Human Pose Estimation using Convolutional Neural Networks

Richard Hsieh {rhsieh91@stanford.edu}

Introduction

Human pose estimation involves identifying key joints and limbs on the human body, which can be useful for activity classification and body movement predictions. Traditionally, this has been a challenging problem due to joints being small/obscured and the need for contextual understanding of the human body in question. Convolutional Neural Networks (CNN) provide an end-to-end learning approach for tackling this problem without the need for hand-crafted features. In this case, the input is a RGB image containing a human figure(s) and the outputs are (x,y) coordinate pairs for key joints on the figure (e.g. elbows, shoulders, hips, etc.).

Data

- Leeds Sports Pose Dataset and its extension
- Contains 12,000 images of humans engaging in a sport or activity
- Labeled with (x,y) coordinates for 14 key joints
- Images rescaled to consistent 96 x 96 squares
- Dataset augmented to 24,000 images by flipping about vertical axis

Figure 1: Sample images from Leeds Sports Dataset

Model

- Formulated as a regression task
- Most promising candidate is a 11-layer CNN with Batch Normalization
- Xavier weight initialization
- Adam optimizer for adaptive learning rate
- Mean squared error loss function

Figure 2: Most promising model with Batch Normalization

Results

- Model was difficult to train and final results unfortunately show high bias and variance
- Batch normalization significantly improved bias
- Final training loss = 139.29
- Final validation loss = 427.41

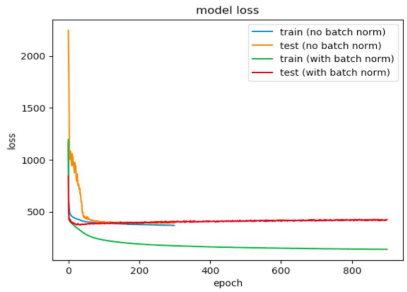

Figure 3: Training loss comparisons with and without Batch Normalization

Discussion

- Predictions illustrate human figures but unable to match exact pose shown in image
- Performs especially poorly on contorted poses
- Model may incorrectly identify main subject if image contains multiple humans

Figure 4: Example model predictions on training set

- Saliency maps indicate that convolutional layers are able to roughly identify edges
- Need deeper layers to distinguish finer body features and joints

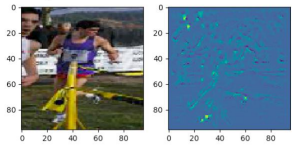

Figure 5: Saliency map for convolutional layer showing regions of high activation

Future Work

- Deeper CNN and more regularization to improve bias and variance
- Merge pose estimation CNN with image classification CNN to tackle activity recognition

References

[1] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks.
[2] Amy Bearman and Catherine Dong. Human pose estimation and activity classification using convolutional neural networks.