Tamil Handwritten Character Recognition Using Artificial Neural Network

Ms.G.Thilagavathi, Ms.G.Lavanya, Dr.N.K.Karthikeyan

Abstract: Conversion of Handwritten English documents into editable digital documents has been done in many parts of the world. Though this conversion is being expanded to many other languages such as Spanish, French, Greek, Indian Languages are still untouched. Our Proposed System aims to convert handwritten characters of South Indian Languages (mainly Tamil) to digital characters which can help converting those handwritten documents into digital ones. It uses Artificial Neural Networks (ANN) with Stochastic Gradient Learning Algorithm with Back-Propagation as learning methodology. Neural Network Model is a learning model which mimics human art of learning using primitive components named neurons. The Neural Network, that this system uses are made up of sigmoid neurons. Previous implementation of such conversion systems has shown accuracy over 90%. It can be used in Banking Sectors, Answer Script Evaluation Systems for Subjective Answer Evaluation and areas that involve intense Human-Computer Interaction Areas.

Keywords: Artificial Neural Networks, Back-Propagation, Neurons, Stochastic Gradient Learning

1. INTRODUCTION

1.1 Character Recognition

Character Recognition is the mechanical or electronic conversion of typed, handwritten or printed characters form a scanned document, a photo of a document, a scene-photo, or a subtitle text. It is widely used of information entry from printed data records, invoices, bank statements, computerized receipts, business cards. It is common method of digitizing printed texts so that they can be electronically edited searched, stored more compactly, displayed on-line and used in machine process such as cognitive computing, machine translation, text-to-speech, key data and text mining. OCR is a field of research in pattern recognition, artificial intelligence, and computer vision. Character Recognition is mainly of two types Optical Character Recognition for printed characters and Handwritten Character Recognition. Handwriting recognition principally entails optical character recognition. However, a complete handwriting recognition system also handles formatting, performs correct segmentation into characters and finds the most plausible words.

1.2 Tamil Language

Tamil Language is a Dravidian Language being used mostly in Tamil Nadu, India. It behaves as a root language for other Dravidian languages. It has a total of 247 characters in total in which 12 are vowels, 18 are consonants, 1 is a special symbol, others are compound ones from combination of vowels and consonants.

- *Ms.G.Thilagavathi, Ms.G.Lavanya, Dr.N.K.Karthikeyan*
- *Ms.G.Thilagavathi, Assistant Profesor,Coimbatore Institute of Technology,Coimbatore.Email:thilaga.apr@gmail.com*
- *Ms.G.Lavanya, Assistant Professor,Sri Ramakrishna College ofEngineering,Coimbatore.Email:lavanyagangadharan@gmail.com*
- *Dr.N.K.Karthikeyan, Professor and Head of Information Technology,Coimbatore Institute of Technology,Coimbatore.*
- *Email: karthikeyan.nk@cit.edu.in*

Tamil Language is known for wide use of symbols due to which pattern matching with Tamil has proved as inefficient techniques for production. It also comprises of separate symbols for numerals but they are not used officially, even in Tamil Nadu, India.

1.3 Artificial Neural Networks

Artificial Neural Networks are computing systems vaguely inspired by the biological neural networks that constitute animal brains. Such systems learn tasks by considering examples, generally without task-specific programming. An ANN is based on collection of connected units or nodes called artificial neurons, each connection between artificial neurons can transmit a signal from one to another. The artificial neuron that receives the signal can process it and then signal artificial neurons connected to it. This is a simple feed forward flow of signal. In order for a network to learn a function, it has to be trained by a learning algorithm. The commonly used such algorithm is Back propagation. In this algorithm, backward flow of information is made so that each neuron can optimize its parameters based on a loss function and the desired output and the current output. There are many types of networks including Dense Networks and Convolutional Networks. In Dense Networks, an input in a layer can trigger all the neurons in the next layer. In Convolutional Layers, only a portion of neurons in an image can trigger a single neuron. Pooling layers are also used for the same purpose as convolutional ones. There are many models have come up serving different set of problems. Some of them are Recurrent Networks, Long Short Term Memory Networks, Generative Adversarial Networks, Auto Encoders, Variational Auto encoders.The original goal of the ANN approach was to solve problems in the same way that a human brain would. However, over time, attention focused on matching specific tasks, leading to deviations from biology. ANNs have been used on a variety of tasks, including computer vision, speech recognition, machine translation, social network filtering, playing board and video games and medical diagnosis. Because of their ability to reproduce and model nonlinear processes, ANNs have found many applications in a wide range of disciplines. Application areas include system identification and control (vehicle control, trajectory prediction, process control, natural resource management), quantum chemistry, game-playing and decision making (backgammon, chess, poker), pattern recognition (radar systems, face identification, signal classification, object recognition and

more), sequence recognition (gesture, speech, handwritten and printed text recognition), medical diagnosis, finance (e.g. automated trading systems), data mining, visualization, machine translation, social network filteringand e-mail spam filtering.

2. LITERATURE SURVEY

2.1Optical Character Recognition Technique Algorithms - N. Venkata Rao and Dr. A.S.C.S.Sastry - 2016 Journal of Theoretical and Applied Information Technology

This paper explains various Optical Character Recognition techniques for both printed and handwritten characters. This paper details about the advantage of Artificial Neural Networks over other methods. This method proposes a 3 layered an architecture in which neurons have only transfer function but not activation function. Activation function is applied at the end of the network before the output layer. This is to reduce the computation time caused by differentiation. Experiment results are based on the 3 class network. They are able to achieve an accuracy of 94 % for 3 characters.

Advantages: The rigid architecture makes good prediction with less number of classes.

Disadvantages: The Model performs good only for 3 characters.

2.2Writer Recognition for South Indian Languages using statistical Feature Extraction and Distance Classifier - Aravinda C.V and Dr. Prakash H.N - International Journal of Natural Language Computing 2016

This paper uses Handwritten Character Recognition techniques to identify the author of a handwritten text. It uses statistical feature extraction and Similarity Edge Detection Techniques. It uses Manhattan distance and Euclidian distance classifiers. Features proposed by it include height from a baseline to upper edge and bottom edge, end points and Loop. It uses angle strings as cost function. Advantages: Due to efficient preprocessing of data, the features make prediction easy through simple methods. Disadvantages: The project proposes ideas for finding features to identify author not the character. So the proposed model cannot be used for character recognition.

2.3 Handwritten Tamil Character Recognition using Artificial Neural Networks - P Banumathi and Dr. G. M. Nasira - 2011 IEEE

The paper details about using Artificial Neural Networks for Tamil handwritten Character Recognition. This paper proposes architecture based on Kohonenself-Organizing Map(SOM).It uses a 4 * 4 Kohenen map. The Kohen Map is a Self-Organizing Map which updates the elements based on a Kohenen Rule. This Papers results are based on the classifications for only 8 classes. The experiments gave an accuracy of over 80% for 5 different hand writings. Advantages: Since it uses a Simple model. Hence it is easier to implement and training. It trains faster due to small number of variables. Disadvantages: The system is only tested for 8 classes. Hence this architecture is not scalable to 247 characters.

2.4 Neural Network based offline Tamil Handwritten Character Recognition System - J. Sujitha, M.E and N. Ramraj, M.E., Ph.d., - 2007 IEEE

The paper talk about the Tamil Handwritten Character Recognition using Neural Networks along with Fourier Descriptors. This paper applies different preprocessing including smoothing, Otsu's histogram thresholding , skeletonization by Hilditch algorithm. Features are then extracted. This paper uses fourier descriptors from the closed boundary trace due to statistical invariance. The experiments were done on a 18 *18 image and output classes were 5. This system can produce accuracy of 97% with 15 hidden layers. Advantages: Since the number of hidden layers is high it is a robust learner. It can handle large variations. Disadvantages: Large number of hidden layers makes learning so slow. It is not feasible to use in production systems.

2.5 An improved Online Tamil Character Recognition using Neural Networks – Ishwarya, M. V, R. Jagdesh Kannan - 2010 IEEE

The paper proposes a Neural Network based on the Convolutional Neural Networks. Paper compares the CNN approach with Principal Component Analysis using a confusion matrix. Proposed Model consists of a 6 layered architecture for input image size 28 * 28. Feature vectors from Normalized images are transformed into 8 dimensional space using 2DCPA. And then a prediction is done over a network. The accuracies obtained using such a network are 95 on average. Advantages: Usage of convolutional promotes high localized accuracy. Spatial relationships between symbols in a character are learnt fast. Disadvantages: The proposed model uses complex preprocessing techniques which are resource intensive.

3. SYSTEM ARCHITECTURE

System for Tamil Handwritten Character Recognition consists of a command line utility which can train the network and allows the user to predict the character in the image. System also consists of several pre-processing modules which are used for pre-processing of training and testing data. The user data will be pre-processed on the fly before it gets feeded into the system. The System is designed to learn vowels of Tamil Language. There are 12 vowels in Tamil Language.

.

3.1 ARCHITECTURE

During preprocessing the image size is reduced to a standard size of 100 * 100 and RGB values are converted into grayscale values.

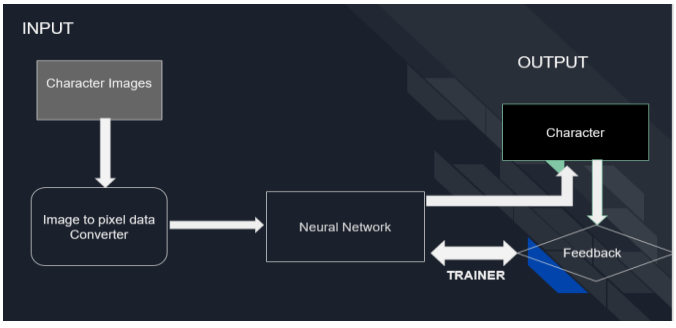


Fig 3.1 *System Architecture*

Image to Pixel Converter - Pre-Processing

It converts the character image which is an input into pixel data for feeding the data into network. Noise in the image should be

reduced. It takes input images of size 100*100 and output is pixel data array in the form of Numpy arrays or Tensors of size 10000.

Network Trainer
Trains the network with the pixel data. Ordinary Sigmoid neural network with 10000 input neurons and 12 output neurons and 4 hidden layers. Input as pixel arrays and their mapped character arrays and it produces no output visibly but optimal network which has been built for this design will be trained.

Character Predictor
Predicts the character in the image to the appropriate character .Inputs the input character image(s) pixel data to Trained Network and gets the output character arrays numbers from 1 to 12 indicating corresponding Neuron has high probability of correct prediction.

4. SYSTEM DESIGN AND IMPLEMENTATION

4.1 DATASET

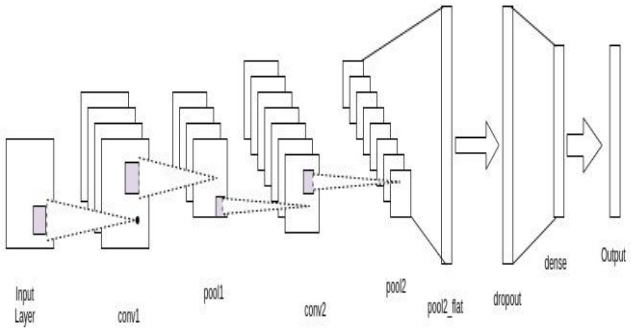


Fig 4.1 *Dataset Distribution*

The Dataset comes from a worldwide competition on handwritten character recognition IFWHR which comprises of images of Tamil Characters written by students on tablet devices. All the images written by a single user are kept in a separate directory which is combined then into a single directory. Labels of characters are converted into numerical values, based on their order in the Tamil Symbol Table. They are used as such for ease of manipulation.
Following is a summary of information about the data.
Total 82934
Number of candidates: 169
Corrupt: 5
Usable: 82929
Number of characters: 156
Min Char Label: 0

Max char Label: 155
Following is a distribution of number of images for each character.

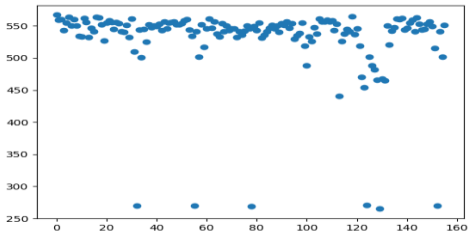


Fig4.2 *Preprocessing Module*

4.1.1 TRAIN DATASET
The Training Dataset consists of a data from all the candidates handwriting but training is done only using a single candidate handwriting. Each Candidate has 10 images for each character on average.

4.1.2 TEST DATASET
The Test dataset consists of randomly selected image for each character. As Current system learns only 12 characters, there are 12 random images in the test dataset.

4.2 DATA PREPROCESSING
The Dataset consists of both png and tiff files and of various sizes. First data from each image is converted into image of size 100*100. A png file has no alpha channel while tiff files does. So alpha channel (if any) in the dataset is removed. And then RGB image is converted into grayscale images. Following is an input and output of pre-processing module.

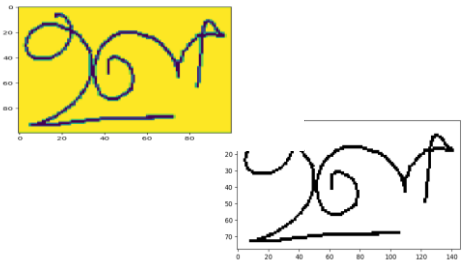


Fig 4.3 *Neural Network Architecture*

4.3 NEURAL NETWORK ARCHITECTURE
Following is a diagrammatic representation of Neural Network used for classification of handwritten characters.

4.3.1 INPUT LAYER
Input layer is the start of the network. It is the layer which receives image as in input. It has no neurons but acts as a virtual neural layer with activations as pixels of characters. This Layer of shape (100,100,1). Any image (from user input) which is of other sizes must be reshaped into this 3D shape. But Dataset (used while training and testing) which consists of large number of images must be reshaped into (x, 100,100,1) where x is number of images in the dataset.

4.3.2 CONV1 and POOL1
The second layer of network is a convolutional layer with number of filters or feature maps as 4 and size of kernel is 5 * 5 with padding as “same”. kernel is the region of image which can collectively trigger a single neuron in the next layer. Padding(“SAME”) fills the zeros in the corners of the image so that size of image previous layer and convolutional layer to be the same. POOL1 is a max Pooling layer which has a pool size of 2*2 with strides as 5. pooling is done to reduce the pick a summary of a region in the previous layers. Pool size denotes

the size of region to be summarized and strides denote the pixel distance between two distinct regions.
The shapes of resultant images after theses layers are as follows
CONV1: 100 * 100 *4
POOL1: 20 * 20 * 4

4.3.3 CONV2 and POOL2

The Third layer of network is a convolutional layer with number of filters or feature maps as 8 and size of kernel is 5 * 5 with padding as “same”. kernel is the region of image which can collectively trigger a single neuron in the next layer. Padding(“SAME”) fills the zeros in the corners of the image so that size of image previous layer and convolutional layer to be the same. POOL2 is a max Pooling layer which has a pool size of 2*2 with strides as 2. pooling is done to reduce the pick a summary of a region in the previous layers. Pool size denotes the size of region to be summarized and strides denote the pixel distance between two distinct regions.
The shapes of resultant images after these layers are as follows
CONV2: 20 * 20 * 8
POOL2: 10 * 10 * 8

4.3.4 POOL2_FLAT and DROPOUT

Image after POOL2 is converted into a flat image combining all the features for further analysis by the upcoming layers. Hence the shape of the image after POOL2_FLAT is (x,800) where x is 1(for a single image) or number of images in the dataset. The flat data from POOL2_FLAT is forced to drop some of its inputs to 0. So image will appear as one with a feature missing. This is made to make sure that neural network can correctly predict the character even if there is a feature missing or noise had corrupt the signal. The dropout rate is 0.2. Resultant image size will remain as the same for POOL2_flat but with a 20% of pixels with incorrect data.

4.3.5 DENSE and OUTPUT

Last two Layers are Dense and Output. Dense Layer is used for analysis based on all the information about the features. So each neuron should be trigger all the neurons. Dense layer has 12 neurons receiving inputs from 800 neurons from previous layer. So this produces an array of 12 values. Each of these values is a real number. This value cannot be used such as those values do not provide any useful information to the user. So this array is transformed into a softmax probability distribution. The index in which the value is maximum in the array is likely to be the output label. Shape of Output Layer is (x,12) where x can be 1 for a single image and number of images in the dataset.

5. CONCLUSION

Using convolutional Neural Networks to recognize handwritten characters can improve efficiency and speed due to way they learn. The proposed system is successful in predicting a character which has been seen by the model during training. And also speed of learning was faster compared to other models. So using such Convolutional models can enable any user to translate his/her handwritten documents into digitized documents. Cloud Vendors can provide service to their client for document conversion uses. It will enable Intelligent machines to communicate without any Language Barrier.

6. SCOPE FOR FUTURE

In this digital age, the variations in hand writing is going to vary over time. So the Model can be scaled to support various other symbols like consonants and numerals. Even the project can be extended for other Languages that share same similar set of feature symbols. Further Neuro Evolution technique can be used to learn the model itself without user having to design it.

7. APPENDIX

7.1 SCREENSHOTS:

TRAIN AND TEST DATA PRE PROCESSING:

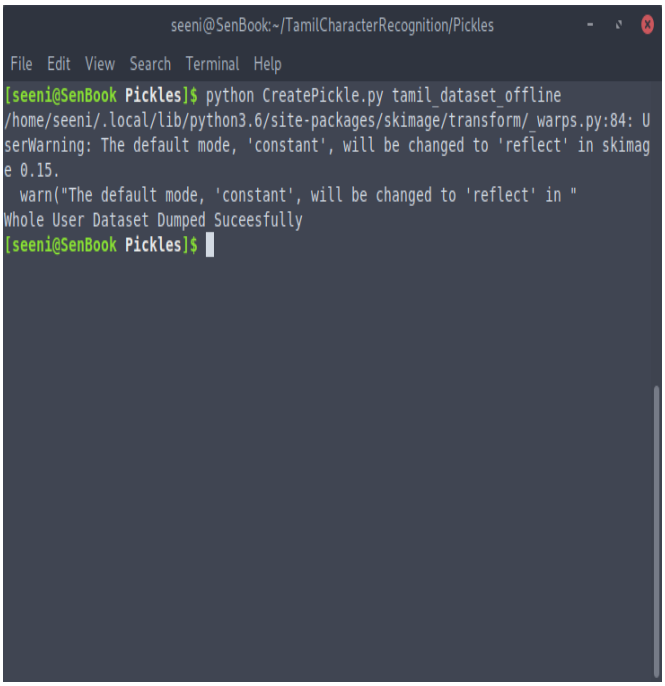
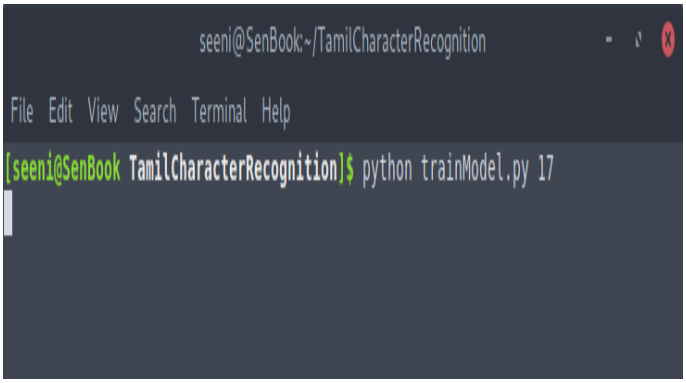


Fig 7.1 *Data Preprocessing Module*

TRAINING:

The system is trained with images of specifed user which is 17 in this case.



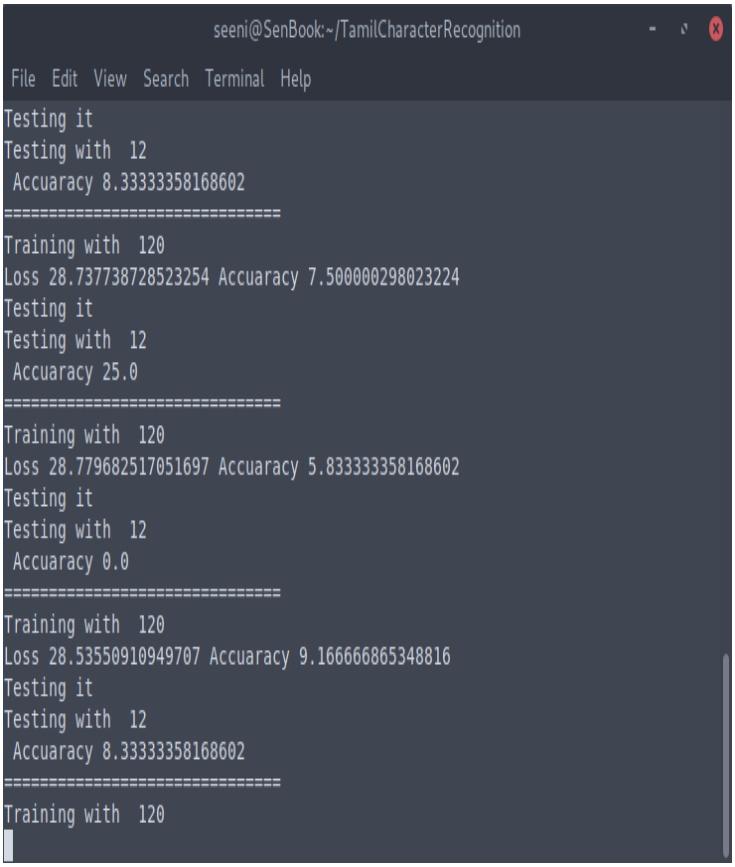


Fig 7.2 *Training*

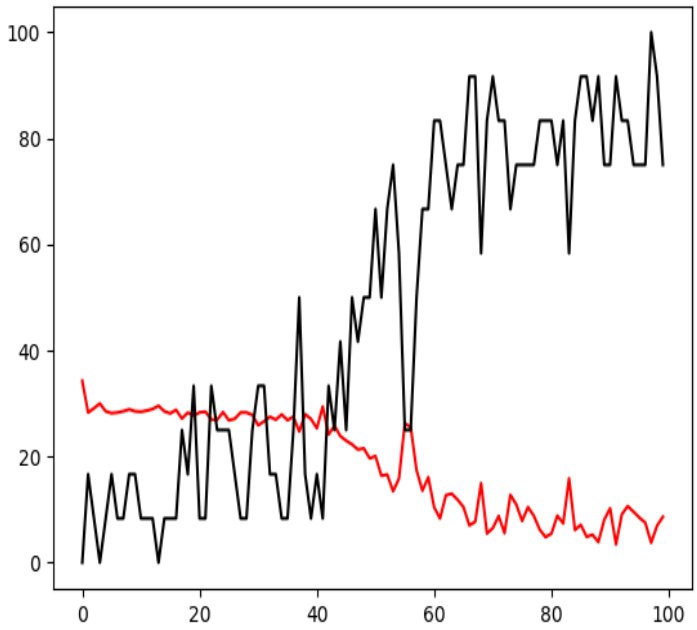


Fig 7.3 *Train Graph*

TRAINER GRAPH

Following graph is plot of loss(red) and accuracy(black) vs step. Those measures are calculated for the candidate data that is being trained. The Candidate number will be specified by the user while invoking trainer.

TESTING

Testing is done along with each step of training. Following graph is plot of loss(red) and accuracy(black) vs step. Those measures are calculated for the test data that were chosen randomly.

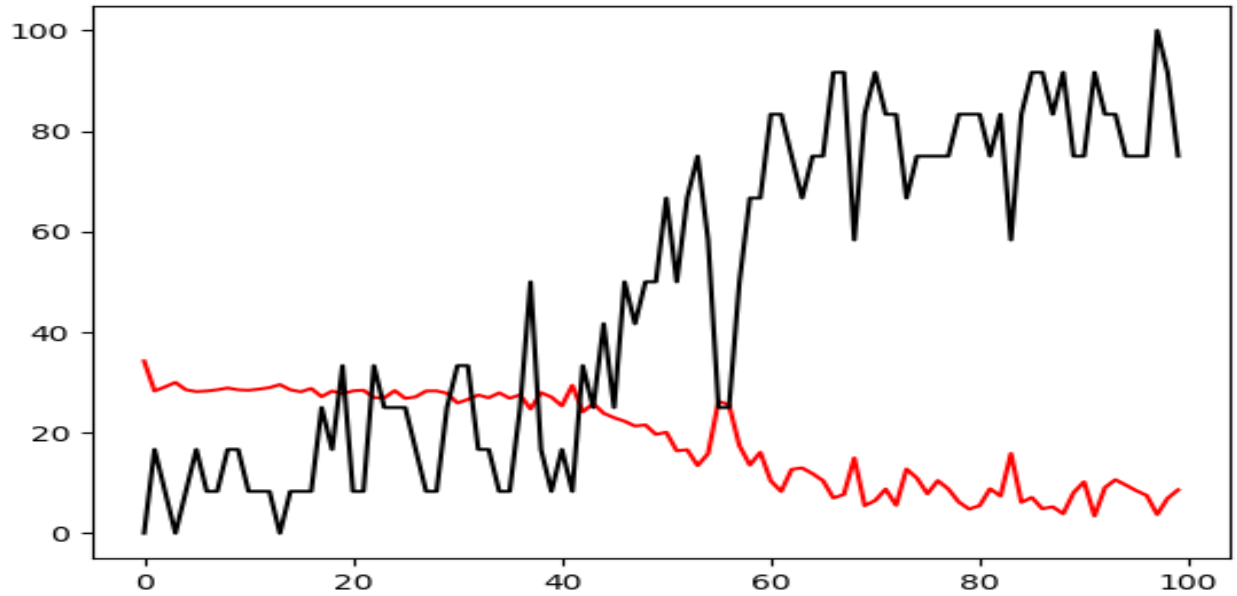


Fig 7.4 *Testing Graph*

PREDICTION

On finished training, user can invoke predict () method which asks absolute image path. On entering it, corresponding prediction will be displayed.

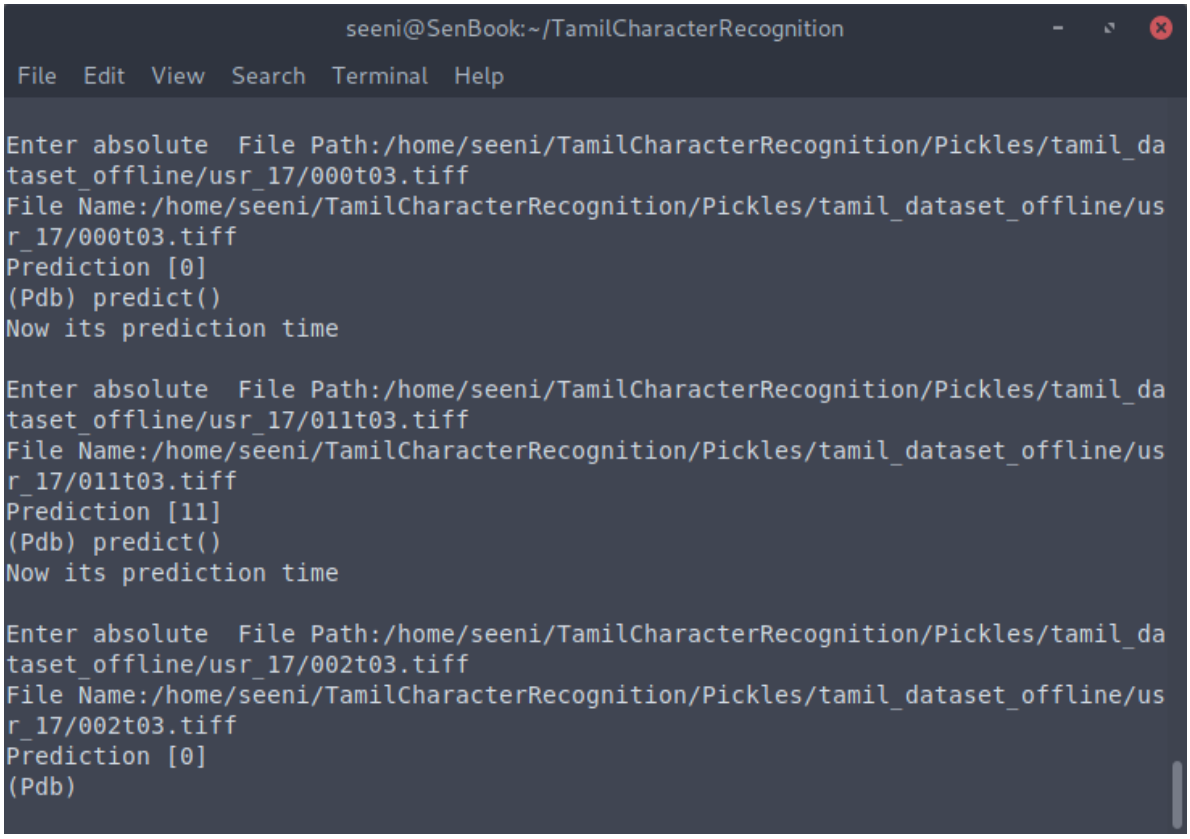


***Fig 7.5** Prediction*

REFERENCES

[1] Aravinda C.V and Dr. Prakash H.N –"Writer Recognition for South Indian Languages using statistical Feature Extraction and Distance Classifier" - International Journal of Natural Language Computing 2016

[2] Pallawi Agarwal and Yashavi Agarwal –"Applications of MATLAB's Toolbox to Recognize Handwritten CharactersPart 2: Experimental Results" –International Journal of Engineering Research and General Science 2014

[3]]N. Venkata Rao and Dr. A.S.C.S.Sastry - "Optical CharacterRecognition Technique Algorithms"-2016 Journal of Theoretical and Applied Information Technology

[4] Youssef Es Saady, Ali Rachidi, Mostafa El Yassa, Driss Mammass - "AMHCD: A Database for Amazigh Handwritten Character Recognition Research" - International Journal of Computer Application 2011

[5] Chirag I Patel, Ripal Patel, Palak Patel –"Handwritten Character Recognition using Neural Network" - International Journal of Scientific & Engineering Research Volume 2, Issue 5, May-2011

[6]] J.Pradeep, E.Srinivasan and S.Himavathi –"Diagonal based feature extraction for handwritten alphabets recognition system using neural network" - International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011

[7] R.Jagadeesh Kannan, R.Prabhakar and R.M. Suresh –"Off-Line Cursive Handwritten Tamil Character Recognition"- 2008 International Conference on Security Technology

[8] P Banumathi and Dr. G. M. Nasira –"Handwritten Tamil Character Recignition using Artificial Neural Networks" - 2011 IEEE

[9] J. Sujitha, M.E and N. Ramraj, M.E., Ph.d., - "Neural Network based offline Tamil Handwritten Character Recognition System" - 2007 IEEE

[10] Ishwarya .M.V, R. Jagdesh Kannan –"An improved Online Tamil Character Recognition using Neural Networks" - 2010 IEEE

[11] http://adventuresinmachinelearning.com/convolutional-neural-networks-tutorial-tensorflow/

[12] https://docs.python.org/3/library/unittest.html

[13] https://in.udacity.com/course/deep-learning-nanodegree-foundation--nd101

[14] https://medium.com/data-science-group-iitr/building-a-convolutional-neural-network-in-python-with-tensorflow-d251c3ca8117

[15] https://medium.com/@awjuliani/visualizing-neural-network-layer-activation-tensorflow-tutorial-d45f8bf7bbc4