BOUNDING BOX PRIORS FOR CELL DETECTION WITH POINT ANNOTATIONS

Hari Om Aggrawal, Dipam Goswami, Vinti Agarwal

1 Independent Researcher, India *
2 Birla Institute of Technology and Science, Pilani, India

ABSTRACT
The size of an individual cell type, such as a red blood cell, does not vary much among humans. We use this knowledge as a prior for classifying and detecting cells in images with only a few ground truth bounding box annotations, while most of the cells are annotated with points. This setting leads to weakly semi-supervised learning. We propose replacing points with either stochastic (ST) boxes or bounding box predictions during the training process. The proposed “mean-IOU” ST box maximizes the overlap with all the boxes belonging to the sample space with a class-specific approximated prior probability distribution of bounding boxes. Our method trains with both box- and point-labelled images in conjunction, unlike the existing methods, which train first with box- and then point-labelled images. In the most challenging setting, when only 5% images are box-labelled, quantitative experiments on a urine dataset show that our one-stage method outperforms two-stage methods by 5.56 mAP. Furthermore, we suggest an approach that partially answers “how many box-labelled annotations are necessary?” before training a machine learning model. The code is available at https://github.com/hariagr/bbox_prior.

Index Terms— object detection, point annotations, bounding box priors.

1. INTRODUCTION
Microscopic urinalysis is a gold-standard diagnostic test to detect urinary tract infections and kidney disorders [1]. Anemia is usually determined by the microscopic examination of the blood [2]. Urine and blood contain various cells that must be accurately detected and classified for medical diagnosis. Deep learning-based supervised models are state-of-the-art for detecting objects. But, they are hungry for high-quality labelled training data. Numerous urine and blood samples can be easily collected in labs, but their annotation is time-consuming and requires extensive medical expert knowledge. Therefore, weakly and/or semi-supervised models are preferable if they perform similarly to a supervised model.

The shape, size, and features of a particular cell do not vary much among humans. However, their appearances differ a lot in microscopic digital images. Improper focus adjustments, lighting conditions, and the quality of image sensors and lenses affect the true features of a cell. Moreover, images contain just two-dimensional (2D) projections of a 3D cell. Therefore, many cell instances are required to train a deep-learning model for classification.

On the contrary, the size of bounding boxes covering cells does not change much, as we notice by drawing joint probability density functions (pdf) of the width and height of bounding boxes; see Fig. 1. In urinary-sediment dataset (USD) [1], the erythrocyte and leukocyte classes are mainly square and have a very narrow distribution. Other classes, except epithelial nuclei, are rectangular and have wider distributions due to both the size and the orientation of cells.

Due to the fewer deviations in size, we hypothesize that the bounding box regression requires fewer ground truth data samples than the samples required for classification. To validate our hypothesis, we trained RetinaNet [3] with only 5% box-labelled images of the USD. We observed that the model predicts 73% of objects accurately at a confidence score 0.05, even with only a few bounding box annotations.

This paper proposes a weakly semi-supervised method that trains with a few box-labelled annotations. Most cells are annotated with points following the annotation strategy in [4]. The point is placed close to the center of the object.

The existing weakly semi-supervised approaches use two-stage training. A teacher model trains with only box-labelled images to generate pseudo boxes for the point-labelled instances. Afterward, a student model trains with both box- and pseudo-box-labelled images [5, 6]. The performance of the teacher model highly depends on the variations in and quality of box-labelled annotations. Moreover, for two-stage methods, it is challenging to handle datasets having both bounding boxes and point annotations in the same image.

We propose a one-stage training method that jointly trains with both box- and point-labelled images. Points are replaced with a stochastic (ST) box or a prediction from the box regression subnet. The proposed “mean-IOU” ST box utilizes the class-specific joint pdf approximated with box-labelled instances available in the dataset.

Sec. 2 describes a point-to-box selection algorithm and a novel loss function. In Sec. 3, we compare one-stage and two-stage methods on urine and blood datasets.

*Email: hariom85@gmail.com
2. METHOD

In this section, we present a one-stage, weakly semi-supervised learning strategy that utilizes both strong and weak annotations jointly in the form of ground-truth bounding boxes and points, respectively, during training.

We describe our method by considering RetinaNet [3] as a base model. However, the proposed method is applicable to any anchor-based detection models with mild modifications.

RetinaNet consists of a backbone network and two sub-networks; classification and box regression. The backbone computes feature maps. The pre-defined object proposals, called anchors, are defined at each spatial position of the feature maps. A set of anchors are assigned to each ground-truth object box, called foreground anchors using an IOU threshold of 0.5 and to the background if their IOU is in [0, 0.4). The classification subnet predicts the probability of object presence at each spatial position for each of the anchors and object classes. The box regression subnet regresses the offset from each anchor box to a nearby ground-truth (GT) object box.

The anchor assignment, as described above, is one of the crucial steps; wrong assignments lead to lower precision. For point annotations, object boundaries are not available; hence anchor assignment is a complex task. A naive approach could assign an anchor as foreground if it contains the annotated point. However, this would select many anchors that cover spatial regions far beyond the ground-truth object and produce many false positives.

This paper proposes a method to build a bounding box around each point annotation during training. We refer to them as stochastic (ST) boxes because they are defined from a class-specific probability distribution of bounding boxes. These boxes change based on the box regression subnet’s output at every iteration. Our idea is simple to implement and requires minor changes to the base RetinaNet model.

2.1. Point to Stochastic Box

During training, one can utilize the bounding box predictions from the box regression subnet for objects having point annotations. However, there are two major issues that need attention. First, how to handle inaccurate predictions at the initial stage of training? second, how to ensure at least one credible prediction corresponding to each point annotation to avoid considering the object as a background?

We propose algorithm 1 that uses predictions only if they are reliable; otherwise, build a ST box utilizing a class-specific joint pdf \( f_c \) of width \( w \) and height \( h \) where \( c \) denotes a class. We measure \( f_c(w, h) \) only from the box-labelled annotations present in the train set using a kernel density estimator with Gaussian kernel; see Fig. 1 for joint pdf and the estimates of ST boxes. We discuss two choices for ST boxes:

S1) **Mean**: We define a “mean” ST box of width \( \hat{w} \) and height \( \hat{h} \) with their expected values, i.e.

\[
\hat{w} = \int wf_c(w)dw, \quad \hat{h} = \int hf_c(h)dh
\]

where \( f_c(w) \) and \( f_c(h) \) are the marginal pdf of width and height, respectively. The center of the ST box is defined at the annotated point for an object.

S2) **Mean-IOU**: The “mean” ST box does not capture the cross-correlation between the width and height, whereas the proposed “mean-IOU” ST box does. It is defined as a box that has maximum overlap with all the boxes belonging to the sample space of bounding boxes, which is a solution to the optimization problem

\[
(\hat{w}, \hat{h}) = \arg \max_{x \cdot y} \int \text{IOU} (B(x, y), B(w, h)) f_c(w, h) dw dh
\]

where \( B(w, h) \) denotes a box of width \( w \) and height \( h \) whose center is at origin without loss of a generality and \( \text{IOU}(A, B) \) denotes the IOU between box A and B.

From the law of the unconscious statistician (LOTUS), the objective function in (1) measures the expected value of IOU. Therefore, we name the measure “mean-IOU”.

We solve (1) by discretizing the integral on \( 100 \times 100 \) rectangular grid and optimizing with the gradient-free patternsearch algorithm. The initial estimate is set to the “mean” ST box for faster convergence. We further reduce the solution space by setting the lower and upper bound with \( \mu \pm 5\sigma \) where \( \mu \) is the mean, and \( \sigma \) is the standard deviation of width and height variables. We verified that the final estimate converges to the same optima irrespective of any initial estimate.

Note, (1) needs to be solved only once at the start of the training process and it is quick to solve; it took 0.04 sec. (on average) per class for the USD dataset on MATLAB running on the Apple M1 Pro (2021) without GPU support.

Fig. 1. Approximated joint pdf of width and height of bounding boxes of all seven classes in USD-50 with only 20% box-labelled instances and the estimates of “mean” (blue circle) and “mean-IOU” (red star) ST boxes. For highly skewed distributions, e.g., for crystal, mycete, and cast, training with “mean-IOU” tends to produce less false positives than with the “mean”.
Algorithm 1 Point to box selection strategy.

Input: Point annotation \( p = (p_x, p_y) \), mean-IOU ST box \( S \) centered at \( p \), predicted boxes from box regression subnet \( D = \{d_1, \ldots, d_N\} \), \( \tau_{iou} > 0 \), \( \tau_s > 0 \).

Output: a box to replace the point for assigning anchors

1. \( D_t = \{d_i \in D, P(d_i) \geq \tau_s, IOU(S, d_i) \geq \tau_{iou}\} \)
2. if \( D_t = \emptyset \) then
3. \( \text{return } S \)
4. else
5. \( f(d_i) = P(d_i) \exp(-||p - u_i||^2) \) \( \forall d_i \in D_t \), \( u_i \) is the center of \( d_i\)
6. \( d_{\text{best}} = d_n \) where \( n = \arg \max_k \{f(d_k)\} \)
7. \( \text{return } d_{\text{best}} \)
8. end if

2.1.1. Algorithm:

We assume that any credible prediction for an object will overlap with the “mean-IOU” ST box. Hence, a prediction is assigned to a point if IOU with the “mean-IOU” ST box is above a threshold \( \tau_{iou} \); otherwise “mean-IOU” ST box is assigned to the point. In this way, we avoid a point-annotated object being treated as a background during training.

Usually, more than one prediction overlaps with the “mean-IOU” ST box. Hence, we need an extra filtering step that assigns one “best” prediction to a point. Motivated from [4], the “best” prediction maximizes the score

\[
s(B) = P(B) \exp(-||p - u||^2)
\]

where \( P(B) \) denotes the probability of object presence, \( i.e. \) the confidence score from the classification subnet, \( u = (u_x, u_y) \) is the center of the box \( B \), and \( p = (p_x, p_y) \) is the location of point annotation. \( ||p - u||^2 \) measures the distance from the center of prediction to the annotated point.

The above-described scheme might provide unreliable boxes at the initial stage of training. Therefore, we select only high confidence score predictions whose score is \( \geq \tau_s \). In our experiments, we use \( \tau_s = 0.2 \) and \( \tau_{iou} = 0.5 \).

2.2. Loss Functions

Following the RetinaNet anchor assignment strategy for GT boxes, a set of anchors is assigned to selected boxes from Algo 1 with probability one to minimize the cross-entropy loss function to learn the classification subnet’s parameters.

The box regression subnet regresses the offset from each foreground anchor to only nearby GT boxes, not the selected boxes. These boxes are just an approximation; the regression with respect to them performed poorly in our experiments. The annotated points are the ground truth; therefore, the distance from the anchor’s center to the point is minimized.

Point annotations are noisier than bounding boxes [4]; annotating a point at the exact center of the object is a cumbersome task. Therefore, the box regression loss function, denoted by \( L \), downweight the contribution from point annotations by a factor of \( (1 + \beta) \), \( \beta \geq 0 \);

\[
L = \frac{1}{N_g} \sum_{i=1}^{N_g} ||\hat{t}_i - (t_g)_i||_1 + \frac{1}{N_p} \sum_{j=1}^{N_p} ||\hat{t}_j - (t_p)_j||_1
\]

where \( \hat{t} \) are the box subnet predictions, \( (t_g)_i \in \mathbb{R}^4 \) and \( (t_p)_i \in \mathbb{R}^2 \) are the offset between a foreground anchor and the associated GT box or the point annotation, respectively. Note that, \( t_p \) are just the centers’ offset. \( N \) denotes the number of foreground anchors and \( \| \cdot \|_1 \), the L1 norm.

The hyperparameter \( \beta \) is a measure of uncertainty in the points; the lesser the uncertainty, the smaller the \( \beta \). The noise analysis in [4] says that the error distance between the true center and the point annotation increases with the object area. Therefore, we set \( \beta = \alpha/\sqrt{A}, \alpha \geq 0 \), where \( A \), is the mean area of bounding boxes associated with the \( c^{th} \) class.

3. EXPERIMENTS AND RESULTS

We present experimental results on two datasets; USD and BCCD [7]. We randomly select 50% of images from USD train set and form a new set USD-50 for training; the trained models with USD and USD-50 attain almost the same mean average precision (mAP) on the val set. USD-50 has 16,294 cell instances from 7 classes. BCCD training set only has 2805 cell instances from 3 classes. The ground truth class label and bounding box are available for each cell instance. We partition the train set into two mutually exclusive subsets; well-labelled and weakly-labelled sets. In the weakly-labelled set, we replace a bounding box with a point. Note that, only the well-labelled set is used to estimate the bounding box priors, not the entire train set.

Points generation: Following the point-annotations error analysis in [4], we introduce noise to the true center of the object bounding boxes. In USD, 99.9% object’s area, \( A \in [18^2, 198^2] \) pixels for which the mean error-distances \( \mu \) vary from 5 to 17 pixels as the object’s area increases; see Fig.4 in [4]. To mimic this behaviour, we assume that the noise follows a normal distribution whose mean \( \mu \) is a linear function of \( \sqrt{A} \) and the standard deviation is 3 pixels. This setting is highly noisy for small-size cells, \( e.g. \) the annotated center of erythrocyte cells in USD is shifted on average by 40% with respect to the cell’s average width and height.

For training, we optimize parameters with SGD for 50 epochs with an initial learning rate 0.01, weight decay factor of \( 10^{-4} \), and momentum 0.9, on one Tesla V100-PCIE-32GB GPU. The learning rate is reduced when the loss has stopped improving. We obtain the best training stability and generalization performance with a batch size of 8 for USD and 2 for BCCD. All experiments are run at a fixed random seed. The ResNet-50 backbone is pre-trained on Imagenet. Both USD and BCCD have a high-class imbalance; therefore, we use the class-balance loss strategy from [8] with \( \beta = 0.99 \).
We compare the proposed one-stage method with a naive method and a so-called two-stage method upper bound. The naive model trains only with the well-labelled images. We use the naive model to generate pseudo boxes for the weakly-labelled set and train with both well- and pseudo box-labelled set for a two-stage method. We assume that a best-performing first stage of a two-stage method will predict the pseudo boxes whose IOU with the ground truth boxes of weakly-labelled set is higher than 0.5. This assumption is followed to generate pseudo boxes; the second-stage performance is reported as an upper bound for the two-stage method.

The results are summarized in Fig. 2. Two-stage method upper bounds can only be constructed for point annotations without noise. Therefore, we also report results for the one-stage method without noise for a fair comparison between methods. We report results only with mean-IOU ST box.

In the most challenging setting, when only 5% images are well-labelled, and the rest 95% images are weakly-labelled, the one-stage method outperforms the two-stage method by 17.52 mAP without noise and 5.56 mAP with noise for USD; similar trend follows for BCCD. As expected, the performance gap decreases with an increase in the number of box-labelled instances in the training set.

We observed the correlation between the naive method’s mAP and the KL divergence of pdfs; see Fig. 2 for details. Noting that, we advise annotating with bounding boxes until the pdf of them starts to converge, the rest of the objects with points, and using a weakly semi-supervised method.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access - the attached licenses confirm that the ethical approval is not required.

6. REFERENCES

[1] Yixiong Liang, Zhihong Tang, Meng Yan, and Jianfeng Liu, “Object detection based on deep learning for urine sediment examination,” Biocybernetics and Biomedical Engineering, 2018.
[2] H Kenneth Walker, W Dallas Hall, and J Willis Hurst, “Clinical methods: the history, physical, and laboratory examinations,” 1990.
[3] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, “Focal loss for dense object detection,” in International Conference on Computer Vision, 2017.
[4] Dim P Papadopoulos, Jasper RR Uijlings, Frank Keller, and Vittorio Ferrari, “Training object class detectors with click supervision,” in Conference on Computer Vision and Pattern Recognition, 2017.
[5] Liangyu Chen, Tong Yang, Xiangyu Zhang, Wei Zhang, and Jian Sun, “Points as queries: Weakly semi-supervised object detection by points,” in Conference on Computer Vision and Pattern Recognition, 2021.
[6] Shilong Zhang, Zhuoran Yu, Liyang Liu, Xinjiang Wang, Aojun Zhou, and Kai Chen, “Group r-cnn for weakly semi-supervised object detection with points,” in Conference on Computer Vision and Pattern Recognition, 2022.
[7] “BCCD dataset,” https://github.com/Shenggan/BCCD_Dataset.
[8] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie, “Class-balanced loss based on effective number of samples,” in Conference on computer vision and pattern recognition, 2019.