**Human Emotion Recognition Model**

*Authors: Om Gaikwad, Ameya Deshmukh*

*University of Cincinnati*

**Abstract**

This research compares four well-known deep learning architectures for the goal of recognising human emotions: CNN, LSTM, VGG-Net, and ResNet50. Every model has particular advantages to offer: While LSTM records temporal dependencies, VGG-Net concentrates on complex visual patterns, ResNet50 tackles vanishing gradient problems, and CNN is the best at extracting spatial features. Performance indicators including accuracy, precision, recall, and F1 score are used in the study to assess these models using well-established emotion detection datasets. By means of this comparison methodology, the research seeks to illuminate the unique contributions of each building and offer insights into their efficacy in identifying a wide range of emotions.

**Introduction**

An essential component of artificial intelligence applications and human-computer interaction is the recognition of human emotions. Conventional methods for obtaining characteristics from facial expressions frequently make use of separate deep learning architectures. In order to compare the performance of four popular architectures—CNN, LSTM, VGG-Net, and ResNet50—and comprehend their respective advantages and disadvantages in the context of emotion recognition, this study adopts a novel strategy.

*Motivation:*

This comparison study was motivated by the realization that various deep learning architectures have distinct advantages when it comes to facial emotion processing. CNNs are good at extracting spatial features, LSTMs are good at capturing temporal dependencies, and designs such as VGG-Net and ResNet50 are good at recognizing complex patterns. We aim to compare these models in order to determine which architecture, or combination of architectures, is best suited for identifying a wide variety of emotions.

*Objectives:*

* This project's main goal is to provide a comprehensive comparative study of four different deep learning architectures in the field of human emotion recognition: CNN, LSTM, VGG-Net, and ResNet50. By putting each architecture through a rigorous review, we hope to determine how well it performs in relation to other well-known emotion identification datasets. Each architecture adds a distinct set of qualities to the task. This examination looks at how effectively each model recognizes a wide range of emotions and how well it can convey the subtleties of facial expressions. It is essential to comprehend the relative advantages and disadvantages of different architectures in order to make well-informed decisions when choosing models for particular applications.
* We use recognized performance criteria, such as accuracy, precision, training vs validation loss plots, CAM to measure how well each architecture performs in identifying different emotional states. We also imported random facial images to see whether our model is making the right predictions. These metrics offer a thorough assessment of the models' performance in accurately identifying emotions while reducing false positives and false negatives. For applications ranging from mental health assessments to human-computer interaction, accurate emotion recognition is essential to their success. We hope to have a comprehensive grasp of each model's performance by employing a range of measures, which will enable us to make insightful comparisons and deductions from the assessment procedure.
* An important part of this endeavour is figuring out each architecture's unique advantages and disadvantages in relation to emotion recognition. For example, CNNs are good at extracting spatial features, which enables them to capture face expressions using static patterns. LSTMs, on the other hand, are adept at identifying temporal relationships and offer insights into how emotions change dynamically over time. VGG-Net and ResNet50 each offer special skills in recognizing complex patterns and solving problems with vanishing gradients. Knowing each of these unique advantages and disadvantages is crucial for choosing the best model for a given set of circumstances and applications, assisting practitioners in making well-informed model selection decisions.
* This project's ultimate goal is to convert the knowledge obtained from the performance measurements and comparison analysis into useful applications. By considering elements like processing efficiency, scalability, and adaptability to various datasets, we aim to examine each architecture's practicality.

**Related Studies**

The application of distinct deep learning architectures for emotion recognition has been thoroughly studied in the past. Research utilizing VGG-Net has shown that it can capture complex patterns in face expressions as well as high-level features (Simonyan & Zisserman, 2014). ResNet50 has also been acknowledged for its ability to resolve vanishing gradient problems, which has resulted in enhanced performance in picture classification applications (He et al., 2016). Multiple-layer CNNs have shown promise in improving facial expression analysis through the extraction of spatial features (Lecun et al., 1998).

To comprehend how emotions change with time, it is imperative to consider the temporal dimension. As a potent tool for modelling sequential data, long short-term memory (LSTM) networks are ideal for capturing the temporal dependencies found in facial expressions (Hochreiter & Schmidhuber, 1997). Promising outcomes have been observed when LSTM is incorporated into emotion recognition models, as it can effectively capture the dynamic nature of emotional states (Fan et al., 2018).

Although successful individual architectures have been shown, subsequent research has investigated integrating numerous models to take advantage of their combined advantages. To recognize facial emotions, Liu et al. (2018) integrated VGG-Net and LSTM, highlighting the significance of both spatial and temporal variables. In a subsequent study, Wang et al. (2020) combined CNNs with ResNet50 and demonstrated enhanced performance because of the complementary effects of several feature extraction techniques.

Comparative studies—like the one used in this project—have become more and more common in recent literature. In their comparison of CNN, LSTM, and hybrid models for emotion recognition, Zhu et al. (2019) emphasized the importance of comprehending the advantages of each architecture. The results of the investigation showed that the integration of temporal and spatial information performed noticeably better than separate models.

*Gap in Literature*

A thorough comparison review of CNN 4 layers, LSTM, VGG-Net, ResNet50, and CNN 4 layers is notably lacking in the literature, even though previous studies have examined a variety of architectures for emotion recognition. By methodically assessing and contrasting these four unique architectures, this study aims to close this gap and offer insights into each one's unique contributions as well as how well they function in terms of identifying a wide range of human emotions.

**Method and Dataset**

*Dataset*  
 Facial Expression Recognition 2013 (FER2013) is a dedicated resource created specifically for the purpose of facial expression recognition, upon which our study is built. This dataset, which consists of a large set of 35,887 grayscale images, is extraordinarily diversified, with people of all ages and races. Deliberate diversity guarantees a portrayal of the intricacy present in real-life scenarios, augmenting the dataset with a range of emotions representative of the wider human experience. All 35,887 grayscale photos in the FER2013 dataset have been painstakingly identified and categorized into seven different facial expression categories: Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Neutral. For the purpose of developing and accessing emotion detection models across a variety of emotional states, this thorough labelling offers an invaluable ground truth.

The FER2013 dataset's well-balanced categories and narrow focus are two of its most notable features. The dataset is a perfect framework for targeted research and improvement in the field of emotion recognition because of its purposeful focus on facial expressions. Because of this specificity, models trained on this dataset are refined to perform very well in the complex job of facial expression analysis. The dataset's benefits are further enhanced by the evenly distributed photos across expression categories, which makes impartial model training and evaluation possible.

Like any dataset, FER2013 has its share of difficulties. It may be difficult to extract fine-grained characteristics of facial expressions due to visual quality, especially when it is grayscale and has different resolutions. While simplifying and providing clarity, the categorization of emotions into seven basic categories might not adequately represent the nuances and complexity of the wide range of human emotional reactions. The FER2013 dataset continues to be a mainstay in emotion recognition research in spite of these difficulties, offering a wealth of annotated facial expressions for model evaluation and training.

*Preprocessing:*

Preprocessing the FER2013 dataset entails several crucial actions to guarantee compatibility with the models of our choice:

1. Processing Image Data:

Every picture is scaled to 48 by 48 pixels and converted to grayscale standard. Image data is reshaped into three-dimensional arrays in order to comply with the models' compatibility requirements.

2. Label Preparation:

For classification purposes, the seven expression categories—0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, and 6=Neutral—are processed into a category format.

3. Splitting a dataset:

With 80% of the dataset designated for training and 20% for testing, the dataset is split into training and testing sets.

The training and testing sets yielded the following shapes: x\_train (22967, 48, 48, 1) and x\_test (5742, 48, 48, 1), in that order.

4. Normalization

Model training is made efficient by scaling the pixel values within the images between 0 and 1. Pixel values are standardized throughout the dataset by using the Min-Max scaling technique for normalization. Our careful preprocessing makes sure that the FER2013 dataset is prepared correctly and ready to be included in our CNN, LSTM, VGG-Net, and ResNet50 model’s comparative study. With respect to the distinct qualities and difficulties connected to every category of facial expression, the standardised dataset offers a basis for insightful analyses.

**Results**

1. VGGNet

With an accuracy of 73.23% (as shown in Figure 1.1), VGGNet emerges as the most effective model in our comparative analysis. Its success can be attributed to the architecture's consistency and simplicity, which enable it to successfully capture complex facial expression patterns. VGGNet performs better because of its capacity to recognize minute details in facial features. This outcome demonstrates the value of a well-balanced architecture and the effectiveness of VGGNet in challenges involving the recognition of facial emotions.

  
(Figure 1)

I’ve also shown the confusion matrix (Figure 2), the example prediction (Figure 3), and the CAM for the same(Figure 4)

A chart with different colored squares

Description automatically generated with medium confidence

(Figure 2)

A screenshot of a video call

Description automatically generated

A screen shot of a graph

Description automatically generated

(Figure 3)

*\*\*Note: The CAM that we got in VGGNet was slightly different and not that detailed than the rest because over here we used FER2013 dataset’s csv file and for the rest we have used ‘.png’ files*.

2. ResNet50

ResNet50 produced a test accuracy of 58.69% and a test loss of 1.1079, as shown in Figure 4. The model's performance indicates that it can recognize and categorize facial expressions, despite its moderate accuracy. ResNet50's effectiveness is demonstrated by its ability to record complex gradients and patterns, which helps it overcome the vanishing gradient problem that deep neural networks frequently face. Even if the accuracy is within a reasonable range, more research may be necessary to tweak the hyperparameters for improved performance.

A screenshot of a computer

Description automatically generated

(Figure 4)

I’ve also shown the confusion matrix (Figure 5), the example prediction (Figure 6), and the CAM of a sample image (Figure 7)

A screenshot of a graph

Description automatically generated

(Figure 5)

A close-up of a person's face

Description automatically generated

(Figure 6)

A screenshot of a computer screen

Description automatically generated

(Figure 7)

*\*\*Note: I have included all the plots, matrices, and other images only for VGGNet and ResNet50 as they were our primary model architectures. I haven’t included it for CNN and LSTM in this report, but if you want to check them out then feel free to visit our codebase.*

3. CNN (4 layers)

The reported accuracy of the four-layer CNN model stands at 53.13%. While this accuracy is marginally lower compared to the other models, it is essential to recognize that CNNs are particularly adept at extracting spatial features. This specialization in spatial feature extraction is invaluable in tasks where the nuances of facial expressions are pivotal for accurate emotion identification. Despite its lower accuracy, the four-layer CNN model may still offer valuable insights, especially in scenarios where understanding intricate spatial details is paramount. The nuanced nature of facial expressions, with subtle cues often playing a crucial role, makes the spatial analysis capability of CNNs particularly relevant. This is especially true in applications such as human-computer interaction or mental health assessment, where identifying the fine-grained details of facial expressions is vital. While the model's performance might benefit from additional optimization, such as tweaking architectural parameters or augmenting the dataset with more diverse examples, its unique strengths in spatial analysis remain noteworthy.

4. LSTM

Test accuracy was 54.01% and test loss was 1.2160 according to the LSTM model. Because LSTMs can capture temporal dependencies, they are a good choice for sequential data, such as changing face expressions over time. The degree of accuracy attained indicates the model's capacity to identify trends in the temporal development of emotions. The performance of the LSTM architecture might be enhanced by fine-tuning it, maybe by changing the length of the sequences or adding attention methods.

**Discussions and Conclusions**

Each model showed advantages and contributions in this comparative research of four different deep learning architectures for human emotion recognition: ResNet50, LSTM, a four-layer CNN, and VGGNet. Their individual achievements are elucidated by the evaluation findings on the Facial Expression Recognition 2013 (FER2013) dataset.

*Model Equivalencies*

With a respectable accuracy of 58.69%, ResNet50 demonstrated its ability to capture complex gradients and patterns, especially resolving the vanishing gradient issue that deep neural networks frequently face. With a test accuracy of 54.01%, the LSTM model—which was created for temporal sequence analysis—proved its capacity to identify patterns in the temporal evolution of emotions. With an accuracy of 53.13%, the four-layer CNN demonstrated its prowess in spatial feature extraction.

But with an astounding accuracy of 73.23%, the VGGNet architecture was the study's best performance. VGGNet's design, with its simplicity and uniformity, demonstrated the power of well-balanced architecture for emotion identification tasks by capturing complex facial expression patterns with remarkable effectiveness.

*Highlighting the Efficiency of VGGNet*

The importance of careful architecture design is highlighted by VGGNet's higher performance. Its successful feature extraction was made possible by its streamlined structure, demonstrating the significance of consistency and simplicity in obtaining effective emotion recognition. Although more intricate models can be beneficial in some situations, the grace of VGGNet's design has a special effect here.

*Prospective Courses*

This work creates opportunities for more investigation. Investigating ensemble models, which combine the advantages of many designs, may result in even more reliable emotion identification systems. Increasing model performance can be achieved by adjusting hyperparameters, experimenting with transfer learning, or investigating sophisticated architecture changes.

*Conclusion*

In conclusion, a detailed grasp of the advantages and disadvantages of ResNet50, LSTM, a four-layer CNN, and VGGNet in the context of human emotion recognition is provided by the comparative study of these models. The most effective model is VGGNet, which highlights the significance of a streamlined and balanced architecture. The results offer significant contributions to the area by helping practitioners and academics choose and optimize models for applications involving emotion recognition. The pursuit of increasingly precise and sophisticated emotion identification models is a constantly developing field of research, given the rapid advancements in technology.

**Citations**

1. The Facial Expression Recognition 2013 (FER2013) dataset:

X. Li, "FER2013: Facial Expression Recognition Dataset," IEEE Dataport, 2013. [Online]. Available: <https://dataport.ieee.org/docs/fer2013/1>.

2. DigitalScreeni, A. (2021, October 13). “239 - Deep Learning training for facial emotion detection” Available: <https://www.youtube.com/watch?v=P4OevrwTq78&t=547s&ab_channel=DigitalSreeni>