Deep Layer and Spatial Aggregation neural network for human pose estimation

Deyuan Zhang¹*, Chao Weng², Haoguang Wang³, Xiangbin Shi⁴

¹Shenyang Aerospace University, Shenyang, Liaoning, China
²Shenyang Aerospace University, Shenyang, Liaoning, China
³Shenyang Aerospace University, Shenyang, Liaoning, China
⁴Shenyang Aerospace University, Shenyang, Liaoning, China

* Deyuan Zhang’s e-mail: dyzhang@sau.edu.cn

Abstract. The simplebaseline model achieves high performance of human pose estimation with simple network structure. But the model lacks the layer and spatial information fusion. In this paper, we propose DLSAnet, which fuse layers and spatial information effectively. DLSAnet uses DLA as backbone which has excellent feature extraction capabilities in the field of object detection. In addition, a modified spatial pyramid pooling is introduced to pool and connect multi-scale local area features, allowing the network to learn object features more comprehensively. Using a four-branch SPP module instead of a single-branch SPP module connected by a single hopping layer. This method is effective in alleviating the problem of slow loss drop late in training. Experiments show that DLSAnet can achieve better accuracy.

1. Introduction

Multi-person pose estimation refers to the recognition and localisation of key points of all people in an image. It is a fundamental research topic for many vision application such as human action recognition and human-computer interaction. It has attracted increasing attention in the last decade and has been widely used in applications such as human-computer interaction, motion analysis, augmented reality, and virtual reality.

There has been a lot of research on this in recent years. Simplebaseline [1] is a simple method among these methods, which uses Resnet50 [2] for feature extraction, followed by three deconvolution operations, The final connected output is 16 feature maps connected to achieve the estimation of the heat map of human nodes. Although resnet50 has incorporated skip connections to combine the information from the layers, these connections are inherently "shallow" and can only be fused in a simple one-step operation. We propose to enhance the standard architecture with a deeper aggregated DLA as the backbone network to better fuse information across layers. The DLA has been shown to be better at obtaining rich semantic and spatial information in networks such as segmentation, target detection, classification, etc. As DLA60 [3] enables deep aggregation the network has better accuracy and feature extraction capabilities compared to the existing branch and merge scheme of resnet50. Moreover, the deep aggregation structure of DLA60 iterates and merges the feature hierarchy in a hierarchical manner, Enables better feature extraction, and for this reason we use DLA60 as the backbone network structure for this experiment. Existing research on the improvement of the simplebaseline family of methods is still scarce and its network structure is not very capable of feature extraction while also failing to take
full advantage of the problem of multi-scale local area features in terms of reduced feature resolution caused by successive convolution operations or convolution spans, which allow DCNNs to learn increasingly abstract feature representations. However, this invariance to local image transformations may hinder dense prediction tasks, where detailed spatial information is required. To overcome this problem, we therefore introduce an improved spatial pyramidal maximum pooling SPP module that collects and joins local area features from different scales in the same convolutional layer to learn multi-scale object features more comprehensively, enabling the model to obtain richer semantic information for more effective feature extraction and human keypoint estimation. Experiments are conducted on the MPII dataset and the COCO human keypoint dataset, and the experimental results show that the method model can achieve better results than the Simplebaseline model.

2. Materials and Methods

2.1 DLSAnet Framework

The DLSAnet structure proposed in this paper is shown in Figure 1, and it consists of three modules respectively, the feature extraction module Deep Layer Aggregation(DLA), the feature enhancement module SPP, and the final prediction module. The DLA network structure [3] used in object detection as a feature extraction module. DLA is composed of Iterative Deep Aggregation (IDA) and Hierarchical Deep Aggregation (HDA). IDA focuses on fusing resolutions and scales while HDA focuses on merging features from all modules and channels. IDA follows the base hierarchy to refine resolution and aggregate scale stage-bystage. HDA assembles its own hierarchy of tree-structured connections that cross and merge stages to aggregate different levels of representation.

Figure 1. Overall Network Structure Map

The DLA backbone is shown in Figure 2. As Figure 2 shows, DLA contains six modules, namely the level0, level1, level2, level3, level4, level5 modules. The DLA is constructed from 60 convolutional layers, level0 and level1 are each composed of one layer of 3*3 convolution with a stride of 1 and 2 respectively. The input feature map size is halved after level1. The red boxes represent HDA, Which is hierarchy linked with a tree structure that enables better propagation of features and gradients. The yellow link represents the IDA, which is responsible for linking the features of two adjacent level modules to allow for better integration of deep and shallow representations. The green box represents the aggregation of the output of the tree structure. The blue link represents the downsampling that was performed, and the network was initially downsampled as quickly as ResNet.
Figure 2. DLA feature extraction network structure [3]

For level2, level3 and up to level5 module, we set a parameter for each level to control the number of recursive iterations. The trees within each level module are connected by a stride of 1, and the trees are linked by a residual structure similar to ResNet [2] to obtain richer semantic information. The different level modules are connected to each other by a stride of 2. For the root module of different levels we used layer hopping to connect them, and fused them in the direction of resolution and scale to improve the model's ability to infer "where", thus achieving spatial fusion. The DLA feature extraction results in an H/32*W/32 feature map of size.

After passing the backbone network DLA we obtained a feature map of size H/32*W/32, followed by bilinear interpolation based upsampling to obtain a higher resolution feature map of size H/16*W/16. SPP feature enhancement module is followed to achieve the fusion of multiple scales, and the concatenate operation is used to add the feature map of each branch to obtain richer semantic and spatial information. The resulting output channel number is 4096, and the resulting feature map of constant size H/16*W/16 is then convolved by a convolution layer of size 3*3 with a stride of 2 with 1024 output channel. The size of the feature map and the number of channels are adjusted to reduce the number of parameters in the model. Finally, a H/4*W/4 feature map is obtained by three times of deconvolution, and then a 1*1 convolution layer is used to output a headmap of the nodes.

2.2 SPP (Spatial Pyramid Pooling Structure) module

The Simplebaseline network structure prediction focuses on the global features of the convolutional layer, ignoring the fusion of multi-scale local area features on the same convolutional layer, and ignores feature enhancement after feature extraction. Therefore, this paper use a spatial pyramid pooling block and introduces it into the network structure we designed for pooling and connecting multi-scale local area features, while exploiting both global and local multi-scale features to improve the accuracy of detecting nodes. The module draws on the idea of spatial pyramids, for which we design the SPP module consisting of pooling layers of different kernel size sizes, and in this experiment we place the SPP module as a feature enhancement module after the feature extraction module Deep Layer Aggregation. the SPP module can consist of multiple parallel branches, and the best branch design structure for the SPP module is illustrated in Figure 3 below is shown.
Figure 3. Space pyramid pooling SPP module diagram

The DLA60 feature extraction results in a feature map of size H/32*W/32, followed by a deconvolution to make the feature map size H/16*W/16, and then the feature enhancement module SPP. The use of concatenation enables the number of channels to be combined, allowing the number of features describing the image itself to be increased, making full use of the semantic information in the feature maps at different scales.

3. Experiment

3.1. Experimental Setup

Two datasets MPII and COCO are conducted in this paper. The MPII Human Pose dataset consists of images taken from a wide-range of real-world activities with full-body pose annotations. There are around 25K images with 40K subjects, where there are 12K subjects for testing and the remaining subjects for the training set. The data augmentation and the training strategy are the same to MS COCO, except that the input size is cropped to 256*256 for fair comparison with other methods. The COCO dataset contains over 200,000 images and 250,000 person instances labeled with 17 keypoints. We train our model on COCO train2017 dataset, including 57K images and 150K person instances. We evaluate our approach on the val2017 set and test-dev2017 set, containing 5000 images and 20K images, respectively.

This model uses data including scaling (±30%), rotation (±40 degrees) and flipping. Our DLA backbone is by pre-training on the ImageNet classification task. In the pose estimation training, there are 140 epochs in total, the base learning rate is 1e-3. It drops to 1e-4 after 90 epochs and to 1e-5 at 120 epochs, the batch size is set to 32, the Adam is used. The GPU server uses a GPU Tesla V100 graphics card with 16G of video memory, the datasets used for the tests were the MPII and COCO test datasets.
3.2 Experimental result analysis

This paper presents a simple human pose estimation method based on a deep aggregation network, which iterates and merges feature levels in a hierarchical manner through a deep aggregation structure, enabling better feature extraction than Resnet50, and uses the SPP module for feature augmentation to better estimate the location of the nodes. To verify the validity of the model, tests were carried out on the MPII dataset.

We have experimented with different combinations of SPP modules for multiple branches and we found that in the case of multiple branches, the final prediction result is better as the size of the maximum pooled kernel_size is selected, the closer the size of the maximum pooled kernel_size is to the input feature map size, the better the prediction result is obtained. For this reason, we made a separate SPP module for the two-branch case with a maximum pooled kernel_size of 15 and 1. The prediction results are basically the same as in the multi-branch case, and the computational effort of the whole model is reduced. However, it is difficult to reduce the loss in the later stages of training, which increases the training time compared to the multi-branch SPP module. We then ran a few more sets of experiments, again using the MPII dataset. We added the SPP module to the simplebaseline model and found that the predictions were lower than the original model. As same as we have also explored kernel size in the spp multi-branch case. In the multi-branch SPP module we made the difference in size of each kernel_size increase. We find that the larger the kernel_size pooled in one of the branches is, the closer the size of the input feature map is to the multi-branch SPP module. The predicted results will be better. Finally, the best multi-branch combinations of 1*1, 9*9, 13*13, 15*15 SPP modules were obtained. By increasing the receptive field and then fusing the outputs under different receptive fields, more feature information can be retained. The spp module is used to achieve the fusion of local features and global features (so the largest pooling kernel of the spatial pyramid pooling structure should be as close as possible to the size of the featherMap to be pooled) at the featherMap level to enrich the expressiveness of the final feature map.

| Table 1: Comparisons of final results on MPII dataset |
|-----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| DLA [3]         | 96.61          | 95.35          | 88.65          | 82.83          | 87.17          | 82.49          | 78.06          | 87.93          | 34.09          |
| Simplebase-Resnet50 | 96.33          | 95.35          | 88.93          | 83.23          | 88.43          | 83.94          | 79.54          | 88.56          | 33.97          |
| Simple+SPP      | 96.83          | 95.26          | 88.47          | 82.78          | 87.12          | 83.31          | 78.86          | 87.71          | 33.05          |
| Yang & Ramanan  | 73.33          | 56.36          | 41.33          | 32.25          | 36.33          | 33.2           | 34.66          | 43.24          | 12.49          |
| Pishchulin et al| 74.37          | 49.01          | 40.85          | 34.23          | 36.55          | 34.5           | 35.17          | 41.43          | 13.67          |
| Tompson & Goroshin | 96.13          | 91.92          | 83.93          | 77.84          | 80.91          | 72.53          | 65.15          | 56.47          | 20.14          |
| DLSAnet         | 97.06          | 95.69          | 89.15          | 83.84          | 88.63          | 84.04          | 80.07          | 88.65          | 34.27          |

With the above experiments we obtained the best design of the SPP model and we compared the prediction results with models such as Simplebaseline on the MPII dataset. We found that the experimental results all outperformed the other models, demonstrating the simple superiority of our DLA+SPP network structure.

For comparison with other models, we further pre-trained our best SPP model on the MS-COCO dataset. Simplebaseline uses Resnet50 as the backbone network, meaning that the method involves extra data for training. Specifically, FAIR Mask R-CNN involves distilling unlabeled data, oks uses AI-Challenger keypoints dataset, bangbangren and G-RMI use their internal data as extra data to enhance performance, "+" indicates results Bathsize is set to 32 and the initial learning rate is set to 0.001. Our model outperforms other models in terms of AP by comparing the prediction results with other models.

| Table 2: Comparisons of final results on COCO test-challenge2017 dataset |
|-----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| Methods         | AP  | AP@.5 | AP@.75 | APm | APt | AR | AR@.5 | AR@.75 | ARm | ARt |
| CMU-Pose[9]     | 61.8 | 84.9  | 67.5  | 57.1 | 68.2 | 66.5 | 87.2  | 71.8  | 60.6 | 74.6 |
| G-RMI [8]       | 64.9 | 85.5  | 71.3  | 62.3 | 70.0 | 69.7 | 88.7  | 75.5  | 64.4 | 77.1 |
| G-RMI* [8]      | 69.1 | 85.9  | 75.2  | 66.0 | 74.5 | 75.1 | 90.7  | 80.7  | 69.7 | 82.4 |
4. Conclusion
In this paper, a simple human pose estimation based on a spatial pyramid pooling SPP module is proposed to address the problem that Simple baseline has low feature extraction capability and does not make full use of multi-scale local area features. In this experiment, a new spatial pyramid pooling is designed and introduced to collect and connect multi-scale local area features for comprehensive learning of object features. A multi-branch SPP is used instead of a two-branch SPP for feature enhancement, alleviating the problem of vanishing gradients and accelerating model training. Experiments on the MPII dataset and the COCO2017 dataset show that our model is more accurate than the simple baseline model, and it is able to achieve higher accuracy in terms of joint point recognition accuracy for the human body.

Acknowledgments
This work was supported in part by Liaoning Provincial Department of Education Science Foundation (JYT2020050)

References
[1] Xiao, B., Wu, H., Wei, Y. (2018) Simple Baselines for Human Pose Estimation and Tracking. In: 16th European Conference on Computer Vision. Munich. pp. 466-481.
[2] He, K., Zhang, X., Ren, S. (2016) Deep residual learning for image recognition. In: the 29th IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas. pp. 770-778.
[3] Yu, F., Wang, D., Shelhamer, E. (2018) Deep Layer Aggregation. In: the 31th IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City. pp. 2403-2412.
[4] Newell, A., Yang, K., Jia, D. (2016) Stacked Hourglass Networks for Human Pose Estimation. In: 14th European Conference on Computer Vision. Amsterdam. pp. 483-499.
[5] Chen, C., Papandreou, G., Kokkinos, et al. (2018) Semantic Image Segmentation with Deep Convolutional Nets. In: Pattern Analysis and Machine Intelligence. J.Sci.Commun.,275: 41-48.
[6] Chen, Y., Wang, Z., Peng, Y., et al. (2017) Cascaded Pyramid Network for Multi-Person Pose Estimation. In: the 30th IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City. pp. 7103-7112.
[7] Adrian, B., Georgios, T. (2016) Human Pose Estimation via Convolutional Part Heatmap Regression. In:14th European Conference on Computer Vision. Amsterdam.pp. 717-732.
[8] Papandreou, G., Zhu, T., Kanazawa, N., Toshev, J., Tompson, C. (2017) Towards Accurate Multi-person Pose Estimation in the Wild. In: the 30th IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City. pp. 4903-4911.
[9] Cao, Z., Simon, T., S, Wei., Sheikh, Y. (2017) Realtime multi-person 2d pose estimation using part affinity fields. In: the 30th IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City.pp.812-825.