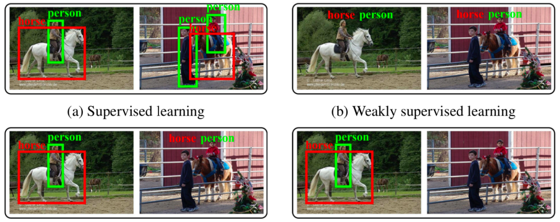
Consistency-based Semi-supervised Learning for Object Detection

Jisoo Jeong*, Seungeui Lee*, Jeesoo Kim and Nojun Kwak, {soo3553, seungeui.lee, kimjiss0305, nojunk}@snu.ac.kr

Seoul National University

Introduction



(a) Supervised learning

(b) Weakly supervised learning

(c) Weakly semi-supervised learning

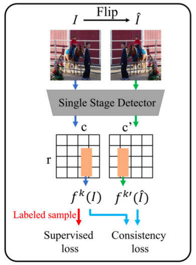
(d) Semi-supervised learning

Figure 1: Different types of object detection settings

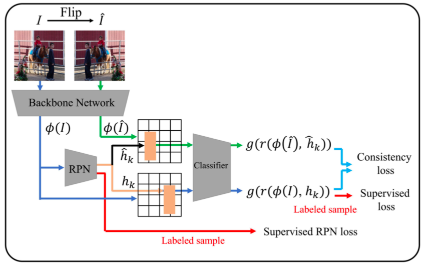
- Making a precise annotation in a large dataset is crucial to the performance of object detection.
- While the object detection task requires a huge number of annotated samples to guarantee its performance, placing bounding boxes for every object in each sample is time-consuming and costs a lot.
- We propose a novel consistency-based semi-supervised learning algorithm for object detection that can be applied not only to single-stage detectors but also to two-stage detectors.
- The proposed consistency constraints for object detection work well for both the classification of a bounding box and the regression of its location.
- We propose the Background Elimination (BE) method to mitigate the effect of background and show improvement of performance in most cases.

Related Work

- Semi-supervised learning
  - Self-training
    - (i) Training a model using labeled data
    - (ii) Predicting unlabeled data with the trained model (sampling and making a pseudo-label)
    - (iii) Retraining the model with labeled and sampled unlabeled data
    - (iv) Repeating the last two steps until meeting stopping criteria
  - Consistency regularization
    - (i) Applying perturbations to an input image x to obtain x’
    - (ii) Minimizing the difference between the outputs predictions f(x) and f(x’)
    - ※ It is known to help smooth the manifold (state-of-the-art performance in semi-supervised classification)



(a) single stage detector



(b) two stage detector

Notations

- $\hat{I}$ is a horizontally flipped version of I
- ϕ(I) is a feature map from backbone network
- $f^{p,r,c,d}(I)$ is corresponding to the pth pyramid, rth row, cth column, and dth default box
- c’ = C − c + 1, (p,r,c,d) = k, (p,r,c’,d) = k’
- h and g are RoI area and classifier, respectively

Consistency loss for classification

- The classification consistency loss used for a pair of bounding boxes in our method is given as below.

$$l_{con\_cls}(f^{k}_{cls}(I), f^{k'}_{cls}(\hat{I})) = JS(f^{k}_{cls}(I), f^{k'}_{cls}(\hat{I}))$$

- The overall consistency loss for classification is then obtained from the average of loss values from all bounding box pairs.

$$\mathcal{L}_{con-c} = \mathbb{E}_k[l_{con\_cls}(f^{k}_{cls}(I), f^{k'}_{cls}(\hat{I}))]$$

Consistency loss for localization

- Since the flipping transformation makes Δcˆx move in the opposite direction, a negation should be applied to correct it.

$$\Delta\, cx^k \Longleftrightarrow -\Delta\, \hat{cx}^{k'}$$
$$\Delta\, cy^k, \Delta\, w^k, \Delta\, h^k \Longleftrightarrow \Delta\, \hat{cy}^{k'}, \Delta\, \hat{w}^{k'}, \Delta\, \hat{h}^{k'}$$

- The localization consistency loss used for a single pair of bounding boxes in our method is given as below:

$$l_{con\_loc}(f^{k}_{loc}(I), f^{k'}_{loc}(\hat{I})) = \frac{1}{4}(\|\Delta cx^k - (-\Delta \hat{cx}^{k'})\|^2 + \|\Delta cy^k - \Delta \hat{cy}^{k'}\|^2 + \|\Delta w^k - \Delta \hat{w}^{k'}\|^2 + \|\Delta h^k - \Delta \hat{h}^{k'}\|^2)$$

- The overall consistency loss for localization is then obtained from the average of loss values from all bounding box pairs.

$$\mathcal{L}_{con-l} = \mathbb{E}_k[l_{con\_loc}(f^{k}_{loc}(I), f^{k'}_{loc}(\hat{I}))]$$

Application to two-stage detector

- The correspondence matching problem between the region proposals occurs.
- To simplify the problem, the flipped area $\hat{h}_k$ is derived by h_k

Background elimination

- The box which has a high probability of background class is excluded

$$m^k = \begin{cases} 1, & \text{if } \operatorname{argmax}(f^{k}_{cls}(I)) \neq background \\ 0, & \text{otherwise.} \end{cases}$$

- Applying the mask to L_{con-c} and L_{con-l}

$$\mathcal{L}_{con-c} = \mathbb{E}_{\mathbb{I}_{m^k=1}}[l_{con\_cls}(f^{k}_{cls}(I), f^{k'}_{cls}(\hat{I}))]$$
$$\mathcal{L}_{con-l} = \mathbb{E}_{\mathbb{I}_{m^k=1}}[l_{con\_loc}(f^{k}_{loc}(I), f^{k'}_{loc}(\hat{I}))]$$

Overall loss for object detection

- The total consistency loss is composed of the losses from L_{con-c} and L_{con-l}

$$\mathcal{L}_{con} = \mathcal{L}_{con-c} + \mathcal{L}_{con-l}$$

- The final loss L is composed of the original object detector and consistency loss

$$\mathcal{L} = \mathcal{L}_c + \mathcal{L}_l + w(t) \cdot \mathcal{L}_{con}$$

Experiments

Table 1: Detection results for PASCAL VOC2007 test set. The first two rows show the performance of each detector by supervised learning. * is the score from [17, 18]. The following experiments use VOC07 as the labeled data and VOC12 as the unlabeled data, and show the results of the proposed CSD with/without $\mathcal{L}_{con-c}$ (cls), $\mathcal{L}_{con-l}$ (loc) and EB. Blue / Red : supervised score (baseline) and Best results. The numbers in the parentheses are the performance enhancement over the baseline.

Labeled data	Unlabeled data	Consistency		Background Elimination	mAP (%)		
		cls	loc		SSD 300	SSD 512	R-FCN
VOC07	-	-	-	-	68.0*/70.2	71.6*/73.3	73.9
VOC0712	-	-	-	-	74.3*/77.2	76.8*/79.6	79.5*/79.4
VOC07	VOC12	✓	-	-	71.6 (1.4)	74.6 (1.3)	74.0 (0.1)
		-	✓	-	72.2 (2.0)	74.6 (1.3)	73.9 (0.0)
		✓	✓	-	72.0 (1.8)	74.8 (1.5)	74.0 (0.1)
VOC07	VOC12	✓	-	✓	71.7 (1.5)	75.4 (2.1)	74.5 (0.6)
		-	✓	✓	71.9 (1.7)	75.2 (1.9)	74.4 (0.5)
		✓	✓	✓	72.3 (2.1)	75.8 (2.5)	74.7 (0.8)

Table 2: Detection results on PASCAL VOC2007 test set. “COCO§”: All 80 classes. “COCO†”: 20 PASCAL VOC classes.

Labeled data	Unlabeled data	CSD Method (mAP)		
		SSD300	SSD512	R-FCN
VOC07	-	70.2	73.3	73.9
VOC07	VOC12	72.3	75.8	74.7
	VOC12+COCO§	71.7	75.1	74.9
	VOC12+COCO†	72.6	75.9	75.1

Table 3: Effects of using Background Elimination (BE) on SSD300 performance.

VOC07(L)+VOC12(U)	mAP
without BE	72.0
BE with m^k	72.3
BE with $m^k \otimes m^{k'}$	71.7

Discussion

Consistency regularization with only labeled data

- The consistency loss does not affect the improvement of performance for labeled data.

Single-stage detector vs. Two-stage detector

- While we can expect to improve performance in the classifier, it is hard to expect additional performance improvement of RPN.
- As a result, the two-stage detector has less performance improvement than the single-stage detector.

Background Elimination

- We apply BE to reduce the effect of the background and show that BE is helpful in improving the performance.
- However, getting rid of too many samples is not helpful in learning

Datasets

- The ratio of labeled/unlabeled class mismatch decides the amount of improvement.

Self-training vs. Consistency regularization

- As it is an iterative method which cycles training, prediction of unlabeled data and changing the training dataset, it is time-consuming and computationally intensive
- Meanwhile, CR method which trains unlabeled data with an additional loss helps the more common and robust learning.