Bahdanau Attention Based Bengali Image Caption Generation

1st MD Sahrial Alam
Department of Computer Science and Engineering
Bangladesh University Of Business And Technology - BUBT
Dhaka-1216, Bangladesh
sahrialalam@gmail.com

2nd MD Sayedur Rahman
Department of Computer Science and Engineering
Bangladesh University Of Business And Technology - BUBT
Dhaka-1216, Bangladesh
srsayed2677@gmail.com

3rd MD Ikbal Hosen
Department of Computer Science and Engineering
Bangladesh University Of Business And Technology - BUBT
Dhaka-1216, Bangladesh
mdiqbalhossain6463@gmail.com

4th Khairul Anam Mubin
Department of Computer Science and Engineering
Bangladesh University Of Business And Technology - BUBT
Dhaka-1216, Bangladesh
anam.mubin1999@gmail.com

5th Sharif Hossen
Department of Computer Science and Engineering
Bangladesh University Of Business And Technology - BUBT
Dhaka-1216, Bangladesh
sharifhossen317@gmail.com

6th M. F. Mridha
Department of Computer Science and Engineering
Bangladesh University Of Business And Technology - BUBT
Dhaka-1216, Bangladesh
firoz@bubt.edu.bd

***Abstract*—In the past few years, many works are done in object detection using images and machine translation. Inspired by those works we introduced Bahdanau Attention Based Bengali Image Caption Generation (BABBICG) that generate automatically bangla caption based on images. The Conventional encoder-decoder architectures performance curse will reduce by Bahdanau Attention and achieving momentous improvements over encoder-decoder architectures. In this work, we extract features from images using InceptionV3 neural network and generate caption using RNN decoder. We used Gated Recurrent Unit (GRU) approach as RNN. We evaluate the model using BanglaLekhaImageCaptions dataset from Mendeley Data that can help to generate bangla caption.**

***Index Terms*—Bahdanau Attention, Bengali Image Caption, Mendeley Data, Gated Recurrent Unit.**

I. Introduction

Generating Image caption is a revolutionary step for Computer Vision and Natural Language Processing sector. Image Captioning has been a very difficult task and more complex than object recognition. For image captioning, the first task is object detection. After that generate caption according to images. Recent image captioning models [8] [10] [7] adopted the transformer architectures. That has introduced dot-product attention for achieving state-of-the-art performance. In recent years, many worked done in image caption generation with different languages. But A few of the work done on bangla languages.

In this paper, we propose a model to explore bangla languages caption knowns as bengali caption. We propose a system that has uses InceptionV3 architecture [14] as CNN for extracting features from images and then applying Bahdanau Attention or local attention mechanism on images. The Badhanau Attention mechanism focus on the specific part of images and accurately detect the object in the images. and built-up features vectors. Using Gated recurrent unit (GRU) [2] as RNN for generating a textual caption that describes images. In bangla caption generation, the main difficulty faced is the scarcity of bangla datasets. There aren't huge datasets available on internet. We use BanglaLekhaImageCaptions dataset. The dataset is available in the dataverse of Mendeley [12]. That dataset has 9,154 images with two Bangali analogous human annotations for each image. We evaluate the performance of our model using the BLEU score metric.

II. Literature Review

Image caption generation is a process of constructing a textual caption using artificial intelligence where we can understand the content of an image from the caption. Sen He et al. [6] introduced an image transformer, which consists of a modified encoding transformer and an implicit decoding transformer. Their design widens the original transformer layer's inner architecture to adapt to the structure of images. With only region features as inputs, their model achieved new state-of-the-art performance on both MSCOCO offline and online testing benchmarks. Vinyals et al. [15] proposed a model that uses a CNN pre-trained on ImageNet [3] to encode the image, then an LSTM [5] model is used to decode the image features into a sequence of words and make image caption. Xu et al. [16] introduced an attention mechanism

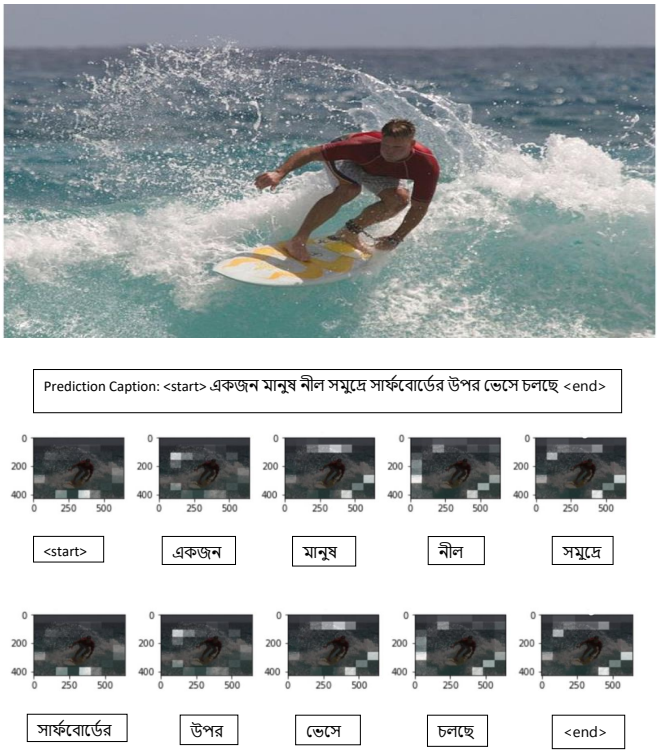

Fig. 1. Visual attention-based model that focuses specific parts of images and it generates a caption – (Image Source; License: Public Domain)

base image caption generation which generate caption during the generation of each word. It is based on the hidden state of their language model and the previously generated word. Their attention module generates a matrix to weight each receptive field in the encoded feature map, and then feeds the weighted feature map and the previously generated word to the language model used to generates the next word of the sentence. Al Momin Faruk et al [4] o proposed a CNN and Bidirectional GRU-based architecture model that generates captions in the Bengali language from an image. They used an encoder-decoder approach to generate Captions. They used a pre-trained model named InceptonV3image embedding model as the encoder for analysis, classification, and annotation of the dataset's images and used the Bidirectional Gated Recurrent unit (BGRU) layer as the decoder to generate captions. Matiur Rahman et al [13] Development of 'Chittron' an automatic image captioning system in Bangla. They have a collection of 16,000 Bangladeshi contextual images that have been accumulated and manually annotated in Bangla. This data set is then used to train a model that integrates a pre-trained model named VGG16 image embedding model with stacked LSTM layers. Their model is trained to predict the caption when the input is an image, one word at a time. AH Kamal et al [9] proposed an automatic image caption model that extracts feature by using VGG16 architecture for CNN . They comprised 25 classes and used LSTM cells as RNN for generating textual caption according to the image.

III. Existing Models

A. Convolutional Neural Network (InceptionV3)

InceptionV3 is one of the most popular convolutional neural network architecture. Convolutional neural network extract different dimensional phases of an image by using feature vectors. In convolutional neural network IncpetionV3 is appealed as the best implementation technique for image classification according to prior literature [9]. So we used InceptionV3 as CNN. InceptionV3 has several improvements including Label Smoothing. According to prior literature, Inception gives the best accuracy for image captioning.

B. Bahdanau Attention Mechanism

The attention mechanism is more effective than encoder-decoder model for machine translation. There are two types of attention mechanisms are global and local attention mechanisms. The local attention mechanisms are also known as the Bahdanau Attention mechanisms. Luong et al [11] proposed a strong distinguish between global and local attention. In the global attention method, attend to all input words. But one the other hand local attention method, attend to only subset of words.

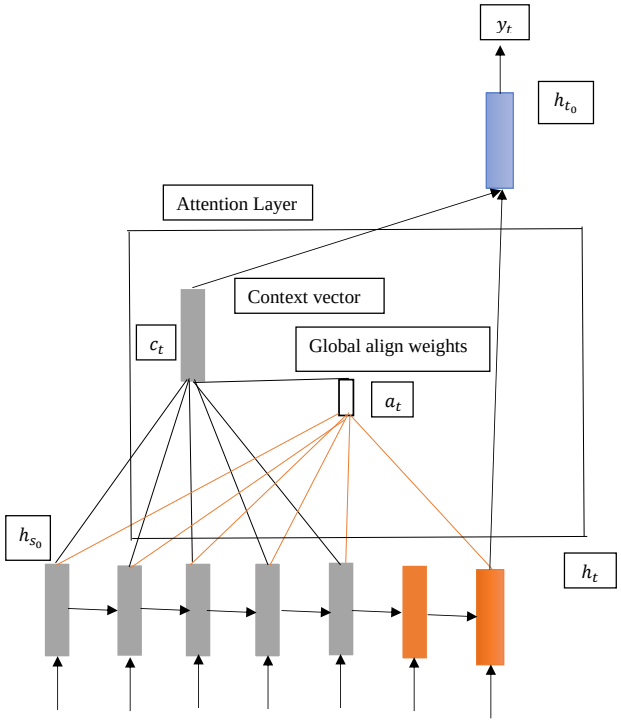

Fig. 2. Global Attentional Model : The global attention model [11] is considered to the encoder's all the hidden layers when deriving the context vector c_t , variable length alignment vector a_t , current target is hidden state h_t , each source hidden state h_{s0}. at each time step t, the model infers a variable-length alignment weight vector at based on the current target state h_t and all source states h_{s0}. A global context vector c_t is then computed as the weighted average, according to a_t, overall the source states

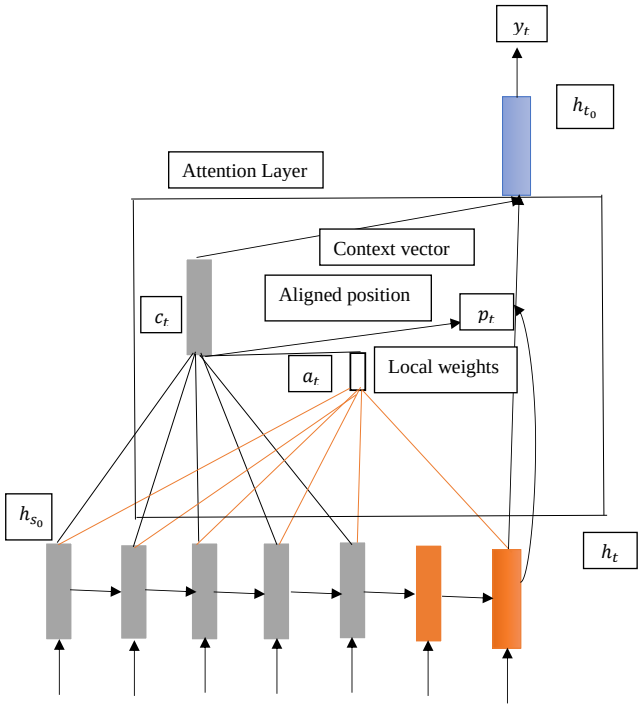

Fig. 3. Local Attentional Model : The Local attentional model [11] first predicts a single aligned position p_t for the current target word. A window centered around the source position p_t is then used to compute a context vector c_t, a weighted average of the source hidden states in the window. The weights at are inferred from the current target state h_t and those source states h_{s0} in the window

C. Word Embedding

Word Embedding is one kind of a method that learn real-valued vector representation for a predefined fixed-sized vocabulary from a corpus of text. There are three techniques that are used to learn word embedding from text data are Embedding layer, word2vec, GloVe. We used the GloVe Word Embedding technique. The GloVe is a way to combine both the global statistics of matrix factorization techniques like LSA with the local context-based learning in word2vec.

D. Recurrent Neural Network

In our paper, we introduced one of the recent RNN known as Gated recurrent unit GRU [1].In recurrent neural networks,GRU is a very gating mechanism. It works like LSTM but the main difference is GRU has fewer parameters than LSTM. GRU gives very good performance in the natural language processing sector. GRU helps to convert vectors using one of word embedding layer is known Embedding layer. GRU is also helps to predict the next word in sequence by previous predicted word.

IV. Dataset

We used BanglaLekhaImageCaptions dataset from Mendeley dataverse. It has contains 9,154 images. Each image has two human annotations description. The dataset has 25 classes. All of the images are not the same size. So we resize all images into 299x299 pixels and keep all RGB channels for information holding as inceptionV3 demand. Here, all the images are stored PNG format. Table I shows that all technical

TABLE I
CHARACTERISTIC OF BANGLALEKHAIMAGECAPTIONS DATASET

		Resizing and Formation	
No of Images	No of Classes	File Format	Image Dimension
9154	25	PNG	299x299

characteristics of our dataset. We split the dataset into three portions are Train data, Validation data, Test, 80% for Train data, 10% for validation data and 10% for testing data.

V. Methodology

In this paper, we divided our model into three parts. Firstly, we collected a dataset from Mendeley Data verse. Then we need to the size all data into 299 x 299 x 3 pixels. We used Inception architecture as Convolution Neural Network. We have 25 classes and inceptionV3 able to pertain 1000 classes. In CNN architecture consist Conv2D, Maxpooling, Dense activation layer. We also used the BatchNormalization layer and SpatialDropout2D layer.The Maxpooling layer is helps us to reduce image dimensions without any noise. Flatten and Fully Connected layers give us feature vectors that helps to generate the next portion of the task. Secondly,

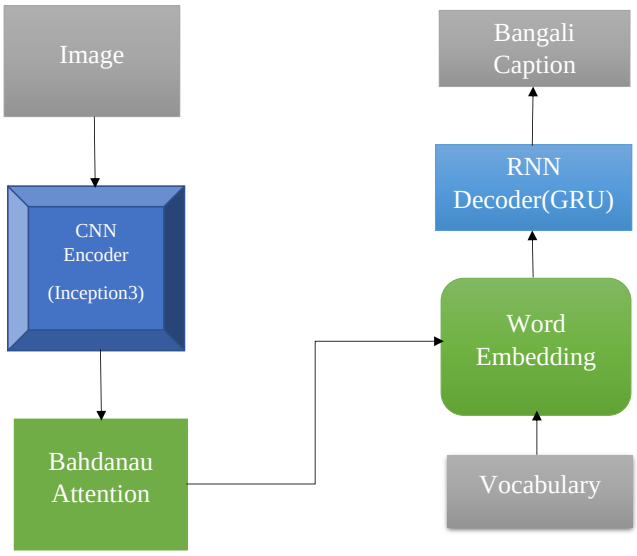

Fig. 4. Methodology of Bengali Image caption with Bahdanau Attention

we used Bahdanau Attention mechanism is only focused on specific portions of images and get weighted image features. Attention hidden layer, Dense and softmax layer helps to get weighted image features We used bangla GloVe Embedding technique for bangla word separation and get a unique index for each unique word. Finally, we used a Gated recurrent unit as a Recurrent Neural Network to generating bengali

caption according to weighted image features. We already include that each image has two bangla description. 256 channels are GRU is used and the dropout layer is set to .25. Batch size = 64, Buffer size = 128, embedding dim= 100 and Optimizer was Adam and estimated loss function used SparseCategoricalCrossentropy. After that, we train our model by 50 epochs. We used kalpurush font for the bangla word. At last we evaluate our caption performance by using BLEU score metrics.

A. Way To Generate Image Caption(Per-Inject)

Inject model is help to generate each word from the text description by an encoded form of the image. There are four types of Inject models are init-inject, pre-inject, per-inject, marge-inject. We choose for our model is per-inject to generate image caption. After finishing the image and word preprocessing phase, word and image get into RNN at the same time. Then RNN generates caption in fig 5 dot box phases according to an image.

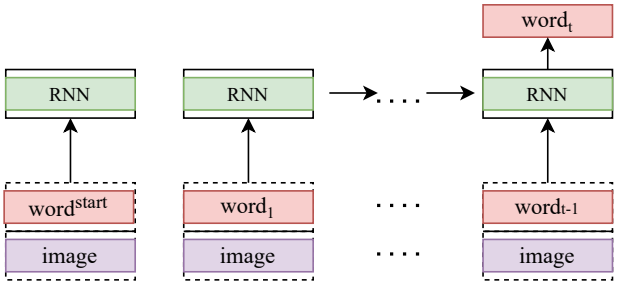

Fig. 5. Way To Generate Image Caption(Per-Inject)

VI. Results And Discussion

Our model has CNN for Encoder, Bahdanau Attention Mechanism for weighed vector features, GloVe Word Embedding for text preprocessing, RNN for Decoder. We trained our model by 50 epochs and we got train loss and validation loss for each of the epochs. Fig 6 shows us our model train-loss decreasing when epochs increase. So we can say that the model is perfectly trains and Validation loss increase when epochs increasing. We evaluate our model by BLEU score metrics. We used BLEU-2, BLEU-3, BLEU-4 for measurement the performance. Fig 7 shows that Bangali caption generation and evaluation scores. However, most of the image captions received the best accuracy and were syntactically correct. This was possible objection detection and generating caption as a following model using Bahdanau Attention mechanism.

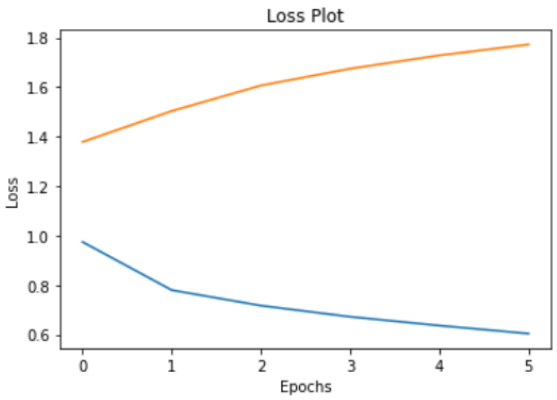

Fig. 6. Blue color is Train Loss and Lite Orange color is Validation Loss

Fig. 7. Illustration of Bengali captions according to images and the evaluation scores using our model

A. Conclusion

In this paper, We have presented the Automatic Bangali caption system. We used 9154 image that has 18,308 human annotations in Bangla. We used Bahdanau Attention mechanism to focus specific portions of an image. Per-inject model is used our model for generating image caption. We evaluate our model using the BLEU score that gave the best possible accuracy. In the future we will collect more images dataset that only base on Bangla Culture Base Images. We think that the dataset will give more accurate and better performance.

References

[1] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[2] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[4] Al Momin Faruk, Hasan Al Faraby, Md Azad, Md Fedous, Md Morol, et al. Image to bengali caption generation using deep cnn and bidirectional gated recurrent unit. *arXiv preprint arXiv:2012.12139*, 2020.

[5] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451–2471, 2000.

[6] Sen He, Wentong Liao, Hamed R Tavakoli, Michael Yang, Bodo Rosenhahn, and Nicolas Pugeault. Image captioning through image transformer. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[7] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. *arXiv preprint arXiv:1906.05963*, 2019.

[8] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4634–4643, 2019.

[9] Abrar Hasin Kamal, Md Asifuzzaman Jishan, and Nafees Mansoor. Textmage: The automated bangla caption generator based on deep learning. In *2020 International Conference on Decision Aid Sciences and Application (DASA)*, pages 822–826. IEEE, 2020.

[10] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. Entangled transformer for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8928–8937, 2019.

[11] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[12] Nafees Mansoor, Abrar Hasin Kamal, Nabeel Mohammed, Sifat Momen, and Md Matiur Rahman. Banglalekhaimagecaptions. *Mendeley Data*, 2, 2019.

[13] Matiur Rahman, Nabeel Mohammed, Nafees Mansoor, and Sifat Momen. Chittron: An automatic bangla image captioning system. *Procedia Computer Science*, 154:636–642, 2019.

[14] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[15] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[16] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.