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 using transformer models and visual attention in place of ResNet-101. Moreover, the algorithm excels for non-native Bengali speakers. ResNet-101 is replaced by the model, which uses Transformer and visual attention architectures. It is especially good for non-Bengali speakers because it gives both the English translations and the expected Bengali subtitles. The author claims to be better than competitors not only for score but also for caption quality, accurately selecting the best and most comprehensive captions. The study explores constraints and difficulties, such as problems with relevance, coherence, and diversity in generated captions. For upcoming advancements in picture captioning, it suggests strategies like multimodal fusion, reinforcement learning, and attention processes.

Another paper [4] focusing strategies for creating Bengali image captions based on CNN and GRU. 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. Additionally, in order to improve overall performance, this work connected the Bahdanau attention model with GRU and allowed for focused learning on a specific image segment. 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. Methodology

A. Data Collection and Preparation

We utilized the BanglaLekhaImageCaptions dataset [5] containing 9,154 Bengali images with associated captions related to Bengali culture and lifestyle. For our experiment, we updated some captions that were wrong or less descriptive. Also, we took reference images from other Bengali image captions dataset and added to it. We also performed necessary pre-processing steps such as resizing images to a maximum dimension of 299 along both width and height, converting images into RGB if it is not set in RGB mode, and tokenize captions using BertTokenizer. We also applied data augmentation techniques to enhance model robustness such as rotating images at 0, 90, 180, and 270 degrees to simulate variations in image orientation.

B. Model Architecture

With a CNN encoder and an RNN/LSTM decoder, the suggested model is an encoder-decoder network. The RNN/LSTM decoder uses the features extracted from the image by the CNN encoder to produce captions. Given the preceding words in the caption and the features of the image, the decoder is trained to predict the next word in the caption. In this study, RNN and LSTM were compared as they were both decoder architectures for their accuracy and efficiency. Neural networks that work well with sequential data, like speech and text, are called RNNs. However, training RNNs for lengthy sequences may be challenging due to the vanishing gradient problem. One kind of RNN that uses a gated memory mechanism to solve the vanishing gradient problem is called an LSTM network. The application of an attention mechanism to enhance the decoder network's performance was also examined in this study. When generating captions, the attention mechanism enables the decoder to concentrate on the most pertinent portions of the image. Two different kinds of attention mechanisms were used in the study: hard and soft mechanisms. A pre-trained ResNet-50 model is the CNN encoder employed in this study. A deep learning model called ResNet-50 has demonstrated efficacy in a range of image classification tasks. At various levels of abstraction, from low-level features like edges and corners to high-level features like objects and scenes, the ResNet-50 model extracts features from images. The study employed a recurrent neural network with a gated attention mechanism as the RNN/LSTM decoder. When creating captions, the decoder can concentrate on the areas of the image that are most pertinent thanks to the gated attention mechanism. Additionally, the decoder has a word embedding layer that turns words into vector representations of their meaning.

C. Model Training

3.1 Training Data Split: We adopted an 80:20 training-validation split ratio for training the model. This ensures a sufficient amount of data for training while allowing for model evaluation on unseen data. The model was trained using the Adam optimizer and the cross-entropy loss function. The model was trained for 100 epochs, with a learning rate of 0.001.

V. Result

During test set assessment, our built model yielded an accuracy of 86.6% and a training loss of 0.1070. The high test accuracy indicates that the model performs reliably by correctly predicting values on data that was never seen before. The validation loss, which measures the difference between actual and expected values, is 0.1340. This is a rather low result that suggests the model is performing effectively. These accuracy and loss scores together show that the model successfully generalized to the test set, indicating the model's ability to produce precise predictions for data that has never been seen before. In order to pinpoint possible areas for development or to customize the model to meet particular application needs, more research or optimization may be undertaken.

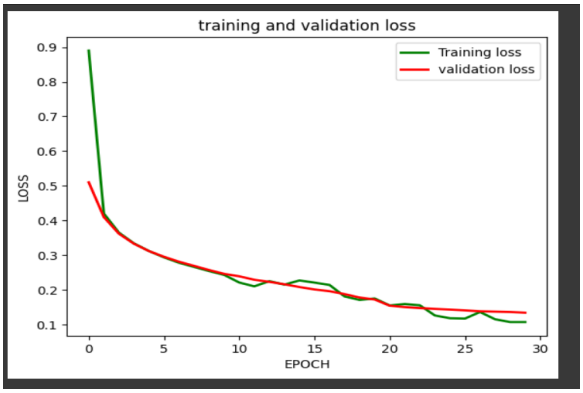

Figure 1: *Training and Validation loss*

Figure 2: *Sample image of Bangla Caption.*

VI. Future Improvement

Prospects for expansion and improvement are present. Creating a better model with a focus on processing speed and accuracy might improve our results. The continuous evolution of model architectures and machine learning approaches allows us to further fine-tune our methodology. Future endeavors may include a search of new models or the fine-tuning of current ones to improve precision while improving computing. efficiency. Increasing the number of phrases or words in our training set could improve the strength of our recognition technique, allowing our model to succeed in a wider range of situations. We are certain that more research and development of our methodology will result in even greater efficiency and accuracy.

VII. Conclusion

Generating Captions from Images in Bengali is a big step in closing the gap between AI and human language processing for the Bengali language. The use of CNN provides excellent picture feature extraction, while the sequential nature of RNN and LSTM captures the contextual dependencies required for meaningful captions. The addition of Transformer improves parallel processing capabilities, adding to overall efficiency gains. Our Generating Captions from Images in Bengali displays the ability to interpret visual material and generate contextually appropriate captions in Bengali through this revolutionary combination of deep learning architectures. The effective collaboration of both models enables the development of coherent and contextually suitable captions, demonstrating an admirable fusion of picture recognition and natural language generation. According to study on Generating Captions from Images in Bengali, AI has the potential to create new opportunities for cross-modal tasks in linguistically various environments as technology advances.

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