EAViT: External Attention Vision Transformer for Audio Classification

Aquib Iqbal[†*], Abid Hasan Zim[‡*], Md Asaduzzaman Tonmoy[§], Limengnan Zhou[¶], Asad Malik[‖], Minoru Kuribayashi[††]

[†] Department of Computer Science, University of Massachusetts, Amherst, MA, USA, E-mail: aquibiqbal@umass.edu

[‡] Department of Mechanical Engineering, Aligarh Muslim University, India, E-mail: abid@zhcet.ac.in

[§] Department of Electrical Engineering, Aligarh Muslim University, India, E-mail: asaduzzaman.tonmoy@outlook.com

[¶] School of Electronic and Information Engineering, UESTC, China, E-mail: dreamzlmn@foxmail.com

[‖] School of Information Technology, Monash University Malaysia, Malaysia, E-mail: asad.malik@monash.edu

[††] Center for Data-driven Science and Artificial Intelligence, Tohoku University, Japan, E-mail: kminoru@tohoku.ac.jp

***Abstract*—This study presents the External Attention Vision Transformer (EAViT) model, a novel approach designed to improve audio classification accuracy in response to the growing need for precise classification systems, particularly for recommendation engines and user personalization in applications such as music streaming. Accurate audio classification is crucial for organizing vast audio libraries into coherent categories, enabling users to find and interact with their preferred audio content more effectively. In this study, we utilize the GTZAN dataset, which comprises 1,000 music excerpts spanning ten diverse genres. Each 30-second audio clip is segmented into 3-second excerpts to enhance dataset robustness and mitigate overfitting risks, allowing for more granular feature analysis. The EAViT model integrates multi-head external attention (MEA) mechanisms into the Vision Transformer (ViT) framework, effectively capturing long-range dependencies and potential correlations between samples. This external attention (EA) mechanism employs learnable memory units that enhance the network's capacity to process complex audio features efficiently. The study demonstrates that EAViT achieves a remarkable overall accuracy of 93.99%, surpassing state-of-the-art models.**

I. Introduction

The increasing availability of digital music resources has led to rapid advancements in multimedia technology. As a result, consumers have progressively shifted towards accessing music through online streaming platforms. AI methods are particularly effective in addressing the complex task of meeting diverse and intricate music retrieval needs from extensive music collections. For instance, consumers may desire to listen to a specific song within a particular genre that conveys a certain emotional tone. In these instances, accurate music labelling is crucial to ensure the precise music is delivered [1], [2]. Additionally, numerous subscription and recommendation systems require the inclusion of music genre preferences to provide customers with more specific and tailored content. Music is a highly diverse art form that encompasses various elements, including rhythm, melody, and harmony. Music media systems use textual labels to categorize or retrieve music, since comprehending music in its exact form requires substantial previous knowledge. Hence, the classification of music genres is an essential component of music information retrieval [3].

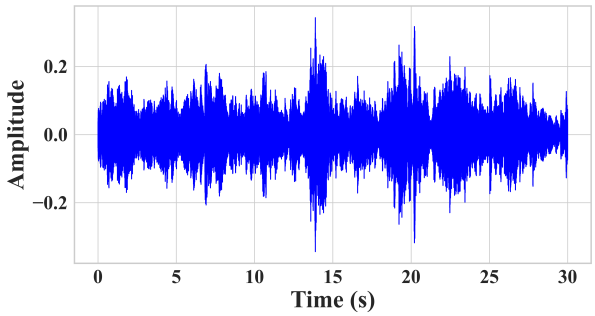

Fig. 1. Raw waveform representation of an audio sample.

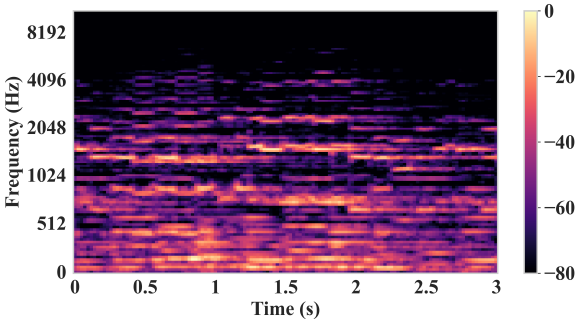

Fig. 2. Mel spectrogram of an audio sample.

Over the past decade, traditional machine learning classifiers have been employed for this purpose. However, these classifiers possess a shallow structure, limiting their capacity to effectively learn and interpret music data. Manual extraction of musical elements has shown limited applicability and lacks robustness. To address the intricate aspects of music, innovative deep-learning approaches have been adopted in recent years. These advanced techniques offer improved accuracy

*Equal contributions.

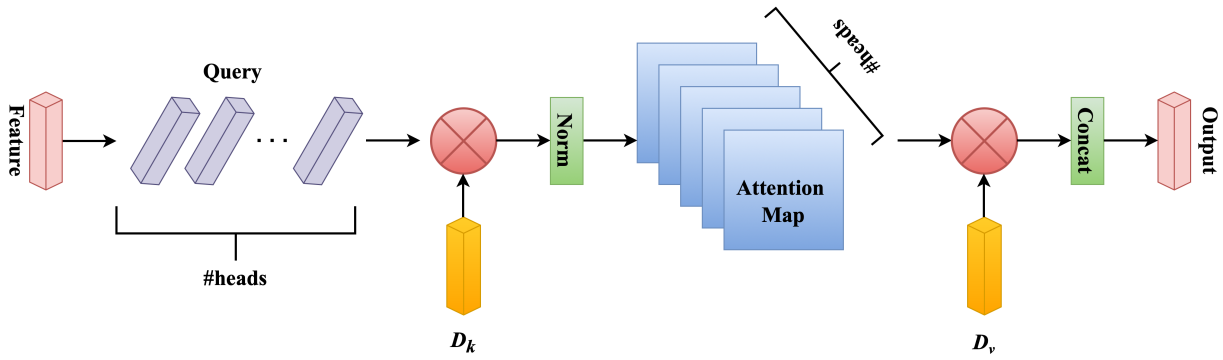

Fig. 3. Multi-head external-attention (MEA).

and efficiency in music genre classification and retrieval [4], [5]. The most recent advancements in audio classification performance are generated using distinct resource-intensive methodologies [6].

Over the years, a large number of classification methods have been created to categorize audio-based data [7]. A study conducted video classification and audio event identification using a convolutional neural network (CNN) on an extensive dataset. By utilizing subsets of varying sizes and initializing convolutional layers with a gammatone filter bank, which is modelled after the human ear, the study examined the impact of these methods on the system's accuracy. This approach allowed for a thorough investigation of model training and performance [8]. Researchers have employed Long Short-Term Memory (LSTM) networks to achieve promising results for the MGC task [9]. Another study introduced a hybrid architecture combining CNNs and RNNs to capture the spatiotemporal characteristics of music [10]. Additionally, a study proposed an attention mechanism-based Bi-LSTM architecture to exploit the varying importance of temporal frames, further enhancing the accuracy and performance of music classification tasks [11]. A research study compared a CNN and a TDSN for environmental sound classification using spectrogram images. The results revealed that the CNN achieved superior performance on the ESC-10 dataset, with an accuracy of 77%, compared to 56% for the TDSN [12]. Additionally, a lightweight CNN utilizing MFCC for environmental sound classification on the UrbanSound8k dataset achieved a notable accuracy of 95.59%, showcasing competitive performance against more complex models [13]. A further investigation concentrated on classifying music across three distinct datasets (ISMIR 2004, LMD, and ethnic African music). The findings revealed that CNN outperformed previous classifiers by achieving an impressive accuracy rate of 92% [14].

Recently introduced transformer-based models offer substantial advancements in AI applications, exhibiting notable performance enhancements compared to current models [15], [16]. Recent advancements include the Causal Audio Transformer (CAT), which employs MRMF extraction and an acoustic attention block for enhanced audio modelling. The proposed causal module mitigates overfitting, facilitates knowledge transfer, and improves interpretability. It demonstrated superior performance on datasets such as ESC50, AudioSet, and UrbanSound8K, and are adaptable to various other transformer-based models [17]. Another study applied a transformer model to music genre classification, converting audio to log-amplitude Mel spectrograms, and achieved promising results on the GTZAN dataset, indicating the potential of transformers in this domain [18].

In this study, EAViT with MEA is proposed to classify music genres in the GTZAN dataset. The proposed approach includes the introduction of an innovative EA mechanism to enhance the classification accuracy of ViT-based models when applied to audio data. EAViT integrates the strengths of traditional vision transformers with specialized attention mechanisms tailored for audio signal processing, achieving superior performance over conventional models. Extensive experiments were conducted on the GTZAN dataset to benchmark our model against state-of-the-art methods, demonstrating significant improvements in classification accuracy. Our contributions in this paper are summarized as follows:

- The EAViT model is proposed for the task of music genre classification, leveraging the integration of MEA mechanisms within its framework.
- A 30-second raw audio clip is segmented into 3-second portions. Each segment is transformed into the frequency domain via Short-Time Fourier Transform (STFT), converted to decibel units to produce a spectrogram, and further processed with a mel filter bank to create a mel spectrogram, which is then saved as an image.
- The proposed models are compared with state-of-the-art methods from existing literature.

The rest of the paper is structured as follows: Section II outlines the methodology, Section III explains the proposed method, Section IV presents the experiments and results, and Section V contains the conclusion.

II. Methodology

A. Data

GTZAN dataset was used which comprises of 1,000 audio excerpts spanning 10 different genres: Blues, Classical, Country, Disco, Hip-hop, Jazz, Metal, Pop, Reggae, and Rock. Each

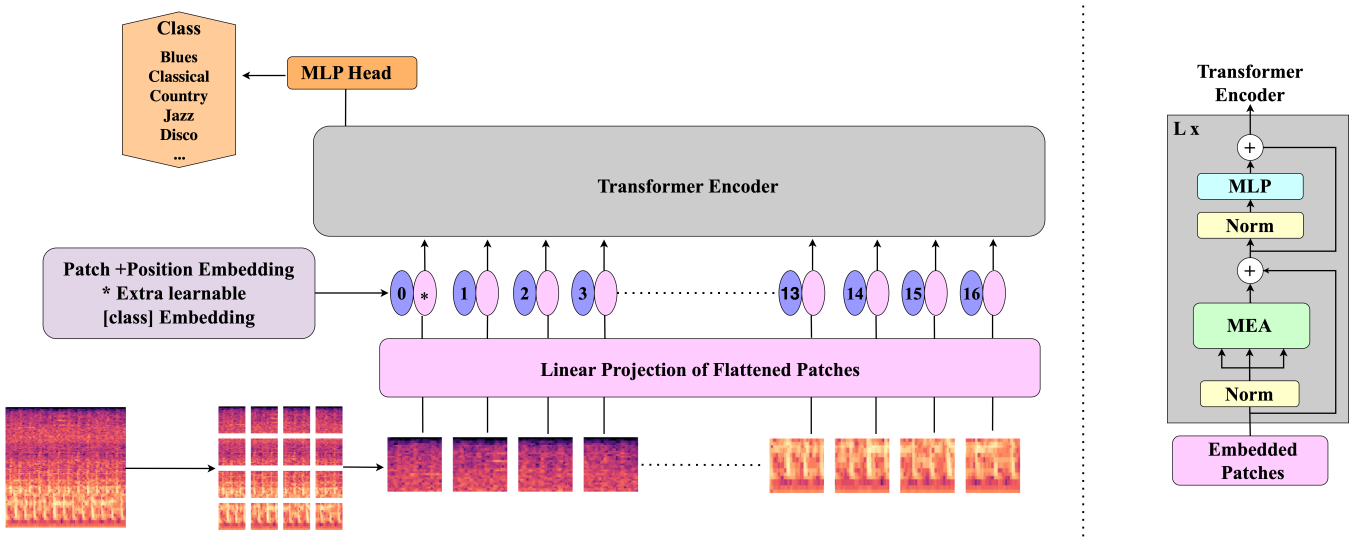

Fig. 4. Model Overview: EAViT.

genre includes 100 excerpts, each lasting approximately 30 seconds and stored as 22,050 Hz, 16-bit, mono audio files [19].

We segmented each 30-second audio excerpt from the GTZAN dataset into 3-second clips, significantly expanding the dataset and mitigating the risk of overfitting. This segmentation allows for a more granular analysis of audio features, enhancing the dataset's robustness. To illustrate the transformation process, we included both the raw audio signals and the corresponding mel spectrogram Fig. 1 and Fig. 2.

Mel Spectrogram represents sound in the time-frequency domain, using the mel scale on the y-axis and the decibel scale to indicate colour intensity. This spectrogram is created by applying a filter bank to the frequency-domain representation of time-windowed audio signals, providing a perception-aligned view of the sound. The conversion process involves transforming each 3-second audio segment into the frequency domain using the STFT. The resulting frequency data is then converted into decibel units to form a standard spectrogram. By applying a mel filter bank, we transform this spectrogram into a mel spectrogram, which is then saved as an image file.

B. Multi-head external-attention (MEA)

In this study, we incorporated the MEA mechanism into the transformer encoder of the ViT. To improve the network's capabilities, the EA mechanism uses two learnable memory units, $M_k \in \mathbb{R}^{d\times S}$ and $M_v \in \mathbb{R}^{d\times S}$, which serve as key and value components. Importantly, M_k and M_v function as memories for all samples in the training set and are independent of the input. Specifically, the flow process of the EA module is defined by Eq.(1) and (2), as follows:

$$A = (\alpha)_{n,m} = \text{Norm}(F_{\text{in}} M_k^T), \tag{1}$$

$$F_{\text{out}} = AM_v, \tag{2}$$

where F_{in} represents the input feature, F_{out} represents the output feature, and Norm$(\cdot)$ denotes the double normalization operator. MEA can be written as:

$$h_i = \text{ExternalAttention}(F_i, M_k, M_v), \tag{3}$$

$$F_{\text{out}} = \text{MultiHead}(F, M_k, M_v) \tag{4}$$

$$= \text{Concat}(h_1, \ldots, h_H)W_o, \tag{5}$$

where h_i represents the i-th head, H denotes the total number of heads, and W_o is a linear transformation matrix ensuring consistent input and output dimensions. $M_k \in \mathbb{R}^{d\times S}$ and $M_v \in \mathbb{R}^{d\times S}$ serve as the shared memory units across different heads [20]. Here Fig. 3 shows the structure MEA.

III. Proposed Method

A. External Attention Vision Transformer (EAViT)

The conventional ViT utilizes multi-head self-attention, which enhances the feature at each position by computing a weighted sum of features based on pair-wise affinities across all positions. This process effectively captures long-range dependencies within a single sample. However, self-attention suffers from quadratic complexity and overlooks potential correlations between different samples. In contrast, EA leverages two small, learnable, shared memories, which can be efficiently implemented using two cascaded linear layers and two normalization layers. In this study, we have utilized the EAViT model, which incorporates MEA within the transformer encoder.

The ViT is specifically designed for image classification tasks by directly applying the transformer architecture to sequences of image patches [15]. It closely adheres to the original transformer design. In order to process 2D images, the input tensor $x \in R^{H\times W\times C}$ is transformed into a set of flattened 2D patches denoted as $x_P \in R^{N\times(P^2\cdot C)}$, where C

represents the number of channels. The main image has a resolution of (H, W), while each image patch has a (P, P) resolution. Consequently, the actual sequence length for the transformer is given by $N = HW/P^2$. The output of a learnable linear projection (as expressed in Eq.(6)) called patch embeddings, associates each vectorized patch with the model dimension D, as the transformer maintains consistent widths across all its layers.

$$\mathrm{z}_0 = \left[\mathrm{x}_{\text{class}}\,; \mathrm{x}_p^1\mathbf{E}\,; \mathrm{x}_p^2\mathbf{E}\,; \ldots\,; \mathrm{x}_p^N\mathbf{E}\right] + \mathbf{E}_{pos}, \tag{6}$$

where $\mathbf{E} \in \mathbb{R}^{\left(P^2 \cdot C\right) \times D}$ and $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1) \times D}$.

Researchers introduce a trainable embedding into the sequence of embedded patches, akin to BERT's $[class]$ token, represented as $(z_0^0 = x_{\text{class}})$. This embedding's state at the output of the Transformer encoder, denoted as (z_L^0), is used as the image representation y (refer to Eq.(9)). During both the pre-training and fine-tuning stages, a classification head is connected to z_L^0. In the pre-training phase, a Multi-Layer Perceptron (MLP) with a single hidden layer guides the classification head. In contrast, a single linear layer is utilized during the fine-tuning phase [21]. Here Fig. 4 shows the overview of the model.

$$\mathrm{z}'_\ell = \mathrm{MEA}\left(\mathrm{LN}\left(\mathrm{z}_{\ell-1}\right)\right) + \mathrm{z}_{\ell-1}, \quad \ell = 1 \ldots L \tag{7}$$

$$\mathrm{z}_\ell = \mathrm{MLP}\left(\mathrm{LN}\left(\mathrm{z}'_\ell\right)\right) + \mathrm{z}'_\ell, \quad \ell = 1 \ldots L \tag{8}$$

$$\mathrm{y} = \mathrm{LN}\left(\mathrm{z}_L^0\right) \tag{9}$$

To retain positional information, position embeddings are added to the patch embeddings. As more complex 2D-aware position embeddings did not significantly enhance performance, the researcher opted for traditional learnable 1D position embeddings. The generated sequence of embedding vectors is then fed into the encoder. The Transformer encoder comprises layers of MEA and MLP blocks, as described in Eq.(7) and (8). Each block is preceded by Layernorm (LN) and followed by residual connections.

In our proposed transformer-based methodology for the classification of 2D Mel spectrogram images, critical hyperparameters are meticulously selected to enhance model efficacy. The input images, resized to 256×256 pixels, are segmented into 16 patches. The learning rate is established at 0.001, accompanied by a weight decay of 0.0001 to mitigate overfitting, and the batch size is set at 256 to facilitate efficient training iterations. The transformer architecture incorporates 16 layers, each with a projection dimension of 32 and 8 attention heads. The MLP classification head comprises two hidden layers with [2048, 1024] units. These hyperparameters collectively define the model's capacity, attentiveness, and overall classification performance.

**The model has been reimplemented.

TABLE I
PROPOSED MODELS COMPARISON WITH SOTA METHODS FROM LITERATURE.

Model	Accuracy
ViT**	91.79%
RNNCA [22]	93.1%
1D-CNN with BiRNN and attention mechanism [23]	91.99%
SA-SLnO with optimization [24]	85.63%
EAViT (our)	**93.99%**

TABLE II
EVALUATION OF CLASS-SPECIFIC CLASSIFICATION RESULTS OF THE PROPOSED EAVIT MODEL

Class	Precision	Recall	F1-Score
Blues	0.94	0.96	0.95
Classical	0.99	0.97	0.98
Country	0.89	0.93	0.91
Disco	0.95	0.97	0.96
Hiphop	0.93	0.90	0.91
Jazz	0.92	0.94	0.93
Metal	0.96	0.93	0.95
Pop	0.99	0.92	0.96
Reggae	0.93	0.92	0.93
Rock	0.90	0.94	0.92

IV. EXPERIMENTS AND RESULTS

A. Evaluation Matrices

To evaluate the performance of the models, we employed key performance metrics, including accuracy, precision, recall, and the F1-score. Accuracy represents the ratio of correct predictions to the total number of predictions. Precision is calculated as the ratio of true positive predictions to the sum of true positive and false-positive predictions across all classes. Recall is the ratio of true positive predictions to the sum of true positive and false-negative predictions across all classes. The F1-Score is the harmonic mean of precision and recall, balancing the consideration of both false positive and false negative predictions. The mathematical expressions for these metrics are as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \tag{10}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{11}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{12}$$

$$\text{F1-score} = 2 \times \left(\frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}\right) \tag{13}$$

Where TP denotes True Positives, TN denotes True Negatives, FN denotes False Negatives, and FP denotes False Positives.

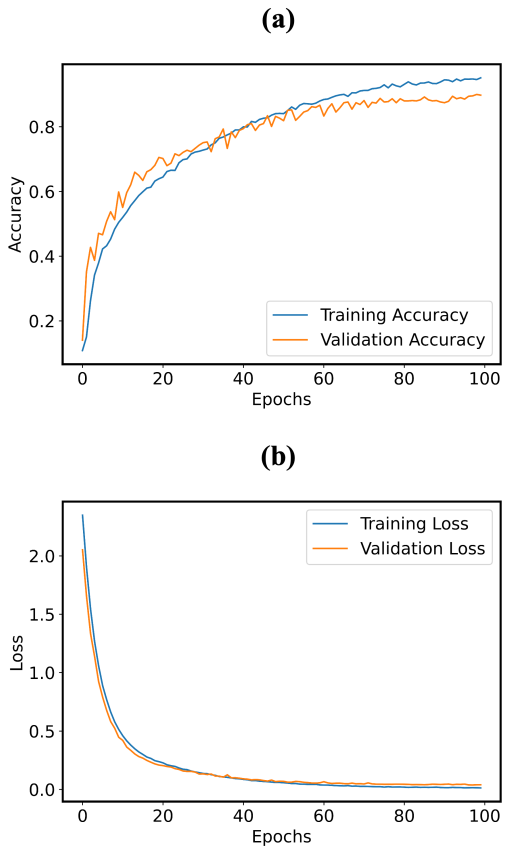

Fig. 5. proposed models: (a) Training and validation accuracy over epochs. (b) Training and validation loss over epochs.

B. Environment

The experiment was conducted using Python 3.11 and TensorFlow 2.16.1. The hardware employed for both training and testing included an AMD Ryzen 9 5900X processor, 64GB of RAM, and an Nvidia GeForce RTX 3090 GPU with 24GB of VRAM.

C. Analysis

In this study, the EAViT model was employed to classify ten different types of music genres. The GTZAN dataset was used in this study. EAViT achieved an impressive overall accuracy of 93.99%. The results underscore the exceptional performance of the EAViT model in music genre classification. Here, Fig. 5 (a) illustrates the training and validation accuracy, and Fig. 5 (b) presents the training and validation loss over a span of 100 epochs respectively. The graphs indicate that the model demonstrates optimal fitting to the dataset. The confusion matrix depicted in Fig. 6 provides a comprehensive overview of the model's performance, illustrating the models classification accuracy.

To demonstrate the robustness and efficacy of the proposed model in this study, we conducted a comparative analysis with state-of-the-art (SOTA) models from the literature, including ViT [15], Recurrent Neural Networks with Channel

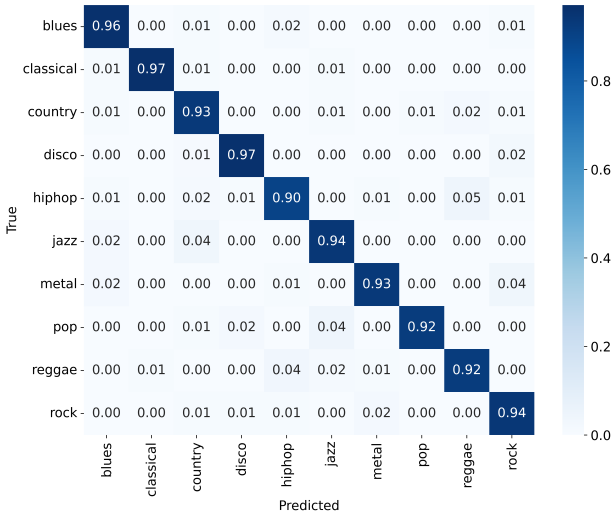

Fig. 6. Confusion matrix of the model.

Attention Mechanism (RNNCA) [22], 1D-CNN with BiRNN and attention mechanism [23], and SA-SLnO with optimization [24]. Table I unequivocally shows that the results of our proposed method surpass the performance of these SOTA models. EAViT surpasses the ViT by 2.37%, RNNCA by 0.95%, and the 1D-CNN with BiRNN and attention mechanism by 2.15%. Furthermore, in comparison to SA-SLnO with optimization model, EAViT achieves a remarkable 9.31% increase in classification accuracy. Here, Table II presents the evaluation metrics, including Precision, Recall, and F1-score, for each class of the proposed EAViT model. These findings underscore the robustness of the proposed model as it demonstrates consistent performance across various classes.

V. Conclusion

In this study, we introduced the External Attention Vision Transformer (EAViT) model, designed to enhance audio classification accuracy by integrating MEA mechanisms within the ViT framework. The EAViT model effectively captures long-range dependencies and correlations between different audio samples, addressing the limitations of traditional self-attention mechanisms. Through extensive experimentation on the GTZAN dataset, comprising 1,000 audio excerpts across ten genres, EAViT achieved a notable accuracy of 93.99%, outperforming state-of-the-art models such. Key metrics, including precision, recall, and F1-score, further demonstrated the model's robustness and reliability. The segmentation of 30-second audio clips into 3-second excerpts enhanced dataset robustness and allowed for detailed feature analysis. These results underscore the potential of the EAViT model to significantly advance audio classification, improving user experiences in various audio-related applications, including music streaming and environmental sound recognition. Future work could explore EAViT's application to other audio classification tasks and datasets, highlighting its versatility and efficacy.

References

[1] A. Elbir and N. Aydin, "Music genre classification and music recommendation by using deep learning," *Electronics Letters*, vol. 56, no. 12, pp. 627–629, 2020.

[2] Y. Li, W. Cao, W. Xie, J. Li, and E. Benetos, "Few-shot class-incremental audio classification using dynamically expanded classifier with self-attention modified prototypes," *IEEE Transactions on Multimedia*, vol. 26, pp. 1346–1360, 2023.

[3] Y. Singh and A. Biswas, "Robustness of musical features on deep learning models for music genre classification," *Expert Systems with Applications*, vol. 199, p. 116 879, 2022.

[4] S. K. Prabhakar and S.-W. Lee, "Holistic approaches to music genre classification using efficient transfer and deep learning techniques," *Expert Systems with Applications*, vol. 211, p. 118 636, 2023.

[5] A. Sterling, J. Wilson, S. Lowe, and M. C. Lin, "Isnn: Impact sound neural network for audio-visual object classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 555–572.

[6] Z. Zhang, S. Xu, S. Zhang, T. Qiao, and S. Cao, "Learning attentive representations for environmental sound classification," *IEEE Access*, vol. 7, pp. 130 327–130 339, 2019.

[7] S. Hershey, S. Chaudhuri, D. P. Ellis, *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, IEEE, 2017, pp. 131–135.

[8] B. Mishachandar and S. Vairamuthu, "Diverse ocean noise classification using deep learning," *Applied Acoustics*, vol. 181, p. 108 141, 2021.

[9] J. Dai, S. Liang, W. Xue, C. Ni, and W. Liu, "Long short-term memory recurrent neural network based segment features for music genre classification," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, IEEE, 2016, pp. 1–5.

[10] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2017, pp. 2392–2396.

[11] Y. Yu, S. Luo, S. Liu, H. Qiao, Y. Liu, and L. Feng, "Deep attention based music genre classification," *Neurocomputing*, vol. 372, pp. 84–91, 2020.

[12] A. Khamparia, D. Gupta, N. G. Nguyen, A. Khanna, B. Pandey, and P. Tiwari, "Sound classification using convolutional neural network and tensor deep stacking network," *IEEE Access*, vol. 7, pp. 7717–7727, 2019.

[13] Y. A. Al-Hattab, H. F. Zaki, and A. A. Shafie, "Rethinking environmental sound classification using convolutional neural networks: Optimized parameter tuning of single feature extraction," *Neural Computing and Applications*, vol. 33, no. 21, pp. 14 495–14 506, 2021.

[14] Y. M. Costa, L. S. Oliveira, and C. N. Silla Jr, "An evaluation of convolutional neural networks for music classification using spectrograms," *Applied soft computing*, vol. 52, pp. 28–38, 2017.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[16] A. H. Zim, A. Iqbal, A. Malik, Z. Dong, and H. Wu, "Tcnformer: Temporal convolutional network former for short-term wind speed forecasting," *arXiv preprint arXiv:2408.15737*, 2024.

[17] X. Liu, H. Lu, J. Yuan, and X. Li, "Cat: Causal audio transformer for audio classification," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[18] Y. Zhuang, Y. Chen, and J. Zheng, "Music genre classification with transformer classifier," in *Proceedings of the 2020 4th international conference on digital signal processing*, 2020, pp. 155–159.

[19] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, 2002.

[20] M.-H. Guo, Z.-N. Liu, T.-J. Mu, and S.-M. Hu, "Beyond self-attention: External attention using two linear layers for visual tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5436–5447, 2022.

[21] A. H. Zim, A. Ashraf, A. Iqbal, A. Malik, and M. Kuribayashi, "A vision transformer-based approach to bearing fault classification via vibration signals," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2022, pp. 1321–1326.

[22] J. Gan, "Music feature classification based on recurrent neural networks with channel attention mechanism," *Mobile Information Systems*, vol. 2021, no. 1, p. 7 629 994, 2021.

[23] K. Zhang, "Music style classification algorithm based on music feature extraction and deep neural network," *Wireless Communications and Mobile Computing*, vol. 2021, no. 1, p. 9 298 654, 2021.

[24] B. Kumaraswamy and P. Poonacha, "Deep convolutional neural network for musical genre classification via new self adaptive sea lion optimization," *Applied Soft Computing*, vol. 108, p. 107 446, 2021.