Bengali Grapheme Classification - Milestone Report

Problem Statement:

The aim of this capstone project is to build a deep neural network which can classify a handwritten Bengali character into three constituent parts - the grapheme root, vowel diacritic and consonant diacritic. Bengali character classification is particularly challenging since along with the 49 Bengali letters, there are 18 potential diacritics and the number of resulting combinations is approximately 13,000. In comparison, the number of graphemic units in English is 250.

Dataset Description:

The dataset, obtained from Kaggle, comes split into training and test datasets. The training dataset contains 200,840 images of the size 137 x 236 pixels. The test set contains 12 images of the same size. Along with the image data, there are csv files which describe the labels for each image.

Data Cleaning:

Since it was a Kaggle dataset, minimal data cleaning was required

Findings from Exploratory Data Analysis:

Before analyzing the dataset, it is important to define what a grapheme and a diacritic are. A grapheme is, according to Wikipedia, “the smallest unit of a writing system of any given language. … a grapheme is a letter or a set of letters that represent a sound (more correctly, phoneme) in a word”. A diacritic is essentially an accent on a character.

Exploring the dataset, the 5 most frequent grapheme roots, vowel diacritics and consonant diacritics are shown below:

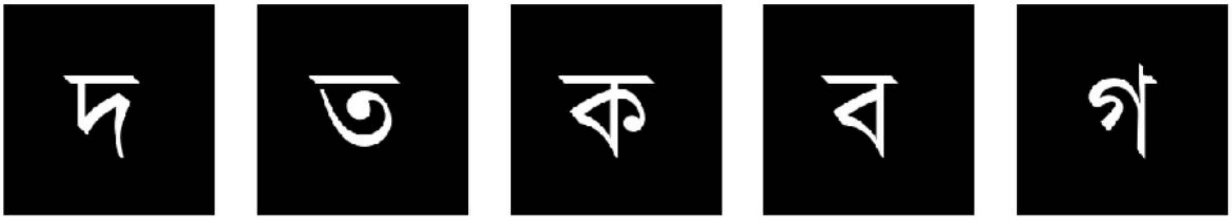


Figure 1: 5 Most Frequent Grapheme Roots

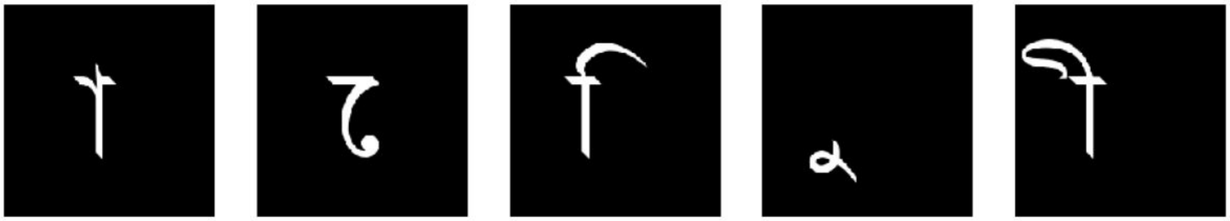


Figure 2: 5 Most Frequent Vowel Diacritics

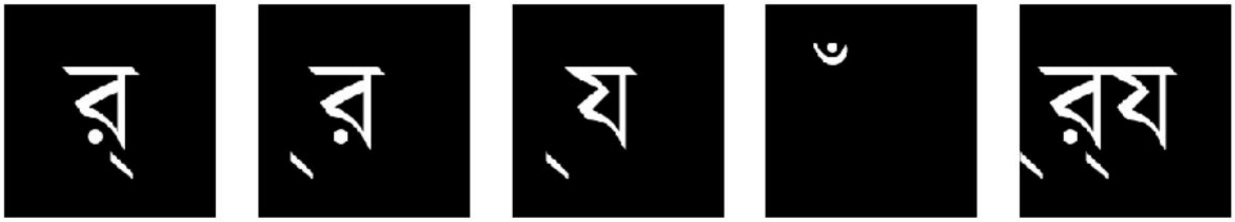
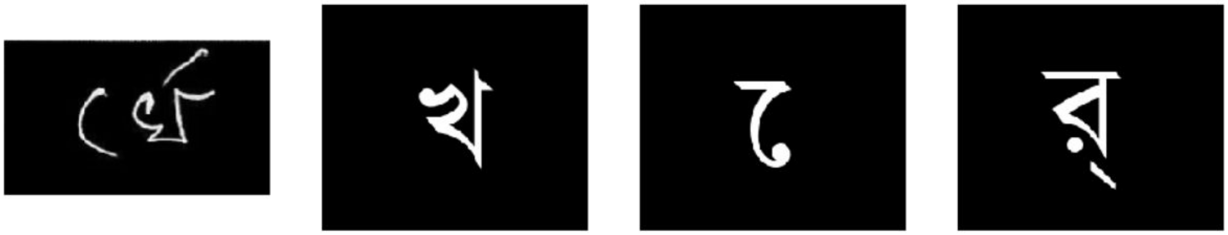


Figure 3: 5 Most Frequent Consonants

The above characters however, are not hand-written characters. They are generated using the computer font *Kalpurush*. The hand-written characters themselves will not be as neat, and will be combinations of grapheme roots, vowel diacritics and consonants (although many characters might not possess all three components).

The figure below is a random image taken from the training data set (far left) compared against its constituent components.



Hand-written Character Grapheme Root Vowel Diacritic Consonant

Figure 3: 6 Hand-written Character Decomposition

From the figure we can make out the grapheme root and vowel diacritic from the hand-written character. But the consonant takes a different shape when integrated into the character. This is a characteristic of Indic languages and is something the classifier we build must be cognizant of.

Next Steps:

The next step is to actually build a deep neural network classifier, train it on the training dataset and conduct experiments to optimize the network performance. We will be implementing a custom convolutional neural network for this purpose.