The Melodies of an Image: Exploring Music Recommendations Based on an Image’s Content and Context

Adwaita Janardhan Jadhav †, Ishmeet Kaur *†

Abstract
Music recommendation systems are widely used in the industry and have traditionally been supported on user behavior and upcoming trends. Concurrently, advancements in sentiment analysis now allow for complex emotional understanding from both text and images. However, a significant gap exists in integrating these areas for a comprehensive user experience. This paper positions a novel approach to address this gap, proposing an image-to-song recommendation system by utilizing the emotional relation between visuals and music.

1. Introduction
Music recommendation has been a focal point in both research and industry (like entertainment and social media), with systems evolving to deliver personalized playlists based on user behaviors, preferences, and broader trends. As the dynamics of user interaction shift towards visual platforms, there’s an expanding research area in understanding the interplay between images and music. With ML models being more sophisticated at extracting various moods from text and image data, sentiment analysis have demonstrated the potential to extract and interpret various emotions. Yet, the convergence of music recommendation with image sentiment remains relatively less explored, representing a gap between these advancing fields. Addressing this, our paper aims to position a novel approach, merging these domains to introduce an image-to-song recommendation system.

2. The Need for Image-Based Music Recommendation
Recommendation systems often rely on user behavior, like count of the times a song is played, the songs a user reacts to on social media posts or type of liked images. [1]. While music and image recommendations typically follow separate historical trends, there's value in combining them for purposes like social media post backgrounds or movie soundtracks. Existing research mainly focuses on recommending music from image content [2], like objects, or context [1], such as facial expressions[2], without merging the two approaches.

Music recommendations can benefit from considering both an image’s subject and its emotional context. [3] For instance, a picture of a girl celebrating her birthday by cutting a birthday cake might lead to a 'happy' song suggestion, even if it may be a wedding song, based on mood alone or a dance party track due to visibility of balloons. A more effective system would combine these approaches, ensuring song choice match both the image’s content and feeling. Hence, there’s a demand for a music recommendation system that understands both aspects of images.

3. Proposed Approach: Mood Melody Mapper (MMM)
This paper proposes a proof of concept of a novel method called Mood Melody Mapper (MMM) to recommend music that maps seamlessly with the content, context, and sentiment of a given image. We explore a dynamic intersection of visual content and auditory experience to enhance user engagement through personalized music recommendations. Using Natural Language Processing (NLP), this method combines techniques such as image description generation and sentiment analysis of music lyrics to identify the mood. Each subsection details each block of the system architecture in Figure 1, discussing underlying algorithms and implementation specifics.

3.1. Image Text Description Generation
The first block of MMM (Block 1 in Figure 1) consists of generating image descriptions on the given user input image. We propose to use the existing encoder-decoder model pair as the base model. The encoder processes an input image and compresses its information into a context vector, and the decoder then uses this vector to
Figure 1: System level architecture of MMM. Given the example of input image here, Block 1 generates current input image’s text description as, “Birthday party of big diverse happy group”. The process is forked to 2 paths. Block 2 uses cosine similarity between output of block1 and songs lyrics to recommend top songs based of image description. For the other path, block 3 will classify the sentiment of image text description into one of the given emotion category like happy, worry, etc.. This will be utilized by block 4 to map image sentiment class with song sentiment class. Then block 5 will recommend songs based of the mapping which will eventually be utilized by block6 along with the output of block2 to aggregate song recommendation.

3.2. Sentimental Classification of Image Text Description

In Block 3 of Figure 1, the second batch of song recommendations is derived from the emotions conveyed in the image. Here we are determining the emotional tone behind the image description text, distinguishing between positive, negative, or neutral tones, and further categorizing into complex emotions like happiness or worryness.[13, 14, 15] A machine learning model, pretrained on a labeled sentiment dataset assigns, sentiment labels to the image descriptions.

Sentiment analysis can be done using models like LSTM[6](for longer text sequences, CNN[16](capture...
local patterns), BERT(capture context in both directions)[12], or hybrid approach. For example, Datasets like Smile twitter dataset[17] can be used to train such model.

3.4. Mapping Sentiments

Sentiment analysis is now conducted on song lyrics, though a pre-categorized music dataset can also be employed. In Block 4 of Figure 1, an emotion mapping between the image text and music text is executed, as their categorizations may differ, as illustrated in Figure 2. Based on this mapping, songs are recommended to align with the desired emotion. In complex mapping scenario, techniques like transfer learning[18] or Canonical Correlation Analysis (CCA)[19] can be used.

3.5. Song Recommendation Aggregation

The songs recommendation from Block 2 and Block 5 are then aggregated or combined to produce the final recommendation song list in Block 6 of Figure 1. The aggregation can be simple like doing a union or weighted union of the two lists. A deduplication filter and user feedback loop can also be used for combining the list. Metadata Integration can be used to refine rankings based on additional data about each song (like user ratings, play count, etc.) available. You can use this metadata to refine your rankings. For instance, songs with higher user ratings might get a boost in the combined list.

4. Evaluation Method

A systematic method is needed to evaluate and select the models used in the different components highlighted in Figure 1. This section provides some details about the proposed evaluation methods:

1. For evaluation on image caption task, BLEU (Bilingual Evaluation Understudy)[20] can be used in combination with CIDEr[21] and ROUGE[22]. BLEU is an algorithm, which has been used for evaluating the quality of machine-translated text. CIDER and ROUGE are metrics for consensus-based image description evaluation. The classification task is assessed on accuracy for five sentiment classes.

2. The classification task can be evaluated on accuracy when classifying into 5 sentiment classes. We can use the weighted F1 metric which is a standard for multi-class classification. Per class F1 score: $2 \cdot \frac{P \cdot R}{P + R}$ (P: Precision and R:Recall)

5. Applications

There are various applications of the Image-Based Music Recommendation System:

1. Social Media Platforms: Adding background music to various social media posts using image-based music recommendations can increase user engagement since the combination of both image and music leads can attract the attention of users [23].

2. Soundtrack generation: Various apps that create digital photo albums, a suitable digital soundtrack could be added based on the images in the album. For example, an album consisting of pictures from childhood to old age can have a digital transitioning soundtrack rather than just one boring related memories soundtrack [3].

3. Event Playlists: Users can upload event images, like parties or road trips, and the system can curate event’s theme specific playlists. [3].

4. Interior Decor Music: Images of home interiors and decor preferences can result in music recommendations that align with the user’s design aesthetic [24].

6. Conclusion and Future Vision

This paper positions a image’s content and context based music recommendation system. We outline the system’s architecture, implementation details, and evaluation criteria. Our findings help assess cutting-edge components and their interplay in developing this sophisticated recommendation model.

Beyond the scope of this paper, there are still many challenges to be overcome. The current system design recommendations by considering the entire music dataset. Thus, some music recommendations might be completely irrelevant to the user even though they are relevant to the image. Another implication on user is, if a particular image is of sad sentiment, then recommending music which is sad might leave user with a more negative feeling which is not good. Context and content-based systems can be enhanced with user-specific features for better recommendation. Another open challenge is the support for music that does not contain lyrics, where only beats are available for analysis. Here, newer transformer-based models may be required to understand the sentiment, based on just the musical notes. And finally, context-based music recommendations have a problem of being subjective, hence it is important to learn how the user associates contexts with different emotions. What one person finds joyful, another might find melancholic, making a one-size-fits-all recommendation challenging.
References


