Optical Text Recognition in Nepali and Bengali: A Transformer-based Approach

S M Rakib Hasan
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
sm.rakib.hasan@g.bracu.ac.bd

Aakar Dhakal
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
aakar.dhakal@g.bracu.ac.bd

Md Mustakin Alam
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
md.mustakin.alam@g.bracu.ac.bd

Md Humaion Kabir Mehedi
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
humaion.kabir.mehedi@g.bracu.ac.bd

Annajiat Alim Rasel
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
annajiat@gmail.com

***Abstract*—Efforts on the research and development of OCR systems for Low-Resource Languages is relatively new. Low-Resource Languages have little training data available for training Machine Translation systems or other systems. Even though, vast amount of text has been digitized and made available on the internet the text is still in PDF and Image format, which are not instantly accessible. This paper discusses typed text recognition for two scripts: Bengali and Nepali; there are about 300 and 40 million Bengali and Nepali speakers respectively. In our study, using convolutional neural networks (CNN) a model was developed, and its efficacy was assessed using a collection of typed text images. The results signify that our suggested technique beats current approaches and achieves high precision in recognizing typed text in Bengali and Nepali. This study can pave the way for advanced and accessible study of linguistics in South East Asia.**

***Index Terms*—Low-Resource Languages, OCR, CNN**

I. Introduction

For the advent of everchanging digital media and technology in recent years, there has been a sharp rise in the demand for automated text detection. Optical Character Recognition(OCR) is a technology that enables the recognition of printed or handwritten text characters from scanned images. It improves access to information, preserves rare texts and enables the development of language technologies such as voice assistants and machine translation systems. Despite numerous works on text recognition in various languages, Bengali and Nepali text recognition has not received enough attention due to its resource deficiency or being morphologically complex. Bengali and Nepali are two extensively spoken languages in South Asia, and understanding them is essential for computer translation, document digitisation, and language processing. In this paper, we suggest an image-based method for identifying Bengali and Nepali written text. We have developed a model using convolutional neural networks (CNN) and its efficacy was assessed using a collection of typed text images. The outcomes demonstrate that our suggested technique beats current approaches and achieves high precision in deciphering typed text in Bengali and Nepali. This study could aid in the advancement of linguistic technology in South Asia and be useful in a number of industries, including automation, administration, and education. This paper discusses the previous researches, some unique characteristics of Bengali and Nepali text, segmentation and feature extraction methods followed by the experimental results and conclusion.

II. Previous Works

There have been many OCR models for low-resource languages like Bengali or Nepali and some remarkable research exists in this field. An end-to-end word identification system for handwritten Bengali words from pictures is introduced in the paper [1]. Deep convolutional neural networks (CNNs) are used by authors as feature extractors, followed by RNNs and a completely connected layer that produces the end prediction. The Connectionist Temporal Classification (CTC) loss function is used to teach the algorithm. The efficacy of four distinct baseline models—Xception, NASNet, MobileNet, and DenseNet—as feature extractors are examined in this article. The writers come to the conclusion that for Bengali handwritten OCR, deeper systems with residuals work better. The BanglaWriting dataset, a reputable Bengali dataset, is used

to assess the suggested technique. With a word recognition accuracy of 90.3%, the authors describe encouraging findings. The work [2] shows an OCR system for printed Bengali and English text that was designed using a single, 128-unit hidden BLSTM-CTC design. The suggested approach operates in two stages. The document vertical strip-based projection profile valleys of the document parts were taken into consideration as the original hypotheses for the line end/beginning in the first step. This hypothesis was further developed in the second step, during which object pictures and other artefacts were eliminated. The individual text lines are then given to the BLSTM-CTC-based OCR system for training and certification after being normalized to a height of 48 pixels. Performance assessment makes use of test line ground truth. The CTC serves as the classifier during testing, generating the most likely classifications for a specific input sequence as the end output. Character level accuracy for the suggested OCR system was 99.32%, and word level accuracy was 96.65%. The difficulties of optical character recognition (OCR) for Bangla text are discussed in the work[3], along with a useful character segmentation method. Using vertical and curved scanning, the segmentation approach uses line, word, and character segmentation. The segmented character extraction procedure utilizing the flood-fill method is also covered in the article. The suggested method had 99% character segmentation accuracy, 99.8% line segmentation accuracy, and 99.5% word segmentation accuracy. In the publication, character segmentation accuracy experimental findings are presented. For five separate pictures, 99% accuracy was attained. The suggested method can enhance the accuracy of OCR for printed Bangla text. A low-cost convolutional neural network design for Bengali handwritten character identification is suggested by the writers in the study [4]. For the training phase, the researchers used openly accessible standard datasets, such as the CMATERdb Bengali handwritten character dataset, and created various dataset forms in accordance with prior research. BengaliNet, the suggested design, outperformed earlier work by a significant margin, achieving a total accuracy of 96-99% for various databases of Bengali characters. The study aids in the creation of an instrument that can automatically and effectively recognize Bengali handwriting symbols. The paper[5] suggests a customized version of a cutting-edge deep learning methodology to identify handwritten Bengali symbols. The suggested approach makes use of transfer learning on the ImageNet dataset along with a pre-trained Resnet-50 deep convolutional neural network model. By altering the input image sizes, the technique also changes the one-cycle strategy to guarantee quicker training. The 84-class BanglaLekha-Isolated dataset is used to assess the suggested methodology. The findings demonstrate that the suggested approach outperforms other contemporary methods, achieving 97.12% accuracy in just 47 epochs. The architecture and regularizations, including batch normalization, dropout, learning rate schedule, and optimizers, used to boost speed are also covered in the article. The authors come to the conclusion that ResNet-50 is successful at classifying Bengali handwritten characters using their suggested technique. A novel feature set for handwritten Bangla alphabet detection is described in the paper[6]. The feature collection comprises 132 features, including quad-tree-based longest run features, distance-based features, modified shadow features, octant and centroid features, and features based on centroid and distance. The features are calculated from binary pictures of Bangla alphabetic letters with a dimension of 64x64 pixels. Four longest-run features are calculated for every sub-image at every point of the quad-tree structure. The novelty of the present study is the partitioning of any character sequence using the CG-based quad-tree structure. From the 75.05% seen in the authors' earlier work to 85.40% on 50 character classes with an MLP-based classifier on the same dataset, the identification performance using this feature set rises significantly. In a study[7], the authors developed a new benchmark for real and synthetic data enriched with noise, for 60 low-resource languages with low resources scripts. The authors evaluated both commercial and research purpose State of the Art OCR models and benchmarked and extracted their most common errors. In this work SOTA OCR errors were used to measure their impact on Machine Translation models by comparing the MT models fine-tuned with OCR-ed data with pre-trained MT models and MT models fine-tuned with initial/non OCR-ed data. The most important takeaway is that OCR-ed monolingual data improves machine translation through backtranslation. This augmentation is robust to most types of errors, except replacements, and in general most current OCR models produce good enough recognition to be able to train MT models, with the exception of a few scripts like Perso Arabic. This work paves the way for future research on data augmentation for Machine Translation based on OCR documents.

III. Characteristics of Bengali and Nepali Texts

By nature, Bengali, Nepali and other oriental languages have quite irregular and unique forms than Latinized languages. These features make the feature extraction and mapping of these languages difficult. However, in its written form, the Bengali language has a total of 50 letters in its alphabet, 11 being vowels and 39 consonants and the Nepali language has the same number of vowels with 33 consonants. Both these languages have some similarities in their written form which make their interpretation harder, described as follows:

A. Dependant Letters

There are vowels and consonants joined with the consonants for producing the sounds in the words. So, the regular shape of the letters gets changed very frequently. These can be in many different shapes(each different for short and long sounds), which can get very confusing for the machine to interpret. These also change the normal distancing between letters which makes it more difficult to segment the letters. Fig-1 and Fig-2 depict these patterns for both languages

B. Overlapping Characters

Often the letters of these alphabets do not remain bounded to a certain region and enter the region of its next character.

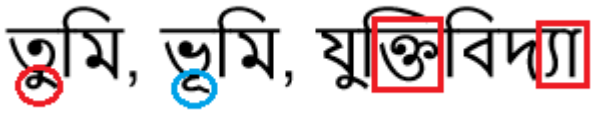

Fig. 1. Dependant letters in Bengali Text.

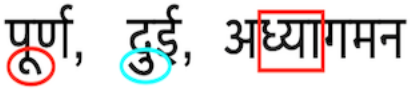

Fig. 2. Dependant letters in Nepali Text.

It happens because the dependent letters join and get into the next letters. It is shown in figure-3 and figure-4:

Fig. 3. Overlapping Character in a Bengali word.

C. Word Separation

There is a line on the letters of each word which keep the word together. This line, known as the "Matra" line can determine the words separately. However, there are letters which skip the line above them, making the feature extraction difficult. Also, above the matra line, there may be parts of a letter and below there may be vowels. The figures below show this phenomenon.

Fig. 4. Separation of words in Bengali texts.

D. Confusing Shapes

Both languages have some letters that are often mistaken to be the same. For a machine, it becomes tough to differentiate between these letters having very low data to train on. The following figures show these letters in both languages.

IV. Methodology

Apart from common OCR techniques in low-resourced languages which used CNNs or RNNs, we propose transformer-based methods. The Transformer models, in contrast to the features retrieved by the CNN-like network, do not contain picture-specific inductive biases but do the image processing as a series of patches, making it simpler for the model to pay attention to either the entire image or the independent patches. We have used the transformer model TrOCR[9] as our base

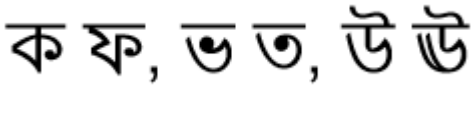

Fig. 5. Some confusing letters in Bengali.

model and extended it to extract features from Bengali and Nepali images and map these features to texts. TrOCR uses typical encoder-decoder methods. It can be divided into two distinct parts

A. Feature Extraction Encoder

An image transformer is used in this part which takes some input images and extracts their features as outputs. We have used the VisionEncoderDecoder model with the pre-trained Vision Transformer (ViT)[10] as the encoder. The google/vit-base-patch16-384 from huggingface was used which is pre trained on 14 million images and fine tuned on one million images. The model receives images as a series of linearly embedded, fixed-size patches with a 16x16 resolution. In order to employ a sequence for classification tasks, a [CLS] token is also added to the beginning of the sequence. Before feeding the sequence to the layers of the Transformer encoder, one may additionally include absolute position embeddings. For example, if you have a dataset of labeled pictures, you may train a typical classifier by stacking a linear layer on top of the pre-trained encoder. By pre-training the model, it learns an inner representation of images that can subsequently be used to extract features helpful for downstream tasks. Since the [CLS] token's final hidden state can be viewed as an outline of an entire image, a linear layer is typically placed on top of it. The architecture of the model is shown in figure-9. After the feature extraction, the language modeling is done by the decoder.

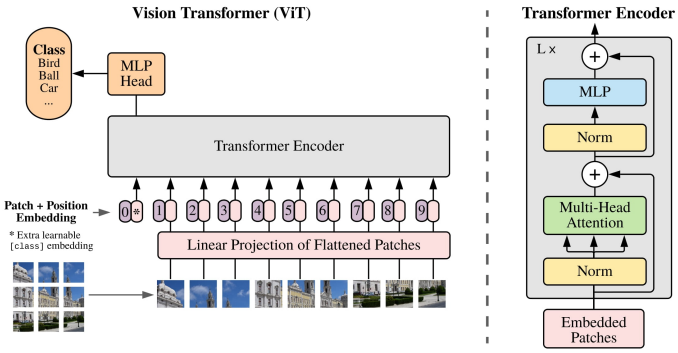

Fig. 6. The ViT architecture[8].

B. Language Modeling

A text transformer is used to map the extracted features to texts. In this instance, the xlm-roberta-base decoder has been chosen. This is a multilingual text transformer model that has been trained on 2.5 terabytes of data from 100

languages. There are few language decoder models specifically for Bengali and Nepali. Therefore, we've selected this model, which can perform the required tasks in both languages. This is an multilingual rendition of RoBERTa. It was pretrained with the objective of Masked language modeling (MLM). Using a sentence as input, the model conceals 15% of the words at random before running the entire sentence through the model and attempting to predict the hidden words. This differs from conventional recurrent neural networks (RNNs), which typically see the words one after the other, and autoregressive models such as GPT, which conceal the future tokens internally. It enables the model to learn a representation of the sentence that is bidirectional. Thus, the model learns an internal representation of 100 languages that can then be used to derive features useful for downstream tasks[11].

V. Experimental Setup

A. Datasets

For the Bengali language, we have used open source BanglaWriting[12] dataset for fine-tuning the transformer model. It includes the single-page handwriting of 260 people of various ages and personalities This data set includes a total of 21,234 words and 32,784 characters. Additionally, this dataset comprises 5,470 unique Bangla words. The dataset also contains 261 comprehensible overwriting and 450 handwritten strikes and errors, in addition to the usual words. The word labels are generated manually. For the Nepali language, we have used a manually prepared dataset which includes 50 images from different people. It contains around 10 thousand words with about 25 thousand characters. It also has many unique words and the word labels are manually generated. We have used both language datasets separately for training and validation.

B. Data Preprocessing

Our chosen dataset had images of different sizes. We resized each to 384x384 pixels for feeding to our encoder, ViT. The images were already denoised so we skipped that part. As the encoder works well with normal images, we did not grayscale the data. We augmented the images with random flipping and random rotation(-5 to 5 degrees).

C. Training

We trained the model on our dataset for 2000 epochs for each language with a batch of 4 images. The training was done on a machine with a 3.6GHz CPU, 16 GB RAM and RTX 3070 GPU. We have used the AdamW algorithm for optimization, which is a stochastic gradient descent technique based on an adaptive prediction of first-order and second-order moments, with an additional method to diminish weights in accordance with the techniques[13]. For the training and validation loss visualization, we used WandB and DataCollator libraries.

VI. Results and Comparison

A. Figures and Tables

a) Comparison: Compared to other methods "Fig. 6", even at the beginning of a sentence.

TABLE I
Table Type Styles

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a]Sample of a Table footnote.

VII. Conclusion

In this paper, we proposed typed text recognition for Low-Resource Language, Bengali and Nepali, by developing CNN model and calculating its efficiency using a collection of typed text images. The extracted texts from images can be further used for Machine Translation, NLP, and other processes to democratize access to Bengali and Nepali text.

References

[1] F. B. Safir, A. Q. Ohi, M. F. Mridha, M. M. Monowar and M. A. Hamid, "End-to-End Optical Character Recognition for Bengali Handwritten Words," 2021 National Computing Colleges Conference (NCCC), Taif, Saudi Arabia, 2021, pp. 1-7, doi: 10.1109/NCCC49330.2021.9428809.

[2] D. Paul and B. B. Chaudhuri, "A BLSTM Network for Printed Bengali OCR System with High Accuracy," CoRR, vol. abs/1908.08674, 2019, [Online]. Available: http://arxiv.org/abs/1908.08674

[3] A. F. Rownak, M. F. Rabby, S. Ismail and M. S. Islam, "An efficient way for segmentation of Bangla characters in printed document using curved scanning," 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), Dhaka, Bangladesh, 2016, pp. 938-943, doi: 10.1109/ICIEV.2016.7760138.

[4] A. Sayeed, J. Shin, Md. A. M. Hasan, A. Y. Srizon, and M. Hasan, "BengaliNet: A Low-Cost Novel Convolutional Neural Network for Bengali Handwritten Characters Recognition," Applied Sciences, vol. 11, no. 15, p. 6845, Jul. 2021, doi: 10.3390/app11156845.

[5] S. Chatterjee, R. K. Dutta, D. Ganguly, K. Chatterjee, and S. Roy, Bengali Handwritten Character Classification Using Transfer Learning on Deep Convolutional Network. Springer Nature, 2019, pp. 138–148. doi: 10.1007/978-3-030-44689-5_13.

[6] N. Das, S. Basu, R. Sarkar, M. Kundu, M. Nasipuri, and D. Basu, "An Improved Feature Descriptor for Recognition of Handwritten Bangla Alphabet," arXiv (Cornell University), Jan. 2015, doi: 10.48550/arxiv.1501.05497.

[7] O. Ignat, "OCR Improves Machine Translation for Low-Resource Languages," arXiv Cornell University), Feb. 27, 2022. https://arxiv.org/abs/2202.13274

[8] Dosovitskiy, Alexey Beyer, Lucas Kolesnikov, Alexander Weissenborn, Dirk hai, Xiaohua Unterthiner, Thomas Dehghani, Mostafa Minderer, Matthias Heigold, Georg Gelly, Sylvain Uszkoreit, Jakob Houlsby, Neil. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

[9] Li, Minghao Lv, Tengchao Cui, Lei Lu, Yijuan Florencio, Dinei Zhang, Cha Li, Zhoujun Wei, Furu. (2021). TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models.

[10] Wu, Bichen Xu, Chenfeng Dai, Xiaoliang Wan, Alvin Zhang, Peizhao Tomizuka, Masayoshi Keutzer, Kurt Vajda, Peter. (2020). Visual Transformers: Token-based Image Representation and Processing for Computer Vision.

[11] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.

[12] Mridha, Dr. M. F.; Quwsar Ohi, Abu; Ali, M. Ameer; Emon, Mazedul Islam; Kabir, Md Mohsin (2020), "BanglaWriting: A multi-purpose offline Bangla handwriting dataset", Mendeley Data, V1, doi: 10.17632/r43wkvdk4w.1

[13] I. Loshchilov and F. Hutter, 'Decoupled weight decay regularization', arXiv preprint arXiv:1711. 05101, 2017.