# Foot Keypoint Estimation using Convolutional Neural Networks

Jordan Nicholson
jnichol@stanford.edu

## Problem Description

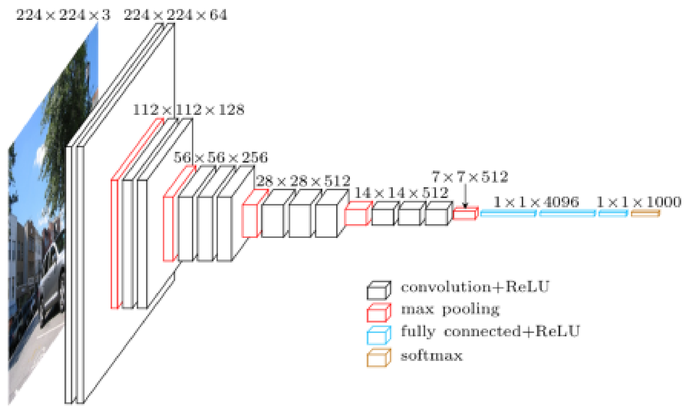
Pose estimation is an interesting and useful area of research in computer vision. Foot keypoint estimation builds upon previous work on human pose estimation, and in this pose estimation task, a VGG-16 [1] convolution neural network gets trained on a labeled dataset of foot keypoints. Pose keypoint estimation specifically has potential useful application in computer vision tasks related to augmented reality.

## Dataset

The foot keypoint dataset [2] is a labeled subset of 14000 training examples and 545 test examples from the COCO dataset. This project uses 900 training samples and 100 test examples. Each image in the dataset is resized to a 224 x 224 scale and each keypoint label is rescaled for compatibility with the VGG-16 network architecture. Each labeled image consists of 6 keypoint locations resulting in a list label of size 12.

Each image has 6 labeled keypoints indicating foot position within the image.

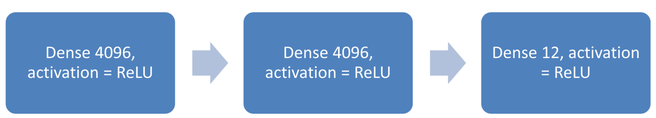
## Model Architecture




This project uses transfer learning to utilize learned parameters from the ImageNet dataset for general visual recognition tasks. This model uses a pre-trained VGG-16 convolutional neural network architecture with trained ImageNet weights and the top 32 layers frozen. The top layer contains two fully connected hidden layers with a total of 8192 network size. The model uses the Adam optimizer with a learning rate of 0.001 and mean squared error loss.

## Methodology

In this project, the neural network is implemented in Keras as a VGG-16 convolutional neural network with the last layer replaced with two hidden layers of size 4096 with ReLu activation functions and a 12 node output corresponding to the size 12 keypoint prediction. The model architecture was trained for 15 epochs with a 150 batch size.




The last layer trained to predict the size 12 keypoint label

## Results

| | Mean Squared Error | Accuracy |
|---|---|---|
| Train | 2822 | 0.178 |
| Test | 7742 | 0.11 |

## Discussion

- The mean squared error provides a good measurement of the distance between a model prediction and a ground truth keypoint since the loss is a squared sum of the element-wise difference between the prediction and ground truth keypoints.
- The accuracy does not provide much insight on the performance of the neural network model because it calculates accuracy based on the number of shared values between the prediction vector and ground truth keypoint vector without accounting for order.

## Future Work

Future work on this project involves training on a larger subsample of the dataset to improve generality of the VGG network for higher test set accuracy and implementing this model architecture with a lighter model such as MobileNet for embedded, smartphone applications.

## References

[1] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.
[2] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, “OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields,” in arXiv preprint arXiv:1812.08008, 2018.