Real Time Image Captaioning

Asha G, R. Hema Sumanth, A. China Venkat Chowdary, A. Shashank, T. Sravan

Abstract: Image caption generator means it will generate a description for the images. It will predict what is happing in the images. We make our model using a hybrid CNN-RNN model in which in the CNN part of the model we use inception model for transfer learning and RNN is majorly used for language modeling. We use Flickr8k Dataset for training and testing the model. We use LSTM model in RNN to avoid the problem of vanishing or exploding gradient in the training phase.

Key words- CNN-RNN architecture, LSTM, SOFTMAX, Image caption generator.

I. INTRODUCTION

Image captioning is one of the most challenging task and an active research area in the field of AI. Given an image it is common for a human to describe what is present in the image but making an algorithm to describe an image accurately is a very challenging goal.

There are Two main approaches in Image Captioning: One is to identify the objects in the images and for the objects detected form a meaningful sentence with all the objects in the image. This considers all the objects in the image but it will not take into consideration the relation between the objects for example: A man standing on the grass has two objects man and grass and man pulling the grass also have same objects man and grass but the meaning of the two sentences is very different.

Second approach which is our approach is not only to use the objects for describing but also generating the semantic representation of the image and then generating the image caption using the LSTM which is for language modeling of the semantic representation. Here in this model we can to a certain extent capture the relation between the different objects in the image.

Revised Manuscript Received on April 30, 2020.
***** Correspondence Author

Asha G, R. Hema Sumanth*, Master In Computer Science And Engineering From VTU Belgaum, Karnataka, India.

R. Hema Sumanth, Master In Computer Science And Engineering From VTU Belgaum, Karnataka, India.

A. China Venkat Chowdary, Pursuing Final Year B.Tech Degree In Computer Science And Engineering From GITAM University, Bengaluru, India.

A. Shashank, Pursuing Final Year B.Tech Degree In Computer Science And Engineering From GITAM University, Bengaluru, India.

T. Sravan, Pursuing Final Year B.Tech Degree In Computer Science And Engineering From GITAM University, Bengaluru, India.

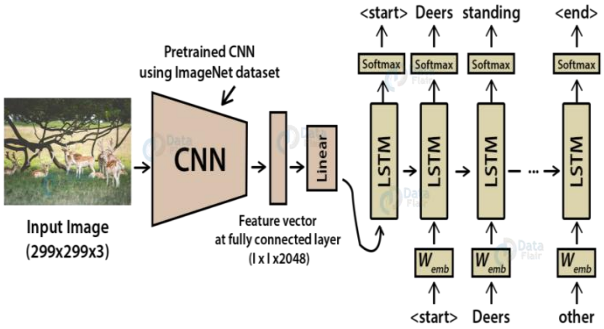

Fig. 1. The Figure above shows the architecture of the model used which is a CNN-LSTM hybrid model.

II. RELATED WORK

As this is an active research area there has been many models for this task, but most of the models will be one of the two approaches.

Top-Down approach: This is straightforward approach in which we first detect all the objects [9] in the image and then use them to form a caption using all the objects detected. This approach is not good because it is not modeled to capture the relation between objects.

In top down approach we identify the objects in the image using an object detection algorithm proposed by fair (Facebook artificial intelligence research) and then use any natural language generation model to form a suitable sentence using the objects detected. This was the first approach trying to describe the image. This method was extensively used in Junhua mao, Wei xu and Jiang wang's paper named deep capturing and recurrent neural network [5] but this method doesn't capture the semantic representation between the objects in the image which is very important to describe the image correctly.

Bottom-up approach: In this model we try to capture the semantic relation between the objects in the image for a better modeling of the system and for more accurate description of the image and objects between them. As this approach tries to captures the relation between objects in the image this is much better than Top-Down approach for images with more complicated object relationship.

The semantic representation which will be captured in the bottom up approach proposed by Moses sob in Stand Ford University and this paper was named as CNN-LSTM architecture for image caption generation [13]. This method use transfer learning methodology. In CNN-RNN architecture. We use both CNN and LSTM as a hybrid. The CNN in these model is inception model in which the last layer will not be a classification layer but provides a linear vector of values which shows the semantic representation of the objects in the image. This inception model trice to learn about the relationship between the objects and then this feature vector is passed to an LSTM.

After being reduced in its dimensionality and the LSTM's trice to convert this encoded linear feature vector into target language which can also be called as decoding but the only complication is that the specificity in describing the image training will directly impact the captions generated. For example, the man is walking is less specific description than a man walking along- side the water in the beach. Due to these the semantic representation produced by the ANN cannot be learned by the LSTM to decode it properly.

There are also other method proposed by Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth called Every picture tells a story: Generating sentences from images [1], In this method they try to map a meaning to the image. This method will have two spaces image space and meaning space meaning space is a triplet of (object, action, scene) and image space is (resp. sentence). We evaluate the similarity between a sentence and an image by mapping each to the meaning space then comparing the results.

Method proposed by Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg, Tamara L. Berg, and Yejin Choi called Collective generation of natural image descriptions [2] use human described captions data base to generate captions for example for a given image it tries to find all the similar images in the data set which has human described captions and then use them to generate a caption by selectively combining thoses images.

Some other method like Composing simple image descriptions using web-scale n-gram proposed by Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi [3] completely describes an image by learning about words in the image rather than using and generating complex phrases.

The paper Learning a recurrent visual representation for image caption generation proposed by Xinlei Chen and C. Lawrence Zitnick [4]. This approach using recurrent neural networks. First, they to generate sentences given a set of visual observations or features. Second, they try to enable the capability of computing the likelihood of the visual features given a set of spoken or read words for generating visual representations of the scene. This method is also used to recreate the image by looking at the sentence. This can be used for performing image search.

In the paper [6], [7] they use Mining Semantic Affordances of Visual Object Categories i.e., they try to determine the action for given object. This is like connecting verb nodes and noun nodes. Krizhevsky, I. Sutskever, G. E. Hinton, Image net classification with deep convolutional neural networks [8] uses traditional method of CNN connected with fully connected at the end to predict the caption.

III. METHODOLOGY

In our model we use an encoder recurrent neural network model in this our model learns to represent the image and then this representation is converted into target language by LSTM.

A. Model Architecture Overview

First we convert the image into a 2048 X 1 feature vector using inception CNN and then pass it to LSTM. The loss function used in LSTM is software.

$$\boldsymbol{\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}} \qquad \textbf{Eq. (A.1)}$$

$$\boldsymbol{for\ i = 1,\ldots,K\ and\ Z = (z_1,\ldots,z_K)\ \epsilon\ \mathbb{R}^K}$$

We use a dense layer to reduce the input layer from 2048 x 1 V vector to 256 x 1 vector and we also do embedding to all the words which is 7577 in our database to 256 embedded state

B. LSTM For Caption Generation

In each LSTM six operations are performed

$$\boldsymbol{f_t = \sigma_g\left(W_f x_t + U_f h_{t-1} + b_f\right)} \qquad \textbf{Eq. (B.1)}$$

$$\boldsymbol{i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i)} \qquad \textbf{Eq. (B.2)}$$

$$\boldsymbol{o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o)} \qquad \textbf{Eq. (B.3)}$$

$$\boldsymbol{c_t = f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c)} \textbf{Eq. (B.4)}$$

$$\boldsymbol{h_t = o_t \circ \sigma_h(c_t)} \qquad \textbf{Eq. (B.5)}$$

The forget gates f_t(B.1) allow the model to selectively ignore past memory cell states and the input gates. It allows the model to selectively ignore parts of the current input. The output gate $o_t \circ \sigma_h(c_t)$ (B.5) then allows the model to filter the current memory cell for its final hidden state. This operation allows long term dependencies for the RNN.

C. Input Representation Of LSTM

The input to our model is [x1, x2] and the output will be y, where x1 is the 2048 features vector of the image, x2 is the combined words of all the previous LSTM's.

First the image is made into 2048 x 1 vector using an inception CNN model and then this is passed to a dense layer to reduce it to 256 x 1 vector and passed to LSTM.

IV. RESULTS

We trained and validated our model using Flickr8k Dataset which contains 6000 images. The max length of the captions is 32 words although all the words are not of length 32 in size. The mean caption length is 15.6805.

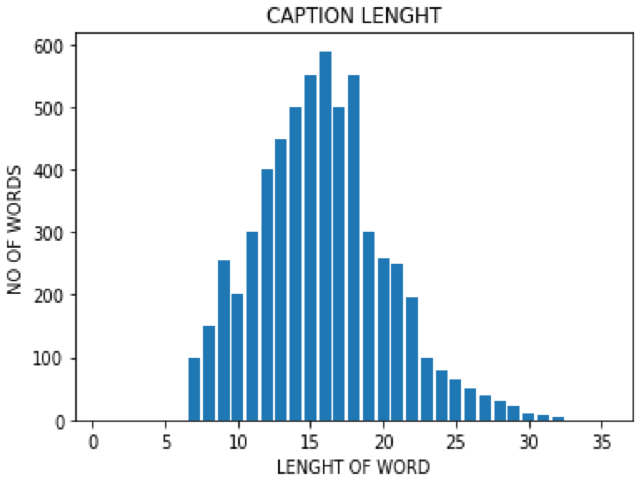

Fig. 2. Distribution of Caption Lengths in the Flickr8k Dataset. The mean caption length is 15.6805. There is a substantial right tail in the empirical distribution.

Sentence Matching: We tried to match the sentences predicate by our model with sentences described by human and considered all the synonyms as a perfect match (for example the word plate and bowl) as the semantic meaning of it is the same in the image. We tested this on 200 captions and the matching was 90 percent accurate.

V. CONCLUSION

We conducted an extensive hyper-parameter search over the CNN-LSTM model architecture,

producing a best model that achieves results that are very accurate with original description of the image and found that at 50 % dropout the models accuracy is very good .Most of the wrong prediction are due to lack of attention to specific details in images caption in training data (for example, A man walking on the beach long side the water is captioned as A man walking)due to this lot of semantic representation in the image is not properly decoded into the target language.We also found that lot of words like plate-bowl etc. will have same semantic meaning in the hidden states although in real world they may have a reasonable difference between them.

REFRENCES

1. Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In Proceedings of the 11th European Conference on Computer Vision: Part IV, ECCV'10, pages 15–29, Berlin, Heidelberg, 2010. Springer-Verlag.
2. Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg, Tamara L. Berg, and Yejin Choi. Collective generation of natural image descriptions. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, ACL '12, pages 359–368, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
3. Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. Composing simple image descriptions using web-scale n-grams. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning, CoNLL '11, pages 220–228, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
4. Xinlei Chen and C. Lawrence Zitnick. Learning a recurrent visual representation for image caption generation. CoRR, abs/1411.5654, 2014.
5. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). CoRR, abs/1412.6632, 2014.
6. L. Bourdev, J. Malik, S. Maji, Action recognition from a distributed representation of pose and appearance, in: IEEE Conference on ComputerVision and Pattern Recognition, Providence, RI, 2011, pp. 3177–3184.
7. Y.-W. Chao, Z. Wang, R. Mihalcea, J. Deng, Mining semantic affordances of visual object categories, in: IEEE Conference on ComputerVision and Pattern Recognition, Boston, MA, USA, 2015, pp. 42594267.
8. A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the 25th International Conference on Neural Information Processing Systems, 2012,pp. 1097–1105.
9. B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database, in: Advances in Neural Information Processing Systems (NIPS), 2014, pp. 487–495.
10. Y. Gong, L. Wang, R. Guo, S. Lazebnik, Multi-scale orderless pooling of deep convolutional activation features, in: European Confrence on Computer Vision, 2014, pp. 392–407.
11. A. Kojima, T. Tamura, K. Fukunaga, Natural language description of human activities from video images based on concept hierarchy of actions,International of Computer Vision 50 (2002) 171–184.
12. P. Hede, P. Moellic, J. Bourgeoys, M. Joint, C. Thomas, Automatic generation of natural language descriptions for images, in: Proc. Recherche Dinformation Assistee Par Ordinateur, 2004
13. Moses Soh Department of Computer Science Stanford University:Learning CNN-LSTM Architectures for Image Caption Generation

AUTHORS PROFILE

Mrs. Asha G received a master in computer science and engineering from VTU Belgaum, Karnataka in 2014. She is working as Assistant Professor in department of computer science for last four years. She has published paper in international journal and conference in the area of IOT and Cloud Computing.

Mr. R. Hema Sumanth, currently pursuing final year B.Tech degree in Computer Science and Engineering from GITAM University, Bengaluru. His area of interest are Python, Artificial Intelligence and Machine Learning.

Mr. A. China Venkat Chowdary, currently pursuing final year B.Tech degree in Computer Science and Engineering from GITAM University, Bengaluru. His area of interest are Python, DBMS and Data Science.

Mr. A. Shashank, currently pursuing final year B.Tech degree in Computer Science and Engineering from GITAM University, Bengaluru. His area of interest are Data Science, Data Structures, C and C++.

Mr. T. Sravan, currently pursuing final year B.Tech degree in Computer Science and Engineering from GITAM University, Bengaluru. His area of interest are DBMS, Cloud Computing and Java.