We don't need no bounding-boxes: Training object class detectors using only human verification[2]

Divyanshu Shende
divush@iitk.ac.in
Rahul Tudu
trahul@iitk.ac.in
Department of Computer Science and Engineering,
IIT Kanpur

Object class detection requires training examples to contain bounding boxes. This is typically done by humans and can be a time consuming task. The authors present a scheme in which the algorithm gives a bounding box, human gives feedback ("Yes/No" or "Yes/Part/Container/Mixed/Missed") and the algorithms learns using that feedback.

Previous work was done using Weakly Supervised Object Learning (WSOL)[5, 6] has been used where only the image classes are provided and not the location of object instances, thus the algorithm first re-trains the object detector and then tries to re-localise the object instances using the current detector. Apart from this, active learning[8] is used in which the learner gives the human annotator a set of data points to be annotated as the most informative by drawing many bounding boxes, which is clearly time consuming.

The algorithm presented in [2] iterates on an image. In each iteration, three steps are performed. First a bounding box is presented to the annotator, who gives feedback. The feedback could be simple yes/no feedback or could be more complicated "Yes/Part/Container/Mixed/Missed Feedback" (*YCPMM*). To keep the annotator honest it is assumed that the annotator uses the IoU measure to label. For two images, a and b, $IoU(a,b) = |l_a \cap l_b|/|l_a \cup l_b|$, where l is the detection and $|\cdot|$ denotes area. Once a bounding box is proposed, the annotator gives feedback "Yes" if $IoU > 0.5$ (with the ground truth bounding box) and "No" otherwise. Let D_n be the set of detections in iteration n. Based on feedback, we split it into D_n^+ and D_n^- corresponding to correct and incorrect predictions. We add D_n^+ to the set D^+ which we use to re-train our object detector (and do not consider their images in further iterations). In re-localizing step, we apply current object detector to the set of bounding box proposals and propose the one with the highest score. However, the search space of all proposals has to be reduced. This is where the feedback comes in. Suppose the feedback is Yes/No. Here, in case of negative feedback, we remove all bounding boxes from our current search space which have $IoU \geq 0.5$ with our (incorrect) hypothesis. This reduces the search space for the next iteration. In case the feedback is *YCPMM*, we follow a more complicated procedure. In case of Part, we eliminate non-supersets of our current proposal (by using IoA). In case of Container, we remove all supersets. In case of Mixed, we impose $IoU \geq 0.5$ criteria to refine our search space. The iteration stops when two subsequent iterations yield the same set of hypothesis. This typically happens in the first 10 iterations. This completes a broad description of the method.

For object detection, Fast RCNNs [4] alongwith EdgeBoxes [3] are used to give object proposals. Bounding box regression is skipped in training but used while testing. The underlying CNN architecture is AlexNet [1]. Another important factor is initialization. Using Multiple Instance Learning [7] a set of detectors is obtained. An SVM is then trained with CNN features.

As for evaluation, the dataset used was *PASCAL VOC 2007*. Measure of localization is given by Correct Localization (CorLoc) which is percentage of images in which final bounding box has $IoU \geq 0.5$ with training set. For object detection, measure used is Mean Average Precision (mAP). The algorithm presented is compared with the fully supervised method with labelled boxes and also with MIL WSOL techniques that have no human verification.

It was found that using *YPCMM* verificaton had 96% CorLoc than retraining object detector with removing all overlaps with negative proposal (95%) which inturn gave better CorLoc that retraining with just D_n^+ (82%). All of them fare much better than MIL WSOL whose CorLoc was 43%. It was also observed that *YPCMM* feedback needs fewer verifications that Yes/No feedback, but Yes/No takes less time (by human annotator) as compared to *YPCMM*. For same mAP (45%), Yes/No gave 83% CorLoc in 5.8 hours while *YPCMM* 81% CorLoc in 7.7 hours. It was concluded that Yes/No is better than *YPCMM* for human verification. Thus, the model outperformed the WSOL state-of-the-art techniques with modest cost in terms of time.

In comparison to fully supervised techniques, the mAP (45%) was close to the state of the are (51%) but the reduction in time was substantial. Fully supervised techniques took about 36 hours at minimum. The proposed method took only 5.8 hours so that's a $6-9\times$ savings in terms of time of human annotators with relatively cheap cost in terms of mAP.

In conclusion, it seems to the authors of this report that the method proposed in the paper [2] was thoroughly tested. It is a good method in that it gives great savings in terms of human annotator time while performing almost equally well (in terms mAP) as the fully supervised setting.

[1] I. Sutskever A. Krizhevsky and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *NIPS, 2012*.

[2] Frank Keller Vittorio Ferrar Dim P. Papadopoulos, Jasper R. R. Uijlings. We don't need no bounding-boxes: Training object class detectors using only human verification. *CVPR2016*.

[3] P. Dollar and C. Zitnick. Edge boxes: Locating object proposals from edges. *ECCV, 2014*.

[4] R. Girshick. Fast RCNN. *ICCV, 2015*.

[5] S. Jegelka J. Mairal Z. Harchaoui H. Song, R. Girshick and T. Darell.

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

[7] J. Verbeek R. Cinbis and C. Schmid. Multi-fold mil training for weakly supervised object localization. *CVPR2014*.

[8] K. Grauman S. Vijayanarasimhan. Large-scale live active learning: Training object detectors with crawled data and crowds. *IJCV 2014*, 2014.