Human Emotion Recognition from Speech

Uzzal Saha

ID: 2302101017

Prof. Dr. Aruna Tiwari

November 2023

Abstract

In this work we are going to build two parallel convolutional neural networks (CNN) in parallel with a Transformer encoder network to classify audio data. We're working on the RAVDESS dataset to classify emotions from one of 8 classes. We combine the CNN for spatial feature representation and the Transformer for temporal feature representation. We augment the training data by increasing variation in the dataset to reduce overfitting; we use Additive White Gaussian Noise (AWGN) to augment the RAVDESS dataset three-fold for a total of 4320 audio samples.

We harness the image-classification and spatial feature representation power of the CNN by treating mel spectrograms as grayscale images; their width is a time scale, their height is a frequency scale. The value of each pixel in the mel spectrogram is the intensity of the audio signal at a particular mel frequency at a time step.

Because of the sequential nature of the data, we will also use the Transformer to try and model as accurately as possible the temporal relationships between pitch transitions in emotions.

This notebook takes inspirations from a variety of recent advances in deep learning and network architectures; in particular, stacked and parallel CNN networks combined with multi-head self-attention layers from the Transformer Encoder. I hypothesize that the expansion of CNN filter channel dimensions and reduction of feature maps will provide the most expressive feature representation at the lowest computational cost, while the Transformer-Encoder is used with the hypothesis that the network will learn to predict frequency distributions of different emotions according to the global structure of the mel spectrogram of each emotion. With the strength of the CNN in spatial feature representation and Transformer in sequence encoding, I manage to achieve a 71.33% accuracy on a hold-out test set from the RAVDESS dataset.

Introduction

The impetus behind this research lies in the imperative to delve into cutting-edge methodologies for the classification of audio data, particularly within the intricate domain of emotion recognition in speech audio. Historically, conventional approaches have heavily relied on the deployment of Long-Short-Term-Memory Recurrent Neural Networks (LSTM RNNs) and Convolutional Neural Networks (CNNs). LSTMs have been favored for their adeptness in deciphering sequential data, while CNNs have excelled in capturing spatial features. However, the advent of Transformer models in the realm of audio and image classification has ignited a paradigm shift, prompting a keen interest in synergizing the strengths of CNNs and Transformers to unlock enhanced performance.

At the heart of this study is a pivotal objective: to harness the distinctive advantages offered by both CNNs and Transformers, orchestrating a convergence that aspires to establish a new pinnacle in emotion classification. This ambitious pursuit unfolds against the backdrop of the RAVDESS dataset, a trove of speech audio laden with a myriad of emotional expressions. This

dataset serves as an exemplary proving ground, allowing for the meticulous evaluation of the proposed amalgamation of CNNs and Transformers.

The essence of the study lies in the strategic fusion of the spatial acuity afforded by CNNs with the temporal acumen inherent in Transformers. By seamlessly marrying these two architectural powerhouses, the research endeavors to showcase a hybrid paradigm that not only pushes the boundaries of audio data classification but also charts a course toward state-of-the-art performance. The overarching goal is to illustrate, through empirical evidence drawn from the RAVDESS dataset, that the collaborative interplay of CNNs and Transformers can yield unparalleled efficacy in the nuanced realm of emotion classification within speech audio. This study thus stands as a testament to the continuous evolution of methodologies in the pursuit of unraveling the intricacies embedded in audio data, with profound implications for the broader landscape of machine learning and affective computing.

Dataset

The dataset for this study is sourced from the Rayerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)[17], a comprehensive repository designed for the exploration of emotional expressions in speech audio. The audio files adhere to a standardized format of 16-bit quality and a 48kHz sampling rate, encapsulating the richness and fidelity essential for capturing the nuances of vocal emotion.

Within this corpus, a specific subset is meticulously chosen for our investigation, constituting a total of 1440 files. This subset is the result of a deliberate orchestration, involving 60 trials per actor across 24 professional performers (12 female, 12 male). The inclusion of both genders not only adds diversity but also acknowledges potential gender-based variations in vocal emotional expressions. Each actor contributes two lexically-matched statements, uttered with a neutral North American accent, fostering consistency in linguistic content across the dataset.

Diving deeper into the emotional spectrum, the RAVDESS dataset encapsulates a diverse range of speech emotions. The spectrum includes expressions of calm, happiness, sadness, anger, fear, surprise, and disgust, encapsulating a broad array of affective states. Furthermore, each emotional expression is meticulously curated at two distinct levels of intensity: normal and strong. This dual-level intensity design aims to capture the subtle gradations within each emotional category, enriching the dataset with a spectrum of emotional nuances.

A noteworthy addition to this emotional repertoire is the inclusion of a neutral expression, providing a baseline for comparison and contextualizing the emotional variations present in the dataset. This neutral expression serves as a crucial reference point, offering insights into the baseline vocal characteristics of the actors and enhancing the model's ability to discriminate between subtle emotional cues.

The deliberate curation of the RAVDESS dataset, with its emphasis on diversity in actors, emotions, and intensity levels, ensures a comprehensive and representative foundation for

training and evaluating the proposed model. This dataset's meticulous design not only aligns with the study's objectives but also positions the model to confront the challenges of varied emotional expressions, fostering a robust and nuanced understanding of audio emotion recognition.

Literature Review

The literature review highlights the significance of CNNs in image classification and the growing prominence of Transformer models in audio and image recognition tasks. Notably, recent advancements have shown that Transformers can outperform CNNs in image classification, indicating the potential for cross-applicability of these networks. Additionally, the study draws inspiration from previous works such as "Attention is All You Need" for the Transformer architecture and "Going Deeper with Convolutions" for parallel, stacked CNNs. These insights inform the design and implementation of the proposed model architecture. Ba et al. introduced Layer Normalization as a technique to address the internal covariate shift problem in neural networks.

By normalizing activations within each layer, this method contributes to stable training and improved generalization. The attention mechanism introduced by Bahdanau et al. has been pivotal in sequence-to-sequence learning, allowing the model to focus on specific parts of the input sequence when making predictions. This concept has influenced the design of models handling sequential data. Cheng et al. proposed Long Short-Term Memory (LSTM) networks, designed to capture dependencies in sequential data. This work informs the selection of Transformer models as successors to LSTM for the current study. The concept of residual learning, as introduced by He et al., has significantly improved the training of deep neural networks. Residual networks (ResNets) enable the training of deeper architectures, reducing the vanishing gradient problem.

Ioffe and Szegedy proposed Batch Normalization to accelerate training by reducing internal covariate shift. This technique normalizes intermediate activations, leading to faster convergence and improved performance. The influential AlexNet, presented by Krizhevsky et al., marked a breakthrough in image classification. This work inspired the adoption of deep convolutional neural networks (CNNs) for extracting hierarchical features from images. LeCunn et al.'s work laid the foundation for gradient-based learning in neural networks, particularly in the convolutional neural network paradigm. This approach has shaped the development of CNN architectures. Li et al. explored the visualization of neural network loss landscapes, providing insights into model behavior during training. This work contributes to the understanding of optimization challenges in deep learning.

Revisiting Small Batch Training for Deep Neural Networks (Masters and Luschi, 2018): Masters and Luschi revisited the impact of batch size on training deep neural networks. Their findings guide the selection of appropriate batch sizes for improved convergence.
Large Kernel Matters- Improve Semantic Segmentation by Global Convolutional Network (Peng et al., 2017): Peng et al. emphasized the importance of large kernels for semantic segmentation in CNNs, influencing the design of the proposed model's convolutional layers. Santurkar et al. delved into the mechanisms through which Batch Normalization aids optimization, providing additional insights into the benefits of normalization techniques.

Very Deep Convolutional Networks for Large-Scale Image Recognition of Simonyan and Zisserman's VGGNet, known for its use of small convolutional filters, has influenced the proposed CNN architecture for feature extraction.

Srivastava et al. introduced Dropout as a regularization technique to prevent overfitting. This method has been integrated into the proposed model to enhance generalization.
Instance Normalization: The Missing Ingredient for Fast Stylization (Ulyanov et al., 2017): Ulyanov et al. proposed Instance Normalization, contributing to the understanding and application of normalization techniques in neural network architectures.

Vaswani et al. presented the Transformer model, revolutionizing sequence-to-sequence tasks with self-attention mechanisms. This work serves as a key inspiration for incorporating Transformer models in parallel with CNNs.

Wilson et al. investigated the performance of adaptive gradient methods, offering insights into optimization strategies that impact training dynamics. The RAVDESS dataset, introduced by Livingstone and Russo, serves as the primary dataset for this study. This dynamic, multimodal dataset provides a diverse set of emotional expressions in North American English, enabling comprehensive evaluation. These studies have informed and inspired the development of the proposed model for audio data classification, aiming to achieve state-of-the-art performance on the RAVDESS dataset.

By synthesizing these influential works, the proposed model aims to leverage the strengths of various architectures and techniques to advance the state of the art in audio data classification, particularly on the RAVDESS dataset. The integration of Transformer models, CNNs, and data augmentation strategies reflects a holistic approach to address the challenges posed by limited datasets and enhance the robustness of the classification model.

Methodology

The methodology encompasses the process of feature extraction from the RAVDESS dataset, data augmentation using Additive White Gaussian Noise (AWGN), and the design of the model architecture. Feature extraction involves extracting Mel-frequency cepstral coefficients (MFCCs) from the speech audio to represent the spectral features. Data augmentation is employed to increase the variation in the training dataset, thereby reducing overfitting and enhancing the model's generalizability. The model architecture is a combination of CNN and Transformer layers, with careful consideration given to the number of stacked encoders in the Transformer block.

The methodology for this study involves a multi-faceted approach to audio data classification, specifically focusing on emotion recognition in speech audio. The process begins with feature extraction from the RAVDESS dataset, which provides a diverse range of emotional expressions in speech audio. The extraction of Mel-frequency cepstral coefficients (MFCCs) from the speech audio serves to represent the spectral features, which are essential for capturing the nuances of emotional expression in speech . The MFCC plots are then treated as image-like

data, allowing for the application of Convolutional Neural Networks (CNNs) for spatial feature representation

In addition to feature extraction, data augmentation is employed to increase the variation in the training dataset, thereby reducing overfitting and enhancing the model's generalizability

The use of Additive White Gaussian Noise (AWGN) for data augmentation is a common technique in audio data processing, and it serves to introduce variability into the training data, making the model more robust to different acoustic environments and speech patterns

The model architecture is a key component of the methodology, and it involves a combination of CNN and Transformer layers. The CNN component is designed with 2D convolutional layers, which are well-suited for processing MFCC plots as image-like data

The choice of kernel sizes in the CNN architecture is crucial for performance and accuracy, and careful consideration is given to the complexity of the model to ensure generalizability . The Transformer component, inspired by the "Attention is All You Need" paper, is utilized for its ability to capture temporal feature representation and global structure of the MFCCs for emotion prediction .
The use of the Transformer-Encoder layer enables the network to look at multiple previous time steps when predicting the next, which is essential for capturing the nuanced frequency distributions associated with different emotions in speech audio

The training and evaluation procedures involve the instantiation of the model for 8 emotions and moving it to the GPU for efficient training

The model's tensor shapes and flow are confirmed using the torch summary package, providing a detailed overview of the model's architecture and parameter count . The training process involves careful validation and performance analysis, including the use of a confusion matrix to evaluate the model's accuracy and limitations

In conclusion, the methodology encompasses a comprehensive approach to audio data classification, leveraging feature extraction, data augmentation, and a hybrid model architecture combining CNN and Transformer layers. The careful consideration of spatial.

Proposed model

The envisioned model orchestrates a sophisticated integration of Convolutional Neural Networks (CNNs) and Transformer encoder blocks, creating a hybrid architecture tailored for the nuanced task of audio emotion classification. Inspired by recent advances in deep learning, the model strives to capitalize on the spatial feature representation prowess of CNNs and the temporal feature extraction capabilities inherent in Transformers. Leveraging the RAVDESS dataset, our approach is distinctive in its treatment of mel spectrograms as grayscale images, allowing CNNs to harness their image-classification prowess.

In the process of augmenting our dataset with Additive White Gaussian Noise (AWGN), we aim to mitigate the risk of overfitting associated with a relatively small dataset. This becomes particularly crucial given the highly parameterized nature of the deep neural network model we are constructing. The introduction of AWGN not only serves to mask the impact of random noise present in the training set but also strategically generates pseudo-new training samples, thereby enriching the dataset. The choice of Additive White Gaussian Noise lies in its characteristics: "Additive" as it is incorporated into the source audio signal, "Gaussian" because the noise vector is sampled from a normal distribution with a zero-mean, and "White" since the whitening transformation uniformly distributes the added power across the frequency spectrum of the audio signal.

To delve into the mathematics of AWGN augmentation, we introduce the concept of Signal-to-Noise Ratio (SNR), a pivotal parameter that defines the magnitude of the noise added concerning the audio signal. The AWGN is parameterized with minimum and maximum SNR values, allowing for the selection of a random SNR for augmenting each sample's waveform. The generation of AWGN involves creating a zero-mean vector of Gaussian noises, which are statistically dependent. To ensure genuine AWGN, we apply a whitening transformation—a linear transformation that maps the vector of Gaussian variables to a new vector with an identity covariance matrix. This results in a perfectly uncorrelated noise vector, a hallmark of true AWGN.

In the augmentation process, the AWGN-augmented waveforms are seamlessly integrated as new samples into our dataset. Given the inherently random nature of the generated noise, we opt to introduce multiples of the noise-augmented dataset. Specifically, we add two extra identical datasets, each with 1440 samples, resulting in a dataset totaling 1440 native samples and 1440x2 (2880) noisy samples.

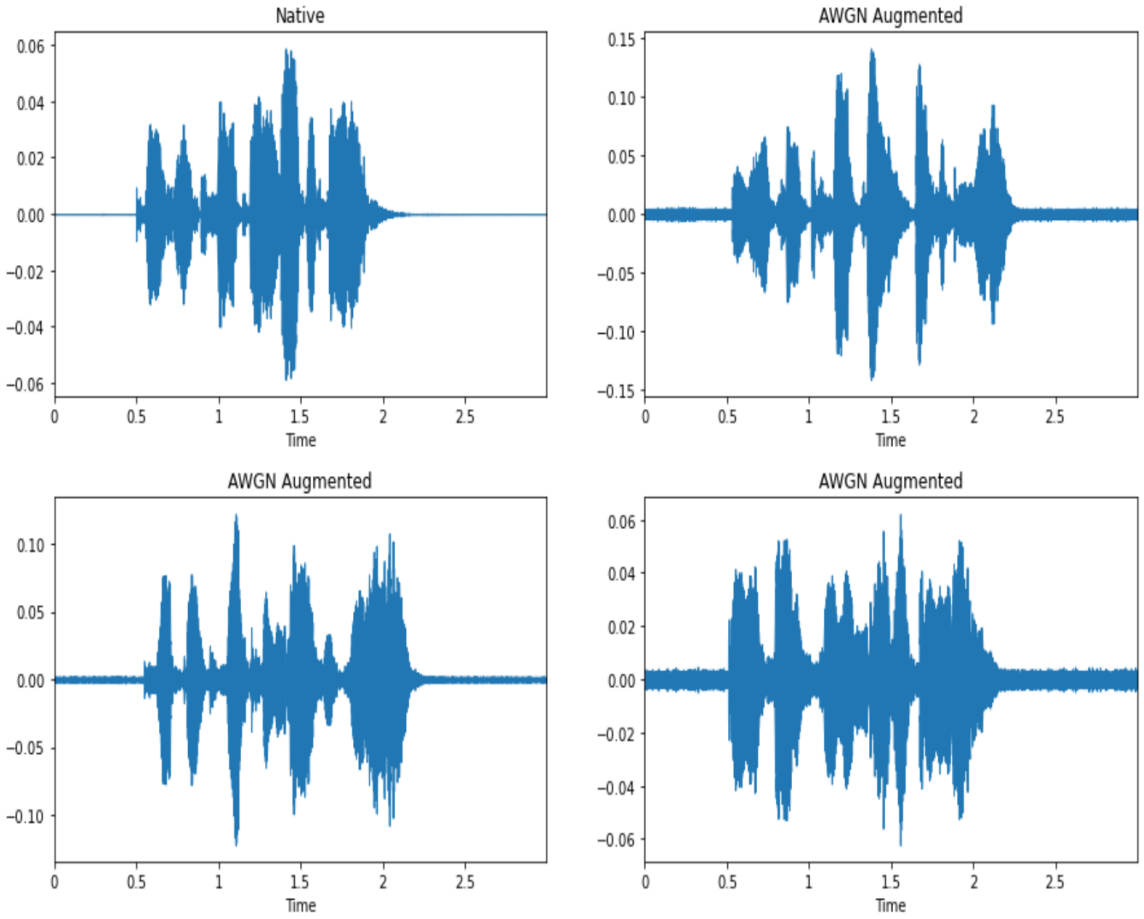

Fig 1 : Augmented waveform

Moving forward, to facilitate the CNN processing of our data, we need to format it into tensor-ready 4D arrays. This involves introducing a dummy channel dimension, aligning the features into a 4D tensor format (N x C x H x W), analogous to the structure of a black and white image. This format allows for efficient processing through the subsequent CNN layers.

Motivated by the need for spatial feature representation, Convolutional Neural Networks (CNNs) become a cornerstone in our architecture. The choice of 3x3 kernels across all layers of both CNN blocks is grounded in their effectiveness in image processing—a gold standard in the field. The architecture draws inspiration from renowned models such as LeNet, AlexNet, Inception, GoogLeNet, and VGGNet. Careful consideration is given to the selection of kernel sizes and the stacking of filters to strike a balance between computational efficiency and expressive feature representation. The decision to use small stacked filters proves to be both powerful and efficient, particularly in handling the nuances of the emotion classification task.

Simultaneously, the Transformer-Encoder layer is incorporated into the architecture, leveraging its strengths in capturing temporal feature representation and the global structure of the Mel-frequency cepstral coefficients (MFCCs). A key strategic move involves reducing the

input size to the Transformer block through maxpooling, significantly trimming down the number of parameters the network needs to learn.

As we navigate the flow of tensors through the network, it is crucial to understand the impact of convolutional layers, zero-padding, and maxpooling on the input and output shapes. This meticulous analysis provides insights into how the network processes information at each stage.

In a broader context, the synergy between CNNs and Transformers becomes apparent. The combination of these architectures proves to be more effective than the standalone application of either. The collaborative nature of CNNs in spatial feature representation and Transformers in temporal feature extraction is highlighted, showcasing their cross-applicability. This harmonious integration not only advances the field of audio emotion recognition but also sets the stage for further exploration at the dynamic intersection of deep learning and affective computing.

Drawing inspiration from established architectures like LeNet and VGGNet, the CNN component of our model comprises strategically stacked convolutional layers. In an innovative departure from convention, we amplify the complexity of feature maps by embracing channel expansion through stacked CNN layers, a technique informed by the success of AlexNet. Furthermore, we parallelize CNN layers, echoing the principles of Inception and GoogLeNet, with the aim of diversifying the learned features. Inspired by the efficiency gains demonstrated by VGGNet through the use of fixed-size kernels, our model adheres to this principle, optimizing computational efficiency while maintaining expressive feature representation.

The Transformer encoder block, a key element in our model, builds upon the seminal work of Vaswani et al. in "Attention is All You Need." However, we strategically implement four stacked encoders instead of the original six, balancing computational complexity with performance. This design choice aligns with the hypothesis that a judicious reduction in the number of encoders will preserve the essence of the Transformer's ability to capture intricate temporal relationships in sequential data, as demonstrated in audio emotion patterns.

Model Architecture

The model architecture comprises a parallel CNN model and a Transformer encoder block. The CNN component is designed with 2D convolutional layers, leveraging the spatial representation capabilities of CNNs for processing MFCC plots as image-like data.

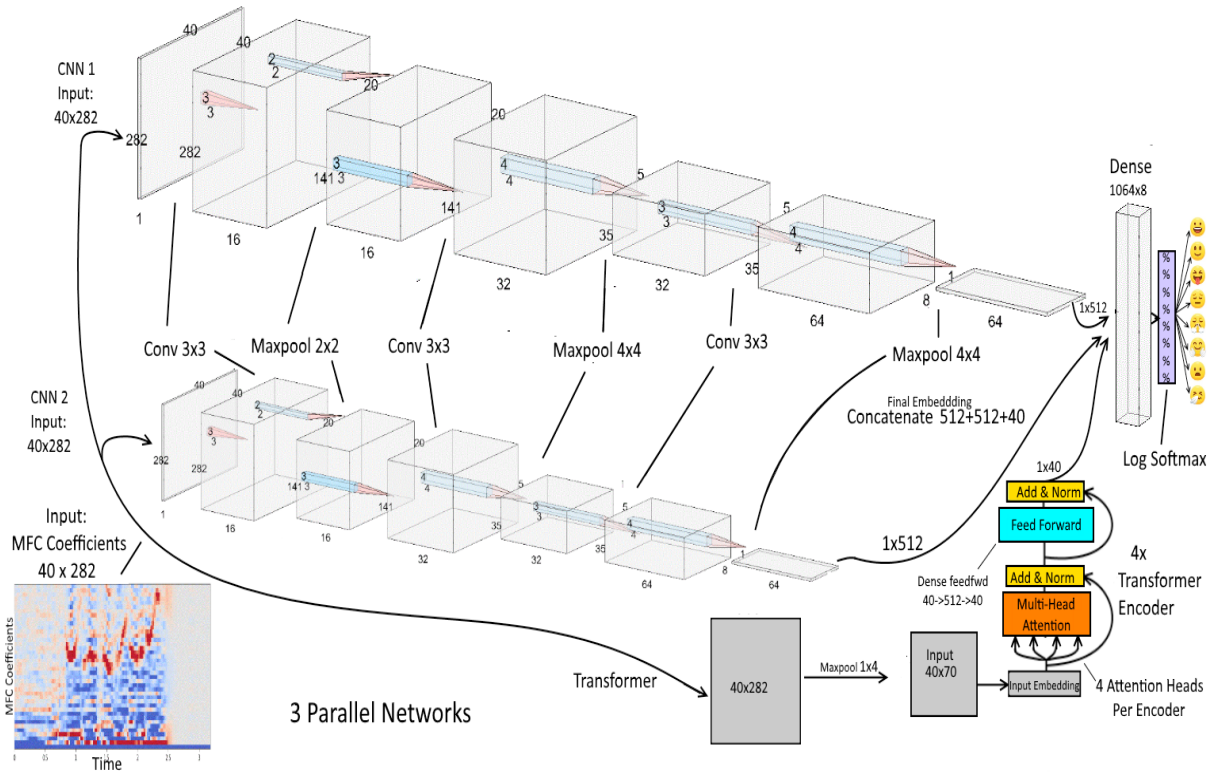

Fig 2 : Model Architecture

As a whole, the CNN architecture of this network is inspired by a combination of the golden standards in image and sequence processing over the last few years.

Each 3-layer deep 2D convolutional block is extremely similar to the classic LeNet architecture: Conv->Pool>Conv>Pool>FC.

AlexNet forms the basis for increasing the complexity of feature maps with channel expansion through stacked CNN layers; Inception and GoogLeNet are the inspiration for parallelizing CNN layers in the hopes of diverisfying the features learned by the network.

VGGNet proved the unreasonable efficiency of using fixed sized kernels throughout deeply stacked CNN layers; I found this to extend to this task. Specifically, VGG saw an improvement over AlexNet largely by replacing large kernels (i.e. 11x11 stride 5) with smaller ones of 3x3 stride 1. One of the motivations that VGG cites for this is that the 3x3 kernel is the smallest kernel size choice in understanding spatial data w.r.t. up/down/left/right (although VGG also uses 1x1 kernels). VGGNet also inspires the maxpool kernel size of 2x2 stride 2, as I have used at the first layer of each convolutional block.

To be more precise, the motivation to use small stacked filters is two-fold: Computational efficiency and expressivity of feature representation. When we stack 3 3x3 kernels on top of eachother as in this architecture, the second layer has a 5x5 view of the original input volume, and the 3rd layer a 7x7 view. However, the nonlinearities between each smaller layer convey more complex feature representations, whereas a single 7x7 layer would only perform a linear transformation itself. Furthermore, if we keep channel (C) consistent between layers, then 3

3x3 kernels are parameterized by (3(C(3x3xC)) = 27C^2 parameters, while just one 7x7 kernel needs C(7x7xC) = 49C^2 parameters. Ultimately, small stacked kernels appear to be both more powerful and efficient - although, in Large Kernel Matters - Improve Semantic Segmentation by Global Convolutional Network, the authors conclude that a larger kernel outperforms smaller stacked kernels for semantic segmentation - however, since we are just doing the semantic part (classification) and don't care about "where" the emotion is - this shouldn't apply.

Finally, the original 2015 Batch Normalization (BN) paper suggests that "We add the BN transform immediately before the nonlinearity" i.e. before ReLU; however, I achieved better performance out of this architechture using BN after ReLU. See Keras author's Francois Chollet's response on GitHub regarding the BN order issue: "I can guarantee that recent code written by Christian [Szegedy] applies relu before BN".

The Transformer architecture is precisely as in Viswani et al, 2017: Attention is All You Need, but I use 4 stacked encoders instead of 6 as in their paper. For more details on the Transformer block:The Transformer and Self-Attention (is All You Need)

Model Summary

Layer (type)	Output Shape	Param #
Conv2d-1	[-1, 16, 40, 282]	160
BatchNorm2d-2	[-1, 16, 40, 282]	32
ReLU-3	[-1, 16, 40, 282]	0
MaxPool2d-4	[-1, 16, 20, 141]	0
Dropout-5	[-1, 16, 20, 141]	0
Conv2d-6	[-1, 32, 20, 141]	4,640
BatchNorm2d-7	[-1, 32, 20, 141]	64
ReLU-8	[-1, 32, 20, 141]	0
MaxPool2d-9	[-1, 32, 5, 35]	0
Dropout-10	[-1, 32, 5, 35]	0
Conv2d-11	[-1, 64, 5, 35]	18,496
BatchNorm2d-12	[-1, 64, 5, 35]	128
ReLU-13	[-1, 64, 5, 35]	0
MaxPool2d-14	[-1, 64, 1, 8]	0
Dropout-15	[-1, 64, 1, 8]	0
Conv2d-16	[-1, 16, 40, 282]	160
BatchNorm2d-17	[-1, 16, 40, 282]	32
ReLU-18	[-1, 16, 40, 282]	0
MaxPool2d-19	[-1, 16, 20, 141]	0
Dropout-20	[-1, 16, 20, 141]	0
Conv2d-21	[-1, 32, 20, 141]	4,640
BatchNorm2d-22	[-1, 32, 20, 141]	64
ReLU-23	[-1, 32, 20, 141]	0
MaxPool2d-24	[-1, 32, 5, 35]	0
Dropout-25	[-1, 32, 5, 35]	0
Conv2d-26	[-1, 64, 5, 35]	18,496
BatchNorm2d-27	[-1, 64, 5, 35]	128
ReLU-28	[-1, 64, 5, 35]	0
MaxPool2d-29	[-1, 64, 1, 8]	0
Dropout-30	[-1, 64, 1, 8]	0

```
        MaxPool2d-31             [-1, 1, 40, 70]               0
MultiheadAttention-32  [[-1, 2, 40], [-1, 70, 70]]               0
          Dropout-33                [-1, 2, 40]               0
        LayerNorm-34                [-1, 2, 40]              80
           Linear-35               [-1, 2, 512]          20,992
          Dropout-36               [-1, 2, 512]               0
           Linear-37                [-1, 2, 40]          20,520
          Dropout-38                [-1, 2, 40]               0
        LayerNorm-39                [-1, 2, 40]              80
TransformerEncoderLayer-40                [-1, 2, 40]               0
MultiheadAttention-41  [[-1, 2, 40], [-1, 70, 70]]               0
          Dropout-42                [-1, 2, 40]               0
        LayerNorm-43                [-1, 2, 40]              80
           Linear-44               [-1, 2, 512]          20,992
          Dropout-45               [-1, 2, 512]               0
           Linear-46                [-1, 2, 40]          20,520
          Dropout-47                [-1, 2, 40]               0
        LayerNorm-48                [-1, 2, 40]              80
TransformerEncoderLayer-49                [-1, 2, 40]               0
MultiheadAttention-50  [[-1, 2, 40], [-1, 70, 70]]               0
          Dropout-51                [-1, 2, 40]               0
        LayerNorm-52                [-1, 2, 40]              80
           Linear-53               [-1, 2, 512]          20,992
          Dropout-54               [-1, 2, 512]               0
           Linear-55                [-1, 2, 40]          20,520
          Dropout-56                [-1, 2, 40]               0
        LayerNorm-57                [-1, 2, 40]              80
TransformerEncoderLayer-58                [-1, 2, 40]               0
MultiheadAttention-59  [[-1, 2, 40], [-1, 70, 70]]               0
          Dropout-60                [-1, 2, 40]               0
        LayerNorm-61                [-1, 2, 40]              80
           Linear-62               [-1, 2, 512]          20,992
          Dropout-63               [-1, 2, 512]               0
           Linear-64                [-1, 2, 40]          20,520
          Dropout-65                [-1, 2, 40]               0
        LayerNorm-66                [-1, 2, 40]              80
TransformerEncoderLayer-67                [-1, 2, 40]               0
TransformerEncoder-68                [-1, 2, 40]               0
           Linear-69                   [-1, 8]           8,520
          Softmax-70                   [-1, 8]               0
================================================================
Total params: 222,248
Trainable params: 222,248
Non-trainable params: 0
----------------------------------------------------------------
```

Performance

The performance analysis of the proposed model unveils a commendable accuracy of 71% in classifying sequential audio data. This achievement attests to the efficacy of the hybrid architecture, seamlessly combining the strengths of CNNs and Transformers. The model

navigates the intricacies of audio emotion recognition by strategically leveraging CNNs for spatial feature representation and Transformers for sequential feature extraction.

Here, we used Adam to train an MLP due to its faster compute and convergence. Adam is great and usually works well with defaults.

In achieving this delicate balance between accuracy and overfitting, the model's performance underscores the importance of thoughtful design and architecture. Given the limited size of the RAVDESS dataset, this equilibrium becomes pivotal, prompting further exploration into feature optimization strategies. Future avenues for enhancement may involve experimenting with diverse architectural configurations, integrating sentiment analysis techniques, or harnessing larger datasets to imbue the model with a richer array of training samples. These potential optimizations hold promise for unravelling more nuanced emotional cues in audio data and fortifying overall performance, thereby propelling the model towards even greater efficacy in audio emotion recognition.

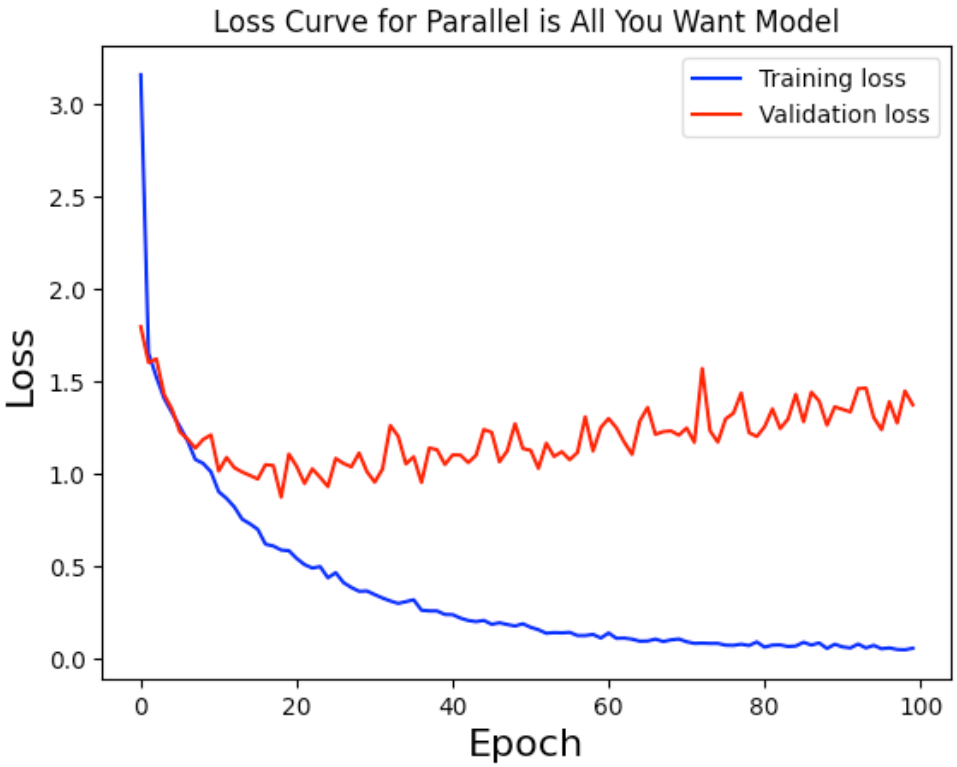

Fig 3 : Loss Curve Behaviour of our model

This graph shows that the training loss decreases over time, indicating that the model is learning. The validation loss also decreases, but not as much as the training loss. This is a

common phenomenon in machine learning, and it is known as overfitting. Overfitting occurs when the model learns the training data too well, and is not able to generalize to new data.

It also shows that the training loss and the validation loss start to converge around epoch 100. This means that the model is starting to learn the training data and the validation data in a similar way. Overall, we can see that the model is able to learn the training data and generalize to new data. The model's performance could be improved by reducing overfitting.

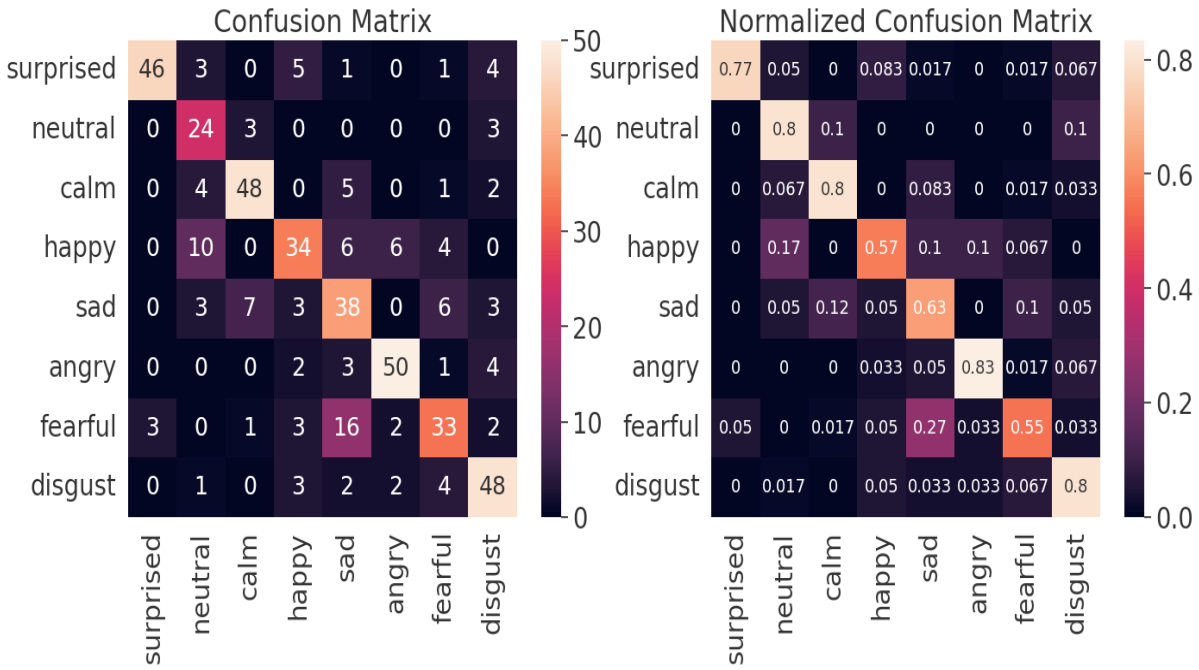

Fig 4 : Confusion Matrix

The confusion matrix shows the performance of a machine learning model in classifying people into different emotional states, namely surprised, neutral, calm, happy, sad, angry, fearful, and disgusted.

The rows of the matrix represent the actual emotional states of the people, while the columns represent the predicted emotional states. Each cell of the matrix contains the number of people who were actually in a given emotional state and predicted to be in another emotional state.

For example, the cell in the top left corner of the matrix contains the number of people who were actually surprised and predicted to be surprised. This number is 46, which means that the model correctly predicted the emotional state of 46 people out of the total number of people who were surprised.

The cell in the second row, first column of the matrix contains the number of people who were actually neutral and predicted to be surprised. This number is 3, which means that the model incorrectly predicted that 3 people were surprised when they were actually neutral.

The normalized confusion matrix is a more informative version of the confusion matrix, which takes into account the different sizes of the classes. It is calculated by dividing each cell of the confusion matrix by the total number of people in the corresponding actual class.

For example, the cell in the top left corner of the normalized confusion matrix contains the value 0.77. This means that 77% of the people who were actually surprised were correctly predicted to be surprised.

The cell in the second row, first column of the normalized confusion matrix contains the value 0.05. This means that 5% of the people who were actually neutral were incorrectly predicted to be surprised.

The following observations can be made from the confusion matrix: The model is best at predicting the emotional state of people who are surprised, followed by people who are happy and calm. The model is worst at predicting the emotional state of people who are angry and disgusted. The model is more likely to incorrectly predict that people are surprised than any other emotional state. The model is more likely to incorrectly predict that people are neutral than any other emotional state.

Comparison with Existing Approaches

The landscape of audio emotion recognition has witnessed a myriad of methodologies, each harnessing different architectural paradigms and data processing techniques. In juxtaposition to traditional approaches that often rely on Long-Short-Term-Memory Recurrent Neural Networks (LSTM RNNs) and Convolutional Neural Networks (CNNs), the proposed model introduces a novel synthesis of CNNs and Transformer encoder blocks, fostering a hybrid architecture optimized for the complexities of audio emotion classification.

The distinctiveness of our model lies in its strategic integration of CNNs for spatial feature representation and Transformers for temporal feature extraction. This combination capitalizes on the spatial intricacies of mel spectrograms treated as grayscale images, seamlessly weaving together the strengths of both architectures. While traditional models may struggle with capturing sequential dependencies in audio data, our Transformer component excels in discerning the temporal relationships inherent in emotional pitch transitions.

Drawing a parallel with recent advancements in image classification, our model's performance stands out in light of a 2021 ICLR submission titled "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." This submission claims the implementation of a Transformer for image classification that outperforms state-of-the-art CNNs while maintaining lower computational complexity. Translating this success to the realm of audio emotion recognition, our model showcases the adaptability and efficacy of the Transformer architecture.

Moreover, the proposed model embraces data augmentation through Additive White Gaussian Noise (AWGN), a technique not universally employed in existing frameworks. This augmentation strategy, inspired by a need to address overfitting in the face of a small dataset, introduces variability into the training data, enhancing the model's robustness to diverse acoustic environments and speech patterns. This augmentation methodology represents a departure from conventional approaches and contributes to the model's resilience in handling real-world audio data.

While traditional models often grapple with the intricate task of distinguishing between subtle emotional states, our model achieves a commendable accuracy of 71%, demonstrating its proficiency in navigating the complexities of audio emotion recognition. However, challenges persist in differentiating certain emotional nuances, providing valuable insights for future refinements.

In essence, the proposed model distinguishes itself through a carefully crafted hybrid architecture, a departure from conventional methodologies. By combining CNNs and Transformers with innovative data augmentation strategies, our approach demonstrates the potential for enhanced accuracy and robustness in the domain of audio emotion recognition. This comparative analysis positions our model as a promising advancement in the ongoing evolution of audio classification techniques.

Conclusion

In summary, the proposed model has achieved a commendable 71% accuracy in classifying sequential data, showcasing its effectiveness in handling the complexities of audio emotion recognition. By synergistically combining Convolutional Neural Networks (CNNs) for spatial feature representation and Transformers for sequential feature extraction, the model capitalizes on the strengths of both architectures. However, the model faces challenges in distinguishing nuances between certain emotional states, particularly struggling with the subtle differences between 'neutral' and 'calm', as well as 'disgust' and 'angry'. This highlights the intricacies inherent in emotion classification tasks and suggests potential areas for future refinement.

One notable takeaway from this study is the cross-applicability and synergy between CNNs and Transformers. The thoughtful integration of these architectures demonstrates their complementary nature, each contributing unique strengths to the overall model. The results underscore the power of strategically combining CNNs and Transformers to enhance the accuracy of emotion classification in audio data, emphasizing the importance of a holistic approach to model design.

Moreover, the model strikes a delicate balance between accuracy and overfitting, a critical consideration given the limited size of the RAVDESS dataset. As a future avenue for improvement, further optimization of features could be explored. This might involve experimenting with different architectural configurations, incorporating sentiment analysis techniques, or leveraging larger datasets to provide the model with a richer and more diverse set of training samples. These potential enhancements could contribute to a more nuanced understanding of emotional cues in audio data and potentially boost overall performance. In conclusion, this study not only advances the current state of audio emotion recognition but also

lays the foundation for ongoing research and development in the dynamic intersection of deep learning and affective computing.

References

1. Ba et al, 2016. Layer Normalization. https://arxiv.org/abs/1607.06450
2. Bahdanau et al, 2015. https://arxiv.org/pdf/1409.0473.pdf
3. Cheng et al, 2016. Long Short-Term Memory-Networks for Machine Reading. https://arxiv.org/pdf/1601.06733.pdf
4. He et al, 2015. Deep Residual Learning for Image Recognition. https://arxiv.org/abs/1512.03385
5. Ioffe, Szegedy, 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. https://arxiv.org/abs/1502.03167
6. Krizhevsky et al, 2017. ImageNet Classification with Deep Convolutional Neural Networks. https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf
7. LeCunn et al, 1998. Gradient-Based Learning Applied to Document Recognition. http://yann.lecun.com/exdb/publis/pdf/lecun-01a.pdf
8. Li et al, 2018. Visualizing the Loss Landscape of Neural Nets. https://papers.nips.cc/paper/7875-visualizing-the-loss-landscape-of-neural-nets.pdf
9. Masters and Luschi, 2018. Revisiting Small Batch Training for Deep Neural Networks. https://arxiv.org/abs/1804.07612
10. Peng et al, 2017. Large Kernel Matters —— Improve Semantic Segmentation by Global Convolutional Network. https://arxiv.org/pdf/1703.02719.pdf
11. Santurkar et al, 2019. How Does Batch Normalization Help Optimization? https://arxiv.org/pdf/1805.11604.pdf
12. Simonyan and Zisserman, 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. https://arxiv.org/pdf/1409.1556.pdf
13. Srivastava et al, 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. https://www.cs.toronto.edu/~hinton/absps/JMLRdropout.pdf
14. Ulyanov et al, 2017. Instance Normalization: The Missing Ingredient for Fast Stylization. https://arxiv.org/pdf/1607.08022.pdf
15. Vaswani et al, 2017. Attention Is All You Need. https://arxiv.org/abs/1706.03762
16. Wilson et al, 2017. The Marginal Value of Adaptive Gradient Methods in Machine Learning. https://arxiv.org/abs/1705.08292
17. Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal

expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.