**Literature review Vocial team**

Micheal Rentz  
Computer science  
University of ConnecticutStorrs, CT  
michael.rentz@uconn.edu

Alexander McLeod  
Compiter science  
universitry of conbnecticutStorrs, CT  
alexander.mcleod@uconn.edu Akhmed Zhanaidarov  
Computer science  
University of ConnecticutStorrs, CT  
 akhmed.zhanaidarov@uconn.edu

Ankith Nagabandi  
Computer science  
University of ConnecticutStorrs, CT  
ankith.nagabandi@uconn.edu

Ronak Sahu  
Computer science  
University of ConnecticutStorrs, CT  
 ronak.sahu@uconn.edu

ABSTRACT

***This literature review examines the recent developments and advancements in audio classification using transformer-based models. Audio classification has gained attention in recent years due to its wide range of applications, such as speech recognition, environmental sound classification, and music genre classification. Transformer-based models have shown promising results in natural language processing tasks and have been adapted for audio classification as well. This review analyzes the effectiveness of different transformer-based models in various audio classification tasks, such as Speech Commands v0.02, UrbanSound8k, and ESC-50 datasets. The review also discusses the impact of various hyperparameters, attention mechanisms, and feature extraction methods on the performance of transformer-based models. The findings suggest that transformer-based models can achieve high accuracy and outperform traditional machine learning models in audio classification tasks. Furthermore, the review identifies future directions for research in this area, including exploring more datasets, optimizing the network architecture, and incorporating self-supervised learning techniques. Overall, this literature review provides valuable insights into the state-of-the-art transformer-based models for audio classification and highlights their potential in various real-world applications.***

# Introduction

Audio classification is the task of identifying and categorizing audio signals based on their content. It is an essential step in many applications such as music genre classification, speech recognition, and environmental sound monitoring. With the advancement of machine learning techniques, audio classification has been approached using various methods, including traditional signal processing techniques, feature-based approaches, and deep learning techniques. This literature review aims to survey and discuss recent advancements in audio classification using deep learning techniques, with a focus on the transformer-based models.

# Background

Audio classification has been a challenging task due to the complexity of audio signals and their high dimensionality. Traditional signal processing techniques such as Fourier Transform, Mel-Frequency Cepstral Coefficients (MFCCs), and Linear Predictive Coding (LPC) have been used for feature extraction in audio classification. These methods have shown promising results in various audio classification tasks, but they have limitations in dealing with the high dimensionality of audio data and lack of modeling complex relationships between audio features.

Deep learning techniques have shown significant improvements in audio classification tasks due to their ability to model complex relationships between audio features and handle high dimensional data. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and their variations have been used for audio classification tasks with notable success. Recently, transformer-based models have emerged as a promising approach for audio classification tasks, mimicking the success of transformer-based models in Natural Language Processing (NLP) tasks.

# Transformers based models

Transformer-based models were initially introduced for language translation tasks, where they have shown remarkable results, outperforming previous models significantly. The transformer model replaces the traditional RNNs with self-attention mechanisms, which enables the model to focus on relevant parts of the input sequence when generating output. The transformer model consists of an encoder and a decoder, where the encoder encodes the input sequence into a fixed-length vector representation, and the decoder generates the output sequence from the encoded representation.

Recent works have shown the effectiveness of transformer-based models in audio classification tasks, where they have achieved state-of-the-art results. Gong et al. [6] proposed the Audio Spectrogram Transformer (AST), a transformer-based model that takes mel-spectrograms as input and uses a self-attention mechanism to generate an encoded representation of the input. Koutini et al. [7] proposed an extension to the AST model, called Patchout, which introduces random masking of attention weights during training to improve the generalization of the model. Chen et al. [8] proposed the Hierarchical Token-Semantic Audio Transformer (HTS-AT), a transformer-based model that uses a hierarchical attention mechanism to model both the temporal and semantic structure of the input.

Liu et al. [9] proposed the Causal Audio Transformer (CAT), a transformer-based model that uses a causal self-attention mechanism, where the output at each position depends only on the inputs at positions before it. Gong et al. [10] proposed the Self-Supervised Audio Spectrogram Transformer (SSAST), a self-supervised transformer-based model that uses contrastive learning to learn a general representation of audio data without labels. Liu and Fang [11] proposed the Multi-Scale Audio Spectrogram Transformer (MS-AST), a transformer-based model that uses multi-scale input representations to capture both short-term and long-term temporal information.

# Audio spectrogram transaformer

The Audio Spectrogram Transformer (AST) is a transformer-based architecture for audio classification that uses the spectrogram as input [6]. Gong et al. proposed this model as a more effective way of processing audio data by adapting the original transformer architecture to spectrograms. The authors propose two variations of the AST architecture: AST and Time-domain Audio Transformer (TAT). AST employs a 2D convolutional neural network (CNN) to transform the spectrogram into patches, which are then processed by the transformer layers. TAT, on the other hand, utilizes 1D convolutions to extract features from the raw audio waveform before transforming them using the transformer layers.

Gong et al. reported that AST outperformed traditional CNN-based approaches for audio classification tasks, such as Environmental Sound Classification (ESC)-50 and Speech Commands v0.02. The model achieved an accuracy of 87.9% on the ESC-50 dataset, which is 7.7% higher than the best-performing CNN model on the same dataset. Similarly, AST achieved an accuracy of 94.4% on the Speech Commands v0.02 dataset, which is 2.2% higher than the best-performing CNN model on the same dataset. These results suggest that AST is a powerful architecture for audio classification tasks audio waveform before transforming them using the transformer layers.

Gong et al. reported that AST outperformed traditional CNN-based approaches for audio classification tasks, such as Environmental Sound Classification (ESC)-50 and Speech Commands v0.02. The model achieved an accuracy of 87.9% on the ESC-50 dataset, which is 7.7% higher than the best-performing CNN model on the same dataset. Similarly, AST achieved an accuracy of 94.4% on the Speech Commands v0.02 dataset, which is 2.2% higher than the best-performing CNN model on the same dataset. These results suggest that AST is a powerful architecture for audio classification tasks.

# Hierarchical Token-Semantic Audio Transformer

The Hierarchical Token-Semantic Audio Transformer (HTS-AT) is a transformer-based architecture for audio classification that uses a hierarchical token-semantic representation of the audio data [8]. Chen et al. proposed this model as a way to capture both the local and global context of the audio signal by leveraging the semantic relationships between tokens. The authors employed a hierarchical architecture consisting of a semantic encoder and a token encoder to learn a token-semantic representation of the audio data. The token-semantic representation is then used as input to the transformer layers for classification.

Chen et al. evaluated HTS-AT on the ESC-50, UrbanSound8k, and Speech Commands v0.02 datasets and reported that it outperformed several state-of-the-art approaches. The model achieved an accuracy of 88.2% on the ESC-50 dataset, which is 4.3% higher than the best-performing approach on the same dataset. Similarly, HTS-AT achieved an accuracy of 90.5% on the UrbanSound8k dataset, which is 0.9% higher than the best-performing approach on the same dataset. These results suggest that HTS-AT is a promising architecture for audio classification tasks that leverages the semantic relationships between tokens to capture both local and global context.

# Causal Audio Transformer

The Causal Audio Transformer (CAT) is a transformer-based architecture for audio classification that uses a causal convolutional layer to process the audio data [9]. Liu et al. proposed this model as a way to capture the temporal information in the audio signal while still leveraging the advantages of the transformer architecture. The authors employed a causal convolutional layer to extract features from the audio signal before transforming them using the transformer layers. The authors also proposed a novel pooling mechanism called causal pooling that leverages the causal structure of the convolutional layer to avoid information leakage from future time steps.

# Evaluation

Another recent study by Gong et al. [6] proposed the Audio Spectrogram Transformer (AST), which aims to improve the efficiency of training audio transformers by reducing the spatial resolution of the input spectrograms through a patch-out strategy. The authors evaluated the AST model on two datasets, ESC-50 and UrbanSound8k, achieving state-of-the-art performance on both datasets.

Similarly, Koutini et al. [7] also proposed a patch-out strategy for training audio transformers, called PatchOut. This method randomly masks small patches of the input spectrogram to improve generalization and avoid overfitting. The authors evaluated the PatchOut model on the ESC-50 and UrbanSound8k datasets and achieved performance comparable to other state-of-the-art models.

Recently, a new hierarchical token-semantic audio transformer (HTS-AT) was proposed by Chen et al. [8] for sound classification and detection. The HTS-AT model uses a two-level hierarchical structure to encode the input spectrograms, which allows the model to learn both low-level and high-level representations of the audio data. The authors evaluated the HTS-AT model on several datasets, including ESC-50 and UrbanSound8k, achieving state-of-the-art performance on both datasets.

Another recent study by Liu et al. [9] proposed the Causal Audio Transformer (CAT) for audio classification tasks. The CAT model is designed to capture the temporal dependencies in the input spectrograms by using a causal convolutional layer in the encoder. The authors evaluated the CAT model on several datasets, including ESC-50 and UrbanSound8k, and achieved state-of-the-art performance on both datasets.

In addition to supervised learning approaches, self-supervised learning has also been applied to audio classification tasks using transformer-based models. Gong et al. [10] proposed the Self-Supervised Audio Spectrogram Transformer (SSAST) for audio classification tasks. The SSAST model is trained using a contrastive self-supervised learning objective, which allows the model to learn representations of the input spectrograms without requiring labeled data. The authors evaluated the SSAST model on several datasets, including ESC-50 and UrbanSound8k, and achieved performance comparable to other state-of-the-art models.

Finally, several studies have explored modifications to the transformer architecture to improve its performance on audio classification tasks. Liu and colleagues [11] proposed the Multi-Scale Audio Spectrogram Transformer (MS-AST) for classroom teaching interaction recognition. The MS-AST model uses multi-scale convolutions and parallel attention modules to capture both local and global information in the input spectrograms. The authors evaluated the MS-AST model on a classroom audio dataset and achieved state-of-the-art performance on this task.

Overall, these studies demonstrate the effectiveness of transformer-based models for audio classification tasks. These models have achieved state-of-the-art performance on several benchmark datasets and have shown promising results for real-world applications. However, there are still challenges to be addressed, such as improving the robustness of these models to noise and environmental variations, and developing more efficient training methods to reduce the computational cost of these models.

# Conclusion

In this paper, we have presented a review of transformer-based networks for audio classification. We have shown that transformer-based networks have achieved state-of-the-art performance on various audio classification tasks, including the ESC-50, Speech Commands v0.02, and UrbanSound8k datasets. The proposed models leverage the attention mechanism in transformers to capture the dependencies in the audio signals and learn better representations of the signals for classification. We have also discussed several avenues for future research, including exploring different hyperparameters and network architectures, investigating other feature extraction methods to improve the quality of the input data, and evaluating the models on more datasets.

In summary, transformer-based networks have shown promising results for audio classification and provide a viable alternative to traditional models such as CNNs and RNNs. As the field of audio classification continues to evolve, it is likely that transformer-based networks will continue to play an important role in advancing the state-of-the-art in this area.

# References

[1**]  Daniel Jurafsky and James H. Martin. 2000. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (1st. ed.). Prentice Hall PTR, USA.**

**[2]  M. Woelfel and J. McDonough, Distant Speech Recognition. Nashville, TN: John Wiley & Sons, 2009.**

**[3]  D. O’shaughnessy, Speech communication: Human and Machine. Universities press, 1987.**

**[4]  K. S. Rao and M. K E, Speech Recognition Using Articulatory and Excitation Source Features. Springer International Publishing, 2017. doi: 10.1007/978-3-319-49220-9.**

**[5]  J.R. Deller, J.H. Hansen, J.G. Proakis, Discrete Time Processing of Speech Signals (Prentice Hall, NJ, 1993).**

**[6]  Y. Gong, Y.-A. Chung, and J. Glass, “AST: Audio Spectrogram Transformer.” arXiv, 2021. doi: 10.48550/ARXIV.2104.01778.**

**[7]  K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, “Efficient Training of Audio Transformers with Patchout,” arXiv, 2021, doi: 10.48550/ARXIV.2110.05069.**

**[8]  K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, “HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection.” arXiv, 2022. doi: 10.48550/ARXIV.2202.00874.**

**[9]  X. Liu, H. Lu, J. Yuan, and X. Li, “CAT: Causal Audio Transformer for Audio Classification.” arXiv, 2023. doi: 10.48550/ARXIV.2303.07626.**

**[10]  Y. Gong, C.-I. J. Lai, Y.-A. Chung, and J. Glass, “SSAST: Self-Supervised Audio Spectrogram Transformer.” arXiv, 2021. doi: 10.48550/ARXIV.2110.09784.**

**[11]  Liu, F.; Fang, J. Multi-Scale Audio Spectrogram Transformer for Classroom Teaching Interaction Recognition. Future Internet 2023, 15, 65. https://doi.org/10.3390/fi15020065.**

**[12]  A. Baade, P. Peng, and D. Harwath, “MAE-AST: Masked Autoencoding Audio Spectrogram Transformer.” arXiv, 2022. doi: 10.48550/ARXIV.2203.16691.**

**[13] N.-C. Ristea, R. T. Ionescu, and F. S. Khan, “SepTr: Separable Transformer for Audio Spectrogram Processing.” arXiv, 2022. doi: 10.48550/ARXIV.2203.09581.**

**[14] E. Rudkowsky, M. Haselmayer, M. Wastian, M. Jenny, Š. Emrich, and M. Sedlmair, “More than Bags of Words: Sentiment Analysis with Word Embeddings,” Communication Methods and Measures, vol. 12, no. 2–3. Informa UK Limited, pp. 140–157, Apr. 03, 2018. doi: 10.1080/19312458.2018.1455817.**

**[15] A. Vaswani et al., “Attention Is All You Need.” arXiv, 2017. doi: 10.48550/ARXIV.1706.03762.**

**[16] K. J. Piczak, “ESC: Dataset for Environmental Sound Classification” Proceedings of the 23rd ACM international conference on Multimedia. ACM, Oct. 13, 2015. doi: 10.1145/2733373.2806390.**

**[17] P. Warden, “Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition.” arXiv, 2018. doi: 10.48550/ARXIV.1804.03209.**

**[18] J.Salamon,C.Jacoby,andJ.P.Bello,“ADatasetandTaxonomyfor Urban Sound Research,” Proceedings of the 22nd ACM international conference on Multimedia. ACM, Nov. 03, 2014. doi: 10.1145/2647868.2655045.**

**[19] Y. Zhang, B. Li, H. Fang, and Q. Meng, “Spectrogram Transformers for Audio Classification,” 2022 IEEE International Conference on Imaging Systems and Techniques (IST). IEEE, Jun. 21, 2022. doi: 10.1109/ist55454.2022.982**

**[20] P. Li, J. Wu, Y. Wang, Q. Lan, and W. Xiao, “STM: Spectrogram Transformer Model for Underwater Acoustic Target Recognition,” Journal of Marine Science and Engineering, vol. 10, no. 10. MDPI AG, p. 1428, Oct. 04, 2022. doi: 10.3390/jmse10101428.**

**[21] S. Park, Y. Jeong, T. Lee, “Many-to-Many Audio Spectrogram Transformer: Transformer for Sound Event Localization and Detection” 6th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2021), Nov. 2021**

**[22] Y. Khasgiwala and J. Tailor, “Vision Transformer for Music Genre Classification using Mel-frequency Cepstrum Coefficient,” 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON). IEEE, Sep. 24, 2021. doi: 10.1109/gucon50781.2021.9573.**