CSCE689 Project

Name: Pei Chen   UIN: 130005135

Main Improvements:

1. Collect much more data from YouTube and manually annotate them. (It is very difficult to collect those videos because the 3 types of foot sentimental actions are very spare in videos. Also, it is time consuming to manually annotate them)
2. Trim the data into pieces and maintain an average of 3 second per video.
3. Turn each frame into size of 240× 416 because the image details are not important for foot actions detection. This will reduce the memory consumption and model size.
4. Try to look at the videos as a sequence of images and regard it as a sequential labeling task. Although this idea fails, but it's good to know it is a bad idea to detect actions using sequential labeling model.

Research Topic

Detecting foot sentiment in videos:

Nervous Pacing: Many people will pace when they are stressed. This acts as a pacifier, as all repetitive behaviors do.

Foot withdrawing: During situations like interviews, interviewees will suddenly withdraw their feet and tuck them in under their chains when they are asked sensitive questions they might not like.

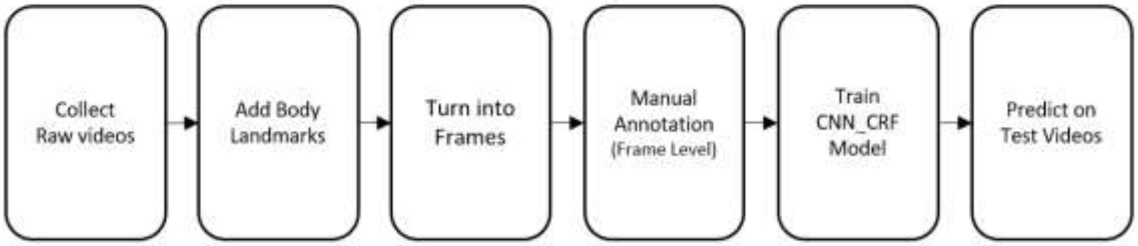
Foot turning away: When we are talking to someone, we might signal that we need to leave by gradually or suddenly pointing one foot toward the door. This is our non-verbal way of communicating "I have to go."

Dataset

I collect 22 videos clips for all the 3 types of actions, for both training and testing. Each clip is about 3 seconds and marked "Body Landmarks" using *OpenPose*. The splitting is as following:

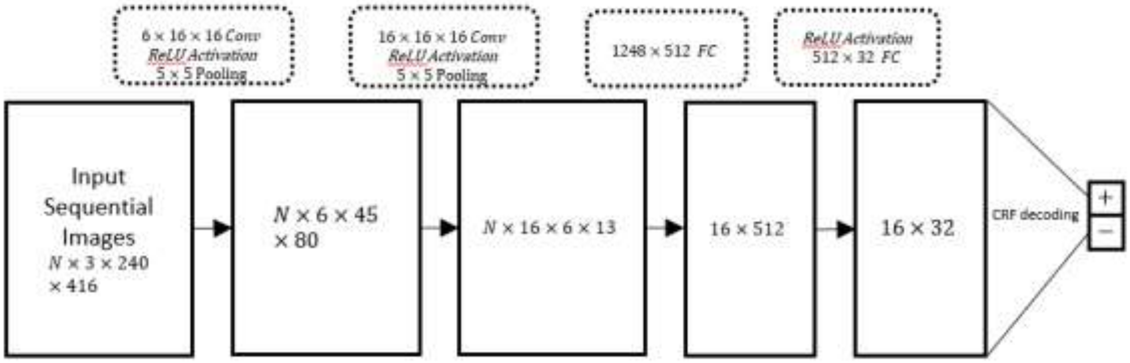
	Training Data	Testing Data
Nervous Pacing	foot_pacing_trian1.mark.mp4 foot_pacing_trian2.mark.mp4 foot_pacing_trian3.mark.mp4 foot_pacing_trian4.mark.mp4 foot_pacing_trian5.mark.mp4 foot_pacing_trian6.mark.mp4	foot_pacing_test1.mark.mp4 foot_pacing_test1.mark.mp4
Foot withdrawing	foot_withdrawing_trian1.mark.mp4 foot_withdrawing_trian2.mark.mp4 foot_withdrawing_trian3.mark.mp4 foot_withdrawing_trian4.mark.mp4 foot_withdrawing_trian5.mark.mp4 foot_withdrawing_trian6.mark.mp4	foot_withdrawing_test1.mark.mp4 foot_withdrawing_test2.mark.mp4 foot_withdrawing_test3.mark.mp4
Foot turning away	foot_turning_away_train1.mark foot_turning_away_train2.mark foot_turning_away_train3.mark foot_turning_away_train4.mark	foot_turning_away_test.mark.mp4

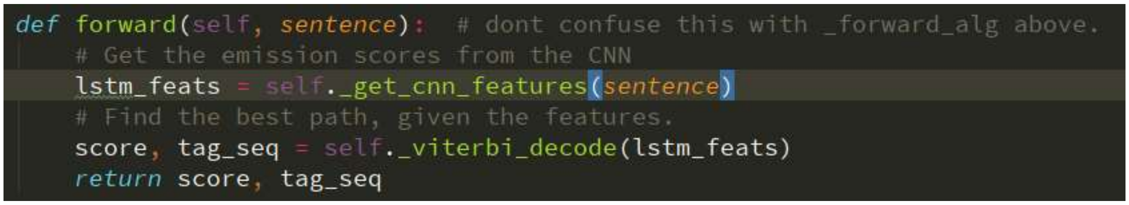
Procedures



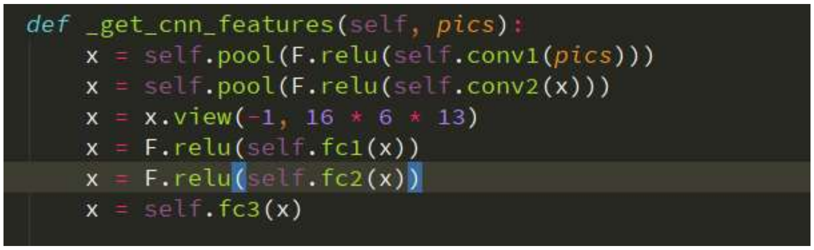
1. Collet videos from YouTube and trim them into 3 seconds per video.
2. Use *OpenPose* tools mark “Body Landmarks” in all clips;
3. Then transform the videos into frames, and store each frame as an image;
4. After that, manually annotate each frame as Positive (with target action) or Negative (without target action);
5. Train separate CNN-CRF models for each action type to classify the frames;
6. Test on the corresponding test data and get the scores for each frame.

Architecture:





```
def forward(self, sentence):  # dont confuse this with _forward_alg above.
    # Get the emission scores from the CNN
    lstm_feats = self._get_cnn_features(sentence)
    # Find the best path, given the features.
    score, tag_seq = self._viterbi_decode(lstm_feats)
    return score, tag_seq
```



```
def _get_cnn_features(self, pics):
    x = self.pool(F.relu(self.conv1(pics)))
    x = self.pool(F.relu(self.conv2(x)))
    x = x.view(-1, 16 * 6 * 13)
    x = F.relu(self.fc1(x))
    x = F.relu(self.fc2(x))
    x = self.fc3(x)
```

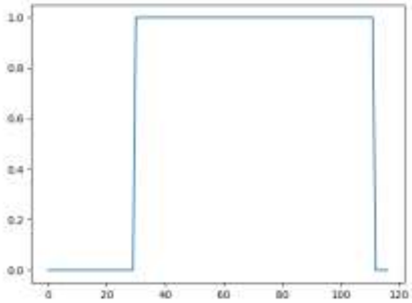
I regard the problem as a Sequence Labeling problem rather than a simple classification problem. The main architecture is a 2-layer CNN (Convolutional Neural Network) plus a CRF (Conditional Random Field) labeling layer. One trick I use is to add mark to original videos instead of only use the landmark video. Another trick is that I use a CRF to capture the connections between adjacent frames so as to label the frame.

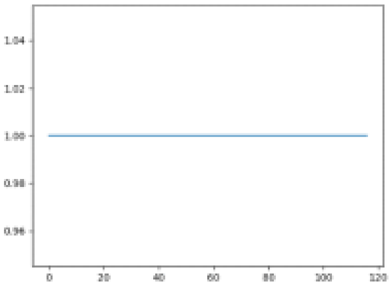
Input

In order to maintain the consistency for all the clips, I turn all the frames into shape of (240, 416, 3).

Output

For each test clip, I predict the score of each frame with the trained model and store them in a corresponding *JSON* file. The following figure are the gold labels and predicted results from foot_pacing_test1 data.





Hyperparameters

Batch_size	Optimization	Learning Rate	Max_epoch
16	SGD	0.001	30

Training and Testing Performance

Unfortunately, the proposed CRF model fails in such a task. It will tend to predict all the frames as positive and the transaction matrix between adjacent frames does not matter for the task.

Next Steps:

Improve the model with such enlarged and well annotated, well organized data.

Links

Videos: https://www.youtube.com/playlist?list=PLlUqXrHW9mC_GlGbgiAUmPIwDMZWuLSP4

Code: https://github.com/brickee/foot_sentiment.git