Joint coco and Mapillary Workshop at ICCV 2019:

COCO Keypoint Challenge Track

Technical Report: COCO Keypoint

Zhang Fangjian
Infinova
GuangDong in China
1198184921@qq.com

Cheng Wei
Infinova
GuangDong in China
244493137@qq.com

Li Liang
Infinova
GuangDong in China
70768416@qq.com

Li liuwu
Infinova
GuangDong in China
liuwu.li@outlook.com

Abstract

To meeting COCO keypoint challenge track, our tream adopt Top-Down detection method to detect all person's keypoints in every COCO test dataset' picture. In this method, we adopt Hybrid Task Cascade to find all person in dataset. And than, base on HRNet, and make some changes on train data and lt's loss function. Our final result on COCO test-challenge 2019 AP value is 0.737.

Overview of Our Approach

There are mainly two methods on Humen keypoints detection, Top-Down method and Bottom-Up method . And our approach falls into the first category. Figure 1 is the mainly process of our approach in COCO keypoint challenge track. In the step of person detection, we use Hybrid Task Cascade[1], which is a accurate person detection model and achieved the state-of-the-art in 2019. In the second step, our programme is base on High Resolution Net (HRNet)[2], which also have achieved the state-of-the-art performance in human keypoints detection. We use both COCO dataset and external data to train this deep convolutional neural networks. Beside, we change it's loss function to balanced different amount of all 17 COCO keypoints. An last, according to format of COCO Keypoint Challenge Track, to make a result .json file, which in include every person's keypoints in COCO test dataset. And our main work lies in the second part.

Figure 1. mainly process map

Person detection

We used the 2019 open source network, Hybrid Task Cascade, which was improved on the basis of cascade rcnn and mask rcnn. The network implements instance segmentation and target detection by cascading. Use the example segmentation to find the optimal cascading structure and improve the mAP of the target detection. Currently, the mAP of HTC can reach 50.7% on the coco dataset. The structure of Hybrid Task Cascade is shown in Figure 2.

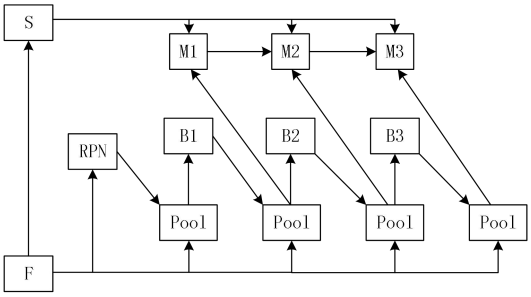

Figure 2. HTC

Keypoints detection

The basics convolutional neural networks of this part's work is HRNet, which have a good perform. HRNet is a new architecture, that is high resolution net. It can maintain high resolution through it's whole process, instead of hourglass shape[3] or network by dilated convolutions[4]. It start from a high-resolution subnetwork as the first stage, radually add high-to-low resolution subnetworks one by one to form more stages, and connect the multi-resolution subnetworks in parallel. And conduct repeated multi-scale fusions by exchanging the information across the parallel multi-resolution subnetworks over and over through the whole process. HRNet estimate the keypoints over the high resolution representations output by this network. The resulting network is illustrated in Figure 3[2].

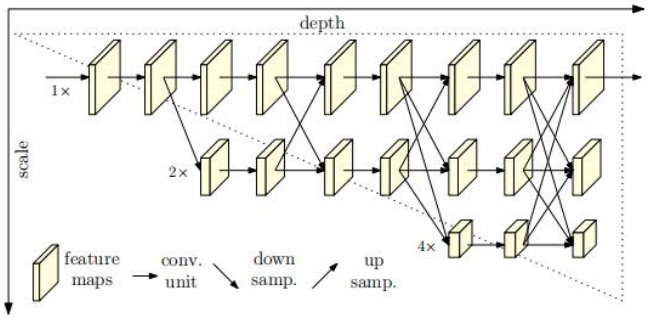

Figure 3. Illustrating the architecture of the proposed HRNet[2].

We train again HRNet by COCO training dataset for COCO Keypoint Challenge Track, in this way get our model1. On the other hand, we also increase extra data AI Challenger[5] to train it. We extracting AI Challenger's keypoints suitable for COCO ,and modify AI Challenger's annotation file with reference to COCO format, and put together their image. Finally combination into a bigger dataset, which have 328176 image, include 248561 person samples. We train HRNet by this dataset and get model2.

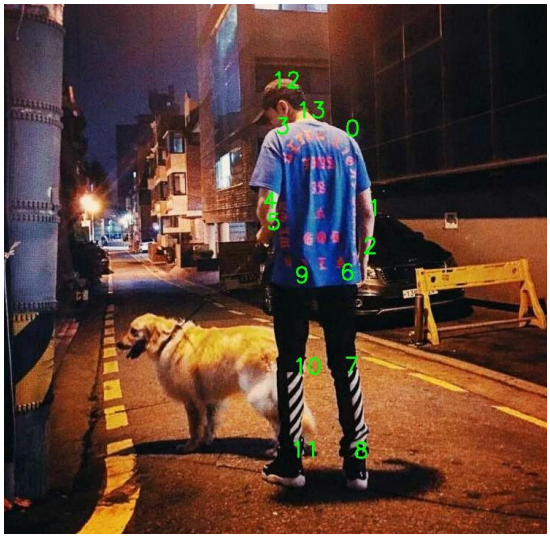

Figure 4. AI Challenger's keypoints

Figure 4 shows AI Challenger's keypoints, we can see that 12 keypoints(0~11) of AI Challenger are adapted to COCO dataset and two keypoints(12,13) are not fit. So our bigger dataset, every person have up to 17 keypoionts, and some of them will not enough 17. Such as, image from AI Challenger have up to just 17 keypoionts, and some person's keypoionts may be obscured ,thus will not enough 17. So the proportion of the total number of keypoiont is different, leading to data imbalance. We modify HRNet's loss function to balance our data. In every train batch, we will count the proportions of 17 keypoionts in this batch. And when calculating the loss corresponding to each key point, simply multiply the reciprocal of the specific gravity. So we can get a different loss function, an finally get our model3.

Result

Through the above work, we can get everyone's box and his keypoints, shown in Figure 5.

Figure 5. show one person's result

And than, we get 3 model from above work. These model's perform in COCO test-challenge in table 1.

Table 1. 3 model's perform

model	test-challenge
Model1	0.734
Model2	0.735
Model3	0.737

Related Attempts

Beside we also have another attempt based on HRNet. such as: 1.change it's basic network to ResNeXt; 2. public HRNet bones network,combining direct measurement of 17 key points with thermal mapping and combining loss on both sides; 3.HRNet add deformable convolution.

But these attempts failure to achieve better performance, so we won't introduce it in this report.

References

[1]. https://arxiv.org/abs/1901.07518

[2]. https://arxiv.org/abs/1902.09212

[3]. A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In ECCV, pages 483 - 499,2016.

[4]. E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi person pose estimation model. In ECCV, pages 34 - 50, 2016

[5]. J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang,S. Zhou, G. Lin, Y. Fu, et al. Ai challenger: A large scale dataset for going deeper in image understanding. arXiv preprint arXiv:1711.06475, 2017.