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

Yuting Zhang*†, Kihyuk Sohn†, Ruben Villegas†, Gang Pan*, Honglak Lee†
* Department of Computer Science, Zhejiang University, Hangzhou, Zhejiang, China
† Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA

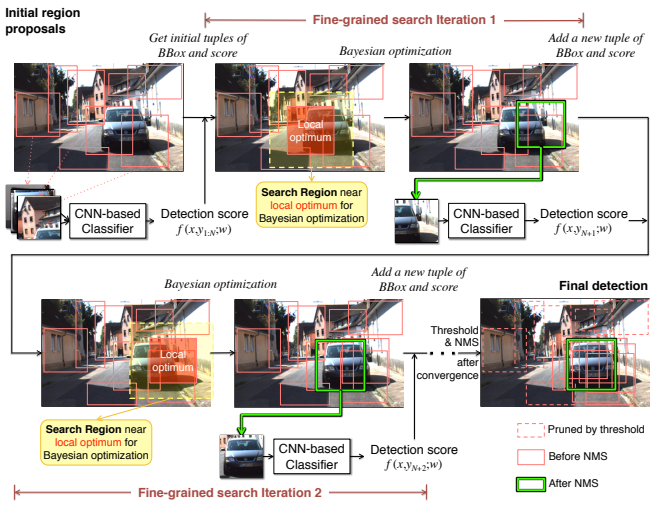

Figure 1: Pipeline of our method.

Motivated by the recent breakthrough of deep convolutional neural networks (CNN) on large scale visual object recognition tasks [3], Girshick et al. [2] proposed the "regions with CNN" (R-CNN) framework for object detection and demonstrated state-of-the-art performance on standard detection benchmarks with a large margin to the previous arts, which are mostly based on deformable part models (DPM). While the features learned by high-capacity neural networks are discriminative for categorization, inaccurate localization is still a major source of error for detection.

In this work, we address the localization difficulty of R-CNN detection framework with two ideas. First, we develop a fine-grained search algorithm to expand an initial set of bounding boxes by proposing new bounding boxes with scores that are likely to be higher than the initial ones. We build our algorithm in the Bayesian optimization framework [5]. Second, we train a CNN classifier with a structured SVM objective that aims at classification and localization simultaneously. We define the structured SVM layer of the CNN whose objective function is defined with a hinge loss that balances between classification (i.e., determines whether an object exists) and localization (i.e., determines how much it overlaps with the ground truth).

Fine-grained search (*FGS*) via Bayesian optimization Let $f(x,y)$ denote a detection score function of an image x at region defined with the bounding box coordinates $y=(u_1,v_1,u_2,v_2)\in\mathcal{Y}$. The object detection problem is to find the local maximum of $f(x,y)$ with respect to y of an unseen image x.

Let $\{y_1,\cdots,y_N\}$ be the set of solutions (e.g., bounding boxes). In the Bayesian framework,

$$p(f|\mathcal{D}_N)\propto p(\mathcal{D}_N|f)\,p(f), \tag{1}$$

where $\mathcal{D}_N=\{(y_j,f_j)\}_{j=1}^N$ and $f_j=f(x,y_j)$. Here, the goal is to find a new solution y_{N+1} that maximizes the chance of improving the detection score f_{N+1}, where the chance is defined as an acquisition function $a(y_{N+1}|\mathcal{D}_N)$ (e.g., expected improvement). Specificially, $p(x)$ is defined as a Gaussian process (GP), and $p(y_{N+1}|,x_{N+1}\mathcal{D}_{N+1})$ can be obtained by GP regression.

The algorithm proceeds by recursively sampling a new solution y_{N+t} from $\mathcal{D}_{N+(t-1)}$, and update the set $\mathcal{D}_{N+t}=\mathcal{D}_{N+(t-1)}\cup\{y_{N+t}\}$ to draw a new sample solution $y_{N+(t+1)}$ with an updated observation.

As it is shown in Figure 1, our pipeline is: 1) Initial bounding boxes are given by methods such as the selective search [6] and their detection scores are obtained from the CNN-based classifier trained with structured SVM objective. 2) The box(es) with optimal score(s) in the local regions are selected by greedy NMS [2], and Bayesian optimization takes the resulting neighborhood of each local optimum to propose a new box with high chance of getting a better score. 3) We evaluate the detection score of the new box and take it as an observation to move to next step until convergence. 4) All the bounding boxes are fed into the standard post-processing stage (e.g., threshold and NMS, etc.).

mean Average Precision		VOC 2007		VOC 2012
Model	BBoxReg	IoU$\geq$ 0.5	IoU$\geq$ 0.7	IoU$\geq$ 0.5
R-CNN (AlexNet)	No	54.2	26.6	49.6
R-CNN (VGGNet)	No	60.6	30.8	59.5
+ StructObj	No	61.2	31.0	-
+ StructObj-FT	No	62.3	33.2	-
+ FGS	No	64.8	37.4	-
+ StructObj + FGS	No	65.9	37.2	-
+ StructObj-FT + FGS	No	**66.5**	**39.8**	-
R-CNN (AlexNet)	Yes	58.5	35.2	53.3
R-CNN (VGGNet)	Yes	65.4	35.2	63.0
+ StructObj	Yes	66.6	40.5	65.1
+ StructObj-FT	Yes	66.9	41.8	-
+ FGS	Yes	67.2	42.7	64.0
+ StructObj + FGS	Yes	**68.5**	43.0	**66.4**
+ StructObj-FT + FGS	Yes	68.4	**43.7**	-

Table 1: Test set mAPs on PASCAL VOC 2007 and VOC 2012

Learning R-CNN with structured loss Suppose the top layer of the CNN is a linear classifier $f(x,y;w)=w^T\phi(x,y)$, where $\phi(x,y)$ denotes the CNN features from the previous layer. Following Blaschko and Lampert [1], we formulate the classifier learning as a structured SVM problem, where the structured loss penalizes both the classification and localization errors. In contrast to their solution, we restrict the output space to regions proposed via selective search, and transform the constraints into hinge loss to make the objective function backpropagatable through the CNN. We alternately perform a gradient-based parameter estimation and hard negative data mining that effectively adapts the number of training examples to be evaluated for updating the parameters. For model parameter estimation, we use L-BFGS to first learn parameters of the classification layer only (***StructObj***) . We found that this already resulted in a good detection performance. Then, we optionally use stochastic gradient descent to finetune the whole CNN classifier (***StructObj-FT***).

Experimental Results In Table 1, we demonstrated that each of the proposed methods improves the detection performance over the baseline method (R-CNN with VGGNet [4]) on PASCAL VOC 2007 and 2012 datasets. Furthermore, two methods are complementary and significantly outperform the previous state-of-the-art when combined.

[1] M. B. Blaschko and C. H. Lampert. Learning to localize objects with structured output regression. In *ECCV*, 2008.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[5] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *NIPS*, 2012.

[6] J. R. R. Uijlings, K. E. A. Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

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