Joint COCO and Mapillary Workshop at ICCV 2019: Pixel Shuffle Module for Human Pose Estimation Challenge Track

Technical Report: keypoints estimation

Shuchun Liu, Feiyun Zhang, Li Long, Weiyuan Shao, Jiajun Wang
The AI Lab of ELEME Inc, China
{shuchun.liu,feiyun.zhang,li.long02,weiyuan.shao,jiajun.wang}@ele.me

Abstract

This report is our submission for the task of keypoints estimation(COCO) 2019. In this task, we use Pixel Shuffle Moudle(PSM) for human pose estimation. We propose PSM to get high resolution heatmaps for generating the final results, which uses smaller channels in down-sampling module. At last, the single model can get mAP 76.2%, which train and test on MS COCO dataset without external data and ensemble models.

1. Introduction

2D Human pose estimation has drawn increasing attention from the research community in recent years by using deep neural networks. Currently one of the best performing methods can be categorized into top-down methods[2, 3, 7, 8, 9, 10] and bottom-up methods[1, 5, **?**]. Firstly, Top-Down methods locate persons. We get persons' bounding boxes by using detectors, then do pose keypoints estimation with a single person box as input. Top-down methods are generally less sensitive to the scale variance of person,since they normalize all the persons to approximately the same scale by cropping and resizing the person bounding boxes. These methods count on independent pedestrian detectors,which are normally computationally intensive and not truly end-to-end systems.On the contrarythe bottom-top methods first locate identity free keypoints for all persons in the images by predicting heat maps of different anatomical keypoints, followed by grouping them into person instances. So bottom-top methods usually are faster and more capable of achieving real-time pose estimation. The top-bottom methods sacrifice time for better accuracy because every person in the image will do a key points inference separatelythe bottom-up methods can be faster by do a key points inference just once which do well in localizing keypoints precisely for large persons while inaccurate for smaller persons.

2. Ours Methods

2.1. Backbone

Top-down methods are used to achieve high accuracy of pose key points estimation in our strategy. We have found that, in each module of Hourglass, the number of convolutional filters (or feature maps) remains constant during repeated down and up sampling steps. This equal-channel-width design results in a relatively poor performance since a lot of information will be lost after every down sampling. On the contrary, in [[11]] the number of feature maps is increased when using a down sampling. [4](MSPN) Using the same smaller number of channels in up sampling modules and using a increase number of feature maps in down sampling. As shown in Fig. 1,We adopt the same small number of feature maps in up-sampling modules as in MSPN. In down sampling modules, the number of feature maps is doubled after every spatial down sampling. It is reasonable since we aim to extract more representative features in the down sampling process and the lost information can hardly be recovered in the up sampling procedure. Therefore, increasing the capacity of down sampling unit is usually more effective. Most existing human pose estimation methods predict Gaussian-smoothed heatmaps by preparing the ground truth headmaps with an unnormalized Gaussian kernel applyed to each keypoint location. However, applying a Gaussian kernel also introduces confusion in precise localization of keypoints. A trivial solution to reduce this confusion is to reduce the std of the Gaussian kernel. However, we empirically find that it makes optimization harder and leads to even worse results. As in MSPN, they use the highest resolution (1/4 of the input image) feature maps for prediction in their experiments and the mean deviation in the x and y directions which may reduce the accuracy. So we use HRNet as backbone together with PSM, which build HRNet on top of the highest resolution feature maps(1/2 of the input image). We also try to use the highest resolution (the same shape of the input image) feature maps for predic-

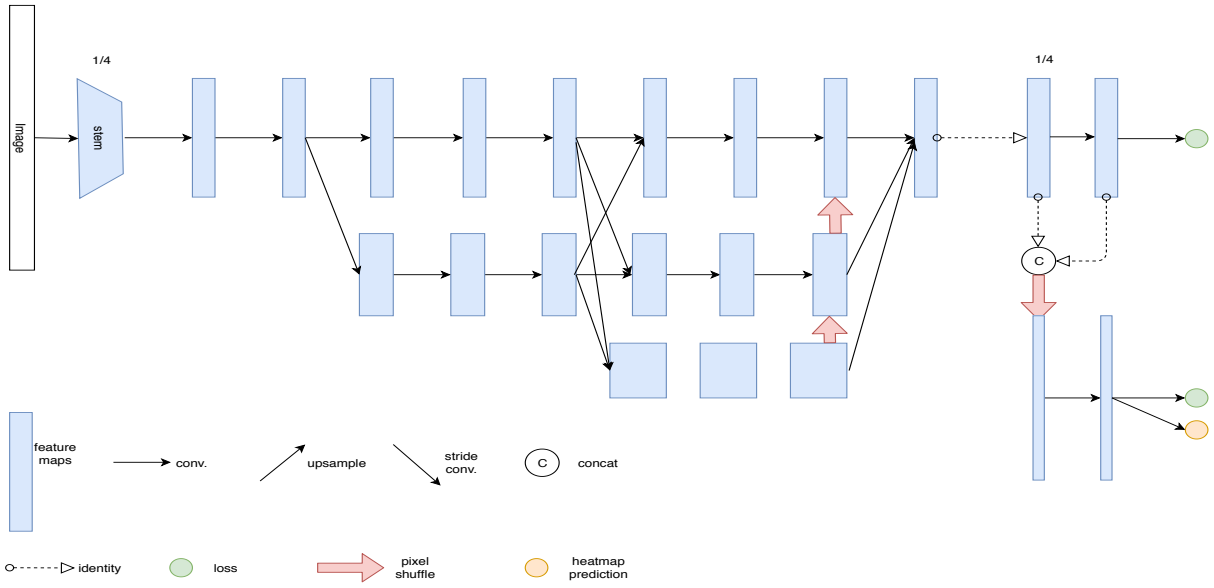

Figure 1: The architecture of PSM.

tion, but worse results are got. We believe that there is an artificial annotation error, and the original image size will lack robustness.

2.2. Pixel Shuffle Moudle

Different from HRNet [8], in fusion step of multi-scale feature maps, we use pixel shuffle module(PSM), instead of upsample or deconvolution. In pixel shuffle module, as shown in Fig. 2 the down-sample feature maps with double channels(2c). Firstly, changing their channels equal to the previous layer(c) by 1x1 convolution, with 2c filters. Then using pixel shuffle to change the shape of the output feature maps to the same shape of the previous layer. In the third step, 1x1 convolution is added in every previous layer and the pixel shuffle maps and using element-wise addition to fuse them. With this design, the current stage can take full advantage of prior information to extract more discriminative representations. In addition, the feature aggregation could be regarded as an extended residual design, which is helpful dealing for with the gradient vanishing problem.

Due to time constraints, we only use MS COCO dataset for training, validation and test. OKS-based mAP (AP for short) is used as our evaluation metric. We adopt our own object detector to generate person boxes, which in trained with full categories of MS COCO dataset. Each image will randomly go through a series of data augmentation operations including cropping, flipping, rotation, and scaling. As for cropping, instances with more than eight joints will be cropped to upper or lower bodies with equal possibility. The rotation range is -45° ∼+45°, and scaling range is 0.7 ∼ 1.35. The image size is set 384x288 for training and testing. Following the same strategy as [6], we average the predicted heat maps of original image with results of corresponding flipped image. Then, a quarter offset in the direction from the highest response to the second highest response is implemented to obtain the final locations of keypoints. The pose score is the multiplication of box score

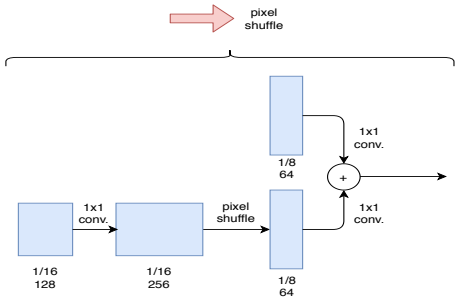

Figure 2: The pixel shuffle moudule.

Method	Backbone	Input Size	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}
CMU Pose[1]	-	-	61.8	84.9	67.5	57.1	68.2	66.5	87.2	71.8	60.6	74.6
Mask R-CNN[3]	Res-50-FPN	-	63	87.3	68.7	71.4	-	-	-	-	-	
G-RMI[7]	Res-152	353x257	64.9	85.5	71.3	62.3	70.0	69.7	88.7	75.5	64.4	77.1
AE[5]	-	512x512	65.5	86.8	72.3	60.6	72.6	70.2	89.5	76.0	64.6	78.1
CPN[2]	Res-Inception	384x288	72.1	91.4	80.0	68.7	77.2	78.5	95.1	85.3	64.6	78.1
Simple Base[10]	Res-152	384x288	73.7	91.9	81.1	70.3	80.0	79.0	-	-	-	-
HRNet[8]	HRNet-W48	384x288	75.5	92.5	83.3	71.9	81.5	80.5	-	-	-	-
MSPN[4]	4xRes-50	384x288	76.1	93.4	83.8	72.3	81.5	81.6	96.3	88.1	77.5	87.1
Ours(PSM)	HRNet-W48	384x288	**76.2**	92.6	83.6	72.4	82.3	81.1	95.6	87.7	76.9	87.0

Table 1: Comparisons of results on COCO test-dev dataset

and average score of keypoints, which is presented in [2]. In MSPN the down-sampling channel are [128,256, 386, 512], but in ours, the channels are [32,64,128,256], which is smller than MSPN. At last, the single model can get mAP 76.2%, which train and test on MS COCO dataset without extra data. The results of PSM are shown in Tab 1 .

3. Results

We propose PSM to get high resolutionWe heatmap for generating the final results, which uses smaller channels. what's more, we only train and test on MS COCO dataset, which is without extra data. Our single model can get map 76.2%. The smaller down-sampling channels may cause only a little rise of our accuracy.

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017. 1, 3

[2] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 3

[3] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 3

[4] Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, and Jian Sun. Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148*, 2019. 1, 3

[5] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in Neural Information Processing Systems*, pages 2277–2287, 2017. 1, 3

[6] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 2

[7] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 3

[8] 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, 2, 3

[9] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 529–545, 2018. 1

[10] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 466–481, 2018. 1, 3

[11] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 1