Consistency-based Semi-supervised Learning for Object Detection

Kobee Raveendran

January 13, 2020

1 Summary

In this paper, the authors propose a novel algorithm for semi-supervised object detection. Their method tackles the problem of the annotation of large datasets for object detection. Manually annotating large datasets is expensive both temporally and monetarily. To somewhat automate this process, weakly- and semi-supervised learning methods have been proposed recently. However, such previous methods either exhibited poor performance on the localization task or were too slow and inefficient to train. To rectify these problems, the authors propose their method, which builds upon a previous method that utilized consistency regulation for image classification, and extend it to object detection.

Additionally, the authors propose a novel method for excluding bounding boxes in which the "background" label is prominent when computing the consistency losses, as they state that inclusion of the background in the computation often hurts performance.

2 Good points

The strong points of this paper are the adaptability within the semi-supervised space. Their method of consistency constraints have been shown to work well in both single- and double-stage detectors, which is surprising considering that one would imagine losses to be heavily impacted by the addition of an RPN and a more extended network (in double-stage detectors). Additionally, I found it impressive that the authors managed to make a single loss function work well with both object classification and bounding box regression. The overall consistency loss is simply a sum of the classification and regression losses, though there seems to be no performance inhibition incurred from simply combining them, which is surprising given that previous methods often mandated the use of separate loss functions (as they were separate tasks).

3 Weak points

One of the weak points of the authors' method is that it has poor adaptability when taken out of the domain of semi-supervised learning. As the authors discuss in their Discussion section, their model experiences performance degradation when it is fed a dataset with **only** ground truth labels. Additionally, the effect of excluding "background" candidate boxes does aid the method by improving the consistency losses, however as mentioned by the authors themselves, it is believed that by masking such boxes (which can often make up a significant portion of an image), learning can be hampered, as shown by their results of using SSD300 on the PASCAL VOC dataset. Additionally, the results presented in Table 1 are very minimal and almost negligible improvements, though still impressive given their semi-supervised nature.

4 Questions

One part of the paper I didn't completely understand was how the flipped horizontal image inputs would be handled. The authors say they do this to enforce consistency of bounding box placement, but would this not result in different bounding box locations if the target object is in a non-centric position in the image, or is this already accounted for?

5 Ideas

I noticed in the architecture diagrams for the single-stage and double-stage CSD models that the region proposal network was only used for the original image input, rather than for both the original and flipped inputs. In the discussion section, the authors mention that CSD has a limitation when used in two-stage detectors, which makes me wonder if applying separate RPNs to both the original and flipped images could help improve performance for double-stage detection.