Tennis Shot Recognition through Spatiotemporal Deep Neural Networks

Video - https://youtu.be/NSkWjPX3jrQ

Ganapathy Sankararaman (ganasank@stanford.edu), Siddharth Buddhiraju (sbuddhi@stanford.edu)

CS230 FINAL PROJECT PRESENTATION, STANFORD UNIVERSITY

Motivation & Objectives

- Recognition of human action in videos has generated substantial attention from the deep learning community in recent years [1-4].
- In this project, we aim apply the video classification problem to detect tennis shots. We aim to build a deep neural networks which takes in an RGB video and classifies it into one of 6 tennis shot classes
- This work is also aimed at making progress towards a larger project that we are working on, where we envision a need to identify and correct player postures while playing various tennis shots.

Dataset

- The THETIS dataset [5] consists of 12 Tennis shots performed by 31 amateurs and 24 experienced players about 3 - 5 times/shot. 1980 RBG videos in total in .avi format, with about 80 frames per video.
- RGB video dataset split 0.8:0.1:0.1 into training (1584), validation (204) and test sets (192) with equal proportions of each shot.

Baseline and Bayes Error

- **Baseline LRCNN**: 'LRCNN' model by Chow and Dibua [6]. To make their shot prediction, they extract features from video frames using the Inception V3 network pre-trained on ImageNet. Then, they feed the features into a many-to-many LSTM network, which is trained to output one of the 6 tennis shots as its prediction. Their results are shown in the first row of Results (baseline).
- **Bayes Error:** We asked 5 people with a good understanding of tennis to classify 24 videos into the 6 different shot categories. By using a voting method to combine their results, they obtained an **87.5%** accuracy. Our Bayes error on these videos is thus **12.5%**.

Models and Methods

- **Loss Function:** Categorical cross-entropy
- **Improved LRCNN Model:** We generated sequences for the THETIS RGB videos for 16 frames from Inception V3 network and used a a Bidirectional LSTM, an extra hidden layer with 128 units, and an increased dropout rate of 50% (baseline dropout was 30%), the train and test accuracies and the F1 score are better than the baseline LRCNN. Results shown in Table in the next column.

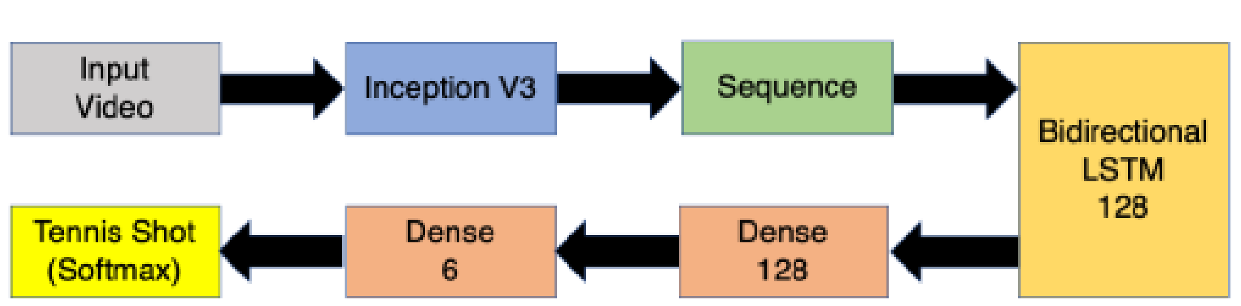


Figure: Improved LRCNN

- **EvaNet + Transfer Learning:** We also explored the models in Ref. [7], where an evolutionary algorithm was employed to find optimal video classification neural architectures. Here, a meta-architecture model was introduced for which the high level connectivity between modules is fixed, but the individual modules can evolve. Two such optimized models ("Model 0" and "Model 1") trained on HMDB and Kinetics-400 were made available in their GitHub repo.
- For transfer learning, we pass the THETIS RGB videos through EvaNet architectures and obtain sequences two layers before the softmax and trained neural networks on the sequences:
- **Shallow network:** The Shallow Network had a Conv3D layer with 32 filters of shape (1; 7; 7), BatchNorm on this Conv3D followed by a Dense with 6 outputs and Softmax to get the output. Model 0 performed poorly on the shallow network.
- **Deep network:** To improve performance on Model 1, we used a deeper architecture shown below. Accuracies on both the networks for both models are tabulated below.

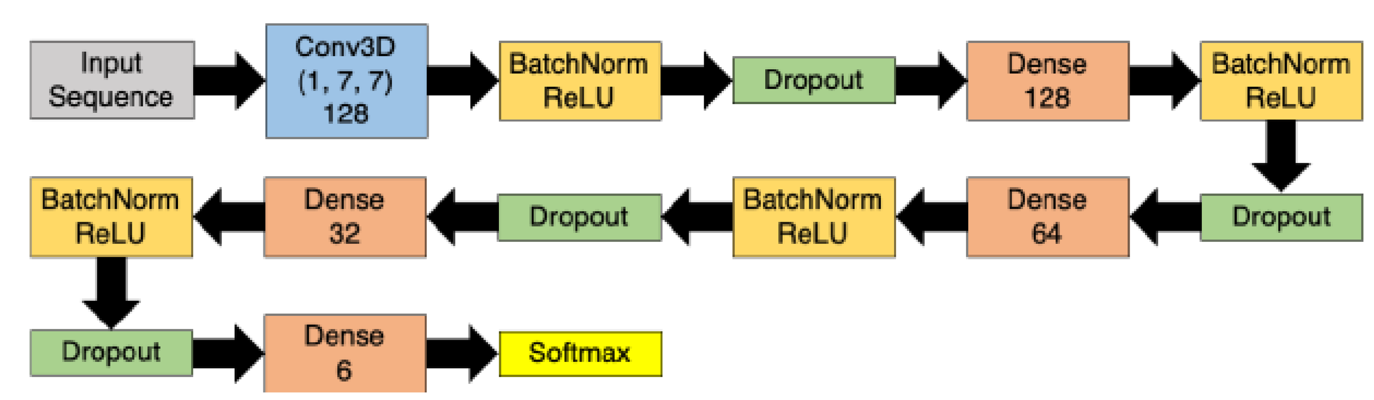


Figure: EvaNet+Transfer with deep neural net

EvaNet Transfer	Model 0	Model 1
Shallow network	71.87%	**80.21%**
Deep network	**77.60%**	78.12%

Table: EvaNet+Transfer with shallow and deep networks

- **Ensembling**: For the best results, we ensembled the outputs of Model 0 on the deep network and Model 1 on the shallow network. Results shown in the Table below.

Results

Model	Train Accuracy	Test Accuracy	F1 Score
LRCNN-Baseline	98.7%	82.3%	0.82
LRCNN-Improved	100.0%	84.4%	0.84
EvaNet+Transfer	99.8%	84.4%	0.84

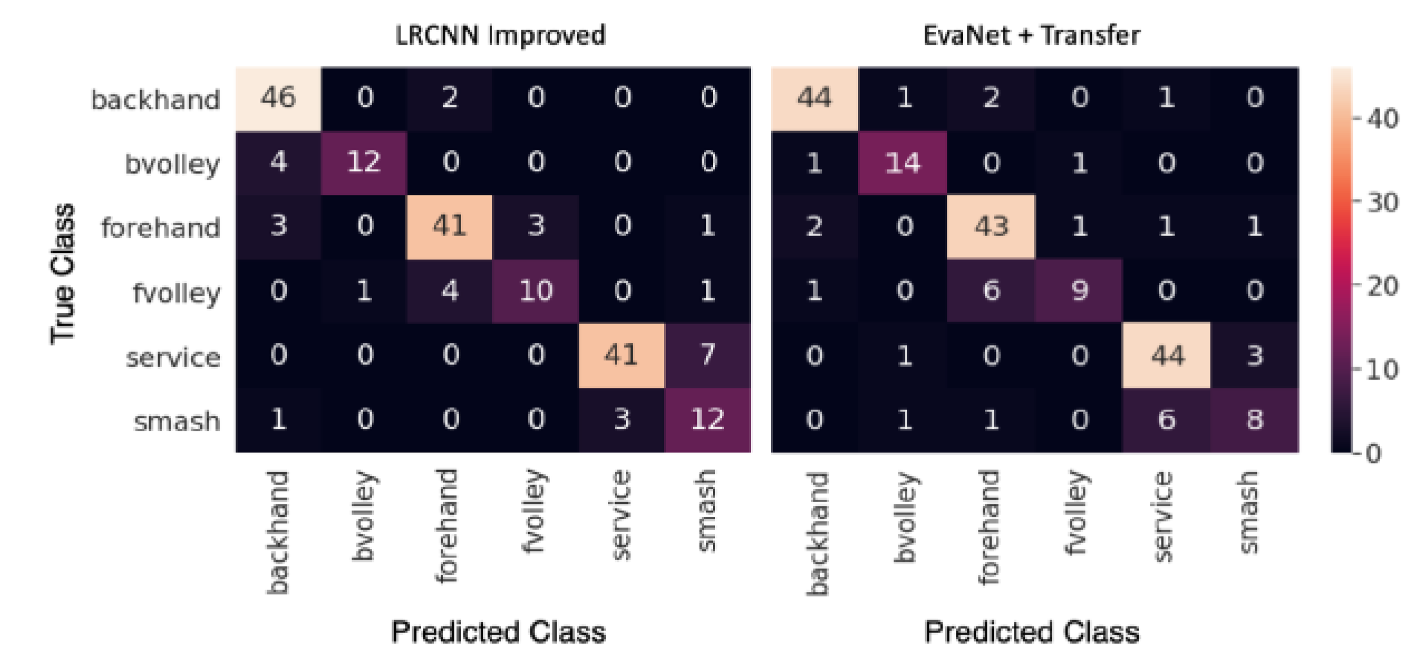
Table: Results from our baseline, improved LRCNN and ensembled EvaNet+Transfer

Discussion and Error Analysis

- Ensembling Shallow Network of Model 0 with Deep Network of Model 1 improved the performance significantly, as can be seen in the results.
- Model 0's poor performance on the shallow network is likely because the output sequences were not sufficiently clustered together based on their true labels when compared to output sequences obtained from Model 1.
- To verify this, we flattened the sequences from Model 0 and Model 1 into vectors and clustered these sequences based on their true labels and computed the normalized centroid distance between classes for the two Models.

Class_i-Class_j	Model 0	Model 1
backhand-forehand	2.83	9.59
forehand-bvolley	4.33	23.25
smash-serve	10.27	19.73

- Much better class separation in Model 1 than in Model 0.
- **Confusion Matrices:** In both models, bvolley-backhand, fvolley-forehand, and service-smash were misclassified often, which is consistent with human error.
- Misclassifications likely due to the nature of similarity in these shots, since most of the players in this dataset are amateurs and the shots are filmed indoors without a tennis ball.



Conclusion and Future Work

- On improved LRCNN and Evanet+Transfer learning, we are able to identify 6 classes of shots with 84.4% accuracy.
- The cause for low accuracy is the shot quality. The large variance in the model performance is due to lack of sufficient data to train the model.
- Generating more quality data and training the model on it will help in removing the variance and improving accuracy, which is needed to further advance in predicting correctness of posture.

References

1. Karpathy, A., et al. Large-scale videoclassification with convolutional neural networks. CVPR 2014 (pp. 1725-1732).
2. Simonyan, K., & Zisserman, A. Two-stream convolutional networks for action recognition in videos. NeurIPS 204 (pp. 568-576).
3. Feichtenhofer, C. et al. Convolutional two-stream network fusion for video action recognition. CVPR 2016 (pp. 1933-1941).
4. Feichtenhofer, C. et al. Spatiotemporal multiplier networks for video action recognition. CVPR 2017 (pp. 4768-4777).
5. Gourgari, S et al. Thetis: Three dimensional tennis shots a human action dataset. CVPR 2013. (pp. 676-681).
6. Chow, V., & Dibua, O. (2018). Course project for CS 230. http://cs230.stanford.edu/files_winter_2018/projects/6945761.pdf.
7. Piergiovanni A. J. et al. Evolving space-time neural architectures for videos. ICCV 2019 (pp. 1793-1802).