MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training

Yizhi Li[1,2]*, Ruibin Yuan[3,4]*, Ge Zhang[4,5,6]*, Yinghao Ma[7]*, Xingran Chen, Hanzhi Yin[3], Chenghao Xiao[8]
Chenghua Lin[1,2]†, Anton Ragni[2], Emmanouil Benetos[7], Norbert Gyenge[2], Roger Dannenberg[3], Ruibo Liu[9],
Wenhu Chen[5], Gus Xia[10,11], Yemin Shi[6,12], Wenhao Huang[6], Zili Wang, Yike Guo[4], Jie Fu[4,6]†

m-a-p.ai [1]University of Manchester [2]University of Sheffield [3]Carnegie Mellon University [4]Hong Kong University of Science and Technology
[5]University of Waterloo [6]Beijing Academy of Artificial Intelligence [7]Queen Mary University of London [8]Durham University
[9]Google DeepMind [10]MBZUA [11]New York University [12]linksoul.ai

@MM_Art_Project

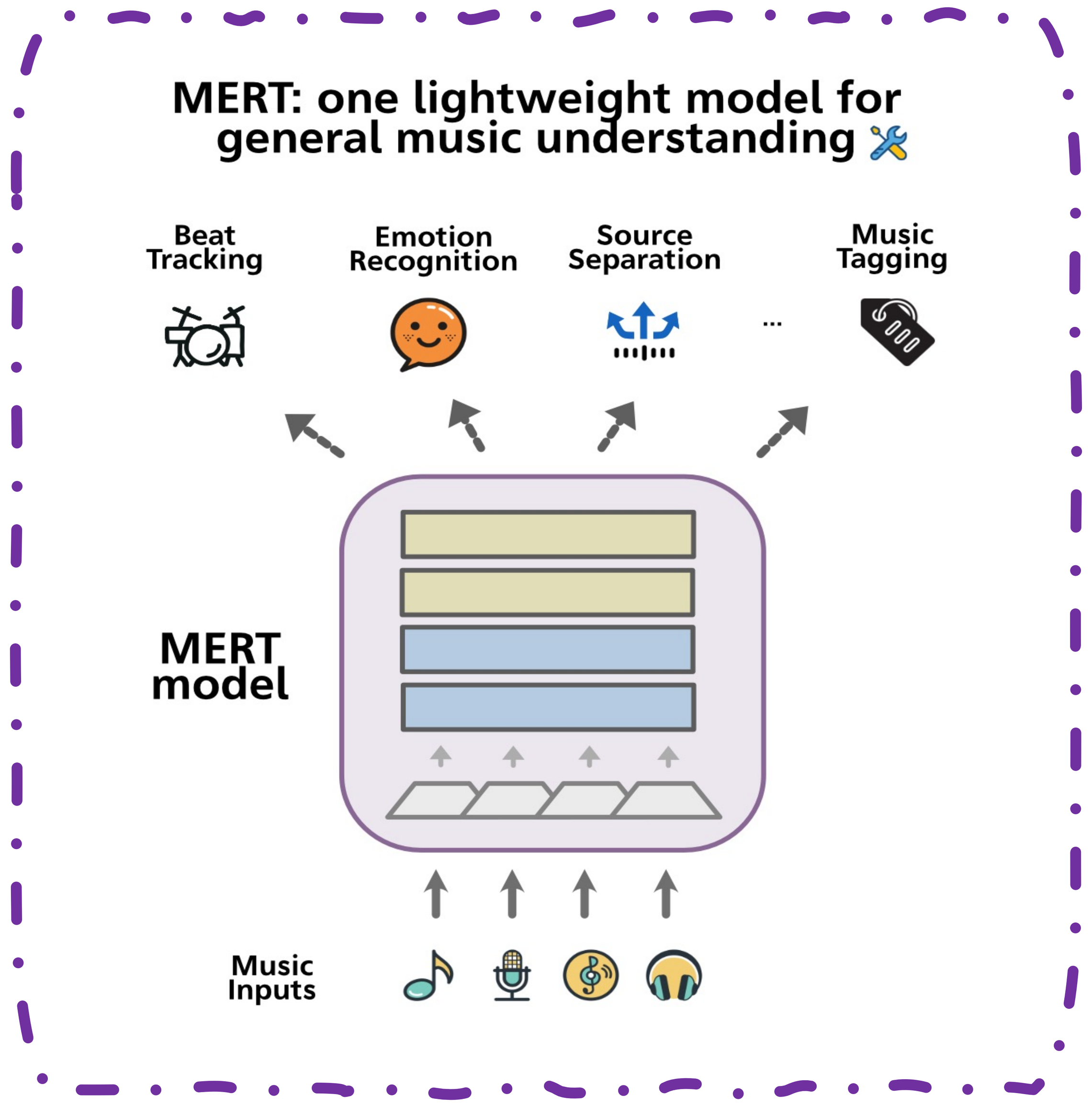

Figure 1: MERT for Multiple Downstream Tasks. We prove that it is possible to inference once to conduct multiple music downstream tasks with ***a single MERT model.***

1. Motivation

- Self-supervised learning is promising for training generalisable models with large-scale data for many domains yet not for music.
- We propose MERT, an **open-source** model incorporating **acoustic and musical** teacher models to provide pseudo labels in the masked language modelling.
- MERT is scaled **from 95M to 330M** parameters and achieve SOTA music understanding performances while remaining efficient (Fig. 1).

*MERT achieve overall SOTA with only **probing** results. We could further **fine-tune** MERT to achieve better performance.*

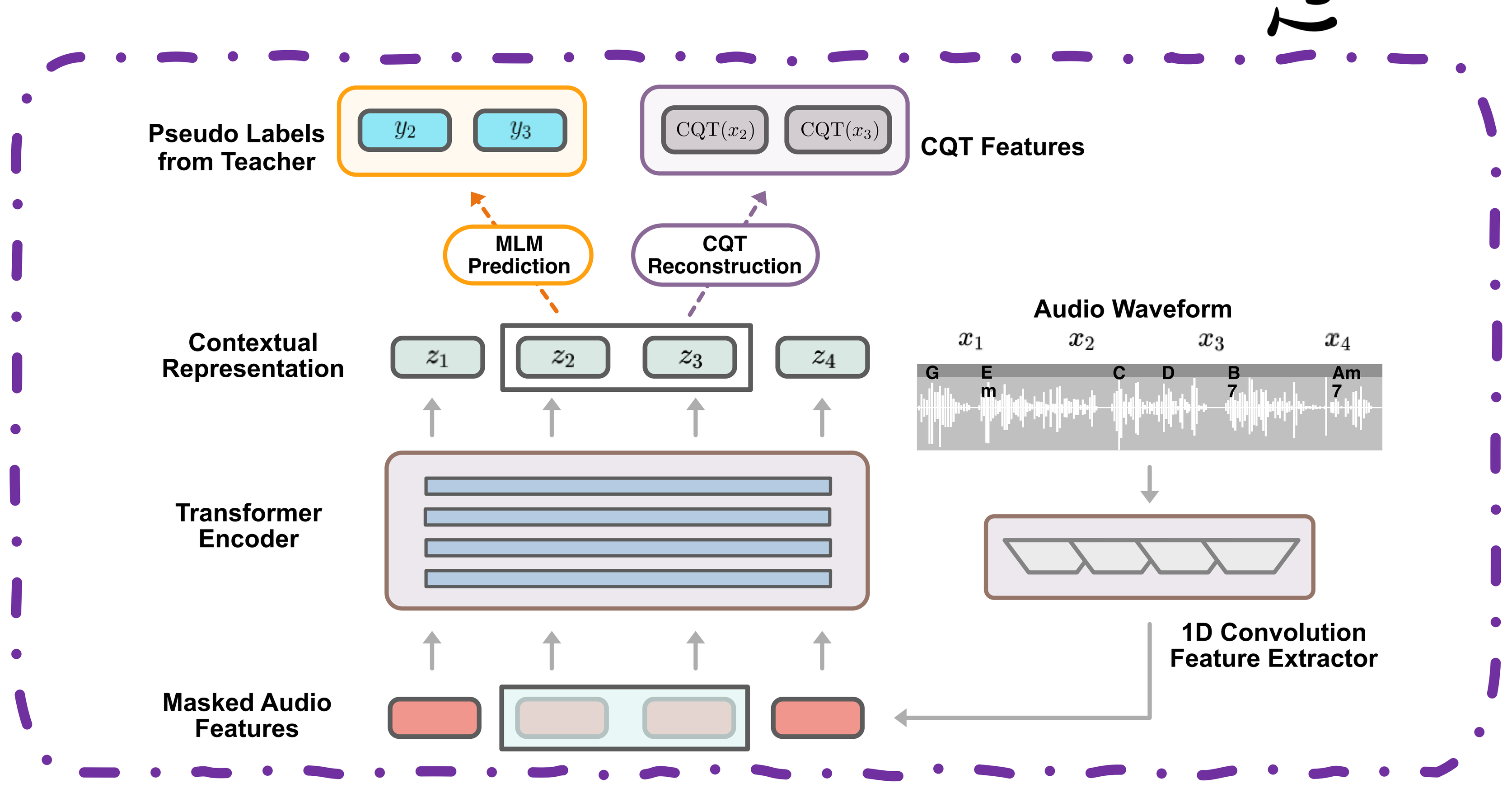

Figure 2: Illustration of the MERT Pre-training Framework.

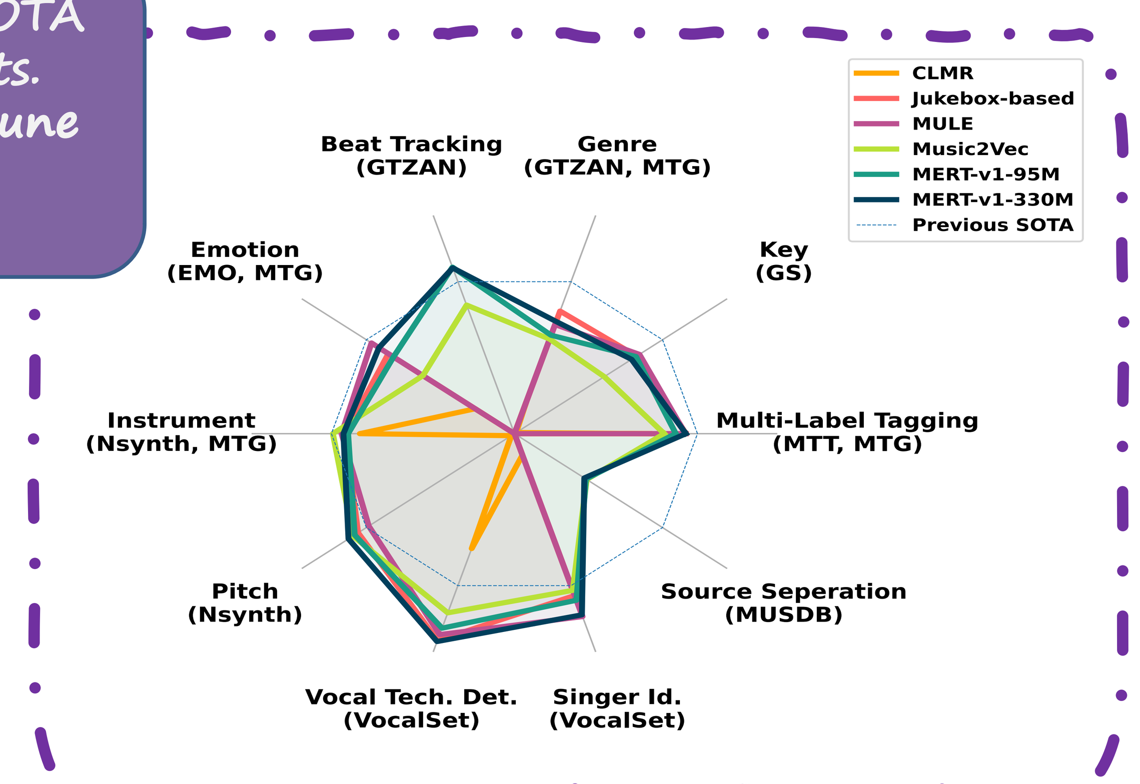

Figure 4: Performance Comparison with Baselines on MARBLE Benchmark.

Dataset Task	MTT Tagging		GS Key	GTZAN Genre	GTZAN Rhythm	EMO Emotion		Nsynth Instrument	Nsynth Pitch	VocalSet Tech	VocalSet Singer
Metrics	ROC	AP	$\mathrm{Acc}^{\mathrm{Refined}}$	Acc	$\mathrm{F1}^{\mathrm{beat}}$	$\mathrm{R2}^{\mathrm{V}}$	$\mathrm{R2}^{\mathrm{A}}$	Acc	Acc	Acc	Acc
MusiCNN [41]	90.6*	38.3*	12.8*	79.0*	-	46.6*	70.3*	72.6	64.1	70.3	57.0
CLMR [48]	89.4*	36.1*	14.9*	68.6*	-	45.8*	67.8*	67.9	47.0	58.1	49.9
Jukebox-5B [15; 57]	91.5*	**41.4***	66.7*	79.7*	-	**61.7***	72.1*	70.4	91.6	76.7	82.6
MULE [36]	91.4*	40.4*	66.7*	73.5*	-	57.7*	70.0*	74.0*	89.2*	75.5	**87.5**
$\mathrm{HuBERT\text{-}base}^{\mathrm{music}}$ [25]	90.2	37.7	14.7	70.0	**88.6**	42.1	66.5	69.3	77.4	65.9	75.3
$\mathrm{data2vec\text{-}base}^{\mathrm{music}}$ [2]	90.0	36.2	50.6	74.1	68.2	52.1	71.0	69.4	93.1	71.1	81.4
$\mathrm{MERT\text{-}95M}^{\mathrm{K\text{-}means}}$	90.6	38.4	65.0	78.6	88.3	52.9	69.9	71.3	92.3	74.6	77.2
$\mathrm{MERT\text{-}95M\text{-}public}^{\mathrm{K\text{-}means}}$	90.7	38.4	67.3	72.8	88.1	59.7	72.5	70.4	92.3	75.6	78.0
$\mathrm{MERT\text{-}95M}^{\mathrm{RVQ\text{-}VAE}}$	91.0	39.3	63.5	78.6	88.3	60.0	**76.4**	70.7	92.6	74.2	83.7
$\mathrm{MERT\text{-}330M}^{\mathrm{RVQ\text{-}VAE}}$	91.3	40.2	65.6	79.3	87.9	61.2	74.7	72.6	**94.4**	**76.9**	87.1
(Previous) SOTA	**92.0** [26]	**41.4** [15]	**74.3 [30]**	**83.5** [36]	80.6 [24]	**61.7**	72.1 [15]	**78.2** [53]	89.2 [36]	65.6 [55]	80.3 [39]

Dataset Task	MTG Instrument		MTG MoodTheme		MTG Genre		MTG Top50		MUSDB Source Seperation				**Avg.**
Metrics	ROC	AP	ROC	AP	ROC	AP	ROC	AP	$\mathrm{SDR}^{\mathrm{vocals}}$	$\mathrm{SDR}^{\mathrm{drums}}$	$\mathrm{SDR}^{\mathrm{bass}}$	$\mathrm{SDR}^{\mathrm{other}}$	
MusiCNN [41]	74.0	17.2	74.0	12.6	86.0	17.5	82.0	27.5	-	-	-	-	-
CLMR [48]	73.5	17.0	73.5	12.6	84.6	16.2	81.3	26.4	-	-	-	-	-
Jukebox-5B [15; 57]	-	-	-	-	-	-	-	-	5.1*	4.9*	4.1*	2.7*	-
MULE [36]	76.6	19.2	78.0	15.4	**88.0**	**20.4**	83.7	30.6	-	-	-	-	-
$\mathrm{HuBERT\text{-}base}^{\mathrm{music}}$ [25]	75.5	17.8	76.0	13.9	86.5	18.0	82.4	28.1	4.7	3.7	1.8	2.1	55.8
$\mathrm{data2vec\text{-}base}^{\mathrm{music}}$ [2]	76.1	19.2	76.7	14.3	87.1	18.8	83.0	29.2	5.5	5.5	4.1	3.0	59.9
$\mathrm{MERT\text{-}95M}^{\mathrm{K\text{-}means}}$	77.2	19.6	75.9	13.7	87.0	18.6	82.8	29.4	5.6	5.6	4.0	3.0	62.9
$\mathrm{MERT\text{-}95M\text{-}public}^{\mathrm{K\text{-}means}}$	77.5	19.6	76.2	13.3	87.2	18.8	83.0	28.9	5.5	5.5	3.7	3.0	63.0
$\mathrm{MERT\text{-}95M}^{\mathrm{RVQ\text{-}VAE}}$	77.5	19.4	76.4	13.4	87.1	18.8	83.0	28.9	5.5	5.5	3.8	3.1	63.7
$\mathrm{MERT\text{-}330M}^{\mathrm{RVQ\text{-}VAE}}$	78.1	19.8	76.5	14.0	86.7	18.6	83.4	29.9	5.3	5.6	3.6	3.0	**64.7**
(Previous) SOTA	**78.8**	**20.2** [1]	**78.6**	**16.1** [36]	87.7	20.3 [1]	**84.3**	**32.1** [36]	**9.3**	**10.8**	**10.4**	**6.4** [44]	64.5

Table 1: Experimental Performances of MERT and Baselines on 14 Downstream Tasks.

3. Pre-training Experiments

- K-Means & RVQ-VAE teachers comparison.
- RVQ codebook ablation study.
- The CQT Musical Loss is effective.

2. Methodology

- We explore an optimal combination of the **teacher models**, which outperforms conventional speech and audio approaches in terms of performance (Fig. 2).
- The combination used for pre-training includes an **acoustic teacher** based on Residual Vector Quantization - Variational AutoEncoder (RVQ-VAE) and a **musical teacher** based on the Constant-Q Transform (CQT).
- We also introduce an in-batch noise mixture augmentation to enhance the representation robustness.
- We explore various settings to **overcome the instability** in acoustic language model pre-training, which allows MERT to scale from 95M to 330M parameters (see Fig. 3).

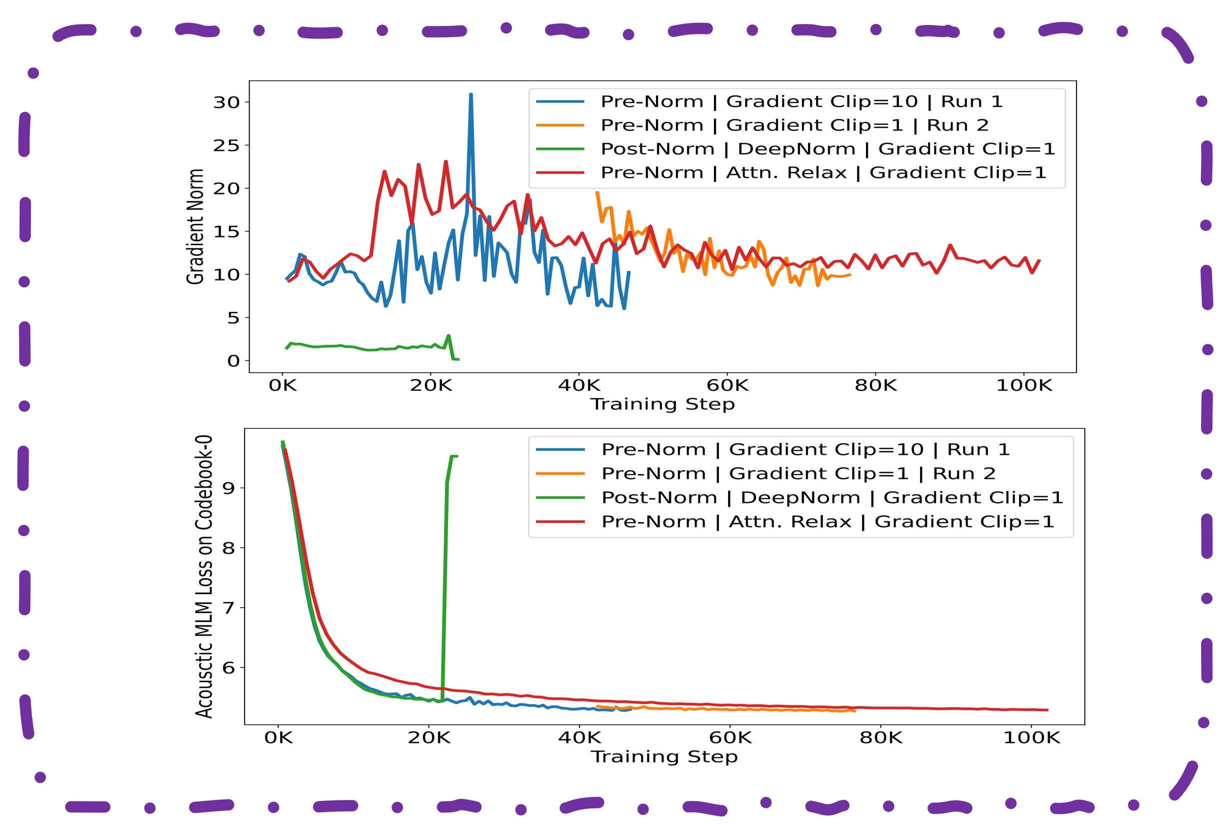

Figure 3: The Gradient Norm and MLM Loss of Different Pre-training Setting.

4. Results

- As suggested by the average scores in Table 1, **MERT-330M** outperforms **the combination of the previous SOTAs** and becomes new SOTA on 4 metrics, while the smaller **MERT-95Ms** still have close performance.
- Generally, MERT models perform well on tasks focusing on **local-level musical information** such as beat, pitch and local timbre such as singer information, and remain competitive on the rest of tasks such as music tagging, key detection, and genre classification, which require more global-level information.
- MERT series models achieve SOTA or comparable performance with only **1.9% (95M) and 6.6% (330M)** parameters compared to the SOTA self-supervised baseline Jukebox-5B.
- Even with probing evaluation, most models could not be trained on **sequence labelling** tasks with affordable computational costs except for MERT-like architectures.

Acknowledgement

This paper is a tribute to our talented friend Anqiao Yang, for his friendship and valuable advice to this work. Yizhi Li is a Ph.D. student fully funded by the Department of Computer Science, University of Manchester, UK. This work is partially funded by Theme based Research Scheme (T45205/21-N), Research Grants Council of Hong Kong. Yinghao Ma is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported by UK Research and Innovation [grant number EP/S022694/1]. Emmanouil Benetos is supported by a RAEng/Leverhulme Trust Research Fellowship [grant number LTRF2223-19-106]. We acknowledge IT Services at The University of Sheffield for the provision of services for High Performance Computing.

Models Codes Paper