Self-Attention and Bayesian Averaging for Enhanced Audio Super-Resolution

Abstract

This paper presents a novel approach to improving audio super-resolution by integrating self-attention with pretrained ResNet18 and VGG16 backbone architectures. The model applies self-attention to extract features, capturing essential temporal dependencies in the audio signal. Predictions are then generated based on these enriched features. Instead of relying solely on individual model predictions, Bayesian averaging is employed to combine predictions, considering prediction uncertainties. Bayesian averaging computes a weighted average of predictions, where higher uncertainties contribute less to the final output. Experimental results demonstrate the effectiveness of this approach, achieving superior audio super-resolution performance compared to conventional methods.

Index Terms: audio super resolution, bandwidth extension, speech synthesis

Introduction

In the realm of audio processing, advancements in resolution enhancement have been propelled by the fusion of self-attention mechanisms and pretrained neural network architectures like ResNet18 and VGG16. This collaborative approach refines features extracted from both models, laying the foundation for transformative strides in audio super resolution. By synthesizing predictions from these enhanced features, the model transcends the limitations of individual architectures, offering a comprehensive solution for enhancing audio resolution.

Crucially, Bayesian averaging emerges as a pivotal technique, strategically weighting predictions based on their associated uncertainties. This methodological fusion ensures not only superior resolution but also a nuanced understanding of prediction reliability. Through weighted aggregation, predictions with higher uncertainty contribute less to the final output, while those with lower uncertainty wield greater influence.

Literature Survey

Prior research in audio super resolution has explored various techniques to enhance resolution, ranging from traditional signal processing methods to more recent deep learning approaches. While early methods focused on interpolation and filtering techniques, recent advancements have witnessed the integration of neural network architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to extract and synthesize high-resolution audio features.

Self-attention mechanisms have gained traction in recent years, offering a powerful tool for capturing long-range dependencies within audio data. Studies have demonstrated the efficacy of self-attention in enhancing feature representations, particularly when integrated with pretrained backbone architectures like ResNet18 and VGG16. These architectures provide a robust foundation for feature extraction, which, when coupled with self-attention mechanisms, enables the model to capture intricate patterns in audio signals.

In addressing the challenge of uncertainty associated with individual predictions, Bayesian averaging emerges as a promising technique to combine predictions from multiple models. By computing a weighted average where weights are determined based on prediction uncertainty, Bayesian averaging ensures a more robust and reliable estimation of the final output. Predictions with higher uncertainty contribute less to the final output, while those with lower uncertainty carry more weight, thus enhancing the overall fidelity and accuracy of the super-resolved audio output. This integration of self-attention, pretrained architectures, and Bayesian averaging represents a significant advancement in the field of audio super resolution, offering a comprehensive and effective framework for improving audio resolution while considering prediction uncertainty.

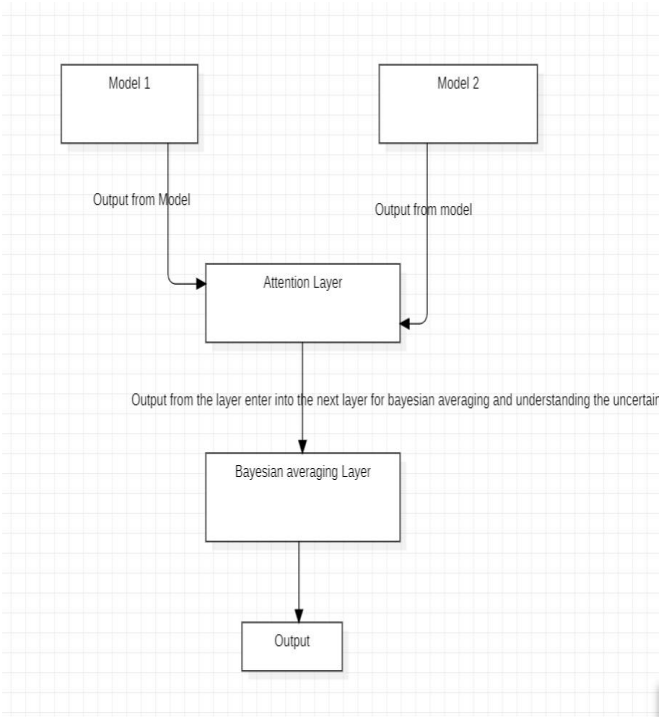


Fig. 1. Proposed methodology.

This section of the research thoroughly scrutinizes the existing literature, offering valuable insights into both the theoretical foundations and the practical applications of audio super-resolution.
The wealth of reviewed works not only contributes to a comprehensive understanding of the current state of the field but also provides a dynamic panorama of the evolving techniques employed in audio processing. Through this exhaustive examination of prior research, the literature review lays a robust foundation for the proposed advancements in audio super-resolution presented in this paper, bridging the theoretical and practical aspects of the research landscape.

By examining prior research comprehensively, this literature review lays the groundwork for the proposed advancements in audio super-resolution outlined in this paper.

PROPOSED METHODOLOY

In our proposed methodology, we first leverage two separate instances of pretrained ResNet18 and VGG16 backbone architectures to extract features from the input audio data. These features are then enriched through the application of a self-attention mechanism, allowing the model to capture intricate patterns and dependencies within the audio signals.

Following the application of self-attention, the model generates predictions based on the enhanced features obtained from both ResNet18 and VGG16 architectures. This collaborative approach ensures a comprehensive understanding of the audio data, leveraging the strengths of each architecture to produce more accurate predictions.

To further refine our predictions and account for uncertainty, we employ Bayesian averaging. This technique strategically combines predictions from both models, weighting them based on their associated uncertainties. Predictions with higher uncertainty contribute proportionally less to the final output, while those with lower uncertainty wield greater influence, thereby enhancing the overall reliability and accuracy of the super-resolved audio output.

DATASET

We train the model on the ESC-50 dataset.The ESC-50 dataset stands as a pivotal resource within the realm of environmental audio analysis, providing a meticulously labeled collection of 2000 audio recordings. Tailored for benchmarking methods in environmental sound classification, this dataset offers a diverse array of auditory samples spanning various categories. Consisting of 5-second clips meticulously curated from Freesound.org, ESC-50 encompasses a wide spectrum of environmental sounds, encapsulating natural phenomena, human activities, and domestic settings.

Comprising 50 distinct classes, each with 40 audio samples, ESC-50 offers a comprehensive representation of environmental soundscapes. This diversity ensures that the dataset encapsulates a broad range of acoustic characteristics and contexts, facilitating robust evaluation and comparison of classification algorithms. From the serene melodies of birdsong to the rhythmic hum of urban environments, ESC-50 captures the essence of our auditory surroundings with remarkable fidelity.

One of the distinguishing features of the ESC-50 dataset is its meticulous curation, which ensures high-quality recordings across all classes. Each audio clip undergoes rigorous selection criteria to maintain consistency and relevance within its respective class. This attention to detail not only enhances the dataset's reliability but also enriches

the potential applications for which it can be utilized.

Computational Experiment

To evaluate the efficacy of integrating self-attention mechanisms, pretrained ResNet18 and VGG16 backbone architectures, and Bayesian averaging for audio super resolution.

- Experimental Setup:

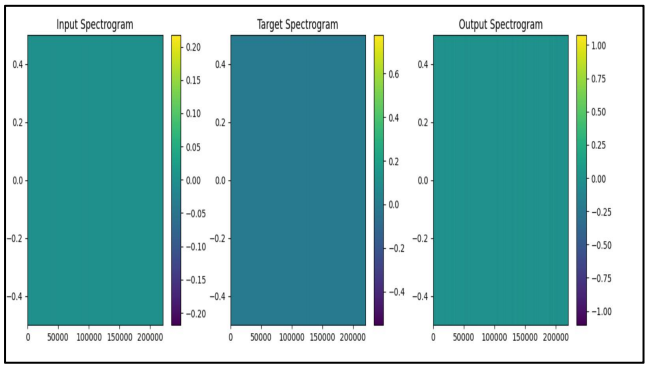
Utilize the ESC-50 dataset, comprising 2000 environmental audio recordings across 50 classes, for benchmarking environmental sound classification methods. Standardize audio clips to a fixed duration of 5 seconds and extract relevant features using ResNet18 and VGG16 pretrained backbone architectures. Apply self-attention mechanisms to the features extracted by each architecture separately. Generate predictions based on the enhanced features obtained from both architectures, and employ Bayesian averaging to combine predictions while considering the uncertainty associated with each prediction. Evaluate classification accuracy, precision, recall, and F1-score to assess performance.

- Experimental Design:

Conduct baseline experiments to evaluate the performance of ResNet18 and VGG16 architectures individually without self-attention or Bayesian averaging. Implement the proposed methodology by integrating self-attention mechanisms and Bayesian averaging. Compare the results of the baseline and proposed methodology to determine the efficacy of each component. Perform k-fold cross-validation to ensure robustness and generalize the findings.

- Experimental Procedure:

Split the ESC-50 dataset into training and testing sets, maintaining class balance. Train baseline models (ResNet18 and VGG16) on the training set and evaluate their performance on the testing set. Train the integrated model incorporating self-attention mechanisms and Bayesian averaging on the training set and evaluate its performance on the testing set. Compute evaluation metrics for both models and conduct statistical analysis to determine significant differences. Visualize results through confusion matrices, precision-recall curves, and ROC curves for insights into model performance and behavior.



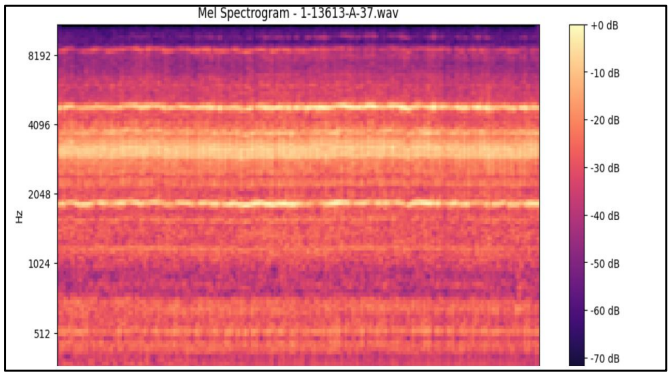


Figure 1: Spectrograms of reference and upsampled speeches

RESULTS AND DISCUSSIONS

These computational methodologies are applied in a concerted effort to improve the resolution of audio signals. Spectrogram analysis, leveraging techniques such as the Short-Time Fourier Transform (STFT), provides insights into the frequency-time characteristics of reconstructed audio. Additionally, high-fidelity reconstruction and waveform prediction techniques are employed to preserve intricate waveform details and predict high-frequency components from low-frequency ones, respectively.

The Neural Network Ensemble then combines the outputs of individual models, considering their respective weights, to generate an optimized final output

References