Performance Analysis and Evaluation of Different Deep Learning Algorithms for Facial Expression Recognition

Iffat Tamanna, Md Ahsanul Haque

Dept of Computer Science& Engineering, Bangladesh University of Business & Technology Dhaka, Bangladesh

DOI: https://doi.org/10.51584/IJRIAS.2023.8719

Received: 24 May 2023; Revised: 24 June 2023; Accepted: 29 June 2023; Published: 05 August 2023

Abstract—Emotions are dynamic biological states that are connected to all of the nerve systems. The problem of facialexpression recognition has been thoroughly investigated, leading to the development of some robust and accurate face recognition algorithms. The effectiveness of three such algorithms (CNN, VGG16, and ResNet50) that have been widely studied and appliedin the research community are investigated and compared inthis paper. The aim is to use grayscale images to train thesetraining models and compare their accuracy and data losses.The system will be able to detect the seven facial expressions Angry, Neutral, Contempt, Disgust, Fear, Happy, and Sad after training these models. To compare their precision, the same batchsize and epoch were used. After reviewing all possible evaluationsbased on these output matrices, it is clear that all three networks produce reliable effect identification, with CNN being the most accurate.

Index Terms—Convolutional Neural Network(CNN),VGG16, ResNet50, Facial Expression, Gabor Filter, Deep Learning

I. Introduction

Every person communicates with others not only through words, but also through body movements, which they useto highlight specific parts of their speech and to express emotions. Emotion can be expressed by many things but facial expression is the best expression to display human emotion [1]. This facial emotion plays a vital role in our communication. Machine learning plays an important in the history of computer science. [2] By using computer-basedtechnology, human facial emotions can be detected. It usedto be able to instantly identify faces, code facial expressions, and recognize emotional states. Using a computer to detect this human emotion is still a huge challenge. The emotion detection method is implemented using many algorithms. The majority of the system accomplishes this by analyzing faces in images or videos through computer-powered cameras embedded in laptops, cell phones, and digital signage systems,as well as cameras mounted on computer screens. Emotion data is now being used by market analysts to deploy product advertisements. Emotion recognition is used in the videogame research process. By using facial expression analysis, game developers can obtain knowledge and draw conclusions about the feelings encountered during game play, and integratethat input into the final product [3]. A lot of recent study is centered on the use of artificial intelligence (AI) and deep learning algorithms to classify emotions. People are working nonstop to close the distance between machine and human contact.

In this paper we have tried to figure out the bestmethodological approach between Convolutional neural network (CNN), VGG16, and ResNet50 for human facial detection. Our machine receives an image as input, and then we use these models to predict the facial expressionmark, which should be one of the following: Angry, Neutral, Contempt, Disgust, Fear, Happy, or Sad. Emotion can be used in a variety of ways in our business and everyday lives. In thisanalysis, we compared three different approaches currently in use to see which one provides the best precision for a vast volume of data, such as FER-2 which is taken from kaggle.

The rest of the paper is summarized as follows.A briefreview of some existing research work is provided in Section I. In Section III, a detailed description of emotion detection techniques is presented. In Section IV, experimental procedureis shown. Result and Performance analysis is presented inSection V. Finally, a conclusion with future work is provided in Section VI.

II. Related Work

Over the last few decades, many kinds of research havebeen conducted and the interest of research on the formulationof methods and systems for the classification and recognition of human facial expressions raises. We have found that mostof the studies and implementation are done to use facial expressions detection from videos and images. Currently, mobile application and social media websites use different techniques to detect emotions.

There has been work conducted on seven classes [4] with ResNet50, VGG16, and SE-Resnet50 to distinguish facial recognition from seven classes. Instead of accuracy,this research focuses on the precision and recall of qualified models. The most critical criterion for determining how well the model works is consistency. They can’t make an informed decision on the right model if they don’t have enough accuracy.

Another research [5] used a convolutional neural network to identify five facial emotions using 10,000 photographs of 156 individuals. This study named this system as FERC. In this scheme, the exponential vector (EV) is used to create various emotions. To get a more precise answer, a durationof 24 EV was used. In comparison to FER-2013, there are fewer samples.

Emotion recognition was shown to be extremely necessary for computer vision in another [6]. They've classified feelings into seven categories. This thesis suggested a video-basedmethod. A current dataset consists of 700 photographs that were used in the creation and testing of this model. Theyused the Resnet50 model, but it's difficult to tell whetherit'll be positive or bad in the long run with such a tiny dataset.

In [7], employs the VGG16 and Bidirectional Long Short- Term Memory (Bi-LSTM) recurrent neural networks to learn spatial-angular features from perspective function sequences, with both forward and backward angular relationships being investigated. This system is designed to handle both vertical and horizontal data. In certain cases, there might be insufficient evidence for certain conditions, making it difficultto obtain accurate results.

Another analysis [8] reveals that Gabor filters are usedto isolate features, and then Convolutional Neural Networks are used to classify six distinct facial emotions. While the machine extracts more images and has greater precision, there is a sample issue since this system did not use large data samples for training models, leaving the question of whether the system would cope with large data samples.

In [9], they explained the hierarchical convolutional neural network (HCNN) can be used to identify three distinct facial emotions: positive or happy, neutral, and negative or sad.Only three emotions are insufficient to declare the system's effectiveness and humans have more than three emotions. They need a training model that can identify a wider rangeof emotions.

III. Emotion Detection Techniques

In this section we will discuss about our proposed methods based on deep learning facial emotion recognition.

A. Convolutional Neural Network (CNN)

A deep feed-forward artificial neural network is a convolutional neural network. Typically, this feed forwardneural network is used to analyze visual imaging. A convolutional neural network is made up of three layers. The reference layer, secret layers, and output layer are the three layers. Since the activation mechanism and final convolution mask the inputs and outputs of every feed-forward neuralnetwork's middle layers, they are referred to as secret layers [10]. Convolutions are done by the hidden layers of a convolutional neural network.

Convolutional layers convolve the inputs and send the output to the next line. There must be an equivalent numberof input and output sources.At the Pooling layers, data dimensions are simplified by integrating the outputs of neuronclusters in one layer into a single neuron in the next layer. At local pooling, small clusters are combined; the most typical tiled scale is 2x2. Global pooling layers work on almost allof the neurons in the featured map.

In fully connected layers, each neuron in one layer is connected to each neuron in another layer. It works in the sameway as a conventional multi-layer perceptron neural network. The flattened matrix passes through a completely connected layer to identify the images [11]

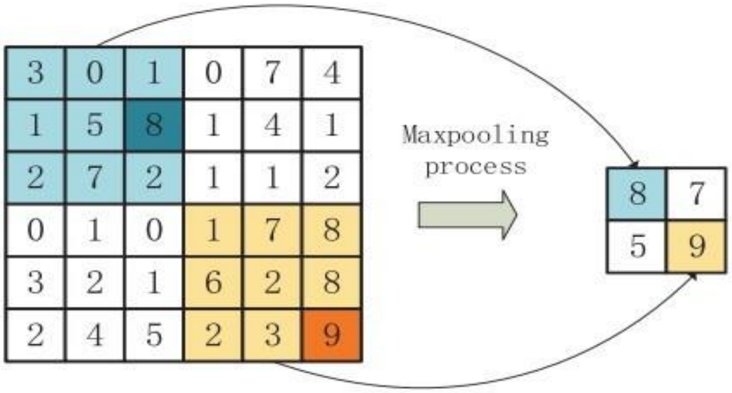


Fig. 1. Max Pooling Layer of CNN [11]

B. VGG16:

The VGG16 model is a type of convolutional neural network. For large-scale picture detection, it is used in very deep 'Convolutional Networks.'

Instead of making a large number of hyper-parameters, this model based on having 3x3 filter convolution layers with a stride 1 and still used the same padding and max-pool layer of 2x2 filter stride 2. And it maintains this convolution and max- pool layer structure throughout the entire architecture. [12]

C. ResNet50:

The term "residual Network" is a shorthand for "residual network." A convolutional neural network is ResNet-50.

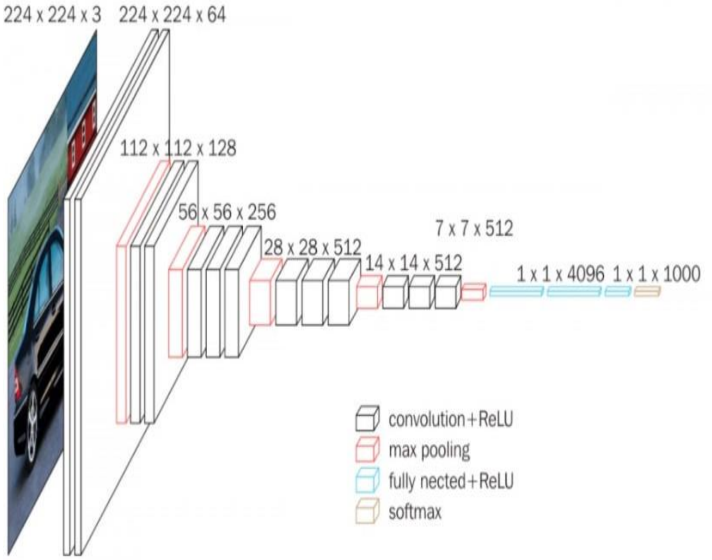


Fig. 2. VGG16 Model [5]

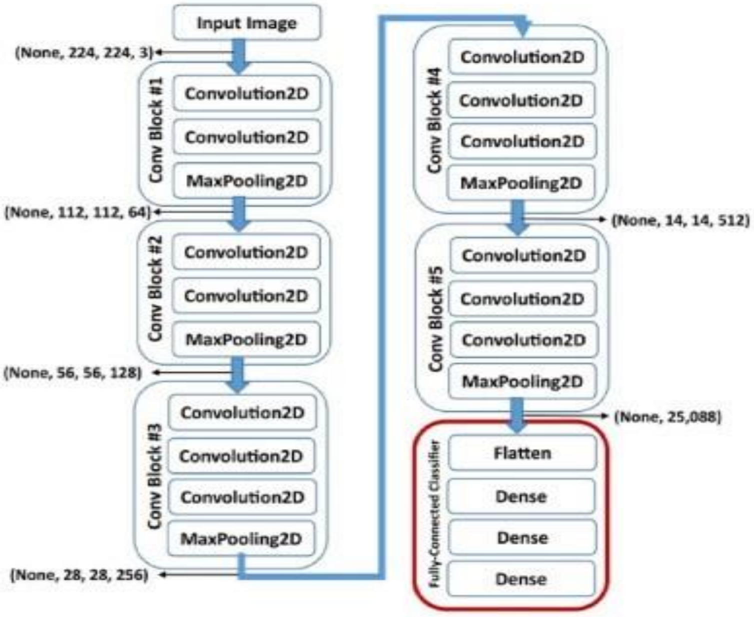


Fig. 3. VGG16 block diagram [12]

There are 50 layers in all. It's a convolutional neural network subclass. With CNN, the most common image classification algorithm is ResNet. The network's image input resolution is 48x48 pixels. It has allowed us to train extremely deep neural networks, which was ResNet's major breakthrough.

The ResNet-50 model is divided into five stages, each with its own convolution and identity block. There are three convo-lution layers in each convolution and identity block. Residual neural networks hop through several layers by skipping links or short-cuts. In the diagram below, the skip relation is called ”identity”. It enables the network to learn the identity function,allowing the input to bypass the other weight layers and move through the stack [13].

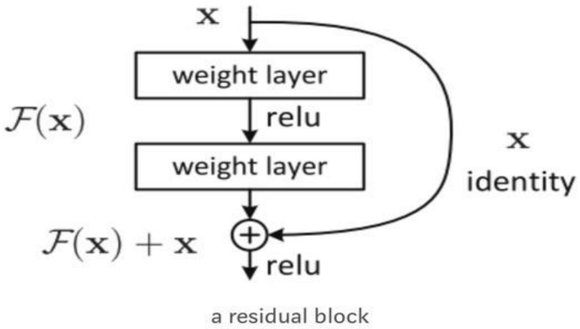

x
weight layer
relu
weight layer
$\mathcal{F}(\mathbf{x})$
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu
a residual block

Fig. 4. ResNet50 skip through layers [13]

IV. Experimental Procedure

In this experiment, we have used FER-2013 data from Kaggle to build the model and used google colab to execute our codes. A set of data consisting 48x48 pixel grayscaleimages of faces is used in our experiments. The faces have been registered as if all the face is more or less centered and occupies about the same amount of space in every image. In this data consists of two different data sets. They are training data sets and validation data sets. The training set consistsof 28,709 examples and the validation set consists of 7,066 examples. And each of the sets have 7 emotion categories (0: Angry, 1: Disgust, 2: Fear, 3: Happy, 4: Sad, 5: Surprise, 6: Neutral).

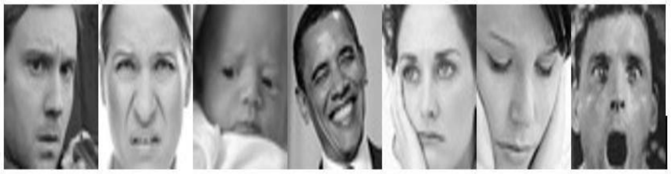


Fig. 5. Seven classes of emotions in the FER-2013 data set.

In our experiment we have categorized the whole data sets in two parts. For every method we have got different trained model. In our first category we have used five emotion setsfor every training (0: Angry, 1: Happy, 2: Sad, 3: Surprise,4: Neutral). Which includes total 24,282 data in train dataset and 5,937 data in validation data set and for test data setsthere was 3006 data.

And on another data set there are 28,709 data in training dataset and 7,066 data in validation dataset and contains allthe seven emotion sets as mentioned before.

While training the model we have used Keras deep learningmethod. Each of the images is zoomed in, zoomed out, sheared, flattered etc. are done in every method to make more data to train and validate. For all training models like “CNN”, “RESNET50”,” VGG16” we have used same constrains. We have used Five and Seven classes (For two different cases)to define emotions to the model in every method. A constant

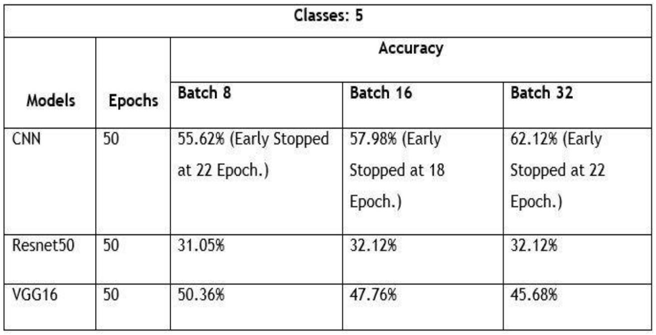
number of epoch (50 epoch) with three different batch sizes (8,16, 32) all the model was trained. After finishing the training of the model there were total 18 number of trained models.We have examined all the model with several random images to know how much the models can detect the emotions, we have used OpenCV for that process.

V. Result & Performance Analysis

For different parameters we got different results in every experiments. We have used 5 and 7 classes to train modelsand got different results for different epochs and batch sizes.

The Following table shows the accuracy of different exper- iments of 5 classes and three different batch sizes with fixed epoch of 50. Some models' preparation was halted early in some epochs for minor changes.

TABLE I ACCURACY OF DIFFERENT TRAINED MODELS WITH 5 CLASSES.



Classes: 5				
		Accuracy		
Models	Epochs	Batch 8	Batch 16	Batch 32
CNN	50	55.62% (Early Stopped at 22 Epoch.)	57.98% (Early Stopped at 18 Epoch.)	62.12% (Early Stopped at 22 Epoch.)
Resnet50	50	31.05%	32.12%	32.12%
VGG16	50	50.36%	47.76%	45.68%

After reviewing the results of our experiments shown in Fig.6, we can see that the accuracy of each model variesdepending on the parameters. As batch size increases for Five Classes and 50 epochs in CNN, accuracy increases as well; similarly, we can see similar types of activity in RESNET50. On the other side, we can see that as the size of the batch grows larger, the accuracy of VGG16 decreases.

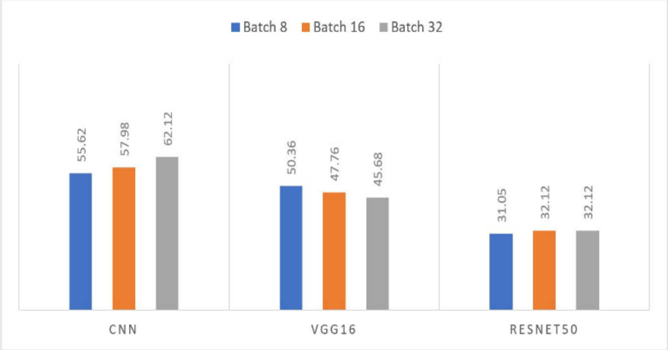
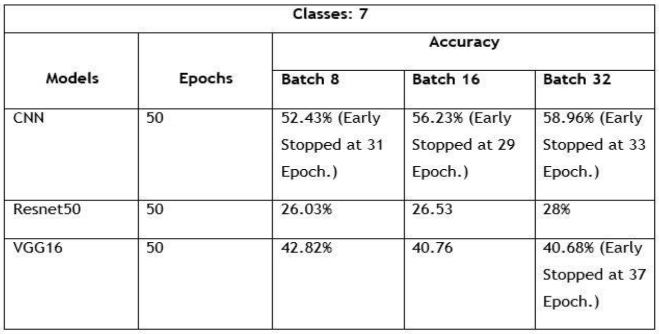


Fig. 6. Accuracy of models in different parameters with 5 emotion classes.

The Following table shows the accuracy of different exper- iments of 7 classes and three different batch sizes with fixed epoch of 50.

TABLE II ACCURACY OF DIFFERENT TRAINED MODELS WITH 5 CLASSES.



Classes: 7				
		Accuracy		
Models	Epochs	Batch 8	Batch 16	Batch 32
CNN	50	52.43% (Early Stopped at 31 Epoch.)	56.23% (Early Stopped at 29 Epoch.)	58.96% (Early Stopped at 33 Epoch.)
Resnet50	50	26.03%	26.53	28%
VGG16	50	42.82%	40.76	40.68% (Early Stopped at 37 Epoch.)

For seven classes, as shown in Fig.8, provide similar charac-teristics. However, the accuracy here is lower than the productof five groups. As a result, the more groups there are, the moredata there is and the lower the accuracy. If we want to improveaccuracy, we must reduce the dataset size, exclude classes, or do both.

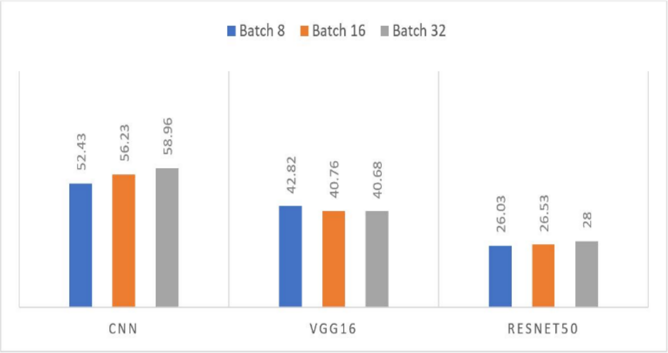


Fig. 7. Accuracy of models in different parameters with 7 emotion classes.

From the results it is clear that CNN is performing best in facial emotion recognition as it has better accuracy than other two. Also, accuracy depends the number of data sets, classes and batch sizes while our epoch was constant.

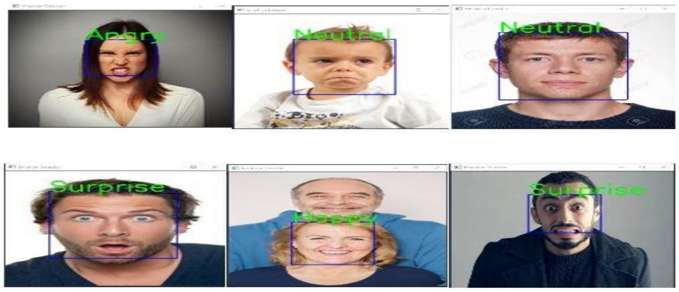


Fig. 8. Different Test of Model

When we have used five classes there were less images usedfor validation than the seven classes. As a result, accuracy improves for every training model. Moreover, changing the batch sizes affects the accuracy of our model. RESNET50and VGG16 are fare away from CNN accuracy. While testing with several other images CNN gave almost every emotion with correct result. So, CNN will be the best choice for Facial emotion Recognition.

VI. Conclusions and Future Work

Using the FER-2013 data collected from Kaggle, we have shown performance analysis of three different methods offacial emotion recognition system. According to the resultsof the present findings, CNN outperforms the other two approaches. CNN does well, but not as well as anticipated. In certain cases, a trained model is unable to classify the correct emotions in some categories.We can work with more methodsand can try to build our own method and use our own image datasets to improve accuracy.

REFERENCES

1. S. Alizadeh and A. Fazel, “Convolutional neural networks for facial expression recognition,” *arXiv preprint arXiv:1704.06756*, 2017.
2. M. A. Rahman, H. Ahmed, and M. M. Hossain, “An integrated hardwareprototype for monitoring gas leaks, fires, and remote control via mobile application,” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 10, 2022.
3. “Face expression recognition - what is it? — how to use it?” https://l- ink.in/sj32e, (Accessed on 05/10/2021).
4. D. Theckedath and R. Sedamkar, “Detecting affect states using vgg16, resnet50 and se-resnet50 networks,” *SN Computer Science*, vol. 1, no. 2, pp. 1–7, 2020.
5. N. Mehendale, “Facial emotion recognition using convolutional neural networks (ferc),” *SN Applied Sciences*, vol. 2, no. 3, pp. 1–8, 2020.
6. B. Li and D. Lima, “Facial expression recognition via resnet-50,” *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 57–64, 2021.
7. A. Sepas-Moghaddam, A. Etemad, F. Pereira, and P. L. Correia, “Facial emotion recognition using light field images with deep attention-based bidirectional lstm,” in *ICASSP 2020-2020 IEEE International Confer- ence on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3367–3371.

8. M. M. T. Zadeh, M. Imani, and B. Majidi, "Fast facial emotion recognition using convolutional neural networks and gabor filters," in *2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI)*. IEEE, 2019, pp. 577–581.
9. J. Li, Z. Zhang, and H. He, "Hierarchical convolutional neural networks for eeg-based emotion recognition," *Cognitive Computation*, vol. 10,no. 2, pp. 368–380, 2018.
10. M. Rahaman, M. Chowdhury, M. A. Rahman, H. Ahmed, M. Hossain, M. H. Rahman, M. Biswas, M. Kader, T. A. Noyan, and M. Biswas, "A deep learning based smartphone application for detecting mango diseases and pesticide suggestions," *International Journal of Computing and Digital Systems*, vol. 13, no. 1, pp. 1–1, 2023.
11. "Convolutional neural network - wikipedia," https://l-ink.in/G90C8, (Ac-cessed on 05/10/2021).
12. "Vgg16 - convolutional network for classification and detection,"https://l-ink.in/Lzv7j, (Accessed on 05/10/2021).
13. "Understanding and coding a resnet in keras — by priya dwivedi —towards data science," https://l-ink.in/hx0r0, (Accessed on 05/10/2021).