Title: Chittron: An Automatic Bangla Image Captioning System

Link: https://arxiv.org/abs/1809.00339

Motivation: Chittron is an automatic image captioning system developed for the Bangla language. It aims to generate accurate descriptions of images in natural language. This paper outlines the development of Chittron and addresses the lack of a proper Bangla image dataset.

Data Set: To overcome the unavailability of a suitable Bangla geo-contextual image dataset, a collection of 16,000 images was produced. These images were gathered from the public domain of the web and cover a diverse range of subjects. Each image was annotated with a single descriptive Bangla caption by a native Bangla speaker.

Model and Training: The proposed model integrates a pre-trained VGG16 image embedding model with stacked LSTM layers. It is trained to predict captions for images one word at a time. The model successfully learns a working language model and generates accurate captions in many cases. The results are evaluated qualitatively, with BLEU scores also measured.

Relevant Work: The paper discusses various existing approaches for image annotation, including search-based and rule-based methods. It also highlights the challenges in ranking descriptions for a given image. The proposed model takes inspiration from successful sequence generation models used in neural machine translation and deep reinforcement learning.

Contribution and Limitations: The results of the proposed system are mainly evaluated qualitatively, comparing the generated captions with human annotations. However, BLEU scores are also reported. The limitations of the current work are discussed, emphasizing the need for a larger and more varied dataset to address grammatical mistakes and inappropriate captions.

Conclusion and Future Works: The paper concludes by highlighting the success of Chittron in generating accurate Bangla image captions. It suggests that better results can be achieved with a bigger and more diverse dataset. Future work includes curating a larger dataset, improving the model's language model, and addressing the limitations identified in the current work.