How well does Single-Stage Detector perform on Face Detection?

Abstract

In recent years, the trade-off between accuracy and speed in face detection has been widely researched. As many single-stage detectors has achieved remarkable performance on both sides, two-stage methods lost its advantages. Several experiments on the obstacles for single-stage face detector are conducted and the findings are summarized. This repo provides an easy-to-adapt face detection baseline with some tricky settings commented in the codes. Hopefully it could provide some basic ideas for the beginners.

1. Introduction

Faster RCNN [7] and its variants have been the dominant detection architecture since proposed. They performed impressively in alleviating pose, illumination and rotation problems due to the large capacity. With appropriate anchor settings and training strategy, tiny and occluded faces can be detected effectively as well. There are two stages in such methods, proposal and detection. The second stage sifts and calibrates the Region of Interests (RoIs) proposed by the first stage. On the one hand, if the proposals fail to assume high recall or the features fail to provide sufficient information, the detector would have poor results. On the other hand, since the detection is conducted in RoI level, the convolution kernel is able to distinguish each facial part more precisely. However, the relatively lower speed (5FPS or so) has hindered its application in real situation.

More recently, single-stage detector increased the speed to over 20FPS while maintaining the accuracy. SSD [4] and YOLO [6] first offered solution to real-time object detection with high accuracy. Many SSD based face detector has also proved the effectiveness on face detection. Nevertheless, these methods performed poorly on small objects or tiny faces. SSD resembles a Region Proposal Network in a lot of ways but their goals are different. The latter one selects top 2000 bounding boxes for training while the former one reserves a small portion as training samples. Therefore, when training a single stage detector, tiny faces can only match few anchors, which are not sufficient to learn a distribution. For this scale problem, many researchers have proposed different strategies. We delve into some of the strategies and report the experimental results. Hopefully it will give some insights for further study. The baseline code is easy to adapt and has been available.

2. Experiment

The basic architecture is SSD. This part reports how the model was modified and how the results were influenced. The models are trained on the Wider Face [10] training set and the reported results is the recall on the FDDB [2] test set at 1000FPs.

2.1. Detection on different layers

State-of-the-art detectors like SSD and FPN [3] detect faces of different sizes in different layers, which was proved again in this method. When training a SSD on the Wider Face training set without any modification, the results is poor (around 94%). It is because Wider Face training set contains nearly 50% tiny faces (10-50 pixels) but the lowest detection layer on SSD is conv_4. Conv_4 layer has a stride of 8, which means the highest resolution for a 16*16 face is 2*2. When moving the lowest layer to conv_3 like s3fd did, the result witnessed an increase of 1.5%. In this way, tiny faces can be learned more efficiently. However, when using conv_2 as the lowest layer, results dropped rapidly. This shows the significance of semantic information.

2.2. Matching strategy

Selecting appropriate matching strategy is important as tiny faces match few anchors. As is analyzed in s3fd [11], lowering the IoU threshold can help increase matched anchors for tiny and outer faces. In the experiments, decreasing the default threshold (0.5) to 0.3 can help increase the recall slightly, because there are only about 1 or 2 new positive anchors added per face in average. But when decreasing the threshold to 0.1, the number of false positives can't be controlled. We use 0.3 as the final setting because all the faces larger than 11*11 can match at least one anchor. The bigger ones match more.

Moreover, after visualizing some of the positive prior anchors and negative prior anchors, the finding was that almost all the negative anchors are around the positive ones with a slightly lower IoU when training with OHEM. It will

be confusing if an anchor with an IoU of 0.31 is positive while the other one with an IoU of 0.28 is negative. Setting a gap like Faster RCNN did will be a good choice. Besides, we also find that OHEM [8] have bigger impact on single-stage detector than two-stage detector because the two-stage detector have embodied methods alike, e.g., RPN has eliminated the most nave anchors.

2.3. Atrous convolution

Atrous convolution [4] is not suitable for face detection, because the context information embodied in the dilated kernel is too discrete to detect tiny faces. This strategy is with a premise which is nearby pixel values are similar. Replacing the atrous convolution with a normal one (kernel size=3, stride=1, padding=1) can lead to some increase. Besides, it depends on the training source. Wider Face training set contains too many tiny faces and is not suitable for this strategy, but it may perform well in some other situations.

2.4. Anchor settings

Tiling different anchor scales and aspect ratios seems to benefit detection to some extent. But it is also responsible for more false positives. Some researches has stated that they don't need a lot of anchors to achieve high accuracy. So we tried using 1 anchor scale and 3 aspect ratios instead of 4-6 aspect ratios for each detection layer. Results showed that accuracy was almost the same while the generated FPs decreased several hundred.

2.5. Detection module and LFPN

The detection module proposed in SSH [5] seems to be interesting and insightful. It includes a context module to increase the receptive field. As analyzed in [1], considering more context information will be more beneficial to recognizing tiny faces than large ones. We added a detection module on the lowest detection layer. To our surprise, it didn't seem to influence the results much. Conv_3_3 layer has a receptive field of 40 pixels and aims to detect faces around 16*16. It may be caused by different architecture. Or the scarce information of tiny faces has not been utilized efficiently enough although the receptive field is large enough.

We also tried Low-Level Feature Pyramid Network (LFPN) [9] but with less improvement. Theoretically, merging low-level features with higher-level features can help detect small objects. More experiments is on the way.

3. Conclusion

Some useful tricks will not work in all the architectures. Hopefully the analysis in this report will be of some use in your study.

References

[1] P. Hu and D. Ramanan. Finding tiny faces. In *IEEE CVPR*, 2017.

[2] V. Jain and E. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.

[3] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In *CVPR*, page 4, 2017.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37, 2016.

[5] M. Najibi, P. Samangouei, R. Chellappa, and L. Davis. Ssh: Single stage headless face detector. In *IEEE ICCV*, 2017.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016.

[7] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[8] A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, pages 761–769, 2016.

[9] X. Tang, D. K. Du, Z. He, and J. Liu. Pyramidbox: A context-assisted single shot face detector. *arXiv preprint arXiv:1803.07737*, 2018.

[10] S. Yang, P. Luo, C.-C. Loy, and X. Tang. Wider face: A face detection benchmark. In *IEEE CVPR*, pages 5525–5533, 2016.

[11] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li. S3fd: Single shot scale-invariant face detector. In *IEEE ICCV*, 2017.