Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/38/20230534

Dense-connected Stacked Hourglass Networks for Human Pose Estimation

H. Liu[1,3,5], and C. Ding[2,4]

[1]McGill University, Computer Science
[2]Georgia Institution of Technology, Biomedical Engineering

[3]haoyuan.liu@mail.mcgill.ca
[4]chengding@gatech.edu
[5]Corresponding author

Abstract. The main idea of this project is to try to improve the accuracy of human pose estimation in previous models. The new model proposed is based on the Stacked Hourglass Network with new structures added. The new structures ensured that the preservation of features of the original data by adding connections across the network, which we refer to as a "Dense-connected Stacked Hourglass" network, and we expected the new structure and the feature preserved could be helpful in the later stages because the Stacked Hourglass network pools down to very low resolution, during which important information may be lost. The data sets used in the project are MPII Human Pose and FLIC (Frames Labelled in Cinema). The final results show that the proposed architecture is able to improve the estimation accuracy to certain extend in identifying head, wrist and hip, while further studies on the architecture and improvements are still required.

Keywords: Stacked Hourglass, Human Pose Estimation, Machine Learning.

1. Introduction

Currently, research teams in both academy and industry in human-computer interaction or animation rely on expensive motion capture devices, which employ motion sensors and motion capture suites to obtain human gestures. The primary objective of human pose estimation is to identify key points of a human body on an image, thus ultimately understanding the exact motion and gesture of the human body in a frame. Nowadays, convolutional neural network (CNN)[1] is widely used in computer vision, also the founding block of human pose estimation. Indeed, some tasks can be solved with a simple model and achieve an astonishing result, yet human pose estimation is much more complex and involves more than simple traits of an object; it contains variables like clothing which cannot be tackled easily [2].

Until now, higher accuracy in estimating human pose was achieved by Hourglass Network and subsequently Stacked Hourglass Network [3]. The Stacked Hourglass Network is primarily a combination of multiple hourglass models, where the outputs of each smaller model are inputted into another smaller model and receive the final result from the last block. Like many other networks employed in computer vision, in each hourglass model, the features are pooled down to lower resolutions, followed by unsampling and combination. The hourglass model is different because it uses a more symmetric structure.

DOI: 10.54254/2755-2721/38/20230534

We followed the steps to construct a "Dense-connected Stacked Hourglass Network", namely DC-hourglass, unlike the standard Stacked Hourglass Network, each single hourglass model receives the combination of output from every previous hourglass model. With the fully-connected feature of the network, we hope that information from the earlier parts of the model can be helpful in later positions since the hourglass model pools down to lower resolutions, where some critical details deciding the key points may be lost. We expect minor improvements in overall accuracy since the original Stacked Hourglass Network has already achieved very high accuracy and around 2-3% improvements on difficult key points such as ankles.

2. Related Works

Human pose estimation composes of two approaches, bottom-up [4]and top-down [5]. The Bottom-up approach involves predicting every instance of a specific key point in an image and subsequently grouping those key points to identify each object. The Top-down approach involves first detecting objects in a given image; after resizing, the object is inputted into the network for pose estimation. The Top-down approach converts pose estimation of multiple into multiple single-person pose estimation and predict k key points of a person [5]. There are two ways to represent the output of the network. The first is to use regression on the k coordinates, while the second is to use k heatmaps. Early works such as DeepPose [5] use a direct regression to the coordinates, whose result is not as precise as using heatmap as the network's output. Convolutional Pose Machine (CPM) [6] has a multi-stage structure, at every stage, the input consists of both the original image features and a belief map from the previous stage. This map can be viewed as an encoding of the spatial context learned in the preceding stage. Thus, the current stage can extract information through convoluted layers and generate a new belief map based on the two kinds of information. After continuous refinement, an accurate result can be obtained. This work came up with the idea that neural networks must learn the image features and spatial context simultaneously, which is indispensable in pose estimation. Another essential part of this model is solving gradient loss with intermediate supervision. As the stages increase and the layers of the network become deeper, gradient loss can occur and prevent the network from converging. Supervision is added to the belief map after each stage ends, laying a foundation for most multi-stage networks to solve the problem of gradient loss.

What inspired us most is the hourglass model and Stacked Hourglass Network [3]. A Stacked Hourglass Network is a multi-stage structure composed of multiple hourglass modules. Every hourglass module contains a bottom-up and a top-down process, the former utilizes convolution and pooling to obtain a low-resolution image, and the latter uses upsample to restore the image to high resolution. The most noticeable advantage of the Stacked Hourglass Network is the ability to extract and combine all characteristics under every scale.

To solve some of the problems encountered in Stacked Hourglass Network, we adopted the technique used in Dense-Net. Dense-Net is an effective solution to the gradient vanishing problem. It achieves this by improving feature propagation, promoting feature reuse, and significantly reducing the number of parameters required, which enables the construction of deeper networks based on it. The network consists of L layers and L(L+1)/2 number of direct connections, where for each layer, the outputs from all previous layers are used as input and the output features are used in all preceding layers. The reason why Dense-Net requires fewer parameters than traditional CNNs is that each layer in Dense-Net has direct access to the gradients from the loss function and the original input signal.

3. Materials and Methods

The data sets used in the project are MPII Human Pose and FLIC (Frames Labelled in Cinema). MPII consists of around 25k images extracted from YouTube videos, including more than 40k annotated human samples, where 28k are for training and 11k are for testing. The dataset covers multiple human activities and focuses on various people's images. The test set also includes annotation of shaded body parts. The FLIC dataset contains 5003 images from extracted movies. Each image is labelled with ten upper body joints. Since images may consist of multiple people, the network is thus trained to only

DOI: 10.54254/2755-2721/38/20230534

annotate one centered labelled person in each image. The data is then resized into 256x256 pixels, using the method from the Stacked Hourglass network.

The proposed DC-hourglass combine both advantages between hourglass and dense connection, as shown in Fig.1 To fulfil the objective, we implemented an augmented architecture of the Stacked Hourglass Network, where each smaller hourglass model is placed with each other connected in an end-to-end way. The main difference between our network and the original model is that it adds more connections in smaller hourglass models. Each hourglass model receives a combination of inputs from all the outputs before it and sends its output to all subsequent models. Through this implementation, we hope that the traits in earlier steps may be reemphasized. The architecture of the model is as follows. For each Hourglass model, it is implemented similarly to the one used in the Stacked Hourglass network.

The proposed DC-hourglass is implemented by PyTorch [8]. Instead of conventional cross-entropy, the mean squared error (MSE) is used as loss function. Adam is leveraged as the optimizer with 0.1 and 0.99 as beta1 and beta2. The DC-hourglass is trained for 200 epochs with a batch size of 20. Batch normalization is used for each block. We start the training process with a learning rate of 1e-4, and 0.1 decay factor after each epoch. The whole process is experimented on a workstation with a 1080TI NVidia GPU.

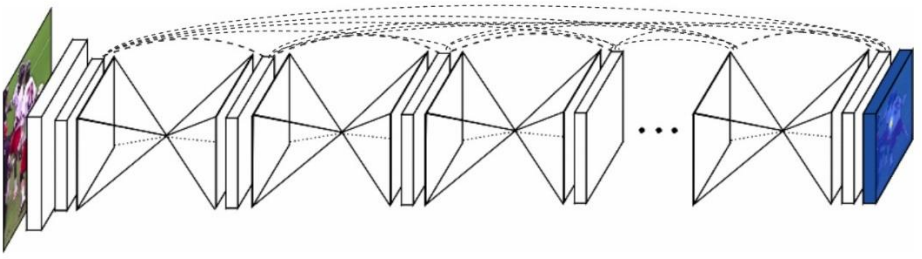

Figure 1. Dense-connected hourglass network [3].

4. Results

Table 1. Results on the MPII Human Pose dataset (PCKh@0.5).

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Stacked-Hourglass	95.46	96.63	88.07	82.37	83.83	80.48	76.95	86.00
Ours	95.77	93.75	87.93	83.13	83.94	80.40	77.30	85.16

The evaluation is based on the Percentage of Correct Key points, which evaluates the percentage of detections that fall within the normalized distance of the ground truth. The accuracy of the network on both FLIC and MPII data sets are reported in Table 1.

Notably, the DC hourglass model exhibited subpar performance relative to the baseline accuracy of 90.9% reported in the original hourglass work, as indicated by the results presented in Table 1. Two factors may account for this deviation. Firstly, we note that our training protocol differed from the approach taken by the original researchers, specifically regarding the number of epochs and the learning rate at initial and its schedule. Regrettably, the precise training regimen used by the original authors is undisclosed, which precludes us from faithfully replicating their methodology. Furthermore, we did not fine-tune the model's training parameters, such as the learning rate or damping factor, to improve the model's accuracy. To obviate any confounding influence of disparate training settings, we employed the same learning strategy across all models compared. Secondly, the original Hourglass network outlines several strategies that were employed to enhance the validation accuracy of stack-HgNets, including the

DOI: 10.54254/2755-2721/38/20230534

use of an inverted image to make predictions and shifting the prediction to the next highest neighbor in the heatmap by 0.25 pixels. These strategies were found to improve the validation accuracy by over 1%. However, we refrained from utilizing these techniques in our evaluation of the DC hourglass model, as our focus was solely on assessing its inherent performance characteristics. Consequently, we contend that our implementation of the hourglass model provides a reasonable benchmark and a valuable reference for evaluating the DC hourglass model.

5. Conclusion

The Dense-Stacked Hourglass Network showcases promising potential in enhancing human pose estimation owing to its innovative structural design. While this network exhibits notable enhancements in detecting key points like the head, wrists, and hips, it is evident that its overall accuracy has seen a decrease compared to the original Stacked Hourglass network upon which it is built. This drop in accuracy could potentially be attributed to the heightened complexity of the network, leading to issues such as degradation. There is a theoretical proposition that removing certain connections might mitigate this problem. However, extensive research is required to determine how to remove certain connections in order to improve the architecture. We theorized that if remove the connection in the later part of the architecture could make it work, because certain features from the very beginning may not be useful in the later part, and may even cause the very problem we currently face. Nonetheless, the network has demonstrated its capability in effectively addressing these tasks, resulting in noticeable improvements. Hence, it remains imperative for future studies to delve deeper into refining the accuracy while upholding the current advancements achieved.

References

[1] Li, Z., Liu, F., Yang, W., Peng, S. and Zhou, J., 2021. A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems.

[2] Andriluka, M., Pishchulin, L., Gehler, P. and Schiele, B., 2014. 2d human pose estimation: New benchmark and state of the art analysis. In Proceedings of the IEEE Conference on computer Vision and Pattern Recognition (pp. 3686-3693).

[3] Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. arXiv preprint arXiv:1603.06937 (2016)

[4] Z. Cao, T. Simon, S. -E. Wei and Y. Sheikh, "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1302-1310, doi: 10.1109/CVPR.2017.143.

[5] Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE (2014) 1653–1660

[6] A Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on (2016)

[7] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261-2269, doi: 10.1109/CVPR.2017.243.

[8] Imambi, S., Prakash, K.B. and Kanagachidambaresan, G.R., 2021. PyTorch. Programming with TensorFlow: Solution for Edge Computing Applications, pp.87-104.