Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 8, No. 1, 2024

Improved Multi-attention Neural Networks for Image Emotion Regression and the Initial Introduction of CAPS

Rending Wang [a], Dongmei Ma *

School of Physics & Electronic Engineering, Northwest Normal University, Lanzhou Gansu, 730070, China
*** Corresponding author:** Dongmei Ma (Email: madongmei@nwnu.edu.cn), [a] 819641724@qq.com

Abstract: Image sentiment analysis is a large class of tasks for classifying or regressing images containing emotional stimuli, and it is believed in psychological research that different groups produce different emotions for the same stimuli. In order to study the influence of cultural background on image sentiment analysis, it is necessary to introduce a dataset of image sentiment stimuli that can represent cultural groups. In this paper, we introduce the Chinese Affective Picture System (CAPS), which represents Chinese culture, and revise and test this dataset. The PDANet model has the best performance among the current image sentiment regression models, but due to the difficulty of extracting cross-channel information from the attention module it uses, image long-distance information correlation and other shortcomings, this paper proposes an image emotion regression multiple attention networks, introduces the SimAM attention mechanism, and improves the loss function to make it more consistent with the psychological theory, and proposes a 10-fold cross-validation for CAPS. The network achieves MSE=0.0188, R^2=0.359 on IAPS, and MSE=0.0169, R^2=0.463 on NAPS, which is better than PDANet; the best training result of CAPS is MSE=0.0083, R^2=0.625, and the paired-sample t-test of the results shows that all the three dimensions are significantly positively correlated, with correlation coefficients r=0.942, 0.895 and 0.943, respectively, showing good internal consistency and excellent application prospect of CAPS.

Keywords: Affective Image Content Analysis; CAPS; SimAM.

1. Introduction

The goal of affective image content analysis (AICA) is to understand semantic information at the cognitive level and to recognize the emotions that an image will induce for a specific viewer or most people. In the field of image sentiment analysis regression, PDANet has the best overall performance in terms of performance metrics, but the SeNet attention mechanism and spatial attention mechanism used by the model have the problems of dimensionality loss of information and difficulty in extracting spatial relations over long distances, and the dichotomous polarity judgement used in its loss function also has room for further optimization.

In the current research in the field of image emotion, the mainstream datasets are all based on the collection of foreign social platforms and picture websites, and the annotators are also more internationalized. In the field of psychology, relevant studies have confirmed that factors such as culture, environment, and even religion can greatly affect people's responses to emotional stimuli, and surveys on the use of IAPS have pointed out that the emotion ratings of subjects from different countries differ significantly from their original data. This means that projects using existing image-emotion datasets are bound to be biased and ineffective in their final application on a country-specific level, such as in a Chinese population or Chinese cultural environment. Therefore, the AICA field needs to introduce image-emotion datasets that are more reflective of local Chinese reality.

The Chinese Affective Picture System (CAPS), which is a localized and standardized set of emotionally stimulating picture systems modeled after IAPS by Luo Yuejia (2005) and others, uses the VAD model to label data and contains 852 pictures with oriental characteristics. Compared with IAPS, it is more suitable for use in Chinese culture and has been widely used in domestic research in the field of psychology.

To address the above problems, the main contributions of this paper are as follows:

(1) Propose an improved image emotion regression multiple attention networks, introduce SimAM attention mechanism based on PDANet model, improve its loss function, and obtain advantages in testing;

(2) Revise and introduce the CAPS dataset, normalize its internal data, study its training technique to prevent overfitting, obtain its theoretical performance, and analyze its training performance based on statistics.

2. Related Work

2.1. Psychological Measures of Emotion

There are two important schools of contemporary emotion theory:

Basic/discrete emotion theory [1] posits that emotions are a number of mutually independent entities, similar to different chemical elements or animal species. Each emotion has a unique adaptive function that distinguishes it from the others, e.g., humans evolved fear to help us escape from the threat of a predator or similar imminent physiological risk. The 4 dimensions of emotion: cognition/appraisal, feelings, physiological responses, and behavior are tightly bound to each other and remain stable and consistent across individuals and cultural groups; basic emotions are universal among humans and manifestations of basic emotions can be seen in other species; basic emotions should have an exclusive and innate expressive system for all humans, communicated through verbal intonation It can be communicated through tone of voice, facial expressions, body movements, and other

behaviors, and it is difficult to be influenced by subjective feelings; in conjunction with the foregoing, such basic emotions should be present in infancy and early childhood, and even earlier.

The core emotion and psychological construct theory [2] suggests that emotions are not discrete, and that there should be some core emotions that are then examined in relation to the rest of the emotions. This approach recognizes emotions in terms of dimensions, where emotions at the same latitude are similar but with different intensities. In addition, the feeling aspects of emotions are primary, rather than cognitive, physiological or behavioral aspects; emotional feelings are best described in terms of continuously varying dimensions, rather than discrete categories; and emotional feelings are described primarily in terms of potency (positive and negative) and degree of arousal or activation.

Based on these ideas, two measures of emotion are proposed.

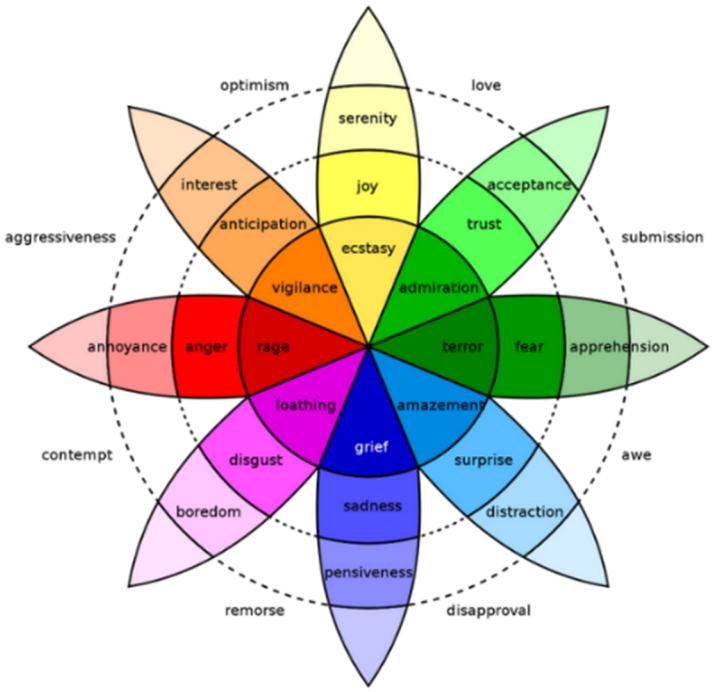


Figure 1. Plutchik's Emotional Roulette Wheel

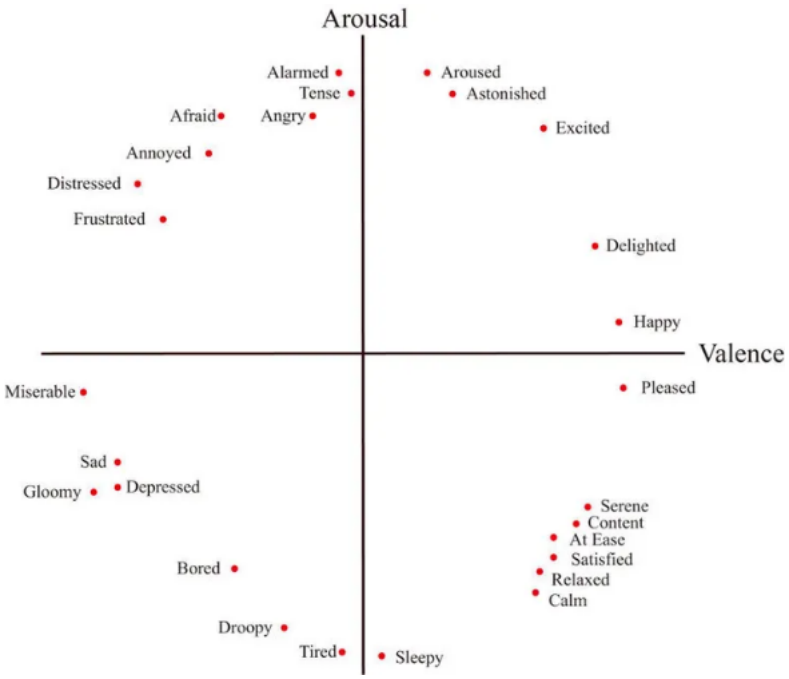


Figure 2. Circumplex modeling of emotional dimensions

Emotion category models, which typically use several basic emotional states to represent emotions. The simplest emotion category model is the binary positive and negative emotion model (polarity) [3,4]. In this case, emotions are often referred to as affects and sometimes include neutral emotions. Given the coarse-grained nature of emotion polarity models, researchers have devised relatively fine-grained emotion models, such as Ekman's six emotions (anger, disgust, fear, happiness, sadness, and surprise) [5], and Mikels's eight emotions (pleasantness, anger, respect, contentment, disgust, excitement, fear, and sadness) [6]. Based on the eight basic emotion categories (anger,

anticipation, disgust, fear, pleasantness, sadness, surprise, and trust), Plutchik [7] extends three different degrees of emotional states to give a richer set. For example, the three degrees of pleasantness are ecstasy-> pleasantness-> tranquility, and the three degrees of fear are dread-> fear-> worry. Another representative classification model is Parrott's tree hierarchical grouping model [8], which divides emotions into three levels. For example, the model contains two emotion categories, positive and negative, at the first level, six categories, anger, fear, pleasantness, fondness, sadness, and surprise, at the second level, and 25 fine-grained emotion categories at the third level. Figure 1 below shows a schematic of Plutchik's Emotion Roulette wheel.

Dimensional modeling of emotion. Affect is represented using a continuous Cartesian space of 2, 3, or higher dimensions, such as the dominance-activation-control model [9], and a range of factors such as degree of likelihood, innovativeness, temperature-activation-weight, and so on [10]. The VAD is the most widely used of the dimensional models of affect [11], with V representing the degree of satisfaction of the emotion ranging from positive to negative, A representing the range of emotion from agitation to calmness degree, and D represents the range from needing control to being in control. Figure 2 below shows a circular model of the emotion dimensions, giving the location of the associated emotions according to the degree of V and A.

These two models mentioned above are not only independently comparable, but also correlated. In some literature [12,13] the relationship between the emotion classification model and the emotion dimension model is analyzed, as well as how the two are transformed. For example, the positive space of the V-axis is associated with the state of being happy, while emotional states such as sadness and anger correspond to the negative space of the V-axis. The state of relaxation is located in the low stimulus level space of the A-axis, while anger corresponds to a greater level of stimulus. In order to further distinguish between the very similar emotions of anger and fear (both belonging to the negative space of the V-axis and the high stimulus space of the A-axis), a specific classification rule is needed (higher stimulus belongs to anger and lower stimulus belongs to fear).

2.2. The Influence of Culture on Emotions

The relationship between reactive emotions and society has a concept called emotional socialization, which means the emotions that a person experiences and expresses over a long period of time in social interpersonal relationships. Individuals are constantly trying to understand the experiences and emotions of other individuals in their social environment, as well as learning how to express their emotions in various situations and interpersonal relationships, due to the demands and influences of socialization during the course of their growth and development. This step has important implications for the development of an individual's empathy. Since human beings cannot exist independently of society, in the process of growth and development, individuals have been expressing their emotions according to the requirements of the social environment and cultural context in which they live, and expressing their emotions according to the norms of the worldview, outlook on life, values, and morality formed under the influence of the society and culture in which they live. To summarize, in the growth and development of an individual, the development of emotions is also complicated and profound because of the influences received from society, culture, history, values and so on. It can be said that it is the combination of social and cultural qualities with the basic emotions innate in human beings that produces the compound emotions that interact with each other, i.e., socialized emotions, and the process of individual growth and development is the process of socialization of emotions.

Although the biological basis of basic emotions has been confirmed, but in specific personalized and even different cultural environments have significant differences in the performance. For example, after the death of a loved one to pay tribute to the grief of all cultures and societies are universal, but the specific expression of emotions have their own different differences, for example, China's funeral culture in the catharsis of emotions, and the Mexican attitude towards the death of a loved one is also different, the Mexican people will be in the Day of the Dead on the Day of the Dead celebrations, songs and dances (which is reflected in the animated film "Finding Dreams and Travels"). It can be argued that there are social behaviors that can transcend cultures but are not universal, and similarly, basic emotions can exist across cultures but are not likely to be expressed and to the same degree. For the individual, social behaviors and emotional experiences and even expressions have been shaped by social influences.

In our lives, it is easy to experience a peculiar thing that is more face-blind to foreigners, and when watching movies and TV shows, if we are not familiar enough with the characters in the plot, it is easy to confuse the characters, and in psychology this phenomenon is known as the Intraclan Dominance Effect of Emotions, or also known as the Alien Race Effect. The meaning of this concept is that individuals tend to be more accurate or responsive in recognizing emotional pictures or expressions associated with their own cultural group, while recognizing and remembering alien emotional faces is more difficult. In a study of expression recognition between Chinese and Australian children [14] it was confirmed that children were more accurate in recognizing expressions of their home group. In addition, before compiling the CAPS dataset, researchers conducted a localized trial study of the IAPS [15], the results of which showed that the Chinese university student subjects participating in the experiment rated the IAPS significantly differently from the original ratings; an analysis of the specific differences revealed that the Chinese subjects tended to be neutral in their ratings of the portrait pictures on the IAPS, preferring to rate the IAPS as neutral, and neither liking positive nor disliking negative faces, compared to the original ratings. faces, in contrast to the original ratings that made a greater distinction between relevant pictures.

Several fMRI [16] and EEG studies [17] have also provided support and evidence for cross-cultural differences in emotions. For example, in one study, Caucasians and Japanese had different activation areas when viewing fearful faces, suggesting that Caucasians may be responding to fearful faces in a more direct and emotional way, whereas for Japanese people there is no need to respond in this way at all, and simply using the template matching system in the brain to emotionally recognize these faces is sufficient, and does not elicit a more pronounced emotional response [16]. . In another project examining the differences between Japanese and French people when receiving emotional stimuli from images, it was found that there were no significant differences in processing early on, but that significant differences

appeared later on, and after analyzing the reasons for this the researchers concluded that the neural coding processes of individuals in the Japanese and French cultures are similar, but since the Japanese culture is somehow more introspective and inhibitory the phenomenon of the Japanese subjects' individual phenomena in the experiment may be a result of their emotional expression was caused by being inhibited in the culture [17].

To summarize, different cultural and social backgrounds affect human reflection of emotional stimuli. In the field of image sentiment analysis, if we want to realize more accurate sentiment calculation and judge the level of emotions generated by images more accurately in practical applications, we must cross the cultural gap.

2.3. Affective Image Content Analysis

The core of image sentiment analysis is how to extract the sentiment features of images. Initially, based on the advancement of image recognition technology, researchers tried to use various types of basic features to extract sentiment information. For example, Machajdik and Hanbury [18] in 2010 combined various features such as color and texture; Lu et al [19] in 2012 explored how visual shapes affect the emotional expression of an image; and Sartori et al [20] in 2015 analyzed the association between different color combinations in abstract paintings and emotion from Itten's color roulette wheel .

Due to the cross-domain nature of image sentiment analysis research, the underlying low-level visual feature extraction was relatively easy to translate from the then-current field of computer vision, but at the time, there was an interpretive divide between computer implementations and psychological theories. In psychological theory, low-level features such as patterns, colors, shapes, and other low-level features are difficult to correlate to human emotional feedback on images and are difficult to explain at a theoretical level. In contrast, mid-level emotional features, such as expressions of human faces in images, are more theoretically explainable, and this information stimulates human emotional mood more strongly. On top of these mid-level features, high-level features provide information that is more intuitive and stimulates emotional responses. The most representative example is SentiBank [21], a large visual-emotional knowledge base containing 1200 concepts, each of which consists of a pair of adjectives and nouns, e.g., "cute kid", which significantly increases semantic richness.

In deep learning-based networks, there are two types of feature extraction: global features and local features. Global features focus on the whole world, do not distinguish between different regions in the image, and are processed in a uniform and consistent manner; while through the results of the study of emotional regions in the field of psychology, another type of local feature method is based on the extraction of local key features, such as human beings in a complex scene, as well as the analysis of human facial expressions and movements. Since there is more relative emotional information in local features, many researchers are working intensively on this feature. pDANet is leading the way in regression tasks based on channel and spatial attention.

2.4. Image Sentiment Dataset

In the field of image sentiment analysis, most of the initial studies were based on datasets and theories from the fields of psychology or art, and the datasets used were small due to factors in the field itself, such as IAPS [22] with 1152 commonly used data and less than 1000 in the early years; in addition, there are, for example, Abstract [23], GAPED [24], and MART [25] which are datasets for different domains, etc. In the early days, among them, IAPS was the most commonly used dataset for image emotion stimuli, in addition to its subset IAPSa [26] labeled by 20 undergraduate subjects with discrete emotion categories. With the development of the Internet, the number of active users of social media, social networking sites, etc. with related data is increasing, and some researchers have collected large-scale datasets in social networks in the form of web crawlers, such as FI [27], VSO [28], Emotion6 [29], T4SA [30], IESN [31] and EmoSet [32].

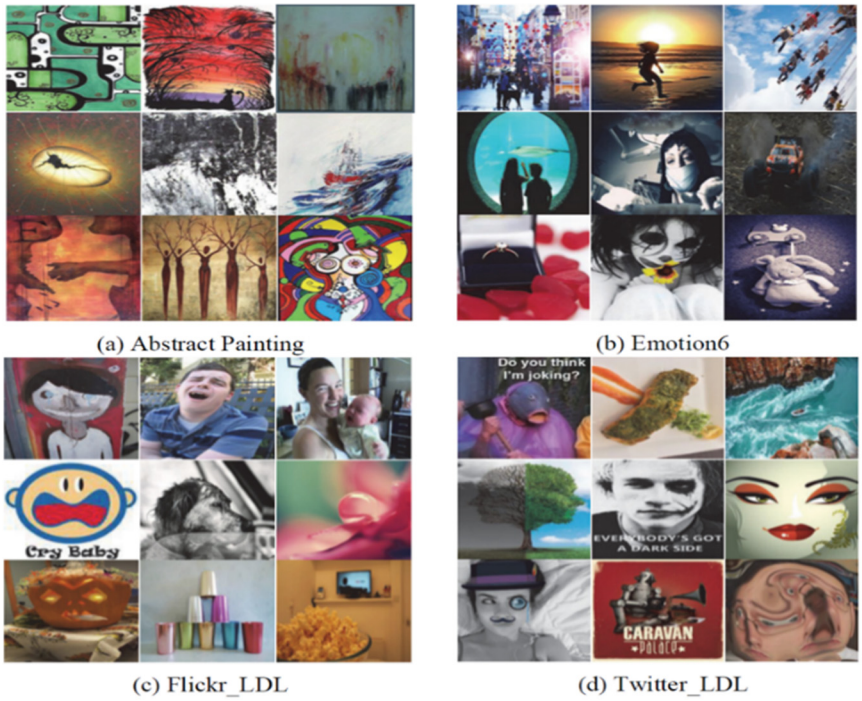


Figure 3. Plutchik's Emotional Roulette Wheel

The datasets compiled by the AICA field in recent years have shown a clear trend of increasing dataset capacity and labeling types, with the expectation of compatibility with more affective models, while the trend of collecting based on social networking sites and picture sites has become more and more obvious, which indicates that researchers have high expectations for the practical application of the research results in the AICA field. Compared with the field of psychology, the data collection and labeling of the AICA domain dataset is obviously not rigorous enough to be applied cross-domain to psychological emotion stimuli, and is not comparable to the scientific degree and consistency level of the experimental process and materials used in IAPS.

In addition, the data collection direction of AICA dataset is basically based on the internationalized environment, mainly on Twitter and Flickr, and the picture data itself is weakly related to the Chinese localization, as shown in Figure 3. below, the pictures in several typical datasets are difficult for Chinese people to experience the emotions. As for the annotation of the data, some datasets explicitly state that the annotation was performed by AMT (Amazon Mechanical Turk), while some datasets do not specify the identity of the annotator. But on the other hand, for an environment like China, which is culturally independent and has large cultural differences with Western countries, its usability validity is difficult to guarantee.

2.5. Attention Mechanism

Attention mechanisms are an important technique in deep learning, designed to enhance the "focus" of a neural network on the important parts of the input data. It mimics the function of human visual attention, allowing the model to dynamically focus on the most critical parts while ignoring other less important information. Attention mechanisms can significantly improve model performance on a variety of tasks such as image recognition, natural language processing, and speech recognition.

Channel attention is used to extract the importance of each channel in the feature map. In convolutional neural networks, a feature map often contains multiple channels, where each channel may contain different feature information about the input data. Channel Attention enables the model to emphasize the more informative channels while suppressing the less important channels from while extracting features efficiently by learning the weights of each channel. An example of a typical channel attention mechanism is the Squeeze-and-Excitation module in SENet [33] (Squeeze-and-Excitation Network) which explicitly the module explicitly models the dependencies between channels. The structure is shown in Figure 4.

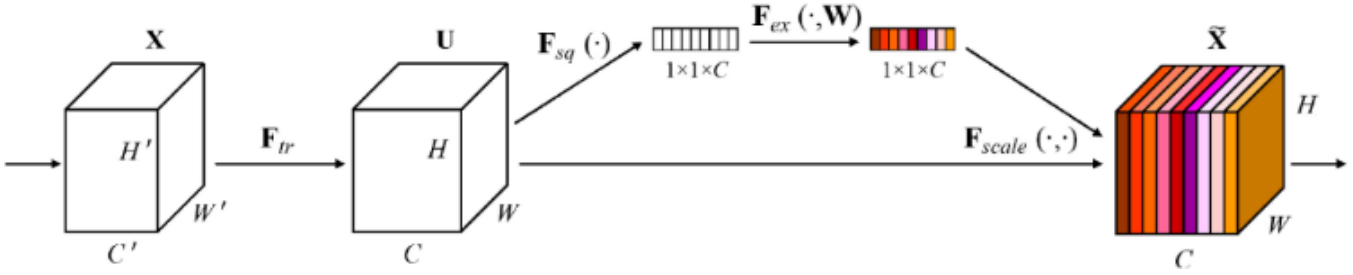


Figure 4. Schematic diagram of SENet model structure

Spatial Attention, on the other hand, focuses on different spatial regions of the input data, enabling the model to recognize which regions are more important. For convolutional neural networks, spatial attention helps the network focus on those regions of the image that contain critical information. Spatial attention can be used not only in traditional image processing tasks, but also in sequence-to-sequence models, such as in natural language processing (NLP), where the model needs to focus on keywords in the input sentence. Spatial attention is usually achieved by weighting different regions of the input feature map, and the weights are usually learned automatically by the network. Typically, CBAM (Convolutional Block Attention Module) [34], CBAM is a model that combines channel attention and spatial attention and is designed to enhance the ability of a convolutional neural network to attend to an image. Figure 5. below shows the schematic structure of CBAM.

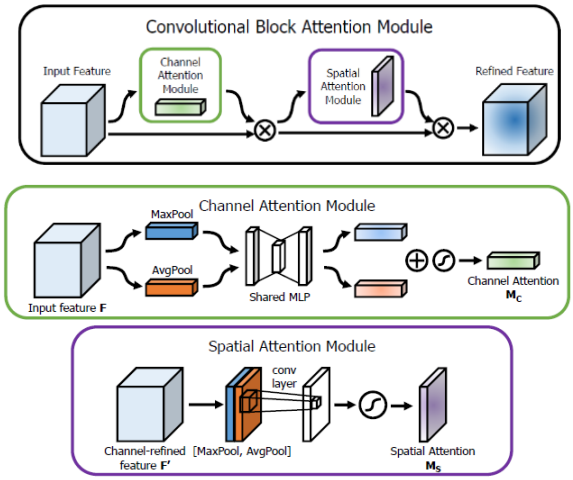


Figure 5. CBAM Schematic

3. Improved Image Emotion Regression Multiple Attention Networks

This section describes the improvement of the multiple attention network, based on the PDANet mentioning the introduction of three-dimensional attention SimAM module expansion to give the model's ability to extract information; introduced the most common triple categorization division in psychological research to improve the accuracy of the regression loss.

3.1. PDANet

PDANet [35] specifically addresses the problem of image sentiment regression based on convolutional neural networks by introducing an attention module to the network and improving the Polarity Consistent Regression (PCR) loss to guide the attention generation and improve the performance of the sentiment regression task. The network structure is shown in Figure 6. below, the incoming images are input into the channel attention module and spatial attention module respectively after the features are extracted in the backbone network of ResNet-101, and the obtained global and local information semantic vectors are subsequently connected and regressed, and the polarity-consistent regression loss is used to guide the attention generation throughout the process.

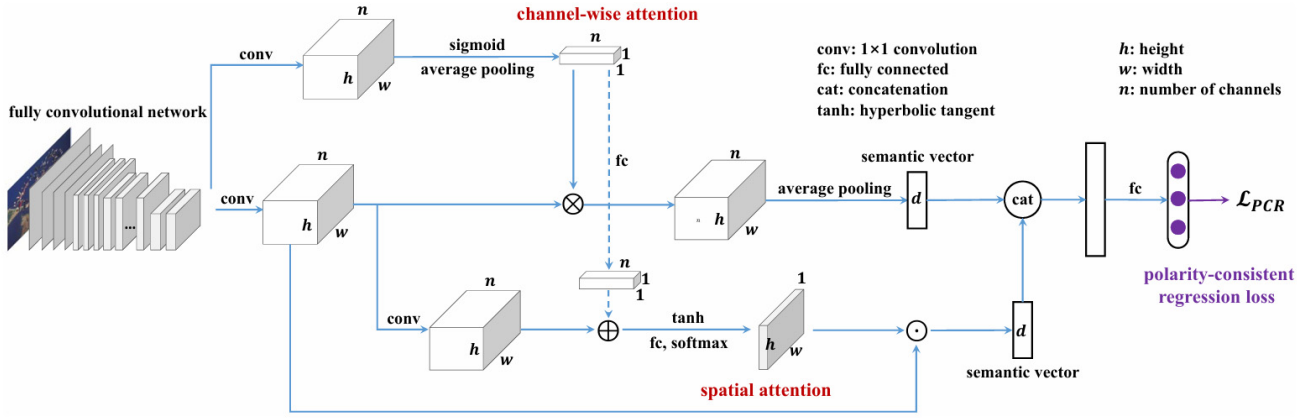


Figure 6. Schematic diagram of PDANet model structure

3.2. Improvements to PDANet

The PDANet model integrates the semantic vectors extracted by the channel attention Spatial Attention and spatial attention SeNet modules for regression through a parallel approach, but the intrinsic connection between the channel and spatial attention cannot be combined in the parallel modules, and the extracted semantic vectors retain the shortcomings of the two modules. To improve the model's ability to extract information from the output of the ResNet-101 backbone network, the SimAM module is introduced.

SimAM [36] is a parameter-free 3D attention module that assigns weights to neurons based on image local self-similarity through a theoretical study of the neuron level. Unlike the traditional 1-dimensional attention, i.e., channel attention, and 2-dimensional attention, i.e., spatial attention, the SimAM module provides attention for 3-dimensional information, a comparison of which is shown in Figure 7. below:

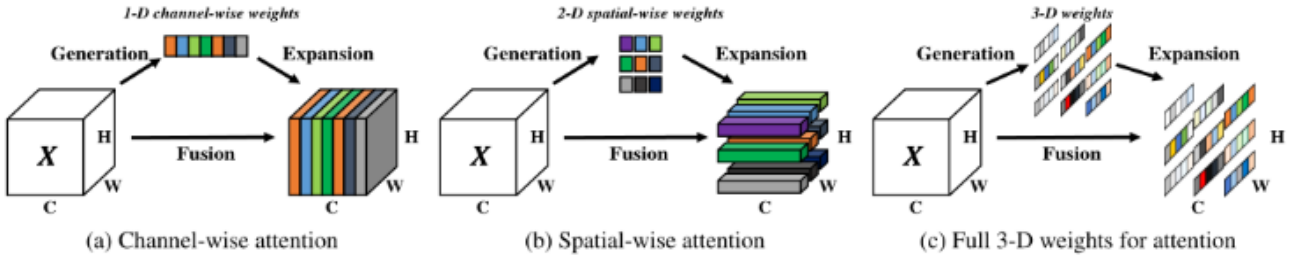


Figure 7. SimAM Module vs. Channel and Spatial Attention

In order to explore the optimal connection between the information extracted by the SimAM module and the final fully-connected layer of the network, this paper conducts experimental comparisons using three methods, namely, SimAM-only extraction of information, parallel connection of the SimAM module with the channel attention spatial attention module, and series connection of the SimAM module, respectively.

3.3. Improvement of the Loss Function

In terms of the improvement of the loss function, this paper proposes a three-category polarity regression loss function based on psychological reality. In psychological research, the emotion evaluation system based on VAD dimensions usually uses positive, medium and negative categories for finer-grained comparisons, and the polarity-consistent regression loss used in PDANet lacks in fineness.

The following equation (1) shows the formula for the polarity consistent regression loss function in PDANet:

$$\mathcal{L}_{\mathrm{PCR}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N_E}\left(fc(f_{Ai})_j - y_{ij}\right)^2\left(1+\lambda g\left(fc(f_{Ai})_j, y_{ij}\right)\right) \quad (1)$$

The part of the loss function that is improved in this paper is $g(\ldots)$, indicating whether a penalty is added or not:

$$g(\hat{y}, y) = \begin{cases} 1, & otherwise \\ 0, & if\ p(\hat{y}), p(y) \in [a, b] \end{cases} \quad (2)$$

where $p(.)$ is a function that calculates the polarity of the sentiment for a given dimension, and the polarity is considered to be the same when the predicted and actual values lie in the same judgment interval. In this paper, the intervals are set as [0, 0.4], [0.4, 0.6], and [0.6, 1].

4. Revision of CAPS

4.1. Introduction to CAPS

The Chinese Affective Picture System (CAPS) [37] was developed by Luo Yuejia's group, and its three dimensions of pleasure, arousal, and dominance were obtained from self-reported 9-point scales rated by 46 Chinese university students. The CAPS consists of 852 Chinese Affective Picture System images, which have been widely used in Chinese Affective research, because of their clear meaning and broad scope.

Since the CAPS is modeled after the IAPS system, which has long been widely used in the field of image sentiment analysis, the CAPS also has the potential to be introduced into this field. In this study, with reference to the specific use of current datasets such as IAPS, we revise the images, annotations and other data of CAPS to meet the needs of image sentiment analysis models, and evaluate and test this dataset using a variety of models and methods to achieve the purpose of introduction.

4.2. The case of the CAPS Dataset

The CAPS dataset's gives two image libraries, a full image library containing all 852 images, and a categorized image library divided into three libraries: positive, neutral, and negative. The classification is based on the fact that those with a pleasantness less than 0.4 are negative, those between 0.4 and 0.6 are neutral, and those greater than 0.6 are positive.

The dataset is provided labeled as an Excel file with statistics including the serial number of the image, the name of the image, and the mean and standard deviation of the three dimensions.

In addition, the dataset was provided with an explanatory document that included an overview of the dataset, a summary of the data with additional thumbnails of the images.

The actual questions for the dataset are as follows:

(1) the total number of pictures in the categorized picture gallery and the number of pictures in the full gallery appeared to be inconsistent, neutral 283, positive 286, negative 288, total 857, which was inconsistent with the number of 852 pictures in the full gallery;

(2) There is a discrepancy between the serial number of the pictures in the categorized gallery and the serial number of the pictures in the full gallery;

(3) the format of the pictures in the categorized gallery is jpg, but the format of the pictures in the full gallery is bmp.

To solve the above problems, it is necessary to compare and de-duplicate the classified gallery and the full gallery, and check against the annotated data, and finally form a standardized CAPS dataset that can be used for AICA.

4.3. Image Comparison Methods

Combining the total amount of CAPS data and the actual characteristics of the images, this paper chooses the method based on pixel comparison.

The principle of this method is to directly compare the pixel values at the corresponding positions of two images. Pixel-based comparisons are usually simple and direct, but they are very sensitive to noise and illumination variations in images. Common pixel-based comparison techniques include:

Absolute Difference: calculates the absolute difference between two images on a pixel-by-pixel basis.

Mean Square Error (MSE): calculates the average of the squares of the corresponding pixel differences between two images.

Structural Similarity (SSIM): evaluates the similarity of image pairs in terms of brightness, contrast, and structure, providing a more comprehensive similarity metric than pixel-based methods alone.

4.4. Implementation of the Pixel Comparison Method

The purpose of this experiment is to verify the consistency of the image data in the full and classified galleries of the CAPS dataset and to standardize the format and name. The pixel comparison method is used to compare and exclude the images one by one.

The results obtained from the above steps are manually checked, and if there is no error, the .jpg format images are selected and placed in the normalized image library, and the images with multiple matches or no matches are manually rechecked, and the no-match images are checked at the end of the dataset annotation description to finally ensure that all the images are sorted out in place.

4.5. Training Strategies

Considering that the CAPS dataset has a data size of only 852, overfitting needs to be minimized in more ways than one. In this paper, we use the method of cross-validation.

The CAPS dataset itself has a total data volume of 852 images, and for the original gallery, there are 286 positive images, 283 neutral images, and 288 negative images, which is a more balanced distribution, so it is considered that layered cross-validation can be disregarded in the division.

Due to the total data volume of 852, for the use of project resources, leave one cross-validation, although it can fully assess the model performance, but the demand for computational resources is too large to be realized; the data set division of the PDANet model is the training set: validation set: test set is 0.7: 0.1: 0.2, at this time, the training set of 596 pictures will be overfitting, training is not sufficient, refer to the NAPS and IAPS The actual division of the training set is 950 and 806 pictures, so it is considered that the use of k-fold cross-validation can take into account the actual computational resources and model training effect.

For the k-fold cross-validation method, in practice, k is usually taken as 5 or 10, and the number of samples in the training set is 682 when taking k=5, and the number of samples in the training set is 767 when taking k=10, in order to ensure the accuracy of the model, the 10-fold cross-validation method is used to alleviate the overfitting problem of the model in this study.

4.6. Paired Samples T-test

From the perspective of psychological research, the paired-sample t-test is used to analyze the predicted results of the CAPS training model with the actual results, to accurately analyze and test the correlation and variability, and to evaluate the value of the model learning results of this paper in the psychological perspective.

The paired-samples t-test is a statistical method used to test whether there is variability in paired quantitative data in paired design experiments. Between the predicted and actual results of the model in this paper, which can be regarded as the results of two different methods of testing for the same sample, the difference and correlation between these two results can be judged.

5. Experiment

5.1. Dataset

In this paper's improved multiple attention network versus the classical model, the IAPS and NAPS datasets are used, and the revised CAPS is described in the previous section.

IAPS: i.e., International Affective Picture System [22] is a quantitatively rated system of emotionally stimulating pictures compiled by the Center for Emotion and Attention Research at NIMH (National Institute of Mental Health). It is a selection of standardized and emotionally arousing color photographs from different domains that have been rigorously screened and rated. Based on Osgood et al.'s theory of emotional dimensions, the IAPS was developed using a 9-point scale based on self-reported ratings of Valence, Arousal, and Dominance to create a standardized emotional stimulus system. Since its introduction, the IAPS has played an important role in research in the field of emotion and attention, and has been widely used in basic and applied research in related directions [38]. The IAPS dataset used in this project totaled 1182 images, using data with VAD averaging and precision to 2 decimal places.

NAPS: i.e., Nencki Affective Picture System (NAPS) [39], which consists of 1356 realistic, high-quality photographs divided into five categories (people, faces, animals, objects, and landscapes). Ratings were collected from 204 participants, most of whom were European. Images were rated according to validity, arousal and dominance using a computerized bipolar semantic slider scale. Normative ratings of categories were given for each dimension. Ratings were validated by comparing them to ratings generated using a self-assessment model and the International Affective Picture System. In addition, the data includes the physical attributes of the photographs, including brightness, contrast, and entropy. This dataset will be used in this paper to validate the effect of cross-cultural influences on model generalization, and the associated data labeling accuracy is consistent with the IAPS.

5.2. Implementation Details

For the software part, the improved model in this paper is based on the Pytorch framework, version Pytorch 1.8; the language used is Python 3.8.5, the operating system is Ubuntu 18.04.5 LTS 64-bit, using Cuda 11.0, cudnn 8.0.4, and the programming platform used is Pycharm Community.

For the hardware part, the CPU used in this experiment is Intel Xeon(R) Gold 6130 @2.10GHz, 12-core processor, with 28GB of RAM, and 2 RTX 2080Ti graphics cards with a total of 22GB of video memory.

For the model training part, the ResNet101 backbone is frozen and the later part is trained; using a weight decay of 0.0005, momentum of 0.9, batch size of 32, and fine-tuning all layers using SGD. The learning rate of the convolutional layer and the last fully-connected layer were initialized to 0.001 and 0.01, respectively. for all datasets, the total number of epochs was 300, and the learning rate was decreased by a factor of 10 for the last 50 epochs.

5.3. Evaluation Indicators

Mean Square Error (MSE) and R-squared (R^2) were used to assess the results of the visual affect regression. MSE is a quadratic scoring rule that measures the average size of the error. It is defined as the mean of the squared difference between the prediction and the true situation:

$$\mathrm{MSE} = \frac{1}{M}\sum_{i=1}^{M}(z_i - \hat{z}_i)^2 \tag{3}$$

where M is the number of test samples, z_i and $\hat{z}_i$ are the original and predicted values of the VAD sentiment, respectively. MSE>=0 and smaller is better.

R^2, also known as the coefficient of determination, shows how well the predictions explain the variability in the underlying true values. It is defined as the:

$$R^2 = 1 - \frac{\frac{1}{M}\sum_{i=1}^{M}(z_i - \hat{z}_i)^2}{\frac{1}{M}\sum_{i=1}^{M}(z_i - \bar{z})^2} \tag{4}$$

where $\bar{z} = \frac{1}{M}\sum_{i=1}^{M} z_i$ is the mean of the true values. The numerator is the MSE and the denominator is the change in mood. R^2<=1, with larger values indicating better results.

5.4. SimAM Module

Different positions of the SimAM module were tested using the IAPS and NAPS datasets, and the loss function used PDANet's polarity-consistent loss function, with the following results, with bolding in the table being the best:

Table 1. IAPS test results MSE ($\times 10^{-2}$)

Module	V	A	D	M
SimAM	5.206	1.961	1.892	3.014
SimAM parallel sp+se	**3.673**	1.445	1.571	2.230
SimAM tandem sp+se	3.731	**1.301**	**1.417**	**2.149**

Table 2. IAPS test results R^2($\times 10^{-1}$)

Module	V	A	D	M
SimAM	-0.056	0.001	-0.371	-0.143
SimAM parallel sp+se	**2.906**	2.629	1.386	2.307
SimAM tandem sp+se	2.792	**3.367**	**2.230**	**2.800**

Table 3. NAPS test results MSE($\times 10^{-2}$)

Module	V	A	D	M
SimAM	4.329	1.772	3.461	3.187
SimAM parallel sp+se	2.720	1.430	**2.187**	2.112
SimAM tandem sp+se	**2.687**	**1.100**	2.249	**2.012**

Table 4. NAPS test results R^2($\times 10^{-1}$)

Module	V	A	D	M
SimAM	-0.205	-0.138	-0.163	-0.168
SimAM parallel sp+se	3.588	1.818	**3.579**	2.995
SimAM tandem sp+se	**3.665**	**3.700**	3.398	**3.588**

The sp+se in the above table is the structure of spatial and channel attention in parallel. From the results of the above experiments, the SimAM module works worst on its own,

improves when used as a parallel branch combining the semantic vectors of the spatial and channel attention modules, but works best when inserted into the ResNet-101 backbone model and between the channel and spatial attention modules and used in tandem. In the actual experiments, the direct use of the SimAM module was achieved by pooling its outputs globally on average and subsequently outputting the regression results using the fully connected layer, but the experiments showed that the SimAM module alone could not be directly relied upon for further feature extraction by the pooling operation. In the parallel experiments, the information obtained by the SimAM module is pooled by global average pooling for subsequent operations, but the results of this approach are also poor, and the performance improvement is likely due to spatial and channel attention. In the tandem experiments, the information output from ResNet-101 is fed into SimAM for one-time attention extraction, followed by the spatial attention branch and the channel attention branch, respectively; the essence of this step is that the information fed into the attention branch undergoes one-time preprocessing, and its information is organized and focused, so that it can be more easily extracted from the spatial attention and the channel attention branches to find the important feature vectors in the spatial and channel attention branches.

5.5. Tricategorical Polar Regression Loss Function

Using the network of SimAM modules in series with the triple categorical polar regression loss function, the test results are shown in Tables 5-8 below, with the bolded results in the tables being the best, and tri is an abbreviation for triple categorical polar regression loss function:

Table 5. IAPS test results MSE ($\times 10^{-2}$)

Module	V	A	D	M
PDANet	3.179	1.279	1.221	1.893
PDANet with Tri	**3.177**	1.253	**1.223**	**1.884**
SimAM parallel sp+se with Polrity	3.731	1.301	1.417	2.149
SimAM tandem sp+se with Tri	3.201	**1.234**	1.230	1.888

Table 6. IAPS test results R^2($\times 10^{-1}$)

Module	V	A	D	M
PDANet	3.859	3.479	**3.305**	3.548
PDANet with Tri	**3.860**	3.511	3.298	3.556
SimAM parallel sp+se with Polrity	2.792	3.367	2.230	2.800
SimAM tandem sp+se with Tri	3.817	**3.704**	3.255	**3.592**

Table 7. NAPS test results MSE($\times 10^{-2}$)

Module	V	A	D	M
PDANet	2.248	0.971	1.793	1.671
PDANet with Tri	**2.245**	0.963	**1.766**	**1.658**
SimAM parallel sp+se with Polrity	2.687	1.100	2.249	2.012
SimAM tandem sp+se with Tri	2.358	**0.932**	1.800	1.697

Table 8. NAPS test results R^2($\times 10^{-1}$)

Module	V	A	D	M
PDANet	4.701	4.443	4.737	4.627
PDANet with Tri	**4.710**	4.521	**4.830**	**4.687**
SimAM parallel sp+se with Polrity	3.665	3.700	3.398	3.588
SimAM tandem sp+se with Tri	4.569	**4.612**	4.703	4.628

The binary loss and triclassification loss have more obvious effects on the network, after PDANet switched to triclassification loss, the related indexes all have some growth, but for example, the MSE index of dimension D in IAPS fluctuates from 1.221 to 1.223, which can be considered as a normal training fluctuation situation. After SimAM tandem network uses triclassification loss, its performance compared to the original binary loss has a Significantly improved, the triple classification loss allows the model to adjust the network parameters more accurately during training and improves the network's ability to extract judgments for each dimension.

5.6. Revision of CAPS

The problems found after counting are:

(1) The same picture with different gallery name is different

(2) The same picture of different gallery size differences

(3) The database image format is not unified

(4) there are pictures repeated in different categories of galleries

Based on the above problems, pixel-by-pixel comparison is carried out against the categorized gallery using the whole gallery as the standard, combined with the instructions for using the dataset and the annotation file. The comparison results are shown in Table 9. below:

Table 9. Results of the revision of the CAPS dataset

Number	Categorie	Matche	Repetition	Unmatched
Positive Pictures	286	283	0	3
Neutral Pictures	283	283	2	0
Negative pictures	288	288	2	0
Total	857	854	2	0

After sorting, it was found that there were three pictures in the image library classified as positive that had no counterparts in the full image library and no related annotation data, so they were deleted; two pictures in the image library classified as neutral and negative were duplicated, but they were retained in consideration of the uniqueness of the annotation data.

For all the retained images, higher pixel versions were selected, and the naming standard was unified in JPG format. The final revised CAPS dataset consisted of a total of 852 images, with labeled data for the three dimensions of the VAD, with an accuracy of 2 decimal places.

5.7. Practical tests for CAPS

The CAPS dataset is tested and analyzed after completing the aforementioned revised CAPS task and model improvement task. Several classical models such as ResNet-101, VGG-16, AlexNet and PDANet models are used for

training and the improved multi-attention image sentiment regression model in this paper is used for training comparison. During the training process, the dataset division ratio is training set: test set: validation set = 0.7:0.1:0.2, and other hyper-parameter settings are the same as in section 3.5.2. The training results are shown in Tables 10-11 below, and the bolded results in the tables are the best:

Table 10. CAPS Training Test MSE ($\times 10^{-2}$)

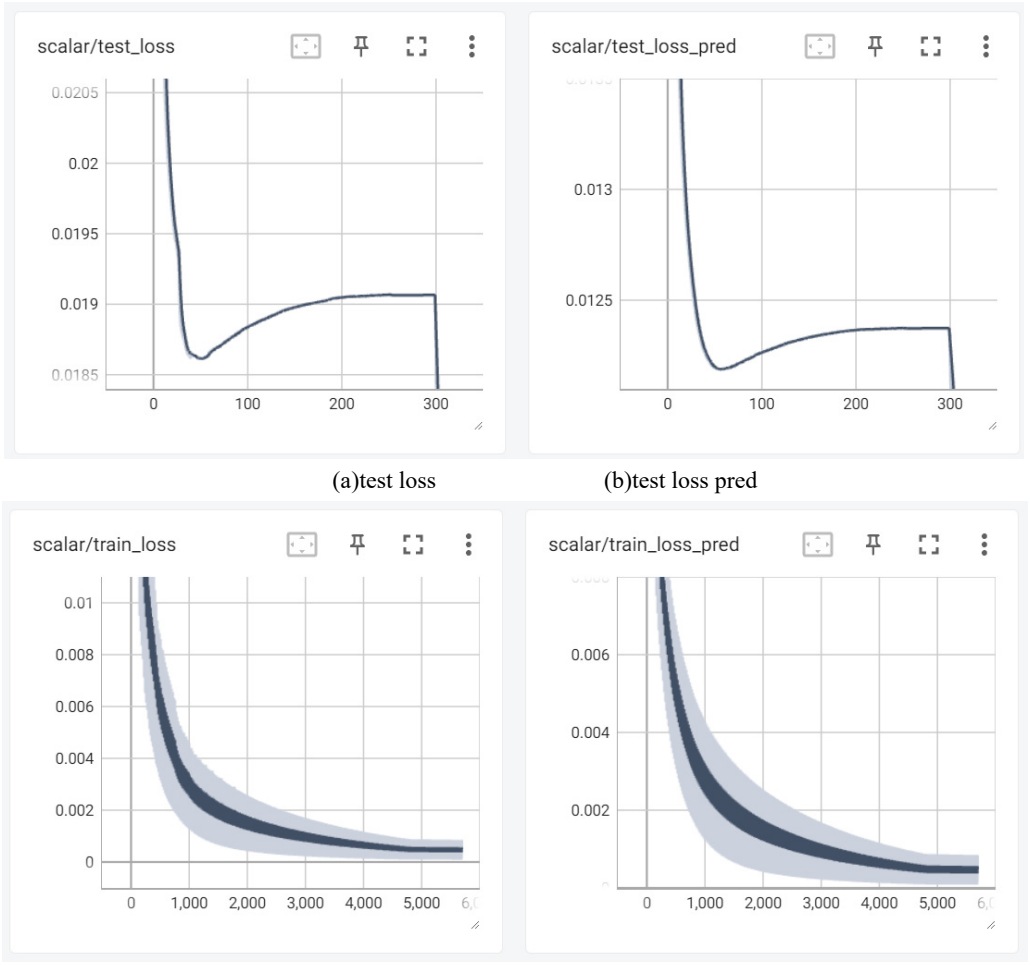
Module	V	A	D	M
AlexNet	3.231	1.259	2.012	2.160
VGG-16	3.150	1.118	2.271	2.180
Resnet-101	1.873	0.836	1.142	1.280
PDANet	1.552	0.537	1.180	1.090
Our Module	**1.524**	**0.533**	**1.024**	**1.027**

Analyzing the above table, it can be seen that the model in this paper achieves optimal results in the training of the CAPS dataset.

Table 11. CAPS Training Test $R^2(\times 10^{-1})$

Module	V	A	D	M
AlexNet	0.265	-0.451	-0.633	-0.273
VGG-16	0.997	-0.812	-0.402	-0.072
Resnet-101	5.817	4.196	6.001	5.338
PDANet	5.801	4.144	5.756	5.234
Our Module	**5.877**	**4.184**	**6.317**	**5.459**

Combined with the experimental results in part 5.5, it is found that the CAPS dataset has the best results in the comparison with a small amount of data, on the one hand, it shows that the data in this dataset is regular and has good internal consistency, and it is more obvious from the excellent R2 results that the model can find the emotional stimulus points common to the whole dataset in the limited training set, and the excellent attentional mechanism plays an important role; on the other hand, when checking the Loss curve of CAPS model training, it is found that the actual training stops early near the 50th epoch, and the subsequent epochs appear to be overfitting on the Loss curve, Figure 8. below shows the Loss curve of one training of the model in this paper:



(a)test loss (b)test loss pred

(c)train loss (d)train loss pred

Figure 8. The loss curve of CAPS on the model of this paper

Due to the division of the training, validation and test sets into 0.7:0.1:0.2, and the fact that the dataset itself is small and the training set contains less data, it is difficult to determine whether the model has adequately learned the information embedded in the dataset.

5.8. 10 Fold Cross Validation

For the 10-fold cross-validation method experiments, this paper adopts the following setup: using the improved model in this paper, the CAPS dataset is disrupted and split into 10 parts, each time, 9 of them are taken as the training set, and 1 part is the validation set, and the validation parameters for all

10 rounds of training are organized as the single 10-fold cross-validation method final test results.

The following tables 12 and 13 are the single test results of the 10-fold cross-validation method, in which the initial test is the experimental results of this paper's model in the aforementioned tables 10 and 11, and the bolded part is the best:

Table 12. Single results MSE ($\times 10^{-2}$)

Rounds	V	A	D	M
1st round	1.117	0.548	0.836	0.834
5th round	**0.954**	0.497	0.883	**0.778**
10th round	1.318	**0.417**	**0.801**	0.845
average	1.130	0.487	0.840	0.819
initial testing	1.524	0.533	1.024	1.027

Table 13. Single results R^2($\times 10^{-1}$)

Rounds	V	A	D	M
1st round	6.863	4.455	6.677	5.998
5th round	**7.320**	4.975	6.490	6.262
10th round	6.298	**5.788**	**6.815**	**6.300**
average	6.827	5.073	6.661	6.187
initial testing	5.877	4.184	6.317	5.459

The CAPS dataset was randomly disrupted multiple times for 10-fold cross-validation method stability tests, which were performed a total of 10 times in this paper, and the results of multiple 10-fold cross-validation method tests are shown in Tables 14 and 15 below.

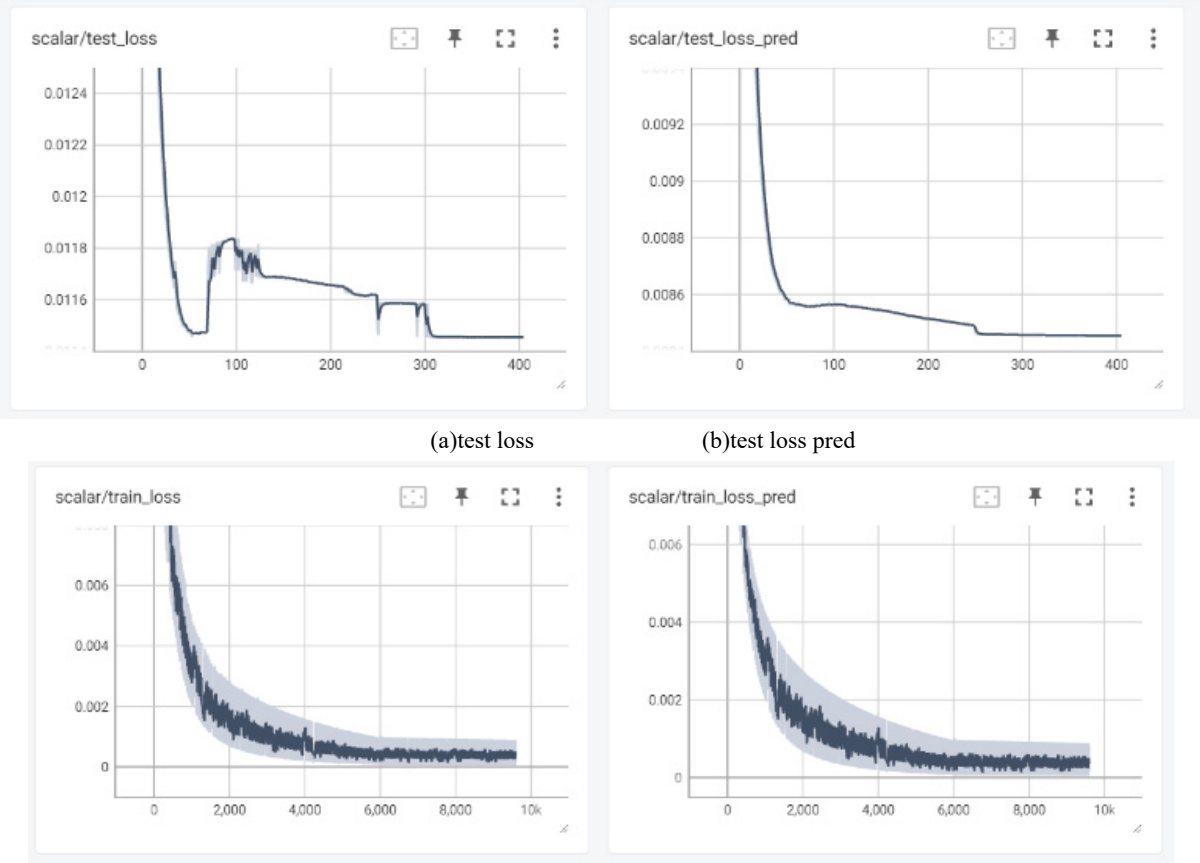
Table 14. Multiple test results MSE($\times 10^{-2}$)

Times	V	A	D	M
1st time	**1.130**	0.487	0.840	**0.819**
5th time	1.277	**0.393**	0.888	0.853
10th time	1.220	0.459	**0.788**	0.823
average	1.209	0.446	0.839	0.832
initial testing	1.524	0.533	1.024	1.027

Table 15. Multiple test results R^2($\times 10^{-1}$)

Rounds	V	A	D	M
1st round	**6.827**	5.073	6.661	6.187
5th round	6.414	**6.024**	6.470	**6.303**
10th round	6.574	5.355	**6.868**	6.266
average	6.605	5.484	6.666	6.252
initial testing	5.877	4.184	6.317	5.459

From Tables 14 and 15, it can be seen that the different divisions of the dataset have a certain impact on the model training results, but on average the impact is small, proving that the model stability is good. From the 10-fold cross validation method training results compared to the initial test results, it can be found that the 10-fold cross validation method training effect is better, and both MSE and R^2 show better results. Analyzing the results, it can be concluded that the model does not learn the information of the dataset completely during the initial test, and it can obtain better results after sufficient training and learning. There is no overfitting in any of the above experiments, and Figure 9 below shows the Loss curve of one of the experiments, which shows that the model is well trained:



(a)test loss (b)test loss pred

(c)train loss (d)train loss pred

Figure 9. Loss curves for CAPS 10-fold cross-validation

5.9. Paired Samples T-test

In this paper, CAPS is introduced into the field of image sentiment analysis, and the experimental results of its model show good performance of indicators, so its overall situation is analyzed by paired sample t-test, and the results are shown in the table below:

Table 16. CAPS paired samples t-tests

Paired group	Relevance	Significance	t	Sig. (2-tailed)
V	0.942	0.000	-1.834	0.067
A	0.895	0.000	-4.752	0.000
D	0.943	0.000	3.292	0.001

From table (b) above, it can be seen that there is a significant positive correlation between the predicted and actual values of CAPS and the degree of correlation is high and the model is well fitted. From table (c) above, it can be seen that the predicted V and A are both lower than the actual values, and the predicted D is higher; the difference is not significant in the V dimension ($p>0.05$), very significant in the A dimension ($p<0.001$), and significant in the D dimension ($p<0.01$).

From the perspective of psychological research, as a reference, Chinese researchers analyzed the paired-sample t-test correlations of emotion reports of different gender groups in localized tests of the IAPS dataset as follows: in the small-group test [7], the pleasantness correlation $r=0.968$, the arousal correlation $r=0.876$, and the dominance correlation $r=0.945$, with all the dimensions being significantly correlated ($p< 0.01$), with a significant difference in variance between V and A ($p<0.01$), and a non-significant difference in dimension D ($p>0.05$); in the large sample test [8], the pleasantness correlation $r=0.895$, the arousal correlation $r=0.636$, and the dominance correlation $r=0.812$, with significant correlations in all dimensions ($p<0.01$), and a significant difference in variance in all dimensions ($p<0.01$).

In psychological research, demographic variables (e.g., age, gender, education, and subject specialization) all differ in their ratings of emotional stimuli, but at the same time, ratings of different categorical groups correlate to a high degree. When as a group, such as in the Chinese cultural context, the correlation of group emotional responses is high, but there is also variability in emotional responses under smaller group divisions such as males and females; when emotional responses are fine-tuned to individual subjects, the same overall correlation and fine-grained differences occur between each subject. Based on the above psychological realities, the predicted values of the CAPS model also show overall correlation and fine-grained differences. In group-oriented applications, the overall high correlation has been able to make good predictions of the group's emotional stimulus levels, but individualized adjustments are still needed in personalized emotion prediction tasks.

6. Conclusion

6.1. Improved Image Emotion Regression Multiple Attention Network

In the above experimental results, the improved model of this paper is slightly better than PDANet in general. Analyzing the results of each dimension, it can be seen that the model of this paper improves from 1.279 to 1.234 for IAPS dataset and 0.971 to 0.932 for NAPS in MSE indicator of the relevant parameter of dimension A; in the indicator of R2, the performance of IAPS dataset is significantly improved from 3.479 to 3.704 and the NAPS dataset also improves from 4.443 to 4.612, it can be said that the improvement of the performance in this dimension is the key to the overall performance improvement; however, the metrics performance of this paper's model in dimensions V and D both have a certain decline compared to PDANet, and its decline is not significant. In the training of CAPS, this paper significantly outperforms PDANet in two metrics in all three dimensions. Overall, this paper's model improves in the overall performance by improving the performance in dimension A, which proves the effectiveness of this paper's improved method.

6.2. Initial Introduction of CAPS

The CAPS dataset is a dataset of image emotion stimuli in Chinese cultural contexts, which has previously been widely used in the field of psychology. In this paper, we introduce this dataset into the image sentiment analysis work and conduct a series of tests on it. The test results show that the data within CAPS are uniformly categorized, reasonably differentiated, and highly consistent, and both outperform IAPS and NAPS in terms of model training metrics. as seen in the paired-sample t-test, the training prediction results of this dataset correlate well with the dataset itself, and it performs excellently from a psychological point of view, and it can be attempted to be applied in psychological research. In addition, due to the unique Chinese cultural attributes of this dataset, it is believed that the models trained using it can achieve better practical results in Chinese cultural applications.

Acknowledgments

Thanks to the research group that produced IAPS, NAPS and CAPS, and the team of PDANet.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] J. Tooby and L. Cosmides, "The evolutionary psychology of the emotions and their relationship to internal regulatory variables," in Handbook of Emotions, M. Lewis, J. M. Haviland-Jones, and L. F. Barrett, Eds., 3rd ed. New York: The Guilford Press, 2008, pp. 114–137.

[2] J. A. Russell, "Emotion, core affect, and psychological construction," Cognition and Emotion, vol. 23, no. 7, pp. 1259–1283, 2009, doi: 10.1080/02699930902809375.

[3] S. Zhao et al., "Personalized emotion recognition by personality-aware high-order learning of physiological signals," ACM TOMM, vol. 15, no. 1s, Art. no. 14, 2019.

[4] S. Zhao, G. Ding, J. Han, et al., "Personality-aware personalized emotion recognition from physiological signals," in Proc. IJCAI, 2018, pp. 1660–1667.

[5] P. Ekman, "An argument for basic emotions," Cognition & Emotion, vol. 6, no. 3-4, pp. 169-200, 1992.

[6] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, et al., "Emotional category data on images from the international affective picture system," BRM, vol. 37, no. 4, pp. 626–630, 2005.

[7] R. Plutchik, Emotion: A Psychoevolutionary Synthesis. Harpercollins College Division, 1980.

[8] W. G. Parrott, Emotions in Social Psychology: Essential Readings. Psychology Press, 2001.

[9] H. Schlosberg, "Three dimensions of emotion," Psychological Review, vol. 61, no. 2, p. 81, 1954.

[10] J. Lee and E. Park, "Fuzzy similarity-based emotional classification of color images," IEEE TMM, vol. 13, no. 5, pp. 1031–1039, 2011.

[11] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," IVC, vol. 31, no. 2, pp. 120–136, 2013.

[12] K. Sun, J. Yu, Y. Huang, et al., "An improved valence-arousal emotion space for video affective content representation and recognition," in Proc. ICME, 2009, pp. 566-569.

[13] S. M. Alarcão and M. J. Fonseca, "Identifying emotions in images from valence and arousal ratings," MTA, vol. 77, no. 13, pp. 17413–17435, 2018.

[14] R. Markham and L. Wang, "Recognition of emotion by Chinese and Australian children," Journal of Cross-Cultural Psychology, vol. 27, no. 5, pp. 616-643, 1996.

[15] Yuxia Huang, Yuejia Luo, "A pilot study of the International Affective Picture System in China," Chinese Journal of Mental Health, vol. 18, no. 9, pp. 631-635, 2004.

[16] Y. Moriguchi, T. Ohnishi, T. Kawachi, et al., "Specific brain activation in Japanese and Caucasian people to fearful faces," Neuroreport, vol. 16, no. 2, pp. 133-136, 2005.

[17] P. Hot, Y. Saito, O. Mandai, et al., "An ERP investigation of emotional processing in European and Japanese individuals," Brain Research, vol. 1122, no. 1, pp. 171-178, 2006.

[18] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in Proc. ACM Multimedia, 2010.

[19] X. Lu, P. Suryanarayan, R. B. Adams Jr, et al., "On shape and the computability of emotions," in Proceedings of the 20th ACM international conference on Multimedia, 2012, pp. 229-238.

[20] A. Sartori, D. Culibrk, Y. Yan, et al., "Who's afraid of itten: Using the art theory of color combination to analyze emotions in abstract paintings," in Proceedings of the 23rd ACM international conference on Multimedia, 2015, pp. 311-320.

[21] D. Borth, T. Chen, R. Ji, et al., "SentiBank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content," in Acm International Conference on Multimedia, ACM, 2013, doi:10.1145/2502081.2502268.

[22] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (iaps): Technical manual and affective ratings," NIMH Center for the Study of Emotion and Attention, 1997, pp. 39-58.

[23] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in Proceedings of the 18th ACM international conference on Multimedia, 2010, pp. 83-92.

[24] E. S. Dan-Glauser and K. R. Scherer, "The geneva affective picture database (gaped): a new 730-picture database focusing on valence and normative significance," Behavior Research Methods, vol. 43, no. 2, pp. 468-477, 2011.

[25] X. Alameda-Pineda, E. Ricci, Y. Yan, et al., "Recognizing emotions from abstract paintings using non-linear matrix completion," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5240-5248.

[26] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, et al., "Emotional category data on images from the international affective picture system," Behavior Research Methods, vol. 37, no. 4, pp. 626-630, 2005.

[27] Q. You, J. Luo, H. Jin, et al., "Building a large scale dataset for image emotion recognition: The fine print and the benchmark," in Proceedings of the AAAI Conference on Artificial Intelligence, 2016, pp. 308-314.

[28] D. Borth, R. Ji, T. Chen, et al., "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in Proceedings of the 21st ACM international conference on Multimedia, 2013, pp. 223-232.

[29] K.-C. Peng, A. Sadovnik, A. Gallagher, et al., "A mixed bag of emotions: Model, predict, and transfer emotion distributions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 860-868.

[30] L. Vadicamo, F. Carrara, A. Cimino, et al., "Cross-media learning for image sentiment analysis in the wild," in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 308-317.

[31] S. Zhao, H. Yao, Y. Gao, et al., "Predicting personalized emotion perceptions of social images," in Proc. ACM MM, 2016, pp. 1385-1394.

[32] J. Yang, Q. Huang, T. Ding, et al., "EmoSet: A large-scale visual emotion dataset with rich attributes," in Proc. of the 2023 IEEE/CVF International Conference on Computer Vision, 2023, pp. 20326-20337.

[33] J. Hu, L. Shen, S. Albanie, et al., "Squeeze-and-excitation networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 8, pp. 2011-2023, 2020.

[34] S. Woo, J. Park, J. Y. Lee, et al., "CBAM: Convolutional block attention module," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11211, pp. 3-19, 2018.

[35] S. Zhao, Z. Jia, H. Chen, et al., "PDANet: Polarity-consistent deep attention network for fine-grained visual emotion regression," in Proc. ACM MM, 2019, pp. 192-201.

[36] L. Yang, R. Y. Zhang, L. Li, et al., "SimAM: A simple, parameter-free attention module for convolutional neural networks," in International Conference on Machine Learning. PMLR, 2021.

[37] Bai Lu, Hui Ma, Yuxia Huang, and Yuejia Luo, "Development of the Chinese Affective Picture System - a trial among 46 Chinese college students," Chinese Journal of Mental Health, vol. 19, no. 11, pp. 719-722, 2005.

[38] Liu Xiaonan, Xu Aoxiang, Zhou Renlai, "A localized study of the International Affective Picture System: a rating in a Chinese college student population," Chinese Journal of Clinical Psychology, 2009(6), pp. 4.

[39] A. Marchewka, Ł. Żurawski, K. Jednoróg, et al., "The Nencki Affective Picture System (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database," Behavior Research Methods, vol. 46, no. 2, pp. 596-610, 2014.