

Classification of Social Signals Using Deep LSTM-based Recurrent Neural Networks

Himanshu Joshi, Ananya Verma, Amrita Mishra

DSPM International Institute of Information Technology Naya Raipur, INDIA - 493661



1 Introduction

- Problem Statement
- State-of-the-art
- Motivation

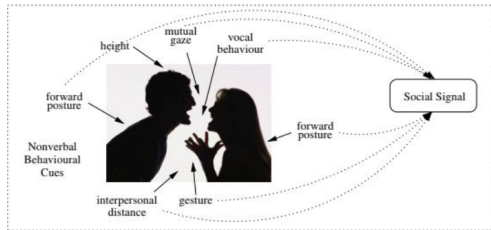
2 Deep-LSTM based Social Signals Classification Approaches

- MFCC + Cluster based Feature Extraction Approach
- Recurrent Neural Network
- Deep-LSTM based Classification Predictions
- Generic Work-Flow of Deep-LSTM based Classification Approach

3 Simulation Results

4 Conclusion

Problem Statement




- Individual's decision-making is strongly affected by nonverbal behaviour. Almost 60% of human communication is based on non-verbal cues¹
- Many important characteristics may reveal during decisions/interactions such as thought process, etc.

In the technology driven era, social intelligence is important in order to recognize social behaviours like turn taking, politeness and disagreement.

¹M. Pantic, R. Cowie, F. D'Errico, D. Heylen, M. Mehu, C. Pelachaud, I. Poggi, M. Schroeder, and A. Vinciarelli, "Social signal processing: The research agenda," in *Visual analysis of humans.*, Springer, 2011, pp. 511–538.

- Many combinations of Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) have been used such as HMM+statistical tools ², GMM+Support Vector Machines (SVM) ³
- **Shortcomings**
 - Models fail to leverage the inherent temporal correlation associated with speech data owing to their conditional independence assumption between different operating modules.

²H. Salamin, A. Polychroniou, and A. Vinciarelli, "Automatic detection of laughter and fillers in spontaneous mobile phone conversations," in *2013 IEEE International Conference on Systems, Man, and Cybernetics*, 2013, pp. 4282–4287.

³A. Janicki, "Non-linguistic vocalisation recognition based on hybrid GMM-SVM approach." in *INTERSPEECH*, 2013, pp. 153–157. 

Motivation

- Shortcomings of traditional HMM & GMM based classification methods.
- Shortcomings of RNN based classification models + inspired by the Recurrent Neural Networks ideology.



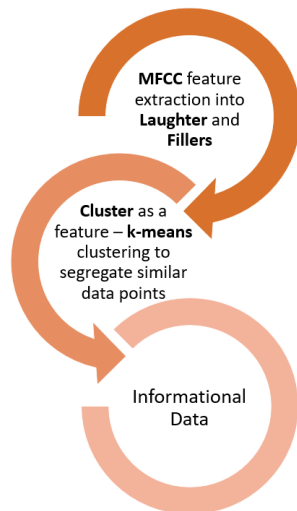
- Deep LSTM network feature of storing long sequences.
- Combining MFCC feature extraction method + Clustering as a feature.
- K-means clustering algorithm.



Deep-LSTM employed with Cluster as a feature classification approach

MFCC + Cluster based Feature Extraction Approach

- 1 Why MFCC? : Equal spacing of the frequency bands closely approximates the human auditory response.
- 2 Clustering as a feature effectively segregates a given set of data points with similar traits and helps discover hidden patterns within the dataset.



Recurrent Neural Network

- Recurrent Neural Network (RNN) is a type of Neural Network where the output from previous step is fed as input to the current step.
- RNN have “memory” i.e. it remembers all information about what has been calculated.
- It uses the same parameters for each input as it performs same task on all the inputs or hidden layers to produce the output.
- **Shortcomings**
 - Cannot process very long sequences.
 - Unstable
 - Cannot be stacked into very deep models.

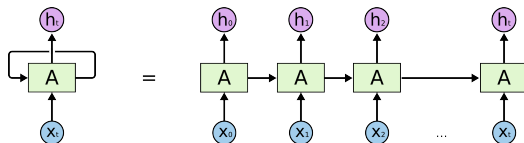


Figure: Diagram of Recurrent Neural Network⁴

Deep-LSTM based Classification Predictions - 1

- LSTMs are designed straightforward to avoid the long-term dependency problem faced in RNNs.
- The memory units remembers information for long periods of time practically, giving it as it's default behaviour.
- Key to LSTMs is the cell state.
- LSTM does have the ability to remove or add information to the cell state, carefully regulated by structures called gates. ⁵

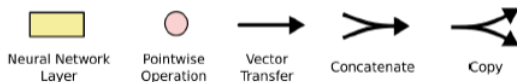


Figure: LSTM Notations

⁵URL: <http://colab.github.io/posts/2015-08-Understanding-LSTMs/>

Deep-LSTM based Classification Predictions - 2

- Gates are a way to optionally let information through. There are 3 types of gate- forget gate, input gate and output gate.
- The sigmoid layer outputs numbers between zero and one, describing how much of each component should be let through. A value of zero means “let nothing through,” while a value of one means “let everything through!”

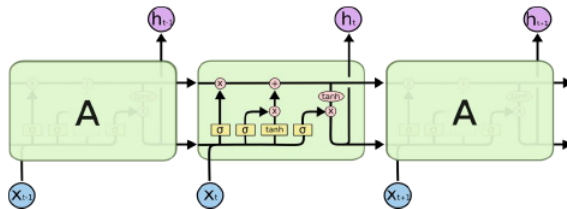
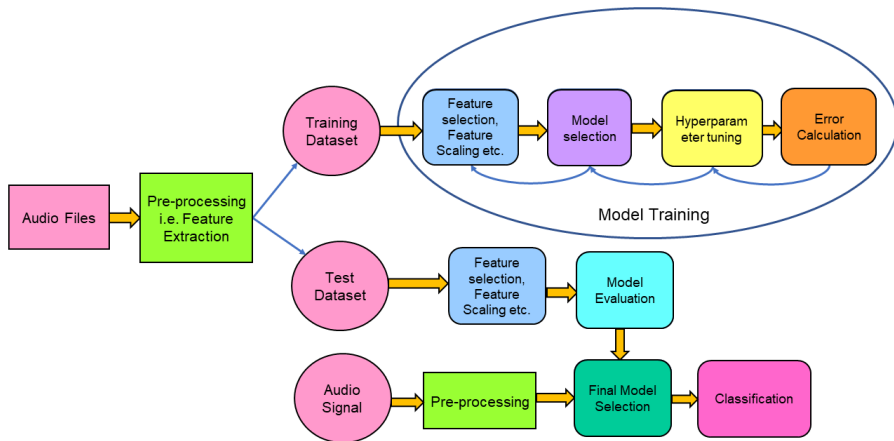


Figure: Repeating Layers in LSTM containing four layers

Generic Work-Flow of Deep-LSTM based Classification Approach



- SSPNet Vocalization Corpus (SVC) whose description is provided in the Social Signals sub-challenge of the 2013 Interspeech Computational Paralinguistics Challenge⁶ is employed to investigate the accuracy of the Machine Learning and Deep-LSTM-based PL prediction schemes
- The corpus comprises of 2763 audio clips of 11 seconds time-frame.
- Input features : 57 males 63 female subjects, summing up to 120 subjects.
- **Labels** : Laughter, Filler.

⁶ B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Wengler, F. Eyben, E. Marchi et al., "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013

Performance Comparison of Machine Learning and Deep-LSTM Classification Approaches

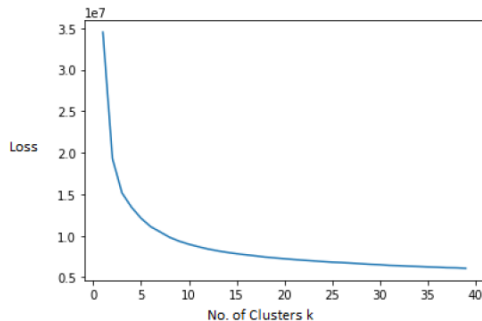


Figure: Loss versus No of Clusters k

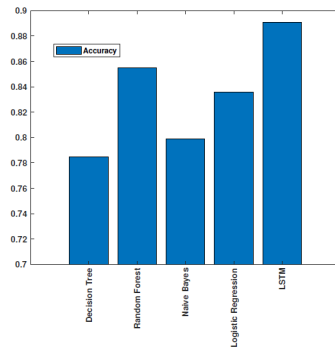


Figure: Accuracy Comparison between classification models

Accuracy of Deep-LSTM + Cluster as a feature is higher than other classification models
 \implies Deep-LSTM based classification superior

Result - Takeaways I

LSTM Network Topology	Training Accuracy	Testing Accuracy
20-30-2	93.49%	87.06%
20-40-2	93.90%	86.17%
20-50-2	95.04%	88.18%
20-60-2	95.14%	86.50%
20-80-2	96.35%	86.17%

Table: LSTM based model comparison without clustering as a feature

- Less number of hidden layers : less amount of learning by the network and extremely high number of hidden layers : over-fitting of the model \implies poor accuracy.
- The best result in terms of testing accuracy is obtained for the network with a hidden layer size of 50.

Result - Takeaways II

LSTM Network Topology	Training Accuracy	Testing Accuracy
21-30-2	93.66%	86.25%
21-40-2	94.87%	87.14%
21-50-2	95.55%	89.15%
21-60-2	95.07%	87.22%
21-80-2	96.35%	86.5%

Table: LSTM based model comparison with clustering as a feature

- Adding clustering as a feature resulted in the size of the input layer to be increased from 20 to 21 since the additional cluster feature information is included.

- Deep-LSTM model outperforms traditional machine learning models for classification of social signals into laughters and fillers.
- Further, incorporating cluster as a feature for the LSTM-based classification via a k-means clustering algorithm demonstrated significant improvement in both training and testing accuracies.

Any Questions?

himanshuj16101@iiitnr.edu.in, ananyav16100@iiitnr.edu.in, amrita@iiitnr.edu.in