Report

Group 22

Group members: .

- Raluca Bolozan (2070485)
- Zala Breznik (2080933)
- Iulia Lupascu (2072964)
- Ina Vlad (2069760)

Data loading and processing

The dataset was unzipped, and each audio file was normalized to a mean of 0 and standard deviation of 1. To standardize duration, the 75th percentile of audio file durations was calculated to be 5.74 seconds, rounded up to 6 seconds. Files longer than 6 seconds were trimmed, and shorter ones were padded. Each file was then converted into a Mel Spectrogram with a sampling rate of 8000 Hz, FFT window length of 2048 samples, hop length of 512 samples, and 128 Mel bands, following the preprocessing steps from Tripathi et al. (2019).

Architecture design

In this study, we tested several architectures based on research articles. Initially, we used the architecture from Yang et al. (2018) with 2 convolutional layers and 2 BLSTM layers. However, the best performance was achieved with a model inspired by Tripathi et al. (2019), featuring four parallel convolutional layers, each with 200 filters of different kernel sizes, ranging from 12x16 to 30x40, followed by 2x2 max pooling layers for capturing diverse audio features. Subsequently, the outputs of the layers are flattened and concatenated into a singular vector. This vector is passed through two fully connected layers with 400 and 200 units, respectively, each being followed by batch normalization and ReLu activation, dropout regularization being performed between the previous processes.

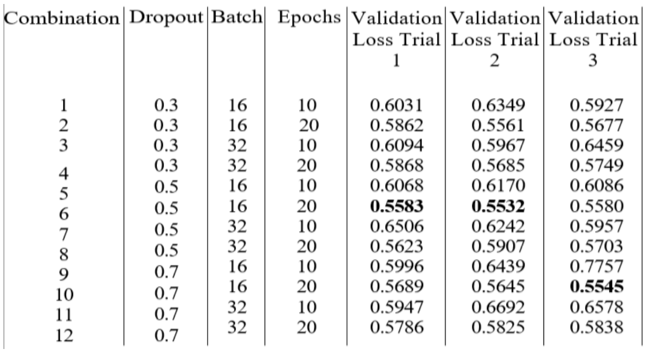
Experiments

In this study we employed an 80-20 train-val data split and experimented with 12 different hyperparameter combinations to optimize our model for predicting valence scores from MEL-spectrograms. The hyperparameters investigated included testing different dropout rates (0.3, 0.5, 0.7), batch sizes (16, 32, 64), and epochs (10, 15, 20). The selected values reflect common practices in the field and considerations for computational efficiency and resource availability. Each combination was tested across 3 random seeds to ensure a thorough evaluation of different model configurations, ultimately identifying the best parameters to improve model accuracy and robustness. Furthermore, Adam optimizer was selected for its adaptive learning rate capabilities, allowing efficient training and faster convergence.

Results

The combinations tested in the hyperparameter tuning show that the combination that yielded the lowest validation loss is the one with a dropout rate of 0.5, batch size of 16, and 20 epochs, with the best validation MAE of 0.5532. These parameters were then used to make predictions on the test set, which yielded a score of 0.516.

An error analysis was conducted by identifying 10 MEL spectrograms with the highest prediction errors. Many spectrograms showed dense low-frequency content (below 1024 Hz) and broad frequency ranges, challenging the model's accuracy. For example, the spectrogram with True: 4.2500, Pred: 2.1462 had energy spread across almost the entire frequency range. The model often over-predicted values, such as True: 1.2500, Pred: 3.2173 and True: 1.5000, Pred: 3.6490, with errors around 2.0, indicating specific audio characteristics might influence these outliers.



Combination	Dropout	Batch	Epochs	Validation Loss Trial 1	Validation Loss Trial 2	Validation Loss Trial 3
1	0.3	16	10	0.6031	0.6349	0.5927
2	0.3	16	20	0.5862	0.5561	0.5677
3	0.3	32	10	0.6094	0.5967	0.6459
4	0.3	32	20	0.5868	0.5685	0.5749
5	0.5	16	10	0.6068	0.6170	0.6086
6	0.5	16	20	**0.5583**	**0.5532**	0.5580
7	0.5	32	10	0.6506	0.6242	0.5957
8	0.5	32	20	0.5623	0.5907	0.5703
9	0.7	16	10	0.5996	0.6439	0.7757
10	0.7	16	20	0.5689	0.5645	**0.5545**
11	0.7	32	10	0.5947	0.6692	0.6578
12	0.7	32	20	0.5786	0.5825	0.5838

Table 1: Table presenting the results

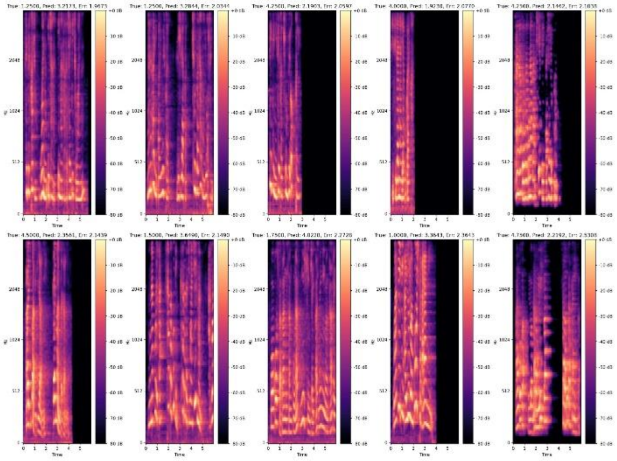


Figure 1: 10 spectrograms with the highest error

Conclusions

• The model architecture inspired by Tripathi et al. (2019), using parallel convolutional layers, max pooling, and fully connected layers proves to be superior when compared to other architectures.

• Improvements could include dynamic padding, advanced models like attention mechanisms, broader hyperparameter searches, more random seed trials, efficient cross-validation techniques, and increasing dataset size through augmentation. Additionally, incorporating techniques to better handle low-frequency content and broad frequency ranges could further enhance prediction accuracy. Addressing over-prediction tendencies by refining the loss function or incorporating more balanced training data may also improve model performance.

Proposed Architecture

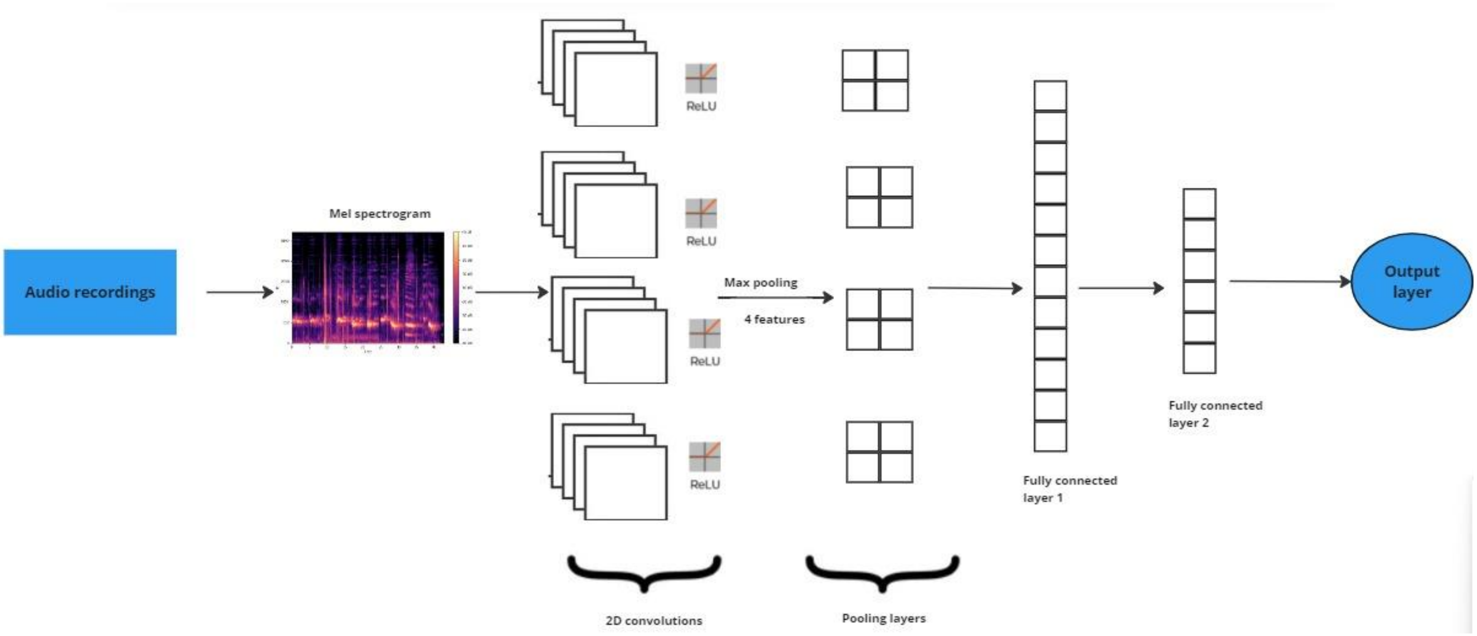


Figure 1: Diagram summarizing the proposed architecture.

Appendix

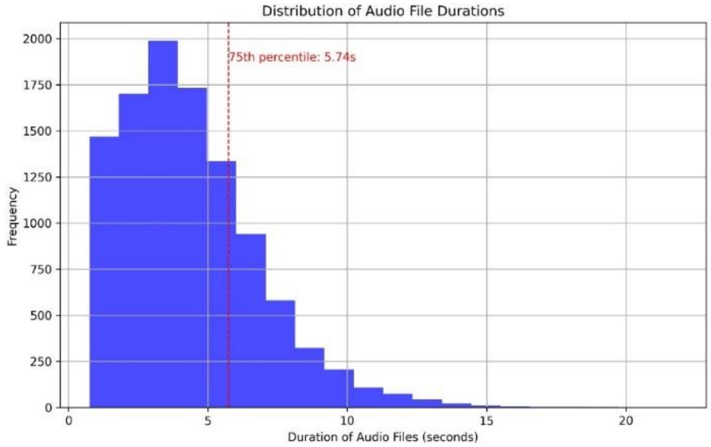


Figure 2: Plot of the audio file durations with the 75th percentile before duration standardization

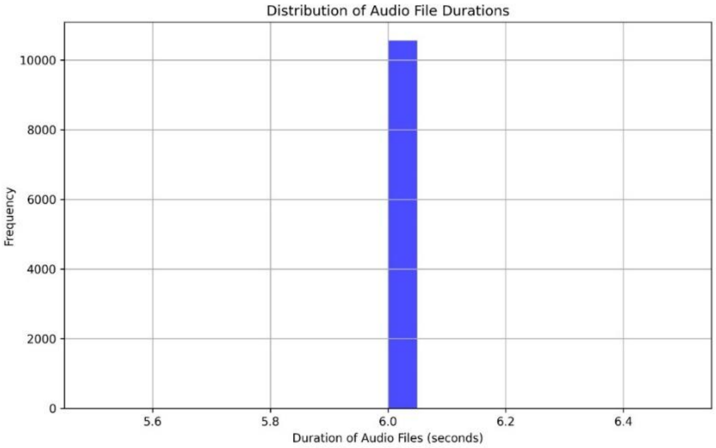


Figure 3: Plot of the audio file durations with the 75th percentile after duration standardization

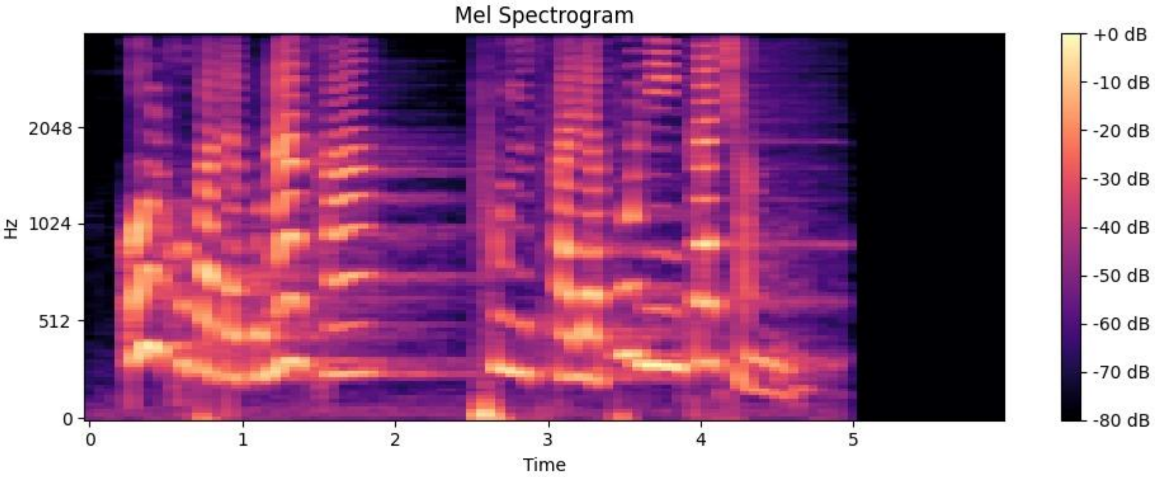


Figure 4: representation of one MEL spectrogram from the dataset

References

1) Tripathi, S., Kumar, A., Ramesh, A., Singh, C., & Yenigalla, P. (2019). Deep learning based emotion recognition system using speech features and transcriptions. arXiv preprint arXiv:1906.05681.

2) Yang, Z., & Hirschberg, J. (2018). Predicting Arousal and Valence from Waveforms and Spectrograms Using Deep Neural Networks. *Interspeech*.