Chord Progression Imitator

Zehao Wang
School of Mathematical Sciences, Peking University

Abstract

In this project, we mainly focus on VAE-based disentanglement representation learning methods for music applications. We introduce the proposed framework, purified-VAE, for sub-task in music generation—-chord progression generation/analogy. To decompose the basic information (chord qualities) and tensions of given chord progressions, we use a two-stage training strategy for a two-part VAE framework using chromagram and TIV-gram (tonal interval vector) as input. Both subjective experiment and ablation study showed that our method is effective for chord progression generation.

A demonstration for this framework is shown in Fig. 1 For the sake of brevity, we will not introduce detailed knowledge of music theory.

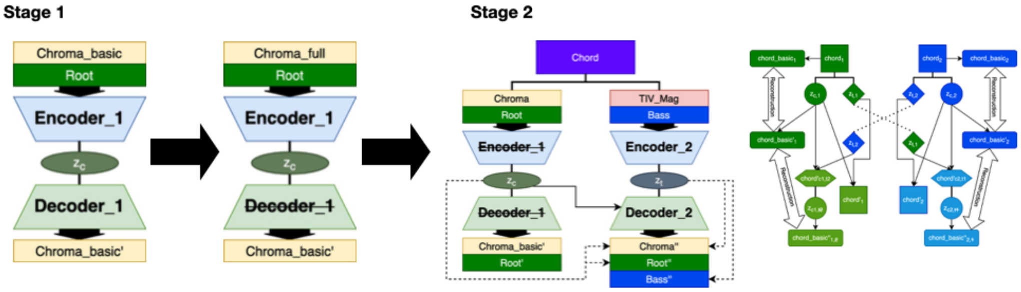


Figure 1: Purified-VAE framework.

1 Data format

The input data for the model has two dimensions (time× something). For y-axis, it has three formats: chroma, TIV, and one-hot vector which denotes a musical note (mod 12).

1.1 Chroma

Chroma(1) is a projection of pianoroll which is a digitalized piano keyboard. Common pianoroll p is a 0-1 vector which has 128 dimensions for 128 different keys. If we modulo 12 for all indices of pianoroll, and project all the notes in a new 0-1 12-d vector c as follows:

$$c[i] = \prod_{0\leq k\leq 128,\ k\equiv i \pmod{12}} p[i], \quad i = 0, 1, ..., 11. \tag{1}$$

Combining a series of chroma in chronological order will get a chromagram.

1.2 Tonal Interval Vector (TIV)

Tonal interval vector aims to describe the periodicity in tonal harmony theory(2). This feature mainly focus on modeling the pitch relatedness and musical consonance which fits our goal for modeling tensions of chord progression very much. This feature can be derived from chroma through Discrete Fourier Transform (DFT) as follows,

$$T(k) = w_*(k) \sum_{n=0}^{N-1} \bar{c}(n) e^{\frac{-j2\pi kn}{N}} \quad ,k \in \mathbb{Z} \quad with \quad \bar{c}(n) = \frac{c(n)}{\sum_{n=0}^{N-1} c(n)}, \tag{2}$$

where $N = 12$ is the dimension of the input chroma vector c and $w_*(k)$ are weights derived from empirical ratings of dyads consonance used to adjust the contribution of each dimension k of the space. It can assume $w_s(k) = 2, 11, 17, 16, 19, 7$ and $w_a(k) = 3, 8, 11.5, 15, 14.5, 7.5$, for symbolic and audio inputs, respectively. We set k to $1k6$ for $T(k)$, since the remaining coefficients are symmetric.

Consider the mathematical properties of TIV $T \in \mathbb{C}^6$, the magnitude of TIV $TIV_{mag} = ||T||$ and the phase of TIV $TIV_{phase} = \angle T$ can represent the chord qualities (translation-invariant) and the relative position of this chord, respectively.

The definition of TIV-gram is similar to chromagram in previous subsubsection.

1.3 One-hot vectors for root and bass

In music theory, root and bass are two major components for describing a chord. Root denotes the lowest note for the original chord without inversion and tensions, and bass denotes the lowest note for the actual chord in a musical piece. Therefore, we use a 12-d one-hot vector to represent the absolute position of root note, and a 12-d one-hot vector for the relative position of bass note (relative to root) which makes sense in harmony theory.

1.4 Input format for the proposed model

In Fig. 1, we have two types of input: chord_* + Root $\in \mathbb{R}^{t\times(12+12)}$ and TIV_Mag+Bass $\in \mathbb{R}^{t\times(6+12)}$. + means concat through y-axis and * $\in$ {basic, full, N/A}. All the input is in gram type, which means it has two dimensions (time$\times$ note). In this project, we use 8-bar musical pieces for both training and inference, and each bar is divided to quarterlength time unit, which means our input time dimension $t = 4 * 8 = 32$.

2 Dataset

We use two datasets for this project: Lead-sheet-dataset[1] and Weimar Jazz Database[2](3). The common characteristics of these two datasets are (1) their musical pieces are from pop songs; (2) all the chord progressions have at least 10% tensions. Fig. 2 is a statistics for Lead-sheet-dataset. Because we said our input is 8-bar musical pieces above, we will only use all 8-bar music in these two datasets for this project.

3 Model

A framework of our final model is shown in Fig. 1. For each encoder and decoder, the basic architecture is a Gated recurrent unit (GRU):

$$\begin{aligned}
&\mathbf{R}_t = \sigma\left(\mathbf{X}_t\mathbf{W}_{xr} + \mathbf{H}_{t-1}\mathbf{W}_{hr} + \mathbf{b}_r\right)\\
&\mathbf{Z}_t = \sigma\left(\mathbf{X}_t\mathbf{W}_{xz} + \mathbf{H}_{t-1}\mathbf{W}_{hz} + \mathbf{b}_z\right)\\
&\tilde{\mathbf{H}}_t = \tanh\left(\mathbf{X}_t\mathbf{W}_{xh} + \left(\mathbf{R}_t \odot \mathbf{H}_{t-1}\right)\mathbf{W}_{hh} + \mathbf{b}_h\right)\\
&\mathbf{H}_t = \mathbf{Z}_t \odot \mathbf{H}_{t-1} + (1 - \mathbf{Z}_t) \odot \tilde{\mathbf{H}}_t,
\end{aligned} \tag{3}$$

where $\mathbf{X}_t$ is the t-step of the input feature in time dimension, $\mathbf{H}_t$ is the hidden state, $\mathbf{Z}_t$ is the gate and others are parameters. We sequentially feed the data into each model and also sequentially decode

[1] https://github.com/wayne391/lead-sheet-dataset

[2] https://jazzomat.hfm-weimar.de/dbformat/dboverview.html

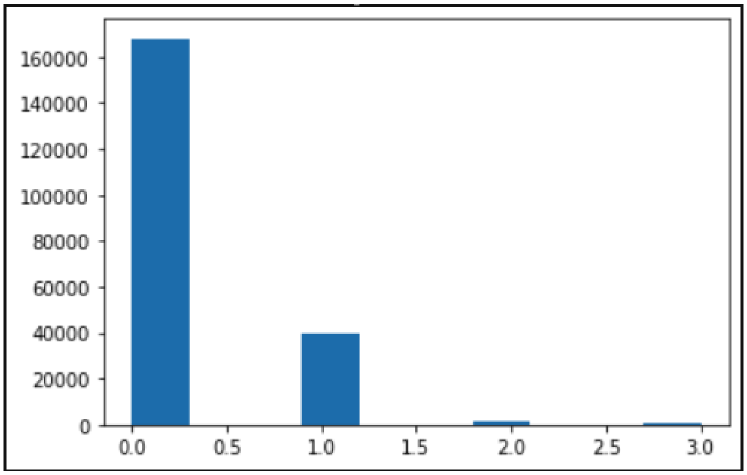


Figure 2: Statistics for Lead-sheet-dataset. x-axis means the total number of notes for a given chord, if it is more than 0, it must have tensions. y-axis means the total amount of the class.

them from the decoder. The representation for disentanglement is got from a fully-connected layer in $\mathbb{R}^{1024\times 128}$ after the last hidden state of our time series $\mathbf{X} = (\mathbf{X}_0, \mathbf{X}_1, ..., \mathbf{X}_3 1)$, i.e. $\mathbf{H}_{31}$. The hidden state is 1024-d.

3.1 A two-stage training strategy

Here we introduce our training strategy. The prior distributions for two encoder-decoder are both Gaussian. The operations of feature extraction like chroma(gram) and TIV(_Mag) will be ignored in the following content.

Notations:

- $\cancel{F}$ means freezing the parameters of F during this training process. Heuristically, the newly trained Enc_1 has the function of purification due to the frozen parameters of Dec_1.
- $\bar{x}$ denotes the reconstructed x.
- c' denotes the basic-version chord progression of c.

The first stage:

The first stage aims to train the left encoder-decoder to get a Encoder_1 which can purify an input chord progression to its basic version (which means only have basic three-note chord qualities: maj, min, dim, aug, sus). The basic version of chord progression can be extracted by rule-based algorithms, so the first stage is a supervised algorithm. There are two sub-stages for the first stage.

The first sub-stage:

- Input: basic-version chord progression c', Output: basic-version chord progression $\bar{c'}$
- Loss: $recon(Dec_1(Enc_1(c')), c') + KL$
- Optimizer: Adam

where the reconstruction losses for one-hot vectors (root and bass), 0-1 chroma(gram) and TIV(gram) are cross-entropy, binary cross entropy and mean square loss, respectively.

After sufficient training, the ideal desired Enc_1, z_c, and Dec_1 are fit for the distribution of basic-version chord progression, i.e., three-note chord progression. Then we freeze the trained parameters of Dec_1, and start the second sub-stage:

- Input: full-version chord progression c, Output: basic-version chord progression $\bar{c'}$
- Loss: $recon(\cancel{Dec_1}(Enc_1(c)), c') + KL$

Model	rec_basic for mixed chord progression	rec_full for original chord progression
Baseline	0.67	0.97
Purified-VAE	0.88	0.91

Table 1: Test result of ablation study

- Optimizer: Adam

The second stage:

Then we freeze parameters of Enc_1 and Dec_1, and conduct an in-the-loop latent-variable-swapping during training process. The strategy is as follows,

- Input: full-version chord progression c_A, c_B
- In-the-loop latent-variable swapping:
  - for each iteration, we first get $z_{Ac} = \cancel{Enc_1}(c_A), z_{Bc} = \cancel{Enc_1}(c_B)$ and $z_{At} = Enc_2(c_A), z_{Bt} = Enc_2(c_B)$,
  - then derive $\bar{c'}_A = \cancel{Dec_1}(\cancel{Enc_1}(_2([z_{Ac}, z_{Bt}]))) = \cancel{Dec_1}(\cancel{Enc_1}(c_{\bar{A}B}))$, and $\bar{c'}_B = \cancel{Dec_1}(\cancel{Enc_1}(_2([z_{Bc}, z_{At}]))) = \cancel{Dec_1}(\cancel{Enc_1}(c_{\bar{B}A}))$
- Original reconstruction: $\bar{c}_A = Dec_2([z_{Ac}, Enc_2(c_A)]), \bar{c}_B = Dec_2([z_{Bc}, Enc_2(c_B)])$
- Loss: $recon(\bar{c'}_A, c'_A) + recon(\bar{c'}_B, c'_B) + recon(\bar{c}_A, c_A) + recon(\bar{c}_B, c_B) + KL$
- Optimizer: Adam

Every stage and sub-stage needs sufficient training, and only do one round for all stages orderly. And the baseline model used for evaluation is the model without in-the-loop latent-variable-swapping but with full TIV information.

4 Evaluation

In this section, we will report some evaluations for the proposed model.

4.1 Ablation study

After training our models, we test for the performance of bassline and current models. The baseline model removes the swapping operation, so this comparison is an ablation study. We calculate the test reconstruction accuracy for original full-version chord progression and purified result of mixed chord progression after swapping to test the effectiveness of the proposed training strategy. Result is shown in Table. 1.

The result shows that training with disentanglement by swapping will slightly decrease the reconstruction ability, but the disentanglement ability has an obvious incremental compared to the baseline model without swapping. To some extent, it illustrates the effectiveness of our proposed framework.

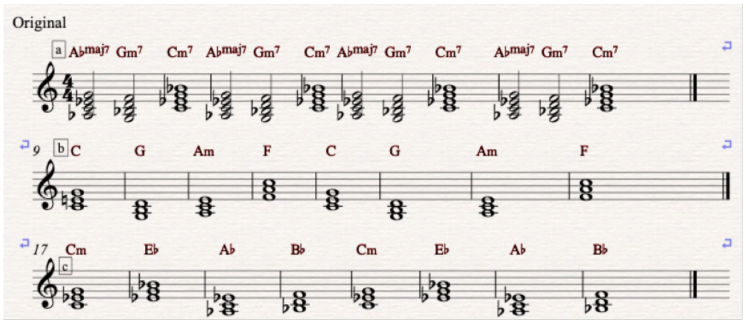
4.2 Evaluating the disentanglement

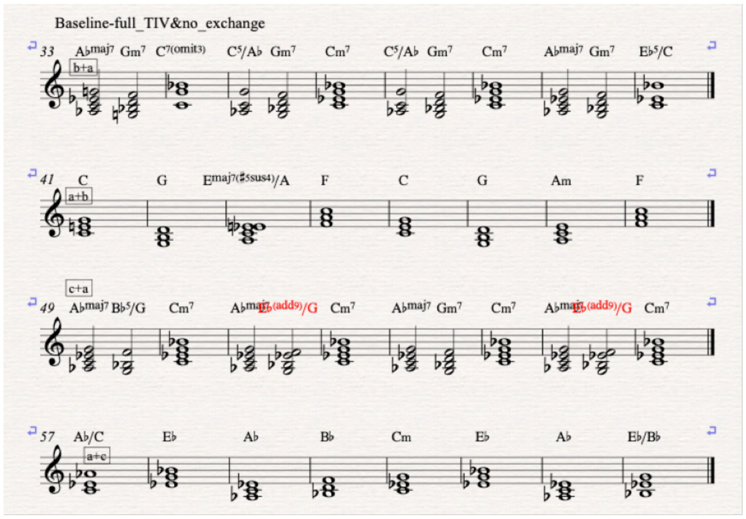
We notice that there is a paper (4) called *Evaluating the Disentanglement of Deep Generative Models through Manifold Topology* which is just accepted by ICLR2021 only one week ago. Due to time limitation in final exam weeks, we failed to complete the implementation of this evaluation to our model. It has to be left as a future work...

4.3 Demo

Subjective Analysis: For the final result, what we want is the left progression's roots and qualities and the right one's tensions.

- Consider three original progressions, all chords in progression[a] is seventh which means they are four-note chord, [b] and [c] are all three-note chord progressions. And these three



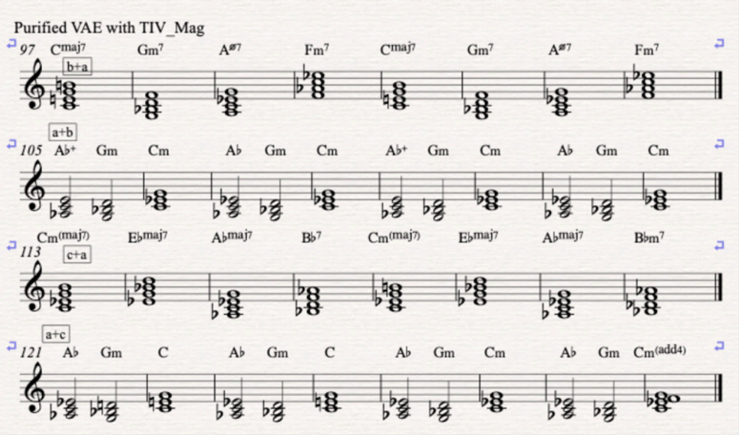

Baseline-full_TIV&no_exchange
33
b+a
Abmaj7 Gm7 C7(omit3) C5/Ab Gm7 Cm7 C5/Ab Gm7 Cm7 Abmaj7 Gm7 Eb5/C
41
a+b
C G Emaj7(#5sus4)/A F C G Am F
c+a
49
Abmaj7 Bb5/G Cm7 Abmaj7 Eb(add9)/G Cm7 Abmaj7 Gm7 Cm7 Abmaj7 Eb(add9)/G Cm7
57
a+c
Ab/C Eb Ab Bb Cm Eb Ab Eb/Bb

progressions are not in the same key, which means if you want to analyze the progressions, you need to translate them to one same key first.

- The result of baseline model shows that the roots is not constantly same as the left input chord progression. For example, roots of the first row of baseline result are [A,G,C,C,...], but roots of its left ([b]) and right ([a]) input chord progression are [C,G,A,F,...] and [A, G, C, A,...], respectively, which means the result is not our desired one (roots of final result are similar to the right one's!).
- The purified-VAE has different result, which is similar to our desired result.

5 Conclusion and Discussion

In this project, we explored the existing methods for controllable music generation and (disentanglement) representation learning methods with application on music. Then based on the current trend, we designed a VAE-based disentanglement representation learning framework with two-stage training strategy and in-the-loop latent-variable-swapping for chord progression generation, which is a sub-task in music generation/style transfer. Both ablation study and subjective analysis showed that our proposed method could be effective on disentangling the basic information (including chord qualities and roots) and tensions of a given chord progression. But the evaluation methods are not enough to prove the effectiveness of our model in all perspectives. Therefore, a possible future work is to implement the evaluation (4) to the proposed algorithm. It is also necessary to give further theoretical guarantee of in-the-loop latent-variable-swapping method.


Purified VAE with TIV_Mag
97 Cmaj7 Gm7 Aø7 Fm7 Cmaj7 Gm7 Aø7 Fm7
b+a
a+b
105 A♭+ Gm Cm A♭ Gm Cm A♭+ Gm Cm A♭ Gm Cm
Cm(maj7) E♭maj7 A♭maj7 B♭7 Cm(maj7) E♭maj7 A♭maj7 B♭m7
113
c+a
a+c
121 A♭ Gm C A♭ Gm C A♭ Gm Cm A♭ Gm Cm(add4)

References

[1] R. N. Shepard, “Circularity in judgments of relative pitch,” *The Journal of the Acoustical Society of America*, vol. 36, no. 12, pp. 2346–2353, 1964. [Online]. Available: https://doi.org/10.1121/1.1919362

[2] G. Bernardes, D. Cocharro, M. Caetano, C. Guedes, and M. E. Davies, “A multi-level tonal interval space for modelling pitch relatedness and musical consonance,” *Journal of New Music Research*, vol. 45, no. 4, pp. 281–294, 2016.

[3] M. Pfleiderer, K. Frieler, J. Abeßer, W.-G. Zaddach, and B. Burkhart, Eds., *Inside the Jazzomat - New Perspectives for Jazz Research*. Schott Campus, 2017.

[4] S. Zhou, E. Zelikman, F. Lu, A. Y. Ng, and S. Ermon, “Evaluating the disentanglement of deep generative models through manifold topology,” *arXiv preprint arXiv:2006.03680*, 2020.