Review on “YOLO9000: Better, Faster, Stronger”

Authors: J. Redmon, A. Farhadi

# Short Summary

This paper investigates improvements to the YOLO object detection system to yield an improved model called YOLOv2. These improvements include: adding batch normalization on convolutional layers, fine-tuning the initial classification network on higher resolution data before tuning the network for detection (high resolution classifier), adding anchor boxes in place of fully connected layers, using k-means to obtain good training set bounding box priors (tackles anchor box issue #1), predicting coordinates relative to the grid-cell (tackles anchor box issue #2), adding a passthrough layer that brings higher-resolution features from an earlier layer to access fine-grain features, and down sampling training data to predict well across a variety of different resolutions (multi-scale training). These improvements are shown to cumulatively improve the performance in object detection.

To increase the speed of the model, a new classification model called Darknet-19 is used as the base of YOLOv2. This model is heavily inspired by the VGG but includes batch normalization, global average pooling for predictions and 1x1 filters to compress feature representations between 3x3 convolutions. This new model reduces the number of operations required for a single pass over an image from 30.69 billion in VGG-16 to 5.58 billion. Varying image resolution, YOLOv2 is shown to demonstrate state of the art performance with a high FPS.

The paper then builds upon this improved model by proposing YOLO9000, a result of a joint training algorithm that allows training of object detectors on both detection and classification data. This is significant because it leverages the vastly more labelled data available for classification tasks. However, in order to facilitate the algorithm, the authors describe the implementation of a hierarchical classification model referred to as WordTree. This model takes advantage hyponym relations defined in WordNet to join labels in a tree-like structure based on structural concepts in language. Consequently, the model can effectively take advantage of training data in different types of computer vision datasets. Utilizing the WordTree, the joint training algorithm can intelligently backpropagate loss by the type of image (classification or detection) and the corresponding level of the label. As a net result of this model, the YOLO9000 is able to detect over 9000 object classes in real-time, including object classes where it has not seen any labelled detection data. The model scored a 19.7 mAP overall and 16.0 mAP on the 156 classes it had not seen in the ImageNet detection task.

# Main Contributions

* Proposed numerous improvements to the base YOLO system, that improve classification performance and speed, to yield YOLOv2
* Proposed the Darknet-19 network architecture upon which YOLOv2 is built
* Proposed a hierarchical classification model WordTree capable of merging image classification and detection datasets and thus closing the gap between available dataset sizes for the tasks
* Proposed a joint training algorithm that can be simultaneously trained for image classification (expanding the number of classes) and detection (bounding box coordinate prediction)

# High-Level Evaluation of Paper

Immediate features of this paper that stood out were the writing style, creative structure outlining the goals, and the name of the final algorithm. This made the paper a little bit more entertaining to read and easier to digest. Furthermore, the proposed hierarchical model and corresponding joint training algorithm seems like significant contributions object detection; ostensibly, both can be applied outside of YOLO9000 to further the field. In terms of the YOLOv2 algorithm, the paper did an excellent job summarizing the numerous modifications being made to the original algorithm, their theoretical benefits and quantitative benefits. However, I think reorganizing the paper to discuss the proposal of the Darknet-19 network architecture prior to delving into results would have made more sense in terms of understanding the changes that influenced the YOLOv2 results. While the paper does an excellent job describing WordTree, the contrasting models visualized in Figure 5 do not maintain the same labels when illustrating differences. This may lead to unnecessary confusion. The evaluation section on the final model, YOLO9000, is also very brief in comparison to the development of YOLOv2.

# Discussion on Evaluation Methodology

The paper outlines many other detection frameworks and their mAP scores on VOC2007 as baselines for quantifying YOLOv2. This list includes Fast R-CNN and its variants, YOLO, and SSD plus its variants. This is helpful in terms of understanding how YOLOv2 performs, and further describes the impact of changing the image resolution on overall performance. The inclusion of an FPS comparison also helps demonstrate the trade-offs between speed, accuracy and the overall quality of the model. This testing is also extended to the COCO test-dev 2015 dataset. Although augmentation strategies are claimed to be similar to baselines, this is a source of potential noise in the model comparison. Oddly, perhaps the biggest contribution in the paper, the final YOLO9000 model is only evaluated on the ImageNet detection task and isn’t compared to any other works. I am also not clear on how well the YOLO9000 model would practically perform if detecting 9000 object categories. As mentioned in the paper, there are subsets of categories where the model fails. It is not clear what proportion of the 9000 classes the model has actually learned well.

# Possible Directions for Future Work

Now that a way to apply image classification data to other computer vision tasks exists, it would be interesting to see how the joint training algorithm can be improved or applied to algorithms other than YOLO. As mentioned in the paper, one possible direction is to apply these techniques to object segmentation.