A Comparison of Music Input Domains for Self-Supervised Feature Learning

Siddharth Gururani [1] **Alexander Lerch** [1] **Mason Bretan** [2]

Abstract

In music using neural networks to learn effective feature spaces, or embeddings, that capture useful characteristics has been demonstrated in the symbolic and audio domains. In this work, we compare the symbolic and audio domains, attempting to identify the benefits of each, and whether incorporating both of the representations during learning has utility. We use a self-supervising siamese network to learn a low-dimensional representation of three second music clips and evaluate the learned features on their ability to perform a variety of music tasks. We use a polyphonic piano-performance dataset and directly compare the performance on these tasks with embeddings derived from synthesized audio and the corresponding symbolic representations.

1. Introduction

A Large body of research in music information retrieval (MIR) aims to reduce the dimensionality of musical data and categorize it into higher-level descriptors such as genre, composer, or tempo. In the past, this was achieved by extracting a low-dimensional intermediate representation from audio involving spectral, dynamic, and temporal hand-crafted features (Tzanetakis & Cook, 2002; Casey et al., 2008; Mandel & Ellis, 2005) followed by classification.

The dimensionality reduction pipeline has now been replaced by the layers of deep neural networks (DNN), where low-dimensional representations are directly learned from labeled data. DNNs are the state-of-the-art in several tasks (Han et al., 2017; Choi et al., 2017a; Hawthorne et al., 2018).

However, these approaches rely on labeled datasets preventing models from working with, for example, new styles of music unrepresented in the dataset. Examples of approaches tackling this challenge are transfer learning (Van Den Oord et al., 2014; Choi et al., 2017b) and student-teacher learning (Wu & Lerch, 2017). In the symbolic music domain, learning a low-dimensional feature or 'embedding' space using unlabeled data and autoencoders has been explored by Bretan et al. (2017). In our work, we evaluate whether such an approach may be applied to learn effective feature spaces using unlabeled musical audio by comparing models trained with audio and symbolic data.

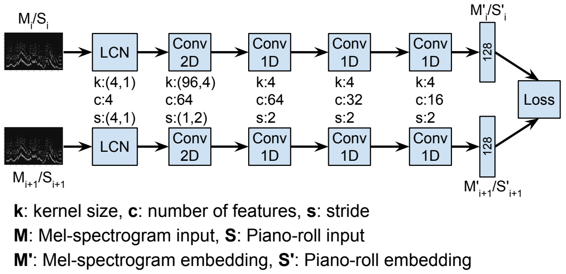

Figure 1. Block Diagram for Proposed Model

The benefits of models capable of extracting meaningful features from audio are: (i) less dependence on labeled data for new tasks, (ii) applicability to generative music tasks in the audio domain, and (iii) ability to design interactive music systems utilizing audio as opposed to symbolic data.

2. Proposed Method

We compare audio and symbolic music inputs for learning effective feature spaces. To the best of our knowledge, a direct comparison of input representations better suited for feature learning has not been performed. We compare by training embedding spaces with audio and symbolic data and evaluating them. Siamese convnets are trained to minimize the distance between embeddings of neighboring clips using a custom loss function, as depicted in Figure 1.

For direct comparison, we use MIDI files to obtain a symbolic representation as well as to synthesize audio using fluidsynth. We use a set of MIDI files of performances of classical compositions from 25 composers as used in (Bretan et al., 2017) as well as improvised piano performances on 92 jazz standards from the Real Book. We use 3 s clips as inputs. For the audio representation, we use log

[1] Center For Music Technology, Georgia Tech, Atlanta, Georgia, USA [2] Samsung Research America, Mountain View, California, USA. Correspondence to: Siddharth Gururani <siddgururani@gatech.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

	Rank@100		Composer		Key Signature		Tempo		Note Density	
	AM	GM	microF	macroF	microF	macroF	8% acc	4% acc	MAE	R^2
MEL_MEL	**7.966**	**3.77**	**0.330**	**0.178**	0.264	0.183	**0.265**	**0.127**	**6.799**	**0.613**
SYM_SYM	13.52	6.18	0.228	0.085	**0.340**	**0.317**	0.172	0.075	11.14	0.134
MEL_SYM	12.30	5.52	0.228	0.089	**0.384**	**0.353**	0.189	0.095	7.467	0.540
SYM_MEL	12.35	5.53	**0.255**	**0.099**	**0.383**	**0.350**	0.169	0.086	8.328	0.458

Table 1. Results for embedding spaces trained as per Sect. 3. Bold indicates statistically significant best performance (Bonferroni correction leads to $p < 0.0083$). Multiple bold numbers in a column indicate that the corresponding models are not significantly different.

mel-spectrograms with 96 mel-bands and for the symbolic representation, we use piano-rolls with 8 octaves.

Each clip's mel-spectrogram M_i and piano-roll S_i are paired with its neighboring clip's mel-spectrogram M_{i+1} and piano-roll S_{i+1}. These pairs are inputs to the network.

2.1. Semantic Relevance Loss Function

Given a pair of semantically relevant/similar vectors $\bar{M}$ and $\bar{N}$ and a set of four vectors D containing $\bar{N}$ along with three randomly chosen irrelevant/dissimilar vectors, the semantic relevance of $\bar{N}$ and $\bar{M}$ is computed as:

$$P(\bar{N}|\bar{M}) = \frac{\exp(\mathrm{sim}(\bar{M},\bar{N}))}{\sum_{\bar{d}\in D}\exp(\mathrm{sim}(\bar{M},\bar{d}))}, \tag{1}$$

where sim is cosine similarity. $|D| = 4$ controls how many dissimilar vectors are used. During the forward pass, a batch of neighboring clips are passed through the network. The pairs of 128-dimensional embeddings obtained correspond to $\bar{M}$ and $\bar{N}$ in Eq. (1). Next, $|D| - 1$ dissimilar embedding vectors are picked randomly from the remaining batch items. Finally, we use Adam optimizer (Kingma & Ba, 2014) to minimize: $-log \prod_{\bar{M},\bar{N}} P(\bar{N}|\bar{M})$ where $\bar{M}, \bar{N}$ are the embeddings of all the neighboring pairs of clips in the dataset. This loss function has also been applied to learning embeddings for symbolic music (Bretan et al., 2017).

3. Evaluation

We perform four experiments to compare the performance of models trained with audio and symbolic representations.

1. **Mel-spectrogram (MEL_MEL)**: M_i and M_{i+1} input
2. **Piano-roll (SYM_SYM)**: S_i and S_{i+1} input
3. **Mix inputs 1 (MEL_SYM)**: M_i and S_{i+1} input. During testing, mel-spectrogram is used for embedding.
4. **Mix inputs 2 (SYM_MEL)**: S_i and M_{i+1} input. During testing, piano-roll is used for embedding.

We test mixed input representations to check if maximizing relevance using a different representation helps the network learn better features. For mixed experiments, the siamese network does not share weights between the sub-networks.

3.1. Metrics

- **Ranking**: For each test clip, we pick its neighboring clip and 99 random test clips. We compute cosine distances between the embeddings of the test clip and the 100 selected clips. Rank of the neighboring clip is the rank@100. We report arithmetic and geometric mean rank@100. Neighboring clips should be nearest neighbors in the embedding space as well.
- **Musical Tasks**: We evaluate the performance of the models on classification and regression tasks involving musical characteristics. We extract embeddings for the training and test set using the four models and train fully connected networks with two hidden layers for each task. The tasks are: 25-way composer classification, 12-way key signature classification, tempo estimation and note-density estimation. We use standard metrics for these tasks such as F1-scores for classification, mean absolute error (MAE) and R-squared for regression. For tempo estimation we compute accuracy with 4% and 8% tolerance.

4. Results and Conclusion

Based on the results in Table 1, the general trend we observe is that models trained using mel-spectrogram outperform those trained only with symbolic data. Models trained with only mel-spectrogram outperform the other models in all metrics except key signature classification. An explanation is the difficulty in resolving pitch in a mel-spectrogram due to harmonics overlapping, compared to piano-roll. The performance of all models is poor for tempo estimation possibly due to the short length of input clips.

To study the impact of our results, possible future directions are as follows: (i) comparing our method of self-supervised feature learning with other methods such as autoencoders, (ii) training with real-world, polyphonic music, and (iii) exploring the benefits of using synthesized audio instead of symbolic music for generative music systems.

Finally, while there is still potential for improvement in supervised learning for MIR, with this work we hope to move further in the direction of leveraging vast amounts of unlabeled data to solve problems in MIR and generation.

References

Bretan, M., Oore, S., Eck, D., and Heck, L. Learning and evaluating musical features with deep autoencoders. *arXiv preprint arXiv:1706.04486*, 2017.

Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., and Slaney, M. Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4):668–696, April 2008. ISSN 0018-9219. doi: 10.1109/JPROC.2008.916370.

Choi, K., Fazekas, G., Sandler, M., and Cho, K. Convolutional recurrent neural networks for music classification. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 2392–2396, New Orleans, 2017a.

Choi, K., Fazekas, G., Sandler, M., and Cho, K. Transfer learning for music classification and regression tasksn. In *Proceedings of the International Society of Music Information Retrieval Conference (ISMIR)*, pp. 141–149, Suzhou, 2017b.

Han, Y., Kim, J., Lee, K., Han, Y., Kim, J., and Lee, K. Deep convolutional neural networks for predominant instrument recognition in polyphonic music. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 25(1):208–221, 2017.

Hawthorne, C., Elsen, E., Song, J., Roberts, A., Simon, I., Raffel, C., Engel, J., Oore, S., and Eck, D. Onsets and frames: Dual-objective piano transcription. In *Proceedings of the International Society of Music Information Retrieval Conference (ISMIR)*, pp. 50–57, Paris, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL http://arxiv.org/abs/1412.6980.

Mandel, M. I. and Ellis, D. P. Song-level features and support vector machines for music classification. In *Proceedings of the International Society of Music Information Retrieval Conference (ISMIR)*, pp. 594–599, London, 2005.

Tzanetakis, G. and Cook, P. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, July 2002. ISSN 1063-6676. doi: 10.1109/TSA.2002.800560.

Van Den Oord, A., Dieleman, S., and Schrauwen, B. Transfer learning by supervised pre-training for audio-based music classification. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 29–34, Taipei, 2014.

Wu, C.-W. and Lerch, A. Automatic drum transcription using the student-teacher learning paradigm with unlabeled music data. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 613–620, Suzhou, 2017.