Semi-supervised Object Detection with Unlabeled Data

Nhu-Van Nguyen, Christophe Rigaud and Jean-Christophe Burie
Laboratory L3i, SAIL joint Laboratory, Université de La Rochelle, 17042 La Rochelle CEDEX 1, France
{*nhu-van.nguyen, christophe.rigaud, jean-christophe.burie*}*@univ-lr.fr*

Keywords: Object Detection, Semi-supervised.

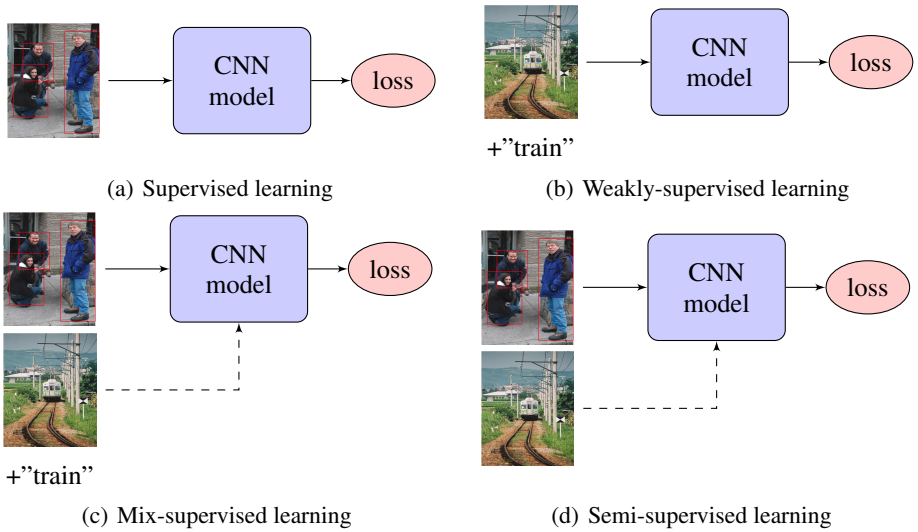
Abstract: Besides the fully supervised object detection, many approaches have tried other training settings such as weakly-supervised learning which uses only weak labels (image-level) or mix-supervised learning which uses few strong labels (instance-level) and many weak labels. In our work, we investigate the semi-supervised learning with few instance-level labeled images and many unlabeled images. Considering the training of unlabeled images as a latent variable model, we propose an Expectation-Maximization method for semi-supervised object detection with unlabeled images. We estimate the latent labels and optimize the model for both classification part and localization part of object detection. Implementing our method on the one-stage object detection model YOLO, we show that like the weakly labeled images, the unlabeled images also can boost the performance of the detector by empirical experimentation on the Pascal VOC dataset.

1 INTRODUCTION

Given an individual image, the object detection task aims at detecting all object instances in the image. Nowadays, the most successful detectors use the Convolutional Neural Network (CNN) models which can be classified into two different types of model: two-stage models (He et al., 2017; Ren et al., 2015) and one-stage models (Liu et al., 2016; Redmon and Farhadi, 2016; Lin et al., 2017). The main difference is that while two-stage models use at first a proposal network to generate candidate boxes and then classify these boxes, one-stage models use anchors boxes and directly predict bounding boxes without generating region proposals. While the CNN approach gives the best results on every standard benchmarking detection datasets such as Pascal VOC (Everingham et al., 2015) or COCO (Lin et al., 2014), it requires a large amount of instance-level labeled images with bounding box annotations. This is a problem in real-case applications as the bounding box labels are very difficult to obtain. To overcome this problem, many approaches have tried other training settings for CNN models, including weakly-supervised, mix-supervised and semi-supervised learning. We would like to insist that many works consider mix-supervised learning, which uses both image-level labels (weak labels) and instance-level labels (strong labels), as semi-supervised learning. In our work, we differentiate semi-supervised from mix-supervised learning; where semi-supervised learning involves instance-level labeled and unlabeled data. In Figure 1, we visualize these different settings.

In this paper, we develop an online method for training semi-supervised object detection by using EM-approach (Expectation-Maximization) for one-stage models which directly predict bounding boxes "without generating region proposals". Unlike most of the previous works, we deal with the labeled and unlabeled data together. Our proposed algorithm estimates latent instance-level labels for unlabeled data and based on this estimation, optimizes the CNN model parameters using mini-batch Stochastic Gradient Descent (SGD). We show that with fewer instance-level labeled data we can train competitive object detectors with the help of many unlabeled data. Moreover, we find out that our semi-supervised setting can almost match the performance of mix-supervised learning in (Yan et al., 2017) which requires image-level labels. There are two works (Yan et al., 2017; Papandreou et al., 2015) have used EM-approach for weakly-supervised or semi-supervised settings, but our work is distinct from them. The difference between (Yan et al., 2017; Papandreou et al., 2015) and our approach is discussed in Section 2.

In resume for our contributions, we present a novel semi-supervised EM algorithm for training one-stage object detection models which does not require image-level labels and focuses on both localization task and classification task in an object detection model. We show that with additional unlabeled data, our semi-supervised training gives better performance



(a) Supervised learning (b) Weakly-supervised learning

(c) Mix-supervised learning (d) Semi-supervised learning

Figure 1: Training settings for object detection using CNN models. (a) is supervised learning which uses instance-level labels (red bounding boxes in the image). (b) is weakly-supervised learning where only image-level labels ("train") are used. (c) is mix-supervised learning where both image-level labels ("train") and instance-level labels (red bounding boxes) are used. (d) is semi-supervised learning where instance-level labeled images and unlabeled images are used.

than the supervised training alone with labeled data, by experimenting on the widely used Pascal VOC dataset (Everingham et al., 2015). We show that our approach almost matches the performance of the mix-supervised training with instance-level labels and image-level labels (Yan et al., 2017).

2 RELATED WORK

Many methods for object detection have been proposed; they can be classified into two categories: one-stage models and two-stage models. All the best object detection models require a significant amount of instance-level labeled data which poses problems in real-case applications where instance-level labels are difficult to obtain. Since the unlabeled data or weakly labeled (image-level) data are larger and easier to acquire, many researches have proposed different settings which can profit weakly labeled data or unlabeled data.

Currently, existing semi-supervised methods usually use both the image-level labels and some of the instance-level labels (as mix-supervised in our definition). From now on, we use the term "mix-supervised" for this kind of setting to differentiate with the semi-supervised setting where instance-level labeled data are used together with unlabeled data. In (Häusser et al., 2017; Lee, 2013; Sajjadi et al., 2016), authors have proposed semi-supervised training methods for image classification task in which very few image-level labeled data and much larger unlabeled data are used. Hausser *et al.*(Häusser et al., 2017) propose to tune the CNN parameters by optimizing association circles which aims at finding the optimal embedding to represent the different classes. Lee (Lee, 2013) improves the classification model performance with additional labeled data which are generated from the model's predictions. In (Sajjadi et al., 2016) authors use regularization technique which profits the unlabeled data to enforces the classes mutual exclusivity. It also helps to push decision boundaries of neural networks to less dense areas of the decision space. However, they only focus on the classification task with image-level labels in which the loss function contains just a classification loss. It is not straightforward to use these methods for detection task which use the instance-level labels and the loss including a classification loss and a localization loss. Some works (Yan et al., 2017; Papandreou et al., 2015) have used EM-approach for mix-supervised training of object detection model or object segmentation model. Yan *et al.*(Yan et al., 2017) use the EM algorithm to train weekly label (image-level) and strong labels (bounding box) to train object detectors. The approach is similar to (Papandreou et al., 2015) but (Yan et al., 2017) run only three training iteration of the EM (E-Step and M-Step). Every M-step consists of 40k SGD iterations. Papandreou *et al.*(Papandreou et al., 2015) use the EM approach for training segmentation model with two settings: weakly-supervised and mix-supervised. In this method, the EM (E-Step and M-Step) is applied for every iteration. In (Rosenberg et al., 2005), authors have proposed a semi-supervised

training method for object detection. The authors train an initial detector using labeled data then retrain the detector by incremental adding the inferred label of unlabeled examples using the initial trained model. The unlabeled samples are selected based on a similarity measure to all instances of the same class in the labeled set. The incremental process ends when all unlabeled examples are added to the training set.

Most of the existing works propose methods for weakly-supervised training or mix-supervised learning while we deal with the semi-supervised training (as our definition, this setting uses unlabeled data instead of weakly-labeled data). Methods in (Papandreou et al., 2015; Yan et al., 2017) have used EM algorithm for object segmentation and object detection which are close to our approach. We also use an EM-like approach, but our method is distinct from them. The difference between our method and (Papandreou et al., 2015) is that (Papandreou et al., 2015) deals with object segmentation and it uses image-level labels and instance-level labels to leverage the training process, while we deal with object detection using instance-level labeled data and unlabeled data. (Yan et al., 2017) proposed an EM approach for weakly-supervised training and semi-supervised learning for object detection. However, (Yan et al., 2017) applied EM update for two-stage object detection models such as (He et al., 2017; Ren et al., 2015) which have the region proposal step. The EM-algorithm in (Yan et al., 2017) focus on maximizing the probability of each object proposal (classification step in object detection models), therefore it has not taken into account the localization step in these object detection models. Our method is applied for one-stage models, to enhance the optimization of both classification part and localization part. In addition, similar to (Papandreou et al., 2015), our method continuously makes the EM update while (Yan et al., 2017) used disjointed approach (only 3 EM updates for the training process).

3 OBJECT DETECTION WITH UNLABELED DATA

In this section, we present our EM-based semi-supervised approach for one-stage CNN object detectors. Our method provides end-to-end semi-supervised training which integrates EM in each iteration for classification optimization and in each epoch for localization optimization. In early stages, our model uses instance-level labeled data to train the detector and then in later stages, uses the combination of labeled and unlabeled data to improve the detector's performance. We first introduce some basic notations and then demonstrate the formulation of our method. Finally, we describe the details of the algorithm and the implementation of our approach with the YOLOv2 object detector proposed in (Redmon and Farhadi, 2016).

Notation: we denote by x the training images (pixel values), and z the instance-level labels (bounding box and class). In particular, $z_i = (b_i, c_i)$ is a pair of a bounding box $b_i = (x_i, y_i, w_i, h_i)$ and a category $c_i \in \{0, ..., C\}$ (C categories and the background). An image x can have the bounding box annotation z (labeled images) or not (unlabeled images). Let θ as the model parameters. We denote L the set of training images with labels (x, z) and U the unlabeled images set (x). Note that the instance-level labels may not be visible in the training set.

3.1 Semi-supervised Training

In the case of supervised learning for object detection, the general objective function is:

$$J(\theta) = \log(P(z|x;\theta) = \sum_{i=0}^{B} P(z_i|x;\theta) \ . \qquad (1)$$

where θ is the parameters vector of the model. As we train with one-stage detectors, we maximize $P(z|x;\theta)$ by maximizing the $P(z_i|x;\theta)$ of each anchor box in the anchor boxes set B of the image. In CNN object detection models, $J(\theta)$ is often optimized by mini-batch SGD.

For semi-supervised learning, we use both labeled images L and unlabeled images U. For the unlabeled images set U, we have image values x and the instance-level labels z as latent variables with the probabilistic graphical model:

$$P(x, z;\theta) = P(x) \prod_{z \in B} P(z|x;\theta) \ . \qquad (2)$$

In our method, we propose an EM-approach to learn model parameter θ from training data which includes both labeled and unlabeled images.

Expectation Step (E step): Calculate the expected value of the log-likelihood function, with respect to the conditional distribution of z given x under the current estimate of the parameters $\theta^{(t-1)}$. As explained in (Bishop, 2006), chapter 9, the expected complete-data log-likelihood given the previous parameter estimate $\theta^{(t-1)}$ is

$$\begin{aligned} Q^{(t)}(\theta;\theta^{(t-1)}) &= E_{z|x,\theta^{(t-1)}}[\log P(x, z;\theta)] \\ &= \sum P(z|x;\theta^{(t-1)}) \log P(z|x;\theta) \end{aligned} \qquad (3)$$

Kumar *et al.* (Kumar et al., 2010) have found that EM-based approaches work better when presented with the training data a meaningful order that facilitates learning and the order of the samples is determined by how easy they are. In our method, we have adopted this technique by selecting the easy instance-level labels estimated from unlabeled data (see 3.2).

In our method, we use a soft-EM approximation, estimating in this E-step the latent label variable z by

$$Q^{(t)}(\theta;\theta^{(t-1)}) = \sum P(z|x;\theta^{(t-1)}) \log P(z|x;\theta) \approx \sum_{\hat{z}\in\hat{Z}} P(\hat{z}|x;\theta^{(t-1)}) \log P(\hat{z}|x;\theta) \ . \quad (4)$$

where we create the set $\hat{Z}$ by remove z with $P(z|x;\theta^{(t-1)}) < \beta$ from all instance-level labels estimated.

$$\hat{Z} = \{z : P(z|x;\theta^{(t-1)}) \geq \beta\} \ . \quad (5)$$

In the CNN model, this step can be understood as follows: we apply the forward pass to unlabeled images and keep high confidence output detections. These detections are then used as the ground-truth of the unlabeled images together with the labeled images in the next step: Maximization.

Maximization Step (M step): we find the model parameters θ that maximize $Q^{(t)}(\theta;\theta^{(t-1)}) \approx \log P(\hat{z}|x;\theta)$ (terms that do not depend on θ are ignored). We optimize $Q^{(t)}(\theta;\theta^{(t-1)})$ by mini-batch SGD similarly to eq. (1), treating $\hat{z}$ as ground truth bounding boxes.

$$\theta^t = \arg\max_\theta Q^{(t)}(\theta;\theta^{(t-1)}) = \arg\max_\theta \log P(\hat{z}|x;\theta) \ . \quad (6)$$

One of the reasons why the other works for object detection and object segmentation have adopted EM-approach together with the weak labels is that the additional information from weak labels can help to optimize the classification part. In our method, we aim at applying the EM algorithm for both classification part and localization part. One difficulty we have had is that the expected complete-data log-likelihood with regard to the parameter estimate $\theta^{(t-1)}$ for the localization part (the bounding box b of the latent variable z) will not help the optimization of $Q^{(t)}(\theta;\theta^{(t-1)})$ using mini-batch SGD because the regression localization loss for unlabeled data will always be zero. To overcome this problem, we do not estimate the bounding box b of the latent variable z on each iteration of the mini-batch SGD but on every epoch of the optimization process, with regard to the parameter estimate $\theta^{(t0)}$ of the model at the first iteration of the epoch. Hence, on every iteration within each epoch, we can optimize $Q^{(t)}(\theta;\theta^{(t-1)})$ using the expected complete-data log-likelihood for the localization part computed at the begin of the epoch using $\theta^{(t0)}$. The detailed algorithm is presented in the next Section 3.2.

Our approach is similar to the work of Papandreou *et al.* (Papandreou et al., 2015). However, we treat the semi-supervised object detection task while they deal with the task of object segmentation using mix-supervised setting. In their work, they use EM to find the maximum likelihood of the parameters with regard to pixel classification, but in our case, we use EM to find the maximum likelihood with regard to object classification and bounding box regression. Another work by Yan *et al.* (Yan et al., 2017) also uses EM-approach. This work deals with the mix-supervised object detection by using weak labels together with strong labels. However, the authors propose a method which focuses on object classification and the method contains only three E-Steps and M-Steps.

In our approach, we focus on the semi-supervised object detection by training together labeled images and unlabeled images. Similar to (Papandreou et al., 2015), we use a continuous EM updates in the training process. Notably, we apply for both tasks: object classification and object localization. Our method can be used for any one-stage CNN detector. In this work, we have implemented the approach for the YOLOv2 model (Redmon and Farhadi, 2016) which is one of the fastest detectors with the state-of-the-art performance.

3.2 Incorporate Semi-supervised EM into YOLO

We use the object detection YOLOv2 model proposed in (Redmon and Farhadi, 2016). The YOLOv2 model considers object detection as a regression problem. Unlike other models such as (He et al., 2017; Ren et al., 2015), it unifies separated components of object detection into a single neural network by combining the loss for both classification part and localization part. The main idea is to divide the image into an even grid and simultaneously predicts bounding boxes from a set of predefined anchors, confidence in those boxes, and class probabilities. In model YOLOv2, each grid cell can have k anchor boxes with $k = 5$ or $k = 9$. However, to simplify the equations, we suppose $k = 1$. Generalizing to $k > 1$ anchor boxes is straightforward.

For each anchor box b, the YOLOv2 model predicts the coordinates $\{x, y, w, h\}$ of the box and its confidence $P(b|x;\theta)$. For each grid cell containing an anchor box b, the YOLOv2 model predicts C conditional class probabilities $P(c_i|x, b;\theta)$ for C categories.

YOLOv2 multiplies the conditional class probabilities and the individual box confidence prediction to

get class-specific confidence scores for each box b:

$$P(c_i|x;\theta) = P(c_i|x,b;\theta)P(b|x;\theta) \ . \quad (7)$$

where θ is the vector of CNN model parameters. The above probabilities are computed using softmax function as

$$P(c_i|x,b;\theta) = \sigma(f_j(c_i|x,b;\theta)) \ . \quad (8)$$

$$P(b|x;\theta) = \sigma(f_j(b|x;\theta)) \ . \quad (9)$$

where $f_j(c_i|x,b;\theta)$ and $f_j(b|x;\theta)$ are the output of CNN model for the anchor box b at the grid cell j.

Our optimization process which relies on the bounding boxes estimation at the first iteration of each epoch and the class probabilities estimation on every iteration is summarized in the Algorithm **??**.

The EM-based algorithms have been widely used in many latent variable models, and its correctness has already been proved (Little and Rubin, 1986). Besides that, in another perspective, we believe that the EM algorithm can help the problem of imbalanced data of the CNN object detectors. A typical data imbalance problem appears in all CNN object detection models where there are more likely background boxes than foreground boxes (boxes with object). Since the background boxes are easier to classify, too many of them will dominate the optimization process (Lin et al., 2017). Similar to techniques like Hard Negative Mining in (Liu et al., 2016) and (Ren et al., 2015) which try to down-sample easy background samples to balance the training data, EM algorithms can also balance the training data by estimating the missing labels of unlabeled data and over-sampling the foreground samples. By using the E-Step, our algorithm provides many more foreground samples, from easy one at the early stage to more difficult one at the later stage. This over-sampling of foreground examples can balance the data between background class and foreground class which should help the optimization process. In Algorithm 1, we set the parameter $\beta = 0.6$ to take into account only 'easy' foreground samples (ignore the 'difficult' foreground samples, and the background samples of the unlabeled data).

Another problem to handle is the confidences of instance-level labels z and the estimated $\hat{z}$ are different, especially at early stages. Hence, in Algorithm 1, we introduce a weighting factor $\alpha \in [0,1]$ to balance them. The parameter α is defined as a function of time $f(t)$ where t presents the number of epochs and $f(t)$ is computed by Equation 10. We limit α in $[0,1]$ by setting $\alpha = \min(f^+(t),1)$. In our experiment, $T_1 = 20$ and $T_2 = 120$ works best.

$$f(t) = \frac{t - T_1}{T_2 - T_1} \ . \quad (10)$$

Input : Initial parameters vector θ^{T_1} of the CNN model which is trained with only labeled data by T_1 epochs, $c \in \{0,...,C\}$, labeled training images set $L = (x_l, z_l)$, unlabeled training images set $U = (x_u)$, parameters α and β.

foreach *epoch* **do**
  foreach $x \in U$ **do**
    $\hat{z}^x_{epoch} = EStep(x, \theta^{(t_0)})$
  end
  $\hat{z}_{epoch} = \{\hat{z}^x_{epoch}\}$ *where* $x \in U$
  foreach *mini-batch* (U_m, L_m) **do**
    $\hat{z}_u$ = E-Step ($U_m, \theta^{(t-1)}$)
    replace $\{b\} \in \hat{z}_u$ by $\{b\} \in \hat{z}_{epoch}$
    M-Step ((x_l, z_l), $(x_u, \hat{z}_u), \theta^{(t)}, \theta^{(t-1)}$)
  end
end
Procedure `E-Step` (x, θ')
  foreach $b \in B$ **do**
    foreach $c_i \in C$ **do**
      $\hat{f}(c_i) = f(c_i|x,b;\theta')$
    end
    $\hat{c} = \arg\max_c \hat{f}(c)$
    $z_b = (b, \hat{c})$
    $\hat{f}(z_b) = f(b|x;\theta')\,\hat{f}(\hat{c})$
  end
  $\hat{z} = \{z_b\} = \{(b,\hat{c})\}$, *where* $\sigma(\hat{f}(z_b)) \geq \beta$
  return $\hat{z}$;
Procedure `M-Step` $((x_l, z_l), (x_u, \hat{z_u}), \theta, \theta')$
  $Q^{(t)}(\theta;\theta') =$
  $\sum_{z \in z_l} \log P(z|x;\theta) + \alpha \sum_{z \in \hat{z}_u} \log P(z|x;\theta)$
  Optimize YOLOv2 model using SGD to update θ
  return;

Algorithm 1: Semi-supervised EM algorithm for object detection on YOLOv2 model.

4 EXPERIMENTAL EVALUATION

4.1 Experimental Protocol

4.1.1 Dataset

We evaluate our method on the widely used Pascal VOC dataset, a well-known benchmark for object detection in computer vision (Everingham et al., 2010; Everingham et al., 2015), consisting of 20 foreground object classes and one background class. We use the Pascal VOC dataset to investigate the influence of the

proportion of labeled data in the training set. We train the system on the union set of VOC 2012 trainval and VOC 2007 trainval $(07+12)$. We divide the training set randomly following different settings: the quantity of labeled images contains 10 to 50% of the training set, the images left in the training set are considered as the unlabeled images. The VOC 2007 test set is used for evaluation.

We compare our semi-supervised model with the baseline supervised object detection on the COCO dataset (Lin et al., 2014). The COCO dataset has 120K unlabeled images (*unlabeled2017*) which is useful for semi-supervised learning on COCO. We train our semi-supervised model with the union of 80k train images and a 35k sub-set of val images (*trainval35k*) as the labeled set and with 120K *unlabeled2017* as the unlabeled set. We report results on the remaining 5k subset of val images (*minival*) by comparing with the baseline supervised model trained on *trainval35k*.

To evaluate detection performance, we report the mean average precision (mAP%) (Everingham et al., 2015).

4.1.2 Training

We have experimented with the Darknet architecture with parameters initialized from the ImageNet pretrained model used in (Redmon and Farhadi, 2016). We train the YOLOv2 model with the scale of (416x416). For SGD, we use a mini-batch of 64 images and initial learning rate of 0.0001, multiplying the learning rate by 0.1 at 80 and 120 epochs. We use momentum of 0.9 and a weight decay of 0.0005. We present the ratio of unlabeled images vs. labeled images in the mini-batch as parameter γ. We set $\gamma = 0$ (no unlabeled images) for the first 12 epochs; from the 13th epoch we set $\gamma = 0.5$ (32 unlabeled + 32 labeled images). We used the same setting for both Pascal VOC and COCO datasets.

4.2 Results

4.2.1 Pascal VOC Dataset

We investigate the influence of the proportion of labeled data in the training set. As we state that the semi-supervised setting requires fewer labeled images and many unlabeled images, we have experimented with five different proportions of labeled images: 10, 20, 30, 40 and 50%. Results are summarized in Figure 2. We can see that the semi-supervised learnings with additional unlabeled images give better results compared to the supervised learning with labeled images alone in every proportion.

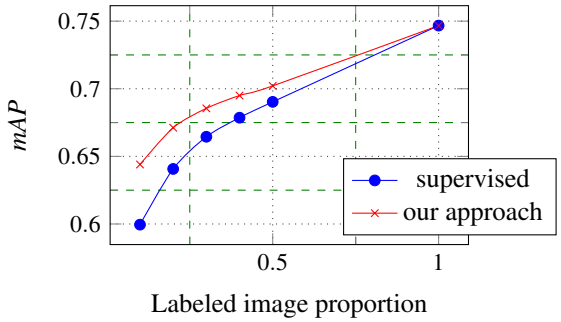


Figure 2: Results of semi-supervised detection on the Pascal VOC test set for different settings of the proportion of labeled data.

Table 1: Comparison between our method with the mix-supervised method in (Yan et al., 2017). Results represent the improvement of the two methods compared to their fully supervised versions.

	Improvement of mAP
(Yan et al., 2017)	1.80%
Semi-supervised (our)	1.64%

Existing works often focus on the mix-supervised object detection by using instance-level labels and image-level labels (they call it semi-supervised learning, but it is different from our setting where unlabeled images are used instead of image-level labeled images). Therefore, we compare our method with a recent mix-supervised method (Yan et al., 2017) in a fair condition on the Pascal VOC dataset (Table 1). In (Yan et al., 2017), authors report the good result of the detector training with 40% instance-level labels and 60% weak labels compared to the fully supervised training with 100% instance-level labels. In our experiment, using the same proportion of 40% labeled images and 60% unlabeled images, we achieve an improvement of 1.64% mAP compared to the supervised detector using only 40% labeled images. This improvement almost matches the 1.8% mAP improvement of the mix-supervised approach reported in (Yan et al., 2017).

In Table 2, we show the results for different versions of our semi-supervised learning compared to supervised learning which confirm the advantages of the proposed method. With 10% of labeled training images out of the total training images, we obtain detection performance improvement on the Pascal VOC test set from 59.95% to 64.40% mAP. In our Algorithm 1, if we remove the bounding boxes estimation and the localization optimization for unlabeled data, we have a lower detection performance of 62.19% mAP with the same setting.

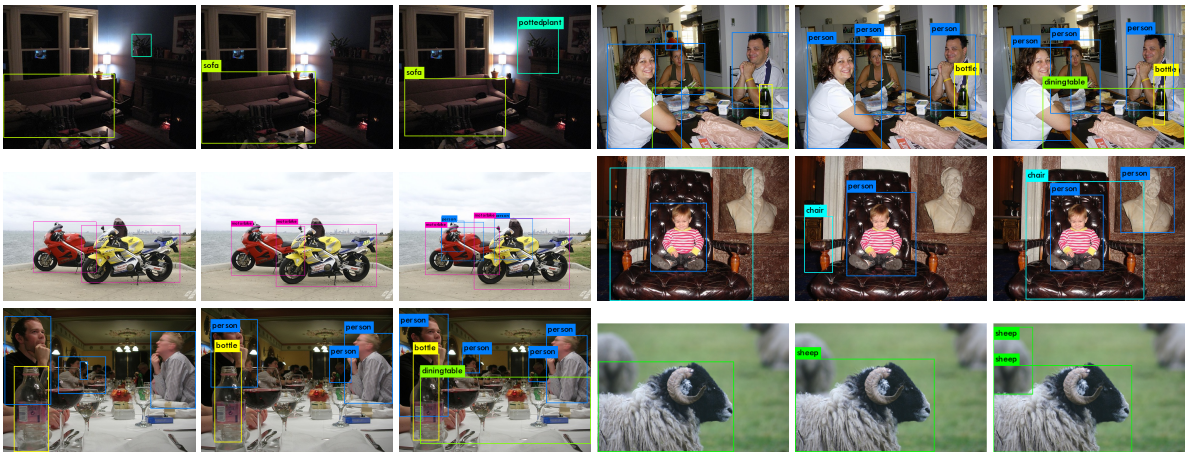


Figure 3: First image is the ground-truth; second image is the detection of the supervised model; third image is the detection of the semi-supervised model. The predictions may be right (first row), or wrong (wrong class, the next row), or maybe right but the objects are not included in the ground-truth annotations (last row).

Table 2: Detection results for different versions of our semi-supervised models (v1, v2) and supervised learning on the Pascal VOC dataset.

	Supervised	v1	v2
Cls optimization		√	√
Loc optimization			√
VOC2007 mAP%	59.95	62.19	64.40

Table 3: Comparison between semi-supervised learning (*trainval35k+unlabeled2017*) with supervised learning (*trainval35k*) on the COCO dataset. Results on *minival* test set.

	Supervised	Semi-supervised
mAP[0.50:0.95]	22.4%	23.1%

4.2.2 COCO Dataset

In the Table 3, we compare our semi-supervised model with the baseline supervised model trained on 115K labeled images (*trainval35k*). Our semi-supervised model is trained on the same labeled images with the additional 120K unlabeled images (*unlabeled2017*). Our semi-supervised model has 23.1% mAP on the *minival* test set (5K images), higher than the supervised model which has 22.4% mAP, but with only an improvement of 0.7%. The reason for this small improvement due to the large size of the labeled images in the training set. In semi-supervised learning, we often have fewer labeled images but require many unlabeled images to have an important impact of the unlabeled images.

4.3 Qualitative Detection Result

In Figure 3 we show examples of the results obtained by the semi-supervised model and the supervised model on the Pascal VOC 2007 test set. The supervised model is trained on 10% labeled images of the training set. The semi-supervised model is trained on 10% labeled images and 90% unlabeled images of the training set. We have found that while the supervised model merely recognizes what it is taught, the semi-supervised model can guess additional objects. However, it can guess them right (first row), or wrong (wrong class, the next two rows), or maybe right but the objects are not included in the ground-truth annotations (last two rows). The problem in the last two rows can lead to a lower mAP, that means our semi-supervised model may have better performance than presented in Table 2.

5 CONCLUSIONS

We have investigated a semi-supervised object detection approach with instance-level labeled images and unlabeled images. We treat the object detection model as a latent variable model in which the instance-level labels are missing values. Our experiments on the Pascal VOC dataset have shown that: (1) Even the unlabeled images can improve the detection performance as the weakly labeled images. (2) Using the unlabeled images to optimize both the classification task and localization task is better than optimizing the classification task alone. (3) In the real-case scenario, we can save effort by annotating fewer images and then use the raw images to improve the detection performance.

ACKNOWLEDGEMENTS

This work is supported by the Research National Agency (ANR) in the framework of the 2017 LabCom program (ANR 17-LCV2-0006-01).

REFERENCES

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Everingham, M., Eslami, S. M., Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vision*, 111(1):98–136.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338.

Häusser, P., Mordvintsev, A., and Cremers, D. (2017). Learning by association - A versatile semi-supervised training method for neural networks. In *CVPR 2017, Honolulu, HI, USA, 2017*, pages 626–635.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. B. (2017). Mask R-CNN. *CoRR*, abs/1703.06870.

Kumar, M. P., Packer, B., and Koller, D. (2010). Self-paced learning for latent variable models. In *Neural Information Processing Systems - Volume 01*, NIPS'10, pages 1189–1197, USA. Curran Associates Inc.

Lee, D.-H. (2013). Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks.

Lin, T.-Y., Goyal, P., Girshick, R. B., He, K., and Dollár, P. (2017). Focal loss for dense object detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollr, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, Zrich. Oral.

Little, R. J. A. and Rubin, D. B. (1986). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., New York, NY, USA.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. *arXiv preprint arXiv:1512.02325*.

Papandreou, G., Chen, L.-C., Murphy, K. P., and Yuille, A. L. (2015). Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *IEEE International Conference on Computer Vision*, ICCV '15, pages 1742–1750, Washington, DC, USA. IEEE Computer Society.

Redmon, J. and Farhadi, A. (2016). YOLO9000: Better, Faster, Stronger.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc.

Rosenberg, C., Hebert, M., and Schneiderman, H. (2005). Semi-supervised self-training of object detection models. In *WACV/MOTION'05 - Volume 01*, pages 29–36, Washington, DC, USA. IEEE Computer Society.

Sajjadi, M., Javanmardi, M., and Tasdizen, T. (2016). Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1163–1171.

Yan, Z., Liang, J., Pan, W., Li, J., and Zhang, C. (2017). Weakly- and semi-supervised object detection with expectation-maximization algorithm. *CoRR*, abs/1702.08740.