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 are relatively new. Low-Resource Languages have little training data available for training Machine Translation systems or other systems. Even though a 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 text recognition for two scripts: Bengali and Nepali; there are about 300 and 40 million Bengali and Nepali speakers respectively. In our study, using encoder-decoder transformers, a model was developed, and its efficacy was assessed using a collection of optical text images, both handwritten and printed. The results signify that our suggested technique corresponds with current approaches and achieves high precision in recognizing text in Bengali and Nepali. This study can pave the way for the advanced and accessible study of linguistics in South East Asia.

Index Terms—Low-Resource Languages, OCR, Transformers, Machine Translation, Digitized text

I. INTRODUCTION

With the advent of ever-changing 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 Microsoft TrOCR, an encoder-decoder based transformer architecture as the base model and its efficacy was assessed using a collection of text images. The outcomes demonstrate that our suggested technique 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 study [2] presents an optical character recognition (OCR) framework for printed Bengali and English text. The system was developed utilizing a solitary, 128-unit hidden bidirectional long short-term memory (BLSTM) architecture with connectionist temporal classification (CTC). The proposed methodology consists of a dual-phase process. In the initial stage, the vertical strip-based projection profile valleys of the document components were regarded as the primary suppositions for the line end/beginning. The aforementioned hypothesis underwent further development in the subsequent stage, wherein depictions of objects and additional artefacts were excluded. The text lines are subjected to normalization to achieve a height of 48 pixels before being presented to the BLSTM-CTC-based OCR system for training and validation. Performance evaluation employs test data that serves as a reference standard. During the testing phase, the CTC functions as the classifier and produces the most probable classifications for a given input sequence as the final outcome. The OCR system's accuracy at the character level was 99.32%, while its accuracy at the word level 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.

The authors of the study [7] established a novel benchmark for 60 low-resource languages that possess low-resource scripts. This was accomplished through the utilization of authentic and synthetic data that was augmented with noise. The researchers conducted a benchmarking study to identify the most common errors in Optical Character Recognition (OCR) models that are utilized for both academic research and commercial applications. This study utilized state-of-the-art OCR mistakes to evaluate the impact of fine-tuning machine translation (MT) models with OCR-ed data in comparison to pre-trained MT models and MT models fine-tuned with initial data. The study aimed to contrast the MT models and determine the effect of SOTA OCR mistakes on them. The principal finding is that the utilization of back translation in optical character recognition-processed monolingual data results in an improvement in machine translation. Overall, the majority of current optical character recognition (OCR) models exhibit a high level of recognition accuracy, which allows for the training of machine translation (MT) models. However, there are certain scripts, such as Perso Arabic, that present challenges to OCR models. Additionally, while this augmentation process is generally robust to various types of errors, it may be susceptible to substitutions in particular. This work will enable future research endeavours focused on data augmentation for machine translation utilizing OCR documents.

The study [14] introduces a novel Optical Character Recognition (OCR) system designed for scripted Bangla characters. The system has been constructed utilizing a Convolutional Neural Network (CNN) for the purpose of character recognition. The system also integrates various techniques such as binarization and noise reduction to enhance the precision of the results. The results of the performance evaluation indicate that

the proposed system exhibits superior accuracy in comparison to existing methods with respect to line detection, word segmentation, and character recognition. The study involved conducting an experiment on the recognition of a maximum number of distinct characters, which was 202. Additionally, the number of training and test images utilized in the experiment exceeded those used in prior research. A precision rate of 99.33% was attained. The system exhibits robust performance in diverse conditions, including images that are affected by noise or blurriness. The duration of computation may pose a challenge contingent upon the capacity of the system in which it is executed, and the system's efficacy may be compromised in instances where the image is excessively rotated.

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
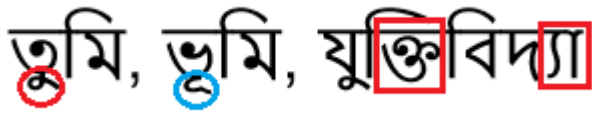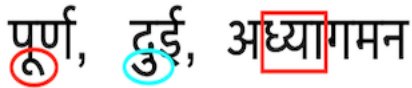

Fig. 1. Dependant letters in Bengali Text.

Fig. 2. Dependant letters in Nepali Text.

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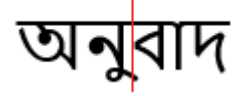

Fig. 3. Overlapping Character in a Bengali word.

Fig. 4. Overlapping Character in a Nepali 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. 5. Separation of words in Bengali texts.

Fig. 6. Separation of words in Nepali 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.
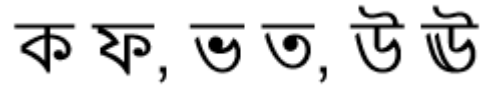

Fig. 7. Some confusing letters in Bengali.

Fig. 8. Some confusing letters in Nepali.

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 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. An instance of utilizing a dataset of labelled images involves training a conventional classifier through the stacking of a linear layer onto the pre-existing encoder. Through pre-training, the model acquires an internal representation of images that can be leveraged to extract features that are advantageous for subsequent tasks. The final hidden state of the [CLS] token is often regarded as a comprehensive representation of an entire image and is therefore commonly augmented with a linear layer. Figure 9 depicts the architectural design of the model. After the feature extraction, the language modelling is done by the decoder.
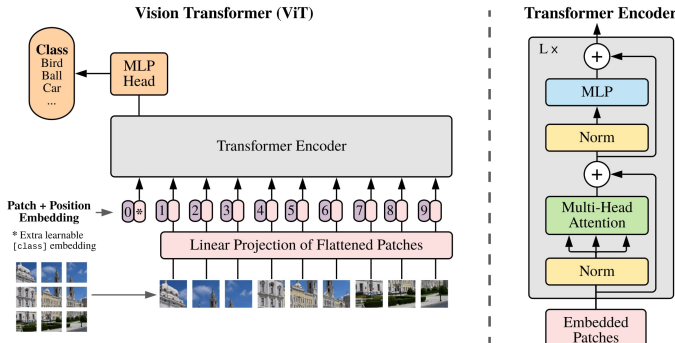

Fig. 9. The ViT architecture [8].

B. Language Modelling

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 not many 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 a multilingual rendition of RoBERTa. It was pretrained with the objective of Masked language modelling (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].
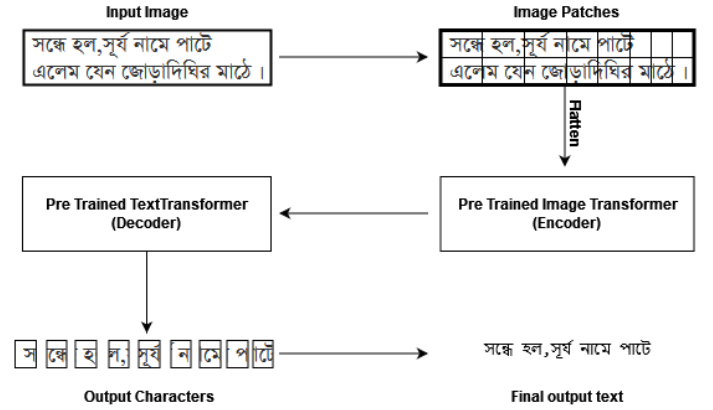

Fig. 10. Proposed workflow

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 7 thousand words with about 10 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.

D. Testing

The performance of our trained Optical Character Recognition (OCR) model was assessed by analyzing a set of 10 images obtained from various sources. The chosen images were curated to exemplify a broad spectrum of font styles, sizes, and orientations that are frequently encountered in practical situations. The initial step in the testing procedure involved image preprocessing, which entailed resizing the images to a uniform size of 512x512 pixels and converting them to grayscale. Subsequently, the preprocessed images were inputted into our OCR model, and the resulting recognized text was documented. In order to assess the precision of our optical character recognition (OCR) model, we computed the subsequent metrics:

- The metric of character-level accuracy evaluates the proportion of accurately identified characters within the completely recognized text. The computation involves the division of the count of accurately identified characters by the overall count of characters present in the reference text.
- The metric of word-level accuracy evaluates the proportion of accurately identified words within the complete recognized text. The computation involves the division of the count of accurately identified words by the overall count of words present in the reference text.

The ground truth text for each of the 10 images was manually annotated and compared with the recognized text generated by our OCR model to calculate the metrics.

VI. RESULTS AND EVALUATION

The text recognition performance of our suggested model in Bengali and Nepali languages was encouraging. During training, the model achieved low Word Error Rates (WER) of 0.10 and 0.14 for Bengali and Nepali, respectively, and Character Error Rates (CER) of 0.04 and 0.10 for Bengali and Nepali, respectively. The Levenshtein distance between the recognized and ground truth texts was used to determine the CER and WER values.

We plotted the CER and WER values over 2000 epochs to show the model's performance. The data demonstrated that
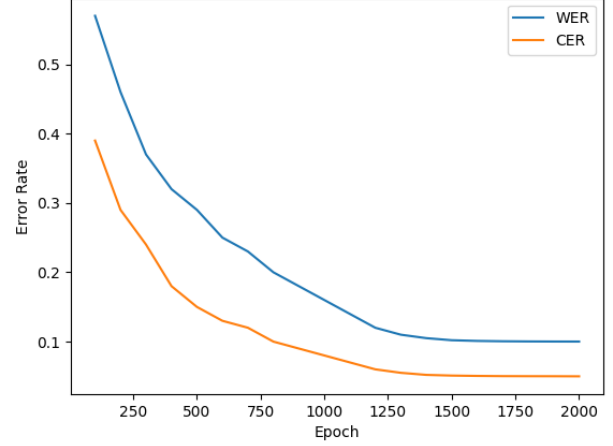

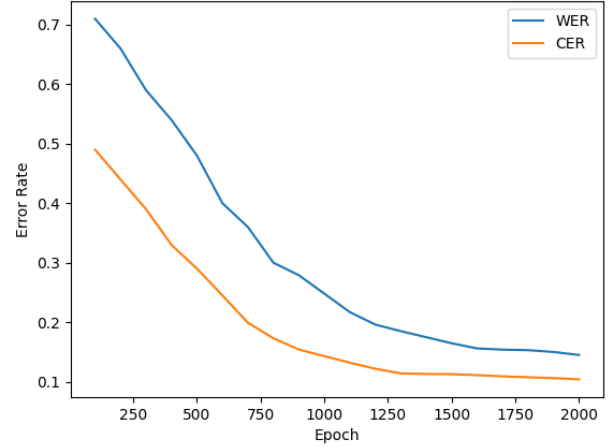Fig. 11. CER and WER for Bengali on Training Dataset

Fig. 12. CER and WER for Nepali on Training Dataset

the model, for both Bengali and Nepali languages, reached a consistently low CER and WER after roughly 1500 epochs.

The CER and WER values were marginally higher than those attained during training, though, when the model was tested on the test set. Bengali had an average CER and WER of 0.07 and 0.12, while Nepali had an average CER and WER of 0.11 and 0.15.

The modest variation in the CER and WER values between training and testing may be explained by the test set's greater diversity and changing image quality when compared to the training set. The test set may have included images with more intricate backgrounds or uncommon font styles that weren't present in the training set, despite the fact that the model was trained on a varied selection of images with different font styles, sizes, and orientations.

VII. Conclusion

In this research, we suggested a model for Bengali and Nepali text recognition. Images with handwritten and printed text in both languages were used to train the model. According to our experimental findings, the suggested model trained with a low Character Error Rate (CER) and Word Error Rate (WER) and performed well on the test set. During training, the suggested model obtained CERs of 0.04 and 0.09 and WERs of 0.10 and 0.14 for Bengali and Nepali, respectively. The model's average CER and WER for Bengali and Nepali, respectively, when tested on the test set, were 0.07 and 0.12 and 0.11 and 0.15, respectively. These outcomes show how well our suggested model performs when it comes to correctly identifying text in Bengali and Nepali. The model has the potential to be used in practical applications like document digitization and text extraction because of its low CER and WER values during training and ability to perform well on the test set. Our suggested OCR methodology shows its potential for practical uses like text extraction and document digitisation. The model's ability to correctly identify text from a variety of images in Bengali and Nepali languages demonstrates how well it can manage complicated and varied input.

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.

[14] T. Ahmed, M. N. Raihan, R. Kushol, and M. S. Salekin, 'A Complete Bangla Optical Character Recognition System: An Effective Approach', in 2019 22nd International Conference on Computer and Information Technology (ICCIT), 2019, pp. 1–7.