Pose Estimation using Attention Based Mechanisms

Shyam Subramanian [1] **Ajith Kumar Ganesan Govindasamy** [1] **Archit Kumar** [1]
Joshua Amrith Raj Caleb Chanthi Raj [1]

Abstract

Human pose estimation problem tackles the issue of localization of human body joints. There are many complex architectures specifically for Human Pose Estimation problem with significant differences but similar accuracy. Several complex architectures will still come up. So there is a need to set a new baseline with minimal changes to existing, problem independent, architectures that are suitable for any problem in computer vision. We aim to address the problem of estimating the poses of single and multi human instances in multiple images existing in the MS-COCO (Common Objects in Context) dataset. We explored the possibility of adding deconvolution layers at the end of a ResNet architecture. We also explored various attention based mechanisms that can provide better localization in the context of Pose Estimation.

1. Introduction

Our project seeks to solve the problem of "Human Pose Estimation". Here pose refers to both the position as well as the orientation of the body of a person with respect to a coordinate frame. Pose estimation is used in a wide variety of applications including but not limited to motion capture in the gaming industry, and in Robotics - humans and robots working together in proximity.

Human body pose estimation is a challenge as the human body consists of approximately 230 joints with about 244 degrees of freedom. To approach the problem significant generalization has to be done by approximating the human body into a simpler form. Also, another major difficulty is that one particular pose may not look the same for every other human. For instance, the type of clothing worn by a person, or the hairstyle etc. would affect the appearance of a pose from person to person. In addition, the pose may not be clearly visible due to occlusions. These occlusions may be due to self or external objects/other humans. These difficulties offer a significant challenge to solving the problem of human pose estimation.

Figure 1. Pose Prediction of MS-COCO

COCO (Common Objects in Context) is a modern dataset for benchmarking different modern deep learning applications like pose estimation, semantic segmentation, panoptic segmentation, object detection, etc. The dataset which contains about 250,000 people instances with keypoints proves to be an extremely well formulated dataset with keypoint annotations for each human instances. Further, the COCO-API provides a refined method to utilise the dataset's annotations effectively. Since the model is trained using the COCO dataset which specializes in object categories in different contextual environments, our model will also perform well for test images of humans in different contexts.

The problem is set to predict individual target gaussian heatmaps for each of the 17 keypoints - Nose, two eyes, two ears, shoulders, two elbows, two wrists, two hips, two knees and two ankles. We develop our architecture upon ResNet-50 architecture by adding deconvolution lay-

[1]Worcester Polytechnic Institute. Correspondence to: Ajith Kumar <aganesangovindas@wpi.edu>, Shyam Subramanian <ssubramanian2@wpi.edu>, Archit Kumar <archit@wpi.edu>, Joshua Amrith Raj <jcalebchanthiraj@wpi.edu>.

CS541 Special Topics Course Final Project, Worcester Polytechnic Institute.

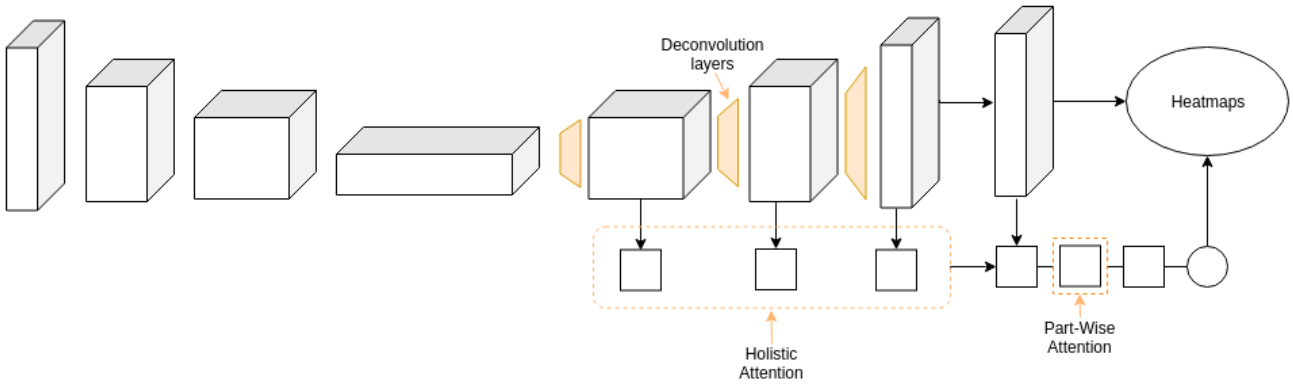


Figure 2. Resnet with attention

ers. The task of predicting keypoints require high performance in localization. A potential solution to better localization can be achieved through attention mechanisms which is only partially explored in context of Pose Estimation. We explore various attention mechanisms in-depth in our work. Each individual attention map generated provides a holistic understanding of how the attention mechanism works and how it contributes to a better overall architecture than a normal ResNet-50 architecture.

1.1. Research contributions

Our project focuses on exploring the various attention mechanisms proposed so far in Deep Learning research and how efficiently these approaches can be applied in various layers and various scales to retain necessary contextual information corresponding to each joint. Attending only parts of the image corresponding to each keypoint can lead to better localized results. Further, attention will help us ignore the background and other non-useful information from the image. We also provide reproducible code at Pose-Estimation.

2. Related Work

[1] is one of the pioneer research works in Single Person Human body pose estimation using Deep Neural Networks. They posed the problem as a regression problem directly regressing the keypoint coordinates. They also proposed a stacked model which works as refinement of coordinates obtained in previous network in the stack.

Directly regressing the joint coordinates leaves the model with no enough degrees of freedom to afford small errors leading to over-training. [2] proposed using gaussian target heatmaps considering that a valid keypoint can be present in multiple spatial locations. Recent approaches to Pose Estimation [3], [4], [5], [6], [7] use this method of predicting gaussian heatmaps and employ an L-2 loss. In all these research, the single person pose estimation can be extended to multiple person pose estimation by first running the image through an object detection model to find bounding box for each person and then finding their poses separately.

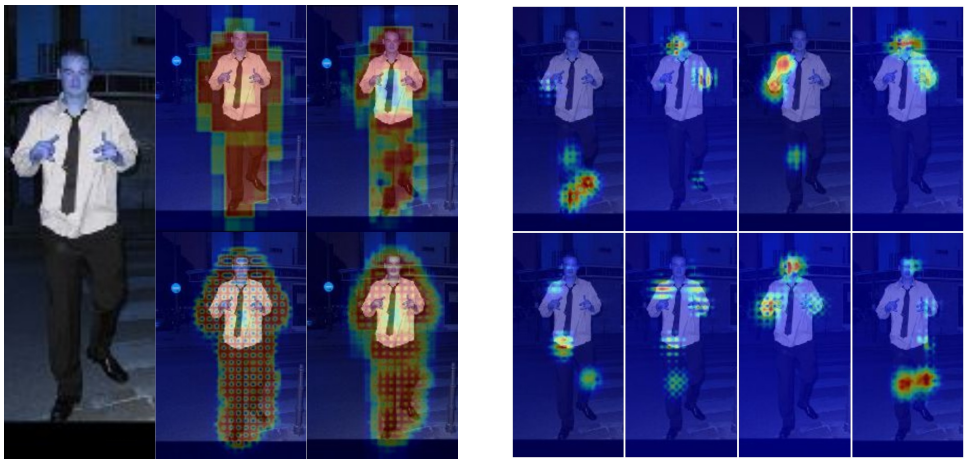
This model increases the computational costs linearly for the number of persons in the image. [3], [8] solves this problem by taking a bottom-up approach and jointly predicting the keypoints of multiple persons. Various graphical models were employed that encode the joints of multiple persons together in a convenient way to predict the poses together. Other complex models from [4], [9] involve getting more contextual information by retaining features from multiple scales using skip connections.

With the advent of very deep neural networks like [10], [11], and [12], we take the opposite approach to check predictive power of such straight-forward networks in the context of Pose Estimation. We explore this model further by using attention mechanisms. Attention has been proven useful in various tasks like Image Captioning, Neural Machine Translation. Apart from [5], attentions with regards to Pose Estimation hasn't been explored extensively in previous research to our knowledge. We experiment with attention mechanisms with the notion of providing better localized results for all keypoints. We explore and compare various Soft attention mechanisms proposed in [13], [14], [15] and [5].

3. Proposed Method

3.1. Basic Architecture

We employ ResNet-50 Architecture as our basic architecture. We add 4 deconvolution layers to the end of the ResNet-50 architecture to produce heatmaps of 64x48 for each keypoint fig 2. To scale this architecture to multiple people we employ an object detection algorithm to



(a) Holistic CRF Attention map (b) Part-wise CRF Attention

Figure 3. Holistic and Part-wise Attentions Maps

predict bounding boxes for each human in an image. We then train our ResNet architecture to predict the pose from each bounding box separately. COCO dataset has bounding boxes built-in which makes it easier for us to train. While for testing, we use YOLO-V2 object detection architecture pretrained on COCO dataset images for generating bounding boxes.

3.2. Attention

We employ a two way attention model namely a holistic attention and a Part-wise attention model. Holistic Attention model will aim to attend the whole body of the human ignoring the background while the part-wise attention model will focus on certain body parts of the image that can help in localizing each body joint. This type of attention model for Pose estimation is inspired from [5]

3.2.1. Softmax Attention

Let s(l) be the feature at location l=(x,y) in the feature map s, obtained from a prior convolution layer with non-linear activation. Then the softmax attention at l is given by,

$$\phi(l) = \frac{e^{s(l)}}{\sum_{l' \in L} e^{s(l')}}, \tag{1}$$

where, L = $\{$(x,y),$\forall(x, y)$ in s$\}$.

3.2.2. Gaussian Attention

Gaussian Attention works by exploiting two parameterized one-dimensional gaussian vectors to create image-sized attention maps. Let $a_y \in R^H$ and $a_x \in R^W$ be the two parameterized vectors which tells us the parts to attend in y and x axis respectively. The image-sized attention map is created as $a_y a_x^T$. Typically, the number of Gaussians is equal to the spatial dimension and each vector is parametrised by three parameters: centre of the first Gaussian μ, distance between centres of consecutive Gaussians d and the standard deviation of the Gaussians σ.

3.2.3. Spatial Transformer Network

Spatial transformers works like attention in the sense that they crop a particular part of the image thus making only that part of the image available to the next stage. The transform is differentiable and can be trained using backpropogation without requiring reinforcement learning techniques. It consists of three parts, a localizer, a grid generator and a sampler. First the localizer generates transformation parameters, then the grid generator generates a sampling grid by applying the transformations and finally the sampler produces the sampled output map.

3.2.4. Spatial Conditional Random Field

Spatial CRF model uses mean field approximation to recursively learn the spatial kernel. The attention map is modeled as a two-class problem. Denote y_l = $\{0, 1\}$ as the attention label at the i-th location. The probability for y_l = 1 is obtained iteratively using the mean-field approximation as follows:

$$\phi(y_l = 1)_t = \sigma(\psi_u(l) + \sum w_{l,k}\phi(y_k = 1)_{t-1}), \tag{2}$$

where $\sigma(a)$ = 1/(1 + exp(-a)) is the sigmoid function. $\psi_u(l)$ is obtained by convolution from previous layer features maps. $\sum w_{l,k}\phi(y_k = 1)_{t-1}$ is obtained by convolving the estimated attention map ϕ_{t-1} at the stage t-1 with the fil-

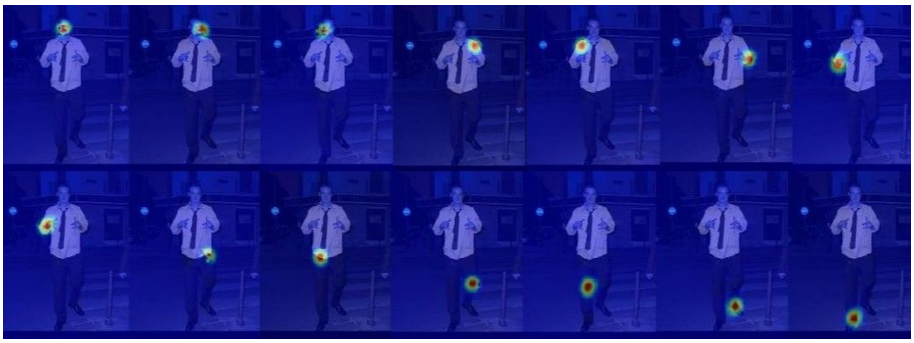


Figure 4. Keypoint Heatmaps from Attention Resnet

ters.

3.3. Holistic Attention

Holistic attention works by attending the whole image of the human body ignoring the background. From previous research work, it is evident that more contextual information leads to better localization of key points. We achieve this by attending images in various scales. Specifically, we use attention in final 4 layers of deconvolution each of shape 8x6, 16x12, 32x24, 64x48. We use k-nearest neighbor upsampling to make each attention map of shape 64x48. We finally add these attention-maps and then attend the final deconvolution layer. Note that here although we create attention maps from each deconvolution layer, we do not attend every layer. Instead, we attend only the output from the final deconvolution layer. By doing this we preserve contextual information from various scales as well as the information flowed through deconvolution layers.

For holistic attention, we experimented with softmax and conditional random field attention. Gaussian attention is not a good candidate since we want to attend the whole human body with similar weights. Also, spatial transforms are not great candidates since it only does cropping and our network would still have non-useful background information. CRF attention in fig 3 (a) takes into consideration the local pattern spatial correlations unlike global spatial softmax. Since we want to attend continuous parts of image, softmax attention did not work as expected. CRF attention works best for holistic attention.

3.4. Part-wise Attention

Part-wise attention works by attending only parts of the human body. This, combined with Holistic attention, works like a hierarchical attention model wherein we first attend the whole body and then attend parts of the human body. We introduce part-wise attention model by taking input from the holistic attention and then attending them for 17 key points.

Gaussian attention and Spatial Transforms attention worked as expected but the predictive power of the network did not increase significantly. We believe this is due to the fact that when localizing a keypoint, looking only at one keypoint might not help as much as looking at 2 or 3 successive or symmetric keypoints would. Although softmax attention captures this information, since it does not take spatial correlations into account, the network takes more time than Spatial CRF to achieve comparable performance. Therefore, Spatial CRF in fig 3 (b) works the best for Part-wise attention maps as well.

4. Experiment

For training our model we used Microsoft's Common Objects in Context (MS-COCO) dataset. Due to computational complexities involved in utilizing a large dataset, we have taken a subset of this dataset for training, testing and validation of our model. The dataset has a total of 113,000 images out of which we used 80,000 where utilized as the training set, 5,000 as the validation set and 28,000 as the testing set. In regards to metrics, we used the "L2 Loss" loss function to compute the error between the generated and targeted heat map images. Performance was evaluated using the metric "Average Precision" (AP) at different "Object Key point Similarities" (OKS). We optimized our hyper-parameters with a subset of over 500 images on 5 epochs.

Also to improve model performance in generalization we augmented our data by scaling on a factor of $\pm 30\%$ and by rotating on a factor of $\pm 40^{\circ}$.

Model training was done using "TensorFlow" by running the model for 30 epochs. The learning rate was set as 5×10^{-4} with a mini-batch size of 32. The optimizer used is Adam. We used a v100 GPU.

After all the hyper-parameters were tuned, we ran the model for 30 epochs, taking about 30 hours. Model training was proceeding as expected and the learning by our model

did not reach an asymptote. However, since the process was computationally expensive we could not proceed further with larger scale performance evaluation. As a result, we did a comprehensive comparison between two models - with "attention" and "without attention".

5. Results

The table below shows the comparison of the performance with and without attention. The average precision and recall values achieved using attention is higher than without attention. We aimed to acheive an AP score of 65%. We acheived a maximum AP score of 56.9% without attention and 65.77% with attention. The AP score would be better had we trained for a longer time.

Average Precision / Recall	IOU	Area	MaxDets	Without Attention	With Attention
Average Precision	0.5:0.95	all	20	0.468	0.567
Average Precision	0.5	all	20	0.773	0.839
Average Precision	0.75	all	20	0.489	0.617
Average Precision	0.5:0.95	medium	20	0.444	0.538
Average Precision	0.5:0.95	large	20	0.508	0.608
Average Recall	0.5:0.95	all	20	0.514	0.606
Average Recall	0.5	all	20	0.796	0.852
Average Recall	0.75	all	20	0.546	0.659
Average Recall	0.5:0.95	medium	20	0.480	0.571
Average Recall	0.5:0.95	large	20	0.563	0.655

fig 5 shows the estimated pose with different backgrounds. Our method works with different background and when multiple people are present. The first two figures show cases where only one person is present, the third figure (bottom left) contains two people and the last figure (bottom right) has multiple people in the image.

Figure 5. Top images: Single person, Bottom Left: Two persons, Bottom right: Multiple persons

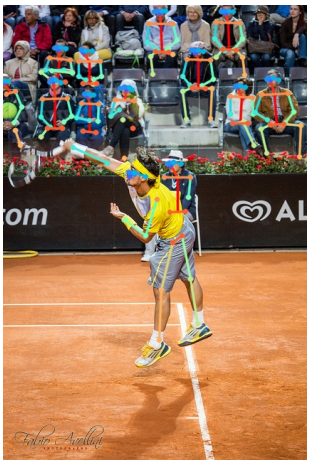


Figure 6. Example of missing Keypoints

6. Discussion

The performance is affected by the YoLo threshold. We set a limit on the number of bounding boxes. We therefore are not able to detect the pose of all the people in a few cases. The fig 6 shows one such case where the pose of some people in the background is not estimated.

In the images with occlusions (see fig 7 bottom images), some of the predictions made by our model were accurate while some predictions were unrealistic. We believe that regularizing by penalising more for such poor results for images with occlusions would help improve the performance of the model in these cases.

Another problem we face is in cases where there are multiple persons or multiple keypoints are close together in the image. In this former, the model confuses between keypoints of one person with another and in the latter the model confuses between closer keypoints. (see fig 7 bottom images).

Holistic and part wise attention model gives better accuracy proving that attention is helpful in pose estimation. Further Spatial CRF works best for attending to the human body parts since they are all related and have spatial orientation.

7. Conclusions and Future Work

Several other network architectures can be explored to tackle the problem of human pose estimation. One possible architecture that can be used is "MobileNet", designed to run on mobile phones it can be useful in implementing human pose estimation as it is computationally less expensive. It also has the advantage that it can predict a higher number of frames per second aiding in real time predictions. Also, "Dilated Residual Networks" can be an alternative. It uses dilated convolutions to help in preserv-

Figure 7. Top left and Top right images provide good result despite occlusions. Bottom left an bottom right images provide unfavorable results due to occlusions.

ing more contextual information using a higher receptive field. Additionally, we could use "hard attention" which involves the use of reinforcement learning. [6] proposed optical flow based key-point similarity between successive frames in posetrack. A possible solution is to come up with another similarity metric based on attention.

Lastly, over-fitting in human pose estimation has not been explored enough. Experimental work in this direction might yield interesting results.

Through our work we were able to provide end to end pose estimation by building upon the Resnet architecture. While we have used Deconvolution layers at the end of the Resnet architecture, this set up can be extended to any other suitable architecture for images such as Xception, Inception, VGG, or MobileNet etc. Also, we explored various other mechanisms and provided a study about how each attention mechanism will affect the network's predictions. Our work on human pose estimation has functioned successfully and given us the desired average precision values. We believe this research work will serve as a new baseline for human pose estimation problems and will aid similar future development and experiments.

References

[1] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1653–1660.

[2] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 648–656.

[3] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1812.08008*, 2018.

[4] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*, Springer, 2016, pp. 483–499.

[5] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1831–1840.

[6] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 466–481.

[7] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," *CoRR*, vol. abs/1902.09212, 2019. arXiv: 1902.09212. [Online]. Available: http://arxiv.org/abs/1902.09212.

[8] J. Tompson, A. Jain, Y. LeCun, and C. Bregler, *Joint training of a convolutional network and a graphical model for human pose estimation*, 2014. arXiv: 1406.2984 [cs.CV].

[9] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7103–7112.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[12] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[13] D. Bahdanau, K. Cho, and Y. Bengio, *Neural machine translation by jointly learning to align and translate*, 2014. arXiv: 1409.0473 [cs.CL].

[14] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, *Draw: A recurrent neural network for image generation*, 2015. arXiv: 1502.04623 [cs.CV].

[15] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, *Spatial transformer networks*, 2015. arXiv: 1506.02025 [cs.CV].