POLITECNICO DI TORINO

Master's Degree in Data Science Engineering

Master's Degree Thesis

A Multimodal Encoder of Music and Image for Valence Arousal Prediction

Supervisors

Prof. GIUSEPPE RIZZO

Dr. LUCA BARCO

Dr. ANGELICA URBANELLI

Candidate

TIANMING QU

DECEMBER 2023

Abstract

Emotion analysis, a fundamental component of human-computer interaction, influences various domains, including content recommendation, image generation, and psychological research. Images and music, as crystallizations of human culture, inherently carry the emotions embedded by their creators. Analyzing the emotions conveyed in these works has long been a prominent direction of exploration in the field. Recent research in emotion analysis can be broadly categorized into two main streams: emotion label classification and valence-arousal prediction. My work primarily focuses on valence-arousal prediction. Valence represents the pleasure or displeasure elicited by a stimulus, while arousal indicates the degree of excitement or calmness. Both these metrics are crucial for the expression of human emotions. In recent years, with the rapid development of computer vision research, people have made breakthroughs in image and audio analysis. At the same time, multimedia applications that combine music and images have become increasingly popular, from advertising to movies to virtual reality experiences. Multi-modal analysis holds great promise in these contexts. In this context, my research endeavors to construct a multi-modal emotion prediction model employing metric learning. Throughout the experiments, I compare two different architectures for the encoders, one based on CNN (i.e. ResNet) and one based on more recent transformers. Different types of training losses are also applied with the aim of not only facilitating the model to acquire a shared latent embedding space but also allowing the model to learn the label space of the corresponding modality. I assess the performance across two types of encoders under this architecture, aiming to establish a foundation for subsequent research.

Table of Contents

List of Tables

List of Figures

Acronyms

AI
artificial intelligence

SOTA
State of the art

MER
Music emotion recognition

VA
Valence and Arousal

CNN
Convolutional neural network

RNN
Recurrent Neural Network

LSTM
Long Short-Term Memory network

ViT
Vision Transoformer

Chapter 1

INTRODUCTION

Emotion, which is one of the most important attributes that define human nature. Likewise, the importance of emotion in human creation cannot be ignored. Through the expression of music and images, people convey their emotions, transforming inner experiences into tangible works of art. Music and images, as carriers of emotion, have become powerful tools for expressing and sharing emotions.

Since the last century, people have been analyzing and predicting emotions. In the process of modeling emotions, two directions have become the focus. They are the emotion category model and the valence-arousal dimensional model, both are widely used in research. My research focuses on the valence-arousal model, in which valence represents one dimension in the emotion space, representing the positive or negative evaluation of a stimulus, while arousal is another key dimension, representing the energy level in the emotion. The model provides a systematic approach to describing and understanding complex emotional experiences, providing a structured framework for emotion analysis.

With the continuous development of deep learning research, data processing methods are no longer limited to a single modality, and multi-modal models have become a prominent trend. In the early days, people used data from another modality (such as text) as additional information to image data to construct multi-modal models. These studies were mainly based on CNN encoders. However, with the breakthrough of the Transformer-based architecture in the field of computer vision, it can process data of various modalities. This shift has made the Transformer-based architecture a highly regarded research direction, showing great potential for development, especially in multi-modal tasks.

Nowadays, research on multimodal models has been increasing. Can multimodal models effectively process data from different modalities based on emotional labels? Is there a noticeable difference in performance based on encoders with different architectures? These have become interesting research directions. For these reasons, I initiated my thesis work.

My thesis work studies a multi-modal valence arousal label prediction model, aiming to explore the performance comparison of CNN-based and Transformer encoders on multi-modal emotion prediction tasks and attempt to identify the differences in the capabilities of different encoders when handling data from various modalities.

The experiment consists of two frameworks: a single-modal model for predicting image or music valence-arousal labels and a multi-task multi-modal model that aggregates single-modal models to predict similarity scores for image-music pairs and the VA labels. Among them, the training of multi-modal models is based on metric learning, which is optimized through the combination of multiple training losses, aiming to lead the model to learn how to extract different modalities data features under the shared embedding space. At the same time, the model also needs to learn the connection between the valence arousal label and the image or music separately.

The selection and training of encoders is the core of my thesis work. For the CNN-based encoder, I chose ResNet pre-trained on ImageNet. Then, for Transformer-based encoders, I selected separate encoders for images and music. For image data, I used ViT (Vision Transformer), which was also pre-trained on ImageNet. It was released in 2021 and achieved SOTA performance on a variety of image processing tasks. For music data, I chose BEATs as the encoder, which applies the same mechanism as ViT and is pre-trained on a large audio dataset-AudioSet. A large number of experiments have proven that it has excellent audio data processing capabilities.

In order to place the outputs of different transformer-based encoders in a shared embedding space, I designed two strategies for aggregating the fully connected layer after the feature extractor. The first strategy incorporates all information from the output sequence of the transformer encoder into the fully connected layer, but the cost is increasing model complexity. The second strategy reduces complexity but sacrifices some information from the sequence. Additionally, comparative experiments were conducted to assess the merits and drawbacks of each strategy.

Through rigorous experiments, the results indicate that the encoder based on the Transformer architecture outperforms CNN in handling multimodal data. Additionally, in the task of predicting valence-arousal labels for music, the Transformer architecture exhibits a significant performance improvement. Although the performance enhancement in image emotion prediction is not as pronounced, it still reaches a level comparable to that of CNN-based encoders, highlighting the superiority of Transformer-based architectures in multi-modal tasks.

Following the completion of the experiment, I conducted a detailed visualization of the experimental process and the performance of the models. This visualization not only provides a more intuitive presentation of the training results but also

offers a richer understanding of the performance of the two different encoders in this experiment.

Through this work, I hope to provide valuable experience to the emerging field of multimodal emotion prediction, laying the foundation for its future advancement and applications.

Chapter 2

STATE OF THE ART

In this chapter, I will have a comprehensive introduction of the background that relates to my thesis work and state-of-the-art methods in this field. This chapter will be divided into three different parts:

In the beginning, I describe in detail the research of emotion recognition, which includes how researchers understand emotions and the mainstream methods for quantifying emotions. Next, I will discuss music and image emotion recognition, respectively, and introduce the development process and SOTA methods in this field.

The second part will center on the discussion of metric learning, which constitutes a fundamental component in many existing cross-modal retrieval tasks. I will go over a complete discussion of the mechanism of metrics learning and principles. Then, it includes the evolution of metrics learning from single-modal to multi-modal models and SOTA methodology.

In the final section of this chapter, I will elaborate on the core research direction of my work. I will detail findings in multi-modal models and provide an overview of recent research advances.

2.1 Emotion Recognition

Emotion is one of the most important aspects of human experience, and its identification and prediction have been the focus of research. In this section, I explore in detail how researchers model emotions and work to analyze the emotional content embedded in images and music.

2.1.1 Emotion Model

As the basis of emotion recognition, people first model emotions; in recent research, emotion modeling can be roughly divided into two approaches: categorical label and dimensional label:

Categorical label: it can also be called discrete labels, which means people use words to describe emotions, like "happy," "sad," "fear," "angry," and so on. Researchers have been working on this approach since the last century, and the model recognized by the academic community is based on Hevner's affective ring [1] in 1935 as Fig 2.1 shows. Hevner et al. interviewed several professional musicians for their opinions on different musical chords. They conducted many experiments, and they defined 67 emotional adjectives, which can be classified into eight categories: dignified, sad, dreamy, serene, graceful, happy, exciting, and vigorous. Most of the subsequent research on discrete labels is based on this model; Farnsworth et al. [2] invited 200 unprofessional students and showed them musical phrases to get the adjective. During the research, they redefined and regrouped the adjective into ten groups. In 2003, Schubert et al.[3] attempted to refine Havner's 67 emotion adjectives into 46 and reclassify them into nine groups.

Language is the core way for humans to express themselves; there are many obvious benefits to using adjectives to model emotions: first, adjectives as labels are easy to explain, and people can receive intuitive information from labels and match it with the corresponding data. Second, a suitable number of adjectives are easy to annotate because the annotator can use their language as the label when they check the data. However, as the research work went further and more detailed, the drawback of this manner appeared. Compared with the richness of human emotions, there needs to be more adjective categories for existing modeling. But using a finer granularity does not necessarily solve the problem since the language for describing emotions is inherently ambiguous and varies from different person [4], and a larger number of emotion categories will increase the burden on the subjects; recent research can reach 26 discrete categories [5] and even more, accurately annotating these data is undoubtedly a heavy burden.

Dimensional label: different from the classification manner, the dimensional approach focuses on the internal changes of human emotion. Researchers employ various emotional dimensions represented by named axes to quantify the intensity of

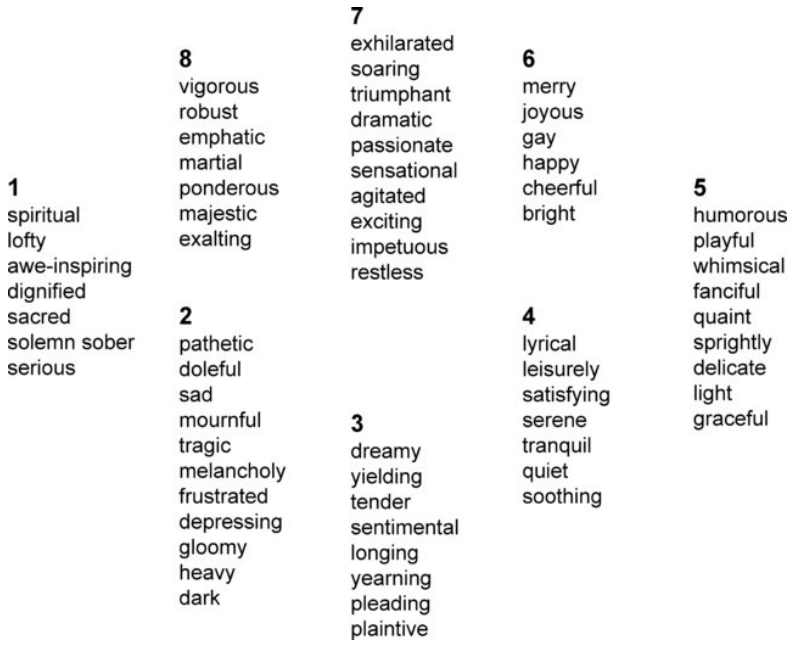


Figure 2.1: Hevners eight clusters of affective terms

emotions. Participants are instructed to employ a broader range of descriptors when characterizing the emotional aspects of music, which in turn yields foundational factors (dimensions). Despite variations in terms, these factors yield similar outcomes, as evidenced in psychological research [6]. Mainstream studies correspond to the following three dimensions: *valence*, the extent of pleasure, from negative to positive, *arousal*, the level of energy and stimulation, and *dominance*, the degree of being controlled.

Russell proposed the pioneering study on this in 1980 [7]; they proposed a *circumplex* model of emotion, constructed by valance-arousal dimensions. As Fig 2.2 shows, the positive emotions are distributed on quadrants 1 and 3 of the two-dimensional coordinates, while negative emotions are the opposite. This model can also be called the two-dimensional emotion space. Many studies based on this model and define their objective as a regression task. This model allows for a direct comparison of different emotions in two dimensions; however, this method is still controversial since some people think it blurs important aspects of the emotion process, like anger and fear; for this reason, some researchers introduced the third dimension into the emotional space, the dominant. With an additional dimension, the subject must annotate emotion in 3D, which will be more difficult. In summary, the two-dimensional model offers a better balance and has become popular.

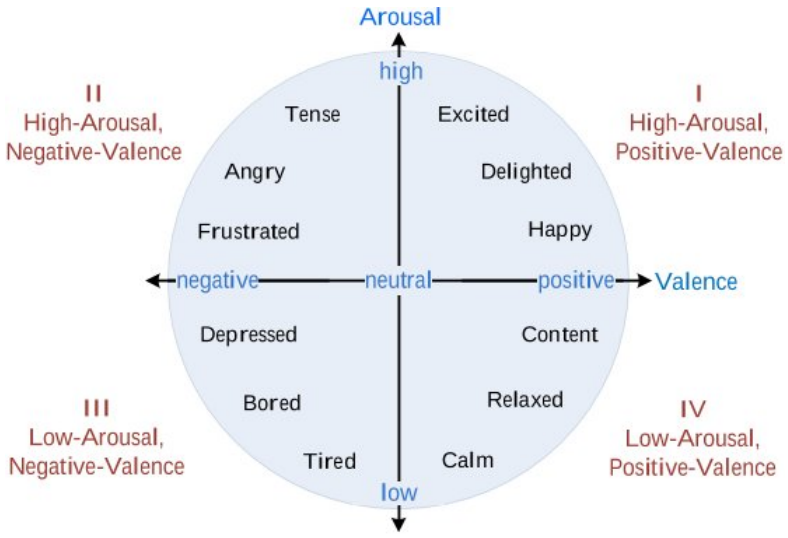


Figure 2.2: Russell circumplex model

2.1.2 Music Emotion Recognition

As Chapter 1 mentions, emotional information becomes an important feature for music retrieval. During the research, researchers found music stimulates people in many ways, and emotional perceptions are always related to different musical features [8]. So, the main issue of MER tasks is feature extraction; the accuracy of the tasks will be directly affected by the quantity of data. The type of feature extraction can be roughly classified into four groups: audio feature, symbolic feature, lyric feature, and biological feature.

Audio features are the most widely and earliest studied in MER tasks. People found that the valence dimension is usually associated with mode (major/minor) and harmony (consonant/dissonant); on the other hand, the arousal dimension is related to the pitch (high/low), loudness level, tempo, and timbre. The most commonly used audio features include pitch, energy (i.e., loudness level), and timbre. The work of Barthet et al. shows the timbre feature provides the best performance in the MER system when used as an individual feature [9]. The most commonly used timbre feature is MFCC, which represents the peak of the spectrum; spectral features are considered to be manifestations of correlations between vocal tract shape changes and articulatory frame movements. Nowadays, it has become the most commonly used feature.

Symbolic features mostly refer to features extracted from music scores. Characteristics of symbolic musical scores are often represented by MIDI (Musical Instrument Digital Interface); MIDI is a music file format that contains precise sequences of pitch, intensity, etc. In the research of Chen et al. [10] built a neural

network-based architecture and tried to extract features from MIDI files. Finally, they obtained an acceptable result and made the model sufficiently robust. However, compared with audio features, there are still fewer studies based on symbol features.

Lyric feature has always been an important feature in music. From classical opera to current pop music, most artists use lyrics as emotional carriers. With the rapid development of natural language processing, researchers began to turn their attention to lyrics. For the lyrics extraction, people usually base on NLP technical, like BOW, n-grams [11], etc. It is worth mentioning that Wang et al. used Tf-idf technical to model rhyme information and got a significant result [12]. Some studies try to use lyrics as an additional music feature and construct a multimodal model, like the research of Laurier et al.; they built a machine-learning model and trained it with audio-only and lyrics-only and mixed them. The result shows that even though lyrics-only performance is worse than audio, it has a complementary relationship with music and can be combined to improve a classification system [13]. This result also shows the drawback of the lyrics feature. Different types of language may distort the meaning of the lyrics, and some music changes the normal word order to suit the melody and rhythm [14]. So, in recent research, the lyrics feature is always used as an additional feature and combined with other audio features.

Biological feature: Music stimulation brings people many feelings, which include biological signal. In recent years, researchers have made bold attempts to generalize and collect data from subjects. [15] used functional magnetic resonance imaging (fMRI) data collected while participants listened to various film soundtrack excerpts. The work of Keelawat et al. [16] used electroencephalography (EEG) as a feature to identify music emotions; it can effectively capture information about emotions from the brain and has the characteristics of high temporal resolution and low cost. EEG has become the earliest and most commonly used biological signature. With the development of medical technology and wearable devices, collecting peripheral physiological signals such as heart rate (HR) and body surface temperature (TEMP) has become convenient and fast. Although compared to the research on other music features, the study based on biological features is still in its infancy, but it has shown enough value to drive people to research it.

From the beginning of this century, people have already started studying machine learning techniques to achieve music emotion recognition. Li et al. [17] considered the MER a classification task. They tried to extract timbre, rhythmic, and pitch features and used SVM as the multi-label classifier, but the results show that different emotion labels have a significant difference. Liu et al.'s [18] study extracted feature values from training music files through PsySound2, generated a music model from the generated feature data set through a classification algorithm, and used it to detect the emotion perceived in music clips. The model achieved satisfactory results. On the other hand, Yang et al. treat MER as a regression task

[19], and their study is also considered one of the earliest works to regard MER as a regression problem; this study uses existing toolkits to extract 114 audio features and input them into SVR.

With the popularity of neural networks, researchers have started constructing the MER model based on it. [20] proposed a model based on CNN trained with the original time and frequency domain information. [21] uses various feature extraction methods to convert the original data into spectrograms and then inputs the spectrograms into CNN for emotion recognition. Dong et al. [22] introduced BCRSN, called Bidirectional Convolutional Recurrent Sparse Network, which combines the advantages of CNN and RNN. CNN can learn features adaptively, while RNN is more suitable for processing sequence data. Meanwhile, with the lyrics feature, people try to combine it with other music features, [23] proposed a bimodal deep Boltzmann machine, which consists of two 2-layer DBM (deep Boltzmann machine) networks, one for audio and one for lyrics. They conducted a lot of experiments and proved this method is effective.

During the research, people realized the importance of time series to MER since the emotion does not stay the same in a single song. For this reason, researchers applied functional blocks that are good at handling long sequence data, like LSTM. [24] Segmented the music into seconds and extracted the super-segmentation features of each segment, which were then input into LSTM to obtain the dynamic change process of VA values; they found that LSTM correlated better with song-level annotations than machine learning methods. Meanwhile, people also tried to apply the attention mechanism to the model. [25] applied LSTM combined with an improved attention mechanism and successfully proved that the performance of the combined attention mechanism is better than that of LSTM only. The work of Ma et al.[26] used the attention mechanism to dynamically integrate different time scales to learn music's temporal and hierarchical information and obtained advanced results. Moreover, the attention mechanism is also applied to the feature extraction process; the work of [27] applies the attention mechanism to extract emotion-related features and then inputs the automatically extracted features into GRU-SVM to obtain the classification results. The experimental results show that the manner is superior to most compared methods. With the introduction of the Transformer architecture based on the attention mechanism, it can now handle various data forms, and its performance in the field of MER has also become a direction of my thesis work.

2.1.3 Image Emotion Recognition

Developing convolutional neural networks (CNN) has made them indispensable in computer vision, especially in image emotion recognition. [28] CNNs are good at extracting hierarchical features from images through iterative convolution and

pooling operations. This hierarchical feature extraction enables the model to abstract information from low-level edges and textures to high-level semantic details, thereby comprehensively capturing the emotional cues embedded in images. However, when researchers delved deeper into artistic imagery, they encountered additional complexities.

While convolutional neural networks (CNNs) have proven effective at identifying emotions in human facial images, artistic expression's inherent diversity and subjectivity require more nuanced approaches. In this regard, three features are proposed for image data in this task [29]:

Low-level features, which are derived from elements of art, including color and texture, etc. [30] proposed a novel CNN model that learns and integrates content information from higher layers of a deep network and style information from lower layers. They conducted extensive experiments on benchmark datasets to demonstrate the superiority of the proposed representation.

Medium-level features act as mediators between high-level semantic content and low-level elements and are typically applied in the context of more artistic photos. Compared with high-level semantics and detailed information on low-level elements, mid-level features provide a more abstract and general way to describe images. In image emotion recognition, mid-level features may include artistic principles, compositional styles, or more abstract visual elements that help capture emotional expressions in artistic photos. Compared with semantically rich high-level features, mid-level features are more suitable for processing images with certain artistic and aesthetic value.

High-level features is the semantic content contained in an image. People can easily understand the emotions an image conveys by recognizing the semantics. [31] attempts to bridge the emotional gap between the image's content and the viewer's emotional response through high-level concepts (HLC). It provides high-level semantic and contextual information about images, and the results show a high correlation with emotional categories.

Some researchers have also studied the task: [32] formulated the image emotion recognition task as a probability distribution learning problem. Since image emotions can be conveyed through visual features (e.g., aesthetic and semantic), Zhao et al. proposed a new framework to solve this problem by fusing multi-modal features. And they applied unsupervised domain adaptation technical. Likewise, people have also tried to apply visual attention techniques to image emotion recognition tasks.[33] proposed PDANet, which integrated spatial and channel-wise attention into a CNN with an emotion polarity constraint and got a SOTA performance. [34] developed a hierarchical attention mechanism in which polarity and emotion-specific attended representations are aggregated for discriminative feature embedding. They weighed the sample pairs adaptively under the guide of the attention module and achieved a good performance.

With the introduction of Vision Transformer (VIT) [35], transformer-based architecture has begun to be applied to more and more computer vision tasks. Its performance in image emotion recognition is also one of the focuses of my thesis work.

2.2 Deep Metric Learning

Metric learning is a distance metric-based method that quantifies the similarity or dissimilarity between objects. The goal is to reduce the distance between similar objects while increasing the distance between dissimilar objects, as shown in Figure 2.3. This could be an important strategy when precise data features are crucial for accurate classification [36]. However, metric learning usually uses linear projection, which is limited in solving real-world problems with nonlinear characteristics. In recent years, deep metric learning has provided better solutions for nonlinear data through activation functions, which has attracted the attention of researchers in many different fields. [37] through learning multiple fine-grained deep localized metrics and proposed a deep localized metric learning for visual recognition. Meanwhile, deep metric learning has also proven to be an effective method in audio recognition. [38]. Most of these studies are inspired by Siamese and Triplet networks. We will introduce these two networks below.

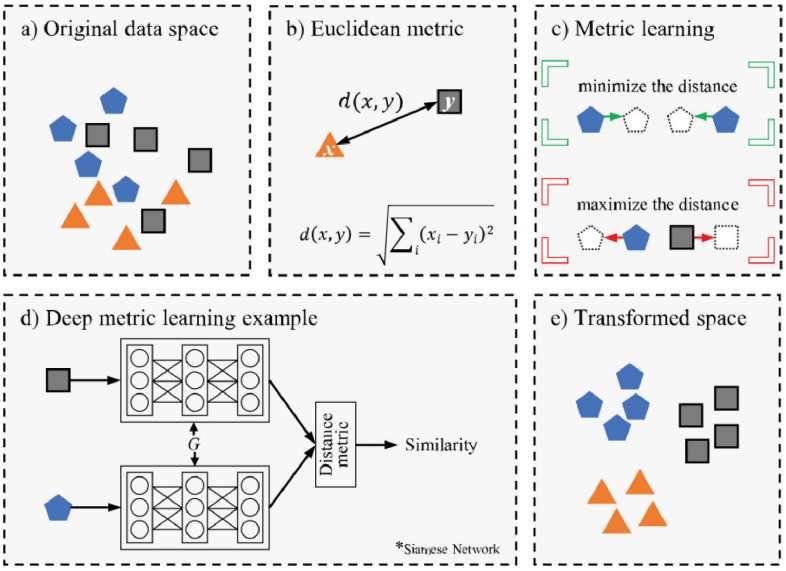


Figure 2.3: Deep Metric Learning

As a metric learning method, the Siamese network receives pairs of images (including positive samples and negative samples) and calculates the distance

between image pairs through a loss function. We reduce the distance between positive samples and expand the distance between negative samples to train the network model. Initially, the Siamese network was used with neural networks for signature verification [39]; the result shows the Siamese network is a very successful model for optimizing the distance between objects, thereby improving classification performance.

The triple network is an extension of the concatenated network principle, which contains three objects - anchors, positive samples, and negative samples [40]. The triple loss function aims to minimize the distance between anchors and positive samples while maximizing the distance between anchors and negative samples. This approach improves the discriminative power of the learned metric space and proves to be effective in tasks such as video-based embedding models [41].

Recent advancements extend metric learning into the realm of cross-modal scenarios. Xu et al. proposed the Deep Adversarial Metric Learning approach(DAML), which maps labeled data of different modalities into a shared latent feature subspace through nonlinear mapping. This subspace is aimed to minimize intra-class variation and maximize inter-class variation to ensure that the differences for each data pair captured by the two modalities from the same class are minimized separately [42]. Narrowing the gap between different modalities by identifying the embedding space with maximum correlation is one of the key approaches in cross-modal retrieval using metric learning. Meanwhile, people also consider the cross-modal heterogeneous issue, for example, the differences between the features of audio and visual modalities. [43] designed a novel adversarial metric learning (AML) model for audio-visual matching. It generates modality-independent representations via adversarial learning while learning robust similarity measures for cross-modal matching via metric learning. Inspired by previous work, zhao et al. proposed a novel cross-modal loss based on triplet loss and log-ratio loss to accurately optimize the distance of multi-modal embeddings based on emotional similarity. This is the way that I handle cross-modal data in my thesis work.

2.3 Multi-modal Model

As mentioned in the previous chapter, As the research continues to deepen, people are no longer limited to tasks involving single-modal data. In the first subsection, the researchers aim to enhance the performance of music emotion recognition by incorporating lyrics as additional features, marking an early stage of multi-modal research. With the continuous advancement of metric learning and computer vision, people have tried to develop multi-modal models based on CNN and achieved good results.

2.3.1 CNN-based Multi-modal Model

In 2017, Arandjelovic et al. The audiovisual correspondence (AVC) task was studied [44]. Given a video clip and an audio clip, the model needs to predict whether they are related, which constitutes a cross-modal retrieval task. To this end, they designed two independent CNN-based encoders for video images and audio and reorganized the encoding features into a decoder composed of two fully connected layers. After extensive experiments, the results show that the model can learn the characteristics of different modal data simultaneously and achieve excellent performance. Based on their work, [45] developed an end-to-end image-music emotion retrieval model by integrating deep metric learning. They designed similarities for images and music based on dimensional labels. Similarly, they employed CNN as an encoder. They designed a series of loss functions according to deep metric learning, aiming to enable the model to learn a shared latent embedding space for different modalities. This approach achieved state-of-the-art (SOTA) performance. As an improvement, [46] tried to design two different encoders for image and audio. Specifically, they integrated spatial, channel-wise, and temporal attention into a visual 3D CNN and temporal attention into an audio 2D CNN and achieved good results. Time series is also an important feature for audio data, so the attention mechanism has also been valued. [47] uses the attention-based long short-term memory (LSTM) model to select audio chunks and uses it to retrieve the entire audio with the corresponding video. In the latest research, people have begun to try to design recommendation models based on the matching of images and music based on dimensional labels [48].

At the same time, people have tried to adopt various methods to design multi-modal models based on image-text tasks. [49] designed two CNN-based image and text encoders, respectively. The model minimizes discriminative loss through supervised learning in label space and shared embedding spaces. The weight-sharing strategy is implemented to mitigate cross-modal differences in the common embedding space of multimedia data. The results demonstrate the effectiveness of the method.

2.3.2 Transformer-based Multi-modal Model

With the ongoing in-depth research into transformer architectures, researchers can now utilize data from various modalities as input to this framework. This has positioned transformers as popular encoders in current multi-modal models and an increasing number of studies are now attempting to construct transformer-based multi-modal models. For example, CLIP, launched by Openai in 2021 [50], uses data in two modalities: text and image. Researchers collected 400 million pairs of text-image data on the Internet and pre-trained the model to learn the representation of the image. The results proved that this method could compete

with a variety of fully supervised baselines with zero-shot and demonstrate the development potential of the transformer architecture in the cross-modal field.

At the same time, people are also studying whether this architecture is equally competitive in cross-modal emotion recognition tasks. Huang et al. [51] conducted a study on a multimodal transformer architecture for continuous emotion recognition. They utilized a transformer encoder to encode audio and visual modalities, employing multi-head attention to generate a multimodal emotional intermediate representation from their shared semantic feature space. The model's self-attention mechanism facilitated effective learning of long-term temporal dependencies. Additionally, they improved performance by integrating the Transformer model with LSTM, achieving superior results compared to other methods. Their study demonstrated the effectiveness of the attention mechanism in the Transformer architecture for continuous emotion recognition. [52] Use the transformer-based GPT model and RNN to model the data of three different modalities: audio, video, and text, respectively, and then use cross-modal fusion technology to obtain excellent results and make the model have sufficient robustness. [53] focuses on speech emotion recognition; they introduce a Transformer-based multi-modal learning framework customized for conversational emotion analysis. The framework models speech patterns and conversation content, and experimental results demonstrate the effectiveness of the method. A large number of studies have proven the role of transformer architecture in cross-modal emotion recognition, which has also become a strong basis for my thesis work.

Chapter 3

METHODOLOGY

3.1 Task Definition

My study builds on the work of Zhao et al. (2020) [45], who proposed a cross-modal model based on metric learning, focusing on image music matching and emotion label prediction. Specifically, for an input image-music pair, their model outputs similarity scores and continuous valence-arousal labels for the image and music. They adopted a CNN-based encoder in their work. However, since CNNs are designed for processing image data and may not effectively capture temporal features in music data, I decided to explore the effectiveness of the Transformer architecture, which incorporates an attention mechanism and is better suited for capturing temporal features. To distinguish them, I name the two methods "CNN-based encoder" and "Transformer-based encoder".

To do this, I built two model frameworks: one using a CNN-based encoder and the other using a Transformer-based encoder. In addition, I conducted experiments on continuous valence-arousal label prediction for music and image branches. In the following sections, I will detail the selection of the dataset and the construction of the model architecture.

3.2 Dataset

Given that my thesis work was based on deep neural networks, the necessity of a large dataset became critical. On the other hand, in order to ensure that the model effectively learns the features of multi-modal data, the selection of music and image corpora should also be fully considered.

For music corpora, it must cover a wide range of music styles, including pop, classical, electronic, jazz, etc. This diversity is crucial for the model's emotional expression in different musical contexts. At the same time, detailed emotional

annotation is equally important. Annotation can focus on the precise classification of emotions such as joy, happiness, and sadness; or the dimensional annotation of valence and arousal. This information allows the model to capture different emotional dimensions skillfully during the learning process. Finally, a wide range of artists and eras need to be considered in the dataset. This inclusiveness ensures that the model can adapt to the changing trends of different musical cultures and time periods.

Ensuring that the image corpora are diverse and representative is crucial for effective training of deep neural networks. The dataset should contain a wide range of images, from natural scenes and cityscapes to portraits and everyday objects. In terms of contextual diversity, images should cover a range of scenes, including indoor and outdoor environments, different lighting conditions, and different viewing angles. This broad contextual representation ensures that the model can consistently recognize and interpret emotions in different situations, thus contributing to its adaptability. Furthermore, displaying images of different artistic styles enables the model to identify and integrate emotional elements in various visual forms. This inclusiveness enhances the model's flexibility, allowing it to learn and generalize image information effectively.

Finally, in the design of multi-modal labels, unbalanced distribution should be avoided, and there should be enough samples for each emotional state to prevent the model from learning bias in certain emotions. At the same time, the rationality of label design should also be ensured. Multimodal labeling systems should be scalable and able to accommodate growing data sets and label types without compromising performance and efficiency.

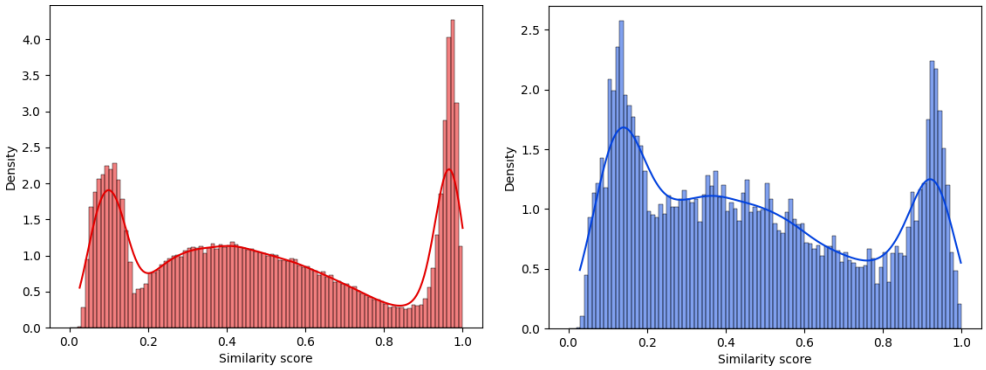
Based on the above considerations, I decided to use the image-music large-scale cross-modal emotion dataset launched by Zhao et al. in 2020, Image-Music-Emotion-Matching-Net (IMEMNet) [45], which consists of an image corpus and a music corpus based on continuous emotion labels. The dataset contains 140k image-music pairs. For multi-modal labels, the author calculated the similarity score based on the continuous emotion labels of its single-modal data, as Formula 3.1 shows.

$$S(I_i, M_j) = \exp(-\frac{d(y^{I_i}, y^{M_j})}{\sigma_n^m}), i = 1, ..., n, j = 1, ..., m, \tag{3.1}$$

Where y^{I_i} and y^{M_j} stands for the continuous emotion labels (valence-arousal label) for image I_i and music M_j. $d(\cdot)$, is the Euclidean distance between labels from different data. σ_n^m set as the average Euclidean distance between all music and image labels. The design of this multi-modal label integrates the emotional label features of images and music to provide a comprehensive and consistent multi-modal emotional representation.

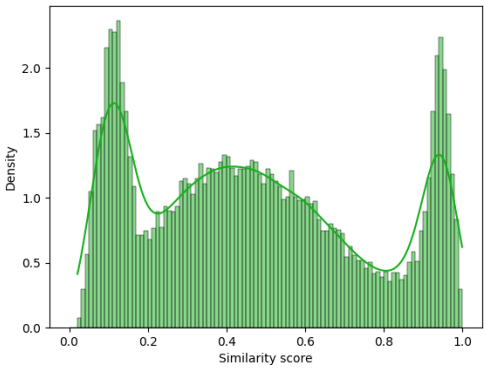
For the image-music pairs construction, since the number of all possible pairs is $m \times n$, it will lead to hundreds of millions of pairs. In order to avoid the

scale explosion, for a single music clip, the author selected 50 images, 10 of them with the highest similarity score and 10 with the lowest score; the remaining 30 were randomly selected. The distribution of similarity score as Fig 3.1 shows, the sampling strategy causes a high peak on the high matching score (> 0.8) and low matching score (< 0.2). This kind of distribution automatically offers an adequate number of positive and negative pairs. For the train-test split, the training set, verification set, and test set do not intersect.



(a) Training set distribution

(b) Validation set distribution



(c) Test set distribution

Figure 3.1: Similarity score distribution.

The dataset was constructed by music corpora, i.e., the Database for Emotional Analysis in Music (DEAM), and the image corpora is composed of three different sub-datasets, which are The International Affective Picture System (IAPS), Nencki Affective Picture System (NAPS), and EMOTIC. Table 1 shows the details of the dataset and how image-music pairs are allocated on the training, validation, and test sets.

	Training	Validation	Test	Total
Songs Num.	1,442	90	270	1,802
Songs clips Num.	21,804	1,750	6,759	30,313
Images Num.	19,770	1,275	4,918	26,026
Paires Num.	109,129	8,741	26,018	143,888

Table 3.1: Statistics of IMEMNet dataset

Subsequently, I will provide a detailed introduction to the information and composition of these sub-datasets.

3.2.1 Image Corpora

As mentioned before, the image corpus contains 20,496 images from three different data sets. The labels of all sub-datasets are annotated using valence-arousal to ensure the consistency of the labels.

The International Affective Picture System (IAPS) [54] Contains 1,182 documentary-style natural color images widely used in psychological research to evoke emotions. Each image in the set has been painstakingly annotated by approximately 100 college students on the Valence, Arousal, and Dominance (VAD) dimensions on a 9-point scale. In my task, in order to avoid a complex label space, only the valence and arousal labels are taken, and discarded the dominance label.

(a) Shopping scene (b) Theft scene

Figure 3.2: Examples of IAPS dataset

From natural scenery to life scenes, from human portraits to animal photos, this database includes a wide range of photos of various types. It is also worth noting that this database also contains some carefully designed images to recognize completely different or even opposite emotions in the same people and scenes. As shown in Fig 3.2, the scenes of the two images are the same: a woman in

the supermarket; (a) shows her walking in the supermarket, conveying a leisurely feeling (valence: 5.31, arousal: 3.26), while the content of (b) is that the woman was stealing in the same supermarket. She secretly put the products on the shelves into her bag. This picture conveys a sense of tension (Potency: 3.91, Arousal: 5.17). This intelligent design helps force the model to discern and learn subtle high-level features related to the emotion depicted in the image, going beyond simplistic predictions based solely on low-level features such as color and lines. By presenting contrasting scenes within the same visual context, the model is challenged to transcend the limitations of basic visual elements and grasp the complexity of human emotion interpretation in complex visual scenes.

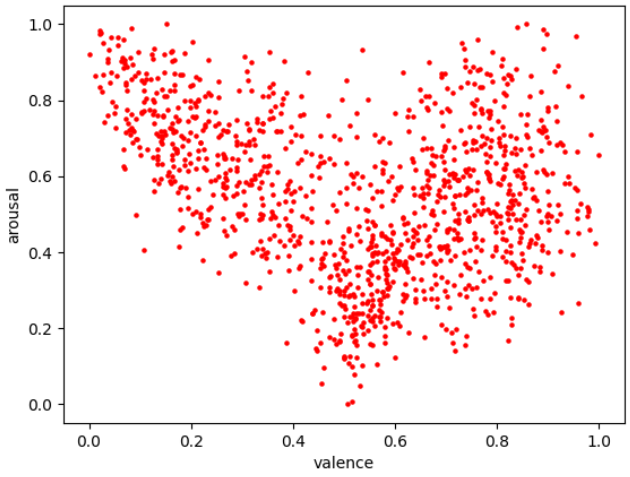


Figure 3.3: IAPS valence-arousal label distribution

In terms of labels, IAPS provides a complete and detailed continuous emotion label, and each picture is annotated with a valence-arousal label ranging from 0 to 9. In order to ensure the labels from different datasets are on the same scale, the labels are normalized in the range 0 to 1. Fig 3.3 shows the distribution of labels: Most labels are assigned in the central region of the distribution and lack high-valency low-arousal image labels and low-valency low-arousal image labels.

Nencki Affective Picture System (NAPS) [55] is a set of image databases based on continuous emotion labels launched by the Laboratory of Brain Imaging (LOBI) of the Polish Academy of Sciences in 2014. It also extends a variety of databases for different tasks, such as NAPS BE [56] based on discrete emotion labels, NAPS ERO [57] based on cross-sexual comparison study and SFIP dataset [58] based on the study of different phobias. Today, this dataset has been recognized by the academic community and is widely used in various studies.

The dataset contains a total of 1,356 real high-fidelity photos, carefully divided into five different categories: individuals, facial portraits, animals, inanimate objects, and landscapes. Each category contains a variety of visual stimuli covering a wide range of human experience and environmental contexts. The inclusion of these carefully curated categories helps enhance the representative richness of the dataset, enabling comprehensive exploration of all aspects of visual content.

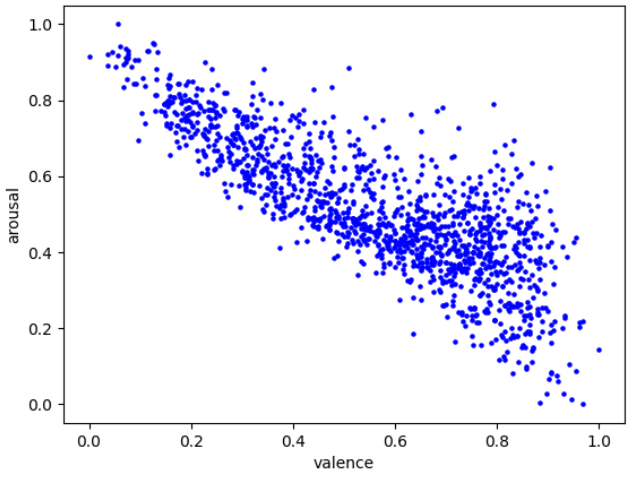


Figure 3.4: NAPS valence-arousal label distribution

These labels are continuous emotion labels annotated by 204 participants, most of whom are from Europe. The scale of the labels is in the range from 0 to 9. Fig 3.4 shows the label distribution of NAPS. Affective space in NAPS compared to IAPS. NAPS ratings demonstrated a more linear correlation between valence and arousal dimensions, in contrast to the "boomerang-shaped" relationship observed in IAPS. This difference is mainly attributed to the arousal dimension. Marchewka et al. [55] studied this phenomenon: in IAPS, both positive and negative images were rated as arousing, while neutral images were placed at the opposite extreme. In contrast, NAPS is characterized by negative images being rated as arousing, positive images being rated as relaxing, and neutral images being in the middle of the arousal scale. Selecting these two databases as image corpora at the same time can make them complementary in label space, allowing the model to learn more comprehensively.

EMOTIC, named after EMOTions In Context. This is a dataset of human images in different natural situations [59] [5], which is also the largest image sub-dataset in the image corpora with 23,082 images. The Emotic dataset exhibits a distinctive domain characteristic, given that all contained images are human-centric;

fig 3.5 shows some examples of the Emotic dataset. This divergence renders Emotic somewhat distinct from the other two sub-datasets and contributes to a more comprehensive perspective for the model.

Figure 3.5: Examples of EMOTIC dataset

Emotic offers 26 discrete emotion categories and continuous emotion labels in three dimensions (Valence, Arousal, and Dominance) from 0 to 10. The annotation work was provided by the Amazon Mechanical Turk (AMT) platform. The creator of the dataset employs a large number of non-professionals (about 20,000) from the platform, which provides a strong guarantee for the label consistency and stability of the data set. Interestingly, the label distribution of the Emotic dataset is significantly different from the other datasets, as shown in figure 3.6:

After label normalization, I found that the labels are evenly distributed in the label space. The Emotic label is noteworthy for its comprehensive nature, covering a wide range of emotional expressions, including extreme cases. This balanced label distribution has significant advantages for the model. This diversity ensures that the model is exposed to the nuances of a variety of emotions, allowing it to learn powerful features that generalize well to unseen data. Furthermore, the inclusion of extreme cases in labels poses a challenge for models to capture subtle distinctions and complex patterns in emotional expressions. Additionally, a balanced label distribution helps mitigate biases that may arise from imbalanced datasets. Models trained on uniformly distributed labels are less likely to be biased toward predicting general emotions, thus promoting a fairer learning experience.

In summary, these three image databases achieve extensive coverage of images from different domains, scenes, and subjects. Simultaneously, as fig 3.7 shows, the comprehensive continuous emotion labels encompass nearly all conceivable scenarios. This not only provides an ample amount of data for the models but also ensures that the models have sufficient transfer ability and robustness to handle

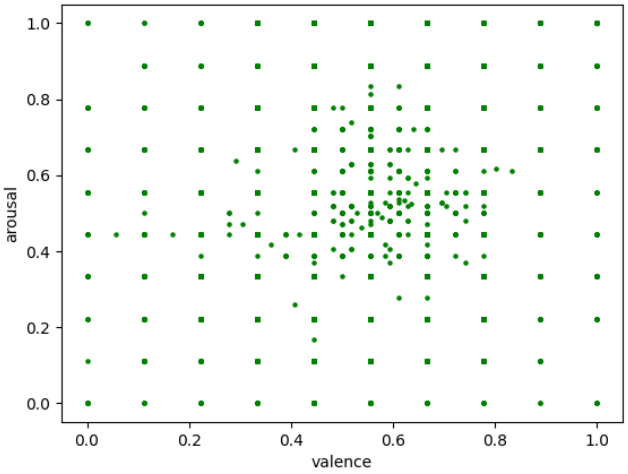


Figure 3.6: EMOTIC valence-arousal label distribution

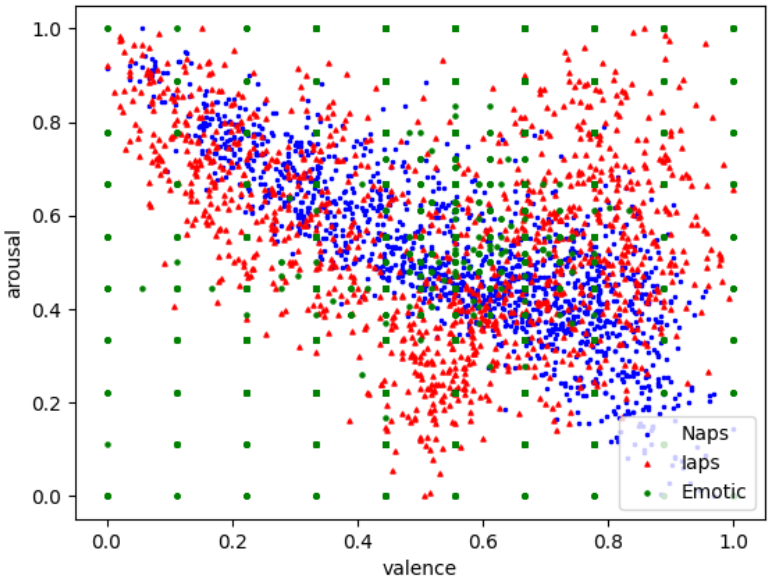


Figure 3.7: Image corpora label distribution

diverse emotional expressions in various application scenarios effectively.

3.2.2 Music Corpora

Considering the music corpora of the dataset, the author selected MediaEval Database for Emotional Analysis in Music (DEAM) [60], which was released in 2018. This dataset is constructed by several royalty-free music sources: freemusicarchive.org (FMA), jamendo.com, and the medleyDB dataset [61]. It contains 2058 pieces of music, and all the music in the dataset is re-encoded to have the same sampling rate, i.e., 44100HZ. The length of songs is diverse; most of them are around 45 seconds long, but some songs can reach a maximum length of 600 seconds. The coverage of music types is wide, from classical music to pop music, from instrumental solos to operas. The richness of its diversity contributes significantly to enhancing the model's generality and robustness.

For the labels of the music, which are annotated by 195 different raters. Each rater is responsible for annotating a part of the song. This database provides diverse emotional annotations: the valence-arousal labels for a single song and dynamically. The song-level labels are annotated in the range of 0 to 9, and the dynamical labels are generated every 500ms; the range is from -1 to +1. It is worth noticing that they removed the labels from the first 15 seconds due to the instability of the annotation at the start of the clips. The music corpora are based on dynamic annotations; they cut the song into two-second music clips and average the dynamic labels. In order to ensure the labels from different corpora are on the same scale, the range is re-scaled in the range from 0 to 1.

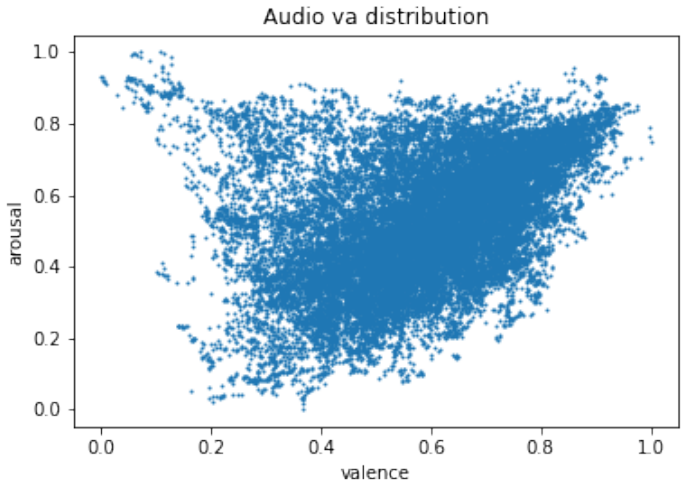


Figure 3.8: Music valence-arousal label distribution

Fig 3.8 shows the valence-arousal label distribution of music corpora. The majority of labels are allocated in the range from 0.5 to 0.8 on valence and 0.4 to

0.7 on arousal. It can also be observed that the low price high arousal area and the high price low arousal area have no distribution labels, and the low price low arousal represents boredom, which is an emotion that music should avoid making the listener appear. And high prices always bring a certain level of arousal, so both situations are acceptable.

In addition, this dataset also offers some features extracted by openSMILE [62], which contains the features for 500ms windows, with multiple kinds of audio features. Since my task is focused on cross-modal learning, I constructed the feature in another way.

In summary, in order to obtain a sufficient amount of music data, each piece of music is cut into 2-second clips. Excluding clips at the beginning and end of the song, the total number of clips is 35,817. The training set contains 28,825 clips, and the validation set and test set have 1,759 and 5,223 clips, respectively. These clips do not overlap each other.

3.2.3 Data Preprocessing

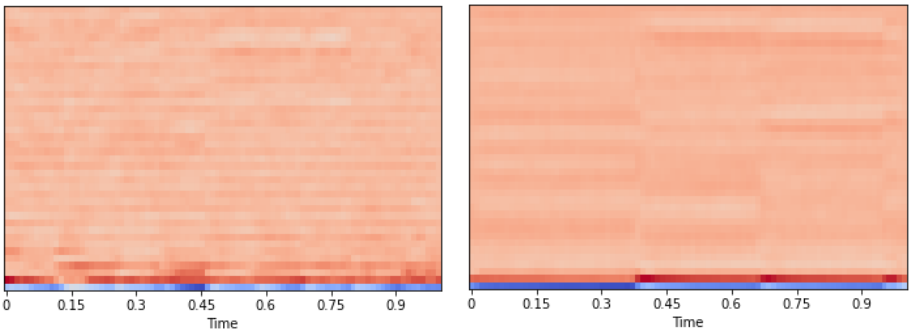
Data preprocessing serves as a critical conduit in model training, facilitating a better understanding of input information. Thoughtfully designed and selected types of data preprocessing enable the model to effectively capture key information within the input data, thereby enhancing the model's expressive capabilities.

These carefully crafted Data preprocessing manners not only aid the model in learning inherent associative information within the data but also, particularly in cross-modal training, support the model in comprehending correspondences between different modalities, consequently improving performance in cross-modal tasks.

Music Preprocessing

For music data, to match it with the encoder, a variety of audio features are extracted: Mel Frequency Cepstral Coefficients (MFCCs), Chroma features, Spectral contrast, Tonal centroid, and Mel spectrogram. Subsequently, I will elaborate on these features and provide a visual representation for better clarity.

Mel Frequency Cepstral Coefficients (MFCCs) mainly represent the spectral characteristics of audio signals, especially the frequency distribution in the Mel frequency domain. Specifically, MFCC captures the energy distribution of the audio signal on the spectrum, as well as the important frequency components on the Mel frequency axis. The calculation process of MFCC covers framing, Fourier transform, Mel filter bank application, logarithmic operation, and discrete cosine transform (DCT). This series of operations converts the audio signal from the time domain to the Mel frequency domain, ultimately generating a set of MFCC

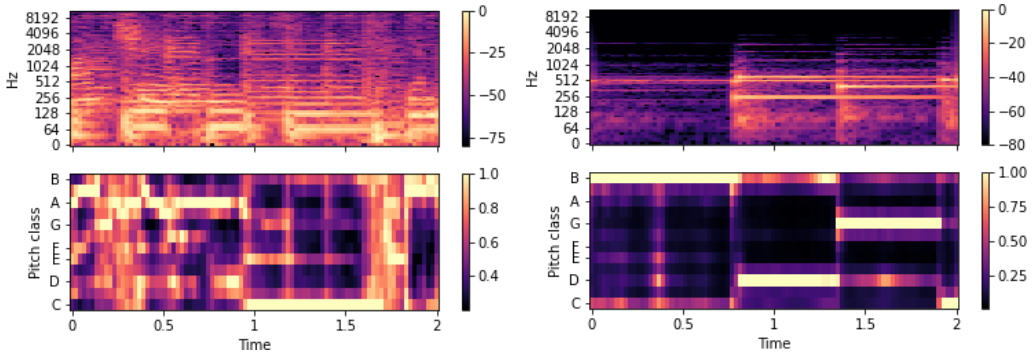


(a) High valence-arousal label (b) Low valence-arousal label

Figure 3.9: Visualization of MFCC with the opposite level of valence-arousal labels

coefficients that reflect important frequency characteristics in audio related to the human auditory system. Fig 3.9 shows the visualization of MFCCs, which contains audio for two levels of valence valance-arousal labels; high-level is va label higher than 0.7 and low-level means lower than 0.3.

Chroma features are a set of features used to represent tonal information in audio. Chroma features focus primarily on the distribution of tones and chords without taking into account the pitch of the audio signal. Chroma features can be used to analyze chord structures, melodic contours, and pitch changes in audio. Similarly, Fig 3.10 shows the visualization of Chroma features at different levels of VA levels.

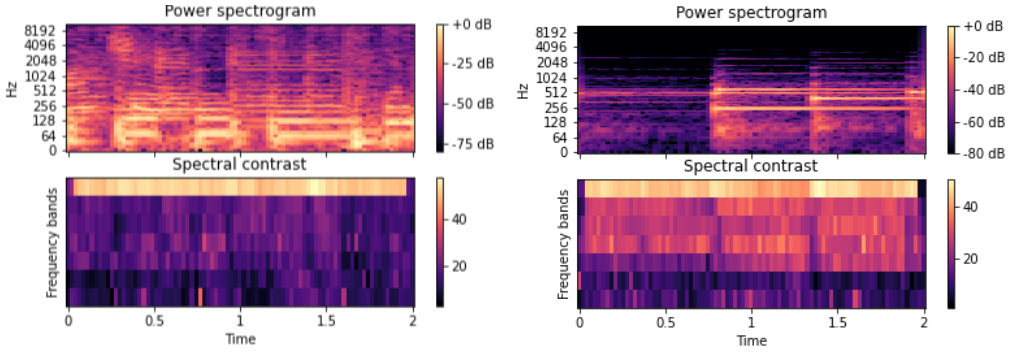


(a) High valence-arousal label (b) Low valence-arousal label

Figure 3.10: Visualization of Chroma features with the opposite level of valence-arousal labels

Spectral contrast is an audio characteristic that describes the difference in

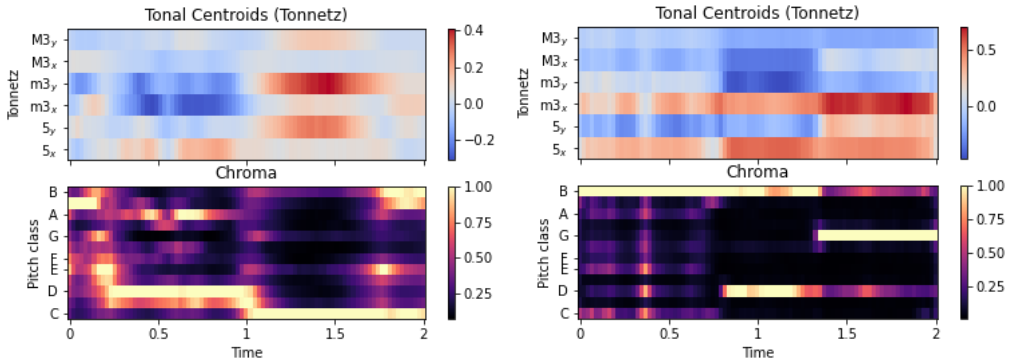
intensity between different frequency bands in the audio spectrum. It is useful for capturing harmonic structure, resonant properties, and other changes in the frequency spectrum in audio. In the field of music, it can be used to distinguish different musical instruments or identify different styles of music. In speech processing, it can also be used to extract speech features. Its visualization is shown in Figure 3.11.



(a) High valence-arousal label (b) Low valence-arousal label

Figure 3.11: Visualization of Spectral features with the opposite level of valence-arousal labels

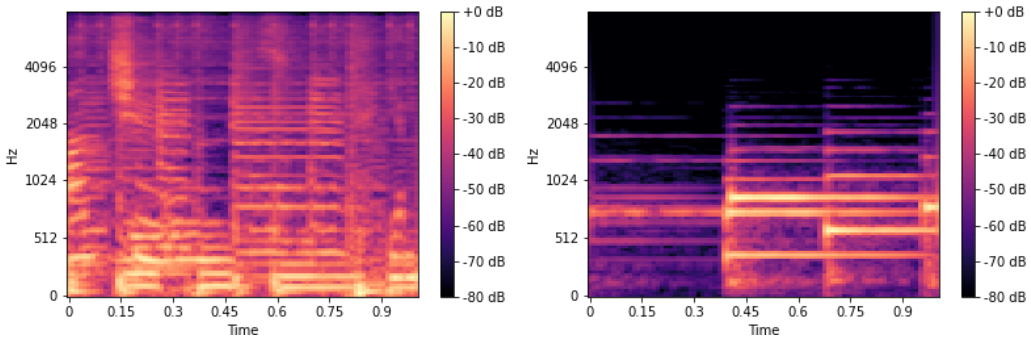
Tonal centroid is a feature used to characterize the distribution of pitch in the audio spectrum. It is obtained by computing the center frequencies of various frequency bands in the spectrum along with their corresponding energy weights. The tonal centroid provides the average pitch position of the audio spectrum, reflecting the distribution of pitch in the audio. In music analysis, the tonal centroid is often employed to differentiate the pitch characteristics among different musical pieces—fig 3.12 shows the visualization.



(a) High valence-arousal label (b) Low valence-arousal label

Figure 3.12: Visualization of Tonal centroid with the opposite level of valence-arousal labels

Mel spectrogram is a spectral representation method used for audio signal analysis. It uses Mel filter banks to process signals in the frequency domain. The Mel spectrogram mainly represents the spectral distribution of the audio signal on the Mel frequency axis rather than the traditional linear frequency axis. This representation is more consistent with the way the human ear perceives pitch, so it is widely used in speech processing and music analysis. The visualization of the mel spectrogram is shown in Fig 3.13.



(a) High valence-arousal label (b) Low valence-arousal label

Figure 3.13: Visualization of Mel spectrogram with the opposite level of valence-arousal labels

As shown from the visualization above, there are clear differences in these characteristics for music with different valence-arousal levels. These features can more effectively build high-performance emotion recognition models. This helps improve the model's sensitivity to music-emotional information, making it more

reliable and effective in practical applications.

Musical features were generated by extracting 12 chroma features, 7 spectral contrast features, 40 MFCCs, 6 tonal centroids, and 128 features from the mel spectrogram. Assuming that the duration of the music clip is 2 seconds, the characteristic length along the second dimension is 87. In order to align the music data with the input dimensions of the CNN-based encoder, the data needs to be reshaped into the format of (H, W, C). Therefore, these features are connected and tiled to achieve the shape of (193, 87, 3). To omit the feature extraction step during training, I pre-extracted the music features and stored them in a Numpy matrix using the Numpy library. as shown in Figure 1.

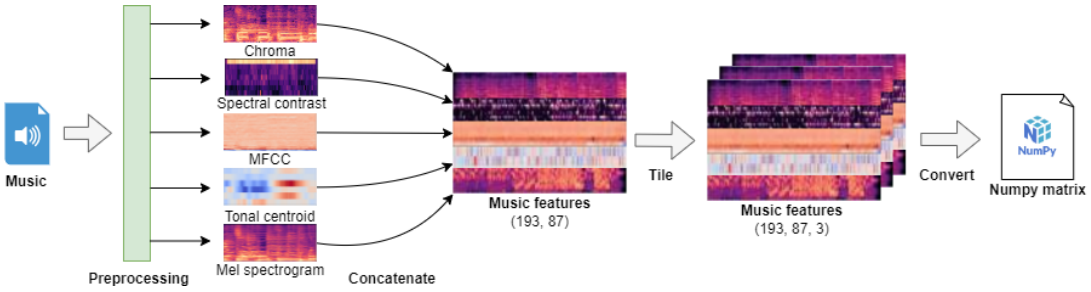


Figure 3.14: Music features extraction

For transformer-based encoders, the extraction of music data is simplified and only the waveform data of the music is required. Given a sample rate of 44100, the length along the second dimension is determined by the duration of the music (2 seconds) multiplied by the sample rate. In summary, the shape of the musical feature will be (1, 88200). Likewise, for the music features in this case, I used the TorchAudio library to pre-extract them and stored them using the Pytorch library.

Image Preprocessing

For the feature extraction of image data, current methods are relatively mature. Specifically, different-sized images are first resized to dimensions of 224 by 224. Subsequently, RGB features are extracted, and the features are normalized based on the statistical data (mean, standard deviation) from the pre-training database (ImageNet 1K) of the image encoder. The final result is data with a shape of (224, 224, 3). The current methods for feature extraction of image data are relatively mature. Specifically, images of different sizes are first resized to 224×224 size. Subsequently, RGB features are extracted and normalized based on the statistics (mean, standard deviation) in the image encoder pre-training database. The final result is data of shape (224,224,3). For CNN-based encoders, the data will be stored in a NumPy matrix, while for Transformer-based encoders, the data will be

preprocessed using a pre-trained preprocessor and will be stored in a tensor matrix. The process is shown in Figure 3.15.

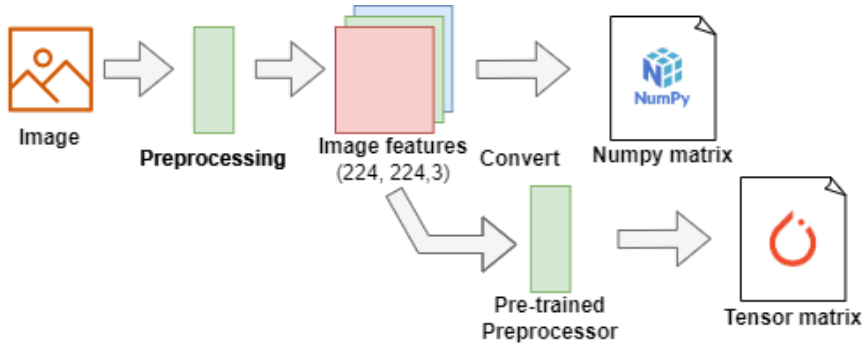


Figure 3.15: Music features extraction

3.3 Model Structure

A reasonable model structure has many advantages during the training process and plays a crucial role in the performance, convergence speed, and generalization ability of the model. First, it enables the model to better capture key features in the input data, thereby improving the model's performance. With carefully designed hierarchies and activation functions, the model can learn nonlinear relationships in the data more effectively, thereby improving its expressive power.

Secondly, a reasonable model structure can balance the complexity of the model and avoid overfitting. Appropriate regularization techniques, judicious selection of layers and nodes, and effective dropout methods help control the complexity of the model and enhance its ability to generalize to unseen data.

In the following sections, I will go ahead and introduce the model structure used in my thesis work in detail.

3.3.1 Model Encoders

As mentioned before, I attempted to use CNN-based encoders and transformer-based encoders. For a CNN-based encoder, its output is a two-dimensional tensor (includes batch size as the first dimension), which can be easily connected to subsequent fully connected layers.

When moved to the transformer-based encoder, unlike that of a ResNet, the output is a three-dimensional tensor. This poses a challenge when integrating it with a fully connected layer. To overcome this. I've tried two strategies:

The first method is to flatten the data in two dimensions, ultimately transforming it into a 2D tensor. While this method retains all sequence information, it leads to an exceptionally large input dimension for the subsequent fully connected layer, approximately 150,000.

The second strategy draws inspiration from the Vision Transformer. In this approach, a Class sequence (CLS) token is added to the first position of the output sequence for classification tasks. The embedding of the CLS token aggregates information from all other tokens, ensuring it retains the most relevant information compared to other tokens. This method leverages the CLS token to capture and consolidate essential information for downstream tasks. During the experiment, I tried both strategies and compared their performance.

Next, I will detail the components of the model and its parameters.

Image Encoder: My image encoder is based on CNN and Transformer. For CNN-based, I used an image encoder that is now widely used in the field of computer vision, ResNet [37]. ResNet (residual network) is a deep convolutional neural network architecture. Its main feature is the introduction of residual learning, which directly adds the input to the output of the middle layer of the network through a shortcut connection, thereby effectively solving the gradient disappearance and gradient explosion problems in deep network training. For the image part, ResNet50 is selected.

For aspects of transformer-based encoders, I selected ViT (Vision Transformer) [35], which is an image classification model based on a self-attention mechanism, which breaks through the conventional framework of traditional CNN in processing image tasks. Compared with CNN, ViT pays more attention to the capture of global information. By dividing the input image into small blocks and flattening these blocks into a one-dimensional sequence, it achieves global interaction with the overall image. Vision Transformer performs well in image classification tasks, especially showing strong generalization capabilities on large-scale image data sets; fig 3.16 shows the mechanism of ViT. To put the features from different modalities into a shared embedding space, I added a fully connected layer after the image encoder to reshape features into 512 dimensions. Table 3.2 shows the details of the image encoder parameters.

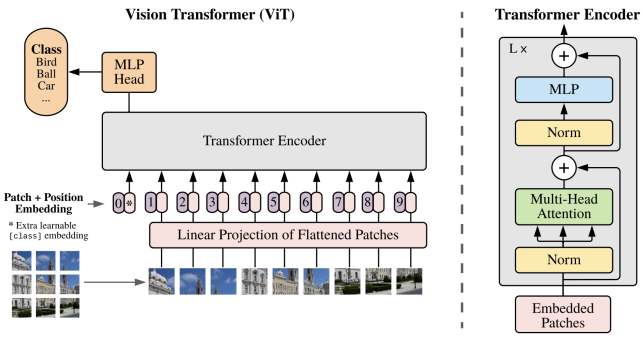


Figure 3.16: Vision Transformer

Encoder	Parameters	
CNN-based	Input dimension	Output dimension
Resnet50 (ImageNet) [1]	(224, 224, 3)	(2048,)
Fc	(2048,)	(512,)
Transformer-based		
ViT (ImageNet)	(224, 224, 3)	(197, 768)
Fc (v1) [2]	(151296,)	(512,)
Fc (v2)	(768,)	(512,)

Table 3.2: Image encoder parameters

Music Encoder: As the same as the images encoder, I selected ResNet18 for CNN-based encoder. For the transformer-based encoder, I decided to use BEATs [63], which was published in 2020 by Microsoft. I use this model based on the following considerations: First, BEATs stand out among many transformer-based encodings and achieve sota performance on multiple audio tasks. Secondly, as a model based on self-supervised learning (SSL), it is pre-trained on a large-scale database and optimized for multiple iterations, especially on AudioSet, which contains thousands of hours of audio extracted from YouTube. The data covers more than two hundred different audio categories, including quite a lot of music data. This is relevant to a music dataset from my thesis work. The detailed parameters of the music encoder are presented in Table 3.3

[1]Pre-trained dataset

[2]Fc is the abbreviation of the fully connected layer, v1 represents the flatten strategy, and v2 represents the CLS token embedding strategy.

Encoder	Parameters	
CNN-based	Input dimension	Output dimension
Resnet18 (ImageNet)	(224, 224, 3)	(512,)
Fc	(512,)	(512,)
Transformer-based		
BEATs (AudioSet)	(1, 88200)	(96, 768)
Fc (v1)	(73728,)	(512,)
Fc (v2)	(768,)	(512,)

Table 3.3: Music encoder parameters

3.3.2 Model Decoders

As mentioned before, my thesis work is designed for two different tasks, namely image-music similarity prediction and valence-arousal prediction of these two different modal data; two decoders are needed.

In the context of similarity prediction, a concatenation of data from various modalities is fed into the decoder architecture. The decoder consists of three fully connected layers, which reduce the dimensionality of the input features to 256 and 128, respectively. The last fully connected layer will output the result, i.e., it reshapes the output dimension as 1. To reduce the risk of overfitting, normalization, and dropout mechanisms with a dropout rate of 0.5, are incorporated between each fully connected layer. In addition, the middle layer uses ReLU (rectified linear unit) as the activation function to introduce nonlinearity into the model. The final layer utilizes a sigmoid function to ensure that the output specification falls within the constrained range of 0 to 1. This design choice imposes constraints on the output values, enhancing the interpretability and relevance of similarity predictions.

The decoding mechanism for continuous emotion labels follows the same structure as the similarity predictor. However, the input method is slightly different. Specifically, different features extracted from various modalities are fed into the decoder individually. After the same decoding step, its output dimension is 2, corresponding to the valence and arousal values respectively. This customized input configuration enables the decoder to process data from different modalities, thereby facilitating the generation of predictions related to continuous emotion labels.

3.4 Losses Design and Combination

Metric learning takes center stage in my thesis, so the experimental design involves a combination of multiple losses. Within this framework, it is crucial to discuss the right combination of design and loss, as they directly impact the effectiveness of

model learning.

First, the choice of loss must fully consider the nature of the task and the characteristics of the data. Different loss functions show different effects when processing different types of tasks and data. By carefully analyzing the goals of the task and the properties of the data, I can choose loss functions suitable for metric learning, ensuring their positive impact on model learning.

Second, the combination of losses needs to consider their contributions and their interaction within the overall framework. Some losses may prioritize model stability, while other losses may contribute more to improving the model's generalization performance. By considering these factors, I can design a more comprehensive and effective combination of losses, thereby improving the performance of the model when facing data in different modalities.

The selection of losses involves two main tasks: cross-modal similarity loss and valence-arousal prediction loss. The goal of the cross-modal similarity loss is not only to enable the model to learn the similarity between different modalities but also to reduce the distance in the shared embedding space between feature data from different modalities with high similarity while increasing the distance between data with low similarity. On the other hand, the single-modal valence-arousal prediction loss aims to enable the model to learn the relationship between different modalities and their respective continuous emotion labels space, facilitating effective prediction of valence-arousal labels. Next, I will explain in detail the loss designs used in the paper and how they are combined.

Cross-modal Feature-Ratio Loss. As mentioned in chapter 2, traditional metrics learning is based on increasing the distance between the distinct classes and reducing the distance between the same class, which works well on the majority of classification tasks. Since my thesis work is a regression task, it is hard to define different classes, and the now widely used metric learning-based loss is difficult to apply to my experiments. Kim et al. proposed a novel triplet loss in 2019 [64], which enables the model to learn the similarity of different features. Following his research, Cross-modal feature-ratio loss was used. Its formula is shown in 3.2.

$$\begin{aligned} L_{CFR} = &\sum_{i=1}^{N} \left\{ \log \frac{D(f^{I_i}, f^{M_i})}{D(f^{I_i}, f^{M_j})} - \log \frac{S(I_i, M_i)}{S(I_i, M_j)} \right\}^2 \\ &+ \sum_{i=1}^{N} \left\{ \log \frac{D(f^{M_i}, f^{I_i})}{D(f^{M_i}, f^{I_j})} - \log \frac{S(f^{M_i}, f^{I_i})}{S(M_i, I_j)} \right\}^2 \end{aligned} \tag{3.2}$$

For given image i and music i, f^{I_i} and f^{M_j} stands for the image feature and music feature after encoding, where $D(\cdot)$ is the Euclidean distance between the features, and $S(\cdot)$ is the similarity score between the current image and music. In the first half of the formula, image i serves as the anchor, and music i serves as the second term. Simultaneously, a random selection of music j is chosen as the third

term for image i. The same principle applies to the second half of the formula.

This loss aims to increase the distance between image-music pairs with lower similarity scores and reduce the distance between image-music pairs with higher similarity scores. However, during the experiment, I found that the summation operation caused extremely high losses from the beginning, and there was a high risk of gradient explosion, which forced me to train the model with an extremely low learning rate. In order to overcome this problem, I tried to change the summation operation to the mean operation, which brought the loss to an acceptable range and allowed the model to be trained at a normal learning rate, as shown in Formula 3.3.

$$\begin{aligned} L_{CFR} = & \frac{1}{N}\sum_{i=1}^{N}\left\{\log\frac{D(f^{I_i}, f^{M_i})}{D(f^{I_i}, f^{M_j})} - \log\frac{S(I_i, M_i)}{S(I_i, M_j)}\right\}^2 \\ & + \frac{1}{N}\sum_{i=1}^{N}\left\{\log\frac{D(f^{M_i}, f^{I_i})}{D(f^{M_i}, f^{I_j})} - \log\frac{S(f^{M_i}, f^{I_i})}{S(M_i, I_j)}\right\}^2 \end{aligned} \tag{3.3}$$

Cross-modal Feature-Margin Loss. This is a classic marginal loss, which can help the model increase its stability. As shown in the formula 3.4.

$$L_{CFM} = \sum_{i=1}^{N}\left[\left\|f^{I_i} - f^{M_i}\right\|_2 - \alpha\right]_+ \tag{3.4}$$

Where $[\cdot]_+$ represent the $max(0, \cdot)$, it will return 0 when the difference between the image and music feature is less than 0, and the α is the threshold for the maximum tolerable distance. The introduced loss function L_{CFR} imposes constraints on the encoders, forcing them to extract features that do not exceed a predetermined maximum distance. This strategically imposed constraint helps the model effectively incorporate the unique characteristics inherent to the two different modal data types within the scope of the shared embedding space.

Essentially, the proposed loss function acts as a regularization mechanism, guiding the learning process to encourage the generation of embeddings that maintain a specified level of consistency between image and music features. By imposing an upper bound on acceptable dissimilarity, the model becomes adept at seamlessly integrating information from different modalities, thereby enhancing its ability to fuse cross-modal features. This regularization technique helps improve the overall robustness and effectiveness of the model when dealing with heterogeneous data sources.

Single-modal Feature-Ratio Loss, which belongs to the single-modal valence-arousal prediction losses family, calculates the loss of image features and music features, respectively. The formulas are shown in 3.5 and 3.6. where y^{I_*} and y^{M_*} are the valence-arousal labels of Images and Music. $D(\cdot)$ stands for the Euclidean

distance, and for each anchor I_i and M_i, their neighbors j, k are selected for computing the loss.

$$L_{SFR\_I} = \sum_{i=1}^{N} \left\{ \log \frac{D(f^{I_i}, f^{I_j})}{D(f^{I_i}, f^{I_k})} - \log \frac{D(y^{Ii}, y^{Ij})}{D(y^{Ii}, y^{Ik})} \right\}^2 \tag{3.5}$$

$$L_{SFR\_M} = \sum_{i=1}^{N} \left\{ \log \frac{D(f^{M_i}, f^{M_j})}{D(f^{M_i}, f^{M_k})} - \log \frac{D(y^{Mi}, y^{Mj})}{D(y^{Mi}, y^{Mk})} \right\}^2 \tag{3.6}$$

The mechanism is the same as L_{CRF}. The proposed loss function aims to group features with similar valence-arousal labels while dispersing features with different labels. This addition helps emotion predictors establish more effective connections between single-modal data features and their emotional information. This loss function enhances the model's ability to effectively predict emotion labels by encouraging the model to discern relationships in the data related to specific emotional states and preventing over-generalization between different emotion categories. Taken together, it serves as a guide, promotes a detailed understanding of emotional characteristics, and improves the overall accuracy of Emotion recognition.

Meanwhile, the summation operation in the process has the inherent risk of causing a gradient explosion during the training process. To alleviate this concern, I chose to replace it with an average operation. This modification helps stabilize the training process by preventing excessive gradients, thereby enhancing the overall robustness and convergence of the model during optimization.

Mean Squared Error (MSE) loss. Given that the focus of my work revolves around regression tasks, specifically the prediction of similarity scores and valence-arousal values, specifically tailored training losses had to be incorporated to optimize performance on these tasks. Therefore, the mean square error (MSE) loss, also known as L2 loss, has been integrated into the training framework. Their formulas are shown in 3.7, 3.8, and 3.9, where the parameters with hat stand for the prediction result, and the opposite is the ground truth.

$$L_{Sim} = \frac{1}{N} \sum_{i=1}^{N} (S(I_i, M_i) - \hat{S}(I_i, M_i))^2 \tag{3.7}$$

$$L_{MSE\_I} = \frac{1}{N} \sum_{i=1}^{N} (y^{I_i} - \hat{y}^{I_i})^2 \tag{3.8}$$

$$L_{MSE\_M} = \frac{1}{N} \sum_{i=1}^{N} (y^{M_i} - \hat{y}^{M_i})^2 \tag{3.9}$$

This choice of loss function is ideal for regression scenarios because it quantifies the average squared difference between predicted and actual values. By introducing

an MSE loss, the training process is tailored to specifically address the nuances of regression, aligning the model's optimization goals with the complex task of similarity score and valence-arousal prediction.

Loss Normalization and Combination. Throughout the experiment, I was dealing with the challenge of varying scales between various loss functions. Taking Loss CFR as an example, the loss is calculated based on the Euclidean distance between features extracted by the encoder from different modal data, ranging from 10 to 10^2. In contrast, the MSE loss is tailored to model predictions in the range 0 to 1, resulting in a scale of approximately 10^{-1}.

The difference in scale between different loss functions may lead to instability in the training process, as the optimization process may be more sensitive to larger-scale losses, and there may even be a risk of gradient explosion. To solve this problem, a feasible approach is to introduce appropriate scaling factors or weight adjustments to balance the contributions of different loss functions. This adjustment ensures that the impact of each loss function on the overall optimization goal is relatively balanced.

Another strategy is to normalize the loss functions so that they operate within a similar range of values. This can be achieved through standardization or normalization of loss values, ensuring a more consistent impact on model parameter updates.

In my study, I tried to implement a normalization strategy. Specifically, at the beginning of training, I recorded the exact value of each loss in the first batch and assigned it as the maximum value. Subsequently, during the next training session, I normalized each batch's loss to a range of 0 to 1 by dividing it by the loss in the first batch. Eventually, when combining all the losses, I calculated their average value to ensure that the final combined loss also remained within an acceptable range. The formula 3.10 shows the total loss of a single batch. N represents the number of combination losses, and $\hat{L}$ stands for the loss of the first batch.

$$\begin{aligned} L_{TOT} = \frac{1}{N}\Big(&\frac{L_{CFR}}{\hat{L}_{CFR}} + \frac{L_{CFM}}{\hat{L}_{CFM}} + \frac{L_{Sim}}{\hat{L}_{Sim}} + \frac{L_{SFR\_I}}{\hat{L}_{SFR\_I}} \\ &+ \frac{L_{SFR\_M}}{\hat{L}_{SFR\_M}} + \frac{L_{MSE\_I}}{\hat{L}_{MSE\_I}} + \frac{L_{MSE\_M}}{\hat{L}_{MSE\_M}}\Big) \end{aligned} \tag{3.10}$$

Taking these steps addresses the potential training instability caused by the varying sizes of the different loss functions. By normalizing each loss to operate within a comparable range of values, I aim to balance their impact on the training process more effectively. Additionally, averaging the losses helps ensure that the scale of the individual loss functions does not overly influence the final composite loss.

3.5 General Pipeline

After exhaustively detailing all the basic components of model training, the training process studied in this paper comes into focus, as shown in Figure 1. In this comprehensive framework, various components interact with each other to build an efficient training process with metric learning as the core. The goal is to achieve precise adjustment of model parameters to maximize optimization of predefined training objectives.

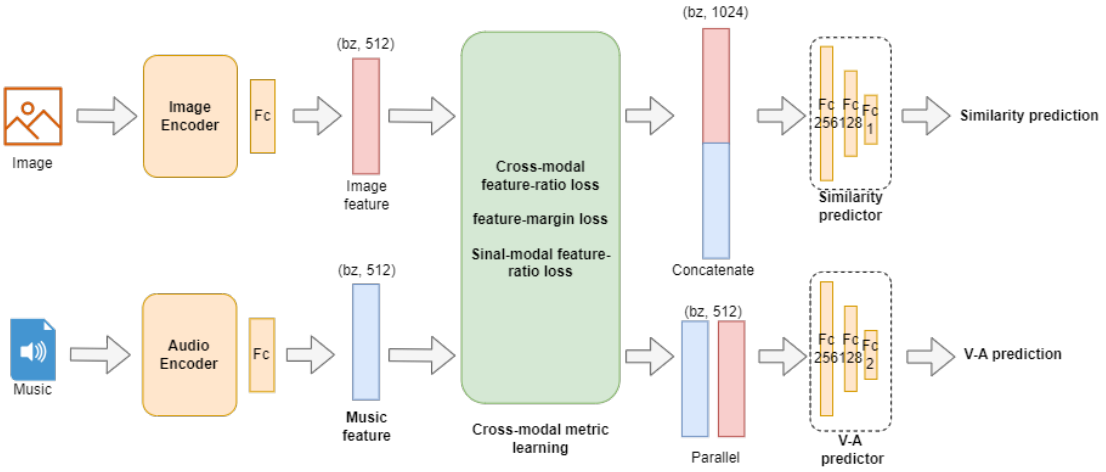


Figure 3.17: Pipeline of the training process

This is an end-to-end cross-modal emotion recognition model. In this model, preprocessed image data and music data are input into two predefined encoders respectively. The outputs of these encoders are then encoded and integrated into 512-dimensional features. This step aims to map data from different modalities into a shared embedding space to achieve more effective information fusion.

Subsequently, I employed a metric learning approach using Cross-modal Feature-Ratio Loss, Cross-modal Feature-Margin Loss, and Single-modal Feature-Ratio Loss to optimize the encoder. This optimization enables the encoder to learn the underlying shared embedding space of the data and the corresponding label space of its modalities. The introduced loss functions ensure the encoder extracts meaningful representations of correlations between modal data in the shared embedding space.

The encoded data features are then processed by a decoder consisting of three fully connected layers to handle different tasks, outputting similarity scores and their respective continuous emotional labels. To optimize these backend tasks, I also leverage MSE Loss to ensure that the decoder can perform its task efficiently and generate predictions that are as close to the true labels as possible.

While conducting cross-modal experiments, I also separated the model structure

according to the data of different modalities, using the same encoder and decoder structure. They constitute the image branch and the music branch respectively, taking single-modal data as input and using MSE loss for optimization. The final objective is to predict the valence-arousal emotional label of single-modal data.

Chapter 4

EXPERIMENT AND SETUP

4.1 Experiment Details

In this section, I will discuss in depth the experimental part of the thesis work. This includes an introduction to the development environment, hyperparameter settings, training procedures, and evaluation metrics. This section aims to provide a transparent and thorough understanding of the experimental setup for subsequent analysis and interpretation of results. Then I also drew and analyzed the loss curve in the experiment to explore the effect of training.

4.1.1 Development Environment

This experiment is written in Python with version 3.9 and uses the PyTorch deep learning framework. For the extraction of parameters of CNN-based pre-trained models, the Torchvision library provides a rich set of computer vision operations and pre-trained models. For Transformer-based pre-training models, the Hugging Face Transformers library provides corresponding resources. During the testing phase, I used the NumPy and SciPy libraries to perform statistics on performance results and combined the Scikit-Learn and Matplotlib libraries to visualize the results.

In terms of data feature extraction, as mentioned earlier, I used Librosa and Torchaudio to extract two versions of music features for their respective encoders. As for image features, I used Pillow to extract features. Table 4.1 summarizes the main libraries used in the experiment and their corresponding versions.

Library	Version
Pytorch	2.1.0
Torchvision	0.16.0
Transformers	4.34.1
NumPy	1.26.1
SciPy	1.11.3
Scikit-Learn	1.3.2
matplotlib	3.8.0
librosa	0.10.1
Torchaudio	2.1.0
Pillow	10.1.0

Table 4.1: Development environment

Since multiple deep networks need to be trained simultaneously in the experiment, a large amount of computing resources are required. The training phase of the experiment was initially conducted on Google Colab, which provides users with free Tesla T4 GPUs and access to NVIDIA Tesla A100 GPUs on a paid basis. However, even in the paid version of Colab, stability issues persist, with frequent disconnections from the virtual environment causing model training progress to be lost. The Links Foundation then provided me with an NVIDIA RTX 3090 GPU. Its performance is slightly better than the Tesla T4 GPU and supports long-term stable training, allowing my experiment to be completed. Next, I will introduce the training process and hyperparameters setup in the experiment.

4.1.2 Hyper-parameter Setup

The effectiveness of model training critically depends on the choice of hyperparameters. Striking a balance and carefully tuning these parameters is critical to achieving optimal performance across a range of tasks. This requires careful consideration of nuances in task characteristics, dataset properties, and chosen model architecture. Systematic exploration of hyper-parameter configurations ensures a comprehensive understanding of their impact on model learning dynamics and ultimate generalization capabilities.

As described in section 3, my experimental framework consists of two basic components: the training of a multi-modal model for multiple tasks and the training of a single-modal model for valence-arousal label prediction. Throughout the experiment, I considered the following hyper-parameters for each experiment and kept them as consistent as possible for subsequent comparison.

First, choosing an appropriate optimizer and learning rate is a pivotal decision in the training step of deep learning models, as it directly influences the model's

convergence speed and performance. The optimizer is responsible for adjusting model parameters to minimize the loss function, while the learning rate determines the step size of each parameter update.

The choice of optimizer can significantly affect the performance of the model. Commonly used optimizers include Stochastic Gradient Descent (SGD), SGD with Momentum, and Adam. Each optimizer has unique advantages and is suitable for specific scenarios. SGD, as a basic optimizer, may exhibit slow convergence in some cases. By introducing a momentum term in SGD speeds up convergence and provides greater stability when dealing with noisy gradients. Adam combines momentum and adaptive learning rate features, which may have better performance in some specific scenarios.

Equally important is the choice of learning rate. A learning rate that is too large will cause the model to oscillate in the loss function space or even fail to converge. Conversely, a learning rate that is too small may result in slow convergence or oscillation around a local minimum. A common approach is to implement a learning rate scheduling strategy that gradually reduces the learning rate during training to maintain balance. The choice of learning rate usually requires experimenting on the validation set and monitoring the performance of the model.

On the other hand, when choosing a batch size in the data loader, you must carefully consider its impact on model performance. The batch size directly determines the number of samples processed each time the model parameters are updated. Choosing a larger batch size can increase training speed by allowing more data to be processed in parallel. However, it can also lead to increased memory pressure, especially if the GPU memory is limited. On the other hand, smaller batch sizes may reduce memory requirements but may also slow down training.

In practice, adjusting the batch size can observe the convergence speed and generalization performance of the model. Larger batches may lead to faster convergence, but they may carry the risk of overfitting. Smaller batches can improve model robustness, but training may require more iterations.

Unimodal Model Setup

For the unimodal model dedicated to valence-arousal label prediction in image and music data, I set up both models using the same hyperparameters. In terms of optimization, I chose stochastic gradient descent (SGD) with a momentum of 0.9 to enhance the model's ability to explore the parameter space more effectively.

Given that the model is optimized using only mean squared error (MSE) loss and the initial loss values fall within a relatively small range, I set the learning rate to 1e-3, which is relatively high. To address potential non-convergence issues associated with large learning rates in SGD, I integrated the stepLR scheduler during training, which is configured to reduce the learning rate by a factor of 10

every ten epochs.

In terms of batch size, I used two different sizes depending on whether the feature extractor in the encoder was trained or not. When I choose to freeze the feature extractor and train only the fully connected layers in the encoder, I set the batch size to 128. However, as I continued training all the parameters in the encoder, the demand for GPU memory increased. Therefore, I had to reduce the batch size to 64 to ensure that the experiment ran smoothly.

Multi-modal Model Setup

As mentioned earlier, for the sake of facilitating subsequent comparative experiments, I try to align the training of multi-modal models with the unimodal models by using similar hyperparameters. Likewise, in the selection of the optimizer, I opted for SGD with a momentum value of 0.9 with an initial learning rate of 1e-3. StepLR is used as a scheduler, which gradually reduces the learning rate as training progresses.

The situation becomes more intricate when it comes to setting the batch size. The multi-modal models necessitate the simultaneous training of four deep neural networks and involve intricate gradient propagation, placing higher demands on hardware resources compared to training single-modal models. While the existing computational devices can handle a batch size of 128 when freezing the parameters of the feature extractor, the scenario changes when training all parameters of the entire framework.

In this context, the CNN-based encoder can only be trained on a data loader with a batch size of 16. On the other hand, due to the greater number of parameters in the transformer architecture compared to ResNet, training the transformer-based encoder is constrained to a batch size of 8. Such a diminutive batch size inevitably results in an extremely low number of samples, significantly impacting the training process.

Loss Hyperparameter

In Chapter 3, the loss functions employed for specifying and training cross-modal data labels have been thoroughly described. However, the author did not provide a detailed report on the parameters required for these functions during training. Specifically, it is necessary to determine the value of the average distance σ_n^m in Formula 3.1, which is used for computing similarity scores, and to select an appropriate maximum margin value α in Formula 3.4. This is crucial to ensure that the encoder can guarantee that the features of different modalities in the shared embedding space do not have excessive distances.

Firstly, regarding the similarity scores, there are some missing images in the experimental dataset I used compared to the one utilized by the author for various

reasons. This discrepancy has resulted in slight variations between the similarity scores provided by the author and those obtained in my experiments. As the calculation of similarity scores is based on the average distance, σ, between all images and the emotional labels of music, I iterated through all images and music in the dataset. Then, I recalculated the value of σ and employed it in the subsequent experiments.

Regarding the parameter α, the author designed it to impose a penalty on the encoder for extracting feature representations with excessively large distances in the shared embedding space. During the experimental process, I conducted distance calculations on the features extracted by the image encoder and music encoder. After careful consideration, I have decided to set α to 15. This choice is aimed at achieving a balance in penalizing the encoder for extracting features with distances that are too large, aligning with the original intent of the parameter in the experimental design.

Table 4.2 shows the value of loss parameters I set during the experiment.

Parameter	**Value**
σ_n^m	0.397071
α	15

Table 4.2: Loss Parameters

4.1.3 Evaluation Metrics

After obtaining the model's predictions for similarity and VA labels, it becomes crucial to evaluate its performance. To achieve this goal, mean square error (MSE) and mean absolute error (MAE) are used as the main evaluation metrics. The formulas for MSE and MAE, shown in equations 4.1 and 4.2 respectively, provide a quantitative assessment. Here, $\hat{l}_i$ represents the predicted value, l_i is the ground truth label, and t represents the number of samples in the test set. These metrics facilitate a comprehensive and objective evaluation of the model in the regression task of predicting similarity and VA labels.

$$MSE = \frac{1}{t}\sum_{i=1}^{t}(l_i - \hat{l}_i)^2 \tag{4.1}$$

$$MAE = \frac{1}{t}\sum_{i=1}^{t}\left|l_i - \hat{l}_i\right| \tag{4.2}$$

These metrics were chosen based on their ability to quantify the difference between predicted and actual values, thus providing a quantitative assessment of

the accuracy of the model. Lower MSE and MAE values indicate closer proximity between predicted and true labels, which means superior model performance.

4.1.4 Experiment Process

To systematically investigate the impact of training the encoder on overall model performance, I employed two distinct approaches. Initially, I took a conservative stance, preserving the parameters of the feature extractor within the encoder while exclusively training the last fully connected layer. Subsequently, I opted for a more extensive training strategy, which involved adjusting the entire encoder, including the feature extractor.

Given the substantial time investment required for training epochs in multimodal models (approximately 3.5 hours per epoch with the transformer encoder on the NVIDIA RTX 3090), I decided to limit all training procedures to a maximum of 60 epochs. Recognizing the challenges posed by extended training periods, such as potential interruptions, I implemented a training resumption mechanism. This mechanism enables my model to seamlessly resume training from any selected checkpoint, ensuring resilience against disruptions.

Throughout the training process, the validation set is used for model evaluation after each epoch. This evaluation produces a validation loss that guides subsequent training cycles and serves as the basis for establishing checkpoints. I save the model's state at the checkpoint corresponding to the epoch with the lowest validation loss. To account for potential later improvements, I also retained the model state from the most recent epoch, even if its validation loss did not reach an absolute minimum. These practices are rooted in empirical considerations, ensuring a methodologically sound and effective training approach.

Additionally, I recorded both training and validation losses for each epoch, laying the foundation for a comprehensive assessment of the model's training effects in subsequent analyses. This documentation provides not only a historical record of the model's progression but also facilitates a detailed retrospective analysis of its performance characteristics across different stages of training.

After the model training is completed, the evaluation of the model performance needs to be tested on a dedicated test set. I performed different evaluations using the model associated with the minimum validation loss and the model for the most recently ended epoch.

While using metrics to evaluate model performance, I used the SciPy library to record and store the specific output of the model systematically. Recording its output results is not only for retention but also to prepare for the visualization of model performance.

A structured approach to testing and recording results is essential to gain reliable insights into a model's generalization and predictive capabilities. In addition to

cold numerical values, this method can also show the performance of the model more comprehensively. Recorded outputs combined with visualization techniques help provide a richer description of the model's efficacy.

4.2 Analysis of Loss Curves

During the process of training the model, I systematically recorded the training and validation losses for each epoch. This approach holds significant importance for gaining in-depth insights into the model's training process, the progress of experiments, and the interpretability of results.

Unimodal Model

To investigate the effectiveness of the training designed for the unimodal model, I recorded the training and validation losses for both the image and music branches.

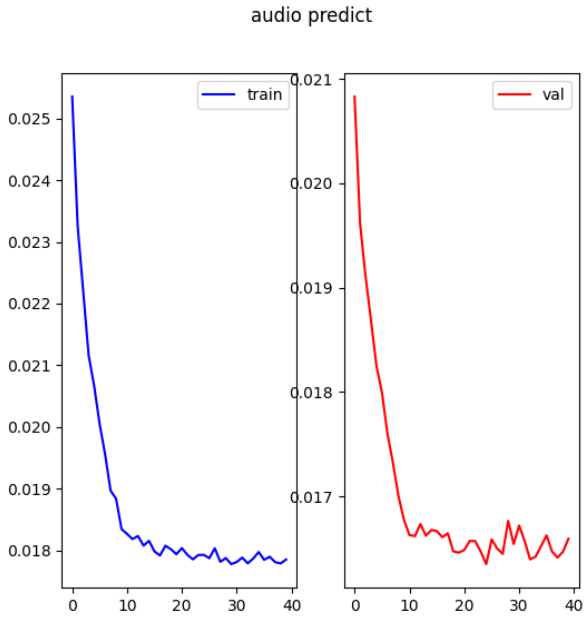
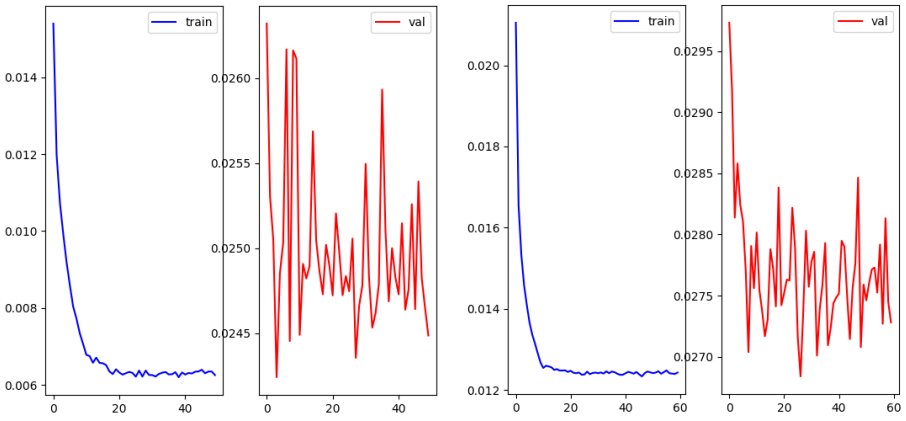


Figure 4.1: CNN-based music encoder training and val loss

Figure 4.1 shows the training and validation losses of the CNN-based music encoder. The training process went smoothly, and the training loss began to converge to around 0.018 after the 20th epoch. Likewise, the validation loss, while exhibiting some oscillations, remains within acceptable limits, around 0.017. This observation demonstrates the effectiveness of our experimental setup. Relatively

stable convergence of training and validation losses means that the model is learning from the data effectively and that the observed fluctuations are within reasonable limits.

Things changed when the experiment progressed to using a transformer-based encoder, as shown in Figure 4.2.



(a) Flatten strategy (b) CLS token embedding strategy

Figure 4.2: Transformer-based music encoder train-val loss

As mentioned in Chapter 3, I implemented two strategies for the input dimension of the fully connected layer aggregated after the transformer-based encoder, with their respective training and validation losses depicted in Figure (a) and Figure (b). It becomes evident that, when using a transformer-based encoder, the model rapidly enters an overfitting state. In the flattening strategy, the model's training loss converges to an extremely low value, approximately at 6×10^{-3}, after 20 epochs. In contrast, its validation loss fluctuates significantly over a large range and fails to converge. This pattern, where the model performs exceptionally well on the training set but fails to converge on the validation set, aligns with the characteristics of overfitting.

The performance of extracting only CLS (class) token embeddings is relatively good. The training loss starts to converge after 20 epochs, while the validation loss gradually slows down after the 8th epoch. Although there are some remaining oscillations, validation losses are reduced by about 7% compared to the initial epoch. This shows an improvement in validation set performance for the extraction-only CLS token embedding strategy.

Regarding the image encoder, the CNN-based encoder also demonstrated a smooth training process, as illustrated in Figure 4.3. The training loss converged to around 0.033 after 20 epochs of model training, while the validation loss converged

to 0.031.

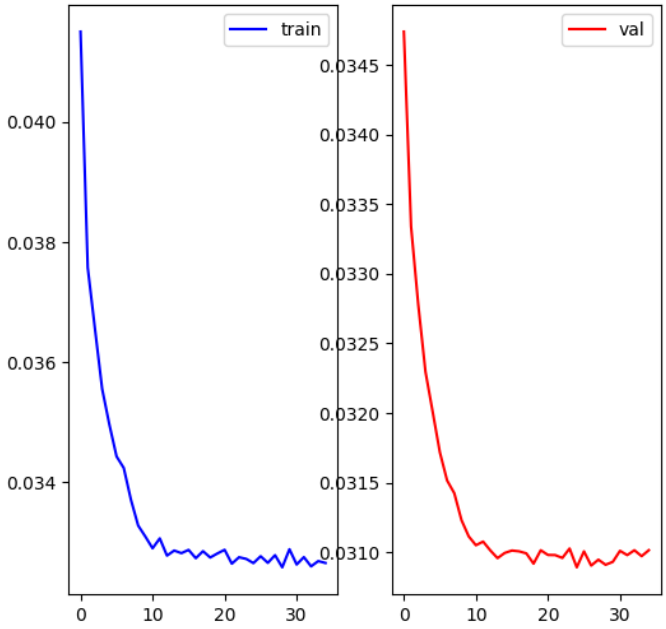
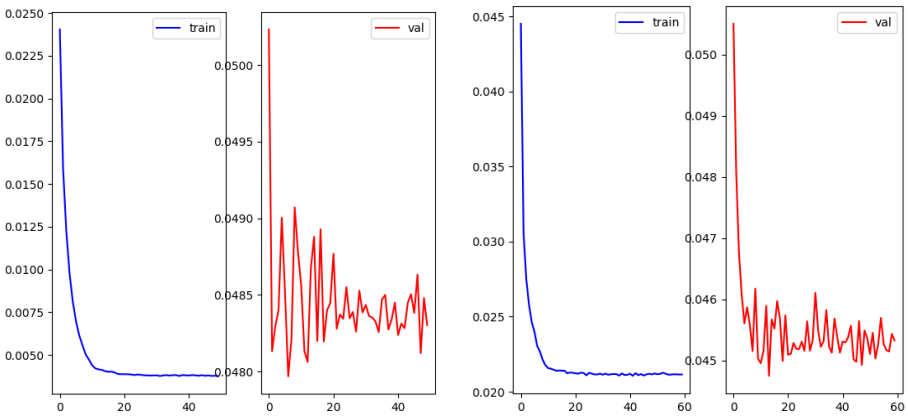


Figure 4.3: CNN-based image encoder training and val loss

When transitioning to the transformer-based image encoder, the situation, compared to the music encoder, shows some improvement but still exhibits signs of overfitting. Firstly, as shown in Figure (a), under the first strategy, the training loss converges to around 5×10^{-3} after approximately 20 epochs. In contrast, the validation loss reaches its minimum in the initial epochs and experiences significant fluctuations. With the second strategy, although the training loss is relatively higher compared to the first, the range of fluctuation in the validation loss substantially decreases, and it remains at a lower value than the first validation loss.

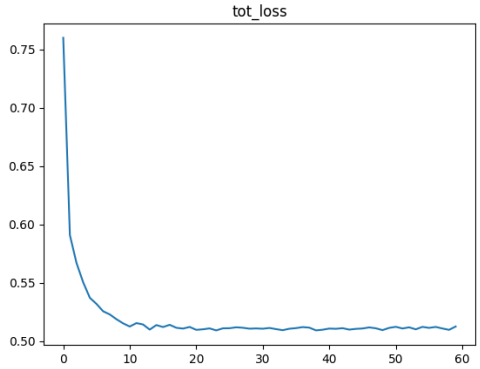


(a) Flatten strategy (b) CLS token embedding strategy

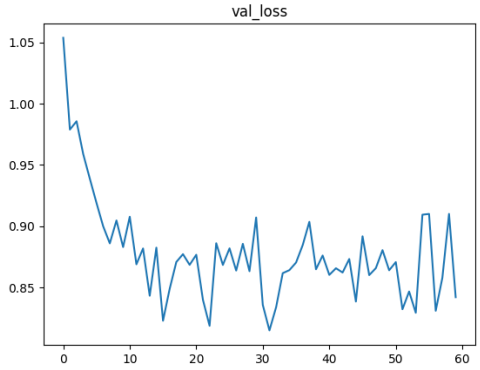
Figure 4.4: Transformer-based music encoder train-val loss

Multi-modal Model

Training of multi-modal models introduces higher complexity compared to single-modal models. The complexity arises from the combination of seven different losses, each designed to fulfill a specific role in the training process. During the initial stages of the experiment, a recurring challenge emerged in the form of exploding gradients. This problem is particularly relevant for L_{CRF} 3.2 and L_{SRF} 3.6 based on the summation operation, which sums the distances of all image features and audio features, resulting in values exceeding 4000. Therefore, the optimization process requires the use of a very small learning rate (approximately 1e-7). To address this challenge, I replaced the summation operation with an averaging operation. Furthermore, to promote numerical stability and facilitate optimization, all losses during the training phase are normalized, ensuring that their values are limited to around 1, allowing experiments to proceed smoothly.



(a) Training Loss



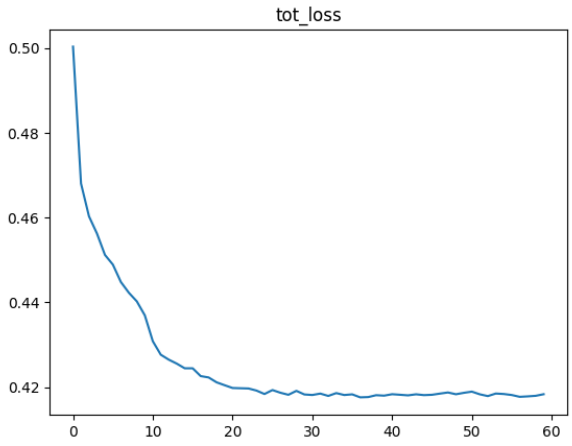
(b) Validation Loss

Figure 4.5: CNN-based encoder, train only fully connected layer

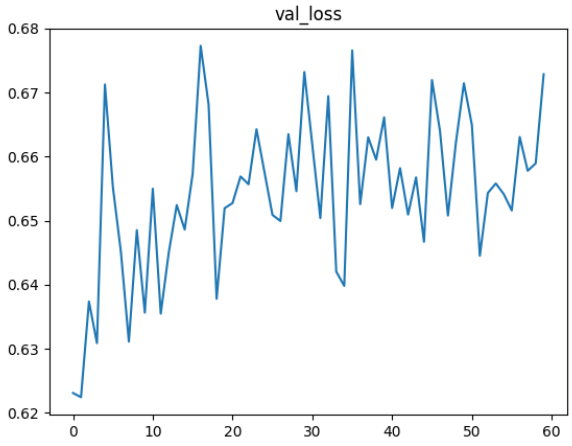
Figure 1 illustrates the training and validation losses when training a fully connected layer using only frozen encoder parameters. The experimental process went relatively smoothly, and the training loss converged from the initial value of 0.75 to about 0.5. At the same time, the validation loss fluctuates around 0.85, a decrease of about 20% compared to the initial value.

Simultaneously, when attempting to train all parameters in the model framework, as depicted in Figure 2, the training loss exhibited a descending trend. However, the magnitude of the reduction was not as pronounced as observed previously. However, the validation loss failed to converge, displaying characteristics indicative of overfitting. It is worth noting that the largest loss in the validation set remained smaller than that observed during training the fully connected layer alone, suggesting the possibility of early convergence in the training process.

In experiments on transitioning to a transformer-based encoder, the first strategy



(a) Training Loss

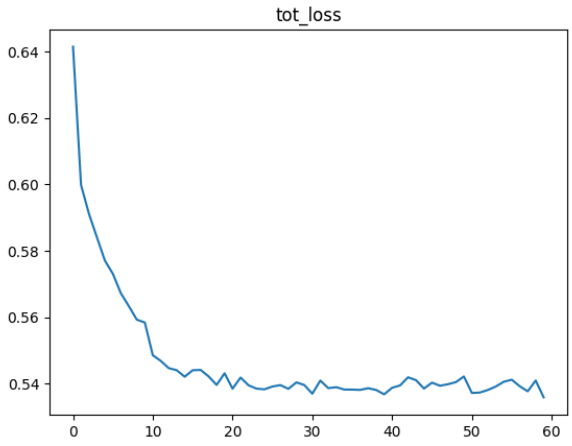


(b) Validation Loss

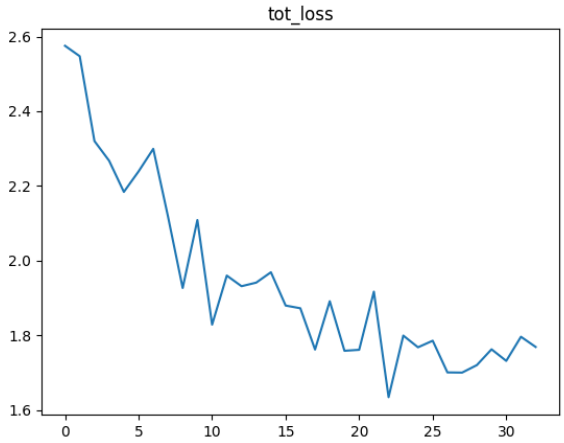
Figure 4.6: CNN-based encoder, train with feature extractor

showed faster convergence than the second strategy, considering the scenario where the feature extractor was not trained. The training loss of the first strategy is relatively low, as figure 4.8, but both strategies show signs of overfitting. Visualization of validation loss is omitted given that it reaches a minimum within the first five epochs.

When training the parameters of the entire framework, due to the large GPU memory consumption, experiments can only be conducted with a batch size of 8. Existing research has proven that the Vision Transformer framework requires a



(a) Flatten strategy Training Loss

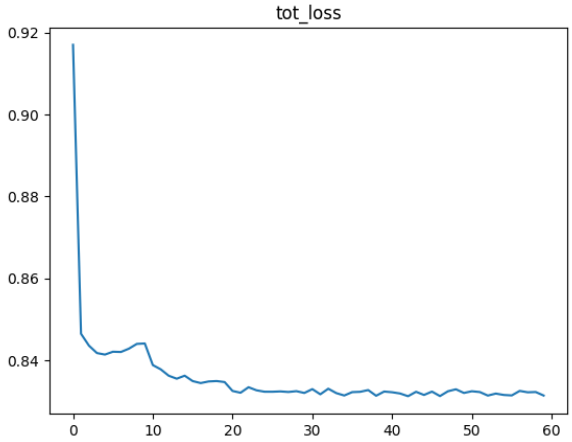


(b) CLS token strategy Training Loss

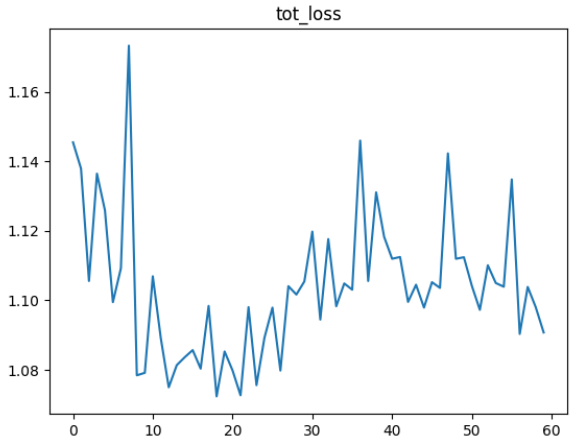
Figure 4.7: Transformer-based encoder, train only fully connected layer

large amount of data to obtain optimal performance, so it has a lower reference value than previous experiments. It is worth noting that when both strategies of the Transformer-based encoder are trained with full parameters, the flattening strategy exhibits a gradient explosion-like phenomenon around the 14th epoch. This causes both training and validation losses to become NaN, and this is consistent across multiple experiments. Preliminary inference is that the model complexity increases due to the high input dimension of the fully connected layer, which may lead to gradient explosion. Figure 4.8 shows the training and validation losses when

training parameters (including those of the feature extractor) under the second strategy.



(a) CLS token strategy Training Loss



(b) CLS token strategy Validation Loss

Figure 4.8: Transformer-based encoder, train with feature extractor

It is evident from the figure that compared to the single-modal case in the transformer-based experiments, a similar phenomenon occurs: the training loss shows a smooth decrease, but the validation loss fails to converge. Considering the early experiments, this occurrence may be attributed to the possibility of the model initiating convergence in the early stages of training.

Chapter 5

RESULTS

In this chapter, I will provide a thorough analysis and comprehensive summary of the experimental results, aiming to extract meaningful insights and illustrate the effectiveness of the experimental approach. I will put forward a comprehensive and objective conclusion through a systematic discussion of various aspects. This will help readers gain a deeper understanding of the performance of the proposed method and provide valuable insights and references for research in related fields.

5.1 Performance Evaluation

5.1.1 Unimodal model

I will initiate the analysis by focusing on the unimodal model. It comprises two branches, each designed to predict valence-arousal labels for images and music, respectively.

Image Branch

The model in the image branch has the same structure as the image encoder and continuous emotion label predictor in the multi-modal model. I conducted experiments using CNN (ResNet50) and transformer-based (ViT) encoders, optimized using Mean Squared Error (MSE) loss. Since training unimodal models does not require extensive computational resources, I trained the entire model in all experiments.

Table 5.1 shows the model's performance in predicting valence arousal labels in the image branch. Among them, I trained and tested two strategies applied to the fully connected layer aggregation after the Transformer encoder. V1 represents the operation of flattening the sequence data, while V2 represents the strategy of only taking the CLS token in the sequence.

Encoder	V MSE	V MAE	A MSE	A MAE
ResNet50	0.023	0.114	0.038	0.156
ViT V1	0.021	0.11	0.036	**0.151**
ViT V2	**0.021**	**0.11**	**0.036**	0.153

Table 5.1: Image valence-arousal prediction on unimodal model

From the table, it is clear that the encoder with Transformer-based architecture shows improvement (around 3%) compared to the CNN-based encoder. This shows that under a well-designed Transformer architecture, its performance can be comparable to or even slightly better than traditional CNN architectures. However, for the ensemble of fully connected layers after the Transformer encoder, both strategies achieved similar results, and it is not possible to judge which strategy performs better based on these results.

Audio Branch

Like the image branch, the music branch underwent training and testing with encoders based on both CNN (ResNet18) and Transformer (BEATs) architectures. Additionally, experiments were conducted with the flattening strategy and the strategy of extracting the CLS token. The results are presented in Table 5.2.

Encoder	V MSE	V MAE	A MSE	A MAE
ResNet18	0.019	0.112	0.013	0.092
BEATs V1	0.02	0.106	0.012	**0.086**
BEATs V2	**0.019**	**0.105**	**0.012**	0.088

Table 5.2: Music valence-arousal prediction on unimodal model

The results show that the encoder based on the transformer architecture achieves better results on the test set and has a non-negligible improvement compared to ResNet performance. This improvement is attributed to BEAT's clever design in processing audio data within the Transformer framework, allowing it to capture feature information that traditional CNNs may ignore, especially temporal information. This enhancement of the Transformer-based music encoder achieves

superior performance in music Emotion recognition tasks. The findings highlight the potential advantages of the Transformer architecture in processing multi-modal data.

5.1.2 Cross-modal

Next, I will show the test results of the multi-modal model. The experiment with the multi-modal model was a multi-task experiment involving predicting similarity scores for image-music pairs while simultaneously predicting arousal-evaluated continuous emotion labels for images and music. In this experiment, I integrated the image and music branches to form the overall model framework and adopted various loss functions for optimization. The experimental results include full parameter training of CNN-based and Transformer-based encoders and training of only parameters of the fully connected layer.

		Similarity score prediction	
	check point	**MSE**	**MAE**
Only train FC	max epoch	0.061	0.209
	min val loss	0.061	0.209
Full-parameter training	max epoch	**0.054**	**0.196**
	min val loss	0.058	0.204

Table 5.3: Similarity prediction on multi-modal model based on CNN-based encoder

Table 5.3 presents the performance of the multi-modal model with a CNN encoder on the task of predicting similarity scores for image-music pairs. It can be observed that the model achieves its optimal performance at the checkpoint of the maximum number of epochs under full-parameter training. This phenomenon indicates that the model still has the potential for improvement in similarity prediction tasks after reaching the minimum validation loss.

Table 2 presents the performance of the model using the Transformer encoder in the image-music similarity prediction task, as mentioned in Chapter 4, when both strategies of the Transformer-based encoder are trained with full parameters, a phenomenon similar to gradient explosion occurs in the training of the flattening

strategy. Therefore, I decided to exclude full parameter training of the flattening strategy. The results show that the best performance is obtained at the maximum epoch checkpoint when only fully connected layer parameters are trained. At the same time, compared with the CNN-based encoder, the Transformer-based encoder showed significant improvement in the similarity prediction task under the same conditions (i.e., only training fully connected layer parameters), showing a performance gain of up to 10%. This highlights the excellent capabilities of the Transformer architecture in cross-modal tasks.

On the other hand, this result demonstrates that, compared to the strategy of only taking the CLS token from the sequence, the flattening strategy, which involves inputting the entire output sequence of the Transformer encoder into the fully connected layer, can capture more information and achieve superior performance during training.

		Similarity score prediction	
	check point	**MSE**	**MAE**
Only train FC	max epoch V1	**0.05**	**0.188**
	min val loss V1	0.052	0.192
	max epoch V2	0.056	0.198
	min val loss V2	0.056	0.198
Full-parameter training	max epoch V2	0.057	0.204
	min val loss V2	0.057	0.204

Table 5.4: Similarity prediction on multi-modal model based on Transformer-based encoder

For the full-parameter training of the Transformer encoder, limitations in device capacity mandated training with a batch size of 8, a significant deviation from the batch size of 128 used in training only the fully connected layer. Nevertheless, even under these circumstances, it exhibits performance comparable to ResNet, showcasing the robust capabilities of Transformer.

When predicting image-music similarity, the cross-modal model is additionally trained to independently predict the valence-arousal labels for both images and

music, aiming to recognize their emotions. The performance is outlined in Table 3:

	Image Emotion recognition				**Music Emotion recognition**			
	V MSE	**V MAE**	**A MSE**	**A MAE**	**V MSE**	**V MAE**	**A MSE**	**A MAE**
CNN-based	**Only train FC**							
Max epoch	0.056	0.176	0.076	0.227	0.03	0.145	0.034	0.15
Min val loss	**0.056**	**0.176**	0.076	0.227	0.03	0.145	0.034	0.15
	Full-parameter training							
Max epoch	0.068	0.196	0.082	0.229	0.03	0.141	0.031	0.147
Min val loss	0.062	0.187	**0.074**	**0.223**	**0.029**	**0.14**	**0.03**	**0.142**
Transformer-based	**Only train FC**							
Max epoch V1	0.059	0.181	0.075	0.223	0.034	0.155	0.034	0.152
Min val loss V1	**0.056**	**0.176**	**0.074**	**0.222**	**0.033**	**0.153**	**0.033**	**0.149**
Max epoch V2	0.06	0.184	0.077	0.226	0.033	0.153	0.035	0.151
Min val loss V2	0.06	0.184	0.077	0.226	0.033	0.153	0.035	0.151
	Full-parameter training							
Max epoch V2	0.067	0.197	0.077	0.228	0.07	0.223	0.046	0.167
Min val loss V2	0.061	0.185	0.077	0.23	0.055	0.197	0.04	0.163

Table 5.5: Valence-Arousal label Prediction on Multi-modal model

As can be seen from the table, for CNN-based encoders, the model trained with full parameters shows better performance. Since training involves optimizing multiple losses simultaneously, the model sacrifices some ability to identify sentiment in single-modal data while enhancing similarity score prediction. This phenomenon could explain why the checkpoint with the best performance in similarity prediction coincides with the final epoch reached, while the best performance for emotion recognition on single-modal data is at the checkpoint where the verification loss is minimal.

Similarly, the Transformer-based encoder exhibits similar patterns in the testing phase for this task, further confirming my hypothesis. The model based on the Transformer architecture achieves performance comparable to the CNN-based model in the single-modal emotion recognition task. Both strategies yield highly similar results, providing evidence to some extent that the training strategies devised

for CNNs may not necessarily be suitable for Transformers. More specifically, it underscores that a tailored approach, including loss functions and training procedures, is essential to unleash the full potential of the Transformer architecture.

Finally, in the experiment involving the Transformer-based model with full-parameter training, the extremely low batch size resulted in comparatively lower test results when compared to other models.

5.2 Results Visualization

In this section, I will perform a visualization of the model performance and the phenomena observed during the experiments. The purpose is to facilitate a comprehensive comparison of the advantages and disadvantages between different encoder architectures and to conduct a thorough analysis of the experimental results.

5.2.1 Visualization of Predict Labels Distribution

To begin, in terms of predicting the valence-arousal labels for the single-modal model, the distribution of predictions on the test set can be visualized by recording the model's predicted labels and corresponding true labels. The distribution of labels for the image branch is illustrated in Figure 5.1, where the orange points represent the true label distribution of the test set, the green points depict the labels predicted by the CNN encoder, and the blue points indicate the labels predicted by the Transformer encoder.

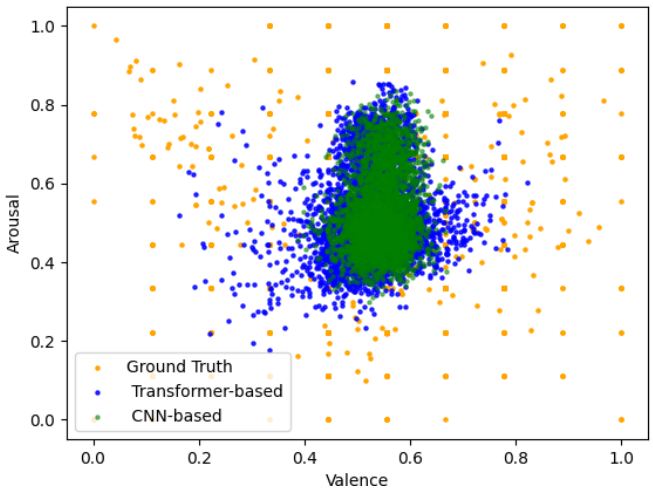


Figure 5.1: Image emotion label distribution

It is evident that both encoders tend to predict labels around the central part of the distribution, indicating a relatively robust strategy. However, when comparing the distribution of predicted labels between encoders, the Transformer architecture produces a sparser distribution of predicted labels compared to the CNN-based distribution. This shows that the Transformer model is more sensitive to data representing different valence-arousal labels and exhibits superior performance.

Figure 5.2 illustrates the distribution of predicted emotion labels for the single-modal model on the music branch. Unlike the image distribution, the ground truth distribution of music emotion labels is denser. This prompts both encoders to adopt relatively bold strategies in prediction, presenting a different distribution than the image labels. This distribution also provides a basis for why the single-modal model performs better in music emotion recognition than in image emotion recognition.

On the other hand, when observing the prediction distributions of the two encoders, a similar phenomenon to the image emotion label distribution appears: the distribution predicted by the Transformer encoder is sparser. This shows that the Transformer encoder also exhibits stronger performance in music emotion recognition tasks.

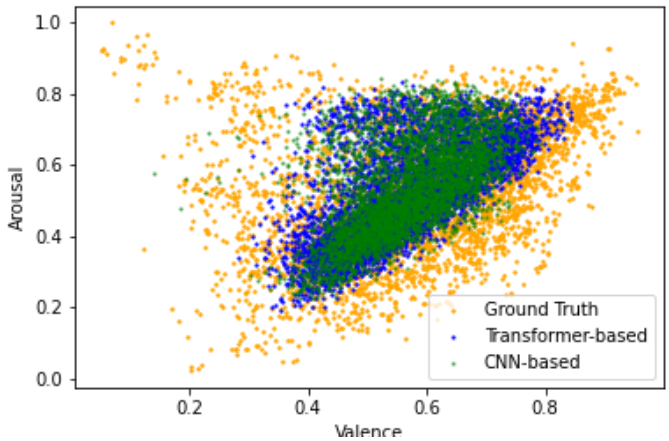


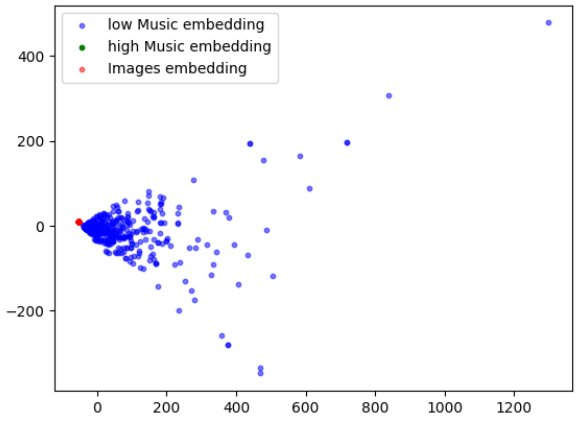
Figure 5.2: Music emotion label distribution

Visual analysis of the distributions generated by these two tasks makes phenomena that are difficult to discern in the table very apparent. It can be seen at a glance that the encoder based on the Transformer architecture shows higher sensitivity to the data features of different emotional labels and superior performance.

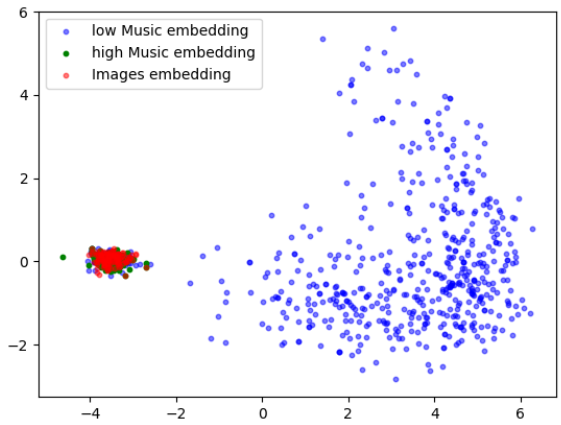
5.2.2 Visualization of Shared Embedding Space

For the similarity score prediction task, considering the model framework designed around metric learning, the main goal is to bring high-similarity image-music pairs

closer while moving away from low-similarity pairs. Visualizing the output of the feature extractor can help provide a deeper understanding of the training dynamics of the model. For the above purpose, I extracted 512 image-music pairs from the test set using both a fully trained model and a zero-shot model (meaning the model was never trained on this dataset). After obtaining the 512-dimensional output features from the respective encoders, I used principal component analysis (PCA) [65] to reduce the dimensionality to 2 dimensions for visualization.



(a) Zero-shot



(b) After-training

Figure 5.3: CNN-based shared embedding space visualization

Notably, I recorded high-similarity image-music pairs (with similarity scores >

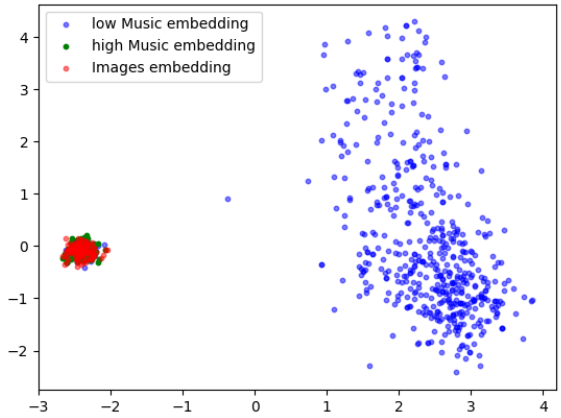
0.5, 257 pairs) and low-similarity pairs (255 pairs), representing them with distinct colors in the visualization to enhance clarity. This approach aims to provide a comprehensive representation of the model's performance across different similarity scenarios.

Figure 5.3 illustrates the shared embedding space when using a CNN-based encoder. In the zero-shot scenario, the embedded features of music and images are distributed in an extremely sparse space (see Figure a). For music with low similarity, the distance from the image in the first dimension is as high as 1200. Due to the pre-training of the image encoder on the ImageNet dataset, the embedded features of the image are relatively concentrated. In such a wide embedding space, the embedded features of images are compressed into single points, which shows that untrained encoders are relatively weak in extracting embedded features from music.

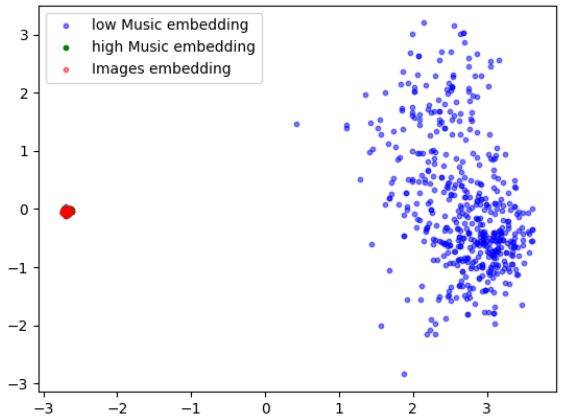
When model training ends, a significant shift in the scene can be observed (see Figure b). First, the distance between images and music in the shared embedding space is placed within an acceptable range. Second, high-similarity music embedding features are placed close to the image, while most low-similarity music embedding features are placed within a certain distance from the image. The emergence of this phenomenon indicates that the model is able to learn the differences and connections between different modal data, providing clear evidence for the validity of the experiment. This change in the embedding space reflects the model's successful acquisition of semantic relationships between images and music throughout the training process, confirming the successful execution of the experiment.

In Figure 5.4, I show the features of the shared embedding space extracted by the Transformer-based encoder. In the zero-shot scenario (Figure a), I observe that the embedded features are relatively evenly distributed, similar to the trained CNN encoder. This trend is partly due to my choice to pre-train the Transformer model on the audio dataset. The self-attention mechanism in the Transformer architecture makes it good at capturing audio-related features to form such a distribution in the shared embedding space.

However, after the model training is completed (Figure b), I observe that the distribution of high-similarity image-music pairs is more compact, and they gradually converge into a point at low-similarity distance scales. This phenomenon may indicate that the model learned to represent similar image-music pairs more closely in the embedding space during training but may also lead to reduced generalization performance. These distribution features provide more information about the effectiveness of model training, further supporting the inference that the Transformer architecture achieves convergence early in training.



(a) Zero-shot



(b) After-training

Figure 5.4: Transformer-based shared embedding space visualization

The visualization of the shared embedding space gave me a deeper understanding of the experimental results. CNN-based encoders show certain limitations when handling cross-modal tasks, but good results can still be obtained with appropriate training. Transformer-based encoders, on the other hand, show superior performance in handling multi-modal data but require more careful training to mitigate the potential risk of overfitting.

5.2.3 Qualitative Results

In order to thoroughly evaluate the performance of the model, qualitative evaluation becomes crucial. Therefore, I designed a mini music-to-image retrieval system specifically for the multi-modal model based on the Transformer encoder. Specifically, I carefully selected three songs of different emotional styles and randomly selected 38 images from Google Images based on categories (e.g., happy, sad, fearful, etc.). I predicted the similarity score of each song to these images separately and selected the most similar image to represent it.

It is worth mentioning that the use of the flattening strategy brings obstacles in the similarity score prediction process. The flattening strategy results in the input size of the fully connected layer after the feature extractor being heavily dependent on the length of the training music (i.e., 2 seconds). It cannot make predictions when faced with music clips longer than 2 seconds. This also shows that the first strategy limits the generalizability of the model.

I designed this approach because it provides a comprehensive evaluation framework, allowing for a more thorough understanding of the model's performance and the discovery of potential shortcomings. This also opens up possibilities for future applications of this multi-modal emotion prediction model.

paganini.mp3. The first piece of music is a track by the violinist Niccolò Paganini, featuring a fresh and cheerful melody that evokes a sense of joy. After using this song to assess the similarity with 38 images, the one with the highest similarity is an image labeled 'joy,' as shown in Figure 5.5. This indicates that our model can recognize different modal data in scenarios characterized by high valence and arousal.

Figure 5.5: The picture with the highest similarity score to Paganini.mp3

Verdi.mp3. The second song is a work by Giuseppe Verdi, with an impassioned melody and strong drum sounds in many places. The arousal experience is high, but the valence is relatively low. The model predicted an image labeled "angry"

that represented the song, a similar feeling I had while listening to the song, as shown in Figure 5.6. This shows that the model can correctly identify multi-modal data at high arousal and low valence levels.

Figure 5.6: The picture with the highest similarity score to Verdi.mp3

wagner.mp3. The third piece comes from a segment of the symphony "Cavalcata delle Valchirie" composed by Richard Wagner. In this segment, the violin performance creates a tense atmosphere with its dense tones, conveying a sense of fear and anticipation. The model predicted images associated with the label 'fear,' as illustrated in Figure 5.7

Figure 5.7: The picture with the highest similarity score to wagner.mp3

This qualitative analysis demonstrates the excellent performance of the multimodal model in capturing different musical emotions. The model successfully associated different emotional styles of music with their corresponding images and showed consistency in its predictions. The model demonstrated the ability to understand a wide range of emotions, from lightheartedness to enthusiasm and excitement, to nervousness and fear, providing strong support for its reliability in emotion recognition tasks.

The consistent performance highlights the model's versatility and adaptability in cross-modal emotion recognition. The success of the qualitative study emphasized the model's deep understanding of musical emotions and laid the foundation for further research.

Chapter 6

CONCLUSIONS AND FUTURE WORK

This chapter is a summary of my thesis work after this long research journey.

Initially, I will provide an overview of the entire research process, emphasizing crucial findings and contributions. Following this, a comprehensive analysis of the experimental results will be conducted, offering insights gained and identifying potential areas for enhancement.

Subsequently, I will delve into exploring prospective avenues for future research, providing readers with potential advancements in the field.

6.1 Conclusion for the emotion prediction model

The goal of my thesis work is to build a framework for emotion recognition in multimodal data. Throughout the process, I experimented with encoders based on CNN and Transformer architectures, aiming to analyze their strengths and weaknesses through experiments.

Experimental results show that Transformer-based encoders outperform CNN-based encoders in emotion recognition tasks on unimodal data, especially in the case of music data. This phenomenon is attributed to the attention mechanism of the Transformer architecture and the researchers' clever handling of non-natural language data. The former enables the model to capture temporal features in music data, resulting in more effective predictions. The latter allows audio data to be used as input to the Transformer architecture, allowing pre-training on large-scale audio datasets. Compared to CNN architectures which focus on image processing, encoders that are pre-trained on audio are better at extracting audio features, resulting in superior performance.

On the image emotion prediction task, both encoders achieved similar performance. However, subsequent analysis of label distribution visualizations shows that the Transformer-based encoder still outperforms the CNN architecture to some extent. This confirms that the Transformer architecture can compete with traditional CNN architectures in the image processing task.

When the experiments extended to cross-modal tasks, specifically the task of predicting similarity scores for image-music pairs, the performance gap between these two encoders widened. Visualizations of the shared embedding space reveal that the untrained Transformer encoder exhibits comparable capabilities in extracting features from multimodal data as the CNN encoder trained on relevant datasets. After training, although sacrificing some performance in unimodal emotion label prediction, the Transformer encoder demonstrates a non-negligible improvement in cross-modal tasks compared to CNN.

This indicates that while the Transformer architecture is originally designed for tasks like natural language processing, it shows considerable potential when dealing with multimodal data. Multimodal models based on the Transformer architecture have now become a prominent direction in current research, demonstrating its adaptability and effectiveness beyond its initial design scope.

When using a transformer-based encoder, the design of its training is critical. Although the current training has improved its performance compared to the CNN encoder. However, during the training process, the phenomenon that it begins to converge in the early stages of training, even before completing an epoch, shows that there is still room for improvement in the design of its experiments.

At the same time, a large number of studies have shown that image encoders based on transformer architecture, Vision Transformer, usually require more data for effective training because of their large number of parameters, and learning meaningful representations from visual data is very complex and richer. Data can also effectively prevent overfitting and enhance its versatility. This also means that if you want to realize the full potential of the transformer encoder, it will require a higher cost compared to CNN-based encoders. Therefore, different considerations need to be carefully weighed when choosing an encoder architecture.

To optimize the shared embedding space of multiple modal encoders, I devised two strategies for fully connected layers integrated after their feature extractors. During the experiment, each strategy shows its advantages and disadvantages. The flattening strategy enables subsequent fully connected layers to receive all information from the output sequence simultaneously, thus achieving better performance. However, its huge input dimension also leads to a significant increase in model complexity, leading to excessive convergence, and even exhibits the characteristics of gradient explosion when the model framework is trained with all parameters. At the same time, as mentioned before, the input dimension of the fully connected layer after the music encoder is highly dependent on the length of the music. This

limits it to encoding only 2 seconds of music, reducing the generality of the model.

On the other hand, the strategy of only taking CLS tokens limits the information received by the fully connected layer thus limiting its performance. However, this approach does not increase the complexity of the model and allows it to encode music of arbitrary length. Processing of the transformer output requires careful balancing of the effects of different strategies. Only through careful design can the transformer architecture reach its full potential.

At the end of the experiment, a detailed qualitative analysis of the model's results demonstrated its ability to understand emotions conveyed in images and music. When provided with music, the model successfully identified and retrieved images with similar emotions. This achievement highlights the development potential of multi-modal models in the field of emotion recognition and lays a solid foundation for subsequent related research.

6.2 Future work and improvements

For future work, further optimization of the experiment is needed. Currently, experiments with multi-modal models have achieved high performance in cross-modal tasks but at the expense of some performance in unimodal emotion prediction. One potential direction for improvement is to improve and rationally combine loss functions to achieve high performance across all tasks. Furthermore, there is room for improvement in the training of the Transformer-based encoder. The existing strategies have their pros and cons, impacting the entire model architecture to varying extents.

The music image retrieval system I designed may pave the way for future research directions. Using the similarity score as a basis, the system automatically retrieves images or music with similar emotions from the given data. This system opens avenues for various applications such as enhanced content recommendations, emotion-based image retrieval, and personalized emotion-aware systems. Furthermore, exploring the integration of this system into real-world applications could provide valuable insights and contributions to the broader field of affective computing.

In addition, in the field of music emotion recognition, lyrics are an important element for creators to express their emotions. Moreover, lyric data has natural language attributes, and through appropriate preprocessing, the performance of Transformer-based encoders can be enhanced. On this basis, the three-modal (image, text, music) emotion prediction model has great potential for further development.

Last but not least, the direct expression of human emotions is still based on language. Therefore, the mapping of valence-arousal labels to emotion categories

becomes crucial. Previous research in this area has been undertaken, but the results have not always been satisfactory. Establishing an effective mapping of valence-arousal labels to emotion categories at different scales can significantly reduce the cost of experiments and facilitate the integration of these two different emotion modeling approaches. Moreover, it can enhance the efficiency of human-computer interaction, allowing computers to more effectively recognize human emotions, making them more human-like.

Bibliography

[1] Kate Hevner. «The affective character of the major and minor modes in music». In: *The American Journal of Psychology* 47.1 (1935), pp. 103–118 (cit. on p. 5).

[2] Paul R Farnsworth. «A study of the Hevner adjective list». In: *The Journal of Aesthetics and Art Criticism* 13.1 (1954), pp. 97–103 (cit. on p. 5).

[3] Emery Schubert. «Update of the Hevner adjective checklist». In: *Perceptual and motor skills* 96.3_suppl (2003), pp. 1117–1122 (cit. on p. 5).

[4] Patrik N Juslin and Petri Laukka. «Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening». In: *Journal of new music research* 33.3 (2004), pp. 217–238 (cit. on p. 5).

[5] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. «Context based emotion recognition using emotic dataset». In: *IEEE transactions on pattern analysis and machine intelligence* 42.11 (2019), pp. 2755–2766 (cit. on pp. 5, 20).

[6] Yi-Hsuan Yang and Homer H Chen. «Machine recognition of music emotion: A review». In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 3.3 (2012), pp. 1–30 (cit. on p. 6).

[7] James A Russell. «A circumplex model of affect.» In: *Journal of personality and social psychology* 39.6 (1980), p. 1161 (cit. on p. 6).

[8] Alf Gabrielsson and Erik Lindström. «The influence of musical structure on emotional expression.» In: (2001) (cit. on p. 7).

[9] Mathieu Barthet, György Fazekas, and Mark Sandler. «Music emotion recognition: From content-to context-based models». In: *From Sounds to Music and Emotions: 9th International Symposium, CMMR 2012, London, UK, June 19-22, 2012, Revised Selected Papers 9*. Springer. 2013, pp. 228–252 (cit. on p. 7).

[10] Peilin Chen, Lei Zhao, Zongyu Xin, Yumeng Qiang, Ming Zhang, and Tiemeng Li. «A scheme of MIDI music emotion classification based on fuzzy theme extraction and neural network». In: *2016 12th International Conference on Computational Intelligence and Security (CIS)*. IEEE. 2016, pp. 323–326 (cit. on p. 7).

[11] Hui He, Jianming Jin, Yuhong Xiong, Bo Chen, Wu Sun, and Ling Zhao. «Language feature mining for music emotion classification via supervised learning from lyrics». In: *Advances in Computation and Intelligence: Third International Symposium, ISICA 2008 Wuhan, China, December 19-21, 2008 Proceedings 3*. Springer. 2008, pp. 426–435 (cit. on p. 8).

[12] Xing Wang, Xiaoou Chen, Deshun Yang, and Yuqian Wu. «Music Emotion Classification of Chinese Songs based on Lyrics Using TF* IDF and Rhyme.» In: *ISMIR*. 2011, pp. 765–770 (cit. on p. 8).

[13] Cyril Laurier, Jens Grivolla, and Perfecto Herrera. «Multimodal music mood classification using audio and lyrics». In: *2008 seventh international conference on machine learning and applications*. IEEE. 2008, pp. 688–693 (cit. on p. 8).

[14] Donghong Han, Yanru Kong, Jiayi Han, and Guoren Wang. «A survey of music emotion recognition». In: *Frontiers of Computer Science* 16.6 (2022), p. 166335 (cit. on p. 8).

[15] Norberto Eiji Nawa, Daniel E Callan, Parham Mokhtari, Hiroshi Ando, and John Iversen. «Decoding music-induced experienced emotions using functional magnetic resonance imaging-Preliminary results». In: *2018 international joint conference on neural networks (ijcnn)*. IEEE. 2018, pp. 1–7 (cit. on p. 8).

[16] Panayu Keelawat, Nattapong Thammasan, Boonserm Kijsirikul, and Masayuki Numao. «Subject-independent emotion recognition during music listening based on EEG using deep convolutional neural networks». In: *2019 IEEE 15th International Colloquium on Signal Processing & Its Applications (CSPA)*. IEEE. 2019, pp. 21–26 (cit. on p. 8).

[17] Tao Li and Mitsunori Ogihara. «Detecting emotion in music». In: (2003) (cit. on p. 8).

[18] Chia Chu Liu, Yi Hsuan Yang, Ping Hao Wu, and Homer Chen. «Detecting and Classifying Emotion in Popular Music». In: *Proceedings of the 9th Joint International Conference on Information Sciences (JCIS-06)*. Atlantis Press, 2006/10. ISBN: 978-90-78677-01-7. DOI: 10.2991/jcis.2006.325. URL: https://doi.org/10.2991/jcis.2006.325 (cit. on p. 8).

[19] Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H Chen. «A regression approach to music emotion recognition». In: *IEEE Transactions on audio, speech, and language processing* 16.2 (2008), pp. 448–457 (cit. on p. 9).

[20] Xin Liu, Qingcai Chen, Xiangping Wu, Yan Liu, and Yang Liu. «CNN based music emotion classification». In: *arXiv preprint arXiv:1704.05665* (2017) (cit. on p. 9).

[21] Pei-Tse Yang, Shih-Ming Kuang, Chia-Chun Wu, and Jia-Lien Hsu. «Predicting music emotion by using convolutional neural network». In: *International Conference on Human-Computer Interaction*. Springer. 2020, pp. 266–275 (cit. on p. 9).

[22] Yizhuo Dong, Xinyu Yang, Xi Zhao, and Juan Li. «Bidirectional convolutional recurrent sparse network (BCRSN): an efficient model for music emotion recognition». In: *IEEE Transactions on Multimedia* 21.12 (2019), pp. 3150–3163 (cit. on p. 9).

[23] Moyuan Huang, Wenge Rong, Tom Arjannikov, Nan Jiang, and Zhang Xiong. «Bi-modal deep Boltzmann machine based musical emotion classification». In: *Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25*. Springer. 2016, pp. 199–207 (cit. on p. 9).

[24] Felix Weninger, Florian Eyben, and Björn Schuller. «On-line continuous-time music mood regression with deep recurrent neural networks». In: *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2014, pp. 5412–5416 (cit. on p. 9).

[25] Sanga Chaki, Pranjal Doshi, Priyadarshi Patnaik, and Sourangshu Bhattacharya. «Attentive RNNs for Continuous-time Emotion Prediction in Music Clips.» In: *AffCon@ AAAI*. 2020, pp. 36–46 (cit. on p. 9).

[26] Ye Ma, Xinxing Li, Mingxing Xu, Jia Jia, and Lianhong Cai. «Multi-scale context based attention for dynamic music emotion prediction». In: *Proceedings of the 25th ACM international conference on Multimedia*. 2017, pp. 1443–1450 (cit. on p. 9).

[27] Meixian Zhang, Yonghua Zhu, Ning Ge, Yunwen Zhu, Tianyu Feng, and Wenjun Zhang. «Attention-based joint feature extraction model for static music emotion classification». In: *2021 14th International Symposium on Computational Intelligence and Design (ISCID)*. IEEE. 2021, pp. 291–296 (cit. on p. 9).

[28] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. «Building a large scale dataset for image emotion recognition: The fine print and the benchmark». In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 30. 1. 2016 (cit. on p. 9).

[29] Sicheng Zhao, Guiguang Ding, Qingming Huang, Tat-Seng Chua, Björn Schuller, and Kurt Keutzer. «Affective image content analysis: A comprehensive survey». In: (2018) (cit. on p. 10).

[30] Wei Zhang, Xuanyu He, and Weizhi Lu. «Exploring discriminative representations for image emotion recognition with CNNs». In: *IEEE Transactions on Multimedia* 22.2 (2019), pp. 515–523 (cit. on p. 10).

[31] Afsheen Rafaqat Ali, Usman Shahid, Mohsen Ali, and Jeffrey Ho. «High-level concepts for affective understanding of images». In: *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE. 2017, pp. 679–687 (cit. on p. 10).

[32] Sicheng Zhao, Guiguang Ding, Yue Gao, and Jungong Han. «Learning visual emotion distributions via multi-modal features fusion». In: *Proceedings of the 25th ACM international conference on Multimedia*. 2017, pp. 369–377 (cit. on p. 10).

[33] Sicheng Zhao, Zizhou Jia, Hui Chen, Leida Li, Guiguang Ding, and Kurt Keutzer. «PDANet: Polarity-consistent deep attention network for fine-grained visual emotion regression». In: *Proceedings of the 27th ACM international conference on multimedia*. 2019, pp. 192–201 (cit. on p. 10).

[34] Xingxu Yao, Dongyu She, Sicheng Zhao, Jie Liang, Yu-Kun Lai, and Jufeng Yang. «Attention-aware polarity sensitive embedding for affective image retrieval». In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 1140–1150 (cit. on p. 10).

[35] Alexey Dosovitskiy et al. «An image is worth 16x16 words: Transformers for image recognition at scale». In: *arXiv preprint arXiv:2010.11929* (2020) (cit. on pp. 11, 30).

[36] Mahmut Kaya and Hasan Şakir Bilge. «Deep metric learning: A survey». In: *Symmetry* 11.9 (2019), p. 1066 (cit. on p. 11).

[37] Yueqi Duan, Jiwen Lu, Jianjiang Feng, and Jie Zhou. «Deep localized metric learning». In: *IEEE Transactions on Circuits and Systems for Video Technology* 28.10 (2017), pp. 2644–2656 (cit. on pp. 11, 30).

[38] Vivek Sivaraman Narayanaswamy, Jayaraman J Thiagarajan, Huan Song, and Andreas Spanias. «Designing an effective metric learning pipeline for speaker diarization». In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 5806–5810 (cit. on p. 11).

[39] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. «Signature verification using a" siamese" time delay neural network». In: *Advances in neural information processing systems* 6 (1993) (cit. on p. 12).

[40] Elad Hoffer and Nir Ailon. «Deep metric learning using triplet network». In: *Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*. Springer. 2015, pp. 84–92 (cit. on p. 12).

[41] Joonseok Lee, Sami Abu-El-Haija, Balakrishnan Varadarajan, and Apostol Natsev. «Collaborative deep metric learning for video understanding». In: *Proceedings of the 24th ACM SIGKDD International conference on knowledge discovery & data mining*. 2018, pp. 481–490 (cit. on p. 12).

[42] Xing Xu, Li He, Huimin Lu, Lianli Gao, and Yanli Ji. «Deep adversarial metric learning for cross-modal retrieval». In: *World Wide Web* 22 (2019), pp. 657–672 (cit. on p. 12).

[43] Aihua Zheng, Menglan Hu, Bo Jiang, Yan Huang, Yan Yan, and Bin Luo. «Adversarial-metric learning for audio-visual cross-modal matching». In: *IEEE Transactions on Multimedia* 24 (2021), pp. 338–351 (cit. on p. 12).

[44] Relja Arandjelovic and Andrew Zisserman. «Look, listen and learn». In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 609–617 (cit. on p. 13).

[45] Sicheng Zhao, Yaxian Li, Xingxu Yao, Weizhi Nie, Pengfei Xu, Jufeng Yang, and Kurt Keutzer. «Emotion-based end-to-end matching between image and music in valence-arousal space». In: *Proceedings of the 28th ACM international conference on multimedia*. 2020, pp. 2945–2954 (cit. on pp. 13, 15, 16).

[46] Sicheng Zhao, Yunsheng Ma, Yang Gu, Jufeng Yang, Tengfei Xing, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. «An end-to-end visual-audio attention network for emotion recognition in user-generated videos». In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 01. 2020, pp. 303–311 (cit. on p. 13).

[47] Donghuo Zeng, Yi Yu, and Keizo Oyama. «Audio-visual embedding for cross-modal music video retrieval through supervised deep CCA». In: *2018 IEEE International Symposium on Multimedia (ISM)*. IEEE. 2018, pp. 143–150 (cit. on p. 13).

[48] Rushabh Chheda, Dhruv Bohara, Rishikesh Shetty, Siddharth Trivedi, and Ruhina Karani. «Music recommendation based on affective image content analysis». In: *Procedia Computer Science* 218 (2023), pp. 383–392 (cit. on p. 13).

[49] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. «Deep supervised cross-modal retrieval». In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 10394–10403 (cit. on p. 13).

[50] Alec Radford et al. «Learning transferable visual models from natural language supervision». In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763 (cit. on p. 13).

[51] Jian Huang, Jianhua Tao, Bin Liu, Zheng Lian, and Mingyue Niu. «Multimodal transformer fusion for continuous emotion recognition». In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 3507–3511 (cit. on p. 14).

[52] Baijun Xie, Mariia Sidulova, and Chung Hyuk Park. «Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion». In: *Sensors* 21.14 (2021), p. 4913 (cit. on p. 14).

[53] Zheng Lian, Bin Liu, and Jianhua Tao. «CTNet: Conversational transformer network for emotion recognition». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 985–1000 (cit. on p. 14).

[54] Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al. *International affective picture system (IAPS): Affective ratings of pictures and instruction manual*. NIMH, Center for the Study of Emotion & Attention Gainesville, FL, 2005 (cit. on p. 18).

[55] Artur Marchewka, Łukasz Żurawski, Katarzyna Jednoróg, and Anna Grabowska. «The Nencki Affective Picture System (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database». In: *Behavior research methods* 46 (2014), pp. 596–610 (cit. on pp. 19, 20).

[56] Monika Riegel et al. «Characterization of the Nencki Affective Picture System by discrete emotional categories (NAPS BE)». In: *Behavior research methods* 48 (2016), pp. 600–612 (cit. on p. 19).

[57] Małgorzata Wierzba, Monika Riegel, Anna Pucz, Zuzanna Leśniewska, Wojciech Łukasz Dragan, Mateusz Gola, Katarzyna Jednoróg, and Artur Marchewka. «Erotic subset for the Nencki Affective Picture System (NAPS ERO): cross-sexual comparison study». In: *Frontiers in psychology* 6 (2015), p. 1336 (cit. on p. 19).

[58] Jarosław M Michałowski, Dawid Droździel, Jacek Matuszewski, Wojtek Koziejowski, Katarzyna Jednoróg, and Artur Marchewka. «The Set of Fear Inducing Pictures (SFIP): Development and validation in fearful and non-fearful individuals». In: *Behavior Research Methods* 49 (2017), pp. 1407–1419 (cit. on p. 19).

[59] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. «EMOTIC: Emotions in Context dataset». In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2017, pp. 61–69 (cit. on p. 20).

[60] Anna Alajanki, Yi-Hsuan Yang, and Mohammad Soleymani. «Benchmarking music emotion recognition systems». In: *PLOS ONE* (2016). under review (cit. on p. 23).

[61] Rachel M Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello. «Medleydb: A multitrack dataset for annotation-intensive mir research.» In: *ISMIR*. Vol. 14. 2014, pp. 155–160 (cit. on p. 23).

[62] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. «Recent developments in opensmile, the munich open-source multimedia feature extractor». In: *Proceedings of the 21st ACM international conference on Multimedia*. 2013, pp. 835–838 (cit. on p. 24).

[63] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. «Beats: Audio pre-training with acoustic tokenizers». In: *arXiv preprint arXiv:2212.09058* (2022) (cit. on p. 31).

[64] Sungyeon Kim, Minkyo Seo, Ivan Laptev, Minsu Cho, and Suha Kwak. «Deep metric learning beyond binary supervision». In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 2288–2297 (cit. on p. 33).

[65] Rasmus Bro and Age K Smilde. «Principal component analysis». In: *Analytical methods* 6.9 (2014), pp. 2812–2831 (cit. on p. 60).