Assignment 5

Neural Networks and Deep Learning

CSCI 5922

Fall 2017

Student: **Bo Cao**

bo.cao-1@colorado.edu
boca7588@colorado.edu

Paper: **YOLO9000: Better, Faster, Stronger** https://arxiv.org/pdf/1612.08242.pdf

General Comments:

1. With 19 convolutional layers and 5 maxpooling layers, the the architecture of Darknet-19 is so simple that other people could easily replicate the experiment. I liked the idea of training on data not just from one dataset, but also from different datasets even with various scales such as different resolutions, like Detection datasets having only common objects and general labels while Classification datasets and ImageNet having different depths and labels. In addition, I also liked how the datasets were combined with WordTree hierarchy.

2. In terms of its previous research of <<You Only Look Once: Unified, Real-Time Object Detection>>, I liked the idea of instead of feeding the whole image into convolutional layer, it let part of the whole image feed this layer to detect objects. The activation function used leaky rectified linear activation. Compared to Fast R-CNN, YOLO eliminated background detections. <<Network In Network>> replaced the fully connected layer with global average pooling to boost performance. Likewise, YOLO9000 removed the fully connected layers from YOLO and used ancher boxes to predict objects. Comparatively, <<SSD: Single Shot MultiBox Detector>> used offsets in bounding box locations with separate predictors.

Critique & Limitations:

1. Classification datasets still requires labels as well as the anchor boxes, which is labor-expensive to have these datasets.

2. After training, not sure if the network still performs well when adding new datasets with uncommon object labels.

Inspirations

1. Two paralleled networks with convolutional network whose number of layers is less than 19. The number of layers of each network does not necessarily to be exactly the same. We might need to do some experiments with different settings to see how the difference of number of layers could affect the performance. The purpose of this is to make the detection and classification faster. When these two networks are trained, let's assume network 1 is for detection and network 2 is for classification. Both of these two networks will be fed with data simultaneously. The output of network 1 could be the location of the centroid of the objects in the picture. The output of network 2 could be softmax output with WordTree hierarchy. Then as a whole of the 2 networks, a simple and fast mechanism should be designed to output the result such as only output the shared object from both networks.

2. Multi Softmax layers. I liked the multiple softmax operations over co-hyponyms (Figure 5). How about trying to have a softmax operation on a higher and more abstract such as common categories first, and then have a second softmax operation over co-hyponyms.

3. A transferable common neural network for images with common knowledge in the field, so that people can design and build their specific network based on this common neural network.

My Questions

In figure 2, I am not sure how k = 5 is determined as a good tradeoff for recall vs complexity. By empirically? In page 5, the author mentioned they used standard data augmentation like random crops, rotations. I am not clear about how this was integrated into the the main architecture in Table 6 Darknet-19 and how this improved the training specifically. I assumed that this data augmentation was done at the beginning of the training.