SentiMozart: Music Generation based on Emotions

Rishi Madhok^{1,*}, Shivali Goel^{2,*} and Shweta Garg^{1,*}

[1]*Department of Computer Science and Engineering, Delhi Technological University, New Delhi, India*

[2]*Department of Information Technology , Delhi Technological University, New Delhi, India*

Keywords: LSTM, Deep Learning, Music Generation, Emotion, Sentiment.

Abstract: Facial expressions are one of the best and the most intuitive way to determine a persons emotions. They most naturally express how a person is feeling currently. The aim of the proposed framework is to generate music corresponding to the emotion of the person predicted by our model. The proposed framework is divided into two models, the Image Classification Model and the Music Generation Model. The music would be generated by the latter model which is essentially a Doubly Stacked LSTM architecture. This is to be done after classification and identification of the facial expression into one of the seven major sentiment categories: Angry, Disgust, Fear, Happy, Sad, Surprise and Neutral, which would be done by using Convolutional Neural Networks (CNN). Finally, we evaluate the performance of our proposed framework using the emotional Mean Opinion Score (MOS) which is a popular evaluation metric for audio-visual data.

1 INTRODUCTION

It was only a few decades ago that Machine Learning was introduced. Since then the phenomenal progress in this field has opened up exciting opportunities for us to build intelligent computer programs that could discover, learn, predict and even improve itself given the data without the need of explicit programming. The advancements in this field have allowed us to train computer not only to mimic human behavior but also perform tasks that would have otherwise required considerable human effort.

In our paper we present two such problems. First, we try and capture the emotions of people from their images and categorize the sentiments into 7 major categories: Angry, Disgust, Fear, Happy, Sad, Surprise and Neutral using Convolutional Neural Network (CNN).

Second, we try to generate music based on these emotions.Music composition, given its level of complexity, abstractness and ingenuity is a challenging task and we aim to generate music that not necessarily approximates to the exact way people play music but one which is pleasant to hear and is appropriate according to the mood of the situation in which it is being played.

A challenge accompanied with music generation is that music is a flowing melody in which one note can be followed by several different notes, however, when humans compose music they often impose a set of restrictions on notes and can choose exactly one precise note that must follow a given sequence of notes. One music pattern cant repeat itself forever, apart from adding a local perspective to our music generation algorithm, global view must also be considered. This shall ensure correct departure from a pattern at the correct instant of time. Magentas RL Tuner (Jaques et al., 2016) is an advancement in this direction however more concentrated effort is required for overcoming such difficulties while developing a near-to-ideal music generation algorithm based on rule discovery. Music experts and machine learning enthusiasts must work hand in gloves with each other so as to create complete orchestral compositions, orchestrating one instrument at a time. Another challenge is the development of flexible programs. Extensive training with large number of iterations and epochs tend to produce more accurate results but there is a need of periodically updating and adding additional information to our program so as to get a better understanding of music composition and artificial intelligence.

So far, fusion of analyzing sentiments from images and generation of music on the basis of that hasn’t been explored much. Previous work done in

*These authors contributed equally to this work.

Madhok, R., Goel, S. and Garg, S.
SentiMozart: Music Generation based on Emotions.
DOI: 10.5220/0006597705010506
In *Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART 2018) - Volume 2*, pages 501-506
ISBN: 978-989-758-275-2

this field includes (Sergio et al., 2015) in which the author analyses the image using the HSV color space model. The author converts the RGB format image to HSV format and then scans the image to find the dominant colors in certain patches of the image. This results in the generation of a color map which is then directly mapped to musical notes which were calculated using the harmonics of a piano. This notes map is then used to generate music. The problem with this approach is that it does not work for a dataset in which every image is a face of a person. The most dominant color every time, would be similar in all the images namely the skin color hence the same set of notes would be generated.

Our model can be incorporated as a feature in various social networking applications such as Snapchat, Instagram etc. where when a person is uploading his/her Snapchat story, depending on the image obtained from the front camera of the electronic device, a sentiment analysis of the image can be done which can then be used to generate music for the user which the user can add to his/her Snapchat story. For example, we would generate a party type peppy melody for a happy expression and a soft romantic melody for a sad expression.

In this paper, we discuss the previous work done in the field of music generation in the next section Section II. Section III describes the proposed model and the details of the training of the model, which is then followed by the analysing of the results obtained in Section IV. We use the Mean Opinion Score (MOS) approach to showcase our results. We finally conclude our paper in Section V.

2 RELATED WORK

2.1 Music Generation

Since last many years a lot of work has been done in generating music. (Graves, 2013) uses the Recurrent Neural Network (RNN) architecture, Long Short Term Memory (LSTM) for generating sequences such as handwriting for a given text. (Chung et al., 2014) extend this concept to compare another type of RNN which is Gated Recurrent Unit (GRU) with LSTM for music and speech signal modeling. Inspired by this approach our model uses LSTM architecture to generate music. (Boulanger-Lewandowski et al., 2012) introduce a probabilistic approach, a RNN-RBM (Recurrent Neural Network Restricted Boltzmann Machine) model which is used to discover temporal dependencies in high space model. The authors introduce the RBM model so that representation of the complicated distribution for each time step in the sequence is easily possible where parameters at a current time step depends on the previous parameters. Currently, the Google Brain Team is working on its project Magenta which would generate compelling music in the style of a particular artist. While magenta also generates music, our model does it based on emotions (facial sentiments). Magentas code has been released as open source as they wish to build a community of musicians, coders and machine learning enthusiasts. (Jaques et al., 2016).

2.2 Music Playlist Generation

Our work shares the high-level goal of playing music based on emotions with (Zaware et al., 2014) however while we aim to generate music, they present a playlist of songs to a user based on the users current mood. They converted images into a binary format and then fed into a Haar classifier for facial detection. Important features were then extracted such as eyes, lips etc. to classify the emotion of the person. (Hirve et al., 2016) use the same thought for generating music as done by (Zaware et al., 2014), they use a Viola-Jones algorithm for face detection and Support Vector Machine (SVM) for emotion detection. While the work by these authors provide music based on a persons emotion, they present a list of songs as a playlist which is a recommender system rather than generating new music.

2.3 Mapping Colors to Music

Similar work also includes the research project by Gregory Granito (Gregory, 2014) that links colors used in the image with music. The author uses a method to calculate the average color value of the image, searches for areas of high concentration of a color and of major changes in hue, this is then used to associate certain colors to motions such as the Yellow color invokes a cheerful emotion in a person. (Sergio et al., 2015) analyses image using the HSV color space model. The author scans the image and generates a color map for the image which is then used to obtain the notes map and music is subsequently generated. While this direct mapping approach is very popular, it has a shortcoming we cannot use this for music generation based on peoples facial expressions because each image in the dataset focuses only on the face of a person and hence would generate the same dominant color for every image.

2.4 Speech Analysis and Hand Gestures to Generate Music

Apart from using images, music has previously been generated by many other ways such as after doing the sentiment analysis of speech of a person or by making use of hand motions. (Rubin and Agrawala, 2014) is one such example in which emotion labels on the users speech are gathered using methods such as hand-labeling, crowd sourcing etc. which is then used to generate emotionally relevant music. A novel application for generation of music was introduced by (Ip et al., 2005) in their work in which they make use of motion-sensing gloves to allow people even non-musicians to generate melodies using hand gestures and motions.

3 PROPOSED FRAMEWORK

3.1 Research Methodology

The proposed framework consists of two models, the first is the Image Classification model and the second is the Music Generation model. The former model classifies the image of the person into one of the seven sentiment class i.e. Angry, Disgust, Fear, Happy, Sad, Surprise and Neutral and the latter model then generates music corresponding to the identified sentiment class.

In the Image Classification model, a Convolutional Neural Network (CNN) is used. The input to the CNN is a 48X48 grayscale image which is then passed onto the first Convolutional layer consisting of 64 filters of size 3X3 each. The output of this layer is then passed onto the second convolutional layer consisting of 128 filters of size 3X3 each. After these two layers, the feature maps are down sampled using a Max Pooling Layer having a window size of 2X2. These set of layers are repeated again with the same specification of layers and filter sizes as shown in Figure 1. Following the second Max Pooling layer, the output is now passed onto the Fully Connected Layer or the Dense Layer having 7 nodes, one for each sentiment class. The output of this Dense layers goes through a Softmax layer at the end, which then finally predicts the sentiment of the image. More details about this model are explained in the section below.

In the Music Generation Model, a Doubly Stacked LSTM Architecture is used as shown in Figure 1. A doubly stacked LSTM architecture is essentially an architecture in which one LSTM layer is stacked over the other LSTM layer and outputs from the first layer is passed onto the second layer. After the Image Classification task, a vector of length 7 is obtained in one hot encoded form, whose each value corresponding to respective sentiment class. This vector is then converted to a new vector of length 3 for Happy, Sad and Neutral classes respectively to choose the dataset for the Music Generation Model as shown above in the Figure 1 Certain classes are merged for the Music Generation Task such as Fear, Disgust and Angry were merged into the Sad Sentiment Class. Also, the Surprise sentiment class was merged into the Happy sentiment class. The dataset contains 200 MIDI files for each sentiment class i.e. Happy, Sad and Neutral. As there was no such dataset of MIDI files available for emotions, these MIDI files were labeled by a group of 15 people. This is explained diagrammatically in Figure 1 and more details about this model are explained in the section below.

3.2 Image Classification Model

The Convolutional Neural Network (CNN) model was implemented for the classification of images into its respective sentiment class. The dataset was split into 80% of the images used for training, and remaining 20% used for cross validation. The model used batch training with a batch size of 256, in which the images were processed in batches of the given size. Consequently, all the training samples were processed similarly through the learning algorithm which comprised one forward pass. After finishing one forward pass, a loss function, here categorical cross entropy loss function, was computed which was minimized by back propagating through all the images and updating the weight matrices using the Adaptive Gradient Descent Optimizer(Adagrad), this comprised one backward pass. One forward pass and one backward pass comprise a single epoch. There were a total of 100 epochs with 1 epoch taking an average of 13 minutes to run on a standard 2.5 GHz i5 processor. The CNN Model had 6 hidden layers, with 4 convolutional layers and 2 max-pool layers which were alternated and 1 fully connected layer with 7 neurons each for the respective sentiment class. Two dropout layers were also added which helped to prevent overfitting.

3.3 Music Generation Model

A doubly stacked LSTM model is trained for the Music Generation Model. The input given to the LSTM is the one Hot Representation of the MIDI files from a particular dataset belonging to a sentiment class. The one hot representation of a MIDI file has

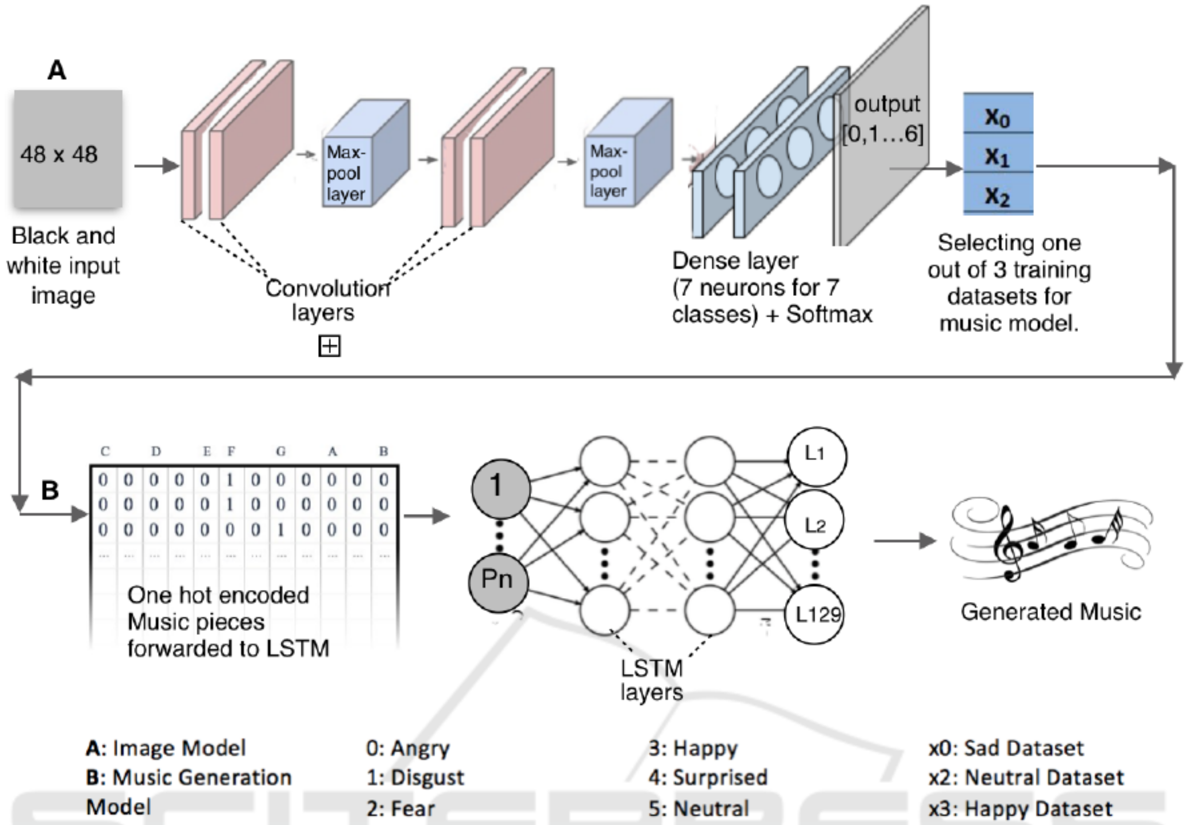

Figure 1: SentiMozart Model.

a shape as (sequence_length, Notes_in_MIDI) where sequence_length is the number of notes in the MIDI file and Notes_in_MIDI is 129 (128 for each MIDI Note number + 1 for EOS). Thus, the overall shape of the dataset which is given as input to the LSTM is (Num_files, Max_Sequence_Length,Notes_in_MIDI) where Num_files is 200 MIDI files in the dataset, Max_Sequence_Length is the maximum sequence length of one of the files in the dataset and Notes_in_MIDI is 129 as explained earlier.

The input is then passed through the first LSTM Layer which outputs sequences, which are then passed to the second LSTM Layer. The Second LSTM layer then outputs a vector which is passed finally through the Fully Connected Layer having 129 nodes. After each layer except for the last layer,there is a Dropout Layer which prevents overfitting of the model. Categorical cross entropy is used as the Loss function of the model and Adaptive Moment Estimation (Adam), a variant of Gradient Descent is used as the Optimizer of the model. The model is run for 2000 epochs following which a MIDI file is generated for the particular sentiment.

3.4 Why MIDI?

The proposed model uses MIDI (Musical Instrument Digital Interface) files to generate music. MIDI format is one of the best options present today to generate new music as the file contains only synthesizer instructions hence the file size is hundreds of times smaller than the WAV format (Waveform Audio Format) which contains digitized sound and hence has a very large size (hundreds of megabytes).

One advantage of WAV format over MIDI is that the WAV format has a better quality of sound because sound depends on the sampling rate and hence will be same for different computers. This is not the case for MIDI files, and hence the quality of sound is different for different computers. Hence we have a trade-off.

4 RESULTS

4.1 Dataset Used

Dataset used for image sentiment classification was the collection of 35,887 grayscale images of 48X48 pixel dimensions prepared by Pierre-Luc Carrier and

Aaron Courville for their project (Goodfellow et al., 2013).

MIDI files corresponding to all 7 categories of facial expressions could not be found. Hence, the following sentiment classes - sad, fear, angry and disgust were merged into the sad sentiment class, happy and surprise were merged into the happy sentiment class and left neutral class as it is. Exploring better evaluation metrics for judging emotion of generated music remains a challenge. Self-prepared dataset of 200 midi files for each sentiment class - Happy, Sad and Neutral was used for the training of music generation model. The music files were mostly Piano music MIDI files and annotated manually by a group of 15 people. The MIDI files dataset is available by request from the authors.

4.2 Result of Image Classification Model

The accuracy obtained in the proposed model was 75.01%. Figure 2 shows the gradual decrease in the loss number obtained for each epoch in the Image Sentiment Classification model till it finally tends to converge after 100 epochs.

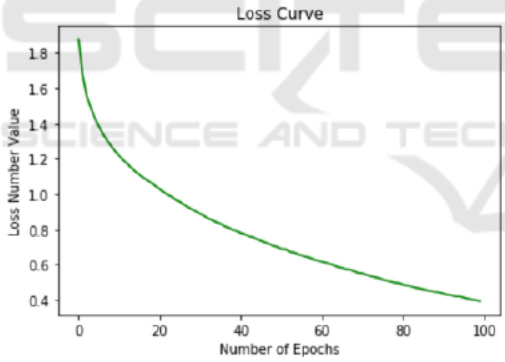

Figure 2: Loss Curve for CNN Model.

4.3 Evaluation Model

To evaluate the quality of our model we use emotional Mean Opinion Score (MOS) which is a common measure for quality evaluation of audio-visual data. Since we classify the songs based on three emotional categories, we choose a rating scale of 0 to 10, where 0 represents sad emotion, 5 represents neutral and 10 represents happy emotion. The rating scale is shown in Figure 3.

30 people gave scores based on the rating scale for 30 randomly chosen images (10 of each class) and their corresponding music generated. The average of the scores was taken and plotted on the graph shown in Figure 4. The images and music was presented before the participant in no particular order so as to avoid any bias between the type of image shown and the likeliness of music belonging to the same category.

Correlation between the two curves was calculated to be 0.93. It is calculated according to the formula given below. As expected this must be high as the facial expression and the music generated must express the same sentiment. During the experiment we observed extreme scoring for facial data but conservative scoring for scoring generated music. Other important observation was for both neutral images and songs most participants tend to assign a perfectly balanced score 5, while deviated the scores conservatively for sad and happy emotion.

Where x is the Image MOS, y is the Music MOS and r determines the correlation. The value of r is always between 1 and 1. $\bar{x}$ and $\bar{y}$ are the means of the x and y values respectively.

When we used upbeat and peppy songs in our training set for the happy emotion, an output file with a similar pattern of notes was generated. Also, when slow and soothing melodies were used in the training set for a sad emotion, consequently, a melody having similar patterns was generated. The model generates the type of songs it is supposed to, that is, depending upon the emotion detected, for instance sad ones for a sad expression and happy ones for a happy expression.

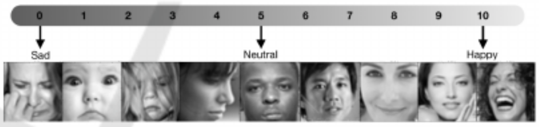

Figure 3: MOS Scale.

5 CONCLUSION

In this paper, the SentiMozart framework is presented which generates relevant music based upon the facial sentiments of a person in the image. It is able to do so by first classifying the facial expression of the person in the image into one of the seven sentiment classes by using the Image Classification Model and then it generates relevant music based upon the sentiment of the person using the Music Generation Model. Finally, we evaluated the performance of our proposed framework using the emotional Mean Opinion Score (MOS) which is a popular evaluation metric for audio-visual data. The high correlation value and the user analysis on the generated music files shows

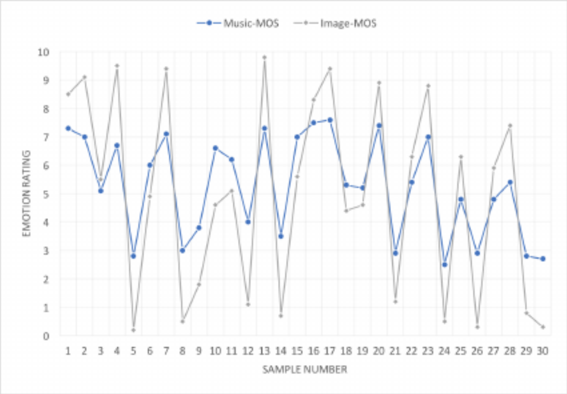

Figure 4: MOS Graph.

the self-sufficiency of the number of training samples in the Music Generation Model.

Music is not just art and creativity but is also based on a strong mathematical background. So, often computer creativity is doubted in its ability to produce fresh and creative works. New notions of computer creativity can evolve by amalgamation of different techniques, use of high performing systems and bio inspiration.

Also, computer scientists and music composers must work in synergy together. Making music composers able enough to use the program by having a basic understanding of the program and learning various commands would enable them to give constructive feedback and radically change the process of music composition, and consequently the way market for music operates. Market opportunities would include incorporating such features in instant multimedia messaging applications such as Snapchat, Instagram and any other application which deals with images.

ACKNOWLEDGEMENTS

We whole heartedly thank our mentor Dr. Rajni Jindal, Head of Department, Computer Science Department at Delhi Technological University for guiding us through this project.

REFERENCES

Boulanger-Lewandowski, N., Bengio, Y., and Vincent, P. (2012). Modeling Temporal Dependencies in High-Dimensional Sequences: Application to Polyphonic Music Generation and Transcription. *ArXiv e-prints*.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *ArXiv e-prints*.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., Ionescu, R., Popescu, M., Grozea, C., Bergstra, J., Xie, J., Romaszko, L., Xu, B., Chuang, Z., and Bengio, Y. (2013). Challenges in Representation Learning: A report on three machine learning contests. *ArXiv e-prints*.

Graves, A. (2013). Generating Sequences With Recurrent Neural Networks. *ArXiv e-prints*.

Gregory, G. (2014). Generating music through image analysis.

Hirve, R., Jagdale, S., Banthia, R., Kalal, H., and Pathak, K. (2016). Emoplayer: An emotion based music player. *Imperial Journal of Interdisciplinary Research*, 2(5).

Ip, H. H. S., Law, K. C. K., and Kwong, B. (2005). Cyber composer: Hand gesture-driven intelligent music composition and generation. In *11th International Multimedia Modelling Conference*, pages 46–52.

Jaques, N., Gu, S., Bahdanau, D., Hernández-Lobato, J. M., Turner, R. E., and Eck, D. (2016). Sequence Tutor: Conservative Fine-Tuning of Sequence Generation Models with KL-control. *ArXiv e-prints*.

Rubin, S. and Agrawala, M. (2014). Generating emotionally relevant musical scores for audio stories. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14, pages 439–448, New York, NY, USA. ACM.

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, pages 213–216, New York, NY, USA. ACM.

Zaware, N., Rajgure, T., Bhadang, A., and Sapkal, D. (2014). Emotion based music player. *International Journal of Innovative Research and Development*, 0(0).