National University of Science and Technology POLITEHNICA Bucharest

Doctoral School of Electronics, Telecommunications and Information Technology

Decision No. 203 from 21-09-2024

Ph.D. THESIS SUMMARY

Ing. Radu-Daniel BOLCAȘ

Contributions to Emotion Recognition Using Artificial Intelligence

---

THESIS COMMITTEE

Prof.dr.ing. Ion MARGHESCU NUSTPB	President
Prof. Dr. Ing. Mihai CIUC NUSTPB	PhD Supervisor
Prof.dr.ing. Dan-Marius DOBREA Gheorghe Asachi Technical University of Iaşi	Referee
Prof. Dr. ing. Laurenţiu-Mihail IVANOVICI Transilvania University of Braşov	Referee
Conf.dr.ing. Eduard POPOVICI NUSTPB	Referee

BUCHAREST 2024

Content

Chapter 1

Introduction

1.1 Presentation of the field of the doctoral thesis

The PhD thesis addresses a subject in the fields of psychology and artificial intelligence: automatic emotion recognition from both facial expressions observed in images and text.

Emotion recognition using Artificial Intelligence (AI) has become a major area of interest in recent years due to advances in image processing, machine learning, and deep learning techniques. These technologies enable the identification and classification of human emotions through the analysis of facial expressions, voice, gestures, and even written text. Essentially, emotion recognition refers to the process of detecting and interpreting emotional signals expressed by a person, whether visual, auditory, or linguistic, with applications in diverse fields such as healthcare, education, customer service, and security.

An important aspect of emotion recognition is the analysis of facial expressions, where CNN technologies are often used to identify specific facial features that reflect emotional states. Similarly, emotion recognition from voice involves spectrogram analysis and the use of RNN or Long Short-Term Memory (LSTM) networks to capture the temporal dynamics of speech, applicable in contexts such as call centers or virtual assistants. Additionally, emotion recognition from text is facilitated by NLP combined with deep learning models, used for sentiment analysis on social media or for evaluating customer feedback.

1.2 Scope of the doctoral thesis

This paper presents research on facial emotion recognition and how it can be applied in practice. The need for human communication, knowledge, and understanding directs this study towards multiple applications in fields such as medicine, security, psychology, and education.

Multimodal models can be applied across various domains, including medical diagnostics and psychology. For example, analyzing a person's written responses along with their facial expressions can provide valuable insight into their mental state.

Emotion recognition in education can offer numerous benefits, helping to improve the learning experience and supporting students' emotional development. Additionally, early identification of stress or anxiety can prevent potential worsening of emotional issues.

1.3 Content of the doctoral thesis

This doctoral thesis explores the automatic detection of emotions from images (containing facial expressions) and text. The thesis is organized into seven distinct chapters and proposes an interdisciplinary approach combining psychology with artificial intelligence technologies.

Chapter 2 provides a detailed overview of the field of psychology and the methods for emotion recognition, both through the analysis of facial expressions in images and the classification of emotions from text. This chapter also introduces the field of artificial intelligence, presenting essential algorithms and methods in machine learning and deep learning that will be used in subsequent chapters.

Chapter 3 presents the current state of facial emotion recognition in the context of machine learning, analyzing methods and techniques used in the literature and highlighting both the advantages and challenges associated with them.

Chapter 4 explores the use of image filters to accelerate image processing tasks. By applying filters, the aim is to speed up image processing while maintaining model accuracy. This chapter investigates various filtering techniques and considers their effects on processing speed.

Chapter 5 investigates the use of ChatGPT (Generative Pre-trained Transformer) in the field of facial emotion recognition, making development processes more efficient, easier, and faster. With ChatGPT, coding and debugging are achieved quickly, leading to the rapid creation of high-performance models in just a few minutes. This chapter also demonstrates the importance of word choice and developer skills in achieving the right balance between speed and accuracy.

Chapter 6 explores the development of an advanced sentiment recognition model using multimodal learning, integrating text and image data. The proposed multimodal model and dataset aim to offer a new perspective on the simultaneous classification of text and image data. The study highlights the benefits of multimodal learning in handling ambiguous information and improving classification tasks.

In the final Chapter 7, the thesis conclusions are formulated, summarizing the obtained results and original contributions. Additionally, the implications of these results are analyzed, and possible directions for future research are suggested.

Chapter 2

General Overview

2.1 Emotion Recognition

Interpersonal communication has always been essential for humans, yet human abilities to interpret others' moods can lead to errors. Emotions are mental processes triggered by internal or external stimuli. Although people can learn to mask their gestures, facial expressions can reveal genuine emotions, and technology can aid in detecting these [1]. The literature defines a limited number of primary emotions, universal regardless of race, society, or culture [2].

2.2 Artificial Intelligence, Machine Learning and Deep Learning

2.2.1 Artificial Intelligence

Artificial Intelligence (AI) is a field of computer science focused on developing machines and software capable of performing tasks that require human intelligence. There are two main categories of AI: weak AI and strong AI.

2.2.2 Machine Learning

Machine Learning (ML) is a subfield of AI where computers can learn from data and generalize to unknown cases. There are three main types of machine learning: supervised learning, unsupervised learning, and reinforcement learning.

2.2.3 Deep Learning

Deep Learning (DL) is a subfield of ML that uses complex artificial neural networks to learn and make predictions from data. This process enables the model to learn abstract representations. Techniques used include Convolutional Neural

Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer networks (BERT [3], GPT [4]).

2.3 Artificial Neural Networks

In DL, the model utilizes artificial neural networks with multiple layers. These networks adjust their parameters through a feedback mechanism, using a cost function to assess the accuracy of the results obtained. Activation functions, like ReLU, transform data from one layer to pass it to the next, while optimizers help minimize error functions and adjust the model's weights.

2.4 Convolutional Neural Networks (CNN)

CNNs are used for processing images, specifically, images containing facial expressions in this context. Convolutional layers are utilized to isolate features such as eyebrows, nose, and mouth from other irrelevant elements. This process enables the system to classify facial expressions efficiently.

2.4.1 Convolutional layers

Convolutional layers process images by recognizing and extracting relevant features. These layers use kernels that apply a dot product to the image, and the resulting matrix is sent to the next layer of neurons. Neurons receive information from the previous layer, apply a mathematical function, and pass the result onward.

2.4.2 Pooling layer

The pooling layer is typically used after a convolutional layer and aims to reduce the data while also extracting the most important features by using a sliding window that calculates the maximum or average of the values within the window with each slide.

2.5 Hardware and Software Description

The implementation was carried out on a Linux-based architecture using Ubuntu 22.04 LTS and the TensorFlow and Jupyter platforms. The models were trained using an NVIDIA GeForce GTX 1070 graphics card.

Chapter 3

Facial Emotions Recognition in Machine Learning [1]

This chapter is adapted from the author's article titled "Facial Emotions Recognition in Machine Learning" [1].

This chapter provides a description of the psychological aspects of Facial Emotion Recognition (FER) and offers an overview of the datasets and algorithms that enable neural networks. It also includes a literature review of recent studies in FER, detailing the methods and algorithms used to enhance the capabilities of systems that employ machine learning. Challenges related to machine learning, such as overfitting, potential causes, solutions, and issues concerning datasets, including discrepancies like head orientation, lighting, and class imbalance within the dataset, are discussed.

3.1 Introduction

The psychological community defines a small number of basic emotions that are expressed similarly, regardless of race, gender, origin, society, or geographic region from which they come [2]. The basic emotions include anger, happiness, disgust, surprise, sadness, contempt, and fear.

In addition to these basic emotions, another model used to identify emotions is the Facial Action Coding System (FACS). Developed by Ekman and Friesen, it consists of coefficients called Action Units (AUs), which can define most possible facial expressions [5]. By combining different AU values, it is possible to determine the emotion at a specific moment.

3.2 Algorithms and databases

Different methodologies have been used by researchers, ranging from classical machine learning to deep learning. Classical machine learning relies on the initial steps of processing data in a way that extracts relevant features from the image and then provides these features as input to a classifier.

Deep learning is a subset of machine learning designed to minimize human intervention. For this reason, it tends to be more complex and requires more hardware resources than the classical approach, while providing better results.

Convolutional Neural Networks (CNNs) are a popular choice of deep artificial neural networks capable of identifying patterns in input data. Their most common application is in image analysis and classification, but they can also classify videos or texts. They can work with 1D or 3D data types, depending on whether images are grayscale or color. They are designed based on the human brain's model and attempt to mimic the same learning process.

Databases are crucial when discussing Facial Emotion Recognition (FER). They represent a key factor that determines the accuracy of recognition. Databases contain varying numbers of images or videos and have specific characteristics [6].

Some of these characteristics include the tilt and orientation of the face, lighting, the number of actors performing, the context in which the emotions were recorded (e.g., laboratory conditions), accessories or facial hair, and, of course, the number of emotions.

The FER2013 database [7] is a large-scale collection of images obtained using the Google image search API. The images were processed to 48x48 pixels and resized. It contains 35,887 grayscale images with six basic emotions and a neutral state.

3.3 Literature review

Researchers are working to enhance the capabilities of FER systems either by increasing accuracy or by reducing training time or the processing power required for training, finding ways to utilize data more efficiently. While these are not the only avenues, they are the most researched.

In a macro-level approach to facial emotion detection, Rzayeva and Alasgarov [8] utilized the CK+ and RAVDESS databases. After collecting and preprocessing the data, a CNN was employed. In 50 epochs, the accuracy reached 80%. Their second approach used VGG16 [9], resulting in an accuracy of 82%, but also significant overfitting. To address this issue, dropout layers were introduced. The proposed model achieved an accuracy of 88% for CK+, 92% for RAVDESS, and 92% for both CK+ and RAVDESS databases.

In a study conducted by Tarnowski et al. [10], Action Units (AUs) were utilized to determine the characteristics of facial expressions using a three-dimensional facial model. The methodologies employed were MLP with the k-NN classifier. Microsoft Kinect was used to model a 3D face and to map points on the face to the edges of facial features. The points were used in conjunction with the Facial Action Coding System (FACS) [5] to determine facial characteristics (corner of the mouth, eyebrows, nose, cheekbones). The final accuracy was 96% for 3-KNN and 90% for MLP.

In a study by Yang et al. [11], a new approach for FER was proposed using an expressive component and a neutral component. By researching the psychological literature, a facial expression was decomposed into an expressive and a neutral component. The network receives two images of the subject in a neutral state and one image with a specific emotion as input. The model creates a second image of the same person in a neutral state. The proposed CNN was pre-trained on the Binghamton University 3D Facial Expression (BU-3DFE) and BP4D-Spontaneous datasets, achieving an accuracy of 75.23% on the MMI database, 88% on the Oulu-CASIA set, and 97.30% on the CK+ database.

3.4 Discussion

Among the observed challenges are class imbalance, which can create bias in predictions, actors trained to express a specific emotion that may deviate from a natural expression, as well as expressions with low intensity. Additionally, a small number of subjects in the dataset can lead to overfitting if the images are too similar. Further difficulties arise in recognizing tilted or rotated faces, as well as images with poor lighting, shadows, and contrast. Subjects wearing glasses, having facial hair, or wearing clothing that obscures important facial features can also present additional obstacles. It is also important to avoid both overfitting and underfitting, as these affect the network's efficiency in determining an optimal result. Furthermore, the time required to train a CNN is generally high due to the considerable number of layers and neurons, making the processing intensive.

3.5 Conclusions

Studies show that models increase in accuracy without necessarily becoming more complex. For simpler models, approaches measuring the distance to facial features or mapping a 3D view of the face combined with FACS have been used. This facial mapping has managed to overcome the simplicity of the model while maintaining high accuracy.

An interesting approach in FER was also achieved through the use of an expressive component and a neutral component [11].

This method yielded good results, although the model is complex and may require an even larger dataset for further analysis.

This chapter has evaluated the challenges in developing a FER system to provide insights into potential research directions that can be considered.

In the article by Diana Dranga and **Radu-Daniel Bolcaș** [12], a possible implementation of a text-based convolutional neural network model was analyzed to facilitate the debugging of digital circuits in the field of functional verification. This

article reviews how Artificial Intelligence can alleviate this bottleneck, considering the time spent implementing the verification environment and the time required to achieve the desired coverage percentage.

The achievements, challenges, and machine learning techniques such as CNN, RNN, etc., for both software and hardware domains were highlighted.

A potential research direction involves analyzing the objectivity of requirements, where various strategies can be adopted. Typically, technical documents are written in a neutral language regarding emotions.

In the testing domain, the later a problem is discovered, the more costly it becomes, as once the necessary changes are made, all steps must be repeated involving many engineers, resulting in costs and time loss. This analysis can create a new prioritization of requirements in testing to have an initial validation followed by comprehensive validation. In this approach, potential major issues can be detected as early as possible in the development and verification process. This research direction is promising and will be explored in future papers and articles.

Chapter 4

Enhancing Training Efficiency in Facial Emotion Recognition [13]

This chapter is adapted from the author's article titled „Enhancing Training Efficiency in Facial Emotion Recognition” [13].

This chapter explores the use of image filters to accelerate image processing tasks. By applying filters, the goal is to speed up image processing while maintaining the model's accuracy.

4.1 Introduction

Convolutional Neural Networks (CNNs) have demonstrated the potential to deliver good results in FER. While performance is excellent, the training process is time-consuming.

One solution to improve both the accuracy and training time of the model is to filter information from the dataset before inputting it. One idea is to identify facial features from an image and then use this image to initiate training. In this way, all irrelevant features are filtered out.

This chapter presents an approach to reduce the amount of irrelevant features while simultaneously speeding up the training process. To enhance performance and seek to improve accuracy for a specific model (developed by Y. Khaireddin and Z. Chen [14]), an additional preprocessing step consisting of filters was added to enhance the data sent to the model. Furthermore, using this new approach significantly reduced the time required to train a model.

4.2 Implementation

4.2.1 Overview

The use of image filters has become a powerful technique, transforming the way CNNs perceive and manipulate visual content. In the field of computer vision,

the application of image filters has emerged as a robust tool capable of seamlessly removing unwanted information from images. The presence of undesirable elements, such as noise, objects, and backgrounds, can significantly compromise the quality of features that a CNN can extract from an image.

4.2.2 Preprocessing methods

The preprocessing techniques applied involve the use of four filters: a simple threshold filter, two adaptive threshold filters, and Otsu's method. These filters are included in the OpenCV library [15], which is widely recognized software used for image and video processing, encompassing a multitude of functionalities.

The first filter is a simple threshold filter. A uniform threshold value is used for each pixel. Pixels with values below the threshold are set to 0, while those exceeding it are set to a predefined maximum value [16].

The second and third filters belong to the category of adaptive threshold filters. The algorithm used by these filters calculates the threshold based on the local area surrounding each pixel. The second filter computes the threshold value by taking the arithmetic mean of the neighboring area and subtracting a constant C [15,16]. The third filter utilizes a "weighted Gaussian sum of the neighboring values minus the constant C" [15].

The fourth filter involves an automatic threshold determination using the image itself. This method, known as Otsu's method, utilizes the image histogram to identify peaks in the graph and selects a value located between these peaks [16]. The advantage in this case is the minimal need for parameter tuning in the filtering process.

4.2.3 CNN architecture and data preprocessing

In a 2021 study conducted by Y. Khaireddin and Z. Chen, the FER2013 dataset was utilized to achieve high performance with a medium-sized network. They employed a convolutional neural network model and fine-tuned its hyperparameters [14].

4.3 Dataset

FER2013 [7] consists of 35,888 images representing 7 different emotions. These images are divided into 3 categories: training, validation, and testing. This division was established at the time of its publication in ICML. In this study, the division for training, validation, and testing was done differently, adding more images to the training data, thus improving the training time. For this reason, the accuracy displayed may vary compared to the literature.

4.4 Results and analysis

By adaptating the FER2013 dataset, a result of 65.0124% accuracy was achieved. The time required for one epoch is 194.0375 seconds. This value will serve as a reference point for comparison with the filters used. The usual values obtained in the literature vary around 67% ± 5%. This decrease is due to the way the data is structured.

The application of the simple Threshold filter resulted in an accuracy of 57.0452%, obtained after running 30 epochs. The time required for one epoch was 128.9716 seconds.

The second filter introduced is the Adaptive Thresh Mean filter. This filter led to an accuracy of 56.6406% after 30 epochs. The time required for one epoch was 129.1821 seconds.

The third filter used is the Adaptive Thresh Gaussian filter. Through the tests conducted, the highest accuracy achieved for the validation data was 62.9883%, coming very close to the reference value. This result is visually illustrated in Figure 4.7. The time elapsed for one epoch is 128.7093 seconds.

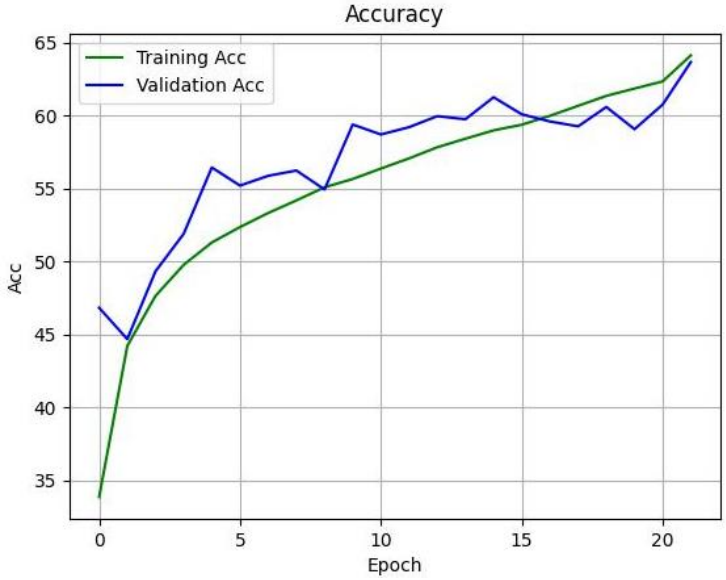

Figure 4.7 *Use of Adaptive Thresh Gaussian Filter*

The visual transformation of an image using the Gaussian threshold is observable in Figure 4.8.

Figure 4.8 *Applying the Adaptive Thresh Gaussian Filer to a photo*

The last filter used in the experiments was based on Otsu's method. By automatically determining the threshold, this filter is capable of producing an accuracy of 53.6830% with a training time per epoch of 128.9333 seconds.

The most significant effect of the filters is observed through the quantification of the time required for the standard training of the model and comparing it with the use of filters. When comparing the unfiltered approach with the adaptive Gaussian filter, a notable reduction of 33.6678% in the duration of training each epoch is evident. Therefore, training using filters can significantly improve the time required for training a model. Particularly for complex models, this approach can serve as a powerful tool for accelerating the training process.

4.5 Conclusions

By carefully selecting and applying filters, the image processing time can be significantly reduced. The experiments conducted demonstrated the effectiveness of various filtering techniques in improving processing speed by approximately 33% without compromising accuracy (with a variation of about 2%).

The results highlight the importance of filter selection and parameter optimization to achieve the desired trade-offs between processing speed and accuracy. The insights gained from this study contribute to a deeper understanding of how image filters can be leveraged to process images more quickly and efficiently.

Chapter 5

Generating FER models using ChatGPT [17]

This chapter is adapted from the author's article titled „Generating FER models using ChatGPT” [17].

This chapter explores the use of ChatGPT in the field of facial emotion recognition, streamlining development processes, making them easier and faster. With ChatGPT, code writing and debugging are done quickly, leading to the rapid creation of high-performance models in just a few minutes. Additionally, the choice of words and the developer's skill in finding a balance between speed and accuracy are also essential.

5.1 Introducere

In the field of Facial Emotion Recognition (FER), supervised learning is predominantly used, while exploration of unsupervised learning has been limited. Developing models involves several stages: selecting an appropriate dataset, preprocessing for standardization and data preparation, choosing the model architecture, training the model, tuning hyperparameters, and finally, evaluating the model.

The proposed approach utilizes OpenAI's ChatGPT [4] to generate the initial code, allowing researchers to make further adjustments. Using ChatGPT as a support tool significantly reduces the time required for development and initial validation. After completing the process and obtaining results, it is evaluated whether the model serves as a solid starting point or if another architecture is necessary.

5.2 General Overview

5.2.1 Large Language Models (LLMs)

Large language models (LLMs) are artificial intelligence systems trained on vast amounts of text data to understand and generate human language. Examples like ChatGPT, developed by OpenAI, can process and generate text that closely mimics

human language. These models understand context, semantics, and syntax, with evaluation focused on their ability to produce coherent and relevant text in specific contexts.

5.2.2 FER architecture and databases

The database used in this study, FER2013 [7], is a widely-used dataset within the research community. In facial emotion recognition, unsupervised learning approaches often involve grouping similar facial expressions or extracting relevant facial features from images, which can be beneficial when labeled data is scarce.

5.3 Implementation and results

ChatGPT was used to help establish the baseline model and for further investigations. The way questions are formulated can influence the responses obtained. Therefore, a zero-shot approach was adopted to generate the basic code. In this way, ChatGPT was prompted to generate models using both supervised and unsupervised learning approaches.

5.3.1 ChatGPT with unsupervised learning

The initial question posed to ChatGPT was to "Create a FER model using unsupervised learning with the FER2013 dataset." It generated a model that integrates StandardScaler for data standardization, PCA for dimensionality reduction and identifying primary and secondary components, and K-Means as the clustering algorithm. The resulting clusters are illustrated in Figure 5.1.

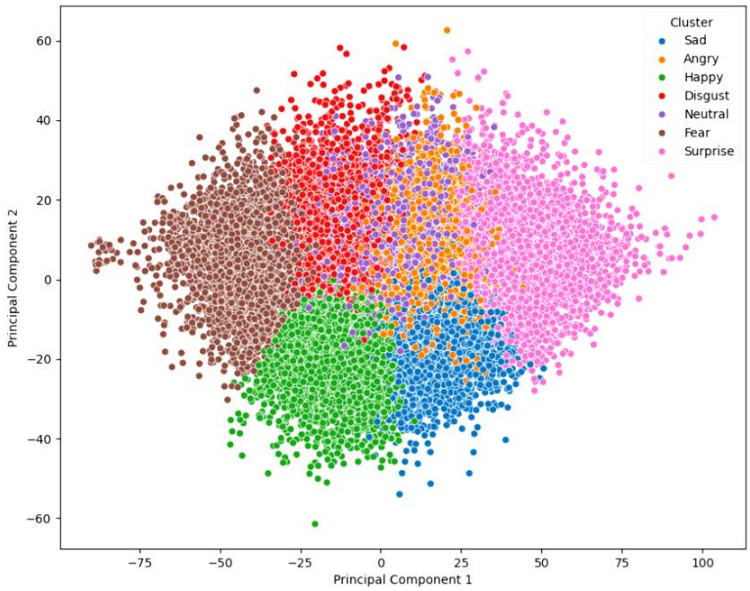

Figure 5.1 *K-Means Cluster for FER2013*

A second question posed to ChatGPT was to use a different algorithm that would offer a similar model, replacing K-means with Gaussian Mixture Models (GMM) for label clustering. This algorithm resulted in substandard performance. After multiple interactions with ChatGPT, errors became more pronounced, and the need for regenerating code increased. As the generated models began to show even weaker results and errors became more frequent, it was decided to conclude this research direction and transition to a supervised approach.

5.3.1 ChatGPT with supervised learning

The initial question asked was: "Create a machine learning model for emotion recognition in images using the FER2013 dataset." Although the design was simple, ChatGPT-3.5 provided all the necessary code to execute the model according to the specifications. However, the metrics obtained were suboptimal, with an accuracy of 51.14%.

Multiple subsequent messages largely yielded theoretical information that, while accurate in a general sense, did not directly apply to the specific scenario. The models proposed by ChatGPT required the author's intervention to make them functional. Due to the large volume of interactions, earlier initial details became less relevant. Thus, each following message contained comprehensive information for each question, proving to be more efficient.

With this approach and multiple regenerations, a distinct model with residual blocks emerged. After training for 30 epochs, it achieved an accuracy of 62.49%.

Using the same approach, a new model with three blocks was trained for 20 epochs, resulting in an accuracy of 60.08%. Although this accuracy is lower than that of the model with residual blocks, this model avoids issues of overfitting or underfitting, as shown in Figure 5.6 and Figure 5.7.

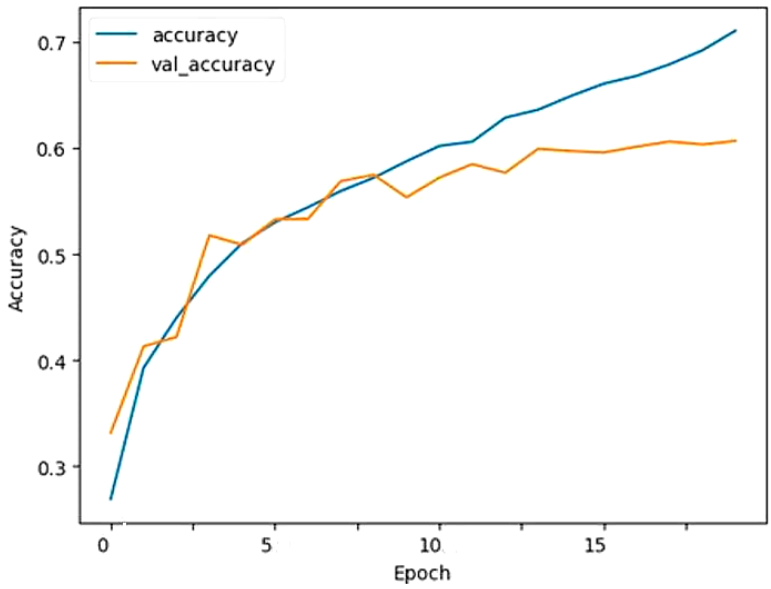

Figure 5.6 *Three Blocks Model accuracy with 20 epochs*

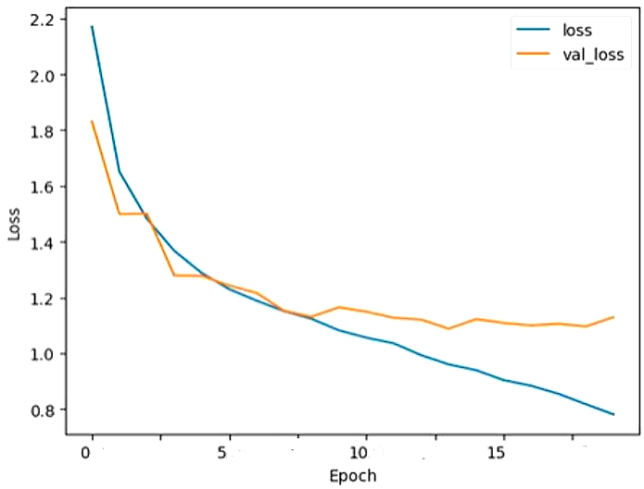

Figura 5.7 *Three Blocks Model loss with 20 epochs*

5.4 Discussions

The process of working with ChatGPT to generate models highlighted both successes and challenges, especially in the case of unsupervised learning, which showed low accuracy. Various strategies, including modifying architectures and data preprocessing, led to variable results, and in some cases, required author intervention to make them functional. Nevertheless, repeated interactions with ChatGPT accelerated development, although the quality of results depended on the clarity of prompts.

The obtained results, with accuracy ranging from 60.08% to 62.49% in facial emotion recognition, demonstrate rapid model development efficiency, although slightly below state-of-the-art models (73.28%-75.97%). The main advantage lies in the ability to iterate quickly and experiment with various configurations, making this approach ideal in fast-paced environments. ChatGPT's use has showcased the potential of large language models (LLMs) to boost efficiency and agility in developing AI applications for facial emotion recognition.

5.5 Conclusions

The study explores the use of ChatGPT to accelerate model development, significantly reducing time spent on coding and debugging. While results in the unsupervised setting were weaker, the supervised environment achieved an accuracy of 60.08%, with the architecture validated in just a few minutes. This provides researchers with a solid foundation for further improvements. The results highlight the importance of word selection and developer expertise in balancing development speed with accuracy, showcasing the potential of LLMs to streamline processes and become a valuable tool for developers.

Chapter 6

Ensemble models for multimodal sentiment analysis using textual and image fusion [18]

This chapter is adapted from the author's article titled „Ensemble models for multimodal sentiment analysis using textual and image fusion” [18].

Sentiment analysis is an evolving field that attracts significant research interest. Multimodal sentiment analysis (MSA) integrates different forms of data, such as text for emotion recognition and images for facial emotion recognition (FER), processing various modalities of input. This paper introduces ImaText, a new dataset for emotion recognition that combines texts and images from DailyDialog and FER2013. The proposed multimodal model and dataset provide a fresh perspective on the simultaneous classification of text and image data.

6.1 Introduction

In this chapter, the two main forms of emotion recognition will be facial emotion recognition (FER) and emotion recognition from text. Multimodal sentiment analysis (MSA) involves integrating different forms of data, including images, text, audio, or video, to process multiple input or output modalities. By integrating various modalities, the capabilities of the model can be significantly enhanced.

In the context of multimodal learning, the existing literature typically outlines two levels of fusion: feature-level fusion, often referred to as early fusion, and decision-level fusion, also known as late fusion.

6.2 Datasets and data preprocessing

This new study proposes a fusion between two datasets, one containing text labeled with emotions (DailyDialog [19]) and the other containing images labeled with facial emotions (FER2013). By correlating the emotions from both datasets, the result consists of a CSV file and a directory structure where the images are located.

This new dataset, named **ImaText**, includes 25,780 entries for six emotions and has been saved for multimodal learning. The final distribution is as follows: 4,865 for "anger," 555 for "disgust," 1,255 for "fear," 8,910 for "happiness," 6,190 for "sadness," and 4,005 for "surprise."

6.3 Architectures and proposed model

The chosen architecture is a convolutional neural network (CNN) optimized for analyzing both text and images. For the text input, a tokenizer maps words to indices, followed by padding to standardize the text length. The data is split into images, texts, and labels, and then further divided into training and testing sets. The text model uses input layers, embedding layers, one-dimensional convolution, activation functions (ReLU), dropout to prevent overfitting, max pooling, and a final fully connected layer that classifies the data into six distinct classes.

The image model begins with an input layer similar to that of the text model. This is followed by a block consisting of a two-dimensional convolutional layer and a max pooling layer, which extract and compress the features of the image. This block is repeated three times. A flatten layer restructures the data for the fully connected layer, which is responsible for classifying the emotions.

To develop a multimodal model, the outputs from the text and image models are combined through a "concatenate" layer. After concatenation, two fully connected layers are added, with the final layer classifying the data into six emotion classes.

6.4 Results and Analysis

The proposed multimodal model performed well on the created dataset. Various experiments were conducted during which the model underwent adaptations and improvements, including testing different optimizers and adjusting hyperparameters. The multimodal model was trained and achieved a validation accuracy of 70.19%, as illustrated in Figure 6.4, while the loss graph is presented in Figure 6.5.

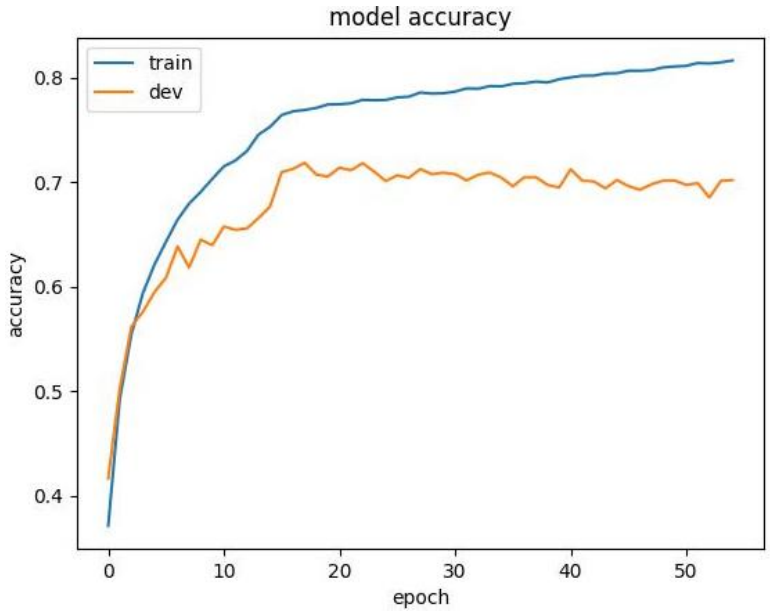

Figure 6.4 *Accuracy of multimodal model*

The model trained with both image and text data showed slight signs of overfitting starting from 15 epochs, becoming pronounced from 25 epochs onward. Optimal performance was observed around the 25th epoch.

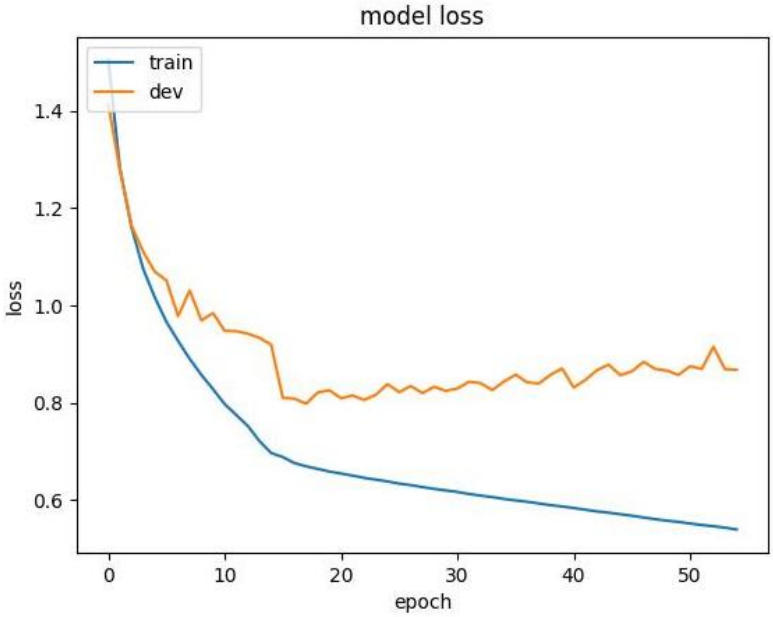

Figure 6.5 *Loss of multimodal model*

The model faced challenges in recognizing certain emotions due to class imbalance, as some emotions represented less than 5% of the data, leading the model to overlook them. In the future, techniques to mitigate this imbalance, such as data augmentation, could be employed. Existing classification models on the DailyDialog dataset achieved around 59% accuracy [20], while on FER2013, the accuracy is about 70% ± 5% [14,21]. The proposed model achieved 70.19% on the new dataset, highlighting the potential of the multimodal approach, where the integration of information from text and images can enhance learning and performance.

6.5 Conclusions

This chapter presents the successful implementation of a new dataset, ImaText, created by combining DailyDialog and FER2013, along with a multimodal model that achieved an accuracy of 70.19%. ImaText is the first multimodal dataset specifically designed for sentiment analysis, which includes only text and images. It contains 25,780 entries distributed across six emotions. The author designed and adjusted the model to adapt to this innovative dataset.

The results are groundbreaking, opening new avenues for research in multimodal sentiment analysis, with applications in medical diagnosis and enhancing human-computer interaction. The study also proposes leveraging existing data to augment model performance, offering a solution to the challenge of the lack of multimodal datasets.

Chapter 7

Conclusions

The doctoral thesis investigates the automatic recognition of emotions using machine learning algorithms, focusing on facial expressions from images and expressive text. It utilizes datasets such as FER2013 and DailyDialog and introduces a proprietary dataset called ImaText, which combines images and text labeled with emotions. The study explores an advanced multimodal recognition model, which combines images and text, that achieves an accuracy of 70.19%. Additionally, it evaluates filtering techniques to optimize image processing speed and examines the use of ChatGPT as a tool to accelerate model development, highlighting the importance of word selection and developer expertise for achieving an optimal balance between speed and accuracy.

7.1 Results obtained

Chapter 3 describes the psychological aspects of facial emotion recognition (FER) and provides a description of the datasets and algorithms that enable neural networks. It then reviews the existing literature on recent studies in FER, discussing challenges related to machine learning and possible solutions.

Chapter 4 focuses on the integration of image filters to accelerate image processing. By carefully selecting and applying filters, the processing time of images can be significantly reduced. The experiments conducted demonstrated the effectiveness of various filtering techniques in improving processing speed by approximately 33% without compromising accuracy, achieving a variation of around 2% for the Adaptive Thresh Gaussian Filter.

Chapter 5 investigates the use of ChatGPT to accelerate the model development process, highlighting the significant reduction in time for development and debugging tasks. In the unsupervised environment, the generated models showed lower quality, however, in the supervised approach, an accuracy of 60.08% was achieved, with the architecture validated in a few minutes. This provides researchers with a solid starting point for further improvements. The chapter emphasizes the importance of word selection and the expertise of developers in balancing the speed of development with model accuracy.

In Chapter 6, the author proposes a study that implements a new dataset created by combining DailyDialog and FER2013, along with a multimodal model capable of achieving an accuracy of 70.19% on the unique dataset, named ImaText. The ImaText dataset consists of combining text with images and contains 25,780 entries distributed across six emotions. The final distribution includes 4,865 entries for "angry," 555 for "disgust," 1,255 for "fear," 8,910 for "happy," 6,190 for "sad," and 4,005 for "surprised." Using this new dataset, various experiments led to the achievement of an accuracy of 70.19%. To the author's best knowledge, ImaText is the first multimodal dataset created specifically for sentiment analysis that includes only text and images.

7.2 Original contributions

Among the original contributions of this work are:

1. The creation of analyses that describe the psychological aspects of FER and provide a description of the datasets and algorithms that enable neural networks. A study of the existing literature on facial emotion recognition is conducted, detailing the main aspects of the research to highlight the novelty, concepts, and related strategies that enable accurate recognition. Additionally, challenges related to machine learning and possible solutions are emphasized. These challenges provide insight into potential directions for developing better FER systems.

2. By carefully selecting and applying filters, the processing time of images can be significantly reduced. This accelerates the training process, while the experiments conducted have highlighted the effectiveness of various filtering techniques in improving processing speed by approximately 33% without compromising accuracy, achieving a variation of around 2% in accuracy for the Adaptive Thresh Gaussian Filter. This research underscores the potential benefits of integrating filters into image processing workflows.

3. The use of ChatGPT, which significantly reduces the time required for development and debugging tasks, especially aiding the early stages of development by providing suggestions for error resolution. In the unsupervised environment, it generated fewer models of lower quality with more errors. For the supervised approach, the experimental results underline its effectiveness in rapidly creating high-performance models with minimal time investment, achieving a solution with an accuracy of 60.08% and an architecture validated within minutes. Thus, the researcher has a solid starting point for further improving and adjusting the model for the necessary

application. The results demonstrate the importance of word selection and the expertise of developers in achieving a balance between development speed and accuracy.

4. The introduction of a multimodal dataset named ImaText. This dataset consists of the combination of text and images and contains 25,780 entries distributed across six emotions. The final distribution includes 4,865 entries for "angry," 555 for "disgust," 1,255 for "fear," 8,910 for "happy," 6,190 for "sad," and 4,005 for "surprised." To the author's knowledge, ImaText is the first multimodal dataset created specifically for sentiment analysis that includes only text and images.

5. Training multimodal networks to determine emotions from ImaText. Various experiments (using different layer configurations, optimizers, parameters, etc.) led to an accuracy of 70.19%.

7.3 List of original publications

This list includes only the published works and communications for which the doctoral candidate is the author or co-author. It also includes research reports from the doctoral program and contracts on which the doctoral candidate has worked. All these works can be found in the Bibliography. All mentioned works are related to the theme of the doctoral thesis.

1. **Radu-Daniel Bolcaș** and Diana Dranga, "*Facial Emotions Recognition in Machine Learning*," Electrotehnică, Electronică, Automatică (EEA) Journal, vol. 69, no. 4, pp. 87-94, DOI:10.46904/eea.21.69.4.1108010, 2021. **Article indexed in Scopus.**
2. Diana Dranga and **Radu-Daniel Bolcaș**, "*Artificial Intelligence Enhancements in the field of Functional Verification*," Electrotehnică, Electronică, Automatică (EEA) Journal, vol. 69, no. 4, pp. 87-94, DOI:10.46904/eea.21.69.4.1108011, 2021. **Article indexed in Scopus.**
3. **Radu-Daniel Bolcaș**, Mihai Ciuc, and Eduard Popovici, "*Enhancing Training Efficiency in Facial Emotion Recognition*," 2023 31st Telecommunications Forum (TELFOR), pp. 1-4, DOI: 10.1109/TELFOR59449.2023.103726002023, 2023. **Article indexed in IEEExplore**.
4. **Radu-Daniel Bolcaș,** "*Generating FER models using ChatGPT*," Romanian Journal of Information Technology and Automatic Control (RRIA), vol. 34, no. 2, pp. 85-96, 2024. **WOS:001253386000007**. **ISI journal article**.
5. **Radu-Daniel Bolcaș**, Mihai Ciuc, and Eduard-Cristian Popovici, "*Ensemble models for multimodal sentiment*," U.P.B. Sci. Bull., Series C, vol. 86, no. tbd, p. tbd, 2024. **ISI journal article** – **accepted for publication**

7.4 Perspectives for further developments

The research explores the impact of image filters on reducing training time, highlighting the importance of optimizing parameters to balance speed and accuracy. Advanced technology demands efficient image processing methods, and this research paves the way for the use of advanced filters and their optimization in real-world scenarios.

The use of ChatGPT and LLMs for rapid validation of promising architectures suggests potential for future investigations in other domains and datasets, although the generated models in the unsupervised environment were inferior. Nonetheless, the K-Means based model performed well and will be further investigated.

The presented multimodal model achieves an accuracy of 70.19% but faces challenges in recognizing certain emotions due to class imbalance. Data augmentation or integration of additional datasets is proposed to address these issues. The integration of filters into the ImaText model is also suggested to enhance performance and reduce training time.

Bibliography

[1] Radu-Daniel Bolcaș and Diana Dranga, "Facial Emotions Recognition in Machine Learning," *Electrotehnică, Electronică, Automatică (EEA) Journal*, vol. 69, no. 4, pp. 87-94, DOI:10.46904/eea.21.69.4.1108010, 2021.

[2] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of personality and social psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[3] C. Alberti, K. Lee, and M. Collins, "A bert baseline for the natural questions," *arXiv preprint arXiv:1901.08634*, 2019.

[4] OpenAI. (2024) ChatGPT, last acessed in April 2024. [Online]. https://openai.com/blog/chatgpt

[5] P. Ekman and W. V. Friesen, "Facial Action Coding System," *Consulting Psychologists Press, Stanford University, Palo Alto*, 1977.

[6] Wafa Mellouka and Wahida Handouzi, "Facial emotion recognition using deep learning: review and insights," *Procedia Computer Science*, vol. 175, pp. 689-694, 2020, DOI: 10.1016/j.procs.2020.07.101.

[7] I. J. Goodfellow et al., "Challenges in Representation Learning: A Report on Three Machine Learning Contests," *Neural Information Processing*, pp. p. 117-124, DOI: 10.1007/978-3-642-42051-1_16., 2013.

[8] Z. Rzayeva and E. Alasgarov, "Facial Emotion Recognition using Convolutional Neural Networks," *IEEE 13th International Conference on Application of Information and Communication Technologies (AICT)*, pp. 1-5, DOI: 10.1109/AICT47866.2019.8981757, 2019.

[9] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint, DOI: 10.48550/arXiv.1409.1556*, 2014.

[10] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak, "Emotion recognition using facial expressions," *International Conference on Computational Science*, 2017.

[11] H. Yang, U. Ciftci, and L. Yin, "Facial Expression Recognition by De-expression Residue Learning," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2168-2177, DOI: 10.1109/CVPR.2018.00231, 2018.

[12] Diana Dranga and Radu-Daniel Bolcaş, "Artificial Intelligence Enhancements in the field of Functional Verification," *Electrotehnică, Electronică, Automatică (EEA) Journal*, vol. 69, no. 4, pp. 87-94, DOI:10.46904/eea.21.69.4.1108011, 2021.

[13] Radu-Daniel Bolcaş, Mihai Ciuc, and Eduard Popovici, "Enhancing Training Efficiency in Facial Emotion Recognition," *2023 31st Telecommunications Forum (TELFOR)*, pp. 1-4, DOI: 10.1109/TELFOR59449.2023.103726002023, 2023.

[14] Yousif Khaireddin and Zhuofa Chen, "Facial Emotion Recognition: State of the Art Performance on FER2013," *arXiv preprint, DOI: 10.48550/ARXIV.2105.03588*, 2021. [Online]. https://arxiv.org/abs/2105.03588

[15] OpenCV. (2022). [Online]. https://docs.opencv.org/4.x/d7/d4d/tutorial_py_thresholding.html

[16] R. C. Gonzalez and R. E. Woods, *Digital image processing*. Upper Saddle River, N.J.: Prentice Hall, 2008.

[17] Radu-Daniel Bolcaș, "Generating FER models using ChatGPT," *Romanian Journal of Information Technology and Automatic Control (RRIA)*, vol. 34, no. 2, pp. 85-96, 2024.

[18] Radu-Daniel Bolcaș, Mihai Ciuc, and Eduard-Cristian Popovici, "Ensemble models for multimodal sentiment," *U.P.B. Sci. Bull., Series C*, vol. 86, no. tbd - 3, p. tbd, 2024.

[19] Yanran Li et al., "DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset," *Proceedings of The 8th International Joint Conference on Natural Language Processing (IJCNLP 2017)*, pp. 986-995, 2017.

[20] Shen Weizhou, Siyue Wu, Yunyi Yang, and Xiaojun Quan, "Directed Acyclic Graph Network for Conversational Emotion Recognition," *Annual Meeting of the Association for Computational Linguistics*, 2021.

[21] S. Vignesh, M. Savithadevi, M. Sridevi, and R. Sridhar, "A novel facial emotion recognition model using segmentation VGG-19 architecture," *International Journal of Information Technology*, vol. 15, no. DOI: 10.1007/s41870-023-01184-z, pp. 1777–1787, 2023.