COSE474-2024F: Final Project Proposal
Image-based Playlist Recommendation using DAPT

Jonghyeon An

1. Introduction

In today's digital world, personalized recommendations are crucial to enhancing user experience. Music and video streaming platforms rely heavily on recommendation systems based on user preferences or historical data. However, I want new approach to recommend contents based on a user's emotional response based an image.

2. Problem definition & chanllenges

The system aims to extract emotions from user-uploaded images using CLIP and match these emotions with suitable music or video genres. By leveraging DAPT, it will be employed to improve the accuracy of emotion matching for specific domains, playlist. The final goal of this project is extract specific emotions correctly and recommend playlist precisely.
One of the key challenges is emotion extraction accuracy. Since interpreting emotions from images can be highly subjective, tuning the performance of CLIP or utilizing additional data may be necessary to improve accuracy. Another challenge lies in genre matching. Selecting the right music or video genre based on the extracted emotion is a complex task, requiring the model to be finely tuned to better understand subtle emotional differences.

3. Related Works

This project utilizes CLIP to extract emotions from images uploaded by users. Previous studies have demonstrated CLIP's strength in linking images with text, and it can also be applied to emotion-based recommendation systems. By analyzing images, the system identifies the emotional atmosphere and recommends content, such as music or videos, that aligns with the extracted emotion. Also, DAPT can be employed for domain-specific emotion-to-music matching through additional fine-tuning. In previous research, DAPT has proven useful in refining large pre-trained models for specific domains, enhancing the accuracy of emotion recognition by adjusting the model to better fit domain-specific nuances.

4. Datasets

For emotion extraction from images, the Open Images datasets are needed. These will train the model to recognize emotions or atmospheres from various images. For matching music to these emotions, the Dataset provides emotional metadata and music details to align with the extracted feelings are also needed. Optionally, a custom dataset of user-uploaded images and curated music can be created for further refinement similar with few-shot learning. Not too much datasets will not be nessesary.

5. State-of-the-art methods and baselines

In comparison with state-of-the-art (SOTA) methods, this project leverages CLIP to achieve a more seamless and accurate connection between visual stimuli and emotional responses, outperforming traditional models that focus separately on either image recognition or text sentiment analysis. While most conventional recommendation systems rely on user history, preferences, or facial recognition, CLIP enables a more holistic approach by directly extracting emotional content from images, regardless of the context.Furthermore, DAPT enhances CLIP's capacity to handle specific domains where more refined emotional nuances are necessary. In contrast, baseline models typically do not possess the adaptability or precision required to operate in such domain-specific settings.

6. Schedule

First, gathering the necessary datasets and set up the project environment is needed. Then, fine-tune the CLIP model for emotion extraction and develop the algorithm for matching emotions to music. Next, integrate and optimize the image-based emotion extraction with the music matching system. After that, conduct user testing to evaluate system performance, create a custom dataset based on feedback, and finally, complete the final optimization and prepare the system for deployment.