MusicGen Representations for MIR transfer learning

Advika Jayanti, Jaehee Kim, Chris Donahue *(Carnegie Mellon University)*

Introduction

- Music Information Retrieval (MIR) is vital for **analyzing music content** to extract meaningful information used for auto-tagging databases, generating recommendations, and enhancing music systems.

- This study explores how **MusicGen**, a computationally efficient model, leverages music generation insights for MIR tasks, comparing it with the more complex **Jukebox** model.

Methodology

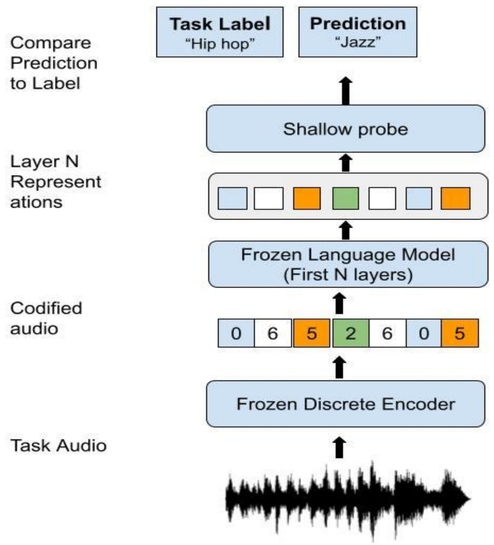
- We pretrain MusicGen with its native language model and encoder to process and encode audio into discrete tokens.

- Codify a high-rate continuous audio signal into lower-rate discrete codes

- We then train a language model on the resulting codified audio and optional metadata:

$$\text{learn } p(\text{codified audio} \mid \text{metadata})$$

- Measure performance of shallow models trained on genre and key classification tasks using musicGen representations as input features.



Results

Approach	Genre	Key	
	GTZAN Acc	Giantsteps Score	Average
(No pre-training) Probing CHROMA	0.328	0.565	0.447
(No pre-training) Probing MFCC	0.448	0.146	0.297
(Tagging) Probing CHOI	0.759	0.131	0.445
(Tagging) Probing MUSICNN	0.790	0.128	0.459
(Contrastive) Probing CLMR	0.686	0.149	0.418
(CALM) Probing JUKEBOX	0.797	0.667	0.732
(CALM) Probing MusicGen	**0.687**	**0.536**	**0.612**
State-of-the-art	0.821	0.796	0.808
Best using any form of pre-training	0.821	0.758	0.790
Best trained from scratch	0.658	0.743	0.701

- MusicGen achieved a **68.7% and 53.6% classification accuracy** on the GTZAN and Giantsteps dataset (**>60%** indicates strong ability in classification; **<80%** considered the goal for optimal representation encoding in MIR tasks).

- Ranked as the **second-best** after Jukebox across other pre-trained MIR models.

Discussion

- Validates **novel** approach to MIR that leverages **inherent capabilities** of music generation models.
- Demonstrates that **MusicGen** can effectively reduce the need for **manual data tagging** using automatic feature extraction.
- Confirms potential for using these **representations in varied MIR tasks** beyond genre classification.
- Highlights the **reduced computational demand**, enhancing scalability and practical application in music analysis.

Future Works

- Plans to compare **audio vs. text conditioning** effects to help identify which conditioning method provides more useful features for MIR tasks.

- Compare effect of hidden layer subset in representation encoding as input

- Run probing tasks on emotion recognition and tagging

Contact
ajayanti@andrew.cmu.edu
jaeheek@andrew.cmu.edu