Photong: Generating 16-Bar Melodies from Images

Yanjia Zhang,[1] Haohan Wang [2]

[1] United World College of South East Asia
[2] School of Information Sciences, University of Illinois Urbana-Champaign
zhang100465@uwcsea.edu.sg, haohanw@illinois.edu

Abstract

This work aims to study the possibility of melody generation based on any arbitrary image using the power of deep-learning neural networks. We suggest a VAE-based pipeline that generates cohesive 16-bar MIDI melodies from images through emotion detection and modality transfer using feature embeddings. To implement this pipeline, we used an image encoder, a MIDI VAE and three bridging computer vision models. We then evaluate the system by examining the musical features of four distinct outputs to see how well they have captured the features of the input images.

Introduction

In the age of machine learning, there have been multiple attempts to connect music with images. A few focus on data from album covers (Chao et al. 2011; Libeks and Turnbull 2011; Oramas et al. 2017), natural scenes (Qiu and Kataoka 2018), paintings (Rivas Ruzafa 2020) and videos (Wang et al. 2012; Yu, Shen, and Zimmermann 2012; Wu et al. 2016, 2012). Nevertheless, most of these works target specific images and are not ideal for generalised use. Learning features from raw audio files also poses a challenge, as audio features may not necessarily have connections with musical features such as melody and rhythm.

Inspired by the methods in (Tham and Kim 2021) and (Zhang 2021), we propose a novel VAE-based system that can generate coherent 16-bar melodies from any image based on its emotion profile. To elaborate, a feature embedding is created from the raw pixels and converted to a MIDI embedding. The arousal and valence values (Mehrabian 1995) are deduced to obtain the tempo and tonality of the generated melody, and this information is added to the decoded MIDI embedding to generate a playable MIDI file.

Method

As seen in Fig. 1, the proposed pipeline contains five key models: an input (image) encoder, an embedding generator, an arousal generator, a valence classifier, and an output (MIDI) decoder. Additionally, to generate the training dataset, another MIDI encoder is used.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

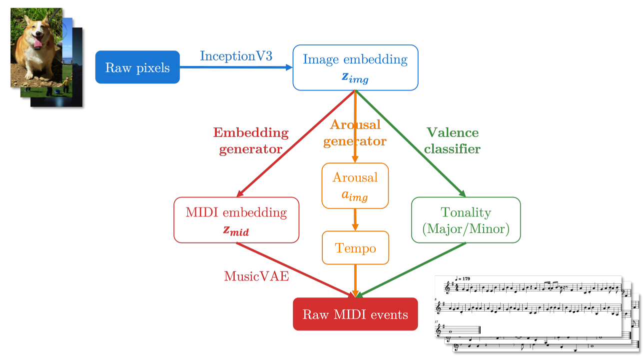


Figure 1: An overview of the pipeline.

Input encoders

Image For this work, we have selected the well-known InceptionV3 model (Szegedy et al. 2016) trained on ImageNet (Deng et al. 2009) as the image encoder. We preprocess the image and pass it to the Inception model without the fully connected layer to generate an image embedding $\mathbf{z_{img}}$ of size $(8, 8, 2048)$.

MIDI We use the pre-trained 16-bar variational autoencoder (`hierdec-mel_16bar`) from the MusicVAE project (Roberts et al. 2018) developed by TensorFlow Magenta, which includes a bidirectional LSTM encoder. We read each MIDI file as a "note sequence" and pass it to an `OneHotMelodyConverter`. It extracts the melody and converts multiple randomly-chosen 16-bar slices to one-hot tensors $\mathbf{x_{mid}}$. Then, we encode them with the MIDI encoder model to obtain an output embedding $\mathbf{z_{mid}}$ of size (512).

Bridging models

Embedding generator We train a model with a 2D convolution layer and multiple dense layers to generate a MIDI embedding $\mathbf{z_{mid}}$ of size (512) from an image embedding $\mathbf{z_{img}}$ of size $(8, 8, 2048)$. This embedding can be passed on to the output decoder to generate a melody, which is saved as a MIDI file. To augment the dataset, we have clustered the image and MIDI embeddings into 10 abstract categories by running the K-nearest neighbour algorithm on arousal and valence as two dimensions. During each training step, em-

Algorithm 1: Training function at step q.

Data: $\mathbf{z}_{\text{img}_q}$ as x, $\mathbf{z}_{\text{mid}_q}$ as y, $\mathbf{c}_q$ as z, list of clusters as $\mathbf{c}_{\text{all}}$.
Result: Loss value at step q.

$\mathbf{x}_q \leftarrow []$;
$\mathbf{y}_q \leftarrow []$;
foreach $c \in \mathbf{c}_{all}$ **do**
    ind $\leftarrow$ vector of positions where $z = c$;
    $\mathbf{x}_c \leftarrow \mathbf{x}[\mathbf{ind}]$;
    $\mathbf{y}_c \leftarrow \mathbf{y}[\mathbf{ind}]$;
    `shuffle`$(\mathbf{y}_c)$;
    $\mathbf{x}_q \leftarrow [\mathbf{x}_q, \mathbf{x}_c]$;
    $\mathbf{y}_q \leftarrow [\mathbf{y}_q, \mathbf{y}_c]$;
$logits =$ `model`$(\mathbf{x}_q, \mathbf{y}_q)$;
$loss =$ `MSE`$(\mathbf{y}_q, logits)$;
`applyGradient`$(loss)$;
return $loss$;

beddings in the same category are shuffled so that the image-MIDI embedding pairs are different every epoch (see Algorithm 1).

Arousal generator We introduce an arousal generator model that can estimate the extent of stimulation of an image, as a probability a_{img}, through binary classification of its embedding. The source dataset is divided into 6 approximately equal classes based on arousal and valence information, and the two most extreme classes are used to train the model . We then apply the following formula to calculate the tempo of the generated melody in BPM (beats per minute):

$$T(a_{img}) = 160 \cdot \frac{1}{1 + e^{-5(a_{img} - 0.5)}} + 40$$

where $0 \leq a_{img} \leq 1$. This gives an output in the range of approximately $[52.14, 187.86]$ centred at $(0.5, 120)$, since 120 BPM is the most common "standard tempo".

Valence classifier We use the valence information to train a binary classification model that can classify whether a given image shows "positive" or "negative" emotions and assign a major or minor tonality accordingly. As a part of the touch-up, "non-diatonic" (out of the scale) notes are randomly moved up or down a semitone to make them so. The tonic is decided based on the first note of the melody and is returned to at the end to establish a sense of completeness. Afterwards, one chord is added to every bar based on the first note of the bar to accompany the melody.

Output decoder

The MusicVAE autoencoder mentioned above comes with a hierarchical LSTM decoder that uses a categorical LSTM decoder at its core. It takes in an embedding $\mathbf{z_{img}}$ of shape (256) and decodes it to a note sequence.

Dataset

Image

Given our request for an image dataset with valence and arousal feature labels, we decide to combine two datasets.

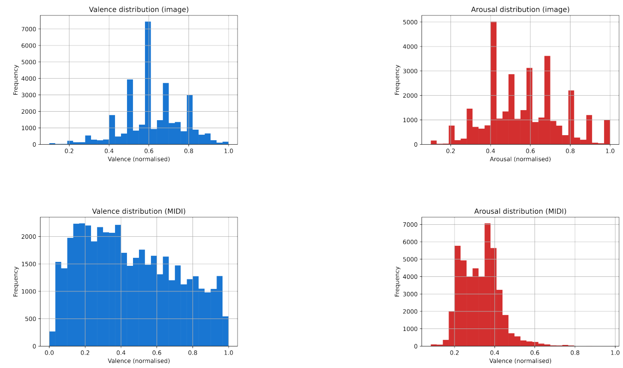


Figure 2: Distribution of the final datasets.

1. **CGnA10766** (Kim et al. 2018), a dataset consisting of $10,766$ images of people, animals and landscapes. Valence and arousal values are provided for the entire image, ranging from $[0, 9]$, and are labelled by volunteer annotators through Amazon Mechanical Turk (AMT).
2. **EMOTIC** (Kosti et al. 2019), a dataset consisting of $23,571$ images of people in the context of their surroundings. For each person, the emotion evoked falls into 26 distinct categories and three continuous dimensions (valence, arousal and dominance) ranging from $[0, 10]$, all labelled by volunteers via AMT.

In particular, to obtain the overall valence and arousal of the images in the EMOTIC dataset, we take the weighted average of all people in an image, where the weight is the relative size of the bounding box of the person to the size of the image itself. In other words, for each image,

$$f_{img} = \frac{\sum_{i=1}^{n} \mathbf{w}_i \mathbf{f}_i}{\sum_{i=1}^{n} \mathbf{w}_i}$$

where $\mathbf{f}$ is a feature vector of each person in the image, and

$$\mathbf{w} = \frac{(\mathbf{b}_3 - \mathbf{b}_1)(\mathbf{b}_4 - \mathbf{b}_2)}{s}$$

where $\mathbf{b}$ is the coordinates of the bounding boxes for each person, given in the form of (x_1, y_1, x_2, y_2), and s is the size of the image.

We then proceed to normalise the arousal and valence values of both datasets to $[0, 1]$ and combine them. Duplicated images and images that produce embeddings with NaN values are removed.

Result In the end, we obtained a total of $33,612$ images with arousal and valence. Fig. 2a reveals that the valence in the combined image dataset is quite unbalanced. In fact, the ratio of high valence (> 0.5) to low valence (≤ 0.5) images is approximately $3 : 1$. One explanation given by Kim et al. (2018) is that "people usually share the positive images, rather than negative images", as happiness is beneficial to health and well-being (Seligman and Csikszentmihalyi 2000). As for arousal, Fig. 2b shows a pretty balanced distribution, with a high-to-low ratio of approximately $1.3 : 1$.

Music

In this work, we use a subset of the Lakh MIDI Dataset (LMD)[1] (Raffel 2016), `LMD-matched`, which contains $45,129$ transcribed MIDI files matched to approximately $31,034$ tracks in the Million Song Dataset (MSD) (Bertin-Mahieux et al. 2011). This allows us to have a huge collection of MIDI files with metadata of the original songs.

Valence MSD has each track labelled with much metadata from the Echo Nest API, but valence is not provided. In addition, the API has been shut down in 2016 and is no longer reachable, so we are unable to retrieve more information using the Echo Nest IDs. Fortunately, AcousticBrainz Labs has provided an archive of mappings between Echo Nest IDs and IDs of other platforms . In this way, we can obtain valence information from Spotify, which uses a proprietary musical analysis tool (from their acquisition of Echo Nest) to deduce a value in the range of $[0, 1]$. For each recognised track, we use the Spotify Web API to query its valence.

Arousal There is no concrete definition for "arousal" in musical terms; in this work, we use the tempo of a song as the intuitive definition. We assume that a faster song leads to stronger positive or negative emotions and vice versa. To incorporate the situation where tempo changes occur, we define the arousal of a MIDI file to be the weighted average of its tempos, where the weight is the relative duration for which a tempo is heard. In other words,

$$a_{mid} = \frac{\mathbf{d} \cdot \mathbf{t}}{\mathbf{d}}$$

where **d** is a vector of the duration of each tempo, the sum of which should add up to the total duration of the MIDI file, and **t** is a vector of tempos expressed in BPM.

Result The final dataset contains approximately $33,380$ MIDI files from $18,395$ tracks. The distribution of valence is reasonable, as shown in Fig. 2c, with the mean at around 0.45. Although the distribution of arousal in Fig. 2d seems quite tilted to the lower range, it is actually because this data is decided by tempo, which ranges from 26 to 310 BPM. Even the 75% percentile is only around 0.39, so the threshold for "low" and "high" arousal should not be set at 0.5.

Training

To prepare the source dataset for training, we have built two training datasets with TensorFlow Dataset.

Embedding dataset To streamline the supervised training process of the embedding generator, we have prepared a dataset with "(image embedding, MIDI embedding, cluster)" triples. To generate this dataset, we first see whether the image or the MIDI dataset has a greater number of samples for each cluster. We then randomly up-sample embeddings from the domain with fewer samples to ensure both are of the same size and write each triple to the dataset. The order of pairing is irrelevant since the embeddings will be shuffled during training anyway. In the end, we have obtained a dataset with $242,200$ examples.

[1]https://colinraffel.com/projects/lmd/

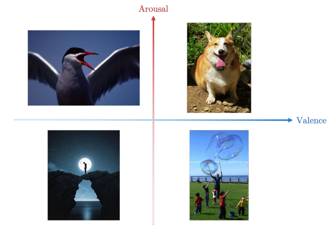
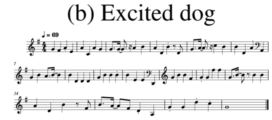


Figure 3: The four images used to test the model.

(a) Angry bird

(b) Excited dog

(c) Sad person



(d) Happy children

Figure 4: Melodies generated for corresponding images, with chords omitted for brevity.

Arousal-valence dataset To integrate the datasets into a useful format that can be used to train arousal and valence models, we have also prepared a dataset with "(image embedding, normalised arousal value, normalised valence value, arousal class, valence class)" information, which all come from the source datasets above. During training, only the image embedding and the target feature are extracted and used.

Results

To analyse the performance of the model musically, we have selected four images with varying arousal and valence, as seen in Fig. 3. The resulting music sheets generated by MuseScore are shown in Fig. 4, and the corresponding audio has been made public[2].

The tempos for images with high and low arousal values are very different. Fig. 4a and Fig. 4b show that the two emotionally stimulating images both lead to a melody of 179 BPM, which is high in the range. Sometimes there are peculiar variations; for instance, Fig. 4c and Fig. 4d show that the happy children image receives a lower tempo (at 69 BPM) than the sad person image (at 83 BPM). Nonetheless, this is understandable as after the image was resized during pre-processing, the detailed features of the children might have been lost and could not be captured accurately by InceptionV3. Another observation is that the two images with higher valence are assigned a major key, while the other two with lower valence are assigned a minor key. This is within our expectations.

The range of the melody generated is quite impressive, almost always spanning more than an octave. Sometimes the melody even changes to the bass clef, as seen in bars

[2]https://v2.photong.ml/samples

6 and 10 of Fig. 4d. In the playback, with chords two octaves lower, this adds an interesting interaction between the melody and the accompaniment. To give an example, the D2 note on the last beat of bar 10 leads nicely into the G chord on beat 1 of bar 11, creating a satisfying perfect cadence. Rhythm-wise, it seems like the melodies mostly contain quarter notes, although there are many interesting variations in the rhythm. For example, dotted notes create a triplet groove, and syncopation (offbeat) can be heard in all four extracts. These are sophisticated features that add rhythmic diversity to the melody. There are also notes with different duration, notably the shorter notes interpreted as staccatos.

Conclusion

In this work, we present a system with three original models to achieve modality transfer between an image and a 16-bar melody using embeddings. They are trained on established emotional features (valence and arousal) to detect the emotions in the image and generate a melody with features that represent these emotions. To demonstrate the capability of the models, we have performed a musical analysis of the generated melodies for four distinct images.

Of course, there are aspects that could be further explored. For example, training the valence model is an imbalanced classification task as there are fewer images with lower valence for reasons described in the dataset section, which can be optimised using methods not yet studied in this work. The source dataset is another element that can be improved on. Although MSD contains (at the time) "contemporary popular music tracks" (Bertin-Mahieux et al. 2011), pop music has evolved notably since then. It would be preferred to have a dataset with newer songs, more genres and artists from different countries so that the model can learn a variety of musical styles. The embedding model can be enhanced by experimenting with techniques such as latent constraints (Engel, Hoffman, and Roberts 2018) and attribute vector arithmetic (Carter and Nielsen 2017). This allows adjustments of certain features of the output embedding, including tonality and note density. An robust evaluation system could also be employed to test the models on more diverse images.

References

Bertin-Mahieux, T.; Ellis, D. P. W.; Whitman, B.; and Lamere, P. 2011. The Million Song Dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*. Columbia University.

Carter, S.; and Nielsen, M. 2017. Using Artificial Intelligence to Augment Human Intelligence. *Distill*, 2(12): e9.

Chao, J.; Wang, H.; Zhou, W.; Zhang, W.; and Yu, Y. 2011. Tunesensor: A Semantic-Driven Music Recommendation Service for Digital Photo Albums. In *Proceedings of the 10th International Semantic Web Conference. ISWC2011 (October 2011)*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-j.; Li, K.; and Li, F.-f. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. Miami, FL: IEEE. ISBN 978-1-4244-3992-8.

Engel, J. H.; Hoffman, M. D.; and Roberts, A. 2018. Latent Constraints: Learning to Generate Conditionally from Unconditional Generative Models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Kim, H.-R.; Kim, Y.-S.; Kim, S. J.; and Lee, I.-K. 2018. Building Emotional Machines: Recognizing Image Emotions Through Deep Neural Networks. *IEEE Transactions on Multimedia*, 20(11): 2980–2992.

Kosti, R.; Alvarez, J.; Recasens, A.; and Lapedriza, A. 2019. Context Based Emotion Recognition Using EMOTIC Dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1.

Libeks, J.; and Turnbull, D. 2011. You Can Judge an Artist by an Album Cover: Using Images for Music Annotation. *IEEE Multimedia*, 18(4): 30–37.

Mehrabian, A. 1995. Framework for a Comprehensive Description and Measurement of Emotional States. *Genetic, Social, and General Psychology Monographs*, 121(3): 339–361.

Oramas, S.; Nieto, O.; Barbieri, F.; and Serra, X. 2017. Multi-Label Music Genre Classification from Audio, Text, and Images Using Deep Features. *arXiv:1707.04916 [cs]*.

Qiu, Y.; and Kataoka, H. 2018. Image Generation Associated With Music Data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2510–2513.

Raffel, C. 2016. *Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching*. Ph.D. thesis, Columbia University.

Rivas Ruzafa, E. 2020. *Pix2Pitch: Generating Music from Paintings by Using Conditionals GANs*. Master's thesis, Universidad Politécnica de Madrid.

Roberts, A.; Engel, J.; Raffel, C.; Hawthorne, C.; and Eck, D. 2018. A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music. In *Proceedings of the 35th International Conference on Machine Learning*, 4364–4373. PMLR.

Seligman, M. E. P.; and Csikszentmihalyi, M. 2000. Positive Psychology: An Introduction. *American Psychologist*, 55(1): 5–14.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818–2826. Las Vegas, NV, USA: IEEE. ISBN 978-1-4673-8851-1.

Tham, I.; and Kim, M. 2021. Generating Music Using Deep Learning. https://towardsdatascience.com/generating-music-using-deep-learning-cb5843a9d55e.

Wang, J.-C.; Yang, Y.-H.; Jhuo, I.-H.; Lin, Y.-Y.; and Wang, H.-M. 2012. The Acousticvisual Emotion Guassians Model for Automatic Generation of Music Video. In *Proceedings of the 20th ACM International Conference on Multimedia - MM '12*, 1379. Nara, Japan: ACM Press. ISBN 978-1-4503-1089-5.

Wu, X.; Qiao, Y.; Wang, X.; and Tang, X. 2012. Cross Matching of Music and Image. In *Proceedings of the 20th ACM International Conference on Multimedia - MM '12*, 837. Nara, Japan: ACM Press. ISBN 978-1-4503-1089-5.

Wu, X.; Qiao, Y.; Wang, X.; and Tang, X. 2016. Bridging Music and Image via Cross-Modal Ranking Analysis. *IEEE Transactions on Multimedia*, 18(7): 1305–1318.

Yu, Y.; Shen, Z.; and Zimmermann, R. 2012. Automatic Music Soundtrack Generation for Outdoor Videos from Contextual Sensor Information. In *Proceedings of the 20th ACM International Conference on Multimedia - MM '12*, 1377. Nara, Japan: ACM Press. ISBN 978-1-4503-1089-5.

Zhang, Y. 2021. Groovy Pixels: Generating Drum Set Rhythms from Images. In *2021 3rd International Conference on Advanced Information Science and System (AISS 2021)*, 1–5. Sanya China: ACM. ISBN 978-1-4503-8586-2.