CSE 641 Project : Human Pose Estimator

Vishal Bansal, Mihir Chaturvedi, Harshal Dev
Indraprastha Institute of Information Technology, Delhi
{ vishal19217, mihir19175, harshal19306}@iiitd.ac.in

Abstract

Although no specific domain knowledge is considered in the design, plain vision transformers have shown excellent performance in visual recognition tasks. In this paper, we work on upon VitPose (Xu et al., 2022) and add our novelty to it. It is not always possible for the body key points to be always present/visible, even though the model is predicting them. Therefore with the help of heatmaps and MSELoss, we predict the same. Also, since we are predicting the key points along with their presence, we have implemented a custom loss function, which is a combination of 2 loss functions, i.e. Binary Cross Entropy and MSELoss. We predict each keypoint presence which is a classification task, and the prediction of the keypoint's coordinate is a regression task. Both these subtasks respectively use Binary Cross Entropy and MSE Loss.

1 Introduction

Human Pose Estimation is a fundamental task in the field of computer vision, which involves identifying and localizing the anatomical key points of the human body, such as the elbows, wrists, head, and ankles. This task has numerous applications, including human-action recognition, animation, and even medical diagnostics. However, due to the complexities involved in accurately identifying these key points, traditional methods have not proven to be entirely successful. Instead, deep-learning methods have emerged as the most promising solution to this challenge. These methods are well-suited for handling variations in occlusion, truncation, scales, and human appearances, which are all factors that can make the human pose estimation task particularly challenging. While CNNs (Convolutional Neural Networks) were the typical model choices for this task, vision transformer architectures are becoming increasingly popular for the same.

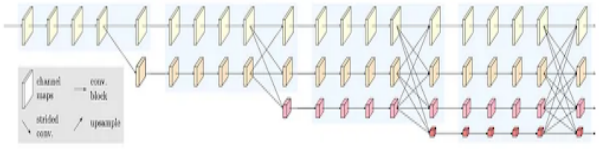

Figure 1: HRNet architecture

The aim of this project is to assess the effectiveness of our novel approach for human pose estimation and compare its performance with existing state-of-the-art models. Additionally, the project will investigate the impact of various hyperparameters and training strategies on the model's performance and provide insights into the strengths and limitations of the proposed method.

2 Related Works

Vision transformer architectures are becoming more and more popular for the same due to the manifold advantages of attention layers in extracting global relationships between image features.

2.1 HRNet

They present a novel architecture, namely High-Resolution Net (HRNet)(Sun et al., 2019), which maintains a high-resolution representation throughout the process. The benefit of the model includes the following:

1. Their approach connects high-to-low-resolution subnetworks in parallel rather than in series, as done in other existing approaches. Thus the network is able to maintain the high resolution instead of recovering the resolution through low to high process, and accordingly, the predicted heatmap is potentially spatially more precise.

2. Most existing fusion schemes aggregate low-level and high-level representations. But they perform repeated multiscale fusions to boost

Model	Backbone	InputSize	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
HRNet-W32	HRNetW32	256x192	74.4.	90.5	81.9	70.8	81.0	79.8
Transpose-R-W32	ResNet-Small	256x192	72.6	89.1	79.9	68.8	79.8	78.0
Reproducible HRNet-W32	HRNetW32	256x192	73.8	89.8	81.2	70.4	80.3	79.3
Reproducible Transpose-R-W32	ResNet-Small	256x192	73.4	90.6	80.7	70.3	77.9	76.1

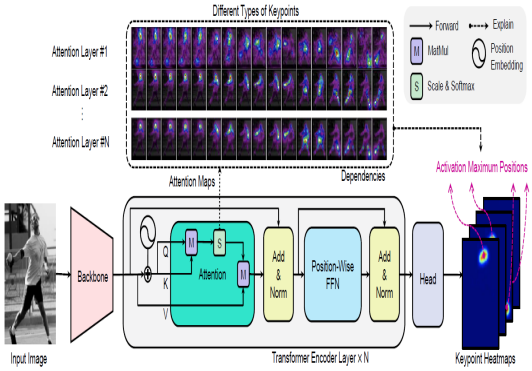

Figure 2: TransPose architecture

the high-resolution representations with the help of the low-resolution representations of the same depth and similar level, and vice versa, resulting in that high-resolution representations are also rich for pose estimation.

2.2 TransPose (Yang et al., 2021)

As proposed by Yang et. al., the model uses visual transformer architecture to capture global and high-level feature dependencies from the low-level features extracted by the CNN backbones from the input images. The output of the transformer encoder layers is then passed to the decoder head, which predicts K keypoint heatmaps using bi-linear interpolation/transposed convolution followed by 1x1 convolution.

The main advantage of this model compared to the pre-existing CNN architectures is that transPose is able to use multi-headed self-attention to capture these global dependencies in images effectively and generate keypoint heatmaps while maintaining explainability. TransPose focuses on explaining model predictions by revealing fine-grained spatial dependencies between body joint variables in the structural skeleton. It exploits attention patterns to uncover which parts of the image contribute most to keypoint localization, with the attention weights of the last attention layer acting as an aggregator. The maximum activation positions in the predicted keypoint heatmaps are the keypoint locations, which can be explained by the locations with a higher attention score than the former, forming its dependency area. Since attention maps can be seen as dynamic weights determined by specific image features, we obtain image-specific and keypoint-specific dependencies.

While transformers have been pivotal for the pose estimation task, existing methods still relied on CNNs or carefully designed, task-specific transformer structures for feature extraction.

ViTPose (Xu et al., 2022) shows the surprisingly good capabilities of plain vision transformers for pose estimation from various aspects, namely simplicity in model structure, scalability in model size, flexibility in training paradigm, and transferability of knowledge between models, through a simple baseline model. This model outperformed the previously mentioned models in all the appropriate evaluation metrics. The model used and fine-tuned pre-trained vision transformers as backbones for feature extraction, and then passes the latter to a decoder which gives predicted keypoint heatmaps of the image. However, the only problem was that solely the output heatmaps of the keypoints couldn't validate the presence of the respective keypoints themselves. Therefore, we propose our novel addition to the architecture described in the next section.

3 Methodology

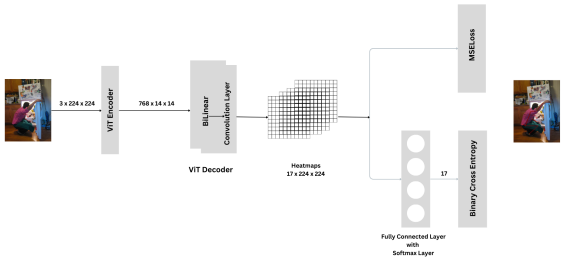

Figure 3: Our Proposed Architecture

Our proposed architecture introduces a novelty

to ViTPose: we use the heatmaps predicted to detect the presence of the corresponding key points in the image.

We use ViT-B as the pre-trained backbone encoder to give us the feature maps of the image. The decoder is also inspired by ViTPose, which upsamples the feature maps 16 times using Bilinear Interpolation to scale the former to image dimensions. This is followed by applying ReLU non-linearity to the outputs, and finally convolving the same with 1x1 convolution kernels to get 17 keypoint heatmaps.

Now, for the novelty; the heatmaps are passed through a max pooling layer which reduces their size by 4 times, after which each heatmap is flattened and sent to separate fully connected layers (with ReLU non-linearity for hidden state and sigmoid for output), which give a score indicating the confidence of the presence of the keypoint which corresponds to the heatmap.

Our model gives 2 outputs: a vector (size 17) denoting the keypoint scores, and the corresponding heatmaps for each keypoint. Since the model is performing 2 tasks, regression (against the ground-truth heatmaps) AND multi-label classification (against the ground-truth vector denoting the presence/absence of keypoints in the image), we come up with a unique loss metric. We use MSE loss for the former task, and Binary Cross-Entropy for the latter, and take the weighted average of the 2 losses to give the one which would be backpropagated along the network. The weights of the losses were experimented with.

4 Experimental Setup and Results

4.1 Dataset

Our model is trained and evaluated on COCO Dataset. The COCO dataset contains over 200,000 images and 250,000 person instances labelled with 17 key points.

- Train Set includes train2017 dataset, including 57K images and 150K person instances.
- Validation Set includes val2017 dataset, including 5K images.
- Test set includes the test-dev2017 set containing 20K images

4.2 Setup Details

We have used the pre-trained model as "vit-base-path16-224" for encoding the embeddings for each patch of the image. Initially, we will have the image, which we will transform to 224x224 size and three colours channels, and then we will do a basic model-specific transformation to the image. Then the transformed image will be passed through the encoder and patch level features with each patch sized 16x16, and feature embedding size = 768 will be received.

Now we passed the decoder to do upsampling and convolution with the encoded output. Decoder will provide us with the heatmaps for each key point in our image. Next, we will process those heatmaps for two tasks

1. We will predict for each key point if it is present or not in the image using heatmaps. Since it is a **Classification task** we used the Binary Cross-Entropy loss function.

2. We will also predict the potential coordinate in the image for each keypoint using specific heatmaps. Since this task is a **Regresssion Task** we used the Mean square Error loss function.

Since the dataset has huge training samples and due to limited GPUs, we used a small subset of data i.e 6000 samples for the training and 500 samples for validating. We trained our model for the 8 epochs with a learning rate of 5e-4. Since we are calculating two different losses so for the final loss we are considering equal weightage to both losses to calculate our final loss. We also add a dropout layer(p=0.4) while calculating potential keypoints using heatmaps. It would prevent our model from overfitting.

4.3 Observation and Results

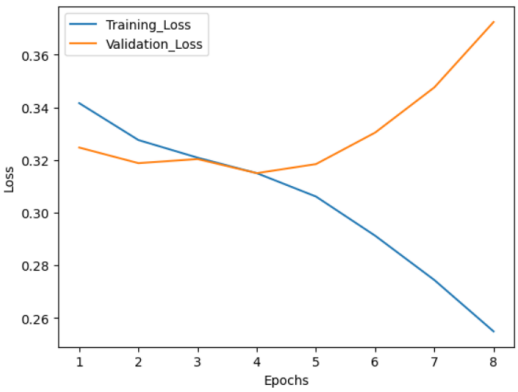

Figure 4: Loss Curve

5 Error Analysis

We can observe from our loss curves that the model starts model trained well till the 4th epoch afterwards validation loss started rising so we used the model state after the 4th epoch for calculating our inference image.

6 Contribution

The contribution of each member is as follows :

1. **Mihir** Coding the pipelines, training the model, baseline testing
2. **Harshal** Report writing, fine-tuning the model and baseline testing.
3. **Vishal** Coding the pipelines, coding the preprocessing steps, baseline testing.

References

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. 2022. Vitpose: Simple vision transformer baselines for human pose estimation.

Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. 2021. Transpose: Keypoint localization via transformer.