Human Pose Estimation

Mothilal Asokan [1] Murtadha Aljubran [1] Santosh Sanjeev [1]

[1]Mohamed bin Zayed University of Artificial Intelligence

Introduction

Human pose estimation (HPE), a practical and useful application where we estimate the pose of a person. We identify 13 different KeyPoints in a human body as shown in the figure 1.

Figure 1. KeyPoints in Human body that helps to identify pose.

Objective

The objective is to train and evaluate our pose estimation models (inspired from stacked hourglass model shown in figure 2) on the MPII Human Pose training set, a benchmark for HPE challenges. Our goal is to reduce the computational complexity of the provided baseline hourglass model without largely sacrificing the accuracy

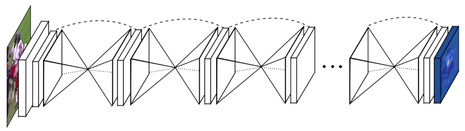

Figure 2. Stacked hourglass modules representing the whole architecture with skip connections.

Experiments

The dataset includes around 25K images containing over 40K people with annotated body joints. The baseline architecture explored in our work is inspired by the stacked hourglass design in [2].

- Weights applied to HG2 outputs.
- **Dilated Convolution** in Bottleneck.
- **Patchify** technique used in ConvNext [1].
- Bottleneck residual block **without skip connection** in Bottleneck.
- **Depthwise Separable Convolutions** in Bottleneck.
- **Convolution Layers** without Bottleneck.

Bottleneck module used in hourglass

X: number of feature maps, NxN: dimensions of image.

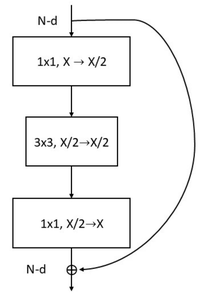

Figure 3. Baseline Bottleneck residual block.

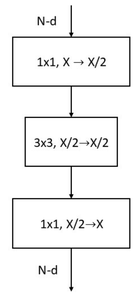

Figure 4. Bottleneck residual block without skip connection.

Different Hourglass Block Designs

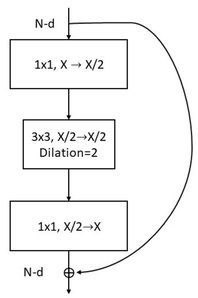

Figure 5. Dilated Convolution.

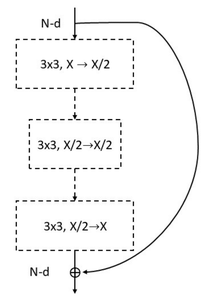

Figure 6. 3x3 Depthwise Separable Convolutions.

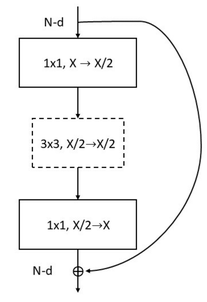

Figure 7. Depthwise Separable Convolutions in only 3x3 layer.

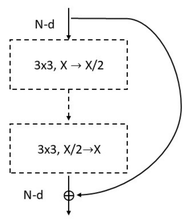

Figure 8. Two 3x3 Depthwise Separable Convolutions.

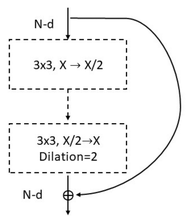

Figure 9. Two 3x3 Dilated Depthwise Separable Convolutions.

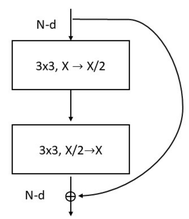

Figure 10. Convolution Layers without Bottleneck.

Results

Experiment Configuration	Validation Accuracy PCKh@0.5 (%)								No. of Parameters	Avg. Training Time(epochs*)
	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean		
Baseline	93.18	89.16	79.10	72.61	76.89	69.72	62.66	77.84	6.73M	2hr 30min(20)
Weights [0.3, 0.7] applied to HG2 outputs	92.50	88.71	77.77	70.64	75.71	67.12	60.04	76.33		
Weights [0.7, 0.3] applied to HG2 outputs	90.35	81.69	69.17	58.88	66.94	56.62	49.46	67.84		
Dilated Convolution	92.84	87.36	76.61	69.18	73.86	68.15	61.22	75.85		
Patchify technique from ConvNext	91.78	87.16	77.57	70.22	73.31	67.48	60.18	75.66	6.72M	2hr 34min(20)
Residual block without Skip connection in Bottleneck	92.84	88.03	78.32	71.42	75.97	68.31	61.36	76.84	6.69M	2hr 28min(20)
3x3 Depthwise Separable Convolutions	93.25	89.86	80.55	72.57	78.05	70.56	64.27	78.67	2.92M	2hr 20min(17)
Depthwise Separable Convolutions in only 3x3 layer	93.11	88.67	78.73	71.50	75.89	68.81	61.83	77.15	2.80M	2hr 30min(20)
Two 3x3 Depthwise Separable Convolutions	92.19	85.09	73.55	63.80	72.17	62.42	54.77	72.19	1.90M	2hr 40min(20)
Two 3x3 Dilated Depthwise Separable Convolutions	91.30	84.17	71.83	62.36	71.59	61.13	54.77	71.28		
Convolution Layers without Bottleneck	92.67	89.50	80.65	73.19	78.26	70.42	62.97	78.52	17.95M	2hr 10min(17)

This table summarizes our experiments in investigation the performance and training time of different hourglass block designs. All experiments had the same settings of no. of epochs, LR scheduler, and batch size.
*20 epochs or minimum epochs to reach baseline mean score.

Conclusion

The number of layers proved to be an important factor to consider. We were able to use a bigger network to achieve better performance with less training time.

References

[1] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022.

[2] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 483–499, Cham, 2016. Springer International Publishing.