HANDWRITTEN CHARACTER RECOGNITION USING CONVOLUTIONAL NEURAL NETWORK IN THE CONTEXT OF TAMIL LANGUAGE

Janotheepan Mariyathas[1]

[1]Department of Computer Science and Informatics, Faculty of Applied Science, Uva Wellassa University

Abstract

English language is having a high impact on handwritten character recognition. It is a difficult task to generate a model for character recognition especially for south Asian languages while they are having curves and dots in their shapes and the collection of characters as compound characters. Among other South Asian languages (e.g.: - Hindi, Malayalam, Telugu, etc.) Tamil characters are unique, because of the curves and dots they are having on each characters. These unique attributes make it hard to build a model to recognize Tamil characters. Another challenging task is to recognize handwritten characters than the printed characters, since the handwriting of each others are differing from each. Because of that little attention gained to build a model for Tamil handwritten character recognition. Convolutional Neural Network (CNN) and Deep Learning concepts are playing a major part in character recognition with more efficient image classification. This study mainly targeting the Tamil handwritten character recognition using CNN. The model was implemented in python programming language using Google colaboratory platform. The performance of the model was evaluated in each training and testing of the dataset. The maximum accuracy level reached when dataset reaches 50 character classes. Finally the overall accuracy was shown as 94.26% for 247 Tamil characters. Around 125 thousand data used for the model. Considering other similar systems this model shows the maximum performance. .

Keywords: Character recognition, Convolutional Neural Network, Deep Learning, Tamil

1 INTRODUCTION

Under the multicultural diversity of Sri Lanka, Tamil culture has been popular and often used culture in Sri Lanka. Around 25% of Sri Lankan diversity was covered by Tamil society. The people in the Tamil society are willing to maintain their official documents and the forms in Tamil language while all are not well educated to fill out them in English. These days the countries all over the world are digitalizing their works due to the pandemic situation faced because of Covid-19. Due to the pandemic situation faced by the people there are less possibility to access the offices and filling out the forms or applications for the government activities to follow. On the other hand digitizing all the documents and maintaining them in the local server or an online server will help to the government to protect them from the disasters [1]

Several algorithms has been developed already to recognize the handwritten characters for several languages like, English, Sinhala, Hindi, Javanese, Gujarati, Latin German, French, etc. [2]. Optical

character recognition (OCR) system and some machine learning techniques especially deep learning techniques was applied for the recognition of the handwritten character recognition in those above mentioned languages [2]. At the same time some algorithms has been already developed to recognize Tamil handwritten characters using deep learning techniques with the low accuracy level and without considering all the characters in Tamil.

So because of the reason mentioned above, model has been generated to recognize Tamil handwritten character recognition using deep learning architecture with the higher accuracy including all the characters in Tamil language. Many researchers under the domain of computer science and machine learning have completely explored the researches on character recognition using deep learning architecture especially CNN because of its algorithm. Among all of the researches, many of them mentioned that handwritten character recognition shows the highest accuracy in deep learning architecture compared to OCR algorithms [3]

These days some character recognition researches done already in the context of Tamil language. Among the concept of handwritten character recognition image classification plays a significant role to classify the characters by its special appearance like the curves and dots. Deep learning techniques help to categorize the images more accurately than the others, it leads to increase the accuracy level of the proposed model. This study focusses on Tamil handwritten character recognition using CNN and test the model to increase the accuracy level of the handwritten character recognition than the earlier methods.

2 RESEARCH METHODOLOGY

The main methodology of the handwritten character recognition system of this study is CNN architecture. CNN architecture is more preferable for the image classification tasks as mentioned above. Once the model has been developed the preprocessed dataset has been used for the training and testing purposes and there the maximum accuracy of the handwritten character recognition will be noticed.

Initially scanned Tamil handwritten character images has been preprocessed for the training. Preprocessing includes the tasks of cropping the images into single characters, removing the noisy data among the cropped images, and categorizing them into each character dataset. Huge amount of dataset is required by CNN to train the model. Data augmentation is makes sense over here, but data augmentation is not applicable for handwritten character images because of the data type. Each and every characters are having around 500 images for the training and testing the dataset. As a whole around 125 000 character images has been used for the model. The above mentioned dataset has been collected by considering all the variations among the handwriting.

Once the preprocessing done, all the images has been resized into 100 * 100 uniform size to consider all character images as same and to avoid getting noisy data. Those resized images are converted into grayscale images to avoid the RGB effect. If the RGB effect has not been converted into grayscale images might mislead the recognition process.

Then all of those images are sent to the convolutional layer, max-pooling layer and fully connected layer respectively. In this architecture 4 convolutional layers and 4 max-pooling layers are used for the enlargement of the character images. Next to that all the images are sent to the hidden layers, here 2 hidden layers are used in the architecture. 4*4 maximization was used in each layers. Rectified Linear Unit (ReLU) has been used in every layers. Figure 1 clearly explain how the enlargement of the images is working in every layer, here are only considered 2 convolutional layers and 2 max-pooling layers.

In this proposed model 80% of data is used for training and the remaining 20% of data used for

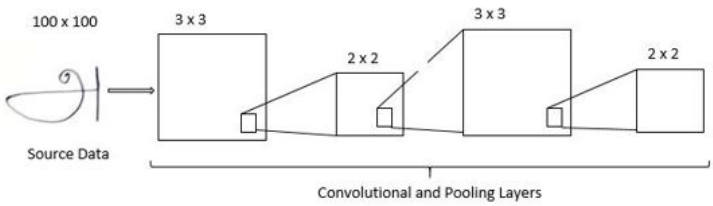

Figure 1. Image Enlargement in Convolutional Neural Network for Tamil Character Recognition

testing the algorithm. In every character, approximately around 400 images were used for training, and 100 images were used for testing the dataset. The random data partitioning was applied on this model, because of this, in each attempt the model will train and test different set of characters. The maximum accuracy was considered as the accuracy of the model, this happened when the clearest data selected as the training dataset. The Google colaboratory with tensor-flow (Online platform) is used to run the model.

3 RESULTS

The experiments has been conducted in Google colaboratory programming environment. Python programming language is used for implementing CNN. 247 different character classes (including vowels, consonants, ayutha ezhuthu and compound characters) of the Tamil language is used for the experiments. The above-mentioned dataset is used for the experiments of character recognition.

At the very first for evaluating the performance of CNN on Tamil handwritten characters, an incremental approach was used for the character classes. First started with 10 character classes of the dataset for the experiments to calculate the accuracy level of the recognition. Then the number of character classes slowly increased (like 25 character classes, 50 character classes, 100 character classes, 247 character classes). When the number of character classes increasing the accuracy level also will increase.

Considering only 5 character classes will lead to the model to reach a considerable accuracy level of the experiment. This experiment let the maximum accuracy level to 90.27%. In this experiment, 80-20 character class grouping was considered for the splitting of training and testing data. 250 iterations (epochs) were held and around 500 data were used for one character class during the experiment. All the time the accuracy received more than 80% for 5 character classes. Table 1 clearly explains the accuracy improvements with the number of character classes.

Table 1. Accuracy improvements according to dataset

Sample Dataset	Accuracy
10 character classes	90.27%
25 character classes	90.79%
50 character classes	91.97%
100 character classes	93.85%
247 character classes	94.26%

4 Conclusions and Suggestions

Optical character recognition is an active research area in handwritten character recognition. Handwritten character recognition always having a research gap as improving the accuracy level, why because 100% accuracy cannot be achieved for handwritten characters. Nowadays deep learning architecture making a high impact on image classification, an important section in character recognition. Considering the deep learning architectures CNN shows the highest performance in character recognition. This research mainly focuses on improving the accuracy level of the Tamil handwritten character recognition.

The developed model is trained with several number of character classes. The model tested with an increasing number of character classes, such as 10, 25, 50, 100, and 247 different character classes. There the reasonable accuracy is received as 90.27% for 10 character classes. Then, the fully trained model with 247 character classes shows an accuracy of 94.26%. For this experiment, 125 thousand data images were used in total (500 images per each).

References

[1] N Sasipriyaa, P Natesan, R Anand, et al. "Recognizing Handwritten Offline Tamil Character by using cGAN & CNN". In: *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*. IEEE. 2022, pp. 509–515.

[2] A Vijay, M Deepak, P Kavin, et al. "Transfer Learning based Offline Handwritten Recognition System using Tamil Characters". In: *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*. IEEE. 2022, pp. 214–220.

[3] Dan Ciregan, Ueli Meier, and Jürgen Schmidhuber. "Multi-column deep neural networks for image classification". In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 3642–3649.