Melody Muse

Karanvir Bath
ksbath@sfu.ca
Simon Fraser University
Burnaby, British Columbia, Canada

Ritchie Kumar
ritchiek@sfu.ca
Simon Fraser University
Burnaby, British Columbia, Canada

Parmveer Dayal
psdayal@sfu.ca
Simon Fraser University
Burnaby, British Columbia, Canada

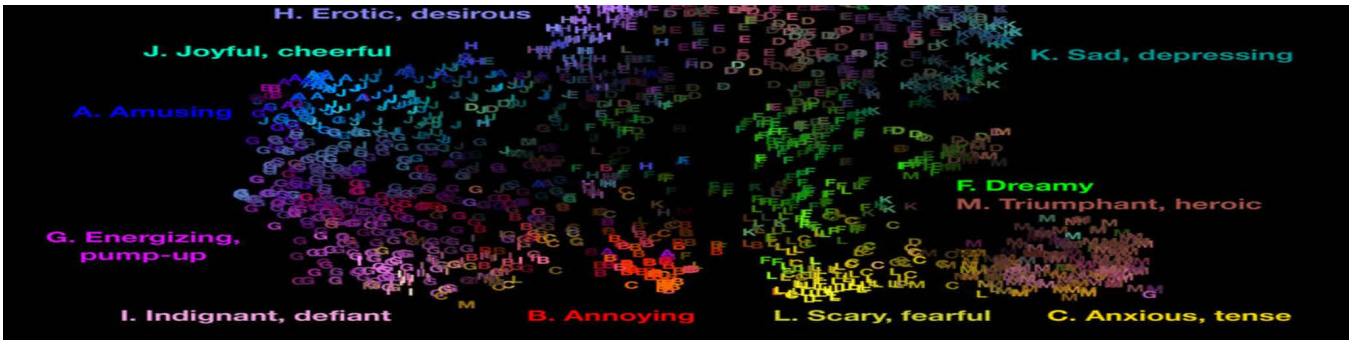

Figure 1: Abstract visualization of datasets consisting of multiple excerpts and full songs annotated with valence and arousal values both continuously (per-second) and over the whole song.

ABSTRACT

In the dynamic fields of artificial intelligence and music creation, generating emotionally expressive music involves a number of tools in a collaborative effort between the two fields. The complexity of this project comes from mapping abstract emotional concepts onto subjective musical elements while maintaining the expressive quality behind the resulting music. To address this, we propose using neural networks to produce MIDI files of piano music. This approach utilizes image labelling to provide a visual stimuli for the user to better associate the generated music with. Through this approach, we integrate the two fields of emotion recognition and generation to provide a practical application behind our project. Our work provides a tool for aspiring musical creators and explores the current capabilities of artificial intelligence and music synthesis.

1 INTRODUCTION

In science fiction, we used to imagine humans being able to produce anything they could imagine within seconds, however, with the explosion of artificial intelligence, this has become reality. With tools such as real-time object detection or AI-powered voice generation, artificial intelligence has become more commonplace in our lives today and leads the charge in enabling regular people to create content across every field [1]. Image detection and labelling has improved significantly over recent years while other fields such as art and music generation is growing in popularity [2, 3]. Building on this, our project explores the capabilities of artificial intelligence when paired with affective labelling and music generation.

There are two fundamental approaches to generating musical pieces with artificial intelligence: rule-based and adaptive. Rule-based methods like Pachet and Roy's, adhere to predefined musical styling rules. Within the context of these rules, there are very strict constraints like how bass leads in cadences which is problematic for generating new compositions as it relies on the developer's understanding of musical rules [4, 5]. On the other hand, the adaptive approach, captures information such as notes or chords and uses it to compose a piece following the same form [4]. Improving on these tools is important because artificial intelligence is having a cultural impact on society, but lacks the tooling for widespread change [6]. Currently, over 70% of AI-powered music generation is done through deep learning, with tools like transformers and GAN-based solutions becoming popular recently [3]. However, the field faces a variety of challenges ranging from the quality and diversity of it's compositions to the ethical implications of the produced music. In addition, these tools struggle with effectively producing music because our perception of music is highly subjective and emotion based music generation is a largely untouched field [3, 7]. AI-generated pieces might lack emotional depth or expressiveness while sounding repetitive and unmemorable. This problem is more prominent in music generation than other art generation tools because these tools lack lyrics. However, adding lyrics makes this task exponentially more difficult. Lastly, in comparison to other fields, there's a lack in high-quality training data that relates music to emotions. Many of these issues require collaboration between different fields for any significant advancements to be made [7].

To address these challenges, our project utilizes an approach that integrates image processing with music generation. The idea behind these two modalities was to give the user a visual stimuli to associate the generated music with. The visual stimuli acts to make the user more biased to the generated music because they unconsciously develop an expectation of the music impacting their impression of it [8]. Our solution utilizes the 'Musical Instrument Device Interface' (MIDI) with Convolutional and Recurrent Neural Networks to create piano pieces alongside single-frame images to reduce the complexity of the task while having a degree of accuracy by utilizing features like Mel-frequency cepstral coefficients or pitch

and tempo [7]. By keeping the complexity of project low from the instrument choice, output length, and feature selection, we could focus on enhancing the diversity of our generated outputs. However, a limiting factor that we could not overcome was from the limited selection of data sets available to us. Despite this, our project acts as a method to aid aspiring musicians entering the field of music creation.

2 APPROACH

Our approach consists of two main components: image labelling and music generation, where the output of the image labelling is the input for the music generation.

2.1 Affective Image Labelling

Our convolutional neural network (CNN) model from the Keras library is illustrated in Figure 2 and takes three inputs: the image's pixel values, brightness, and colorfulness. The image input goes through 2 convolutional layers with max-pooling layers in between to extract hierarchical features. The brightness and colorfulness inputs are then concatenated with the flattened output of the image input. Following the concatenation layer, dropout and batch normalization layers are incorporated to prevent over-fitting and improve model generalization.

The EmoSet dataset was created by Yang et al. for the purpose of facilitating research in visual emotion analysis. Consequentially, the data is annotated with features such as emotion, brightness, colorfulness, scene type, object class, facial expression, and human action. From these features, we chose to use the emotion annotation, brightness and colorfulness data. We did not use the other features because it would've expanded the scope of our project beyond what's relevant as the focus was music reflecting the emotions behind an image [9]. Preprocessing of the dataset involved loading each music sample and extracting features related to brightness and colorfulness using JSON annotation files accompanying the audio files.

The dataset was split into training, validation, and test sets using an 80-10-10 ratio with 50,000 samples used. The dataset contains 118,000 samples, however, as we increased the number of samples, the time it took for the model to run was drastically increasing. A batch size of 64 with 10 epochs was used during training, and the model was trained with early stopping and learning rate reduction callbacks to prevent over-fitting. The Adam optimizer was employed with a categorical cross-entropy loss function. We chose to use the Adam optimizer for its adaptive learning rate method that enables faster convergence without large costs to performance while the loss function was chosen because our samples belong to mutually exclusive classes.

2.2 Music Generation

For music generation, we created a Recursive Neural Network (RNN) model that takes the output of the Affective Image Labelling CNN model, in order to predict and generate a musical clip of that emotion. The RNN model consists of a dropout layer to prevent over-fitting, three Long Short-Term Memory (LSTM) layers to learn the long-term dependencies, and five dense layers, including one being an output layer using a softmax activation function, which converts the raw model outputs into probability scores for each emotional group.

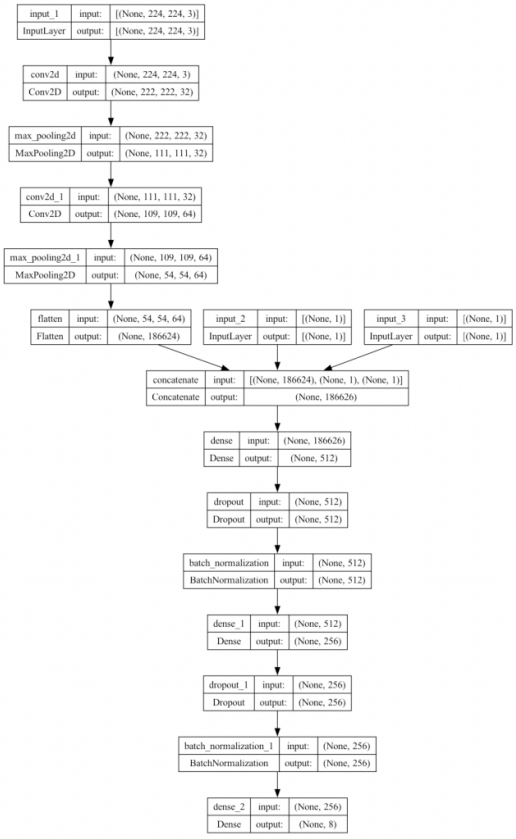

Figure 2: Architecture of the Convolutional Neural Network (CNN) model used for Image Labelling

The RNN model was trained using a Turkish Music Emotion dataset created by Er and Aydilek. This dataset has 400 musical samples and consists of features such as Mel Frequency Cepstral Coefficients (MFCCs), tempo, and chromagram. The dataset also labels each sample with one of four emotions (happy, sad, relaxed, and angry). The 13 MFCC values and corresponding emotional label for each sample were used to train the model. We made sure to focus on only MFCC values because of MFCC values being able to successfully predict human emotions from speech such as in a study conducted by Likitha et al. [10]. In this study, researchers extracted MFCC values from audio clips of human speech and found the system was able to predict happy, sad, and anger about 80% of the time. Based on the findings of this study, we determined that utilizing the approach of MFCC values was the most effective way for us to generate music based on emotion.

The 400 sample dataset, consisting of 100 samples for each emotion, was split into training and test sets using an 80-20 ratio. The model was trained using the training set with a batch size of 32 and epoch value of 75 as these values worked best with the model

and dataset. Just like our CNN image labelling model, this RNN model was compiled using the Adam optimizer and categorical cross-entropy loss function.

After training the RNN model, a random feature from the test set was taken to act as a seed based on whichever emotion was predicted from our image labelling model. Using this seed, the model begins to predict the next features. These newly generated features from the model were turned into a midi file using the "music21" library.

Finally, a human survey made up of 18 participants was conducted in order to validate if the generated music was able to convey the correct emotional label. Each participant was shown one randomly selected musical clip. Participants were then simply asked if the generated music matched the emotion it is supposed to be conveying, and their response was recorded.

3 DATASET

We use the EmoSet dataset for training and evaluation of our image labelling model. The dataset consists of 118,000 annotated samples for the emotions: amusement, awe, contentment, excitement, anger, disgust, fear, and sadness. While we could not find inter-rater agreement scores for the dataset, their process for annotation involved hiring 60 annotators and using the annotation when the agreement rate was over 85% [9]. We used 50,000 randomly selected samples for training and testing with a focus on the features: emotion, brightness and colorfulness.

Turkish Music Emotion dataset was used for the training of our music generation model. This dataset was created by Er and Aydilek with the purpose of investigating emotions by using chroma spectrogram values [11]. The dataset consists of 100 samples each for the emotions happy, sad, angry, and relaxed. The dataset also contains many acoustic features such as MFCC, tempo, fluctuations, and harmonics. However, for the purpose of our study, only MFCC values were used. Since the dataset was very clean and did not contain many extreme outliers, we used all 400 samples to train and validate our model.

4 EXPERIMENTS AND RESULTS

4.1 Affective Image Labelling

Our CNN model was trained on 50,000 samples, employing a batch size of 64 and running for 10 epochs with the Adam optimizer and categorical cross-entropy loss function. During training, we observed over-fitting while training the model on smaller datasets. This led us to incorporating early stopping and learning rate reduction callbacks to prevent over-fitting.

Upon evaluation, our model achieved an accuracy of 31% on the test set as seen in Figure 4. The precision, recall, and F1-score were 35%, 33%, and 31% respectively. The confusion matrix in Figure 3 revealed a weak bias towards the "amusement" emotion, indicating the model was over-fitting images to this emotion. The confusion matrix values (0-8) corresponded to the emotions: amusement, awe, contentment, excitement, anger, disgust, fear, and sadness. Our goal with this model was to accurately label an image with an emotion, which the model was successful in doing so. However, there is room for further improvements by removing the bias in the confusion matrix and raising the accuracy metrics.

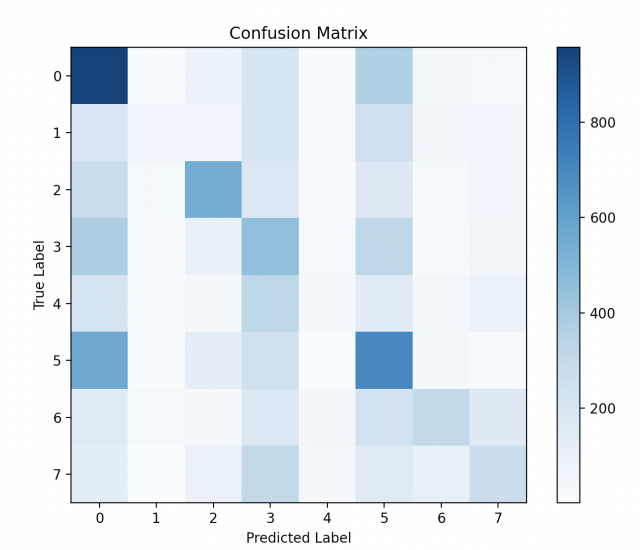

Figure 3: Confusion Matrix for Image Labelling

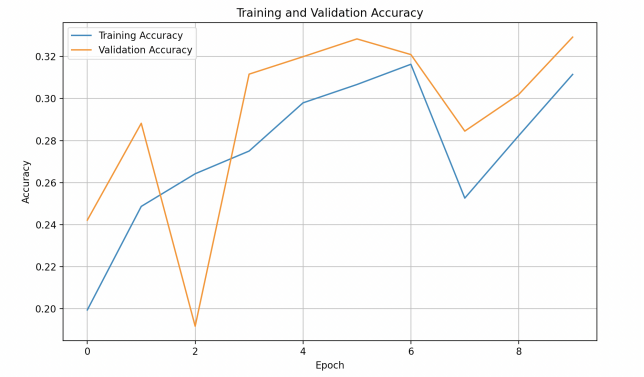

Figure 4: Accuracy graph across epochs for Image Labelling

4.2 Music Generation

After training our RNN model with a batch size of 32 and 75 epochs, we found that our model converged at approximately a validation accuracy of 55%. However, the accuracy of the test dataset, obtained from the classification report, found the accuracy to be about 50%, typically about 5% lower than the training set.

The model was best at predicting the emotion relaxed, as shown in the confusion matrix in Figure 5. Relax consistently has the highest f1-score in comparison to the other three emotions. Both anger and happiness has a lot of true positives and false positives. The model also showed a tendency to predict sadness the least frequently, while it predicted anger the most frequently. The model was the worst at predicting sadness, as it usually has more false positives than true positives for the emotional label. Overall, the confusion matrix is well balanced as it does not exhibit significant biases toward any particular emotional label.

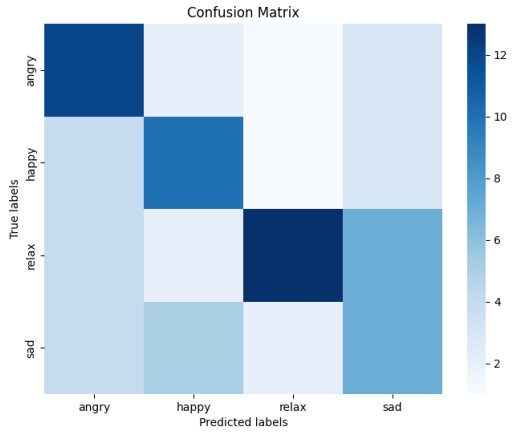

Figure 5: Confusion Matrix for Music Generation

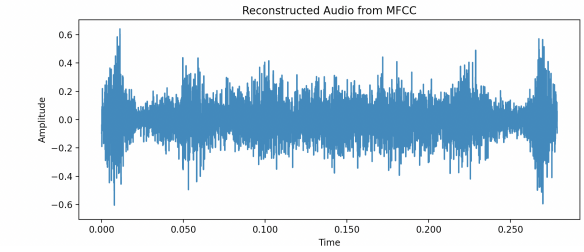

Figure 6: MFCC Spectrogram

Based on the human survey conducted with 18 participants, 11 individuals reported that the generated music matched the emotional label, while the other 7 said it did not. This survey shows that majority of participants perceived the generated music as somewhat conveying the target emotion. However, the degree of which the music effectively conveyed the intended emotion remains unclear, as this was not being tested.

5 DISCUSSION

The observation of an amusement bias in our model's predictions was predictable given the set of features focused on brightness and colorfulness. While it's surprising that it was biased to amusement and not a different positive emotion, it's understandable that many images might feature bright and colorful regions without being a positive emotion. For example, a volcano erupting during a clear day would be more likely to be given a positive label because it's a bright image with a variety of colors from the red lava to blue sky. Moving forward, one method to solve this issue could be involving more features like edge detection and 'blob' (groupings of similar pixels) detection [12].

The results of our RNN model highlight several key findings regarding music generation by emotion. After training the data, we observed a validation accuracy of 55% and it dropped to around 50% for the test dataset. This is consistent with typical trends where test set accuracy is slightly lower than training set accuracy.

When analyzing the confusion matrix, we found that the model showed varying degrees of success in predicting different emotions. For example, the model would predict sadness less frequently compared to the other three emotions, and even when sadness was predicted, it would result is more false positives than true positives. Despite these challenges, the confusion matrix overall exhibited a well-balanced distribution of predictions across the four emotional labels, without showing significant biases toward a particular emotion. This allowed us to generate music that equally represented all emotions without being disproportionately influenced by any emotional bias.

Additionally, our human survey found that a majority of participants (11/18) reported that the generated musical clip conveyed the intended emotion. This shows that we were able to achieve our goal of being able to generate music based on emotion. However, there are a few limitations with this human survey. Firstly, the extent to which the music effectively conveyed the intended emotion remains unclear, as the survey simply asked a yes or no question. So we cannot say for sure, if our model was able to create very specific music based on the emotional label or if it was vague enough to be able to match that label. Another limitation of this survey is the size of participants were very small which can lead to high variance and less precise estimates. Finally, the last and most important limitation of this survey is participants were asked if the emotional label matches the music, instead of letting them tell us which emotion they feel is being conveyed through the music. This can cause biases as participants may be influenced by the provided emotional label and try to match the music to it, rather than expressing their genuine emotional response.

When we consider the levels of accuracy of 55% and 50% for the validation and test dataset of the RNN respectively, the model shows room for approximately doubling its performance metric. However, intricacies lie when we consider the vastly unexplored area of music generation and emotional labeling in unison. Music generation in itself contains multiple subtle features which humans can interpret trivially like dynamics, timbre, and harmony, but our audio datasets were limited to the most fundamental components such as MFCC, tone and harmony. We had hoped to use larger datasets incorporating all of these features as well as more emotional annotations including valence, arousal, and more subjective subtleties into our model. It is also important to consider that the music generation model was adapted to be an input to the output of the affective image labeling model, wherein lies a narrower range of emotions to be analyzed.

6 CONCLUSION

This paper introduces the use of CNN and RNN's for integrating visual emotion analysis and music generation. Our model labels a given image with one of 8 emotions, then produces a short, musical piece to reflect the emotion. Experiments with our model demonstrated that it could reasonably produce music matching an image. However, as a consequence of limitations in the generative tools used and biases in our models, there is still potential improvements to be made to better create music reflective of the emotion in a visual.

REFERENCES

[1] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[2] Jingyuan Yang, Jie Li, Xiumei Wang, Yuxuan Ding, and Xinbo Gao. Stimuli-aware visual emotion analysis. *IEEE Transactions on Image Processing*, 30:7432–7445, 2021.

[3] Miguel Civit, Javier Civit-Masot, Francisco Cuadrado, and Maria J Escalona. A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends. *Expert Systems with Applications*, 209:118190, 2022.

[4] Maximos Kaliakatsos-Papakostas, Andreas Floros, and Michael N Vrahatis. Artificial intelligence methods for music generation: a review and future perspectives. *Nature-Inspired Computation and Swarm Intelligence*, pages 217–245, 2020.

[5] Gaëtan Hadjeres, François Pachet, and Frank Nielsen. Deepbach: a steerable model for bach chorales generation. In *International conference on machine learning*, pages 1362–1371. PMLR, 2017.

[6] Lev Manovich. *AI aesthetics*. Strelka press Moscow, 2018.

[7] Luca Casini, Gustavo Marfia, and Marco Roccetti. Some reflections on the potential and limitations of deep learning for automated music generation. In *2018 IEEE 29th annual international symposium on personal, indoor and mobile radio communications (PIMRC)*, pages 27–31. IEEE, 2018.

[8] Yael Ecker and Yoav Bar-Anan. Conceptual overlap between stimuli increases misattribution of internal experience. *Journal of Experimental Social Psychology*, 83:1–10, 2019.

[9] Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Danny Cohen-Or, and Hui Huang. Emoset: A large-scale visual emotion dataset with rich attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20383–20394, 2023.

[10] M. S. Likitha, Sri Raksha R. Gupta, K. Hasitha, and A. Upendra Raju. Speech based human emotion recognition using mfcc. *International Conference on Wireless Communications, Signal Processing anf Networking (WiSPNET)*, 2017.

[11] Mehmet Bilal Er and Ibrahim Berkan Aydilek. Music emotion recognition by using chroma spectrogram and deep visual features. *International Journal of Computational Intelligence Systems*, 12:1622–1634, 2019.

[12] Stefan Hinz. Fast and subpixel precise blob detection and attribution. In *IEEE International Conference on Image Processing 2005*, volume 3, pages III–457. IEEE, 2005.

7 APPENDIX

A DATASHEETS FOR DATASETS

EmoSet Dataset:

- **For what purpose was the dataset created?** Due to a lack of development in datasets for visual emotion analysis, the creators wanted a dataset that focused on: scale, annotation richness, diversity and data balance. Using this dataset, they wanted to further encourage research and understanding in visual emotion analysis.
- **Who created the dataset and on behalf of which entity?** Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Daniel Cohen-Or, and Hui Huang created this dataset. They were supported by: National Natural Science Foundation of China, DEGP Innovation Team, Guangdong Science and Technology Program, Israel Science Foundation, and Guangdong Laboratory of Artificial Intelligence and Digital Economy.
- **Who funded the creation of the dataset?** This was funded by the Visual Computing Research Center of Shenzhen University in China.

Turkish Music Emotion Dataset:

- **For what purpose was the dataset created?** This dataset was created for the purpose of recognizing human emotion by using chroma spectrogram and deep visual features.
- **Who created the dataset and on behalf of which entity?** This dataset was created by Mehmet Bilal Er and Ibrahim Berkan Aydilek on behalf of the Department of Computer Engineering at Harran University.
- **Who funded the creation of the dataset?** There was no funding for the creation of the dataset.

B CONTRIBUTIONS OF GROUP MEMBERS

(1) Karanvir Bath
   - Found data set for affective image labelling
   - Developed affective image labelling model
   - Contributed to Poster styling and contents

(2) Ritchie Kumar
   - Found data set mapping music to emotions
   - Developed music generation model
   - Contributed to Poster styling and contents

(3) Parmveer Dayal
   - Found data set mapping music to emotions
   - Developed music generation model
   - Contributed to Poster styling and contents