Enhancing Emotion Detection Through CNN-Based Facial Expression Recognition

Jinyang Wang

Johnbapst Highschool, Bangor, ME 04401, USA
jwang26@johnbapst.org

Abstract. Artificial intelligence-based approaches, such as Convolutional Neural Networks (CNN), hold significant promise for emotion detection, particularly in facial expression recognition, offering invaluable insights for various sectors including business, medicine, and psychology. This paper explores the utilization of CNN for facial feature extraction to discern emotions, employing preprocessing procedures to standardize images sourced from the Kaggle website. Methods including blurring, scaling, contour image alteration, and normalization are employed for standardization, facilitating accurate feature extraction and emotion detection. Despite the separation of classification and feature extraction phases in CNN, which necessitated extensive effort to enhance performance, the technique ultimately offers superior accuracy compared to traditional classifiers. However, challenges arise from noisy and deviated images in the dataset, impacting the efficacy of the CNN model. To mitigate these challenges, preprocessing techniques such as grayscale conversion, resizing, and normalization are applied to standardize dataset images. The research aims to identify weaknesses in existing models and develop improvements to conventional emotion detection techniques. Experimental results underscore the value of combining these techniques for precise prediction of facial expressions, contributing to advancements in emotion detection methodologies.

Keywords: Emotion Detection, Convolutional Neural Networks (CNN), Identify Weaknesses, Preprocessing Techniques.

1 Introduction

The main objective of this study is to implement a Convolutional Neural Networks (CNN) approach for emotion detection. Specifically, the CNN approach analyzes and matches the images to the emotion. The characteristics of the methods applied in emotion detection correspond to the problems solved. Facial emotion detection is a crucial feature in businesses and institutions that need more information about people or an additional layer of security. Human emotion detection is applied as an extra step to face detection to infer the visitors' emotions. This is crucial in verifying that the individual visible in the camera is not a 2-dimensional representation [1]. The other vital application of human emotion detection is in business promotions since positive

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_100

consumer response to their offers and products is the primary driver of their success. Therefore, the accuracy of an artificial intelligence-based system in identifying real-time emotions based on facial expressions can aid businesses in making informed decisions on whether their clients like or even dislike their offers.

Studies in this area have investigated different algorithms and databases and have demonstrated different potentials in performance computation and accuracy. Most of the methods applied encountered the problem of adjusting the hyperparameters and determining the model with better experimental optimization. Since this research emphasizes using a dataset from Kaggle, the primary focus would be research that applies the same dataset. Based on the studies above, this study finds the use of CNN, GoogleNet, and the Inception layer to be the most effective. Similarly, Kim et al. did a study with an additional two layers of CNN ensampling 36 networks and obtained an accuracy of 72.71%. This model was later improved by Connie et al. by proposing a hybrid model of Scale-invariant Feature Transform (SIFT) aggregator that combined three models, thereby helping obtain a 73.4% accuracy [2]. A year later, Jun et al. proposed a methodology inspired by the 19-layered Visual Geometry Group Network (VGGNet), which included a variation where images were trimmed by 4 pixels and then mirrored, achieving an accuracy of 73.05% [3]. Hua et al. deviated from the previous studies and offered a 3-model research with 68. 18% accuracy. Their result did not improve the earlier studies but offered a justification for using ensemble methods [4]. Therefore, emotional recognition in the interaction between a computer and a person is integral for institutions and businesses to understand the manifestations of the specific behavior of the individual.

On the other hand, Poru et al. applied the RESNET50 method by training in the VGGFace2 dataset and achieved accuracy of 71.25%. In addition to employing necessary data augmentation, that also utilize Contrast Limited Adaptive Histogram Equalization (CLAHE) to enhance the contour and compare of the image [5]. In 2019, Georgescu and colleagues introduced a sophisticated model that combined Bag of Visual Words (BOVW) with a CNN, achieving a 75.42% accuracy. In the wave of increasing interest in machine learning, Kusuma and team crafted a novel approach using the VGG-16 model, which reached a 69.40% accuracy rate. Their adjustments involved leveraging Global Average Pooling (GAP) and incorporating features like optimizers and batch normalization [6, 7]. Concurrently, Riaz and associates proposed a streamlined eXnet network with 4.57 million parameters, employing advanced techniques like blending and shearing, achieving a 73.54% accuracy [8].

From the previous studies, it suffices that the predictive performance of the different models yields different results in terms of accuracy. Nevertheless, approaches that combined two or more techniques were more likely to obtain better performances. Consequently, this research thoughtfully reflects on earlier studies, particularly focusing on the benefits of ensemble methods and the application of the dataset. It starts with adjustments to the dataset and a suggested approach involving a few layers and transfer learning techniques. The study promotes a more advanced ensemble model to enhance performance. For instance, Kim et al. improved on the previous studies by suggesting an additional two-layered approach to CNN that improved the performance of their approach. However, their approach still suffered

from performance issues [9]. Hua et al. model justifies the use of ensemble methods. The experimental results demonstrate that combining techniques is the best approach to improving the accuracy and efficiency of emotion detection approaches [4]. Therefore, this study is significant because it advocates for adopting a combination of both CNN and DNN to solve the problem of the inefficacy of the existing models in emotion detection.

2 Methodology

2.1 Dataset Description and Preprocessing

The dataset for the model training was obtained from Kaggle, which contained the facial emotions of people received from social networks [10]. The dataset was segmented into three parts: A training set used to train the model, a testing set for evaluating the model, and a validation set used during the tuning of model hyperparameters. The dataset labels encompass seven emotional categories: anger, contempt, disgust, fear, happiness, sadness, and surprise. This dataset poses a challenge with a human-level accuracy benchmark of approximately 65 ± 5%. The highest accuracy achieved to date, as reported in this research, is 75%. Additionally, the dataset exhibits significant imbalance in its distribution, with counts of angry (4,953), neutral (6,198), happy (8,989), be scared (5,121), grieved (6,077), and disgust (547).

2.2 Proposed Approach

Emotion recognition is a field of research that is integral in different human-computer interactions and other fields that require understanding human emotions. However, the recognition of human emotions is an endeavor that requires multiple components. Fig. 1 illustrates the pipeline of the proposed model with the different elements. The dataset/input is obtained from the Kaggle website dataset. Ensuring the model learns and recognizes emotions from images showing different emotions requires extensive data. Since the data received from the website could be more organized, it should first be preprocessed. This requires that the data be standardized before it is supplied to the machine as input for the learning model. Therefore, preprocessing plays an essential role in ensuring that the complexity of the data is decreased to improve the accuracy of the model [1]. Preprocessing will involve resizing the images from Kaggle's website, converting to grayscale, and normalizing. The preprocessed images are then added to the model as input for classifying the different emotion states. Facial emotional recognition faces challenges such as Inter-Class Variation and Intra-Class Variance. Intra-class variance arises from the diverse expressions of the same emotion, complicating standardization. Conversely, Inter-Class Variation stems from the varied expressions of different emotions among individuals. Addressing these challenges necessitates hierarchical representations. Key steps in this pipeline include dataset acquisition in CSV format, containing pixel values and labels, followed by array storage. Given dataset imbalance, balancing techniques like oversampling and

under sampling are employed. Oversampling enhances minority class representation, after which pixel values are normalized.

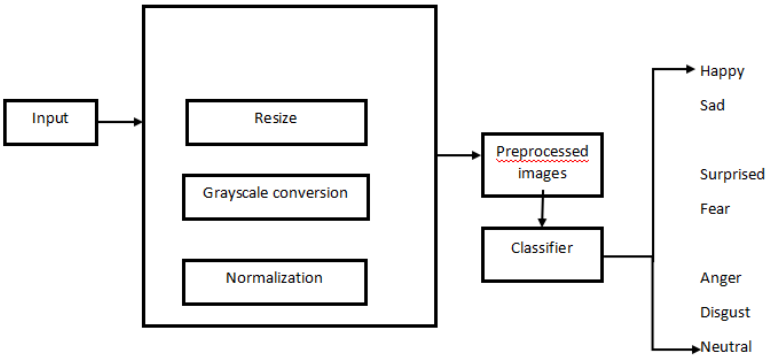

Fig. 1. The pipeline of the model.

Diffusion model. The diffusion model is a novel approach that leverages cross-attention layers to adaptively condition inputs, allowing for the processing of high-resolution images with reduced computational demands while preserving fine details. At its core, the diffusion model introduces an additional layer of detail to facilitate the conceptualization of detection tasks. This model is designed to learn the inverse of the diffusion process during training, iteratively refining images after interference. The key principle behind the diffusion model lies in its ability to enhance image fidelity by training during the latency phase. By iteratively refining images through the diffusion process, the model effectively captures intricate details while minimizing computational overhead.

Operationally, the diffusion model follows a structured workflow. Initially, it conditions inputs using cross-attention layers, enabling the processing of high-resolution images. The model then iteratively refines these images, learning from the diffusion process and incorporating additional layers of detail to improve detection accuracy. Internally, the diffusion model's architecture is characterized by its sophisticated layering and attention mechanisms. Cross-attention layers facilitate adaptive conditioning, while additional layers introduced during training enhance image fidelity and detail. This internal structure enables the model to effectively handle complex image processing tasks with high precision and efficiency.

Loss Function. The loss function is essential for every machine learning algorithm because it is the objective measure of the model performance and guides the learning process. The choice of fitting loss function is an important part of the success of a deep learning model. Therefore, it’s used to predict the error between sentiment and actual values in validation data. This study applied a categorical cross-entropy loss function model. It demonstrates the performance of a classification model by

outputting probability values (between 0 and 1). These output values are converted because of changes between predicted probabilities and actual outputs.

2.3 Implementation Details

This research applied Python 3. 10 and the Scikit-learn library to implement the decision tree models. Seaborn and Matplotlib libraries were used for data visualization. The settings of the decision tree were such that the Gini index. Preprocessing the images reduced both the validation and training losses. The difference between the two loss curves is an indication of how the proposed model underfits or overfits the dataset. Therefore, preprocessing is crucial for solving issues related to overfitting. The dataset used in this study was obtained from the Kaggle. Initially, the dataset was divided at random into three segments with the following distribution: 80% for train, 10% for verify, and 10% for test. Each image in the dataset features a cut short face measuring 48 × 48 pixels. The main application involves prepared models on Kaggle datasets. The process involves manually refining the original dataset by removing images that do not depict human faces and accurately reclassifying mislabeled images into appropriate categories. To further balance the dataset, basic data augmentation techniques such as scaling, rotation, and mirroring were employed.

3 Results and Discussion

Once the 32 epochs had been completed, the accuracy of this model is approximately 83% on the test data with a 0.8 Loss. The implementation of class weighting increased the efficacy of the misclassified emotions. A plot of the model accuracy to the epochs can be summarized in the Fig. 2.

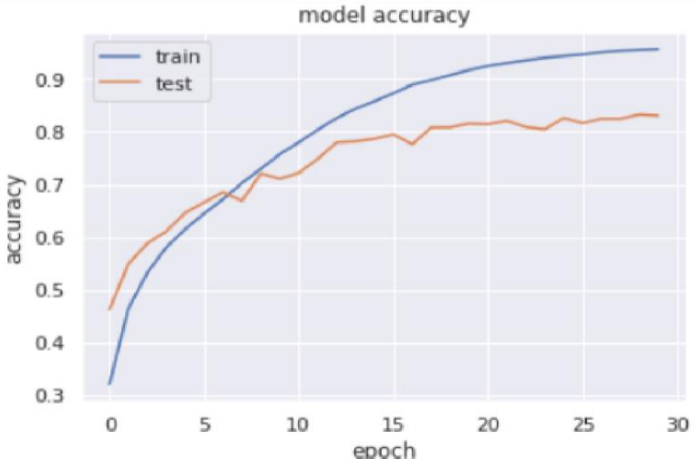

Fig. 2. The accuracy of each epoch

Table 1. Results of training.

S. No	Epoch	Training Loss	Validation Loss
1	5	1.0401	1.5158
2	10	1.6078	1.2239
3	15	1.4065	1.1954
4	20	1.1691	1.1488
5	25	1.1864	1.1173
6	30	1.1274	1.0917
7	35	1.0848	1.8639
8	40	1.0481	1.0323
9	45	1.0173	1.0101
10	50	0.9795	0.9825
11	55	0.8657	0.8659
12	60	0.8462	0.8462
13	65	0.7243	0.8243
14	70	0.8065	0.8542
15	75	0.7723	0.8834
16	80	0.7501	0.9156

Extensive experiments were conducted to evaluate the proposed model's accuracy in classifying facial expressions to their correct emotions. The Experimental results in Fig. 3 show that the model used had the best performance in both experiments, with validation rates of 73.2% and 73.5%.

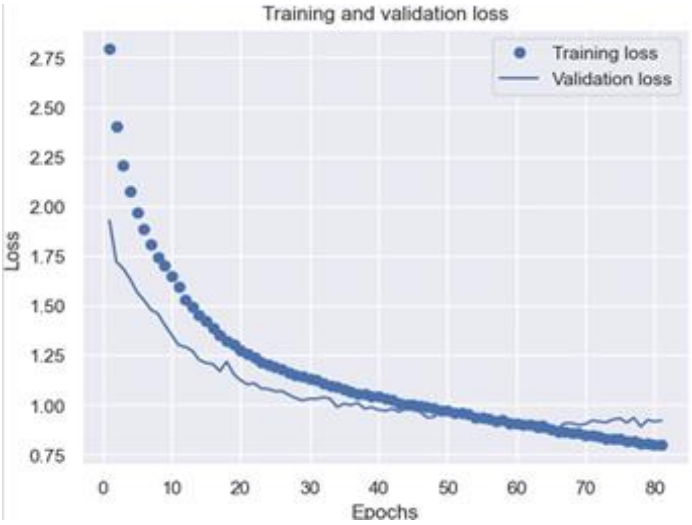

Fig. 3. The loss of each epoch.

Table 1 shows how image preprocessing reduced the difference between the validation and training losses. The image depicts a good baseline for an epoch above 50.

4 Conclusion

This study presents an emotion recognition model designed to classify the seven facial expressions from the Kaggle dataset. Thorough tuning of hyperparameters was conducted to optimize the model for facial recognition, resulting in significant improvements over previous research. Through rigorous experimentation, various enhancements to existing methodologies were explored, ultimately achieving an accuracy of 73.06%, surpassing all previously reported performances.

Preprocessing of images from the dataset involved feature extraction and feeding into a CNN model for facial expression detection. Model evaluation was conducted using accuracy as the primary metric, revealing promising results. Specifically, the combination of normalization procedures with rescaling and grayscale conversion demonstrated notable improvements in performance accuracy. The proposed model represents a simplified yet highly effective approach to facial recognition, outperforming previous models employing similar classification methods. Notably, the training process exhibited considerably reduced time requirements, further enhancing the model's practical utility. In essence, this research underscores the potential of leveraging advanced techniques in emotion recognition to achieve superior accuracy levels. By offering a streamlined and efficient solution, this model enhances the ongoing development of facial recognition technology, with significant implications for fields such as human-computer interaction, security, and healthcare.

References

1. Dabhi, V.K., Prajapati, H.B., Vyas, A.S.: Survey on Face Expression Recognition Using CNN. Presented at the 5th International Conference on Advanced Computing & Communication Systems (ICACCS), pp. 102–106. Coimbatore, India (2019).
2. Al-Shabi, M., Cheah, W.P., Connie, T., Goh, M.: Facial Expression Recognition Using a Hybrid CNN–SIFT Aggregator. Discussed at the Multi-Disciplinary Trends in Artificial Intelligence conference, 139–149 (2017)
3. Jingwei, W., Jinming, S., Jun, H., Peng, J., Shuai, L., Yue, L.: Facial Expression Recognition Based on VGGNet Convolutional Neural Network. Delivered at the 2018 Chinese Automation Congress (CAC), pp. 4146–4151. Xi'an, China (2018).
4. Dai, F., Gui, G., Hua, W., Huang, L., Xiong, J.: HERO: Human Emotions Recognition for Realizing Intelligent Internet of Things. Featured in IEEE Access, 7, 24321–24332 (2019).
5. Leon, F., Miron, C., Porușniuc, G.C., Timofte, R.: Convolutional Neural Networks Architectures for Facial Expression Recognition. Presented at the 2019 E-Health and Bioengineering Conference (EHB), pp. 1–6. Iasi, Romania (2019).
6. Georgescu, M.I., Ionescu, R.T., Popescu, M.: Local Learning with Deep and Handcrafted Features for Facial Expression Recognition, pp. 64827-64836. Published in IEEE Access, (2019).

7. Jonathan, A.P.L., Putra Kusuma, G.: Emotion Recognition on FER-2013 Face Images Using Fine-Tuned VGG-16. Advanced Science, Technology and Engineering Systems Journal, 5, 315–322 (2020).
8. Guo, M., Riaz, M.N., Shen, Y., Sohail, M.: eXnet: An Efficient Approach for Emotion Recognition in the Wild. Published in Sensors, 20, 1087 (2020).
9. Dong, S.Y., Kim, B.K, Lee, S.Y., Roh, J.: Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. Journal of Multimodal User Interfaces, 10, 173–189 (2016).
10. Emotion detection, https://www.kaggle.com/code/vinitkp/emotion-detection, last accessed 2024/1/18.