MULTI-SCALE TRANSFORMER LANGUAGE MODELING FOR MUSIC CLASSIFICATION TASKS

Anonymous Authors
Anonymous Affiliations
anonymous@ismir.net

ABSTRACT

Most large-scale audio classification models process two-dimensional spectral data with convolutional neural network (CNN) or Vision Transformer (ViT) architectures, inheriting a vision inductive bias misaligned with the temporal nature of audio. While neural audio codecs offer a promising alternative by providing discrete, time-native representations, they produce sequences thousands of tokens long, rendering the usage of standard Transformer architectures computationally expensive. In this study, we present Mega-AudioFormer, a multi-scale Transformer-based model pre-trained from scratch on AudioSet with masked codec-token modeling and fine-tuned on music classification tasks. Our architecture features a global encoder over channel-packed sequences for efficient long-range context, augmented by a local encoder for fine-grained detail. This design confers a key advantage: decode-free inference directly in the compressed domain. Promising performance on music genre recognition (GTZAN), instrument classification (NSYNTH), and speech/music discrimination validates our approach. This work establishes a scalable and effective new direction for audio foundation models and is explicitly designed to leverage advancements from pre-trained language models.

1. INTRODUCTION

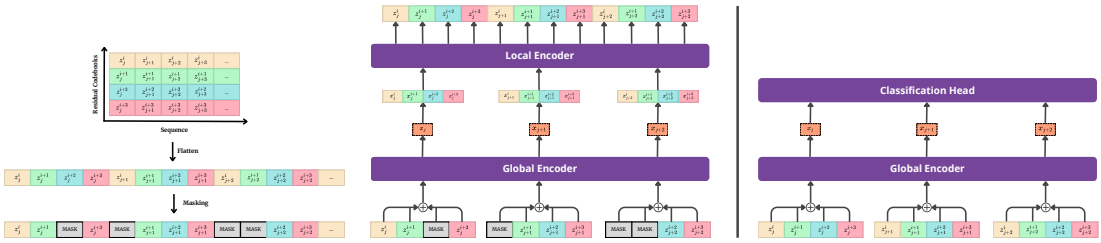
The success of Large Language Models (LLMs) has sparked a paradigm shift across domains, inspiring a move towards unified, sequence-based foundation models in audio processing. This transition is enabled by neural audio codecs like Encodec [1] and SoundStream [2], which use Residual Vector Quantization (RVQ) to convert continuous waveforms into discrete token sequences. This approach unlocks the power of Transformer architectures [3] to model audio as a temporally ordered language, moving beyond the vision-centric inductive biases of earlier models that relied on two-dimensional representations. However, this method confronts a critical bottleneck: the quadratic complexity of self-attention. Since even a few seconds of audio can yield thousands of tokens, the computational and memory demands become prohibitive for standard Transformers, severely hindering their ability to capture the long-range temporal dependencies essential for complex audio understanding.

We address this scalability challenge with Mega-AudioFormer, a multi-scale Transformer designed for efficient audio modeling and inspired by two other works—MegaByte [4], which introduced a multi-scale architecture for modeling million-byte sequences, and UniAudio [5]—our architecture employs a global encoder over channel-packed tokens to tractably model long-range context. We pre-train the model on AudioSet [6] with a masked token objective and demonstrate its effectiveness by fine-tuning on downstream classification tasks. This work establishes a scalable and effective path for creating time-native audio foundation models that directly leverage advancements from the natural language processing domain.

2. RELATED WORK

Our work builds on recent architectural advancements designed to mitigate the computational cost of modeling long sequences. The key solution we adopt is a multi-scale, hierarchical model pioneered by MegaByte [4] large-scale generative modeling of million-byte sequences and successfully adapted to a universal audio generation by UniAudio [5]. While these foundational works firmly establish the architecture's effectiveness for causal, generative tasks, our work reorients its use. Mega-AudioFormer adapts this proven, efficient structure into a bidirectional model for discriminative classification tasks, investigating its effectiveness for audio understanding and analysis.

Masked Modeling has emerged as a dominant pre-training strategy for learning robust audio representations [7]. Models like MERT [8] have demonstrated the value of masked language model pretraining on a wide range of Music Information Retrieval (MIR) tasks. However, they typically rely on standard Transformer architectures. In a parallel line of research focused on audio representation learning, EnCodecMAE [9] demonstrated the value of codec tokens as a learning target. It employs a Masked Autoencoder (MAE) to reconstruct masked audio segments by predicting their corresponding EnCodec tokens. This self-supervised task forces the model to learn perceptually rich features.



(a) Pre-Training Phase (b) Fine-Tuning Phase

Figure 1. High-level diagram of the Mega-AudioFormer architecture.

Further validating the richness of EnCodec's representations, an alternative line of work explores using the continuous latent representation from the EnCodec encoder, prior to the quantization step, as direct input for classification networks [10]. In contrast, our work investigates the viability of using the discrete codebook indices, an approach that aligns with other state-of-the-art models like MERT and EnCodecMAE, thereby making the audio representation directly compatible with language modeling techniques.

3. EXPERIMENTAL EVALUATION AND PERSPECTIVES

Our proposed model is a two-phase training process as presented in Figure 1. The workflow begins by converting an audio waveform into a sequence of discrete tokens using the Encodec neural audio codec, which yields a multi-channel discrete representation of shape (N_q, T), where N_q is the number of residual vector quantizers and T is the number of temporal steps. This representation is subsequently flattened in a time-first manner to create a single interleaved sequence, which is then partitioned into non-overlapping patches where each patch groups all N_q tokens from a single temporal step.

As shown in Figure 1(a), the pre-training phase uses a dual-encoder structure. A Global Encoder, a 12-layer Transformer with a hidden dimension of 768, processes the sequence of patches. To handle temporal order, this encoder utilizes Rotary Position Embeddings (RoPE) [11]; unlike standard methods that add sinusoidal signals to the input, RoPE injects relative positional information by rotating the key and query vectors within the attention mechanism itself, which is particularly effective for modeling long sequences. This is augmented by a lightweight Local Encoder (a 2-layer Transformer) that also utilizes RoPE to operate on the token-level representations and learn fine-grained patterns within each patch. This complete model is pre-trained from scratch on the AudioSet dataset using a Masked Language Modeling (MLM) objective, where 25% of the input tokens are randomly masked and the model learns to predict them.

For downstream tasks the Local Encoder is discarded, and a single linear classification layer is placed on top of the pre-trained Global Encoder. The final hidden states from the Global Encoder are mean-pooled across the patch dimension to produce a fixed-size vector for classification. We also applied a data augmentation method that dropped a contiguous chunk of the tokens in the sequence.

Our preliminary experimental evaluation covers three tasks: music genre classification on the GTZAN dataset [12], instrument classification on the NSYNTH dataset [13], and a binary speech/music discrimination task [14]. We use Encodec at bitrates of 1.5 and 12 kbps, corresponding to patch sizes of 2 and 16.

Our preliminary results, summarized in Table 1, demonstrate the effectiveness of our pre-training strategy. Fine-tuning the pre-trained Global Encoder consistently and significantly outperforms training from scratch across all tasks, with particularly strong gains in speech/music discrimination and genre classification. The addition of data augmentation also provided a performance boost in some cases, notably on the more complex NSYNTH instrument classification task. Notably, the model pre-trained with a lower bitrate Encodec (1.5 kbps) generally achieved superior or comparable performance to the 12 kbps version. This suggests that a more compressed, lower-token-count representation is not only viable but potentially more effective for these classification tasks, offering significant computational advantages.

Table 1. Mega-AudioFormer Performance on Downstream Tasks. We compare fine-tuning against training from scratch, with and without data augmentation.

Dataset	**Type**	**Acc. (%)**		**F1 (%)**	
		1.5	**12**	**1.5**	**12**
GTZAN	Scratch	69.00	63.00	68.10	60.07
	FT	83.00	79.00	82.50	78.01
	FT w/ Aug.	78.00	79.00	78.00	78.91
NSYNTH	Scratch	28.27	26.03	17.69	14.87
	FT	26.59	27.56	17.47	16.51
	FT w/ Aug.	28.91	27.88	18.45	17.61
Speech/Music	Scratch	84.62	84.62	84.52	84.52
	FT	100.00	100.00	100.00	100.00
	FT w/ Aug.	100.00	100.00	100.00	100.00

For future work, we plan to expand on these promising results. Our primary goal is to scale the pre-training phase by utilizing larger and more diverse audio datasets. Subsequently, we will extend our fine-tuning evaluation to a broader range of MIR benchmarks and downstream tasks to further validate the versatility and scalability of the Mega-AudioFormer architecture. Finally, we plan to also validate the knowledge transfer from a pre-trained language model into the audio domain using our framework.

4. REFERENCES

[1] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[2] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[4] L. Yu, D. Simig, C. Flaherty, A. Aghajanyan, L. Zettlemoyer, and M. Lewis, "Megabyte: Predicting million-byte sequences with multiscale transformers," *Advances in Neural Information Processing Systems*, vol. 36, pp. 78 808–78 823, 2023.

[5] D. Yang, J. Tian, X. Tan, R. Huang, S. Liu, X. Chang, J. Shi, S. Zhao, J. Bian, X. Wu *et al.*, "Uniaudio: An audio foundation model toward universal audio generation," *arXiv preprint arXiv:2310.00704*, 2023.

[6] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An Ontology and Human-Labeled Dataset for Audio Events," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2017, pp. 776–780.

[7] Y. Ma, A. Øland, A. Ragni, B. M. Del Sette, C. Saitis, C. Donahue, C. Lin, C. Plachouras, E. Benetos, E. Shatri *et al.*, "Foundation models for music: A survey," *arXiv preprint arXiv:2408.14340*, 2024.

[8] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Lin, A. Ragni, E. Benetos, N. Gyenge *et al.*, "Mert: Acoustic music understanding model with large-scale self-supervised training," in *International Conference on Learning Representations*, 2024.

[9] L. Pepino, P. Riera, and L. Ferrer, "Encodecmae: Leveraging neural codecs for universal audio representation learning," *arXiv preprint arXiv:2309.07391*, 2023.

[10] J. Perianez-Pascual, J. D. Gutiérrez, L. Escobar-Encinas, Á. Rubio-Largo, and R. Rodriguez-Echeverria, "Beyond spectrograms: Rethinking audio classification from encodec's latent space," *Algorithms*, vol. 18, no. 2, p. 108, 2025.

[11] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, p. 127063, 2024.

[12] 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.

[13] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan, "Neural audio synthesis of musical notes with wavenet autoencoders," in *International conference on machine learning*. PMLR, 2017, pp. 1068–1077.

[14] G. Tzanetakis. (1999) Gtzan music/speech collection. [Online]. Available: http://marsyas.info/index.html