Multiple Optimized Deep Learning Models for Effective Facial Expression

Ruoyu Li

Department of Cognitive Science, University of California, California, 95618, USA
email: pryli@ucdavis.edu

Abstract. Facial expression recognition is an essential domain within computer vision, focused on interpreting human emotions through facial cues for enhanced human-computer interaction. This study examines the current state and challenges in facial expression recognition, emphasizing the role of deep learning architectures like CNNs, ResNet, and VGG in driving advancements in this field. These models have improved system performance by enabling more precise feature extraction and efficient pattern recognition. However, the generalization of these systems to diverse, real-world environments remains a significant challenge due to factors like inconsistent lighting, occlusions, and varied facial orientations.This research contributes to overcoming these limitations by proposing a novel deep learning-based architecture that optimizes the recognition process across different scenarios and demographic variances. The study leverages extensive datasets like FER2013 and incorporates advanced model training techniques, including transfer learning, to enhance the robustness and accuracy of facial expression recognition systems. By addressing these challenges, the study aims to refine the technology to be more adaptive and sensitive to a wide array of emotional expressions, thereby supporting the development of more intuitive and engaging user interfaces that can integrate seamlessly into daily human interactions and applications. This will potentially revolutionize interactions within digital environments, making them more humane and responsive to emotional feedback.

Keywords: Facial expression recognition, deep learning, artificial intelligence

1 Introduction

Facial expression recognition is a field of computer vision and artificial intelligence that focuses on the automatic detection and interpretation of facial cues [1-3]. By analyzing facial features and patterns, it aims to identify and classify the emotional states expressed by individuals e.g. anger, disgust, fear, happiness, sadness, surprise, and neutrality. The ability to accurately recognize and interpret facial expressions has far-reaching implications. It is of paramount importance for enhancing human-computer interaction, enabling empathetic and intuitive interfaces that can respond to the emotional states of users. Furthermore, facial expression recognition contributes to advancements in affective computing, where machines can better understand and

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_62

empathize with human emotions, leading to more engaging and personalized interactions.

Previous research in facial expression recognition has made significant strides, particularly with the integration of deep learning techniques like Convolutional Neural Networks (CNNs) and advanced model architectures such as ResNet and VGG [4-6]. These innovations have facilitated sophisticated pattern recognition and improved the performance of expression recognition systems in controlled environments. However, a current gap exists in the ability to generalize these solutions to real-world, uncontrolled scenarios. The implementation of CNNs in the domain of facial recognition has been transformative, allowing for sophisticated pattern recognition capabilities that are critical for interpreting the nuances of human expressions. The convolutional layers of CNNs have facilitated the extraction of distinguishing features from facial data, while advanced models like VGG have provided further refinements, offering nuanced methods to analyze facial expressions through robust feature extraction.

Additionally, the incorporation of residual architectures, such as the ResNet model, which include residual blocks, has aided in avoiding overfitting and enabled the network to perform well even with deeper architectures. This has bolstered the capability of expression recognition systems to function effectively across a multitude of scenarios and datasets, paving the way for emotion-sensitive interfaces that respond to the spectrum of human emotions in real-time.

Challenges persist in handling variations in head pose, lighting conditions, and occlusions, which can significantly impact the accuracy of expression recognition. Additionally, the classification of a wide range of emotional expressions, including subtle and complex ones, remains an ongoing research challenge. Previous studies have laid the foundation for facial expression recognition by demonstrating the effectiveness of deep learning approaches, particularly the use of CNNs, ResNet, and VGG models. These contributions have enabled significant advancements in feature extraction, pattern recognition, and the overall performance of expression recognition systems. Furthermore, the availability of curated datasets, such as the FER2013 and CK+ datasets, has been instrumental in facilitating research and enabling comparative analysis across different methods and architectures. These datasets have provided the necessary ground-truth data for supervised learning and model evaluation, driving progress in this field.

This study aims to address the current limitations in facial expression recognition by developing a robust and accurate system that can effectively categorize a diverse range of emotional expressions, including anger, disgust, fear, happiness, sadness, surprise, and neutrality. By leveraging state-of-the-art deep learning techniques and exploring novel architectural designs, the study seeks to enhance the model's ability to capture the nuances of facial expressions while maintaining generalizability across different environments and demographic groups. Designing a novel deep learning-based architecture that can effectively extract and represent the distinctive features associated with various facial expressions, enabling accurate classification across a broad spectrum of emotional states. Developing strategies to enhance the model's robustness and generalization capabilities, enabling it to perform well in real-world, uncontrolled scenarios with variations in head pose, lighting conditions, and occlusions. By addressing the current limitations and advancing the state of the art in facial expression recognition, this study aims to contribute to the development of more intuitive, emotion-aware, and engagement-centric applications that can seamlessly integrate with the social fabric and psychological paradigms.

2 Methodology

2.1 Dataset Preparation

The FER2013 dataset is a widely used benchmark for facial expression recognition, sourced from the popular data repository Kaggle [7]. The dataset consists of 48×48 pixel grayscale images of faces, with the faces automatically registered to be more or less centered and occupying a similar amount of space in each image. In total, the FER2013 dataset contains 35,887 facial images, with a training set of 28,709 examples and a public test set of 3,589 examples. The task is to categorize each facial image into one of seven emotion categories: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral.

The distribution of the emotion categories in the FER2013 dataset is not entirely balanced, with the Disgust expression having the minimal number of images at 600, while the other labels have nearly 5,000 samples each. This imbalance in the dataset can pose challenges in the development and evaluation of facial expression recognition models. Fig. 1 presents the distribution of different facial expressions.

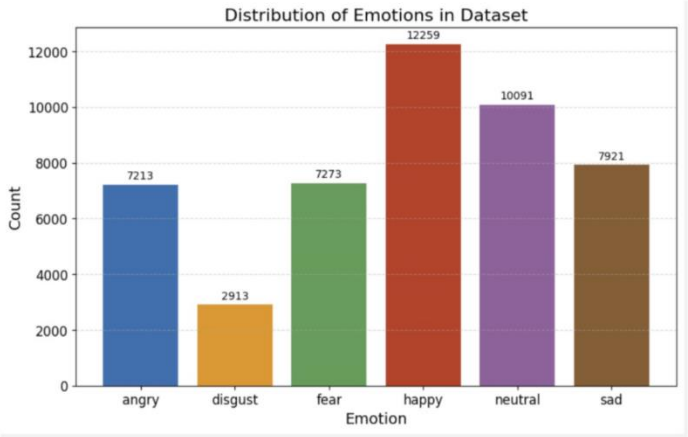

Fig. 1. The distribution of different facial expression.

2.2 CNN Models

MobileNetV2 is a convolutional neural network architecture designed specifically for mobile and embedded vision applications [8, 9]. The model is 53 layers deep and uses an "inverted residual structure" where the residual connections are between the bottleneck layers. The intermediate expansion layer in MobileNetV2 utilizes lightweight depthwise convolutions as a source of non-linearity. The overall architecture of MobileNetV2 consists of an initial fully convolution layer with 32 filters, followed by 19 residual bottleneck layers. Compared to larger CNN models, MobileNetV2 achieves comparable accuracy while using significantly fewer parameters, making it well-suited for mobile device use. In the pursuit of refining facial expression classification, the experimental journey commenced with the utilization of the MobileV2 model, renowned for its expeditious image classification capabilities and proficient performance. ResNet and DenseNet are two influential convolutional neural network architectures that have significantly advanced the field of deep learning. ResNet [10], or Residual Network, introduced the concept of residual blocks which allow for the training of extremely deep networks by incorporating skip connections that help mitigate the vanishing gradient problem. This design facilitates "memory" capabilities and information reuse, making ResNet a robust model for various image recognition tasks. On the other hand, DenseNet, or Densely Connected Network, employs a different approach by ensuring that each layer receives inputs from all preceding layers, fostering feature reuse and reducing the number of parameters. This dense connectivity pattern enhances gradient flow and improves network efficiency. Both models have distinct architectural attributes that make them powerful tools for tasks such as facial expression classification, where capturing subtle details and maintaining robust generalization are crucial.

The initial phase entailed freezing the weights of the convolution layers to capture distinctive photo features, with custom fully connected layers subsequently harnessed for classification. Notably, the model's optimizer was Adam, and for efficient resource utilization, an early stopping function was instituted, monitoring validation accuracy, coupled with the "ReduceLROnPlateau" callback function to modulate the learning rate on plateaus and navigate minima. The training spanned 50 epochs, each lasting approximately 7 to 9 minutes.

Following the initial fitting process, the model underwent a phase of reevaluation, involving the adaptability of the weights acquired in the preceding training session by making the convolution layers trainable. Despite an extended training duration spanning 15 epochs, the model's performance didn't register significant improvements, prompting a comprehensive analysis to identify avenues for enhancement. This led to the utilization of the F1-score and confusion matrix to gain insights into class-specific accuracies.

Given the discerned issues, the study delved deeper into the dataset, unearthing a class imbalance wherein the "disgust" category comprised a mere 436 images juxtaposed with a staggering 7265 images for the "happy" class, sparking concerns about skewed accuracies. Aiming to rectify these imbalances, two pivotal strategies emerged - augmenting the dataset to address class disparities and combatting overfitting. Addressing this imbalance involved diversifying dataset augmentation techniques due to the simplistic nature of the existing data generator. Subsequently, the experimental trajectory transitioned towards configuring the MobileV2 model with the FER2013 dataset, primarily leveraging 'image net weight' and frozen convolution layers. This highlighted potential limitations within the FER2013 dataset, prompting the integration of the Facial Expression Training Data to introduce diverse, colorful images, and enable transfer learning.

The strategic incorporation of the new, diverse dataset presented a transformative opportunity, culminating in a transfer learning approach. This heralded a subtle fine-tuning process, subsequent to the adaptation of previously accrued weights to the new dataset. Thus, conclusively attesting to the effectiveness of integrative transfer learning and dataset diversification in bolstering the model's classification capabilities and performance.

In essence, this comprehensive exploration not only redressed class imbalances and combatted overfitting but also elevated the model's performance through robust regularization techniques and the strategic assimilation of a fresh, diverse dataset. This yielded a substantial uplift in validation accuracy, signifying a substantial evolutionary leap in the realm of facial expression classification techniques.

3 Results and Discussion

In the pursuit of refining facial expression classification, the experimental journey commenced with the utilization of the MobileNet-V2 model, renowned for its expeditious image classification capabilities and proficient performance. The initial phase entailed freezing the weights of the convolutional layers to capture distinctive photo features, with custom fully connected layers subsequently harnessed for classification. The model's optimizer was set to Adam, and for efficient resource utilization, an early stopping function was instituted, monitoring validation accuracy, coupled with the "ReduceLROnPlateau" callback function to modulate the learning rate on plateaus and navigate minima.

The training spanned 50 epochs, each lasting approximately 7 to 9 minutes, culminating in an initial training accuracy of 84% and a validation accuracy of 65%. Following the initial fitting process, the model underwent a phase of reevaluation, involving the adaptability of the weights acquired in the preceding training session by making the convolutional layers trainable. However, despite an extended training duration spanning 15 epochs, the model's performance didn't register significant improvements. This led to the utilization of the F1-score and confusion matrix to gain insights into class-specific accuracies. The analysis unveiled disparities in validation accuracy, with the "happy" class demonstrating the highest accuracy of 85% and the "disgust" class recording the lowest at 48%. Given the discerned issues, the study delved deeper into the dataset, unearthing a class imbalance wherein the "disgust" category comprised a mere 436 images juxtaposed with a staggering 7265 images for the "happy" class, sparking concerns about skewed accuracies. Aiming to rectify these imbalances, two pivotal strategies emerged - augmenting the dataset to address class disparities and combatting overfitting, manifested in a notable 19% gap between training and validation accuracies. Addressing this imbalance involved diversifying dataset augmentation techniques due to the simplistic nature of the existing data generator.

Subsequently, the experimental trajectory transitioned towards configuring the MobileNet-V2 model with the FER2013 dataset, primarily leveraging 'image net weight' and frozen convolutional layers. However, the model's classification performance, marked by an early stopping at around 35 epochs, resulted in a modest validation accuracy of approximately 60%. This highlighted potential limitations within the FER2013 dataset, prompting the integration of the Facial Expression Training Data to introduce diverse, colorful images, and enable transfer learning. The strategic incorporation of the new, diverse dataset presented a transformative opportunity, culminating in a transfer learning approach. This heralded a subtle fine-tuning process, subsequent to the adaptation of previously accrued weights to the new dataset. This iterative approach ensued training for approximately 10 epochs, eventually yielding a notable improvement in validation accuracy, ascending to around 70%. Thus, conclusively attesting to the effectiveness of integrative transfer learning and dataset diversification in bolstering the model's classification capabilities and performance.

In essence, this comprehensive exploration not only redressed class imbalances and combatted overfitting but also elevated the model's performance through robust regularization techniques and the strategic assimilation of a fresh, diverse dataset. This yielded a substantial uplift in validation accuracy, signifying a substantial evolutionary leap in the realm of facial expression classification techniques.

4 Conclusion

Facial expression recognition leverages advanced computer vision and artificial intelligence to discern and categorize human emotions, which is pivotal for augmenting human-computer interaction. By analyzing facial features and expressions, technologies can adapt to emotional cues such as anger, happiness, or sadness, thereby enhancing user experience through empathetic responses. The integration of deep learning techniques, notably CNNs, ResNet, and VGG, has significantly advanced this field, offering robust feature extraction and sophisticated pattern recognition. These technologies have facilitated the development of interfaces that are sensitive to a range of human emotions, responding in real-time and fostering more personalized interactions. Despite these advancements, challenges remain, particularly in adapting these systems to uncontrolled, real-world environments where variations in lighting, occlusions, and head poses can affect accuracy. Moreover, the classification of complex and subtle emotions continues to be an area requiring further research and development. Ongoing advancements aim to refine the accuracy and applicability of these systems, ensuring they are as effective in everyday scenarios as they are in controlled settings.

References

1. Li, S., Deng, W.: Deep facial expression recognition: A survey. IEEE Transactions on Affective Computing 13(3), 1195-1215 (2020).
2. Tian, Y., Kanade, T., Cohn, J.F.: Facial expression recognition. Handbook of Face Recognition, pp. 487-519 (2011).
3. Valstar, M.F., Jiang, B., Mehu, M., Pantic, M., Scherer, K.: The first facial expression recognition and analysis challenge. In: 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG), pp. 921-926. IEEE (2011).
4. Xie, S., Hu, H.: Facial expression recognition with FRR-CNN. Electronics Letters 53(4), 235-237 (2017).
5. Liu, K., Zhang, M., Pan, Z.: Facial expression recognition with CNN ensemble. In: 2016 International Conference on Cyberworlds (CW), pp. 163-166. IEEE (2016).
6. Li, J., Jin, K., Zhou, D., Kubota, N., Ju, Z.: Attention mechanism-based CNN for facial expression recognition. Neurocomputing 411, 340-350 (2020).
7. Kaggle: FER 2013. Available online at https://www.kaggle.com/datasets/msambare/fer2013 (Accessed in 2021).
8. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510-4520 (2018).
9. Dong, K., Zhou, C., Ruan, Y., Li, Y.: MobileNetV2 model for image classification. In: 2020 2nd International Conference on Information Technology and Computer Application (ITCA), pp. 476-480. IEEE (2020).
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778 (2016).