A Bengali Word Identification and Verification Using Machine Learning Approach

BY

Md. Shakibul Hasan
ID: 181-15-10914

Masud Rana
ID: 181-15-11052

Nasim Bin Rahman
ID: 181-15-10530

This Report Presented in Partial Fulfillment of the Requirements for

The Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Abdus Sattar

Assistant Professor

Department of CSE

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2022

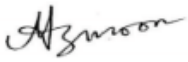
APPROVAL

This Project/internship titled "**Bengali word Identification and verification using machine learning approach"**, submitted by **Md. Shakibul Hasan**, **Md. Nasim Bin Rahman**, **Masud Rana**, ID No: 181-15-10914, 181-15-10530, 181-15-11052 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 4 January,2022.

BOARD OF EXAMINERS

Chairman

Dr. Md. Ismail Jabiullah
Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

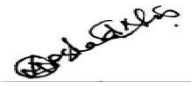


Internal Examiner

Nazmun Nessa Moon (NNM)
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Aniruddha Rakshit (AR)
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



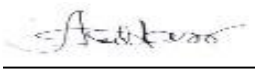
External Examiner

Dr. Md Arshad Ali
Associate Professor
Department of Computer Science and Engineering
Hajee Mohammad Danesh Science and Technology University

DECLARATION

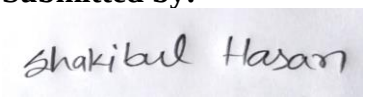
We hereby declare that, this thesis has been done by us under the supervision of **Abdus Sattar, Assistant Professor, Department of CSE** Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for award of any degree or diploma.

Supervised by:



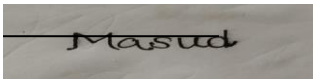
Abdus Sattar
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:



Md. Shakibul Hasan
ID: 181-15-10914
Department of CSE
Daffodil International University

Md. Nasim Bin Rahman
ID: 181-15-10530
Department of CSE
Daffodil International University



Masud Rana
ID: 181-15-11052
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final thesis successfully.

We really grateful and wish our profound our indebtedness to **Abdus Sattar**, Assistant Professor, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "Human Computer Interaction in Education" to carry out this thesis. His endless patience ,scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this thesis.

We would like to express our heartiest gratitude to **Pro. Dr. Touhid Bhuiyan, Professor, and Head,** Department of CSE, for his kind help to finish our thesis and also to other faculty members and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and passion of our parents.

ABSTRACT

The research project is mainly based on image classification. We also showed a path to Optical Character Recognition. If we are to define Optical Character Recognition is basically a technology which can change over a computerized image of content to editable content i.e it can change digital image to editable text. An OCR by using image processing and classification for Bangla dialect is proposed here. The primary step of this project is that to recognize and detect 'text', the usage of image classification and Convolutional Neural Network(CNN) is implemented. Characters are separated from the Corpus and then the most decisive and challenging of the project step is implemented which is the verification of the word or corpus. The Bengali word or corpus is then extracted from the digital image and then it is converted to digital font and then the digital font is compared with our accumulated data set to verify whether the word is absolute or not. To enrich the diversity of Bengali language in the field of technology the project plays a very significant role from its own view.

TABLE OF CONTENTS

CONTENTS

LIST OF FIGURES

FIGURES **PAGE NO.**

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

According to Janendra Mohan Das's lexicon, there are about 150,000 words (approximately) in Bengali Language. However, even after having so many words the recognition of this specific language is comparatively negligible. On the contrary, the development in English Language compared to Bengali is very significant. This research project tries to analyze the approach of OCR (Optical Character Recognition). To detect and recognize 'text', the usage of OCR is implemented. The OCR model method changes digital image to convertible and editable text. The research project focuses specially and solely on the recognition of words i.e identifying it and then extracting the word from the image and then after conversion, verifying the words.

1.2 Motivation

We used to read numerous Bengali books, novels and newspapers. And in often circumstances, we often find that, there are a few troublesome or obscure words in it. Sometimes we ought to know the meaning of that particular word and even justify the correctness of that particular word. For this, we need to inquire about and create a framework to assist us out in certain circumstance. In the modern era, Bangladesh is becoming more and more digitalized and advanced in the field of technology. And moreover, there isn't a lot much work related with Bengali Language and Bangla word and corpus. Furthermore, there has never been a well-organized database for Bengali Transcribed Digits that may be used by the general public. So, for this very reason, we are motivated and enthusiast in working with this project.

1.3 Problem Definition

AI may be an exceptionally imperative term within the field of ICT in this advanced age. Application of AI will offer assistance to develop our Bengali dialect. To grant a legitimate arrangement it's vital to discover out the problems and related necessities in this field. Fathoming

issues with Artificial Intelligence in a competitive way has long been truant in Bangladesh and Bengali-speaking community.AI Computer Vision Challenge. The challenge saw both local and international teams participating with unprecedented efforts. To create and provide easy, simple, acceptable and satisfactory outcomes AI is implemented to relinquish every other aspects.

1.4 Research Questions

The most questions those are centered in this proposal are given below:

- What is the current situation of Optical Character Recognition (OCR) in Bangla Language?
- What are the limitations of working with Bangla OCR?
- How to extract Bangla Words and solve the limitations?

1.5 Research Methodology

The experiment data, data processing model architecture, training model, and outcome performance are all covered in this section of our research report. The proposed paradigm will be demonstrated in action at the end of this chapter.

1.6 Research Objectives

To perform this project more efficiently and smoothly we have some objectives to be achieved. Those are:

- To identify Bengali words of the selective font.
- To detect Bangla word.
- To verify whether the word is correct or not.
- Extract Bangla words or corpus from the digital image

1.7 Research Layout

Chapter 1: will cover the following topics: introduction, motivation, problem definition, research question, research methodology, and our project's predicted conclusion..

Chapter 2: will explain the background of this research, as well as related work and present state from the perspective of Bangladesh.

Chapter 3: We discuss about the methodology of our project in this chapter. The model structure and dataset will be described here.

Chapter 4: In this chapter the train, test and validation of the model will be described.

Chapter 5: The result and the performance will be described here so that we can assume the efficiency of this thesis based project.

Chapter 6: It describes future work and the conclusion of this research.

Chapter 7: states all the references used to accomplish this project.

CHAPTER 2

BACKGROUND

2.1 Introduction

In In the perspective of Bangladesh, there has been absolutely no related work or practices where Bangla word is extracted to verify its exactitude using Artificial Intelligence.

2.2 Related Works

In this research topic, there are numerous ways for correcting the real-word blunder. Andrew, Golding, and Yves presented an approach that could be a hybrid of Trigram and Bayes' hypothesis. The Bayes hypothesis is based on word features, while the trigram hypothesis is based on discourse elements. The POS labeling trigram is used for distinct confounding words, while Bayes hypothesis is used for the same POS labeling of confounded terms. The suitable term is extracted using the Tribayes approach and the likelihood calculation. [1]

POS tagging, Covered Markov Show, and Trigram were used by Davide Fossati. To label the phrase, POS labeling is used, and labeled sentences with unclear words are compared to well-labeled tags to determine the difference. Contrasts are a sign that something has to be adjusted, and trigrams are used to fix the error. Mays has provided a realistic technique for dealing with the real-world error. [2]

Sumit and Swadha offer a demonstration in which the highlights are extracted using the trigram and Bayes technique to deal with the real-world mistake.[3]

Pratip and Bidyut proposed a demonstration that uses Levenshtein Remove to create perplexity and bigram and trigram to determine likelihood. It is unusual to come across a work in Bangla that deals with a real-word gaffe.[4]

Akas Roy has presented a model that uses n-grams to check the correctness of a word in a phrase. He used n-grams of characters in his approach, and there is no substantial evidence that they can compensate for the real-word mistake.[7]

Following that, there isn't any investigation into recognizing and correcting real-word blunders in Bangla dialect. As a result, we accept that our approach would provide a new research direction and broaden the scope of Bangla real-word mistake detection and correction.

2.3 Bangladesh Perspective

If the scenario or perspective of Bangladesh is made into the point of consideration, there isn't simply any related work where a project proposes Bangla words or corpus to be primarily identified from a digital image and then converted to editable text and then extracted to verify the word's absoluteness. With the advancement of technology, Bangladesh is also progressing in her own way day after day. And in this modern era, the recognition of Bengali Language is foremostly ineluctable since there are so many native speakers of Bengali Language who lacks the apprehension of Bangla words and corpuses. There are more than 150,000(approximately) words in Bengali Language. So, in a mainstream consideration it is absolutely impossible for an individual to remember all these huge collection of words and corpuses and also to know their meanings is on another level of toughness. On a regular basis we read Bangla novels, newspapers and so on... and for the ones having inadequate apprehension about Bengali Language like the foreigners, the proposed project is a revolutionary one from its own prospect.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

To achieve the highest accuracy we used 3 models of CNN model which means the convolutional Neural Network. With VGG16 model we achieved the best accuracy.

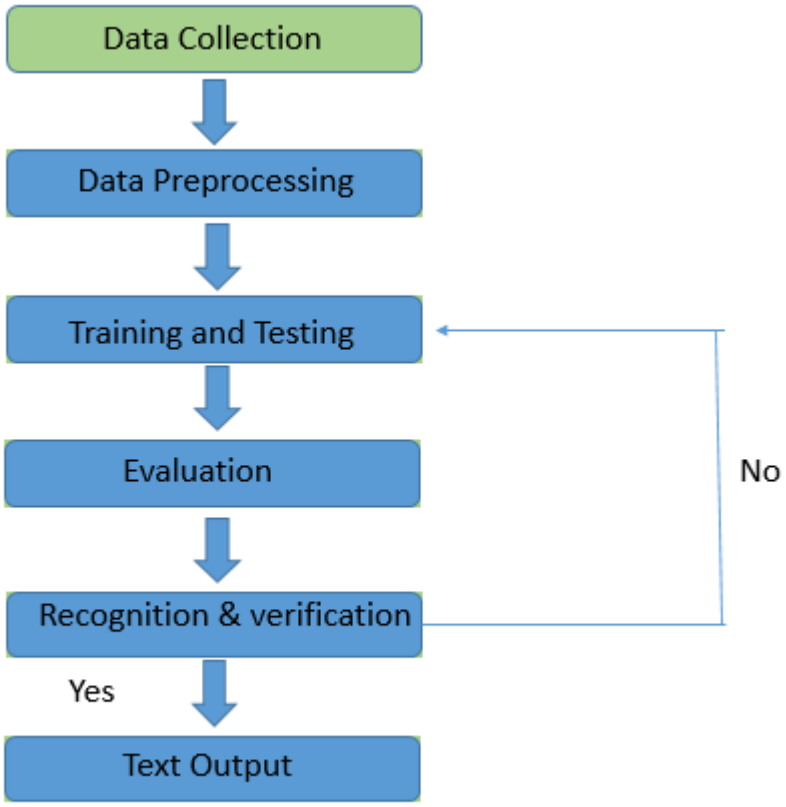


Figure 3.1.1 Steps of Model Methodology

3.2 Experiment Data Set

To perform this project we had to collect images of Bengali words enormously. We have collected nearly 2000 images of 200 Bengali words. There is 10 images for each word. The collection was done randomly from books, poster, signboards, leaflets, Headings etc. by smart mobile camera. Those images were in random shapes and size. In the research, between these 2000 images we picked 1700 for training and 300 for testing.

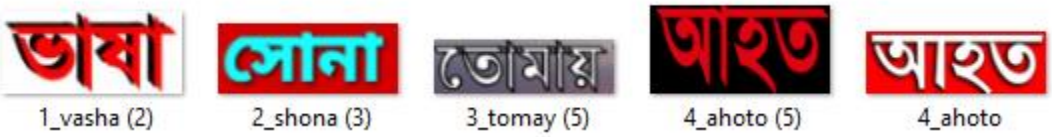


Figure 3.2.1 Collected Dataset

3.3 Data Pre-Processing

Though the data were collected all by ourselves but we need to pre-process the data to perform our research. The images were in various shape and size and their RGB were also different. Firstly we have to clean the data. For his we have to resize all our images into 100x100 shape. Then we converted our image in grayscale for better performance.

Then we labeled our images with their perspective names and save it in a .txt file. Later we converted the .txt file into csv file as well. Labelling was an important part of data pre-processing.

We have used openCV to perform pre-processing of the data. Then our data is ready to train. We created a csv file to validate and verify the words whether it is correct or not. We also provided the corresponding English meaning through this csv file.

3.4 Architecture of the Model

To identify the words from the image, we created three CNN models. There are two standard CNN models and a third that is created using the VGG16 model. VGG16 is a convolutional neural network with 16 convolution layers that is a version of the VGG model. [8]

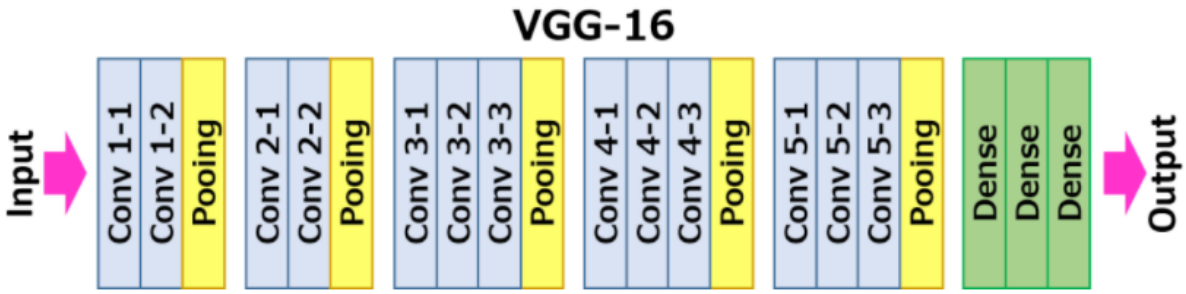
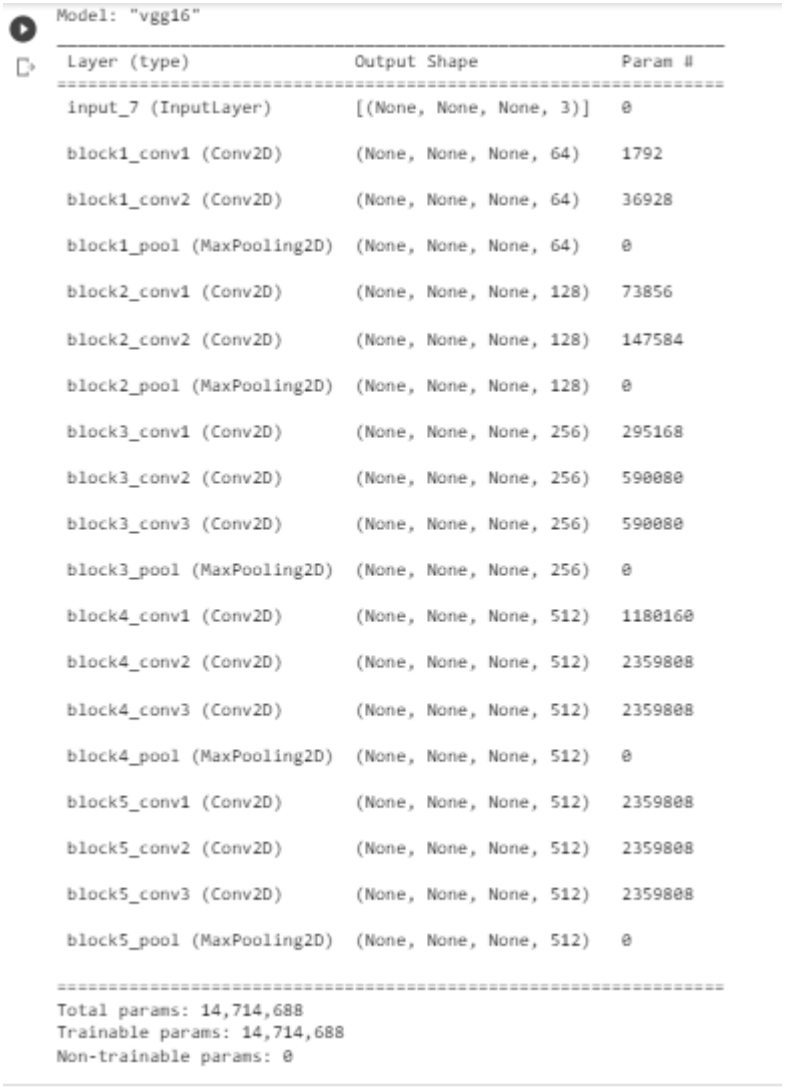


Fig 3.4.1: VGG model Architecture

VGGNet-16 has 16 convolutional layers and a fairly homogeneous architecture, making it quite appealing. It just contains 3x3 convolutions, but it has a lot of filters. The input to the cov1 layer is a fixed-size 100 X 100 RGB image with a kernel size of 3. The image is passed through a series of convolutional (conv.) layers, with the channels being used with a very small open field: 33 (which is the smallest estimate to capture the concepts of left/right, up/down, and center).

It also uses 11 convolution channels in one of the setups, which can be considered as a direct alteration of the input channels (taken after by non-linearity). The spatial cushioning of convolution layer input is such that the spatial determination is protected after convolution, i.e. the cushioning is 1-pixel for 33% of the convolution layers. Five max-pooling layers, which follow a couple of the conv. layers, do spatial pooling (not all the conv. layers are taken after by max-pooling). With walk 2, max-pooling is done over a 22-pixel window.

Finally, Fully-Connected layers are based on a stack of convolutional layers, each of which comprises a distinct depth of totally unique models. Two thick layers are used to depict matrix-vector multiplication. In the first dense layer, there are 256 hidden units, and activation value causes a 50% dropout rate. In order to tackle the issue, the over fitting drop out approach is used. The density in the second dense layer should be 200 without drop out to normalize the model, and the activation should be sigmoid. The model is built in this process.



```
Model: "vgg16"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_7 (InputLayer)        [(None, None, None, 3)]   0

 block1_conv1 (Conv2D)       (None, None, None, 64)    1792

 block1_conv2 (Conv2D)       (None, None, None, 64)    36928

 block1_pool (MaxPooling2D)  (None, None, None, 64)    0

 block2_conv1 (Conv2D)       (None, None, None, 128)   73856

 block2_conv2 (Conv2D)       (None, None, None, 128)   147584

 block2_pool (MaxPooling2D)  (None, None, None, 128)   0

 block3_conv1 (Conv2D)       (None, None, None, 256)   295168

 block3_conv2 (Conv2D)       (None, None, None, 256)   590080

 block3_conv3 (Conv2D)       (None, None, None, 256)   590080

 block3_pool (MaxPooling2D)  (None, None, None, 256)   0

 block4_conv1 (Conv2D)       (None, None, None, 512)   1180160

 block4_conv2 (Conv2D)       (None, None, None, 512)   2359808

 block4_conv3 (Conv2D)       (None, None, None, 512)   2359808

 block4_pool (MaxPooling2D)  (None, None, None, 512)   0

 block5_conv1 (Conv2D)       (None, None, None, 512)   2359808

 block5_conv2 (Conv2D)       (None, None, None, 512)   2359808

 block5_conv3 (Conv2D)       (None, None, None, 512)   2359808

 block5_pool (MaxPooling2D)  (None, None, None, 512)   0

=================================================================
Total params: 14,714,688
Trainable params: 14,714,688
Non-trainable params: 0
```

Fig 3.4.2: Proposed model architecture

3.5 Learning rate and Optimizer of the model

The learning rate determines how much the weights are updated with respect to the estimated error. Choosing too small of a value and the model will train forever. So for this model we have chosen Adam Optimizer .0001 for update network weights and faster result.

The size of the model has no bearing on the learning rate. It works equally well for both 1x and 16x model sizes. The learning rate is the model's most important hyperparameter. The following is the procedure for calculating each update with learning rate decay:

$$LR = initial\_lrate \times (1 / (1 + decay \times iteration))$$

In this case, a learning rate decrease of 0.00001 was employed, and the learning rate was turned low to reduce the error. Keras-supported call back function was used to call the learning rate in order to fine-tune the model weight and reduce the learning rate when the model stopped progressing.

3.6 Data Augmentation

For strengthening our deep learning VGG16 CNN model, we updated our train and test data as well as the width, height shift range of 0.1. The following is an example of a data augmentation algorithm: $y = \text{argmax}_j \max(\{pij\} 1 \leq i \leq r)$

To use the sample's prediction, the classifier is most certain about:

$$Averaging\ all\ predictions: y = \frac{1}{r} \sum_{i=1}^{r} pij$$

In this work, data augmentation has played a key role in achieving the best recent results in a variety of text identification tasks, and it is utilized to expand the augmentation feature. With rescaling the image, we apply horizontal flipping, brightness changes, and random RGB color. We must multiply the data before performing any other processing, and the outcome is referred to as rescaling value. The model's color mode is RGB, and the RGB coefficients vary from 0 to 255, however in this case, such values can be extremely high. As a result, we use a horizontal flip and rescale the image (rescale=1./255).

3.7 Training the model

The VGG16 convolution layer and three fully connected layers are used in our experiment. By using this algorithm when we trained our model, it achieved the better performance with more accuracy. We use the batch size 16 with the number of epochs 50 and the dropout was 50%. This model achieves a fantabulous unimaginable accuracy of 100% with loss 1.749 and test accuracy

is 93.46%. Both the train and test images have a goal size of 100 x 100 pixels. The VGG16 CNN model used here can detect the word by more precisely classifying it..

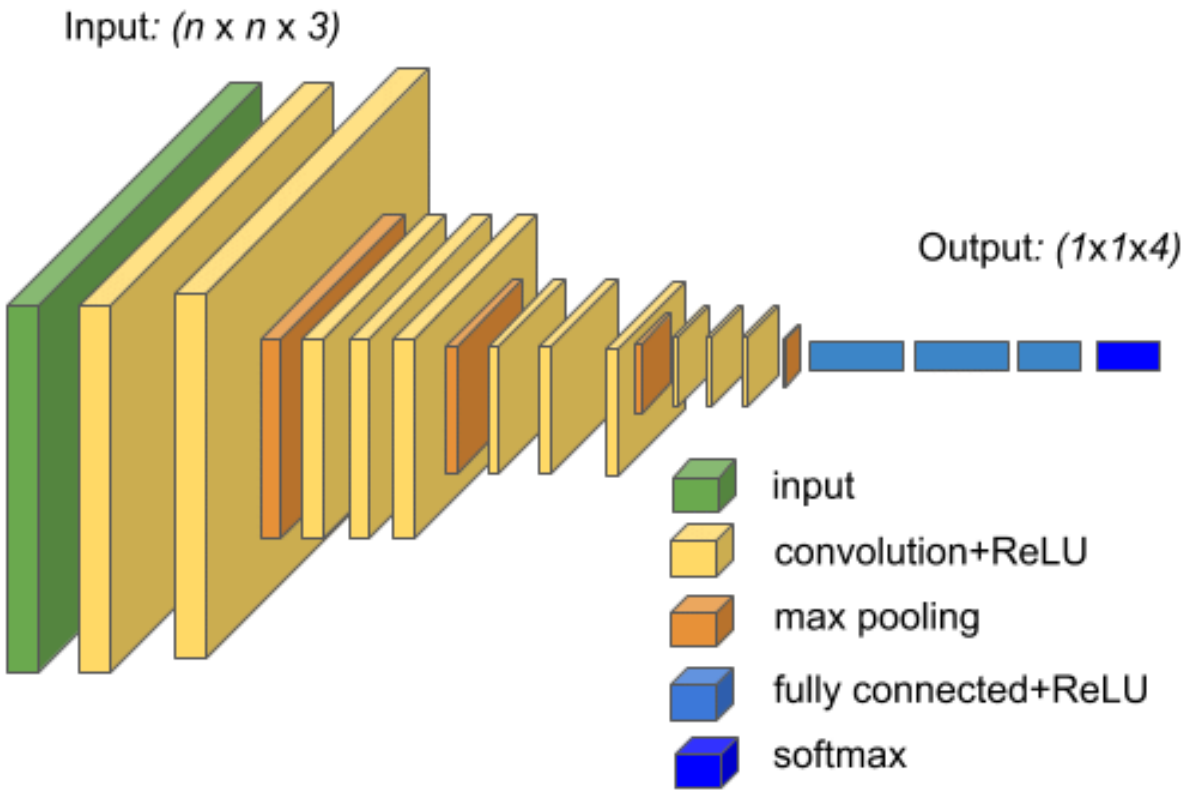


Fig 3.7.1: Training CNN model Using VGG16 algorithm

CHAPTER 4

PERFORMANCE OF THE PROPOSED MODEL

4.1 Training, Testing and the Validation of the model

Our datasets are divided into training and test data, with a total of roughly 2000 photos. We used around 85% of the data (1700 total) for training and 15% for testing the suggested model (300 total). The image processing deep learning frameworks TensorFlow and Keras were used as the backend. For both training and testing the datasets, we employed categorical mode with RGB color mode and a batch size of 16. The proposed model provides excellent performance and accuracy. The test and predicted data are displayed below after training:

Fig 4.1.1: Result of prediction

4.2 Model efficiency

In this model we run 50 epochs. After completing 50 epochs, 98.5% of the accuracy for the training datasets and about 97% of accuracy for the testing datasets was achieved. We got a great successive

accuracy by this approach. So this CNN model can be used to recognize and detect Bengali word accurately. The efficiency of this model is visualizing here:

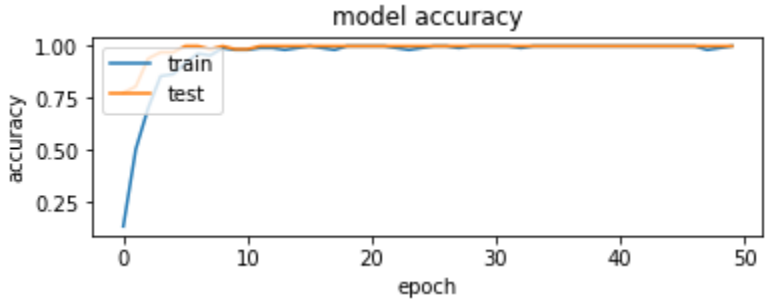


Fig 4.2.1: Model Accuracy

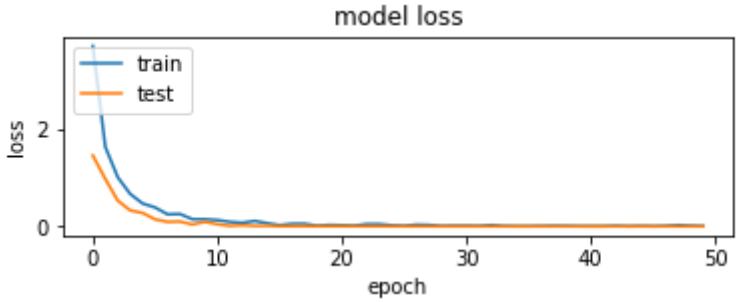


Fig 4.2.2: Model Loss

The accuracy is calculated using the following formula,

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

The confusion matrix is frequently used to describe the classification model's performance. For each type of dataset, the False positive(FP) rate is where an observation is negative and negative values are predicted to be positive, the True Positive(TP) rate is where an observation is positive and is predicted to be positive, the True Negative(TN) rate is where predicted values are correctly predicted as an actual negative, and the False Negative(FN) rate is where positive values are predicted as negatively.

CHAPTER 5

RESULT COMPARISON AND ANALYSIS

In this section we will compare our accuracy of the model with other previous works and model. Many researcher already worked on it and we made 2 more models for this work. The following table will show the comparison.

Algorithm	Accuracy
CNN	70%
Messalodi	54% precision, 91.2% recall
Otsu	77.6% precision, 67.36% recall
Gllavata	84.94% precision 85.94% recall
Kim	80%
Efficient Text Extraction Algorithm Using Color Clustering for Language Translation in Mobile Phone[9]	89.08% precision, 80.15% recall
Proposed Model	98%

Table 5.1: Comparison with previous works

From this analysis part we can see that our model is more effective and accurate. Most of the algorithm's accuracy are nearly 80-90%. But VGG16 algorithm worked more effectively. This is

totally a new approach for text detection and verification. But this has shown an enormous result. This will effectively work to detect Bengali words and verify them.

CHAPTER 6

CONCLUSION AND FUTURE WORK

This The project mainly emphasizes on Image classification and CNN(Convolutional Neural Network) which mainly is the perfect way to implement this kind of recognized work. The models gained an accuracy of when we were able to input the dataset perfectly. This sort of work can bring a great assessment in Bengali Language division for Bangladesh. This work can alter the circumstance of the rancher in Bangladesh.

Anticipating to create and provide something big in the near future to provide more accuracy for the models. Since the models were trained with not a handful of datasets so the accuracy of the project has been . In the future by increasing the number of datasets to huge, it is expected to provide the aftermath more precisely and accurately.

REFERENCE

[1] Andrew Golding, Yves Schabes "Combining Trigram-Based and Feature-Based Methods for Context-Sensitive Spelling Correction", 1996.

[2] avide Fossati, Barbara Di Eugenio, " I saw TREE trees in the park: How to correct real-word spelling mistakes ", 1987.

[3] Sumit sharma, Swadha Gupta, "A Correction Model for Real-word Errors", 2015.

[4] Pratip Samanta, Bidyut B. Chaudhuri ,"A simple real-word error detection and correction using local word bigram and trigram" , 2013

[5] aoguang Shi, Xiang Bai and Cong Yao, " An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition ", 2015.

[6] Joan-Andreu Sánchez, U. Pal "Handwritten Text Recognition for Bengali ", 15th International Conference on Frontiers in Handwriting Recognition (ICFHR) , 2016.

[7] Akash Roy, "AKHCRNet: Bengali Handwritten Character Recognition Using Deep Learning", 2020.

[8] Karen Simonyan & Andrew Zisserman, "VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION", ICLR 2015

[9] Adrián Canedo-Rodríguez1 , Jung Hyoun Kim2 , Soo-Hyung Kim3 , John Kelly2 , Jung Hee Kim2 , Sun Yi2 , Sai Kiran Veeramachaneni2 , Yolanda Blanco-Fernández1," Efficient Text Extraction Algorithm Using Color Clustering for Language Translation in Mobile Phone ", journal of Signal and Information Processing, 2012, 3, 228-237

[10] H. Li and D. Doermann, "Text Enhancement in Digital Video Using Multiple Frame Integration," Proceedings of ACM International Conference on Multimedia, 1999, pp. 19-22..

[11] W. Niblack, "An Introduction to Image Processing," Prentice-Hall, Upper Saddle River, 1986, pp. 115-116

[12] I.-J. Kim, "Multi-Window Binarization of Camera Image for Document Recognition," Ninth International Workshop on Frontiers in Handwriting Recognition, Inzisoft Co. Ltd., 26-29 October 2004, pp. 323-327

[13] R. Lienhart and A. Wernicke, "Localising and Segmenting Text in Images, Videos and Web Pages," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 12, No. 4, 2002, pp. 256-268. doi:10.1109/76.999203

[14] D. Comaniciu, "Nonparametric Robust Methods for Computer Vision," Ph.D. Thesis, Rutgers University, Newark, 2000

[15] T. Perroud, K. Sobottka, H. Bunke and L. Hall, "Text Extraction from Colour Documents—Clustering Approaches in Three and Four Dimensions," Proceedings of International Conference on Document Analysis and Recognition, 10-13 September 2001, pp. 937-941

[16] S. Messelodi and C. M. Modena, "Automatic Identification and Skew Estimation of Text Lines in Real Scene Images," Pattern Recognition, Vol. 32, No. 5, 1992, pp. 791-810.

A Bengali Word Identification and Verification Using Machine Learning Approach

ORIGINALITY REPORT

14%	5%	8%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Md. Imran Hossain, Bidhan Paul, Abdus Sattar, Md. Mushfiqul Islam. "A Convolutional Neural Network Approach to Recognize the Insect: A Perspective in Bangladesh", 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), 2019 Publication	6%
2	Submitted to Daffodil International University Student Paper	3%
3	Submitted to University of Lincoln Student Paper	1%
4	Submitted to Yakın Doğu Üniversitesi Student Paper	1%
5	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
6	N. Jayanthi, S. Indu, Snigdhaa Hasija, Prateek Tripathi. "Chapter 62 Digitization of Ancient Manuscripts and Inscriptions - A Review",	<1%