Day la cai tieu de cua paper Reiew Assistant nhe!!!

Minh Huu Nguyen[1,2] minh.nguyenhuu@pixta.co.jp
Hieu Trong Phung[1,2] hieu.phungtrong@pixta.co.jp
Tung Dinh Nguyen[1] tung.nguyendinh@pixta.co.jp

[1]PIXTA Vietnam, 8th Floor, Truong Thinh Building, Phung Chi Kien, Cau Giay District, Hanoi, Vietnam.
[2]Hanoi University of Science and Technology, 1 Dai Co Viet Road, Ha Noi, Viet Nam.

Abstract

Nowadays, a device which can capture an image in high resolution such as 8K or 16K becomes more popular. Besides, face detection is always an attractive problem for researcher because of its importance in many downstream application. Lots of incredible achievements have been published recent years to solve face detection problem, but these models work on quite low resolution images. The development in the resolution of image requires a machine learning model processes a high resolution image fast with an optimized memory usage. Moreover, there are few open datasets with facial bounding boxes consists high resolution images. To overcome the above problems, in this paper, we present several contributions: (1) We propose RetinaFocus - a novel face detector which can run effectively on 4K resolution image. (2) We share a new idea to create a high resolution validation subset of WIDER FACE dataset to benchmark model performance both accuracy and speed.

Keywords: face detection, high resolution image

1 Introduction

Since face detection becomes a fist and foremost part in lots of face applications like face attributes classification, face identity recognition, face translation etc., intensive attention has been paying to face detection problem. Many solutions have been published and archieve incredible results in localizing in-the-wild faces in different condition. To improve model performance, some works [1, 2, 3, 4, 5, 6, 7] focus on re-designing model architecture while [1, 2, 8, 9] present auxiliary loss function. Besides, some others pay more attention to extra data annotation [9, 10, 11] or data augmentation [4, 3]. In spite of the remarkable results of these works, they can run efficiently on low resolution image. The majority of above models use [12, 13, 14, 15, 16] to benchmark model performance but images in these datasets are quite small, mostly under 1000 pixels. Lacking of public datasets with high resolution images is also an obstacle for researcher to study and improve their model to work with such huge images.

With the improvement of modern cameras, images captured by these devices have resolution as high as 4K (3840×2160 pixels) or 8K (7680×4320 pixels) or even more and they contains much more information. Forwarding such huge images through the deep learning model requires lots of computation cost, memory usage and time and training the deep learning model with these images is almost impossible. Naive idea like resizing these images into smaller size before forwarding through the model will exclude lots of information especially on the tiny object. Specifically, in face detection problem, rescaling the images into popular size for research such as 112×112, 256×256 or even bigger still make all the small faces almost disappear. This raises a problem is that we need a solution to handle the high resolution images efficiently without losing tiny objects information in both training and inference phase. In the training phase, some works [17, 18] propose an idea to train model with high resolution image which requires shorter training time and

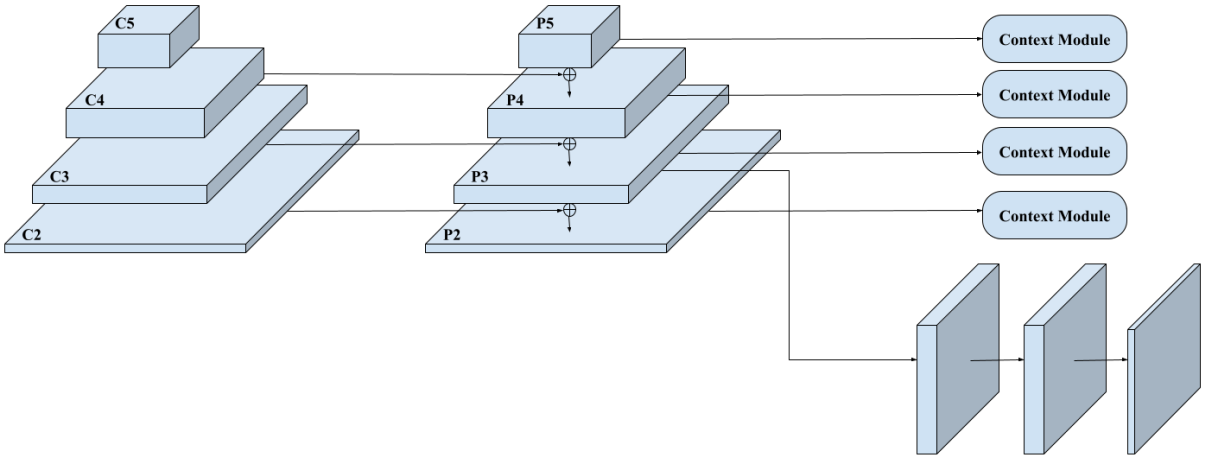

Figure 1: RetinaFocus overall architecture

lower memory usage while maintaining model performance. Moreover, in the inference phase, [19] can ensure memory constraint by cropping an image into multiple regions before forwarding through the model, but processing a dozen of cropped regions still takes lots of time. Some others [20, 21] pay attention to both memory usage and inference time by creating a module to learn which are the best region to be cropped and ignore the majority background on the image.

To solve face detection problem but working with 4K or 8K images, our key contributions can be summarize as follow:
- Inspired by [9] and [20], we propose a novel solution named RetinaFocus, which can handle face detection problem effectively on high resolution images.
- To benchmark the performance of solution on high resolution images, we propose a new idea to generate high resolution validation subset of WIDER FACE dataset.

2 Related Work

Generic object detection. In the deep learning era, to solve generic object detection problem, lots of ideas have been proposed and achieved remarkable results. These methods can be categorized into two groups: the two-stage and the single-stage methods. [22], [23] and [24] which belong to the two-stage methods aim at creating proposals and refining them to achieve higher accuracy. The three models have shown the improvement on model speed by changing region proposals module from [25] to deep learning module named RPN. On the other hand, the single-stage methods [26, 27, 28, 29, 30, 31, 32] sample lots of object location in multiple scale and ratio which help improving detection speed significantly. The single-stage methods have been attracted researchers recent years, compared with the two-stage methods, by the model performance while model accuracy is still improving.

Face detection. With a wide application field, face detection has been developing rapidly to solve some indentical problems such as scale, occlusion, pose, illumination, expression, makeup, age, blur and etc.. TinaFace [33] has shown that *there is no gap between face detection and generic object detection* and face detection methods can be inherited a lot from methods solving generic object detection problem but some proposed ideas are customized to handle face features better. Inspired by [26], [7] proposed a scale compensation anchor matching strategy to improve model accuracy on small faces. [3, 4], based on the architecture of [27], propose a new module to improve performance working

on face. Inherited from [28], [9] proposed five facial landmarks an extra annotation and multi-task loss function for better learning facial features.

Large-scale variance. Although the proposed object detection solutions have achieved remarkable results, the large scale variance problem still be sore because it severely degrade performance of object detector. [17] has shown that the CNN-based models don't perform very well while changing in scale of objects. Some ideas have been proposed to overcome the multiple scales problem such as combining the feature maps from multiple layers [27], modifying training or inference scheme [17, 20], preprocessing training dataset [18].

3 RetinaFocus

Inspired by [9] and [20], our RetinaFocus was built to take advantage of both of the models and fix the problems. Despite having FPN [] in the backbone, [9] still meets difficult while working with small face in the image. With image pyramids inference strategy, [9] archieves a good result on WIDER FACE dataset [12] but it takes time and computation cost. TODO: Prepare figure to explain. Besides, [20] proposes a smart way to run image pyramids inference strategy while reducing number of processing pixels and save time while working with high-resolution images. That's why we come to an idea to combine both of them to be one single RetinaFocus model.

3.1 Model architecture

Our RetinaFocus consists of two branches: detection branch and focus branch shown in figure 1.

Detection branch. Instead of using Faster R-CNN model like in [20], we employ the model from [9] to be the model of the detection branch. The architecture of detection model is unchanged and the output of the model will be used for the multi-task loss function from [9].

Focus branch. The feature maps from FPN backbone not only be forwarded to the Context Module like [9] but one of them also become the input of Focus branch. In figure 1, we show an example of the architecture using $P3$ feature maps of FPN [] as an input of Focus branch. In some other configs, we can use other feature maps such as $C2, C3, C4, P2, P4$ to be the input of Focus branch. Focus branch is simply a series of convolution layers and gives an output predicted mask to compare with a generated focus mask from Focus Pixel like in [20]. To optimize the difference, we employ a Focal loss function from [] to handle imbalance data.

3.2 Hyper-parameters study for Focus Pixel

Based on performance of [9] on WIDER FACE dataset, we study the size and the confidence of prediction to conclude hyper-parameters for the Focus Pixel from [20].

From figure **??**, we found that there are .... % of

3.3 Inference strategy

4 Experiments

4.1 Datasets

WIDER FACE dataset. One of the most popular dataset to benchmark the face detection models is WIDER FACE dataset [12]. The dataset contains 32,203 images with 393,703 face bounding boxes. Face bounding boxes in WIDER FACE dataset are in various scales, poses, expressions, occlusions and illuminations. The WIDER FACE dataset is split into three subsets: training (40%), validation (10%) and testing (50%) with three level of difficulty: Easy, Medium and Hard based on the detection rate of EdgeBox [34]. Moreover, five facial landmarks annotation was added into the training and validation subset of WIDER FACE by [9].

WIDER FACE 4K validation subset. In order to benchmark the performance of face detection solutions while working with high resolution images, we propose an extra validation subset of WIDER FACE dataset. This subset consists of images which was made by concatenating multiple images from original WIDER FACE validation subset.

4.2 Implementation details

Detection branch.

Focus branch.

Data augmentation.

Training configs. We use ... optimizer with .... to train our RetinaFocus. The learning rate scheduler starts from ...., ..., ... It takes ... days to finish the whole ... epoch training process on two NVIDIA RTX 2080Ti (12Gb) GPUs.

Inference configs.

4.3 Ablation Study

Parameters for Focus Pixel.

FPN feature maps for Focus branch.

4.4 Face detection accuracy and speed trade-off

.

5 Conclusion

Acknowledgments

This project is sponsored by PIXTA Inc. We are grateful to , , and for their fruitful comments, corrections, and inspiration. The last author also received the support from Vietnam Institute for Advanced Study in Mathematics in Year 2020.

References

[1] Bin Zhang, Jian Li, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, Yili Xia, Wenjiang Pei, and Rongrong Ji. Asfd: Automatic and scalable face detector. *arXiv preprint arXiv:2003.11228*, 2020.

[2] Shifeng Zhang, Cheng Chi, Zhen Lei, and Stan Z Li. Refineface: Refinement neural network for high performance face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4008–4020, 2020.

[3] Xu Tang, Daniel K Du, Zeqiang He, and Jingtuo Liu. Pyramidbox: A context-assisted single shot face detector. In *Proceedings of the European conference on computer vision (ECCV)*, pages 797–813, 2018.

[4] Zhihang Li, Xu Tang, Junyu Han, Jingtuo Liu, and Ran He. Pyramidbox++: High performance detector for finding tiny face. *arXiv preprint arXiv:1904.00386*, 2019.

[5] Mahyar Najibi, Pouya Samangouei, Rama Chellappa, and Larry S Davis. Ssh: Single stage headless face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 4875–4884, 2017.

[6] Mahyar Najibi, Bharat Singh, and Larry S Davis. Fa-rpn: Floating region proposals for face detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7723–7732, 2019.

[7] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 192–201, 2017.

[8] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: dual shot face detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5060–5069, 2019.

[9] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020.

[10] Samuel WF Earp, Pavit Noinongyao, Justin A Cairns, and Ankush Ganguly. Face detection with feature pyramids and landmarks. *arXiv preprint arXiv:1912.00596*, 2019.

[11] Dmitry Yashunin, Tamir Baydasov, and Roman Vlasov. Maskface: multi-task face and landmark detector. *arXiv preprint arXiv:2005.09412*, 2020.

[12] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5525–5533, 2016.

[13] Vidit Jain and Erik Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical report, UMass Amherst technical report, 2010.

[14] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial

landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 2144–2151. IEEE, 2011.

[15] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 397–403, 2013.

[16] Junjie Yan, Xuzong Zhang, Zhen Lei, and Stan Z Li. Face detection by structural models. *Image and Vision Computing*, 32(10):790–799, 2014.

[17] Bharat Singh and Larry S Davis. An analysis of scale invariance in object detection snip. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3578–3587, 2018.

[18] Bharat Singh, Mahyar Najibi, and Larry S Davis. Sniper: Efficient multi-scale training. *Advances in neural information processing systems*, 31, 2018.

[19] Vit Ruzicka and Franz Franchetti. Fast and accurate object detection in high resolution 4k and 8k video using gpus. In *2018 IEEE High Performance extreme Computing Conference (HPEC)*, pages 1–7. IEEE, 2018.

[20] Mahyar Najibi, Bharat Singh, and Larry S Davis. Autofocus: Efficient multi-scale inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9745–9755, 2019.

[21] Mingfei Gao, Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. Dynamic zoom-in network for fast object detection in large images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6926–6935, 2018.

[22] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation, 2013.

[23] Ross Girshick. Fast r-cnn, 2015.

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.

[25] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[26] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. *Lecture Notes in Computer Science*, page 21–37, 2016.

[27] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[28] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018.

[29] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2016.

[30] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger, 2016.

[31] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement, 2018.

[32] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection, 2020.

[33] Yanjia Zhu, Hongxiang Cai, Shuhan Zhang, Chenhao Wang, and Yichao Xiong. Tinaface: Strong but simple baseline for face detection. *arXiv preprint arXiv:2011.13183*, 2020.

[34] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*, pages 391–405. Springer, 2014.