Spectrogram Transformers for Audio Classification

Yixiao Zhang, Baihua Li*, Hui Fang, Qinggang Meng
Department of Computer Science
Loughborough University
Loughborough, U.K.
{Y.Zhang8, B.Li, H.Fang, Q.Meng}@lboro.ac.uk

2022 IEEE International Conference on Imaging Systems and Techniques (IST) | 978-1-6654-8102-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IST55454.2022.9827729

***Abstract*—Audio classification is an important task in the machine learning field with a wide range of applications. Since the last decade, deep learning based methods have been widely used and the transformer-based models are becoming new paradigm for audio classification. In this paper, we present Spectrogram Transformers, which are a group of transformer-based models for audio classification. Based on the fundamental semantics of audio spectrogram, we design two mechanisms to extract temporal and frequency features from audio spectrogram, named time-dimension sampling and frequency-dimension sampling. These discriminative representations are then enhanced by various combinations of attention block architectures, including Temporal Only (TO) attention, Temporal-Frequency sequential (TFS) attention, Temporal-Frequency Parallel (TFP) attention, and Two-stream Temporal-Frequency (TSTF) attention, to extract the sound record signatures to serve the classification task. Our experiments demonstrate that these Transformer models outperform the state-of-the-art methods on ESC-50 dataset without pre-training stage. Furthermore, our method also shows great efficiency compared with other leading methods.**

***Keywords*—Transformer, Spectrogram, Audio representation, Audio classification**

I. Introduction

Sound is a crucial signifier which contains rich high-level semantic environmental information. Consequently, computerised audio classification, aiming to recognise a various of sound patterns, has been developed for decades [1]. It is still one of the most important tasks in machine learning which is driven by a wide range of real-world applications, including surveillance [2], monitoring [3], intelligent fault diagnosis of machines for safety [4], and animals detection for nature reserve protection [5].

Deep learning based models are the most popular methods used for the audio classification [6–9]. There are many studies using CNN architectures to adapt pre-trained models on audio representations, e.g., Spectrograms [10–12] and Mel-Frequency Cepstral Coefficients (MFCC) [13–15], to significantly improve the performance of audio classification, tagging, and recognition tasks [16, 17]. Recently, transformer based models [9, 18, 19] have been emerging to further boost the performance. Transformer models enable to model long feature dependencies and support parallel processing. Since their success on natural language processing tasks, they have great potential to model the time series signals, i.e. audio, in our work.

*Corresponding author

In this paper, to further design an effective and efficient transformer network architecture, we propose the Spectrogram Transformers, a novel family of Transformer models, for audio classification task. Specifically, we introduce two sampling methods to extract features from audio spectrogram, which include time-dimension sampling and frequency-dimension sampling. These features are further used in four variants of architectures named Temporal Only (TO) Transformer, Temporal-Frequency Sequential (TFS) Transformer, Temporal-Frequency Parallel (TFP) Transformer, and Two-stream Temporal-Frequency (TSTF) Transformer. We test our models and compare them to the state-of-the-art methods on ESC-50 dataset, which is a standard dataset for the audio classification task.

The contributions of this paper can be summarised as follows: first, current transformer networks are based on ViT architecture [20] which uses the patches cropped from spectrogram as input. We are the first method to investigate different sampling and embedding generalisation methods since our feature partitions are more meaningful from a domain knowledge perspective. Secondly, our Transformer framework achieved 11.89% performance uplifted compared to the state-of-the-art method "AST" [9] on accuracy when processing ESC-50 dataset without pre-training. Thirdly, our architectures have gained large margins in model efficiency.

The rest of our paper is organized as follows: Section II introduces related work for audio classification task. Section III describes our novel Spectrogram Transformer architectures. Section IV presents our experimental results compared to other methods. Section V draws the conclusion and discusses our future work.

II. Related Work

Time-frequency representations are the most common mid-level features used for audio processing. A majority of studies project raw audio signals into the time-frequency space before the analysis [21–23]. Among these features, Spectrograms [10] and Mel-Frequency Cepstral Coefficient (MFCC) features [13] are the most representative since the 2D form allows the exploration from recent DNN methods on the interactions between temporal and frequency dimensions.

CNN-based methods are primarily used when analyzing the audio signals during this decade. Many audio classification models deploy standard image classification models, e.g., Inception, ResNet, and VGG [6, 17] on the time-frequency

978-1-6654-8102-1/22/$31.00 ©2022 IEEE

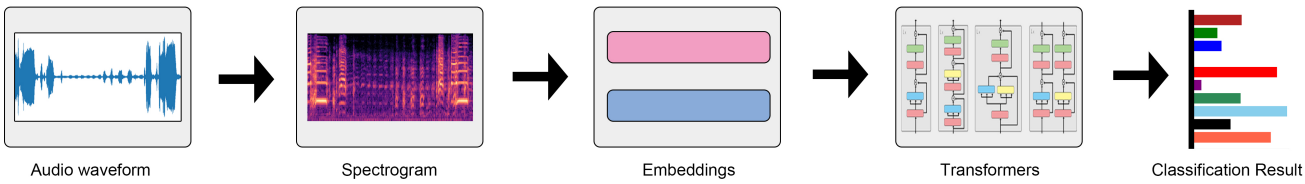

Fig. 1. Pipeline of proposed method. Firstly, audio spectrogram is extracted from audio waveform. Then, time-dimension embedding and frequency-dimension embedding are generated using the proposed sampling methods. Finally, the proposed transformers are used to give classification prediction.

features. To further improve the performance, several new architecture were designed in order to enhance the audio feature extraction and modelling. AclNet [24] presented a VGG architecture with drastically reduced memory to reach a trade-off between accuracy and complexity. SeCoST [25] introduced an audio segment level predictions for its classification. While ERANNs [26] proposed efficient residual audio neural networks for audio pattern recognition.

Recently, the attention mechanism and transformer models are introduced to improve the audio classification with global contextual feature awareness. Attention-augmented convolutional neural network [7] was proposed to enhance the audio features by exploring the relationship between various frequency bands. While [27] discussed to utilise a temporal attention from energy changes over time to improve the representations. In audio classification task, compared with CNN-based methods, one of the advantages of using transformer is these methods support variance of input length since the length of input sequence does not affect the number of parameters in multi-head self-attention or transformer block. When changing the length of input audio, transformer-based method can still capture useful global context information promptly. AST [9] is the first transformer model for audio classification which uses the architecture of the image classification network ViT [20] and adapts the pre-training weights from ViT. PaSST [19] is another method in the leading board which significantly reduces computation and memory complexity of training transformers for the audio domain.

The distinctive merits of our model compared to other transformer models are: (i) we propose a new sampling strategy to extract attentions from audio spectrograms; and (ii) we design temporal multi-head self-attention module and frequency multi-head self-attention module to further investigate effective architectures which integrate these two attentions for better performance.

III. The Spectrogram Transformers

In this section, we explain the details of our proposed spectrogram transformers. We firstly present our overall system pipeline followed by the two sampling mechanisms to extract features which are used in our attention blocks. Then, we introduce the four variants of the transformer architectures and their design logic.

A. Spectrogram Transformer Framework

The processing pipeline of our system is depicted in Fig. 1. When the audio waveform segments are input into the system, they are converted to spectrogram images. Compared to the original audio waveform which is 1D signal, the conversion could potentially boost the DNN performance by exploring the interactions between temporal and frequency features. In our work, we generate 128-dimensional log Mel filter-bank energy features in our system. Specifically, the target sampling rate from audio waveform is 16,000 per second. We use 400 as the length of the Fast Fourier Transform (FFT) window and 160 as the stride step for the sampling. For each second audio, a 100-dimensional feature in time domain are generated. The input feature of our network is $\mathbf{Z} \in \mathbb{R}^{128\times 100t}$ where t is the length of input audio in second. Subsequently, we use time-dimension sampling method and frequency-dimension sampling method (details are introduced in the following subsection) to generate time-dimension embeddings $\mathbf{E_t} \in \mathbb{R}^{100t\times 768}$ and frequency-dimension embeddings $\mathbf{E_f} \in \mathbb{R}^{128\times 768}$. Similar to the ViT [20], we append a learnable classification (CLS) token $\mathbf{CLS} \in \mathbb{R}^{1\times 768}$ to the embeddings for classification task. Since transformer has no access to the sequential information, we also add learnable positional embeddings $\mathbf{E_{pt}} \in \mathbb{R}^{(100t+1)\times 768}$ on the time-dimension embeddings or $\mathbf{E_{pf}} \in \mathbb{R}^{129\times 768}$ for frequency-dimension embeddings in our work. Finally, we input the sequence $\mathbf{E_{pt}} \in \mathbb{R}^{(100t+1)\times 768}$ or $\mathbf{E_{pf}} \in \mathbb{R}^{129\times 768}$ to transformer blocks for the classification task.

B. Time-dimension sampling and frequency-dimension sampling

As illustrated in Fig. 2, we propose two sampling methods to obtain both temporal and frequency features from spectrograms for our transformer models, named time-dimension sampling and frequency-dimension sampling. Different from processing a normal 2D image which has two spatial dimensions, the spectrograms present time and frequency dimensions where we believe it is a more meaningful way to sample them separately for exploring the contextual attentions in our transformer models. Thus, we design time-dimension sampling method which generates embedding $\mathbf{E_t} \in \mathbb{R}^{100t\times 768}$ based on the vectors $\mathbf{z_t^n} \in \mathbb{R}^{1\times 128}, n = 1...100t$ that match to sliding FFT windows at each timestamp n. While for the frequency-dimension sampling method, we produce the frequency-dimension embedding $\mathbf{E_f} \in \mathbb{R}^{128\times 768}$ from $\mathbf{z_f^m} \in \mathbb{R}^{100t\times 1}, m = 1...128$.

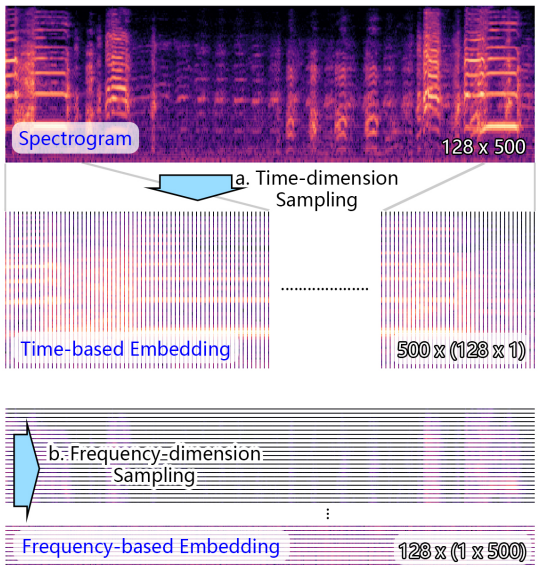

Fig. 2. The sampling methods used for Spectrogram Transformer. For a t-second audio segment, a $128 \times t$ dimension spectrogram could be generated. 128 refers to frequency bins and t refers to sliding FFT windows on time. (a) Time-dimension sampling: By cutting through time-dimension, t time-dimension embedding with shape 128×1 could be generated. (b) Frequency-dimension sampling: By cutting through frequency-dimension, 128 frequency-dimension embedding with shape $1 \times t$ could be generated.

C. Transformer architectures

In our system, we investigate a set of combinations of the use of temporal and frequency attentions to improve the system performance. As illustrated in Fig. 3, we design four transformer-based architectures for audio classification. We firstly start from the simplest model Temporal Only (TO) Transformer which uses a temporal multi-head self-attention. Then, we add a frequency multi-head self-attention into the TO model sequentially and in parallel respectively which are named Temporal-Frequency Sequential (TFS) Transformer and Temporal-Frequency Parallel (TFP) Transformer. Finally, we have the Two-stream Temporal-Frequency (TSTF) Transformer which fuses the attentions before feeding them into a MLP classification head. We use 12 heads self-attention modules and stack six layers of transformer blocks in each transformer encoder.

1) Model1: Temporal Only (TO) Transformer: Since audio is time series data, the sequential information of features are extremely important. Thus, We design the "TO" Transformer for audio spectrograms to capture the attentions between the features at temporal space. Simply put, our "TO" transformer implements the original multi-head self-attention [28] on the dimension of time.

2) Model2: Temporal-Frequency Sequential (TFS) Transformer: Inspired by Wu et al. [7], the attention from frequency bands also contribute to improve the classification accuracy. In this architecture, we explore to use the frequency attention and the temporal attention to further enhance the temporal feature. To diversify frequency features for the attention computation, we propose frequency multi-head self-attention in our transformer block to expand the original frequency features from 128D to 768D via a linear projection layer. Since the temporal features are more reliable compared to the frequency features, we use the temporal multi-head self-attention module before the frequency multi-head self-attention module in the architecture.

3) Model3: Temporal-Frequency Parallel (TFP) Transformer: We make another design to explore the ensemble effect of using both the temporal attention and frequency attention to enhance the time-dimension embeddings. In this architecture, we use temporal multi-head self-attention and frequency multi-head self-attention in parallel with a residual connection as illustrated in Fig. 3.

4) Model4: Two-stream Temporal-Frequency (TSTF) Transformer: The fourth model architecture is a two-stream structure. There are two differences between TSTF Transformer with previous TFP Transformer and TFS Transformer: (i) we design two individual pipelines to enhance both the time-dimension embedding and frequency embedding via self-attention moduels before integrate the two features via an MLP head; and (ii) compared to TFS Transformer and TFP Transformer which fuse the temporal information and frequency information in each transformer block, the two-stream Transformer integrate them till the last step.

IV. Empirical evaluation

A. Experimental Setup

Dataset We evaluate the performance of our proposed models on a audio classification dataset ESC-50 [29]. The ESC-50 dataset is a single-label dataset which consists of 2000 environmental audio recordings. Each audio recording in ESC-50 dataset is 5-second long. There are 50 semantical classes loosely arranged into 5 major categories, including Animals, Natural soundscapes & water sounds, Human, non-speech sounds, Interior/domestic sounds, and Exterior/urban noises. The dataset is a balanced dataset which has 40 examples for each class. We use 5-fold cross-validation for the experiments where the folds are prearranged in the dataset[29].

Training details The spectrogram features are extracted from audio recordings before training the transformer models. For each 5 seconds audio, we extract 128-dimensional log Mel filterbank energy features $\mathbf{Z} \in \mathbb{R}^{128\times500}$ using the setting described in Section III-A. We use SpecAugment [30] for data argumentation. The idea of SpecAugment is using a random length of mask to filter out blocks of frequency channels and time windows. The maximum length of the frequency mask is 24 while the maximum length of frequency mask is 96. The time-dimension embedding $\mathbf{E_t} \in \mathbb{R}^{500\times768}$ and frequency-dimension embedding $\mathbf{E_f} \in \mathbb{R}^{128\times768}$ could be computed by two linear embedding layers based on the masked spectrogram features. Finally, we input embeddings $\mathbf{E_{pt}} \in \mathbb{R}^{501\times768}$ and $\mathbf{E_{pf}} \in \mathbb{R}^{129\times768}$ into our different model architectures.

Notably, most transformer-based audio classification methods are strongly rely on pre-training on ImageNet, e.g., AST [9], and PaSST [19]. Transformer-based method shows a huge

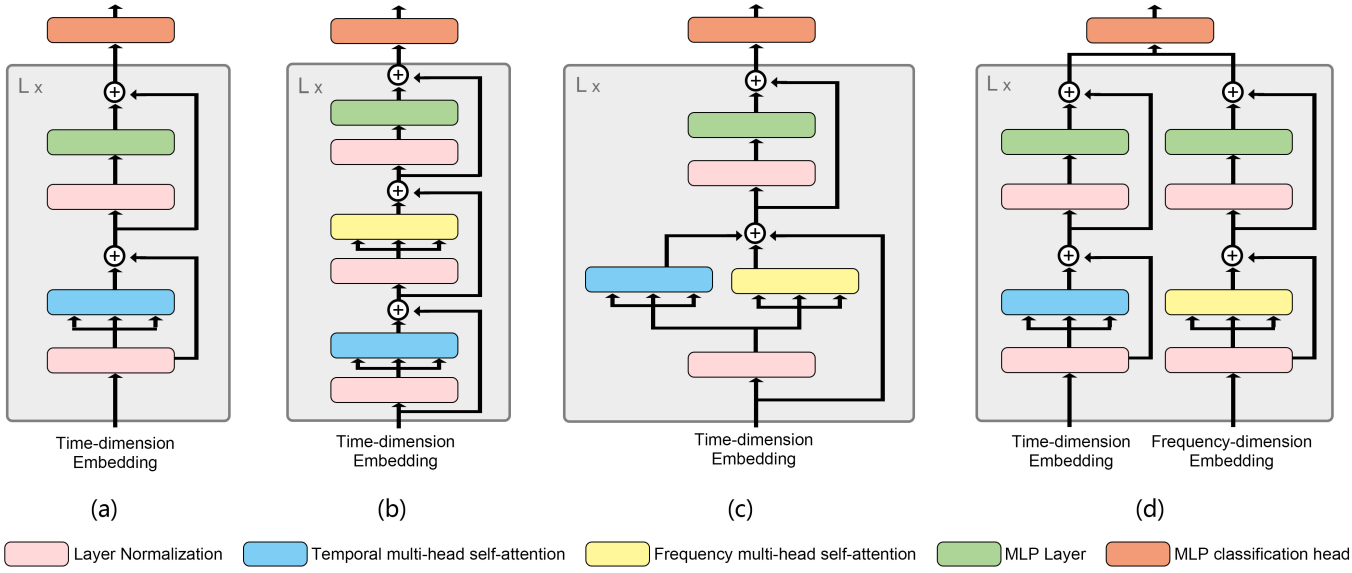

Fig. 3. The 4 Spectrogram Transformer architectures: (a) Temporal Only Transformer, (b) Temporal-Frequency Sequential Transformer, (c) Temporal-Frequency Parallel Transformer, and (d) Two-stream Temporal-Frequency Transformer. L refers to the number of transformer blocks.

gap when pre-traning of ImageNet is not available. In our experiments, we focus on the performance of the transformer model when there is no pre-training model. Thus, all of our models and compared are trained from scratch using random initial weight without any pre-training.

We train our model using batch size 28 with SGD optimizer (momentum 0.9 and weight decay 1e-4) and cross-entropy loss in our experiments. Our model is trained for 50 epochs with an initial learning rate 1e-2 which is decayed to 1e-3 after 30th epoch.

We choose the champion of the ESC-50 leaderboard method AST [9] for comparison. The training code used is from their GitHub repository without any change, except slightly reducing batch size to fit our GPU.

B. Comparison to state-of-the-art methods

We compare our models to the state-of-the-art methods, including kNN, SVM, Convolutional Autoencoder and AST on ESC-50 dataset. Table I shows the results of the benchmark models and ours in terms of Top-1 Accuracy. All of our models outperform these models on the ESC-50 dataset. Specifically, our TSTF transformer achieves 57.24 % on top-1 accuracy which is 11.8 % better compared to the latest AST [9] method.

TABLE I
COMPARISON OF DIFFERENT MODELS PERFORMANCE ON ESC-50 DATASET. THE RESULTS OF KNN, SVM, AND CONVOLUTIONAL AUTOENCODER ARE FROM THE PAPERS [31, 32] DIRECTLY. THE RESULT OF AST IS OBTAINED BY RUNNING THE OFFICIAL CODE [9] USING THE ORIGINAL SETTINGS.

Method	Top-1 Accuracy
KNN [31]	32.20
SVM [31]	39.60
Convolutional Autoencoder [32]	39.90
AST [9]	45.35
Proposed TSTF Transformer	**57.24**

TABLE II
PERFORMANCE COMPARISON OF SPECTROGRAM TRANSFORMER MODELS AND AST ON ESC-50 DATASET. TOP-1 ACCURACY IS PRESENTED.

Model	fold 1	fold 2	fold 3	fold 4	fold 5	Average
AST [9]	45.00	44.50	44.25	48.25	44.75	45.35
TO	56.00	51.25	62.36	61.75	54.50	57.17
TFS	45.75	45.75	52.68	50.25	46.00	48.09
TFP	49.50	47.00	59.13	56.50	52.25	52.88
TSTF	56.25	52.50	61.72	60.25	55.50	**57.24**

C. Ablation studies

We perform an empirical study to understand the performance of our proposed architectures of Spectrogram Transformers when compared to the state-of-the-art transformer-based method AST [9]. The findings are summarised as below:

Model Accuracy We firstly conduct the experiments to evaluate the model accuracy shown in Table II. The Temporal Only (TO) Transformer achieves 57.17% top-1 accuracy on ESC-50 without pre-training which outperforms state-of-the-art method AST by 11.82%. Temporal-Frequency Sequential (TFS) Transformer performs a slightly better result than AST at 48.09%, whereas another temporal-frequency model Temporal-Frequency Parallel (TFP) Transformer shows a higher accuracy at 52.88%. The fourth architecture Two-stream Temporal-Frequency (TSTF) Transformer is the best

variant which achieves 57.24% top-1 accuracy and outperforms state-of-the-art method AST by 11.89%.

From Table II, we find that Temporal-Frequency Serial (TFS) Transformer and Temporal-Frequency Parallel (TFP) Transformer which study frequency information in the transformer block rather than in an individual stream are not perform as good as Two-stream Temporal-Frequency (TSTF) Transformer. This may be caused by the loss of positional information of frequency. In TFS Transformer and TFP Transformer, unlike to temporal dimension, there is no positional embedding for the frequency dimension. The model is difficult to capture the sequential information in the frequency dimension.

TABLE III
COMPARISON OF OUR MODELS WITH AST IN COST. MSA REFERS TO MULTI-HEAD SELF-ATTENTION

Model	Layers	No. of MSA	GFLOPs	Params
AST [9]	12	12	49.40	86.86 M
TO	6	6	21.64	43.05 M
TFS	6	12	27.12	85.56 M
TFP	6	12	27.12	85.56 M
TSTF	6	12	23.46	57.21 M

Model efficiency We observe that our model has a lower inference cost as well as fewer parameters. Table III presents the comparison of model capacity between AST model and our models.

Our Temporal Only (TO) Transformer has the most similar architecture with AST model. We use 6-layer architecture for Temporal Only (TO) Transformer. As the result, the number of parameters of TO Transformer is about 50% of the AST model. We also benefit from time-dimension sampling method which leads less the number of embedding, the FLOPs of TO Transformer is only about 43.8% of the AST model.

Compared with TO Transformer, both TFS Transformer and TFP Transformer have two multi-head self-attentions in the transformer block instead of one. The FLOPs of these Transformers are slightly increased, while the number of parameters nearly doubled.

For Two-stream Temporal-Frequency (TSTF) Transformer, we also use 6-layer architecture. Due to the sequence length of frequency stream 129 is significantly lower than the sequence length of temporal stream 501 and the algorithm complexity of multi-head self-attention is $O(n^2d + nd^2)$, the computational cost of the frequency stream is much less than the temporal stream and the computational cost of 6-layer TSTF Transformer (12 multi-head self-attentions in total) is much less than 12-layer Temporal Only (TO) Transformer.

V. CONCLUSION & FUTURE WORK

In this work, we proposed spectrogram transformers to improve the performance of audio classification. Specifically, we design two sampling mechanisms, including time-dimension sampling and frequency-dimension sampling, and four transformer architectures for the audio classification task. All these architectures achieved state-of-the-art results on ESC-50 dataset when using transformer-based method, with their own distinctive advantages, either on accuracy or efficiency improvement. The main limitation is that there is still a significant performance gap when compared to the transformer with pre-trained model. In future work, we will investigate to integrate pre-trained model in our transformer architectures to improve the accuracy.

ACKNOWLEDGEMENT

The authors would like to thank China Scholarship Council and Loughborough University for supporting his study.

REFERENCES

[1] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for tv baseball programs," in *Proceedings of the eighth ACM international conference on Multimedia*, 2000, pp. 105–115.

[2] N. Almaadeed, M. Asim, S. Al-Maadeed, A. Bouridane, and A. Beghdadi, "Automatic detection and classification of audio events for road surveillance applications," *Sensors*, vol. 18, no. 6, p. 1858, 2018.

[3] R. V. Sharan and T. J. Moir, "An overview of applications and advancements in automatic sound recognition," *Neurocomputing*, vol. 200, pp. 22–34, 2016.

[4] F. Jia, Y. Lei, L. Guo, J. Lin, and S. Xing, "A neural network constructed by deep learning technique and its application to intelligent fault diagnosis of machines," *Neurocomputing*, vol. 272, pp. 619–628, 2018.

[5] D. Stowell, M. Wood, Y. Stylianou, and H. Glotin, "Bird detection in audio: a survey and a challenge," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2016, pp. 1–6.

[6] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *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.

[7] Y. Wu, H. Mao, and Z. Yi, "Audio classification using attention-augmented convolutional neural network," *Knowledge-Based Systems*, vol. 161, pp. 90–100, 2018.

[8] H. Liang and Y. Ma, "Acoustic scene classification using attention-based convolutional neural network," DCASE2019 Challenge, Tech. Rep., Jun. 2019.[Online]. Available: http://dcase ..., Tech. Rep., 2019.

[9] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[10] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 6964–6968.

[11] X. Zhang, Y. Zou, and W. Shi, "Dilated convolution neural network with leakyrelu for environmental sound classification," in *2017 22nd international conference on digital signal processing (DSP)*. IEEE, 2017, pp. 1–5.

[12] Z. Chi, Y. Li, and C. Chen, "Deep convolutional neural network combined with concatenated spectrogram for environmental sound classification," in *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*. IEEE, 2019, pp. 251–254.

[13] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 1128–1132.

[14] Z. Mushtaq and S.-F. Su, "Environmental sound classification using a regularized deep convolutional neural network with data augmentation," *Applied Acoustics*, vol. 167, p. 107389, 2020.

[15] D. M. Agrawal, H. B. Sailor, M. H. Soni, and H. A. Patil, "Novel teo-based gammatone features for environmental sound classification," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 1809–1813.

[16] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Esresnet: Environmental sound classification based on visual domain models," in *2020 25th*

International Conference on Pattern Recognition (ICPR). IEEE, 2021, pp. 4933–4940.
[17] K. Palanisamy, D. Singhania, and A. Yao, "Rethinking cnn models for audio classification," *arXiv preprint arXiv:2007.11154*, 2020.
[18] L. Pepino, P. Riera, and L. Ferrer, "Study of positional encoding approaches for audio spectrogram transformers," *arXiv preprint arXiv:2110.06999*, 2021.
[19] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," *arXiv preprint arXiv:2110.05069*, 2021.
[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.
[21] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time–frequency representations for audio scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 142–153, 2014.
[22] F. Lieb and H.-G. Stark, "Audio inpainting: Evaluation of time-frequency representations and structured sparsity approaches," *Signal Processing*, vol. 153, pp. 291–299, 2018.
[23] M. Huzaifah, "Comparison of time-frequency representations for environmental sound classification using convolutional neural networks," *arXiv preprint arXiv:1706.07156*, 2017.
[24] J. J. Huang and J. J. A. Leanos, "Aclnet: efficient end-to-end audio classification cnn," *arXiv preprint arXiv:1811.06669*, 2018.
[25] A. Kumar and V. K. Ithapu, "Secost:: Sequential co-supervision for large scale weakly labeled audio event detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 666–670.
[26] S. Verbitskiy, V. Berikov, and V. Vyshegorodtsev, "Eranns: Efficient residual audio neural networks for audio pattern recognition," *arXiv preprint arXiv:2106.01621*, 2021.
[27] X. Li, V. Chebiyyam, and K. Kirchhoff, "Multi-stream network with temporal attention for environmental sound classification," *arXiv preprint arXiv:1901.08608*, 2019.
[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
[29] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia*. ACM Press, pp. 1015–1018. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2733373.2806390
[30] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.
[31] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.
[32] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," *Advances in neural information processing systems*, vol. 29, 2016.