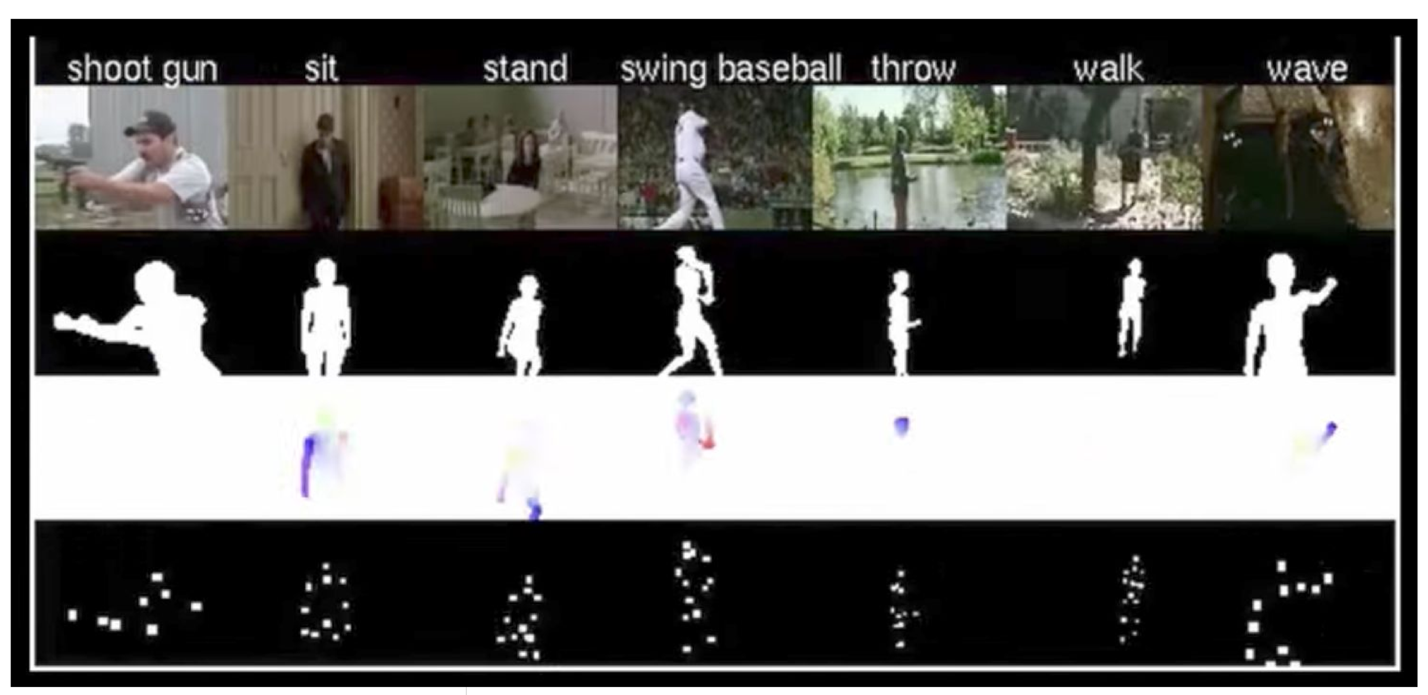
Human Motion Reconstruction from Action Video Data Using 3-Layer-LSTM

Jihee Hwang[1], Danish Shabbir[2]
[1]Stanford, Computer Science, [2]Stanford, Electrical Engineering

Introduction

Automatic motion generation from data is a challenging problem. Typically the model is trained using motion capture (MOCAP) data, which is limited in scope and diversity. Solving this problem using motion data extracted from Youtube videos would enable motion generation pipelines to feed from a wider range of action data, allowing for greater variability in the actions generated. Human action is expressed through a sequence of bodily movements. Hence, we seek to implement several variations of Recurrent Neural Networks (RNNs) to process our sequential motion data. Ultimately, robust generation of a wide range of actions could be made possible by incorporating adversarial training with other contextual data provided from the input video.

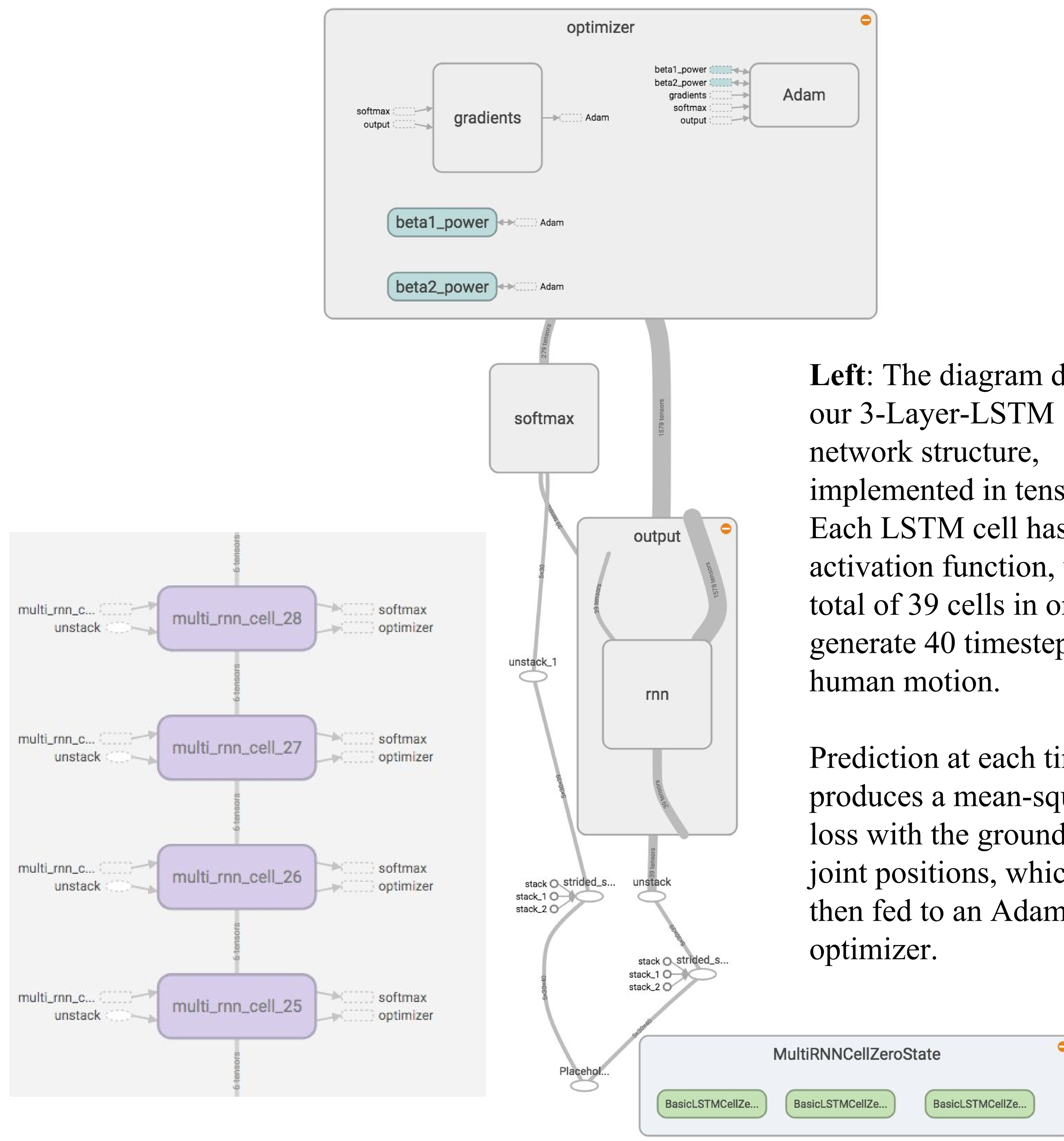
Data



We use the JHMDB dataset for our problem, it is a subset of the popular HMDB51 motion video dataset. Each of the 928 clips, divided into a total of 21 action classes, includes not only 15 annotated joint positions but also other information such as visible body parts, number of people, frame rate, and camera orientation.

While providing rich contextual background and a much wider data availability and spectrum of actions, motion data generated from videos are different from traditional motion capture data in that they are two dimensional, lack the number of frames (per individual motion instance) and are generally more prone to noise. To simplify our problem, we decided to focus on videos that i) contains only one person, ii) facing one direction per action category, iii) and have more than 40 frames per a clip's action instance.

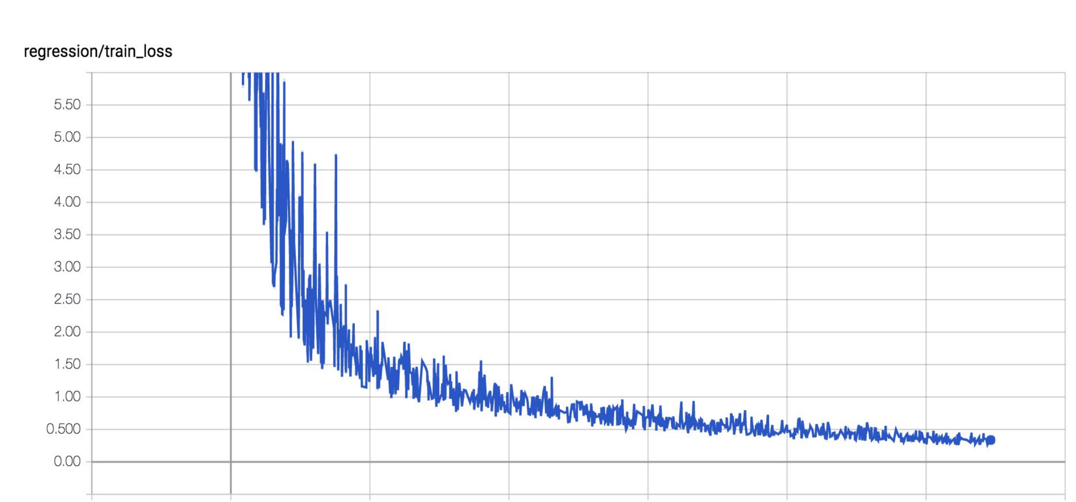
Network Design



Left: The diagram depicts our 3-Layer-LSTM network structure, implemented in tensorflow. Each LSTM cell has a tanh activation function, with a total of 39 cells in order to generate 40 timesteps of human motion.

Prediction at each timestep produces a mean-square loss with the ground truth joint positions, which were then fed to an Adam optimizer.

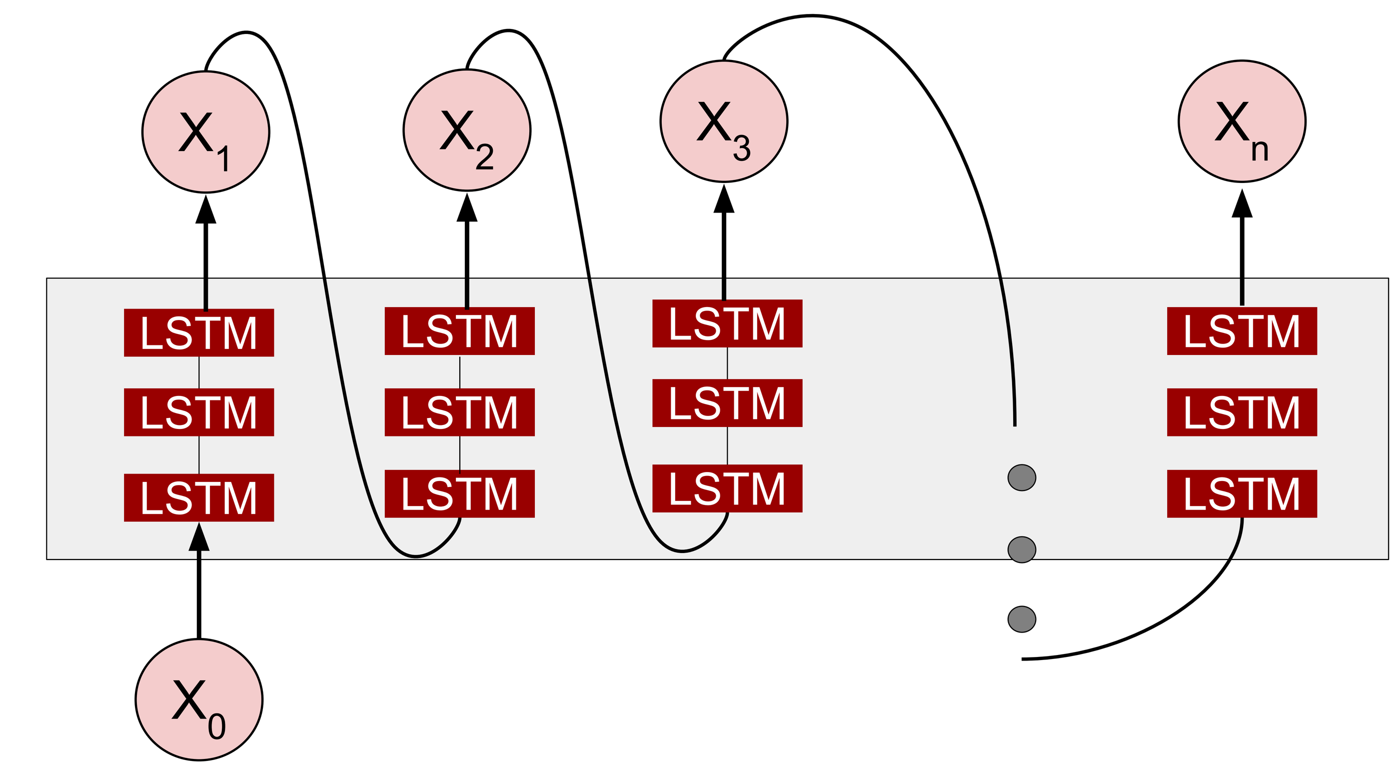
Loss Optimization

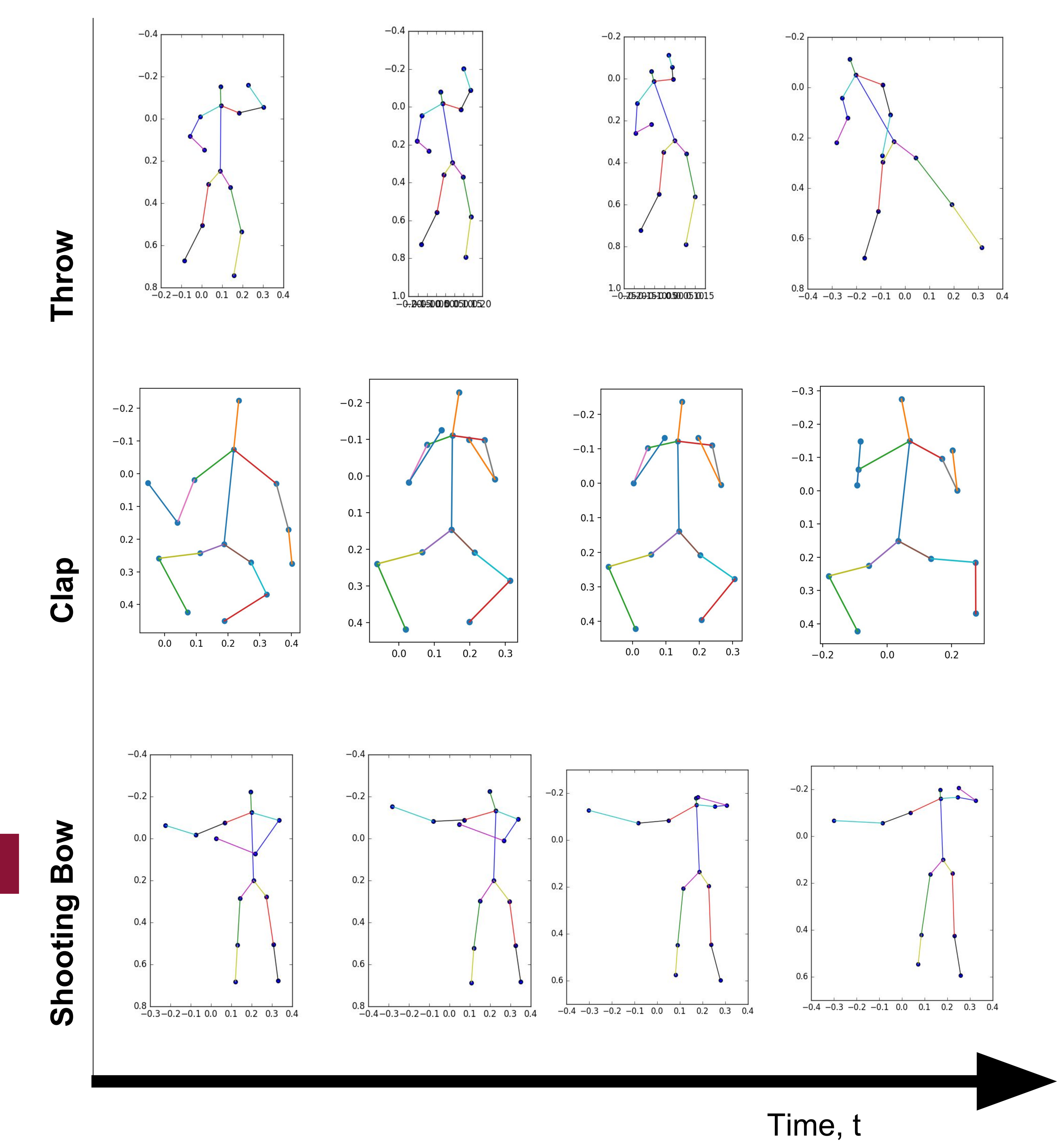


Top: Within 50,000 epochs, loss quickly goes down under 6e-4

Hyperparameter tuning was performed across three variables: learning_rate, hidden_state_size, and layer density.
Learning rate performs best around 1e-3 with training errors around 6e-4. The model performance improved as we increased state_size and num of layers; however, for computational efficiency and best performance we trained model with 3 layers and state_size of 200.

Motion Generation Pipeline





Top: After training our LSTM model, we then sample each action class using the first 5 seed frames of an action video clip that the model has not been trained on before. The results are surprisingly realistic, proving that even with noisy data such as motion annotation from videos, a wide range of human body movements can be reconstructed in a generalizable and robust way.