Open-Vocabulary Object Detection Using Captions
Supplementary Material

Alireza Zareian[1,2], Kevin Dela Rosa[1], Derek Hao Hu[1], Shih-Fu Chang[2]
[1] Snap Inc., Seattle, WA [2] Columbia University, New York, NY
{azareian, kevin.delarosa, hao.hu}@snap.com {az2407, sc250}@columbia.edu
github.com/alirezazareian/ovr-cnn

In this section, we provide statistical and qualitative analysis to gain additional insights about the performance of the proposed method. Since one of the most critical issues of deep learning is bias, we start by analyzing the effect of training data bias on our per-class performance. Since we have two training phases, the class frequency during pretraining and downstream training should be separately analyzed. Figure 1 shows our per-class performance (right), along with the frequency of bounding box instances during downstream training (left), and the frequency of words during pretraining (center).

Our first observation is that our performance is not affected by the bias in downstream training data. As we move down the list, classes become exponentially less frequent, but the performance does not drop at all, except target (red) classes which have exactly zero examples during downstream training, and are inevitably less accurate. Our robustness to data bias is most likely due to the fact that we fix the classification head during downstream training, including both the V2L layer and the class embeddings. This is in contrast with conventional classifiers which fully adapt the classifier parameters, including an explicit *bias* term, to the biased training data.

Nevertheless, when we compare the performance to word frequency during pretraining, we do observe a correlation between the least frequent words and the lest accurate classes. This correlation is not very strong, but it motivates our future work on bias mitigation mechanisms that can be used in naturally supervised (image-caption) settings.

Furthermore, we observe that smaller objects such as `knife` and `tie` have lower performance, which is to some extent consistent with supervised object detection, but is fueled by the fact that our grounding mechanism is weakly supervised, and is less likely to correctly align smaller objects to words, because they take a smaller portion of the feature map.

To get a qualitative look at the performance, we deploy our model on the COCO validation set and visualize its detection outputs in Figure 2. We use the generalized version

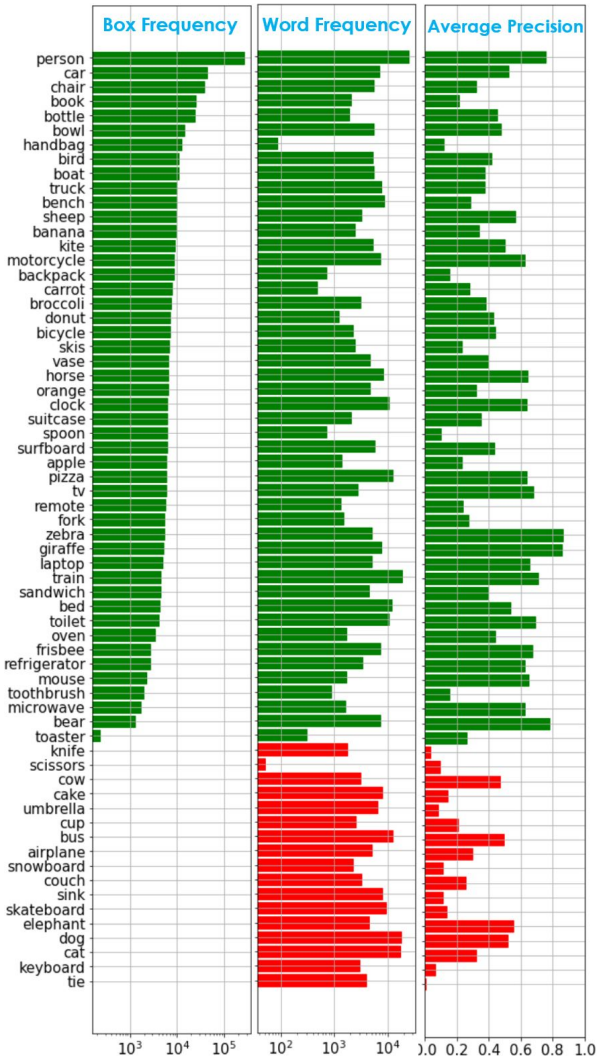

Figure 1. Performance for each class along with data frequency during pretraining and downstream training. Green and red show base and target classes respectively.

Figure 2. Qualitative results of our OVR-CNN model, detecting both base and target classes. Target classes are shown with larger font, thicker border, and uppercase.

which selects the category of each object from the union of base and target classes. We emphasize target classes for better visibility, and analyse the quality of the predictions. Based on our observation, the main limitation of our method is localization accuracy for target classes. There are several cases of overly loose or overly tight bounding boxes, which is due to the fact that we have no ground truth bounding boxes for target classes. This motivates future work on class-agnostic boundary refinement.