BUTTER: A Representation Learning Framework for Bi-directional Music-Sentence Retrieval and Generation

Yixiao Zhang **Ziyu Wang** **Dingsu Wang** **Gus Xia**
Music X Lab, NYU Shanghai
{yz6492, zz2417, dw1920, gx219}@nyu.edu

Abstract

We propose BUTTER, a unified multimodal representation learning model for **Bi**-directional m**U**sic-sen**T**ence Re**T**rieval and Gen**ER**ation. Based on the variational autoencoder framework, our model learns three interrelated latent representations: 1) a latent music representation, which can be used to reconstruct a short piece, 2) keyword embedding of music descriptions, which can be used for caption generation, and 3) a cross-modal representation, which is disentangled into several different attributes of music by aligning the latent music representation and keyword embeddings. By mapping between different latent representations, our model can search/generate music given an input text description, and vice versa. Moreover, the model enables controlled music transfer by partially changing the keywords of corresponding descriptions.[1]

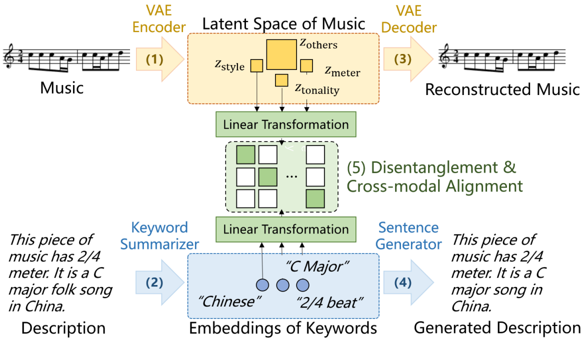


Figure 1: An overview of the model architecture.

1 Introduction

The ability to relate natural language descriptions with music is of great importance. It is useful for cross-modal *music analysis*, such as automatic music captioning and music retrieval based on natural language queries. It also has considerable research value in cross-modal *controlled music generation*, say, automatically compose a piece of music according to text descriptions.

While traditional machine-learning algorithms mostly consider analysis (from data to labels) and controlled generation (from labels to data) two completely different tasks, recent progress in multimodal representation learning (Baltrušaitis et al., 2018) suggests that the two tasks can be unified into a single framework. Specifically, music and the corresponding text descriptions can be regarded as data (of two different modalities) with shared latent representations. Therefore, if we can successfully learn the shared cross-modal representation, music analysis and generation would simply refer to various ways of mapping between such representation and data of different modalities.

Inspired by the idea above, we contribute a multimodal representation learning model for bi-directional music-sentence retrieval and generation. Figure 1 shows the overall model architecture. Here, the yellow path shows music representation learning, the blue path shows keywords summarization and text generation of music description, and the green part represents the cross-modal alignment between the latent music space and keyword embeddings. The cross-modal alignment helps disentangle the latent music representation into four factors: meter, key, style, and others, in which the first three factors have corresponding keywords descriptions. During the inference time, this model enables a number of applications:

- **Task 1: Music retrieval by text description**, where the information flow is (1)→(5)←(2). That is, to search the music segments whose latent representations best correlated with the keywords of the text description in the cross-modal space.
- **Task 2: Music captioning**, where the informa-

[1] Codes are available at https://github.com/ldzhangyx/BUTTER

tion flow is (1)→(5)→(4). E.g., to output "this is a British style song in C major of 4/4 meter" given a segment of music.

- **Task 3: Controlled music generation**, where the information flow is (2)→(5)→(3). This is very similar to task 1, except that we are now generating/sampling new music rather than searching existing music in the dataset.
- **Task 4: Controlled music refinement (style transfer)**, where the information flow is (2)→(5)→(3)←(1). That is, to learn the latent representation of piece and then refine it by partially changing the keywords of the text description. This task helps us answer the imaginary questions, such as "*what if* a piece is composed in a different style" .

In sum, the contributions of our paper are:

- We designed the first multimodal representation-learning framework which unifies music analysis and controlled music generation.
- We show that music-text alignment in the latent space serves as an effective inductive bias for representation disentanglement. Such disentanglement leads to controllable attributes of music via natural language under weak supervision with no need to do feature engineering for each separated attribute.

2 Related Work

Cross-modal retrieval task has attracted researchers for decades(Gudivada and Raghavan, 1995). Frome et al. (2013) uses a ranking cost to map images and phrases into a common semantic embedding. Yu et al. (2019) uses CCA to model cross-modal relation between audio and lyrics for bi-directional retrieval. Feng et al. (2014) learns multi-modal representations by correlating hidden representations of two uni-modal autoencoders and minimizes a linear combination error. Karpathy et al. (2014) proposed a bidirectional image-sentence mapping method by extracting local fragments. Compared with cross-modal retrieval, cross-modal controlled generation is in general a more difficult task since it requires reconstruct or sample new data from the latent representation. Recent works include automatic image captioning(Xu et al., 2015; Chen and Lawrence Zitnick, 2015; Jia et al., 2015) and text-to-image generation(Hinz et al., 2019; Zhu et al., 2019; El-Nouby et al., 2019). However, very few of them consider the bi-directional generation problem.

Another related area to this study is representation disentanglement. Locatello et al. (2019) shows that the key to a successful disentanglement is to incorporate the model with proper inductive biases. Speaking of the disentanglement for music, Deep Music Analogy (Yang et al., 2019) is very relevant to this study as it can disentangle pitch and rhythm factors. However, the inductive bias comes from a supervision (an explicit rhythm loss), while our study uses text descriptions as a much weaker supervision as well as a more natural form of inductive bias.

3 Method

3.1 Music Modality

We use a similar data representation as in MusicVAE (Roberts et al., 2018). Each 16-beat melody segment x is represented as a sequence of 64 one-hot vectors. Each vector represents a 16th note and has 130 dimensions, representing 128 MIDI pitches, hold and rest, respectively.

We use the VAE framework to learn the latent code z of a melody segment (as shown in (1) and (3) of Figure 1). We assume z conforms to a standard Gaussian prior (denoted by $p(z)$), and can be partitioned into four disentangled factors $z = [z_{\text{key}}, z_{\text{meter}}, z_{\text{style}}, z_{\text{others}}]$, where z_{others} represents the music information not covered by key, meter or style. The VAE encoder uses a single layer bi-directional GRU to encode the melody and emit the mean and variance of the approximated posterior $q_\theta(z|x)$. We assume $q_\theta(z|x)$ is isotropic Gaussian and denote its mean as $e = [e_{\text{key}}, e_{\text{meter}}, e_{\text{style}}, e_{\text{others}}]$. For the VAE decoder, we apply a 2-layer GRU which outputs $p_\theta(x|z)$.

We define the *reconstruction objective* by the ELBO (evidence lower bound) (Kingma and Welling, 2013) as follows,

$$\mathcal{L}_r(\phi, \theta; x) = -\mathbb{E}_{z\sim q_\phi} \log p_\theta(x|z) + \alpha \text{KL}\Big(q_\phi || p(z)\Big), \tag{1}$$

where α is a balance parameter.

3.2 Language Modality

3.2.1 Keywords Representations

We define the keywords of a music description as a triplet $[w_{\text{key}}, w_{\text{meter}}, w_{\text{style}}]$, where $w_{\text{key}} \in D^{\text{key}}$, $w_{\text{meter}} \in D^{\text{meter}}$, and $w_{\text{style}} \in D^{\text{style}}$. Here D^{key}, D^{meter}, and D^{style} are the dictionaries of the three corresponding attributes. We define the overall dictionary $D = D^{\text{key}} \cup D^{\text{meter}} \cup D^{\text{style}}$ and embed every keyword $w \in D$ to a $|D|$ dimensional one-hot vector e'_w.

3.2.2 Summrizer and Generator Module

We apply two GRU-based encoder-decoder models as the keyword summarizer and the description generator, respectively. Both models are pretrained with sentence-keywords pairs directly retrieved from the dataset. This procedure is shown in (2) and (4) in Figure 1.

3.3 Cross-modal Alignment

We use two linear transformations $f(\cdot), g(\cdot)$ to map latent melody representation and keywords to a shared latent space. We employ a *similarity objective* $\mathcal{L}_a$ to align two representations by maximizing correlation of corresponding attributes while minimizing the correlation of irrelevant attributes. Formally,

$$\mathcal{L}_a = 3 - \sum_{i \in \mathcal{I}} \Big(\frac{\langle u_i, v_{w_i} \rangle}{|u_i| \cdot |v_{w_i}|} - \beta \sum_{\substack{w \in D \\ w \neq w_i}} \Big| \frac{\langle u_i, v_w \rangle}{|u_i| \cdot |v_w|} \Big| \Big), \tag{2}$$

where $\mathcal{I} = \{\text{key}, \text{meter}, \text{style}\}, u_i = f(e_i)$, and $v_w = g(e'_w)$. Here, 3 is a constant to keep the loss term non-negative and β is a balance factor. Hence, the overall loss $\mathcal{L}$ is calculated by $\mathcal{L} = \mathcal{L}_r + \gamma \mathcal{L}_a$, where γ is a parameter for balancing two losses.

In theory, the cross-modal alignment module allows bi-directional music-sentence retrieval and generation by inference-time optimization of eq. 2. In practice, we find the keyword combination that best describes a given melody (i.e., minimizes eq. 2) by a brute-force search. Conversely, we compute the latent music code corresponding to a keyword by averaging the latent codes of all music samples aligned with the same keyword.

4 Experiments

We conduct two experiments to demonstrate that the proposed model can be applied to the four tasks mentioned in the introduction.

The former two tasks, i.e., music retrieval by text description and music captioning are both about *music analysis*. The core of the two tasks requires the latent code being able to classify to the correct keywords. Our first experiment (Section 4.2) focuses on this classification accuracy.

The latter two tasks, controlled music generation and controlled music refinement, require that by changing the latent codes (e.g., from minor to major key), the generated samples also have the corresponding change (e.g., key change) *while still preserving high music quality*. We conduct subject evaluations regarding this aspect in Section 4.3.

4.1 Dataset and Training

Our dataset contains 16,257 folk songs paired with metadata collected from the abc notation homepage[2]. From metadata we select *key*, *meter* and *style* as keywords, and we synthesize diverse description sentences by human craft and paraphrasing tools[3]. We associate them with 4-bar music segments of corresponding songs. We use 80% for training, 10% for validation and 10% for testing.

We train our model on two keyword settings. In the *full version*, the key keyword contains 25 classes including 24 major/minor keys and *others*; the meter keyword contains 6 classes including *2/4*, *3/4*, *4/4*, *6/8*, *9/8* and *others*. In the *easy version*, the key keyword contains *major*, *minor*, and *others*; the meter keyword contains *triple*, *duple*, and *others*. In both modes, the style keyword contains 3 classes, including *Chinese*, *English*, *Irish*.

We set the size of GRU hidden states, latent variables and attribute variables to 512, 256 and 32 respectively. We map the latent z and words to a shared 32-D embedding space. During training, we set batch size to 4 and learning rate to $1e{-}3$ with weight decay of 0.999. We set balancing factors $\alpha = 0.01, \beta = 0.2$ and $\gamma = 0.1$.

4.2 Objective Measurements for Cross-modal Music Information Retrieval

We design a classification task to evaluate whether the latent codes e_{key}, e_{meter} and e_{style} can predict the corresponding keywords by the similarity objective eq. 2. If so, it follows that with simple algorithms the model is capable of the task *music retrieval by text description* and *music captioning*.

To this end, we compare our models with 2 baseline classifiers under both *full version* and *easy version*. Both baseline models uses the same GRU encoder ((1) in Figure 1) and replace the alignment module ((5) in Figure 1) by three separate MLP classifiers for the three keyword attributes accordingly. Each MLP has 3 linear layers with 128 hidden dimensions. The first baseline method (**GRU-MLP**) trains the whole network from scratch, and the second baseline (**Latent-MLP**) method trains only the MLP classifiers and fixes the GRU encoder

[2] http://abcnotation.com/

[3] https://quillbot.com/

Model	Key		Meter		Style
	Full	Easy	Full	Easy	
GRU-MLP	**0.63**	0.76	**0.44**	0.54	0.84
Latent-MLP	0.45	**0.83**	0.38	0.72	0.90
Ours	0.60	0.77	0.40	**0.76**	**0.92**

Table 1: Performance of models in classification task.

parameters. Table 1 shows the evaluation results.

When all the keywords of the melody are determined, our model generates complete sentences through the description generator. Figure 2 provides two generated examples.

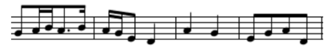
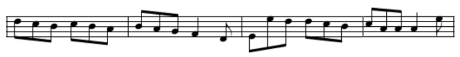
Melody	
Caption	This is a song in **G**. Is has a **2/4** meter and it is a **Chinese** song.
Melody	
Caption	This is a **6/8**-meter composition in **G major**. It is an **Irish** song.

Figure 2: Generated descriptions for input melodies.

4.3 Subjective Evaluation for Controlled Music Generation

We wish that when we change one or more keyword attributes, the generated music would also make corresponding change while still preserving good musicality. The *controlled music generation* and *controlled music refinement* tasks directly follow from this desired property.

We invite people to subjectively rate the quality of the generated music. In particular, we ask the subjects to listen to 30 samples randomly picked from the test dataset with three types of processing, and each type contains 10 samples:

1. **(Original)** No processing: identical to the data sample.

2. **(Ours)** Randomly change the latent code (among e_{key}, e_{meter} and e_{style}) into a target latent code. E.g. to substitute $e_{\text{key=major}}$ as $e_{\text{key= minor}}$

3. (Prior) Randomly change the latent code into Gaussian noise sampled from the prior distribution.

The subjects are asked to:

1. Rate the musicality of the processed sample based on a 5-point scale from 1 (low) to 5 (high).

2. Select the keyword attributes that best describe the music. That is, 1) whether the *key* is major, minor or others, 2) whether the *meter* is duple, triple or others, and 3) whether the *style* is Chinese, Irish or English.

Model	Musicality	Human Accuracy		
		Key	Meter	Style
Original	3.44	0.57	0.60	0.38
Prior	2.68	0.33	0.43	0.31
Ours	**3.25**	**0.35**	**0.48**	**0.49**

Table 2: Performance of models in generation task. Musicality means the overall quality of music.

A total of 30 subjects (9 female and 21 male) participated in the survey. 10% of them are at the professional level of musicological knowledge and the 37% are over the average level. In table 2, the left column shows the rating of musicality and the right column shows the accuracy of selecting the correct keywords. The results show that our proposed method has higher musicality and achieves better control of generation than randomly sampling from the prior in all three factors. However, we still see a gap between our method and the original samples. This is probably because the selected factors deal with deep music structure which remains a challenging task for existing methods. which we leave for future work. Moreover, due to the cultural background of subjects, they generally have difficulty in distinguishing between Irish and English songs. If we combine these two categories into one, then the scores of the original songs, ours, and baseline are: 0.72, 0.59 and 0.32, respectively. Figure 3 shows a transfer example:

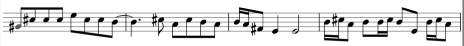
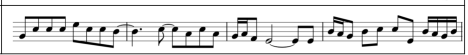
Origin (Chinese Style)	
Culture Transfer (to English)	

Figure 3: An example of music refinement.

5 Conclusion and Limitations

In conclusion, we contributed a unified multimodal representation learning model allowing bidirectional retrieval and generation between music and sentences. The cross-modal alignment serves as an effective inductive bias to disentangle latent representations of music according to text.

We see that the current text description is still very rigid, limited to three keywords and the text descriptions have to cover the exact keywords. In the future, we will make the text description more flexible, covering more music attributes while allowing synonyms. In addition, human descriptions of music may be subjective and ill-defined, making the learning process difficult. It will be the biggest challenge our model faces in the future.

References

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence, 41(2):423–443.

Xinlei Chen and C Lawrence Zitnick. 2015. Mind's eye: A recurrent visual representation for image caption generation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2422–2431.

Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W Taylor. 2019. Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. In Proceedings of the IEEE International Conference on Computer Vision, pages 10304–10312.

Fangxiang Feng, Xiaojie Wang, and Ruifan Li. 2014. Cross-modal retrieval with correspondence autoencoder. In Proceedings of the 22nd ACM international conference on Multimedia, pages 7–16.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In Advances in neural information processing systems, pages 2121–2129.

Venkat N Gudivada and Vijay V Raghavan. 1995. Content based image retrieval systems. Computer, 28(9):18–22.

Tobias Hinz, Stefan Heinrich, and Stefan Wermter. 2019. Semantic object accuracy for generative text-to-image synthesis. arXiv preprint arXiv:1910.13321.

Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. 2015. Guiding the long-short term memory model for image caption generation. In Proceedings of the IEEE international conference on computer vision, pages 2407–2415.

Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In Advances in neural information processing systems, pages 1889–1897.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In international conference on machine learning, pages 4114–4124.

Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. 2018. A hierarchical latent vector model for learning long-term structure in music. In ICML.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning, pages 2048–2057.

Ruihan Yang, Dingsu Wang, Ziyu Wang, Tianyao Chen, Junyan Jiang, and Gus Xia. 2019. Deep music analogy via latent representation disentanglement. arXiv preprint arXiv:1906.03626.

Yi Yu, Suhua Tang, Francisco Raposo, and Lei Chen. 2019. Deep cross-modal correlation learning for audio and lyrics in music retrieval. ACM Trans. Multimedia Comput. Commun. Appl., 15(1).

Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. 2019. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5802–5810.