Image generation associated with music data

Yue Qiu
University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki, Japan
s1830151@s.tsukuba.ac.jp

Hirokatsu Kataoka
National Institute of Advanced Industrial
Science and Technology (AIST)
1-1-1 Umezono, Tsukuba, Ibaraki, Japan
hirokatsu.kataoka@aist.go.jp

Abstract

Recently, the development of music appreciation device has made it possible to listen to various kinds of music regardless of location. On the other hand, if it is possible to express visual contents that best matches music, we can expect a more expressive music appreciation experience by not only "listening to" the music but also "watching" the music. In this paper, we address the problems below: (1) learning a correlation between music data and images; (2) generating images from music data automatically. The experiments show that our proposed method can effectively generate proper images from music data.

1. Introduction

In music appreciation, recent years it has become possible to select a preferred means from among all options such as YouTube and other video sharing sites, portable music players and other music devices. Nowadays, we can listen to various kinds of music regardless of location. Meanwhile, although we can enjoy the music contents by listening using our hearing sense among the five senses, recently it is thought that if we can combine with other senses such as sighting, it will be possible to experience a more expressive music appreciation experience to have.

TextAlive [1] can be mentioned as an example combing music expression and vision. TextAlive is a system that can perform, produce and share lyrics animation on a browser by synchronizing the music contents on the web with previously extracted lyrics, and expanded to not only listening but also visually enjoyable contents. However, for next-generation music contents, we think that not only displaying lyrics according with the tune, it will be more expressively to automatically generate images and videos that best match the music. Therefore, we propose a method that automatically generates images associated with music, which make it possible for users to visually enjoy images while listening to music. With the development of CNNs and RNNs technologies, it has become possible to extract meaningful information from music data and images. Furthermore, with the proposal of the GAN (Generative Adversarial Networks) [2], it became easier to automatically generate images under some conditions, such as labels of images. In this research, we extract information from images and music data using CNNs and RNNs methods, and learn relations between images and music data based on those extracted information, also, by utilizing GANs method, we automatically generate related images from music data.

2. Related work

In recent years, simultaneously learning images and descripitive texts corresponding to images, so-called image captioning, have been extensively studied. Such research can be used for image retrieval and automatically generating human-level image descriptions. Among them, the method [3] proposed by Reed et al is widely used. Reed proposed a framework to learn correlation between detailed labeled images and images captioning. Framework proposed by Reed uses CNNs to extract features from images, and uses CNN-RNNs based methods to extract time-related features from image captioning. The network will be trained to maximize inner production of extracted features of image and correlated image captioning. In our research, based on Reeds work, we encode music data instead of image captioning, and learn the correlation between music data and images.

In 2014, Goodfellow introduced GAN which can simultaneously train two networks, an image generation network G and a discrimination network D that aims to classify the Real and Fake images. GAN can be used to learn from real images and generate images similar to them. In 2015, Radford proposed DCGAN (Deep Convolutional GAN) [4], which utilizing Deep Neural Network to generate images with high realism. Also, Conditional GAN [5] is proposed by Mirza which can generate images under some appointed conditions such as image labels. In our work, after learnt correlations between music data and corresponding images,

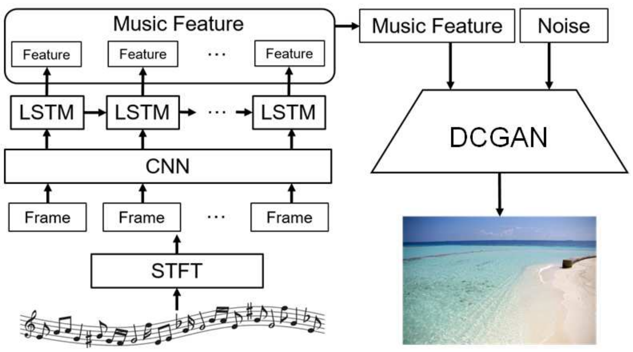

Figure 1. Overview of proposed image generation process.

we use music data as conditions to generate images by Conditional DCGAN. As result, we can generate corresponding images from input music data.

3. Proposed method

In our research, we propose a method to generate images under the conditions of input music data. Figure 1 shows the overview of our image generation process. We extract features from music data and images separately, and learn correlation of extracted features of music data and images. And based on the learnt correlation, automatically generate images from music data.

3.1. Music feature extraction

For the preprocessing, we use STFT (Short Time Fourier Transform) to transfer raw music data to power and phase vectors as formulation below, where x(t) represents the input music data, and τ, ω indicate the phase and power vectors to be extracted.

$$STFTx(t)(\tau, \omega) = \int_{-\infty}^{+\infty} x(t)w(t-\tau)e^{-j\omega t}\,\mathrm{d}t \qquad (1)$$

When implementing, we actually input raw music data for about 1 minute long, a frame of 16 seconds is randomly extracted for each piece of music to obtain the two 1024 dimension vectors.

Next, we use CNNs to extract features from the two vectors separately and concatednate extracted features by a final fully connect layer. Through these processes, we can obtain a 1024 dimension music feature. In addition, the feature vectors extracted by CNN are related in terms of time, we then input the extracted features to LSTM to generate time-series features.

3.2. Image feature extraction

We use AlexNet [6] to extract 1024 dimension representation of every image. In this paper, as we divide the images into only four categories, the ability of feature extraction of AlexNet would satisfy our needs. As a future work, we believe that using a deeper neural network will be useful for further information extraction of images.

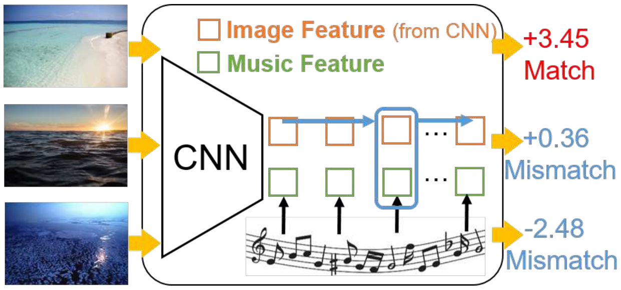

Figure 2. Correlation learning of music data and images.

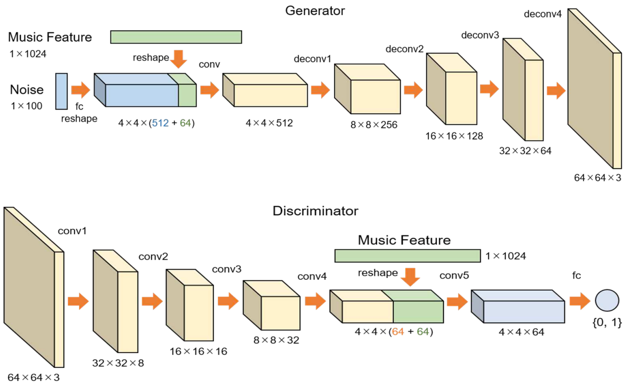

Figure 3. Generator and Discriminator network of image generation.

3.3. Correlation learning of music data and images

After those processes above, as shown in Figure 2, we use method also introduced in paper [3] by Reed et al. to learn the correlation of images and music data. We compute the inner production (1024 dimension vector) of two kinds of features: music features obtained by section 3.1 and image features obtained by section 3.2. During training, we use loss function proposed by Reed et al. to train network parameters to maximize the inner production of correlated image and music data of full training data.

3.4. Image generation from music data

After learnt correlations of images and music data, we can obtain music features that can correspond to images through our trained CNN-LSTM model. We then fuse those extracted music features into DCGAN to generate images. We show detail of our network in Figure 3. Throughout the network, we can obtain color image (64*64) from 1024 dimension input music feature.

4. Experiments and results

4.1. Data collection

In order to find out what kinds of images people prefer to watch while listening to the music, we did a survey that let 50 people (almost 20s year old contains 33 man and 17 women) to choose the genre of images they most want to watch while listening to the music from “Natural Scene”, “Picture of Singer”, “Urban City” and “Picture of Anime Character”. The result shows that 54 percent of them (27 people) prefer “Natural Scene”. Therefore we gathered images of 4 kinds of natural scene (“sky”, “water”, “mountain” and “desert”) from the Places-205 Dataset [7], each scene category contains 2,000 images. We show some examples of images we picked from Places dataset in Figure 4-Figure7.

For the music data corresponding to those 4 kinds of natural scene, we used keyword “sky”, “water”, “mountain” and “desert” to search free music from Internet and collected about 30 music data for each category (almost 1minute long music).

4.2. Learning strategies

For the music data and images correlation learning, we set learning rate to 10^{-5} and learnt 200 epochs. For the learning of image generation, we set learning rate to 10^{-4}, and learnt 100 epochs. Several generated images of our learned model are shown in Figure 8. We can find that our proposed method could effectively generate 4 kinds of natural scene.

4.3. Results

In order to evaluate if our proposed method can generate proper images corresponding to input music, we invited 12 people to assist our experiment. We played 4 pieces of music (one music per category) to each experimenter, and for every piece of music, we also shown them four images (only one of those images was really generated by the music, others are generated from other categories of music). They will pick one image for each piece of music that matches the music they heard the best. Table 1 shows the result of this experiment. The diagonal of Table 1 indicates the images generated from input music (Vertical axis) are chosen to be the most matches one. The non-diagonal of table 1 indicates the images generated from other kinds of music (Horizontal axis) are chosen by experimenter to be the most matched one to the input music. From the result, we can know that our proposed method can generate correct images from correlated input music well. However, the result of sky and water is a little mutual confused. For the future work, its important to collect more data in order to generate these two categories well.

Figure 4. Examples of sky’s images picked up from Places dataset.

Figure 5. Examples of water’s images picked up from Places dataset.

Figure 6. Examples of mountain’s images picked up from Places dataset.

Figure 7. Examples of desert’s images picked up from Places dataset.

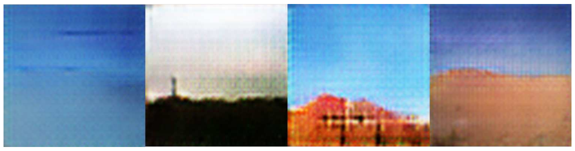

Figure 8. Generated images of proposed method(left to right: generated image of “sky”, “water”, “mountain”, “desert”).

	sky	water	mountain	dessert
sky	(9/12)	(3/12)	(0/12)	(0/12)
water	(4/12)	(7/12)	(1/12)	(0/12)
mountain	(1/12)	(2/12)	(9/12)	(0/12)
dessert	(1/12)	(0/12)	(0/12)	(11/12)

Table 1. User evaluation results of the consistency of music and generated images.

5. Conclusions

In this paper, we proposed a method that automatically generates images associated with music data, which make it possible for users to visually enjoy images while listening to music. Currently, our experiments are conducted using four kinds of music data and natural images, further researching the relevance between music data, video data and images, generating images for more categories will be our future task.

References

[1] J. Kato, T. Nakano, and M. Goto. Textalive: Integrated design environment for kinetic typography. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3403–3412. ACM, 2015.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, and et al. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[3] S. Reed, Z. Akata, H. Lee, and et al. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–58, 2016.

[4] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[5] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[6] A. Krizhevsky, I. Sutskever, and et al. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[7] B. Zhou, A. Lapedriza, J. Xiao, and et al. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.