DMRN+17: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2022

QUEEN MARY UNIVERSITY OF LONDON
TUE 20 DECEMBER 2022

Large-Scale Pretrained Model for Self-Supervised Music Audio Representation Learning

Yizhi Li[1], Ruibin Yuan[2,4*], Ge Zhang[2,5*], Yinghao Ma[3*], Chenghua Lin[1†], Xingran Chen[5], Anton Ragni[1], Hanzhi Yin[4], Zhijie Hu[6], Haoyu He[7], Emmanouil Benetos[3], Norbert Gyenge[1], Ruibo Liu[8], Jie Fu[2†]*

[1]Department of Computer Science, University of Sheffield, UK {yizhi.li, c.lin}@sheffield.ac.uk
[2]Beijing Academy of Artificial Intelligence, China fujie@baai.ac.cn
[3]Centre for Digital Music, Queen Mary University of London, UK yinghao.ma@qmul.ac.uk
[4]School of Music, Carnegie Mellon University, PA, USA
[5]University of Michigan Ann Arbor, USA
[6]HSBC Business School, Peking University, China
[7]University of Tübingen & MPI-IS, Germany
[8]Dartmouth College, NH, USA

***Abstract*— Self-supervised learning technique is an under-explored topic for music audio due to the challenge of designing an appropriate training paradigm. We hence propose MAP-MERT, a large-scale music audio pre-trained model for general music understanding. We achieve performance that is comparable to the state-of-the-art pre-trained model Jukebox using less than 2% of parameters.**

***Index Terms*—** Self-supervised learning, Music representation learning, Music information retrieval

I. INTRODUCTION

Deep learning is undergoing a paradigm shift with the rise of large-scale pre-trained models. In recent years, self-supervised learning (SSL) has achieved significant results in domains like computer vision, natural language processing, and speech processing. SSL leverages large-scale unlabelled data to obtain general representations, which could benefit a wide range of resource-restricted downstream tasks.

Although such a large-scale pre-training paradigm is of potential to improve annotation-limited music information retrieval (MIR) tasks, it is not well-studied in the community. Jukebox, the state-of-the-art SSL model learns music representations by reconstructing the raw audio [1, 2]. But it can barely be fine-tuned or efficiently adapted to more downstream tasks due to the enormous number of 5 billion parameters. To this end, we propose a novel representation learning method for music understanding.

Inspired by HuBERT [4], we obtain discrete pseudo labels by K-Means to conduct the mask prediction pre-training. Apart from only focusing on distinguishing sound textures like HuBERT, we use additional Chroma-based pseudo labels and design a CQT reconstruction task to help Mu-BERT learn the significant pitch information for music tasks. Moreover, we explore and analyse masking strategies and data augmentation techniques appropriate for music audio pre-training. To conclude, the aim and potential innovations of this work include:

1. developing self-supervised methods for music understanding;
2. providing a general music pre-trained model with trainable size; and
3. establishing a user-friendly and extendable MIR benchmark.

Approach	MTT		GTZAN	GS	EMO		Average
	AUC	AP	Acc	Score	$R2_{Arousal}$	$R2_{Valence}$	
CHOI	89.7	36.4	75.9	13.1	67.3	43.4	51.9
MUSICNN	90.6	38.3	79.0	12.8	70.3	46.6	53.7
CLMR	89.4	36.1	68.6	14.9	67.8	45.8	50.8
Music2Vec [3]	89.5	35.9	76.6	50.1	69.4	57.4	63.2
Jukebox (5B)	**91.5**	**41.4**	79.7	66.7	**72.1**	**61.7**	**69.9**
MERT (90M)	90.8	38.4	**80.7**	**67.0**	71.2	52.1	66.7

Table 1: Preliminary Results on MIR Tasks. The baseline results (except Music2Vec) and probing protocol are adopted from JukeMIR [2]. All results are produced with probing settings. Our model with 768-D representations under the probing setting achieves performances comparable to the SOTA Jukebox with 4800-D representations on the auto-tagging, genre classification, key detection and emotion regression tasks.

II. REFERENCES

[1] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *arXiv preprint arXiv:2005.00341*, 2020.

[2] R. Castellon, C. Donahue, and P. Liang, "Codified audio language modeling learns useful representations for music information retrieval," *arXiv preprint arXiv:2107.05677*, 2021.

[3] Y. Li, R. Yuan, G. Zhang, Y. MA, C. Lin, X. Chen, A. Ragni, H. Yin, Z. Hu, H. He, *et al.*, "Map-music2vec: A simple and effective baseline for self-supervised music audio representation learning," in *ISMIR 2022 Hybrid Conference*, 2022.

[4] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

* The authors contributed equally to this work.
† Corresponding authors.