ZoomNet and Lean Fast R-CNN: Speeding up Single Object Detection

Bijan Fakhri, Meredith Moore

Abstract—In this project we attempt to trim the complexity associated with a many-class problem and turn it into a two class problem (presence or absence of the object of interest), creating what we call Lean Fast-RCNNs, Lean FRCNN and ZoomNet. Zoom. Our second approach is to create a novel neural network architecture that ouputs a 3D vector corresponding to the offset of the object of interest to the center of the image. ZoomNet essentially starts by looking at the whole picture, and iteratively looks at smaller windows until the scale of the object of interest (the z component of the 3D vector corresponding to depth), it maximized.

Index Terms—Regional Convolutional Neural Networks, Fast-Regional Convolutional Neural Networks, Pedestrian Detection,

1 Introduction

REGIONAL convolutional neural networks (R-CNNs) and its derivatives represent the state-of-the-art in object detection techniques in terms of bounding box accuracy [1]. Currently, R-CNNs take a very long time (relative to other methods) to produce bounding boxes of objects of interest [2], [3]. There have been several incremental improvements to the R-CNN framework which make them perform more efficiently, however they still have significant computational limitations [4], [5]. This makes it very difficult to use for cyber-physical systems.

Real-time applications where only a single type of object is of interest (ie. birds or pedestrians), are thus unable to take advantage of the increases in accuracy afforded by R-CNN and its derivatives. We propose a two-prong strategy to attack this problem. We will take R-CNN, and trim the complexity associated with a many-class problem and turn it into a two class problem (object of interest is present, or not present), creating what we call Lean Fast-RCNN (Lean-FRCNN). Our second approach is to create a novel neural network architecture that outputs a 3D (x,y,z) vector corresponding to the offset of the object of interest to the center of the image. The first two components of the vector (x, y) represent the offset in the traditional dimensions of the image (height, width). The third component (z) represents the offset in the 'depth' dimension. In other words, it represents the amount of 'zoom' we must apply to the frame in order to center the object of interest as well as have it fill the frame completely. We call this ZoomNet. ZoomNet is designed to be used in an iterative manner, where the network is given a portion of an image (the subimage) and the network returns the 3D offset vector. The subimage is shifted and scaled according to the offset vector, and the new image is sent into the ZoomNet. In this way, the tradeoff between computational intensity and IoU can be directly managed.

Several papers have looked at the application of R-CNNs to a popular two class object detection problem—pedestrian detection[6][7]–[11], however we are looking to futher decrease the bounding box computation, while maintaining a sufficiently high bounding box accuracy, and minimizing the network memory usage.

2 Research Question

In developing these two approaches, we are looking to answer if it is possible to retain the accuracy of object detection afforded by R-CNNs while limiting it to a domain specific twp-class problem. We are also curious as to how we can increase the hspeed of R-CNNs using domain specific constraints that allow for the problem to be reduced to a two-class classification problem for use in cyberphysical systems.

3 Experiments

To test the efficiency and accuracy of ZoomNet and Lean-FRCNNs, we will first compare ZoomNet to FR-CNNS. Then we will compare lean-FR-CNNs to Faster R-CNNS, and then compare Lean FR-CNNs to ZoomNet. We will test the performance of these algorithms will be measured using F1 scores, bounding box computation time, network memory usage, and bounding box accuracy (average IoU).

The algorithm that performs the best will then be implemented into a cyber-physical system.

TABLE 1
Proposed Timeline

Due Date	Task	Partner Responsible
3/6/17	Decide on a dataset to use	Meredith and Bijan
3/7/17	Install Girshick's "Faster R-CNN"	Meredith
3/7/17	Create 3D vector labels for bird dataset	Bijan

3/8/17	Decide on Architecture for 3D Vector Net	Bijan
3/10/17	Train/Test Faster R-CNN on Dataset (Baseline)	Meredith
3/17/17	Implement, Train/Test Scale-Aware R-CNNs on dataset	Meredith
3/17/17	Train/Test 3D Vector Net	Bijan
3/24/17	Create lean-FR-CNN by simplifying the best performing R-CNNs to work on ARM	Meredith
3/24/17	Couple 3D Vector Net with Iterative Localizer	Bijan
3/31/17	Compare against (state-of-the-art) Fast R-CNN	Bijan
4/14/17	Compare against lean-FR-CNN	Meredith and Bijan
4/20/17	Draft paper and presentation	Meredith and Bijan
	If Time Permits: Implement best performing algorithm on ARM (RPi3)	Meredith and Bijan
4/21/17	Project is Due	
4/28/17		

REFERENCES

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *ArXiv13112524 Cs*, Nov. 2013.

[2] J. Yan, X. Zhang, Z. Lei, S. Liao, and S. Z. Li, "Robust Multi-resolution Pedestrian Detection in Traffic Scenes," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3033–3040.

[3] D. Park, D. Ramanan, and C. Fowlkes, "Multiresolution Models for Object Detection," in *Computer Vision – ECCV 2010*, 2010, pp. 241–254.

[4] R. Girshick, "Fast R-CNN," presented at the Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 91–99.

[6] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware Fast R-CNN for Pedestrian Detection," *ArXiv151008160 Cs*, Oct. 2015.

[7] W. Ouyang and X. Wang, "Joint Deep Learning for Pedestrian Detection," in *2013 IEEE International Conference on Computer Vision*, 2013, pp. 2056–2063.

[8] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian Detection with Unsupervised Multi-Stage Feature Learning," *ArXiv12120142 Cs*, Dec. 2012.

[9] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3258–3265.

[10] Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian Detection aided by Deep Learning Semantic Tasks," *ArXiv14120069 Cs*, Nov. 2014.

[11] Y. Xu, T. Xiao, J. Zhang, K. Yang, and Z. Zhang, "Scale-Invariant Convolutional Neural Networks," *ArXiv14116369 Cs*, Nov. 2014.