Human Pose Detection

Anowarul Kabir

April 2020

Abstract

Human Pose estimation is one of the critical tasks that has been around the Computer Vision community for the past few decades. Understanding people's pose in images and videos can unlock many real life application, such as pedestrians move detection, players pose estimation and so on. DeepPose was the first major paper that applied Deep Learning to Human pose estimation. This studies pose estimation as keypoints regression problem guided by regression loss function. They leverages AlexNet model for building the network. Following their work, we implement the neural network network as regression problem and use PCKh for quality estimation of the model. Finally we compare our implementation with some existing implementation and show some outputs generated by our model and the ground truth.

1 Introduction and related work

Human Pose Estimation is defined as the problem of localization of human joints (also known as keypoints - elbows, wrists, etc) in images or videos. It is also defined as the search for a specific pose in space of all articulated poses. 2D Pose Estimation - Estimate a 2D pose (x,y) coordinates for each joint from a RGB image. 3D Pose Estimation - Estimate a 3D pose (x,y,z) coordinates a RGB image. In this work, we want to understand human pose given a 2D image that contains human.

Understanding people in images and videos is critical for developing many real world applications; for instance, human surveillance in garage or office, detecting pedestrians, human movement detection for automatic driving cars and so on. Therefore, this problem has been drawn attention of the computer vision community in a great scale. Many of the previous studies focused on graphical model from handmade features. However, by the arrival of deep learning models [1, 2, 3], computer vision community has dived into looking for solutions of many computer vision problems by efficiently designing and optimizing neural network architectures and loss functions. One of the benefits of deep learning approach is that it can extract features guided by the loss function that could be specific to the problems or general from the top view of the solution approach [1, 4].

In this study, we look for an approach human pose estimation using deep neural networks. The input of the model is an image and the output is keypoints predictions. This study leverages regression problem given some ground truth images and their keypoints set. The dataset used in this study is MPII Human Pose Dataset [5]. MPII Human Pose dataset is a state of the art benchmark for evaluation of articulated human pose estimation. The dataset includes around 25K images containing over 40K people with annotated body joints. The images were systematically collected using an established taxonomy of every day human activities. Overall the dataset covers 410 human activities and each image is provided with an activity label. Each image was extracted from a YouTube video and provided with preceding and following un-annotated frames. In addition, for the test set we obtained richer annotations including body part occlusions and 3D torso and head orientations [5].

This study is basically a reconstruction of a previous work titled as *DeepPose: Human Pose Estimation via Deep Neural Networks* [6]. Deeppose is the first major article that leverages deep learning approach to the problem of 2D human pose estimation and beats existing models. This approach formulates pose estimation as a CNN[1]-based regression problem towards body joints. Finally they use a cascade of such regressors to refine the estimated pose estimation and showed getting better results.

A study titled as *Stacked Hourglass Networks for Human Pose Estimation* is a landmark paper that introduced a novel architecture that beat all previous methods. The naming of stacked hourglass network is as the network consists of steps of pooling and upsampling layers which looks like an hourglass and these are stacked together. The design of the hourglass is motivated by the need to capture information at every scale. While local evidence is essential for identifying features like faces hands, a final pose estimate requires global context. The person's orientation, the arrangement of their limbs, and the relationships of adjacent joints are among the many cues that are best recognized at different scales in the image.

2 Methodology

Given a pair (X, Y) where X denotes an input image and Y denotes as ground truth pose vector, the network estimates Y' as estimated pose vector. [6] encodes a pose vector as k key body joints where each joint is x and y coordinates. So $Y = (Y_1, Y_2, ...Y_k)$ denotes our pose vector which encodes pose estimation, where Y_i denotes (x, y) coordinates. Then the input pair is normalized with respect to a bounding box b. Let, b is defined as box center b_c, box height b_h and box width b_w. Then the image is cropped sized of the bounding box where the image contains the human on the middle and finally normalizes the image pixel values; this normalizing operation is denoted as $N(X, b)$. Then each key point is normalized as translated as the box center and scaled by the box size which is denoted in the article as $N(Y_i, b)$. The following equation shows the

Model id	Learning rate	#epochs	Batch size	Final train loss
Model21	1e-3	40	30	0.2515
Model22	1e-4	40	30	0.2471
Model23	1e-5	40	30	0.0260
Model30	1e-6	40	30	0.0202
Model31	1e-7	40	30	0.0235
Model32	1e-8	40	30	0.0308

Table 1: Model with parameters

normalization process:

$$N(Y_i, b) = \begin{bmatrix} 1/b_w & 0 \\ 0 & 1/b_h \end{bmatrix} (Y_i - b_c)$$

The article [6] defined the function as $\psi(X, \theta)$ where X is the input image, θ is the learned parameters and ψ regresses k joints. So in addition to the boxed normalization, the prediction of the points on the absolution image will be as follows:

$$Y^* = N^{-1}(\psi(N(X, b), \theta))$$

where Y^* denotes the predicted joint points and N^{-1} represents the opposite operation of $N(X, b)$ or $N(Y_i, b)$.

The function ψ is denoted as deep learning neural network inspired from AlexNet [1]. This is a convolutional neural network with several layers. Each layers consist of linear transformation followed by non-linear layers excepts for the last three layers. Last three layers are fully connected layers with 50% dropout layer. Dropout layers have advantages, for instance general regularization, encourage all neural nodes to learn something and each dropout of some node generate a fully different model than others which indirectly encourages to work as ensemble methods where multiple models do same thing and final output is the average or maximum of them. For elaborate discussion the reader may refer to AlexNet [1] original paper. The general motivation of using such deep neural network comes from the successful evidence of DNN in the field of classification and object localization problems.

3 Result and Analysis

Before going into result, lets look at the trained model's final loss and learning rate in Table 1. Learning rate with 1e-6 works better with our model with the same number of epochs and min-batch size. Throughout the whole training for different hyperparameters, we leveraged adam optimizer provided by the pytorch framework. Note that, the model id indicates nothing specific except for indicating for some visualization works.

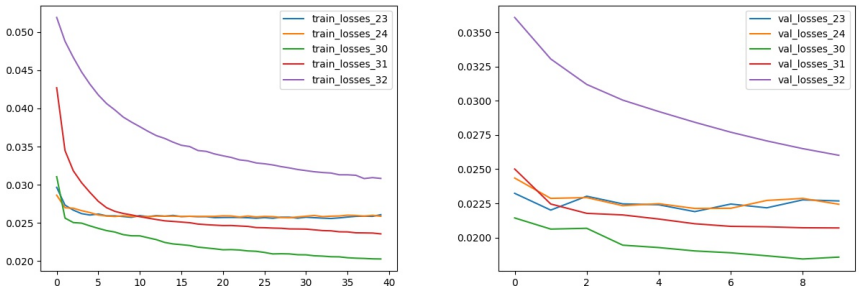
We run our models on MPII [5] dataset and optimize the learnable parameter for different hyperparameters and optimize the model using adam optimizer. Figure 1 shows train and validation losses and we see Model30 dominates in both cases. For measuring the quality of out models we used Percentage of Correct Key-points (PCKh) at 50% threshold of the head bone link. A detected joint is considered correct if the distance between the predicted and the true joint is within a certain threshold [7]. The main deeppose article [6] did not evaluate

Model	PCKh
Model23	36.07
Model24	34.25
Model30	51.13
Model31	36.42
Model32	27.96
Deeppose [8]	54.20

Table 2: Evaluation of the models

their method on MPII [5] dataset. So we overview some codes in github and found [8] which is mentioned to achieve 54.20% PCKh. On the other hand, none of our implementation could not achieve this far. Our best PCKh is 51.13% for Model30 and other models are far behind that. For looking the reasons behind that, we finally see we used AlexNet [1] architecture for our model whereas [8] used ResNet [3].

Finally, we like to look at some predicted vs ground truth keypoints. Figure 2 shows some images that are cropped, clipped or rotated with the ground truth krypoints. Figure 3 shows predicted keypoints for each images. From predicted keypoints, we can see that they are almost making a straight line. In very few cases, the model is able to generate keypoints that did not follow straight line.



(a) Train losses

(b) Validation losses

Figure 1: Losses for different models

Figure 2: Ground truth Keypoints

4 Conclusion

Human pose estimation is a challenging task. Deep learning imposes end-to-end learning whereas classical methods uses handmade features. DeepPose was the first major paper that applied Deep Learning to Human pose estimation

Figure 3: Predicted Keypoints

and beats the existing models. Even though the result is not quite expected, however CNN models proves to the removal of handmade features and able to

extract the required features guided by the loss function.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[2] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[5] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele.

[6] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, page 1653–1660, USA, 2014. IEEE Computer Society.

[7] Sudharshan Chandra Babu. guide to Human Pose Estimation with Deep Learning. https://nanonets.com/blog/human-pose-estimation-2d-guide/#CPM, 2020. [Online; accessed 15-April-2020].

[8] Naman-ntc. Deeppose github implementation. https://github.com/Naman-ntc/Pytorch-Human-Pose-Estimation, 2020. [Online; accessed 15-April-2020].