Paper Summary

A Large Multi-Target Dataset of Common Bengali Handwritten Graphemes
Link: https://arxiv.org/abs/2010.00170

Abstract

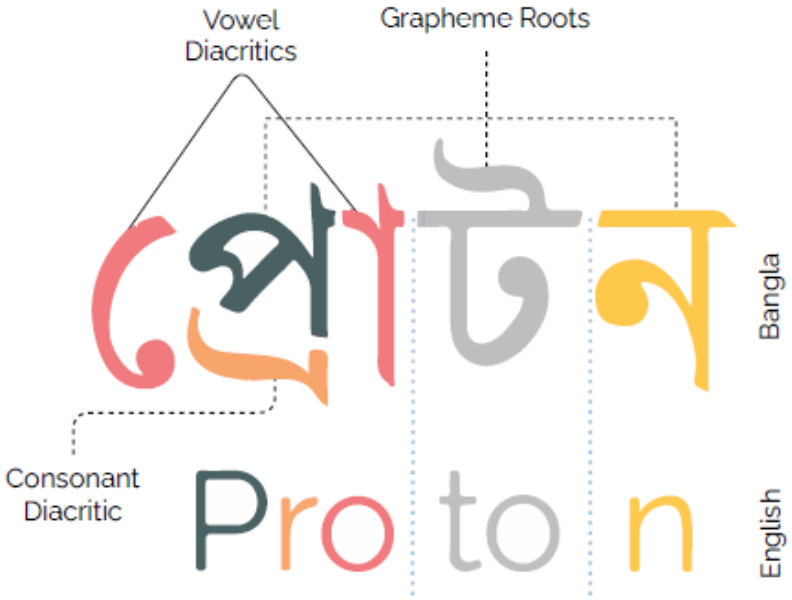

Figure 1. Grapheme components.
Source: Samiul *et al.* [1]

- Handwriting optical character recognition of alpha-syllabary languages is harder than Latin due to cursive writing system and frequent use of diacritics. Particularly, the segmentation of the graphical constituents is much harder as shown in Figure 1.
- The work proposes grapheme-based labeling scheme to create the first dataset of Bengali handwritten graphemes.
- 411000 samples of 1295 unique commonly used Bengali graphemes.

- About 900 uncommon Bengali graphemes in test set for out of dictionary performance evaluation.
- Open-source benchmark released on Kaggle platform and a competition took place based on this new benchmark.
- Unique graphemes are selected based on commonality in the Google Bengali ASR corpus.

Introduction

- Unlike English, in alpha-syllabary languages like Bengali and Hindi, each character may consist of sequence of symbols that do not correspond to a linear arrangement of phonemes.
- So, non-linear positioning must be considered when developing an OCR for Bengali handwriting.
- The grapheme-based scheme proposed by the authors bypass the complexities of character segmentation inside handwritten alpha-syllabary words.
- Problem has been posed as multi-target classification with three targets (grapheme root, consonant diacritic, vowel diacritic).

Challenges

- **Consonant conjuncts** may have **second order conjuncts**, e.g., ষ্ট = শ + ট {sta = śa + ta} or,
- **Third order conjuncts**, e.g., ক্ষ্ণ = ক + ষ + ণ {kṣṇa = ka + ṣa + na}.
- **Allographs** are cases where same grapheme can have multiple writing styles. Examples are shown in Figure 2.

- 3883894 unique graphemes possible by different combinations of vowels, consonant conjuncts, vowel diacritics, etc.
- However, only a small amount is prevalent in modern Bengali.

Figure 2. Examples of allograph pairs.
Source: Samiul *et al.* [1]

Dataset

- Popular graphemes were selected by using Google Bengali ASR dataset as reference corpus.
- 1295 commonly used Bengali graphemes are selected from 2111256 graphemes available in the dataset.
- To be selected, each grapheme had to occur more than twice in the entire corpus or used in at least two unique words.
- The dataset has three target variables based on their co-occurrence: vowel diacritics, consonant diacritics, grapheme roots.
- Dataset collected from 2896 volunteers.

Kaggle competition

- The metric for the challenge was a hierarchical macro averaged recall: $R = \frac{1}{4}(2R_r + R_{vd} + R_{cd}).$, where the components are for the roots, vowel diacritics and consonant diacritics.
- The winner took a grapheme classification approach where the input images were classified into 14784 (168 x 11 x 8) classes, that is, all the possible graphemes that can be created using the available grapheme roots and diacritics.
- EfficientNet model is used to classify the graphemes.
- Private test set score of the best model was 0.976.

References

[1] S. Alam *et al.*, "Multi-label classification of common Bengali handwritten graphemes: Dataset and challenge," *arXiv*. 2020.