Automated Music Generation for Visual Art through Emotion

Xiaodong Tan
Independent Researcher
frances.tanx@gmail.com

Mathis Antony
CognAI, Hong Kong
mathis.antony@gmail.com

Abstract

We explore methods of generating music from images using emotion as the connection between the visual and auditory domains. Our primary goal is to express visual art in the auditory domain. The resulting music can enrich visual art or provide a form of translation to facilitate enjoying visual art without reliance on the visual system. We use pre-trained image representations and explore two different types of music modelling methods based on RNN and Transformer architectures to build models capable of generating music given an image as input. To evaluate the performance of these methods, preliminary human and machine evaluation are conducted. The results suggest that both music generators are able to express music with an emotional connection.

Introduction

Given a computational method to translate or supplement art in other domains, we could make art accessible to a wider audience and in general enrich the experience of consuming art. Artists express themselves in various domains and a single piece of art may exist in one or more of these domains. Artists can be inspired by art or styles in one domain and create in another domain. The audience can link the art works in different domains (Ranjan, Gabora, and O'Connor 2013) and enjoy the artistic ideas in one domain by being exposed in another domain. These connections might be established through synesthesia, in which perception with one sense is perceived through other senses simultaneously (Ramachandran 2003). It can also be argued that artistic styles exist across domains (Hasenfus 1978) and that art works in two domains can trigger the same perception. While the process of generating and consuming art involves potentially all senses, individual works of art often only exist in one or a few domains (painting, music, ...). Can we use modern computational methods to express art across domains? Different cross-domain art has been studied such as image to music, sculpture to painting, body movement to sound (Marković and Malešević 2018), image to 3-D printable vase (Horn et al. 2015).

In this paper, we focus on the cross-domain image and music translation. Some studies have been conducted on the mapping between visual features (such as color) and acoustic features (such as pitch) (Xiaoying Wu and Ze-Nian Li 2008; Sergio et al. 2015) while others explored existing linked data, such as video and background musics, and used supervised classification to learn their connections (Martín-Gómez et al. 2019). These ideas have been used in applications such as suggesting background sound (Solèr et al. 2016) or descriptive music (Martín-Gómez et al. 2019).

Motivated by the observation that both music and paintings may provoke an emotional response, we build the connection between the two worlds around emotions and use emotions to create in the new domain. (Verma, Dhekane, and Guha 2019) studied the affective relationship between the signals from the two domains. However, this links are not used in generation directly. (Sergio et al. 2015) proposed a mapping of image features to sound features that can potentially trigger the same emotional response from the audience. (Madhok, Goel, and Garg 2018) generates music based on the sentiment detected in people's facial expression. However, the visual information is not integrated into the model.

The rapid development in AI and deep learning technologies provide new opportunities for multi-modal creation. Transfer-learning and deep image representations are already commonly used in computer vision tasks. These representations are usually pre-trained with models based on CNN architectures on large-scale image datasets, and can be fine-tuned on downstream tasks such as image classification. Cross-domain generative tasks, such as image captioning make good use of these ideas to infuse different signals into generative models (Vinyals et al. 2017). Our methods are greatly inspired by these ideas.

The purpose of this paper is to fill in the gaps in the image-music generation areas with new technologies. The major contribution of the paper is as follows.

1. We propose the task of generating emotionally connected music from images.
2. We use deep learning methods to fuse both the information contained in the image and music into the music generator.
3. We explore the performance of different music generators in terms of how well they convey the emotions in images in their generated music.

The proposed cross-domain generation has applications

Proceedings of the 11th International Conference on Computational Creativity (ICCC'20)
ISBN: 978-989-54160-2-8

as a creative tool, and can also help people with different perceptional preferences to enjoy art from different perceptions. In particular, this kind of system may give visually-impaired people another way to enjoy visual art.

Results and media related to this work are available via a git repository at `https://github.com/sudongtan/synesthesia`.

Methodology

From a given image and piece of music labeled the same emotion, we extract image and audio features. Then we fuse the two sources of data into a music generator for music generation. An overview of the architecture is shown in figure 1.

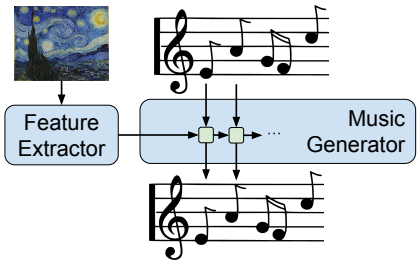

Figure 1: Image to music generator architecture.

Image Feature Extraction

An *ImageNet* (Deng et al. 2009) pre-trained *ResNet* (He et al. 2016) CNN model is used to extract images feature representations. In practice we use the ResNet-18 model from *pytorch-torchvision* (Paszke et al. 2019; Marcel and Rodriguez 2010) and extract as features the output of the last pooling layer.

Music Modelling

Throughout this project, music is captured in the format of MIDI. Event-based feature extraction is used to capture the timing and dynamics of MIDI files (Simon and Oore 2017), including if there is note or not at a certain timestamp (if it is a note-on or note-off event). For the note-on event, it also included how hard the note is played, i.e., the velocity information.

Multi-modal Encoder-decoder

After image information and music information are encoded separately, they are fused into music decoders, based on the model architecture of the sequence decoder. This method is inspired by other multi-modal generative tasks, such as image captioning (Vinyals et al. 2017) where the image and language information are fused to text decoder and generate text that is related to the image.

Model 1: RNN Decoder The decoder is composed of a multi-layer RNN architecture. The initial hidden states of the RNN are initialized to vectors that represent the image information. For this the image features are transformed via a matrix multiplication such that they are compatible with the size of the hidden state. The initial input to the RNN is the music representation. By doing this, the decoder has both the image and music information.

In particular, this model takes the velocity information into consideration, so the generated music focus on the timing and dynamics of the music (Simon and Oore 2017).

Model 2: Transformer Decoder Here a Transformer architecture is used as decoder. Transformers are a self-attention based sequence models. Here we also initialize the initial hidden state of the decoder based on the features extracted from the images as inspired by Zhu et al. (2018).

The Transformer architecture is better than the RNN at capturing the long-term structure of the music (Huang et al. 2018), i.e., the coherence of the music.

Integration of Emotion into the Model

Each training sample consists of a pair of an image and a MIDI file with the same emotion label. In this way, the image information, the music information, as well as their connection are presented to the model. During inference, the image is fed into the model and the music encoding input to the decoder to the is randomly initialized.

Experiments and Results

Training Dataset

Image We use a dataset built for emotion recognition (You et al. 2016) as the source of our images. The images are classified into eight types of emotions, namely amusement, anger, awe, contentment, disgust, excitement, fear and sadness.

Music in MIDI The MIDI files with emotion labels in the Multi-modal MIREX-like emotion dataset (Malheiro et al. 2013) are used as music data. Originally the dataset partitioned according to emotions into five clusters, which are 1: passionate, rousing, ..., 2: cheerful, good nature, ..., 3: poignant, wistful, ..., 4: humorous, campy, ..., 5: aggressive, tense/anxious, ...

Pairing Image and Music We think that the best possible mapping between the labels in the two datasets above is to map music clusters 1 - 5 to image emotion labels excitement, contentment, sadness, amusement and anger respectively. In total there are 17,349 images and 196 midi files.

Model Training and Music Generation

Model Training Both models were trained using stochastic gradient descent with Adam optimizer until the validation loss stopped improving.

Music Generation During inference, around 600 art photos of emotions amusement, excitement, contentment, sadness and anger from (Machajdik and Hanbury 2010) are used to generate music. The music generated is around 5 - 15 seconds for the RNN model and 10 - 20 seconds for the Transformer model. As expected, the music generated by the RNN model sounds more dynamic but less coherent.

The music generated by the Transformer model is more coherent but also lacks variation.

Evaluation

For evaluations we focus on if an image and a piece of music generated from it invoke similar emotions. To simplify the task of labeling, we consider only their overall positive (contentment, amusement, excitement) or negative (sad, angry) aspect.

Human Evaluation Evaluators are six human beings. The data to be evaluated are 14 pieces of generated music, including 8 from the RNN generator and 6 from the Transformer generator, together with the 14 images from the (Machajdik and Hanbury 2010) dataset from which the pieces of music are generated. Half of the images from each model are labeled positive, the other half are negative. The pieces of music are selected at random but those that were less than 6 seconds long, very repetitive or repeating the training data were discarded. The evaluators are asked to evaluate the emotion of the music and image on a scale from 1 (very negative) to 10 (very positive) without knowing which piece of music is related to which image. To remove subjective biases the absolute ratings are transformed to relative rankings for each participant and medium (Yannakakis, Cowie, and Busso 2017). The results are summarized in figure 2.

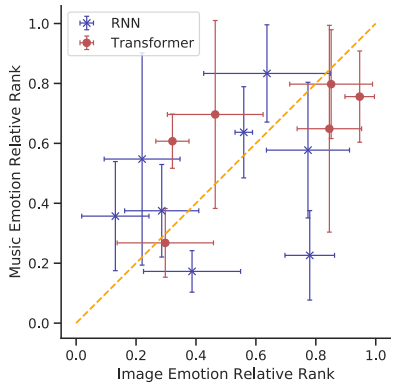

Figure 2: Human evaluation results. Each point denotes the mean relative emotion rating rank for a given image, music pair. The error bars denote the standard deviation. The dashed line is $y = x$, the ideal response.

The correlation coefficient between two mediums is 0.49 indicating that on average the emotion ratings of the generated music is correlated with the emotion rating of the source image. We note there is one obvious outlier towards the bottom right corner where there is consensus that the image conjures positive emotion while the corresponding music conjures negative emotions. We expect that it would take a significantly larger evaluation experiment to determine if there is a significant difference in performance between the two music generators. We also observed that there was better agreement on the images compared to the music among evaluators. In particular there are four pieces of music where the ratings were very dissimilar.

A danger with this evaluation method is that evaluators may express their fondness for a piece of work, instead of the emotion it triggers. For example, a sad but well liked song may be given a high rating instead of the expected low rating in accordance with the triggered negative emotion. As a result, the quality of the generated music may strongly influence the evaluation because songs of good quality are more enjoyable.

Machine Evaluation Here we propose an automated evaluation method and briefly mention a preliminary result for the RNN generator. In order to evaluate the models on a larger scale, we propose to train an emotional correspondence classifier. This classifier takes the image and music features as input and predicts if they express the same emotion (positive label) or not (negative label).

To construct a training sample we choose an image from the image dataset and select a crop from a randomly chosen piece of music labeled with the same emotion (different) for positive (negative) samples. The cropping is done to artificially inflate the number of samples due to the lack of available data. During training we construct positive and negative samples at random and feed them to the classifier. Training this kind of classifier has proven more difficult than anticipated. Preliminary experiments suggest that it is hard to beat 60% validation accuracy.

To evaluate the RNN music generation model we generate 10 pieces of music for 600 images and feed these 6000 pairs into the classifiers. We evaluated this data with one of the better performing classifiers. It predicted 63% of the samples have similar emotions. Therefore we think that even though this seems to be a difficult problem and there is much room for improvement, this method does have potential for automated evaluation of our music generators.

Discussions and Future Work

Technically, one major issue is how to label the emotion of the image and music. Manual browsing of the images and music revealed that some of the labels surprised us. Ideally we would use a stronger connection between the auditory and visual domain and not rely solely on a single word emotion label. Another direction we would like to explore is how to fully use the supervised information. The emotion recognition for image and music respectively can be used to learn better representations of the features.

We explored training a image-music emotion correspondence classifier in the evaluation. We like the fully automated nature and scalability of this evaluation method compared to using human evaluators. It could be incorporated into our workflow of building better music generators. In particular, it could be further explored to play a discriminator role in a Generative Neural Network.

While we would not expect all the music generated with our methods may be suitable for use, we think it provides artists with a exploratory tool to enrich their visual artwork. In practice an artist may generate multiple pieces of music and filter or further process them to their liking. In other

words it could reduce the minimal effort required to find suitable music to supplement a visual art piece from music composition to music curation. We hope this could lead to better accessibility of visual art in the future.

Acknowledgments

We thank the reviewers for valuable criticism and suggestions. In particular for pointing us towards the literature on the ordinal nature of emotions.

References

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Hasenfus, N. A. 1978. Cross-media style: A psychological approach. *Poetics* 7(2):207 – 216.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Horn, B.; Smith, G.; Masri, R.; and Stone, J. 2015. Visual information vases: Towards a framework for transmedia creative inspiration. In *Proceedings of the Sixth International Conference on Computational Creativity (ICCC 2015)*, 182–188. Park City, Utah: Brigham Young University.

Huang, C.-Z. A.; Vaswani, A.; Uszkoreit, J.; Shazeer, N.; Hawthorne, C.; Dai, A. M.; Hoffman, M. D.; and Eck, D. 2018. Music transformer: Generating music with long-term structure. *arXiv preprint arXiv:1809.04281*.

Machajdik, J., and Hanbury, A. 2010. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, 83–92. New York, NY, USA: Association for Computing Machinery.

Madhok, R.; Goel, S.; and Garg, S. 2018. Sentimozart: Music generation based on emotions. In *ICAART*.

Malheiro, R.; Rocha, B.; Oliveira, A.; and Paiva, R. P. 2013. Multi-modal music emotion recognition: A new dataset, methodology and comparative analysis. In *in Proc. Int. Symposium on Computer Music Modeling and Retrieval*.

Marcel, S., and Rodriguez, Y. 2010. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, 1485–1488. New York, NY, USA: Association for Computing Machinery.

Marković, D., and Malešević, N. 2018. Adaptive interface for mapping body movements to sounds. In Liapis, A.; Romero Cardalda, J. J.; and Ekárt, A., eds., *Computational Intelligence in Music, Sound, Art and Design*, 194–205. Cham: Springer International Publishing.

Martín-Gómez, L.; Pérez-Marcos, J.; Navarro-Cáceres, M.; and Rodríguez-González, S. 2019. Convolutional neural networks and transfer learning applied to automatic composition of descriptive music. In Rodríguez, S.; Prieto, J.; Faria, P.; Kłos, S.; Fernández, A.; Mazuelas, S.; Jiménez-López, M. D.; Moreno, M. N.; and Navarro, E. M., eds., *Distributed Computing and Artificial Intelligence, Special Sessions, 15th International Conference*, 275–282. Cham: Springer International Publishing.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; dAlché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc. 8024–8035.

Ramachandran, V. 2003. *The emerging mind*. London: BBC/Profile Books.

Ranjan, A.; Gabora, L.; and O'Connor, B. 2013. The cross-domain re-interpretation of artistic ideas.

Sergio, G. C.; Mallipeddi, R.; Kang, J.-S.; and Lee, M. 2015. Generating music from an image. In *Proceedings of the 3rd International Conference on Human-Agent Interaction*, HAI '15, 213–216. New York, NY, USA: Association for Computing Machinery.

Simon, I., and Oore, S. 2017. Performance rnn: Generating music with expressive timing and dynamics. `https://magenta.tensorflow.org/performance-rnn`.

Solèr, M.; Bazin, J.-C.; Wang, O.; Krause, A.; and Sorkine-Hornung, A. 2016. Suggesting sounds for images from video collections. In Hua, G., and Jégou, H., eds., *Computer Vision – ECCV 2016 Workshops*, 900–917. Cham: Springer International Publishing.

Verma, G.; Dhekane, E. G.; and Guha, T. 2019. Learning affective correspondence between music and image. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3975–3979.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2017. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* 39(4):652–663.

Xiaoying Wu, and Ze-Nian Li. 2008. A study of image-based music composition. In *2008 IEEE International Conference on Multimedia and Expo*, 1345–1348.

Yannakakis, G. N.; Cowie, R.; and Busso, C. 2017. The ordinal nature of emotions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 248–255.

You, Q.; Luo, J.; Jin, H.; and Yang, J. 2016. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. *CoRR* abs/1605.02677.

Zhu, X.; Li, L.; Liu, J.; Peng, H.; and Niu, X. 2018. Captioning transformer with stacked attention modules. *Applied Sciences* 8:739.