Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/53/20241556

Emotion recognition based on ResNet and transfer learning

Dongwei Liu

Qingdao No.2 Middle School of Shandong Province, Qingdao, 266100, China

196062116@mail.sit.edu.cn

Abstract. In the fields of psychology and artificial intelligence, emotion recognition is a crucial area of research. As an expression of body language, emotion recognition has a profound impact on our life and interpersonal relationships while existing models have the phenomenon of exploding or disappearing gradient indices due to the lack of information. To alleviate the aforementioned limitations, this paper further introduces the idea of residual jumping and transfer learning to conduct further research and exploration for the emotion recognition. Specifically, based on the RAF-DB dataset, the self-created model Awei and the pre-trained VGG model are combined. Numerous tests show that the suggested approach is effective. By reducing the amount of information lost during the convolution process, the residual idea can help with the issue of an exploding or vanishing gradient index brought on by an excessively long backpropagation. This will increase the correctness of the model. The pre-trained VGG model's addition improves the Awei model's parameter modification's efficacy and accuracy.

Keywords: Emotion Recognition, Residual Jumper Structures, Transfer Learning

1. Introduction

Emotion recognition is an important branch in the field of computer vision [1], aiming to identify various parts of a human face to determine emotions. Thanks to the rapid development of artificial intelligence technology represented by convolutional neural networks , facial Emotion recognition methods based on convolutional neural networks have been widely used in various fields such as social intelligence and human-computer interaction, emotion analysis, mental health,emotion-driven education, market research and advertising, safety and emotion detection, autonomous driving, and traffic safety, etc [2]. Especially in social intelligence and human-computer interaction, emotion recognition is used to understand the user's emotions, allowing computers to interact with people. Humans are more natural and provide a more personalized user experience, which is crucial for virtual assistants, emotional robots, and intelligent customer service applications [3].

Face detection, feature extraction, and expression categorization are the three primary phases in facial expression recognition [4]. Most existing research focuses on extracting high-quality expression features, which can significantly improve recognition performance. According to the different design ideas of feature extraction methods, the earlier techniques can be categorized as follows: techniques based on appearance features, techniques based on geometric features, and techniques based on combined features. techniques using geometric characteristics such as a complete label distribution learning (LDL) framework utilizing geometric features and a deep convolutional neural network

DOI: 10.54254/2755-2721/53/20241556

(CNN) [5], switching up the facial points comprise of feature vectors that indicate face geometry values, which are created using picture coordinates on a particular face area [6]. It must extract contour reference points and facial organs, both of which are difficult to accurately analyze in complicated or poor quality backdrop images. This is due to the fact that it reduces feature discriminability by concentrating solely on the geometric aspects of the face and ignoring information about other facial features like texture of the skin or variation in color. The appearance feature-based method uses all face image pixels to capture the underlying information of the face image. For instance, first of all, a facial landmark detector is used to locate, extract, and align facial features. Then, the eigenpictures technique is applied to characterize images. Next, we use the weights retrieved using the eigenpictures technique as input for a clustering of each type of face feature. Lastly, we use human validation to confirm the observed clusterings. [7]. It mainly uses a set of filters to filter the image to extract the relationship between local pixels (such as gradient, correlation, texture, etc.). Geometric features can efficiently reflect changes in the macrostructure of the face, while appearance features focus on extracting local subtle changes. Therefore, methods based on mixed features try to combine the two parts of features for expression recognition. Although previous research has made great progress [8], related fields still lack the practical application of using pre-trained models as guidance to modify their own model parameters. At the same time, some models may have the problem of gradient exponential explosion or backpropagation too far, resulting in information loss.

Focusing on the aforementioned issues, this paper proposes a new emotion recognition method named Awei, which combines residual jumping and transfer learning. Specifically, the proposed Awei introduces a strong pre-trained VGG model to predict the probability of emotion, enhancing the validity of model parameter modification. Awei also uses the residual idea to digitally superimpose the feature map of neighboring convolutional layers, so that can solve the problem of backpropagation gradient exponential explosion. Extensive experiments were carried out on the RAF-DB dataset, which shows the effectiveness of the Awei method.

2. Materials and Methods

2.1. Original data

The data for this research comes from the RAF-DB dataset containing about 30,000 facial images. This dataset contains faces that differ significantly in terms of gender, age, ethnicity, head pose, lighting conditions, occluders (such as hair on the face, self-occlusion, or eyeglasses), and additional processing manipulations (such as a variety of special effects and filters). As a result, the dataset has a greater degree of diversity, and the trained models are also more generalizable. The images in this dataset are reset to 224×224 pixels, which is convenient for the unified convolution operation later, and each image is tagged with the corresponding emotion label (anger, disgust, fear, happiness, neutral, sadness, surprise [9]), three channels are opened to import the RGB color values, and the corresponding pixel value of each image is scaled to the range of [0,1], so as to standardize the images, decrease noise and pointless image modifications, and enhance the stability of the model training. noise and unnecessary changes in the image, while increasing the speed of model training to a certain extent, reducing the risk of overfitting, and helping the model to better generalize the unprocessed data as a way to increase the accuracy of the validation set.

2.2. Method

2.2.1. Residual module. In the model utilization section, a new model (Awei) is created. Two fully connected layers and five convolutional layers make up the Awei model. Within the convolutional layer section, a 3×3 size convolutional kernel with 32, 64, 128, 256, and 512 kernels, respectively, is used for all convolutional operations. The padding of each layer is set to the same, while the following pooling layer with 2×2 kernel reduces the feature map size to a quarter. After flattening the pixel points through the FLAT layer, 1792 neurons are used for the fully connected layer operation, and half of the

DOI: 10.54254/2755-2721/53/20241556

neurons are thrown away through the regularization technique in the DROPOUT layer to reduce the risk of overfitting, so that the machine learns more robust and generalized features during training and the interdependency among neurons can further reduce. However, the process may result in the critical information loss. Therefore, the idea of residual jumper structure is applied, which is one of the main differences between the Awei model as opposed to creating a sequential model directly.

A convolution layer dedicated to the residual jump structure is created during the initialization of the Awei model for the purpose of mathematical computation with the structure generated by the convolution of the fifth layer. As shown in Figure 1, the filters of this convolution layer are the same as the 512 values corresponding to the fifth layer's convolution, while the convolution kernel's size is not changed (3×3), and the pooling layer is changed to a block of 32×32 with a max pooling operation, so that the image through the residual jump structure can be changed into a four-dimensional structure with the same result as that of the convolution of the fifth layer (1, 7, 7, 512). The data processed by the residual jumper structure and the data after five convolutions are mathematically computed in the flattening of the flattening layer in front of the fully connected layer, which are passed into the first fully connected layer (1792 neurons) for computation, reducing the number of fully connected neurons. neurons) for computation to solve the problem of exploding or vanishing gradient indices due to loss of information.

2.2.2. Transfer learning module. Considering the accuracy of the validation set is still relatively low and overfitting still exists, on this basis, transfer learning is introduced to assist the training of the Awei model by applying the pre-trained VGG model. Concretely, the fully connected layer's parameters of the pre-trained VGG model are frozen, which does not serve as the output of the result. For the training of the Awei model, it is initialized using the pre-trained VGG model as a guide. At the same time, the position of the intermediate layer of data to be acquired after the image passes through the pre-trained model is defined in the process of forward propagation of the Awei model. The input image will be fed into the pre-trained model to acquire the data after the pre-trained model is flattened and before it enters into the calculation of the full connectivity layer, which will be mathematically sumed with the output computed in the residual jumper structure. Then, the output will be fed into the dropout layer for regularization calculation. If the data generated by the pre-trained VGG model is thrown in after the dropout layer, it will cause data inequality, because after the dropout calculation. Some neurons in the Awei model have been discarded and the structure of the data has been changed, which cause data errors when adding the same size of the data structure generated by the pre-trained model. All these aspects will lead to inefficiency in model training and a decrease in the validation set's precision. Finally, the normalized intermediate data will be fed into the full connected layer for the classification of the seven emotions.

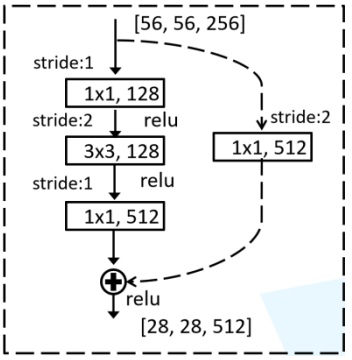


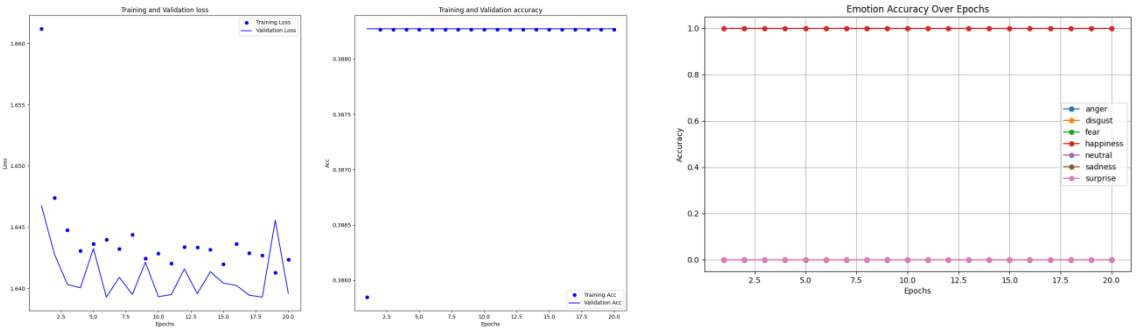
Figure 1. The structure of residual module

DOI: 10.54254/2755-2721/53/20241556

3. Experiment

3.1. Performance of original CNN

Adding the residual bar structure to our model greatly enhances its learning capacity, which is based on the most fundamental five-layer convolutional layer. When passing the same RGB three-channel images to the incoming model, controlling the single variable batch_size size both 8 and epoch both 20, the accuracy and loss function of the training and loss sets of the five original convolutional layers (without residual bar structure), and the accuracy corresponding to each emotion can be observed in Figure 2.

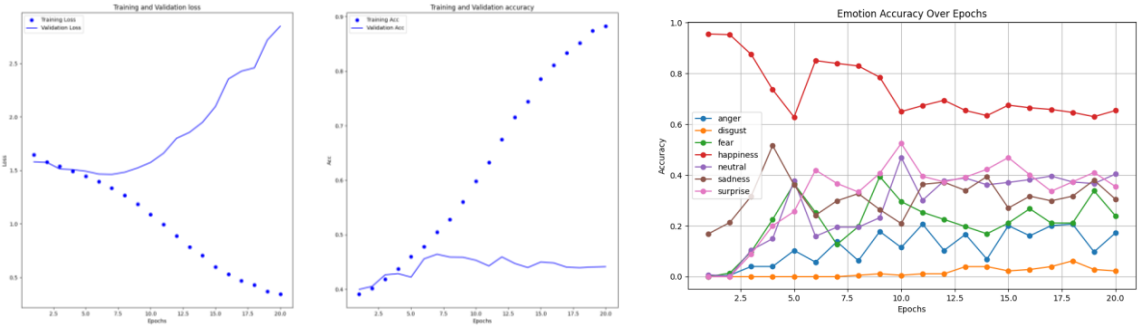


(a) Training accuracy and loss (b) Accuracy of each emotion

Figure 2. Performance of original CNN

3.2. Performance of proposed Awei model

Accuracy and loss functions for the training and loss sets at the time of introducing the residual jumper structure, and the accuracy corresponding to each emotion is shown in Figure 3. For the first two types of data results are analyzed, it can be clearly seen, because the batch_size is too small not to use the residual data structure of the learning ability is almost 0 (or even do not learn), because the data available for reference is too small, after five layers of convolution, the machine can be applied to the effective data can only be 7×7 size, so so that the accuracy of the training set and the validation set has always been a stagnant rate, and the Why the machine only knows to learn happy, part of the reason may be because the number of happy in the training set is too much (more than 4,000), compared to not learn much about disgust only 400 images, thus making the weight of the training data of different emotions imbalance [10], and when the residual jumper structure is introduced, even in a very small batch_size can be continuously learning. Improving the accuracy of emotion recognition is a sufficient proof of the effectiveness of the residual jumper structure in emotion recognition research, i.e., in reducing the gradient exponentially exploding or disappearing due to the loss of important information and thus even the condition of not learning in Figure 2(a).

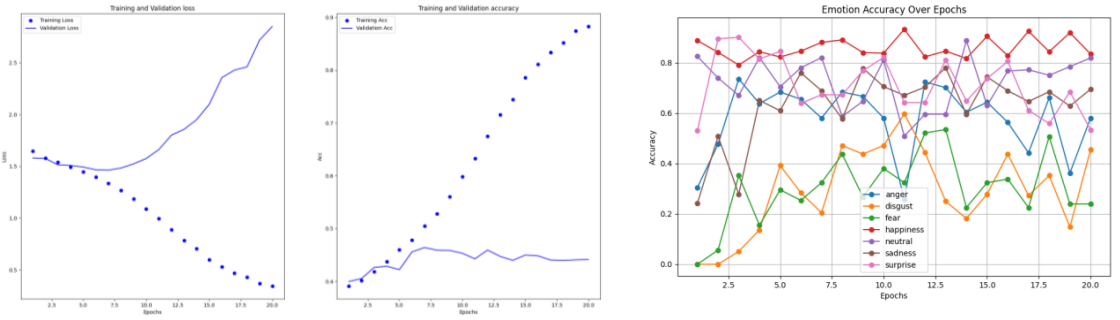


(a) Training accuracy and loss (b) Accuracy of each emotion

Figure 3. Performance after introducing Residual structure

DOI: 10.54254/2755-2721/53/20241556

The accuracy and loss functions for the training and loss sets of the model with the simultaneous introduction of the residual jumper structure and migration learning, as well as the accuracy corresponding to each emotion are shown in the following Figure 4.



(a) Training accuracy and loss (b) Accuracy of each emotion

Figure 4. Performance of proposed Awei

For the analysis of the model with both residual jumper structure and migration learning, after introducing a pre-trained VGG model, the model's learning efficiency has risen drastically, and the accuracy of the validation set is stabilized at 0.7 or above, which is a significant improvement compared to the initial value of less than 0.5, and the value of the loss function has been controlled to be less than 1.6 (compared to the final value of 2.5 or above in the case of the residual jump structure) and solves the problem that the machine only chooses happy in the initial learning process (other types of emotions also have higher accuracy at the beginning), which naturally improves the correctness of the learning) and solved the problem that the machine would only choose happy in the process of learning at the beginning (other types of emotions also had higher accuracy at the beginning), which naturally improved the correct rate of learning, and further optimized the purposefulness of the model's emotion category selection, rather than wasting a lot of training costs and training time on one emotion at the beginning. Moreover, the introduction of a new training model by migration learning effectively standardizes the effectiveness and purposefulness of parameter modification of the Awei model and accelerates the learning progress, so the method of learning by migration is effective for recognizing categorization models such as emotions.

Overall, it further also demonstrates the rapport of the combination of the structure of the residual jumper and transfer learning, where the performance of the model is further enhanced by the introduction of the pre-trained VGG model in a context based on the structure of the residual jumper, and there is no doubt that these two improvements complement each other in the emotion recognition project.

4. Discussion

For this model there is still part of the space that can be discussed, because the residual structure is directly jumped from the first layer to the fifth layer, which is directly reduced by the pooling layer of 32×32, and also lose some important information, so if the residual jump structure is changed to be jumped from the first layer to the third layer in the jump to the fifth layer maybe less important information will be lost, but along with this, we have to consider that exponentially increase the training cost and the training time.

On other hand, we can introduce an image enhancement module [11], which rotates and partially intercepts the training images and passes them into the model to increase the samples for model training, so as to have a stronger ability to analyze emotions. Alternatively, a weakly supervised module could be introduced to simply annotate the facial expressions we want to select, so that the machine can recognize the position to be judged more quickly, if only for the RAF-DB dataset, where most of the images have white space, meaning that the introduction of a weakly supervised module

DOI: 10.54254/2755-2721/53/20241556

will produce better training results. Finally there may be some engineering improvements, such as using a subset of the training set as a test and validation set.

5. Conclusion

In this study, it is found that the introduction of residual jump structure and migration learning will jointly improve the efficiency of machine recognition of emotions, and the two are complementary and mutually reinforcing, and at the same time eliminate the possibility of selecting a certain emotion as the output at the beginning of the original Awei model, and in the process of the initial training, the accuracy rate of the validation set is even higher than that of the training set, which is supposed to be a credit to the pre-training of the model of the VGG. It further proves that the starting point of Awei's model learning is higher than other models, but there is still a situation where the accuracy of judging a certain emotion is lower, which is mainly caused by the smaller samples of that emotion, and the machine has less chance to see it during the training, so the model predicts the result of the emotion with a smaller probability. This study has a partially heuristic guidance for the emotion recognition field, that is, the residual jump structure and migration learning for emotion recognition model is a big improvement, and the two improvement ideas can co-exist to further improve the training efficiency of the model. However, there are still some improvements that can be made, such as further improving the complexity of the residual jump structure, and performing two convolutional pooling operations instead of only one convolutional 32×32 pooling layer for pixel reduction processing as mentioned in this paper, which can further reduce the loss of information. Further, the introduction of image enhancement techniques to vary the diversity of images in the training set increases the ability of the model to face diverse images. Alternatively, a weakly supervised module is introduced to appropriately label the sample content for selective discarding to increase the effectiveness of the learned content. The above methods are used to further strengthen the discriminative ability of the emotion recognition model and lay the foundation for further scientific research development.

References

[1] E. Churaev and A. V. Savchenko, "Multi-user facial emotion recognition in video based on user-dependent neural network adaptation," 2022 VIII International Conference on Information Technology and Nanotechnology (ITNT), Samara, Russian Federation, 2022, pp. 1-5.

[2] M. Dinesh, S. K. G. and A. P. R., "Facial Emotion Recognition based on Attentive Neural Network for the Blind," 2023 International Conference on Control, Communication and Computing (ICCC), Thiruvananthapuram, India, 2023, pp. 1-5.

[3] V. V. Salunke and C. G. Patil, "A New Approach for Automatic Face Emotion Recognition and Classification Based on Deep Networks," 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, India, 2017, pp. 1-5.

[4] N. Mishra and A. Bhatt, "Feature Extraction Techniques in Facial Expression Recognition," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2021, pp. 1247-1251.

[5] S. Liu, B. Li, Y. -Y. Fan, Z. Quo and A. Samal, "Facial attractiveness computation by label distribution learning with deep CNN and geometric features," 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 2017, pp. 1344-1349.

[6] D. Y. Liliana, M. R. Widyanto and T. Basaruddin, "Geometric Facial Components Feature Extraction for Facial Expression Recognition," 2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Yogyakarta, Indonesia, 2018, pp. 391-396.

[7] F. Fuentes-Hurtado, J. A. Diego-Mas, V. Naranjo and M. Alcañiz, "A Holistic Automatic Method for Grouping Facial Features Based on Their Appearance," 2018 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 2018, pp. 1322-1326.

DOI: 10.54254/2755-2721/53/20241556

[8] Y. Zhenyu and Z. Jiao, "Research on Image Caption Method Based on Mixed Image Features," 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chengdu, China, 2019, pp. 1572-1576.
[9] S. Giri et al., "Emotion Detection with Facial Feature Recognition Using CNN & OpenCV," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022, pp. 230-232.
[10] Shan Li and Weihong Deng, "Real world expression recognition: A highly imbalanced detection problem," 2016 International Conference on Biometrics (ICB), Halmstad, 2016, pp. 1-6.
[11] W. Zeng, J. Cao, J. Wang, X. Lai and Z. Lin, "MAM: Mixed Attention Module with Random Disruption Augmentation for Image Classification," 2020 IEEE International Symposium on Circuits and Systems (ISCAS), Seville, Spain, 2020, pp. 1-5.