LEARNING LONG-TERM MUSIC REPRESENTATIONS VIA HIERARCHICAL CONTEXTUAL CONSTRAINTS

Shiqi Wei[1,2] **Gus Xia**[2]
[1] School of Data Science, Fudan University
[2] Music X Lab, Computer Science Department, New York University Shanghai
sqwei19@fudan.edu.cn, gxia@nyu.edu

ABSTRACT

Learning symbolic music representations, especially disentangled representations with probabilistic interpretations, has been shown to benefit both music understanding and generation. However, most models are only applicable to short-term music, while learning long-term music representations remains a challenging task. We have seen several studies attempting to learn hierarchical representations directly in an end-to-end manner, but these models have not been able to achieve the desired results and the training process is not stable. In this paper, we propose a novel approach to learn long-term symbolic music representations through contextual constraints. First, we use contrastive learning to pre-train a long-term representation by constraining its difference from the short-term representation (extracted by an off-the-shelf model). Then, we fine-tune the long-term representation by a hierarchical prediction model such that a good long-term representation (e.g., an 8-bar representation) can reconstruct the corresponding short-term ones (e.g., the 2-bar representations within the 8-bar range). Experiments show that our method stabilizes the training and the fine-tuning steps. In addition, the designed contextual constraints benefit both reconstruction and disentanglement, significantly outperforming the baselines.

1. INTRODUCTION

Deep music representation learning have been proven to be a powerful tool for high-quality symbolic music generation [1]. The learned representations can be directly fed into downstream predictive models such as LSTMs [2] and Transformers [3] to achieve more coherent results than note-based or event-based generation [4–6] Furthermore, when a representation learning model has a probability interpretation, the representation can then be easily interpolated or resampled to create new music pieces. Recently, several studies further disentangle music representations into interpretable factors (such as pitch, rhythm, chord and texture) to achieve a more controllable and interactive music generation [5, 7, 8]. For example, we can keep the pitch factor of a melody while resampling its rhythm factor to achieve theme variation. We can also interpolate the pitch factor for a smooth music morphing [7].

Despite the above mentioned progress [9–11], most existing work applies only to short music segments with a length of several beats, while learning long-term representations remains a challenging task. In particular, studies have shown that even for monophonic melodies, "flat" model designs (e.g., using long-range sequential encoders) have difficulty remembering a complete music phrase at once. Some other studies have attempted to solve this problem by building another layer of hierarchy on top of short-range flat models, learning short-term and long-term representations simultaneously in an end-to-end manner [1, 12]. However, as the model expressivity increases with the number of layers, models also become much more difficult to train.

We argue that the main problem with current methods is the lack of proper inductive bias, and in this paper we propose a new method for learning long-term, phrase-level symbolic music representations through contextual constraints. The method consists of two stages pre-training and fine-tuning, with two steps in each stage. In the pre-training stage, we first adopt EC^2-VAE [7] to learn bar-level, disentangled latent pitch and rhythm representations. Then, we apply the same model to learn phrase-level representation but with contrastive losses to constrain the difference between phrase-level and bar-level representations. It is indeed difficult to learn phrase-level representations directly using bar-level models, but the additional contrastive constraint can serve as a useful inductive bias to help find a reasonable solution that can subsequently be improved by fine-tuning. During the fine-tuning stage, we replace the pre-trained decoder with a hierarchical prediction model that forces the phrase-level representation to reconstruct the bar-level ones. This is achieved by first tuning only the new hierarchical decoder (while fixing the pre-trained encoder) and then tuning the whole network. During these two steps, structured contrastive loss is applied to stabilize the learning process.

Experiments show that the proposed method significantly outperforms the baselines and successfully learns disentangled pitch and rhythm representations for 8-bar long phrases (32 beats in 4/4 meter) without increasing the latent dimensionality. To our knowledge, this is also the first generative model that achieves phrase-level composition style transfer, latent factor interpolation, and theme

variation. In sum, our contributions are as follows:

- We demonstrate the importance of structured contextual constraints in learning long-term disentangled representations. Our approach only requires reasonable amount of data to train and could learn compact latent representation.
- We show that the proposed Structured InfoNCE loss effectively expresses the contextual constraints, stabilizes the training of long-range models and helps the model converge faster.
- Our model achieves phrase-level music style transfer, latent factor interpolation, and theme variation.

2. RELATED WORK

We review two realms of research related to our work on long-term music-representation learning: contrastive learning, which is the main method to stabilize the training process, and hierarchical music modeling, which is related to our fine-tuning model.

2.1 Contrastive Learning

Contrastive learning (CL) is an efficient method in self-supervised learning [13–15], serving as regularization to latent representations. For example, NCE-based contrastive losses [16, 17] have been widely used and achieved good results in natural language processing. Contrastive predictive coding (CPC) [18] and Deep Infomax (DIM) [19] explore the relation between minimizing a contrastive learning loss and maximizing a lower bound of the mutual information. In DIM, global feature is connected with local feature to learn more abstract and informative representations.

2.2 Hierarchical Music Representation Learning

The hierarchical nature of music has been studied for a long time [20–23]. Recently, we see some efforts on learning long-term music representations using hierarchical modeling [12, 24, 25]. The basic idea is that since a flat model design can only effectively learn shot-term representations, we can stack more layers on top of the short-term representations module for long-term representations. Existing works include MusicVAE [1], Music Transformer VAE [12], Jukebox [26], etc. However, experiments show that unless we have a huge amount of data, the model is in general very difficult to train. In this study, we provide a two-stage algorithm with contrastive loss as a better learning strategy. Also, no model so far has achieved disentanglement for long-term representation as done in this study.

3. METHODOLOGY

In this section, we introduce our algorithm in detail. Conceptually, it consists of two stages, each with two steps.The first stage is *pre-training*:

- In step 1, we simply adopt EC^2-VAE [7], an existing music representation disentanglement model, to extract short-term pitch and rhythm representations.
- In step 2, we build Long-EC^2-VAE, a long-term version of the model and train it with an extra contextual constraint using the proposed *Structured* InfoNCE loss. Intuitively, this loss prevents the learned long-term representations from deviating too far from corresponding well-trained short-term representations.

The second stage is *fine-tuning*, in which we build a hierarchical representation-learning model by combining the encoder of Long-EC^2-VAE with a hierarchical decoder. We name this model after Hierarchical-EC^2-VAE.

- In step 1, we only train the hierarchical decoder to ensure the predictive power of the long-term representation.
- In step 2, we train the whole hierarchical network for a better long-term pitch-rhythm disentanglement.

3.1 Pre-training by Contrastive Learning

The model of the pre-training stage, Long-EC^2-VAE, is shown in Figure 1. It is built upon an off-the-shelf music representation model, EC^2-VAE [7], which can effectively disentangle pitch and rhythm factors for short music segments by cutting the latent representation into two parts and pairing one part with a local rhythm decoder. In Figure 1, the right-hand-side part is a literal copy of the EC^2-VAE encoder (with parameters fixed) to extract short-term representations, while the left-hand-side part is a simple adaptation of EC^2-VAE for long-term music by lengthening its temporal receptive field. Note that the left part alone is not able to learn long-term representations, and our goal is to assist it using contrastive learning. Formally, the loss function of Long-EC^2-VAE is:

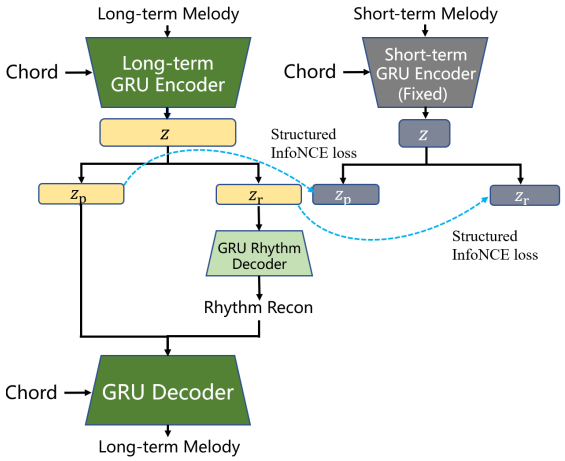
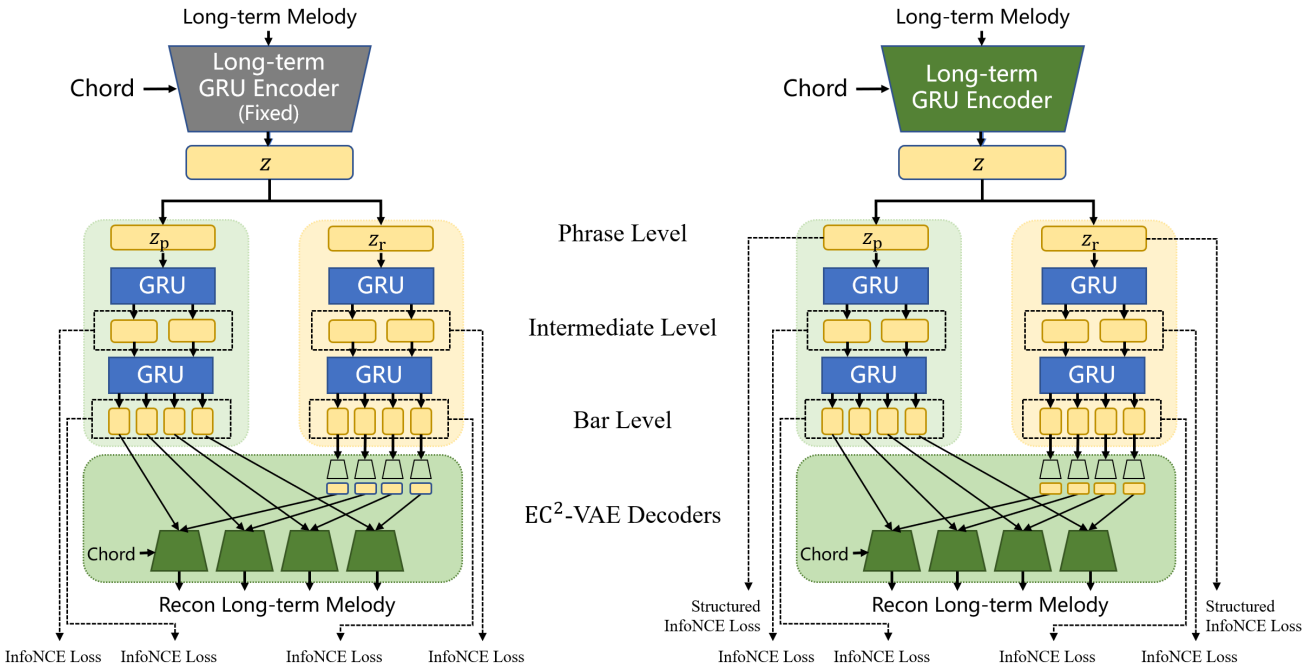


Figure 1: The model architecture of Long-EC^2-VAE in the pre-training stage, where the right-hand-side is short-term model with parameter fixed and the left-hand-side is the long-term model. The dotted lines denote contrastive losses, whose weighting matrices are joined optimized with the parameters on the left-hand-side networks.

$$\mathcal{L} = \mathcal{L}_{\text{Long-EC}^2\text{-VAE}} + \mathcal{L}_{\text{Structured InfoNCE}}, \quad (1)$$

where $\mathcal{L}_{\text{Long-EC}^2\text{-VAE}}$ is the same as in the original EC^2-VAE model (which contains the KL loss, the rhythm loss



(a) tuning only the decoder.

(b) tuning the whole network.

Figure 2: The model architecture of Hierarchical-EC^2-VAE in the fine-tuning stage. The training follows two steps.

and the overall reconstruction loss). The Structured InfoNCE loss expresses the contextual constraint. It is developed from InfoNCE [18] loss, and it is *structured* since the compared representation pairs are extracted from music segments of different length, one is long term and the other is short term. Formally:

$$\mathcal{L}_{\text{Structured InfoNCE}} = -\ln \frac{\exp\left(z_{\mathrm{L},f}^T W \hat{z}_{\mathrm{S},f}^+ / \tau\right)}{\exp\left(z_{\mathrm{L},f}^T W \hat{z}_{\mathrm{S},f}^+ / \tau\right) + \sum_{i=1}^K \exp\left(z_{\mathrm{L},f}^T W \hat{z}_{\mathrm{S},f}^- / \tau\right)}, \quad (2)$$

where $z_{\mathrm{L},f}$ and W are the normalized long-term representations and weighing matrix we need to learn. $f = \{\mathrm{p}, \mathrm{r}\}$ indicates whether it is the pitch or rhythm factor. Likewise, we use $\hat{z}_{\mathrm{S},f}$ to denote the short-term representations extracted by right-hand-side model. K and τ are hyperparameters. K is the amount negative samples and τ is the temperature parameter.

In specific, the short-term melodies are half as long as long-term ones. The positive samples $\hat{z}_{\mathrm{S},f}^+$ are in the cases that the corresponding short-term melody is a part of the long-term melody and f takes the same value as in $z_{\mathrm{L},f}$, while the negative samples are not in this case. Also, the long-term and short-term representations share the *same* dimensionality.

3.2 Fine-tuning with Hierarchical Generation

Figure 2 shows the architecture of the fine-tuning model, Hierarchical-EC^2-VAE, where the two subfigures illustrate the two training steps. Here, the encoder design is the same as in the Long-EC^2-VAE model, while the decoder is a hierarchical predictive model with three layers. The first two layers are new designed and the last layer is an aggregation of several EC^2-VAE decoders sharing the same parameters. Given the disentangled long-term (phrase-level) representations, it first decodes intermediate-level representations, then decodes bar-level representations, and finally reconstructs concrete rhythm and music tokens.

Compared to the phrase-level representation, the temporal receptive fields of the intermediate-level representations all shrink to a half, but at the same time their number doubles in order to cover the same range of music. The same relationship holds between intermediate and bar-level representations. In particular, a phrase means 8 bar (in 4/4 meter, 32 beats) in our design, so that the intermediate-level and bar-level mean 4-bar and 2-bar melody segments (a length which the original EC^2-VAE model can handle), respectively. All levels of latent representations share the same dimensionality.

In the first step of training (Figure 2(a)), the encoder is a literal copy from the Long-EC^2-VAE model and we only train the hierarchical decoder. Formally, the loss function is:

$$\mathcal{L}_{\text{step1}} = \mathcal{L}_{\text{Hierarchical}-\text{EC}^2\text{-VAE Decoder}} + \mathcal{L}_{\text{InfoNCE}}, \quad (3)$$

where the first term refers to the reconstruction losses adopted from the EC^2-VAE model, and the second term is defined as:

$$\mathcal{L}_{\text{InfoNCE}} = -\ln \frac{\exp\left(z_{l,f}^T W \hat{z}_{l,f}^+ / \tau\right)}{\exp\left(z_{l,f}^T W \hat{z}_{l,f}^+ / \tau\right) + \sum_{i=1}^K \exp\left(z_{l,f}^T W \hat{z}_{l,f}^- / \tau\right)}, \quad (4)$$

where $z_{l,f}$ are the normalized hierarchical representations we need to learn with $l = \{\text{intermediate}, \text{bar}\}$ indicating the level of representation and other notations follow the

same meaning as in Eq.(2). Here, both positive $\hat{z}^{+}_{l,f}$ and negative $\hat{z}^{-}_{l,f}$ samples are normalized representations computed from a pre-trained EC^2-VAE, in which the positive samples are in the cases that the $\hat{z}^{+}_{l,f}$ and $z_{l,f}$ are computed based on the same music segment and have the same value of l and f, while $\hat{z}^{-}_{l,f}$ are not in this case.

After the first step achieves a reasonable accuracy, we proceed to step 2 (Figure 2(b)), unfreezing the encoder and training the whole hierarchical representation-learning model with:

$$\mathcal{L}_{\text{step2}} = \mathcal{L}_{\text{step1}} + \mathcal{L}_{\text{Structured InfoNCE}} + \beta\mathcal{L}_{\text{KL phrase}}, \quad (5)$$

where the first two terms are defined in Eq. (3) and Eq. (2) respectively. $L_{\text{KL phrase}}$ is KL divergence to only regularize the phrase-level representations by a normal distribution. The value β controls the degree of KL divergence penalty.

4. EXPERIMENTS

4.1 Dataset and data format

We train our model on Nottingham Database [27] and POP909 database [28]. Our dataset contains 2154 melodies (at song level) in total. We randomly split these pieces into 2 subsets: 90% songs for training and 10% songs for test. The data format is designed as the same as in [7] in which 4 bar or 8 bar melodies are formalized as sequences of 130-dimensional one-hot embedding vectors and 16-beat and 32-beat rhythm pattern is represented by a sequence of 3-dimensional one-hot embedding vectors. Each vector in the melody sequence denotes a $\frac{1}{4}$-beat unit. The first 128 dimensions of this vector denote 128 MIDI-format pitches from 0 to 127, the 129^{th} dimension is the holding state for longer note duration, and the last dimension is kept for rest. The three dimensions of rhythm pattern vectors represent the onset of any pitch, a holding state, and rest, respectively.

4.2 Implementation Details

All of our models are trained using Adam optimizer [29] with a scheduled learning rate from 1e-3 to 1e-5. The batch size is 128 in the pre-training stage and is 64 in the fine-tuning stage. We do normalization on representations in Eq.(2) and (4) to make the training process more stable. The representations fed into decoders are original representation without normalization.

4.2.1 Pre-training

In the pre-training stage, we simply adopt the structure of EC^2-VAE [7] to model 4 bar and 8 bar EC^2-VAE. Each model comprises an encoder with a bi-directional GRU layer, a rhythm decoder with a GRU layer, and a global decoder with a GRU layer. We set the hidden dimension of the GRU in the encoder and decoders to 2048. The latent dimension is 128 for disentangled pitch representations and 128 for disentangled rhythm representations for each range model. For $\mathcal{L}_{\text{Structured InfoNCE}}$ depicted in Eq. (2), we set K to 512 and τ to 1. The positive samples for Eq. (2) and Eq. (4) are the representations of 1-4th, 3-6th and 5-8th bar from well-trained 4 bar EC^2-VAE. Actually, even when training the 4-bar EC^2-VAE (right-hand side of Figure 1), we use a similar constrastive loss as in Eq. (2) where the positive samples are representations of 1-2th, 2-3th, 3-4th bar from well-trained 2 bar (original) EC^2-VAE [7] .

4.2.2 Fine-tuning

Hierarchical-EC^2-VAE model consists of a long-term (8 bar) EC^2-VAE encoder, 4 GRU layers, and an aggregation of 2 bar EC^2-VAE decoders. We first train the hierarchical model with fixed 8 bar EC^2-VAE encoder from pre-trained stage for around 25 epochs. Then we train the whole model without fixing parameters. We set the hidden dimension of 4 GRU layers to 1024. We set K to 256 and τ to 1 for both Structured InfoNCE loss and InfoNCE loss and set β to 0.1 in Eq. (5).

4.3 Objective Evaluation

We objectively evaluate the model in terms of reconstruction accuracy, training stability, and disentanglement.

4.3.1 Reconstruction Accuracy

Table 1 shows that the reconstruction accuracy of the proposed models (2nd an 3rd rows) significantly outperform the baseline, a vanilla EC^2-VAE applied to 8-bar melody (first row). The last two rows show the results of two ablation settings of Hierarchical-EC^2-VAE: one without the contrastive loss and the other without first fixing the parameters of encoder and directly train the model end-to-end. We see that the proposed Structured InfoNCE or InfoNCE losses play a vital role for an accurate reconstruction and the two-step training strategy improves the result marginally.

Method	Recon.Acc	Rhythm Recon.Acc
Baseline	0.772	0.847
Long-EC^2-VAE	**0.992**	**0.995**
H-EC^2-VAE(ours)	**0.997**	**0.995**
H-EC^2-VAE(w-o CL)	0.584	0.599
H-EC^2-VAE(w-o fixed)	0.991	0.989

Table 1: A comparison on reconstruction accuracy of different models.

4.3.2 Training stability

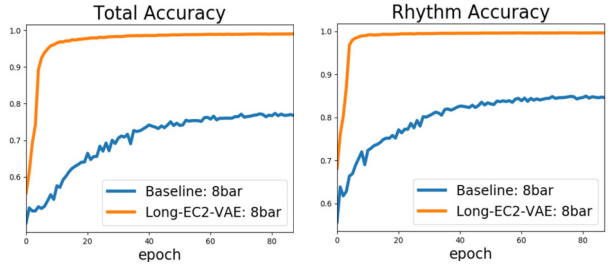
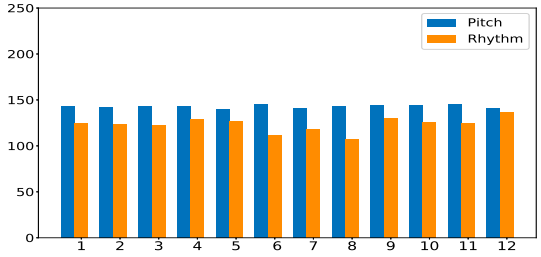


Figure 3: Experimental results of overall reconstruction and rhythm accuracy on the test set.

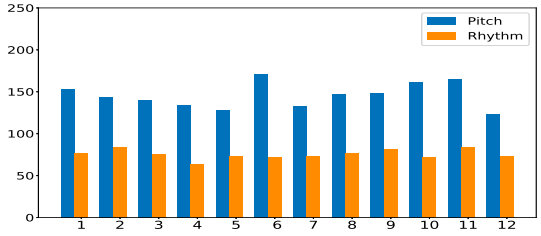
Comparing the accuracy curves of the proposed Long-EC^2-VAE with the baseline as illustrated in Figure 3, we find that the proposed Long-EC^2-VAE converges more quickly during training. This indicates that the proposed training strategy leads to a better initialization and makes the performance of the model fluctuate less during training.

4.3.3 Disentanglement Evaluation

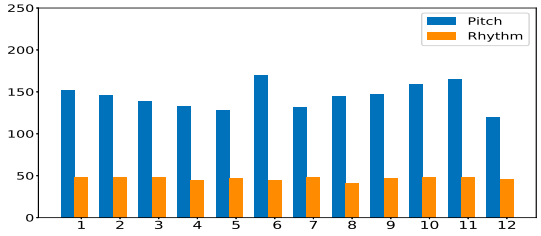
We evaluate the disentanglement performance of models using a disentanglement evaluation method adopted in [7] and [30]. The method randomly transposes all the notes of the input data by $i (i \in [1, 12])$ semitones while keeping the rhythm and underlying chord unchanged and then measures the variation of disentangled representations.We denote $\Sigma|\Delta z_{\mathrm{p}}|$ and $\Sigma|\Delta z_{\mathrm{r}}|$ as the variation of z_{p} and z_{r} .



(a) Baseline model (8 bar)



(b) Long-EC^2-VAE model (8 bar)



(c) Hirarchical-EC^2-VAE model (8 bar)

Figure 4: The comparison between $\Sigma|\Delta z_{\mathrm{p}}|$ and $\Sigma|\Delta z_{\mathrm{r}}|$ after transposition. The numbers show the pitch augmented by 12 semitones in each sub-figure from left to right.

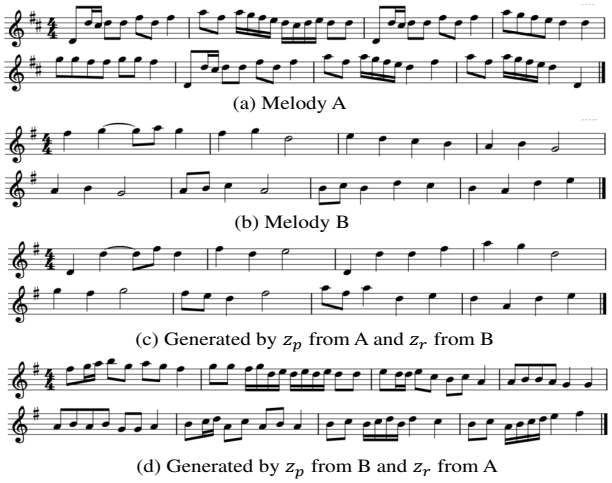
As shown in Figure 4, values of $\Sigma|\Delta z_{\mathrm{p}}|$ of the proposed Hierarchical-EC^2-VAE are relatively high while $\Sigma|\Delta z_{\mathrm{r}}|$ maintains in a significantly low level. This indicates that the pitch and rhythm representations of the proposed Hierarchical-EC^2-VAE are well-disentangled as the change of notes has a tiny impact on z_{r}. Similarly, we can intuitively find in the figure that the disentanglement performance of the proposed Hierarchical-EC^2-VAE is much better than the baseline and also outperforms the proposed Long-EC^2-VAE.

4.4 Music generative examples

In this section, we show some music generation results by manipulating the disentangled phrase-level pitch and rhythm representations in three different ways: style transfer via swapping the representation, rhythm morphing via interpolating the representation, and theme variation via representation posterior sampling.

4.4.1 Phrase-level composition style transfer

We cross-swap the disentangled pitch and rhythm factors z_{p} and z_{r} of two 8-bar melodies A and B and then obtain generative pieces C and D. The results are shown in Figure 5, in which we see that both of the two generative pieces perfectly inherit target rhythm patterns. Besides, these generative melodies vary slightly from the source melody and these variations tend to sound creative, i.e. the appearance of embellished notes.



(a) Melody A

(b) Melody B

(c) Generated by z_p from A and z_r from B

(d) Generated by z_p from B and z_r from A

Figure 5: Style transfer examples by hierarchical-EC^2-VAE model.

4.4.2 Latent z_{r} interpolation

We interpolate rhythm representations z_{r} of two phrases using SLERP [31] while keeping the pitch and chord unchanged. The interpolated latent representations can then be "re-synthesized" using Hierarchical-EC^2-VAE.

As shown in Figure 6, we interpolate z_{r} of the piece A and B with different SLERP weights. The results exhibit a surprising sense of coherence of pitch and rhythm in generative melodies, even in the transition between consecutive bars.This implies that a longer-term representation is also adept at modeling short-term generation and even contains more global harmonic information than a short-term representation.

4.4.3 Theme variation

We can also achieve theme variation by adding a Gaussian noise to z_{r} while keeping z_{p} unchanged . As a sample shown in Figure 7, we find that as the variance of the noise grows larger, the pitch and rhythm of the generative melody are still reasonable smooth, implying that the long-term representations contain the coherence of contextual information and can "control" the generation process.

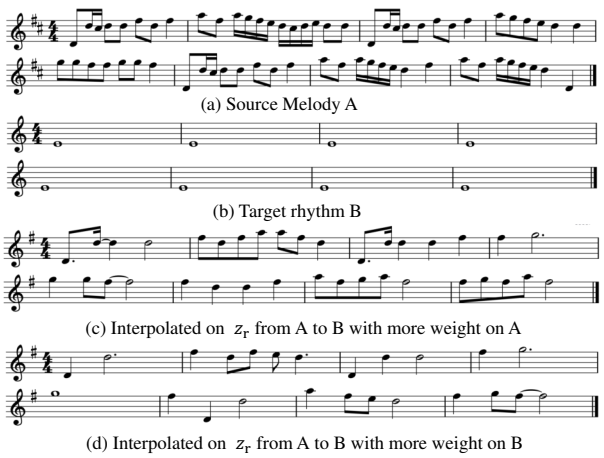


Figure 6: Interpolation examples.

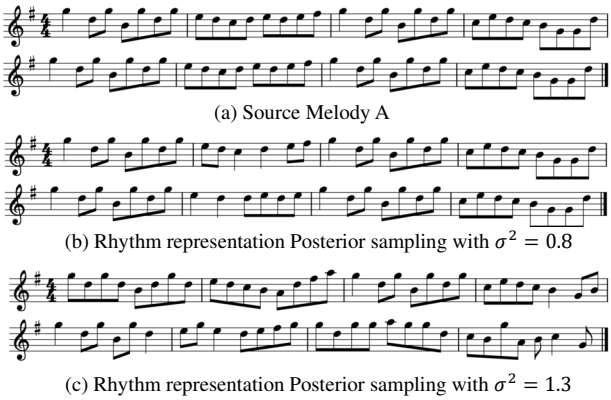


Figure 7: Rhythm representation posterior sampling examples.

4.5 Subjective Evaluation

One may wonder what are the advantages of learning long-term representations since we can always generate the music bar by bar using short-term models and just concatenate the generated samples together. One merit lies in the coherency in controlled music generation. For example, when sampling the long-term rhythm representation, the overall rhythm pattern of a phrase is controlled as an organic whole, while individually sampling the rhythm of different bar may easily lose the rhythm coherency. To better illustrate this idea,we conduct a survey on theme variation (as in Section 4.4.3) to compare the performance of the proposed 8-bar Hierarchical-EC^2-VAE and baseline 2-bar EC^2-VAE.

4.5.1 Survey Configuration

In our survey, each subject is given 5 groups of pieces. Each group contains three 8 bar pieces: a human-composed piece from Nottingham dataset and 2 theme variations generated by a 2-bar EC^2-VAE and Hierarchical-EC^2-VAE, respectively. In each group, the generated pieces use z_p of the human-composed piece and the sampled z_r.

Each subject listens to five randomly arranged groups in turn and is required to rate each melody ranging from 1 (very low) to 5 (very high) according to three aspects: *creativity*, *naturalness* (how human-like the composition is) and overall *musicality*.

4.5.2 Results

A total of 29 subjects (18 females and 11 males) participated in the survey. Experimental results depicted in Figure 8 demonstrate that people prefer melodies generated by the proposed Hierarchical-EC^2-VAE to those generated by the 2 bar EC^2-VAE [7], implying the effects of a long-term coherence learned by our model. The heights of bars represent means of the ratings and the error bars represent the MSEs computed via within-subject ANOVA [32]. The results show that our model performs significantly better than the 2 bar EC^2-VAE in terms of all three dimensions($p < 0.05$). Besides, the qualities of melodies generated by the proposed Hierarchical-EC^2-VAE reach a competitive standard compared to the human-composed pieces, especially in *creativity*.

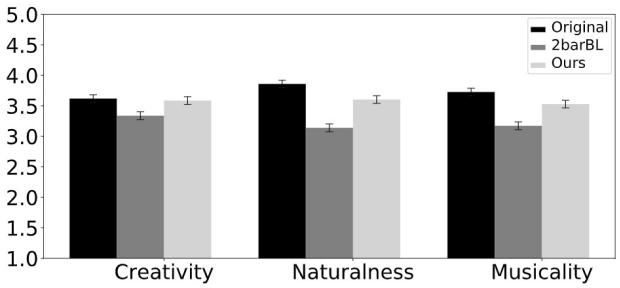


Figure 8: The results of the subjective evaluation.

5. CONCLUSION

In conclusion, we contribute a pipeline of algorithms to learn long-term and disentangled music representations. The main novelty lies in the proposed two inductive biases which constrain the long-term representations using contextual information. The first one requires long-term representation to be not too different from the short-term ones which represent a part of the long-term sequence, and we demonstrate contrastive loss is well-suited for such rough constraint. The second inductive bias is that a good long-term representation should be able to reconstruct the corresponding short-term ones, and we use a hierarchical predictive model to realize this constraint. Unlike most hierarchical models, our purpose is not prediction for its own sake, but rather to leverage the prediction power to learn a well-disentangled long-term representation. Experimental results show that our approach is quite successful, capable of disentangling pitch and rhythmic factors for phrase-level (32 beats) melody without increasing the dimensionality of latent representation compared to bar-level models. Moreover, the learned representations enable high-quality phrase-level style transfer via representation swapping and theme variation by representation interpolation and posterior sampling.

6. REFERENCES

[1] A. Roberts, J. H. Engel, C. Raffel, C. Hawthorne, and D. Eck, "A hierarchical latent vector model for learning long-term structure in music," in *Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden*, 2018, pp. 4361–4370.

[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA*, 2017, pp. 5998–6008.

[4] K. Chen, G. Xia, and S. Dubnov, "Continuous melody generation via disentangled short-term representations and structural conditions," in *14th International Conference on Semantic Computing, San Diego, CA, USA*, 2020, pp. 128–135.

[5] Z. Wang, D. Wang, Y. Zhang, and G. Xia, "Learning interpretable representation for controllable polyphonic music generation," in *Proceedings of the 21th International Society for Music Information Retrieval Conference, Montreal, Canada*, 2020, pp. 662–669.

[6] A. Pati, A. Lerch, and G. Hadjeres, "Learning to traverse latent spaces for musical score inpainting," in *Proceedings of the 20th International Society for Music Information Retrieval Conference, Delft, The Netherlands*, 2019, pp. 343–351.

[7] R. Yang, D. Wang, Z. Wang, T. Chen, J. Jiang, and G. Xia, "Deep music analogy via latent representation disentanglement," in *Proceedings of the 20th International Society for Music Information Retrieval Conference, Delft, The Netherlands*, 2019, pp. 596–603.

[8] Y. Huang and Y. Yang, "Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions," in *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA*, 2020, pp. 1180–1188.

[9] G. Brunner, A. Konrad, Y. Wang, and R. Wattenhofer, "MIDI-VAE: modeling dynamics and instrumentation of music with applications to style transfer," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, Paris, France*, 2018, pp. 747–754.

[10] T. Akama, "Controlling symbolic music generation based on concept learning from domain knowledge," in *Proceedings of the 20th International Society for Music Information Retrieval Conference, Delft, The Netherlands*, 2019, pp. 816–823.

[11] Y. Luo, K. Agres, and D. Herremans, "Learning disentangled representations of timbre and pitch for musical instrument sounds using gaussian mixture variational autoencoders," in *Proceedings of the 20th International Society for Music Information Retrieval Conference, Delft, The Netherlands*, 2019, pp. 746–753.

[12] J. Jiang, G. Xia, D. B. Carlton, C. N. Anderson, and R. H. Miyakawa, "Transformer VAE: A hierarchical model for structure-aware and interpretable music representation learning," in *International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain*, 2020, pp. 516–520.

[13] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning*, 2020, pp. 1597–1607.

[14] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA*, 2020, pp. 9726–9735.

[15] X. Chen, H. Fan, R. B. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *CoRR*, vol. abs/2003.04297, 2020.

[16] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, ser. JMLR Proceedings, vol. 9, 2010, pp. 297–304.

[17] A. Mnih and K. Kavukcuoglu, "Learning word embeddings efficiently with noise-contrastive estimation," in *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, United States*, 2013, pp. 2265–2273.

[18] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018.

[19] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," in *7th International Conference on Learning Representations,New Orleans, LA*, 2019.

[20] F. Lerdahl and R. Jackendoff, "A generative theory of tonal music," *Journal of Aesthetics and Art Criticism*, vol. 9, no. 1, pp. 72–73, 1996.

[21] W. Rothstein, O. Jonas, and J. Rothgeb, "Introduction to the theory of heinrich schenker: The nature of the musical work of art," *Journal of Music Theory*, vol. 27, no. 2, 1983.

[22] M. Hamanaka, K. Hirata, and S. Tojo, "Implementing "a generative theory of tonal music"," *Journal of New Music Research*, vol. 35, no. 4, pp. pp. 249–277, 2006.

[23] A. Marsden, "Schenkerian analysis by computer: A proof of concept," *Journal of New Music Research*, vol. 39, no. 3, pp. 269–289, 2010.

[24] H. H. Tan and D. Herremans, "Music fadernets: Controllable music generation based on high-level features via low-level feature modelling," in *Proceedings of the 21th International Society for Music Information Retrieval Conference, Montreal, Canada*, 2020, pp. 109–116.

[25] G. Hadjeres and L. Crestel, "Vector quantized contrastive predictive coding for template-based music generation," *CoRR*, vol. abs/2004.10120, 2020.

[26] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *CoRR*, vol. abs/2005.00341, 2020.

[27] E. Foxley, "Nottingham database," 2011.

[28] Z. Wang, K. Chen, J. Jiang, Y. Zhang, M. Xu, S. Dai, and G. Xia, "POP909: A pop-song dataset for music arrangement generation," in *Proceedings of the 21th International Society for Music Information Retrieval Conference, Montreal, Canada*, 2020, pp. 38–45.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations*, Y. Bengio and Y. LeCun, Eds., 2015.

[30] H. Kim and A. Mnih, "Disentangling by factorising," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 2654–2663.

[31] S. Hollasch, "Advanced animation and rendering techniques: By alan watt and mark watt, acm press," *Computers & Graphics*, vol. 18, no. 2, p. 249, 1994.

[32] H. Scheffé, "The analysis of variance," in *Architectural Institute of Japan*, 1999.