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

Yuting Zhang*†, Kihyuk Sohn†, Ruben Villegas†, Gang Pan*, Honglak Lee†

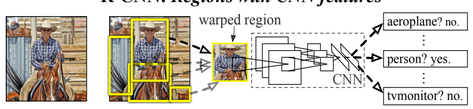
† University of Michigan

* Zhejiang University

CVPR 2015 BOSTON JUNE 7-12

1 Revisiting R-CNN

- Convolutional neural network (CNN) for object detection
- The state-of-the-art: "Regions with CNN" (R-CNN)



[Image taken from Girshick et al, (2015)]

Main technique	Region proposal	CNN trained with classification objective
Pros	Reasonable efficiency	Strong discriminative power
Cons	Proposed bounding boxes may have poor overlap with GT	Insensitive to localization errors
Our solutions	Find better boxes via Bayesian optimization	Improve localization sensitivity via structured objective

2 R-CNN with structured objective (StructObj)

- Training set: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_M, y_M)\}$
  - x_i is the i^{th} image, and y_i is its box annotation (if applicable).
- Linear classifier:

$$f(x, y; w) = w^\top \phi(x, y)$$

A revisit of (Blaschko and Lampert, 2008)

- Structured objective:

$$\min_w \frac{1}{2}\|w\|^2 + \frac{C}{M}\sum_{i=1}^{M} \xi_i, \quad \text{subject to}$$

$$w^\top \phi(x_i, y_i) \geq 1 - \xi_i, \ \forall i \in I_{\text{pos}}, \qquad \text{➢ Classification constraints}$$

$$w^\top \phi(x_i, y) \leq -1 + \xi_i, \ \forall y \in \mathcal{Y}, \forall i \in I_{\text{neg}}$$

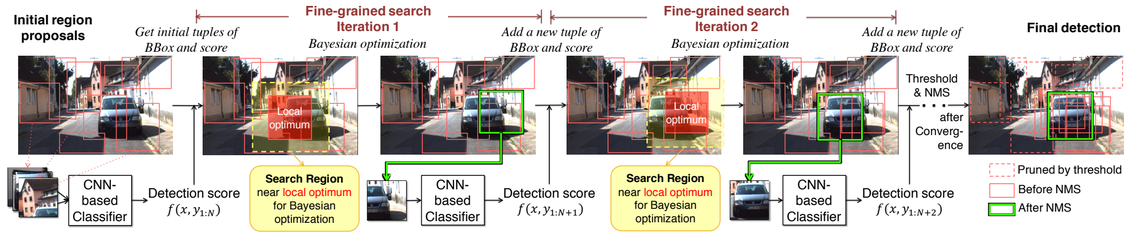
$$w^\top \phi(x_i, y_i) \geq w^\top \phi(x_i, y) + \Delta^{\text{loc}}(y, y_i) - \xi_i, \qquad \text{➢ Localization constraints}$$

$$\forall y \in \mathcal{Y}, \forall i \in I_{\text{pos}},$$

where $\Delta^{\text{loc}}(y, y_i) = 1 - \text{IoU}(y, y_i)$. $I_{\text{pos}}, I_{\text{neg}}$: pos / neg indexes.

- Gradient descent using $\mathcal{Y} \approx \mathcal{Y}_i$ (boxes sampled by selective search). L-BFGS for classification layer only learning, SGD for fine-tuning.
- Hard negative mining is necessary to make the training practical.

3 Fine-grained search (FGS) for bounding boxes via Bayesian optimization



a. Detection framework

- Image x, bounding box $y = (u, v, w, h) \in \mathbb{R}^4$, score function f.
- Detection is to locally (non-maximum suppression) find

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} f(x, y)$$

- Ideally, $\mathcal{Y} = \mathbb{R}^4$ is continuous.
- In practice, $\mathcal{Y}$ consists of boxes sampled
  - densely in a grid (sliding window) or
  - sparsely according to segmentations (region proposal).

b. General Bayesian optimization

- **Given:** existing boxes $y_{1:N} = \{y_1, y_2, \ldots, y_N\}$ and their detection scores $f_j = f(x, y_j)$, say, $\mathcal{D}_N = \{(y_j, f_j)\}_{j=1}^N$
- **Framework**: we model f using the Bayesian framework:

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

- **Goal:** To find a new box y_{N+1} that maximizes the chance of improving the detection score, i.e., $f_{N+1} > \hat{f}_N = \max_{j=1,2,\ldots,N} f_j$ where the chance is measured by expected improvement:

$$a_{EI}(y_{N+1}|\mathcal{D}_N) = \int_{\hat{f}_N}^{\infty} (f_{N+1} - \hat{f}_N) \cdot p(f_{N+1}|y_{N+1}, \mathcal{D}_N) df_{N+1}$$

- **Iterations:** $\mathcal{D}_{N+t} = \mathcal{D}_{N+t-1} \cup \{(y_{N+t}, f_{N+t})\}$

c. Region proposal via Gaussian process regression

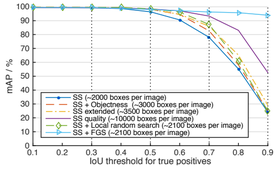
- The prior $p(f)$ is defined as a Gaussian process (GP)
  - Formally $f_{1:N}|y_{1:N} \sim \mathcal{N}([m(y_j)]_{1 \leq j \leq N}, [k(y_i, y_j)]_{1 \leq i,j \leq N})$
- $m(\cdot)$ is a constant mean, $k(\cdot, \cdot)$ is a squared exponential covariance.
- $k(\cdot, \cdot)$ is scale-invariant, as a latent scaling factor z is incorporated. It is determined by maximizing the likelihood of existing observations,

$$\hat{z} = \arg\max_z p(f_{1:N}|y_{1:N}; z)$$

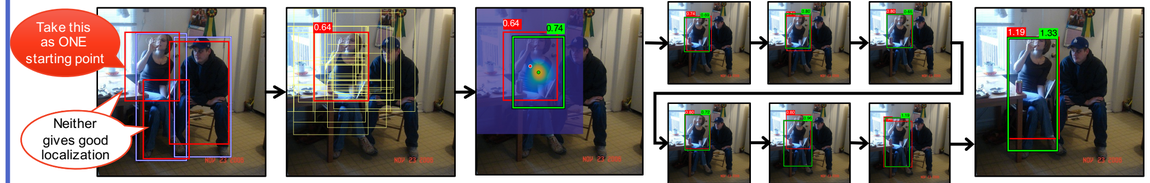
- The expected improvement (EI) and its gradients are in closed forms.

d. Control experiments: FGS with an Oracle detector

- Oracle detector $f_{\text{ideal}} = \text{IoU}(y, y_i)$ where y_i is the i^{th} ground truth.
- x-axis: levels of localization accuracy for accepting TPs
- Perfect search algorithm should lead to 100% mAP at any level.
- If the object detector is accurate enough, our method practically bounds the number of bounding box proposals to ~2000 to achieve near-perfect detection.



Numbers: actual scores White+Blue: ground truth BBs Red: current optimal BBs Yellow: existing BBs Green: newly proposed BBs



Initial detection results | Local search region | EI heat map and its maximum (new box) | Iteration 2 ~ 7 | Iteration 8 (final)

4 Experiments

a. mAP on VOC 2007 & 2012 test set

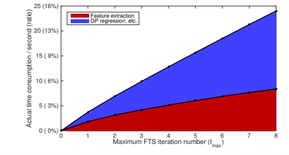
mean Average Precision		VOC 2007		VOC 2012
Model	BBoxReg	IoU> 0.5	IoU> 0.7	IoU> 0.5
R-CNN (AlexNet)	No	54.2	26.6	49.6
R-CNN (VGGNet)	No	60.6	30.8	59.5
+ StructObj	No	61.2	31.0	-
+ StructObj-FT	No	62.3	33.2	-
+ FGS	No	64.8	37.4	-
+ StructObj + FGS	No	65.9	37.2	-
+ StructObj-FT + FGS	No	**66.5**	**39.8**	-
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**	-

R-CNN (VGGNet) + StructObj + FGS on VOC 2007:

# GP iter	0	1	2	3	4	5	6	7	8
mAP	66.6	67.5	67.8	68.2	68.3	68.6	68.4	68.6	68.5

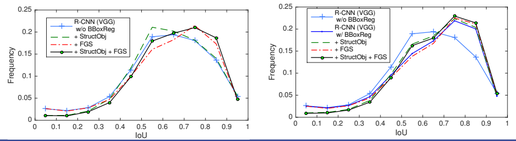
c. FGS efficiency

- "R-CNN (VGGNet) + StructObj + FGS" on VOC 2007 test set
- Baseline: the time consumption of initial feature extraction
- Ablation: Features + GP regression



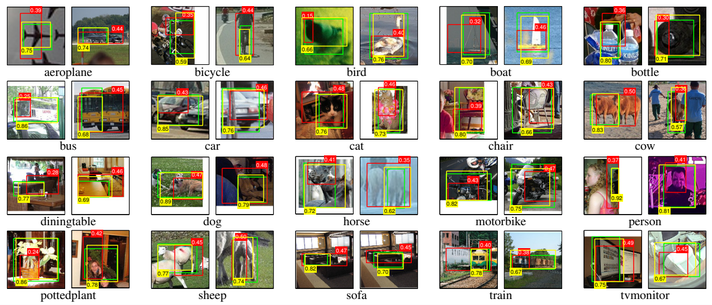
c. Localization accuracy on VOC 2007 test set

- Distribution of the overlap between a ground-truth and the predicted box best localizing it.



d. Examples with large improvement on VOC 2007 test set

Green: Ground truth Red: R-CNN (VGGNet) Yellow: Ours (+ StructObj + FGS)



References

(Girshick et al, 2015) R. Girshick, J. Donahue, T. Darrell, J. Malik, "Region-based Convolutional Networks for Accurate Object Detection and Semantic Segmentation", PAMI, 2015.
(Blaschko and Lampert, 2008) M. B. Blaschko and C. H. Lampert, "Learning to localize objects with structured output regression", ECCV, 2008.

Get code & models: http://bit.ly/fgs-obj