Simultaneous Melody-Accompaniment Generation with musicVAE

Ziheng Chen

Abstract

In this article we propose a method to generate a combination of melody and accompaniment at the same time. The relation between the melody part and the accompaniment part is inferred from the encoded embedding of realistic dataset via a transform network on the latent space. We also discuss several different aspects involved in this procedure.

1. Introduction and Review on Music Generation Techniques

Music generation has been a popular topic and much attention has been drawn to this field (Briot & Pachet, 2017; Briot et al., 2017). One way of music generation is to directly output the audio files and (van den Oord et al., 2016) addresses a technique to generate wideband raw audio waveforms which are signals with very high temporal resolution. The idea is to use the dilated causal convolutional layers which keeps both short-term and long-term correlation in audio profiles. In (Dieleman et al., 2018), an autoregressive model and the argmax encoder is built upon Wavenet structure and the authors show that the convergence can be reliably ensured.

Another way is to generate the symbolic representation rather than the audio profile, which is more preferable since we don't need to deal with the high temporal resolution. One popular approach is to use convolutional based network structure. In (Yang et al., 2017), the authors explore the combination of conditioner and generator CNN with GAN, which leads to music generation without prior information of the melody. In (Dong & Yang, 2018), the authors design a GAN to generate realistic music piece in the multi-track piano-roll representation. The discriminator consists of three CNN parts: a main stream which compress the single-line sequence, an onset/offset stream and a chroma stream in the end.

However, due to the nature of music (repetition of melodic materials, careful arrangement of instruments, sophisticated connection between polyphonic progressions), a recurrent based network is much more advisable. In (Hadjeres & Nielsen, 2017), enforcing positional constraints becomes feasible in the newly proposed Anticipation-RNN, which makes possible a realistic replication of the style of the soprano parts of the 'Bach chorale harmonization'. The model consists a forward-propagating token chain and a back-propagating constraint chain. (Roberts et al., 2018) explores another option to address the 'vanishing latent variable' issue. The idea is to use a hierarchical RNN design for encoder and decoder which keeps the encoded latent variable changing slowly on a bar-ly basis so that it can be kept for a relatively longer time. Based on that, the work (Simon et al., 2018) extends the variational autoencoder design which enables the feature of conditioning and allows a richer instrumentation. In (Brunner et al., 2018), a hierarchical design is proposed in which the base chord LSTM produces a chord line first and the second LSTM generates the polyphonic music upin that, making the model music theory aware. The work (Mao, 2018) is based on the Biaxial LSTM Architecture, which is essentially a sequential probability model. The DeepJ network extends Biaxial LSTM by adding dynamics to the represetation which is controlled by a mean squared loss. (Chu et al., 2019) focuses on designing a hierarchical RNN network with the key, press, chord and drum layer representation. The training set is built upon user-composed pop songs and videogame music. This work also compares different patterns in Major/Minor, Harmonic Minor, Melodic Minor, and Blues chords.

There are also some interests in other aspects in this field. The work (Roy et al., 2017) targets on preserving certain musical structure and melody similarity while at the same time producing realistic samples. The underlying model is still a Markov chain in a melody-chord two-stage separation fashion. (Jaques et al., 2017) takes a RL approach to impose certain structure which ensures the sequence generation RNN is influenced by task-specific rewards as well as information learned from data is retained. In (Lattner et al., 2018), the Convolutional Restricted Boltzmann Machine is used to generate polyphonic music samples. The paper proposes a novel method to work out the challenge of keeping the constraints when performing gradient descent and sampling. (Van Der Weerdt & Schlobach, 2018) works on the generation of music based on given lyrics. The idea is to use a word2vec network as the encoder and use musicVAE proposed in (Roberts et al., 2018) to output music sampls. In (Wang & Yang, 2019), the authors use a combination of contour network and texture network to generate audio performance solely based on the score in the piano roll representation. Tn (Berthelot et al., 2019), the authors improve current autoencoder design by introducing an adversarial critic network which distinguish if the synthesized data is an interpolation or not. (Huang & Yang, 2020) proposes a novel representation of musical sequence which gives better performance in training and testing stages.

2. Proposed Network Structure

2.1. MusicVAE

As described in (Roberts et al., 2018), we can utilize the hierarchical RNN structure to extract the encoded latent information in a note sequence. To be more specific, given the input x, we characterize the encoded latent vector z by a Guassian distribution $q_\lambda(z|x)$ which has mean μ and covariance $\operatorname{diag}(\sigma)$. The two important statistical quantities, μ and σ, are computed through the following network structure:

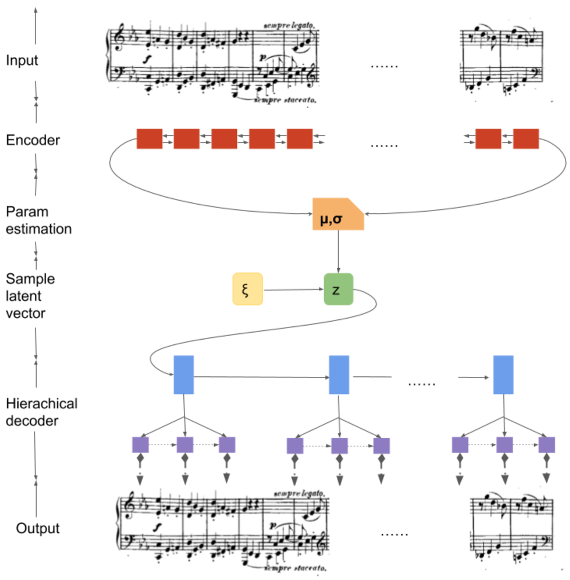

Figure 1. Schematic of the hierarchical recurrent Variational Autoencoder model (MusicVAE, (Roberts et al., 2018)). The red blocks are bi-directional LSTM nodes, which encode the input into two parameters μ and σ. Then with the help of a non-updating noise term ξ, the latent vector z is sampled and passed to the hierarchical decoder (blue and purple) in the bottom.

The posterior distribution $p(x) = \int_Z p(z)\, p(x|z)$ reflects how likely the input can be replicated under current setting of network parameters, which we wish to be as high as possible to complete the encode-decode precedure. Since the $p(x)$ is intractable, we can instead maximize the evidence lower bound (ELBO)

$$\boldsymbol{E}\left[\log p_\theta(x|z)\right] - KL\left(q_\lambda(z|x)\,||p(z)\right) \le \log p(x)\,.$$

2.2. Melody and Accompaniment Generation

Based on the forementioned discussion, we model the problem as follows: given a set of data $\left\{\left(x^1, x^2\right)_\lambda\right\}_\lambda$ and a pre-trained network $\boldsymbol{N}$ producing $q_\lambda^{1,2}(z|x)$ which maximizes the posterior distribution $p(x)$, find another network $\widetilde{\boldsymbol{N}}$ such that we can only use $\left\{x_\lambda^1\right\}_\lambda$ or $\left\{x_\lambda^2\right\}_\lambda$ (not both!) to produce $\widetilde{q_\lambda^{1,2}}(z|x)$.

Since we only allow one component of the input, it is natural to map the latent distribution (for example $q_\lambda^1(z|x)$) into another (correspondingly $q_\lambda^2(z|x)$). It is especially straightforward for gaussian distributions since we can just map the mean and covariance on one part to another.

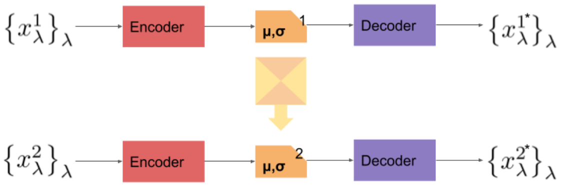
$\{x_\lambda^1\}_\lambda$
Encoder
μ,σ
1
Decoder
$\{x_\lambda^{1^*}\}_\lambda$
$\{x_\lambda^2\}_\lambda$
Encoder
μ,σ
2
Decoder
$\{x_\lambda^{2^*}\}_\lambda$

Figure 2. Network structure for simultaneous melody-accompaniment generation.

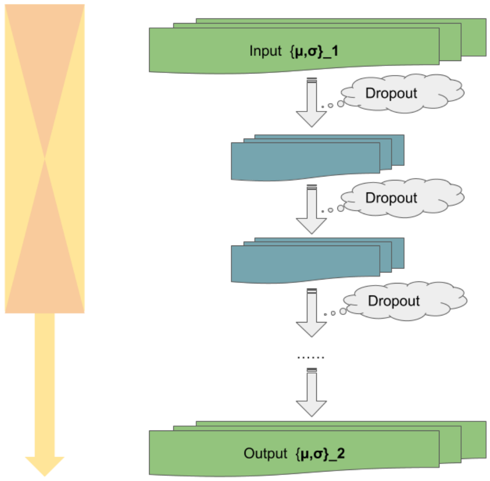

Figure 3. Latent parameter transformer in detail.

As shown in Fig. 2 and 3, the transform is a hour-glass shape network with hidden layers. The network configuration depends on how many hidden layers we have, how wide is the hidden layer and which component we are using ($\{x^1_\lambda\}_\lambda$ or $\{x^2_\lambda\}_\lambda$).

Since the training dataset is small, we can use relatively small hidden layer size. The loss function is the typical mean squared loss of the mean value and covariance.

To prevent the network from overfitting the training data, we have dropout in between two consecutive layers.

3. Numerical Experiment

3.1. Source of Training Data

The training dataset used is from Bach's compositions[1]. We download 467 midi files and 248 of them are valid input since the pre-trained musicVAE model can only handle quadruplemeter midi files.

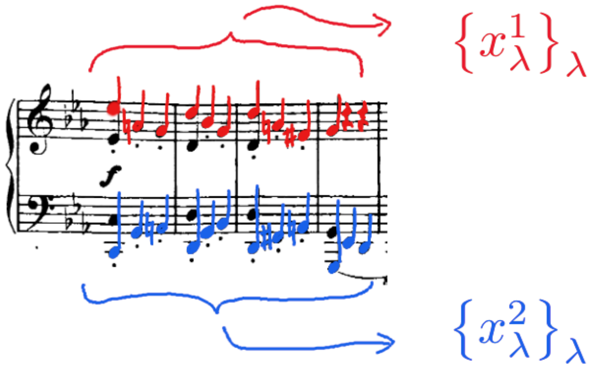

Figure 4. Extract $\{x^1_\lambda\}_\lambda$ and $\{x^2_\lambda\}_\lambda$ from a multi-note sequence. The red sequence consists the upper-most notes at each timestamp while the blue one consists the bottom-most notes.

The first component $\{x^1_\lambda\}_\lambda$(correponding to the "melody" part) is extracted as the highest note at each timestamp and the second component $\{x^2_\lambda\}_\lambda$ (correponding to the "accompaniment" part) is extracted as the lowest note. We will also denote the two components as upper-bottom in the figure labels.

3.2. Conditions and Hyper-Parameters

We have three independent hyper-parameters to adjust:

- The number of hidden layers: N_h can be either 1 or 2;
- The width of each hidden layer: we set them to be the same in each trial as $k = 16, 32, 64, 128$;
- The component we are using as the independent variable: $i = 1, 2$.

In total we will have $2 \times 4 \times 2 = 16$ different conditions and we will compare them in the following section.

3.3. Results

We train each network $\widetilde{\boldsymbol{N}}$ for 500 epochs with dropout rate 10%. Fig. 5 and 6 shows the loss decaying throughout the training procedure.

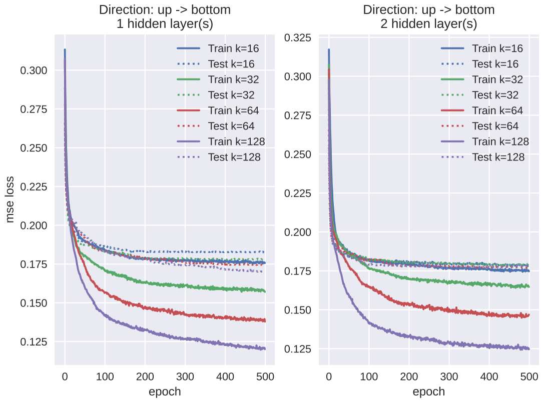

Figure 5. Training (in solid line) and testing (in dashed line) loss under the $\{x^1_\lambda\}_\lambda \Rightarrow \{x^2_\lambda\}_\lambda$ setting.

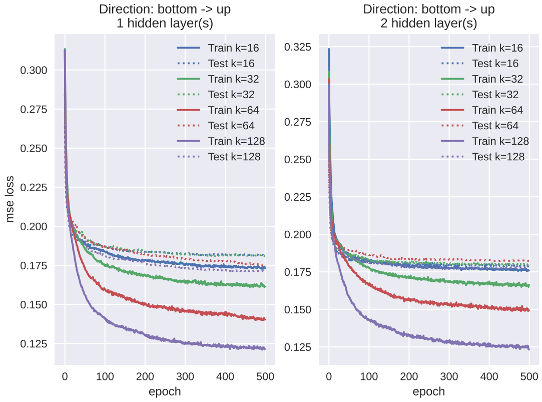

Figure 6. Training (in solid line) and testing (in dashed line) loss under the $\{x^2_\lambda\}_\lambda \Rightarrow \{x^1_\lambda\}_\lambda$ setting.

By comparing the loss curves, we can see that larger hidden layer width can lead to better performance, but only in the 1 hidden layer setting; increasing the width doesn't have such a significant effect when there are two hidden layers.

It is also worth checking the generated samples. We provide some samples[2] for $k = 128$ conditions:

[1] downloaded from Dave's J.S. Bach Page - MIDI Files - Bach MIDI Sequences by John Sankey

[2] The audio files can be found here.

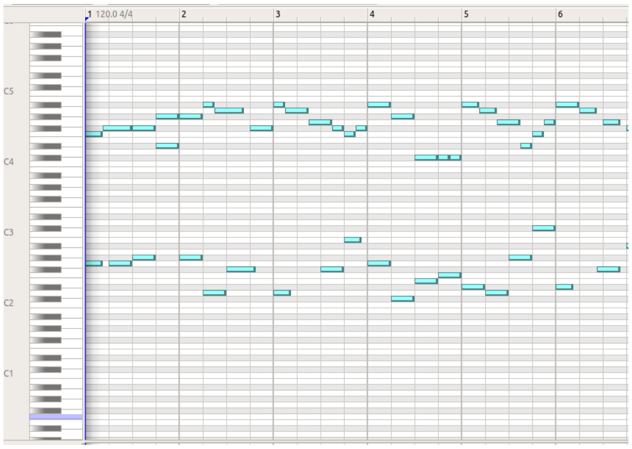

Figure 7. Sample generated by network with 1 hidden layer, $k = 128$ and $\left\{x_\lambda^2\right\}_\lambda \Rightarrow \left\{x_\lambda^1\right\}_\lambda$ setting.

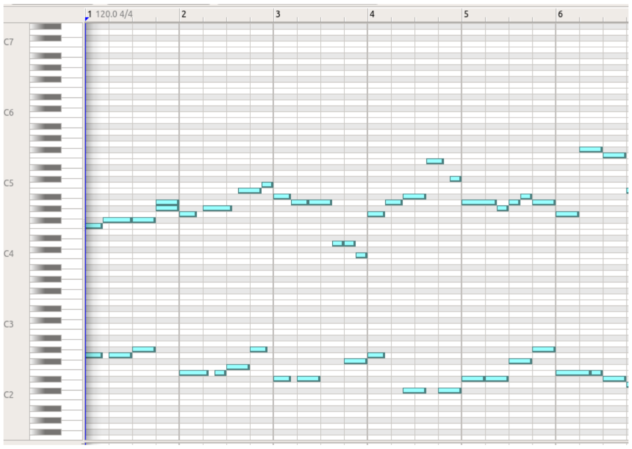

Figure 8. Sample generated by network with 1 hidden layer, $k = 128$ and $\left\{x_\lambda^1\right\}_\lambda \Rightarrow \left\{x_\lambda^2\right\}_\lambda$ setting.

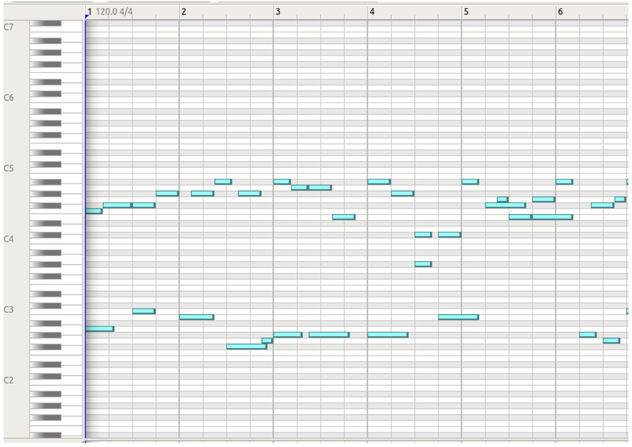

Figure 9. Sample generated by network with 2 hidden layers, $k = 128$ and $\left\{x_\lambda^2\right\}_\lambda \Rightarrow \left\{x_\lambda^1\right\}_\lambda$ setting.

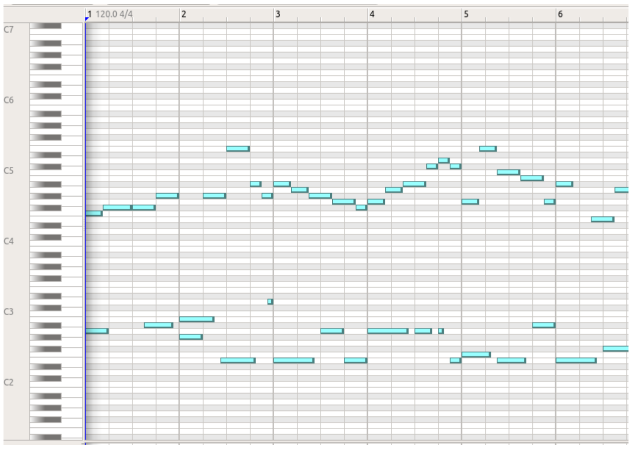

Figure 10. Sample generated by network with 2 hidden layer, $k = 128$ and $\left\{x_\lambda^1\right\}_\lambda \Rightarrow \left\{x_\lambda^2\right\}_\lambda$ setting.

4. Discussion

In this report we demonstrate that it is feasible to generate a melody-accompaniment pair simultaneously with musicVAE architecture. By transforming the latent vector, we can obtain the melody part from the accompaniment part or vice versa.

However, the samples generated from the network are not quite satisfying and don't sound like Bach's music, even produced under the 2-hidden-layer setting. A reasonable conjecture is that the encoder does not capture the chord progression quite well due to the fact that the original network is trained on pop-music dataset. Another potential problem is that the network fail to learn the fuga structure which is heavily used in Bach's compositions. One possible workout is to add connections between LSTM nodes in the encoders to strength certain recurrent pattern.

References

Berthelot, David, Goodfellow, Ian, Raffel, Colin, and Roy, Aurko. Understanding and improving interpolation in autoencoders via an adversarial regularizer. *7th International Conference on Learning Representations, ICLR 2019*, 2019. 1

Briot, Jean-Pierre and Pachet, François. Music Generation by Deep Learning - Challenges and Directions. pp. 1–17, 2017. doi: 10.1007/s00521-018-3813-6. URL http://arxiv.org/abs/1712.04371{%}0Ahttp://dx.doi.org/10.1007/s00521-018-3813-6. 1

Briot, Jean-Pierre, Hadjeres, Gaëtan, and Pachet, François-David. *Deep Learning Techniques for Music Generation*

– *A Survey*. 2017. ISBN 9783319701622. URL http://arxiv.org/abs/1709.01620. 1

Brunner, Gino, Wang, Yuyi, Wattenhofer, Roger, and Wiesendanger, Jonas. JamBot: Music theory aware chord based generation of polyphonic music with LSTMs. *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, 2017-November:519–526, 2018. ISSN 10823409. doi: 10.1109/ICTAI.2017.00085. 1

Chu, Hang, Urtasun, Raquel, and Fidler, Sanja. Song from PI: A musically plausible network for pop music generation. *5th International Conference on Learning Representations, ICLR 2017 - Workshop Track Proceedings*, pp. 1–9, 2019. 1

Dieleman, Sander, Van Den Oord, Äaron, and Simonyan, Karen. The challenge of realistic music generation: Modelling raw audio at scale. *Advances in Neural Information Processing Systems*, 2018-December(NeurIPS): 7989–7999, 2018. ISSN 10495258. 1

Dong, Hao Wen and Yang, Yi Hsuan. Convolutional generative adversarial networks with binary neurons for polyphonic music generation. *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018*, pp. 190–196, 2018. 1

Hadjeres, Gaëtan and Nielsen, Frank. Interactive Music Generation with Positional Constraints using Anticipation-RNNs. pp. 1–9, 2017. URL http://arxiv.org/abs/1709.06404. 1

Huang, Yu-Siang and Yang, Yi-Hsuan. Pop Music Transformer: Generating Music with Rhythm and Harmony. 2020. URL http://arxiv.org/abs/2002.00212. 1

Jaques, Natasha, Gu, Shixiang, Bahdanau, Dzmitry, Hernández-Lobato, José Miguel, Turner, Richard E., and Eck, Douglas. Sequence tutor: Conservative fine-tuning of sequence generation models with KL-control. *34th International Conference on Machine Learning, ICML 2017*, 4:2587–2596, 2017. 1

Lattner, Stefan, Grachten, Maarten, and Widmer, Gerhard. Imposing higher-level structure in polyphonic music generation using convolutional restricted Boltzmann machines and constraints. *Journal of Creative Music Systems*, 2(0), 2018. ISSN 23997656. doi: 10.5920/jcms.2018.01. 1

Mao, H. H. DeepJ: Style-Specific Music Generation. *Proceedings - 12th IEEE International Conference on Semantic Computing, ICSC 2018*, 2018-January:377–382, 2018. doi: 10.1109/ICSC.2018.00077. 1

Roberts, Adam, Engel, Jesse, Raffel, Colin, Hawthorne, Curtis, and Eck, Douglas. A hierarchical latent vector model for learning long-term structure in music. *35th International Conference on Machine Learning, ICML 2018*, 10:6939–6954, 2018. 1, 2.1, 1

Roy, Pierre, Papadopoulos, Alexandre, and Pachet, François. Sampling Variations of Lead Sheets. pp. 1–16, 2017. URL http://arxiv.org/abs/1703.00760. 1

Simon, Ian, Roberts, Adam, Raffel, Colin, Engel, Jesse, Hawthorne, Curtis, and Eck, Douglas. Learning a Latent Space of Multitrack Measures. 2018. URL http://arxiv.org/abs/1806.00195. 1

van den Oord, Aaron, Dieleman, Sander, Zen, Heiga, Simonyan, Karen, Vinyals, Oriol, Graves, Alex, Kalchbrenner, Nal, Senior, Andrew, and Kavukcuoglu, Koray. WaveNet: A Generative Model for Raw Audio. pp. 1–15, 2016. URL http://arxiv.org/abs/1609.03499. 1

Van Der Weerdt, Roderick and Schlobach, K S. Generating Music from Text: Mapping Embeddings to a VAE's Latent Space. 2018. URL https://esc.fnwi.uva.nl/thesis/centraal/files/f189374806.pdf. 1

Wang, Bryan and Yang, Yi-Hsuan. PerformanceNet: Score-to-Audio Music Generation with Multi-Band Convolutional Residual Network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:1174–1181, 2019. ISSN 2159-5399. doi: 10.1609/aaai.v33i01.33011174. 1

Yang, Li Chia, Chou, Szu Yu, and Yang, Yi Hsuan. Midinet: A convolutional generative adversarial network for symbolic-domain music generation. *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017*, pp. 324–331, 2017. 1