Alexandria Engineering Journal (2022) 61, 4435–4444

HOSTED BY

Alexandria University

Alexandria Engineering Journal

www.elsevier.com/locate/aej
www.sciencedirect.com

A-MobileNet: An approach of facial expression recognition

Yahui Nan [a,b], Jianguo Ju [a,*], Qingyi Hua [a], Haoming Zhang [a], Bo Wang [a]

[a] School of Information Science and Technology, Northwest University, Xi'an 710027, China
[b] Department of Computer Science and Technology, Lvliang University, Lvliang 033000, China

Received 26 August 2021; revised 18 September 2021; accepted 27 September 2021
Available online 19 October 2021

KEYWORD

Attention module;
Center loss;
Facial expression recognition;
Channel attention;
Spatial attention

Abstract Facial expression recognition (FER) is to separate the specific expression state from the given static image or video to determine the psychological emotions of the recognized object, the realization of the computer's understanding and recognition of facial expressions have fundamentally changed the relationship between human and computer, to achieve better human computer interaction (HCI). In recent years, FER has attracted widespread attention in the fields of HCI, security, communications and driving, and has become one of the research hotspots. In the mobile Internet era, the need for lightweight networking and real-time performance is growing. In this paper, a lightweight A-MobileNet model is proposed. First, the attention module is introduced into the MobileNetV1 model to enhance the local feature extraction of facial expressions. Then, the center loss and softmax loss are combined to optimize the model parameters to reduce intra-class distance and increase inter-class distance. Compared with the original MobileNet series models, our method significantly improves recognition accuracy without increasing the number of model parameters. Compared with others, A- MobileNet model achieves better results on the FERPlus and RAF-DB datasets.

1. Introduction

With the rapid development of network and communication technologies [1,2], a large amount of data and information every day, and collecting, processing and analyzing data and information has great value in various industries [3,4], for example, analyzing data through cloud services for enterprise risk assessment [5] and resource optimization [6]. Collecting face data on the Internet to build datasets for face recognition and its related research, expressions are a subset of face datasets. Expressions are the reactions of human mental states (such as happy, angry, and sad) on the face. Recognizing and using facial expression information can enable applications to provide users with better communication, more personalized services and help. Therefore, FER has become one of the current research hotspots in artificial intelligence, human–computer interaction (HCI), and image recognition.

The FER process consists of three parts: facial image acquisition, and pre-processing, expression feature extraction and expression classification. Therein, the most important part is

* Corresponding author.
Peer review under responsibility of Faculty of Engineering, Alexandria University.

https://doi.org/10.1016/j.aej.2021.09.066

Feature extraction. Traditional methods mostly use hand-designed features, such as LBP [7], LBP-TOP [8], nonnegative matrix factorization (NMF) [9], and sparse learning [10] for FER. Since 2013, FER competitions such as FER2013 (the Facial Expression Recognition 2013 [11]) and EmotiW [12,13] have collected relatively sufficient information from challenging real-world scenes. The training samples promote the transformation of FER from a controlled environment in the laboratory to a natural environment. From the perspective of research objects, the field of expression recognition is experiencing rapid development from laboratory posing to spontaneous expression in the real world, from long-lasting exaggerated expressions to instantaneous micro-expressions, and from basic expression classification to complex expression analysis. Simultaneously, due to the rapid growth of chip-processing capabilities (such as GPU units) and the elaborate design of network architectures, research in various fields has begun to use deep learning to solve various problems, and recognition results far exceeding the previous methods have been obtained [14-16]. Similarly, deep learning technology is gradually applied to facial expression recognition to deal with complex interference factors.

The model of deep learning can automatically extract features and has high-performance feature expression capability. However, with the continuous development of deep neural network, some shortcomings have gradually appeared. The disadvantages of the complexity of the network and numerous parameters make these models only applicable in some specific scenarios. It is difficult for mobile terminals and embedded devices to meet the hardware requirements they need. The high hardware requirements of complex models limit its application scenarios. Therefore, it is natural to think that we must design a model, which can reduce the parameters and computation under the premise of ensuring accuracy so that it can run on the mobile terminal, embedded device, or universal terminal. Since 2017, Google has successively proposed MobileNetV1 [17], MobileNetV2[18], and MobileNetV3[19], which all of these can be applied to mobile and embedded devices. These models have achieved high accuracy on the ImageNet dataset. Therefore, the MobileNet series models are also used to recognize human faces and expressions. Wang et al. [20] mentioned that the original MobileNetV1 and V2 models were directly applied to facial expression recognition tasks, but the recognition results were not significantly improved. Cotter et al. [21] proposed that using MobileNetV2 (Where the hyperparameter alpha = 0.35) as the feature extraction network, followed by a dense network for expression recognition. However, compared with the original MobileNetV2 model, the parameter quantity of this model is not significantly reduced when the recognition accuracy rate has decreased. Zhang et al. [22] proposed the combination of MobileNetV2 and SSD models for FER. This model can not only reduce the parameters of the convolutional network but also improve the recognition accuracy of FER. However, the article did not mention specific comparative experiments. Hu et al. [23] proposed that using island loss and softmax loss to jointly optimize the MobileNetV2 model for expression recognition tasks to improve recognition accuracy. However, most of these methods are improving the model structure or optimizing the loss function, and little attention is paid to the particularity of local information when various facial expressions change. That is, they did not pay attention to the differences in the local features of the expression.

In this paper, by analysing the characteristics of facial expression, when the same person shows different expressions, the differences between them are relatively small, that is, little difference inter-class. When different people show the same expression, the differences between them are large, in other words, much difference intra-class. The key of FER is to extract fine-grained features of local expression changes (for example, eyes and mouth). Attention module was introduced into the lightweight MobileNetV1 model to pay more attention to local expression features. Meanwhile, softmax loss and central loss function are used to optimize model parameters for expression recognition. The Softmax function increases the distance between classes and reduces the cross problem between them. The central loss function reduces internal differences in the same emoticon category and makes the samples of the same expression more similar. From experimental results, compared with MobileNetV1, V2, V3-small models, the proposed A-MobileNet model can effectively improve the expression classification accuracy of FER on the FERPlus and RAF-DB datasets without increasing the amount of model parameters. Compared with others, our method achieves better results.

2. Related work

In this section, we mainly introduce the core features of the MobileNetV1 model to be used, the advantages of the attention module, the optimization of the loss function and propose our improved model architecture combined with the above three parts.

2.1. The related theory

Introduction to the MobileNetV1: Since AlexNet used a convolutional neural network (CNN) for image classification and won the first place in the ImageNet competition in 2012. Researchers have designed more and more deep neural network models, such as classical VGGNet16/19, GoogleNet, ResNet50, and so on. Compared with traditional classification algorithms, these have been excellent. However, as people continue to deepen the network, huge storage pressure and computational burden caused by the model calculations have begun to limit the application field of the deep learning models. Traditional CNN has large memory requirements and a large computational amount, which makes it impossible to run on mobile devices and embedded devices. To this end, Google has proposed a lightweight deep neural network called MobileNetV1. It is a CNN with a smaller model size, less trainable parameters and calculation amount, and is suitable for mobile devices. It takes full advantage of its computing resources and improves the accuracy of the model to the greatest extent.

The core idea of MobileNetV1 network is to replace the standard convolution operation with depthwise separable convolution (DSC) to reduce model parameters. Specifically, DSC is to use the 3×3 convolution kernel with only one layer thickness, sliding layer by layer on the input tensor, and generate an output channel after each convolution. When the convolution was completed, use 1×1 pointwise convolution to adjust the

thickness, as shown in Fig. 1. However, the MobileNetV1 introduces a ReLU activation function after the deep convolution layer, which cannot change the number of channels. The extracted features are single channel, and the operation of the ReLU in the convolution layer output with fewer channels will lead to information loss.

Attention module: In the aspect of FER, the conclusion of literature [24] is that the features of the eye region more influence recognition accuracy than the mouth region. That is to say, different expression changes mainly concentrate on eyes and mouth. Naturally, we can pay more attention to these parts when extracting facial features from the MobileNet network, so we introduce an attention module in the MobileNetV1 network model. Attention mechanism can improve the feature expression ability of the deep network.

The attention module CBAM [25] in the convolution layer is a lightweight attention module, which can conduct attention in the channel and spatial dimensions, as shown in Fig. 2. In most cases, the parameters and computation cost of the CBAM module can be ignored. The CBAM mainly performs the following two operations on it:

$$\begin{aligned} F' &= M_c(F) \otimes F \\ F'' &= M_s(F') \otimes F' \end{aligned} \tag{1}$$

where F represents the input feature. The operator $\otimes$ represents element-wise dot multiplication. M_c and M_s represent the attention extraction operation in the channel and spatial dimension, respectively.

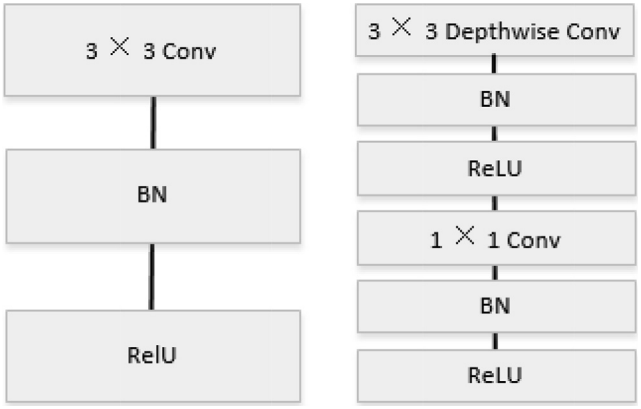

Fig. 1 Standard convolution with batch normalization (BN) and ReLU (left) against DSC with depthwise and pointwise followed by BN and ReLU (right) [17].

The CAM module first performs global max pooling (MaxPool) and global average pooling (AvgPool) on the input feature F to obtain two one-dimensional feature maps and then sends them to a two-layer multi-layer perceptron (MLP). Then, the MLP output features are added based on element-wise, and then the sigmoid activation operation is performed to generate the final channel attention feature map M_c. Finally, the Mc and the input feature F are subjected to an element-wise multiplication operation to generate the input feature F' required by the SAM. The specific calculation of the Attention module on the channel is shown in Eq. (2). The CAM module is shown in Fig. 3.

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma\left(W_1\left(W_0\left(F^c_{avg}\right)\right) + W_1\left(W_0\left(F^c_{\max}\right)\right)\right) \end{aligned} \tag{2}$$

where F represents the input feature map, F^c_{avg} and $F^c_{\max}$ represent the feature calculated by global AvgPool and MaxPool. W_0 and W_1 represent the two-layer parameters of MLP.

The CAM module output feature map F' is used as the input of the SAM module. The SAM module first conducts channel-based global MaxPool and AvgPool, and then concatenates the two feature maps over the channel. The spatial attention feature map M_s is generated by activation of sigmoid function. Finally, the M_s and the input of this module are multiplied to get the final generated feature map. As shown in Fig. 4. The processing of the SAM module is expressed by mathematical Eq. (3):

$$\begin{aligned} M_s(F) &= \sigma\left(f^{7\times7}([AvgPool(F); MaxPool(F)])\right) \\ &= \sigma\left(f^{7\times7}\left(\left[F^s_{avg}; F^s_{\max}\right]\right)\right) \end{aligned} \tag{3}$$

where σ represents the sigmoid activation function, the convolution layer shown in this section uses a 7×7 convolution kernel. In this experiment, we use the sequential arrangement of the two modules. The CBAM module embedded in the MobileNet network is shown in Fig. 5.

The Softmax loss (L_s): The L_s function is a generalization of the logistic function [26], which was mainly used to solve multiple classification problems. The results obtained by softmax represent the probability that the input image was classified into each class, if it is a K classifier, the output is a k-dimensional vector (the sum of elements in the vector is 1).

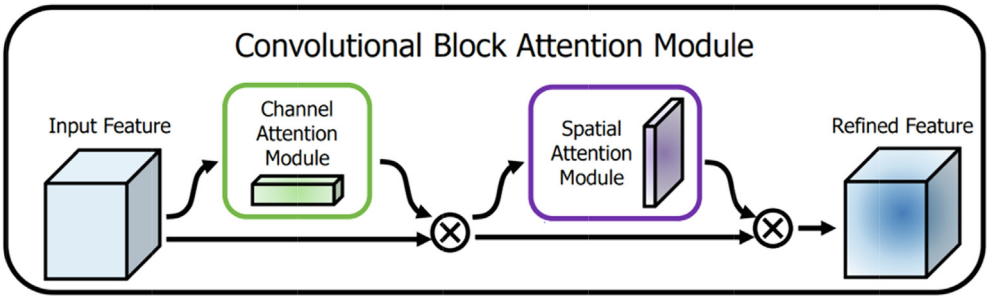

Fig. 2 The overview of convolution block attention (CBAM). This module contains channel attention module (CAM) and spatial attention module (SAM) [25].

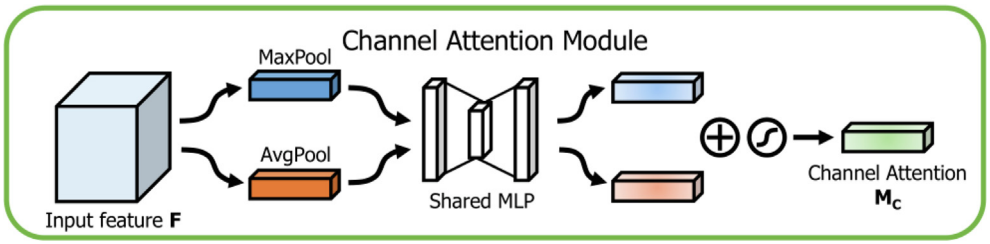

Fig. 3 CAM structure diagram [25]

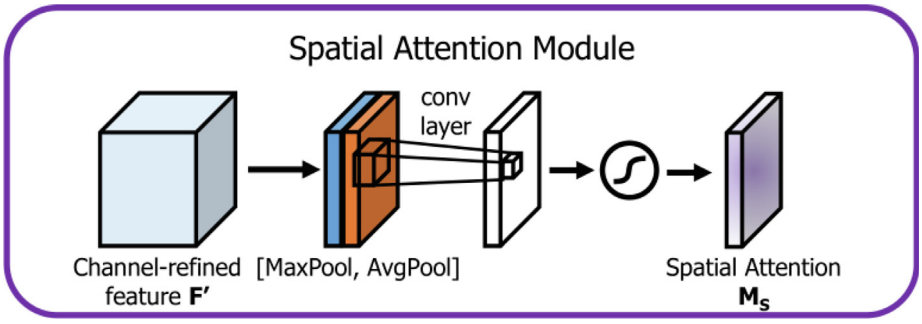

Fig. 4 SAM structure diagram [25]

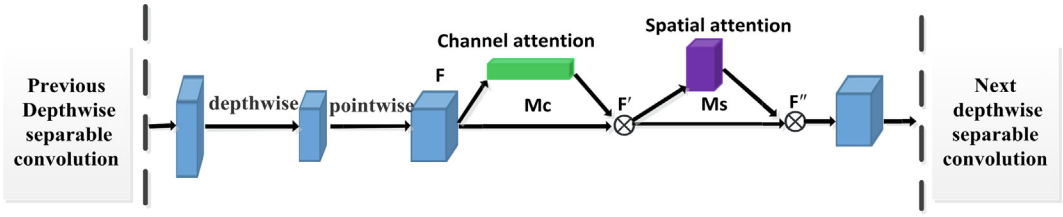

Fig. 5 CBAM integrated with a depthwise separable convolution Block in MobileNet.

For m samples, the training set and its corresponding label are $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})\}$, and K estimated probabilities of each sample are shown in Eq. (4):

$$h_\theta(x^{(i)}) = \begin{bmatrix} p(y^{(i)} = 1|x^{(i)}; W) \\ p(y^{(i)} = 2|x^{(i)}; W) \\ \vdots \\ p(y^{(i)} = k|x^{(i)}; W) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} e^{W_j^T x^{(i)}}} \begin{bmatrix} e^{W_1^T x^{(i)}} \\ e^{W_2^T x^{(i)}} \\ \vdots \\ e^{W_k^T x^{(i)}} \end{bmatrix} \tag{4}$$

where $W_i (i = 1, 2, \ldots, k)$ is the parameter of the network, K is the number of categories, $\frac{1}{\sum_{j=1}^{k} e^{W_j^T x^{(i)}}}$ normalizes the result, and the sum of all the probabilities is 1. In the training phase, the L_s function adopts the gradient descent method to make the result reach convergence. The loss function is shown in Eq. (5):

$$L_s = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\{y^{(i)} = j\} \log \frac{e^{W_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{W_l^T x^{(i)}}}\right] \tag{5}$$

where $\{y^{(i)} = j\}$ is an indicative function. When its value is false, the function value is 0, or else is 1. Eq. (5) can be simplified to Eq. (6):

$$L_s = -\frac{1}{m}\left[\sum_{i=l}^{m} \log \frac{e^{w_{y(i)}^T x(i)}}{\sum_{l=1}^{k} e^{w_l^T x(i)}}\right] \tag{6}$$

In practical application, in order not to make arbitrary $(W_1, W_2, W_3, \ldots, W_k)$ with a parameter was 0, we usually add a weight attenuation to the loss function. The larger the loss function, the smaller the probability that the classifier will be a real label. The minimum value of the loss function was calculated iteratively to obtain the best target result.

Center loss (L_c): Wen et al. [27] first proposed the central loss function, which was a typical clustering algorithm. For a CNN with multiple features, multiple feature centers can be calculated in each batch, and the loss function can be calculated at the same time. The loss function was calculated according to the distance between the eigenvalue and its corresponding center, so the central loss function Eq. (7) is as follows:

$$L_c = \frac{1}{2m}\sum_{i=1}^{m} ||x_i - c_{yi}||_2^2 \tag{7}$$

The updated L_c gradient and c_{yi} are shown in Eq. (8) and (9).

$$\frac{\partial L_C}{\partial x_i} = x_i - c_{yi} \tag{8}$$

$$\Delta c_j = \frac{\sum_{i=1}^{m} \delta(yi = j) \cdot (c_j - x_i)}{1 + \sum_{i=1}^{m} \delta(yi = j)} \tag{9}$$

where x_i represents the eigenvalue of the i-th picture, c_{yi} is the center of the classification to which the i-th picture belongs (the center of the eigenvalue of the classification); and Δc_j represents the change of the classification center. Eq. (8) describes the change intra-class, and the feature center of the classificationc_{yi} changes with the change of depth features. In Eq. (9), when $\delta(y_i = j)$ is 0 or 1, that is, when the condition$y_i = j$is satisfied, $\delta(y_i = j) = 1$,otherwise, $\delta(y_i = j) = 0$. Center loss hopes that the sum of squares of the distance between the feature of each sample in a batch and the center of the feature should be as small as possible, that is, the intra-class distance should be as small as possible.

L_s and L_c joint supervision function: Through the joint supervision of L_s and L_c , increase the feature distance inter-class and reduce the feature distance intra-class. The obtained features have stronger recognition ability, as shown in Eq. (10).

$$L = L_s + \lambda L_c = -\frac{1}{m}\left[\sum_{i=1}^{m} log \frac{e^{W_{y(i)}^T x^{(i)}}}{\sum_{l=1}^{k} e^{W_l^T x^{(i)}}}\right] + \frac{\lambda}{2}\sum_{i=1}^{m} ||x_i - c_{yi}||_2^2 \tag{10}$$

where coefficient λ is used to balance L_s and Lc and the appropriate value of λ is helpful to improve the classification ability of network; C_{yi} is the sample center of class y_i . When $\lambda = 0$, the function can be regarded as the case of only softmax loss.

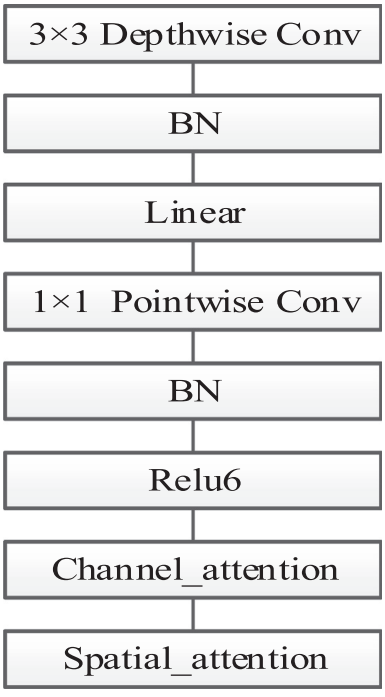

Fig. 6 Improved DSC.

2.2. The improved model A-MobileNet

In the MobileNetV1 network, in order to better show the nonlinear modeling ability of the network, simultaneously, to alleviate the occurrence of over-fitting, the Relu is used after the depthwise convolution and the pointwise convolution. To prevent the gradient explosion, accelerate the convergence speed of the model and improve the efficiency of the model, the BN layer is added before the Relu. In the process of deep convolution, the feature extracted by depthwise convolution is single-channel because it cannot change the number of channels. However, when the Relu operates in the convolution layer output with fewer channels, it may lead to the loss of information, and even affect the modeling ability. To this end, we propose an improved depthwise separable convolution layer; that is, after the depthwise convolution, the linear output was adopted, and the attention module is added after pointwise convolution, as the Fig. 6 shows. The rest is consistent with the DSC layer in MobileNetV1.

Inspired by the MobileNet and attention mechanism, combined with the characteristics of FER, this paper designs an improved MobileNet based on the DSC layer, and use the pre-training parameters on the ImageNet. The network structure is shown in Table 1.

In the model, the input image is first passed through a standard convolution layer, then through 9 DSC layers and 4 improved DSC layers in turn. Finally, features are extracted through the AvgPool layer and full connection layer. In order

Table 1 A-MoblieNet body architecture.

Type/Stride		Filter Shape	Input Size
Conv/s2		3×3×3×32	224×224×3
Conv dw/s1		3×3×32 dw	112×112×32
Conv/s1		1×1×32×64	112×112×32
Conv dw/s2		3×3×64 dw	112×112×64
Conv/s1		1×1×64×128	56×56×64
Conv dw/s1		3×3×128 dw	56×56×128
Conv/s1		1×1×128×128	56×56×128
Conv dw/s2		3×3×128 dw	56×56×128
Conv/s1		1×1×128×256	28×28×128
cbam	channel_attention		28×28×256
	Spatial_attention		28×28×256
Conv dw/s1		3×3×256 dw	28×28×256
Conv/s1		1×1×256×256	28×28×256
Conv dw/s2		3×3×256 dw	28×28×256
Conv/s1		1×1×256×512	14×14×256
cbam	channel_attention		14×14×512
	Spatial_attention		14×14×512
5×	Conv dw/s1	3×3×512 dw	14×14×512
	Conv/s1	1×1×512×512	14×14×512
cbam	channel_attention		14×14×512
	Spatial_attention		14×14×512
Conv dw/s2		3×3×512 dw	14×14×512
Conv/s1		1×1×512×1024	7×7×512
Conv dw/s1		3×3×1024 dw	7×7×1024
Conv/s1		1×1×1024×1024	7×7×1024
cbam	channel_attention		7×7×1024
	Spatial_attention		7×7×1024
Avg Pool/s1		Pool 7×7	7×7×1024
FC/s1		1024×7	1×1×1024
Softmax + center loss		classifier	1×1×7/8

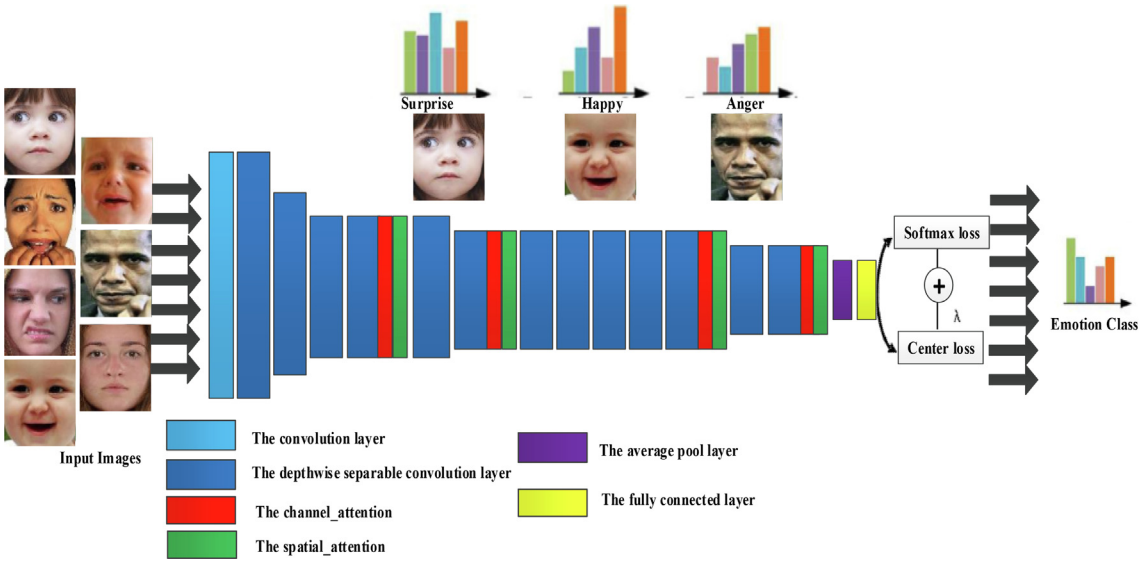

Fig. 7 Basic framework of A- MobileNet network model.

to better identify features and make the network converge quickly, it is necessary to perform BN layer and Relu6 nonlinear activation function processing on the output of each pointwise to the convolutional layer to improve the nonlinear expression ability. In the depthwise convolution layer, to retain information as much as possible, it only passes through the BN layer without adding a nonlinear activation function. It adopts a linear output, that is, all of the original MobileNet model and the improved deep separable convolution layers were adopted. In terms of classifier design, MobileNet uses a softmax classifier. Because the discrimination between different facial expressions is not high, using a softmax classifier may lead to misjudgement, the softmax classifier is not suitable for FER. To solve this problem, the A-MobileNet uses L_s and L_c to jointly supervise and to optimize the parameters. While increasing the distance between categories and reducing the distance within the same category, the features obtained have stronger recognition ability. The architecture of A-MobileNet is shown in Fig. 7.

3. Experiment

In this section, firstly, the environment for experimentation, related data sets and preprocessing of training data are introduced. Secondly, the experimental results are analyzed through confusion matrix and compared with other excellent algorithms.

3.1. Preparation for experiment

***Experimental environment*:** The experiment was based on the Keras framework, with the TensorFlow framework as its backend, the programming language used Python 3.6 in the PyCharm integrated development environment. Experiments on Windows 10 and 64-bit operating systems. The hardware platform was: GPU NVIDIA 2080 Ti, graphics memory 11 GB. In the experiment, optimize the loss using Adam, the learning rate was 0.0001, decay was 1e-5, the epoch was 100, dropout was 0.5, the batch size was 64.

Experimental datasets: The experimental datasets used FERPlus [12,28] and RAF-DB [13], both collected from the Internet and contain unconstrained facial expression images. Sample images of each database are shown in Fig. 8. Table 2 illustrates the amount of each expression taken from FERPlus and RAF-DB.

The predecessor of the FERPlus is the FER2013[29]. Due to non-face data and false tags in the FER2013 , the recognition rate of human expression is only 65±5%. For the FER2013 to play a more important role, researchers have relabelled the FER2013, using 10 types (happy-HA, neutral-NE, sad-SA, surprise-SU, fear-FE, contempt-CO, dispute-DI, anger-AN, unknown and non-face). The maximum-voting method is used to remove some uncertain images. Therefore, the FERPlus consists of 8 emotion categories and 31,412 facial grayscale images. These were split into the training set, private test set, and public test set with 25,060, 3,152, and 3,199 samples, respectively.

Real-world Affective Faces Database (RAF-DB) consists of 7 emotion categories (except for contempt) and 15,539 facial color images. These were split into training and test sets with 12,271 and 3,068 samples, respectively.

Facial expression image pre-processing: Both the original FERPlus and RAF-DB provided cropped face images. As shown in Fig. 8, the image sizes in the FERPlus and RAF-DB are 48 × 48 × 3 and 100 × 100 × 3 (width × height × channel), respectively. We know that the training dataset is unbalanced from Table 2. Models trained with data will result in poor prediction performance for categories with few samples, and even the category of samples cannot be predicted. So do data augmentation. Perform operations such as rotation (0° - 20°), crop (0–0.2), rescale, and translation on the category pictures with few samples to expand the samples so that the number of samples in each category is as balanced as possible.

Fig. 8 Examples Images in FERPlus (top), RAF-DB (bottom).

Table 2 Number of selected images per expression from FERPlus and RAF-DB.

Dataset	Expression							
	AN	DI	FE	HA	SA	SU	NE	CO
FERPlus	2100	119	532	7287	3014	3149	8740	119
RAF-DB	705	717	281	4772	1982	1290	2524	–

3.2. Experiments on two popular datasets

This paper conducts training according to the network model designed in Table 1, input the pre-processed pictures into the network and use the test set classification effect to evaluate the network performance. We tested the MoblieNetV1, V2, V3-small, and our model on RAF-DB and FERPlus.

The confusion matrices of these models on two datasets are illustrated in Fig. 9. Therein,(a)–(d) are experimental results on the FERPlus, implemented by the MoblieNetV1, V2, V3-small, and our model A-MobileNet respectively. (e)–(h) are experimental results on the RAF-DB with these models. As demonstrated in these figures, our model has higher performance on most labels than the MobileNet family model. In the two datasets, the recognition accuracy of happy expression is the highest. The possible reason is that happy expressions have richer features and are easy to distinguish from other expression features. The recognition rate of surprised, neutral, angry, and sad expressions decreased sequentially. Common

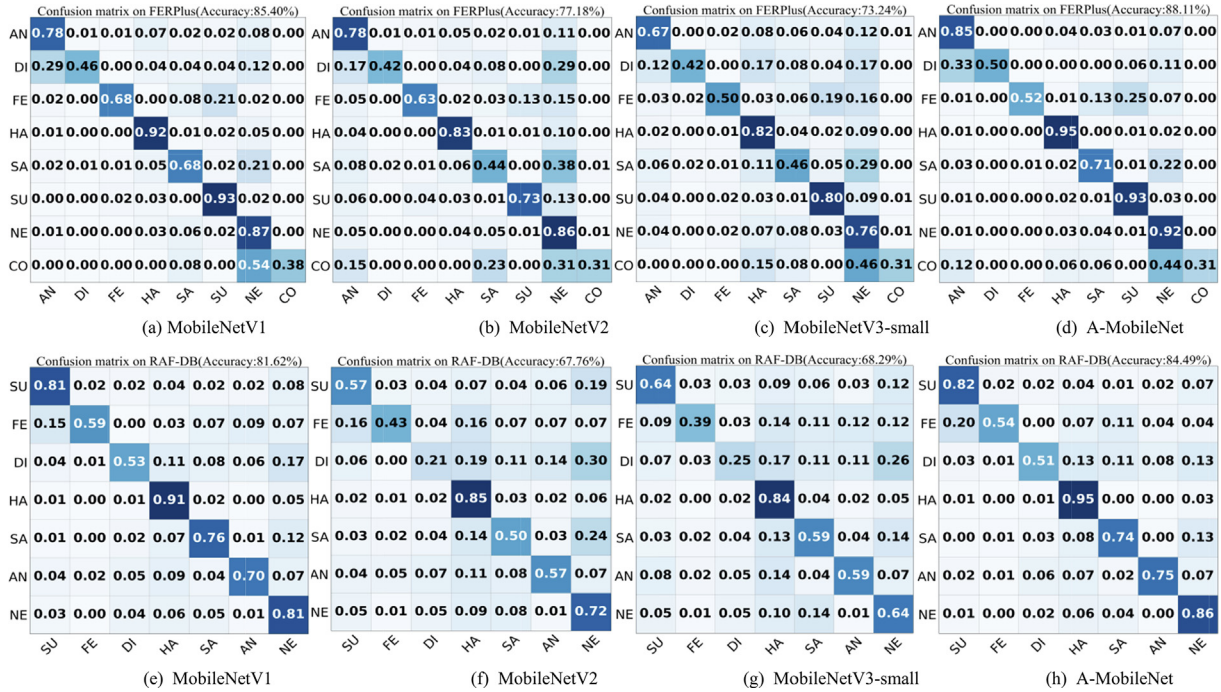

Fig. 9 Confusion matrices on two expression databases. (a)–(d) are confusion matrices for MobileNet V1, V2, V3-small and our model A-MobileNet on FERPlus. (e)–(h) are confusion matrices for the four networks on RAF-DB.

facial expressions in life are relatively easy to recognize, and the expressions that are easily misidentified are fear, disgust and contempt. For example, we can see that on the FERPlus, the expression of fear is easily misidentified as surprise and the expression of disgust is easily misidentified as angry. On the RAF-DB, disgust is easily misidentified as neutral or happy. Regarding the emergence of this problem, we believe that one is because the number of different expressions in the dataset is unbalanced, the other is that anger, disgust, fear and sadness have certain similarities in themselves. In real life, it is difficult to correctly recognize the expressions of strangers who suddenly appear.

Table 3 Performance evaluation on FERPlus.

Model	Accuracy (%)	Parameters	note
VGG13(MV) [12]	83.86	–	
TFE-JL [32]	84.29	–	
CNNs and BOVM + global SVM [33]	87.76	–	
ResNet + VGG [34]	87.4	–	
SCN + ResNet18[30]	88.01		
MobileNetV1	85.40	3.2 M	
MobileNetV2	77.18	2.3 M	alpha = 1.0
MobileNetV3-small	73.24	1.5 M	
A-MobileNet	**88.11**	3.4 M	

Table 4 Performance evaluation on RAF-DB.

Model	Accuracy (%)	Parameters	note
PG-CNN [35]	82.27	–	
Capsule-based Net[36]	77.48	–	
DLP-CNN [31]	84.13	–	
Mean + ASL + L2SL [37]	84.69	–	
DBA-Net (DenseNet-161) [38]	79.37	42.9 M	
MobileNetV1	81.62	3.2 M	
MobileNetV2	67.77	2.3 M	alpha = 1.0
MobileNetV3-small	68.29	1.5 M	
A-MobileNet	**84.49**	3.4 M	

On the FERPlus and RAF-DB datasets, the experimental results of the A-MobileNet network model and other deep models and MobileNet V1, V2, V3-small network models are shown in Table 3 and Table 4. In the MobileNet series models, the MobileNetV1 model has the highest recognition rate. Its performance is 2.71% and 2.87% higher than the MobileNet V1 baseline, and the parameter amount is only increased by 0.2 M.

In addition to comparing our implementation of the baseline, compared with other excellent deep models. On the dataset of FERPlus, SCN + ResNet18[30] adopted a self-cure network (SCN) to effectively suppress the uncertainty of the expression annotation of the training data, so as to learn more different expression features. In order to solve the huge differences within the same expression class caused by occlusion, illumination, head posture change, individual differences, and so on. Li and Deng [31] proposed deep locality-preserving CNN. In this network, a new supervisory layer (i.e., locality preserving loss) is added to the basic CNN to make the local neighborhood features in each class as cohesive as possible. However, these networks are not lightweight models and are limited in real-time applications. To improve the ability to recognize facial features and reduce parameters. The A-MobileNet model hardly increases any reasoning cost. The performance of the A-MobileNet model on the RAF-DB and FERPlus datasets is 84.49% and 88.11%, respectively.

This paper designs a PC version of the player that simply recognizes user expressions and automatically searches for

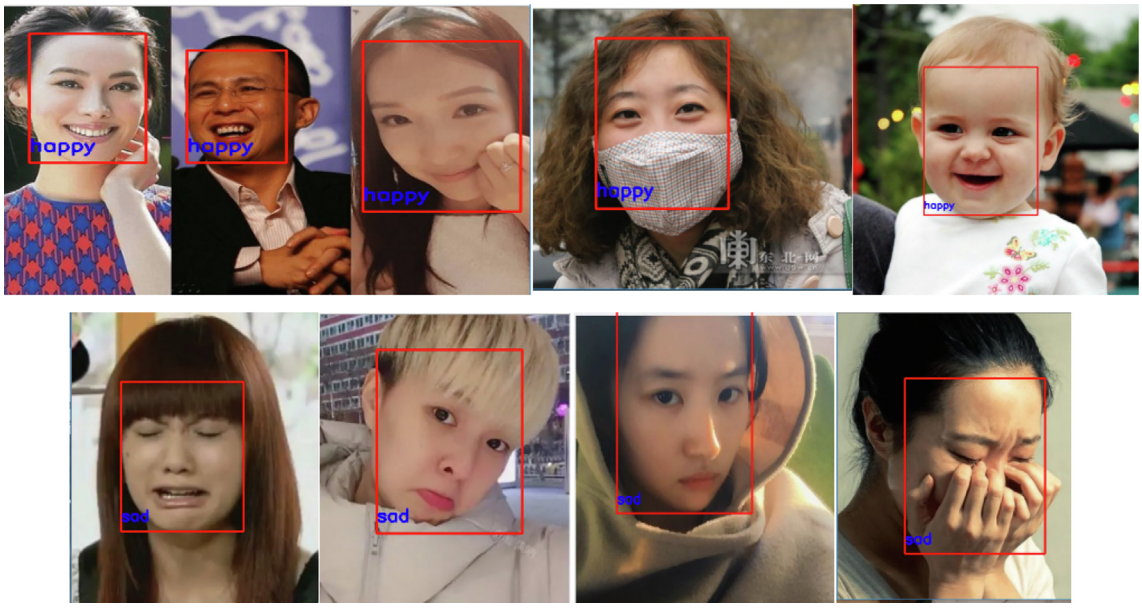

Fig. 10 Facial expression recognition effect of the A-MobileNet network.

expression-related music. First, the camera is opened, and face detection is performed by the MTCNN algorithm to obtain the facial image, and the facial expression is recognized by the model in this paper; the recognition effect is shown in Fig. 10. Then, according to the identification results in the local disk search for the corresponding mood classification song list to play. The emotional tag of the song has been processed in advance. This simple experiment proves that emoticons can be used as an interactive method.

CRediT authorship contribution statement

Yahui Nan: Conceptualization, Methodology, Software. **Jianguo Ju:** Software, Validation, Revision draft. **Qingyi Hua:** Put forward suggestions for revision, Supervision. **Haoming Zhang:** Investigation, Visualization, **Bo Wang:** Writing - reviewing and editing.

4. Conclusion

In this paper, the A-MobileNet model is proposed for FER. Specifically, in the A-MobileNet network model, an attention module is introduced to enhance the model's ability to extract fine-grained features of facial expressions, and dropout technology is added to prevent overfitting. Experimental results on the FERPlus and RAFDB show that our improved model achieves better results than the lightweight MoibleNet series models and other excellent methods. The recognition accuracy is 84.49% and 88.11% on the RAF-DB and FERPlus, respectively. In the following work, we will analyze the misrecognized expression data and find out the reasons, pay attention to the detailed information of specific expressions, and further improve the classification ability of the model.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M.S. Omar, S.A. Hassan, H. Pervaiz, Q. Ni, L. Musavian, S. Mumtaz, O.A. Dobre, Multiobjective optimization in 5G hybrid networks, IEEE Internet Things J. 5 (3) (2018) 1588–1597.

[2] Z. Zhou, X. Chen, Y. Zhang, S. Mumtaz, Blockchain-empowered secure spectrum sharing for 5G heterogeneous networks, IEEE Netw. 34 (1) (2020) 24–31.

[3] M. Zhang, F. Conti, H. Le Sourne, D. Vassalos, P. Kujala, D. Lindroth, S. Hirdaris, A method for the direct assessment of ship collision damage and flooding risk in real conditions, Ocean Eng. 237 (2021) 109605.

[4] Z. Xie, J. Wang, L. Miao, Big data and emerging market firms' innovation in an open economy: The diversification strategy perspective, Technol. Forecast. Soc. Chang. 173 (2021) 1–14.

[5] R. Jiang, Z. Fei Ma J. Yang, An assessment model for cloud service security risk based on entropy and support vector machine, Concurrency and Computation: Practice and Experience. (2021).

[6] W. Shu, K. Cai, N. Xiong, Research on strong agile response task scheduling optimization enhancement with optimal resource usage in green cloud computing, Future Generation Comput. Syst. 124 (2021) 12–20.

[7] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on local binary patterns: a comprehensive study, Image Vis. Comput. 27 (6) (2009) 803–816.

[8] G. Zhao, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 915–928.

[9] R. Zhi, M. Flierl, Q. Ruan, W.B. Kleijn, Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition, IEEE Trans. Systems, Man, Cybern. Part B (Cybernetics) 41 (1) (2010) 38–52.

[10] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, D.N. Metaxas, Learning active facial patches for expression analysis, in: In 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2562–2569.

[11] J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park. R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, Y. Bengio, Challenges in representation learning: A report on three machine learning contests, In International Conference on Neural Information Processing, Springer, Berlin, Heidelberg. (2013) 117–124.

[12] E. Barsoum, C. Zhang, C.C. Ferrer, Z. Zhang, Training deep networks for facial expression recognition with crowd-sourced label distribution, in: In Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016, pp. 279–283.

[13] S. Li, W. Deng, J. Du, Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild, in: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2852–2861.

[14] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2012) 1097–1105.

[15] K. Simonyan, A. Zisserman, A very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv: 1409.1556 (2014).

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.

[17] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv: 1704.04861. (2017).

[18] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.C. Chen, Mobilenetv2: inverted residuals and linear bottlenecks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.

[19] A. Howard, M. Sandler, G. Chu, L. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Hartwig Adam, Searching for mobilenetv3, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1314–1324.

[20] Y. Wang, J. Wu, K. Hoashi, Lightweight deep convolutional neural networks for facial expression recognition, in: 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP), 2019, pp. 1–6.

[21] S. Cotter, Low complexity deep learning for mobile face expression recognition, in: Proceedings of the 3rd International Conference on Vision, Image and Signal Processing, 2019, pp. 1–5.

[22] F. Zhang, Q. Li, Y. Ren, H. Xu, S. Liu, An expression recognition method on robots based on mobilenet v2-ssd, in: 2019 6th International Conference on Systems and Informatics (ICSAI), 2019, pp. 118–122.
[23] L. Hu, Q. Ge, Automatic facial expression recognition based on MobileNetV2 in Real-time, J. Phys. Conf. Ser. 1549 (2) (2020).
[24] J. Olivares-Mercado, K. Toscano-Medina, G. Sanchez-Perez, J. Portillo-Portillo, G. Benitez-Garcia, Analysis of hand-crafted and learned feature extraction methods for real-time facial expression recognition, in: 2019 7th International Workshop on Biometrics and Forensics (IWBF), 2019, pp. 1–6.
[25] S. Woo, J. Park, J.Y. Lee, I.S. Kweon, Cbam: convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.
[26] G. King, L. Zeng, Logistic regression in rare events data, Political Analysis. 9 (2) (2001) 137–163.
[27] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, Eur. Conf. Comput. Vis. (2016) 499–515.
[28] FERPlus emotion label, https://github.com/Microsoft/FERPlus. 2021(accessed July14, 2021).
[29] I.J. Goodfellow, D. Erhan, P.L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, Challenges in representation learning: a report on three machine learning contests, International Conference on Neural Information Processing. (2013) 117–124.
[30] K. Wang, X. Peng, J. Yang, S. Lu, Y. Qiao, Suppressing uncertainties for large-scale facial expression recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6897–6906.
[31] S. Li, W. Deng, Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition, IEEE Trans. Image Process. 28 (1) (2018) 356–370.
[32] M. Li, H. Xu, X. Huang, Z. Song, X. Liu, X. Li, Facial expression recognition with identity and emotion joint learning, IEEE Trans. Affective Comput. (2018).
[33] M.I. Georgescu, R.T. Ionescu, M. Popescu, Local learning with deep and handcrafted features for facial expression recognition, IEEE Access. (2019) 764827–764836.
[34] C. Huang, Combining convolutional neural networks for emotion recognition, 2017 IEEE MIT Undergraduate Research Technology Conference (URTC) (2017) 1–4.
[35] Y. Li, J. Zeng, S. Shan, X. Chen, Patch-gated CNN for occlusion-aware facial expression recognition, International Conference on Pattern Recognition (2018) 2209–2214.
[36] S. Ghosh, A. Dhall, N. Sebe, Automatic group affect analysis in amages via visual attribute and feature networks, International Conference on Image Processing (2018) 1967–1971.
[37] P. Jiang, G. Liu, Q. Wang, J. Wu, Accurate and reliable facial expression recognition using advanced softmax loss with fixed weights, IEEE Signal Process Lett. 27 (2020) 725–729.
[38] C.H. Hua, T. Huynh-The, H. Seo, S. Lee, Convolutional network with densely backward attention for facial expression recognition, in: 2020 14th International Conference on Ubiquitous Information Management and Communication (IMCOM), IEEE, 2020, pp. 1–6.