Compact LSTM-based model for classifying high dimensional multi-modality data

Barath Mohan U [*] **Vidhi Sinha** [*]

Abstract

High dimensional data requires feature reduction or selection methods while using traditional classification methods. The available dataset combines EEG, facial and audio modalities during speaking and imagined part of speech in experimental task, which increases the number of features drastically. We propose an LSTM-based architecture which shares weights across features to counter this problem. We see that this performs better than the baseline SVM and feed-forward neural network methods in the multi-class classification problem.

1. Technical Details

EEG recording was done by the authors on 14 participants where each participant was looked at screen for 30-40 minutes(Zhao & Rudzicz, 2015). One of 7 selected phonemic/syllabic prompts and 4 words from Kent's list of phonetically-similar pairs were shown on the screen. During each trial the subject had to go through 4 successive stages i.e. **rest**, **stimulus**, **imagined**, **speaking**. Each EEG data trial was segmented into those 4 stages, windowing to 10% of segment and 50% overlap between consecutive windows was done. For each window, features like mean, median, standard deviation, etc along with first and second derivatives of the features were calculated, giving 1197 features for 62 channels. For audio recording and facial data, a subset of features were calculated. Pearson correlation coefficient was computed between the class labels and data to rank the features, selecting $N \in [5, 100]$ of them. Training and test set had data from 13 and 1 subjects respectively. Using those set of features, a feed forward neural network (FFNN) with 2 hidden layers of 40 nodes each, having exponential linear unit activation was trained. The output was softmax activation for 7 classes. The FFNN has 5967 parameters. SVM-rbf kernel model was trained on training set using 'one vs all multi-class method'.

The proposed architecture is shown in Fig. 1. The thinking and speaking EEG states are fed in parallel to the network. They are each of size $time(19) \times electrodes(62) \times features(63)$. At each state, a convolutional LSTM with a filter of size (62×1) is used. The convolution computes a linear combination of the electrodes and the filter is shared across features. The weights in each gate of the LSTM are in turn shared across time. This is followed by another LSTM layer. The audio data $(time(19) \times features(63))$ is passed through an a single LSTM. To the facial data $(features(42) \times locations(6))$, a convolution of size (1×6) is applied to calculate a linear combination of the locations and these weights are shared across the features. The model has 5509 parameters, which is close to the FFNN model.

2. Results

The test accuracies for the multi-class classification for the different models for different subjects as test subjects is shown in Fig. 2. The average test accuracies across subjects were as follows - **SVM - 0.1422** ; **Feed-forward neural network - 0.1385** and **proposed method - 0.1698**. The accuracies of all models are poor, but the proposed model performs marginally better across subjects (more data could improve this).

3. Novel Contributions

The novelty lies in the proposed model architecture. It does not require the features to be hand-picked before feeding the data into it. Instead, it exploits the meaning of these features in order to share weights across them. The LSTM layer allows to use the same weights across time points, while the convolutions allows to share weights across the different computed features. The different modalities and EEG states are processed in parallel before concatenating; this decreases the number of weights considerably as well. Coding, report writing and presentation work was done with equal contribution for this project

4. Tools Used

Tensorflow-gpu (tf) - For building and training the FFNN and the proposed LSTM-based architecture; **Scikit-learn (sci)** - For training the SVM. The codes used in this project are available on the github link - `https://github.com/BarathMohanU/MLSP-Project`

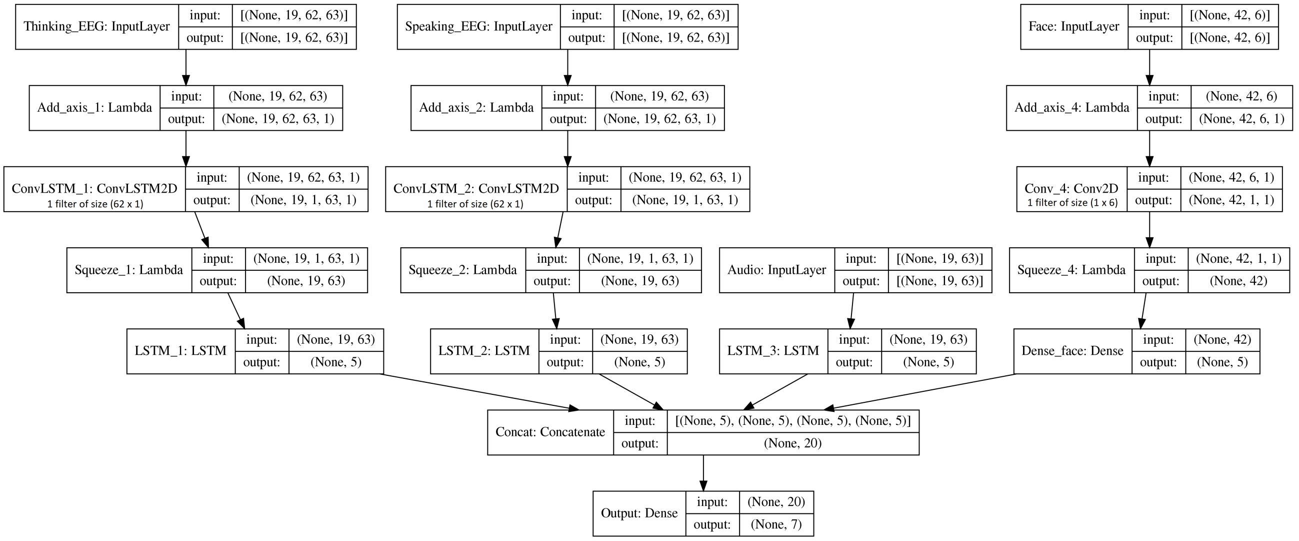

Figure 1. Proposed model architecture which processes the modalities in parallel while sharing weights within them to reduce the number of parameters.

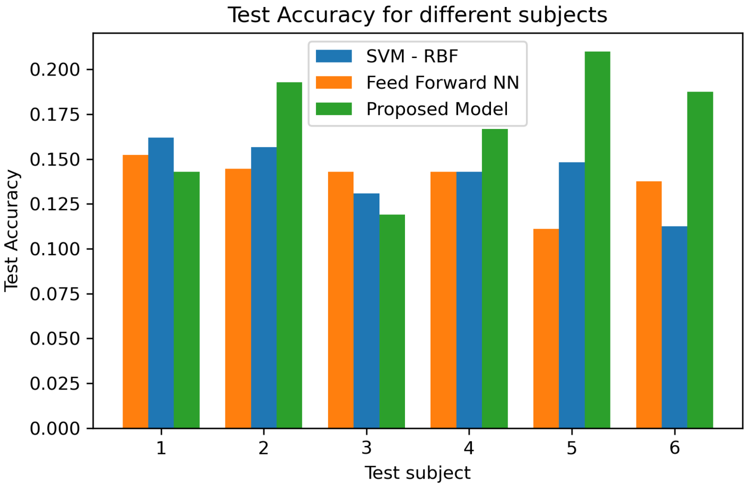

Figure 2. The average test accuracies for the different models for different subjects as test data.

References

https://scikit-learn.org/stable/.

https://www.tensorflow.org/.

Zhao, S. and Rudzicz, F. Classifying phonological categories in imagined and articulated speech. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 992–996. IEEE, 2015.