Thesis for Degree of Master

Supervisor: Prof. Dugki Min

Multi-task Emotion Recognition Based on Context-aware and Attention Module

Submitted by

Lina Zhang

August, 2022

Department of Computer,Information&

Communications Engineering

Graduate School of Konkuk University

Multi-task Emotion Recognition Based on Context-aware and Attention Module

A Master's Thesis

submitted to the Department of Computer,

Information &Communications Engineering

and the Graduate School of Konkuk University

in partial fulfillment of the

requirements for the degree of

Master of Science in Computer&

Communication Engineering

Submitted by

Lina Zhang

April,2022

This certifies that the Thesis of

Lina Zhang is approved.

Approved by Examination Committee

Chairman 김 학 수

Member 조 기 춘

Member 민 덕 기

June, 2022

Graduate School of Konkuk University

TABLE OF CONTENTS

List of Tables

List of Figures

ABSTRACT

Multi-task Emotion Recognition Based on Context-aware and Attention Module

Lina Zhang

Department of Computer,Information&Communications Engineering

Graduate School of Konkuk University

Emotion recognition is an important research topic in the field of computer vision, and is the primary problem in affective computing, as well as the focus and difficulty of research. It would have a better impact on our lives if we could use inexpensive machines to monitor and understand the emotional information of others. However, there is currently no system that can do such a job. Because human emotional states are expressed in various ways, such as speech, expressions, actions, the environment in which a person lives, and various physiological signals, it is difficult to accurately reflect human emotions by relying on a single feature parameter and its characteristics. Therefore, multi-task-based feature fusion is an effective method to achieve accurate emotion recognition. With the continuous development of convolutional neural networks (CNN) and machine learning technology, human emotion recognition has entered a new stage. Currently standard CNNs are mostly applied to general facial expression recognition, but are limited in their feature extraction capability.

In this paper , we propose a framework, Muti-task Emotion Recognition (MTER) which designs a lightweight CNN model for multitasking based on visual context. METR has 4 main models: face feature extraction model, body feature extraction model and context (scene) feature extraction model, and fusion classification model. It is used to analyze images containing multiple people and recognize fused emotions based on face facial features, body features, and context information. The face feature and body feature extraction module takes the face and body parts of the image as input and the information implicit in the image such as facial expression, head position and body pose is extracted. In order to make the emotion recognition actively applied to real life, mobilenet lightweight network is utilized to reduce the computational effort and increase the recognition speed. An attention model, Squeeze-and-Excitation, is added to make the network focus more on the relationship between channels. add Batch Normalization (BN) after depthwise separable convolution to speed up the network convergence, use Global Average Pooling (GAP) instead of fully connected layer to reduce the number of parameters in the network. We use Fine-tuned Mobilenet network pre-trained on ImageNet, and the scene feature extraction module takes as input the entire image, which may contain more than one person, and these features reflect and encode the main aspects of the image. We used only the convolutional layer of ResNet and pre-trained on the ImageNet. For feature extraction, the size of the input image is reduced to reduce the complexity of the model.

To deal with the data imbalance, we develop a small sample dataset about faces, EMOTIC_Face, which is a sample balanced face dataset considering the context, with 26 emotions. Our experimental results show that our proposed framework is feasible compared to other field recognition work.

The overall robustness of the model is improved and the effectiveness of the improved algorithm is demonstrated on the EMOTIC dataset of real scenarios, with an average accuracy improvement of 10.07% over the base convolutional neural network.

Keyword: CNN, Emotion Recognition ,Context, Multi-task

Chapter 1.Introduction

Emotion recognition refers to artificial intelligence[1] that automatically discriminates an individual's emotional state by acquiring physiological or non-physiological signals of the individual and is an important component of affective computing. Emotion recognition studies include facial expression, speech, heart rate, behavior, text, and physiological signal recognition to determine the emotional state of a user the above. With the continuous development of technology, human emotion recognition has entered a new stage. Many researchers have proposed many different techniques and methods such as signal processing [2], machine learning [3], speech processing [4], and computer vision [5]. Different methods and techniques are also used to analyze emotion representations, such as Bayesian network [6], Gaussian mixture models [7], Hidden Markov model [8], and deep neural network [9]. At present, there is already a high level of emotion recognition, and at the same time, the hardware and software infrastructure conditions for large-scale popularization are also available, the application market and field needs are great, and the market development and specific applications based on this technology are showing a booming trend. Emotion recognition, as an important component of recognition technology, has received wide attention in the fields of human-computer interaction [10], robotics, automation, mental health, and medical care, and has become a research hotspot in academia and industry.

Emotion is a state that integrates human feelings, thoughts, and behaviors, and plays an important role in human-human communication. It includes a person's psychological response to external or self-stimuli, including the physiological responses that accompany that psychological response. At present, there are three major types of human emotions: facial expressions, vocal expressions, and body gestures, among which, facial expressions have become the most important way of communication for people to convey information in daily life. The role of emotions is everywhere in our daily communication. In medical care, if we can know the emotional state of patients, especially those with expression disorders, we can make different care measures according to their emotions and improve the quality of care. In the process of product development, if we can identify the emotional state of users in the process of using the product and understand the user experience, we can improve the product functions and design products that are more suitable for the needs of users. In various human-machine interaction systems, if the system can recognize a person's emotional state, the interaction between humans and machines will become more friendly and natural.

It is widely assumed that a person's emotional state can be easily inferred from his or her facial movements, often referred to as emotional expressions or facial expressions. This assumption has influenced legal judgments, policy decisions, and educational practices; guided the diagnosis and treatment of mental illness, and the development of commercial applications. In the future, intelligent robots will certainly be used in all aspects of human life. To improve the experience of human-machine interaction, it is necessary for machines to be able to accurately determine the true inner thoughts of human beings based on their behavioral characteristics and expression features, so the study of emotion recognition is crucial, and the analysis and recognition of emotions allows people to further deepen their understanding of human inner emotions.

In this paper, we investigate examples of this widespread assumption, and then we examine the scientific evidence that tests this view, focusing on the seven emotion categories most commonly used by consumers of emotion research: anger, disgust, fear, happiness, sadness, surprise, and neutrality. The available scientific evidence suggests that people do sometimes smile when they are happy, frown when they are sad, frown when they are angry, and so on, more than might be expected by chance, as suggested by the common view. However, how people communicate anger, disgust, fear, happiness, sadness, surprise, and neutrality varies widely across cultures, across situations, and even across people in the same situation. Furthermore, similar facial action configurations will express instances of more than one emotional category to varying degrees. Indeed, a particular facial action configuration, such as a frown, often conveys more than one emotional state. Scientists agree that facial actions convey a range of information that is important for social communication, emotional or otherwise. But to recognize people's emotions without facial information remains a difficult problem and research direction for emotion recognition nowadays. There is an urgent need to study how people actually move their faces to express emotions and other social information in the various contexts that make up everyday life, as well as to scrutinize the mechanisms by which people perceive instances of each other's emotions if they do not have faces. In this paper, we present specific research proposals that will yield a more effective result on how people can recognize human emotions without faces and how they can infer the meaning of emotions, from everyday life. This research is critical to providing consumers of emotion research with the translational information they need.

In recent years, based on the rapid development of computer vision technology, significant progress has been made in the field of facial expression recognition. Convolutional neural networks(CNN) [11] are one of the most popular deep learning algorithms, which have achieved excellent results in image classification [12] and target detection [13] with end-to-end learning. CNN combines feature extraction and classification by training samples with labels to learn features. B. Fasel et al [14] designed a 6-layer shallow network structure and used convolutional kernels of different sizes for face expression feature extraction, which can extract face expression features of different scales, resulting in a significant improvement in facial expression recognition accuracy. In 2012, Alexnet [15] convolutional neural network achieved first place in the ImageNet [16] large-scale visual recognition competition, which opened up the research boom of convolutional neural networks. In 2014, Simonyan proposed the VGG [17] convolutional neural network, which won first place in the localization project at the ImageNet

Challenge in the same year. The VGG network investigates the relationship between network depth and network expressiveness and increases the network depth with small-scale convolutional kernels while improving the model expressiveness. Due to a large number of parameters and computations, the VGG network is slow in practical applications.

Szegedy et al. proposed the GoogleNet [18], whose core module is the Inception module, and used three different scales of convolutional kernels and a pooling layer to extract features, and then stitched together the features extracted from different branches. together, which increases the width of the network. Quan et al [19] found that in real scenarios, facial expression recognition is disturbed by many extraneous factors, and the shortage of training data and the unbalanced data distribution of existing face expression datasets make it difficult to improve the expression recognition accuracy significantly. They learn to use a new clustering-weighted loss function in the fine-tuning phase of network training by borrowing from deep transfer learning techniques [20]. The clustering weighted loss function improves both intra-class tightness and inter-class separability by learning the class centers of each expression class. Considering the unbalanced nature of the facial expression data set, each emotion class is given a weight based on its proportion of the total number of images, and good results are achieved in expression recognition. As the number of layers of the convolutional network increases, the training cost becomes higher and higher, and the gradient disappears. In order to solve this problem, Kaiming He et al. proposed the residual network model (ResNet) [21], which learns the residual mapping by shortcut connection.

When convolutional networks learn to get good recognition results, some people focus their attention on lightweight networks. 2017, google proposed MobileNetv1 [22], proposing depthwise separable convolutions, which reduces the computational effort of general convolutions. 2018, and then proposed MobileNetv2 [23], proposed inverted residual block and linear bottlenecks, which can remove both the redundant information of highdimensional features and the information collapse of low-dimensional features. Also, EfficientNet [24] proposed a new scaling method that uses simple but efficient composite coefficients to uniformly scale all dimensions of depth, width and resolution. Pablo et al [25] also used a lightweight network for training and used a suppression layer in the last layer of the network, which helps to form features for facial expression learning in the last layer of the network. Facial emotion recognition plays an important role in our emotion recognition, but there are other factors that influence our emotion recognition, such as body pose, speech intonation, and the environment we are in.

Facial expression recognition has achieved great research results, but emotion recognition of other features still faces many problems in terms of accuracy and real-time:

Most emotion recognition only focuses on facial emotion recognition, but in the daily complex environment, the general recognition of facial expressions is affected by lighting and occlusion, and the face posture can also change at will, which greatly affects the accuracy of expression recognition. How to overcome the influence of these factors on expression recognition and improve the accuracy of expression recognition in complex environments is a key issue in current face expression recognition.

With the deepening of the network, the standard of hardware is raised, and the arithmetic power of our devices in daily life is not as high as that in research laboratories. Moreover, facial expressions are often instantaneous, and it is difficult to improve real-time to a level that can be practical.

With the increase in computer processor speed, the innovation of algorithms for emotion feature extraction has become an important breakthrough;

With the in-depth study of psychology, the benchmark of emotion calibration in emotion recognition still varies and we still need a large number of samples to train.

This paper focuses on emotion recognition classification. Since too many parameters of deep neural networks inevitably lead to a network model whose real-time performance is not guaranteed, an emotion recognition method with depthwise separable convolutional channel features is proposed. First, a CNN structure is designed, then features are extracted by multiple tasks simultaneously, and finally, the extracted features are fused to produce a 2-dimensional output containing multiple labels.

In this section, we introduce the importance of emotion recognition in life and the research background and significance of emotion recognition, describe the concept of emotion recognition and the main application requirements and scenarios. It also introduces the classical convolutional neural networks based on deep learning in recent years. In the existing research work on emotion recognition, these existing problems cannot be effectively solved, so this paper proposes a context-aware approach based on the use of deep neural networks to investigate the above existing problems, and this thesis is structured as follows.

In Chapter 2, the types of emotion recognition as well as methods are introduced, as well as examples of what context-awareness is, and also the techniques and concepts involved in this thesis are introduced. It focuses on the basic concepts of neural networks and introduces some variants in convolutional neural networks, describes the background and introduces the details of the existing convolutional neural networks in the proposed thesis approach, which are MobileNet[23], ResNet[21], and Squeeze-and-excitation(SE)[26]. using depthwise separable convolution[22] containing an Inverted Residual Block[23], reduces the number of parameters significantly, which improves the detection speed and reduces the space occupied by the model, and makes it easier to be applied to mobile emotion recognition, and analyzes the difference between deep separable convolutional operations and traditional convolutional operations, followed by an understanding of structures that enhance network robustness such as linear bottlenecks and inverted residuals.

In Chapter 3, we propose an overall network framework for problem solving and implementation, using the stacking of Inverted-SE residul blocks, to achieve our main task. To make the emotion recognition actively applied to real life, we use MobileNet lightweight network to reduce the computational effort and increase the recognition speed. We pre-trained on ImageNet using Fine-tuned MobileNet. The scene feature extraction module takes as input the whole image, and these features reflect and encode the main aspects of the image. We used only the convolutional layer of ResNet and pre-trained it on the location database.

Chapter 4 is the experimental part, in order to demonstrate the effectiveness of our proposed network, we devolop a small dataset on faces, EMOTIC_Face, to train the network and give a comparison of the accuracy and performance of the proposed method at the experimental results. The comparison includes a variant of MTER, and a comparison between the original CNN[27] and MTER.

Chapter 5 summarizes the results of the full paper and proposes subsequent improvements based on them, and provides an outlook on future work.

Chapter 2.Background and Related Work

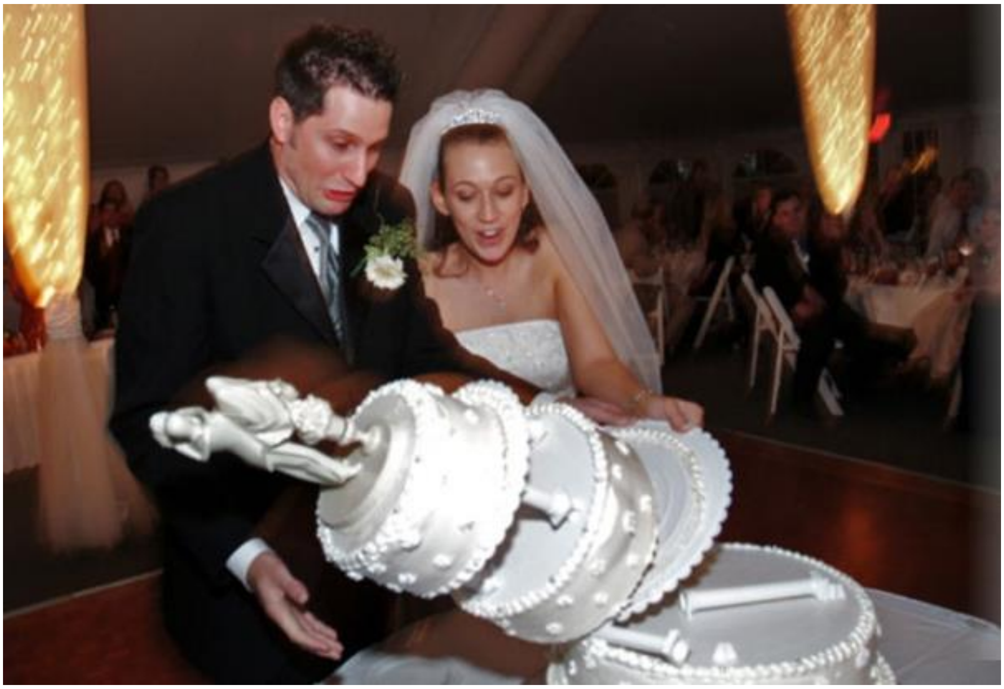
2.1.Emotion Recognition

Since the development of artificial intelligence, emotion recognition has also become more and more good recognition results, traditional expression recognition, which is recognized by facial emotion [28,29,30,31] of human face, and emotion recognition based on speech[32,33,34], multimodal-based emotion recognition[35,36,37],based on ElectroencePhalography(EEG)[38,

39,40], emotion recognition based on spatial-temporal [41,42,43], etc. Most techniques for recognizing emotions in images are based on the analysis of people's facial expressions, with the hidden condition that these expressions are considered to best convey human emotional responses. Therefore, most datasets used for training and evaluation of emotion recognition tools contain only cropped images of human faces. The main limitation of traditional emotion recognition tools is that they fail to achieve satisfactory performance when people's facial expressions are blurred or indistinguishable. In contrast to these approaches, humans are able to recognize the emotions of others based not only on their own facial expressions but also on contextual cues [27][44][45][46][47] (e.g. the action they are performing, their interaction with others, their location).

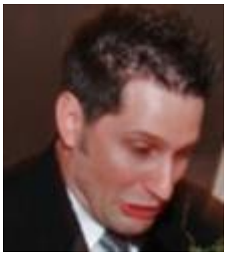
2.2.Context-aware in Emotion Recognition

In our daily life, objects do not exist in isolation from the scene according to the semantic understanding of visual content. The information that directly or indirectly affects the perception and discrimination of a visual object can be called contextual information. For example, in figure 2.1.



<Figure 2-1> A sample from the EMOTIC dataset(a)

The semantically relevant information about the "couple undergoing a wedding" includes: interaction objects (bride, groom, cake), scene information (audience, place of wedding), and component information (wedding dress, boutonniere, etc.) These elements, which have a symbiotic, inclusive, and positional relationship with the target visual content, are all important for understanding the visual content. These elements are important for understanding the visual content. As shown in Figure 2-2.



<Figure 2-2> A sample from the EMOTIC dataset(b)

In the absence of contextual information, it is difficult to accurately determine the real behavior of the objects by relying on a single target and to understand the message of the whole image. For the object in Figure 2.2, it can be interpreted as uncertain information such as surprise or fear, while in Figure 2.1, based on the rich contextual information such as interaction (bride looking at the cake together), scene information (place of marriage), etc., it is possible to give an accurate picture of a couple who is getting married and the cake has fallen but still feels happy.

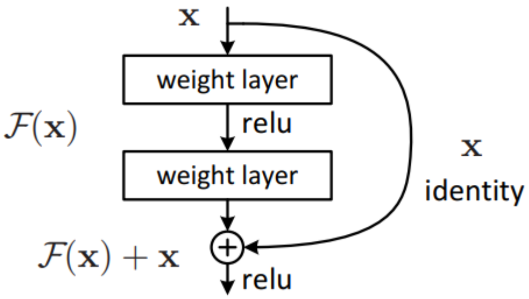
2.3.Context-based Emotion Recognition methods

In recent years, many studies have considered the use of multiple up and down texts to improve the performance of different tasks. A similar architecture was proposed by Kosti et al [27]. They have the architecture of a basic CNN followed by a fusion network. One network concentrates on the body and the other model focuses on capturing the context. lee et al [45] consider everything except the face as the context and therefore mask the face from the image and feed it into the fusion network. Thus, the face is masked from the image to provide to the context stream. On the other hand, [47] used a region proposal network (RPN) to extract contextual elements from the images. These elements become the nodes of the sentiment map and are fed into a graphical convolutional network (GCN) to encode the context. Emoticon [46] proposed three interpretations of context for perceptual emotion recognition. They used Openface[48] to recognize faces, Openpose[49] to recognize gait, and fused multiple patterns, background visual information, and social dynamics of interactions between agents to infer perceived emotions.

2.4.ResNet

The Residual Network (ResNet) was proposed by Microsoft Research, which deepens the network structure and makes the features extracted by the network more effective by introducing residual blocks. In the 2015 ILSVRC competition.ResNet won the first place in the 2015 ILSVRC competition with an error rate of 3.57% that exceeded the accuracy of normal human eye recognition compared with the VGG network.ResNet models have fewer parameters and therefore higher recognition rates than VGG networks.It can be found from the previously introduced models that in the development of deep learning, the number of layers of the model and the depth of its network structure are increasing.The depth of the network structure is increasing, so it can be assumed that by deepening the structure of the network.

Therefore, it can be assumed that by deepening the structure of the network,the model must show better results than the relatively shallow network. However, practice again shows us that the training error also increases of the model structure. In response to this phenomenon, Kaiming He et al. propose a solution.The ResNet network adopts the idea of inter-layered connectivity by adding directly connected channels in the network layers to map the data to subsequent layers.The ResNet adopts the idea of skip connection, adding direct connection channels in the network layers, mapping the data to the subsequent network layers, and adding the results after the operation. By this way of jumping over the connections.The problem of gradient disappearance with increasing number of layers can be moderately alleviated. Figure 2-3 shows the basic structure of the ResNet.

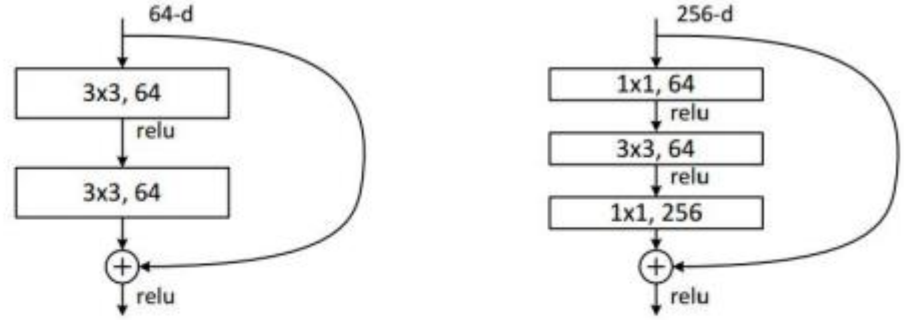

x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

<Figure 2-3> Residual Network building block

The structure shown is also called the residual block, and the model proponent, through experiments, proposes a network ResNet with 34, 50, 101, and 152 layers, and this structure is proposed so that the effect of adding more layers to the network does not deteriorate.The effect of adding more layers to the network does not deteriorate, and the error rate is greatly reduced. Equation 2-1 is the residual structure formula.

$$Y=F(x)+x \tag{2-1}$$

In Eq. 2-1, F(x) is the output after the convolution operation and x is the original input. Figure 2-4 shows the specific design of the residual block.One of them is a two-layer 3*3 convolutional network connected in series. The other one uses a three-layer convolutional network with 1*1, 3*3 and 1*1.



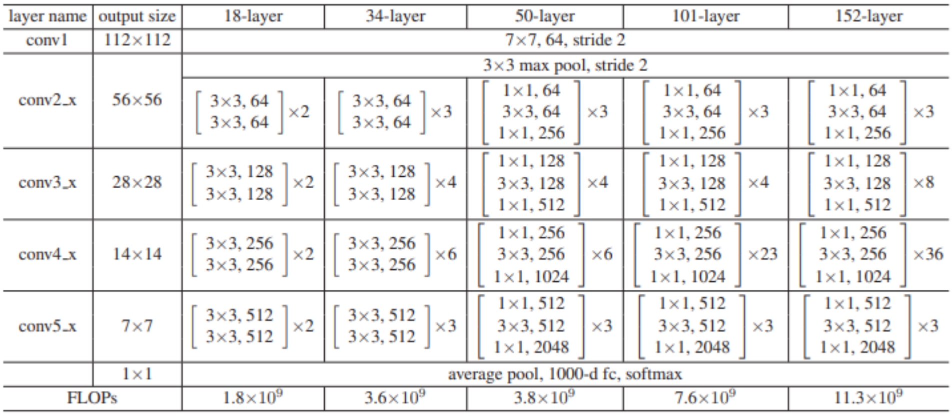
< Figure 2-4> ResNet blocks

The gradient is directly related to the training procedure and generalization of the network, and it is difficult to maintain a good error profile after multiple training sessions, which may cause the model to fail to back propagate. Currently, there are initialization and gradient shearing and BN layers to solve such problems, which may alleviate the problem of gradient explosion or gradient dispersion to some extent, but if the network structure can be modified, it may have better performance.

The main idea of the ResNet is the Highway Network, which allows to preserve part of the output of the previous network layer, so that the original input information can be passed to the later network layers, reducing the learning pressure of the later neural network layers and directly learning the residuals of the output of the previous network layer.

The formula mainly adopts the idea of problem transformation, in the structure of the residual neural network before, the solution F(x) is required, with a certain depth increase, the network structure is optimized, to the appropriate depth, the gradient of the network can not be increased, otherwise the update of the weights will become very complex, resulting in an explosion of parameters, so there is no way to train. The residual structure can ensure that the optimal state of this layer is passed through the residual structure, but the optimal solution here is not a global optimal solution, but a suboptimal solution that is very close to the optimal solution, so the residual structure can be used to achieve a suboptimal solution, so that the network can be continuously updated, because each process is a process close to the optimal solution. However, it is worth mentioning that the residual structure is not without drawbacks, as the residual structure cannot solve the optimal solution, and in fact there is a barrier to improve the accuracy by just increasing the depth. As the Table2-1, is the overall architectures of ResNet.

<Table 2-1>ResNet Architectures



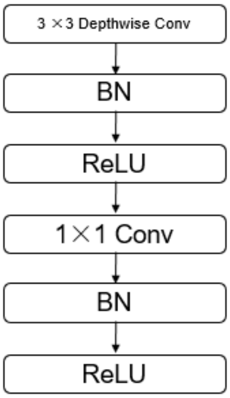
layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix}3\times3, 64\\3\times3, 64\end{bmatrix}\times2$	$\begin{bmatrix}3\times3, 64\\3\times3, 64\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 64\\3\times3, 64\\1\times1, 256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 64\\3\times3, 64\\1\times1, 256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 64\\3\times3, 64\\1\times1, 256\end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix}3\times3, 128\\3\times3, 128\end{bmatrix}\times2$	$\begin{bmatrix}3\times3, 128\\3\times3, 128\end{bmatrix}\times4$	$\begin{bmatrix}1\times1, 128\\3\times3, 128\\1\times1, 512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1, 128\\3\times3, 128\\1\times1, 512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1, 128\\3\times3, 128\\1\times1, 512\end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix}3\times3, 256\\3\times3, 256\end{bmatrix}\times2$	$\begin{bmatrix}3\times3, 256\\3\times3, 256\end{bmatrix}\times6$	$\begin{bmatrix}1\times1, 256\\3\times3, 256\\1\times1, 1024\end{bmatrix}\times6$	$\begin{bmatrix}1\times1, 256\\3\times3, 256\\1\times1, 1024\end{bmatrix}\times23$	$\begin{bmatrix}1\times1, 256\\3\times3, 256\\1\times1, 1024\end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix}3\times3, 512\\3\times3, 512\end{bmatrix}\times2$	$\begin{bmatrix}3\times3, 512\\3\times3, 512\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 512\\3\times3, 512\\1\times1, 2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 512\\3\times3, 512\\1\times1, 2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 512\\3\times3, 512\\1\times1, 2048\end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

2.5.MobileNet

The original purpose of deep learning is to help machines serve people better by training models, so the application of CNNs has been a key factor in considering whether deep image learning can be properly implemented. However, the problems of too large network and difficult to deploy and inconvenient to transplant have been the chronic problems of convolutional neural networks, and the research direction of convolutional neural networks is clear. In the previous chapter of Googlenet, a 1*1 convolutional kernel was used for dimensionality reduction, which is the initial development of lightweight CNNs. The development of lightweight CNNs has been rapid and fruitful, from SqueeNet, Xception to MobileNetV2, and now lightweight CNNs have become an important research direction of CNNs. In this chapter, lightweight CNNS are used for pre-training and then for feature extraction, which can greatly reduce the feature extraction time. This section describes in detail the lightweight network MobileNet.

2.5.1.Depthwise Separable Convolution

In order to ensure the real-time requirements of the network model, it is necessary to adopt a lightweight network, so that the network parameters can be reduced. In this paper, we adopt the depthwise separable convolution based on a CNN to effectively reduce the parameters of the neural network model, because it uses the information of each channel of the input image and the information between channels separately, so under the same conditions, the parameters of the neural network using depthwise separable convolution are greatly reduced than those of the ordinary convolution.



<Figure 2-5> Depthwise Separable Convolution with BN and ReLU

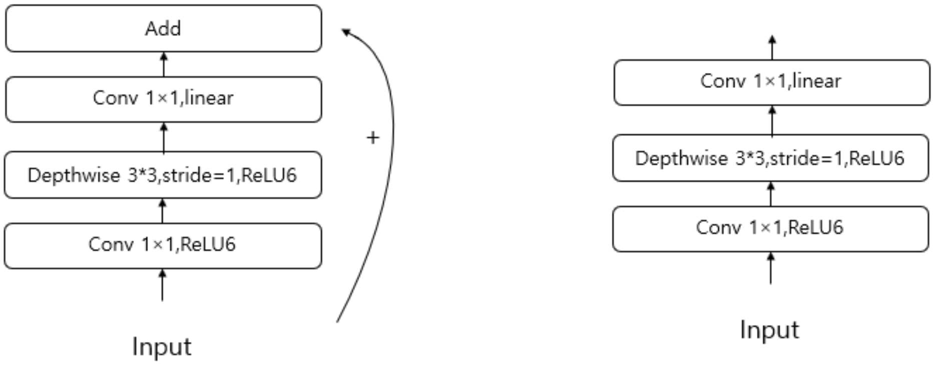
In general, the same convolution kernel is required for all channels in the same layer, but the depthwise separable convolution changes this. It is clear from the structure diagram that the whole depthwise separable convolution step can be divided into two convolution operations. The first convolution operation is to extract the features of each channel of the feature map with different convolution kernels, called depthwise(DW) convolution, which extracts the image feature information of each channel. The feature information of each channel is already available, and it is necessary to fuse all channels into one feature map, after which the second convolution operation is to use 1×1 convolution kernels to achieve fusion. This step is called pointwise(PW) convolution. Depthwise convolution followed by pointwise convolution can greatly reduce the computational effort of the network model, and these two convolutions are followed by a Batch Normalization[48](BN) layer that can normalize the data, which is used to solve the problem of unbalanced data distribution and make the network convergence speed increase significantly. The activation layer is also added to introduce nonlinearity and enhance the ability of the network model to deal with nonlinearity and improve the ability to fit the training parameters, which is also helpful to solve the gradient disappearance problem.

Residual block with bottleneck structure, which contains two convolutional layers for dimensionality reduction and expansion, and an intermediate convolution for feature extraction. This structure is not suitable for lightweight networks because the number of parameters and computational effort of the intermediate convolution is very large.

Depthwise separable convolutions are decomposed into depthwise convolutions, which are used for extracting single-dimensional features, and pointwise convolutions, which are used for linearly combining multi-dimensional features, in order to solve the problem of the number of parameters and the computational effort caused by standard convolutions.

2.5.2.Inverted-Residual block

Inverted residual convolution is a structure proposed in MobileNet V2, which borrows the structure of ResNet block and also uses skip connection, while Resnet uses standard convolution to extract features, and Inverted residual convolution always uses depthwise to extract features. ResNet first descends, convalesces, and then ascends, so the structure is called Inverted Residual Block. This is also done because of the adaptation using DW convolution, hoping that feature extraction can be performed in higher dimensions. The final PW convolution is performed with linear activation, i.e., the Linear Bottleneck mentioned above.The Inverted-Residual block is constructed as follows in Figure 2-6:



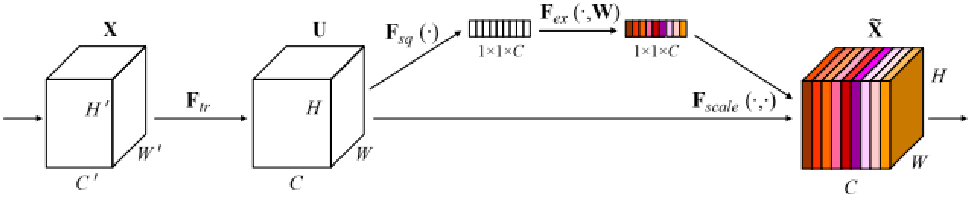
<Figure 2-6> Inverted-Residual block

2.6.Attention Model

The attention module was first used on the Encoder-Decoder model for machine translation (or natural language processing), allowing the network to focus on specific semantic vectors or inputs that have been encoded when processing different parts of speech. In computer vision, there are three general types of attentional mechanisms: channel domain, spatial domain and hybrid domain. The principle is to give different weights to different channels or regions in the space, instead of treating a position in the space ,a position in the space, corresponding to a narrow column of length C in all CxWxH. Or all channels as having the same weight when performing convolution/pooling operations as before, thus allowing the network to focus on the extraction of more important information. The different weight parameters on the convolutional kernel are domain specific, i.e., they can only form a small attention within its receptive field, which becomes larger as the depth increases, but the update of parameters in the shallow and middle layers is indirect and weak. The attention mechanism, on the other hand, focuses on the information of all channels/the whole WxH space and selects the most important parts.

2.6.1.Channel Domain

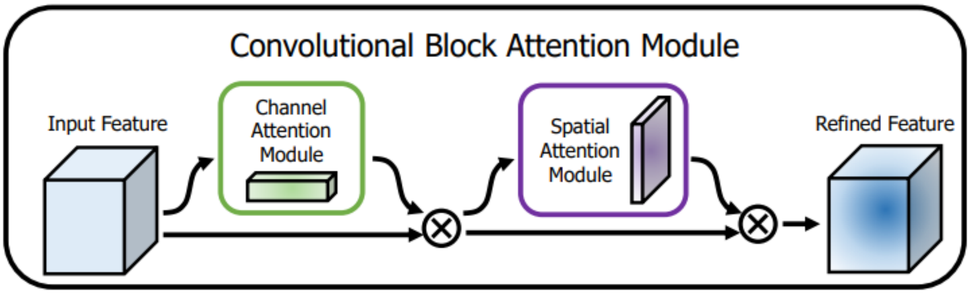
The masterpiece of channel domain attention is SE Net[26] which won the last ImageNet 2017 competition classification task, surpassing last year's winner by 25%, and the model dramatically improved its performance by simply adding a plug-and-play branch of the SE module for modeling correlations between different channels. SE Net is more like a concept that can be applied to all types of network structures. The structure of SE Net is shown in Figure 2-7, which improves feature extraction by reallocating the weights of the feature channels. Firstly, the eigenfunctions of channel number c1 are transformed to obtain a eigenfunctions of channel number c2, and then the eigen compression is performed to compress each two-dimensional eigenfunctions into a real number, which can be interpreted as a global sensing field. Then, a weight vector w is proposed to characterize the correlation between the channels. Finally, this weight vector is used to rescale each feature of the passages so that the passages that contribute more to the classification effect have more weight. SE Net can be embedded into various classification networks, and better classification results are achieved compared to the original model.



<Figure 2-7>A Squeeze-and-Excitation block

2.6.2.Spatial Domain and Hybrid Domain

Attention in the spatial domain is relatively intuitive, and the human visual system has this feature of focusing on a certain region in the visual field to help us capture the most important information quickly. The spatial attention needs to assign a weight to each point in size WxH, so the generated attention is a matrix of attention scores. CBAM [51] is an improvement of SE Net, which adds a spatial attention module to the original channel attention, it is actually a hybrid domain approach, as shown figure2-8, except that we introduce its SAM module here to introduce the spatial domain.

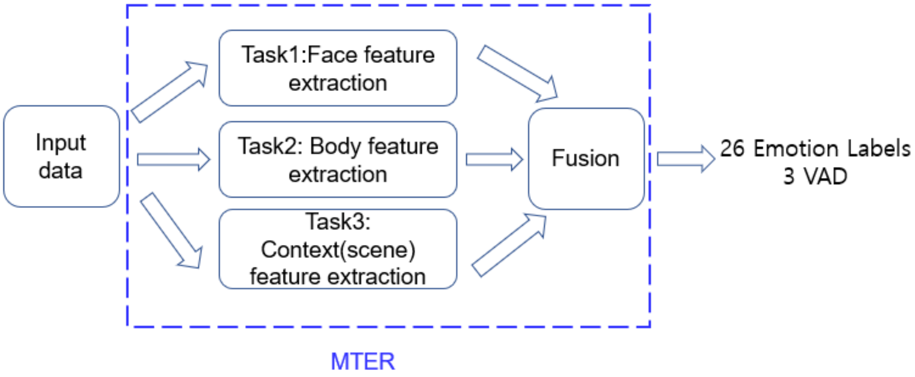


<Figure 2-8>CBAM network structure

Chapter 3.Model Design and Implementation

3.1.Overview

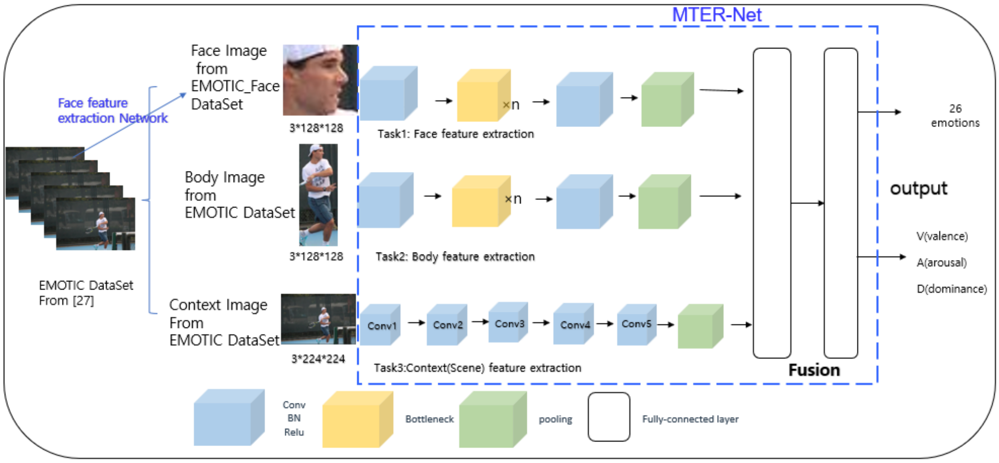
In this section,we describe the CNNs-based feature extraction study enables the simultaneous acquisition of global features and local features through a complete feature extraction framework. The global features mainly reflect the overall information of the sample, the broad contour information of the objects, and the interrelationship between the objects; the local features usually reflect the local information of the objects or the close-up information. With the help of a typical deep learning method, a more robust representation of features can be obtained than traditional manual features. However, how to obtain optimal global and local features is not a natural process, and this issue is the focus of this paper. In this thesis, a lightweight network for CNN feature extraction based on inverted residual convolution is proposed, and we describe our model framework, which mainly contains three independent sub-networks and a fusion network, Our system flowchart is shown in Figure 3-1:


Input data
Task1:Face feature extraction
Task2: Body feature extraction
Task3: Context(scene) feature extraction
Fusion
26 Emotion Labels
3 VAD
MTER

<Figure 3-1>The flowchart proposed for the emotion recognition

3.2.Multi-task Emotion Recognition

Three CNN for visual cues Considering the characteristics and differences of different visual cues, we use different CNNs to extract the features of visual cues. The detailed network structure is shown:


MTER-Net
Face feature extraction Network
Face Image from EMOTIC_Face DataSet
3*128*128
Task1: Face feature extraction
×n
Body Image from EMOTIC DataSet
3*128*128
Task2: Body feature extraction
×n
EMOTIC DataSet From [27]
Context Image From EMOTIC DataSet
3*224*224
Conv1
Conv2
Conv3
Conv4
Conv5
Task3:Context(Scene) feature extraction
Fusion
26 emotions
output
V(valence)
A(arousal)
D(dominance)
Conv BN Relu
Bottleneck
pooling
Fully-connected layer

<Figure 3-2>MTER:Multi-task emotion recognition network structure

3.2.1.Task1:Face-based CNN model

A face features extraction module, which is a pre-trained modified Mobilenet network.The facial emotion of each face in the image is an important cue for emotion recognition. An important cue for emotion prediction is landmark aligned faces. We also consider the case where the face is occluded, so the overall structure has the face part as auxiliary information for recognition. For better recognition of emotions, we use a modified mobilenetv2 as the backbone network. It was fine-tuned on the EMOTIC dataset and pre-trained on the EMOTIC_Face dataset using this model.

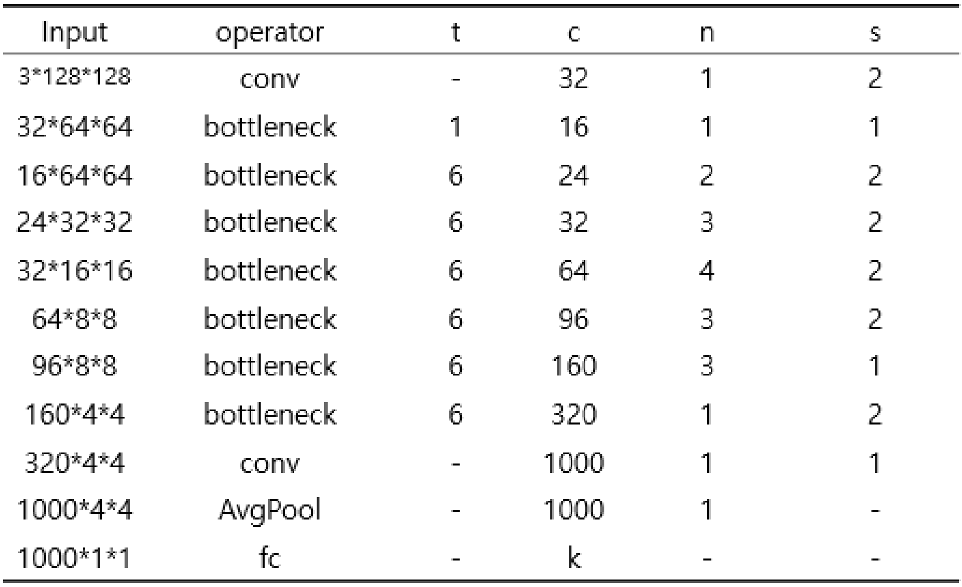
3.2.2.Task2:Body-based CNN model

A model is trained with body regions containing more semantic information, which helps the CNN to learn body features to predict more accurate emotions. And crop the largest external rectangular region of the human body with the largest rectangular region as the body region. Since body information is more complex than face information, we chose to modify it to train the mobilenetV2 body-based CNN model.

The face-based CNN model and the body-based CNN model can solve the problem of too many network parameters by using a normal convolutional layer to extract low-dimensional features in the initial layer, and then placing the depth-separable convolution in the middle layer of the network, and introducing a compression excitation module to make a channel with strong representational ability stronger and a channel with weak representational ability naturally weaker. Since the number of channels in different convolutional layers may be different, the compression excitation module needs to choose different compression rates for different channels, so that the compression excitation module can be used reasonably to improve the accuracy of the network model. To improve the network model recognition effect.

The face feature extraction network and body feature extraction network use a modified network based on mobilenetV2[23], as shown in Table3-1:

<Table 3-1> Task1/2 Structural Details



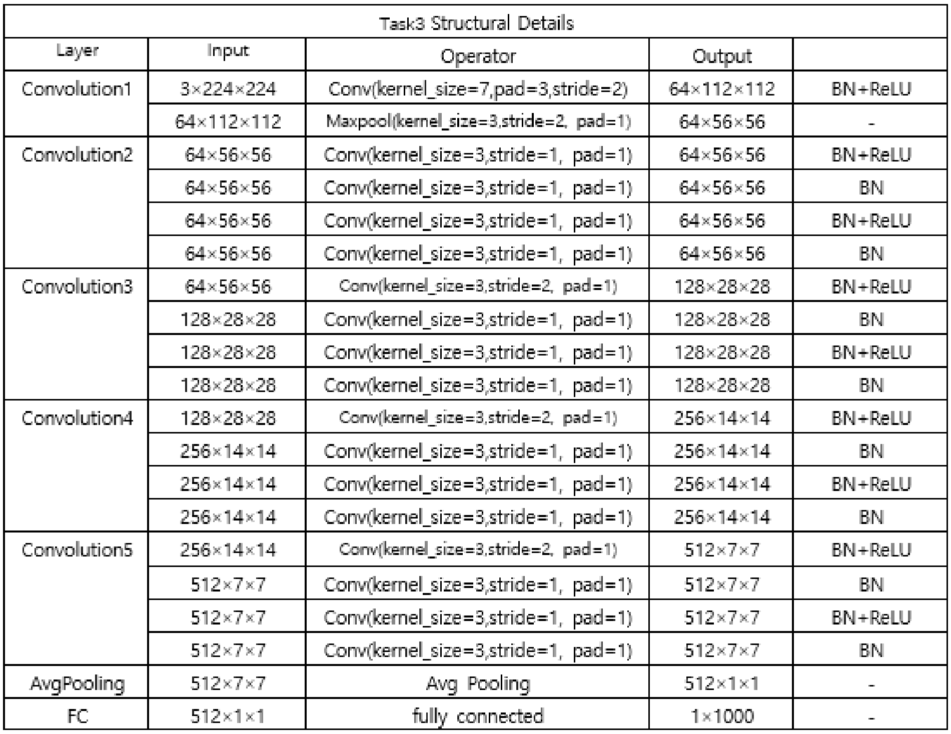
Input	operator	t	c	n	s
3*128*128	conv	-	32	1	2
32*64*64	bottleneck	1	16	1	1
16*64*64	bottleneck	6	24	2	2
24*32*32	bottleneck	6	32	3	2
32*16*16	bottleneck	6	64	4	2
64*8*8	bottleneck	6	96	3	2
96*8*8	bottleneck	6	160	3	1
160*4*4	bottleneck	6	320	1	2
320*4*4	conv	-	1000	1	1
1000*4*4	AvgPool	-	1000	1	-
1000*1*1	fc	-	k	-	-

In Table3-1, t is the expansion factor, in order to extend the dimensionality. c is the number of output channels, n: is the repeating number, s is the stride. 3×3 kernels are used for spatial convolution.The bottleneck of which will be described in detail in section 3.3.

3.2.3.Task3:Context-based CNN model

The global image provides the overall scene and contextual information of the group. The global image provides an overall scene and contextual information. It is more complex and contributes more than the face and body information. Therefore, we come to train our image-based CNN model. We found after experiments that the more lightweight CNN: mobile net does not handle the overall scene information well. Therefore, ResNet18 is chosen as the scene recognition network, using a model that has been pre-trained on the ImageNet. As shown in Table3-2:

<Table 3-2> Task3 Structural Details

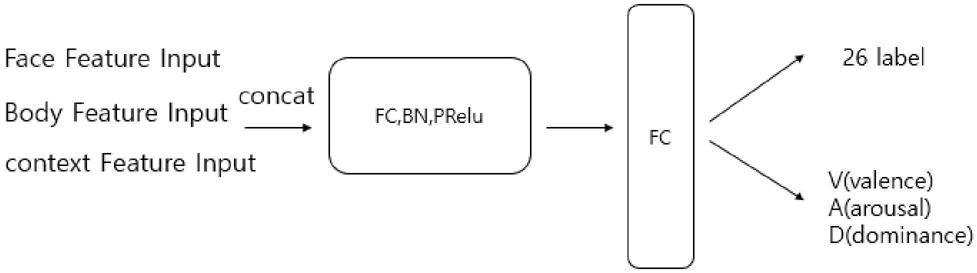


Task3 Structural Details

Layer	Input	Operator	Output	
Convolution1	3×224×224	Conv(kernel_size=7,pad=3,stride=2)	64×112×112	BN+ReLU
	64×112×112	Maxpool(kernel_size=3,stride=2, pad=1)	64×56×56	-
Convolution2	64×56×56	Conv(kernel_size=3,stride=1, pad=1)	64×56×56	BN+ReLU
	64×56×56	Conv(kernel_size=3,stride=1, pad=1)	64×56×56	BN
	64×56×56	Conv(kernel_size=3,stride=1, pad=1)	64×56×56	BN+ReLU
	64×56×56	Conv(kernel_size=3,stride=1, pad=1)	64×56×56	BN
Convolution3	64×56×56	Conv(kernel_size=3,stride=2, pad=1)	128×28×28	BN+ReLU
	128×28×28	Conv(kernel_size=3,stride=1, pad=1)	128×28×28	BN
	128×28×28	Conv(kernel_size=3,stride=1, pad=1)	128×28×28	BN+ReLU
	128×28×28	Conv(kernel_size=3,stride=1, pad=1)	128×28×28	BN
Convolution4	128×28×28	Conv(kernel_size=3,stride=2, pad=1)	256×14×14	BN+ReLU
	256×14×14	Conv(kernel_size=3,stride=1, pad=1)	256×14×14	BN
	256×14×14	Conv(kernel_size=3,stride=1, pad=1)	256×14×14	BN+ReLU
	256×14×14	Conv(kernel_size=3,stride=1, pad=1)	256×14×14	BN
Convolution5	256×14×14	Conv(kernel_size=3,stride=2, pad=1)	512×7×7	BN+ReLU
	512×7×7	Conv(kernel_size=3,stride=1, pad=1)	512×7×7	BN
	512×7×7	Conv(kernel_size=3,stride=1, pad=1)	512×7×7	BN+ReLU
	512×7×7	Conv(kernel_size=3,stride=1, pad=1)	512×7×7	BN
AvgPooling	512×7×7	Avg Pooling	512×1×1	-
FC	512×1×1	fully connected	1×1000	-

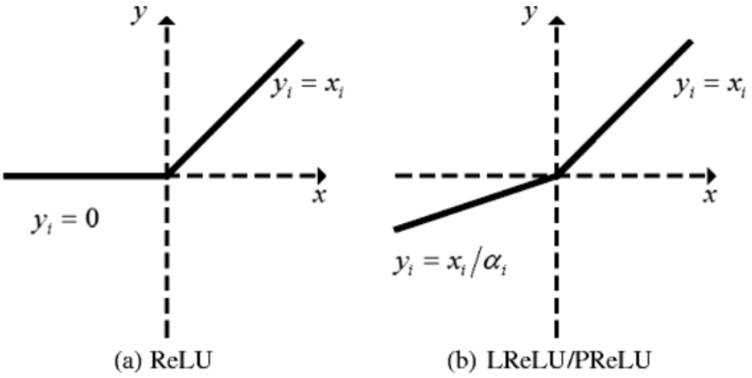
3.2.4.Fusion model

Finally, all the extracted features are fused and concat.



<Figure 3-3> Fusion Network Details

The PReLU function is selected as the nonlinear activation function of the fusion network in this module. PReLU, the ReLU function with parameters, has the function image shown in Figure 3-4.



<Figure 3-4> (a)ReLU and (b)PReLU

When xi ≥ 0, the function image of PReLU is the same as that of the ReLU function, and the gradient $f'(xi)$ is 1, which effectively avoids the problem of gradient explosion and disappearance. In contrast, when the input is negative, i.e., xi < 0, ReLU will fall into an inactive state, and the gradient will become completely 0 during the backpropagation of the convolutional network with a negative input, which has the same problem as the traditional Sigmoid-like unit.PReLU is different from ReLU in the negative region, and there is a gradient parameter ai, which can solve the problem of disappearing neurons in the negative region. The expression of PReLU is as follows.

$$PReLU = \begin{cases} \boldsymbol{x_i} & \boldsymbol{x_i \geq 0} \\ \boldsymbol{a_i x_i} & \boldsymbol{x_i < 0} \end{cases} \qquad (3.1)$$

where xi is the input of the nonlinear activation function on the ith channel and ai is the slope coefficient of the negative region. When ai = 0

the PReLU function degenerates to the ReLU function. If ai is small and fixed, the PReLU function becomes LReLU (ai = 0.01). the original intention of LReLU is also to avoid zero gradients. However, after an experimental comparison between the LReLU function and the ReLU function, the difference in accuracy is found to be minimal, while PReLU effectively improves the accuracy after introducing a very small number of additional parameters. PReLU can be trained together using the backpropagation algorithm and optimized simultaneously with the other layers. ai's update formula can be derived from the chain rule, and the gradient of ai for one layer is:

$$\boldsymbol{\frac{\partial \varepsilon}{\partial a_i}} = \sum\nolimits_{x_i} \boldsymbol{\frac{\partial \varepsilon}{\partial f(x_i)} \frac{\partial f(x_i)}{\partial a_i}} \qquad (3.2)$$

where ε is the objective function; ∂ε/(∂f(x_i)) is the gradient propagated from the update layer. The expression of the gradient of the activation function is calculated as follows.

$$\boldsymbol{\frac{\partial f(x_i)}{\partial a_i}} = \begin{cases} \boldsymbol{0\ x_i \geq 0} \\ \boldsymbol{x_i\ x_i < 0} \end{cases} \qquad (3.3)$$

$\boldsymbol{a_i}$ of the update using the momentum method, expressed as follows:

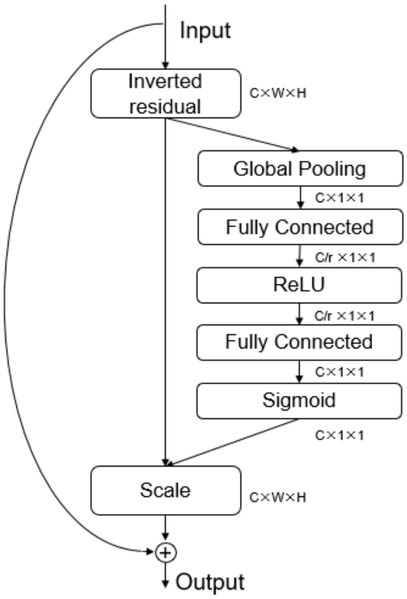
$$\Delta \boldsymbol{a_i} := \boldsymbol{\mu} \Delta \boldsymbol{a_i} + \boldsymbol{\epsilon \frac{\partial \varepsilon}{\partial a_i}} \qquad (3.4)$$

where μ is the momentum coefficient and ϵ is the learning rate. Since the learning rate is mostly less than 1, the PReLU function is commonly initialized with a_i= 0.25.

3.3.Block of Fused Attention Module

If we want a machine to read a picture like a human, we prioritize attention to local information, so the attention mechanism is applied to enhance the feature extraction network. We add the attention mechanism to the effective feature layer extracted from the backbone network, and the model proposed in this paper can increase the category output of the detection, and implement the detection of faces being occluded in a model in order to directly and effectively recognize emotions and process them accordingly, reduce the impact of occlusion on emotion recognition, improve the impact of context on emotion recognition, speed up the model recognition process, and improve the overall recognition efficiency of the model.

The compressed excitation module is essentially a kind of attention network, and there is a certain correlation between the channels of the convolutional neural network, according to which a model can be built and the relationship between the channels can be understood at a glance. Each channel is a value in the model, and the proportion of weights is assigned according to the size of each channel in the model, and the stronger the feature, the larger the corresponding model value. By placing the compressed excitation module in the middle layer of the network structure, the perceptual field can be enlarged and the extracted features can be more prominent. The structure of the compressed excitation module incorporating the inverse residual block is shown in Figure 3-5:



<Figure 3-5> Inverted-SE residual block

As shown in the figure, the structure of the Squeeze and Excitation module has three main parts, namely Squeeze, Excitation, and Scale. It is given an input of feature map X, which is a feature map of size W×H with C channels and is compressed by the excitation module to produce a strong feature map with more prominent features. This strong feature map is mainly because it can effectively combine the information extracted by each convolution kernel because the common convolution operation is fixed by the convolution kernel, that is, its perceptual field is already determined and localized, if the feature map extracted by all convolution kernels can be combined to achieve a global effect. Specifically, the compression part is the global average pooling (GAP) of the feature map, which compresses the C feature maps of size W×H into a series of C values, each representing a channel feature, which is equivalent to the compression of the spatial dimension of the feature map. After the compression, two fully connected layers are used to model the series, find the correlation between each channel, and output the proportional weight of each channel according to this correlation. The first fully connected layer is used to reduce the feature dimensionality to 1/r of the original one, where r represents the compression rate, and a ReLU activation function is indirectly applied to both fully connected layers. The second fully connected layer is used to raise the feature dimension back to the initial dimension, which is the operation step of the excitation part. It is possible to use one fully connected layer for the excitation part, but it is better to use two fully connected layers.

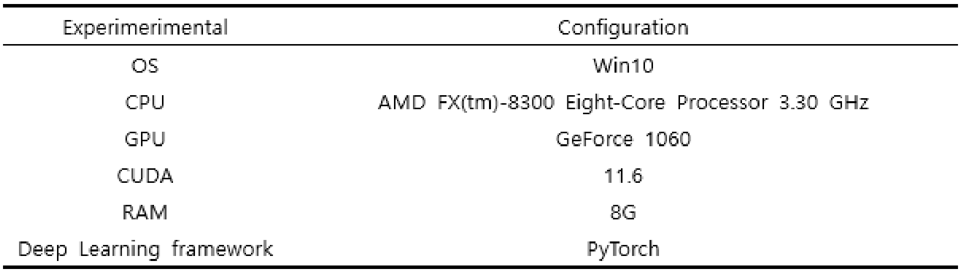
Chapter 4.Experimental procedure

Most of the current commonly used face datasets are focused on small pictures of human faces, while EMOTIC is more suitable for realistic situations without constraints, in a free environment, and where there are large and small parts of human faces, and EMOTIC in which psychologists compress more than 400 emotions into 26 emotions, and there is a disjunction between 26 emotions, expressing all human expressions, which can better reflect human expressions.

4.1.Experimental environment

The experimental environment is shown in Table 4-1:

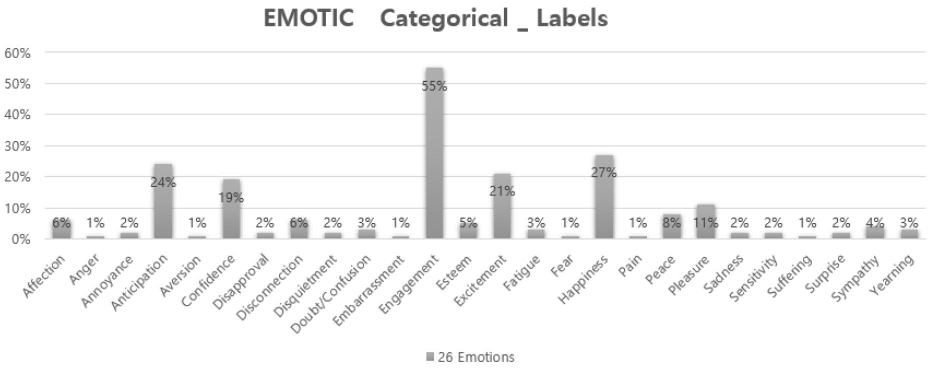
<Table 4-1> Experimental environment



Experimerimental	Configuration
OS	Win10
CPU	AMD FX(tm)-8300 Eight-Core Processor 3.30 GHz
GPU	GeForce 1060
CUDA	11.6
RAM	8G
Deep Learning framework	PyTorch

4.2.EMOTIC Dataset

Dataset of images with people in real environments, annotated with their apparent emotions. The EMOTIC dataset Collected 23,571 images and 34,320 annotated people. Some of the images were manually collected from the Internet by the Google search engine and some from public datasets: COCO and Ade20k.It annotated on 26 emotions and their degree of arousal,valence,and dominance.And containing various places, social environments, different activities, and a variety of keywords on emotional states. Overall, the images show a wide diversity of contexts, containing people in different places, social settings, and doing different activities. These 26 expressions include all the expressions that people may have. Human emotions are complex and continuous.So EMOTIC dataset is suitable to ues it.But the dataset is unbalanced(see Figure 4-1) that it is challenging to classify these many of categories in an accurate manner.

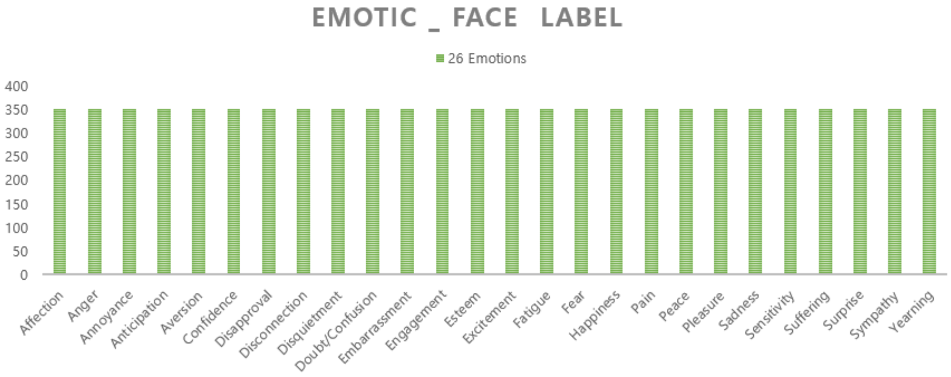


<Figure 4-1>Label distribution of the EMOTIC Dataset

4.3.EMOTIC_Face Dataset

In our proposed method, there are face features extracted to help context recognition, but in the EMOTIC dataset, there is no face feature information, so a new dataset is made: EMOTIC_Face. at this time, there are only 7 emotions in the general emotion dataset, while EMOTIC_Face is a face emotion dataset that takes into account the context information. Face is a dataset based on EMOTIC, which is the face part cut from the EMOTIC dataset.

The specific operation is to iterate through all the photos in the EMOTIC dataset [27], use YOLOv5-Face to detect them, if there is a face, crop the face down and save the face location information. If there is no face, the index is saved by default, but the content is empty. In the EMOTIC_Face dataset, because about 25% of the faces in the original dataset EMOTIC are obscured or small, so in order to alleviate the problem of unbalanced labels(see Fig4-1) in the original dataset, the training set is set to 70% of the EMOTIC _face, the validation set is set to 10%, and the test set is set to 20%, the same proportion as the original dataset EMOTIC.



<Figure 4-2>Label distribution of the EMOTIC_Face Dataset

4.4.Training process

1） The training process uses a stochastic gradient descent optimization algorithm with momentum, with a momentum size of 0.9, a weight decay coefficient of 0.0005, and a batch size of 16 . The learning rate decreases in a warm-up manner and the learning rate is set to 0.001 in the initial 10 rounds, decreasing by a factor of 10 at rounds 75, 100, and 115, respectively, and stopping the training at round 120.

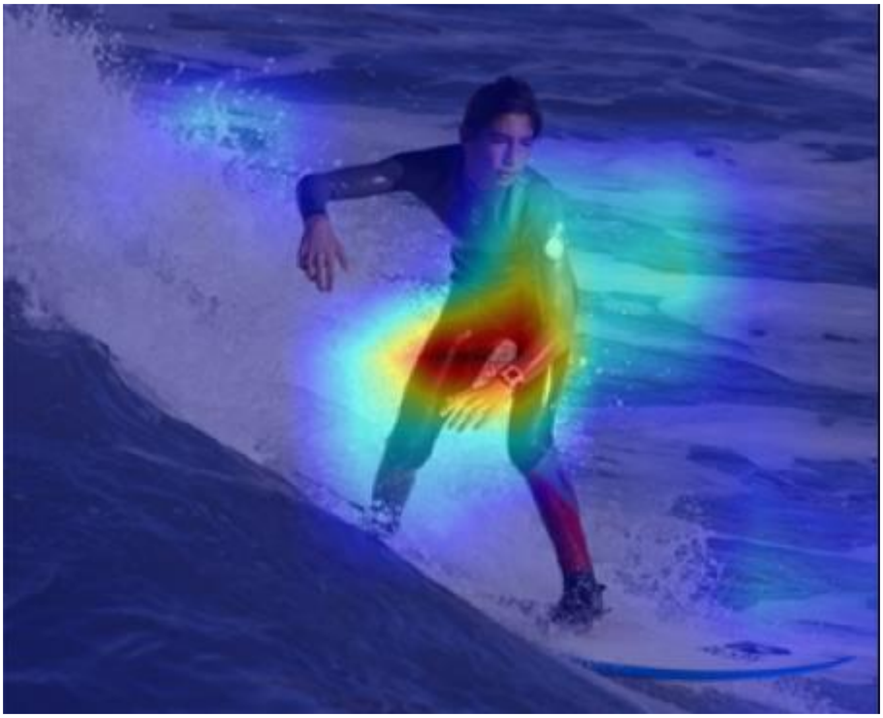
2）During the training process, when the intersection over union (IoU) ratio between the anchor frames and the body frames in the labels is greater than 0.5, it is considered a positive sample, and when the IoU is less than 0.3, it is considered a negative sample, and the rest of the anchor frames are ignored during training. This strategy will lead to more than 99% of the anchor boxes being negative samples, and there is sample imbalance. Therefore, this paper adopts the online hard case mining method to control the ratio of positive and negative samples in the training process.

3) Because data widening can effectively improve the model performance, this paper crops a square area by the size of 0.3~1 of the short side length of the original image, and scales the image to 128×128 to produce more face images in this way to improve the robustness of the model.

4) For each training sample, this paper performs random horizontal mirroring or color distortion with 50% probability. The algorithm with a backbone network of fine-tuned MobileNet trained using the same training method and data set to obtain a model is the new model.

4.5. CNN interpretability study

The core idea of deep learning is to teach the human mind to the network, and since the machine can only accept numbers as input, in fact the machine can only accept strategies that tend to be more in the direction of operational priorities, etc. Deep learning can guide feature ranking and importance ranking in certain domains, which coincides with interpretability. For example, when people analyze a lot of data, they can have an a priori ranking of the importance of features, which is also the focus of interpretability research. In general, data interpretability is well understood, and it is mainly about ranking the input features of the network, which can help decision making to a certain extent. The interpretability of images is different from the interpretability of data, because the images need to be extracted during the training process, and the convolutional neural network is originally a black-box model, so the extracted features cannot be interpreted well, so the development of interpretability of convolutional neural network is slow. In order to study the interpretability of convolutional neural networks, the main method used nowadays is the feature mapping method, which maps the previously extracted features by a global average pooling layer. Selvaraju et al.[52] have adopted the Gradient-weighted Class Activation Mapping (Grad-CAM) approach to study the interpretability of convolutional neural networks to address the shortcomings of the class activation mapping approach. The Grad-CAM approach is consistent with the class activation mapping, but it uses the global average gradient to obtain the weight, sums this weight with the feature map to obtain the feature importance, and finally visualizes the feature importance of the convolutional neural network in the form of a heat map. The interpretive description of Grad-CAM is shown in Figure 4-3.

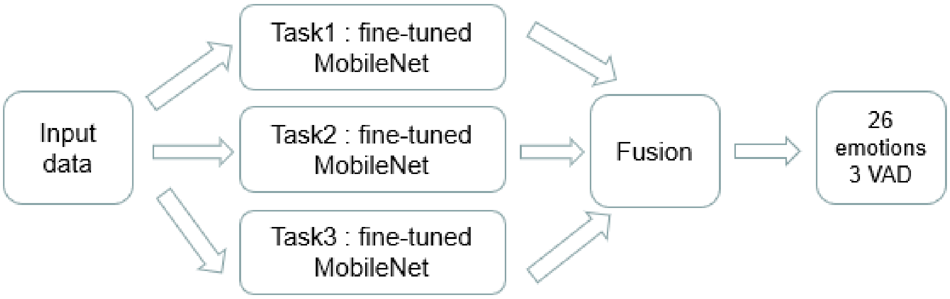


<Figure 4-3> Description of Grad-CAM

4.6.Ablation Experiments

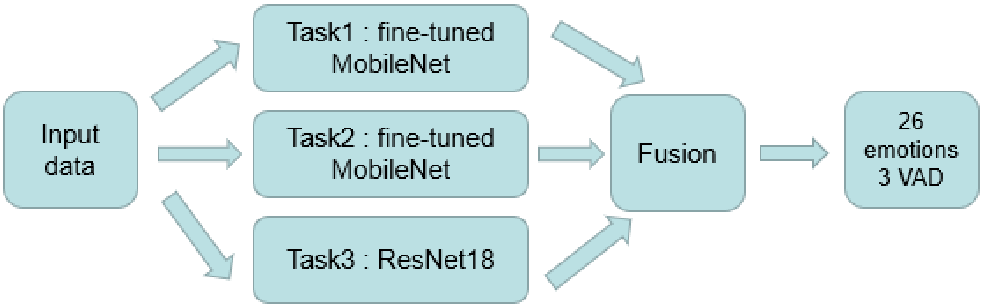
We have done comparative experiments in this paper to propose MTER1,MTER2 for the lightweight model.

In the 3 tasks of MTER1(Figure 4-4), extraction of face and body features and context are used with fine-tuned MobileNet, and then fusion is performed, yielding the result 35.42.

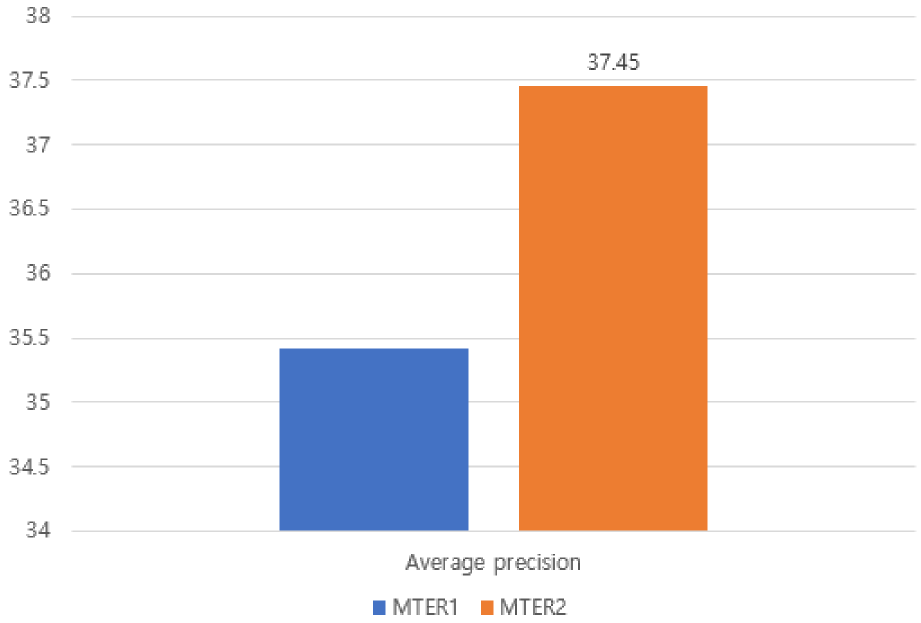

Input data
Task1 : fine-tuned MobileNet
Task2 : fine-tuned MobileNet
Task3 : fine-tuned MobileNet
Fusion
26 emotions 3 VAD

<Figure 4-4> MTER1 Network Structure

In the three tasks of MTER2(Figure 4-5)., the fine-tuned MobileNet is used for extracting face and body features, and ResNet is used for extracting features for context, and then fusion is performed, resulting in 37.45.



<Figure 4-5> MTER2 Network Structure

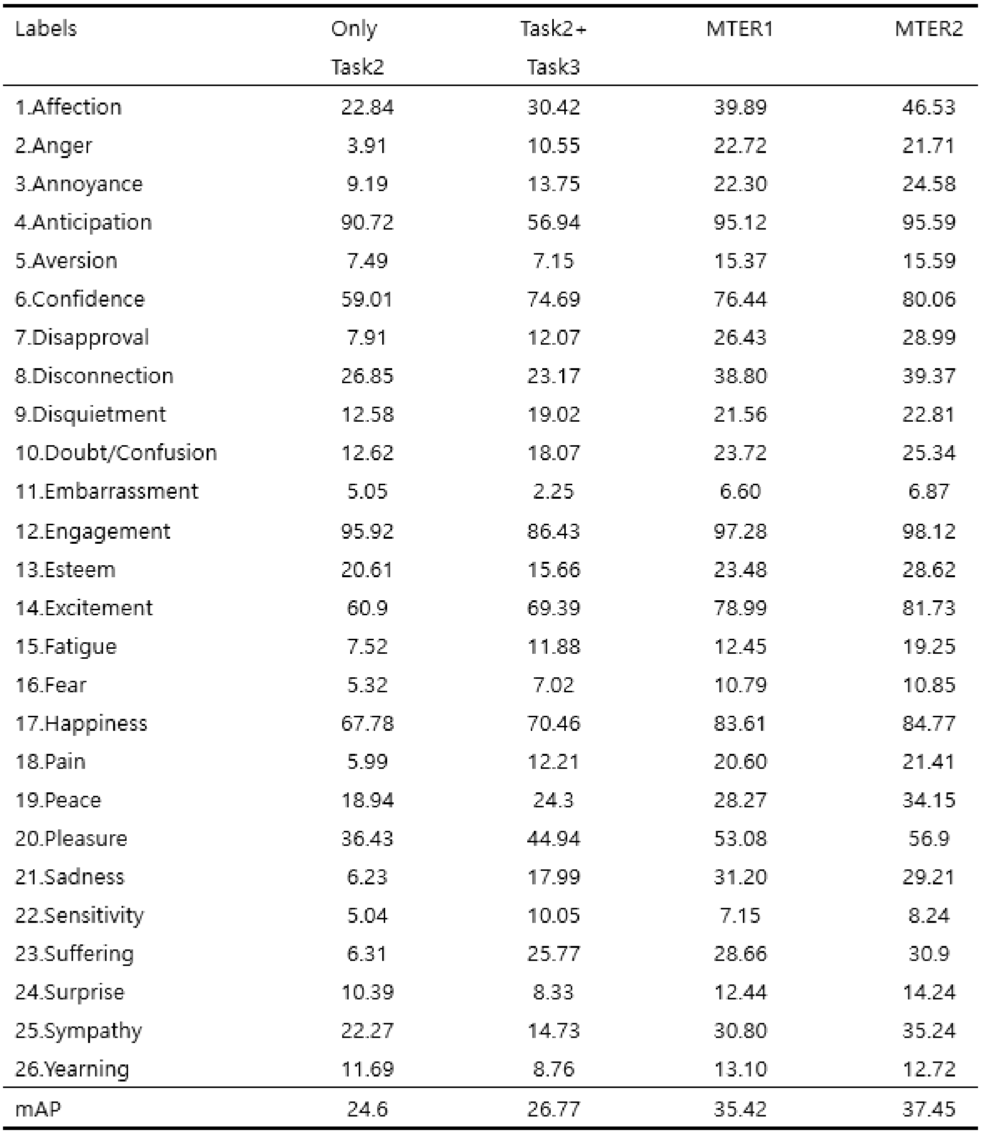


<Figure 4-6> Average Precision of MTER1 and MTER2

Because the face and body are input data, the pre-processed size is 3*128*128 figure, while the full figure is 3*224*224. The calculation volume and parameters are larger than normal. Figure 4-6 are the accuracy comparison results of MTER1 and MTER2.

We have done ablation experiments in this paper to verify the importance of each task. We used the standard metric Average Precision (AP) to evaluate all our methods. The results are shown in Figure 4-7. Because task1 was trained on the EMOTIC_Face dataset, task1 achieved a score of 61% on the EMOTIC_Face test set, because about 25% of the EMOTIC dataset has photos without faces, so simply recognizing face features does not have good performance on the EMOTIC dataset. The main idea of our task1 is to be able to weight if there is a human face, and to ignore the effect of the face if no face is recognized.

<Table 4-2> Ablation Experiments

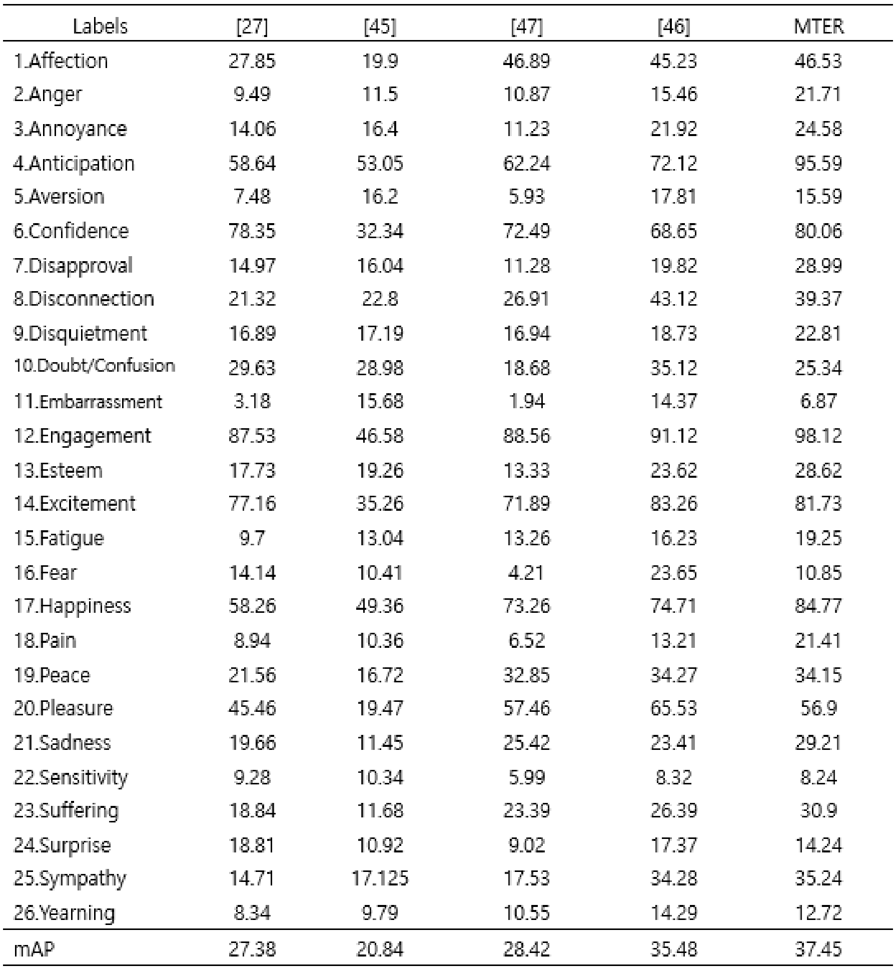


Labels	Only Task2	Task2+ Task3	MTER1	MTER2
1.Affection	22.84	30.42	39.89	46.53
2.Anger	3.91	10.55	22.72	21.71
3.Annoyance	9.19	13.75	22.30	24.58
4.Anticipation	90.72	56.94	95.12	95.59
5.Aversion	7.49	7.15	15.37	15.59
6.Confidence	59.01	74.69	76.44	80.06
7.Disapproval	7.91	12.07	26.43	28.99
8.Disconnection	26.85	23.17	38.80	39.37
9.Disquietment	12.58	19.02	21.56	22.81
10.Doubt/Confusion	12.62	18.07	23.72	25.34
11.Embarrassment	5.05	2.25	6.60	6.87
12.Engagement	95.92	86.43	97.28	98.12
13.Esteem	20.61	15.66	23.48	28.62
14.Excitement	60.9	69.39	78.99	81.73
15.Fatigue	7.52	11.88	12.45	19.25
16.Fear	5.32	7.02	10.79	10.85
17.Happiness	67.78	70.46	83.61	84.77
18.Pain	5.99	12.21	20.60	21.41
19.Peace	18.94	24.3	28.27	34.15
20.Pleasure	36.43	44.94	53.08	56.9
21.Sadness	6.23	17.99	31.20	29.21
22.Sensitivity	5.04	10.05	7.15	8.24
23.Suffering	6.31	25.77	28.66	30.9
24.Surprise	10.39	8.33	12.44	14.24
25.Sympathy	22.27	14.73	30.80	35.24
26.Yearning	11.69	8.76	13.10	12.72
mAP	24.6	26.77	35.42	37.45

4.7.Results on EMOTIC Dataset

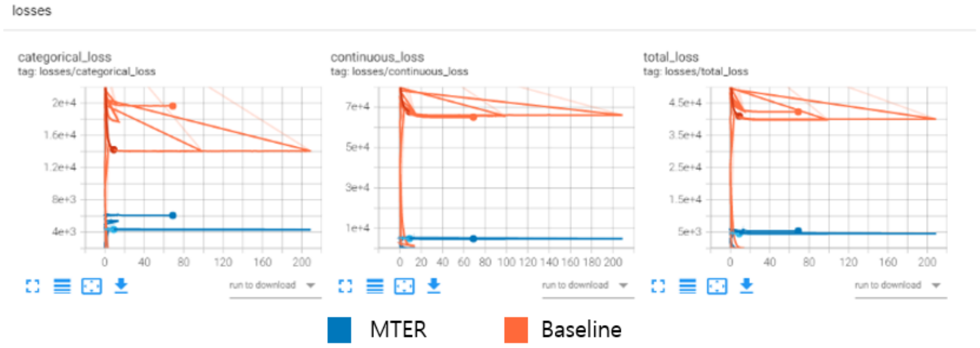
In the research paper on contextualization, we show the superiority of our proposed method and demonstrate that a single method is inferior to our proposed method. And in the context-based emotion recognition study, we compared the recognition effects of existing context-based emotion recognition, as shown in Table4-3:

<Table 4-3> The Comparison on EMOTIC Dataset



Labels	[27]	[45]	[47]	[46]	MTER
1.Affection	27.85	19.9	46.89	45.23	46.53
2.Anger	9.49	11.5	10.87	15.46	21.71
3.Annoyance	14.06	16.4	11.23	21.92	24.58
4.Anticipation	58.64	53.05	62.24	72.12	95.59
5.Aversion	7.48	16.2	5.93	17.81	15.59
6.Confidence	78.35	32.34	72.49	68.65	80.06
7.Disapproval	14.97	16.04	11.28	19.82	28.99
8.Disconnection	21.32	22.8	26.91	43.12	39.37
9.Disquietment	16.89	17.19	16.94	18.73	22.81
10.Doubt/Confusion	29.63	28.98	18.68	35.12	25.34
11.Embarrassment	3.18	15.68	1.94	14.37	6.87
12.Engagement	87.53	46.58	88.56	91.12	98.12
13.Esteem	17.73	19.26	13.33	23.62	28.62
14.Excitement	77.16	35.26	71.89	83.26	81.73
15.Fatigue	9.7	13.04	13.26	16.23	19.25
16.Fear	14.14	10.41	4.21	23.65	10.85
17.Happiness	58.26	49.36	73.26	74.71	84.77
18.Pain	8.94	10.36	6.52	13.21	21.41
19.Peace	21.56	16.72	32.85	34.27	34.15
20.Pleasure	45.46	19.47	57.46	65.53	56.9
21.Sadness	19.66	11.45	25.42	23.41	29.21
22.Sensitivity	9.28	10.34	5.99	8.32	8.24
23.Suffering	18.84	11.68	23.39	26.39	30.9
24.Surprise	18.81	10.92	9.02	17.37	14.24
25.Sympathy	14.71	17.125	17.53	34.28	35.24
26.Yearning	8.34	9.79	10.55	14.29	12.72
mAP	27.38	20.84	28.42	35.48	37.45

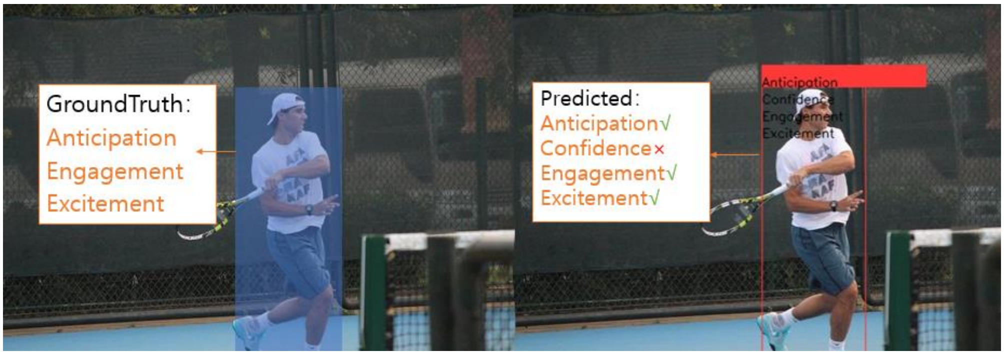
We use [27] as a baseline, and our method not only reduces the number of parameters and computation, but also improves by 10.07%.Figure 4-7 show that our model can converge faster in the same time.



<Figure 4-7> The compared with [27]CNN and MTER losses

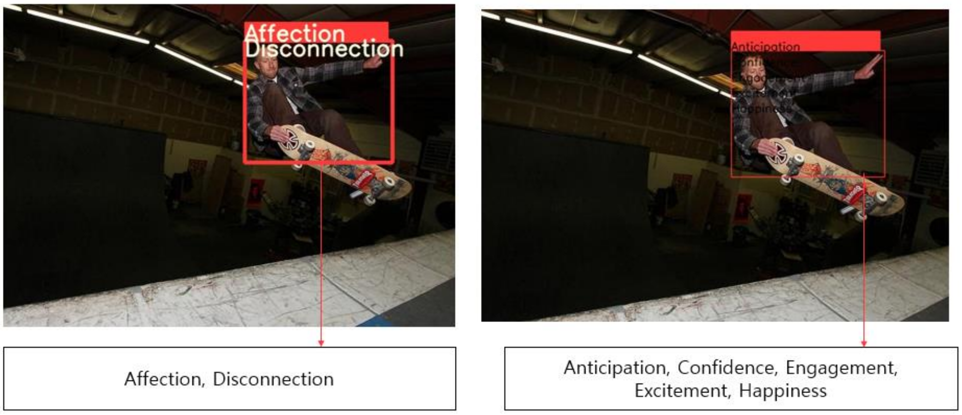
4.8. Analysis of results

Figure 4-8 shows the GroundTruth of the image in the EMOTIC dataset and the prediction results of MTER-Net.



<Figure 4-8> The compared with GroundTruth and MTER Predicted

Even though the recognition is only over 37%, he still has a good result. If without GT, we just look at the recognition results, it is hard to be sure that his sentiment is not right.



<Figure 4-9>[27]CNN (left) and MTER(right)

Back to our initial problem of emotion recognition. Because human emotions are complex, and because of the different cultures and experiences, the emotion calculation of emotion recognition is still a challenge and a difficult problem.Also, because the data set sample labels are uneven, resulting in categories with more labels being more easily recognize.

Chapter 5.Conclusion

In this paper, we proposed a multi-task emotion recognition network (MTER-Net) based on context-aware and attention model, used deep learning techniques to extracted emotion-related information of face, body and context through three tasks, and fused and classified them at the decision level. And it was validated on the EMOTIC dataset. To alleviated the sample imbalance problem in the dataset, the EMOTIC_Face dataset is proposed to provide balanced samples for multi-label emotions, and our network obtained higher recognition accuracy when compared with the CNN model without attention mechanism. Moreover, by used depthwise separable convolution, it makes our model size smaller and the time on inference was reduced with the deep network. For the comparison of fused features, we focus more on natural emotion recognition in real situations than facial expression recognition played in laboratory situations, and by designing a network architecture and a data-driven automatic feature learning approach, we can effectively extract emotion features to improve the recognition accuracy. The experimental results show that the method has good generalization ability and applicability to real scenes; the emotion information of expression, gesture expression and contextual representation has good complementary effects, and the combined used can improve the reliability of emotion recognition. For emotion recognition, the automatic extraction of features used deep learning method has achieved good results, and the analysis of three main streams of emotion cue data based on AI technology: face, body movement and contextual perception has achieved some success in practice.

Future work will focus on extracting complex emotional features to make the machine more intelligent in recognizing human emotions, combining multimodal emotion recognition, thus segmenting and subclassing complex emotions in temporal order, and pairing this model with other hardware devices such as sensors, which will quantitatively assign values to individual mental states (emotions). This model will provide a reliable scientific basis for aiding future emotion recognition technology, thereby recognizing human emotions more intelligently, as well as improving the accuracy of emotion recognition.

References

Copeland, B.J.. "artificial intelligence". Encyclopedia Britannica, 18 Mar. 2022, https://www.britannica.com/technology/artificial-intelligence. Accessed 6 June 2022.

Fifty Years of Signal Processing: The IEEE Signal Processing Society and its Technologies, 1948–1998. The IEEE Signal Processing Society. 1998.

K. Murphy, Machine Learning: A Probabilistic Perspective (MIT Press, Cambridge, MA, 2012).

Juang, B.-H.; Rabiner, L.R. (2006), "Speech Recognition, Automatic: History", Encyclopedia of Language & Linguistics, Elsevier, pp. 806–819, doi:10.1016/b0-08-044854-2/00906-8, ISBN 9780080448541.

https://en.wikipedia.org/wiki/Computer_vision.

Miyakoshi, Yoshihiro, and Shohei Kato. "Facial Emotion Detection Considering Partial Occlusion Of Face Using Bayesian Network". Computers and Informatics (2011): 96–101.

Hari Krishna Vydana, P. Phani Kumar, K. Sri Rama Krishna, and Anil Kumar Vuppala. "Improved emotion recognition using GMM-UBMs". 2015 International Conference on Signal Processing and Communication Engineering Systems.

B. Schuller, G. Rigoll M. Lang. "Hidden Markov model-based speech emotion recognition". ICME '03. Proceedings. 2003 International Conference

Efficient Processing of Deep Neural Networks: A Tutorial and Survey, 20 November 2017, IEEE,2295 – 2329.

Hewett; Baecker; Card; Carey; Gasen; Mantei; Perlman; Strong; Verplank. "ACM SIGCHI Curricula for Human–Computer Interaction". ACM SIGCHI.

Fukushima, K., 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 36(4), 193-202.

Krizhevsky, Alex; Sutskever, Ilya; Hinton, Geoffrey E. (June 2017). "ImageNet classification with deep convolutional neural networks" (PDF). Communications of the ACM. 60 (6): 84–90. doi:10.1145/3065386. ISSN 0001-0782.

Chen, V. (February 2011). Micro-Doppler Effect in Radar. Norwood, MA: Artec House. pp. 18–21. ISBN 9781608070589.

Beat Fasel. Head-Pose Invariant Facial Expression Recognition Using Convolutional Neural Networks[P]. Multimodal Interfaces,2002.

Krizhevsky A， Sutskever I， Hinton G E， et al. ImageNet classification with deep convolutional neural networks[C]. neural information processing systems， 2012: 1097-1105.

Russakovsky, Olga; Deng, Jia; Su, Hao; Krause, Jonathan; Satheesh, Sanjeev; Ma, Sean; Huang, Zhiheng; Karpathy, Andrej; Khosla, Aditya; Bernstein, Michael; Berg, Alexander C. (December 2015). "ImageNet Large Scale Visual Recognition Challenge". International Journal of Computer Vision. 115 (3):211252. doi:10.1007/s112630150816y. hdl:1721.1/104944. ISSN 0920-5691. S2CID 2930547.

K. Simonyan, and A. Zisserman, “Deep Convolutional Networks for Large-Scale Image Recognition,” ICLR, pp. 1-14, 2015.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich; Going Deeper With Convolutions.Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1-9.

Quan T. Ngo，Seokhoon Yoon. Facial Expression Recognition Based on Weighted-Cluster Loss and Deep Transfer Learning Using a Highly Imbalanced Dataset[J]. Sensors，2020.20(9).

Hoffman J, Tzeng E, Darrell T, et al. Simultaneous Deep Transfer Across Domains and Tasks[M]. IEEE, 2017.

K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual Learning for Image Recognition," Computer Vision and Pattern Recognition (CVPR) Las Vegas, pp. 770-778, November 2016.

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv preprint arXiv:1704.04861.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., " MobileNetV2: Inverted Residuals and Linear Bottlenecks", Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510-4520.

Tan, M., & Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946.

Pablo Barros ， Nikhil Churamani ， Alessandra Sciutti. The FaceChannel: A Fast and Furious Deep Neural Network for Facial Expression Recognition[J]. SN Computer Science，2020，1(6) .

Hu J., Shen L. and Sun G., “Squeeze-and-excitation networks,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132-7141, 2018.

Ronak Kosti,Jose Alvarez,Adria Recasens,and Agata Lapedriza.Context-Based Emotion Recognition Using EMOTIC Dataset， IEEE transactions on pattern analysis and machine intelligence,2019.

Dhall A, Goecke R, Lucey S, et al. Collecting large, richly annotated facial-expression databases from movies[J]. IEEE Multimedia, 2012 (3): 34-41.

Meng Z, Liu P, Cai J, et al. Identity-aware convolutional neural network for facial expression recognition[C]//2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, 2017: 558-565.

Canedo D, Neves A J R. Facial Expression Recognition Using Computer Vision: A Systematic Review[J]. Applied Sciences, 2019, 9(21): 4678.

Li S, Deng W. Deep facial expression recognition: A survey[J]. IEEE Transactions on Affective Computing, 2020.

M. Singh, M. M. Singh, and N. Singhal, "Ann based emotion recognition," Emotion, no. 1, pp. 56–60, 2013.

C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, “Emotion recognition using a hierarchical binary decision tree approach,” Speech Communication, vol. 53, no. 9, pp. 1162–1171, 2011.

N. E. Cibau, E. M. Albornoz, and H. L. Rufiner, “Speech emotion recognition using a deep autoencoder,” Anales de la XV Reunion de Procesamiento de la Informacion y Control, vol. 16, pp. 934– 939, 2013.

Siddharth, Tzyy-Ping Jung, and Terrence J Sejnowski. Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing. IEEE Transactions on Affective Computing, 2019.

F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari. 2019. Audio-Visual Emotion Recognition in Video Clips. IEEE Transactions on Affective Computing 10, 1 (2019), 60–75.

Hatice Gunes and Massimo Piccardi. 2006. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In 18th International Conference on Pattern Recognition (ICPR’06), Vol. 1. IEEE, 1148– 1153.

Soroush, M.Z., Maghooli, K., Setarehdan, S.K. & Nasrabadi, A.M. 2017, A Review on EEG Signals Based Emotion Recognition.

David Hairston, W., Whitaker, K.W., Ries, A.J., Vettel, J.M., Cortney Bradford, J., Kerick, S.E., and McDowell, K.: 'Usability of four commercially-oriented EEG systems', Journal of Neural Engineering, 2014, 11, (4), pp. 046018.

Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. 2018. EEG emotion recognition using dynamical graph convolutional neural networks. IEEE Transactions on Affective Computing (2018).

Ziyu Jia, Youfang Lin, Jing Wang, Ronghao Zhou, Xiaojun Ning, Yuanlai He, and Yaoshuai Zhao. Graphsleepnet: Adaptive spatial-temporal graph convolutional networks for sleep stage classification. In Christian Bessiere, editor, Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, pages 1324–1330. ijcai.org, 2020.

Yang Li, Wenming Zheng, Lei Wang, Yuan Zong, and Zhen Cui. From regional to global brain: A novel hierarchical spatial-temporal neural network model for eeg emotion recognition. IEEE Transactions on Affective Computing, 2019.

Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In proceedings of the IEEE International Conference on Computer Vision, pages 5533–5541, 2017.

Batja Mesquita and Michael Boiger. Emotions in context: A sociodynamic model of emotions. Emotion Review, 6(4):298–302, 2014.

Jiyoung Lee,Seungryong Kim,Sunok Kim,Jungin Park,and Kwanghoon Sohn.Context-Aware Emotion Recognition Networks.arXiv preprint arxiv:1908.05913,2019.

Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, Dinesh Manocha; EmotiCon: Context-Aware Multimodal Emotion Recognition Using Frege's Principle.Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14234-14243.

Minghui Zhang,Yumeng Liang,and Huadong Ma.Context-Aware Affective Graph Reasoning for Emotion Recognition.In 2019 IEEE International Conference on Multimedia and Expo(ICME),pages 151-156,IEEE,2019.

T. Baltrusaitis, P. Robinson and L.-P. Morency, "OpenFace: an open source facial behavior analysis toolkit" in IEEE WACV, 2016.

Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., and Sheikh, Y. (2018). Openpose: realtime multi-person 2D pose estimation using part affinity fields. arXiv preprint arXiv:1812.08008. doi: 10.1109/CVPR.2017.143.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167, 2015.

Sanghyun Woo , Jongchan Park , Joon-Young Lee, CBAM: Convolutional Block Attention Module. In So Kweon; Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3-19.

Selvaraju, Ramprasaath R., et al. ”Grad-cam: Visual explanations from deep networks via gradient-based localization.” Proceedings of the IEEE international conference on computer vision. 2017.

Abstract (in Korean)

상황인식 및 어텐션 모듈 기반의 다중 태스크 감정 인식

장이나
컴퓨터정보통신공학과
건국대학교 대학원

감정 인식은 컴퓨터 비전 분야에서 중요한 연구 주제이며, 연구의 초점과 어려움뿐만 아니라 감정 컴퓨팅의 주요 문제이다. 만약 우리가 다른 사람들의 감정 정보를 감시하고 이해하기 위해 저렴한 기계를 사용할 수 있다면 그것은 우리의 삶에 더 좋은 영향을 미칠 것이다. 하지만 현재 그런 일을 할 수 있는 시스템이 없다. 인간의 감정상태는 말, 표정, 행동, 사람이 사는 환경, 다양한 생리학적 신호 등 다양한 방식으로 표현되기 때문에 하나의 특징 파라미터와 그 특성에 의존해 인간의 감정을 정확하게 반영하기 어렵다. 따라서, 다중 작업 기반 기능 융합은 정확한 감정 인식을 달성하는 효과적인 방법이다. Convolutional Neural Network (CNN)과 머신 러닝 기술의 지속적인 발전으로 인간의 감정 인식은 새로운 단계로 접어들었다. 현재 표준 CNN 은 대부분 일반적인 얼굴 표정 인식에 적용되지만 정보 추출 기능이 제한적이다.

본 논문에서 우리는 시각적 맥락에 기초한 멀티태스킹에 대한 경량 CNN 모델을 설계하는 프레임워크인 Muti-task Emotion Recognition (MTER)을 제안한다. METR 에는 얼굴 특징 추출 모델, 신체 특징 추출 모델 및 컨텍스트(장면) 특징 추출 모델, 융합 분류 모델의 4 가지 주요 모델이 있다. 여러 사람이 포함된 이미지를 분석하고 얼굴 특징, 신체 특징, 맥락 정보를 기반으로 융합된 감정을 인식하는 데 사용된다. 얼굴 특징 및

신체 특징 추출 모듈은 이미지의 얼굴과 신체 부위를 입력으로 취하고 얼굴 표정, 머리 위치, 신체 자세 등 이미지에 내재된 정보를 추출한다. 감정인식을 실생활에 적극적으로 적용하기 위해 모바일넷 경량 네트워크를 활용해 계산 노력을 줄이고 인식 속도를 높인다. 네트워크가 채널 간의 관계에 더 집중하도록 하기 위해 attention model 인 SE(Squeeze-and-Excitation) 여기가 추가되었다. 네트워크 수렴 속도를 높이기 위해 깊이 분리 가능한 컨볼루션 후 배치 정규화(BN)를 추가하고, 완전히 연결된 계층 대신 GAP(Global Average Pooling)을 사용하여 네트워크의 매개변수 수를 줄인다. 우리는 ImageNet 에서 사전 훈련된 미세 조정된 Mobilenet 네트워크를 사용하며 장면 특징 추출 모듈은 두 명 이상을 포함할 수 있는 전체 이미지를 입력으로 사용하며 이러한 특징은 이미지의 주요 측면을 반영하고 인코딩한다. 우리는 ResNet 의 컨볼루션 레이어만 사용하고 ImageNet 에서 사전 교육을 받았다. 형상 추출의 경우 모델의 복잡성을 줄이기 위해 입력 이미지의 크기를 줄인다.

데이터 불균형을 해결하기 위해, 우리는 26 개의 감정을 가진 문맥을 고려한 샘플 균형 얼굴 데이터 세트인 이모티콘_페이스에 대한 작은 샘플 데이터 세트를 개발한다. 우리의 실험 결과는 우리가 제안한 프레임워크가 다른 현장 인식 작업과 비교하여 실현 가능하다는 것을 보여준다.

모델의 전반적인 견고성이 향상되고 실제 시나리오의 EMOTIC 데이터 세트에서 개선된 알고리듬의 효과가 입증되며, 기본 컨볼루션 신경망에 비해 평균 10.07%의 정확도가 향상된다.

주제어 : CNN, 감정 인식, 컨텍스트, 다중 태스크