Generation of lyrics lines conditioned on music audio clips

Olga Vechtomova, Gaurav Sahu, Dhruv Kumar
University of Waterloo
{ovechtom, gsahu, d35kumar}@uwaterloo.ca

Abstract

We present a system for generating novel lyrics lines conditioned on music audio. A bimodal neural network model learns to generate lines conditioned on any given short audio clip. The model consists of a spectrogram variational autoencoder (VAE) and a text VAE. Both automatic and human evaluations demonstrate effectiveness of our model in generating lines that have an emotional impact matching a given audio clip. The system is intended to serve as a creativity tool for songwriters.

1 Introduction

Creative text synthesized by neural text generative models can serve as inspiration for artists and songwriters when they work on song lyrics. Novel and unusual expressions and combinations of words in generated lines can spark an idea and inspire the songwriter to create original compositions. In contrast to systems that generate lyrics for an entire song, our system generates suggestions in the form of individual lyrics lines, and is intended to serve as a creativity tool for artists, rather than as a standalone songwriting AI system.

In a song, musical composition, instrumentation and lyrics act together to express the unique style of an artist, and create the intended emotional impact on the listener. Therefore, it is important that a lyrics generative model takes into account the music audio in addition to the textual content.

In this paper we describe a bimodal neural network model that uses music audio and text modalities for lyrics generation. The model (Figure 1) generates lines that are conditioned on a given music audio clip. The intended use is for an artist to play live or provide a pre-recorded audio clip to the system, which generates lines that match the musical style and have an emotional impact matching the given music piece.

The model uses the VAE architecture to learn latent representations of audio clips spectrograms. The learned latent representations from the spectrogram-VAE are then used to condition the decoder of the text-VAE that generates lyrics lines for a given music piece. Variational autoencoder lends itself very well for creative text generation applications, such as lyrics generation. It learns a latent variable model of the training dataset, and once trained any number of novel and original lines can be generated by sampling from the learned latent space.

Three main groups of approaches towards stylized text generation in natural language processing (NLP) include: (1) embedding-based techniques that capture the style information by real-valued vectors, and can be used to condition a language model (Tikhonov and Yamshchikov, 2018) or concatenated with the input to a decoder (Li et al., 2016); (2) approaches that structure latent space to encode both style and content, and include Gaussian Mixture Model Variational Autoencoders (GMM-VAE) (Shi et al., 2020; Wang et al., 2019), Conditional Variational Autoencoders (CVAE) (Yang et al., 2018), and Adversarially Regularized Autoencoders (ARAE) (Li et al., 2020); (3) approaches with multiple style-specific decoders (Chen et al., 2019).

All of the above papers infer style from only one modality, text. Our work belongs to the first category of approaches: embedding based techniques, and is different from all of the above works in learning the style information from audio and text. Furthermore, previous embedding-based approaches use embeddings from a discrete vector space. This is the first work that uses a continuous latent variable learned by a spectrogram-VAE as a conditioning signal for the text-VAE.

A number of approaches have been proposed towards poetry generation, some focusing on

rhyme and poetic meter (Zhang and Lapata, 2014), while others on stylistic attributes (Tikhonov and Yamshchikov, 2018). Cheng et al. (2018) proposed image-inspired poetry generation. Yang et al. (2018) generated poetry conditioned on specific keywords. Yu et al. (2019) used audio data for lyrics retrieval. Watanabe et al. (2018) developed a language model for lyrics generation using MIDI data. Vechtomova et al. (2018) generated author-stylized lyrics using audio-derived embeddings. To our knowledge this is the first work that uses audio-derived data to generate lyrics for a given music clip.

The main contributions of this work are: (1) A probabilistic neural network model for generating lyrics lines matching the musical style of a given audio clip; (2) We demonstrate that continuous latent variables learned by a spectrogram-VAE can be effectively used to condition text generation; (3) Automatic and human evaluations show that the model can be effectively used to generate lyrics lines that are consistent with the emotional effect of a given music audio clip.

2 Background: unconditioned text generation with VAE

The variational autoencoder (Kingma and Welling, 2014) is a stochastic neural generative model that consists of an encoder-decoder architecture. The encoder transforms the input sequence of words $\boldsymbol{x}$ into the approximate posterior distribution $q_\phi(z|x)$ learned by optimizing parameters ϕ of the encoder. The decoder reconstructs $\boldsymbol{x}$ from the latent variable $\boldsymbol{z}$, sampled from $q_\phi(z|x)$. Both encoder and the decoder in our work are recurrent neural networks, specifically, Long Short Term Memory networks (LSTM). The reconstruction loss is the expected negative log-likelihood of data:

$$J_{\text{rec}}(\phi, \theta, x) = -\sum_{t=1}^{n} \log p(x_t|\boldsymbol{z}, x_1 \cdots x_{t-1}) \tag{1}$$

where ϕ and θ are parameters of the encoder and decoder, respectively. The overall VAE loss is

$$J = J_{\text{rec}}(\phi, \theta, \boldsymbol{x}) + \text{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) \tag{2}$$

where the first term is the reconstruction loss and the second term is the KL-divergence between $\boldsymbol{z}$'s posterior and a prior distribution, which is typically set to standard normal $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

3 Approach

The audio clip conditioned generation model (Figure 2) consists of a spectrogram-VAE to learn a meaningful representation of the input spectrogram, and a text-VAE with an audio-conditioned decoder to generate lyrics.

In order to train the spectrogram-VAE, we first split the waveform audio of songs into small clips, and transform them into MEL spectrograms. For this mode of generation, we generate data with two levels of audio-lyrics alignment: high-precision and low-precision, which are described in more detail in Section 4. The spectrogram-VAE follows a encoder-decoder architecture. The encoder consists of four convolutional layers followed by a fully connected layer, and the decoder architechture is mirrored by using a fully-connected layer followed by four deconvolutional layers. This model was trained for 100 epochs.

We then feed all the spectrograms in the dataset to obtain their respective spectrogram embeddings. More precisely, we first obtain the $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ for every data point, and then sample a latent vector from the learned posterior distribution using a random normal noise $\epsilon \in \mathcal{N}(\mathbf{0}, \mathbf{I})$.

The audio clip conditioned VAE, which has the same architecture as described in Section 2, is then trained to generate lyrics befitting the provided piece of music by concatenating the spectrogram embedding with the input to every step of the decoder. The reconstruction loss is calculated as:

$$J_{\text{rec}}(\phi, \theta, {z_k}^{(s)}, x^{(t)}) = -\sum_{i=1}^{n} \log p({x_i}^{(t)}|\boldsymbol{z_k}^{(t)}, {z_k}^{(s)}, {x_1}^{(t)} \cdots {x_{i-1}}^{(t)}) \tag{3}$$

where ${z_k}^{(s)}$, spectrogram embedding of the k-th data point. At inference time, a latent vector ${z_k}^{(t)}$ sampled from the text-VAE's prior is concatenated with the corresponding spectrogram embedding ${z_k}^{(s)}$, and fed to the LSTM cell at every step of the text-VAE's decoder.

4 Evaluation

We collected a dataset of lyrics by seven Rock artists: David Bowie, Depeche Mode, Nine Inch Nails, Neil Young, Pearl Jam, Rush, and Doors. Each of them has a distinct musical and lyrical style, and a large catalogue of songs, spanning many years. We intentionally selected artists from the sub-genres of Rock as this models a real-world

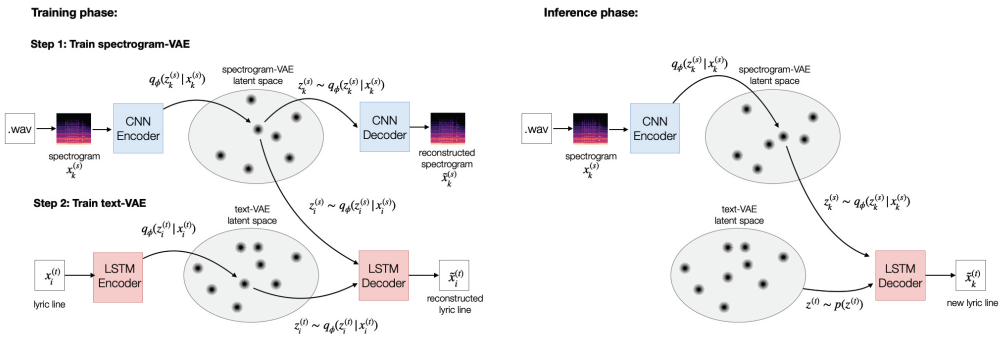


Figure 1: Music audio clip conditioned lyrics generation.

scenario, when a given songwriter might use the model to get influences from the genre congruent with their own work.

Since we do not have access to aligned data for these artists, we manually created a high-precision aligned dataset for two artists (Depeche Mode and Nine Inch Nails, 239 songs), and did an automatic coarse-grained alignment for the other five (518 songs). To create a manually aligned dataset, we annotated the original waveform files in Sonic Visualizer (Cannam et al., 2006) by marking the time corresponding to the start of each lyrics line in the song. The automatic alignment process consisted of splitting each song into 10-second segments and splitting lyrics lines into the same number of chunks, assigning each to the corresponding 10-second clip. In total, the training dataset consists of 18,210 lyrics lines and 14,670 spectrograms.

The goal of the developed model is to generate lyrics lines for an instrumental music piece. Our test set, therefore, only contains instrumental songs: 36 songs from an instrumental album "Ghosts I-IV" by Nine Inch Nails[1] and eight instrumental songs from three other albums by two artists (Depeche Mode and Nine Inch Nails). Each song was split into 10-second clips, which were then converted into spectrograms (807 in total).

First we evaluate the quality of the latent space learned by the spectrogram-VAE. For every spectrogram in the test set, we computed pairwise cosine similarity between its embedding $z^{(s)}$ and the embedding of every spectrogram in the training and test set. We then calculated the proportion of clips in the top 50 and 100 that (a) are part of the same song, (b) are part of the same album, and (c) belong to the same artist. The results (Table 1) indicate that large proportions of clips most similar to a given clip are by the same artist and from the same album. This demonstrates that spectrogram-VAE learns representations of an artist's unique musical style.

Top-n	same song	same album	same artist
n=50	0.1707	0.4998	0.7293
n=100	0.0988	0.4462	0.7067

Table 1: Clustering effect in the spectrogram-VAE latent space.

We divided songs in the test set into two categories: "intense" and "calm". The peak dB differences between tracks in these two categories are statistically significant (t-test, $p<0.05$). A spectrogram for each 10-second clip was used to generate 100 lines according to the method described in Section 3. The songs in these two categories evoke different emotions and we expect that the lexicon in these two categories of generated lines will be different, but more similar among songs within the same category.

Automatic evaluation of generated lines conditioned on an instrumental audio clip is difficult, since there is no reference ground-truth line that we can compare the generated line to. For this reason, we cannot use n-gram overlap based measures, such as BLEU. Secondly, style-adherence metrics, such as classification accuracy w.r.t. a certain class, e.g. style, in the dataset are inapplicable, since there is no categorical class variable here.

We calculated KL divergence values for words in each song. KL divergence measures the relative entropy between two probability distributions. It was defined in information theory (Losee, 1990) and was formulated as a word ranking measure

[1]Ghosts I-IV. Nine Inch Nails. Produced by: Atticus Ross, Alan Moulder, Trent Reznor. The Null Corporation. 2008. Released under Creative Commons (BY-NC-SA) license.

in (Carpineto et al., 2001). Given word w in the generated corpus G_k conditioned on clip k and generated corpus N conditioned on all other clips in the test set, KL divergence is calculated as word-KL$(w) = p_{Gk}(w) \cdot \log(p_{Gk}(w)/p_N(w))$.

We then calculated rank-biased overlap scores (RBO) to measure pairwise similarity of word-KL ranked lists corresponding to each pair of songs. RBO (Webber et al., 2010) is a metric developed in Information Retrieval for evaluating ranked search results overlap, and handles non-conjoint ranked lists. The RBO score falls in the range [0,1] where 0 indicates a disjoint list and 1 - identical.

Figure 2 shows that "calm" songs have higher RBO values with other songs in the same category, indicating similar generated word distributions, and low RBO values w.r.t. "intense" songs. RBO values for lines generated conditioned on "intense" songs are not as high, suggesting that they have less word overlap. This is likely because there are more songs in this category in the training set with widely different lyrics, therefore the model may be picking up more subtle musical differences, which make it correspondingly generate lyrics that have lyrical influences from different songs.

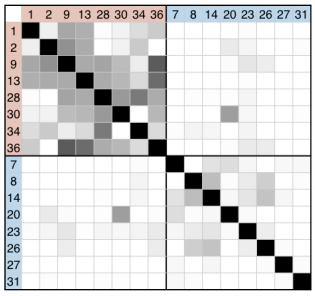
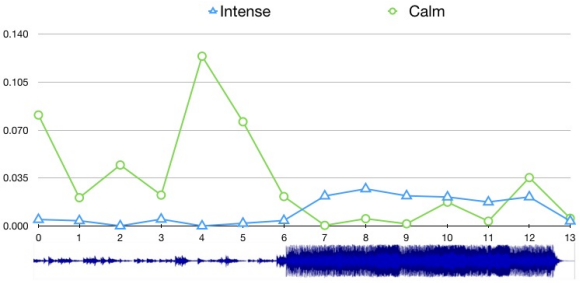


Figure 2: Rank-biased overlap (RBO) between the KL-divergence ranked lists of generated words for songs in "Ghosts I-IV" album. Pink-highlighted songs are calm, and blue - intense. The darker grey cells indicate higher overlap. The row/column headings correspond to the numbers in the song titles (e.g. 2 is for "2 Ghosts I") .

The difference between word distributions is also evident at the audio clip level. Figure 3 shows RBO values for each 10-second clip in the song "12 Ghosts II"[2]. The x-axis is the timeline. We first calculated word-KL for every clip w.r.t. all other clips in the test set. Then pairwise RBO was computed between the given clip's ranked word list and the ranked word lists for "intense", and "calm" generated corpora, respectively. The original song's waveform is given for reference, showing correlation with the change in the lexicon being generated for calm and intense sections of the track.



clip	generated line
0	shadow of the world that's gone
1	tomorrow's all over
2	as everything's a mirror
3	fly in the light
4	ghosts i've lost in the world
5	i can't see the answers
6	i can't shake this

clip	generated line
7	all i suffer
8	it must be breaking
9	you'd learn to see my pain
10	the measure of the moment
11	breathe out of dreams
12	all you couldn't control
13	floating in the end

Figure 3: Rank-biased overlap (RBO) values for 10-second clips of "12 Ghosts II" song and examples of lines generated for each clip.

We have also conducted a human evaluation. Six participants (3 female and 3 male), none of whom are members of the research team, were asked to listen to ten instrumental music clips from the test set. For each clip they were given two lists of 100 generated lines. One list was generated conditioned on the given clip in either "calm" or "intense" category, the other list was generated based on a clip from the opposite category. The participants were asked to select the list that they thought was generated based on the given clip. The average accuracy was 78.3% (sd=9.8), which shows that participants were able to detect emotional and semantic congruence between lines and a piece of instrumental music.

5 Conclusions

We developed a bimodal neural network model, which generates lyrics lines conditioned on an instrumental audio clip. The evaluation shows that the model generates different lines for audio clips from "calm" songs compared to "intense" songs. Also, songs in the "calm" category are lexically more similar to each other than to the songs in the "intense" category. A human evaluation shows that the model learned meaningful associations between the semantics of lyrics and the musical characteristics of audio clips captured in spectrograms.

[2]Demos of generated lines are available at: https://sites.google.com/view/nlp4musa-submission/home

References

Chris Cannam, Christian Landone, Mark B Sandler, and Juan Pablo Bello. 2006. The sonic visualiser: A visualisation platform for semantic descriptors from musical signals. In *ISMIR*, pages 324–327.

Claudio Carpineto, Renato De Mori, Giovanni Romano, and Brigitte Bigi. 2001. An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems (TOIS)*, 19(1):1–27.

Cheng-Kuan Chen, Zhufeng Pan, Ming-Yu Liu, and Min Sun. 2019. Unsupervised stylish image description generation via domain layer norm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8151–8158.

Wen-Feng Cheng, Chao-Chung Wu, Ruihua Song, Jianlong Fu, Xing Xie, and Jian-Yun Nie. 2018. Image inspired poetry generation in xiaoice. *arXiv preprint arXiv:1808.03090*.

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003.

Yuan Li, Chunyuan Li, Yizhe Zhang, Xiujun Li, Guoqing Zheng, Lawrence Carin, and Jianfeng Gao. 2020. Complementary auxiliary classifiers for label-conditional text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8303–8310.

Robert M Losee. 1990. *The science of information: Measurement and applications*. Academic Press New York.

Wenxian Shi, Hao Zhou, Ning Miao, and Lei Li. 2020. Dispersed em-vaes for interpretable text generation. In *In Proceedings of the 37th International Conference on Machine Learning (ICML)*.

Alexey Tikhonov and Ivan P Yamshchikov. 2018. Guess who? multilingual approach for the automated generation of author-stylized poetry. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 787–794.

Olga Vechtomova, Hareesh Bahuleyan, Amirpasha Ghabussi, and Vineet John. 2018. Generating lyrics with variational autoencoder and multi-modal artist embeddings. *arXiv preprint arXiv:1812.08318*.

Wenlin Wang, Zhe Gan, Hongteng Xu, Ruiyi Zhang, Guoyin Wang, Dinghan Shen, Changyou Chen, and Lawrence Carin. 2019. Topic-guided variational auto-encoder for text generation. In *HLT-NAACL*, pages 166–177.

Kento Watanabe, Yuichiroh Matsubayashi, Satoru Fukayama, Masataka Goto, Kentaro Inui, and Tomoyasu Nakano. 2018. A melody-conditioned lyrics language model. In *HLT-NAACL*, pages 163–172.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38.

Xiaopeng Yang, Xiaowen Lin, Shunda Suo, and Ming Li. 2018. Generating thematic chinese poetry using conditional variational autoencoders with hybrid decoders. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4539–4545.

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).

Xingxing Zhang and Mirella Lapata. 2014. Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 670–680.