Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 218 (2023) 383–392

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Machine Learning and Data Engineering

Music recommendation based on affective image content analysis

Rushabh Chheda[a], Dhruv Bohara[a], Rishikesh Shetty[a], Siddharth Trivedi[a,*], Ruhina Karani[a]

[a]*Department of Computer Engineering, Dwarkadas J. Sanghvi College of Engineering, University of Mumbai, Mumbai, 400056, Maharashtra, India.*

Abstract

Music has the ability to invest even the tritest scenes with so much meaning when added to them. Human perceptions of music and image can be closely related to each other, as both can incite similar sensations and emotions. Advertising agencies often make use of audio and music over their visuals to engage more audiences and to convey the emotions associated with their content more effectively. Matching visuals and music to comparable feelings might help people perceive emotions more vividly and strongly. This paper proposes an effective cross-modal neural network that provides music recommendations to a user by generating matches between images and music over a common emotional vector space. Using the valence and arousal values, a combined image-music pair dataset has been created. The images incorporated in this dataset are leveraged from the OASIS dataset while the music part is queried using Spotify API and YouTube. A Transfer Learning approach is proposed with Convolution Neural Network architecture for training on this dataset using MobileNetV3, ResNet-18 and EfficientNetB4 for the images and SampleCNN for the raw audio clips. For any given image input, a list of top-n music recommendations shall be outputted. This concept thus aims to generate music and image matching based on various deep hidden features over the emotion space of the two modalities.

Keywords: Emotion Recognition; Valence-Arousal Space; Cross-Modal Matching; Deep Learning;

1. Introduction

The goal of music recommendations is to broaden a listener's musical horizons beyond what they are currently familiar with and enjoy. When listeners have exhausted all of their song/artist search possibilities, it presents them with more navigational alternatives [22]. This concept is expanded upon by image-based music recommendation, which suggests songs to the user based on the image provided [2]. People face a variety of scenarios in their everyday lives when they can listen to music while doing something else, such as commuting, eating, exercising, or socializing.

* Corresponding author. Tel.: +91-981-928-8033 ;
E-mail address: siddharthtrivedi19@gmail.com

1877-0509
10.1016/j.procs.2023.01.021

Image-based music recommendation tries to propose appropriate music from the available corpus when an image capturing any of the above-mentioned scenarios is entered into the system.

Different forms of music elicit different feelings in people, which is one of the main reasons why people listen to music so much. There have been several research studies [12, 16] that show that music may elicit or trigger emotions in listeners, as well as transmit or express feelings to listeners. As a result, research into how computers and technology may be used to interpret the feelings that music can transmit is gaining popularity. Music Emotion Retrieval [7] is the name given to this area of research, which is a branch of Music Information Retrieval (MIR). The interdisciplinary study of retrieving information from music is known as music information retrieval (MIR). Various applications and use cases of MIR require combinations of various backgrounds such as musicology, psychoacoustics, academic music study, signal processing, informatics, machine learning, optical music recognition, and artificial intelligence.

Analogous to music, images also play a major role in expressing and affecting people's emotions. Visual Sentiment Analysis is a field of research that aims to understand how various images affect people in terms of emotions [18]. Diverse approaches addressing various data sources and challenges have been developed during the previous few years. These techniques include analyzing textual contents of the image or visual contents for sentiment analysis. Furthermore, great progress has been made in recognizing emotions based on a person's facial features [5, 23] and responding by providing a specialized, tailored service to that user. However, this research goes beyond identifying face emotions to extract characteristics for other types of images, such as landscape and situational images, utilizing deep learning techniques to extract visual elements.

The task of Emotion Recognition in images and music has various approaches and variegated solutions are debated. As both modalities have to be matched based on a set of factors, it should be possible to represent them over shared space. The shared space can be partitioned into categories or dimensional systems for emotion-based matching. A variety of emotional categories (adjectives) are employed to categorize music snippets in the categorical method. While, in the dimensional approach, emotion is described using dimensional space where the dimensions are represented by valence, arousal, and dominance. [9].

Recommending Music based on Image can be considered a Cross-modal task, and therefore requires cross-modal matching and retrieval. The capacity to recognize entities presented in two separate sensory modalities is referred to as cross-modal matching. Cross-modal retrieval [30] is a study topic in the fields of multimedia, information retrieval, computer vision, and database that tries to retrieve data in one modality using a query in another mode. The cross-modal retrieval task's fundamental problem is to learn joint embeddings from a common subspace to compute similarity across multiple modalities. The proposed model aims to generate recommendations in one modality (music) based on the input of different modalities (image).

The main contribution of this paper is as follows:

- A novel dataset is created for image-music emotion mapping by making use of Spotify API for the metadata of 500 popular English songs and YouTube for their audio clips. Images from the OASIS dataset are mapped with the songs depicting similar emotional contents by minimizing the Euclidean distance between image and music's respective valence and arousal values to generate image-music pairs.
- Multiple Cross-Modal Deep Neural Networks were trained on the above-mentioned dataset using transfer learning to find similarity scores between the Image modality and the Music modality. This was further used as the basis for a novel music recommendation system based on an image's emotional content.

The organization of this paper starts with the discussion of the past related work published in the fields of image and music emotion recognition as well as cross-modal retrieval done in section 2. The datasets leveraged for the image and music side along with the preprocessing and formation of the image-music pairs based on Valence-Arousal values are discussed in section 3. The methodology proposed for building the model for image-based music recommendations is elaborated in section 4. Section 5 presents the evaluation and results of the experimentation and discusses its interpretation. Finally, the paper concludes in section 6 and discusses some future scope of the research topic.

2. Related Work

Emotion recognition can be done in two ways: classification in categorical/discrete groups or displaying it in a dimensional space. Ekman [8] states that emotion can be classified into six basic emotions that are anger, disgust, fear, happiness, sadness, and surprise. Cross-modal matching needs a dimensional approach to find an accurate similarity between the two modals. Schlosberg [21] named three dimensions of emotion: ”pleasantness–unpleasantness”, ”attention–rejection” and ”level of activation”. Dimensional models of emotion try to define where human emotions belong in two or three dimensions in order to conceptualize them.

Image emotion recognition is a field that works on finding out what sort of emotions a particular image evokes in various people. Hanna's paper [4] is one that works on classifying an image through Bayesian Model Averaging, using both Early and Late fusion. The method uses text as well as deep-visual features from a pre-trained AlexNet and classifies Images into emotion categories. It shows the power of Model Averaging in Cross-Modal tasks. It considers both visual and textual features of the image when classifying. It works on a Large Scale Dataset for Emotion Classification [26] and gives out 80% accuracy using late BMA on deep features. The International Affective Picture System (IAPS) [3] is a database of photos that have been validated as consistently eliciting a specific emotional response in viewers. A valence scale (ranging from pleasant to unpleasant), an arousal scale (ranging from calm to stimulated), and a dominance/control scale (ranging from ”in control” to ”dominated”) are used to create IAPS. Open Affective Standardized Image Set (OASIS) [14] is also a database of photos categorized on the Valence-Arousal scale. Unlike IAPS it is open access and is also not restricted by copyright issues. Fig. 1 displays a few images from the OASIS dataset along with the Valence and Arousal values for the images as provided by the dataset.

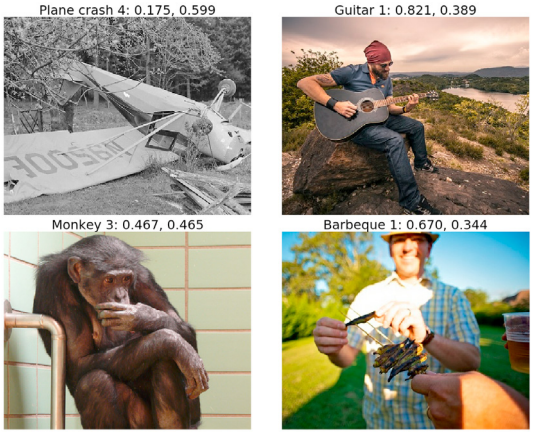

Fig. 1. Sample images from the OASIS dataset with their respective Valence and Arousal values.

Music emotion recognition (MER) can be said to be a subdomain of Music Information retrieval (MIR). There are a lot of features to be considered when talking about music. Audio Feature Engineering has been successfully used together with a Gaussian-RBF Kernel SVM to classify music into the 4 emotion quadrants with an accuracy of 76.4% when 29 novel and 71 baseline features are considered [19]. This work showcases how domain knowledge can be used for creating better emotion classifiers. Such a technique can be extended to be used within a Cross-Modal classification system. Hizlisoy's paper [11] is another paper that represents emotion in two-dimensional space having valence and arousal as its axis. It works on a Turkish music dataset created from Free Music Archive [6]. Convolutional Neural Network (CNN) is used for feature extraction from log-Mel filterbank energies and Mel-frequency cepstral coefficients (MFCCs). Long short-term memory (LSTM) and Deep Neural Network (DNN) are used for classification. The classification performance of the model is 87.09 for the first set of features without correlation-based feature selection (CFS) and 93.54 with CFS

Humans use images, songs, videos, and text together on social media to share their emotions. Emotion-based matching between Images and Songs has been performed before by matching Sentiment Polarity between Image and Music [24]. FC layers are used together with music features and subnetworks for prediction. The result is improved in [29] using Continuous Emotions Matching. It has been demonstrated to yield a much better result. CNN architecture

is used together with a combination of single-modal and cross-modal losses. Both papers do not account for lyrical information and have not used a Cluster CCA with TNN based approach. The papers are also focused on matching Songs and Images based on Emotions and have not been demonstrated to be used in Cross-Modal Retrieval and Recommendation tasks.

Cross-modal retrieval is a task that focuses on retrieving data across different modalities by bridging the modality gap between the two modes. This is often achieved by representing both modalities in a shared subspace where distance functions like cosine or Euclidean can be applied. The past works in cross-modal retrieval have focused and had big successes on image-text [25], audio-text [27], and audio-video cross-modal tasks [28]. Such a task is yet to be applied to the music-image modality pairs. A recommendation system based on images and music would require cross-modal matching and retrieval. The paper by Pang [20] researches various modalities and makes use of the Deep Boltzmann machine and restricts Boltzmann to represent them together in a joint representation layer. Its network architecture is made up of several routes for visual, textual, and auditory modalities, each of which is made up of numerous RBM. The overall prediction accuracy for E-MDBM-VA (the joint representation using both visual and auditory modalities) is 0.4. The paper by Zhao [29] does cross-modal emotion matching using IMEMNET dataset created from random sampling and ground truth Valence-Arousal labels by using NAPS [17], IAPS [3], EMOTIC [13] for Image Data and DEAM [1] for Music Data. It uses Cross-Modal Feature-ratio loss, Cross-Modal Feature-margin loss, and Single-Modal Feature-ratio loss to learn shared latent embedding space. To predict the concrete Valence-Arousal values and similarity between music and image, the model further uses Cross-Modal similarity MSE loss and Single Modal Valence-Arousal loss. Thus, the total CDCML loss is a sum of all the above losses. It achieves better Similarity, Image Emotion, and Music Emotion scores than the previous state-of-the-art held by ACP-Net. The paper by Verma, Dhekane, and Guha [24] proposes a deep neural network architecture that learns to compare the emotional content present in the two modalities without explicitly requiring emotion labels. To accomplish so, Image-Music Affective Correspondence (IMAC), a vast database with over 3,500 music samples and 85,000 pictures divided into three emotion classes (positive, neutral, negative), was constructed. For the effective correspondence prediction challenge, this method obtains a 65.7 percent accuracy.

3. The Imuse Dataset

In this section, the IMUSE dataset is introduced, created by continuous emotion-based music-image mapping. It contains a multitude of image-music pairs which are mapped using the OASIS Image dataset and a Spotify queried dataset for music, from the Spotify API (`https://api.spotify.com/v1`). This section expands on image and music data selection, followed by image-music mapping.

3.1. Image and Music Data Selection

The images used for training and testing of the models were leveraged from the OASIS dataset [14]. This dataset provides 900 images divided into 4 categories with 200 Object images, 346 Person images, 134 Animal images, and 220 Scene images spread across Valence-Arousal space. The images in this dataset have been gathered from Pixabay and Wikipedia. 822 participants from MTurk were employed to rate the images for Valence and Arousal values over a 7-point Likert scale. Thus, the mean Valence-Arousal values for the images range from 1-7. OASIS dataset is a suitable replacement for other popular datasets which are not open source or easily available online.

A list of 500 varied songs with English lyrics was strategically chosen to cover the entire emotional spectrum. Features of these 500 popular songs were queried using Spotify API. The Valence-Arousal values obtained from these features ranged from 0-1. Additionally, the audio clips for these songs were fetched using Pytube. To extract the most meaningful parts from the audio, Pychrous was used, and the timestamp of the first chorus of the songs was detected. From here, the next 30 seconds of the songs were clipped out using Pydub.

Fig. 2 displays the distribution of the 900 images from the OASIS dataset and 500 songs accumulated from Spotify for building the Imuse Dataset on the Valence-Arousal Space.

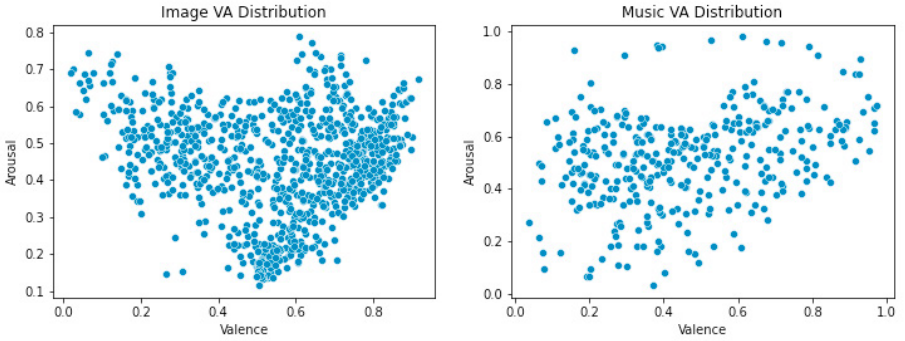

Fig. 2. Distribution of the Image and Music Data in the Valence-Arousal Space

3.2. Image-Music Pairs Formation

For continuous emotion-based image-music mapping Valence and Arousal values are used. Since image and music data is labeled in different scales, the Valence-Arousal values are normalized into [0,1]. The Euclidean distance between their Valence-Arousal ground truth labels is calculated for every image-music pair to find the degree of similarity. The degree of similarity represents the emotion matching label for the corresponding image-music pair. Eq. 1 below is the euclidean formula used for computing the similarity between the image I and music M where V represents the Valence value and A represents the Arousal value.

$$S(I, M) = 1 - \sqrt{(I_V - M_V)^2 + (I_A - M_A)^2} \tag{1}$$

For every image, best five, worst five, and random five image-music pairs are chosen and added to the IMUSE dataset. This sampling lets us take into account the closest and furthest pairs of the dataset while also maintaining a good representation of other pairs due to the random sampling performed. This type of sampling is done to avoid making a huge dataset that would slow down training as it would contain every unique image-music pair. This results in a fast and efficient training of the model as it is fed with the best and worst image-music pairs, to take into account maximum possible scenarios. The worst five samplings result in a distribution where the furthest values are not skipped which makes them look like outliers even when they are not.

4. Methodology

4.1. Transfer Learning

Transfer learning is a well-known approach for transferring models trained on large datasets to be used by smaller datasets. It reduces the time taken to build a model from the ground up and makes use of patterns learned before. For CNN this is achieved by removing the final fully-connected layers and instead repurposing them for your task. Pretrained models such as ResNet-18, EfficentNetB4, and MobileNetV3 have been used for training the images and SampleCNN [15] is used for extracting music features.

4.2. Proposed Architecture

As illustrated in Fig. 3, the model consists of two subnetworks joined by Fully Connected layers at the top to generate Top-N recommendations. The Image subnetwork consists of pre-processing of images obtained from the

OASIS database followed by a deep-CNN model to generate Intermediate Image Vectors. Pretrained models such as EfficientNet, MobileNet, and ResNet-18 are used for this task.

The music subnetwork makes use of Spotify API to fetch the songs for the music Database. These fetched songs are pre-processed, and their Valence-Arousal values are normalized before storing. Random sampling is done to obtain a digitized version of raw audio signals. Sample CNN architecture takes the raw audio signal as input and produces a part of the Intermediate Music Vector. During Training, the immediate music and image vectors are concatenated. These concatenated vectors are used as a basis for training the FC layers using the Huber loss function.

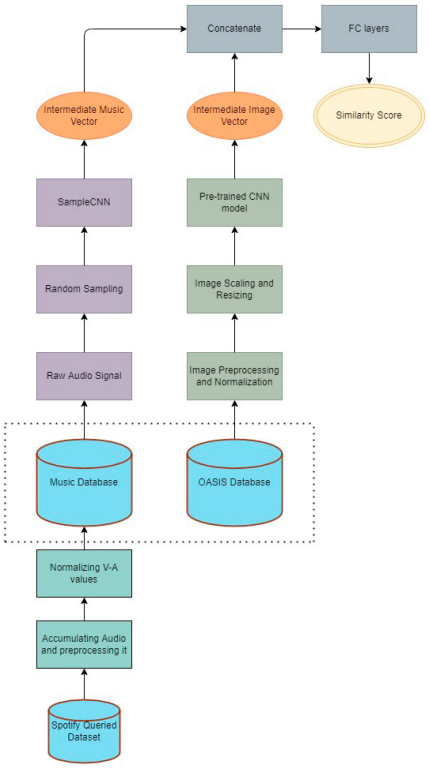

Fig. 3. Architecture diagram for the proposed methodology

During Deployment, the Intermediate Music Vectors for all the Songs in the database are pre-calculated and stored in a database. More songs can be added by maintainers using the music subnetwork to the intermediate music vector database. The input image from the user is passed through the image subnetwork and compared using the clusters and FC layers against the music vectors to find the similarity score between music and Image.

The beauty of this architecture is that it is symmetric between Images and Songs. While here the architecture is used to recommend the best Songs based on an Image, The architecture can be easily inverted to be used to recommend the best Image corresponding to a Song by accepting the Song as input and maintaining a database of Images instead.

4.3. Model Description

Three transfer learning models, namely ResNet-18, EfficientNetB4, and MobileNetV3, pre-trained on the ImageNet dataset, were employed in the Image side of the model for transfer learning. Their weights were frozen, to not train or fine-tune them further to keep computational time low, and the head layer was removed. The output at the end of the remaining CNN architecture was a linear vector of size 512. This vector was passed through a Fully Connected layer with ReLU activation, to generate a one-dimensional vector of size 256. The standard method was followed for all three pre-trained CNNs in use and was used on the Image side to obtain a 256-sized vector at the end of the Image arm.

For the Music side of the model, SampleCNN pre-trained on audio files from the MagnaTagATune dataset was used. The parameters used for SampleCNN were trained for 10000 epochs with a batch size of 96. Like the Image side, the fine-tuned head for SampleCNN was removed and replaced with a Fully Connected layer with ReLU activation, input dimensions of 512, and output dimensions of 128. The vectors obtained from both arms of the model were concatenated to form a single 384-sized vector. This vector was passed through a ReLU activated fully connected layer with output dimensions of 128. The head used were two separate Fully Connected Sigmoid layers to generate a total of 4 outputs corresponding to Valence-Arousal values of Image and Music respectively. Table 1 displays the Training and Validation times taken for the various model combinations employed for the task. It is observed that MobileNetV3 + SampleCNN model takes the least time for training on the dataset.

Table 1. Training and Validation time trends

Model	Training time per epoch (in seconds)	Validation time per epoch (in seconds)	Total training time (in minutes)
ResNet-18 + SampleCNN	65	28	76
MobileNetV3 + SampleCNN	61	24	70
EfficientNetB4 + SampleCNN	101	126	191

4.4. Working of the Model

The dataset was shuffled and stratified based on Song Code, i.e, a unique ID assigned to each song. It was then divided into 75-25% train-test splits. The stratification ensured that both splits were a good representation of the dataset. A seed value was used to ensure a deterministic and reproducible outcome. The model was implemented using PyTorch. For training, the music was normalized to 22050 Hz and the two channels of the audio were combined into one by the mean method, and the audio was Randomly Cropped into a 59049 size vector which corresponds to about 2.5 seconds of audio. The images were randomly cropped to 224x224 size, followed by a random flip with a probability of 0.5. A batch size of 32 with shuffle was used. The Loss Criterion used for Training was Huber Loss which is formulated as shown in Eq. 2. Huber loss handles outliers better than MSE by significantly reducing the weight given to outliers while training. The model was then optimized using Adam with a learning rate of 0.01 and a weight decay of 1e-4.

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for } |y - f(x)| \le \delta \\ \delta|y - f(x)| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases} \tag{2}$$

All of the models were trained using Python 3.7.12 with Ubuntu 18.04.5 LTS as the underlying Operating System on Google Colab. The data loading was done using an Intel Xeon CPU and 13 GB of Ram. Each model was trained for 50 epochs on a Tesla K80 GPU. For each epoch, the model was set to train mode and trained on the training split. Following that the model was validated on the testing split and the losses from both were recorded.

5. Results

5.1. Evaluation Metrics

Three metrics have been primarily employed to measure the recommendation results produced by the models employed to generate the Image-Music similarity. Root Mean Square Error (RMSE), formulated as $\sqrt{\frac{1}{|\hat{R}|}\sum_{\hat{r}_{im}\in\hat{R}}(r_{im}-\hat{r}_{im})^2}$ and Mean Absolute Error (MAE), formulated as MAE $= \frac{1}{|\hat{R}|}\sum_{\hat{r}_{im}\in\hat{R}}|r_{im}-\hat{r}_{im}|$, are effective in measuring the model's performance in computing the Image-Music emotional similarity. Here, r_{im} is the ground truth similarity between the image and song pair and $\hat{r}_{im}$ is the predicted similarity value for the same.

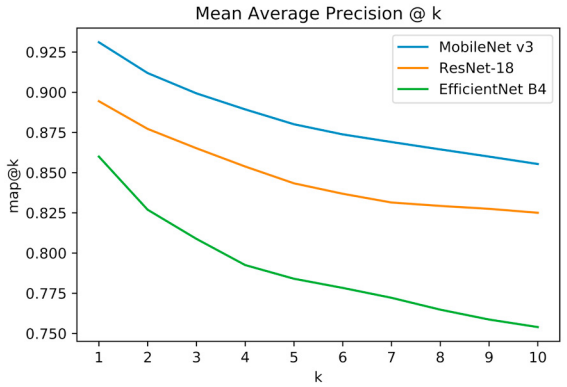

Fig. 4. MAP@K value graph for the three models with x-axis as k for top-k recommendations.

For computing the recommendation strength of the proposed models, MAP@K is employed. MAP@K is a typical evaluation metric used in a recommendation system where an ordered list of recommendations is provided to each user in the test set [10]. It calculates the average precision (AP) across all of your users. The AP is a metric that compares a ranked list of your K suggestions to a list of the "right" or "relevant" recommendations for that particular user.

5.2. Model Performance

The Image-Music similarity results obtained after testing several transfer learning models used for training on the IMUSE Dataset were evaluated based on the above-mentioned metrics i.e., MAP@K, MAE, and MSE. The values of the metrics mentioned above, computed based on the test data results of the four pre-trained models deployed for the training of the image branch, are shown in Table. 2. It was observed that MobileNetV3 + SampleCNN model produces the best results for all metrics from the models taken into consideration giving a MAP@K value of 0.855, an RMSE of 0.246, and MAE of 0.225.

Table 2. Performance results of the proposed methodology on the IMUSE Dataset

Model	Similarity Evaluation Results		
	MAP@K (k = 10)	MAE	RMSE
ResNet-18 + SampleCNN	0.825	0.227	0.250
MobileNetV3 + SampleCNN	**0.855**	**0.225**	**0.246**
EfficientNetB4 + SampleCNN	0.754	0.225	0.248

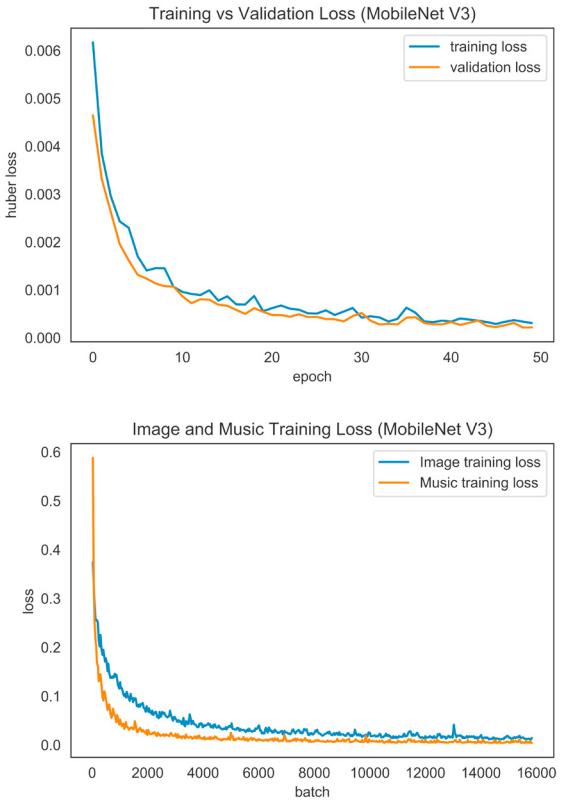

Fig. 5. Training and Validation loss graphs for MobileNetV3 + SampleCNN model.

Fig. 4 shows the comparison of map@k values in all 3 models across values of k ranging from 1 to 10. The low slope of MobileNetV3 showcases its ability to understand the dataset and provide good recommendations even while going lower down the list of recommendations. Fig. 5 shows the training vs validation loss for MobileNetV3 + SampleCNN model. The training loss is likely lower than validation lower due to the use of dropout layers. It also displays the change in different parts of training loss for the MobileNet v3 + SampleCNN model with every batch. From the graph, it can be inferred that minimizing the image loss has been harder than the audio loss. The low MAE and RMSE result imply that the emotions conveyed by the recommended music are relatively near to, and hence similar to, the emotional content of the input image in the VA emotional space. This also demonstrates the models' capacity to bridge the inter-modal gap reliably.

6. Conclusion and Future Scope

This paper proposes a cross-modal neural network system that takes an image as an input to recommend suitable music based on the emotional content of the image to the user. Various methods are studied to determine the best technique to extract features from images and music based on emotions, context, or other deep features over a shared latent space for cross-modal matching and retrieval. Raw audio is leveraged as an input to the model to obtain the music vectors. Similarly, an image is passed through a pre-trained CNN to obtain the image vectors. These music-image vectors are concatenated and then sampled and passed through Fully connected layers for training the whole model. Three transfer learning models are employed and compared for this task, from which MobileNetV3 + SampleCNN model generates the best results with MAP@K, MAE and RMSE values of 0.855, 0.225 and 0.246 respectively.

In the Future, there is a scope for expanding the research by adding lyrics on top of the emotion-based model to provide contextual recommendations with matching emotional content. Another angle for furthering the research in

this domain can be considering the third axis of emotional representation, i.e., Dominance and dataset curation based on 3-D emotional space for more accurate Image-Music Emotion Matching.

References

[1] Aljanaki, A., Yang, Y.H., Soleymani, M., 2017. Developing a benchmark for emotional analysis of music. PLOS ONE 12, e0173392.

[2] Baijal, A., Agarwal, V., Hyun, D., 2021. Analyzing images for music recommendation, in: 2021 IEEE International Conference on Consumer Electronics (ICCE), pp. 1–6.

[3] Bradley, M.M., Lang, P.J., 2017. International Affective Picture System. Springer International Publishing, Cham. pp. 1–4.

[4] Corchs, S., Fersini, E., Gasparini, F., 2017. Ensemble learning on visual and textual data for social image emotion classification. International Journal of Machine Learning and Cybernetics 10, 2057–2070.

[5] De Silva, L., Miyasato, T., Nakatsu, R., 1997. Facial emotion recognition using multi-modal information, in: Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications (Cat., pp. 397–401 vol.1.

[6] Defferrard, M., Benzi, K., Vandergheynst, P., Bresson, X., 2017. Fma: A dataset for music analysis.

[7] Deng, J.J., Leung, C.H.C., 2012. Music emotion retrieval based on acoustic features, in: Hu, W. (Ed.), Advances in Electric and Electronics, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 169–177.

[8] Ekman, P., 1992. An argument for basic emotions. Cognition and Emotion 6, 169–200.

[9] Grekow, J., 2016. Music emotion maps in arousal-valence space, in: Saeed, K., Homenda, W. (Eds.), Computer Information Systems and Industrial Management, Springer International Publishing, Cham. pp. 697–706.

[10] Hanna, P., 2018. Considering durations and replays to improve music recommender systems. ArXiv abs/1711.05237.

[11] Hizlisoy, S., Yildirim, S., Tufekci, Z., 2021. Music emotion recognition using convolutional long short term memory deep neural networks. Engineering Science and Technology, an International Journal 24, 760–767.

[12] Koelsch, S., 2014. Brain correlates of music-evoked emotions. Nature Reviews Neuroscience 15, 170–180.

[13] Kosti, R., Alvarez, J., Recasens, A., Lapedriza, A., 2019. Context based emotion recognition using EMOTIC dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence , 1–1.

[14] Kurdi, B., Lozano, S., Banaji, M.R., 2016. Introducing the open affective standardized image set (OASIS). Behavior Research Methods 49, 457–470.

[15] Lee, J., Park, J., Kim, K., Nam, J., 2018. SampleCNN: End-to-end deep convolutional neural networks using very small filters for music classification. Applied Sciences 8, 150.

[16] de Leeuw, R.N.H., Janicke-Bowles, S.H., Ji, Q., 2021. How music awakens the heart: An experimental study on music, emotions, and connectedness. Mass Communication and Society , 1–23.

[17] Marchewka, A., Żurawski, Ł., Jednoróg, K., Grabowska, A., 2013. The nencki affective picture system (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database. Behavior Research Methods 46, 596–610.

[18] Ortis, A., Farinella, G.M., Battiato, S., 2020. Survey on visual sentiment analysis. IET Image Processing 14, 1440–1456.

[19] Panda, R., Malheiro, R., Paiva, R.P., 2020. Novel audio features for music emotion recognition. IEEE Transactions on Affective Computing 11, 614–626.

[20] Pang, L., Zhu, S., Ngo, C.W., 2015. Deep multimodal learning for affective analysis and retrieval. IEEE Transactions on Multimedia 17, 2008–2020.

[21] Schlosberg, H., 1954. Three dimensions of emotion. Psychological Review 61, 81–88.

[22] Song, Y., Dixon, S., Pearce, M., 2012. A survey of music recommendation systems and future perspectives, in: 9th international symposium on computer music modeling and retrieval, pp. 395–410.

[23] Szwoch, M., Pieniażek, P., 2015. Facial emotion recognition using depth data, in: 2015 8th International Conference on Human System Interaction (HSI), pp. 271–277.

[24] Verma, G., Dhekane, E.G., Guha, T., 2019. Learning affective correspondence between music and image, in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3975–3979.

[25] Wang, L., Li, Y., Lazebnik, S., 2016. Learning deep structure-preserving image-text embeddings, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5005–5013.

[26] You, Q., Luo, J., Jin, H., Yang, J., 2016. Building a large scale dataset for image emotion recognition: The fine print and the benchmark, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI Press, Phoenix, Arizona. p. 308–314.

[27] Yu, Y., Tang, S., Raposo, F., Chen, L., 2019. Deep cross-modal correlation learning for audio and lyrics in music retrieval. ACM Transactions on Multimedia Computing, Communications, and Applications 15, 1–16.

[28] Zeng, D., Yu, Y., Oyama, K., 2020. Deep triplet neural networks with cluster-CCA for audio-visual cross-modal retrieval. ACM Transactions on Multimedia Computing, Communications, and Applications 16, 1–23.

[29] Zhao, S., Li, Y., Yao, X., Nie, W., Xu, P., Yang, J., Keutzer, K., 2020. Emotion-based end-to-end matching between image and music in valence-arousal space, in: Proceedings of the 28th ACM International Conference on Multimedia, Association for Computing Machinery, New York, NY, USA. p. 2945–2954.

[30] Zhen, L., Hu, P., Wang, X., Peng, D., 2019. Deep supervised cross-modal retrieval, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10386–10395.