CCBR 2017

Detecting Face with Densely Connected Face Proposal Network

Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi
Xiaobo Wang, Stan Z. Li

CBSR&NLPR, Institute of Automation, Chinese Academy of Sciences
University of Chinese Academy of Sciences, Beijing, China

1. Introduction

➢Motivation

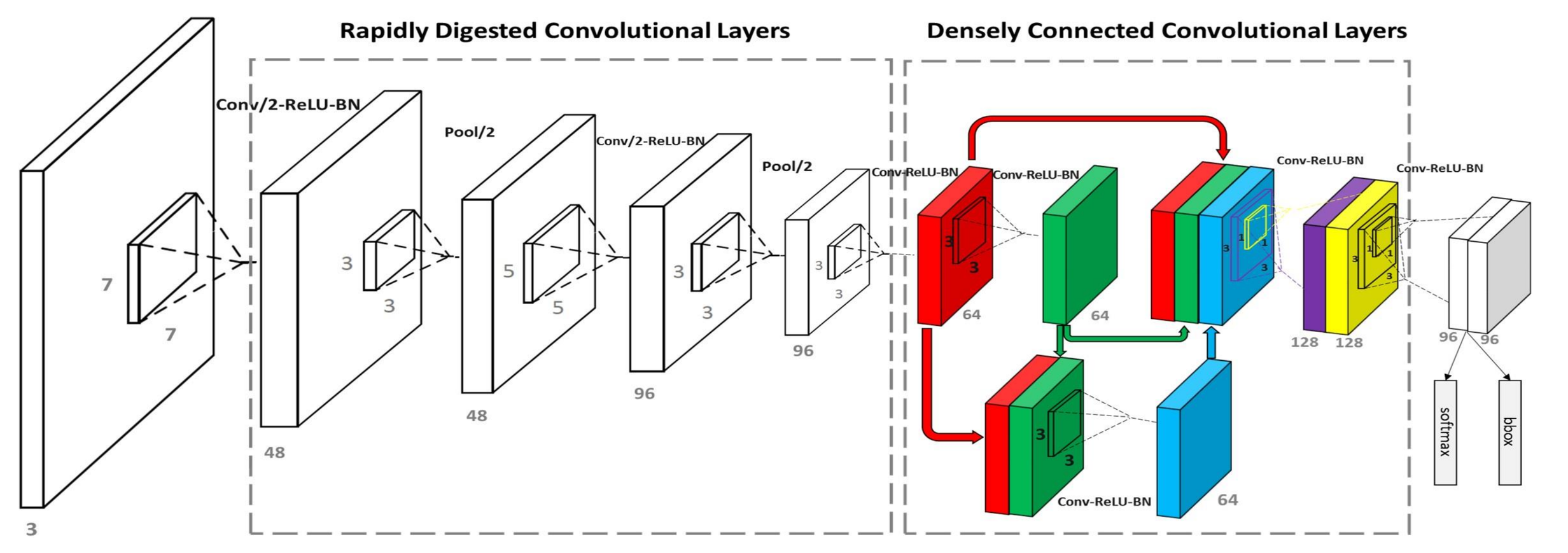
Accuracy and efficiency are two conflicting challenges for face detection, since effective models [2] tend to be computationally prohibitive. To address these two conflicting challenges, our core idea is to **shrink the input image and focus on detecting small faces**.

➢Contribution

- We develop a novel face detector, named Densely Connected Face Proposal Network (DCFPN), with high performance as well as CPU real-time speed
- We subtly design a lightweight-but-powerful fully convolution network with the consideration of efficiency and accuracy for face detection
- We propose a fair L1 loss and use dense anchor strategy [1] to handle small faces well
- We achieve state-of-the-art performance on the AFW, PASCAL face and FDDB datasets at the speed of 30 FPS on a single 2.60GHz CPU core and 250 FPS using a GPU for the VGA-resolution images

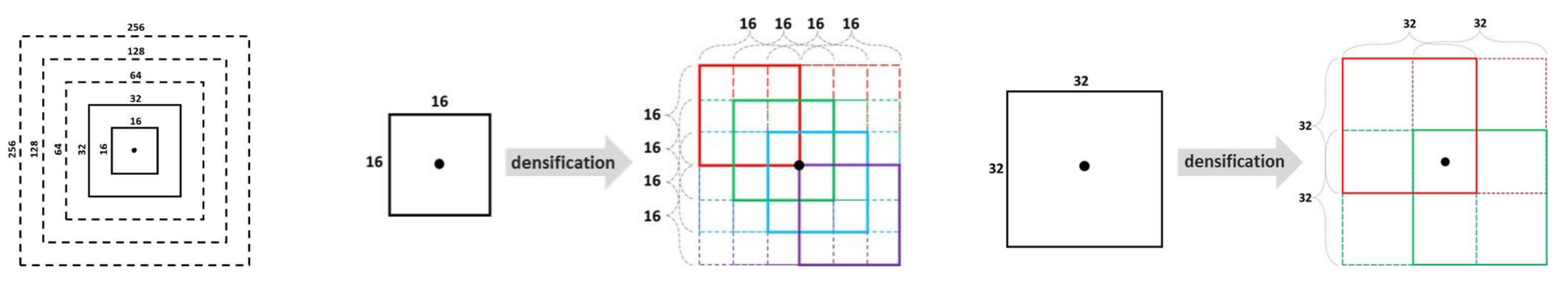
2. Proposed Method

➢Architecture



- **Rapidly Digested Convolutional Layers (RDCL)**: It is designed to achieve real-time speed on the CPU devices via quickly reducing the image spatial size by 16 times with narrow but large convolution kernels.
- **Densely Connected Convolutional Layers (DCCL)**: It aims at not only enriching the receptive field to learn visual patterns for different scales of faces, but also combining coarse-to-fine information across deep CNN models to improve the recall rate and precision of detection.

➢Dense anchor strategy



- **Problem:** the last conv layer has 5 default anchors whose tiling interval are 16 pixels. Comparing with large anchors (64, 128, 256), small anchors (16, 32) are too sparse, which results in low recall rate of small faces.
- **Solution:** dense anchor strategy is proposed by [1] to solve this tiling density imbalance problem. As illustrated in above figure, it uniformly tiles several anchors around the center of one receptive field instead of only tiling one.

➢Fair L1 loss

- The regression target of Fair L1 loss is as follows:

$$t_x = x - x^a,\ t_y = y - y^a,\ t_w = w,\ t_h = h;\ t_x^* = x^* - x^a,\ t_y^* = y^* - y^a,\ t_w^* = w^*, t_h^* = h^*$$

- where x, y, w, h denote center coordinates and width and height, x, x_a, x^* are for predicted box, anchor box, and GT box (likewise for y, w, h). The scale normalization is implemented to have scale-invariance loss value as follows:

$$L_{reg}(t, t^*) = \sum_{j\in\{x,y,w,h\}} fair_{L_1}(t_j - t_j^*),\quad where\ fair_{L_1}(z_j) = \begin{cases} |z_j|/w^*, & if\ j \in \{x, w\} \\ |z_j|/h^*, & otherwiese \end{cases}$$

- It equally treats small and big face by directly regressing box's relative center coordinate and width and height.

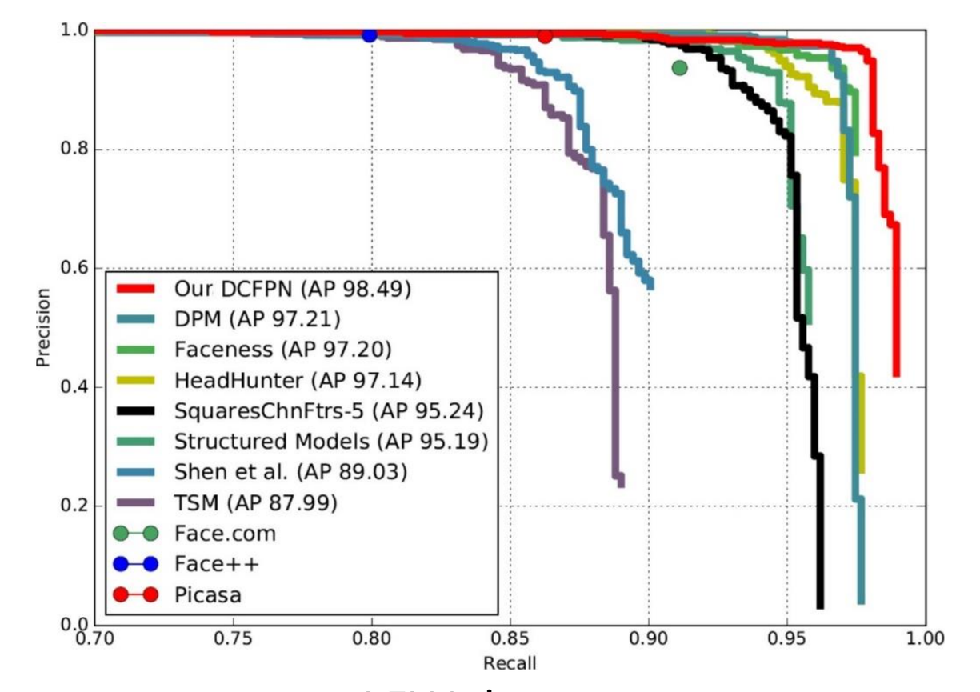
3. Experiments

➢ Model analysis

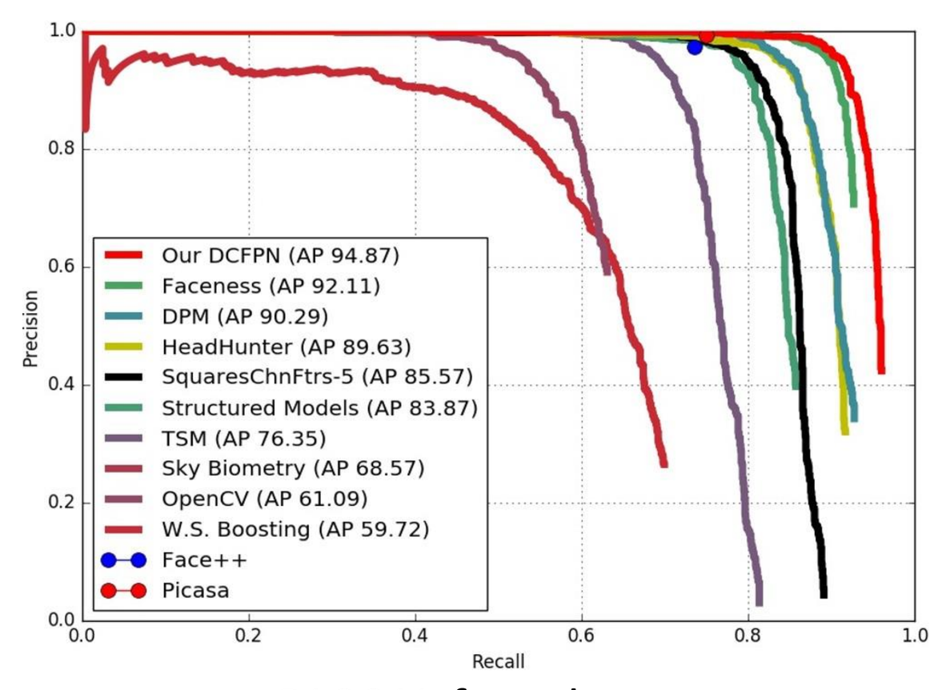
Component	DCFPN			
Designed architecture?	✓	✓	✓	
Dense anchor strategy [1]?	✓	✓		
Fair L1 loss?	✓			
Accuracy (mAP)	95.2	94.5	93.7	93.2

- **Fair L1 loss is promising**: it effectively increases the mAP performance by 0.7%, owning to locating small faces well
- **Dense anchor strategy is effective**: ablating the dense anchor strategy results in 0.8% decline, showing the importance of this strategy
- **Designed architecture is crucial**: the 0.5% decline demonstrates that enriching the receptive fields and combining coarse-to-fine information across different layers are useful to handle faces of various scales

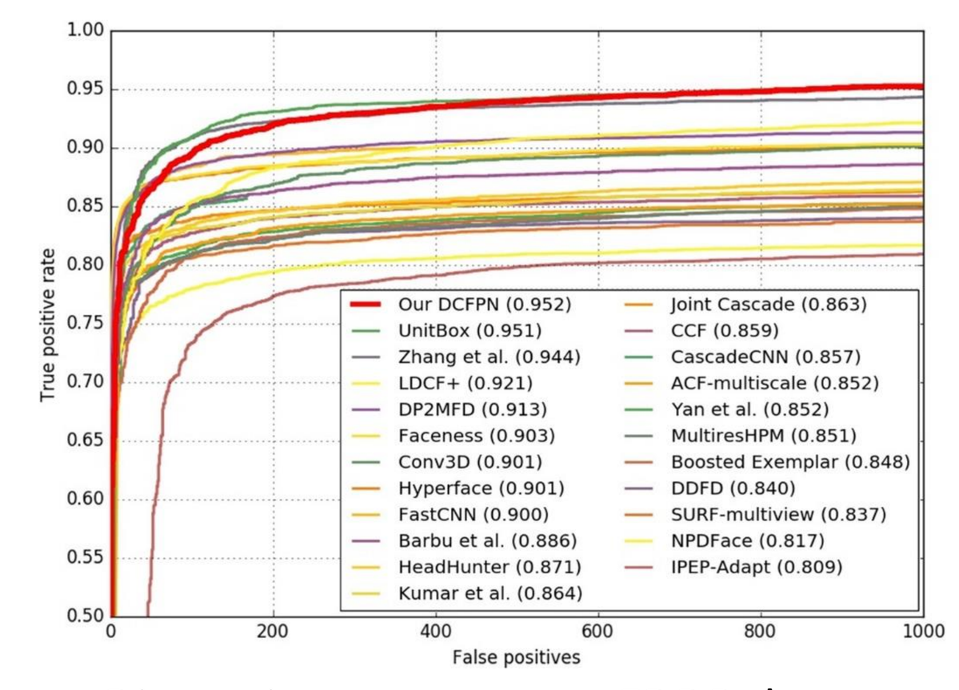
➢ Evaluation on benchmark



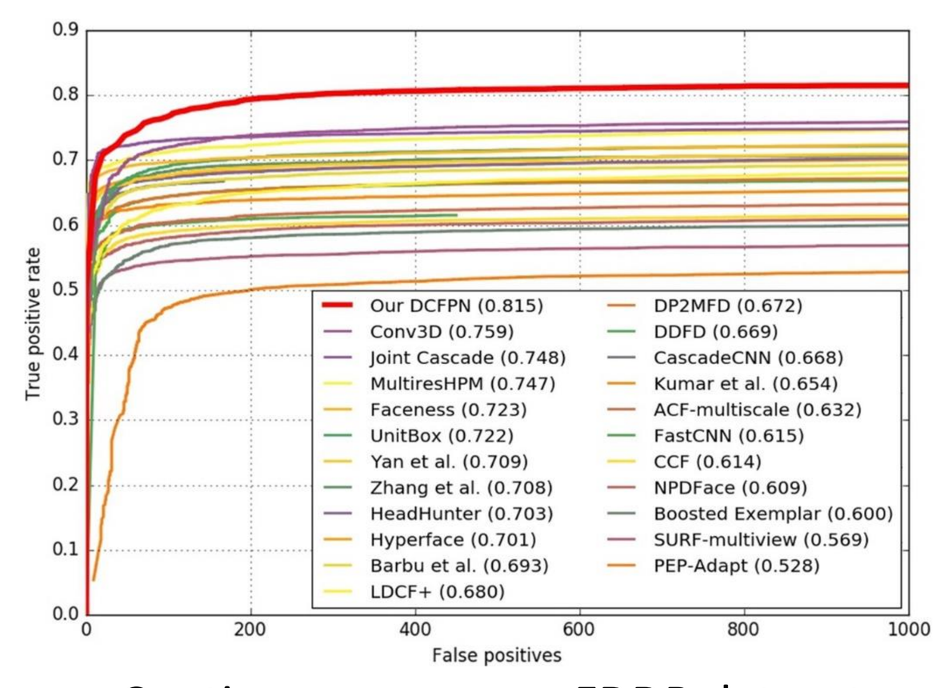
AFW dataset



PASCAL face dataset



Discontinuous score on FDDB dataset



Continuous score on FDDB dataset

➢ Runtime efficiency

As for the practicability, DCFPN is efficient and accurate enough to detect faces bigger than 40 pixels for the VGA-resolution images：

- **30** FPS on a 2.60GHz CPU
- **250** FPS using a GPU card
- **3.2** MB in model size

4. Summary

In this paper, we propose a novel face detector with real-time speed on the CPU devices as well as high performance.

On the one hand, our DCFPN has a lightweight-but-powerful framework that can well incorporate CNN features from different sizes of receptive field at multiple levels of abstraction.

On the other hand, we use the dense anchor strategy and propose the fair L1 loss function to handle small faces well.

The state-of-the-art performance on three challenge datasets shows its ability to detect faces in the uncontrolled environment.

The proposed detector is very fast, achieving 30 FPS to detect faces bigger than 40 pixels on CPU and can be accelerated to 250 FPS on GPU for the VGA-resolution images.

[1] Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., Li, S.Z. Faceboxes: A CPU Real-time Face Detector with High Accuracy. IJCB (2017)

[2] Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., Li, S.Z. S³FD: Single Shot Scale-invariant Face Detector. ICCV (2017)

CASIA CBSR

2017 深圳