Image Captioning

A thesis report submitted for BTP phase II
by

Rishabh Sharma
(Roll No. 180102058)

Under the guidance of
Prof Prithwijit Guha

DEPARTMENT OF ELECTRONICS & ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
April 2020

Abstract

Captioning of an image with proper descriptions has become an fascinating problem in computer vision and natural language processing. In this report, we describe a model based on a recurrent design that mixes the recent developments in these fields which are often accustomed to generate caption describing a picture. Our aim is in the direction of advancement of the state of the art models to achieve more promising results . This model is trained to maximise the chance of the description of the target sentence in any language based on the dataset of image. At the end, we have also evaluated the performance of model using standard evaluation matrices. Finally, through this report we have highlighted some major challenges in the image captioning task.

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 is a difficult errand, yet it could have a huge benefit, 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. Thus, the generation of captions has long been viewed as a difficult problem. One of the main motivation behind image captioning could 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. After analysing that composing an assorted, clear inscription straightforwardly is difficult, we worked on a captioning model with two-stage decoding [8] which initially produces a primer inscription (less assorted and graphic) given the visual information, and afterward rewords it into a more different and enlightening inscription utilizing these visual rework sets. Our model gains from visual-semantic data as well as uses connections from various phrasings of visual paraphrases. The fundamental motivation for our work comes from ongoing advances in machine interpretation, where the important task is to change a caption C wrote in the source language, into its interpretation T in the objective language, by maximizing p($T|C$).

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 intermediate (conjectural) sequence [8] [3]. Traditionally, these captions are considered to be the objective captions but here propose a two-step decoding by firstly generating a preliminary caption sequence and then work on its enhancement in order to generate better or one can say a more diverse and descriptive captions.

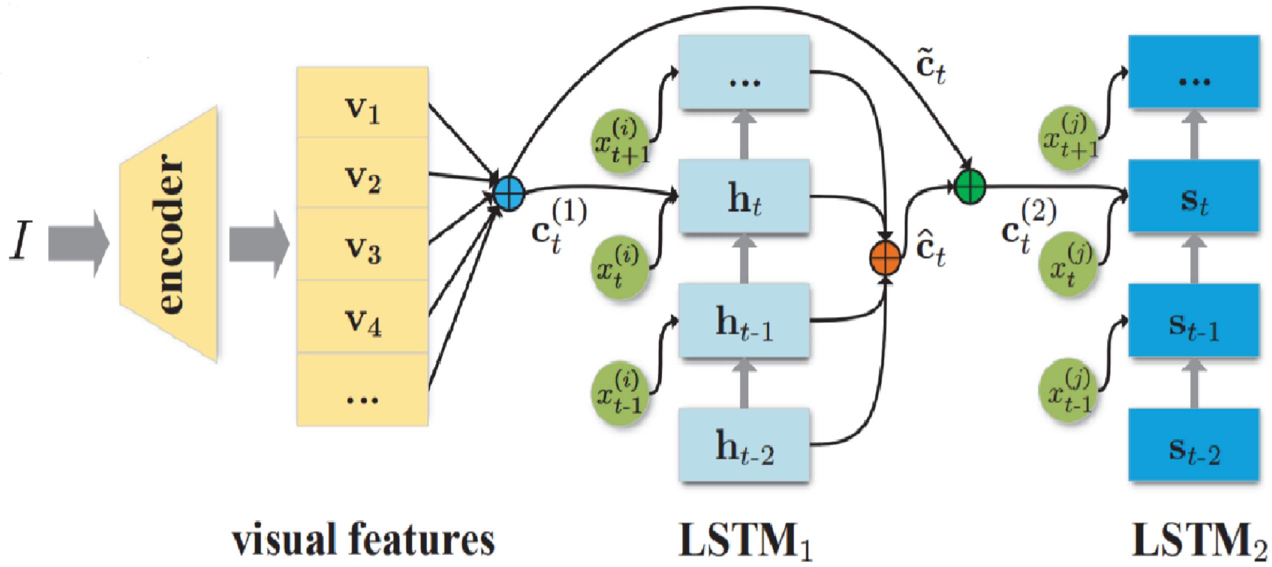


Figure 1.1: Basic Structure of our Model

Chapter 2

Literature Review

A lot of work has been done on the picture inscription problem. 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. The most revolutionary works include O.Vinyals [1], presented a novel methodology of utilizing CNNs and RNNs for picture inscription tasks. O.Vinyals and Alexander Toshev in their Show And tell paper [1] where they consolidated profound convolutional nets for picture grouping with intermittent organizations for succession displaying, to make a solitary organization that creates depictions of pictures. The RNN was prepared with regards to this single "start to finish" organization. The model was propelled by late triumphs of succession age in machine interpretation, with the distinction that as opposed to beginning with a sentence, they gave a picture which was at that point handled by a convolutional net.

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. Most recent works include Guojun Yin's "Context and Attribute Grounded Dense Captioning" [7]. They have planned a unique circumstance and characteristic grounded dense captioning model that licenses multi-scale (i.e., nearby, neighboring, worldwide) logical data sharing and message passing, where the information combination is based on a non-neighborhood comparability diagram among cases in the info picture. And the model is as complex as it sounds. The basis of their work is based on coarse-to-fine semantic attribute supervision which is proposed to enhance the discriminative-ness of the created inscriptions, wherein the ground-truth progressive linguistic attributes are matched to the anticipated watchwords through a new coarse-to-fine way.

A lot of these models have been proposed some have a very complex structures some have a very simple architecture,but our aim is to optimise the results as well as complexity of our architecture which would require a very less computation power and can also generate captions which are as good as the state of the art caption generators.

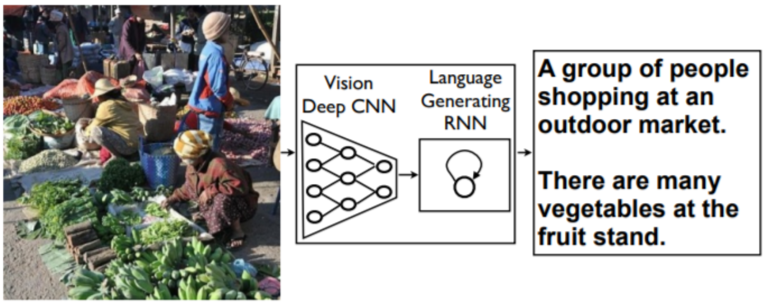

Vision
Deep CNN
Language
Generating
RNN
A group of people shopping at an outdoor market.
There are many vegetables at the fruit stand.

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.

Figure 2.1: Basic structure of state of the art models

Chapter 3

Model Implementation and Evaluation

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 simply depicted in three main sections. In the fundamental segment,the design of data that includes the picture and the train captions are encoded as included vectors. The model utilizes RNN that decodes the varying length information from these vectors(with dimensions fixed) and utilizes this portrayal to decipher it to the target(intermediate) sentence. The final section of our model includes another RNN which takes a mixed type multi-data input which includes the sequence generated in last section and the encoded features of the picture generated in the first section and giving weights to both of these it produces more refined captions. Thus, we have to boost the probability of right depiction in the given picture by utilizing the formula:

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

here theta is the argument of our model, I is a picture, and C 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 of C_0,C_1,...,C_N ,here N is length of the specific model

$$\log p(C|I) = \sum_{i=0}^{N} \log p([C_t/I, C_0, C_1, C_2.., C_{t-1}]) \tag{3.2}$$

For the implementation of the above model, we need to use two Recurrent Neural Networks (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 g. RNN is a group of neural organizations that permit past yields to be utilized as data sources while having shrouded states.

$$k_{t+1} = g(k_t, x_t) \tag{3.3}$$

here k_t represents a fixed state or memory (hidden length). The memory-block is to be replaced by a network ,that consists of time delays or has feed-back loops. For the implementation this function g in above equation we are utilizing a specific type of RNN, i.e. LSTM, which was previously tried with incredible accomplishment to translation [3] and seq. formation and has already shown some impeccable results.

The working of LSTM is based on three gates (See in the figure 3.1). The arrows represent the recurrent associations - the result at time t - 1 is given-back to the memory at time t through three entryways.The block esteem is given back as input through the forget-gate. The anticipated word at time t-1 is given back with the memory yield m at time t into the soft-max for word forecast.

Specifically by discharging the "STOP" word the LSTM signs that a entire sequence 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.

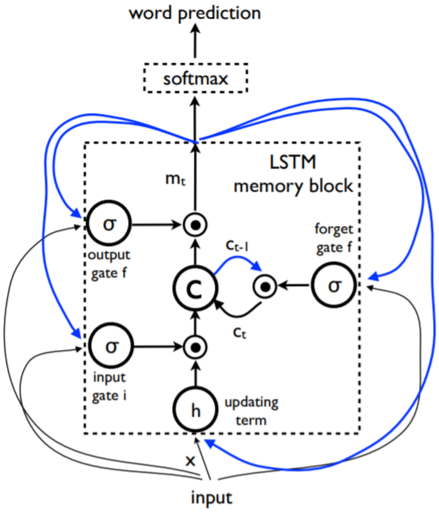


Figure 3.1: Description of LSTM

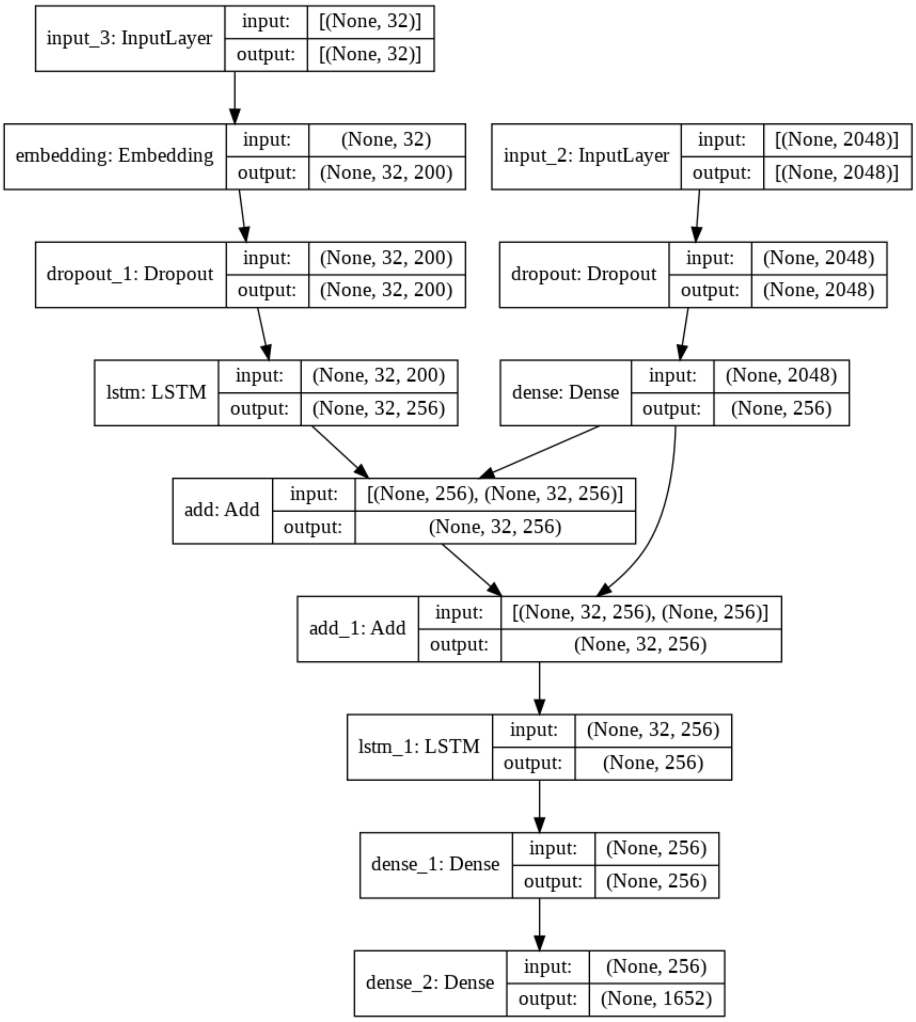


Figure 3.2: Model Architecture

Training Details

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. We have used the Flickr-8k dataset for evaluation.

The assessment of picture subtitling models is commonly performed utilizing measurements, for example, BLEU [4], METEOR [6], 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.

To encode the image i as a feature vector, we use a Inception-V3 pre-trained on Imagenet with fixed weights. Except CNN, all other weights were randomly initialized. Adam optimizer was used with elementary learning rate equal to 1×10^{-4}.

Evaluation and Results using perfomance matrices 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.Loss of Inception-V3 was used for encoder part and LSTM and LSTM for Second and the third section for the experiment. The results obtained were as follows :

Performance Matrices	Bleu-1	Bleu-2	Bleu-3	Bleu-4
Model Score	39.5581	24.0672	17.4412	8.6210

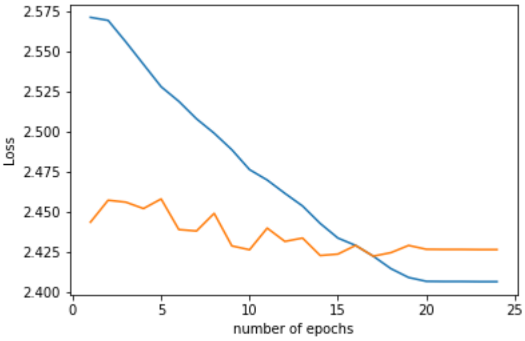
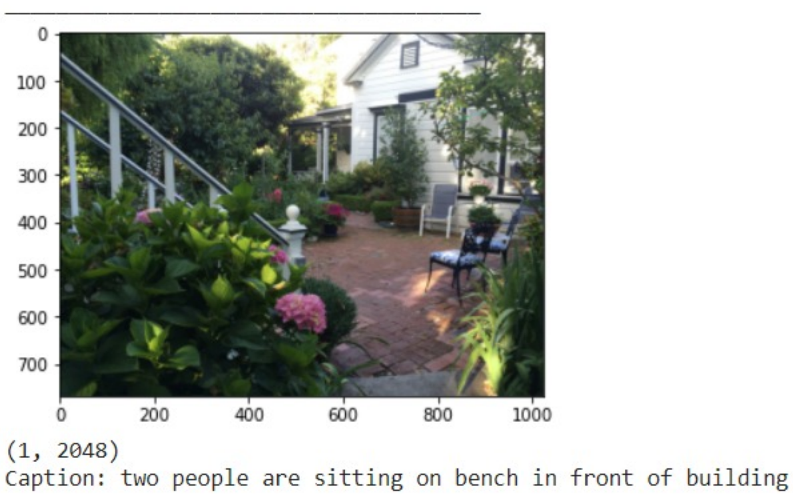
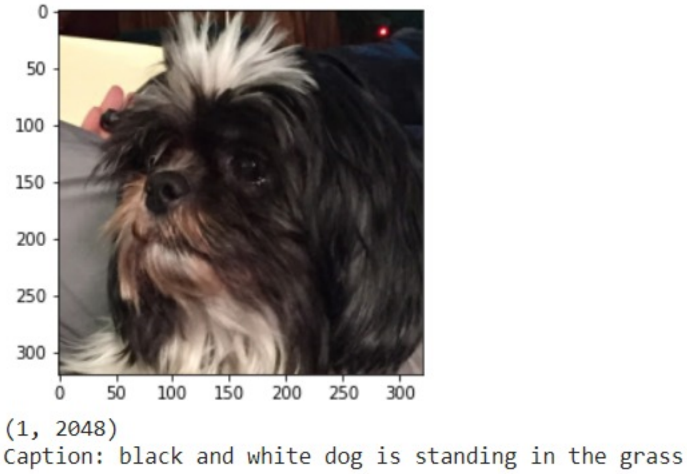


Figure 3.3: Loss on training and validation set



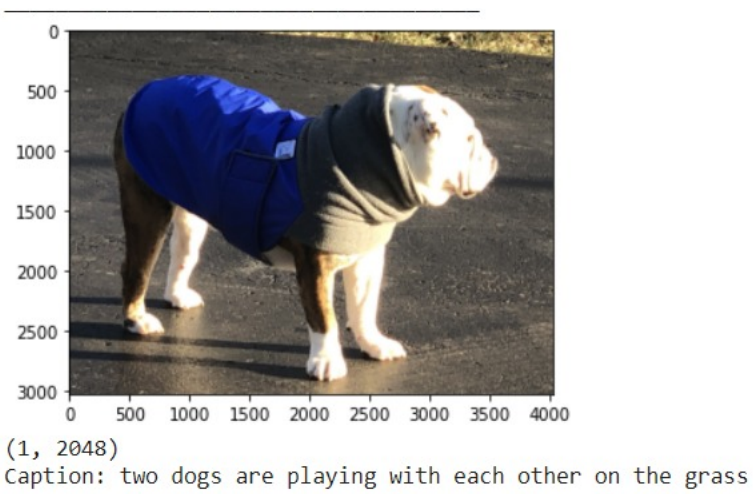
```
(1, 2048)
Caption: two people are sitting on bench in front of building
```

Figure 3.4: Example-1



```
(1, 2048)
Caption: black and white dog is standing in the grass
```

Figure 3.5: Example-2



```
(1, 2048)
Caption: two dogs are playing with each other on the grass
```

Figure 3.6: Example-3

Chapter 4

Conclusions and Future Intends

In this Report, We extended our previous work [1] to describe an image with the idea of hierarchial RNNs for the refinement of a sequence (caption in our case). Our Model depends on a CNN that encodes a picture to a minimal portrayal, trailed by a RNN that produces a relating sentence which is further refined by consideration of weights of image features by another RNN. The model is prepared to boost the probability of the sentence for the given picture.There have been a lot recent development in the field of image captioning using transformers and many other complex-architectural models which require a lot of computational power as well time for the processing of a simple image giving highly accurate results. One heading for future work could be an intend to prepare a light-weighted model for the image captioning task which could work on low computational power devices such as mobile-phones. We could also work on catching the heterogeneous idea of human explained descriptions and to join such data for assessment of captions. Lack of computational power becomes a major issue for heavy models. Our complete experimentation was done on Flickr-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 Flickr-30k which have far larger sizes as compared to flickr-8k dataset. We could also implement the visual attention features [2] instead of just one dimensional feature vector for every image in fundamental segment of our model and a massive improvement in the results can be expected.

Bibliography

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

[2] "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," *Xu K et al*, ICML 2015.

[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] "Context and Attribute Grounded Dense Captioning," *Guojun Yin's*, Available at ArXiv 1904.01410v1.

[8] "Generating Diverse and Descriptive Image Captions Using Visual Paraphrases," *Lixin Liu, Jiajun Tang, Xiaojun Wan, Zongming Guo*, ICCV 2019.