Iterative Subnetwork with Linear Hierarchical Ordering for Human Pose Estimation

#1955

Shek Wai Chu[1], Chaoyi Zhang[1], Yang Song[2] and Weidong Cai[1]

[1] School of Computer Science, The University of Sydney, NSW 2006, Australia

[2] School of Computer Science and Engineering, University of New South Wales, NSW 2052, Australia

Introduction

Most recent developments in Human Pose Estimation (HPE) can be summarized into two main ideas: 1) refinement subnetwork to improve predictions iteratively and 2) exploitation of human joint graphical relations. In this work, we present how efficient and simple iterative subnetworks (IS) with linear hierarchical ordering (LHO) based on the ideas can help to improve accuracy on strong backbone models.

Overviews

LHO:

The prediction of a particular keypoint should depend on previous predictions of its adjacent keypoints: (head top, upper neck, thorax, left shoulder, right shoulder, pelvis, left elbow, right elbow, left hip, right hip, left wrist, right wrist, left knee, right knee, left ankle, right ankle).

IS:

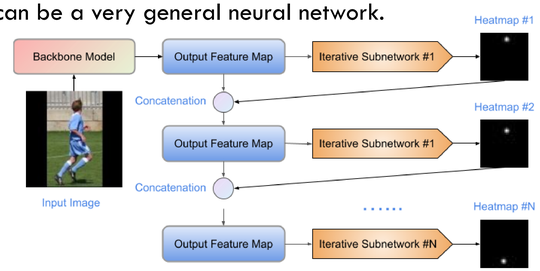
The output feature maps generated from the backbone network will be passed to an iterative subnetwork for production. of a single heatmap.

$$H_i = F_i^{IS}(O_i)$$
$$O_i = concat(O_{i-1}, H_{i-1}) \quad (1)$$

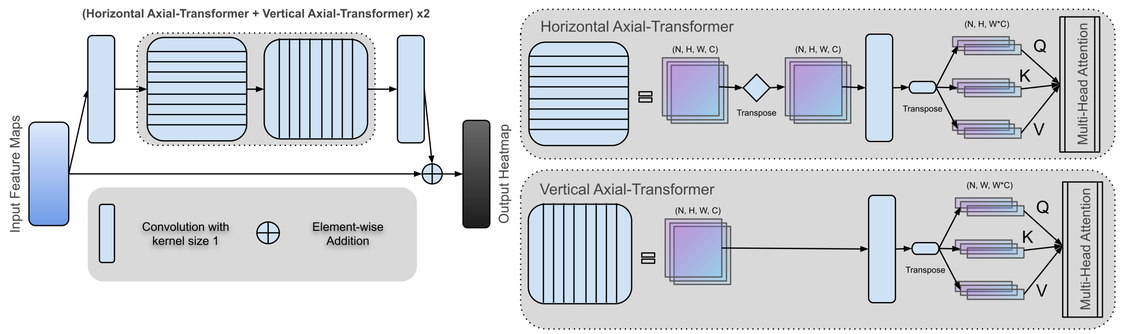
where H_i is the i^{th} output heatmap for joint keypoint i, O_i is the input feature maps for the i^{th} iterative subnetwork, and F_i^{IS} is the iterative subnetwork for i^{th} heatmap. Note that O_1 is the output feature map generated by a backbone network and $N = 16$ for human pose estimation.

Methods and Results

The iterative subnetworks can be a very general neural network.



In this study, we consider three different versions for IS block construction: 1) Simple Convolution, which represents basic building layers in deep learning studies; 2) Residual Block, which is a successful and very popular design for various computer vision tasks; and 3) Axial-Transformer, which is one of the efficient ways to apply Transformer in computer vision.



Below is our result on MPII held-out test set:

Comparisons of PCKh@0.5 on MPII held-out test set. LHO represents linear hierarchical ordering. HRNet is adopted as the backbone network for all our IS designs.

Method	Head	Shoulders	Elbows	Wrists	Hips	Knees	Ankles	Total
HRNet	97.5	96.1	90.5	86.0	88.8	85.6	81.6	89.9
IS_{Conv} w/o LHO	**97.7**	96.2	90.8	**86.3**	89.8	86.2	81.6	90.2
IS_{Conv} w/ LHO	**97.7**	**96.3**	90.6	**86.3**	89.8	86.5	81.8	90.3
IS_{Res} w/ LHO	**97.7**	96.2	90.8	**86.3**	89.5	86.6	**82.1**	90.3
IS_{AT} w/ LHO	97.3	96.0	**91.2**	86.2	**90.1**	**87.2**	81.5	**90.4**

Conclusions

In this work, we demonstrate that a linear hierarchical ordering of the predicted keypoints with simple iterative subnetwork can improve the prediction accuracy of human pose estimation. The experimental results suggest that we can achieve similar improvement to accuracy on strong backbone networks without an unnecessarily large complex refinement network. Significant improvements were observed in the difficult keypoints such as hips and knees. Both keypoints have strong dependence on their connecting keypoints, which shows that IS and LHO are effective at improving the prediction by utilizing adjacent keypoint heatmaps. Larger receptive field subnetwork such as Axial-Transformer can be used to further improve the accuracy.

Acknowledgments

I would like to express my special thanks of gratitude to my supervisors Prof. Cai and Dr. Song as well as my mentor Chaoyi for their valuable suggestions.

Further information

If you have any questions or comments, please feel free to contact me via schu9751@uni.sydney.edu.au.

Introduction

- Most recent developments in Human Pose Estimation (HPE) can be summarized into two main ideas: 1) refinement subnetwork to improve predictions iteratively and 2) exploitation of human joint graphical relations.
- In this work, we present how efficient and simple iterative subnetworks (IS) with linear hierarchical ordering (LHO) based on the ideas can help to improve accuracy on strong backbone models.

Overviews

LHO:

The prediction of a particular keypoint should depend on previous predictions of its adjacent keypoints: (head top, upper neck, thorax, left shoulder, right shoulder, pelvis, left elbow, right elbow, left hip, right hip, left wrist, right wrist, left knee, right knee, left ankle, right ankle).

IS:

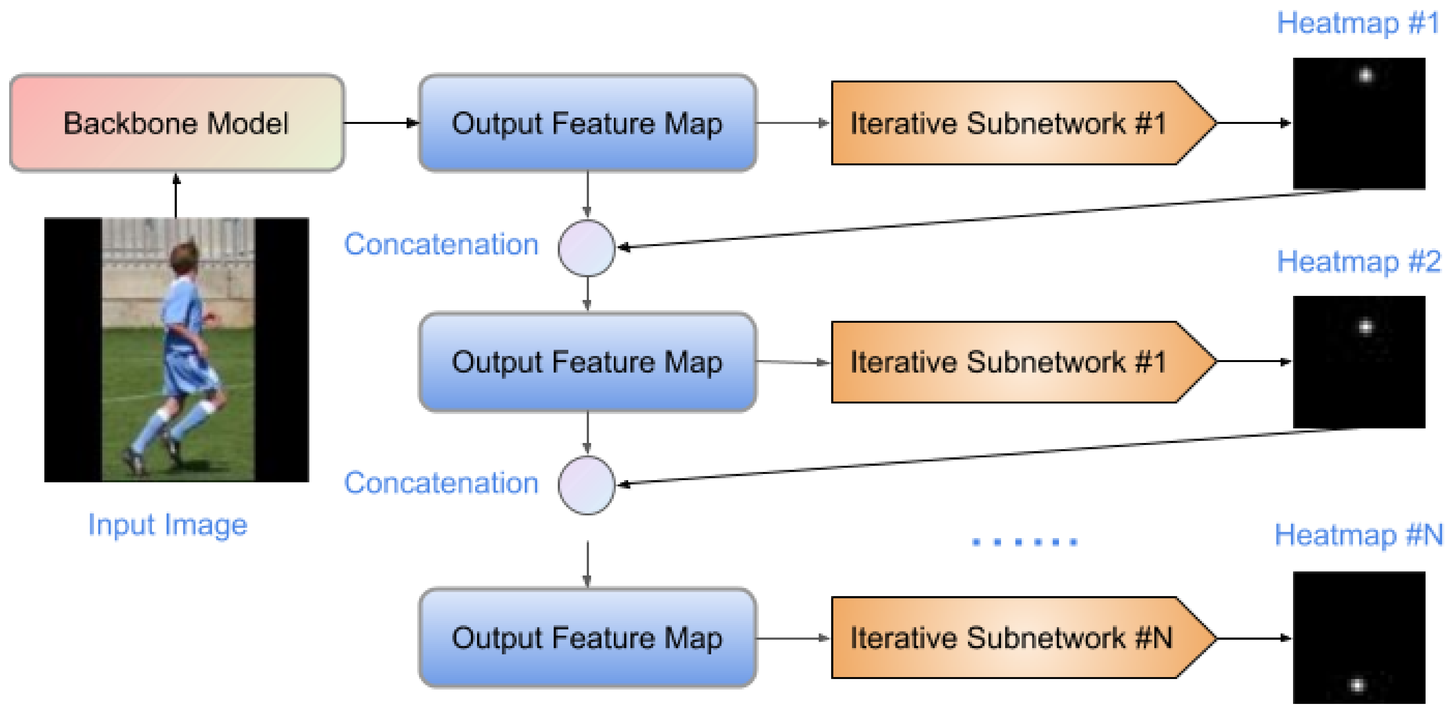
The output feature maps generated from the backbone network will be passed to an iterative subnetwork for production. of a single heatmap.

$$
\begin{aligned}
H_i &= F_i^{IS}(O_i) \\
O_i &= concat(O_{i-1}, H_{i-1})
\end{aligned}
\tag{1}
$$

where H_i is the i^{th} output heatmap for joint keypoint i, O_i is the input feature maps for the i^{th} iterative subnetwork, and F_i^{IS} is the iterative subnetwork for i^{th} heatmap. Note that O_1 is the output feature map generated by a backbone network and $N = 16$ for human pose estimation.

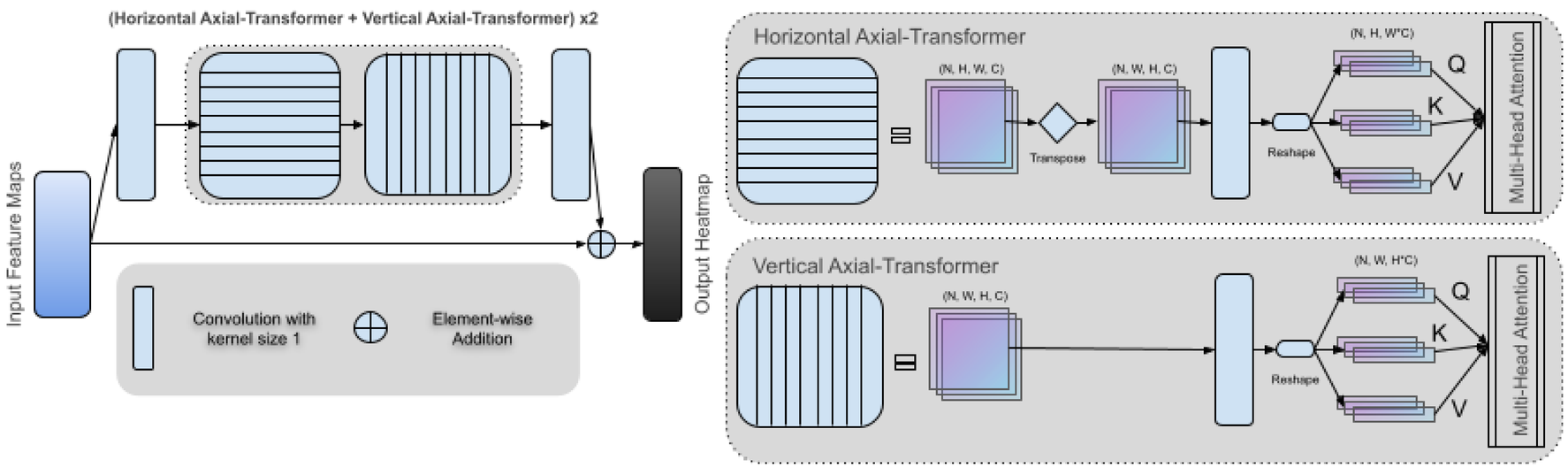
Methods

- The iterative subnetworks can be a very general neural network. Below is the pictorial representation:



Methods

- In this study, we consider three different versions for IS block construction: 1) Simple Convolution, which represents basic building layers in deep learning studies; 2) Residual Block, which is a successful and very popular design for various computer vision tasks; and 3) Axial-Transformer, which is one of the efficient ways to apply Transformer in computer vision.



Results

- Below is our result on MPII held-out test set:

Comparisons of PCKh@0.5 on MPII held-out test set. LHO represents linear hierarchical ordering. HRNet is adopted as the backbone network for all our IS designs.

Method	Head	Shoulders	Elbows	Wrists	Hips	Knees	Ankles	Total
HRNet	97.5	96.1	90.5	86.0	88.8	85.6	81.6	89.9
IS_{Conv} w/o LHO	**97.7**	96.2	90.8	**86.3**	89.8	86.2	81.6	90.2
IS_{Conv} w/ LHO	**97.7**	**96.3**	90.6	**86.3**	89.8	86.5	81.8	90.3
IS_{Res} w/ LHO	**97.7**	96.2	90.8	**86.3**	89.5	86.6	**82.1**	90.3
IS_{AT} w/ LHO	97.3	96.0	**91.2**	86.2	**90.1**	**87.2**	81.5	**90.4**

Conclusions

- hierarchical ordering of the predicted keypoints with simple iterative subnetwork can improve the prediction accuracy of human pose estimation.
- The experimental results suggest that we can achieve similar improvement to accuracy on strong backbone networks without an unnecessarily large complex refinement network. Significant improvements were observed in the difficult keypoints such as hips and knees. Both keypoints have strong dependence on their connecting keypoints, which shows that IS and LHO are effective at improving the prediction by utilizing adjacent keypoint heatmaps.
- Larger receptive field subnetwork such as Axial-Transformer can be used to further improve the accuracy.

Acknowledgments

- I would like to express my special thanks of gratitude to my supervisors Prof. Cai and Dr. Song as well as my mentor Chaoyi for their valuable suggestions.

- If you have any questions or comments, please feel free to contact me via schu9751@uni.sydney.edu.au.