University of British Columbia

From Exploring to Building Machine Learning Models

EECE 568

Enhancing Speech Emotion Recognition Accuracy with Strategic Data Pre-processing and Ensemble CNN-RNN Model

Author: Jason Li
Author: Kevin Chu

Group 2

December 12, 2023

1) Goal of Report

Our research focuses on enhancing the accuracy of speech emotion recognition, a crucial component in human-computer interaction and affective computing. To address the limitations of limited datasets, we aim to implement advanced data preprocessing techniques. Our strategy involves leveraging Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to create a more precise and resilient model for emotion recognition.

2) Previous Work Done

In the landscape of human-computer interaction and affective computing, prior research has laid the groundwork for understanding and recognizing speech emotions, contributing to the transformative potential of technology in responding to human emotions. The precision in speech emotion recognition is recognized as a key factor in reshaping how we interact with technology [1].

A consistent challenge faced by researchers in this domain has been the limitation of datasets. To address this, various studies have focused on innovative data preprocessing strategies. These strategies are crucial for enhancing the quality and diversity of datasets, thereby improving the reliability of emotion recognition models [2-3].

In the context of our research, the meticulous curation and enhancement of two prominent datasets, EmoDB and IEMOCAP [4], exemplify a trend observed in previous works. Emphasis on data quality, authenticity, and consistency has been a common thread in multiple studies, underscoring the importance of a robust dataset foundation [4].

The methodological approach adopted in our research aligns with trends seen in earlier works. The synergistic utilization of the Radial Basis Function Network (RBFN) [5], Convolutional Neural Networks (CNNs) [6-7], and Recurrent Neural Networks (RNNs) [8] reflects a growing interest in combining diverse neural network architectures for improved performance. The ensemble model employed in our study [9-10] is an extension of this trend, aiming to transcend the limitations associated with individual methods and provide a more comprehensive understanding of emotional nuances in speech.
Building on the achievements of previous research, our work not only promises a substantial improvement in recognition accuracy but also highlights the importance of resilience and adaptability across diverse datasets. This aligns with the evolving narrative in the field, emphasizing the need for models that can generalize well to varying data conditions.

In conclusion, our research stands on the shoulders of prior contributions in human-computer interaction and effective computing. By integrating and extending methodologies observed in previous works, we aim to propel the field forward, contributing to the advancement of technology in perceiving and responding to human emotions. The implications of our findings span a spectrum of applications, from virtual assistants to mental health applications, echoing the broad impact that this line of research can have on society [1-10].

3) Strategy

Our strategy for achieving our goal involved a multifaceted approach, primarily focusing on optimizing the model architecture, enhancing data representation through augmentation and feature extraction techniques, and fine-tuning hyperparameters. The overarching goal was to create a robust and versatile model for audio data processing.

We experimented with various configurations of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) layers. Which includes adjusting neurons in each layer, batch sizes, and training epochs.

To augment our training dataset and improve the model's generalization, we employed several techniques, including adding noise, altering pitch, and stretching the audio signals. These augmentations aimed to expose the model to a wider range of variations in the input data, helping it become more resilient to real-world scenarios.

We leveraged a combination of feature extraction methods to represent the audio data comprehensively. Mel-frequency cepstral coefficients (MFCC), chroma features, zero-crossing rate (ZCR), and root mean square (RMS) were among the features extracted. This diverse set of features was chosen to capture distinct aspects of the audio signal, enabling the model to learn intricate patterns and nuances.

In the optimization of hyperparameters, we conducted thorough tuning for dropout rates, output units, and batch size. Specifically, refining the model's dropout rate was a pivotal measure to mitigate overfitting and enhance generalization. Following extensive experimentation with different dropout rates, we identified that a rate of 0.6 struck the optimal balance between regularization and enabling effective learning from the training data.

4) What Methods Worked

a. Pseudo-code Description

Project Overview:

Define the purpose of Speech Emotion Recognition (SER), focusing on recognizing human emotions from speech. Highlight SER's applications, such as in call centers for emotion-based call classification, in-car systems for driver safety, etc.

Step 1: Setup and Data Loading

i. Import necessary Python libraries for data handling, audio processing, and machine learning, such as Pandas, Numpy, Librosa, Seaborn, Matplotlib, Sklearn, Keras.
ii. Mount Google Drive if using cloud platforms like Google Colab for accessing datasets.
iii. Specify file paths for datasets: RAVDESS, TESS, SAVEE, CREMA-D, EmoDB.

Step 2: Data Preprocessing

i. For each dataset, read audio files and extract associated emotions to create

individual data frames with file paths and emotions.

ii. Merge all individual datasets into a single comprehensive data frame for analysis.
iii. Visualize data distribution to understand the balance of different emotions.

Step 3: Data Augmentation

i. Implement 3 audio augmentation techniques such as noise injection, stretching, and pitch changing.
ii. Augment data to improve the model's ability to generalize and perform under various conditions.

Step 4: Feature Extraction

i. Extract 5 audio features such as Zero Crossing Rate, Chroma STFT, MFCC, RMS value, and Mel-spectrogram.
ii. Normalize these features and prepare them for input into the deep learning model.

Step 5: Model Architecture

i. Design a neural network model combining convolutional neural network (CNN) and long short-term memory (LSTM) layers.
ii. Use Conv1D layers for initial feature extraction from audio data, followed by MaxPooling1D layers to reduce dimensionality.
iii. Apply LSTM layers to capture the temporal dynamics in the audio features.
iv. Apply dropout layers to reduce overfitting.
v. Include dense layers for the final classification with a SoftMax activation function.

Step 6: Data Preparation for Model Training

i. Normalize the feature data.
ii. Split the data into training and testing sets.
iii. Implement one-hot encoding for target labels.

Step 7: Model Compilation and Training

i. Compile the model using Adam as the appropriate optimizer, and Categorical Cross-entropy as the loss function.
ii. Use accuracy as the metric for performance evaluation.
iii. Train the model on the training set, using a portion of the data for validation.
iv. Implement callbacks like ReduceLROnPlateau for dynamic adjustment of learning rates during training.

Step 8: Model Evaluation

i. Evaluate the model's performance using the test dataset.
ii. Analyze results through metrics like accuracy and generate a confusion matrix and classification report for detailed evaluation.

Step 9: Demonstration

i. Create a function to preprocess self-recorded audio samples for prediction.

ii. Load the trained model and predict the emotion of self-recorded audio samples.
iii. Display the predicted emotion labels for the demonstration.

Step 10: Conclusion and Model Saving

i. Summarize the performance of the model and its potential applications.
ii. Save the trained model and extracted features for future use.

b. Explanation of Method Used and Results:

The methodology adopted in our speech emotion recognition project was carefully crafted, prioritizing efficient feature extraction and model optimization, leading to a promising accuracy of 62%. This section explains the rationale behind the chosen methods and their connection to the data used.

1. CNN Architecture Implementation:

The Convolutional Neural Network (CNN) architecture was selected for its proficiency in extracting critical features from our audio data. This choice was rooted in the ability of CNNs to utilize convolutional layers, which employ filters to effectively distill necessary features from input training samples. In our project, the CNN architecture was pivotal in isolating five key features essential for training the subsequent Long Short-Term Memory (LSTM) network. This methodological decision was driven by the nature of our data, comprising complex audio samples where feature extraction plays a crucial role in accurate emotion recognition. The CNN's ability to handle such data complexity and its proven track record in similar tasks made it an ideal choice.

Ultimately, this approach yielded an accuracy of 62%, affirming the effectiveness of CNNs in handling the nuanced task of emotion recognition from audio inputs.

2. Dropout Integration and Optimization:

To address the challenge of overfitting, a common hurdle in deep learning models, we integrated dropout layers into our network. The strategic placement and fine-tuning of these layers were key in enhancing the model's ability to generalize. The rationale behind using dropout lies in its functionality: it randomly deactivates a subset of neurons during training, thus preventing the network from becoming overly reliant on any specific set of features. This randomness encourages the network to find robust patterns that generalize better to unseen data.

However, optimizing the dropout rate was crucial. A rate too high might lead to underfitting, losing essential information, whereas a rate too low wouldn't adequately address overfitting. Our tuning process considered these factors, ensuring a balance that contributed significantly to achieving our final accuracy figure.

3. Epoch Optimization for Resource Efficiency:

In machine learning, particularly with complex models, resource management is a key concern. We optimized the number of training epochs to balance computational efficiency with model performance. The choice to focus on epoch optimization was

driven by the nature of our audio data, which could potentially lead to lengthy training times due to its complexity and size. By fine-tuning the number of epochs, we aimed to prevent overtraining and excessive resource consumption. The implementation of an early stopping strategy played a crucial role here. It allowed our model to cease training once it reached a point of diminishing returns in terms of accuracy improvements. This approach ensured that our model reached its desired performance level efficiently, without unnecessary strain on computational resources.

In summary, the methodological choices in our project were heavily influenced by the nature of our data and the specific challenges it presented. The integration of CNN for feature extraction, careful optimization of dropout layers, and strategic management of training epochs collectively contributed to building a resource-efficient model with commendable accuracy in emotion recognition from speech.

5) What Did Not Work

Our journey to enhance the CNN-LSTM model for speech emotion recognition has been insightful, revealing several challenges that have informed our understanding of the complexities involved.

1. In-depth Analysis of Model and Training Parameter Adjustments:

Our extensive experimentation with the model's parameters, including adjusting neurons in each layer, batch sizes, and training epochs, didn't yield significant improvements in accuracy. This outcome has led us to believe that the core issues may lie in the data itself. Redirecting our focus towards dataset refinement and preprocessing methods could be the key. We anticipate that employing sophisticated data augmentation and advanced feature extraction techniques might be the breakthrough needed for enhancing model accuracy and robustness.

We are considering the implementation of 2D convolution layers as a potential method for improvement. This change is expected to provide a more in-depth analysis of spatial-temporal features in audio data, potentially enhancing the model’s ability to discern complex emotional states. Although untested in our current research, transitioning to 2D convolution layers is a promising avenue for future exploration that could bring significant advancements in our model's performance.

2. Detailed Challenges with Custom Audio Recordings:

Our model's performance with self-recorded audio revealed a disparity in accurately identifying certain emotions. We noticed that it was more successful in recognizing emotions like anger and happiness, which we attribute to our implementation of the root mean square (RMS) feature extraction. RMS is closely linked to the energy level in sound waves, which likely explains the model's proficiency with these higher-energy emotions. However, this focus on energy-related features might have contributed to the model's limitations in detecting lower-energy emotions such as sadness, fear, and boredom. To address this, expanding our feature extraction to include elements that capture the nuances of these subtler emotions is crucial. Analyzing the top results from the SoftMax output can also provide a more nuanced

view of the emotional content in speech, recognizing the complex interplay of emotions often present.

In summary, our research has revealed the intricacies of speech emotion recognition, underscoring the importance of a multifaceted approach. This includes not only refining the model's architecture but also deepening our understanding and processing of the data. Exploring new dimensions such as 2D convolution layers, coupled with a more balanced feature extraction strategy, offers promising directions for future work.

6) Conclusions and Future Work

We have successfully utilized a CNN-LSTM model in datasets encompassing up to nine emotions, achieving a noteworthy 62% accuracy in speech emotion recognition. Additionally, our introduction of dropout layers and the optimization of the dropout rate have significantly reduced overfitting. This enhancement has made the model more robust and reliable. While these results are promising, they also highlight areas where further improvements can be made.

1. Dataset Refinement:

We plan to expand and balance our dataset, potentially removing less represented emotions like calm or boredom, to improve the model's training and accuracy.

2. Enhanced Data Augmentation:

Implementing additional data augmentation methods, such as audio time shifting, will enhance the model's ability to handle diverse speech patterns and improve its real-world application.

3. Broadening Feature Extraction:

By extracting a wider range of features, we aim to achieve more nuanced emotion recognition, capturing a broader spectrum of emotional expressions in speech.

4. Implementing 2D Convolution Layers:

Transitioning to 2D convolution layers could provide deeper insights into speech characteristics, potentially increasing the accuracy of emotion detection.

5. Multi-Emotion Analysis:

Considering the top three predictions will allow us to address the complexity of human emotions, acknowledging the coexistence of multiple emotions in speech.

Through these targeted enhancements, we aim to elevate the performance of our model, making it a more effective tool for nuanced and accurate speech emotion recognition.

7) References

[1] ZHAO, Jianfeng; MAO, Xia; CHEN, Lijiang. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomedical signal processing and control, 2019, 47: 312-323.

[2] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in IEEE Access, vol. 7, pp. 117327-117345, 2019, doi: 10.1109/ACCESS.2019.2936124.

[3] M. Kavitha, B. Sasivardhan, P. M. Deepak and M. Kalyani, "Deep Learning based Audio Processing Speech Emotion Detection," 2022 6th International Conference on Electronics, Communication and Aerospace Technology, Coimbatore, India, 2022, pp. 1093-1098, doi: 10.1109/ICECA55336.2022.10009064.

[4] Niu, Y., Zou, D., Niu, Y., He, Z. and Tan, H., A breakthrough in speech emotion recognition using deep retinal convolution neural networks. arXiv preprint, 2017, arXiv:1707.09917.

[5] D. Bertero and P. Fung, “A first look into a convolutional neural network for speech emotion detection,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2017, pp. 5115–5119.

[6] Mustaqeem, M. Sajjad and S. Kwon, "Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM," in IEEE Access, vol. 8, pp. 79861-79875, 2020, doi: 10.1109/ACCESS.2020.2990405.

[7] G. Trigeorgis et al., "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," 2016, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 2016, pp. 5200-5204, doi: 10.1109/ICASSP.2016.7472669.

[8] K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” in Proc. Conf. Empirical Methods Nat. Lang. Process., 2014, pp. 1724–1734.

[9] Latif, S., Rana, R., Khalifa, S., Jurdak, R., & Epps, J., “Direct modelling of speech emotion from raw speech,” arXiv preprint, 2019, arXiv:1904.03833.

[10] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir and B. Schuller, "Survey of Deep Representation Learning for Speech Emotion Recognition," in IEEE Transactions on Affective Computing, vol. 14, no. 2, pp. 1634-1654, 1 April-June 2023, doi: 10.1109/TAFFC.2021.3114365

.

8) Tables, Mathematics and Figures

8.1 Tables and Figures

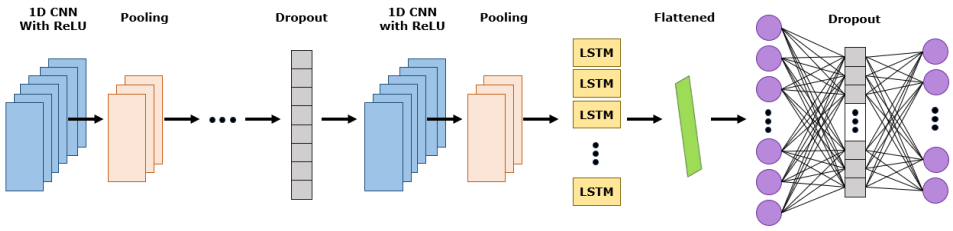


Figure 1: CNN-LSTM Model

Variable	CNN Layers (256 Neurons)	LSTM Layers (128 Neurons)	Accuracy	Plot	Variable	CNN Layers (256 Neurons)	LSTM Layers (128 Neurons)	Accuracy	Plot
CNN Layer	2	1	58.23%		LSTM Layer	4	1	63.65%	
	4	1	63.65%			4	2	62.29%	
	8	1	62.27%			4	4	62.13%	

Figure 2: Changing Number of Layers for CNN and LSTM

Variable	CNN Neurons	LSTM Neurons	Accuracy	Plot	Variable	CNN Neurons	LSTM Neurons	Accuracy	Plot
CNN Neurons	64/ 64/ 64/ 64	128	60.02%		LSTM Neurons	256/ 256/ 256/ 256	64	61.41%	
	128/ 128/ 128/ 128		62.25%				128	63.65%	
	256/ 256/ 256/ 256		63.65%				256	63.58	
	512/ 512/ 512/ 512		63.34%						

Figure 3: Changing Number of Neurons for CNN and LSTM Layers

Variable	CNN Neurons	LSTM Neurons	Dropout Rate	Accuracy	Plot	Dropout Rate	Accuracy	Plot
Dropout Rate	256/ 256/ 256/ 256	128	0	63.65%		0.6	62.15%	
			0.3	61.12%		0.9	36.70%	

Figure 4: Changing Dropout Rate

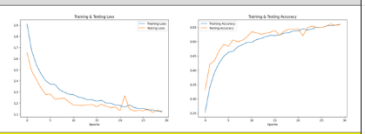
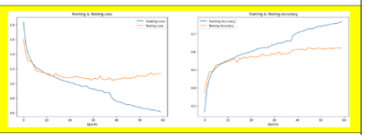
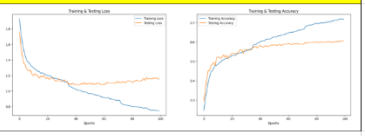
Variable	CNN Neurons	LSTM Neurons	Epochs	Accuracy	Plot
Epochs	256/ 256/ 256/ 256	128	30	55.97%	
			60	62.15%	
			100	60.60%	

Figure 5: Changing Epochs

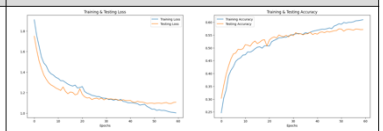
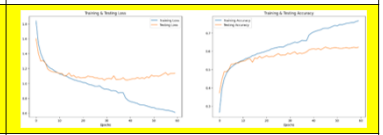
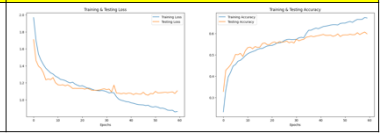
Variable	CNN Neurons	LSTM Neurons	Batch Size	Accuracy	Plot
Batch Size	256/ 256/ 256/ 256	128	32	57.14%	
			64	62.15%	
			128	59.68%	

Figure 6: Changing Batch Size

8.2 Mathematics

1. Convolutional Neural Networks (CNNs):

$$(f \cdot g)(t) = \int_{-\infty}^{\infty} f(\tau) \cdot g(t-\tau) d\tau$$

2. Long Short-Term Memory (LSTM) Network:

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \qquad i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i)$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \qquad \overline{c}_t = \sigma_c(W_c x_t + U_c h_{t-1} + b_c)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \overline{c}_t \qquad h_t = o_t \odot \sigma_h(c_t)$$