International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
Impact Factor (2020): 6.733

Comparative Analysis of CNN Based Face Detection

Chenxing Ma

North China Electric Power University, School of Control and Computer Engineering, No. 2 Beinong Road, Changping District, Beijing, China
Mail: *1434728517[at]qq.com*

Abstract: ***The face detection task has been studied in depth. There are may efficient face detectors which utilize specialized designs in different aspects for the detection task for faces, making the detection algorithms and models more and more complex. As a result, the computational and time cost becomes higher. In recent years, many studies are carried out aiming at reducing the algorithm and model complexity. These simpler face detectors make the detection faster while ensuring detection accuracy. In this paper, we select three different face detection models that simplify the face detection algorithms or model structures based on common CNN networks and YOLO structures, they are, YOLO5Face, DSFD and TinaFace. We first analyze the algorithm and model structure of the selected face detectors and then test them on several datasets to evaluate the generalization ability of the models. The experiment result show that the selected face detectors can efficiently complete the face detection task while YOLO5Face has the best performance on the datasets.***

Keywords: Face detection, CNN, YOLO, datasets

1. Introduction

Face detection is a fundamental task in the area of computer vision, which is the first step of various face involved vision tasks, including face alignment, face tracking, and key point detection. A number of existing deep neural network model-based face detection methods utilize specialized designs in the aspects like supervision data or network structures. Some approaches use extra supervision data with the help of annotated landmarks information, aiming at improving the model performance by providing the extra supervision data0[2]. While some other approaches directly pay attention to the network structure[3][4][5][6], they improve their model detection accuracy by designing specialized module into the network so that enhancing the model's ability to extract features and estimate face landmarks.

The approaches mentioned above gradually separate the face detection task from the general object detection, even face detection is just a subproblem of the object detection task. As a result, the face detection algorithms and the model structures are getting more and more complex. Therefore, some recent studies re-examine the face detection task and treat this task as just a general object detection task[7][8]. These methods consider the properties of faces, such as pose, scale, occlusion and illumination also exist in general objects, while the particular properties such as makeup and expression can also correspond to the common object properties such as color and distortion. And the face landmarks can be treated as the key points of common objects. Moreover, the challenges encountered in the face detection task like multi-scale, small faces and dense scenes also exist in generic object detection task and have been properly solved by previous work. Based on the above analyzes, the face detection task can be accomplished through generic object detection.
The face detection task has a strong practical application significance, which requires the face detection model to have a sufficiently high accuracy rate and a high calculation speed. In practical applications, the input images obtained by the face detection model have high uncertainty, with variables such as the number and size of faces. Therefore, this requires the model to have high robustness and generalization in practical applications, so that it can achieve sufficiently high detection accuracy in other datasets while ensuring high accuracy in the test dataset.

Therefore, in this paper, we select several existing SOTA face detector to evaluate their detection accuracy and the generalization ability. They are DSFD[6], YOLO5Face[7] and TinaFace[8]. DSFD expand the vanilla VGG 16 network as the backbone network, and design a dual shot detector for both coarse and fine feature maps extracting to improve the accuracy and speed of face detection. The YOLO5Face detector treat the face detection task as a general object detection task which is based on the You Only Look Once (YOLO) architecture. The method proposes modifications to the YOLOv5 model to make it more suitable for face detection, including an anchor-free detector, an improved feature pyramid structure and a new data augmentation strategy. The Tinaface is a face detector that achieves strong performance while maintaining a simple architecture. Tinaface also treat the face detection task as a general object detection task, while it is constructed by a two-stage framework that first generates candidate regions using a lightweight backbone network, followed by refining the regions using a feature pyramid network.

We carry out comparative experiment for the above three face detection method on different datasets, the WiderFace dataset[9], the MAFA dataset[9] and the UFDD dataset[11]. Among them, the Wider-face dataset has a total of 32,203 pictures, a total of 393,703 faces, which is 10 times larger than the FDDB dataset, and there are great changes in the size, posture, occlusion, expression, makeup, and lighting of the face. The algorithm not only labels Boxes, which also provide occlusion and pose information, have been widely used since their publication to evaluate convolutional neural networks

that outperform traditional methods. The MAFA dataset is a face detection dataset with occlusion. The dataset is an occluded face detection dataset, which contains a total of 30,811 images and 35,806 occluded faces, including occlusions in various directions and scales. The UFDD dataset is a face detection data set in an unrestricted scene. It contains a total of 6425 images and 10897 faces, including Rain, Snow, Haze, Blur, and Illumination, Lens impediments and Distractors.

The comparative experiment shows that the selected face detector can achieve high detection accuracy in the datasets. However, the processing efficiency of DSFD is so low that the method cannot support real-time face detecting application in real world.

2. Face Detection Methods

This section analyzes the model structure and innovations of the three selected face detection methods.

2.1 Dual Shot Face Detector

The Dual Shot Face Detector (DSFD) proposed by Li et al. is a state-of-the-art face detection method that achieves high accuracy and efficiency by using two stages of detection. The detection network accepts the ResNet152 network as the backbone and replace the full-connection layers in ResNet152 with other assistant convolutional layers. In the first stage of the network, the DSFD algorithm employs a coarse-grained detector to localize potential face regions in the input image. This detector is based on a feature pyramid network (FPN), which extracts multi-scale features from the input image and generates region proposals using anchor boxes. The proposed regions are then refined using a regression network to obtain more accurate bounding boxes for potential faces. While in the second stage, the DSFD algorithm first utilize a feature enhance module wo obtain enhanced fine-grained feature maps, and then uses a fine-grained detector to further refine the candidate face regions generated in the first stage. The fine-grained detector is also based on an FPN architecture and uses a similar regression network to refine the face bounding boxes. However, unlike the first stage, the fine-grained detector is designed to be more accurate and has a smaller detection range, allowing it to better capture the details of small faces in the input image. The model structure of the DSFD is show in Fig. 1.

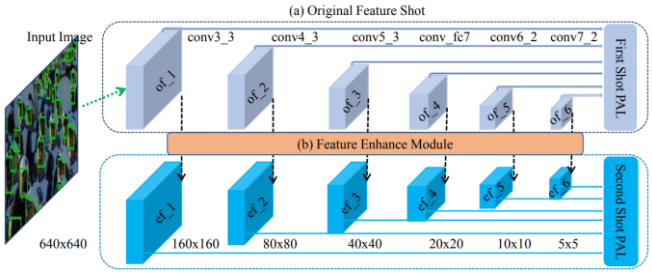

Figure 1: The structure of DSFD network.[6]

To improve the detection performance, DSFD utilizes an improved anchor matching strategy. The One-stage detector has a dense anchor assigned to the output layer, and the matching between the anchor and the face directly affects the training effect. In the process of data augmentation, DSFD fully considered the relationship between faces of different sizes and each anchor, and proposed a new data augmentation method, which combines the anchor division strategy and the anchor-based data augmentation method to provide better initialization to the regressor, so that the anchors and ground-truth faces match as much as possible.

A Feature Enhance Module (FEM) is designed to enhance the original feature maps directly extracted from the input image. The FEM combines the advantages of the FPN module in some previous work and thus can improve the discriminability and robustness of the algorithm. Specifically, FEM accept feature maps from both the current layer and the upper layer as input, firstly normalize them using 1x1 convolutional layers, then upsample the normalized upper feature map to perform element-wise product with the current feature map. The multiplied feature map is then separated into three parts and input the parts into three dilation convolution blocks. Finally, the output of FEM is obtained by concatenate the output of the dilation convolution blocks. The structure of the FEM is shown in Fig. 2.

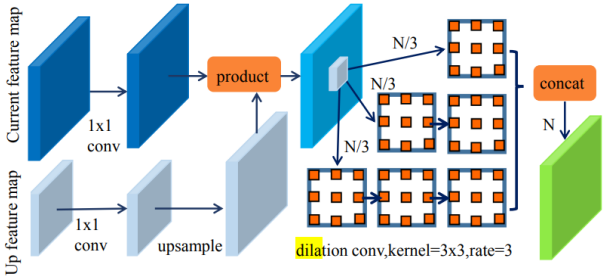

Figure 2: The structure of the Feature Enhance Module[6]

In order to further improve the model's performance, a loss strategy called Progressive Anchor Loss (PAL) is proposed to supervised the training of the dual shot detection model. The model adopts 2 hierarchies, based on the difference between the first layer (low-level) and the second layer (high-level), tiling smaller anchors to the higher-level feature map cell can get more Semantic information for classification and more high-resolution localization information for detection. In other words, it is through a set of smaller anchors to calculate the auxiliary supervised loss to assist feature learning. During the training process, PAL forms a more effective supervision of the entire model.

2.2 YOLO5Face

Different from some existing face detector such as DSFD mentioned above, the YOLO5Face treat the face detection task as a general object detection task. Following the generic object detection methods, the YOLO5Face model is constructed of a YOLOv5 network as the backbone, a neck and a head, whose overall structure is shown in Fig. 3. The neck module utilizes an SPP and a PAN for the feature aggregation while the head module plays a role to output detection regression result, which constructs of a bounding box, confidence, classification and landmarks.

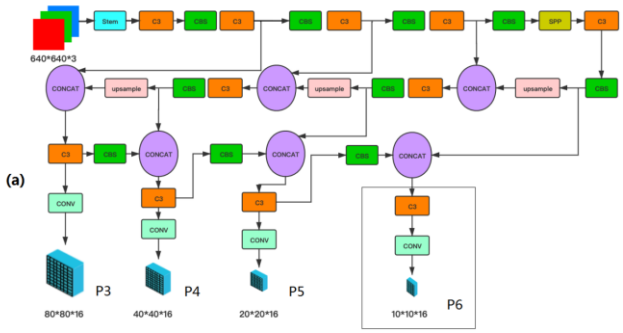

Figure 3: The structure of the YOLO5Face[7]

In YOLO5Face, a tailored convolutional block called CBS is designed as a basic block in the network. The CBS block consists of a convolutional layer, a Batch normalization layer and a SILU layer. Based on the CBS basic block, there are several key modifications performed on YOLOv5 network to construct YOLO5Face. First, a regression head for landmark obtaining is added into vanilla YOLOv5 network, with a Wing loss applied in the overall loss function to supervise the regression head during model training. Comparing with the L1 and L2 loss functions, the response of Wing loss in the small error region close to zero is improved. Thus, the overall loss function consists of the Wing loss and the general object detection loss function of YOLOv5. This extra loss function helps to make the estimated landmark locations more accurate and thus improve the face detector accuracy.

Then the Focus layer of YOLOv5 network is replaced by a stem block, which helps to improve the generalization ability of the model so that it can adapt to more application scenarios, and at the same time reduce the computation cost while keep the model performance not to degrade. In the spatial pyramid pooling block, a smaller kernel size is used instead of the original larger size, making the model more suitable to the face detection task. Considering the possibility of existence of large face in an image, for example, in a portrait image, a P6 output block with a stride of 64 is added so that the model can be sensitive to the large face.

There are many data augmentation strategies in general object detection, while these strategies are often not suitable for the face detection, which is a specialized object detection task. Therefore, a data augmentation performed before the model training, removing up-down flipping and mosaic in general object detection augmentation while introducing random cropping into the augmentation. This strategy helps the performance improving.

Further, considering the usage of face detection on embedded and mobile devices, a lightened face detection model based on ShuffleNetV2 is designed, which perform the SOTA performance on embedded and mobile devices with smaller parameter amount.

2.3 TinaFace

Face detection methods of recent years are getting more and more complex, with a number of modules specially designed for the face detection task, like FEM and PAL in the DSFD face detector mentioned in 2.1, and the PA and OAM in HAMBox. Although these complex designs largely improve the accuracy of face detecting, the overall complexity of the model is inevitably increased, resulting in a significant increase in computing power and time overhead, which makes it difficult for these models to adapt to practical application scenarios.

The TinaFace face detector network is constructed based on the RetinaNet network. Comparing with the RetinaNet network, the TinaFace network makes some modifies to achieve higher detection accuracy. First, Group Normalization is employed as the Normalization layers in the TinaFace network, as it is a simpler substitute to the Batch Normalization which plays an important role in a convolutional network encouraging the model to converge. This replacement makes the model more lightweight and reduces computational overhead, while ensuring the stability of model performance.

In order to further reinforce the detection ability of the model, a Deformable Conv Net (DCN) architecture is introduced into the backbone network, thus overcoming the challenge that traditional convolution operation cannot learn and encode the complex geometric transformations, resulting in the low capability of the model.

Considering a common problem in object detection task, that is, the problem of the classification score mismatches the localization accuracy of a single-stage object detector, an IoU-aware regression head, implemented using a single 3x3 conv layer with a sigmoid layer, is added into the network to predict the IoU error between the estimated detected box and the ground-truth. Correspondingly, a Distance-IoU loss function is employed in the overall loss function instead of smooth L1 loss as the box regression loss function.

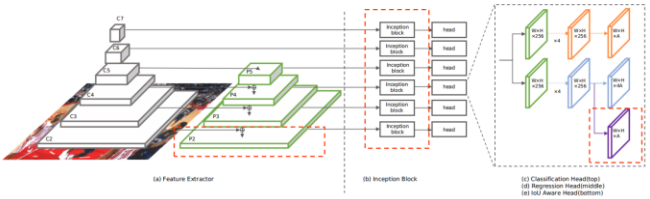

Figure 4: The structure of the TinaFace[8]

3. Comparative Experiments

3.1 Datasets

In this paper, we carry out comparative among the three face detection methods, DSFD, YOLO5Face and TinaFace, on different face detection datasets, thus evaluating the detecting accuracy, processing efficiency and generalization ability of the above methods. The datasets we selected are the WiderFACE dataset**Error! Reference source not found.**, the MAFA dataset[9] and the UFDD dataset[11]. the Wider-face dataset has a total of 32,203 pictures, a total of 393,703 faces, which is 10 times larger than the FDDB dataset, and there are great changes in the size, posture, occlusion, expression, makeup, and lighting of the face. The algorithm not only labels Boxes, which also provide occlusion and pose information, have been widely used since their publication to

evaluate convolutional neural networks that outperform traditional methods. The MAFA dataset is a face detection dataset with occlusion. The dataset is an occluded face detection dataset, which contains a total of 30,811 images and 35,806 occluded faces, including occlusions in various directions and scales. The UFDD dataset is a face detection data set in an unrestricted scene. It contains a total of 6425 images and 10897 faces, including Rain, Snow, Haze, Blur, and Illumination, Lens impediments and Distractors. Fig. 5 to Fig. 7 gives some sample data from the MAFA, WiderFace and UFDD dataset, respectively.

Figure 5: Sample data from the MAFA dataset

Figure 6: Sample data from the WiderFace dataset

Figure 7: Sample data from the UFDD dataset

3.2 Experiment Result and Analysis

We carry out comparative experiment of the DSFD, YOLO5Face and TinaFace face detector on the three datasets above. Like most studies do, we define the three datasets into three levels of difficulty progressively more challenging, that is, easy, medium and hard, therefore, the result in level hard is most convincing in terms of model detection accuracy. The result is shown in Tab. 1, in which YOLO5FaceFull indicates the YOLO5Face detector with YOLOv5-CSPNet as backbone, which is the full model of YOLO5Face, while YOLO5 Face Light indicates the lightened YOLO5Face model with ShuffleNetv2 as backbone. At the same time, Tab. 2 shows the complexity of the face detector networks.

Table 1: Comparative experiment result

Detector	Backbone	Easy	Medium	Hard
DSFD	ResNet152	94.35	91.26	71.58
YOLO5FaceFull	YOLOv5-CSPNet	**96.64**	**95.12**	**86.61**
YOLO5FaceLight	ShuffleNetv2	93.56	91.42	80.68
TinaFace	ResNet50	95.48	94.33	81.26

Table 2: Complexity of the face detector networks

Detector	Backbone	Params(M)	FLOPS(G)
DSFD	ResNet152	120.06	259.55
YOLO5FaceFull	YOLOv5-CSPNet	141.158	88.665
YOLO5FaceLight	ShuffleNetv2	0.447	91.42
TinaFace	ResNet50	37.98	0.571

It can be seen from Tab. 1 that YOLO5FaceFull achieve the highest detection accuracy in the three level with a high processing efficiency with the help of TOLOv5-CSPNet as the backbone, inferring that the method has a strong potential for practical applications. However, the full model of YOLO5Face, the YOLO5FaceFull has an extremely large parameters amount of 141.158M, meaning that this model cannot be used in mobile and embedded devices, while the face detection demands on these devices are becoming higher and higher. Therefore, according to Tab. 2, the YOLO5FaceLight and TinaFace face detector, with a smaller complexity, are more suitable to the application on mobile and embedded devices or the application scenarios with a high processing speed demand.

4. Conclusions

In this paper we select several face detection methods, the DSFD, the YOLO5Face and the TinaFace, to evaluate the detection accuracy, generalization ability and complexity. In order to obtain a more convincing result, we use WiderFACE, MAFA and UFDD datasets together as the test datasets. As the comparative experiment results show, the YOLO5Face with a backbone of YOLOv5-CSPNet get the highest detection accuracy while with the highest network complexity, meaning that this model cannot support the real-time face detecting task. On the other hand, TinaFace and the YOLO5Face model with ShuffleNetv2 as backbone can achieve relatively high detection accuracy without the network being complex, allowing them to be applied in the real-world real-time face detection task, and to be applied on mobile and embedded devices.

References

[1] Jiankang Deng et al. “Retinaface: Single-stage dense face localisation in the wild”. In: arXiv preprint arXiv: 1905.00641 (2019).
[2] Samuel WF Earp et al. “Face Detection with Feature Pyramids and Landmarks”. In: arXiv preprint arXiv: 1912.00596 (2019).
[3] Bin Zhang et al. “ASFD: Automatic and Scalable Face Detector”. In: arXiv preprint arXiv: 2003.11228 (2020).
[4] Shifeng Zhang et al. “Refineface: Refinement neu□ ral network for high performance face detection”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).
[5] Zhihang Li et al. “Pyramidbox++: High performance detector for finding tiny face”. In: arXiv preprint arXiv: 1904.00386 (2019).
[6] Li J, Wang Y, Wang C, et al. DSFD: dual shot face detector[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 5060-5069.
[7] Qi D, Tan W, Yao Q, et al. YOLO5Face: why reinventing a face detector[C]//Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V. Cham: Springer Nature Switzerland, 2023: 228-244.
[8] Zhu Y, Cai H, Zhang S, et al. Tinaface: Strong but simple baseline for face detection [J]. arXiv preprint arXiv:2011.13183, 2020.
[9] Shuo Yang et al. “Wider face: A face detection benchmark”. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, pp. 5525–5533.
[10] Ge S, Li J, Ye Q, et al. Detecting masked faces in the wild with lle-cnns[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2682-2690.
[11] Nada H, Sindagi V A, Zhang H, et al. Pushing the limits of unconstrained face detection: a challenge dataset and baseline results[C]//2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS). IEEE, 2018: 1-10.