# Multi-scale Transformer Language Modeling for Music

**Alef Iury Ferreira , Lucas Rafael Gris , Luiz Fernando de Araújo Vidal, Arlindo Galvão Filho**

AKCIT Federal University of Goiás

## The Problem and Motivation

- **The Challenge in Audio Analysis:**
  - **Large-scale audio classification** models often process **2D spectral data** using Computer Vision architectures (CNNs or ViTs), which is not ideal for the temporal nature of audio. On the other hand, the use of neural audio codecs, while promising, generates **extremely long token sequences** (thousands of tokens for just a few seconds of audio). This makes the use of standard Transformers **computationally infeasible** due to the quadratic complexity of the self-attention mechanism, hindering the ability to capture long-range temporal dependencies.

## Current Challenges

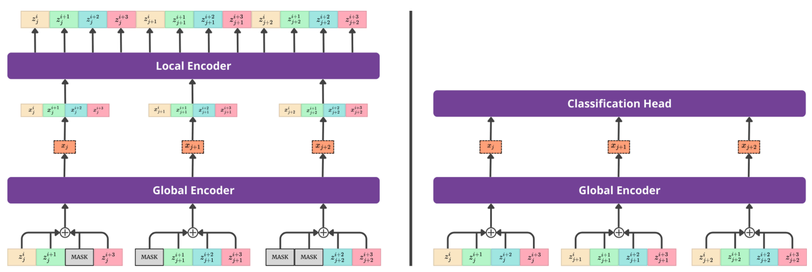
- **High Computational Cost**: The complexity of standard Transformers prevents the efficient processing of long audio sequences.
- **Architectural Bias**: Models adapted from the vision domain are not inherently suited for the sequential nature of audio.
- **Loss of Context**: The difficulty in processing long sequences limits the model's ability to understand complex musical contexts.

## Why this Matters?

- This work presents a **scalable and efficient** solution that treats audio natively as a temporal sequence, **similar to language**. By overcoming computational barriers, we pave the way for creating more powerful and efficient foundation models for audio, capable of understanding everything from fine details to long-term musical structures.

## Mega-AudioFormer

- We introduce **Mega-AudioFormer**, a multi-scale Transformer model pre-trained from scratch on the AudioSet dataset.
  - **Multi-scale Architecture**: It uses a Global encoder to efficiently capture long-range context and a Local encoder to extract short-term details.
  - **Masked Token Modeling**: It learns robust representations by predicting audio tokens that were randomly masked during pre-training.
  - **Efficient Inference**: The design allows for inference directly in the compressed codec domain, eliminating the need for decoding.




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

## Methodology

- **Phase 1: Pre-training (a)**
  - Audio is converted into tokens by a neural codec (Encodec).
  - 25% of the tokens are masked.
  - The complete model (Global and Local encoders) is trained to predict the masked tokens, learning the fundamental structure of audio.
- **Phase 2: Fine-Tuning (b)**
  - **The Local encoder is discarded.**
  - A single classification layer is added on top of the Global encoder.
  - The model is fine-tuned for specific tasks (e.g., genre classification).
- **Experimental Evaluation**
  - **Datasets:** Pre-training on AudioSet. Fine-tuning on GTZAN (music genre), NSYNTH (instruments), and a Speech/Music discrimination dataset.
  - **Metrics:** Accuracy and F1-Score.

## Results

- **Superiority of Fine-Tuning:**
  - Fine-tuning the pre-trained model consistently and significantly outperformed training from scratch in all tasks, validating the effectiveness of the pre-training strategy.
- **Efficiency with Low Bitrate:**
  - The model pre-trained with a more compressed representation (1.5 kbps Encodec) achieved superior or comparable performance to the 12 kbps version. This indicates that a leaner representation is not only viable but potentially more effective, offering enormous computational advantages.

| Dataset | Method | Accuracy (%) | | F1 Score (%) | |
|---|---|---|---|---|---|
| | | 1.5 kbps | 12 kbps | 1.5 kbps | 12 kbps |
| GTZAN (Genre Classification) | Scratch | 69.00 | 63.00 | 68.10 | 60.07 |
| | Fine-tuned | **83.00** | **79.00** | **82.50** | **78.01** |
| | FT + Augmentation | 78.00 | **79.00** | 78.00 | **78.91** |
| NSYNTH (Instrument Classification) | Scratch | 28.27 | 26.03 | 17.69 | 14.87 |
| | Fine-tuned | 26.59 | 27.56 | 17.47 | 16.51 |
| | FT + Augmentation | **28.91** | **27.88** | **18.45** | **17.61** |
| Speech/Music (Binary Classification) | Scratch | 84.62 | 84.62 | 84.52 | 84.52 |
| | Fine-tuned | **100.00** | **100.00** | **100.00** | **100.00** |
| | FT + Augmentation | **100.00** | **100.00** | **100.00** | **100.00** |

## Contributions

- **A scalable and efficient audio model** (Mega-AudioFormer) that handles long temporal sequences.
- Validation of a pre-training strategy that significantly improves performance on music classification tasks.
- Demonstration of the **effectiveness of compressed audio representations** for classification tasks, with computational benefits.