Image Captioning in Indian Languages

A thesis report submitted for BTP phase I

by

Rishabh Sharma

(Roll No. 180102058)

Under the guidance of

Prof Prithwijit Guha

DEPARTMENT OF ELECTRONICS & ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

November 2020

Abstract

Captioning of an image with proper descriptions has become an fascinating problem. It connects computer vision and natural language processing. There are many APIs and automatic captioning bots for captioning an image in the English language but we aim to extend it to a multilingual level. We explore the show and tell image caption generation model along with translater APIs which attempt to generate captions for any language, ignoring the English output during evaluation. In this report, we describe a model based on a recurrent design that mixes the recent developments in computer-vision and machine learning which are often accustomed to generate caption describing a picture. This model is trained to maximise the chance of the description of target sentence in any language based on the dataset of image.

Contents

Chapter 1

Introduction

Naturally producing descriptions of a picture is a near the core of scene understanding — one of the essential objectives of computer vision. Having the option to generate narration of the contents of any picture alongside appropriately framed sentences in various dialects is a difficult errand, yet it could have a huge effect, for example by aiding outwardly disabled individuals to better comprehend the content of pictures on the web. Not exclusively should the captions training models be incredible enough to comprehend the computer-vision difficulties of deciding which objects are present in the picture, they should also be equipped for catching and communicating their connections in that particular language. Thus, the generation of captions has long been viewed as a difficult problem. One of the main motivation behind image captioning would be to produce a gadget for the visually-challenged individuals which would lead them in traveling on the roads without the help of any other individual. We can achieve this by converting the view into captions and then the captions into audio outputs. Both of these are now well-known applications of Deep Learning.

There have been numerous papers which have proposed to join together existing arrangements of the above sub-issues, to go from a picture to its portrayal. Conversely, we might also want to introduce in this work a solitary joint the model that accepts a picture I as information and is prepared to amplify the probability p($S|I$) of delivering an objective succession of words S = S_1, S_2, S_3 . . . where each word S_t comes from guaranteed word reference, that portrays the picture enough. The fundamental motivation for our work comes from ongoing advances in machine interpretation, where the important task is to change a sentence S wrote in the source language, into its interpretation T in the objective language, by maximizing p($T|S$).

An encoder RNN peruses the source sequence and changes it into a rich vector(fixed-

length) portrayal, which thus is utilized as an underlying shrouded condition of a decrypter RNN that produces the objective sequence [2] [3]. Here, we propose to follow this formula, by supplanting the encoder recurrent neural network with a profound convolution neural network. For a long time, machine interpretation was likewise accomplished by a progression of independent errands (deciphering words separately, adjusting words, reordering, and so on), however ongoing work has indicated that interpretation should be possible in a lot less complex way utilizing Recurrent Neural Networks and still arrive at best in class execution. Thus, it is clear to utilize a convolutional neural organization as a picture encoder by pre-preparing it for an image arrangement undertaking and afterward utilizing its back layer(hidden) as a contribution to the Recurrent neural organization decoder that produces captions. This model is otherwise called the Neural Image Caption (NIC).

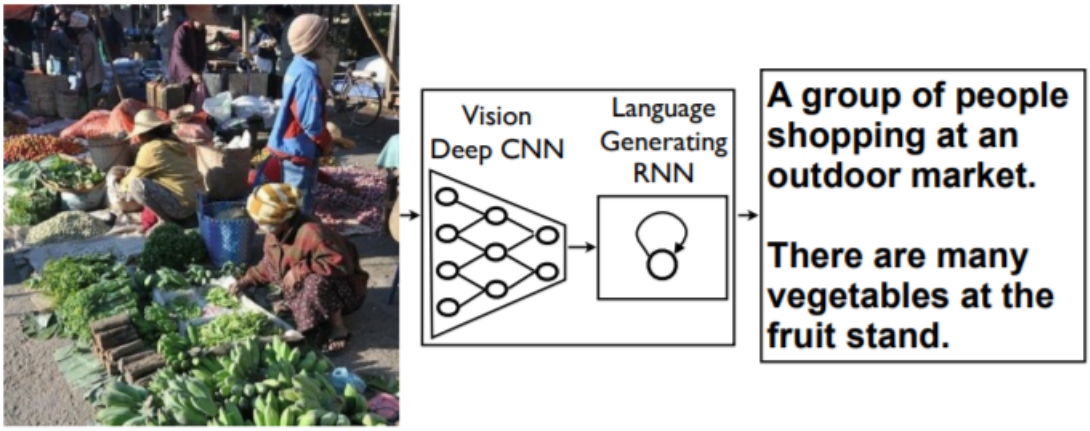


Figure 1. NIC, our model, is based end-to-end on a neural network consisting of a vision CNN followed by a language generating RNN. It generates complete sentences in natural language from an input image, as shown on the example above.

Chapter 2

Literature Review

A lot of work has been done on the picture inscription problem. The primary critical work in explaining picture inscribing undertakings was finished by Ali Farhadi [8] where three spaces are characterized in particular the picture space, which is the meaning space, the sentences space where the planning is accomplished from the separate picture and sentence space to the meaning space. With the assistance of planning, the closeness linking the photos and the caption is assessed, the implications are put away as trios of (picture, activity, object) and a score is assessed by foreseeing the picture and sentence trios. On the off chance that an image and sentence have an elevated level of similitude as far as the anticipated trios then they will be profoundly viable and have a high score. Along these lines, fitting sentences can be produced. This model has numerous downsides, for example, the prerequisite of the central importance space and the outcomes acquired from it are not in any way exceptionally precise.

It is a significant test for ML algorithms, as it sums to emulating the wonderful human capacity to pack immense measures of notable visual data into graphic language. Undoubtedly, a depiction must catch not just the objects contained in a picture, however, it likewise should communicate how these articles identify with one another also as their qualities and the exercises they are engaged with.

Different works were presented yet later work utilizes the procedure of neural organizations for settling the undertaking. With the approach of Convolutional Neural Networks and Recurrent Neural Networks, a decent presentation was accomplished and discovered applications in different fields of study. Cho and O.Vinyals, in different works [**?**] [3], presented a novel methodology of utilizing CNNs and RNNs for picture inscription tasks. Convolutional Neural Networks (CNN) were utilized to separate highlights from the pictures. Along these lines, CNN

goes about as an encoder, fundamentally for the order of errands, and the rear layer's yield is given as the info to RNN. (RNN) goes about as a decoder that produces a sequence of words. LSTM networks (Long Short Term Memory) was the kind of RNN utilized..

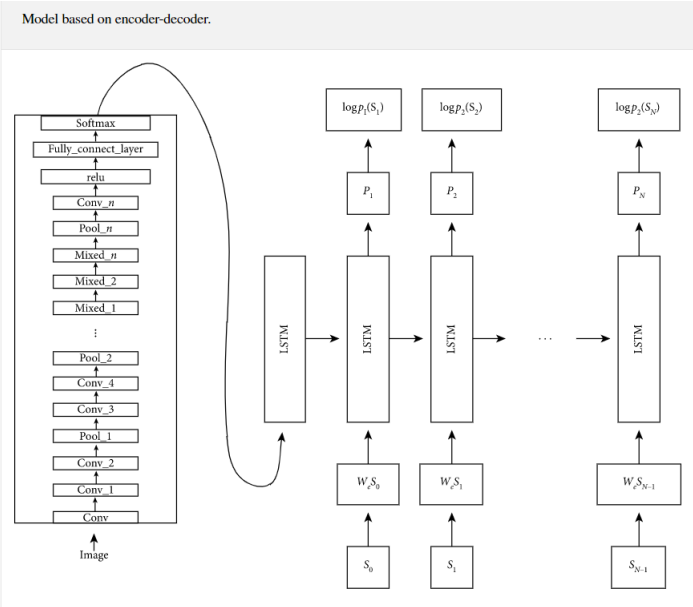


Figure 2.1: This is an image from a text that uses color to teach music.

Chapter 3

Implementation of Show and Tell

In this report, we describe a neural and probabilistic structure to create depictions from pictures [1] and our implementation of the same. Ongoing advances in machine interpretations have demonstrated that it is absolutely possible to accomplish best-known results by just straightforwardly boosting the likelihood of the right interpretation given a sentence in an "start to finish" design – both for training and deduction. The proposed model can be commonly depicted in two main sections. In the fundamental segment,the design of data that includes the picture and the train captions are encoded as include vectors. The model utilizes RNN that encodes the varying length information to a vector(with dimensions fixed) and utilizes this portrayal to decipher it to the wanted yield sentence. We utilize a similar methodology where, given a picture (rather than an input sequence of words in the source language), one applies a similar guideline of "translate" it to its appropriate depiction. Thus, we have to boost the probability of the correct depiction in the given picture by utilizing the following formula:

$$\theta^* = \arg max_\theta \sum_{I,S} \log p(S/I;\theta) \tag{3.1}$$

here theta is the argument of our model, I is a picture, and S is training caption data. S describes any sentence(unbounded length). Subsequently, it is entirely expected to apply the chain rule to show the joint Probability over S_0,S_1,...,S_N ,here N is length of the specific model

$$\log p(S|I) = \sum_{i=0}^{N} \log p([S_t/I, S_0, S_1, S_2.., S_{t-1}]) \tag{3.2}$$

For the implementation of the above model, we need to use a Recurrent Neural Network (RNN), with a different count of words and conditions upon t-1. the memory at every stage must be updated using a non-linear function f. RNN is a group of neural organizations that permit past

yields to be utilized as data sources while having shrouded states.

$$h_{t+1} = f(h_t, x_t) \tag{3.3}$$

here h_t represents a fixed state or memory(hidden length). The memory-block can also be replaced by another network or graph, if that incorporates time delays or has feedback loops. For the implementation of the function f in above equation we can utilize a specific type of RNN, i.e. LSTM, was presented and tried with incredible accomplishment to translation [3] and seq. formation.

In LSTM the memory block contains a cell c which is constrained by three gates. (See in the figure 3.1) . In blue color we represent the recurrent associations – the yield m at time t - 1 is fed-back to the memory at time t through the three entryways. The cell esteem is fed back through the forget-gate. The anticipated word at time t-1 is fed back with the memory yield m at time t into the softmax for word forecast.

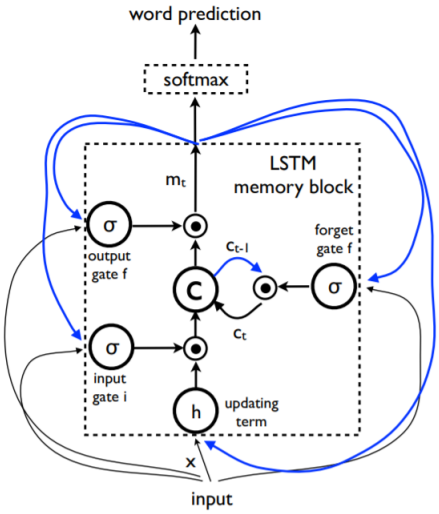


Figure 3.1: Description of LSTM

Training Details

To encode the image i as a feature vector, we use a ResNet-34 pre-trained on ImageNet with fixed weights. Except CNN, all other weights were randomly initialized. Adam optimizer with initial learning rate of 1×10^{-3}. Learning rate annealed by 0.8 after every 8 epochs of no

improvement on BLEU-4. Training for 120 epochs with early stopping on BLEU-4. At test, We used the BeamSearch approach with a beam size of 20.

The LSTMs used to encode the two inscriptions share similar weights. The weights of underlying word implanting just as of the LSTM are constantly refreshed while training. When the encoded highlight vectors are figured, they are joined into a solitary vector. Specifically by discharging the stop word the LSTM signs that a complete sentence has been created. Both the picture and the words are mapped to a similar space, the picture by utilizing CNN, the words by utilizing word embeddings W_e.

In the end, the sentence sequence resulted by the LSTM is converted to Hindi by the utilization of Googletrans library similarly it can also be translated from any source language to any other target language.

Chapter 4

Experimentation & Results

Dataset

A decent dataset to utilize when beginning with picture captioning is the Flickr-8K dataset. The explanation is that it is reasonable and moderately little so you can download it and manufacture models on your own PC. Many datasets consist of images and sentences in English describing the images. We have used the Flickr8k dataset for evaluation. It contains a sum of 8092 pictures in JPEG design with various shapes and sizes. Of which 6000 are utilized for training, 1000 for testing, and 1000 for advancement. Each picture in these datasets is related to five unique inscriptions that portray the elements and occasions portrayed in the picture that were gathered by means of publicly supporting (Amazon Mechanical Turk). The size of the preparation-vocab is 7371. By partner, each picture with numerous, autonomously created sentences, this dataset catches a portion of the etymological assortment that can be utilized to depict a similar picture.

Performance Metrics

The assessment of picture subtitling models is commonly performed utilizing measurements, for example, BLEU [4], METEOR [6], ROUGE [7] or CIDEr [5], all of which fundamentally measure the word cover among created and reference captions.These are calculations, which have been utilized for assessing the nature of machine deciphered content.

Model Implemented	Bleu1	Bleu2	Bleu3	Bleu4	ROUGE	METEOR	CIDEr
ResNet34 + LSTM	46.1	29.1	17.8	11.2	34.8	15.7	17.8

Results using Performance Metrics Discussed

In our experiments, we used the predefined splits of Flickr-8k dataset. 6000 are utilized for training, 1000 for testing, and 1000 for advancement. ResNet34 and loss of Inception-v3 could be used for encoder part and LSTM and GRU for decoder part.We chose Resnet34 and LSTM for the experiment. The results obtained are as follows :

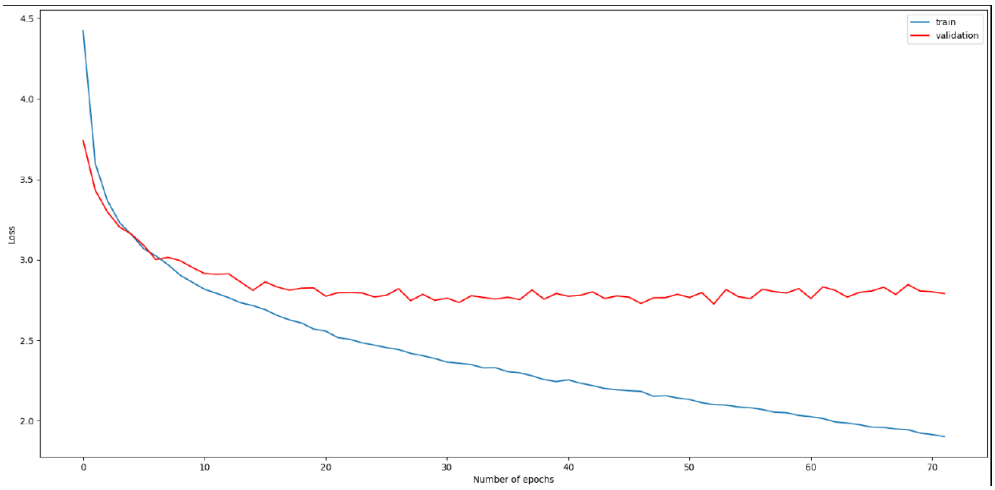
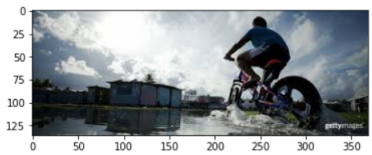


Figure 4.1: Loss Resnet32 + LSTM

Figure 4.2: Example-1

a black and white dog is standing on the grass
Translated(src=en, dest=hi, text=एक काला और सफेद कुत्ता घास पर खड़ा है, pronunciation=ek kaala aur saphed kutta ghaas par khada hai

Figure 4.3: Example-2



a man on a bike is doing a trick on a dirt bike
Translated(src=en, dest=hi, text=बाइक पर एक आदमी गंदगी बाइक पर एक चाल चल रहा है

Figure 4.4: Example-3

a little boy in a red shirt is sitting on a red slide
Translated(src=en, dest=hi, text=लाल शर्ट में एक छोटा लड़का लाल स्लाइड पर बैठा है,

Figure 4.5: Example-4

Chapter 5

Conclusions and Future Intends

In this Report, We described a simple neural image caption generator which could be used to describe the contents of a picture with appropriately formed sentences in a particular language based on the dataset availability. NIC depends on a CNN that encodes a picture to a minimal portrayal, trailed by a RNN that produces a relating sentence. The model is prepared to boost the probability of the sentence for the given picture. One heading for future work could be an intend to catch the heterogeneous idea of human explained descriptions and to join such data for assessment of captions. Then the captions can be converted to any other Indian language using Googletrans API. Our future aims include searching for better models for image captioning which are more accurate with respect to human-annotated captions and we could aim for the preparation of image captioning datasets in indian languages so as to avoid the error margin of the Googletrans API. Our work this semester was limited due to the unavailability of captions in Indian languages. Our complete experimentation was done on Flckr-8k dataset and these results were on that basis. Our results were just fine but for future we could work on some other datasets such as MS-COCO or Fickr-30k which have far larger sizes as compared to flick-8k dataset. We could also implement the visual attention mechanism [].

Bibliography

[1] "Show and tell: A neural image caption generator." *O. Vinyals, A. Toshev, S. Bengio, and D. Erhan.* , In CVPR, 2015

[2] "Neural machine translation by jointly learning to align and translate," *K. Cho, D. Bahdanau, and Y. Bengio*, Available at arXiv:1409.0473, 2014

[3] "K. Cho, H. Schwenk, B. van Merrienboer, F. Bougares, C. Gulcehre, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation," *In EMNLP, 2014.*,

[4] "BLEU: a method for automatic evaluation of machine translation," *P. Kishore, S. Roukos, T. Ward, and W.-J. Zhu,*, 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, July 2002.

[5] "Cider: consensus-based image description evaluation," *R. Vedantam, C. Lawrence Zitnick, and D. Parikh*, IEEE Conference, pp. 4566–4575, Boston, MA, USA, June 2015.

[6] "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," *S. Banerjee and L. Alon,*, pp. 65–72, Ann Arbor, MI, USA, June 2005.

[7] "ROUGE: a package for automatic evaluation of summaries," *C.-Y. Lin*, Barcelona, Spain, July 2004.

[8] "Generating image descriptions using dependency relational patterns," *A. Aker and R. Gaizauskas*, In ACL, 2010.