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 being 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

Fifth Knowledge-aware and Conversational Recommender Systems (KaRS) Workshop @ RecSys 2023, September 18–22 2023, Singapore.

[*] Corresponding author.

[†] These authors contributed equally.

adwaitas28@gmail.com (A. J. J.); ishmeet3kk@gmail.com (I. K.)

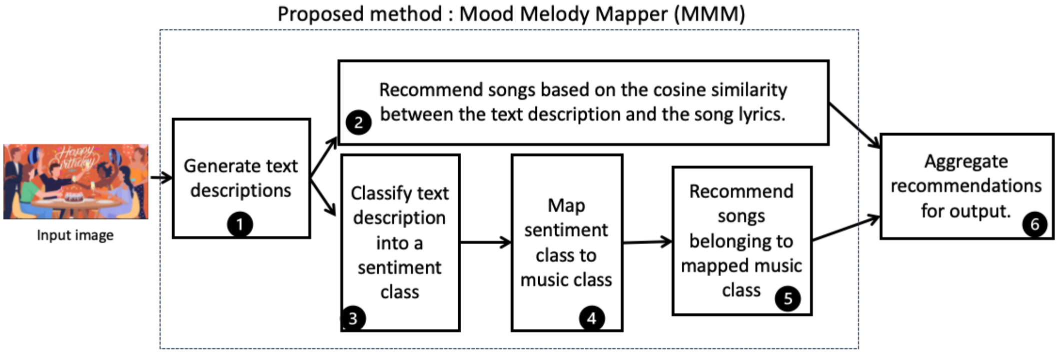
Proposed method : Mood Melody Mapper (MMM)
Input image
Generate text descriptions
1
Recommend songs based on the cosine similarity between the text description and the song lyrics.
2
Classify text description into a sentiment class
3
Map sentiment class to music class
4
Recommend songs belonging to mapped music class
5
Aggregate recommendations for output.
6

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.

produce an output sequence which is the text description of the image.

Here encoders such as CNN (VGG16 or ResNet152)[4, 5] and decoders (LSTM, transformer, GRUs)[6, 7] are used. For example,the ResNet152 encoder, pretrained on ImageNet, is modified by removing its softmax layer, producing fixed-length vector embeddings from images which is then used by decoder to produce text description. Training such encoder-decoder model leverages datasetlike the VizWiz-Caption dataset[8], which offers image-caption pairs. Given that images contain multiple objects and features, an attention mechanism like soft attention[9] is incorporated into the decoder to ensure nuanced image details are considered during caption generation.

3.2. Song Recommendation based on Cosine Similarity

The initial set of song recommendations is derived using cosine similarity[10], comparing the text description generated in Block 1 to the lyrics of each song. In Block 2 of Figure 1, we represent both the text description and song lyrics as vectors and calculate the cosine similarity between them. The system then recommends the top songs—let's say, the top 10—that exhibit the highest cosine similarity values.

To compute cosine similarity here, both the song lyrics and image text description goes through preprocessing(tokenization, removing stop words/punctuation) followed by conversion to a vector using methods like Doc2Vec[11] or BERT[12].

3.3. 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

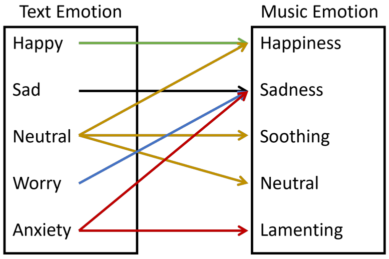

Figure 2: This example illustrates the mapping of emotion categories derived from image text descriptions with those from music's lyrics, assuming five distinct emotion categories in each domain. This alignment serves as a crucial step in bridging the sentiment analysis gap between image and song based emotions, central to our research focus. For instance, an image text description evoking feelings of anxiety might lead to song recommendations with sentiments of sadness or lamentation.

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 ROGUE 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: $\frac{2\times P\times R}{P+R}$($P$: 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 a 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

[1] C. Hsia, K. Lai, Y. Chen, C. Wang, M. Tsai, Representation learning for image-based music recommendation, CoRR abs/1808.09198 (2018). URL: http://arxiv.org/abs/1808.09198. arXiv:1808.09198.

[2] O. Ghosh, R. Sonkusare, S. Kulkarni, S. Laddha, Music recommendation system based on emotion detection using image processing and deep networks, in: 2022 2nd International Conference on Intelligent Technologies (CONIT), 2022, pp. 1–5. doi:10.1109/CONIT55038.2022.9847888.

[3] Y. Yu, Z. Shen, R. Zimmermann, Automatic music soundtrack generation for outdoor videos from contextual sensor information, in: Proceedings of the 20th ACM International Conference on Multimedia, MM '12, Association for Computing Machinery, New York, NY, USA, 2012, p. 1377–1378. URL: https://doi.org/10.1145/2393347.2396493. doi:10.1145/2393347.2396493.

[4] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2015. arXiv:1409.1556.

[5] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015. arXiv:1512.03385.

[6] R. C. Staudemeyer, E. R. Morris, Understanding lstm – a tutorial into long short-term memory recurrent neural networks, 2019. arXiv:1909.09586.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2023. arXiv:1706.03762.

[8] D. Gurari, Y. Zhao, M. Zhang, N. Bhattacharya, Captioning images taken by people who are blind, 2020. arXiv:2002.08565.

[9] M. Kulkarni, A. Abubakar, Soft attention convolutional neural networks for rare event detection in sequences, 2020. arXiv:2011.02338.

[10] R. H. Singh, S. Maurya, T. Tripathi, T. Narula, G. Srivastav, Movie recommendation system using cosine similarity and knn, International Journal of Engineering and Advanced Technology 9 (2020) 556–559.

[11] Q. V. Le, T. Mikolov, Distributed representations of sentences and documents, 2014. arXiv:1405.4053.

[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[13] A. G. M. Meque, N. Hussain, G. Sidorov, A. Gelbukh, Guilt detection in text: A step towards understanding complex emotions, 2023. arXiv:2303.03510.

[14] P. Nandwani, R. Verma, A review on sentiment analysis and emotion detection from text, Social Network Analysis and Mining 11 (2021) 81.

[15] Z. Wang, C. S. Chong, L. Lan, Y. Yang, S. Beng Ho, J. C. Tong, Fine-grained sentiment analysis of social media with emotion sensing, in: 2016 Future Technologies Conference (FTC), 2016, pp. 1361–1364. doi:10.1109/FTC.2016.7821783.

[16] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, et al., Recent advances in convolutional neural networks, Pattern recognition 77 (2018) 354–377.

[17] B. Wang, A. Tsakalidis, M. Liakata, A. Zubiaga, R. Procter, E. Jensen, SMILE Twitter Emotion dataset (2016). URL: https://figshare.com/articles/dataset/smile_annotations_final_csv/3187909. doi:10.6084/m9.figshare.3187909.v2.

[18] S. Ruder, M. E. Peters, S. Swayamdipta, T. Wolf, Transfer learning in natural language processing, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials, 2019, pp. 15–18.

[19] S. Faridani, Using canonical correlation analysis for generalized sentiment analysis, product recommendation and search, in: Proceedings of the fifth ACM conference on Recommender systems, 2011, pp. 355–358.

[20] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[21] R. Vedantam, C. Lawrence Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575.

[22] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

[23] J.-H. Su, W.-Y. Chang, V. S. Tseng, Personalized music recommendation by mining social media tags, Procedia Computer Science 22 (2013) 303–312. URL: https://www.sciencedirect.com/science/article/pii/S1877050913009009. doi:https://doi.org/10.1016/j.procs.2013.09.107, 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems - KES2013.

[24] T. Xia, Y. Sun, Y. An, L. Li, The influence of music environment on conceptual design creativity, Frontiers in Psychology 14 (2023). URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1052257. doi:10.3389/fpsyg.2023.1052257.