A study on performance comparison between human pose estimation models

Suyoung Yoon
Aiffel Research
Republic of Korea
y8534797@naver.com

Abstract

This study is a basic study conducted to evaluate the performance of deep learning-based models in the field of Human Pose Estimation, and a performance analysis was performed using the same dataset and metrics for the Stacked Hourglass model and the Simple Baseline model. The dataset used in this study was the MPII Human Pose Dataset, and PCKh was used as the evaluation metric. As a result, it was confirmed that the Stacked Hourglass model was about 5-10% superior in accuracy, but the Simple Baseline model was about 83% less in terms of learning time. In future research, we plan to evaluate and compare the performance of the models using the larger COCO-Pose Dataset.

1. Introduction

Human pose estimation is a key technique to enable human-machine interaction, and is used to accurately identify and track key points on the body in a single RGB image. Advances in deep learning have led to the devising of various network architectures that have significantly improved performance in this field, but performance comparison and accurate evaluation between models is essential. Therefore, this study aims to compare the performance of pose estimation models by utilizing the same challenge and evaluation metrics, and accurately evaluate the performance of the models through detailed analysis and ablation study.

2. Related works

In recent years, research on pose estimation that incorporates deep learning techniques has been actively conducted. Toshev *et al*. [5] presented a method to solve the pose estimation problem by utilizing a regression model based on deep learning. Tompson *et al*. [4] utilized the MPII Human Pose Dataset to learn a heatmap representing the keypoint probability distribution, combined coarse and fine heatmap regression models to improve accuracy, and introduced span dropout. Wei *et al*. [6] proposed a multi-stage structure capable of end-to-end learning and extended the receivetive field, and Newell *et al*. [2] effectively extracted features from various scales through a stacked hourglass network and residual connections. Xiao *et al*. [7] proposed a baseline model that combines a backbone network and a deconvolution module to provide a simple and well-performing model.

3. Method

3.1. Dataset

The evaluation is conducted using the MPII Human Pose Dataset. This dataset contains approximately 25,000 images depicting various daily activities, with annotations for around 40,000 individuals. Each image includes keypoint coordinates representing 16 major body parts, precisely labeled based on the physical characteristics of the subjects. For training, the label data was transformed into a heatmap of size 64 x 64 centered on the keypoint coordinates and fitted to the following 2D Gaussian distribution:

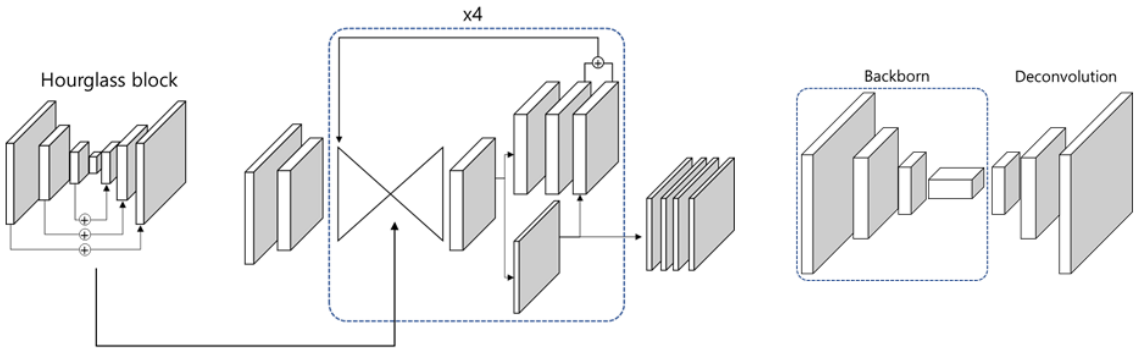
$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{1}$$

where any point on a 2D plane can be represented by its coordinates (x, y), and standard deviation of Gaussian distribution σ.

3.2. Model and Network Architectures

The comparison of pose estimation models is based on the Stacked Hourglass Networks proposed by Newell *et al*. [2] and the Simple Baselines Networks introduced by Xiao *et al*. [7]. As can be seen in Fig. 1a, the Stacked Hourglass model consists of a stacked architecture as multiple hourglass modules based on the successive steps of pooling and upsampling, with residual connections. Also, as shown in Fig. 1b, the Simple Baseline model consists of a backborn module and a deconvolution module.

As a basic study to evaluate the performance of pose estimation models, the parameters of each model were configured with basic specifications. In the case of the Stacked



(a) Stacked Hourglass Network (b) Simple Baseline Network

Figure 1. Schematic diagram of the pose estimation network used in the study.

Hourglass model, the residual connection was performed once, and the number of stacks was set to 8 for reproducibility with previous research [7]. In the Simple baseline model, the backbone model used the ResNet-50 model by removing the classifier also for reproducibility with previous research [7], and the weights were initialized by pretraining on ImageNet classification task [3].

The input data for each model is a 256 x 256 RGB image from MPII Human Pose Dataset, and the output consists of a single-channel heatmap image of 64 x 64 size. Each model was trained for 20 epochs to find the optimal joint, with a batch size of 8, with approximately 2776 train batches and 369 validation batches per epoch. The loss function for the models was MSE, the optimizer was Adam, and the learning rate was set to 0.0007.

4. Result

The results were analyzed by comparing the metrics and time taken at the optimal fit point for each model, where time taken was time per epoch and time taken to reach optimal fit.

4.1. Metrics

In this study, we used PCKh from Andriluka *et al*. [1] to measure the accuracy of the position of the key points. PCKh is an articulation independent improvement of PCK by Yi *et al*. [8]. To determine keypoint matching with ground truth in PCK, we define the partial length of the object's head as a threshold, which can be confirmed by the following equation:

$$d_i = \sqrt{(x - \hat{x})^2 + (y - \hat{y})^2} \tag{2}$$

and

$$PCKh = \frac{1}{N} \sum_{N} \delta(d_i \leq T d_h) \tag{3}$$

where keypoint coordinates of ground truth (x, y) and predicted $(\hat{x}, \hat{y})$, d_h is object head size, N is number of objects, T is an artificially set threshold, and if the bracketed condition holds, then δ 1; otherwise,it is 0.

In this study, we set PCKh @ 0.25, PCKh @ 0.5, PCKh @ 0.75, and PCKh (then PCKh values at T = 0.25, 0.5, 0.75, and 1) from Eq. (3).

4.2. Comparison Result

As shown in [**?**], the Stacked Hourglass model reached optimal fit in 18 epochs, while the Simple Baseline model reached optimal fit in 8 epochs. As can be seen in [**?**] and [**?**], the learning results for optimal fit show that the Stacked Hourglass model is about 5-10% more accurate than the Simple Baseline model in all indicators, but the time it took to learn to reach the optimal fit point was about 28,297 seconds, and the average epoch per second was about 1,870 seconds for the Stacked Hourglass model and about 817 seconds for the Simple Baseline model, taking about 1,053 seconds. Figure 4 shows the results of keypoint inference using the test sample, and confirms that there is not much difference between the two models except in dynamic situations. Ultimately, the Simple Baseline model was found to be more accurate with a quantitative metric difference of about 10% in accuracy. However, the Simple Baseline model was found to be the more efficient model, taking about 83% less training time to reach the optimal fit point and about 56% less per epoch.

5. Discussion

In this study, we compared and analyzed the performance of both the Stacked Hourglass and Simple Baseline models based on the MPII Human Pose Dataset. We confirmed that the Stacked Hourglass model was superior in terms of accuracy, but the Simple Baseline model was more efficient in training time. These results highlight the importance of model selection considering the interplay between accu-

Method	Backbone	Stack	PCKh@0.25	PCKh@0.5	PCKh@0.75	PCKh@1	Optimal epoch	Sec
stackedhourglass	-	4	0.55	0.64	0.69	0.73	18	33721
simplebaseline	ResNet-50	-	0.48	0.58	0.64	0.68	7	5424
distance	-	-	0.16	0.11	0.05	0.05	10	28297

Table 1. Comparison optimal model with Stacked Hourglass and Simple Baseline on MPII Human Pose Dataset.

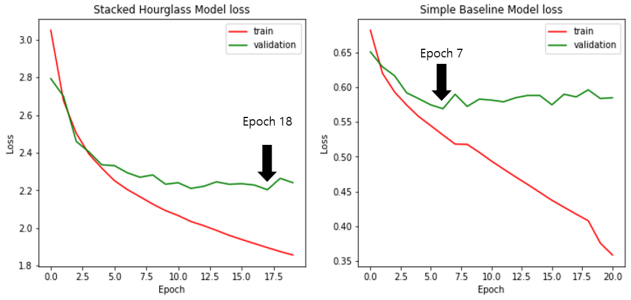


Figure 2. Learning curve and optimal epoch each model.

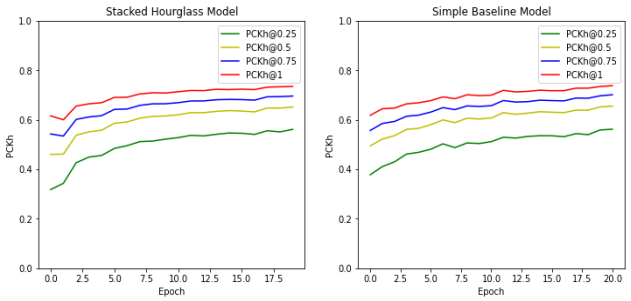


Figure 3. metric per epoch on each model.

Figure 4. sample results on test set.

racy and efficiency. In future studies, we plan to utilize the larger-scale COCO-Pose Dataset to evaluate model performance in various situations and explore advanced learning methods and scalable model designs.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, pages 3686–3693, 2014. 2

[2] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499, 2016. 1

[3] Russakovsky Olga, Deng Jia, Su Hao, Krause Jonathan, Ma Sean, Huang Zhlihenh, Karpathy Andrej, Khosla Aditya, Bernstein Michael, Berg Alexander C., and Fei-Fei Li. Imagenet large scale visual recognition challenge. *IJCV*, 115(3): 211–252, 2015. 2

[4] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *CVPR*, pages 648–656, 2015. 1

[5] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, pages 1653–1660, 2014. 1

[6] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, pages 4724–4732, 2016. 1

[7] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, pages 466–481, 2018. 1, 2

[8] Yang Yi and Ramanan Deva. Articulated human detection with flexible mixtures of parts. *IEEE TPAMI*, 35(12):2878–2890, 2013. 2