Bounding out-of-sample objects

A weakly-supervised approach

Li Quan Khoo
lqkhoo@stanford.edu
Stanford Center for Professional Development

Abstract

Convnets, by design, have spatial awareness built into their architectures. In the context of image processing, one the most salient parts of the image tend to trigger the largest activations.

Our goal is to train a network to perform out-of-sample object bounding, on entire classes of images that do not have bounding box information. For example, we might train a network to put bounds around cats, but we also want it to be to be able to put bounds around dogs, too, without being trained beforehand to do so. This means the network has to be indifferent to the class label, and operate solely on spatial information, which, one might hope, generalizes sufficiently across samples in the universe of all images in order for this method to work.

Method

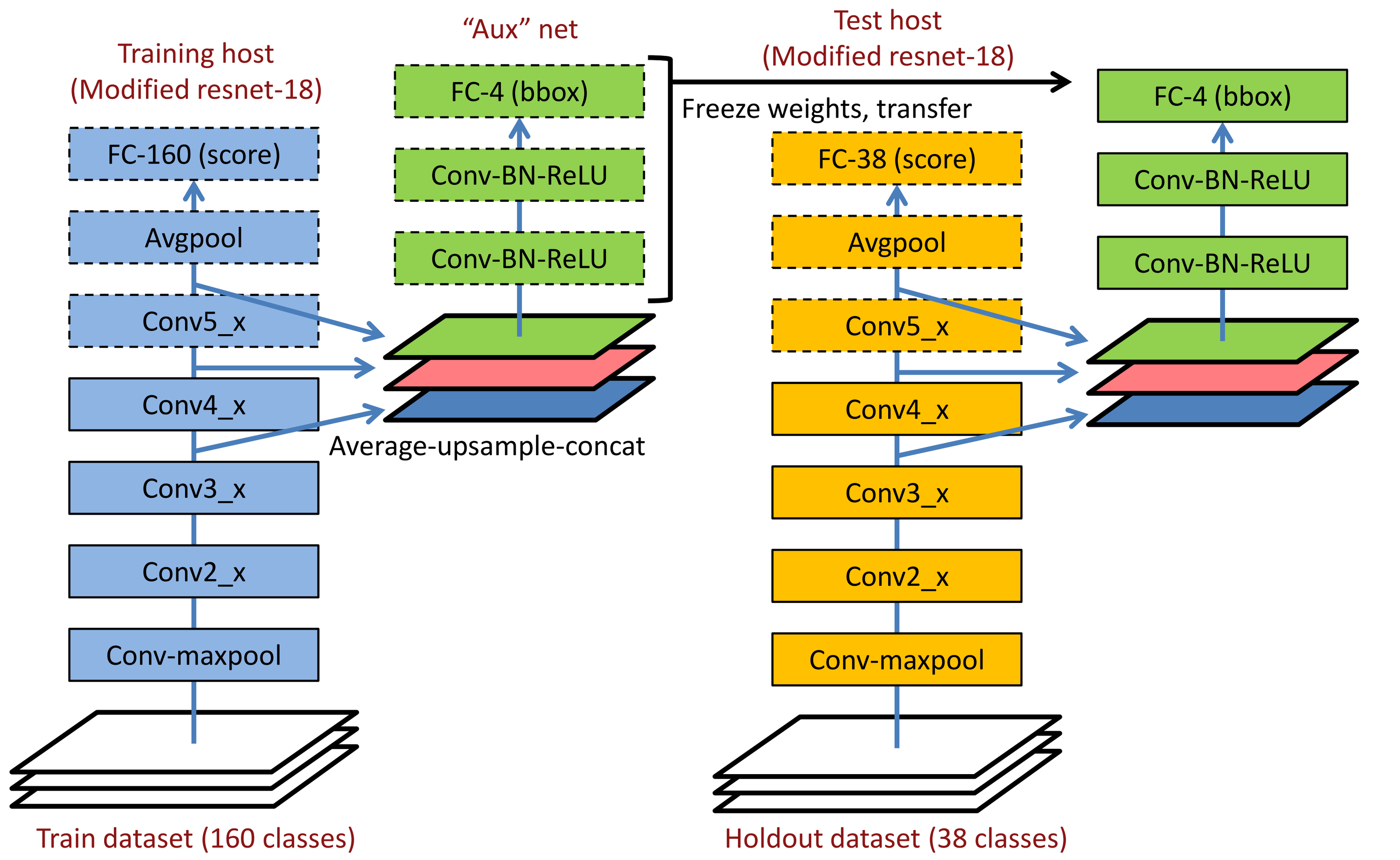
1. Train a Resnet-18 to classify images. This is the *training host.*
2. Freeze host's weights, train the *Aux network* which outputs bounding boxes given a subset of host network's weights. This requires ground-truth bounding boxes.
3. Train a new Resnet-18 to classify images on the holdout dataset. This is the *testing host.*
4. Attach the trained *Aux network* to the testing host and evaluate.

Data

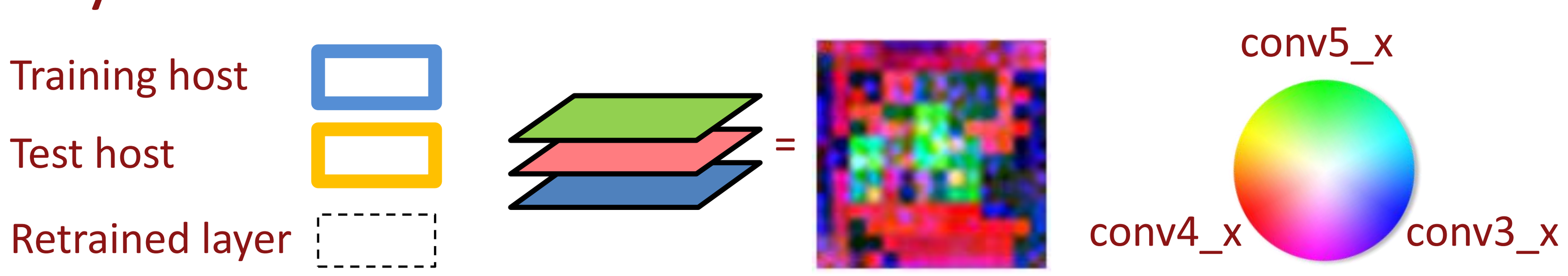
198 ImageNet synsets with named labels, containing at least 400 images in 3-channel RGB at least 224 pixels in each dimension, each associated with one bounding box.

Train dataset: 160 most populated synsets, 50 validation images per synset
Holdout dataset: 38 remaining synsets, 50 validation images per synset
N = 115,064 images across both datasets

Architecture



Key



Validation performance

We evaluate the classifier using top-1 accuracy. For bounding box annotations, we evaluate using both the intersection-over-union (IoU) and CorLoc summary statistics. CorLoc is defined as the percentage of images with IoU >= 0.5

We create two baseline models to evaluate against, which consist of both the classifier and the aux network trained end to end (the Aux in a weakly-supervised manner).

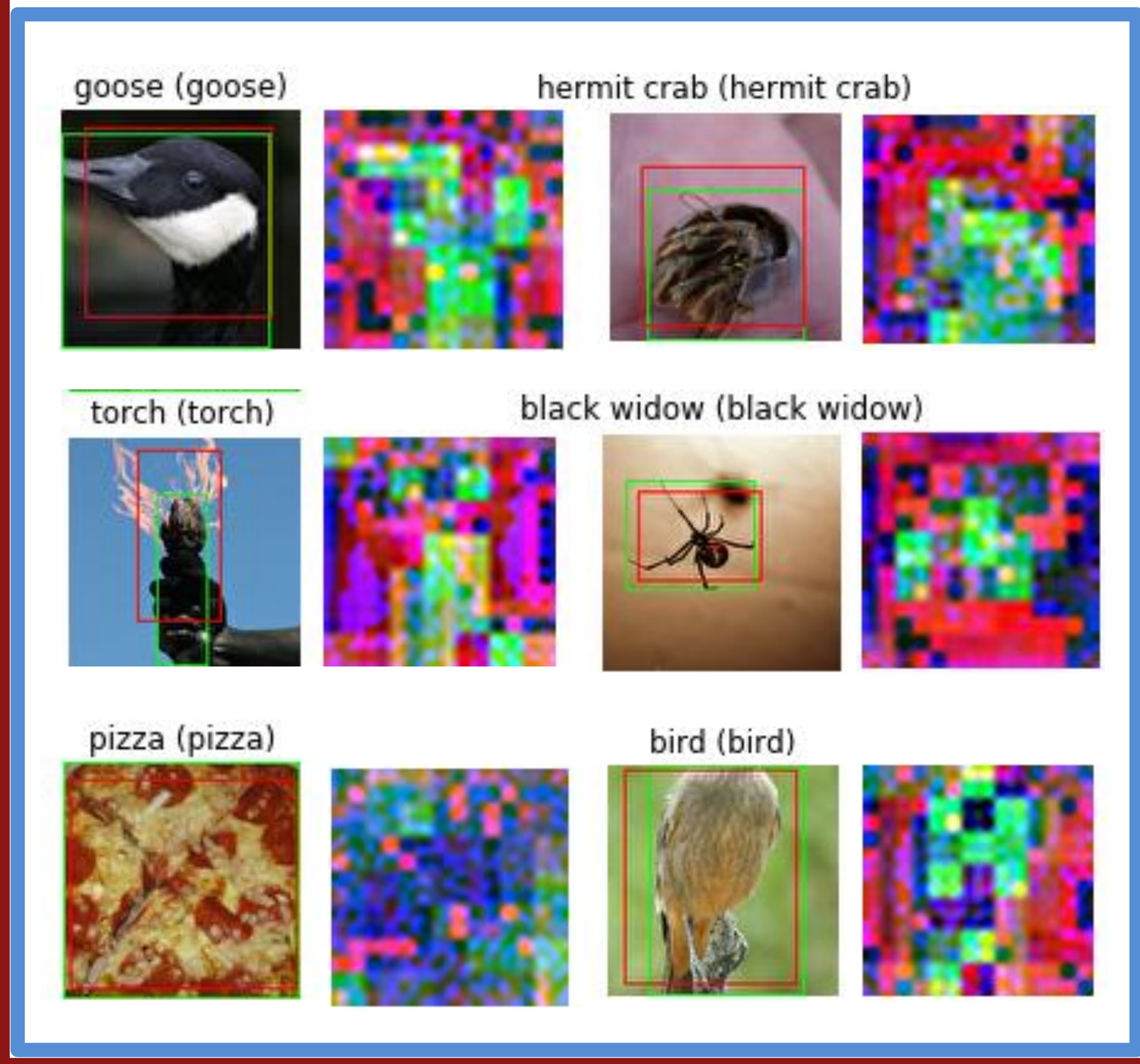
		Top-1 acc	Mean IoU	CorLoc
Baseline 1	Training host + Aux	0.80	0.555	0.463
Baseline 2	Test host + Aux	0.82	0.512	0.403
Model	Test host + Transferred Aux	Same as above	0.511	0.399

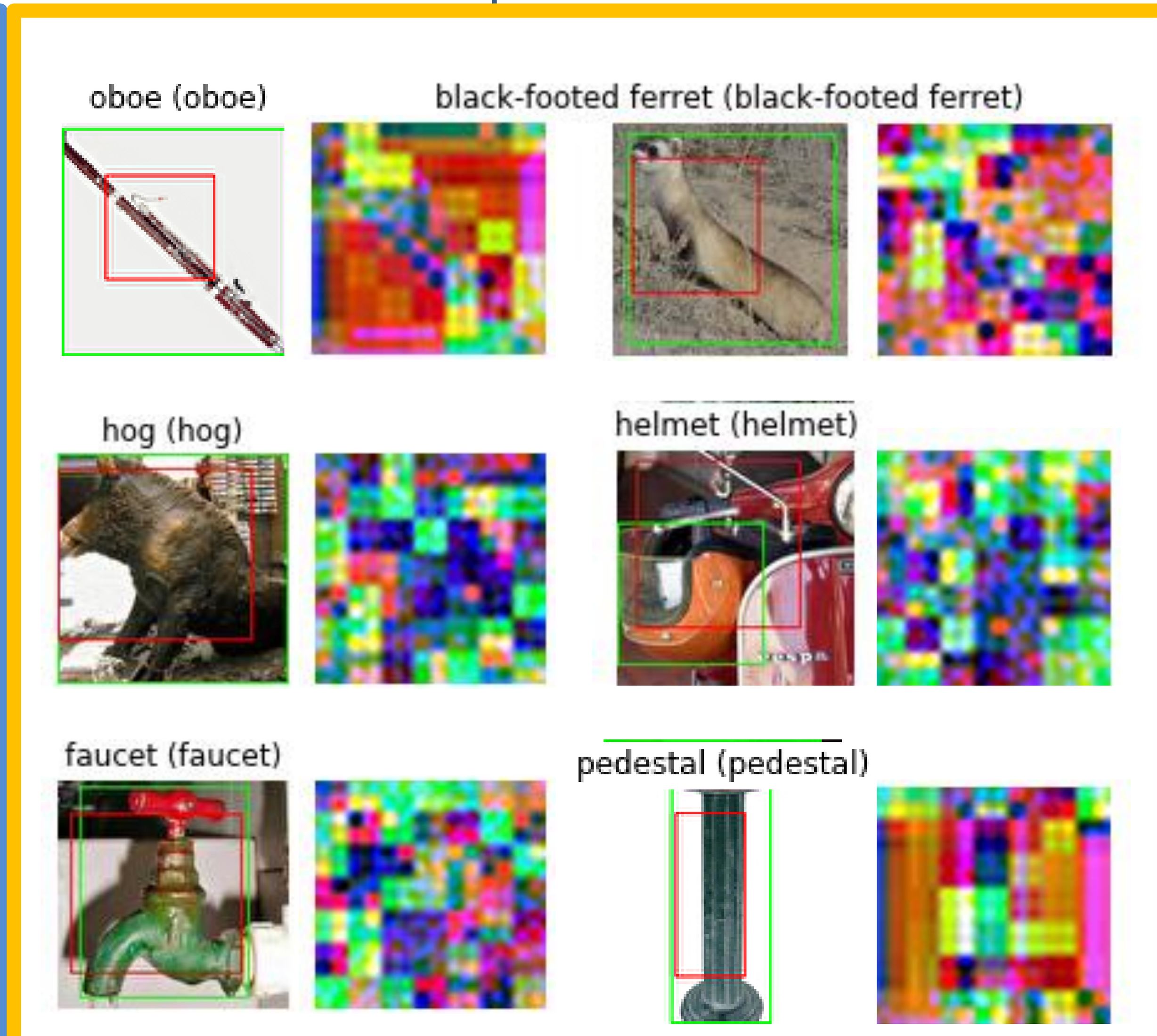
In comparison, the state of the art model in fully-supervised localization on 1,000 ImageNet categories (N=1.2M images) has a CorLoc of 0.923, as of ILSVRC 2017. Semi-supervised models tend to have roughly half that performance [3], which is about 0.46.

Findings and further work

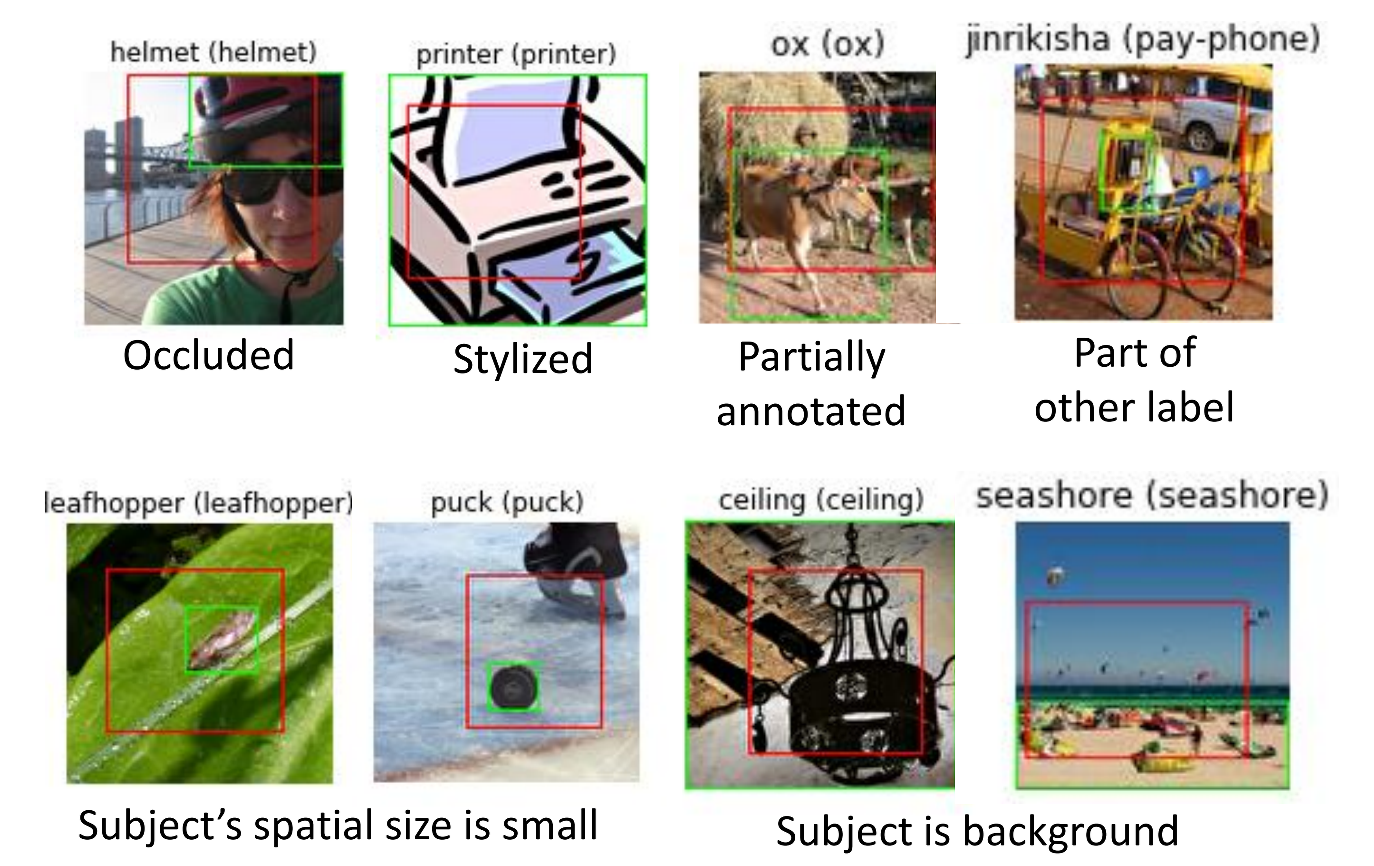
We showed that in our setting, the bounding boxes annotated by the transferred aux network is competitive with a weakly-supervised model trained end-to-end. Averaging Resnet-18's feature volumes also seems to give better performance in terms of IoU than using the maximally-activated layer.

The Aux net's architecture is general enough to try transfer learning *across* models, e.g. trained on a ResNet and transplanted onto a modified VGGNet. It would be interesting to confirm the hypothesis that, as long as its inputs have the same spatial size, channels, and are derived from batchnorm outputs, the transferred model would have comparable performance to a weakly-supervised one trained end-to-end.





Model does poorly when subject is:



Bibliography

[1] J. L. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. context, 2(l3):l4.
[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 770–778, 2016.
[3] D. P. Papadopoulos, J. R. Uijlings, F. Keller, and V. Ferrari. We don't need no bounding-boxes: Training object class detectors using only human verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 854–863, 2016.
[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. 2014.
[5] E. W. Teh, M. Rochan, and Y. Wang. Attention networks for weakly supervised object localization.
[6] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In International Conference on Machine Learning, pages 2048–2057, 2015.
[7] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In European conference on computer vision, pages 818–833. Springer, 2014.