World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:17, No:2, 2023

Deep-Learning Based Approach to Facial Emotion Recognition Through Convolutional Neural Network

Nouha Khediri, Mohammed Ben Ammar, Monji Kherallah

Abstract—Recently, facial emotion recognition (FER) has become increasingly essential to understand the state of the human mind. However, accurately classifying emotion from the face is a challenging task. In this paper, we present a facial emotion recognition approach named CV-FER benefiting from deep learning, especially CNN and VGG16. First, the data are pre-processed with data cleaning and data rotation. Then, we augment the data and proceed to our FER model, which contains five convolutions layers and five pooling layers. Finally, a softmax classifier is used in the output layer to recognize emotions. Based on the above contents, this paper reviews the works of facial emotion recognition based on deep learning. Experiments show that our model outperforms the other methods using the same FER2013 database and yields a recognition rate of 92%. We also put forward some suggestions for future work.

Keywords—CNN, deep-learning, facial emotion recognition, machine learning.

I. Introduction

EMOTION Recognition (ER) is an essential role in human communication and has become a hot spot in human-computer interaction. ER is a valuable branch of affective computing and can be expressed using different modalities such as speech, text, or facial expression [1]. Despite the available range of communication modalities, the mainstream research on emotion recognition has mostly focused on facial expression. FER is one of the key components of human-computer interaction. In recent years, deep learning has been used with great success in determining emotional states. Deep learning is a machine learning technique that has the capacity to solve complex problems with access to big data. It is an Artificial Intelligence function that teaches computers how to emulate a human brain's operation. Deep learning is likewise called deep neural networks. In the past few years, it has achieved enormous success and has been applied in many fields like self-driven cars, speech recognition, vision recognition, automatic handwriting generation, natural language processing, and many others. Deep means many intermediate layers between input and output wherein each node in hidden layers represents a different configuration of inputs through feature identification and processing.

In the rest of this paper, we will focus on deep supervised learning algorithms. Among various network architectures, Convolutional Neural Network (CNN) is the most mainstream technique in supervised deep learning methods. It is responsible for processing data that come in the structure of arrays. The architecture of CNN contains four layers:

- Convolution Layer: This layer aims to find the features and apply filters. The most activation function used in this level is Rectified Linear Unit (ReLU) [4].
- Pooling Layer: also called subsampling or downsampling, it reduces the dimensionality of the feature map and keeps only the important ones. It has various types like Max, Average and Sum.
- Flattening Layer: It takes the layers obtained after the max pooling and transforms them into one dimension array (feature vector).
- Full Connection Layer: This layer aims to build all the needed connections.

For more details about emotion detection in human faces, readers can refer to [2] and [3], which describe a comparative study of various algorithms and techniques of literary works. In this work, we focus on the problem of emotion recognition from image channels. In this scenario, we start by pre-processing the image. Then, we have implemented a FER model. The proposed model has achieved an emotion recognition accuracy of 92% on the FER2013 dataset. Finally, a softmax classifier is used to classify the emotions into seven labels (happy, fear, angry, sad, surprise, Disgust and neutral).

The following paper is as follows. The next section briefly summarizes the related works. Section III describes our facial emotion recognition approach, and Section IV presents the experimental results and analysis. Finally, Section V concludes this paper and describes potential future guidance.

II. Related Work

Since the facial expression is the earliest and most common technique to perceive emotions, many approaches have been introduced to classify emotions based on facial expressions. The facial expression can be expressed in two different channels. One is an image channel, and the other is a video channel. Our approach is based on image channel. Deep Learning models have made incredible progress in automatically recognizing facial emotion in images thanks to the multilayered architecture of neural networks.

Singh et al. [5] used a deep neural network on the FER2013 dataset, and they obtained 67.7% test accuracy. A new bilinear pooling model with CNNs for facial emotion recognition is proposed in [6]. The idea of this work was to capture additional discriminant information compared to ordinary CNN, and the

N. Khediri is with the Department of Computer Sciences, Faculty of Sciences of Tunis, El Manar University, Tunisia (e-mail: Nouha.khediri@fst.utm.tn).

M. Ben Ammar is with the Department of Information Systems, Faculty of Computing and IT, Northern Border University, KSA.

M. Kherallah is with Faculty of Sciences, University of Sfax, Tunisia.

accuracy of this model using FER2013 is 77.81%. But the bilinear aggregation function with the square root function of the matrix consumes extreme memory and processor, which decreases the performance of the model.

The authors in [7] use viola-jones for face detection. In addition to Gabor features and LBP features, they propose Joint geometric features. Then, these features are used as input for the multilayered CNN. Finally, an SVM classifier is used for expression recognition.

Niu et al. [8] focused on facial expressions from static images. They used the Dlib library to detect faces. Then, they proposed fused features using Local Binary Pattern (LBP) and Oriented Fast and Rotated Brief (ORB) descriptors. Finally, the combined features are classified by Support Vector Machine (SVM) to recognize seven facial emotions (six basic facial emotions and one neutral emotion), in which the accuracy is 79.8%.

Saroop et al. [9] propose a deep learning approach for emotion recognition that uses Facial Action Units (AU) to extract features from the face and apply CNN to classify the seven emotions of the FER-2013 dataset with an accuracy equal to 67.91%. Another work of deep learning approach for emotion recognition is proposed in [10]. It is based on an attentional convolutional network that focuses on feature-rich areas and takes advantage of the visualisation technique of [11] to highlight the salient part of the face. It achieves 70.02% accuracy on FER-2013.

III. Facial Emotion Recognition Approach

Based on Transfer Learning (TL), we propose our facial emotion recognition approach named **CV-FER**. The name refers to the networks used, **C**NN and **V**GG16, in our **F**acial **E**motion **R**ecognition system. Transfer learning's key advantage is that it shortens the time needed to create and train a model by reusing the weights of previously created models. The current section presents the proposed CV-FER architecture, as shown in Fig. 1.

The input of the model was in CSV format, and each element on the table was converted to a vector and then to an image. In the pre-processing stage, we apply rotation to some images with a range equal to 10. Then, we flip the image to the horizontal direction, and we shift the image to the left or right with a width and height range equal to 0.1, and we choose the nearest as fill-mode. Then, we proceed to data augmentation in order to feed into the network with a larger number of images. After data generation, we define our model that is based on the VGG16 model [13]. We then train our model to track how much improvement the CNN provides in classifying the emotion.

Our FER model, named CV-FER, contains five blocs of convolutions layers, each bloc followed by a max pooling layer. The max-pooling approach has been shown to result in speedier convergence and improved generalization; hence it is often employed for the sub-sampling layer. The filter numbers of each convolution bloc are respectively from 256 to 16 (256,128, 64, 32, 16) and stride (1,1).

Our CNN parameters used are derived from multi-class classification as our interest in this paper. On the first hand, and for the activation function, we used Rectified Linear Unit (ReLU) [4] to convert all the input values to positive numbers, which requires minimal load. On the other hand and for loss function, we used Categorical Cross-entropy because we have seven label classes. This function computes the cross-entropy loss between the labels and predictions. For the optimizer, we use Adam, which is the abbreviation of adaptive moment estimation [14]. Adam is the best optimizer to train an ANN in less time and more efficiently converge rapidly. For the classifier, we used Softmax which gives a more normalized class as output and converts them from weighted sum values into probabilities.

IV. Results

For CV-FER, our system is trained, tested, and validated using Facial Expression Recognition (FER-2013) [12]. Some samples of images in FER-2013 dataset are presented in Fig. 2. In this dataset, there are seven emotions recorded, which are angry, disgust, fear, happy, sad, surprise and neutral. It contains 35,887 images of 48x48 resolution with 28,709 samples for the training set, 3589 samples for the public test set, and 3589 samples for the private test set.

We chose this dataset as it is well-defined and enormous, and it is considered as a challenge. On the one hand, due to the variations in the head position and the illumination of images like low-contrast images, as well as, this dataset contains facial occlusions such as partial faces or hands in front of faces, eyeglasses like those mentioned in Fig. 3. Note that facial occlusion is one important aspect that has an impact on face recognition performance. On the other hand, it is imbalanced in category distribution where some classes have more examples than others, like Happy, which has 8.989 samples of image. However, Disgust has only 547 images.

Our model was trained for 100 epochs from scratch. As shown in Fig. 5, in epoch 0, the loss is 1.0, and the accuracy is about 62%. In epoch 100, the loss is about 0.2, and the accuracy is about 92%. So, we conclude that the lower the loss, the better the model. The comparison between the proposed algorithm and some of the previous works on FER-2013 dataset, are provided in Table I. We can conclude from this table that our method remarkably outperforms the other FER methods, which are investigated in [5], [6], [9] and [10] and outlined in section II.

TABLE I
Performance Comparison between Our Method with Other Methods with FER-2013

Method	Accuracy
Bilinear Pooling [6]	77.81%
CNN [5]	67.7%
Deep Learning [9]	67.91%
Attentional Convolutional [10]	70.02%
The proposed method	**92%**

The confusion Matrix figured in Fig. 4 helps us identify the correct prediction of a model. The model shows a good classification of *happy* 82% compared to *disgust* 60%, and this is related to the number of samples in each category, as

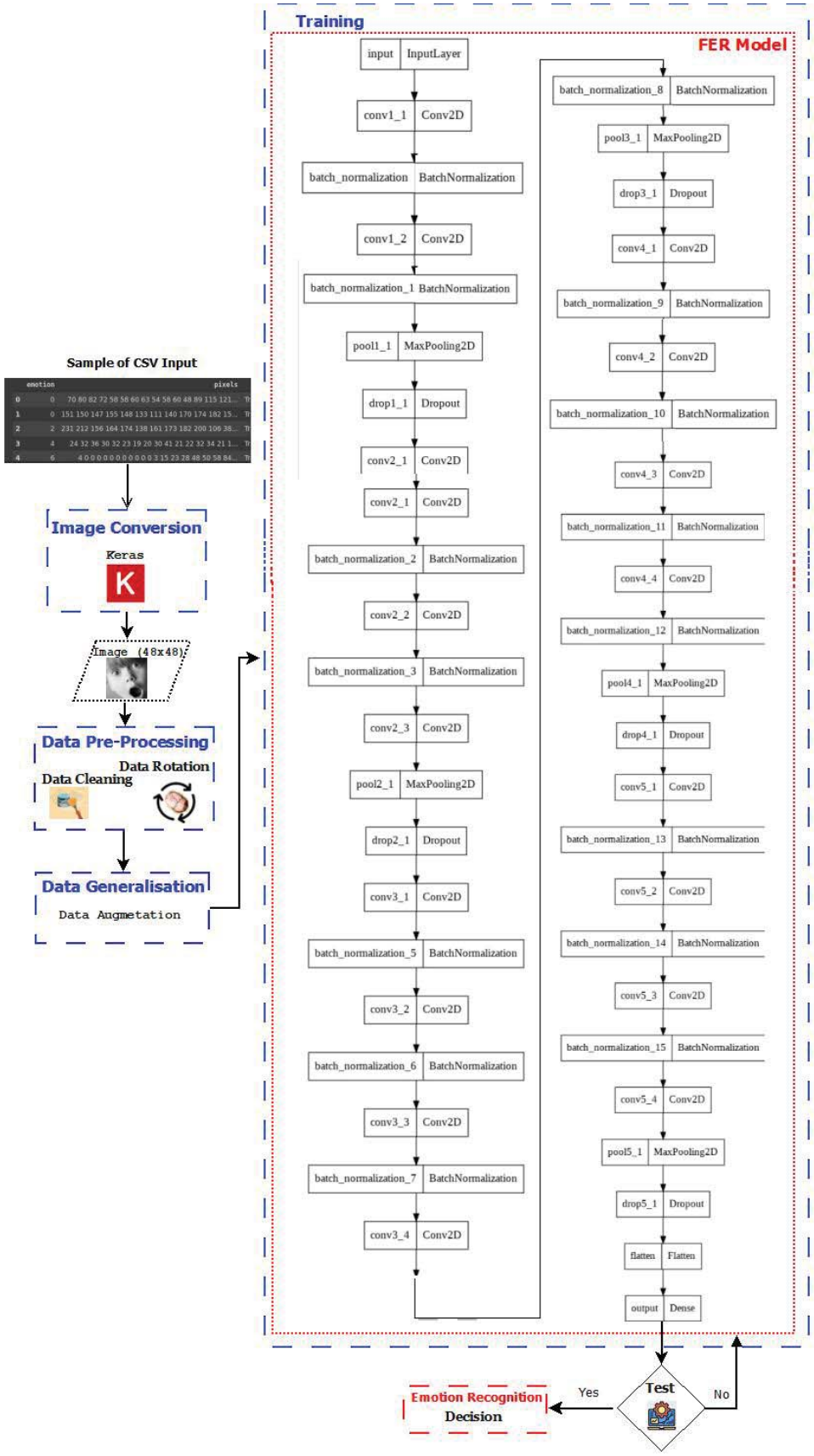

Fig. 1 Our CV-FER Architecture

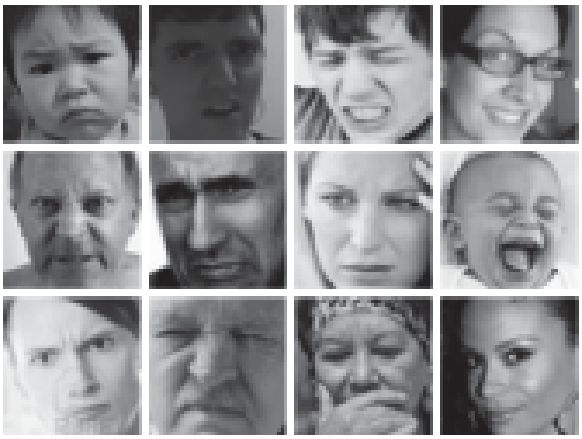

Fig. 2 Samples of FER-2013 Images

Fig. 3 Samples of facial occlusions from the FER-2013 dataset

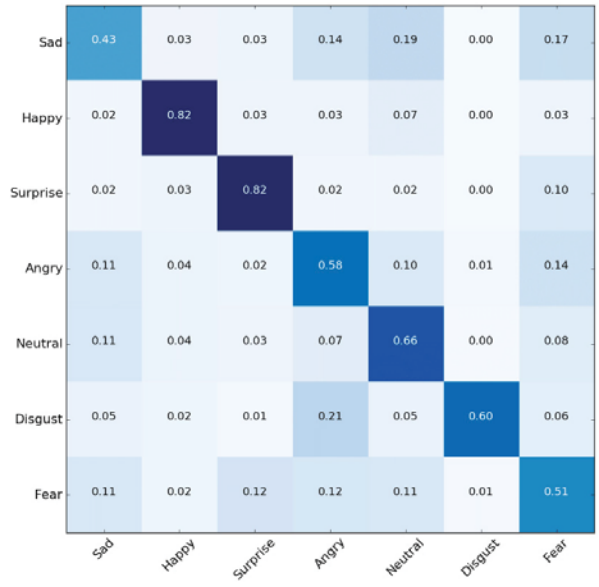

Fig. 4 Confusion Matrix for CV-FER

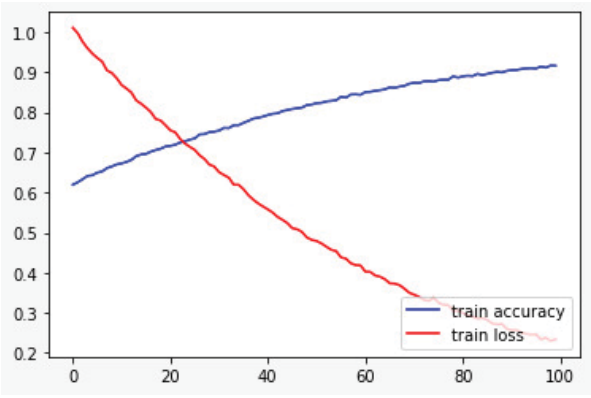

Fig. 5 Accuracy and Loss Plot for CV-FER

Fig. 6 Image samples with confusion in labels

discussed prior. As we can see also, the model is making more mistakes for *fear* 51% and *sad* 43% due to the ambiguity and confusion of image's interpretations in some cases in which the same image can have two possible labels, and this is caused by Bayes error as shown in Fig. 6. The two first images in Fig. 6 show confusion between anger and sadness as emotions. At the same time, the ambiguity between fear or sadness as an emotion is figured in the two last ones.

V. Conclusion and Future Work

In this paper, we proposed an approach to emotion recognition through facial expression named CV-FER. We presented a deep learning-based methodology that uses a CNN. The experimental evaluation carried out on the FER2013 dataset shows the benefit of our approach. Our work is not without limitations. In some cases, we cannot recognize the right emotion due to ambiguities found in some images. In the future, we plan to tackle this limitation by trying another dataset. As we mentioned before that, facial expression can be expressed in an image channel or video channel, and this paper was focused on an image channel only. So, our future directions may include utilizing video channels as well as various communication modalities.

References

[1] N. Khédiri, M. Ben Ammar and M. Kherallah, "Towards an online Emotional Recognition System for Intelligent Tutoring Environment.", (ACIT'2017) The International Arab Conference on Information Technology, Yassmine Hammamet, Tunisia, 22-24 December 2017.
[2] M. Moolchandani, S. Dwivedi, S. Nigam and K. Gupta, "A survey on: Facial Emotion Recognition and Classification," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 1677-1686.
[3] Naga, P., Marri, S.D. and Borreo, R., 2021. "Facial emotion recognition methods, datasets and technologies: A literature survey". Materials Today: Proceedings.
[4] Nair, V., Hinton, G.E.: "Rectified linear units improve restricted boltzmann machines". In: ICML. pp. 807–814 (2010).
[5] Singh M., Sharma S., Paul S., Sajeevan J., and Paul S., "Facial Emotion Recognition system" Journal of Scientific Research and Advances, volume 6, issue 6,2020.
[6] Mahmoudi, M., Chetouani, A., Boufera, F., and Tabia, H. (2020), "Improved Bilinear Model for Facial Expression Recognition", Pattern Recognition and Artificial Intelligence, 1322, 47 - 59.
[7] Hao Meng, Fei Yuan, Yue Wu and Tianhao Yan, "Facial Expression Recognition Algorithm Based on Fusion of Transformed Multilevel Features and Improved Weighted Voting SVM", Mathematical Problems in Engineering,1-117,2021.
[8] Ben Niu, Zhenxing Gao and Bingbing Guo, "Facial Expression Recognition with LBP and ORB Features", Computational Intelligence and Neuroscience,2021.
[9] Saroop, A., Ghugare, P., Mathamsetty, S., and Vasani, V. (2021), "Facial Emotion Recognition: A multi-task approach using deep learning". ArXiv, abs/2110.15028.

[10] Minaee, S., and Abdolrashidi, A. (2021). Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network. Sensors (Basel, Switzerland), 21.
[11] Zeiler, D.M.; Fergus, R., "Visualizing and understanding convolutional networks", In European Conference on Computer Vision;Springer: Cham, Switzerland, 2014.
[12] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee et al., "Challenges in representation learning: A report on three machine learning contests," in International Conference on Neural Information Processing. Springer, 2013, pp. 117–124.
[13] Simonyan K. and Zisserman A., "Very Deep Convolutional Networks for Large-Scale Image Recognition", 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.
[14] Diederik P. Kingma, Jimmy Ba, "Adam: A Method for Stochastic Optimization", 3rd International Conference for Learning Representations, San Diego, 2015.

Nouha Khédiri received her Master degree in Intelligent Information Systems from Higher Institute of Informatics and Management of Kairouan. She is currently, a Lecturer of the Department of Information Systems, Faculty of Computing and IT, Northern Border University, Rafha, Kingdom of Saudi Arabia and is pursuing her Doctoral degree at Faculty of Mathematical, Physical and Natural Sciences of Tunis. Her research interests include affective computing, intelligent tutoring system and Emotion recognition.

Mohammed Ben Ammar is currently, an Assistant Professor of the Department of Information Systems, Faculty of Computing and IT, Northern Border University, Rafha, Kingdom of Saudi Arabia. He received his PhD. Degree in engineering of Information system from Sfax University-Tunisia, National Engineering School of Sfax (ENIS). His research interests include affective computing, intelligent tutoring system and Emotion recognition.

Monji Kherallah is currently, an Assosiate Professor in Faculty of Science of Sfax. He received his HDR from National Engineering School of Sfax (ENIS), Sfax University-Tunisia. His research interests include the Handwritten Documents Analysis and Recognition. He is one of the developers of the ADAB-Database (used by more than 50 research groups from more than 10 countries). He is co-organizer of the Online Arabic Handwriting Recognition Competitions at ICDAR 2009 and ICDAR 2011. He has more than 70 papers, including journal papers and book chapters. He is a member of IEEE and IEEE AESS Tunisia Chapter Chair, 2010 and 2011. He is reviewer of several international journals.