A comparative study on generating captions from images in Bengali through an encoder-decoder network using a CNN-RNN/LSTM model

Fahim Faisal Rafi
ID:19201081
Dept: CSE
Brac University
Dhaka, Bangladesh
fahim.faisal.rafi@g.bracu.ac.bd

Md. Fahim Haque
ID:20101014
Dept: CSE
Brac University
Dhaka, Bangladesh
md.fahim.haque@g.bracu.ac.bd

Reak Roy
ID:22301776
Dept: CSE
Brac University
Dhaka, Bangladesh
reak.roy@g.bracu.ac.bd

***Abstract*—In recent years, in the field of computer vision, image processing and natural language processing has been combined to shorten the gap between visual information and text information becoming a prominent area of research. This paper focuses on comparing the approaches of different methods in generating descriptive captions in Bengali from images, employing a deep learning approach. As we will see in the comparisons, various methods such as using convolutional neural networks (CNNs) in the encoder block to extract images and using recurrent neural networks (RNNs) in the decoder block to generate the text captions are employed. Our paper endeavors to compare the visual content of images and the accuracy of the generated captions in Bengali that these models have produced.**

***Index Terms*—CNN, RNN, LSTM, encoder, decoder, captions, images**

I. Introduction

The convergence of natural language processing and computer vision has opened up new avenues for image recognition and language production in this age of fast technological advancement. Among them, image captioning is a particularly interesting field that tries to close the semantic gap between textual descriptions and visual data. The rapid advancement of technology necessitates the investigation and adaptation of such innovations to varied linguistic contexts, promoting inclusivity and broadening the scope of these revolutionary technologies. With an emphasis on the Bengali language, this research takes a fascinating look into the field of picture captioning. Even though a great deal of research has been done on caption generation from images using encoder-decoder networks—specifically, using CNN for image feature extraction and RNN or LSTM networks for sequence generation—the peculiarities of Bengali pose special opportunities and challenges. Bengali is a language with unique linguistic traits, cultural quirks, and script complexities that call for a customized strategy for efficient picture captioning. This study aims to perform a comparative analysis to assess how well encoder-decoder networks with CNN-RNN/LSTM architectures perform when producing Bengali captions for images. Through a methodical examination of the efficacy of different model configurations, our goal is to unearth knowledge that advances the capacity for Bengali image captioning, leading to more precise and culturally appropriate descriptions. This comparative study addresses the unique linguistic difficulties presented by the Bengali language in addition to contributing to the expanding body of knowledge in the fields of computer vision and natural language processing in general. In this paper, we hope to expand the use cases of generated captions from Bengali.

II. Literature review

This paper [1] discusses about developing automated image annotation in Bangla called "Chittron". The system uses a pre-trained VGG16 image embedding model alongside stacking LSTM layers to generate captions for images. The model is trained on a dataset of 16,000 images with one descriptive caption for each image. The resulting model is able to generate accurate captions in most of the cases, although there are still some limitations and areas for improvement. Firstly, the data set used for training the model consists of 16,000 images but these are not intended to capture the diversity of Bangla geo-contextual images. Secondly, the evaluation of the model's performance is mainly qualitative, with BLEU scores being measured as an additional metric. However, the BLEU scores are discussed to have limitations in accurately assessing the model's performance.

This paper [2] discusses the qualitative and quantitative analysis of a model for generating captions for images. The qualitative analysis involves evaluating the generated captions based on their quality and similarity to human descriptions. The quantitative analysis compares the performance of different models using metrics such as BLEU, ROUGE, CIDEr, and SPICE. The document also describes the architecture of the proposed model, which includes a CNN-ResNet-50 merged model for image feature extraction and a one-dimensional CNN for sequence processing. The model achieves superior

performance in generating captions, as indicated by the evaluation scores. The document presents the qualitative evaluation of generated captions, comparing them to human descriptions. It also includes the quantitative analysis of different models using metrics such as BLEU, ROUGE, CIDEr, and SPICE. The proposed model uses a CNN-ResNet-50 merged model for image feature extraction and a one-dimensional CNN for sequence processing. The model achieves superior performance in generating captions, as indicated by the evaluation scores.

This paper [3] introduces Transformer-based architecture for automatic Bengali image captioning, achieving excellent results on the BanglaLekhaImageCaptions dataset. Evaluation using BLEU and METEOR scores supports the feasibility of the proposed approach, with both quantitative and qualitative analyses. The study suggests potential improvements in Bengali image captioning and aims to enhance feature extraction and caption precision by adopting visual attention and transformer models instead of ResNet-101. Moreover, the algorithm excels for non-native Bengali speakers by providing both predicted Bengali captions and their English translations. The author claims it outperformed competitors not only in scoring but also in caption quality. Compared to other models, this model accurately identified the most suitable and well-detailed captions. The paper also discusses the challenges and limitations of these algorithms, such as the lack of diversity, coherence, and relevance of the generated captions. It suggests some possible solutions and future directions for improving image captioning, such as using attention mechanisms, multimodal fusion, and reinforcement learning.

Another paper [4] discusses CNN and GRU based attention mechanisms for generating image captions in the Bengali language. An encoder and decoder framework was implemented. The encoder uses a pre-trained CNN to encode the picture, and the decoder uses RNN to generate captions for that image. It combines a transformer-based architecture with a pre-trained ResNet-101, InceptionV3, VGG16, DenseNet169 model for extracting image features. The experiments conducted demonstrate that this approach surpasses existing Bengali image captioning methods and achieves impressive scores on evaluation metrics like Rogue, BLEU, METEOR, CIDER while comparing. This document also integrated the Bahdanau attention model with GRU, enabling focused learning on a particular image segment to enhance overall performance. Moreover, MSCOCO dataset was used, comprising 82,783 training, 40,504 validation, and 40,775 test images, stands as the most extensive benchmark dataset for image captioning tasks. Finally, Inception demonstrated superior performance, surpassing other models, with ResNet101 following closely.

III. Dataset

BanglaLekhaImageCaptions dataset [5] is a comprehensive dataset containing 9,154 images and each of them have been assigned with a caption written in Bengali. These images are also related to Bengali culture and lifestyle. Comparing this dataset with other image caption datasets like Flicker8k, we can see a strong western cultural bias and the captions associated with it are mainly English. So, to generate captions in Bengali, many models have used this dataset to train and better refine the generation captions.

IV. Proposed Methodology

In this paper, we will compare the different models used to generate Bengali captions from images like the CNN-ResNet-50 merged model. We will analyze the generated captions, observe how accurate it is and compare its results with other models using the same dataset. In addition, we will try to employ a model with an augmented dataset to see if it can generate captions any better than the existing models.

References

[1] Rahman, M., Mohammed, N., Mansoor, N., & Momen, S. (2018). Chittron: An Automatic Bangla Image Captioning System (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1809.00339

[2] Khan, M. F., Shifath, S. M. S.-U.-R., & Islam, Md. S. (2021). Improved Bengali Image Captioning via deep convolutional neural network based encoder-decoder model (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2102.07192

[3] Palash, M. A. H., Nasim, M. A. A., Saha, S., Afrin, F., Mallik, R., & Samiappan, S. (2021). Bangla Image Caption Generation through CNN-Transformer based Encoder-Decoder Network (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2110.12442

[4] Khan, R., Islam, M. S., Kanwal, K., Iqbal, M., Hossain, Md. I., & Ye, Z. (2022). A Deep Neural Framework for Image Caption Generation Using GRU-Based Attention Mechanism. arXiv. https://doi.org/10.48550/ARXIV.2203.01594

[5] Mansoor, N., Kamal, A. H., Mohammed, N., Momen, S., Rahman, M. M. (2019), "BanglaLekhaImageCaptions ", Mendeley Data, V2, doi: 10.17632/rxxch9vw59.2