Semi-Supervised Object Detection with Sparsely Annotated Dataset

Jihun Yoon
hutom
Republic of Korea
yjh2020@hutom.io

Seungbum Hong
hutom
Republic of Korea
qbration21@hutom.io

Sanha Jeong
hutom
Republic of Korea
jeongsanha@hutom.io

Min-Kook Choi
hutom
Republic of Korea
mkchoi@hutom.io

Abstract

In training object detector based on convolutional neural networks, selection of effective positive examples for training is an important factor. However, when training an anchor-based detectors with sparse annotations on an image, effort to find effective positive examples can hinder training performance. When using the anchor-based training for the ground truth bounding box to collect positive examples under given IoU, it is often possible to include objects from other classes in the current training class, or objects that are needed to be trained can only be sampled as negative examples. We used two approaches to solve this problem: 1) the use of an anchorless object detector and 2) a semi-supervised learning-based object detection using a single object tracker. The proposed technique performs single object tracking by using the sparsely annotated bounding box as an anchor in the temporal domain for successive frames. From the tracking results, dense annotations for training images were generated in an automated manner and used for training the object detector. We applied the proposed single object tracking-based semi-supervised learning to the Epic-Kitchens dataset. As a result, we were able to achieve ***runner-up*** *performance in the* ***Unseen*** *section while achieving the* ***first place*** *in the* ***Seen*** *section of the Epic-Kitchens 2020 object detection challenge under IoU* > 0.5 *evaluation.*

1. Introduction

Thanks to the rapid development of CNN (Convolutional Neural Networks), the performance of object recognition networks using CNN has also been improved dramatically [16]. As the performance of the object detection network has been improved, the dataset for evaluating it was also started from a dataset with low complexity such as PASCAL VOC [8] and developed to have a high complexity such as MS-COCO [15]. Among the object detection datasets, the relatively recently released Epic-Kitchens dataset has the following characteristics different from other object detection datasets [7].

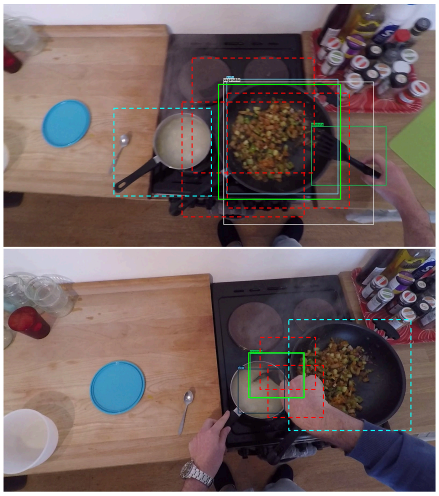

Figure 1. **An example of anchor-based detector training on a sparsely annotated dataset.** The solid green line represents the label information in each training image, and the red dotted line is an example of positive examples. Light blue dashed lines indicate objects that are included in the label in other training images (top), but are not labeled in the given image to train (bottom). As such, in the Epic-Kitchens object detection dataset, it is an object to learn when training an anchor-based detector, but training performance is impaired because label information is missing.

- Images for training detector are collected from the original video, and corresponding frame sequences are provided.
- In a training image, only some of the trainable objects are sparsely annotated.
- The difference in the amount of annotations between the few and many shot classes is large, depending on the distribution of the appearance of the objects in the training dataset.

As described above, the annotation of Epic-Kitchens for object detection is provided in a different way from the existing dataset, and has a characteristic that it is difficult to apply the method of training the existing object detection model as it is. Typically, in the case of detectors that train positive examples based on anchors [21, 17] or detectors