An Adaptive Supervision Framework for Active Learning in Object Detection

Sai Vikas Desai[†1]
cs17mtech11011@iith.ac.in
Akshay Chandra Lagandula[†1]
akshaychandra@iith.ac.in
Wei Guo[2]
guowei@isas.a.u-tokyo.ac.jp
Seishi Ninomiya[2]
snino@isas.a.u-tokyo.ac.jp
Vineeth N Balasubramanian[1]
vineethnb@iith.ac.in

[1] Indian Institute of Technology Hyderabad, Kandi, 502285, India
[2] Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo 1880002, Japan

Abstract

Active learning approaches in computer vision generally involve querying strong labels for data. However, previous works have shown that weak supervision can be effective in training models for vision tasks while greatly reducing annotation costs. Using this knowledge, we propose an adaptive supervision framework for active learning and demonstrate its effectiveness on the task of object detection. Instead of directly querying bounding box annotations (strong labels) for the most informative samples, we first query weak labels and optimize the model. Using a switching condition, the required supervision level can be increased. Our framework requires little to no change in model architecture. Our extensive experiments show that the proposed framework can be used to train good generalizable models with much lesser annotation costs than the state of the art active learning approaches for object detection.

1 Introduction

State-of-the-art performance of deep neural networks in computer vision tasks such as object detection and semantic segmentation has been largely achieved using fully supervised learning methods [8, 10], which demand large amounts of strongly annotated data. However, it is known that obtaining labels for vast amounts of data is expensive and time-consuming. In this work, we focus on the problem of training efficient object detectors while minimizing the required annotation effort.

Active learning has been shown to be efficient in reducing labeled data requirement for image classification [9, 14, 16, 23, 28]. However, fewer efforts have been proposed to attempt active learning for object detection using deep neural networks [5, 13, 22]. In these approaches, an oracle is asked to provide accurate bounding box labels for the most informative set of images, which are selected by the corresponding methodology. These methods mostly vary by the nature of the methodology used to choose the query images, or in case of

†Both authors have equally contributed.

object detection, by the nature of the underlying object detection framework. In this work, we propose a highly effective approach to leverage weak supervision for active learning in object detection.

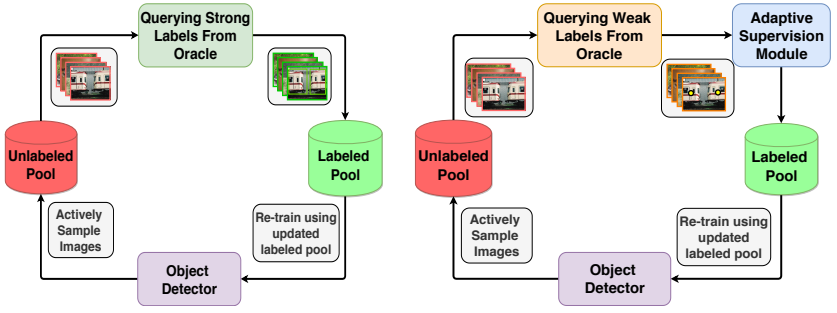
Learning with weak supervision has grown significantly in importance over the last few years [1, 2, 3, 12, 15, 18, 19, 20, 25, 31]. Achieving desired generalization performance with a lower labeling budget has been achieved using image-level labels [1, 2, 3, 12, 25, 31], object center clicks [18] and answering yes/no questions [20]. On the other hand, active learning is a set of methods where the model systematically queries labels for the most informative subset of a given dataset. There has been no effort so far, to the best of our knowledge, that leverages weak supervision for better performance in active learning. While weak supervision focuses on learning with cheaper labeling methods, active learning focuses on reducing the number of samples required to label, with full supervision. These two classes of methods differ in their approach of reducing annotation costs. We propose that a combination of weak supervision and active learning can result in greater savings in annotation costs since both the label quality and the size of labeled data can be optimized. In this work, we propose an adaptive supervision framework for active learning and show its effectiveness in training object detectors. We use the standard pool based active learning approach, but instead of querying strong bounding box annotations (which are time consuming), we query a weaker form of annotation first and only query bounding box labels when required. We propose variants in how weak and strong supervision can be interleaved to show the flexibility of the proposed methodology. An overview of our framework is shown in Figure 1. We validate the proposed methodology on standard datasets such as PASCAL VOC 2007 and VOC 2012, as well as in a real-world setting, agriculture, where labeling expertise is expensive, and the proposed methodology can provide significant savings in labeling budgets.

2 Related Work

Previous work on reducing labeling efforts for training object detection methods can be broadly divided into two categories: Weak Supervision and Active Learning. Weakly supervised learning methods focus on reducing the labeling effort for each label, but however result in lower performance due to the imprecise supervision. Active learning methods focus on selecting appropriate image data for querying for labels in an iterative manner, but require fully supervised labels in each iteration.

Weak Supervision. Image-level labels, i.e. the class names of the objects present in the image, are the most common form of weak supervision used in object detection. There have been several efforts on Weakly Supervised Object Localization (WSOL) [1, 2, 3, 12, 25, 31], in which the task is to localize objects in an image given only image-level labels. However, models trained on image level labels typically do not reach the performance level of their fully supervised counterparts. Recently, alternative methods for annotating objects such as center-clicking [18], clicking on the object extremes [19] and bounding box verification [15, 20] have been proposed, which show promising savings in annotation time. However, to the best of our knowledge, weak supervision methods have by far not integrated active learning into their training methodology.

Active Learning. This is a class of techniques used to pick the most beneficial samples to train a model (please see [24] for a detailed survey). Active learning has been shown to be very effective in image classification [9, 14, 16, 23, 28]. In a deep network setting, there have been limited efforts however in active learning for object detection [6, 13, 22]. Active selection metrics such as localization uncertainty [13], margin sampling based on



(a) Standard PBAL framework (b) Proposed Framework

Figure 1: (a) Standard pool-based active learning (PBAL) framework; (b) Proposed framework which interleaves weak supervision in the active learning process. Our framework includes an adaptive supervision module which allows switching to a stronger form of supervision as required when training the model.

convolutional layers [22] and 1-vs-2 margin sampling [5] have been proposed. However, these methods directly query for full supervision during active learning.

In this work, we leverage the advantages of both these categories of methods by introducing an adaptive supervision framework for active learning in object detection. Our framework allows switching between weak and strong supervision to obtain significant savings in annotation cost when compared to earlier works.

3 Methodology

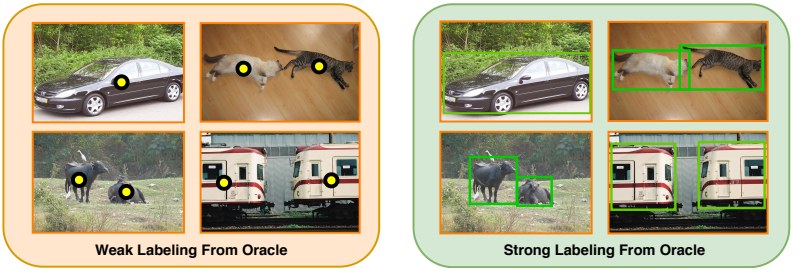
We first present an overview of the proposed framework, before explaining each component in detail. For the rest of the paper, we interchangeably use the terms weak supervision, weak labels and weak annotations.

3.1 Overview

Figure 1a shows the standard pool-based active learning (PBAL) setting for object detection, in which a batch of informative images is queried for bounding box annotations every episode, using which the object detector is updated. In the proposed method, instead of directly querying for time-consuming bounding box annotations, we first query for just weak labels and generate pseudo labels to train the model. Secondly, we introduce an adaptive supervision module to allow switching to strong supervision when required. We introduce two variants of supervision switching, namely *hard switch* and *soft switch*. A *hard switch*, also called *inter-episode switch*, causes the model to permanently switch to a stronger form of supervision at a certain stage of the training process, and after the switch, our framework reduces to a standard PBAL setting (Figure 1a). In contrast, a *soft switch*, also called *intra-episode switch*, allows the model to query both forms of supervision in each round of active learning all through the training process. Based on a switching criterion, in a given batch of actively selected images, the model asks for weak supervision on some images and asks for strong supervision on the other images. More details are provided in Section 3.4. In brief, we show in our experiments that our adaptive supervision module results in substantial savings in annotation time.

3.2 Active Learning Setup

To begin with, we consider a deep object detection model $\mathcal{M}$ (e.g. Faster R-CNN [21]) and a dataset $\mathcal{D}$ which is initially unlabeled. Our aim is to maximize the model performance under



(a) (b)

Figure 2: Illustration of: (a) Weak supervision using center clicks, (b) Strong supervision using bounding box annotations and (c) Soft Switching Mechanism.

a given labeling budget $\mathcal{B}$. We assume, like any other active learning setup, that an initial (randomly chosen) subset of $\mathcal{D}$ is queried for strong labels and a labeled pool of samples, $\mathcal{L}$, is generated. The remaining images form the unlabeled pool $\mathcal{U}$. We also consider a weakly labeled pool $\mathcal{W}$ which is initially empty. As a common practice in active learning, we begin with training our model $\mathcal{M}$ on the initial labeled pool.

The choice of query technique is a key design decision in any active learning method. We study the use of multiple standard query techniques in this work, and show that given any of these querying techniques, our framework can achieve annotation savings when compared to the standard fully-supervised PBAL setting.

3.3 Labeling Techniques

In our framework, the oracle (e.g. a human annotator) can be queried for two types of annotations:

- **Strong Labels:** Strong labeling involves drawing tight bounding boxes around objects in an image, and is the conventional form of labeling used for object detection datasets. Since the annotation times of the datasets we used were unavailable, we use the statistics of ImageNet [26] for consistency, as the difficulty and quality of annotations of PASCAL VOC and ImageNet are quite similar. Su *et al*. [26] and Papadopoulos *et al*. [18] report the following median annotation times on ImageNet: 25.5s for drawing one box, 9.0s for verifying its quality and 7.8s for checking whether there are other objects of the same class yet to be annotated. Hence, we take 34.5s (25.5s + 9.0s) to be the median time taken to draw an accurate bounding box around an object and additionally add 7.8s for every image annotated.
- **Weak Labels:** In our method, we use the recently proposed center-clicking [18] as our weak labeling method. For each object in a given image, the annotator clicks approximately on the center of the imaginary bounding box that encloses the object. Papadopoulos *et al*. [18] report that the maximum median time to click on an object's center is 3.0s.

Figure 2 illustrates both the above mentioned labeling techniques used in our framework.

3.4 Adaptive Supervision

We use the adaptive supervision module which helps in deciding when its time to make a switch from weak to strong supervision. A stronger supervision method takes up more annotation time but provides greater information to the model. We propose two variants of supervision switching, which are at different levels of granularity:

1. **Hard (Inter-Episode) Switch:** As the name suggests, in this method, we define a switching criterion at the end of a given episode based on the change in model's performance on the validation set. Let d_n be the difference between mAP of the model in episode n and $n-1$, and d_{max} be the maximum difference of mAP between any two consecutive episodes until episode n and $\gamma \in [0,1]$ be a suitably chosen threshold value. The criterion can then be written as follows:

$$S_{hard}(n) = \begin{cases} 1 & \text{if } \frac{d_n}{d_{max}} \leq \gamma \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

When the above condition evaluates to 1, we perform a *hard switch* (and hence the name) to strong supervision i.e., our model would query only strong bounding box annotations in later episodes of active learning thus reducing to a standard PBAL setup then on (Figure 1a).

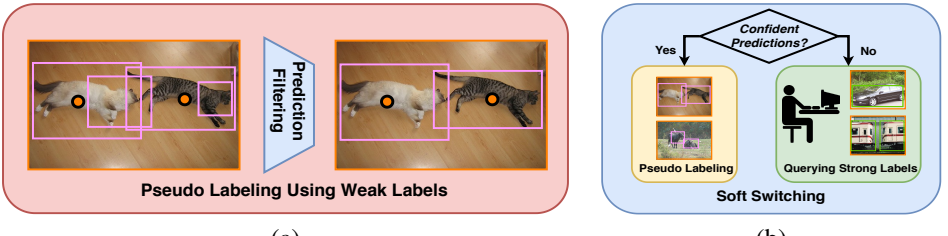
2. **Soft (Intra-Episode) Switch:** In each episode of active learning, we use the obtained weak labels for the actively selected batch to *pseudo-label* these selected images with a bounding box. Pseudo-labeling the images is a simple low-cost step as described in Section 3.5. For each image, we obtain a confidence score c which is the mean probability score obtained for each predicted object. Given the confidence score c_i for a selected image i and a suitably chosen threshold $\delta \in [0,1]$, we perform the soft switch when the following condition evaluates to 1:

$$S_{soft}(i) = \begin{cases} 1 & \text{if } c_i < \delta \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

In other words, we query an image for strong supervision if the model's average confidence on its object predictions is below a threshold δ. Otherwise, we pseudo label the image using its current predictions. This intuitively makes sense because we query for strong labels only when the model is very unsure of its current bounding box predictions, else manage with the weak annotations. We note that this switch is carried out episode-wise, and each new episode starts afresh with seeking weak labels again for images with a reasonably high confidence (and hence, the name *soft switch*).

3.5 Pseudo Labeling using Weak Labels

We use a low-cost pseudo-labeling approach to train object detectors with weakly labeled data. We do not explicitly consider complex training methods for learning with weak supervision [18, 27, 30] to avoid introducing significant computational overhead in the training methodology and to keep our approach as model-agnostic as possible. In this approach, we first use the trained model, $\mathcal{M}$, to predict bounding boxes (which may be imprecise) for all



(a) (b)

Figure 3: We illustrate (a) pseudo labeling using center clicks and (b) the soft switching mechanism used in the adaptive supervision module.

possible classes on the weakly labeled images. We then use the weak labels provided by the oracle to filter and choose the best possible bounding box for each object as follows.

In a given weakly labeled image, each center-click (our weak annotation) location corresponds to an object. For every click location, we pseudo-label that object with a bounding box with center closest to the click location. The object is classified as the class with the highest probability for the chosen bounding box. Computationally, this method involves a forward pass through the network for each image followed by computation of pairwise distances (2 dimensions) between the click locations and centers of predicted bounding boxes. Figure 3a illustrates our pseudo-labeling strategy. We finally use labeled data (from $\mathcal{L}$) and pseudo-labeled data (from $\mathcal{W}$) to retrain our object detection model in an end-to-end manner. Our overall methodology is summarized below in Algorithm 1.

Algorithm 1: Adaptive Supervision for Active Learning

Input : Unlabeled pool $\mathcal{U}$, Labeled pool $\mathcal{L}$, Weak labeled pool $\mathcal{W}$, Model $\mathcal{M}$, episode num n, sample size b, soft switch threshold δ, hard switch threshold γ, query function `ActiveSampling`

Output: Updated model $\mathcal{M}$

// Actively sample b valuable images
S = `ActiveSampling` (from $\{\mathcal{U} \cup \mathcal{W}\}$, sample size = b)
// Query weak annotations on S
W_S = `QueryWeakAnnotations` (S)
// Obtain pseudo labels for S using W_S, as described in Sec 3.5
P_S = `PseudoLabels` (model = $\mathcal{M}$, sample = S, weak supervision = W_S)
if *soft switch* **then**
    $S_{high} := \{i : i \in S \ni \text{confidence}(P_S^i) > \delta\}$
    $S_{low} := \{i : i \in S \ni \text{confidence}(P_S^i) \leq \delta\}$
    // Use pseudo labels for S_{high}
    $S_{high}^{Pseudo} := P_S^i : \text{i} \in \text{S}_{high}$
    // Query strong annotations on S_{low}
    $S_{low}^{Strong} := (S_{low})$
    $\mathcal{L} \leftarrow \mathcal{L} \cup S_{low}^{Strong}$; $\mathcal{W} \leftarrow \mathcal{W} \cup S_{high}^{Pseudo}$
else if *hard switch* **then**
    d := difference in *mAP* between last two episodes
    d_{max} := maximum difference in *mAP* between episodes so far
    if $\frac{d}{d_{max}} \leq \gamma$ **then**
        Use fully supervised pool-based active learning from next episode
    $\mathcal{W} \leftarrow \mathcal{W} \cup P_S$
Train model $\mathcal{M}$ on $\{\mathcal{L} \cup \mathcal{W}\}$
return $\mathcal{M}$

4 Experiments and Results

4.1 Implementation Details

Active Sampling Techniques. The choice of query technique is a key design decision in any active learning method. Considering our framework is independent of the query method, We study the following query techniques to actively sample images: **(i) Max-Margin:** For a predicted bounding box, margin is calculated as the difference between the first and the

second highest class probabilities. For each image, margin is chosen to be the summation of margins across all the predicted bounding boxes in the image, as in Brust *et al*. [4]. **(ii) Avg-Entropy:** Samples with high entropy in the probability distribution of the predictions are selected, as in Roy *et al*. [22]. **(iii) Least Confident:** Confidence for an image is calculated as the highest bounding box probability in that image. Images with least confidence are selected. This criterion is taken from the minmax method specified in [22].

Evaluation Metrics: In our experiments, we measure the annotation effort required to reach a certain level of test performance. Annotation effort is measured in terms of time taken to annotate images through the active learning cycle. As discussed in 3.3, to have a consistent measure, we follow the previous work on click supervision[18] and utilize the median annotation times reported on ImageNet [26] to compute time taken for bounding box annotations and weak annotations. In our experiments, we use object center clicks as the chosen form of weak supervision. To get weak supervision for our datasets, we obtain the centers of the ground truth bounding boxes and perturb the center location by a small zero mean Gaussian random noise for robustness. Given an image I with b_I objects, we hence calculate annotation time (in seconds) as:

$$Time(I) = \begin{cases} 7.8 + 34.5 \times b_I & \text{for bounding box annotations} \\ 7.8 + 3 \times b_I & \text{for center click annotations} \end{cases} \tag{3}$$

We use mean Average Precision (mAP) to evaluate the performance of the detection itself.

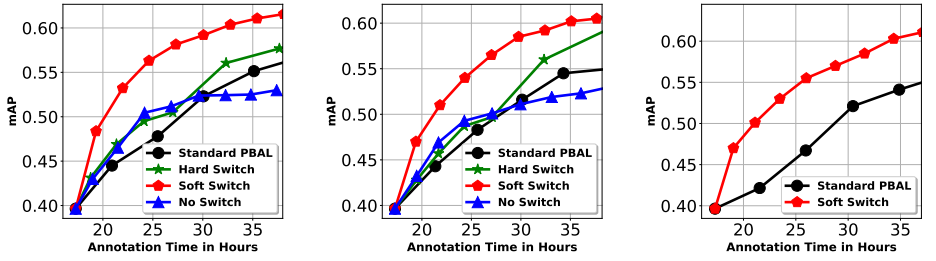
4.2 Results on PASCAL VOC 2007

Setup. We show results on PASCAL VOC 2007 [7] with 20 object classes. We use the trainval set of 5011 images as our training set $\mathcal{D}$ and evaluate our model's performance on the test set of 4952 images. In all our experiments, we use Faster R-CNN [21] with ResNet-101 [11] backbone as our object detection model, and extend PyTorch implementation by [29]. As in Section 3.2, we follow the standard pool-based active learning (PBAL) setup. We choose 500 images (around 10% of dataset) as the initial labeled pool $\mathcal{L}$ and train our model on it.

Active Learning. With the same initial model, we use different query techniques: max-margin sampling [4], least-confident [22] and avg-entropy [22]. For each query method, we implement the standard PBAL framework, our adaptive supervision framework with hard switching and soft switching. We do not do hard switching for the least confident sampling method because pseudo labeling the least confident samples using the model is counter-intuitive. We fix an annotation budget $\mathcal{B}$ of 35 hours. Until this budget is exhausted, we run multiple episodes of active learning. In any given episode, if we're querying for strong supervision, we query 250 images (around 5% of the dataset size). If we're querying for weak supervision instead, we query 500 images (around 10% of the dataset size).

Adaptive Supervision. While performing active learning with our adaptive supervision module, we set $\gamma = 0.3$ as our *hard switch* threshold, i.e. we switch to strong supervision when the test mAP increase in the last episode is less than 30% of the maximum test mAP increase in any previous episode. While using *soft switch*, we set the probability threshold $\delta = 0.75$ i.e., if a model's average confidence on an actively sampled image is less than 0.75, that image will be queried for strong labels.

Evaluation. Figure 4 shows the performance of various training methods for three different active sampling methods. In the figure, 'Standard PBAL' represents the standard PBAL with strong supervision in every episode. The graphs corresponding to hard switch and soft switch



(a) Avg-Entropy (b) Max-Margin (c) Least-Confident

Figure 4: PASCAL VOC 2007: For each active query method, we show performance of our adaptive supervision methods against standard PBAL framework (budget = 35 hours).

represent our adaptive supervision methods. Finally, *no switch* represents active learning using only weak supervision. It can be observed that our soft switch method significantly outperforms the standard PBAL method. For example, as seen in Figure 4a, to achieve a test mAP close to 0.55, standard PBAL requires $\approx$ 35 hours of annotation time whereas soft switch requires only 24.6 hours (30% savings) and hard switch requires around 30 hours (14% savings). Similarly the graphs for the other two metrics show that soft switch method achieves significant reduction in annotation efforts. A few qualitative results on VOC 2007 are shown in Figure 7, the top row images were pseudo labeled using just weak labels, the bottom row images were queried for strong labels. A significant difference in prediction quality can be observed between them.

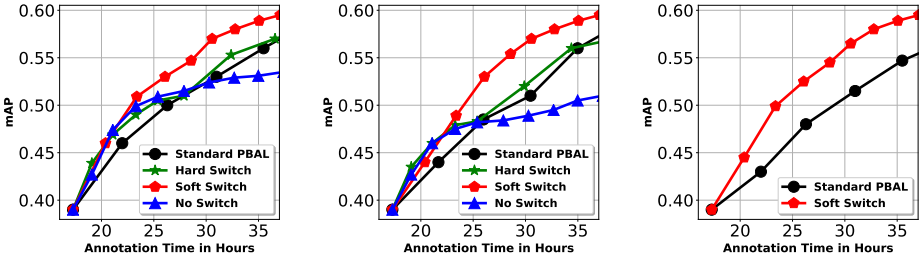
4.3 Results on PASCAL VOC 2012

Setup. We perform similar experiments on PASCAL VOC 2012 [8] which also has 20 object classes. We use the training set of 5717 images as our training set $\mathcal{D}$ and evaluate our model's performance on the validation set of 5823 images. The experimental setup is same as that of PASCAL VOC 2007, with an annotation budget $\mathcal{B}$ of 35 hours. We query 250 images in a strong supervision episode and 500 images in a weak supervision episode. We use the same threshold values for adaptive supervision used for the experiments on PASCAL VOC 2007.

Evaluation. We show the performance comparison of different supervision techniques for three different active sampling techniques in Figure 5. Once again, soft switching outperforms all other compared methods. For example, in Figure 5a, for achieving a test mAP close to 0.55, soft switching requires around 29 hours of annotation time (17.1% savings) compared to 32.5 for hard switching (7.1% savings) and 35 hours for the standard PBAL method. Also, hard switching slightly outperforms the standard PBAL method as seen in Figure 5a and 5b. Thus, hard switching can be used in a case where obtaining weak and strong supervision at the same time is not feasible.

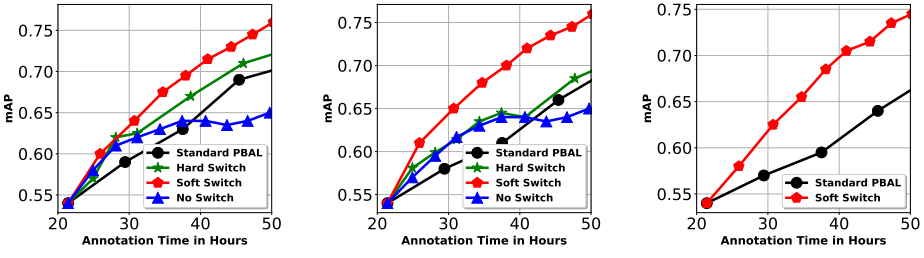
4.4 Results on Wheat

Setup. In addition to standard datasets, we show the effectiveness of our adaptive supervision framework on a real world agriculture dataset of wheat images. Obtaining expert level labels on agricultural datasets is generally expensive. In this case, we show that our framework can result in significant savings in labeling efforts. We use the Wheat dataset by Madec *et al.* [17], which contains high definition images of wheat plants with objects of a single class: wheat head. To create a dataset suitable for training a deep object detection network, we preprocess the original 4000×6000 images as follows. We first downsample the images by a factor of 2 (to 2000×3000) using a bi-linear aggregation function. We then split these



(a) Avg-Entropy (b) Max-Margin (c) Least-Confident

Figure 5: PASCAL VOC 2012: For each active query method, we show performance of our adaptive supervision methods against standard PBAL framework (budget = 35 hours).



(a) Avg-Entropy (b) Max-Margin (c) Least-Confident

Figure 6: Wheat: For each active query method, we show performance of our adaptive supervision methods against standard PBAL framework (budget = 50 hours).

downsampled images into tiles of 500×500 images with no overlap (this has no impact on the study due to the nature of these images). We split the set of obtained 5663 images to use 4530 images (80%) for training and 1133 images (20%) for testing our methods. We choose 450 images (around 10% of the dataset size) as the initial labeled pool $\mathcal{L}$ and train our model on it.

Active Learning and Adaptive Supervision. Since this is a dataset with a highnumber of object instances in each image, we use an annotation budget $\mathcal{B}$ of 50 hours. Until this budget is exhausted, we run multiple episodes of active learning. In any given episode, if we're querying for strong supervision, we query 250 images. If we're querying for weak supervision instead, we query 500 images. While performing active learning with our adaptive supervision module, we set $\gamma = 0.3$ as our *hard switch* threshold. While using *soft switch*, we set the probability threshold $\delta = 0.85$.

Evaluation. Figure 6 shows the performance comparison of various supervision techniques for three different active sampling techniques. It can be observed that soft switch performs better than all other supervision techniques. As an example, for avg-entropy sampling (Figure 6a), to attain a test mAP of around 0.68, soft switch method requires 34 hours of annotation (24% savings), hard switch requires 38 hours (15.5% savings) whereas the standard PBAL method requires around 45 hours.

5 Discussion

Given an annotation budget, the choice of hard switch and soft switch thresholds is crucial in getting optimum performance out of the model. A higher value of γ (hard switch threshold) results in a quicker switch to strong supervision, which would quickly deplete the annotation budget. Similarly, a lower γ would result in a delayed switch to strong labeling. This would reduce strong label requirement but at the cost of providing a lot of noisy labels to the model.

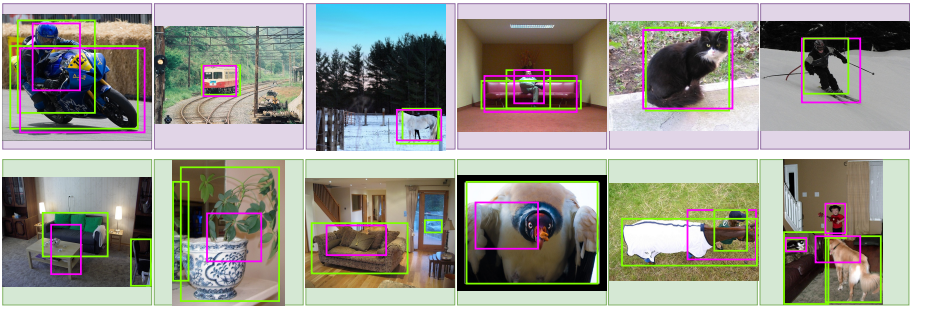


Figure 7: Soft switch mode on VOC 2007: (Top row) Images pseudo labeled using weak labels vs (Bottom row) images queried for strong labels. Boxes in pink were predicted while boxes in green denote ground truth.

Similarly, a lower δ (soft switch threshold) would provide a lot of noisy labels to the model and a higher δ might mostly query only for strong labels. In other words, the hard switch and soft switch thresholds can be seen as knobs to adjust label quality, annotation costs and the number of training episodes taken to reach a desired level of performance.

To understand the effect of adaptive supervision alone without active sampling, we conducted an ablation study to evaluate the performance of our framework in the context of a passive learning (random sampling) setting. It can be seen in Figure 8 that our adaptive supervision methods still outperform the standard PBAL method. To attain a test mAP of 0.53, standard PBAL requires around 35 hours of annotation time whereas hard switch requires 31 hours (11.4% savings) and soft switch requires 30.4 hours (13.1% savings). From this experiment, we observe that our methods reduce annotation effort even in the passive sampling case, albeit a lower percentage of savings when compared to performance on top of active sampling techniques.

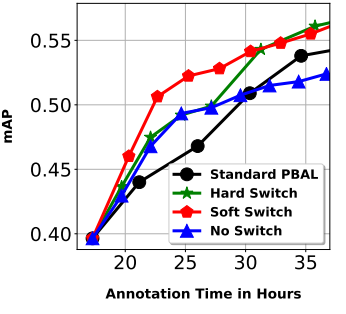


Figure 8: Performance comparison of different supervision techniques in passive learning on PASCAL VOC 07

6 Conclusions

Using our proposed adaptive supervision framework, we empirically show that active learning approaches can be interleaved with multiple levels of supervision to achieve significant savings in annotation effort required to train deep object detectors. By only using the prediction outputs of the object detection model, we develop two supervision switching techniques: hard switch (inter episode switch) and soft switch (intra episode switch). Our experiments show that our adaptive supervision methods outperform standard PBAL on standard active query techniques. We believe that our work could open up a range of possibilities in fusing weak supervision techniques with active learning such as: using other forms of weak supervision with active learning, posing the problem of combining weak and strong supervision as an optimization problem under given budget constraints, combining active learning techniques with data programming based weak supervision techniques to name a few.

References

[1] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with convex clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1081–1089, June 2015. doi: 10.1109/CVPR.2015.7298711.

[2] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2846–2854, 2016.

[3] Hakan Bilen, Marco Pedersoli, and Tinne Tuytelaars. Weakly supervised object detection with posterior regularization. 2014.

[4] Clemens-Alexander Brust, Christoph Käding, and Joachim Denzler. Active learning for deep object detection. *CoRR*, abs/1809.09875, 2018. URL http://arxiv.org/abs/1809.09875.

[5] Clemens-Alexander Brust, Christoph Käding, and Joachim Denzler. Active learning for deep object detection. *CoRR*, abs/1809.09875, 2018. URL http://arxiv.org/abs/1809.09875.

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, apr 2018. doi: 10.1109/tpami.2017.2699184. URL https://doi.org/10.1109%2Ftpami.2017.2699184.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, .

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, .

[9] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *ICML*, 2017.

[10] K. He, G. Gkioxari, P. DollÃąr, and R. Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, Oct 2017. doi: 10.1109/ICCV.2017.322.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

[12] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. 09 2016.

[13] Chieh-Chi Kao, Teng-Yok Lee, Pradeep Sen, and Ming-Yu Liu. Localization-aware active learning for object detection. *CoRR*, abs/1801.05124, 2018.

[14] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Active learning with gaussian processes for object categorization. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, Oct 2007. doi: 10.1109/ICCV.2007.4408844.

[15] Ksenia Konyushkova, Jasper Uijlings, Chris Lampert, and Vittorio Ferrari. Learning intelligent dialogs for bounding-box annotation. 2018. URL https://arxiv.org/abs/1712.08087.

[16] X. Li and Y. Guo. Adaptive active learning for image classification. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 859–866, June 2013. doi: 10.1109/CVPR.2013.116.

[17] Simon Madec, Xiuliang Jin, Hao Lu, Benoit de Solan, Shouyang Liu, Florent Duyme, Emmanuelle Heritier, and Baret Frederic. Ear density estimation from high resolution rgb imagery using deep learning technique. *Agricultural and Forest Meteorology*, 264: 225–234, 01 2019. doi: 10.1016/j.agrformet.2018.10.013.

[18] Dim Papadopoulos, Jasper Uijlings, Frank Keller, and Vittorio Ferrari. Training object class detectors with click supervision. In *CVPR*, 2017. URL https://arxiv.org/abs/1704.06189.

[19] Dim Papadopoulos, Jasper Uijlings, Frank Keller, and Vittorio Ferrari. Extreme clicking for efficient object annotation. In *ICCV*, 2017. URL https://arxiv.org/abs/1708.02750.

[20] Dim P. Papadopoulos, Jasper R. R. Uijlings, Frank Keller, and Vittorio Ferrari. We donâĂŹt need no bounding-boxes: Training object class detectors using only human verification. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 854–863, 2016.

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015. URL http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-reg pdf.

[22] Soumya Roy, Asim Unmesh, and Vinay P. Namboodiri. Deep active learning for object detection. In *BMVC*, 2018.

[23] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR 2018*, 2018.

[24] Burr Settles. Active learning literature survey. Technical report, 2010.

[25] Hyun Oh Song, Ross Girshick, Stefanie Jegelka, Julien Mairal, Zaid Harchaoui, and Trevor Darrell. On learning to localize objects with minimal supervision. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1611–II–1619. JMLR.org, 2014. URL http://dl.acm.org/citation.cfm?id=3044805.3045072.

[26] Hao Su, Jia Deng, and Li Fei-Fei. Crowdsourcing annotations for visual object detection. In *HCOMP@AAAI*, 2012.

[27] Chong Wang, Weiqiang Ren, Kaiqi Huang, and Tieniu Tan. Weakly supervised object localization with latent category learning. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 431–445, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10599-4.

[28] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Trans. Cir. and Sys. for Video Technol.*, 27(12):2591–2600, December 2017. ISSN 1051-8215. doi: 10.1109/TCSVT.2016.2589879. URL https://doi.org/10.1109/TCSVT.2016.2589879.

[29] Jianwei Yang, Jiasen Lu, Dhruv Batra, and Devi Parikh. A faster pytorch implementation of faster r-cnn. *https://github.com/jwyang/faster-rcnn.pytorch*, 2017.

[30] D. Zhang, D. Meng, C. Li, L. Jiang, Q. Zhao, and J. Han. A self-paced multiple-instance learning framework for co-saliency detection. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 594–602, Dec 2015. doi: 10.1109/ICCV.2015.75.

[31] Yi Zhu, Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Soft proposal networks for weakly supervised object localization. *CoRR*, abs/1709.01829, 2017. URL http://arxiv.org/abs/1709.01829.