Paper title: A Visual Attention-Based Model for Bengali Image Captioning

Paper link: https://dl.acm.org/doi/abs/10.1007/s42979-023-01671-x

1. **Summary**

1.1 Motivation:

This model aims to establish the best model to target to exhibit a rigid model that can convert the actual meaning of any image into a flawless textual form like other language captioning as the availability of Bengali images in several social media or the internet is growing day by day.

1.2 Contribution:

They have used two available Bengali datasets which they have benchmarked by curating all the human errors that were present in these datasets and have shown that their applied algorithm InceptionV3 performed better than other models that can satisfactorily generate error-free captions from images.

1.3 Methodology:

The system has comprised two main components: an encoder and a decoder. The encoder, relying on convolutional neural networks (CNN) and featuring an attention module, has processed an RGB image which was provided and transformed into a one-dimensional vector. The CNN has been pre-trained on the Imagenet dataset, serving as a feature extractor. Subsequently, a visual attention mechanism has been applied to highlight essential parts of the image. Finally, a decoder based on recurrent neural networks (RNN) has interpreted the caption, and a combined loss function has been utilized to learn the caption generation process.

1.4 Conclusion:

This paper developed a deep-learning model for generating Bengali image captions. By utilizing two datasets and comparing its algorithm with other algorithms, this model outperformed in BLEU1 and BLEU4 evaluation metrics. It proves to be an efficient choice for Bengali image captioning.

2. **Limitations:**

2.1 First limitation:

The datasets used in this model generating captions contain human bias where this bias poses a limitation, as it hinders the model's capability to accurately create captions for non-human subjects. In other words, the bias in the data impacts the model's generation of non-human entities when generating captions.

2.2 Second limitation:

The captions that are provided are sometimes not in a detailed manner which impacts the training and evaluation accuracy using these datasets. This divergence from expectations suggests a limitation in the dataset's ability to support effective model learning and evaluation.

3. **Synthesis:**

The general findings suggest that this model has the potential to make valuable contributions to regional image-captioning research. Moreover, it can be extended for generating questions from images in regional languages, which shows its versatility and applicability where it opens up a platform to do more research to delve deeper into this topic and discover the more plausible way to create Bengali captions from images.