Chittron: An Automatic Bangla Image Captioning System

Name : Md. Musfiqur Rahman

Id: 1620277042

Abstract: Automatic image caption generation aims to produce an accurate description of an image in natural language automatically. However, Bangla, the fifth most widely spoken language in the world, is lagging considerably in the research and development of such domain. Besides, while there are many established data sets to related to image annotation in English, no such resource exists for Bangla yet. Hence, this paper outlines the development of “Chittron”, an automatic image captioning system in Bangla

Critical analysis: This paper reports the development of “Chittron”, an automatic image annotating system in Bangla. As an initial effort to encounter the unavailability of a proper Bangla geo contextual image dataset, the data set of 16,000 images has been produced. It is worthwhile to mention that the images are collected from the public domain of the web with relatively diversified subject matters.

Dataset: Here, used two type of datasets. One is used for pictures and one for the captions which is written in Bengali language.

This is the second data set collected in the BanglaLekha series. The first is BanglaLekha-Isolated, which concentrated on images of isolated Bangla characters. This offering is quite different in its nature, but no less important. Data sets like MSCOCO has 200,000 images, with 5 captions per image.

Roadblock: This system is not fully functional when there are more than two people in same picture. Sometimes it gives caption correctly sometimes not fully correct. So, this is one of major thing to concern.

Conclusions:

This paper reports on the development of “Chittron” , a system to automatically generating Bangla image captions using Deep Neural Networks and the collection of BanglaLekha Image Captions data set consists of 16000 images, with a single overly-descriptive captions per image. This data set is used to train a model employing a pre-trained VGG16 model and stacked LSTM layers. The VGG16 model was used to extract a rich description of the image content as an image embedding vector. This is then used in a model with stacked LSTM layers, which in turn was trained to predict the captions one word at a time. As LSTM layers are effective for sequential data, they are employed to learn a working Bangla language model so that the generated captions can appear to be in natural language.

Literature Cited: Motiur Rahman1, Nabeel Mohammed, Nafees Mansoor and Sifat Momen Department of Computer Science and Engineering, University of Liberal Arts Bangladesh (ULAB), Bangladesh 2,4 Department of Electrical and Computer Engineering, North South University (NSU), Bangladesh Email:nabeel.mohammed@northsouth.edu2, nafees@nafees.info3, [sifat.momen@northsouth.edu](mailto:sifat.momen@northsouth.edu)