Towards Fine-grained Sampling for Active Learning in Object Detection

Sai Vikas Desai, Vineeth N Balasubramanian
Indian Institute of Technology, Hyderabad, India

Abstract

We study the problem of using active learning to reduce annotation effort in training object detectors. Existing efforts in this space ignore the fact that image annotation costs are variable, depending on the number of objects present in a single image. In this regard, we examine a fine-grained sampling based approach for active learning in object detection. Over an unlabeled pool of images, our method aims to selectively pick the most informative subset of bounding boxes (as opposed to full images) to query an annotator. We measure annotation efforts in terms of the number of ground truth bounding boxes obtained. We study the effects of our method on the Feature Pyramid Network and RetinaNet models, and show promising savings in labeling effort to obtain good detection performance.

1. Introduction

State-of-the-art deep object detectors such as Faster R-CNN [18], Feature Pyramid Networks [15], RetinaNet [16] and YOLO [17] have been shown to achieve excellent performance in visual object detection. However, it is well-known that training such networks requires large amounts of bounding box labeled data. Acquiring a large-scale annotated dataset for object detection is both time-consuming and expensive. It has been reported [24] that drawing bounding boxes is at least ten times costlier than labeling an image for the classification task. While large-scale datasets propel detection performance, not all bounding box samples in such datasets are equally valuable. In object detection, each image can have multiple labels depending on the number of objects. Also, the labeling cost of each image can vary based on the number of objects present in a single image. Nonetheless, existing methods for active learning in object detection [19, 12, 3, 5] treat a single image as an instance to be labeled. The annotation cost is also measured in terms of number of images labeled. While these assumptions enable one to construct image-level selection metrics (similar to classification), it is clear that object detection labeling is done at a bounding box level. Each bounding box drawn contains a piece of information useful to the model. We consider these assumptions and propose fine-grained sampling, a bounding box level active learning method where each potential bounding box is considered as a sample to be labeled and adjusted. Figure 1 summarizes our approach.

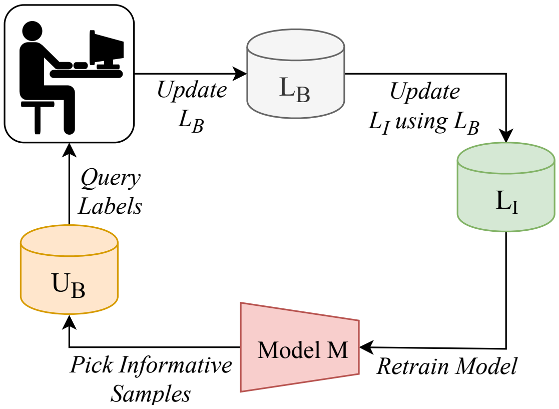


Figure 1: **Active Learning Framework:** Small informative subsets of bounding boxes are picked from U_B and queried for labeling. Subsequently, L_B and L_I are updated. L_I is used to retrain the model.

We train a baseline object detector on a small pool of randomly selected images, which is used to predict bounding boxes for each image, to effectively create a bounding box dataset. Each bounding box is now a single entity to be labeled and adjusted. Keeping the current state-of-the-art object detection models in mind, we point out a few challenges in using fine-grained bounding box sampling for active learning: *(i) Covering all objects:* It is possible that some objects might be missed while creating the bounding box dataset since the model is trained on a small subset of data. We address this issue by choosing a low detection threshold so as to maximize recall while making bounding box predictions to generate the box dataset. In addition, we choose the size of the initial subset based on the difficulty and size of the full dataset. *(ii) Training with partially labeled images:* During active learning, sampling bounding boxes instead of images can result in partially labeled images at a given episode, since picking all boxes in a given image is not guaranteed. We aim to alleviate this problem by devising a dampening parameter to reduce the effect of the non-sampled bounding boxes in an image on the training loss. The key contributions of our work can hence be summarized as follows:

- We propose a simple approach for fine-grained sampling approach for active learning to train object detectors, and introduce a new methodology built on existing object detection models to achieve the same.
- We show promising results on the standard benchmark PASCAL VOC-12 and MS-COCO datasets.

We hope that our work will promote open discussions around the broader issues of active learning in training object detection models with limited labels.

2. Related Work

Several active learning (AL) frameworks have been proposed including stream-based sampling [2], membership query synthesis [23] and pool-based active learning [13]. AL has been applied to a variety of machine learning algorithms [9, 11, 25, 14]. For a detailed survey, we request the interested reader to refer to [22].

When compared to the area of active learning on classification, active learning for object detection is a relatively less explored direction of research. Prior to deep learning, Abramson and Freund [1] proposed a boosting based active selection method for pedestrian detection. Vijayanarasimhan and Grauman [26] suggested using a margin based active selection method for training an SVM based object detector. These methods most often use uncertainty information from classifiers to select instances for labeling. In a deep neural network setting, there have been very few efforts [19, 12, 3, 5] in the space of active learning for object detection. Roy *et al*. [19] propose a query-by-committee technique based on the layered structure of object detection networks. Kao *et al*. introduced a novel sampling strategy based on predicting the localization performance of the detector. Brust *et al*. [3] suggest effective aggregation metrics for getting image level uncertainty scores and also combine active learning with an incremental learning approach. Desai *et al*. [5] propose a novel scheme to effectively leverage the combination of weak supervision and strong supervision in the active learning process. These works perform active learning at an image level i.e., consider each image as an instance to be queried and labeled. We argue that, unlike classification, the annotation cost in object detection should be measured in terms of number of bounding boxes drawn. To the best of our knowledge, none of the works in this space have attempted active selection queries at the level of bounding boxes. To bridge that gap, we introduce the concept of bounding box queries in our work.

3. Methodology

We study a bounding box based querying method for active learning in object detection. In a standard active learning setup [3, 12, 20], images are queried and labeled. In contrast, our proposed method generates queries which ask for bounding box labels.

3.1. Initial Setup

The initial setup comprises of training an initial baseline model using a small subset of images and consequently generating a bounding box dataset. A pool-based active learning framework is applied on this dataset in the next steps.

Given a dataset D_I of images which is completely unlabeled at the beginning, our method randomly selects a small subset of images L_I and queries their labels. The size of this subset is suitably chosen based on the overall difficulty of the dataset, the number of classes, etc. In a typical active learning setup, the initial subset size is around 5-10% of the dataset size. The remaining unlabeled images constitute U_I. An object detection model is trained on L_I to obtain an initial baseline model M. Note that each image in L_I is fully labeled i.e., all objects of interest are annotated with bounding boxes. This allows for training our initial model in a robust manner, with a reasonably small annotation effort. This initial model is used to generate the bounding box dataset D_B which is used in further steps.

To generate D_B, the initial model M is run on each image from U_I, and a suitable detection threshold t_d is chosen. A predicted bounding box b is added to U_B when the following membership function evaluates to 1:

$$M(b) = \begin{cases} 1, & \text{if } prob(b) \geq t_d \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Ground truth bounding boxes of each image in L_I are aggregated to form L_B. Finally, D_B is created as the union of L_B and U_B.

3.2. Active Learning Framework

The bounding box dataset D_B consists of: (1) originally labeled bounding boxes from the initial pool; and (2) bounding boxes predicted by the initial model on the unlabeled pool of images U_I. A pool-based active learning framework [22] can be readily applied on such a dataset. In contrast to existing active learning methods for object detection, we employ our framework on a bounding box dataset instead of an image dataset. In this framework, active learning consists of multiple learning episodes. As shown in Figure 1, each episode mainly comprises of three steps: (1) Sampling from Unlabeled Pool, (2) Updating the training dataset and (3) Retraining the model. Typically, active learning episodes are carried out until the desired detection performance is reached.

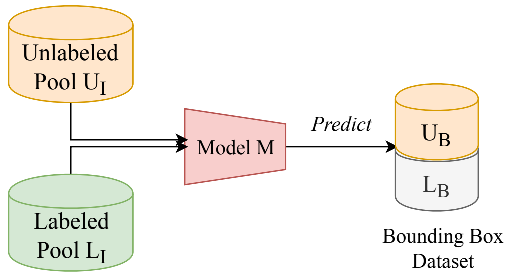


Figure 2: **Creating the Bounding Box Dataset:** The model M is run on images from unlabeled image pool U_I and the labeled image pool L_I to obtain bounding box predictions to generate the bounding box dataset D_B.

Sampling from Unlabeled Pool: The bounding boxes in U_B might not be accurately drawn or classified since they are obtained as predictions from a model trained on an small randomly chosen subset of training data. The aim of our active learning method is to pick and adjust the minimum number of bounding boxes to obtain a desired level of detection performance. The aim of this step is to choose a subset of most informative K samples from the unlabeled pool U_B to be queried and adjusted by the annotator. Here, K is the size of the subset of samples queried in each episode. We solve the following optimization K times:

$$b_{picked} = \arg\max_{b \in U_B} E_b \tag{2}$$

Here, E_b denotes the informativeness measure of a bounding box b which measure how useful its label could be, to the model.

Updating the Training Dataset: In an episode, once the samples are picked from the unlabeled pool U_B, they are queried for labeling. A bounding box is labeled by an oracle as follows: if it partially encloses an object, then its dimensions are adjusted such it tightly encloses the object and the class of the object is labeled appropriately. If the bounding box doesn't enclose any object, then it is labeled as a background and its dimensions are not adjusted. After adjusting the picked bounding boxes, they are added to the labeled box pool L_B.

Since contemporary object detection networks are trained using images, we use the labeled box pool L_B to update the labeled image pool L_I as follows. Each bounding box in the labeled box pool is drawn on its corresponding image. All images with bounding boxes in the labeled box pool, L_B, are added to the labeled image pool, L_I by default. In addition, we include all the bounding boxes in U_B which correspond to the images present in L_I. This means that each image in the updated training set L_I now contains: (1) the labeled ground truth bounding boxes from L_B; and (2) unlabeled bounding boxes from U_B which are present on the image but are not sampled for active learning yet. Such bounding boxes act as noisy labels. They are used as part of L_I for training the model but they still are eligible to be queried in further stages of active learning.

Retraining the Model: Since the updated labeled pool of images L_I can now consist of both labeled ground truth boxes as well as other objects that are unlabeled, we modify the training loss function to minimize the amount of noise involved in training the model. In our training method, we define the loss function for an image as follows:

$$L = L_{labeled} + \lambda_d L_{noisy} \tag{3}$$

Here, $L_{labeled}$ is the average loss over all the anchors assigned to a labeled bounding box and L_{noisy} refers to the average loss over all the anchors assigned to a noisy label. We define λ_d as the dampening parameter which controls the effect of noisy labels on the loss. Our experiments show that setting $\lambda_d = 0.5$ gives good results in most cases.

3.3. Informativeness Measures

An informativeness measure is a metric used in actively sampling a subset of data. This is an important design decision in any active learning framework. We experiment with the following informativeness measures in our work:

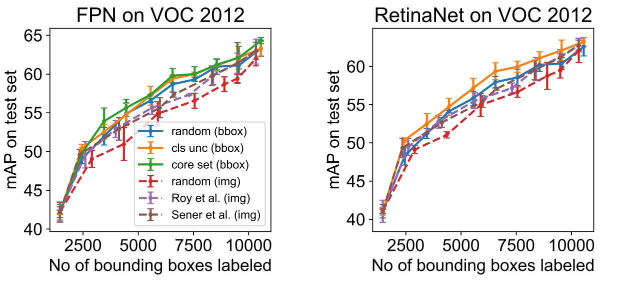
1. Random selection: In this method, samples are chosen randomly from the unlabeled pool. This method acts as a baseline for other active learning metrics.

2. Mean Classification Uncertainty: In case of bounding boxes, the boxes with the least predicted class probability are picked. In case of images, the average of the classification uncertainties of the predicted bounding boxes is computed. The classification uncertainty of a predicted bounding box is calculated as $(1 - P)$ where P is the highest predicted probability score for the box. Images with high mean classification uncertainty are chosen for querying. This criterion has been used for object detection in Roy *et al.* [20].

3. Coreset Greedy selection: This metric is based on the feature geometry of data samples. Sener and Savarese [21] have proposed this method for classification problems, which we repurpose for object detection (please refer the appendix for further details). For image-level coreset selection, the distances are calculated based on image-level feature representations, typically produced by a backbone network such as ResNet [7].

4. Experiments

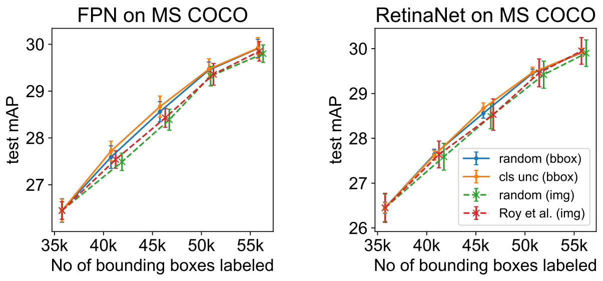
Implementation Details: We use Detectron2 [28] implementations of Faster R-CNN + Feature Pyramid Network [15] and RetinaNet [16] in all our experiments. ResNet-50 [7] is used as the backbone network, which is pretrained on ImageNet1k [4]. To implement the greedy algorithm for core-set selection of bounding boxes in FPN+Faster R-CNN, we make use of the FC2 feature representations of each bounding box. We use the L2 distance between the



(a) FPN + Faster R-CNN

(b) RetinaNet

Figure 3: AL results on PASCAL VOC 2012 showing number of bounding boxes labeled vs test mAP.



(a) FPN + Faster-RCNN

(b) RetinaNet

Figure 4: AL results on MS COCO showing number of bounding boxes labeled vs test mAP.

feature representations as the distance metric. We omit coreset bounding box selection for RetinaNet as it is a single stage network with no intermediate bounding box representations. For all querying methods, we use $\lambda_d = 0.5$ (in Eqn 3) for retraining the model in each episode.

In our experiments, in case of overlap with multiple ground truth boxes, a selected bounding box is labeled as that ground truth box with which it has maximum IoU score. However, if it does not have at least $IoU \geq 0.5$ with a ground truth bounding box of some class label, the bounding box is classified as *background* and its coordinates are not adjusted during the oracle's annotation. To evaluate the performance of object detection, we use the standard mean average precision (mAP) metric.

Datasets:

PASCAL VOC 2012: We use the train set of 5717 images as the unlabeled pool of data. The validation set of 5832 images is used to test the performance of the model. An initial baseline model is trained on a random subset of 500 images from the unlabeled pool. To ensure fairness of comparison, we use the same initial baseline model for all active learning methods. For image-based baseline methods, 500 images are queried in each round. In our proposed bounding box querying method, we set the sample batch size equal to 1000 bounding boxes. In each round of AL, the model is retrained for 15 epochs using SGD with a learning rate of 0.002. We use a batch size of 2 images. A learning rate decay of 0.1 is applied after every 8 epochs of training.

MS COCO: We use the train2017 subset for training and val2017 for reporting the results. An initial baseline model is trained on a random subset of 5000 images. We execute 4 AL rounds and measure the performance. For image-based baseline methods, 1000 images are queried in each round. In our proposed bounding box querying method, we set the sample batch size equal to 5k bounding boxes. The model is trained using SGD for 8 epochs with a learning rate of 0.0025. We use a batch size of 2 images.

Results: We assess the performance of various AL metrics in an image-based querying setting and a bounding box-based querying setting. Figure 4 shows mAP on the test set plotted against the number of bounding boxes labeled for models trained using various active learning methods on the MS COCO dataset. Similar results on PASCAL VOC 2012 are shown in Figure 3. All results shown are averaged over three runs of each experiment. In all the four settings, as a general trend, it can be seen that just by using bounding box querying instead of image-based querying, the performance on AL metrics are improved. While the curves may not demonstrate this explicitly, while training FPN on PASCAL VOC 2012, attaining an mAP of 60 points now requires less than 6500 bounding box labels by the best bounding box querying method. To achieve the same mAP, image-based querying methods require close to 9000 bounding box labels (Figure 3a). This reduction could be significant when annotation cost is high. In fact, 85% of the fully supervised test mAP has been reached using just 50% of the bounding box labels. For MS COCO, a large dataset with 80 classes, the "random" image-based querying method seems to perform close to bounding-box querying methods. Nevertheless, bounding box querying methods still consistently perform better than image-level methods. Picking in terms of bounding boxes allows for a more granular selection approach.

5. Conclusion

We studied a fine grained sampling method for active learning in object detection which seeks to query at the level of bounding boxes instead of an image. We experiment with various query strategies in a bounding box querying setting as well as an image based querying setting. Our experiments on the popular PASCAL VOC-12 and MS COCO datasets show the effectiveness of bounding box queries in reducing human annotation efforts when compared to image-level queries. Our future work includes studying this method in real-world settings from application domains (such as healthcare or agriculture) where annotation costs are prohibitive for the object detection task, as well as study alternate strategies for bounding box-informativeness in object detection frameworks.

References

[1] Yotam Abramson and Yoav Freund. Active learning for visual object recognition. 2006. 2

[2] Les E. Atlas, David A. Cohn, and Richard E. Ladner. Training connectionist networks with queries and selective sampling. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems 2*, pages 566–573. Morgan-Kaufmann, 1990. 2

[3] Clemens-Alexander Brust, Christoph Käding, and Joachim Denzler. Active learning for deep object detection. *ArXiv*, abs/1809.09875, 2018. 1, 2

[4] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. 3

[5] Sai Vikas Desai, Akshay L Chandra, Wei Guo, Seishi Ninomiya, and Vineeth N. Balasubramanian. An adaptive supervision framework for active learning in object detection. In *BMVC*, 2019. 1, 2

[6] Ross B. Girshick. Fast r-cnn. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 3

[8] Dorit S. Hochbaum and David B. Shmoys. A best possible heuristic for the k-center problem. *Math. Oper. Res.*, 10(2):180–184, May 1985.

[9] A. Holub, P. Perona, and M. C. Burl. Entropy-based active learning for object recognition. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, June 2008. 2

[10] Wen-Lian Hsu and George L. Nemhauser. Easy and hard bottleneck location problems. *Discrete Applied Mathematics*, 1(3):209 – 215, 1979.

[11] A. J. Joshi, F. Porikli, and N. Papanikolopoulos. Multi-class active learning for image classification. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2372–2379, June 2009. 2

[12] Chieh-Chi Kao, Teng-Yok Lee, Pradeep Sen, and Ming-Yu Liu. Localization-aware active learning for object detection. In *ACCV*, 2018. 1, 2

[13] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pages 3–12, New York, NY, USA, 1994. Springer-Verlag New York, Inc. 2

[14] 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. 2

[15] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, July 2017. 1, 3

[16] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017. 1, 3

[17] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015. 1

[18] 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. 1

[19] Soumya Roy, Asim Unmesh, and Vinay P. Namboodiri. Deep active learning for object detection. In *British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018*, page 91, 2018. 1, 2

[20] Soumya Roy, Asim Unmesh, and Vinay P. Namboodiri. Deep active learning for object detection. In *BMVC*, 2018. 2, 3

[21] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR 2018*, 2018. 3

[22] Burr Settles. Active learning literature survey. Technical report, 2010. 2

[23] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 287–294, New York, NY, USA, 1992. ACM. 2

[24] Hao Su, Jia Deng, and Li Fei-Fei. Crowdsourcing annotations for visual object detection. In *HCOMP@AAAI*, 2012. 1

[25] Sudheendra Vijayanarasimhan and Kristen Grauman. Cost-sensitive active visual category learning. *International Journal of Computer Vision*, 91(1):24–44, Jan 2011. 2

[26] S. Vijayanarasimhan and K. Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. In *CVPR 2011*, pages 1449–1456, June 2011. 2

[27] Gert W. Wolf. Facility location: concepts, models, algorithms and case studies. series: Contributions to management science. *International Journal of Geographical Information Science*, 25(2):331–333, 2011.

[28] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019. 3