Joint COCO and Mapillary Workshop at ICCV 2019: COCO Keypoint Challenge Track

Technical Report: ss_de

Zixuan Huang
Beijing University of Posts and Telecommunications
huangzixuan0508@bupt.edu.cn

Yuan Chang
Beijing University of Posts and Telecommunications
changyuan@bupt.edu.cn

Abstract

In recent years, convolution neural networks have significant progress on keypoint detection. However, lots of approaches require multiple upsampling of small featuremaps to produce the final output. The huge information loss in the upsampling process is the key to restrict the accuracy. So we use HRNet and dilated convolution with stride of one as two type backbone to do experiments. But what is most different from other methods is that we not only use keypoint heatmaps regressions, but also let the network predict trunk heatmaps and two values which are trunk lengths and angles to help keypoints detection. Using the length and position of trunk can help model to predict the difficult points and invisible points, and improve the accuracy of the final results.

1. Introduction

We firstly use dilated convolution to overcome the disadvantage of upsampling. Multi-scale resampling and attention mechanism are adopted to improve network's performance[1]. Based on the same idea, we find HRNet[2] has the same characteristics, less computation and better precision. So we finally use HRNet as backbone, and incorporate some dilated convolutions to improve the receptive field.

1.1. Data Generation

In order to avoid the influence of occluded objects, we use trunk information to assist keypoints prediction. The length and angle of trunk and the coordinates of joint point are used to infer the coordinates of the corresponding joint point.

$$S_{len} = \|S\| \tag{1}$$

$$S_{\theta} = arccos\left(\frac{S \cdot i_x}{\|S\| \, \|i_x\|}\right) \tag{2}$$

Nineteen skeletons are defined as follows. "skeleton": [(point16 point14); (point14 point12); (point17 point15); (point15 point13); (point12 point13); (point6 point12); (point7 point13); (point6 point7); (point6 point8); (point7 point9); (point8 point10); (point9 point11); (point2 point3); (point1 point2); (point1 point3); (point2 point4); (point3 point5); (point4 point6); (point5 point7)].

Firstly, heatmaps are generated for each skeleton. At the same time, the length and angle of each skeleton are calculated by the coordinates of the two endpoint. We normalize the length and calculate angle relative to the horizontal vector. Assume that the key points at ends of the skeleton s are a and b. Then

$$b_x = S_{len} * cos\left(S_{\theta}\right) + a_x \tag{3}$$

$$b_y = S_{len} * sin\left(S_{\theta}\right) + a_y \tag{4}$$

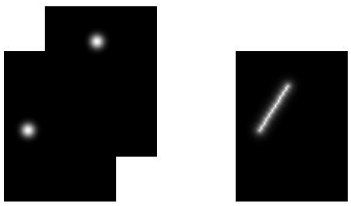

Figure 1: The skeleton heatmap from keypoints.

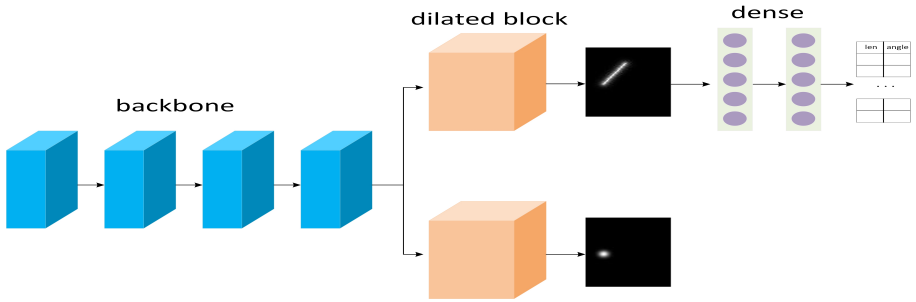

Figure 2: Our network architecture.

1.2. Network Architecture

The network backbone is based on HRNet with two head branches to predict keypoints and trunk respectively. These two branches share the same feature extraction layer. As shown below, the lower branch predicts heatmaps of keypoints. In order to predict the coordinates of keypoints and enable the network to backward, we replace maximum with mathematical integration to get the coordinates xj and yj, where j from 1 to m.

$$J = \int_{p \in \Omega} p \cdot H\left(p\right) \tag{5}$$

$$H\left(p\right) = \frac{e^{H(p)}}{\int_{q \in \Omega} e^{H(q)}} \tag{6}$$

The upper branch is responsible for the skeleton. After predicting heatmaps, the length and angle of the trunk are regressed through fullly connection layers. The starting keypoints coordinate matrix SP of each skeleton is obtained by multiplying the coordinates of the keypoints coordinate matrix P predicted by the lower branch and the transfer matrix T1 which indicate the start points of each skeleton. The coordinates of the end keypoints matrix EP can be obtained by adding the SP and skeleton matrix predicted S by the upper branch. The EP is multiplied with the transfer matrix T2 to get the revised coordinates of the key points which calculated by the skeleton. Then we average the revised coordinates with the original coordinates to get the final results. In order to improve the accuracy of the network and accelerate the convergence, we use skeletal heatmaps as intermediate supervision to the upper branch.

$$P_{new} = \left(PT_1 + S\right) T_2 \tag{7}$$

$$p_{final} = \frac{P + P_{new}}{1 + W} \tag{8}$$

2. Experiment

Our models are only trained on COCO2017 dataset, no extra data involved. The experiment is divided into two parts, only using keypoints regression and supplementing with trunk information. The experiment results are shown as below.

Table 1: The Comparison of two methods on the validation dataset using GT bbox.

Model	Input Size	AP
HRNet-W32 without skeleton branch	384×288	0.776
HRNet-W32 with skeleton branch	384×288	0.785

It can be seen that the AP has been improved after adding skeleton branch.

3. Conclusion

Due to the late participation in the competition, we have not completed the improvement of the whole network. We eventually plan to design an end-to-end model which can fuse the two outputs of network and enable network to learn information through trunk and keypoint. In this way model can predict the final result without manual computation. We will continue this research after the competition.

References

[1] Yuan Chang, Zixuan Huang, and Qiwei Shen. The same size dilated attention network for keypoint detection. In *International Conference on Artificial Neural Networks*, pages 471–483. Springer, 2019. 1

[2] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. *arXiv preprint arXiv:1902.09212*, 2019. 1