Improving Object Detection with Deep Convolutional Networks via Bayesian Optimization and Structured Prediction

Xuewen Yang

August 3 2018

Abstract

Object detection systems based on the deep convolutional neural network (CNN) have recently made groundbreaking advances on several object detection benchmarks.While the features learned by these high-capacity neural networks are discriminative for categorization, inaccurate localization is still a major source of error for detection.Building upon high-capacity CNN architectures,the author address the localization problem by 1) using a search algorithm based on Bayesian optimization that sequentially proposes candidate regions for an object bounding box, and 2) training the CNN with a structured loss that explicitly penalizes the localization inaccuracy. In experiments, they demonstrate that each of the proposed methods improves the detection performance over the baseline method on PASCAL VOC 2007 and 2012 datasets. Furthermore, two methods are complementary and significantly outperform the previous state-of-the-art when combined.

1. Introduction

Object detection is one of the long-standing and important problems in computer vision. Motivated by the recent success of deep learning [2] [6] [1] on visual object recognition tasks, significant improvements have been made in the object detection problem. Most notably, Girshick *et al.* proposed the regions with convolutional neural network (R-CNN) framework for object detection and demonstrated state-of-the-art performance on standard detection benchmarks (e.g., PASCAL VOC, ILSVRC) with a large margin over the previous arts, which are mostly based on deformable part model (DPM) [4].

There are two major keys to the success of the R-CNN. First, features matter. In the R-CNN, the low-level image features (e.g., HOG [3]) are replaced with the CNN features, which are arguably more discriminative representations. One drawback of CNN features, however, is that they are expensive to compute. The R-CNN overcomes this issue by proposing a few hundreds or thousands candidate bounding boxes via the selective search algorithm [44] to effectively reduce the computational cost required to evaluate the detection scores at all regions of an image.

Despite the success of R-CNN, it has been pointed out through an error analysis that inaccurate localization causes the most egregious errors in the R-CNN framework. For example, if there is no bounding box in the close proximity of ground truth among those proposed by selective search, no matter what we have for the features or classifiers, there is no way to detect the correct bounding box of the object. Indeed, there are many applications that require accurate localization of an object bounding box, such as detecting moving objects (e.g., car, pedestrian, bicycles) for autonomous driving [5], detecting objects for robotic grasping or manipulation in robotic surgery or manufacturing, and many others.

2. Fine-grained search for bounding box via Bayesian optimization

In this section, the author develop a fine-grained search (FGS) algorithm based on Bayesian optimization that sequentially proposes a new bounding box with a higher expected detection score than previously proposed bounding boxes without significantly increasing number of region proposals.

2.1. General Bayesian optimization framework

In the Bayesian optimization framework, $f = f(x; y)$ is assumed to be drawn from a probabilistic model:$p(f|D_N)\infty p(D_N|f)p(f)$.Bayesian optimization is efficient in terms of the number of function evaluation [24], and is particularly effective when f is computationally expensive.

2.2. Local finegrained search

In this section, the author extend the GPR-based algorithm for global maximum search to local fine-grained search (FGS).They perform the FGS by pruning out easy negatives with low classification scores from the set of regions proposed by the selective search algorithm and sorting out a few bounding boxes with the maximum scores in local regions. Then, for each local optimum y_{best}, they propose a new candidate bounding box.Specifically, they initialize a set of local observations D_{local} for y_{best} from the set given by the selective search algorithm, whose localness is measured by an IoU between y_{best} and region proposals.

References

[1] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE TPAMI*, 2013. 1

[2] Y. Bengio, P. Lamblin, P. Dan, and H. Larochelle. Greedy layer-wise training of deep networks. In *NIPS*, 2007. 1

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1

[4] P. F. Felzenszwalb, R. B. Girshick, D. Mcallester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 2010. 1

[5] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012. 1

[6] M. A. Ranzato, Y. L. Boureau, and Y. Lecun. Sparse feature learning for deep belief networks. In *NIPS*, 2007. 1