Emotion Classification Using Deep Learning

Md. Mihal

ID: 2018-1-60-020

Md. Muyeen – Ul – Islam

ID: 2018-1-60-022

Daniel Deep Pereira

ID: 2018-1-60-023

Md. Mominur Rahman Emon

ID: 2018-1-60-136

A thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering

Department of Computer Science and Engineering
East West University
Dhaka-1212, Bangladesh

June, 2022

Declaration

I, hereby, declare that the work presented in this thesis is the outcome of the investi- gation performed by me under the supervision of Dr. Taskeed Jabid, Professor, Department of Computer Science and engineering, East West University. I also declare that no part of this thesis/project has been or is being submitted elsewhere for the award of any degree or diploma.

Countersigned

. . . .
Dr. Taskeed Jabid

Supervisor

Signature

. . . .
Md. Mihal
(2018-1-60-020)
Signature

. . . .

Md. Muyeen – Ul – Islam
(2018-1-60-022)

Signature

. . . .

Daniel Deep Pereira
(2018-1-60-023)
Signature

. . . .

Md. Mominur Rahman Emon
(2018-1-60-136)

Abstract

Emotion classification using deep learning classifier (DNN) namely convolutional neural network (CNN) is a very popular topic for research in this era. In this paper, we proposed two CNN based models are applied on FER-2013 dataset and our very own dataset that is also based on FER-2013 but it has a large number of augmented images. Both methods are applied on two phases, first on FER-2013 dataset and secondly on our dataset. Our first proposed method achieved an accuracy of 94.71% on FER-2013 and 99.16% on our dataset. Our second method achieved an accuracy of 88.3% on FER-2013 and 89.8% on our dataset. Finally results of both approaches are compared with previous researches.

Acknowledgments

For the better understanding, a sample Acknowledgment is given below.

As it is true for everyone, I/We have also arrived at this point of achieving a goal in my/our life through various interactions with and help from other people. However, written words are often elusive and harbor diverse interpretations even in one’s mother language. Therefore, I/We would not like to make efforts to find best words to express my thankfulness other than simply listing those people who have contributed to this thesis itself in an essential way. This work was carried out in the Department of Computer Science and Engineering at East West University, Bangladesh.

First of all, I/We would like to express my deepest gratitude to the almighty for His blessings on me/us. Next, my/our special thanks go to my/our supervisor, Dr. Taskeed Jabid, who gave me/us this opportunity, initiated me/us into the field of Emotion Classification Using Deep Learning, and without whom this work would not have been possible. His encouragements, visionaries and thoughtful comments and suggestions, unforgettable support at every stage of my/our B.Sc. study were simply appreciating and essential. His ability to muddle me/us enough to finally answer my/our own question correctly is something valuable what I/We have learned and I/We would try to emulate, if ever I/We get the opportunity.

I/We would like to thank Md. Mihal for his excellent collaboration during performance evaluation studies; Daniel Deep Pereira for his overall support; Md. Mominur Rahman Emon for her helpful suggestions in solving tricky technical problems. Last but not the least, I/We would like to thank my/our parents for their unending support, encouragement and prayers.

There are numerous other people too who have shown me their constant support and friendship in various ways, directly or indirectly related to my/our academic life. I/We will remember them in my/our heart and hope to find a more appropriate place to acknowledge them in the future.

Md. Mihal
June, 2022

Md. Muyeen – Ul – Islam
June, 2022

Daniel Deep Pereira
June, 2022

Md. Mominur Rahman Emon
June, 2022

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Computer vision, the concept of this term was developed in 1959. Since then, many progresses have been made until now. The first real time face recognition system appears in 2001 and after that it became a very popular area for researchers from all over the world. Computer vision is a field of artificial intelligence that derives meaningful information from digital images by using computer or computer-based systems. Convolutional Neural Network (CNN), a type of Deep Learning Classifier (DNN) is arguably the main technology behind computer vision [1].

Human being has a common set of emotions that can be recognized all across the world. In this set of emotions, all the emotions are different from one another. These emotions can also be described as discrete. There have been many studies to determine the most basic emotions. In a study of Paul Ekman and his colleague in 1992, they concluded six basic emotions which are anger, disgust, fear, happiness, sadness, and surprise [2]. In this project, we used seven different types of emotions including these six which are anger, disgust, fear, happiness, sadness, surprise and added another emotion which is neutral.

Deep Neural Network can also be considered as a stack of neural networks. Deep neural network contains at least two layers. Deep nets process data in complex ways by employing sophisticated math modeling. In order to build deep neural network, a few things needed to be built first which are artificial intelligence, machine learning, artificial neural networks [3]. In order to deep neural network to exist. Machine learning needed to be developed. Machine learning is a framework to automate statistical models through algorithms to make better predictions. A model that learns takes all its bad predictions and tweaks the weights inside the model to create a model that makes fewer mistakes. Then comes the development of artificial neural network. ANN uses hidden layers to store and evaluate how significant one of the inputs is to the output. Then comes the part of deep neural networks. Here deep means the number of hidden layers in the model. The more number of hidden layer there is in the model, the deeper the neural network is [4].

Convolutional neural network is a type of artificial neural network used in image recognition and processing that is specifically designed to process pixel data. CNN uses deep learning to perform generative and descriptive tasks including image and video recognition. A neural network is a system of hardware or software patterned after the operation of neurons in the human brain. In CNN, neurons are designed in such a way as to cover the entire visual field avoiding the piecemeal image processing problem of traditional neural networks. CNN uses multilayer perception. The layers of CNN includes an input layer, an output layer and a hidden layer containing multiple convolutional layers, pooling layers, fully connected layers and normalization layers [5].

Chapter 2

Literature Review

Emotion detection and classification is one of the most researched topics in the modern era of machine learning. The ability to accurately detect and identify a facial expression, human emotion solely improves the performance of computer vision and open up new ways of advanced human computer interaction. Many researchers have worked on the aspect of facial emotion and expression recognition. A handful of approaches have been made to improve facial emotion recognition on the FER-2013 dataset. Mollahosseini proposed a deep neural network that uses FER-2013 across 7 public datasets. His proposed network contained and exclusive inception layer architecture. Liu at el. Proposed a boosted deep belief network (BDBN). The boosted network iteratively performs feature learning, feature selection, and classifier construction within a unified framework. Later, Liu at el. Established the AU-inspired deep neural network (AUDN) [6]. In the CNN perspective, Khorramis have demonstrated that CNNs are better performer in emotion detection as they can intelligently calculate the weights of different facial regions thus perform more competitively. Sun et al. proposed a fast and simple architecture called ROI-KNN for FER. Cui et al. proposed a multiple CNN ensembled algorithm for FER. Xiao Sun proposed a ROI-guided deep architecture for robust facial expression recognition that has a 40.13% accuracy on FER-2013 dataset. LeNet-5 was the earliest convolutional neural network that could detect handwriting. VGG Net used very small convolution filters (3*3) to increase the architecture depth [7]. There are 16-19 weighted layers in the network that helps the algorithm to achieve significant improvement on the prior art configurations. GoogleNet is a deep-wide neural network that has 22 layers. Inception layer is the distinguishable factor here comparing with other neural networks. Inception layers have different sizes of convolution filter, so that the input images can convolve at different scales of feature maps. ResNet offers a network that increases the training speed and improves the training effect and also avoids gradient disappearance. Fernandez et al. proposed an end to end network with an attention model for facial expression recognition. Thanh-Hung Vo proposed a pyramid with super-resolution network architecture to deal with the different-image-size problem for in the wild facial emotion recognition task. Emotions are classified based on biophysical signals too [8]. Oana Balan showed a study of comparison of different machine learning algorithms on the DEAP dataset. The DEAP database was created with the purpose of developing a music video recommendation system based on the user’s emotional responses [9].

In our proposed method, we proposed two CNN based methods on two versions of FER-2013 dataset. We applied all the algorithms of stock FER-2013 dataset and our very own modified FER-2013 dataset.

Chapter 3

Background

TensorFlow:

TensorFlow is an open-source library for large scale machine learning and numerical computation[4]. It was created by the google brain team and was released in 2015. TensorFlow has a large number of machine learning and deep learning models. It also has a large number of pre-trained models. TensorFlow can train and run deep neural networks for many purposes, in our case image classification. TensorFlow supports a high level API and both CPU and GPU computing devices [10].

Keras:

Developed by google and written in python, Keras is a deep learning API for implementing neural networks. Keras supports multiple backend neural network computation. Keras also supports TensorFlow back ends. Keras is officially embedded in TensorFlow and it provides inbuilt modules for all neural network computation. Keras is extremely user friendly specially for beginners. Keras is modular in nature and therefore it is expressive, easy and flexible for innovative research [11].

PyTorch:

PyTorch is a Tensor library which highly optimized based on python. It uses dynamic computation graphs. Torch is mainly used for such applications that uses GPUs and CPUs. It allows an individual to run and test the whole code or potions of it in real time. By doing so, it saves time and cost. PyTorch has a strong GPU acceleration support. As it uses dynamic computation graphs, the process is more flexible than static graphs. Users can can make interleaved construction and valuation of the graph. It allows line by line code execution and so it is very debug friendly [12].

OpenCV-Python:

OpenCV stands for open-source computer vision library. OpenCV-Python is a library of python bindings designed to solve computer vision problems. OpenCv was first invented by intel in 1999. The first release of OpenCV came out in 2000. Generally, OpenCV supports a wide range of programming languages like C++, Python, java. OpenCV-Python makes use of Numpy and other libraries that optimizes machine learning algorithm [13].

CNN:

CNN stands for convolutional neural network. It is also known as ConvNet. It is a class of neural networks. It specializes in processing data that has a grid like topology, for example an image. CNN is like human brain, consists of numerous numbers of neurons. Every neuron is connected with each other. Each neuron works on its respective fields. Data is processed in its own respective field by the neurons in CNN. The layers are built in such way that the first layers detect simpler patterns and the deeper the network goes, it detects more complex patterns. A CNN consists of several layers [5].

Chapter 4

Methodology

CNN:

Convolutional neural network is an essential technology in terms of emotion classification. CNN follows a specific work flow- input image, convolutional layer (kernel), pooling layer, fully connected layer (classification). CNN is made up of neurons like brain cells. The neurons can bear learnable weights and biases. Neurons can receive various inputs and calculates a weighted sum over the inputs. The sum is then passed through an activation function that generates an output. Each neuron is connected to other neurons and each neuron in a CNN processes data only in its receptive field [14]. Different layers are used in CNN to detect complex patterns, in our case facial emotion from live objects and images.

An image is basically a matrix of pixel values. RGB images have three planes and grayscale images have one single plane. CNN can reduce the dimension of the image to the point that it is easier to process. CNN also maintain all the features of the image into one piece. By doing so a more accurate prediction can be obtained [15].

The very first layer of CNN is the convolutional layer. The first layer is often recognized as the core layer of CNN. Convolutional layer extracts useful features from the images. The process is performed by convolution filtering. At first a portion of the image is selected that represents the feature of that image, it can be compared with a drag window [16]. After that the feature and each portion of the scanned image is used to calculate the convolution product. The feature is seen then as a filter. The convolution product is a dot product of two matrices, the first matrix is the kernel which is the set of learnable parameters, the other matrix is the selected portion of the image. The kernel of CNN works on the basis of this formula-Image Dimensions = n1 x n2 x 1, here n1 is the height of the image, n2 is the breadth of the image, and l is the number of channels. Size of the kernel is pre-defined. During the forward pass, the kernel goes across the height and width of the image, and on its way it generates convolutional product. The sliding size of the kernel is called stride. The kernel parses the image until it completes the breadth of the image, the breadth is fixed for example the breadth of an RGB image is 4. This process goes on row by row until the whole image is covered. The parsing continues until every part of the image is traversed. This process an two dimensional representation of the image which is known as activation map or feature map. The feature map shows the location of the features of the image, the greater the value, the more the corresponding place in the image resembles the feature. The feature map gives the response of the kernel at each spatial position of the image. If we consider an input size of n=n1*n2*l, a spatial size of F with stride S, padding amount P, then the output volume can be determined by the following formula: Nout= ((n-F+2P)/S)+1. If there is lout number of kernels, then the output volume size will be Nout*Nout*Lout [17]. There can be several numbers of convolutional layers. The first layer usually extracts basic features such as horizontal or diagonal edges. The output of the first layer is then passed to next layer to detect more complex features like complex edges and corners. As more layers are implemented, the network can detect faces and expressions.

The second layer of a CNN is pooling layer. Pooling layer reduce the spatial size of the convolved feature. This decrement is required to lessen the computational power for the process. It is also very useful the extract the dominant features. As it reduces dimension, pooling layer always preserve the important characteristics of the image. Pooling layer also eliminates over-learning.

Pooling layer can be placed between two layers of convolution, doing so takes several feature maps and applies pooling operation on each of them. Pooling layers typically have two hyperparameters: 1. Spatial dimension and 2. Stride. There are mainly two types of pooling: 1. Max Pooling and 2. Average Pooling. Max pooling is the more used pooling, it provides the maximum value within the covered image by the kernel. Average pooling calculates and return the average value within the covered image by the kernel. Max pooling layer also performs as a noise suppressant. Max pooling discards the activation which contain noisy activation, it also performs dimensionality reduction while de-noising. If we apply a max pooling layer that is (2*2*2) on a convolved layer that is (26*26*32), the output volume will be 13*13*32 [18].

The fully connected layer does the flattening part of the process. This layer takes the high-level filtered images than translate those into votes. In traditional neural network, fully connected layer is considered as the building block of the network. The feature map from pooling layer is converted into a single column matrix. Then it is fed to the network for processing. Linear combinations and activation function are applied in the processing. It returns a vector size of N, here N is the number of classes in the image. If there are two classes, Softmax activation function is applied. Each input is multiplied by wight and then summed. After that activation function is applied. Weights are learned by backpropagation in the process.

When all features are connected to fully connected layer, the process can cause overfitting. That means the model might have an excellent performance on the training data, but as soon as new data are fed, the performance drops drastically. To overcome this problem, dropout layers are used. It drops a few neurons from the neural network during training process. This reduces the size of the model.

Activation function adds non-linearity to the network. Activation functions are used to learn and approximate all kinds of complex and continuous relationship between variables of the network. It mainly decides which information should pass forward and which are not. ReLU, Softmax, Sigmoid are some popular activation functions. They have specific usage according to specific scenario. For multi class-classification, Softmax activation is used [16].

Proposed Method 1:

Our first proposed method is based on a simple CNN concept. After importing important libraries for learning such as TensorFlow and Keras, batch size is defined. A mini batch gradient descent (64) is used here. After declaring 28709 images of 7 different classes as training set and 7178 images of 7 different classes as testing set, the model was built. In the model there is a convolutional layer, and it convolutes the input total of 8 times. In convolution layer the padding set to same to preserve the spatial dimensions of the input volume. First three convolutions used 96 filters to learn, and rest of the layers used 192 filters. Kernel size remained 3*3. As for activation, ReLU is used to preserve the properties that make linear models easy to use. First three convo2D has a dropout rate of 0.2 and the others has a dropout rate of 0.5. A dense layer is used at the end of the network. As for optimizer Adam is used having a learning rate of 0.0001.

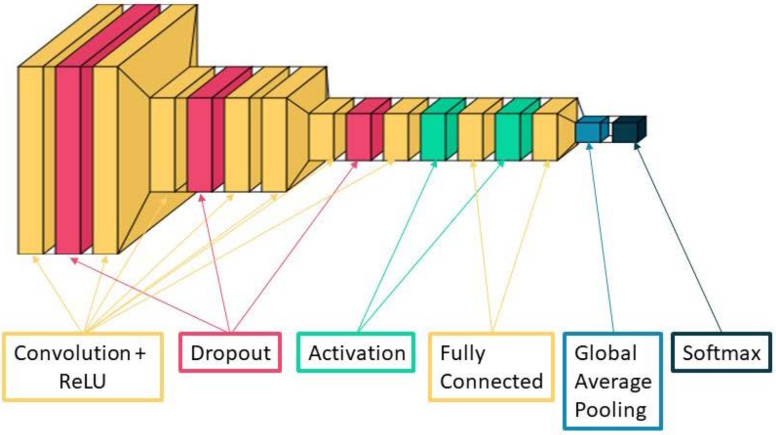

Figure 1: Structural View of Proposed Method 1

Proposed Method 2:

The second algorithm we applied is a complex variant of CNN architecture. The core model has some similarities with AlexNet. After importing necessary libraries from Tensorflow, Keras and PyTorch, training and testing datasets were defined. Training contains 28709 images and testing contains 7178 images both containing 7 classes. Training images were augmented using the following parameters:

Rotation	40 degrees
Shear	0.2
Zoom range	0.2
Horizontal flip	True
Brightness range	(0.2, 1.5)

Table 1: Properties and Values of Augmentation

This model has 4 convolutional blocks or layers. The architecture is stated below:

		Kernel	Padding	Batch Normalization	Activation	Max Pooling2D	Dropout
1st layer	Convo2D	3*3	Same	Yes	ReLU	2*2	0.25
2nd layer	Convo2D	5*5	Same	Yes	ReLU	2*2	0.25
3rd layer	Convo2D	5*5	Same	Yes	ReLU	2*2	0.5
4th layer	Convo2D	7*7	Same	Yes	ReLU	2*2	0.5

Table 2: Architecture of Convolutional Layers

After convolutions there is a flatten layer. The comes two fully connected layers. Each contains a dense layer of 192 neurons. It also has batch normalization. As for activation function ReLU is used. Dropout rate here is 0.25. For the activation layer we have used Softmax. As for optimizer Adam is used at a learning rate of 0.0001.

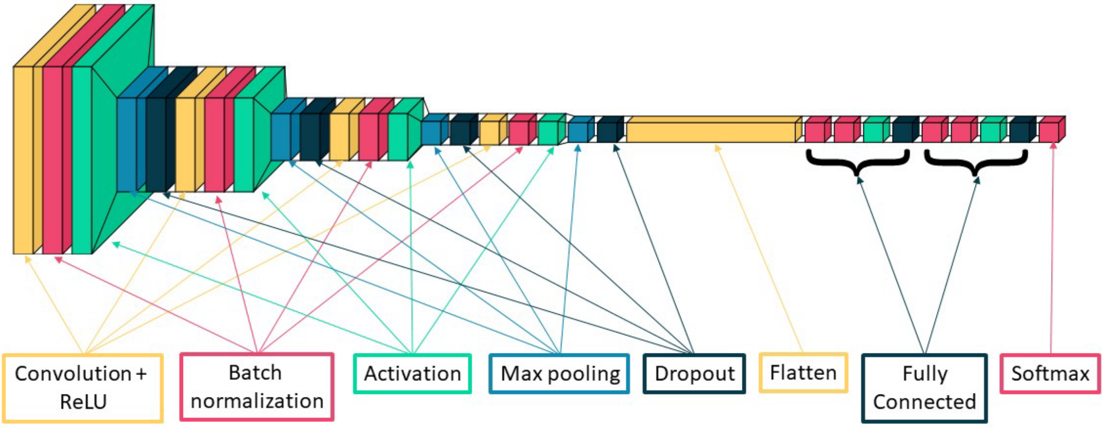

Figure 2: Structural View of Proposed Method 2

Chapter 5

Dataset

FER-2013

The facial expression recognition dataset 2013, namely FER-2013 is a very popular and of the most used dataset in emotion recognition using machine learning. This dataset consists of 48*48 pixels grayscale images of different faces. All the faces have been automatically registered and for that the faces of the images are more or less centered and occupied about the same amount of space in each image. Each face is categorized based on the emotion. There are seven different categories regarding facial emotions, 0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral. In the dataset, the training portion contains 28709 images and test set contains 7178 images [19].

Pre-augmented FER-2013

This is our own version of FER-2013 dataset. This dataset contains training images and testing images. We have gathered some extra images of the seven different classes and added them to the dataset. The images are grayscale and 48*48 pixel. Later there were some augmentations applied on the dataset. We have applied +90 degrees rotation, -90 degrees rotation, +70% scaling on all the images. The original images of FER-2013 are also included in the database. Doing so increases the number of images in the disgust and surprise section. The original dataset has 111 images of disgust expression. Our dataset has 444 images of disgust classification. Total number of train images were 115281 and total number if test images were 28263.

Chapter 6

Results

1. On Normal FER-2013

We have applied three approaches of CNN in the FER-2013 dataset. Total number of train images were 28709 and total number if test images were 7178.

a) Proposed Method 1:

We have applied the CNN algorithm the FER-2013 dataset. We set the batch size to 64. The input size of the images are 48*48pixels. We applied the augmentation while declaring the train set. We augmented the images through shear, zoom, horizontal flip and brightness range. The number of classes are 7 as there are 7 emotions in the dataset. We used three different kernel sizes which are (3,3), (5, 5) and (7, 7). We added 3 dropout layers and global average pooling on the model. The learning rate was set to 0.0001 and the optimizer we used is Adam optimizer. We user the Softmax activation and ran our tests. After running the algorithm the accuracy we got was 94.71%.

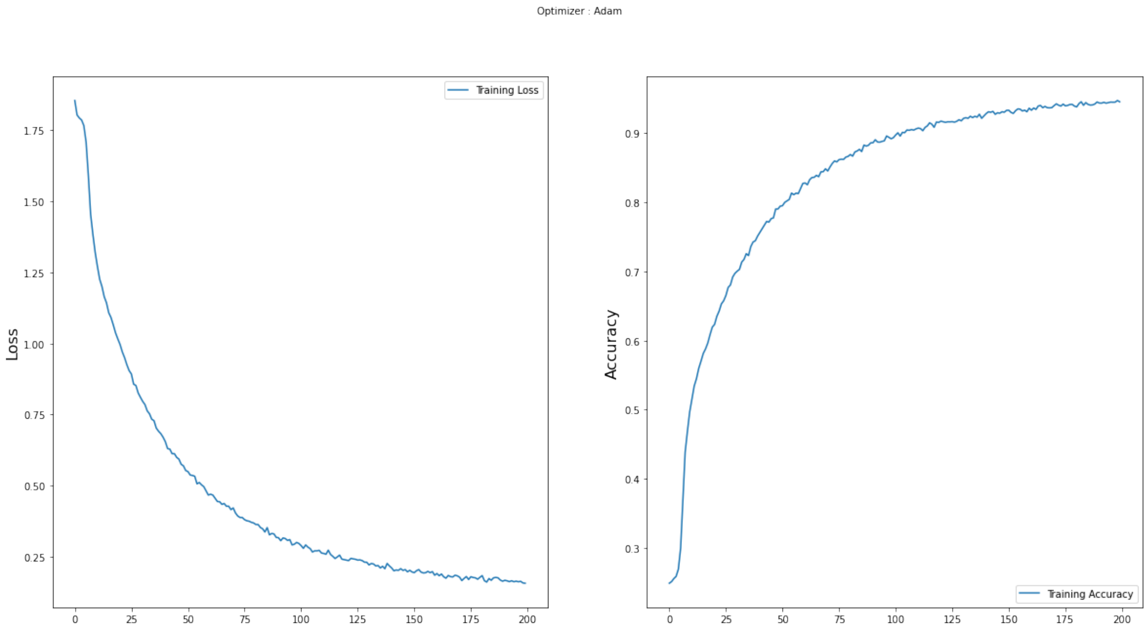

Figure 3: Accuracy and Loss Plot of Proposed Method 1

b) Proposed Method 2:

We have applied the pure CNN algorithm the FER-2013 dataset. We set the batch size to 64. The input size of the images are 48*48pixels. We applied the augmentation while declaring the train set. We augmented the images through shear, zoom, horizontal flip and brightness range. The number of classes are 7 as there are 7 emotions in the dataset. We used three different kernel sizes which are (3,3), (5, 5) and (7, 7). We added dropout layers, max pooling 2D layers and batch normalization layers with every convolutional layer and added a flatten layer on these layers. The learning rate was set to 0.0001 and the optimizer we used is Adam optimizer. We user the Softmax activation and ran our tests. After running the algorithm the accuracy we got was 88.3%.

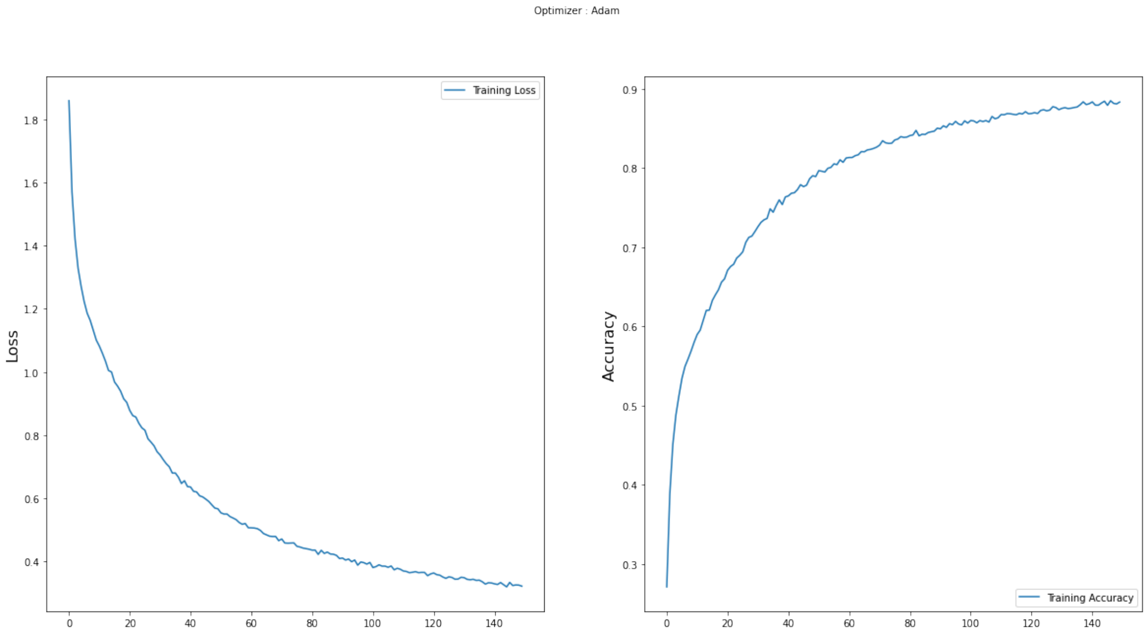

Figure 4: Accuracy and Loss Plot of Proposed Method 2

2. On Modified FER-2013

We have applied three approaches of cnn in the modified FER-2013 dataset where we increased the number of images in the dataset and manually augmented the images. Total number of train images were 115281 and total number if test images were 28263.

a) Proposed Method 1:

We have applied the CNN algorithm the modified FER-2013 dataset. We set the batch size to 128. The input size of the images are 48*48pixels. The number of classes are 7 as there are 7 emotions in the dataset. We used three different kernel sizes which are (3,3), (5, 5) and (7, 7). We added 3 dropout layers and global average pooling on the model. The learning rate was set to 0.0001 and the optimizer we used is Adam optimizer. We user the softmax activation and ran our tests. After running the algorithm the accuracy we got was 99.16%.

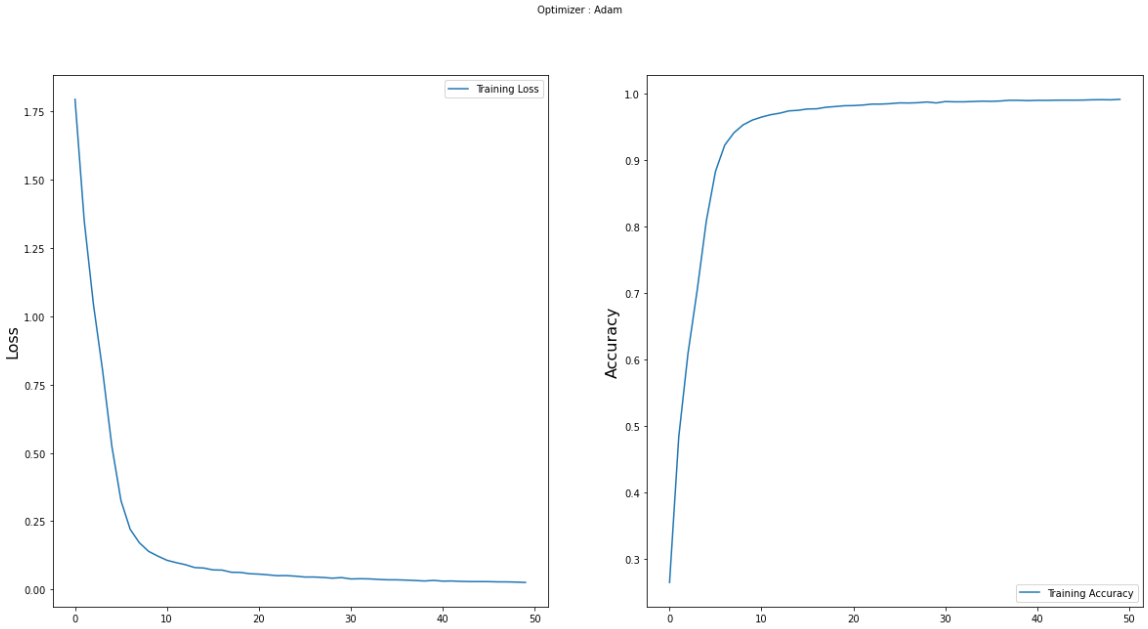

Figure 5: Accuracy and Loss Plot of Proposed Method 1 of Modified FER - 2013 Dataset

b) Proposed Method 2:

We have applied the modified CNN algorithm on the modified FER-2013 dataset. We set the batch size to 128. The input size of the images are 48*48pixels. We used three different kernel sizes which are (3,3), (5, 5) and (7, 7). We added dropout layers, max pooling 2D layers and batch normalization layers with every convolutional layer and added a flatten layer on these layers. The learning rate was set to 0.0001 and the optimizer we used is Adam optimizer. We user the Softmax activation and ran our tests. After running the algorithm the accuracy we got was 89.8%.

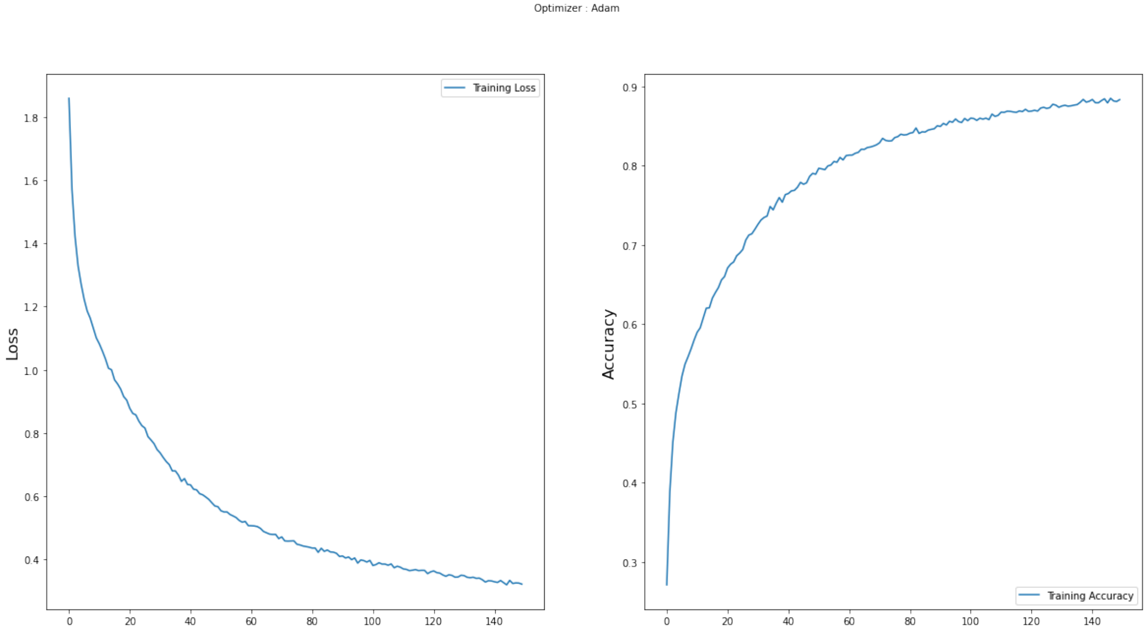

Figure 6: Accuracy and Loss Plot of Proposed Method 2 of Modified FER - 2013 Dataset

Overview:

Methods	Accuracy (**FER-2013**)	Accuracy (**Modified FER-2013**)
Proposed Method 1	**94.71%**	**99.16%**
Proposed Method 2	**88.3%**	**89.8%**

Table 3: Overview of the Accuracies of the Methods

Comparison with Other Approaches:

References	Methods	Dataset	Accuracy
Zhou et al [20].	CNN + MVFE + LightNet	FER2013	68.4%
Ziyang et al [21].	CNN + Music Algorithm	FER2013	62.1%
N. Christou and N. Kanojiya [22].	CNN	FER2013	91%
Hongh Zhang et al [23].	CNN + image edge computing	FER2013 + LFW	88.56%
Pham and Quang [24].	CNN + FPGA	FER2013	66%
Zuheng Ming et al [25].	FaceLiveNet	FER2013	68.60%
Khalid Bhatti et al [26].	Deep features + Extreme learning	FER2013	62.7%
Tanay et al [25].	CNN + LBp + ORB + ConvNet	FER2013	91.01%
Our Method	**Proposed Method 1**	FER2013	**94.71%**
	Proposed Method 2		**88.3%**

Table 4: Comparison with Other Approaches

Chapter 7

Future Work

In this project, we applied three different approaches of CNN on the FER – 2013 dataset. In each approach, we got over 80% accuracy on this dataset. In future, we plan to test these approaches on different datasets and datasets created by us to know how these algorithms holds up against those. Due to the limitation of time and appropriate hardware needed to run these algorithms quicker and efficiently, we could not run more tests as we intended.

Chapter 8

Conclusion

Facial expression recognition is an indispensable technique in the field of machine learning and artificial intelligence. FER – 2013 is an old dataset and it lacks in proper images in some emotion sets. Due to the limited availability of the facial expression datasets, we had to work with this dataset and it is the largest one available. However, we got decent enough accuracies by applying three different approaches of CNN on this dataset. We also tried augmenting the datasets manually and ran the algorithms on that dataset. Results of all the approaches are quite close to one another which means that our method works efficiently and competitively. We also did the real time facial expression recognition test on ourselves and our models were working quite accurately. Hopefully there will be better datasets on facial expression in future and we are looking forward to apply our methods on them.

Chapter 9

References

[1] "What is Computer Vision? | IBM." https://www.ibm.com/topics/computer-vision (accessed Jun. 28, 2022).

[2] A. F. Bulagang, N. G. Weng, J. Mountstephens, and J. Teo, "A review of recent approaches for emotion classification using electrocardiography and electrodermography signals," *Informatics Med. Unlocked*, vol. 20, p. 100363, 2020, doi: 10.1016/j.imu.2020.100363.

[3] "What's a Deep Neural Network? Deep Nets Explained – BMC Software | Blogs." https://www.bmc.com/blogs/deep-neural-network/ (accessed Jun. 28, 2022).

[4] J. Schmidhuber, "Deep Learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015, doi: 10.1016/j.neunet.2014.09.003.

[5] "What is convolutional neural network? - Definition from WhatIs.com." https://www.techtarget.com/searchenterpriseai/definition/convolutional-neural-network (accessed Jun. 28, 2022).

[6] X. Sun, P. Xia, L. Zhang, and L. Shao, "A ROI-guided deep architecture for robust facial expressions recognition," *Inf. Sci. (Ny).*, vol. 522, pp. 35–48, 2020, doi: 10.1016/j.ins.2020.02.047.

[7] J. Li, K. Jin, D. Zhou, N. Kubota, and Z. Ju, "Attention mechanism-based CNN for facial expression recognition," *Neurocomputing*, vol. 411, pp. 340–350, 2020, doi: 10.1016/j.neucom.2020.06.014.

[8] T. H. Vo, G. S. Lee, H. J. Yang, and S. H. Kim, "Pyramid with Super Resolution for In-the-Wild Facial Expression Recognition," *IEEE Access*, vol. 8, pp. 131988–132001, 2020, doi: 10.1109/ACCESS.2020.3010018.

[9] O. Bălan, G. Moise, L. Petrescu, A. Moldoveanu, M. Leordeanu, and F. Moldoveanu, "Emotion classification based on biophysical signals and machine learning techniques," *Symmetry (Basel).*, vol. 12, no. 1, pp. 1–22, 2020, doi: 10.3390/sym12010021.

[10] "What is TensorFlow? The machine learning library explained | InfoWorld." https://www.infoworld.com/article/3278008/what-is-tensorflow-the-machine-learning-library-explained.html (accessed Jun. 28, 2022).

[11] "What is Keras and Why it so Popular in 2021 | Simplilearn." https://www.simplilearn.com/tutorials/deep-learning-tutorial/what-is-keras (accessed Jun. 28, 2022).

[12] P. Mishra, *PyTorch Recipes*. 2019. doi: 10.1007/978-1-4842-4258-2.

[13] S. Gollapudi, "Learn Computer Vision Using OpenCV," *Learn Comput. Vis. Using OpenCV*, pp. 97–117, 2019, doi: 10.1007/978-1-4842-4261-2.

[14] "Different types of Deep Learning models explained." https://roboticsbiz.com/different-types-of-deep-learning-models-explained/ (accessed Jun. 28, 2022).

[15] "Introduction to how CNNs Work. Introduction to how CNNs work | by Simran Bansari | DataDrivenInvestor." https://medium.datadriveninvestor.com/introduction-to-how-cnns-work-77e0e4cde99b (accessed Jun. 28, 2022).

[16] "Basic CNN Architecture: Explaining 5 Layers of Convolutional Neural Network | upGrad blog." https://www.upgrad.com/blog/basic-cnn-architecture/ (accessed Jun. 28, 2022).

[17] "Understand the architecture of CNN | by Kousai Smeda | Towards Data Science." https://towardsdatascience.com/understand-the-architecture-of-cnn-90a25e244c7 (accessed Jun. 28, 2022).

[18] "Convolutional Neural Networks, Explained | by Mayank Mishra | Towards Data Science." https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939 (accessed

Jun. 28, 2022).

[19] "FER-2013 | Kaggle." https://www.kaggle.com/datasets/msambare/fer2013 (accessed Jun. 28, 2022).

[20] J. Zou, X. Cao, S. Zhang, and B. Ge, "A Facial Expression Recognition Based on Improved Convolutional Neural Network," *Proc. 2019 IEEE Int. Conf. Intell. Appl. Syst. Eng. ICIASE 2019*, pp. 301–304, 2019, doi: 10.1109/ICIASE45644.2019.9074074.

[21] Z. Yu, M. Zhao, Y. Wu, P. Liu, and H. Chen, "Research on Automatic Music Recommendation Algorithm Based on Facial Micro-expression Recognition," *Chinese Control Conf. CCC*, vol. 2020-July, pp. 7257–7263, 2020, doi: 10.23919/CCC50068.2020.9189600.

[22] N. Christou and N. Kanojiya, *Human facial expression recognition with convolution neural networks*, vol. 797. Springer Singapore, 2019. doi: 10.1007/978-981-13-1165-9_49.

[23] H. Zhang, A. Jolfaei, and M. Alazab, "A Face Emotion Recognition Method Using Convolutional Neural Network and Image Edge Computing," *IEEE Access*, vol. 7, pp. 159081–159089, 2019, doi: 10.1109/ACCESS.2019.2949741.

[24] Trường đại học bách khoa TP. Hồ Chí Minh, IEEE Vietnam Section, and Institute of Electrical and Electronics Engineers, "2019 International Symposium on Electrical and Electronics Engineering : proceedings : October 10-12, 2019, Ho Chi Minh City, Vietnam," *2019 Int. Symp. Electr. Electron. Eng.*, pp. 1–4, 2019.

[25] T. Debnath, M. M. Reza, A. Rahman, A. Beheshti, S. S. Band, and H. Alinejad-Rokny, "Four-layer ConvNet to facial emotion recognition with minimal epochs and the significance of data diversity," *Scientific Reports*, vol. 12, no. 1. 2022. doi: 10.1038/s41598-022-11173-0.

[26] Y. K. Bhatti, A. Jamil, N. Nida, M. H. Yousaf, S. Viriri, and S. A. Velastin, "Facial Expression Recognition of Instructor Using Deep Features and Extreme Learning Machine," *Comput. Intell. Neurosci.*, vol. 2021, 2021, doi: 10.1155/2021/5570870.