Face-to-Music: Emotion-Based Music Generation

Group members:

- Hansen Wang, whansen1
- Mingjun Ma, mma47
- Zhikai Zhao, zzhao130

Introduction

In this project, we built a system, which takes an image of a face as input, detects the emotional state of the person, and generates music that matches that emotion. We chose this project because music is a powerful medium for emotional expression. We believe that linking music directly to a person's facial emotions creates a more intuitive and immersive experience. This project consists of two main parts: the first part involves facial emotion recognition (multi-class classification) and the second part contains music generation (sequence generation). We think this is a challenging project because we can explore the creative side of generative deep learning and learn how it can explain abstract emotional states.

Methodology

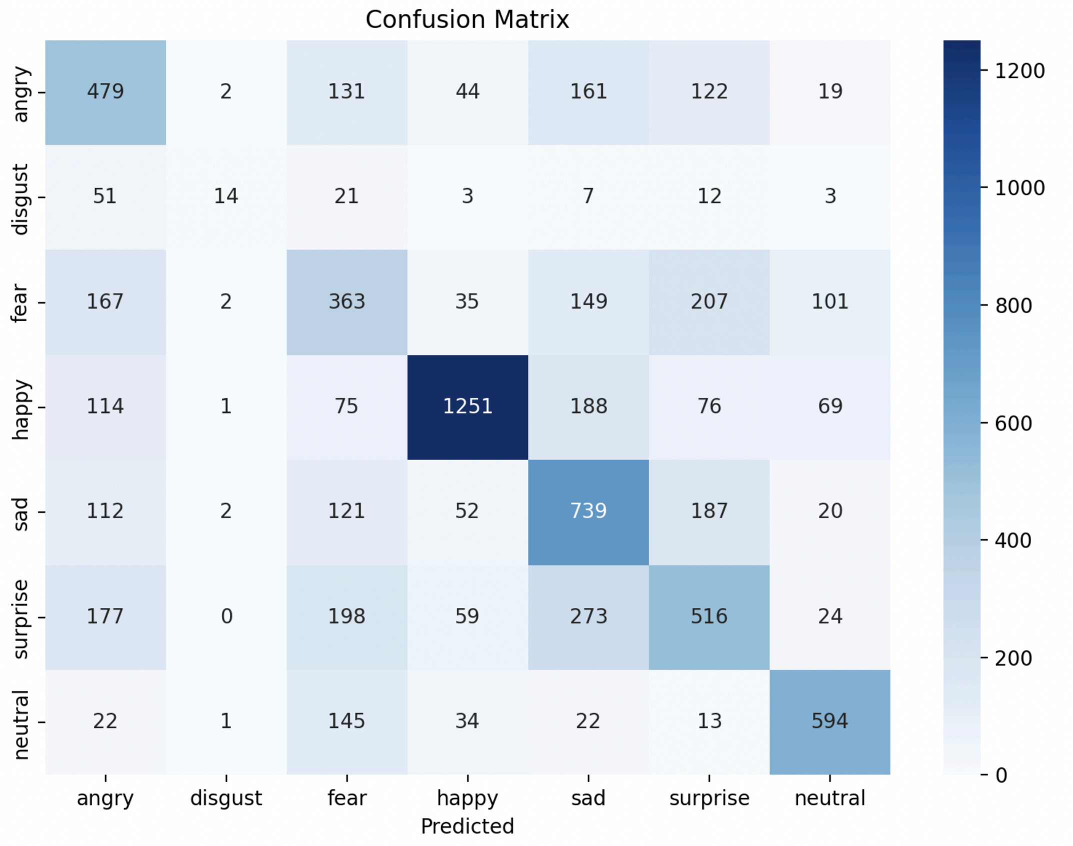
In reference to facial expression recognition, we used the FER-2013 dataset which contains 35,887 images of grayscale faces, sized 48x48 pixels and labeled with 7 emotions (angry, disgust, fear, happy, sad, surprise, neutral). In reference to music, we used the DEAM dataset which consists of 1802 music snippets and full songs with average valence and arousal values on the entire song. We define the corresponding emotion of the music based on the values.

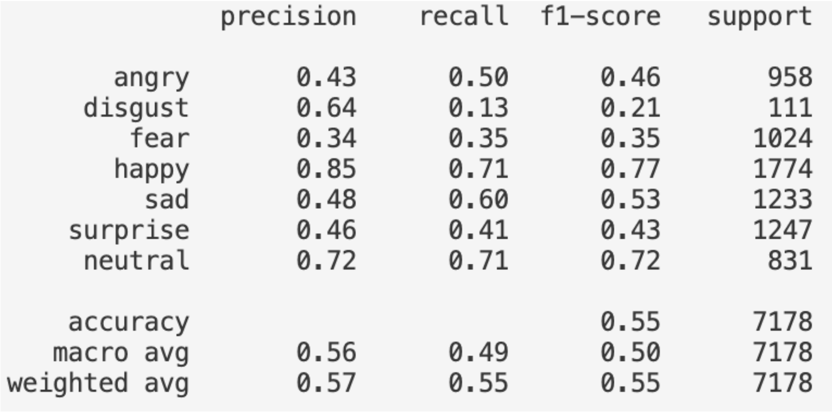
We created a facial emotion recognition system using a custom-trained Convolutional Neural Network (CNN) that we built from scratch in TensorFlow. The input images must be pre-processed by an initial cropping to the size we needed (i.e., 48×48 pixels) and per-pixel normalization into a range of [0, 1]. The CNN consists of a series of convolutional layers which have a ReLU activation function. We also used layers for batch normalization and maximum-pooling layers in between the convolutional layers. We would have a dense classification head at the end of the Convolutional layers with a softmax activation to predict classes for the emotion. The model was trained using Adam optimizer and categorical cross-entropy loss. To avoid overfitting and to maximize accuracy, we used early stopping and checkpointing based on the validation accuracy.

In the music generation stage, we will begin by extracting musical features using the librosa library. We will analyze each track to extract key musical attributes that are relative to emotional expression. Tempo will be estimated using librosa.beat.beat_track, which will provide a min-max estimate where tempo will fall, pitch will be based on mean and standard deviations of chroma features, velocity (loudness) will be inferred using root-mean-squared (RMS) energy since it reflects MIDI velocity, note density will be inferred from onset strength, mode will be determined from correlations with primary templates using chroma features, chord progression will simplified to "I-IV-V-I" for major keys and "i-VI-III-VII" for minor keys. Rhythms are labelled as "irregular" for emotions correlated with tension (e.g., anger, fear, surprise) and "regular" otherwise. For music Generation we will be using the mido library to create a MIDI file. To define the tempo track, a random tempo will be sampled within the estimated minimum and maximum tempo range, the chord track translates the abstract representation of chords into MIDI notes, pitch will be generated by randomly selecting melodic notes for the melodic range defined by the theorized pitches, velocity will be sampled for dynamic expression, and finally, note density will allow for controlling the number of notes per-bar, so that the phrasing aligns with the constructs of emotional tone.

Results

Our facial emotion recognition model achieved an overall accuracy of 55% on the FER2013 test set (baseline accuracy 24.71%). It performs best on the “happy” and “neutral” classes, and worst on the "disgust" and "fear" classes. The confusion matrix reveals frequent misclassifications between similar emotions such as “fear” vs. “surprise” and “sad” vs. “angry”.





```
              precision    recall  f1-score   support

       angry       0.43      0.50      0.46       958
     disgust       0.64      0.13      0.21       111
        fear       0.34      0.35      0.35      1024
       happy       0.85      0.71      0.77      1774
         sad       0.48      0.60      0.53      1233
    surprise       0.46      0.41      0.43      1247
     neutral       0.72      0.71      0.72       831

    accuracy                           0.55      7178
   macro avg       0.56      0.49      0.50      7178
weighted avg       0.57      0.55      0.55      7178
```

For the music generation part, we tested a total of 42 images evenly distributed across 7 emotional categories: angry, disgust, fear, happy, neutral, sad, and surprise. The system demonstrated generally accurate emotion-to-music mappings, but some confusion patterns were observed, for example, the music generated for fear and sadness tended to sound more similar. Overall, the system successfully demonstrated the ability to generate emotion-aligned symbolic music from facial expressions, while also revealing areas for further refinement in distinguishing subtle emotional nuances.

Challenges

The hardest part of this project is music generation. We initially planned to use Transformer-based models for music generation. However, due to limited computational resources and the complexity of training such models on symbolic music data, we faced significant implementation and training challenges.

To address this, we pivoted to a more lightweight and controllable approach: extracting musical features (e.g., tempo, pitch range, chord progression) from real emotion-labeled music, and using a rule-based MIDI generator to compose new music conditioned on these features. This method allowed us to generate emotion-aligned music efficiently without relying on large-scale model training.

Reflection

Overall, I think our project made a strong initial step into the area of emotion-driven music generation. We achieved our target goal of identifying emotions from faces and using them to generate corresponding music, while the stretch goal of generating high-quality, expressive music is still a space for improvement. The facial emotion detection model we created with a custom CNN performed as we anticipated from an architectural and training behaviour standpoint, however, its performance yielded an accuracy of approximately 55%, which revealed a critical limitation: the implementation quality of detecting emotional expressions is highly impactful upon music generation. This bottleneck partially limited the reliability of and emotional congruence with the generated music.

Our methodologies have changed over the weeks. For our project we originally intended to utilize deep learning mechanisms such as LSTMs, Transformers, or VAEs designed for creating music. Nevertheless, we experienced a lot of different implementation challenges as well as issues training these denser deep learning models on symbolic music data primarily due to limited computational resources due to the denser representations and subsequent model complexity. Thus, we leaned more toward deterministic mapping from explicit emotion to established musical rules, or simply, were to extract musical features (e.g., tempo, pitch range, chord progression) from actual emotion-labeled music, and then get a rule-based MIDI generator to compose new music conditioned by these features. If we were to start the project again we would likely prioritize pretrained emotion recognition models to reach a higher accuracy, and aim to collect more music content for model training.

With more time, we could enhance the system by enabling emotional dynamics over time, allowing the music to evolve (e.g., from sadness to joy) in a more natural and expressive manner. We would also invest in a better labeling and evaluation scheme for music quality, given how subjective and fuzzy musical "correctness" can be.

This project taught me many things technically and conceptually. Technically, I understood image preprocessing, CNN architectures, and music feature extraction using librosa and mido in greater depth; conceptually, I gained appreciation for the difficulty in tying together a perception (facial emotions) to a generative creativity (music) especially considering the subjectivity and abstraction. This project elucidated a distinction between engineering models based on defined rules versus expressiveness and subtleties achieved through data-driven, human-design thinking.