Human Pose Estimation and Activity Classification using Convolutional Neural Networks

Amy Bearman, Catherine Dong

Background

The problem of **human pose estimation** involves the identification of the location of keypoints of the body, which include major body parts and joints. There are various applications associated with this problem, such as action classification and body movement prediction.

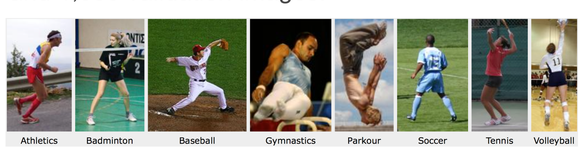
Pose estimation has made significant progress in the last few years, especially with the introduction of **Deep Convolutional Neural Networks (CNN)**. The identification of body keypoints has proved to be a challenging problem due to small joints, occlusions, and the need to capture context.

Data & Preprocessing

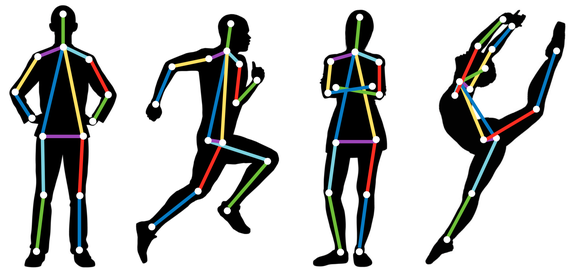
We are using the MPII Human Pose Dataset (above) and the Leeds Sports Pose Dataset (below). Images from both datasets include labels denoting the pixel location of certain joints in the body, referred to as body keypoints. The MPII database also includes 410 activity and 20 activity category labels.

The data is **preprocessed** by resizing the images to squares and subtracting the mean, and the dataset is augmented by taking random crops and performing horizontal flips of the image.

After **data augmentation**, the MPII dataset has 180,000 training and 30,000 validation images, and the Leeds dataset consists of about 16,500 training and 1,000 validation images.

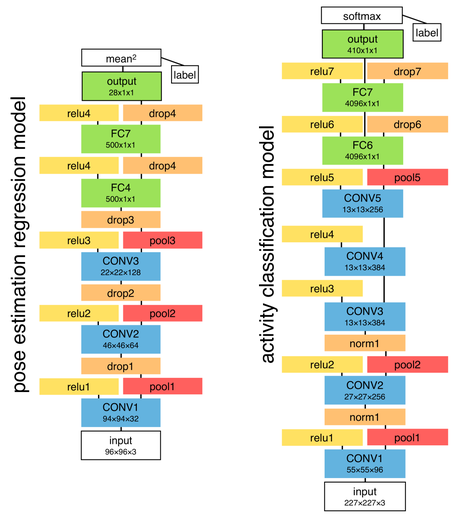


Problem



Our objective is two-fold: pose estimation and activity classification. For the former task, we aim to **identify x-y pixel coordinates for 14 body joints** depicted in the figures above. For the latter, we aim to **label the images based on activity category** (20 classes, e.g. “winter activity”) and specific activity (410 classes, e.g. “downhill skiing”).

Models



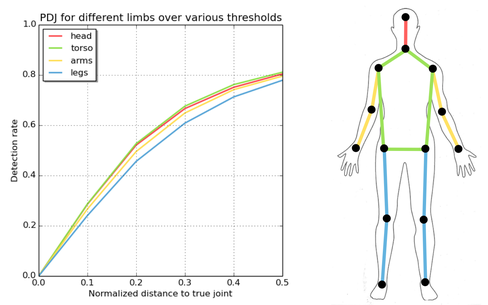
Pose estimation is a **regression** problem, and thus the CNN model for this problem uses **mean squared error** as the top layer. The output is a vector of 28 numbers representing the x-y pixel coordinates of the 14 keypoints.

Activity classification is a standard multi-class classification problem which uses cross-entropy loss. We **fine-tuned** through all layers of a pre-trained Caffe reference model based on AlexNet.

Results & Analysis

POSE ESTIMATION

(Above) Predicted human pose trees on images from the Leeds Sports Pose Dataset. Each limb is colored the same across multiple images.



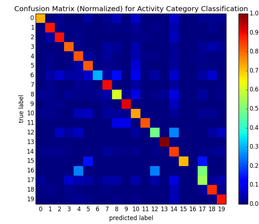
We evaluate using the **Percentage of Detected Joints (PDJ) metric**, which considers a joint correctly detected if the distance between the predicted and ground truth joint locations is within a certain fraction (which we define and vary) of the torso diameter. Using the PDJ metric means that all joint accuracies are evaluated using the same error threshold, as opposed to other metrics which tend to penalize shorter limbs.

Above, we present results over a range of normalized distances between predictions and ground truth labels. Results are accumulated into four categories: head (head and neck keypoints), torso (neck, shoulders, hips), arms (wrist, elbow, shoulders), and legs (ankles, knees, hips).

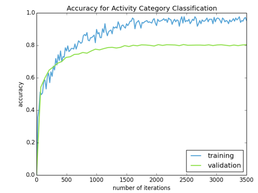
ACTIVITY CLASSIFICATION

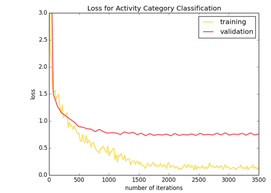
Classification of the images by activity category (20 classes) achieved a maximum validation **accuracy of 80.51%.**

Classification by specific activity (410 classes) performed worse, yielding a validation accuracy of 31.89%.



0: bicycling
1: conditioning
2: dancing
3: fishing and hunting
4: home activities
5: home repair
6: inactivity quiet/light
7: lawn and garden
8: miscellaneous
9: music playing
10: occupation
11: religious activities
12: running
13: self care
14: sports
15: transportation
16: volunteer activities
17: walking
18: water activities
19: winter activities





Future Work

We must further fine-tune the parameters of our model based on trends in our loss and training/validation accuracy data.

Our major objective involves a combination of the classification and regression tasks. The final model should take as input both the original image and the keypoint location estimations obtained from the pose estimation model and use this to determine the activity being performed. We predict that this additional input to our model will significantly increase classification accuracy.

References

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks."

Toshev, Alexander, and Christian Szegedy. "Deeppose: Human pose estimation via deep neural networks."

Nouri, Daniel. “Using convolutional neural nets to detect facial keypoints tutorial.”

Frameworks: Caffe (Berkeley Vision and Learning Center), Theano (Frederic Bastien, et al.)

We would also like to acknowledge the CS231N instructors Fei-Fei Li and Andrej Kaparthy for their guidance.