Report

Group 15

Group members:

- Penelope Brenner (2049490)
- Arash Mirshahi (2060358)
- Jurre van Opzeeland (2082072)
- Katherine Williams (2055908)

Data loading and processing

In this project, the raw audio files, initially in pickle format, were transformed into Mel-spectrograms. This transformation utilized parameters including a sampling rate of 8000 Hz, an FFT size of 2048, 128 Mel bands, and a hop length of 512. This preprocessing step was essential for standardizing the varying lengths of the audio files into a uniform format for neural network processing. Mel-spectrograms were specifically chosen due to their ability to encapsulate crucial audio signal features like pitch and tone, which are vital for accurately recognizing emotional valence.

Architecture design

Our chosen architecture is a bidirectional LSTM, which is particularly suited for sequential data analysis like Mel-spectrogram data used in valence prediction from audio signals. Unlike traditional LSTMs, which only capture temporal dependencies in one direction, the bidirectional approach processes data in both forward and backward directions, offering a richer understanding of context and improving the model's ability to discern the emotional nuances in audio sequences. The model consists of a two-layer bidirectional LSTM with 128 hidden units per layer, enhanced by a dropout rate of 0.2 to prevent overfitting. This configuration was selected because the two layers increase the model's ability to learn more complex patterns, and the bidirectional setup ensures comprehensive temporal analysis, crucial for capturing the full emotional range expressed in audio. The final layer is a fully connected linear layer with an input size of 256 (double the LSTM's output due to bidirectionality), leading to a single output representing the predicted valence, a measure of emotional content in the audio. This architecture was preferred over others due to its effectiveness in similar sequence learning tasks and its robustness in handling variable-length input, crucial for our dataset where audio clip lengths vary significantly.

Experiments

Training: The final model was trained using the Adam optimizer with Mean Squared Error (MSE) as the loss function over 20 epochs.

Hyperparameter Tuning:

LSTM1: We Started with a basic single-layer LSTM model, setting learning rates initially at 0.001. (128 hidden units).

CNN: We also tried a simple CNN structure, similar to the one used in Module 8 (2 convolutional layers, ReLU, and pool_stride=1).

LSTM2: After testing various configurations, the final model was a bidirectional dual-layer LSTM with 128 hidden units each. A dropout of 0.2 was implemented, alongside L2 regularization and a learning rate scheduler starting at 0.001, and reducing every 5 epochs.

Results

The model was evaluated using MSE on both the training and testing sets. The **LSTM2** achieved an MSE of 0.6798 on the test set.

	Validation set	Test Set
MSE	0.6738	0.6798

Table 1: Table presenting the results

The error analysis showed that a bidirectional LSTM has a clear advantage over a more standard CNN or unidirectional LSTM model architecture. This is clearly shown by a lower MSE on the validation set.

	LSTM2	CNN	LSTM1
MSE	0.6738	0.6937	0.7536

Table 2: Table presenting error analysis

Conclusions

Discussion Section:

- **Insights from Results:** The bidirectional LSTM model effectively captured the emotional valence from audio signals, as seen from the MSE scores on the validation and test sets of 0.6738 and 0.6798, respectively. This demonstrates that sequential models are suitable for this type of task. However, some variations in extreme valences indicate potential areas for improvement.

Conclusion Section:

- **Model Effectiveness:** The bidirectional LSTM architecture, with its ability to handle sequential data, proved to be effective for emotion recognition from speech.
- **Future Improvements:** Incorporating additional audio features (e.g., pitch, volume ,intensity) and experimenting with more complex architectures like CNN-LSTM hybrids (Lilhore et al., 2023) could further enhance the performance.
- **Challenges:** Handling varying audio lengths and optimizing hyperparameters were significant challenges, mainly due to scarcity of time and computing power. This was successfully mitigated through careful preprocessing and experimentation.
- **Theoretical Connections:** The results align with the theoretical strengths of bidirectional LSTM networks in capturing temporal dependencies(Du et al., 2020), reinforcing their suitability for audio-based emotion recognition tasks.

Proposed Architecture

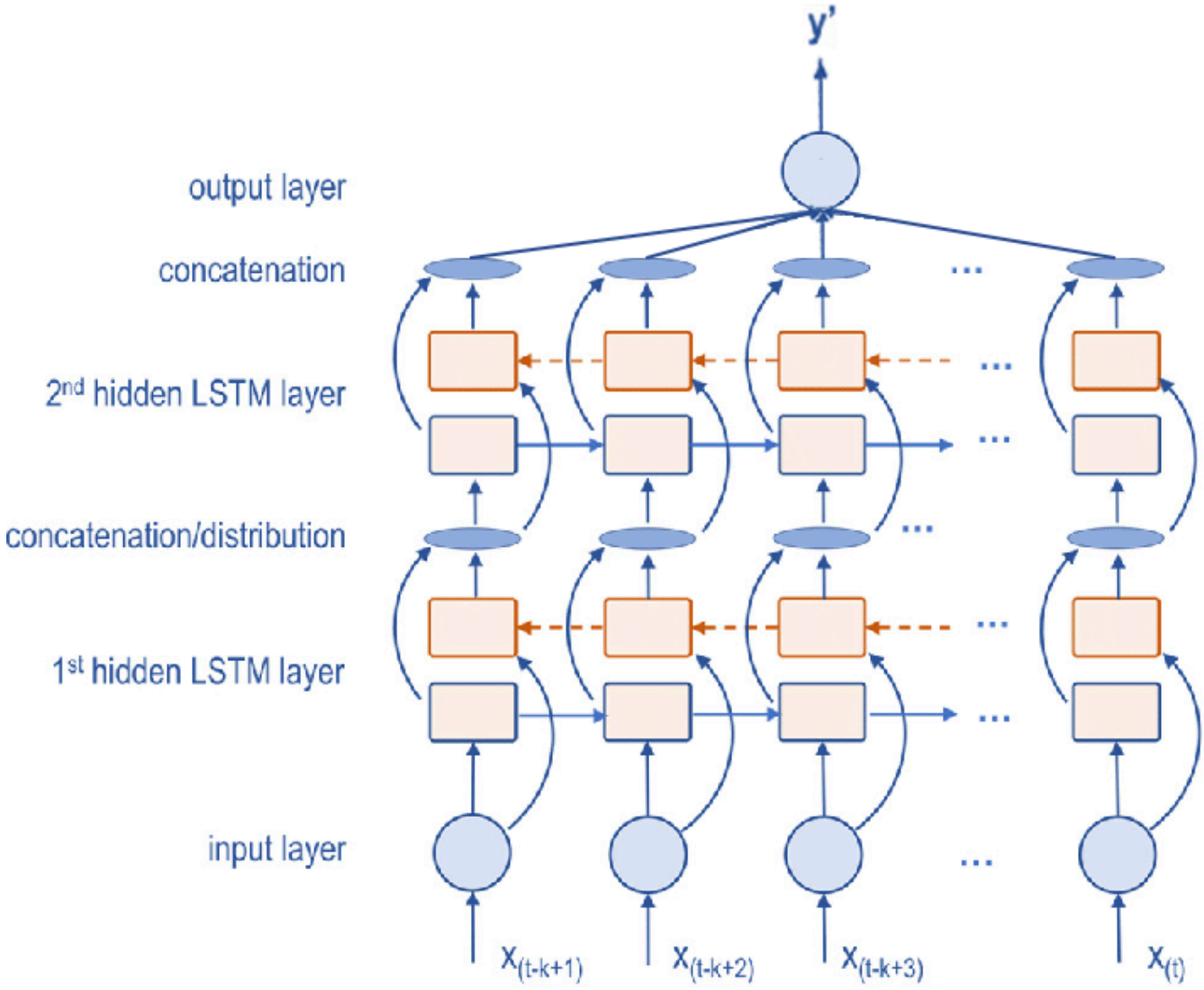

Figure 1: Diagram summarizing the proposed architecture.

References

Du, X., Ma, C., Zhang, G., Li, J., Lai, Y., Zhao, G., Deng, X., Yong-Jin, Liu, & Wang, H. (2020). *An efficient LSTM network for emotion recognition from multichannel EEG signals*. https://www.semanticscholar.org/paper/An-Efficient-LSTM-Network-for-Emotion-Recognition-Du-Ma/c453f25a440e65c70cb128deb1538524c60a8b39

Lilhore, U. K., Dalal, S., Faujdar, N., Margala, M., Chakrabarti, P., Chakrabarti, T., Simaiya, S., Kumar, P., Thangaraju, P., & Velmurugan, H. (2023). Hybrid CNN-LSTM model with efficient hyperparameter tuning for prediction of Parkinson's disease. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-41314-y