times W}$ is the intermediate result which will be further used to generate semantic meaningful feature map:

$$Z = \mathbb{H}(\mathbb{G}(X)) \in \mathbb{R}^{C\times H\times W}. \tag{3}$$

3. Conclusion

In this paper, the author propose a novel single shot object detector named Detection with Enriched Semantics (DES). To address the problem that low level detection feature map does not have high level semantic information, they introduce a segmentation branch, which utilize the idea of weakly supervised semantic segmentation, to provide high semantic meaningful and class-aware features to activate and calibrate feature map used in the object detection.

References

[1] J. Dai, Y. Li, K. He, and J. Sun. [r-fcn]: Object detection via region-based fully convolutional networks. In *NIPS*, pages 379–387, 2016. 1

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1