Semi-supervised Object Detection with Unbiased-Teacher v2

Patrick Collins Xiangkun Fang Yiqing Lu Shuyu Liu

Abstract

In this project, we used the Unbiased-Teacher v2 model to conduct semi-supervised object detection with 5.5% labeled image data. The model has a fully-supervised burn-in stage and a semi-supervised teacher-student mutual learning stage. We chose Fully Convolutional One-Stage Detector (FCOS) for the burn-in stage and trained from scratch. With 97695892 parameters, the model reached 0.487 and 0.413 mean Average Precision (mAP) on the test and validation set respectively.

1. Introduction (Literature Review)

Object Detection. Object Detection is a task to find objects in images and return object classes together with bounding boxes. In fully supervised object detection, all training data is labeled; while in semi-supervised object detection (SS-OD), only a small portion of the data is labeled. Unlabeled data, which is abundant and almost free to have access to, takes the majority and is expected to be exploited to improve model performance.

Semi-supervised Methods. Early SS-OD methods such as CSD (Jeong et al., 2019) apply consistency regularization to enforce the model to generate consistent predictions with respect to data argumentation. Teacher-student framework, which can be viewed as hard consistency regularization, becomes popular in the state-of-the-art methods. Such models first train a teacher with labeled data in a fully supervised manner and then generate pseudo-labels for the unlabeled data with the teacher model. Introducing data argumentation to the unlabeled data in the student training improves model robustness and generalization. STAC (Sohn et al., 2020) is one of the fundamental work in this streamline, where the pseudo-labels are only generated once. Instant-Teaching (Feng Zhou et al., 2021), an effective end-to-end method, performs model training and pseudo-labeling at the same time. Unbiased teacher (Liu et al., 2021; 2022) deals with the problem of biased class distribution by introducing focal loss and a teacher-student mutual learning algorithm. Consistent-Teacher (Wang et al., 2022) focuses on the problem of oscillating pseudo bounding boxes and proposed consistent Adaptive Sample Assignment, 3-D Feature Alignment Module as well as Gaussian Mixture Model to promote bounding boxes stability. Other methods take very different approaches, for example, DETReg (Bar et al., 2022) uses the unlabeled dataset for pretrain.

Supervised Methods. In the teacher-student framework, a supervised model needs to be chosen to initialize the teacher. Depending on whether there is a separate regional proposal stage, fully supervised model can be classified as one-stage or two-stage. One of the most prominent supervised methods with regional proposal is R-CNN, which bases on convolutional neural network. The first version of R-CNN (Girshick et al., 2014) used selective search to generate category-independent region proposals. It has high accuracy, yet suffers from inefficiency. The single-stage method Fast R-CNN (Girshick, 2015) uses shared convolutional features to greatly speed up the model. Faster R-CNN (Ren et al., 2015) introduces a Region Proposal Network (RPN) that shares the full-image convolutional features with the detection network and further accelerates the running time. Apart from the R-CNN streamline, YOLO (Redmon et al., 2016), FCOS (Tian et al., 2019), DETR (Carion et al., 2020) are some other popular methods for supervised object detection. Most state-of-the-art methods (e.g. Faster R-CNN, SSD, YOLOv5) are anchor-based and require regional proposal. FCOS is one of the few anchor-free methods that avoids numerous hyper-parameters related to anchor boxes and thus is simple and fast.

Train from Scratch. Most state-of-the-art models use pre-trained weights from ImageNet, this issue was relooked (He et al., 2019). The anthors showed that random initialization can achieve competitive results under carefully tuning.

2. Methods (Unbiased-Teacher v2)

Given 20k labeled and 512k unlabeled images, we have a 5.5% SS-OD task. Checking the performance of state-of-the-art models on the COCO dataset, we find Unbiased-Teacher v2 (Liu et al., 2022) perform well on task with small portion of labeled data. The published code implements faster R-CNN and FCOS for the supervised stage. Considering the limited training time, we choose the efficient anchor-free detector FCOS (Tian et al., 2019). The unbiased teacher v2 generalizes v1 by a novel mechanism Listen2Student, which includes Student model feedback in training and improves robustness in bounding boxes locations significantly.

2.1. Pseudo-Labeling Method: Unbiased Teacher

The pseudo-labeling method contains two stages: 1) the burn-in stage; 2) the mutual learning stage. In the burn-in stage, an initial object detector is trained on the labeled images $\boldsymbol{D}_s = \{\boldsymbol{x}_i^s, \boldsymbol{y}_i^s\}_{i=1}^{N_s}$, with the standard supervised losses $\mathcal{L}_{sup} = \sum_{i=1}^{N_s} \mathcal{L}(\boldsymbol{x}_i^s, \boldsymbol{y}_i^s)$. At the beginning of the mutual learning stage, the pretrained object detector is duplicated to a Student model and a Teacher Model. Iteratively in the mutual learning stage, unlabeled images are augmented weakly and are passed through the Teacher model to obtain pseudo labels and bounding boxes if the box scores go above a confidence threshold τ. Then those pseudo labeled images $\hat{\boldsymbol{D}}_u = \{\boldsymbol{x}_i^u, \hat{\boldsymbol{y}}_i^u\}_{i=1}^{N_u}$ are augmented strongly to train the Student model with an combination of the supervised loss $\mathcal{L}_{sup}$ and the unsupervised loss $\mathcal{L}_{unsup} = \sum_{i=1}^{N_u} \mathcal{L}(\boldsymbol{x}_i^u, \hat{\boldsymbol{y}}_i^u)$. The Student Model weights are updated with the gradient of $\mathcal{L}_{sup} + \boldsymbol{\lambda}_u \mathcal{L}_{unsup}$ and the Teacher model weights are updated with the Student model weights via Exponential Moving Average (EMA) (Liu et al., 2021).

2.2. FCOS and Pseudo Labeling

Comparing with the anchor-based detectors (like Faster-RCNN) which adjust those predefined anchor boxes during training, the anchor-free detectors, looking for centrality and relative distance to the object boundaries of each pixel, achieve competitive accuracy, computational efficiency and great adaptability to new datasets when conducting the task of fully supervised object detection (Tian et al., 2019).When applied to the SS-OD tasks through pseudo-labelling methods, the anchor-based detectors benefit from the unlabeled data and have increased in accuracy. However, the anchor-free detectors, such as FCOS, have shown less improvement (Liu et al., 2022). The possible reasons, suggested in the paper (Liu et al., 2022), are that 1) Unreliable centerness scores, due to lack of supervision for background instances, will predict false-positive pseudo boxes. 2) Unrobustness of the label assignment technique to the localization noise, generated by the pseudo boxes, may results in opposite labelling of foreground or background.

The unbiased teacher v2 succeeds in adapting FCOS to the unbiased teacher framework by alleviating mislead from pseudo labels. When indicating objectness, this pseudo-labelling method only adopts classification scores and ignores the unreliable centerness scores. Also, instead of using those soft labeling methods that improve FCOS performance in the fully supervised scenario, the standard label assignment method is applied, which assigns all pixels inside the bounding box with foreground, to enhance robustness.

2.3. Listen2Student: Improve Pseudo-Bounding Box

To provide more accurate bounding boundaries and further remove misleading instances, a mechanism named *Listen2Student* is developed (Liu et al., 2022), which assigns uncertainty score to each boundary of the bounding box, instead of defining confidence for the entire bounding box. The score describes the extent of mislead, and will facilitate decision on bounding boxes location. Briefly, under this mechanism, the student network will trust and learn a pseudo boundary only when the teacher has a significant lower uncertainty score at that boundary.

We would like the boundary location learnt by the Student model to be as close as to the ground truth. However, unlabeled images cannot provide ground truth. Therefore, the localization uncertainty for each boundary was proposed to predict the error to the ground truth. This parameter is derived by adding another branch with the same output size as the boundary distance branch and is joint trained with it, using the negative power log-likelihood loss(NPLL) as the regression loss(Lee et al., 2020):

$$\mathcal{L}_{reg}^{sup} = \sum_{i}^{N_s} \eta_i (\sum (\frac{(d_s^i - d_g^i)^2}{2(\delta_s^i)^2} + \frac{1}{2}\log(\delta_s^i)^2) + 2\log 2\pi),$$

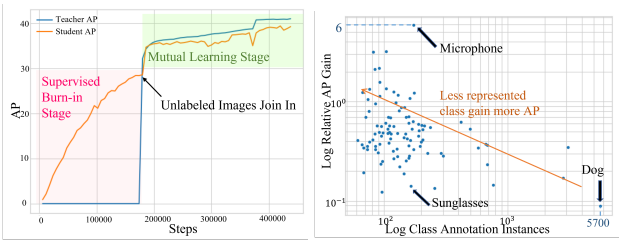
where η_i is the IoU score between the predicted box and ground-truth box, δ_s^i is the localization uncertainty of the Student model on the specific boundary, and $d_s^i,\ d_g^i$ are the distances to the object center predicted by the student model and the ground truth. When the prediction is bad, i.e. $(d_s^i - d_g^i)^2$ is large, uncertainty δ_s^i is large.
Under the estimation of uncertainty, to reduce misleading boundary given by the Teacher, a selection mechanism, which takes both of the uncertainty of the Teacher Model and the Student Model into consideration, determines whether the Student should learn from the Teacher in the boundary regression:

$$\mathcal{L}_{reg}^{unsup} = \sum_{i}^{N_u} ||\tilde{d}_t^i - \tilde{d}_s^i|| \mathbb{1}\{\tilde{\delta}_t^i + \sigma \leq \tilde{\delta}_s^i\},$$

where $\tilde{d}_t^i,\ \tilde{d}_s^i$ are the Teacher and the Student predictions of boundary distance for the unlabeled images, $\tilde{\delta}_t^i,\ \tilde{\delta}_s^i$ are the corresponding uncertainties, and σ is a threshold for the uncertainty difference that guarantees the boundary will be learnt only when the teacher is confident enough.

3. Results

We observed continuing gains in validation mAP with purely-supervised training up until roughly 170,000 steps ($\approx$ 280 GPU-hours on 4 NVIDIA T4 GPUs). After switching to the semisupervised training strategy at 175,000 steps, we continued training for a total of 475,000, representing a total of roughly 790 GPU-hours, or 1 GPU-month of training. Our data indicates that semisupervised training leads to better detection of classes that are under-represented in the training data, and better bounding box predictions for well-represented classes.



(a) AP vs Steps (b) Relative AP gain vs class annotation count

Figure 1. **AP improvement during training and for each classes.** We saw a sharp increase of AP at the begining of the mutual learning stage after unlabeled data joining in (1(a)). The under-represented classes(less labeled) got more increase in AP compared to the well-represented classes(1(b)).

As shown in figure 1(a), the model gained roughly 12.6 points of AP during the semisupervised phase, increasing from 28.4 at step 174,000 to 41 at step 474,000. We note that the architecture is very resilient against over fitting, with AP continuing to climb even after 474,000 steps.

Next, we define AP_{sup} to be the greatest AP achieved for each class during the supervised phase of training, and AP_{unsup} to be the greatest AP achieved for each class during the semisupervised phase of training, and call $\frac{\mathrm{AP}_{unsup}}{\mathrm{AP}_{sup}} - 1$ the "relative AP gain" for each class. Looking at the relative AP gain for each class vs the number of annotations for that class in the validation dataset in figure 1(b), we see a clear trend, with well-represented categories benefiting relatively little from semisupervised training, and under-represented categories benefiting greatly from the semisupervised phase, with one class, "microphone," seeing a 5.87x improvement. There are some classes got no improvement like "sunglasses", which will be discussed in section 4.2.

4. Discussion

We will show in images how the model improves prediction on the well-represented and under-represented class, give an example on failure at small objects detection and discuss some implementation details at the end.

4.1. Visualization on classes

We compare the validation set ground truth with outputs of a fully-supervised model (inference performed with a checkpoint at the $110,000^{\text{th}}$ step) and the final semisupervised model (inference performed with the final checkpoint at the $435,000^{\text{th}}$ step).

We first examine the model output on the most common class, "dog", which accounts for roughly 20% of the total annotations. In the left column of figure 2, we see that semisupervised training greatly increases the model confidence in correct predictions, and reduces the number of spurious bounding boxes. We also note that the validation set often appears to be missing bounding boxes (e.g. the right-hand dog in this example), which we believe, explains the model's 7.4% gain in AP between the validation and test sets.

Figure 2. Ground truth (top left) vs supervised checkpoint (middle left) vs unsupervised checkpoint (bottom left) inference outputs for "dog"; Ground truth (top right) vs supervised checkpoint (middle right) vs unsupervised checkpoint (bottom right) inference outputs for "microphone"

Next, we examine the model output on the "microphone" class, which accounts for only 0.6% of annotations. As discussed in section 3, this class saw the largest improvement during the semi-supervised phase.

Here, we see a similar pattern as in the "dog" example, where the confidence of correct predictions is greatly increased, and many spurious bounding boxes are removed. We also note that a spurious "monkey" prediction has been replaced with a spurious "drum" prediction: based on similar examples we observe in inference outputs, we note that during the unsupervised phase, the model seems to have developed a strong prior regarding the types of objects that tend to go together – so here we see that the presence of a strong microphone prediction seems to have biased it towards (correctly) interpreting this scene as part of a musical event, likely to contain "drum" objects, rather than a scene likely to contain "monkey" objects.

4.2. Performance Relies on Sizes

From the statistics of the final model, we observe that the model has extremely low average precision on the small objects (7.16%), compared to the larger ones (47.56%). A possible reason could be that small objects are more vul-

nerable to the noise in the pseudo boxes provided by the teacher. Given the same amount of noise in the pseudo boxes (quantified by pixels of shifts), the small objects lose more information compared to the larger ones. Moreover, in the data images, small objects lose more details than the larger ones under the same resolution.

An example is the under-represented "sunglasses" class shown in figure 3(a). Even though the model has recognized more "people" than the ground truth, most of the extremely small sunglasses objects (only around 10 pixels on the shortest edge) in the ground truth image are missed in the model prediction. And we can see that among the 3 sunglasses the largest one is detected.

(a) Ground Truth (b) Semi-supervised

Figure 3. **Performance on small objects.** There are 3 sunglasses in the ground truth, but only one is detected in our model.

4.3. More discussion on implementation details: energy function collapse with default settings

Using the default value of $\boldsymbol{\lambda}_u$ from (Liu et al., 2022), we found that AP fell of dramatically after the beginning of the semisupervised phase, typically reaching ≈ 0 within a few thousand steps. Looking into loss curves, we discovered that $\mathcal{L}_{unsup} \gg \mathcal{L}_{sup}$ – indeed, the loss `teacher_better_student` is as much as 1000x larger than any other losses calculated by the model, as depicted in figure 4.

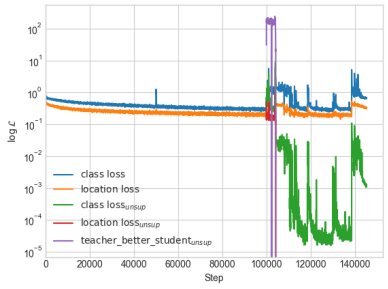


Figure 4. Loss curves indicating root-cause for collapse

This loss, which is not described in (Liu et al., 2022), counts the number of bounding boxes in a batch where the teacher has higher confidence than the student. Since this loss is not normalized according to the total number of bounding boxes predicted, it almost-necessarily overpowers all other loss functions (which, as shown in figure 4, generally lie in the range $(0, 1)$), punishing the model for high-confidence predictions and pushing it into a collapsed state that assigns near-zero confidence to all bounding boxes.

We believe that this loss is most likely a bug introduced while open sourcing the unbiased teacher V2 code base – likely it was originally intended to be exported as a metric and accidentally transformed into a loss while scrubbing the model for internal code. As a result, we believe that this paper describes the only successful training of the unbiased teacher v2 architecture outside of Meta AI.

We solved this issue by changing the model to remove the `teacher_better_student` loss, but continued to observe that the model would enter a collapsed state soon after the beginning of the unsupervised phase, with a similar loss curves indicating that $\frac{\mathcal{L}_{unsup}}{\mathcal{L}_{sup}} \approx 10$. After reducing $\boldsymbol{\lambda_u}$ from the default, 5, to .01, we were able to reproduce the results from (Liu et al., 2022), as we described in 3.

References

Bar, A., Wang, X., Kantorov, V., Reed, C. J., Herzig, R., Chechik, G., Rohrbach, A., Darrell, T., and Globerson, A. Detreg: Unsupervised pretraining with region priors for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14605–14615, 2022.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Feng Zhou, Q., Yu, C., Wang, Z., Qian, Q., and Li, H. Instant-teaching: An end-to-end semi-supervised object detection framework. 2021 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4079–4088, 2021.

Girshick, R. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

He, K., Girshick, R., and Dollár, P. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4918–4927, 2019.

Jeong, J., Lee, S., Kim, J., and Kwak, N. Consistency-based semi-supervised learning for object detection. *Advances in neural information processing systems*, 32, 2019.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lee, Y., Hwang, J., Kim, H., Yun, K., and Park, J. Localization uncertainty estimation for anchor-free object detection. *CoRR*, abs/2006.15607, 2020. URL `https://arxiv.org/abs/2006.15607`.

Liu, Y.-C., Ma, C.-Y., He, Z., Kuo, C.-W., Chen, K., Zhang, P., Wu, B., Kira, Z., and Vajda, P. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*, 2021.

Liu, Y.-C., Ma, C.-Y., and Kira, Z. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9819–9828, 2022.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Sohn, K., Zhang, Z., Li, C.-L., Zhang, H., Lee, C.-Y., and Pfister, T. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020.

Tian, Z., Shen, C., Chen, H., and He, T. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9627–9636, 2019.

Wang, X., Yang, X., Zhang, S., Li, Y., Feng, L., Fang, S., Lyu, C., Chen, K., and Zhang, W. Consistent targets provide better supervision in semi-supervised object detection. *arXiv preprint arXiv:2209.01589*, 2022.