Improvement for Fast Object Detection Based on Regression Method

Jing Shi
University of Rochester
Rochester 14623
j.shi@rochester.edu

1. Introduction

Object detection for 2D image based on convolutional neural networks(CNN) has witnessed a conspicuous development in recent five years. Since a famous frame for CNN to achieve image classification, location and detections is developed by Overfeat [19], a main branch for object detection based on region proposals has gradually come into being. Inspired by such work, Ross Girshick *et al*. [6] established the architecture for region based CNN(R-CNN), whose contribution including initializing the CNN by pre-trained parameters on a large auxiliary dataset, and fine-tuning the CNN to fit our target dataset, employing SVM for classification, using regression to modify the location of bounding boxes. However, R-CNN warped each proposal so as to meet the input sizes of fully connected (fc) layer. To remedy this, SPPnet [8] developed spatial pyramid pooling layers which can map different size of feature maps into uniform ones. In addition, SPPnet computed a convolutional feature map for the entire input image and then classifies each object proposal using features extracted from the the portion of the shared feature map inside the proposal, allowing extraction of high level feature on proposal windows much faster. Built upon this work, Fast R-CNN [5] simplified the spatial pyramid pooling layers to region of interest(RoI) pooling strategy and integrated classification and location into one multi-task loss and thus realized training in single-stage except for nominating region proposals. Nevertheless, one main issue of Fast R-CNN was that it relied on selective search [23]. Such kind of traditional methods [15][3][16] attempted to generate region proposals by merging multiple segments or by scoring windows that are likely be included in objects, usually time-consuming, rendering region proposals generation step consumes as much running time as the detection network. To combine object proposal and detection into a unified network, Faster R-CNN [18] added two additional convolutional layers called region proposal network(RPN) on top of traditional ConvNet output, to compute high qualified proposals and shared features with Fast R-CNN.

Many types of variants of faster R-CNN were researched latter because a close scrutiny of those work reveals that the R-CNN series left huge space for improvement. One issue of the faster R-CNN was that the last layer output of a very deep CNN is too coarse for classification of some instances with small size, Hence a good object detection system should include features from both deep and shallow layers [20][12][13][7][10]. In order to combine both worlds, HyperNet [10] concatenated the features sampled from feature maps in various layers to one single output cube that contained information and resolution from multi-layers. Aside from such modification, Inside-Outsie Net [1] further made use of spatially varying contextual information by adding four spatial Recurrent Neural Networks (RNNs) layers that move in the cardinal directions and exhibited better results. Another major change is Fully Convolution Network (FCN), not only demonstrated effective on semantic segmentation task [13][7], but also more efficient in computing. However, da detector that based on FCN usually struggled with location inaccuracy because the higher layers were position insensitive, which conflicted with translation-variant property required by precise location. R-FCN [2] addressed this problem by introducing position selective pooling layer that shepherd the last convolution layer to learn the position information of each feature map channel. With the better features of image produced by ResNet[9], R-FCN was both fast and accurate for object detection. From another perspective, LocNet [4] further boosted the localization accuracy by introducing more detailed description for bounding box coordinates regression function. To be specific, it divided a region of interest into regular grids, and assigned a probability to each row and column of the search region for being the left, fight, top, or bottom borders of the bounding box.

Contrary to the region-based method, another kind of method was proposed based on one-step regression, where both locations and classes of the bounding boxes were simultaneously obtained without region proposals [22][17][14][11]. YOLO [17] has launched a tentative work on the possibility of the one shot regression idea and received relatively accurate but extremely fast detection results. SSD [11] modified the coarse grids in YOLO and combined the region proposal idea in RPN [18] that

instructed the selection for default bounding boxes. Specifically it sampled both the deep and shallow features and correspond to anchors with different scales and aspect ratios. SSD made the regression-based detection method practical due to its fast speed and high accuracy. G-CNN [14] modeled object detection as finding a path from a fixed grid to boxes tightly surrounding the objects, and slacked the regression process to several iterations for the reason that one step regression cannot handle the nonlinearity of the coordinates of bounding boxes. Although G-CNN was independent with region proposals, it sacrifice its speed by iterating the bounding box regression. Therefore, as we can see, the regression based detection method is less developed compared with region-based one. In fact, region based methods plays a dominant role in detection accuracy currently, but the one shot regression detection method is more suitable for real time detection due to its speed advantage.

2. Research Issue

Based on aforementioned discussions, the defeats of object detection method still remain unsolved. The main limitations of once-regression based method includes:

1. Difficulty in small-size object detection and precise localization.
2. Limited recognition in new or unusual aspect ratios.

In this case, how to explore a good way to tackle these obstacles still require further research.

3. Research Evidence

3.1. Identify different aspect ratios

It is obvious that the once-regression method encounters the problem that the default boxes [11] must be defined manually. But when look back at the RPN that can firstly regress the object proposals, it is not hard to find that the region based method equals to do twice regression, once for RPN, once for detection net. And G-CNN [14] regress the predefined grids in multiple times to approach the precise bounding boxes because the author argues that it is hard to regress the nonlinear problem in one step. Thus it remains hard to solve this problem for once-regression method. However, we attempt to improve the ability to detect object with unusual aspect ration in region-based method, namely, the RPN. The way to get the features within default boxes projected on feature maps is to use convolution kernel to sliding on each location of the projected area on feature maps [18][10], However, their kernel are always square. Object with unusual aspect ratio may not be obtained only by square kernels, because the empirical perceptive field corresponding to a cell of feature map is smaller than its theoretical perceptive fields, in this case, even if the perception fields are large, the cell can only see a small portion of it and a square perception field may fail to detect unusual aspect ratio. A heuristic way to get rid of this quandary is to leverage kernels with different aspect ratios, e.g. 3*3, 1*3, 3*1 on the feature maps, thus can get better object proposals.

3.2. Locate more precise

Two ideas to enhance the location precise:

1. **Combine information from multiple layer**. HyperNet [10] embodies this idea obviously by concatenating features extracted from feature maps in different layers to one feature tube.

2. **Reinforce the higher layers to learn location information**. This idea is incardinated in R-FCN [2], where a position-sensitive pooling layer is added after the last convolution layer to offset the blur of position information on upper layers. Differently, Inside-Outside Net [1] absorb this idea by embedding four RNNs to depict the position context.

It is intuitive for us to apply these ideas to one-regression method. While SSD makes use of information from multiple layers by adding more feature extraction layers, the features it extract are built on very high layers, which are less sensitive to position translation. In addition, SSD assign feature maps in each layer with default boxes just in a fixed scale. E.g. the lower layer features predicts boxes with small scale and higher layer features predicts large ones; thus it does not combine both high layers and low layers together to predict boxes in various scale. Therefore, it is supposed to lead better performance for once-regression method to concatenate the feature from multiple layers and adding a position sensitive layers.

4. Methods

My approach only regresses for once, which is based on a feed-forward convolutional network that produces a fixed-size collection of bounding boxes and scores for the presence of object class instances in those boxes, followed by a non-maximum suppression step to produce the final detections. The early network layers are based on a standard architecture used for high quality image classification (such as VGG [21] or ResNet [9], truncated before any classification layers). We then add auxiliary structure to the network to produce detections with the following parts:

Default boxes and aspect ratios: We set a series of associate default boxes with different scale and aspect ratios across the whole image, similar to SSD [11]. Each default box will generate k boxes sharing an center location, correspond to which we further regress the box offsets $\delta cx, \delta cy, \delta w, \delta h$ as well as the possibility of an object's existence.

Pooling from multiple layers: We add a max pooling layer on the lower layer to carry out subsampling. For higher layers, we add a deconvolutional operation to conduct upsampling. A convolutional layer (Conv) is applied to each sampled result. We normalize multiple feature maps out of such Conv layer using local response normalization (LRN) and concatenate them to one single output cube, called Hyper Feature. The proportion of projected from the default boxes in Hyper Feature maps are then send through RoI pooling layer to the final fully connecting (FC) layer.

Loss layer: The FC layer is followed with two sibling output layers for each box, one for prediction the location offset and one for objectness score, just like Fast R-CNN [5]. Similarly, we minimize a multi-task loss function similar to what in SSD [11].

Based on our model, the locations and classes of each object can be obtained in one regression, and the result is supposed to overshadow SSD's due to our improvement.

5. Resorces

The resources include two computers with four Maxwell Titan X GPUs.

6. Conclusion

The regression based detection method will play an important role in real-time detection. By combining the feature both at deep and shallow layers and using convolution kernels with aspect ratios, the detection results are expected with a leap forward.

References

[1] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. *arXiv preprint arXiv:1512.04143*, 2015.

[2] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. *arXiv preprint arXiv:1605.06409*, 2016.

[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[4] S. Gidaris and N. Komodakis. Locnet: Improving localization accuracy for object detection. *arXiv preprint arXiv:1511.07763*, 2015.

[5] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[7] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 447–456, 2015.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European Conference on Computer Vision*, pages 346–361. Springer, 2014.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[10] T. Kong, A. Yao, Y. Chen, and F. Sun. Hypernet: Towards accurate region proposal generation and joint object detection. *arXiv preprint arXiv:1604.00600*, 2016.

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed. Ssd: Single shot multibox detector. *arXiv preprint arXiv:1512.02325*, 2015.

[12] W. Liu, A. Rabinovich, and A. C. Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.

[13] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[14] M. Najibi, M. Rastegari, and L. S. Davis. G-cnn: an iterative grid based object detector. *arXiv preprint arXiv:1512.07729*, 2015.

[15] C. Papageorgiou and T. Poggio. A trainable system for object detection. *International Journal of Computer Vision*, 38(1):15–33, 2000.

[16] B. Pepikj, M. Stark, P. Gehler, and B. Schiele. Occlusion patterns for object class detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3286–3293, 2013.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *arXiv preprint arXiv:1506.02640*, 2015.

[18] 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*, pages 91–99, 2015.

[19] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[20] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3626–3633, 2013.

[21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[22] C. Szegedy, A. Toshev, and D. Erhan. Deep neural networks for object detection. In *Advances in Neural Information Processing Systems*, pages 2553–2561, 2013.

[23] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.