LEARNING MUSIC AUDIO REPRESENTATIONS VIA WEAK LANGUAGE SUPERVISION

Ilaria Manco[*†], *Emmanouil Benetos*[*], *Elio Quinton*[†] & *György Fazekas*[*]

[*]School of EECS, Queen Mary University of London, London, U.K.
[†]Music & Audio Machine Learning Lab, Universal Music Group, London, U.K.

ABSTRACT

Audio representations for music information retrieval are typically learned via supervised learning in a task-specific fashion. Although effective at producing state-of-the-art results, this scheme lacks flexibility with respect to the range of applications a model can have and requires extensively annotated datasets. In this work, we pose the question of whether it may be possible to exploit weakly aligned text as the only supervisory signal to learn general-purpose music audio representations. To address this question, we design a multimodal architecture for *music and language pre-training* (MuLaP) optimised via a set of proxy tasks. Weak supervision is provided in the form of noisy natural language descriptions conveying the overall musical content of the track. After pre-training, we transfer the audio backbone of the model to a set of music audio classification and regression tasks. We demonstrate the usefulness of our approach by comparing the performance of audio representations produced by the same audio backbone with different training strategies and show that our pre-training method consistently achieves comparable or higher scores on all tasks and datasets considered. Our experiments also confirm that MuLaP effectively leverages audio-caption pairs to learn representations that are competitive with audio-only and cross-modal self-supervised methods in the literature.

Index Terms— audio and language, multimodal learning, music information retrieval, audio representations

1. INTRODUCTION

With the increasing demands for annotated data and compute required by task-specific models trained on dedicated datasets, pre-training to learn general-purpose and transferable representations is becoming an increasingly important alternative [1]–[3]. Since pre-training is only carried out once and in a task-agnostic way, it allows to solve downstream tasks in a more sample-efficient way. This is particularly crucial in fields like Music Information Retrieval (MIR) where fully annotated datasets are notoriously costly to obtain and difficult to scale, since data is often copyrighted and annotations require expert knowledge [4]. In MIR it has now become relatively common to pre-train convolutional neural networks (CNN) in a supervised fashion on a source task such as auto-tagging and then transfer their representations to downstream tasks [5]–[7]. This approach however still requires fully annotated datasets.

An important but often neglected source of supervision can instead be found in noisy natural language text associated to music audio. Examples of this are music reviews [8] and user-generated descriptions found on the internet or provided in private collections such as production music libraries. In this work, we explore whether this type of noisy companion text, only weakly aligned to the audio, can be used to learn music audio representations. We do so via *music and language pre-training* (MuLaP), a method for multimodal pre-training on audio and language data through which we learn audio representations that can be transferred to diverse MIR tasks. [1] Compared to the supervised pre-training methods mentioned above, MuLaP does not require ad-hoc annotations and strong alignment of text and audio for pre-training. In our setting, the text and audio modalities are only aligned at the track level, without strong correspondence between text tokens and audio frames. For example, a caption may contain the word *guitar* at the beginning of the sentence, but guitar sounds may only appear in a limited number of audio frames in a later section of the track.

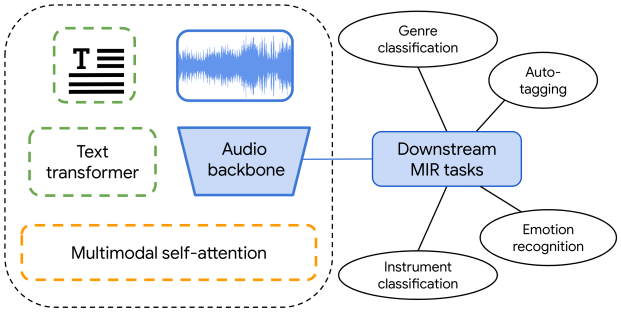


Fig. 1: **Illustration of MuLaP**, our music and language pre-training framework for music information retrieval tasks.

Our work shares a similar goal to recent MIR pre-training methods [9]–[11], but aims to extend the range of downstream tasks that the model can generalise to via multimodal learning. Our study also contributes to a growing body of work on multimodal pre-training, pioneered in computer vision and NLP with large-scale visio-linguistic models [12], [13] and more recently introduced in the machine listening field. To our knowledge, this is the first work on audio-linguistic pre-training for the music domain. Although multimodal learning has yet to see widespread adoption in the field, some MIR literature has begun to explore cross-modal learning on audio-tag inputs. Some notable examples are [14]–[16], which learn audio embeddings through cross-modal contrastive approaches, and [17], which explores multimodal metric learning for music audio retrieval. However, unlike our work, these only use tags and metadata as the text modality. Finally, the idea of using noisy natural language for weak supervision has been explored in [18] for music recommendation and tagging. However our work differs from [18] in several ways: we do not introduce additional supervision in the form of co-listen statistics, we process the text input through a language model instead of using it simply to extract labels, and we use a much smaller training set (by ~ 50 times).

I. Manco is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported jointly by UK Research and Innovation [grant number EP/S022694/1] and Universal Music Group.

[1]Code is available at https://github.com/ilaria-manco/mulap

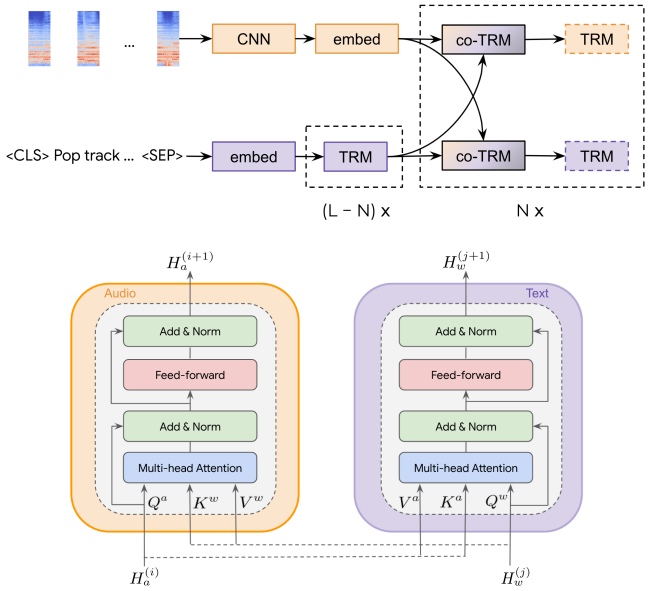


Fig. 2: **Architecture of our multimodal transformer**. Top: audio-caption pairs are first processed through modality-specific layers: CNN and one embedding layer (embed) for the audio sequence, one embed and $(L-N)$ transformer (TRM) layers for the text modality. Bottom: intermediate representations $H^{(i)}$ from each branch are passed through N co-attentional transformer layers (co-TRM).

2. AUDIO AND LANGUAGE PRE-TRAINING

2.1. Architecture

We consider an extension of the Bidirectional Encoder Representations from Transformers (BERT) [19] architecture, based on ViLBERT [12], as our model for audio-linguistic pre-training, adapting its design to the multimodal scenario where the non-text modality is audio. This is primarily motivated by the wide success of transformer-based multimodal approaches for visio-linguistic tasks [12], [13] and by the possibility of initialising part of the network with a pre-trained language model to speed up training. At a high level, our model consists of two branches, operating on text and audio respectively, which interact via co-attentional layers (Fig. 2).

Text and audio branches The text branch closely follows the standard design of BERT: the input text sequence is first tokenized, embedded and passed through $(L-N)$ multi-head self-attention layers consisting of standard transformer blocks. We initialise this branch with pre-trained $\text{BERT}_{\text{BASE}}$ weights [19].

The audio input is first processed by a CNN audio backbone, which operates on short audio clips, obtained by splitting the input audio sequence into T non-overlapping segments. The goal of the audio CNN is to capture short-range dynamics and produce a sequence of d-dimensional local feature vectors $\boldsymbol{a} = \{\boldsymbol{a}_1, ..., \boldsymbol{a}_T\}, \boldsymbol{a}_i \in \mathbb{R}^d$. Similarly to the text input, audio features are also processed through an embedding procedure and then summed to positional embeddings to obtain the final input representations to be passed to the transformer layers, which model long-range and cross-modal dynamics.

In both the audio and text streams, the first element of the embedding sequence, corresponding to the mean-pooled convolutional feature $\mathbf{a}_0$ and the <CLS> token respectively, assume a special role and their final representations, $\mathbf{h}_0^a$ and $\mathbf{h}_0^w$, are taken as a summary of the sequence of each modality.

Co-attentional layers Intermediate audio and text representations are processed by N co-attentional layers [12]. These are given respectively by the output of the $(L-N)$-th transformer layer for the text branch and by the output of the embedding layer for the audio branch. These are identical to standard encoder transformer layers, with the only difference that key and value vectors are exchanged between modalities, as illustrated in Fig. 2. For each query vector $\mathbf{q}_i^\alpha$ of modality α, the output of the attention module A is obtained from key and value matrices K^β and V^β of all the tokens of modality β:

$$A\left(\mathbf{q}_i^\alpha, K^\beta, V^\beta\right) = \text{softmax}\left(\frac{\mathbf{q}_i^\alpha K^\beta}{\sqrt{d_K}}\right) V^\beta, \tag{1}$$

where d_K is the dimension of the key vectors.

2.2. Pre-training objectives

We consider three learning objectives aimed at solving proxy tasks which model intra-modal relationships within the audio and text modalities, and inter-modal relationships between the two. The two intra-modal objectives are designed as extensions of the Masked Language Modelling (MLM) objective commonly used to train language transformers [19].

For the language component, similarly to the standard MLM objective, we swap some of the S text tokens $\boldsymbol{w} = \{w_1, ..., w_S\}$ with a <MASK> token and task the model with predicting these based on the unmasked input. The associated loss function is:

$$\mathcal{L}_{\text{MLM}} = -\mathbb{E}_{(\boldsymbol{w},\boldsymbol{a})\sim\mathcal{D}} \log p_\theta^w \left(w_m \mid \boldsymbol{w}_{\setminus m}, \boldsymbol{a}\right), \tag{2}$$

where w_m and $\boldsymbol{w}_{\setminus m}$ are masked and unmasked text tokens sampled from the training set $\mathcal{D}$, and p_θ^w the probability distribution over the vocabulary, estimated by the transformer parametrised by θ.

For the audio component, we adopt an equivalent of MLM, which we refer to as Masked Audio Modelling (MAM), where instead of text tokens, we mask a subset of the feature vectors in the audio sequence $\boldsymbol{a}$ by replacing them with zeros. The network is then trained to reconstruct the masked features $\boldsymbol{a}_{i\in M}$ via feature regression by minimising

$$\mathcal{L}_{\text{MAM}} = -\mathbb{E}_{(\boldsymbol{w},\boldsymbol{a})\sim\mathcal{D}} \sum_{i\in M} \|h_\theta(\boldsymbol{a}_i) - \boldsymbol{a}_i\|_2^2, \tag{3}$$

where M is the set of indices of masked features and $h_\theta(\boldsymbol{a}_i)$ the transformer output of the i-th feature, passed through a linear layer to obtain a vector of the same dimensions as the input.

To promote cross-modal learning, a third task, which we call audio-text matching (ATM), is added to pre-training to encourage the model to learn whether items in an audio-text pair match. The model is trained to minimise the binary cross-entropy between an output score $s_\theta(\boldsymbol{w}, \boldsymbol{a})$ and the ground-truth label y:

$$\mathcal{L}_{\text{ATM}} = -\mathbb{E}_{(\boldsymbol{w},\boldsymbol{a})\sim\mathcal{D}} y \log s_\theta + (1-y)\log\left(1-s_\theta\right), \tag{4}$$

where s_θ is obtained by taking the element-wise product of the output representations of the first item of the audio and text embedding sequences $\mathbf{h}_0^a$ and $\mathbf{h}_0^w$, passed through a linear layer and the sigmoid function. Negative pairs are created by replacing the associated caption with a random one in the mini-batch with a probability of 0.5.

The objectives are combined through linear scalarization:

$$\mathcal{L} = \sum_i \lambda_i \mathcal{L}_i, \tag{5}$$

where λ_i is the weight for the i-th loss $\mathcal{L}_i$. We assign these weights manually before starting the learning procedure.

3. EXPERIMENTAL SETUP

The goal of our experiments is to examine whether audio and language pre-training can help learn transferable music audio representations. Therefore we focus here on assessing the representations produced by the audio backbone and leave the exploration of the multimodal components of the model for future work.

In the following sections we describe our experimental approach, providing details on the training setup, evaluation protocol, datasets used and tasks considered.

3.1. Pre-training dataset and settings

We pre-train our model on a dataset of 250,513 audio-caption pairs from a private production music library, which we randomly split into training, validation and testing sets with an 80/10/10 ratio. The captions cover categories such as genre, instrumentation, mood and tempo; they can convey an overall description of the track (e.g. *"An uplifting and reflective pop track featuring synth, electric guitar, and lyrics with male vocals."*) or focus on one of these categories (e.g. genre) or one component of the track (e.g. the melody). Since no strong alignment is provided between words in the captions and sections of the audio, and since the content covered in the text input is not consistent across the examples, the captions can be considered as weak annotations. In preliminary experiments, we also tested our method on a subset of 115,933 pairs obtained from filtering out noisy captions based on some heuristics (length between 3 and 15 words, absence of text patterns observed to appear in bad captions). Despite the difference in size, this filtering did not seem to have a significant effect on the downstream performance and we report only results on the full data in the rest of the paper. Due to memory constraints, we also truncate the input audio to the first 20s.

Unless otherwise specified, all architectural and training settings are the same as in ViLBERT and we refer to the original paper for more details. As our audio backbone, we use *musicnn* [20], which operates on mel-spectrogram representations of the audio, with an input length of 3 seconds, and is trained from scratch with the rest of the model. In our experiments, we set $L = 6$ and $N = 2$.

3.2. Transferring to MIR tasks

Following standard protocols in transfer learning [21], [22], we train shallow classifiers on the audio representations in a supervised fashion. To this end, we discard the transformer layers after pre-training and keep the audio backbone frozen during downstream training. Specifically, we train a multilayer perceptron (MLP) with one hidden layer of size 512, using the output features of the audio backbone as input. Other training settings are also kept constant across all experiments in order to reduce the cost of hyperparameter tuning. We train the classifiers for 200 epochs using the Adam optimizer with minibatches of size 64 and an initial learning rate of 0.001, reduced linearly when the validation metric stops improving. We take mean accuracy, ROC-AUC and R^2 (see Sec. 3.3) as validation metrics for classification, multi-label classification and regression tasks respectively. Early stopping on the same metric is used together with weight decay (0.01) to impose regularisation. We report the average of each metric across 3 random initialisations of the MLP.

3.3. Downstream tasks and datasets

For downstream evaluation, we select a set of target datasets that are popular in the literature and representative of typical MIR tasks. In all cases, we follow the same pre-processing pipeline: we downmix the right and left channels to produce mono channel audio, downsample it to 16 kHz and apply a 3-second random crop on a dataset-specific basis to reduce training time and memory requirements. At test time, the full input audio is segmented into non-overlapping 3-second clips and predictions are computed for each clip. Track-level predictions are then obtained by averaging results across clips.

Auto-tagging Auto-tagging consists in assigning one or more labels to an audio clip from a set of predefined tags. The labels, or tags, typically have different levels of abstraction and cover various musical concepts, such as genre, instrumentation, mood and era. There are three main datasets for this task: MagnaTagATune (MTAT) [23], MTG-Jamendo [24] and MillionSongDataset (MSD) [25]. We consider only the first two, since audio tracks for the MSD dataset are no longer publicly available. MTAT consists of 30-second previews of around 26k tracks, while MTG-Jamendo contains around 54k full-length tracks. For both datasets we use standard splits[2,3] containing the 50 most frequent tags.

Genre and instrument classification For genre classification, we use the *small* subset of the Free Music Archive (FMA-small, FMA for brevity) [26], containing 30-second clips of 8,000 tracks from 8 different genres. For instrument classification, we adopt NSynth (NS) [27], a popular dataset made of over 300k 4-second monophonic audio samples categorised in 11 instrument families.

Emotion and theme recognition In order to easily compare our method to prior work on this task, we adopt the same protocol as in the Emotions and Theme Recognition in Music task of the MediaEval 2020 Benchmarking Initiative [28]. In this formulation, emotion recognition is a subtask of auto-tagging, where mood and theme annotations are taken as descriptors of the emotional content of the music, and the dataset used is a subset of the MTG-Jamendo dataset containing 18,486 tracks labelled with 56 mood and theme annotations. We use one of the public splits (*split-0*) for training, validation and testing. Additionally, we consider an alternative formulation of emotion recognition as a regression task and evaluate on the Emomusic dataset [29], using the artist-stratified split provided by [11]. In this case, we report an average of the valence and arousal coefficients of determination (R^2).

4. RESULTS

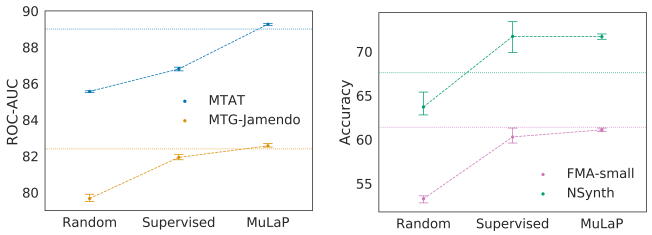
4.1. Comparison to baselines

We compare the test performance of our MuLaP-trained audio backbone to a fully supervised baseline and two transfer learning baselines. The fully supervised baseline consists of an architecturally identical audio backbone jointly trained with the classifier from scratch. In Fig. 3 we illustrate the performance of the end-to-end supervised baseline with horizontal lines and denote the two transfer learning baselines by *Random* and *Supervised*. *Random* indicates that no pre-training is used, while *Supervised* indicates that the backbone is pre-trained in a supervised way for auto-tagging on the MTG-Jamendo dataset (or the MTAT dataset if the downstream dataset is MTG-Jamendo).

A first takeaway is that MuLaP consistently outperforms the random initialisation baseline across all datasets and improves over both supervised pre-training and the fully supervised baseline on

[2] github.com/jongpillee/music_dataset_split/tree/master/MTAT_split

[3] github.com/MTG/mtg-jamendo-dataset



(a) Auto-tagging (b) Classification

Fig. 3: **Downstream performance of the frozen backbone** with different pre-training strategies: *Random*, where no pre-training is used, *Supervised* pre-training with auto-tagging as the source task, and *MuLaP*. The horizontal lines mark the average performance of a supervised baseline trained on the downstream task from scratch.

both datasets for auto-tagging (Fig. 3a) and on music recognition with Emomusic (with a 4.1% improvement, not shown here). It also performs competitively with the supervised pre-training case in both genre and instrument classification (Fig. 3b). Although MTG-Jamendo was chosen over MTAT when reporting final results for supervised pre-training due to its bigger size, similar results were observed on MTAT. This demonstrates that our approach overall learns transferable representations for a wider set of tasks than a supervised auto-tagging model. It should be noted, however, that the margin between MuLaP and the baselines isn't uniform across all datasets. A possible explanation for this may lie in the different degree of semantic overlap between pre-training captions and target labels: words making up tags in both MTAT and MTG-Jamendo are found in the pre-training captions ∼70-85% more often than those in genre and instrument labels contained in NSynth and FMA. The different performance gaps may also be attributed to a possible domain shift in the audio data between pre-training and downstream datasets. Though somewhat surprising, this could also explain why the end-to-end supervised baseline achieves higher accuracy on FMA.

4.2. Comparison to prior work

In Table 1 and 2 we compare MuLaP to relevant prior work. We select primarily transfer learning approaches on the same tasks and datasets considered in our study, reporting results from the literature for direct comparison.

For the auto-tagging task (Table 1), we compare our results to two prior approaches for MIR pre-training: CLMR [9], which trains a SampleCNN audio backbone via self-supervised contrastive learning on MSD, and CALM [11], a transformer trained on codified audio from a private dataset of 1M songs, conditioned on genre and artist labels. In Table 2 we extend the comparison to two cross-modal contrastive models which are trained to learn audio representations that align to the latent representation of corresponding tags (w2v [15]) or metadata and playlist information (Contr_G [16]). In both tables we also include a fully supervised approach trained end-to-end: HCNN [17], which achieves state-of-the-art results on auto-tagging, and MediaEval 20 [31], the MediaEval 2020 highest-scoring method for emotion and theme recognition. We note that the diversity of architectures, supervision mechanisms, training datasets and settings makes a direct comparison to our results difficult, particularly in the case of fully supervised methods, where full annotations are used to train the models from scratch on the target task. However a comparison is still beneficial to fully contextualise our work.

We observe that our approach generally does not match the state-of-the-art performance of fully supervised methods or that of CALM, which however trains on ∼5 times more audio data than ours. However MuLaP outperforms CLMR when this is trained on an out-of-domain dataset and used for transfer learning, indicating that using captions in pre-training can boost downstream performance compared to audio-only self-supervised learning. With the exception of auto-tagging on the instrument subset of MTG-Jamendo, in Table 2 we also observe that our approach achieves comparable performance to cross-modal methods on genre classification, instrument classification and emotion recognition, and improves classification performance on the FMA dataset compared to Park et al. [30], which make use of supervised pre-training.

Table 1: MuLaP auto-tagging performance compared to state-of-the-art MIR pre-training and fully supervised models.

Method	MTAT		MTG-Jamendo	
	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC
MuLaP	89.3	40.2	82.6	27.3
Self-supervised pre-training				
CLMR [9]	86.6	32.0	-	-
Semi-supervised pre-training				
CALM [11]	91.5	41.4	-	-
Fully supervised				
HCNN [17]	91.2	46.1	83.2	29.6

Table 2: MuLaP performance compared to transfer learning and state-of-the-art methods on: (i) genre and (ii) instrument classification as single-label classification on FMA and NS, and as multi-label classification on the corresponding MTG subsets; and (iii) emotion recognition on Emomusic and on the MTG mood subset.

Method	Genre		Instrument		Emotion	
	FMA	MTG	NS	MTG	Emo	MTG
MuLaP	61.1	85.9	71.7	76.8	58.5	76.1
Cross-modal pre-training						
w2v [15]	-	-	70.0	-	-	-
Contr_G [16]	-	84.7	-	79.7	-	73.2
Semi-supervised pre-training						
CALM [11]	-	-	-	-	66.9	-
Supervised pre-training						
Park et al. [30]	57.9	-	-	-	-	-
Fully supervised						
MediaEval 20 [31]	-	-	-	-	-	77.8

5. CONCLUSION

We have presented MuLaP, a framework for audio-linguistic pre-training, and investigated whether weak natural language supervision can be successfully used to learn transferable audio representations for a wide set of MIR tasks. We find that MuLaP can attain similar or better downstream performance when compared to the same audio architecture trained with traditional supervised techniques, confirming that audio captions can be usefully leveraged to enhance the quality of music audio representations produced by a standard CNN audio backbone. While we have focused on assessing audio representations on conventional MIR tasks, MuLaP can also be exploited for zero-shot classification and audio-linguistic tasks such as music captioning and cross-modal retrieval. Future work will explore these aspects and analyse the representations learnt through the multimodal components of the proposed framework.

6. REFERENCES

[1] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[2] M. Tagliasacchi, B. Gfeller, F. D. C. Quitry, and D. Roblek, "Pre-Training Audio Representations with Self-Supervision," *IEEE Signal Processing Letters*, vol. 27, pp. 600–604, 2020.

[3] A. Kolesnikov *et al.*, "Big Transfer (BiT): General Visual Representation Learning," in *Computer Vision-ECCV 2020*, vol. 12350, 2020, pp. 491–507.

[4] P. Hamel, M. E. Davies, K. Yoshii, and M. Goto, "Transfer Learning In MIR: Sharing Learned Latent Representations For Music Audio Classification And Similarity," in *Proceedings of the 14th ISMIR Conference*, 2013.

[5] A. van den Oord, S. Dieleman, and B. Schrauwen, "Transfer learning by supervised pre-training for audio-based music classification," in *Proceedings of the 15th ISMIR Conference*, 2014.

[6] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Transfer Learning for Music Classification and Regression Tasks," in *Proceedings of the 18th ISMIR Conference*, 2017.

[7] J. Lee and J. Nam, "Multi-Level and Multi-Scale Feature Aggregation Using Pre-trained Convolutional Neural Networks for Music Auto-tagging," *IEEE Signal Processing Letters*, vol. 24, no. 8, pp. 1208–1212, 2017.

[8] S. Oramas, L. Espinosa-Anke, A. Lawlor, X. Serra, and H. Saggion, "Exploring customer reviews for music genre classification and evolutionary studies," in *Proceedings of the 17th ISMIR Conference*, 2016.

[9] J. Spijkervet and J. A. Burgoyne, "Contrastive Learning of Musical Representations," *arXiv preprint arXiv:2103.09410*, 2021.

[10] H.-H. Wu *et al.*, "Multi-Task Self-Supervised Pre-Training for Music Classification," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 556–560.

[11] R. Castellon, C. Donahue, and P. Liang, "Codified audio language modeling learns useful representations for music information retrieval," in *Proceedings of the 22nd ISMIR Conference*, 2021.

[12] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks," in *Advances in Neural Information Processing Systems*, 2019, pp. 13–23.

[13] Y. C. Chen *et al.*, "UNITER: UNiversal Image-TExt Representation Learning," in *European Conference on Computer Vision*, 2020.

[14] X. Favory, K. Drossos, V. Tuomas, and X. Serra, "COALA: Co-Aligned Autoencoders for Learning Semantically Enriched Audio Representations," in *International Conference on Machine Learning (ICML), Workshop on Self-supervised learning in Audio and Speech*, 2020.

[15] X. Favory, K. Drossos, T. Virtanen, and X. Serra, "Learning Contextual Tag Embeddings for Cross-Modal Alignment of Audio and Tags," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[16] A. Ferraro, X. Favory, K. Drossos, Y. Kim, and D. Bogdanov, "Enriched Music Representations with Multiple Cross-modal Contrastive Learning," *IEEE Signal Processing Letters*, vol. 28, pp. 733–737, Apr. 2021.

[17] M. Won, A. Ferraro, D. Bogdanov, and X. Serra, "Evaluation of CNN-based Automatic Music Tagging Models," in *Proceedings of the 17th Sound and Music Computing Conference*, 2020, pp. 331–337.

[18] Q. Huang, A. Jansen, L. Zhang, D. P. W. Ellis, R. A. Saurous, and J. Anderson, "Large-Scale Weakly-Supervised Content Embeddings for Music Recommendation and Tagging," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[19] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019.

[20] J. Pons, T. Lidy, and X. Serra, "Experimenting with musically motivated convolutional neural networks," in *14th International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2016.

[21] A. Van Den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[22] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *International conference on machine learning*, PMLR, 2020.

[23] E. Law, K. West, M. Mandel, M. Bay, and J. Stephen Downie, "Evaluation of algorithms using games: The case of music tagging," in *Proceedings of the 10th ISMIR Conference*, 2009.

[24] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, "The MTG-Jamendo Dataset for Automatic Music Tagging," in *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML)*, 2019.

[25] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proceedings of the 12th ISMIR Conference*, 2011.

[26] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A dataset for music analysis," in *Proceedings of the 18th ISMIR Conference*, 2017.

[27] J. Engel *et al.*, "Neural audio synthesis of musical notes with WaveNet autoencoders," in *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[28] D. Bogdanov, A. Porter, P. Tovstogan, and M. Won, "MediaEval 2019: Emotion and theme recognition in music using jamendo," in *CEUR Workshop Proceedings*, 2019.

[29] M. Soleymani, M. N. Caro, E. M. Schmidt, C. Y. Sha, and Y. H. Yang, "1000 songs for emotional analysis of music," *CrowdMM 2013 - Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia*, pp. 1–6, 2013.

[30] J. Park, J. Lee, J. Park, J.-W. Ha, and J. Nam, "Representation Learning of Music Using Artist Labels," in *Proceedings of the 19th ISMIR Conference*, 2018.

[31] D. Knox, T. Greer, B. Ma, E. Kuo, K. Somandepalli, and S. Narayanan, "MediaEval 2020 emotion and theme recognition in music task: Loss function approaches for multi-label music tagging," in *CEUR Workshop Proceedings*, 2020.