Joint Training of Cascaded CNN for Face Detection

Liangjie Cao

Jun 20, 2018

1. Joint Training of Cascaded CNN

It is optimized through back-propagation. Compared to separate networks, the joint network also use threshold control layers to decide which proposals from up branches contribute to the loss of the down branches. Each branch has a face v.s. non-face classification loss and a bounding-box regression loss. Adding them with loss weights, we get the joint loss function:

$$L_{joint} = \lambda_1 L_{x12} + \lambda_2 L_{x24} + \lambda_3 L_{x48} \quad (1)$$

where L_{x12}, L_{x24} and L_{x48}denote different losses of three branches. $\lambda 1$, $\lambda 2$ and $\lambda 3$ are loss weights of the three branches.

2. Implementation details

To prepare training data, they first use sliding windows on each training image to get face candidates. The positive samples are chosen from the candidates that have intersection over union (IoU) overlap with any groundtruth bounding box of larger than 0.8. The negative samples are sampled from the face candidates that have a maximum IoU with ground-truth in the interval [0,0.5). The samples are cropped and resized to network input size. To apply data augmentation, each sample is horizontally flipped. The ultimate ratio of positive samples of the whole training data is about 5%. The input patches are mean removed with mean image from ImageNet [1]. No other pre-processing is used.

Each training image is first built into image pyramids with interval of 5. The smallest pyramid is 125 of the original image. They prepare face proposals by sliding windows with stride 8 over training images. Positive samples are chosen from face proposals whose maximum IoU with ground-truth is larger than 0.8. Negative samples are chosen from the proposals that have a maximum IoU with groundtruth in the interval [0,0.5). For the sample ratio, they keep a very low positive sample ratio during stage one. Because this can decrease false positives , which also accelerates the following negative mining stages. In our experiments, setting the ratio of positive samples as 5% is appropriate. The x_{12} branch threshold is set as 0.1, x_{12}-x_{24} branch threshold is set as 0.003. They are set empirically. Within appropriate threshold range, the training procedure is quite robust. The principle is to make the threshold as high as possible while keeping the recall, so as to reject as many proposals as possible in the earlier stages. Alternatively, they can fix the proposal number, which is exactly what we did in the joint training of RPN and fast R-CNN. During forward, only face proposals that have x_{12} branch scores higher than 0.1 contribute to x_{12}-x_{24} branch. Only face proposals that have x_{12}-x_{24} branch scores higher score than 0.003 contribute to x_{12}-x_{24}-x_{48} branch.

They decrease the positive sample thresholds when training the three stages. So in the later stages, they can train the networks with harder samples. This in turn results in stronger models for face v.s. non-face classification. To make it converge eaisily, they train seperate networks and initialize the joint network with trained weights. They set global learning rate 0.001. After a number of iterations, they lower the learning rate to 0.0001 to train more iterations. The specific iteration number is related to the number of training samples. Generally, 5 to 10 epochs would be appropriate. Following standard practice, they use a momentum term with weight 0.9 and weight decay factor of 0.0005.

3. Datasets

In training joint cascaded CNNs, they use Annotated Facial Landmarks in the Wild (AFLW) [1] and our dataset called $3R$. $3R$ contains about 26000 images that have faces and 27000 images that have no faces. $3R$ is collected from online social network, the image on which is a reflection of the real world images in everyday life. To add negative samples, they also use images in PASCAL VOC2012 that do not contain persons as background image. In total, the dataset contain 47211 images with 82987 faces and about 32000 background images. To avoid confusing circumstances when it is difficult to judge a patch is groundtruth or not, they add ignore regions in our training images. An ignore region is defined as a region where we do not sample negative samples.

To avoid annotation confusion, they do not annotate us-

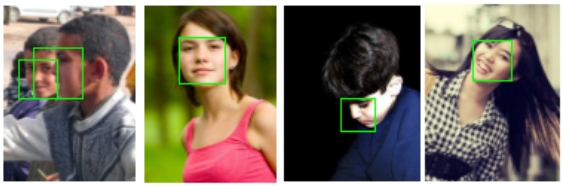

Figure 1. **Face annotation examples**

ing face rectangles. Instead, each face is annotated by 21 facial landmarks. The landmarks are slightly different from those of AFLW official annotations, of which a face may be annotated with less than 21 landmarks. They design a transformation algorithm from facial landmarks to face rectangles. The face rectangles are square annotations. Face examples are shown in Fig. 1. They can see that nose is always in the center of the square annotations.

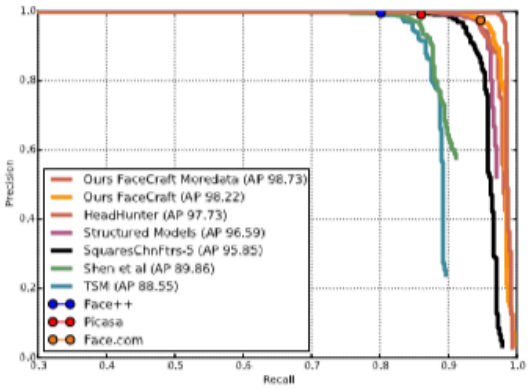

Figure 2. **Precision-Recall Comparisons with state-of-the-art methods on AFW. The methods are HeadHunter [2], Structured Models [4], SquaresChnFtrs-5 [2], Shen *et al.* [3], TSM [5], Face.com, Face++ and Google Picasa.**

4. AFW results

Figure 3. **Qualitative results of FaceCraft on AFW**

They evaluate FaceCraft on Annotated Faces in the Wild (AFW) [5]. AFW contains 205 images collected from Flickr. The images contain cluttered backgrounds and various face viewpoints and appearances. In ground-truth annotation, one specific problem of face detection different from general object detection is how to decide the face bounding box when a face is not frontal. Different rules in face annotations result in various groundtruth. So detectors trained with different training data may get mismatched results on test dataset annotated following different rules. This problem has been pointed out before [18, 16]. In our test results, this is also true. Examples of detection results are shown in Fig. 3. In our test results, non-frontal face bounding-box centred on the nose, which is consistent with our training ground-truth shown in Fig. 1. While in AFW ground-truth, nose is on the bounding-box edge.

References

[1] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Feifei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[2] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *ECCV*, 2014.

[3] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *ECCV*, 2014.

[4] J. Yan, X. Zhang, Z. Lei, and S. Z. Li. Face detection by structural models. *Image and Vision Computing*, 2014.

[5] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012.