MusicBERT: A Shared Multi-Modal Representation for Music and Text

Federico Rossetto
University of Glasgow
fedingo@gmail.com

Jeff Dalton
University of Glasgow
Jeff.Dalton@glasgow.ac.uk

Abstract

Recent advances in deep learning have led to significant advances in both text and music representations. However, the representations and tasks remain largely separate. Most Music Information Retrieval models focus on either music or text representations but not both. In this work we propose unifying these two modalities in a shared latent space. We propose building on a common framework of Transformer-based encoders for both text and music modalities using supervised and unsupervised methods for pre-training and fine-tuning. We present initial results and key challenges that need to be overcome to make this possible. The result will be a new class of models that are able to perform advanced tasks that span both NLP and music.

1 Introduction

Voice-based conversational agents such as Alexa, Google Assistant, are growing in importance and allow interaction with music systems using natural language. However, current approaches only support interactions with text and metadata (e.g. playing a specific song). This research will enable a multi-modal representation that supports conversation *about* music and its concepts. We propose using new state-of-the-art deep learning models that are capable of producing *joint latent spaces* of both music and text.

We propose a new multi-modal representation model we call *MusicBERT*. It is composed of a set of modality-specific encoders that are then fed to a shared model, based on the Transformer. A key challenge is that audio signals are very long and standard Transformer models are limited in the amount of vectors that they can feasibly and effectively process. As a result, a second layer that encodes music (and its derived concepts) is needed to provide some abstraction with current models. A high-level view of this model is provided in Figure 1. Similar to how BERT has been adapted for QA and NLP tasks, the proposed architecture can be adapted for music QA tasks (Sutcliffe et al., 2014) using the pre-trained representations.

Although text representations are proven, music remains challenging. In order to have an effective music representation the proposed model must address two key challenges: 1) What is an effective low-level encoding of music that works effectively with Transformers, and 2) What are effective pre-training loss functions for learning music representations that exhibit transfer learning properties. We hypothesize that this requires having the right level of semantic representation.

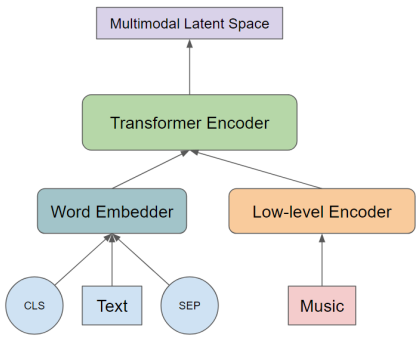

Figure 1: High-level MusicBERT architecture

To train this new multi-modal model we use a standard music datasets the MuMu dataset (Oramas et al., 2017), that maps album reviews to a subset of the Million Song Dataset (Bertin-mahieux et al., 2011). These reviews are not very granular, been at album-level, but still provide insightful discussions about the musical content of the tracks contained in the album.

	MFCC	TLM	MTLMR	VGGish	Audioset	MLM	MIM
GTZAN (accuracy)	59.83%	77.60%	65.80%	85.9%	83.00%	73.80%	85.30%
Deezer (r2 score)	7.18%	18.58%	9.60%	20.65%	18.38%	14.58%	16.45%

Table 1: SVM evaluation using the Music Representations extracted

2 Background and Related Work

Textual Representation Learning - The use of Transformers (Vaswani et al., 2017) paired with a language modeling objective is the current state-of-the-art for most NLP tasks. Models such as BERT and similar are effective for NLP tasks and critically they demonstrate strong transfer learning effectiveness (Devlin et al., 2018). We use this as the base for our text-based representations.

Music Representation Recent work by Kim et al. (2019) explores deep representation learning. They apply *multi-task transfer learning* to test the impact on multiple tasks showing an improvement in effectiveness when pre-training on additional external tasks. Similar work on music representations and transfer learning is (Choi et al., 2017). They demonstrate the potential for pre-training on music tagging to create effective *latent representations*. Recent work with Transformers for monophonic music by Huang et al. (2018) begins to address scalability issues. In contrast, we propose representations that generalize to polyphonic music and raw audio.

Multi-modal Representation This work is inspired by multi-modal representations in the field of Computer Vision, and specifically VisualBERT (Li et al., 2019). They develop a model that uses ImageNet concepts to encode the key-points of an image, and train BERT to translate these visual vectors into a textual description. We propose using both encoded audio and audio concepts as a semantic representation. Instead of ImageNet, we use AudioSet (Gemmeke et al., 2017), a concept detector for general audio. Our early results find that more work is needed in developing an ontology specific to music.

3 Method and Preliminary Experiments

In this section we discuss the methods we use including the low-level music encoding and training objective. For low level encoders for our experiments use a word embedding layer as textual encoder, and the VGGish (Hershey et al., 2017) model for the music encoder.

To train this model on music data, we experiment with three different pre-training approaches. A *Masked Language Modeling* algorithm, a *Mutual Information Maximization* algorithm taken from van den Oord et al. (2018) and a standard classification task on the *AudioSet* ontology (Gemmeke et al., 2017). The first algorithm uses a reconstruction loss after masking some of the music vectors. The second instead aims at maximizing the Mutual Information between the music vectors and the multi-modal representations. The last one is just a multi-label classification of sound events.

Results To evaluate the obtained representations, we use standard music tasks, and evaluate the representations using them in an Support Vector Machine (SVM) on the *latent space* for each target, following Choi et al. (2017). We evaluate on the GTZAN (Tzanetakis and Cook, 2002) and Deezer (Delbouys et al., 2018) datasets.

We report the results of our initial experiments in table 1. The results show, VGGish provides a strong representation, and the MIM and MLM training hurt effectiveness. We also provided the results for three baselines. One using the standard MFCCs (Muda et al., 2010), one base on (Choi et al., 2017) (TLM) and one base on (Kim et al., 2019) (MTLMR).

We found that the VGGish encoder output has very limited variability across time and this makes it much more challenging for the model to be effectively trained. Also, there is a blurring effect on the vectors across time caused by the soft-max self-attention. This suggests that using sparse-attention could improve the model effectiveness.

4 Conclusion

We motivate the need for a *multi-modal representation space* and its application to natural language music conversation. We introduce the MusicBERT model to tackle this problem and present preliminary results on standard music tasks. Our goal is to advance the capabilities of current models and enable them to integrate music and text together in new and more effective representations that generalize to complex tasks.

References

Thierry Bertin-mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. 2011. The million song dataset. In *In Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR*.

Keunwoo Choi, George Fazekas, Mark Sandler, and Kyunghyun Cho. 2017. Transfer learning for music classification and regression tasks. In *The 18th International Society of Music Information Retrieval (ISMIR) Conference 2017, Suzhou, China*. International Society of Music Information Retrieval.

Rémi Delbouys, Romain Hennequin, Francesco Piccoli, Jimena Royo-Letelier, and Manuel Moussallam. 2018. Music mood detection based on audio and lyrics with deep neural net. In *ISMIR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA.

S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson. 2017. Cnn architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135.

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, and Douglas Eck. 2018. An improved relative self-attention mechanism for transformer with application to music generation. *CoRR*, abs/1809.04281.

Jaehun Kim, Julián Urbano, Cynthia C. S. Liem, and Alan Hanjalic. 2019. One deep music representation to rule them all? a comparative analysis of different representation learning strategies. *Neural Computing and Applications*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language.

Lindasalwa Muda, Mumtaj Begam, and I. Elamvazuthi. 2010. Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Sergio Oramas, Oriol Nieto, Francesco Barbieri, and Xavier Serra. 2017. Multi-label music genre classification from audio, text, and images using deep features.

Richard FE Sutcliffe, Tim Crawford, Chris Fox, Deane L Root, and Eduard H Hovy. 2014. The c@ merata task at mediaeval 2014: Natural language queries on classical music scores.

G. Tzanetakis and P. Cook. 2002. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.