A Data-Centric Approach for Detecting a Neutral Facial Expression using Deep Learning

Lúcia Sousa[1]
luciamsousa00@ua.pt
Daniel Canedo[1]
danielduartecanedo@ua.pt
Miguel V. Drummond[2]
mvd@av.it.pt
João Ferreira[3]
joÃčo.ferreira@vision-box.pt
António J. R. Neves[1]
an@ieee.org

[1] The Institute of Electronics and Informatics Engineering of Aveiro
University of Aveiro
Aveiro, PT
[2] The Institute of Telecommunication
Aveiro, PT
[3] Vision-Box Lisboa, PT

Abstract

Neutral facial expression recognition is of great importance in various domains and applications. This study introduces a data-centric approach for neutral facial expression recognition, presenting a comprehensive study that explores different methodologies, techniques, and challenges in the field to foster a deeper understanding. The results show that data augmentation plays a crucial role in improving dataset performance. Additionally, the study investigates different model architectures and training techniques to identify the most effective approach, with the InceptionV3 model achieving the highest accuracy of 72%. Furthermore, the research examines the influence of preprocessing methods on the performance of both InceptionV3 and a simplified CNN model. Interestingly, the results indicate that preprocessing techniques positively affect the performance of the simpler CNN model but negatively impact the InceptionV3 model. The implemented system, used to evaluate the findings, demonstrates promising results, correctly classifying 77% of neutral expressions. However, there are still areas for improvement. Creating a specialized dataset that includes both neutral and non-neutral expressions would greatly enhance the accuracy of the system. By addressing limitations and implementing suggested improvements, neutral facial expression recognition can be significantly enhanced, leading to more effective and accurate results.

1 Introduction

The neutral facial expression, also called a poker face, is the absence of any emotion of any sort. It occurs when the face is mostly relaxed and there is no contraction of the muscles. With other facial expressions occurs the opposite. They indicate a demonstration of emotion and, in turn, contraction of muscles.

Facial Expression Recognition (FER) is the process of identifying and interpreting human emotions from facial cues [2]. It involves analyzing the movements and positions of facial features such as the eyes, eyebrows, mouth, and cheeks to infer emotions such as happiness, sadness, anger, fear, surprise, and disgust.

FER can be performed autonomously using computer algorithms and machine learning techniques. It involves image capture, preprocessing, feature extraction, classification, and result interpretation. By analyzing key facial features, machine learning algorithms accurately classify expressions, providing a label or probability score as the output [6].

There are several reasons why the ability to detect neutral facial expressions can be important, such as in nonverbal communication, psychology research, marketing, security contexts and in the field of human-computer interaction.

Detecting a neutral facial expression with machine learning can be a challenging task. Facial expressions can be influenced by a variety of factors, such as the person's facial structure, the lighting conditions, and the presence of other facial features [4].

This paper is divided into the following sections: Section 2 presents the study conducted on multiple datasets, along with the applied data augmentation techniques; Section 3 presents the conducted study across several models, highlighting their impact; Section 4 provides an analysis of the impact of preprocessing techniques; Section 5 introduces the results of a real-time system implemented to evaluate the effectiveness of the developed work; Section 6 presents the conclusion and future work.

2 Datasets

For the investigation, the following datasets were selected: FER2013, CK+, and JAFFE.

The FER2013 dataset contains grayscale images with seven facial expression categories [3]. To classify emotions as Neutral or Not Neutral, the original categories were combined, resulting in 83% Neutral and 17% Not Neutral images. The CK+ dataset includes various expressions from 123 subjects [5], with 36% Not Neutral and 64% Neutral images. The JAFFE dataset consists of 213 images from 10 subjects [1], with 86% Not Neutral and 14% Neutral images.

The class imbalance between Neutral and Not Neutral images in the FER2013 dataset necessitated the use of data balancing and data augmentation techniques. Data augmentation was found to yield superior results when compared to data balancing. Similar findings were observed when applying these techniques to the CK+ dataset. The scarcity of images in the JAFFE dataset led to unsatisfactory results when trained individually.

To address the limited availability of images, all three datasets were integrated while maintaining a person-independent dataset. Offline and online data augmentation methodologies were explored, with offline augmentation proving effective for the Neutral class and online augmentation for both classes.

In this study, a sequential deep CNN architecture was adopted, shown in Figure 1, which is a popular neural network design for image classification applications. The selection of this model as a starting point was a deliberate choice based on the promising performance in previous studies. However, this model was chosen as a starting point. It is not considered a final solution. It serves as a foundation to study the impact of the datasets.

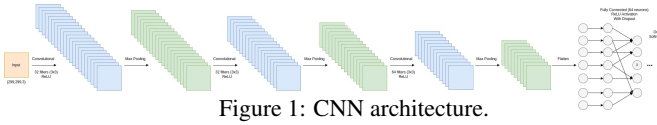

Figure 1: CNN architecture.

The best results were achieved by utilizing the merged dataset with online data augmentation for both classes using class weights. This approach offered benefits such as increased data diversity, balanced augmentation, adaptability to real-world scenarios, and avoidance of biases towards the majority class. These factors contribute to improved model performance, better generalization, and equitable treatment of all classes during training.

3 Models

The selection of an appropriate model is crucial in the field of machine learning to obtain accurate and reliable results. Various models were explored in this research, including InceptionV3, VGG16, ResNet50, and MobileNet.

InceptionV3, a pre-trained deep convolutional neural network, was utilized by removing the top layer responsible for final classification. Additional layers (Global Average Pooling, Dense with ReLU activation, Dropout, and Softmax Output) were added. The model was trained using Adam optimizer, categorical cross-entropy loss, and early stopping. VGG16 and ResNet50 followed a similar approach, achieving overall accuracies of 0.79 and 0.82 respectively on the test set. MobileNet, another

pre-trained network, achieved an overall accuracy of 0.82 after similar modifications.

These models demonstrated varying levels of performance in accurately classifying emotions. While InceptionV3 using early stopping, freezing all layers and incorporating top layers, showed moderate performance, VGG16, ResNet50, and MobileNet achieved higher accuracies, but InceptionV3 exhibited more reliable and favourable outcomes.

4 Preprocessing

Preprocessing is necessary for data preparation for analysis and machine learning activities. It was divided into five phases, shown in Figure 3.

Figure 2: Preprocessing phases.

The first step is to analyze the impact of each preprocessing phase, on the CNN Model.

The CNN model was trained using the selected dataset. Figure 3 illustrates the accuracy and loss trends throughout the epochs for each phase, providing a comprehensive visualization of their respective performances.

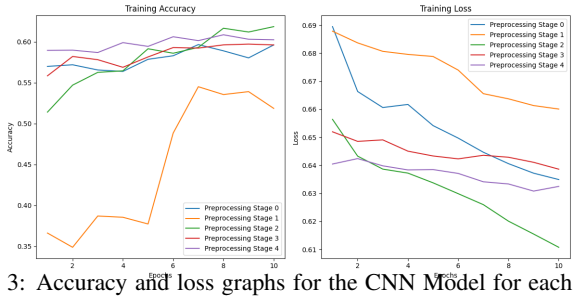

Figure 3: Accuracy and loss graphs for the CNN Model for each of the phases.

The second step is to analyze the impact of each preprocessing phase, on the InceptionV3 Model.

The InceptionV3 model was trained using the selected dataset, and Figure 4 shows the accuracy and loss tendencies throughout the epochs for each phase, providing a visualization of their individual performances. A distinct pattern appears after analyzing the accuracy and loss graphs. During Phase 0, the precision is at its maximum, while the loss is at its lowest. Surprisingly, despite using numerous preprocessing steps, they all resulted in no improvement in the findings. The accuracy stays stable, and the loss does not drop beyond the initial values obtained in Phase 0.

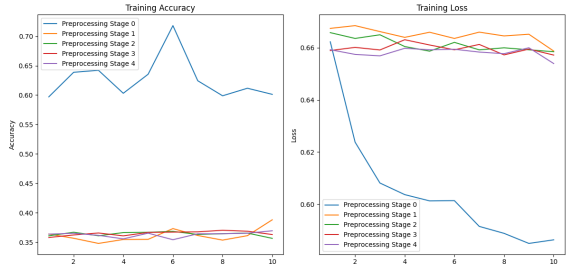

Figure 4: Accuracy and loss graphs for the InceptionV3 Model for each of the phases.

The preprocessing techniques for the simple CNN model enhanced the model's ability to extract significant features from the dataset. This could have resulted in improved accuracy and reduced losses. The simplicity of the model allows for easier manipulation and adaptation to the preprocessing steps, leading to noticeable improvements in performance. On the other hand, the InceptionV3 model has a more complex architecture that incorporates sophisticated convolutional operations and intricate feature extraction mechanisms. The preprocessing techniques may not have significantly impacted the model's performance due to the model's inherent ability to handle a wide range of input variations and extract complex features.

5 Results

An application was developed to evaluate the model. During the data collection process, participants were instructed to sit in front of the camera and begin by displaying a neutral expression, followed by various facial expressions. Of the 32 participants analyzed, 24 had their neutral expressions correctly captured and included in the dataset, accounting for approximately 75% of the participants. However, approximately 16% (five individuals) did not have their neutral expressions detected during the recording. In every frame captured for these individuals, their expressions were classified as Not Neutral. Notably, two of these participants were wearing glasses, which might have affected the accurate detection of neutral expressions. Furthermore, three participants (approximately 9% of the total) had their saved images initially classified as neutral but later identified as Not Neutral, specifically a smiling expression. This occurred because the model assigned a lower confidence value to the neutral classification, indicating higher uncertainty in those specific instances.

The confidence values associated with the frames saved vary from 51% to 79%. These scores belong to participants 4 and 23, respectively, shown in Figure 5.

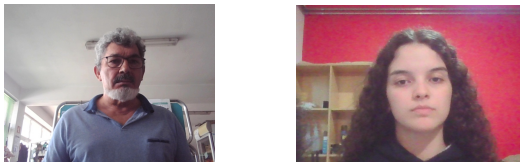

Figure 5: Neutral expressions saved with the lowest and highest confidence values.

6 Conclusion

In conclusion, the study comprehensively investigated neutral facial expression recognition, exploring various datasets, data augmentation techniques, model architectures, and preprocessing methods. The results demonstrated the significance of dataset selection, with data augmentation proving an effective technique for addressing class imbalances. The choice of model architecture, such as InceptionV3, also played a crucial role in achieving high performance. The findings highlight the success and challenges in accurately detecting and capturing neutral expressions, emphasizing the need for further refinement in dataset construction and model training. By incorporating suggested improvements and addressing the identified limitations, neutral facial expression recognition accuracy can be improved.

References

[1] Miyuki Kamachi, Michael Lyons, and Jiro Gyoba. The japanese female facial expression (jaffe) database. *Availble: http://www. kasrl. org/jaffe. html*, 01 1997.

[2] Mohan Karnati, Ayan Seal, Ondrej Krejcar, and Anis Yazidi. Fernet: facial expression recognition using deep neural net. *Neural Computing and Applications*, 33, 08 2021. doi: 10.1007/s00521-020-05676-y.

[3] Yousif Khaireddin and Zhuofa Chen. Facial emotion recognition: State of the art performance on fer2013. *arXiv preprint arXiv:2105.03588*, 2021.

[4] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(3):1195–1215, 2022. doi: 10.1109/TAFFC.2020.2981446.

[5] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101, 2010. doi: 10.1109/CVPRW.2010.5543262.

[6] Najmeh Samadiani, Guangyan Huang, Borui Cai, Wei Luo, Chi-Hung Chi, Yong Xiang, and Jing He. A review on automatic facial expression recognition systems assisted by multimodal sensor data. *Sensors*, 19(8):1863, 2019.