Joint COCO and Mapillary Workshop at ICCV 2019: COCO Keypoint Detection Challenge Track
Technical Report: Multi-Scale Channel Attention R-CNN

Myunggu Kang*
Clova AI
NAVER Corp.
myunggu.kang@navercorp.com

Dongyoon Wee*
Clova AI
NAVER Corp.
dongyoon.wee@navercorp.com

Abstract

Top-down pose estimation approach has achieved the state-of-the-art performance for multi-person 2D pose estimation because of its effective architecture to utilize multi-scale feature maps. However, top-down approach with single-model based method such as Mask R-CNN [2] has not reached the comparable performance with the separated-model methods.

We assume that this comes from an ineffective network structure to handle multi-scale feature maps. Based on this assumption, we design Multi-Scale Channel Attention(MSCA), a method to fuse multi-scale feature maps using channel-wise attention. In addition, we propose MSCA R-CNN framework based on Mask R-CNN [2] using MSCA-RoIAlign, which applies MSCA to extract RoI feature map. Experiments on MS COCO [6] dataset demonstrate that the proposed method improve the performance for multi-person 2D pose estimation.

1. Introduction

The multi-person 2D pose estimation is challenging because it requires to handle not only local information but also global one to detect poses in a given image. In order to manage both local and global information simultaneously, using multi-scale feature maps has become essential for pose estimation. Accordingly, the recent top-down pose estimation models have focused on developing an effective architecture to deal with multi-scale information.

Among top-down pose estimation approaches, separated model-based methods have outperformed single model-based methods with large margin in terms of accuracy. We assume that it comes from an ineffective network structure for exploiting multi-scale information. Based on this assumption, we explore a better method to use multi-scale information for a top-down single model-based method.

In this paper, we propose Multi-Scale Channel Attention(MSCA) R-CNN, which use MSCA for RoIAlign method to extract multi-scale feature maps. Combining with the channel attention method, Squeeze-and-Excitation Network(SENet) [3], MSCA R-CNN shows its effectiveness on multi-person 2D pose estimation task.

2. Related works

2.1. Top-down multi-person 2D pose estimation

There are two methods in the top-down multi-person 2D pose estimation; separated-model based method and single-model based model. [1, 11, 10, 4] are based on the separated-model based method. Those methods are based on single-person pose estimation models whose inputs are a raw image of a single person. Each person image is cropped by an extra human detector from a given images. They usually show the state-of-the-art performance compared to the other approaches by focusing its capacity on predicting a pose of a single person.

[2, 8] are the single-model based methods. Contrary to the separated-model based method, they integrate a human detector model and a single-person pose estimation model into a single model by using feature maps of each person from shared backbone instead of raw images. Accordingly those models don't require an separated human detector so that they have advantages on computational efficiency compared to the separated-model based method.

Mask R-CNN [2] proposed RoIAlign to pass feature maps from shared backbone to head networks. Based on Mask R-CNN, Multi-scale Aggregation(MSA) R-CNN [8] proposed the framework which consists of MS-RoIAlign and MS-KpsNet to use multi-scale information.

[*] indicate equal contribution.

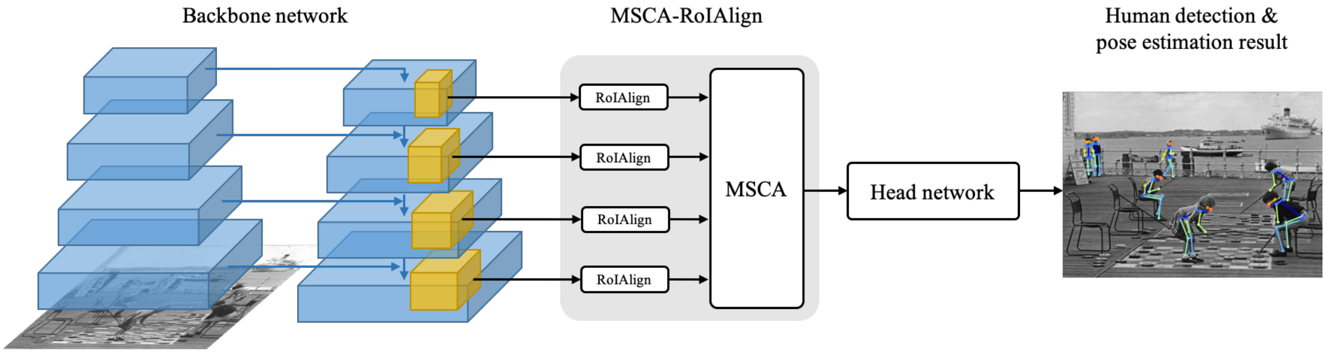

Figure 1: Overall structure of the proposed network. Input is a multi-person image and outputs are the position and pose of people. Feature maps is extracted using RoIAlign for the proposed roi from ResNet-FPN. These multi-scale RoIs are passed through the MSCA module to the head network. For simplicity, only one RoI is marked with an yellow square box.

2.2. Channel attention

Squeeze-and-Excitation Networks(SENet) [3] proposed the Squeeze-and-Excitation block(SE block) to adaptively recalibrate channel-wise feature responses. In the SE block, the squeeze operator aggregates channel-wise context then the excitation operator weight each channel adaptively based on the features aggregated previously.

3. Multi-Scale Channel Attention

For the scale invariant modeling, the ResNet-FPN framework [5] is widely used as backbone network to extract multi-scale feature maps. By passing its feature maps to a head network, it has demonstrated its effectiveness on many computer vision tasks such as image recognition, object detection and pose estimation.

In the Mask R-CNN [2], the RoIAlign block is used to pass feature maps of detected area from backbone network to head networks. The RoIAlign block of the original Mask R-CNN select a single-scale feature maps out of multi-scale feature maps based on the size of proposals.

On the other hand, MSA R-CNN [8] proposes the MS-RoIAlign to aggregate multi-scale feature maps of backbone network regardless of the size of proposals. The MS-RoIAlign add or concatenate the feature maps after convolutional filters and upsampling. Compared with the above RoI assignment strategy [2], it demonstrates the effectiveness of MS-RoIAlign by using all information from multi-scale feature maps of backbone.

Combining the advantages from the above methods, we propose the Multi-Scale Channel Attention-RoIAlign(MSCA), which can be applied to a RoIAlign block. In this paper, we use MSCA only for a RoIAlign block and name it MSCA-RoIAlign. Fig 1 illustrates the overall structure with MSCA-RoIAlign. MSCA-RoIAlign generate a RoI feature map from the backbone network, then pass it to the head network for human detection and pose estimation. Inside of MSCA-RoIAlign, MSCA fuses the multi-scale RoI feature maps extracted from each feature map of the backbone network. Fig 2 shows the detailed architecture of MSCA. In MSCA, each multi-scale features are upsampled to get the same spatial resolutions, then each feature goes through 1x1 convolutional filters to reduce the the number of channels of each feature map. After concatenating those multi-scale feature maps, we apply the channel attention to reweight channels according to the context of the aggregated feature map.

4. Implementation

Our model is based on the official Pytorch implementation of Mask R-CNN [7]. The model is trained in an end-to-end manner. The backbone network of our model is based on the ResNet-FPN and its weights are pretrained by ImageNet [9]. A mini-batch involves 4 images per each GPU and each image sampled 512 RoIs with positive-to-negative ratio of 1:3. For data augmentation, we trained using image scale randomly sampled between 640 and 800 pixels. For COCO dataset, max iteration is 60k, with learning rate of 0.02. Weight decay and momentum are set to 0.0001 and 0.9, respectively. In longer training, learning rate is decreased at 90k, 120k. Other details are identical as in Mask R-CNN [2]

5. Experiment

5.1. Dataset and evaluation metric

The proposed model is trained on MS COCO [6] training set without using any external or extra dataset. The training dataset contains 57K images including 150K person instances with keypoint annotations. The validation is performed on MS COCO validation set which includes 5K images and testing is executed on the MS COCO test-dev set including 20K images. We use the official MS COCO

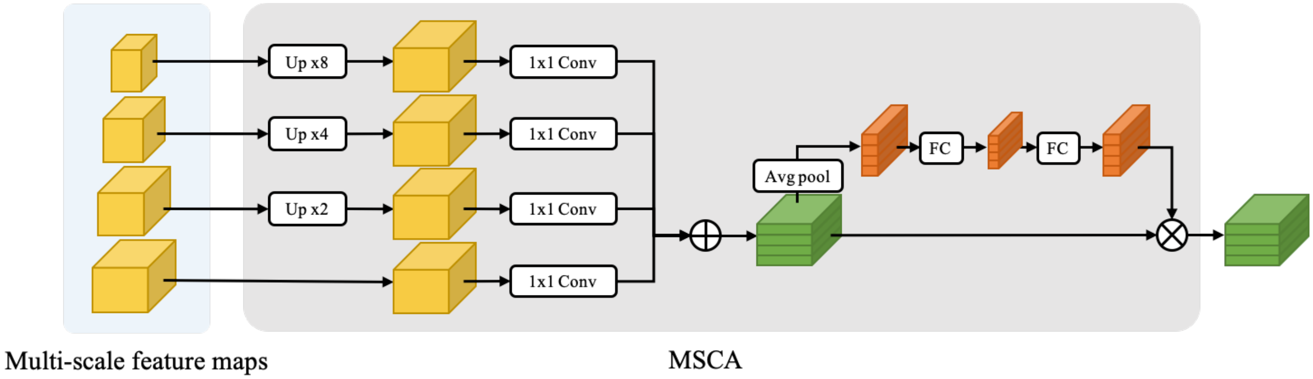

Figure 2: Multi-scale feature maps are fed into MSCA as input. Each feature map goes through upsampling, 1x1 conv, concatenation to get an aggregated feature map. This integrated feature map is multiplied with the branch that generates channel-wise attention, resulting in a reweighted feature map.

evaluation metric using average precision(AP) and average recall(AR). OKS and IoU based scores were used for keypoint and person detection tasks, respectively.

5.2. Ablation study

For ablation study, we trained our model on the MS COCO training set and validated on the MS COCO validation set. To demonstrate the effectiveness of our proposed model, we compare the performance with the baseline model[2] in table 1.

5.2.1 Baseline

We use our Pytorch implementation based on the official Pytorch Mask R-CNN implementation [7] as the baseline. The performance of our baseline is slightly improved compared to that of Mask R-CNN [2].

5.2.2 Rotation

Though it is widely known that rotation augmentation helps to improve the performance object detection and keypoint estimation task, the Pytorch implementation of Mask R-CNN [7] doesn't include rotation augmentation in the dataset augmentation methods. In order to improve the performance for both human detection and keypoint estimation, we apply rotation augmentation to entire each image instead of cropped single-person image. Degrees for rotation is randomly chosen from 0 to 45. Table 1 shows that rotation augmentation helps to improve all performances, especially 1.1 AP for keypoint.

5.2.3 MS-KpsNet

We adopt MS-KpsNet [8] as the keypoint head for this experiment. Experimental result shows that MS-KpsNet improves 1.9 AP for keypoint compared to the original keypoint of Mask R-CNN [2]. We recommend to refer [8] for further details.

5.2.4 Longer training

For stable convergence of larger keypoint head, we scaled the training schedule by approximately 1.5 times. Longer training slightly improves all performances.

5.2.5 MSCA-RoIAlign

For this experiment, we apply SENet [3] as the channel attention method with reduction ratio, 16. Table 1 shows that MSCA-RoIAlign improves all performances both keypoint and bounding box detection.

5.2.6 Test augmentation

To improve the performance, we apply the test augmentation methods including multi-scale augmentation, horizontal flips.

5.3. Comparison with state-of-the-art methods

We compare the performance of the proposed model with that of the recent state-of-the-art methods including the separated-model based methods and the single-model based methods. Methods that involve extra training data or use ensemble technique are excluded from this comparison. Table 2 demonstrates the performance comparison on MS COCO test-dev dataset. The proposed method achieve the state-of-the-art performance among the single-model based methods with ResNet-50-FPN backbone.

Methods	AP^{kps}	$AP^{kps}_{.50}$	$AP^{kps}_{.75}$	AP^{kps}_{M}	AP^{kps}_{L}	AP^{bb}	$AP^{bb}_{.50}$	$AP^{bb}_{.75}$	AP^{bb}_{S}	AP^{bb}_{M}	AP^{bb}_{L}
Baseline	64.3	85.9	69.6	59.3	72.4	53.8	82.9	58.6	36.8	61.6	69.6
+ Rotation	65.4	86.8	71.1	60.4	73.7	54.1	83.1	59.2	36.7	61.9	70.3
+ MS-KpsNet	67.3	87.5	73.3	62.4	75.4	54.1	83.0	59.1	36.6	62.0	70.2
+ Longer training	67.6	87.6	73.8	62.8	75.7	55.0	83.6	60.4	37.5	63.0	71.1
+ MSCA-RoIAlign	68.2	87.8	73.9	63.5	76.2	55.6	84.1	61.0	37.6	63.7	72.0
+ Test augmentation	**70.3**	**88.7**	**77.3**	**65.7**	**78.5**	**57.2**	**84.6**	**62.9**	**40.4**	**65.6**	**72.8**
	+6.0	**+2.8**	**+7.7**	**+6.4**	**+6.1**	**+3.4**	**+1.7**	**+4.3**	**+3.6**	**+4.0**	**+3.2**

Table 1: Effect of various settings in terms of the performance on the MS COCO validation set. AP^{bb} means the average precision of detection task for the human class only.

Methods	Backbone	AP	$AP_{.50}$	$AP_{.75}$	AP_M	AP_L	AR	$AR_{.50}$	$AR_{.75}$	AR_M	AP_L
Separated Top-down: Human detection and single person keypoint detection											
CPN [1]	Res-Inception	72.1	91.4	80.0	68.7	77.2	78.5	95.1	85.3	74.2	84.3
Simple Baseline [11]	Res-152	73.8	91.7	81.2	70.3	80.0	-	-	-	-	-
HRNet [10]	HRNet-W48	75.5	92.5	83.3	71.9	81.5	80.5	-	-	-	-
MSPN [4]	4-Res-50	**76.1**	**93.4**	**83.8**	**72.3**	**81.5**	**81.6**	**96.3**	**88.1**	**77.5**	**87.1**
Single Top-down: Human detection with keypoint detection											
Mask R-CNN [2]	Res-50	63.1	87.3	68.7	57.8	71.4	-	-	-	-	-
MSA R-CNN [8]	Res-50	68.2	**89.7**	75.0	63.8	75.6	74.4	93.4	80.3	69.2	81.5
MSCA R-CNN (our)	Res-50	**68.6**	89.6	**75.5**	**64.1**	**76.3**	**75.5**	**94.0**	**81.6**	**70.5**	**82.3**

Table 2: Comparison with the state-of-the-art methods on MS COCO test-dev dataset. Methods that involve extra training data or use ensemble technique are excluded.

6. Conclusion

We propose MSCA and MSCA R-CNN framework and demonstrate its effectiveness on MS COCO [6] keypoint detection task. Due to its simplicity, we expect to use MSCA for other computer vision tasks, which require to deal with multi-scale feature maps.

References

[1] 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, 4

[2] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 2, 3, 4

[3] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 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, 4

[5] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[6] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 2, 4

[7] Francisco Massa and Ross Girshick. maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. https://github.com/facebookresearch/maskrcnn-benchmark, 2018. Accessed: [Insert date here]. 2, 3

[8] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Multi-scale aggregation r-cnn for 2d multi-person pose estimation. *arxiv:1905.03912*, 2019. 1, 2, 3, 4

[9] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014. 2

[10] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 1, 4

[11] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 4