Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 154 (2019) 636–642

Procedia Computer Science

www.elsevier.com/locate/procedia

8th International Congress of Information and Communication Technology, ICICT 2019

Chittron: An Automatic Bangla Image Captioning System

Matiur Rahman[a], Nabeel Mohammed[b*1], Nafees Mansoor[a], Sifat Momen[b]

[a]*University of Liberal Arts Bangladesh, Dhanmondi, Dhaka, Bangladesh*
[b]*North South University, Bashundhara, Dhaka, Bangladesh*

Abstract

Automatic image caption generation aims to produce an accurate description of an image in natural language automatically. How- ever, Bangla, the fifth most widely spoken language in the world, is lagging considerably in the research and development of such domain. Besides, while there are many established data sets related to image annotation in English, no such resource exists for Bangla yet. Hence, this paper outlines the development of "Chittron", an automatic image captioning system in Bangla. To address the data set availability issue, a collection of 16, 000 Bangladeshi contextual images has been accumulated and manually annotated in Bangla. This data set is then used to train a model that integrates a pre-trained VGG16 image embedding model with stacked LSTM layers. The model is trained to predict the caption when the input is an image, one word at a time. The results show that the model has successfully been able to learn a working language model and to generate captions of images quite accurately in many cases. The results are evaluated mainly qualitatively. However, BLEU scores are also reported. It is expected that a better result can be obtained with a bigger and more varied data set.

Keywords:BanglaLekha-Image-Caption dataset ; Image Annotation ; Automatic Caption Generation ; LSTM ; Deep Learning ; Neural Networks ; Machine Learning*; Bangla; Natural Language Processing*

1. Introduction

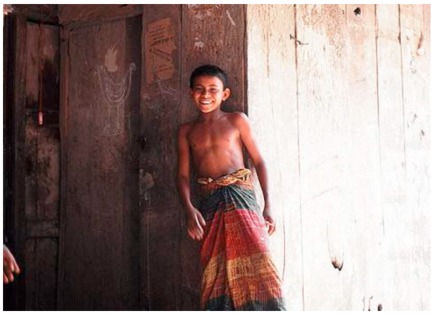
While an image may well be worth 'a thousand words', it is a difficult task to describe an image with a lot of important details. Instead, what is useful in many applications is an adequate description of an image comprising of the essential information. Therefore, the automatic methods of image captioning aim to do just that, already have

[1] Email address: nabeel.mohammed@northsouth.edu (Nabeel Mohammed)

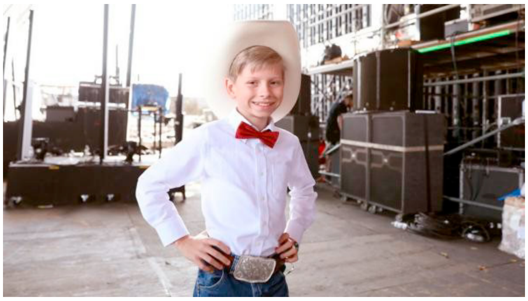
10.1016/j.procs.2019.06.100

major impacts in various fields, e.g. image search. Furthermore, it has the potential to influence positive changes in many different areas, including software for disabled individuals, surveillance & security, human-computer interaction etc.

The stark reality is that most of the works in image captioning have concentrated almost exclusively on English language [1, 2, 3]. Additionally, the relevant data-sets, e.g. the MSCOCO [4], have a prominent western preference which leads to a two-pronged problem: (1) the language in which captions are generated is English only, and (2)the data set is not representative of the cultural peculiarities of non-western countries. These very problems exist for generating image captions in Bangla, particularly for images which have a decidedly Bangla geocultural flavor. A simple example of this can be seen in Figure 1, where a web service is used to generate captions. The service uses the im2txt model trained on the MSCOCO data set and quite clearly the model fails to recognize the image in Figure 1a as a boy wearing a lungi, a very common male garb in Bangladesh. In fact, it incorrectly identifies the subject as a female since the attire is identified as women gown. On the other hand, the model shows quite an impressive performance by very precisely describing the image in figure 1b. As depicted in the figure, the model not only rightly identifies the subject to be a boy, it also accurately describes what the subject is wearing (a bow tie).



(a)tion generated: A woman standing in front of a mirror holding a teddy bear.



(b)tion generated: A young boy wearing a bow tie.

Figure 1: Example of western bias in existing data sets and captioning models

Taking into account the present state and the challenges, this paper reports the development of ‘Chittron’, an automatic image annotating system in Bangla. As an initial effort to encounter the unavailability of a proper Bangla geo-contextual image dataset, the data set of 16, 000 images has been produced. It is worthwhile to mention that the images are collected from the public domain of the web with relatively diversified subject matters. Next, a native Bangla speaker annotated each image with a single overly-descriptive Bangla caption. This data set is then used to train a model similar to the im2txt model [1]. The proposed model integrates a pre-trained VGG16 image embedding model with stacked LSTM layers. The model is trained to predict the caption when the input is an image, one word at a time. The results show that the model is successfully able to learn a working language model and to generate captions of images quite accurately in many cases. The results are evaluated mainly qualitatively. However, BLEU scores are measured and the limitation of the BLEU scores are also discussed. The shortcomings of the current work is discussed, most of which can be addressed by creating a larger more varied data set.

The rest of the paper is organized as follows. In section II, relevant work in image captioning is discussed while the prepared data set is described in section III. The model and training details is presented in section IV. Results of the proposed system are presented in section V. Finally, conclusion and future works have been highlighted in section VI.

1. Relevant Work

A naive approach towards generating the caption of an image is perhaps to predict words from the image regions and link them up. One of the first image annotation system has been developed back in 1999 [5]. Later, in 2002 the image captioning task has been re-cast as that of machine translation [6]. However, it turns out that the technique fails to perform properly for a number of reasons including simple mapping of an image to a word completely overlooks any kind of relationships that exist between the objects in the image. Moreover, arranging the annotations in the caption becomes very difficult in such systems.

Sentences are richer than a set of individual words since a sentence can express actions going on in an image. It also shows the relationships among different entities in the image. Additionally, sentences are also expected to begrammatically correct and natural sounding in the target language. The system presented in [7] shows that using the spatial relationship between objects improves both the quality of generated annotation and its placings.

The work presented in [1] pursued the second machine learning approach where a pretrained convolutional neural network[8] is used to extract a rich image embedding vector, which is then used in a recurrent neural network (RNN) (particularly stacked LSTM layers)[9], for sequence modeling. This model treats the captions as sequences to be predicted one word at a time and aims to learn a languages model directly from the data. This model is inspired by the successes of sequence generation witnessed in neural machine translation [10, 11, 12]. Other similar works include that of [2] which uses a recurrent visual representation of an image and [3] which uses deep reinforcement learning for the image captioning task. These systems are trained end-to-end, as in, no human intervention or human engineering is done except the network architecture. These networks are then trained to generate captions directly from images as inputs, using different strategies.

All the mentioned approaches are developed for English languages. Hence, this research identifies the necessity of a Bangla Captioning System. In this paper, the model trained and used to generate Bangla image captions is similar to that of [1].

2. BanglaLekha-Image Captions: The Data Set

This is the second data set collected in the BanglaLekha series. The first is BanglaLekha-Isolated [13], which con- centrated on images of isolated Bangla characters. This offering is quite different in its nature, but no less important.

Data sets like MSCOCO has 200, 000 images, with 5 captions per image. BanglaLekha-ImageCaptions is significantly smaller in size, with only 16, 000 images, all collected from public domain of the web with relatively diversified subject matters. Almost all the images are related to Bangladesh in some way, with some being relevant to the wider Indian Subcontinental context. For each image, a native Bangla speaker is tasked with writing a caption. As multiple captions per image were not possible due to different constraints, the annotator was instructed to write overly-descriptive captions (see the human captions in Figures 3, 4 for some example captions of images and Tables 1 and 2 for their corresponding English translation).

Careful analyses of the captions show that the data set contains 6035 unique Bangla words, where words with the same root but different prefixes and/or suffixes are counted as different words. Numerals are also counted as individual words under this scheme.

3. Model and Training Details

In this section, we discusses the model used in the proposed system. The discussion has been split into two parts where one concentrates on the model preparations and the other one focuses on the training environment.

1.1. Model description

In our proposed model, the captioning task is divided into two broad parts, (a) extracting relevant image features, and (b) generating a language description using the features. The proposed model has similarities with existing successful models, e.g. im2txt, NeuralTalk [1], in terms of use of pre-trained model to extract image embeddings. Later, the model uses a one-word-at-a-time strategy to predict caption from stacked LSTM layers. The widely used VGG16 [14] model, with slight adjustments, is used as the pre-trained image model in our work.

Figure 2 depicts the details of the model used in the training phase. The last layer of the pre-trained VGG16 model is discarded, as the requirement is to get an image descriptor, not probability distributions. The model has two inputs: the first is the image itself and the second is a sequence of tokens (corresponding to each unique word in the vocabulary). An embedding layer accepts the tokens as input and generates corresponding word embeddings [15]. The output of the second last layer is reduced to 512 dimensions through a fully connected layer. Next, the model concatenates with the output of the embedding layer, whose output is also 512 dimensions. The concatenated data forms the sequence data input for the stacked LSTM layers. In total, the sequence data fed into the stacked LSTM layers is a sequence of n + 1 embeddings, where n is the maximum length of the generated caption.

This work predicts captions of up to 10 tokens in length. Thus, for each image-caption pair, the caption is first truncated or extended to 10 tokens. The extension is done by padding tokens with the unknown token. Once, all the aptions are of the required length, for each image-caption pair, 10 training data points are created. The first data point has no tokens, i.e. all tokens are the unknown token, in which case the model is expected to predict the first word from the image embedding alone. In each data point an increasing number of tokens are added, so that the last data point has the first n 1 tokens of the caption, expecting the model to predict the last nth token.

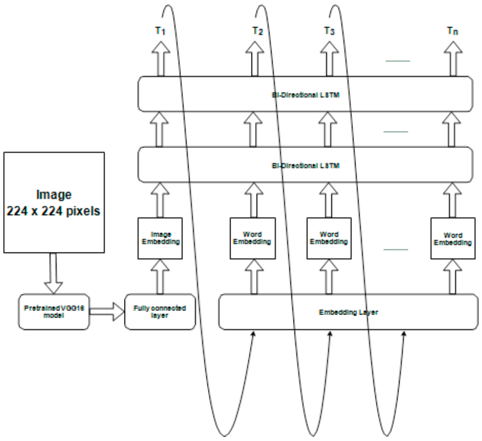


Figure 2: The image captioning model employed

The model is trained on 15, 700 images from the collected data set, resulting in 15, 700 10 training samples. 300 images are considered as the test data.

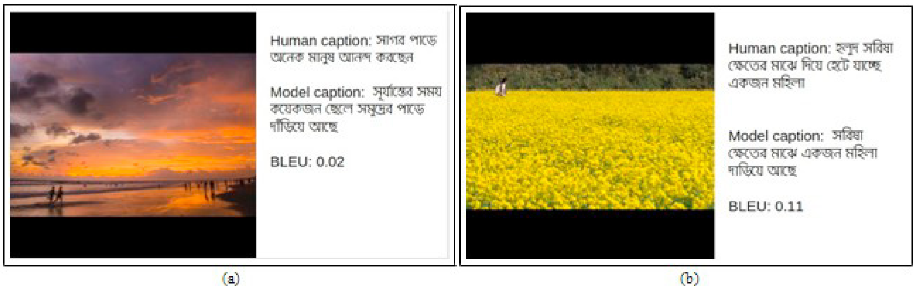
The proposed model is trained end-to-end using the back propagation method. Basic Stochastic Gradient Descent is used to minimize the categorical cross entropy of the output of the stacked LSTM layers.

4. Results and Discussion

Both quantitative and qualitative results are discussed in this section. Quantitative results are presented in terms of the BLEU score [16], which is a widely used metric for evaluating machine translation systems.

The model achieved an average BLEU score of 2.5, which is admittedly not impressive. However, BLEU scores are usually calculated in cases where multiple reference sentences are available. As the current data set only has a single caption per image, BLEU scores are not necessarily a good indication of performance. The qualitative assessment presented below will further bolster this notion.

Figures 3 and 4 show two images each with their corresponding human annotated caption as well as their model generated captions. In all cases, the BLEU score is unsatisfactory. However, for the three images in Figure 3, the generated captions are grammatically correct and appropriate. For those who cannot read Bangla, translations are given in Table 1.


Human caption: সাগর পাড়ে অনেক মানুষ আনন্দ করছেন

Model caption: সূর্যাস্তের সময় কয়েকজন ছেলে সমুদ্রের পাড়ে দাঁড়িয়ে আছে

BLEU: 0.02

Human caption: হলুদ সরিষা ক্ষেতের মাঝে দিয়ে হেঁটে যাচ্ছে একজন মহিলা

Model caption: সরিষা ক্ষেতের মাঝে একজন মহিলা দাড়িয়ে আছে

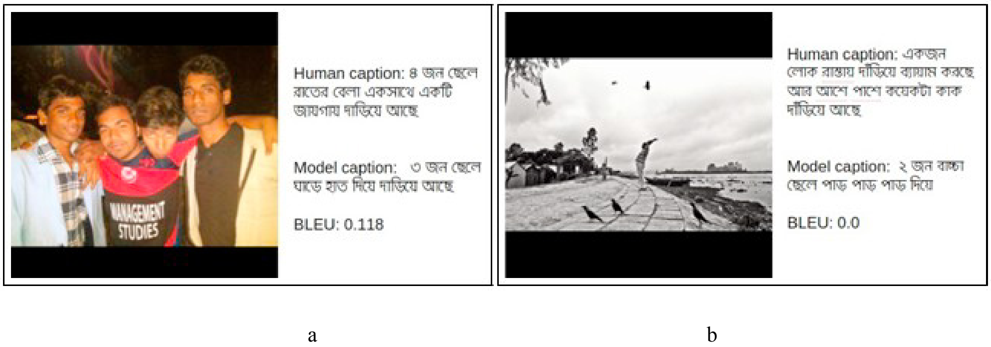
BLEU: 0.11

Figure 3: Examples of appropriatelygenerated captions

Table 1: Translations of captions inFigure3

Subfigure	Human annotation	Model Generated
3a	A lot of people are having fun in the sea shore	A few boys are standing on the sea shore at sunset
3b	A woman is walking through a field of yellow mustard	A woman is standing in he middle of a mustard field

While the BLEU scores are quite low, the model predictions are indistinguishable from what may be generated by a human. Figure 4 shows two images where the generated captions either have mistakes or are entirely inappropriate. These also have low BLEU scores, with the lowest being zero. Table 2 shows the corresponding translations and includes a comment about what is wrong with the generated caption.


Human caption: ৪ জন ছেলে রাতের বেলা একসাথে একটি জায়গায় দাড়িয়ে আছে

Model caption: ৩ জন ছেলে ঘাড়ে হাত দিয়ে দাড়িয়ে আছে

BLEU: 0.118

Human caption: একজন লোক রাস্তায় দাঁড়িয়ে ব্যায়াম করছে আর আশে পাশে কয়েকটা কাক দাঁড়িয়ে আছে

Model caption: ২ জন বাচ্চা ছেলে পাড় পাড় পাড় দিয়ে

BLEU: 0.0

Figure 4: Examples of inappropriate or wrong generated captions

These examples demonstrate that it is possible to learn a working language model of Bangla entirely from human annotated captions. The shortcomings presented in Figure 4 can be addressed by adding more captions per image in the data set and also increasing the number of images so that a larger domain of objects and actions are included for the model to learn. The case of the incomplete sentence can be traced back to the fact that training was only done using the first 10 tokens of the captions and larger captions were truncated. This has lead to a situation where the model learns to create incomplete sentences.

Table 2: Translations of captions in Figure 4

Subfigure	Human annotation	Model Generated	Comment
4a	Four boys are standing together in a place at night.	Three boys are standing together with hands on each other's shoulders.	Wrong count of boys
4b	A man is exercising on the street and a few crows are standing nearby.	Two baby boys on the shore.	The word for shore repeated three times, Incomplete sentence, Not relevant to the image

5. Conclusion

This paper reports on the development of "Chittron", a system to automatically generate Bangla image captions using Deep Neural Networks and the collection of BanglaLekha-ImageCaptions data set consists of 16000 images, with a single overly-descriptive captions per image. This data set is used to train a model employing a pre-trained VGG16 model and stacked LSTM layers. The VGG16 model was used to extract a rich description of the image content as an image embedding vector. This is then used in a model with stacked LSTM layers, which in turn was trained to predict the captions one word at a time. As LSTM layers are effective for sequential data, they are employed to learn a working Bangla language model so that the generated captions can appear to be in natural language.

Both quantitative and qualitative evaluation are done to analyze the results. BLEU scores used for quantitative evaluation is found to be not appropriate for this particular case as the data set only includes a single reference caption per image. Qualitative evaluation demonstrates that in cases the generated captions are on par with human annotations. Such cases clearly demonstrate the capacity of such models to not only learn an effective language model but one which can be conditioned upon image content, making it effective for image captioning. However, as the data set is small, there are also cases where the model makes grammatical mistakes or generates entirely inappropriate captions. These can be remedied by curating a larger, more varied data set with multiple captions per image which will also allow for more effective quantitative assessment of the work. This is scheduled to be done in the future.

6. References

2. Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. IEEE transactions on pattern analysis and machine intelligence 2017;39(4):652–63.
3. Chen X, Lawrence Zitnick C. Mind's eye: A recurrent visual representation for image caption generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015:2422–31.
4. Rennie SJ, Marcheret E, Mroueh Y, Ross J, Goel V. Self-critical sequence training for image captioning. In: CVPR; vol. 1. 2017: 3.
5. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dolla´r P, Zitnick CL. Microsoft coco: Common objects in context. In:
6. European conference on computer vision. Springer; 2014:740–55.
7. Mori Y, Takahashi H, Oka R. Image-to-word transformation based on dividing and vector quantizing images with words. In: First International Workshop on Multimedia Intelligent Storage and Retrieval Management. Citeseer; 1999:1–9.
8. Duygulu P, Barnard K, de Freitas JF, Forsyth DA. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: European conference on computer vision. Springer; 2002:97–112.
9. Gupta A, Davis LS. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In: European conference on computer vision. Springer; 2008:16–29.
10. LeCun Y, Bengio Y, Hinton G. Deep learning. nature 2015;521(7553):436.
11. Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation 1997;9(8):1735–80.
12. Cho K, Van Merrie¨nboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:14061078 2014;.
13. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:14090473 2014;.

14. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. 2014:3104–12.
15. Biswas M, Islam R, Shom GK, Shopon M, Mohammed N, Momen S, Abedin A. Banglalekha-isolated: A multi-purpose comprehensive dataset of handwritten bangla isolated characters. Data in brief 2017;12:103–7.
16. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556 2014;.
17. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In:
18. Advances in neural information processing systems. 2013:3111–9.
19. Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics; 2002:311–8.