ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096513

MULTISCALE AUDIO SPECTROGRAM TRANSFORMER FOR EFFICIENT AUDIO CLASSIFICATION

Wentao Zhu[⋆] *Mohamed Omar*[⋆]

[⋆] Amazon

ABSTRACT

Audio event has a hierarchical architecture in both time and frequency and can be grouped together to construct more abstract semantic audio classes. In this work, we develop a multiscale audio spectrogram Transformer (MAST) that employs hierarchical representation learning for efficient audio classification. Specifically, MAST employs one-dimensional (and two-dimensional) pooling operators along the time (and frequency domains) in different stages, and progressively reduces the number of tokens and increases the feature dimensions. MAST significantly outperforms AST [1] by 22.2%, 4.4% and 4.7% on Kinetics-Sounds, Epic-Kitchens-100 and VGGSound in terms of the top-1 accuracy without external training data. On the downloaded AudioSet dataset, which has over 20% missing audios, MAST also achieves slightly better accuracy than AST. In addition, MAST is 5× more efficient in terms of multiply-accumulates (MACs) with 42% reduction in the number of parameters compared to AST. Through clustering metrics and visualizations, we demonstrate that the proposed MAST can learn semantically more separable feature representations from audio signals.

Index Terms— Audio event classification, audio Transformer, multiscale audio Transformer

1. INTRODUCTION

Audio classification has many applications, such as speaker recognition [2], event, emotion and intent classification [3, 1]. The audio classification has been improved from manually designed feature based approaches [4, 5] and hidden Markov models (HMM) [6] to deep learning based end-to-end solutions [7, 8, 9]. Among these deep learning models, convolutional neural networks have become a *de facto* standard component to model various fixed lengths of dependencies along the time dimension for audio classification [10, 11]. Recently, pure self-attention based deep learning architectures, without convolutional network's inductive bias, *i.e.*, spatial locality and translation equivariance, have outperformed conventional convolutional networks on audio classification [1, 12].

On the other hand, audio can be efficiently perceived in a hierarchical structure [13, 14], *e.g.*, from each individual audio sample to audio activities and semantic audio classes. In the convolutional networks, the hierarchical feature learning can be achieved through various dilation rates and pooling strategies along the time dimension [14]. To the best of our knowledge, there is no multiscale pure-Transformer architecture for audio classification, which can be utilized to learn hierarchical audio representations.

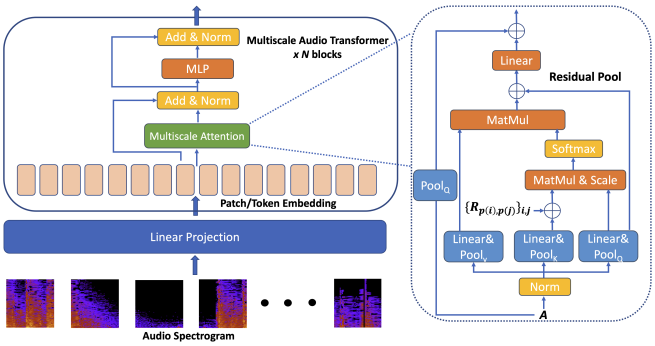

Fig. 1. One block of multiscale audio spectrogram Transformer (MAST). The pooling operator in the block permits to construct representations from dense to coarse resolution and is able to effectively learn hierarchical audio representations.

In this work, we design a multiscale audio spectrogram transformer (MAST) which processes the audio spectrogram for audio classification. We compare our MAST's architecture with widely used AST [1] explicitly in table 1 of section 3. MAST utilizes a multiscale architecture, which is efficient and yields a discriminative representation for audio classification. MAST outperforms AST by a large margin on Kinetics-Sounds [15, 16], Epic-Kitchens-100 [17, 18, 19] and VGGSound [20]. MAST also achieves slightly better accuracy than AST on the downloaded AudioSet [21]. Moreover, MAST is **5**× more efficient based on the number of MACs with only 58% parameter numbers than AST. Through UMAP [22] visualization and clustering metric discussion based on representations, we demonstrate that MAST can learn semantically more separable representations from audio signals. The efficient and light weighted MAST can be an essential component for multimodal architecture design and a strong baseline for audio representation learning.

Table 1. Architecture comparison between AST and MAST on AudioSet [21]. MAST employs multiscale representation learning and uses 58% of the number of AST parameters and 24% MACs of AST.

	AST		MAST	
Block	Feature	Arch./Param.	Feature	Arch./Param.
Input	1×128×1024	0	1×128×1024	0
Patch Embed.	768×(1212=12×101)	768×16×16×1	96×(8192=32×256)	96×7×7×1
Block {0, 1}	768×1212	Attn-MLP	96×8192	Attn-MLP
Block 2	768×1212	Attn-MLP	192×(2048=16×128)	MMSA-MLP
Block {3, 4}	768×1212	Attn-MLP	192×2048	Attn-MLP
Block 5	768×1212	Attn-MLP	384×(512=8×64)	MMSA-MLP
Block {6,···,11}	768×1212	Attn-MLP	384×512	Attn-MLP
Block 12	527	527×768	384×512	Attn-MLP
Block {13,···20}	-	-	384×512	Attn-MLP
Block 21	-	-	768×(256=8×32)	MMSA-MLP
Block {22,23}	-	-	768×256	Attn-MLP
Block 24	-	-	527	527×768

2. RELATED WORK

With large scale and realistic datasets, *e.g.*, AudioSet [21], advanced network architectures have been adopted for audio classification including convolutional neural networks [23, 24], convolutional-attention networks [25, 26], and recent pure-attention based networks [1, 27]. Particularly, AST [1] outperforms previous state-of-the-art audio classification approaches, and obtains widely adoption in many tasks, *e.g.*, multimodal event classification [28] and video retrieval [29]. CMKD [30] further designs cross-modal knowledge distillation between convolutional networks and AST for audio classification. SSAST [12] conducts masked spectrogram patch modeling in self-supervised learning to reduce the need for large amount of labeled data.

There are several hierarchical Transformers for efficient language processing and computer vision. In language processing, Funnel-Transformer [31] gradually compresses the sequence of hidden states to a shorter one and hence reduces the computation cost. Swin Transformer [32] designs a shifted window strategy in an image Transformer. PVT [33] uses a progressive shrinking pyramid to reduce the computations of large feature maps for dense prediction tasks, *e.g.*, object detection and semantic segmentation. Multiscale Transformers [34, 35] adopts several channel-resolution scale stages and hierarchically expands the channel capacity while reducing the spatial resolution. We design a multiscale audio Transformer with one-dimensional and two-dimensional pooling operators along the time dimension and frequency dimension in audio spectrogram for audio classification, which achieves better accuracy than AST with much more efficient number of parameters and MACs.

3. MULTISCALE AUDIO TRANSFORMER

We can perceive an audio sequence in a hierarchical structure, from one signal value at each sampling time point to audio activities and an audio classification category for the whole sequence. Therefore, hierarchical representational learning from an audio spectrogram, which progressively reduces the temporal length and increases the channel dimensions, improves audio-based action recognition. We construct a multiscale audio spectrogram Transformer (MAST) with an audio spectrogram $X \in \mathbb{R}^{h\times T}$ as input, where h is the number of triangular mel-frequency bins, and T is the temporal length. The multiscale audio spectrogram Transformer (MAST) is illustrated in Fig. 1. After the patch embedding, which can be a convolutional block in table 1, conducted in the audio spectrogram, we obtain the embedding token matrix $A \in \mathbb{R}^{d\times N}$, where d is the embedding dimension and N is the number of tokens. One block of MAST can be a stack of multihead multiscale self-attention (MMSA), layer normalization (LN) and multilayer perceptron (MLP)

$$\begin{aligned} A' &= \text{MMSA}(\text{LN}(A)) + \mathcal{P}(A), \\ \text{Block}(A) &= \text{MLP}(\text{LN}(A')) + A', \end{aligned} \tag{1}$$

where $\mathcal{P}$ is a pooling operator, which can be a one-dimensional pooling along the time dimension or a two-dimensional pooling along both the time and frequency dimensions. One head

in multihead multiscale self-attention [35] (MSAttn) can be

$$Q = \mathcal{P}_Q(AW_Q),\ K = \mathcal{P}_K(AW_K),\ V = \mathcal{P}_V(AW_V),$$
$$\text{MSAttn}(A) = Q + \text{Softmax}((QK^T + E^{(rel)})/\sqrt{d})V, \quad (2)$$

where $E^{(rel)}_{ij} = Q_i \cdot R_{p(i),p(j)} = Q_i \cdot (R^t_{t(i),t(j)} + R^f_{f(i),f(j)})$, R^t and R^f are positional embeddings along the temporal and feature axes in the spectrogram.

The multihead multiscale self-attention (MMSA) can be stacked to construct the multiscale audio spectrogram Transformer (MAST) for audio classification. To explicitly demonstrate the details of network architecture, we list and compare the networks of AST and MAST in table 1. In block 21, the pooling is one-dimensional and conducted on the time dimension to retain eight dimensions of spectrogram features, compared with 12 dimensions in AST. MAST employs fewer number of feature dimensions than AST in the first 21 blocks, and it utilizes fewer number of tokens than AST after the 5th block. The multiscale design leads to fewer number of parameters and MACs of MAST than AST. We also experiment other multiscale pooling schedules and strategies, and we find the design in table 1 yields the best accuracy in section 4.

Compared with the previous audio spectrogram Transformer [1], MAST can efficiently extract representation that effectively models hierarchical characteristics of audio signals. In section 4, we demonstrate that MAST significantly reduces the number of parameters and MACs. The efficient MAST is light-weighted and can be used as a component in multimodal networks.

4. EXPERIMENTAL RESULTS

We experiment with four audio event classification datasets – Kinetics-Sounds [15, 16], Epic-Kitchens-100 [17, 18, 19], VGGSound [20] and AudioSet [21].

Kinetics-Sounds is a commonly used subset of Kinetics [16], which consists of 10-second audios from YouTube. As Kinetics-400 is a dynamic dataset and audios may be removed from YouTube, we follow the dataset collection protocol in Xiao *et al.* [36], and we collect 22,914 valid training audios and 1,585 valid test audios.

Epic-Kitchens-100 consists of 90,000 variable length egocentric clips spanning 100 hours capturing daily kitchen activities. The dataset formulates each action into a verb and a noun. We employ two classification heads, one for verb classification and the other one for noun classification.

VGGSound is a large scale action recognition dataset, which consists of about 200K 10-second clips and 309 categories ranging from human actions and sound-emitting objects to human-object interactions. Like other YouTube datasets, *e.g.*, AudioSet [21], some audios are no longer available. After removing invalid audios, we collect 159,223 valid training audios and 12,790 valid test audios.

Table 2. Comparison to state of the art on Kinetics-Sounds. We report top-1 and top-5 classification accuracy.

Models	Top-1	Top-5
AST [28]	52.6	71.5
MAST (Ours)	**74.8**	**93.1**

AudioSet [21] is another YouTube dataset, which consists of almost 2 million 10-second video clips annotated with 527 classes. After removing invalid audios, this gives us 15,818 audios for the test set, which misses 23% audios compared with 20,372 audios in the original test set, 17,823 audios for the balanced training set, which misses 20% audios compared with 22,162 audios in the original balanced training set, and 1,592,753 audios for the full training set, which misses about 25% audios compared with the original 2M unbalanced training set. Over 20% audios are missing on both training and test sets, and rerun the experiments on the downloaded AudioSet based on the official code of AST (https://github.com/YuanGongND/ast) is necessary for a fair comparison. We train MAST with a binary cross-entropy (BCE) loss and report mean average precision (mAP) over all classes for multi-label classification.

Table 3. Comparison to state of the art on VGGSound [20].

Models	Top-1	Top-5
Chen *et al.* [20]	48.8	76.5
AudioSlowFast [37]	50.1	77.9
AST [28]	52.3	78.1
MAST (Ours)	**57.0**	**81.3**

Table 4. Comparison to state of the art on Epic-Kitchens-100.

Models	Verb	Noun	Action
Damen *et al.* [17]	42.1	21.5	14.8
AudioSlowFast [37]	46.5	22.8	15.4
AST [28]	44.3	22.4	13.0
MAST (Ours)	**50.1**	**24.2**	**17.4**

For hyperparameters in MAST, we follow MViTv2-B [35] and use ImageNet-1K publicly available pretrained weights. AdamW [41] is used in the backpropagation and the learning

Table 5. Comparison to state of the art on AudioSet (single model) based on mAP. Our AS denotes evaluation on the downloaded AudioSet. MACs (G), #Params (million).

Models	Balanced	Full	MACs	#Params
Baseline [21]	-	31.4	-	-
PANNs [23]	27.8	43.9	-	-
PSLA [25]	31.9	44.4	-	-
AST [1]	34.7	45.9	103.4	88.1
MBT (AST) [28]	31.3	44.3	103.4	88.1
AST [1] (Our AS)	31.3	38.9	103.4	88.1
MAST (Ours)	**31.4**	**39.0**	**25.6**	**51.3**

Table 6. Statistical metrics for representations of all categories on VGGSound test set.

Metrics	AST [1]	MAST (Ours)
Average silhouette [38]	0.343	**0.527**
Adjusted rand index [39]	0.292	**0.375**
Homogeneity score [40]	0.695	**0.720**

rate is set as 0.00001 with cosine annealing schedule [42]. The numbers of epochs are set as 300, 100, 50, 50 and 10 for Kinetics-Sounds, Epic-Kitchens-100, VGGSound, AudioSet blanced and full sets, respectively. We employ the code of AST to calculate the number of parameters, and ptflops library (https://pypi.org/project/ptflops/) to calculate the number of multiply-accumulates (MACs). Note that one MAC is roughly equal to two floating point operations (FLOPs).

We compare MAST with the previous state-of-the-art single models on the four datasets in table 2, 3, 4, 5. Best scores are in **bold** face. MAST outperforms AST on all the four datasets, with only 24% MACs and 58% parameter numbers as those of AST. Specifically, MAST outperforms AST by 22.2%, 4.7% and 4.4% based on the top-1 accuracy on Kinetics-Sounds, VGGSound and Epic-Kitchens-100 datasets, respectively. Because the downloaded AudioSet has much fewer training samples as the used dataset of AST [1], we rerun the AST using the official code and use (Our AS) to denote the difference. On the downloaded AudioSet, MAST with much fewer MACs and number of parameters achieves slightly better mAP than AST.

We conduct ablation study $w.r.t.$ the number of pooling operators and pooling strategies on the balanced AudioSet. To compare the architecture without 1D pooling, we employ the 2D pooling in the block 21 and obtain 29.7%, which is probably because we reduce the dimension along the triangular mel-frequency bins dimension too much. We also try to use 2 pooling operators and remove the pooling in the block 21, which achieve 28.8% mAP. We only retain the first pooling and remove the rest two pooling operators, and obtain 28.8% mAP. Without multiscale pooling, the architecture achieves 28.0% mAP on balanced AudioSet. The accuracy gap between fewer numbers of pooling operators and MAST is probably because the multiscale pooling benefits the learning of MAST for audio classification.

Fig. 2. UMAP [22] visualizations for test sample representations of the first 30 classes on VGGSound from AST (Left) and MAST (Right). MAST extracts semantically more separable representations than AST from the circled classes.

To further understand the representations learned from MAST and AST, we employ UMAP [22] to visualize the classification tokens in the second last layer. To clearly visualize the representations, we only use the test set of the first 30 classes in VGGSound. For UMAP, we use the default hyperparameters, *i.e.*, the number of neighbors of 15 and the minimal distance of 0.1. From Fig. 2, MAST can learn semantically more separable representations than AST from the circled categories. We further calculate the statistic metric based on the clustering for the representations on the VGGSound full test set in table 6. We utilize average silhouette [38], adjusted rand index [39] and homogeneity score [40] in Scikit-learn package, and MAST achieves the best scores based on all the three metrics. Our MAST learns a compact and discriminative representation.

5. CONCLUSION

In this work, we have presented a multiscale Transformer for audio classification, named multiscale audio spectrum Transformer (MAST). MAST learns hierarchical representations from dense and simple to coarse and complex. It outperforms AST by a large margin on Kinetics-Sounds, Epic-Kitechens-100 and VGGSound. On the downloaded AudioSet, MAST achieves a slightly better mAP than AST, with only 24% MACs and 58% parameter numbers. MAST is efficient, light-weighted and high accurate, and it can be utilized as an essential building component for other applications, *e.g.*, multimodal classification.

6. REFERENCES

[1] Yuan Gong, Yu-An Chung, and James Glass, "AST: Audio spectrogram transformer," in *Proc. Interspeech*, 2021.

[2] Wentao Zhu et al., "Speechnas: Towards better trade-off between latency and accuracy for large-scale speaker verification," in *ASRU*. IEEE, 2021, pp. 1102–1109.

[3] Pengcheng Li et al., "An attention pooling based representation learning method for speech emotion recognition," *Proc. Interspeech*, 2018.

[4] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *ACM MM*, 2013.

[5] Björn Schuller et al., "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. Interspeech*, 2013.

[6] Jeffrey P Woodard, "Modeling and classification of natural sounds by product code hidden markov models," *IEEE Transactions on signal processing*, 1992.

[7] Navdeep Jaitly and Geoffrey Hinton, "Learning a better representation of speech soundwaves using restricted boltzmann machines," in *ICASSP*. IEEE, 2011, pp. 5884–5887.

[8] Sander Dieleman and Benjamin Schrauwen, "End-to-end learning for music audio," in *ICASSP*. IEEE, 2014.

[9] George Trigeorgis et al., "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *ICASSP*. IEEE, 2016, pp. 5200–5204.

[10] Yann LeCun, Yoshua Bengio, et al., "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, pp. 1995, 1995.

[11] Shawn Hershey et al., "Cnn architectures for large-scale audio classification," in *ICASSP*, 2017.

[12] Yuan Gong et al., "SSAST: Self-supervised audio spectrogram transformer," in *Proc. AAAI*, 2022.

[13] Sander Dieleman and Benjamin Schrauwen, "Multiscale approaches to music audio feature learning," in *ISMIR*, 2013.

[14] David Snyder et al., "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*. IEEE, 2018.

[15] Relja Arandjelovic and Andrew Zisserman, "Look, listen and learn," in *Proc. ICCV*, 2017, pp. 609–617.

[16] Will Kay et al., "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[17] Dima Damen et al., "Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100," *IJCV*, 2021.

[18] Dima Damen et al., "Scaling Egocentric Vision: The EPIC-KITCHENS Dataset," in *ECCV*, 2018.

[19] Dima Damen et al., "The EPIC-KITCHENS Dataset: Collection, Challenges and Baselines," *IEEE TPAMI*, 2021.

[20] Honglie Chen et al., "VGGSound: A large-scale audio-visual dataset," in *ICASSP*. IEEE, 2020, pp. 721–725.

[21] Jort F Gemmeke et al., "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*. IEEE, 2017.

[22] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger, "UMAP: Uniform Manifold Approximation and Projection," *Journal of Open Source Software*, 2018.

[23] Qiuqiang Kong et al., "PANNS: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM TASLP*, 2020.

[24] Yun Wang, Juncheng Li, and Florian Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *ICASSP*. IEEE, 2019.

[25] Yuan Gong, Yu-An Chung, and James Glass, "PSLA: Improving audio tagging with pretraining, sampling, labeling, and aggregation," *IEEE/ACM TASLP*, vol. 29, pp. 3292–3306, 2021.

[26] Qiuqiang Kong et al., "Sound event detection of weakly labelled data with cnn-transformer and automatic threshold optimization," *IEEE/ACM TASLP*, vol. 28, pp. 2450–2460, 2020.

[27] Ke Chen et al., "HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection," in *ICASSP*. IEEE, 2022.

[28] Arsha Nagrani et al., "Attention bottlenecks for multimodal fusion," in *NeurIPS*, 2021, vol. 34.

[29] Yan-Bo Lin et al., "Eclipse: Efficient long-range video retrieval using sight and sound," in *Proc. ECCV*, 2022.

[30] Yuan Gong, Sameer Khurana, Andrew Rouditchenko, and James Glass, "CMKD: CNN/Transformer-Based Cross-Model Knowledge Distillation for Audio Classification," *arXiv preprint arXiv:2203.06760*, 2022.

[31] Zihang Dai, Guokun Lai, Yiming Yang, and Quoc Le, "Funnel-transformer: Filtering out sequential redundancy for efficient language processing," in *Proc. NeurIPS*, 2020.

[32] Ze Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. ICCV*, 2021.

[33] Wenhai Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Prof. ICCV*, 2021.

[34] Haoqi Fan et al., "Multiscale vision transformers," in *Proc. ICCV*, 2021.

[35] Yanghao Li et al., "Improved multiscale vision transformers for classification and detection," in *Proc. CVPR*, 2022.

[36] Fanyi Xiao et al., "Audiovisual slowfast networks for video recognition," *arXiv preprint arXiv:2001.08740*, 2020.

[37] Evangelos Kazakos et al., "Slow-fast auditory streams for audio recognition," in *ICASSP*. IEEE, 2021, pp. 855–859.

[38] Peter J Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[39] Lawrence Hubert and Phipps Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

[40] Andrew Rosenberg and Julia Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure," in *Proc. EMNLP-CoNLL*, 2007, pp. 410–420.

[41] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *ICLR*, 2018.

[42] Ilya Loshchilov and Frank Hutter, "Sgdr: Stochastic gradient descent with warm restarts," in *ICLR*, 2017.