Image Captioning

BTP PHASE-II

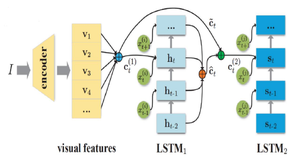
Rishabh Sharma(180102058)
Under the Guidance of Prof. Prithwijit Guha
Department of EEE, IIT Guwahati

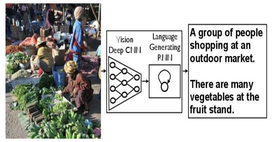
Abstract

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 maximize 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.

Introduction

we propose to follow this formula, by supplanting two decoder recurrent neural networks with a profound convolution neural network.

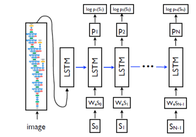




Review of Prior Works

A lot of work has been done on the picture inscription problem.

- The most revolutionary works include O.Vinyals [1], presented a novel methodology of utilizing CNNs and RNNs for picture inscription tasks.



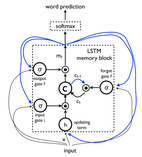


Fig 3: Model architecture of Show and Tell

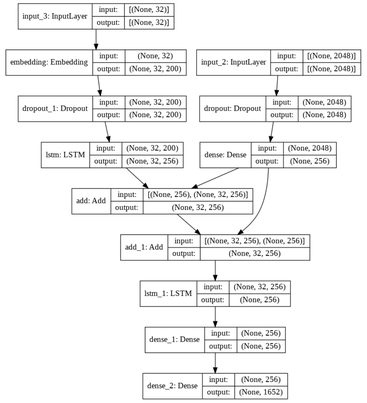
- Most recent works include Guojun Yin's "Generating Diverse and Descriptive Image Captions Using Visual Paraphrases" [8]. They proposed a captioning model which fuses visual and textual information with two-step decoding by firstly generating a preliminary caption and then paraphrasing it into a more diverse and descriptive caption

$$x_{-1} = \text{CNN}(I)$$
$$x_t = W_e S_t, \quad t \in \{0 \dots N-1\}$$
$$p_{t+1} = \text{LSTM}(x_t), \quad t \in \{0 \dots N-1\}$$

$$L(I,S) = -\sum_{t=1}^{N} \log p_t(S_t) .$$

Model Description

- 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 input 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.
- 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.

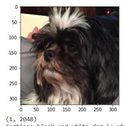


Results

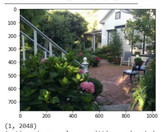
Few examples of captions generated by the our model.

a little boy in a red shirt is sitting on a red slide

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

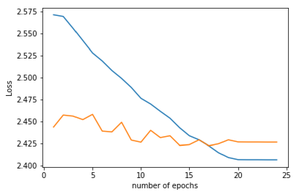


(1, 2048)
Caption: black and white dog is standing in the grass



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

Loss on training and validation set



Performance Evaluation

Evaluated using BLEU-1, BLEU-2, BLEU-3 and BLEU-4. on Flickr8K Database.

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

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) .$$

Future Intends

- 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
- 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.
- Getting more and more accurate with respect to human-annotated captions.

References

[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.