IMAGE CONTENT DESCRIPTION USING LSTM APPROACH

Sonu Pratap Singh Gurjar[1], Shivam Gupta[1] and Rajeev Srivastava[2]

[1]Student, Department of Computer Science and Engineering,
IIT-BHU, Varanasi, Uttar Pradesh, India
[2]Professor, Department of Computer Science and Engineering,
IIT-BHU, Varanasi, Uttar Pradesh, India

ABSTRACT

In this digital world, artificial intelligence has provided solutions to many problems, likewise to encounter problems related to digital images and operations related to the extensive set of images. We should learn how to analyze an image, and for that, we need feature extraction of the content of that image. Image description methods involve natural language processing and concepts of computer vision. The purpose of this work is to provide an efficient and accurate image description of an unknown image by using deep learning methods. We propose a novel generative robust model that trains a Deep Neural Network to learn about image features after extracting information about the content of images, for that we used the novel combination of CNN and LSTM. We trained our model on MSCOCO dataset, which provides set of annotations for a particular image, and after the model is fully automated, we tested it by providing raw images. And also several experiments are performed to check efficiency and robustness of the system, for that we have calculated BLUE Score.

KEYWORDS

Image Annotation, Feature Extraction, LSTM, Deep Learning, NLP.

1. INTRODUCTION

The image is an important entity of our digital system. It contains much useful information like an image of a receipt, an image taken from CCTV footage etc. We can surely say that an image tells a unique story in its way. In today's digital world, one can perform or gather a large information or facts just only after analyzing a digital image. When we are dealing with digital images, we have to gather what each and every part of it wants to contain. For that, we should extract each and every part with optimal care and extract information of that particular region and then gather the whole information to reach out to a conclusion. Here we need to get what the picture have like objects, boundaries, and colour etc. features. Here, we need an accurate description of an image and for digital images, we need an efficient and accurate model that could give accurate annotations of each and every region of that image and can provide a rich sentence form, so that we can understand what's happening in that image.

Image captioning[1] is used by several software giants companies like Google, Microsoft etc. It is used for many other specific tasks like explaining an image for a blind person by giving him a sentence generated form of annotations of an image. Such essential and significant functionalities make image captioning an important field to study and explore.

Dhinaharan Nagamalai et al. (Eds) : SIGEM, CSEA, Fuzzy, NATL - 2017
pp. 01– 12, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70901

Image annotations generation of an image is very much close to scene understanding model. Computer vision involves the complete understanding of an image, so our model should not only just provide image annotations, but should be capable of expressing the scene and what exactly objects are doing in that image. In this way, computer vision and natural language processing go hand to hand for solving this problem of automatic image description by providing suitable generated sentence explaining the scene of an image. Likewise, a human can easily perceive the content of an image just by looking at it, and he can explain the scene in that image accurately. But, when it comes to the computer, it's a difficult task to generate and explain scene of an image by using machine learning algorithms. Human generated annotations have properties like rich, concise and accurate etc. Like, a human can generate a well fit sentence, and that sentence has only relevant things and accurate as it contains all essential region's information of an image.

Many researchers [1, 14] and others have explored this problem and provided few appropriate solutions. There are many advances in this field as large datasets like MSCOCO, Flickr30k etc. are available to train the model more efficiently. The basic model (as shown in figure 1.) for image captioning works like, first gather the features of an image and given captions in the dataset, and based on features provide a suitable annotation to that image. In figure 1, it shows a man lying on the table and a dog sitting near him. As this is a sample input image firstly its features based on colour, objects, boundaries, texture etc. is extracted and also features of given captions, then based on its attributes a common representation is produced. And from that common space embedding representation, an appropriate sentence is generated for this image like a man lying on the table with a dog.

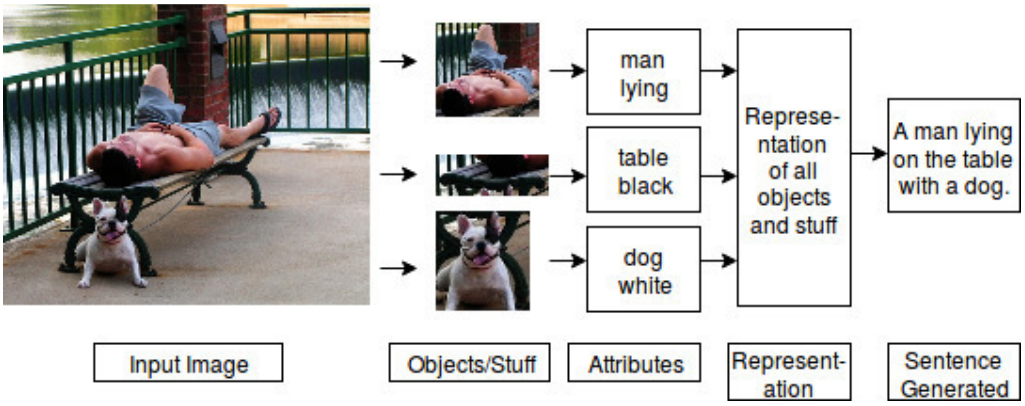

Figure 1. Basic working of Image Captioning model

Most of the works in this field are based on these two approaches: similarities approach and constructive approach. Similarities approach [1], [2], [5] and [6] means taking the model as a retrieval task, after extraction of features of image and captions, this approach provides an embedding representation of information and based on that most suitable annotation is selected as annotation for an image. This approach has few limitations like it doesn't provide good results when a raw image contain an unseen object or thing in it, as due to the limited size of its dictionary, it produce annotations based on previously gathered features and language models. In this way, this approach is not suitable in today's advance in this field. Constructive approach [3], [4], [7], [8], and others mean the generation of sentence based on firstly learning image features and after that sentence generation process occurs in many parts. Like a basic constructive approach based model consists of these basic steps: language modeling part, image features extraction and analysis part, and representation part which combines the previous both parts. In [4], they described an image by using the constructive approach as Convolutional Neural Network (CNN) is used for extraction of image features and after that Recurrent Neural Network

(RNN) is used to learn the representation of space embedding and generated a suitable sentence for an image. Here, language modeling means, it learns dense features information for every word related to the content of that particular image present in the dictionary and it gathers its semantic form in recurrent layers.

In [1], Pan et al. used similarities approach and found similarity measures between the captions keywords and the images. They described an image region in form of blob-tokens, it means a part of image region based on its features like colour, texture, size, position, boundaries, and shape. And [2], applied nearest neighbor approach along with similarities measures, as a set of keywords which are nearer to each other form the sentence. These models have some limitations like they are biased with training dataset, and doesn't give a good result for a new unseen image. In [3,4], they proposed a model based on CNN[3] with RNN[11] method, as this model is based on constructive approach and it produced some good results as compared to models defined with similarities approach. These models provide better scene understanding and able to express the content of the image semantically. But they has few limitations like over fitting of data and not able to give good results when attention is not given to similar type of images.

Our novel work involves most optimal technique with the constructive approach, as we used CNN method to extract image features and then the common representation of whole information and features of images and captions are made by using LSTM[15] model, which then produce appropriate sentence for any new image. Our model produces accurate annotations and also the sentence expresses the scene of the image, along with information about all objects and stuff in that image. Our contributions are as follows:

- Image feature extraction is done using robust CNN technique, which produces features based on colour, texture, and position of objects and stuff in the image.
- The common representation contains all gathered information in layers and cells of LSTM model and then based on input, hidden layers and then final output from output layers are obtained.
- For language modeling, we used n-gram model, so that accurate and rich form of a sentence is generated.
- Blue-4 metric is used for analysis of efficiency and robustness of our method. And the comparison of various previous models with our model.

A further portion of the paper is divided into sections as follows, Section 2 included related work portion which gives detailed information about previous works done in this field. Section 3 provides the complete overview of the technique of image captioning such as problem statement, input and output details, and also includes the motivation behind this work, and all relevant terminology and concepts are discussed in detail. After that our model is described in Section 4, then Section 5 give results and implementation details as dataset used are MSCOCO, Flickr30k, and Flickr8k, also the value of BLUE-4 metric score is provided as the efficiency measure of our model. Section 6 discuss the conclusion and future improvements in this work.

2. RELATED WORK

There are many advances in the field of automatic image captioning. As in earlier works, [1] Pan et al. proposed technique which work by annotating a particular part of image region, in this a word for each image region and then on combining we get the sentence, they discussed about considering the image regions as blobs-token which means an image regions based on its feature such as colour, texture, and position of object in image. But this technique has few limitations

such as it is effective only for a small dataset, this work include much manual work as they have to provide annotated words and blob-tokens with an image, and this approach sometimes give results based on a training set, that means it is biased on a training set.

Jacob et al. [2], provided technique that explores nearest neighbor images in training set to the query image and based on that appropriate k nearest neighbor images, their given captions are imported and based on that only caption is given for the query image. But there is a limitation of this approach as it performs better for highly similar images, but worse for highly dissimilar images. In [5], they used similarities approach as a candidate matching images with the query image are retrieved from the collection of captioned images, then after features are matched and based on the best rank obtained a caption is given to the query image. But this model has few limitations like re-ranking the given captions could create error for training images and related text, and also object and scene classifiers could give erroneous results, so the model could have given faulty results. In [6], Ali et al. proposed model based on computing of a score linking a sentence with an image. And worked based on the semantic distance between some words like two or three for a particular image, and SVM models are trained on it. This model lacks as dataset used is not much used and large, and not much emphasized for checking adjectives and other relevant potential good information from image regions.

In [3], [4], and [7,14], they used the constructive approach, but with different techniques for extracting images and then after that for sentence generation. Karpathy et al. [3], described the common intermodal representation between the visual information and the language models. They used the combination of CNN over image regions and Bi-directional Recurrent Neural Network (BRNN) for sentence generation approach by computing word representation. But this model also have certain limitations as this model didn't focus on attention concept for captioning.

In our model, CNN is used for extraction of features of an image, and it provided that information to the common representation. In [4], this paper focussed on tight connection between image objects and text related to that. As they acquired every detailed information of each and every region of an image by particularly dealt with each region, as they extracted and detected objects/stuff in an image, and their attributes such as adjectives which provide extra useful information about that particular region, also details about the spatial connections between those regions, and based on these Conditional Random Field (CRF) is constructed and labels of graphs are predicted, and finally based on these labelling sentence is generated. This model contains few limitations such as it didn't provide semantically related texts on input images, so the accuracy of the model is compromised in this way.

Most of the work in this field is related to providing a visual interpretation of the image and relation to that to given captions in the dataset. In [7], Desmond et al. introduced representation to contain connections between different objects/stuff of an image, it worked based on similarities between the objects or image regions and based how these regions relate to each other. They used image parser to get information for each region of an image. But this model have certain limitations as the output of an object detector is not used to obtain a fully automated model, and there are various improvements that can be done to the image parser to enhance the efficiency of this model.

3. OVERVIEW OF METHOD

Image captioning involves machine learning algorithms and as well many mathematical simulations so that an accurate annotation can be provided.

3.1. Motivation

As human can perceive an image just by looking at it, we must have a robust and accurate model that can cope up with a human in case of captioning an image and express what the objects are doing in that image. Our model should have all important characteristics like accuracy, rich in sentence generation, consistent as could not be a biased model and concise as it much includes only relevant regions of the image. And there are many real life applications of image captioning like image search, tell stories from the pictures uploaded on the web and helpful for visually impaired people so that they can be aware of relevant information from the web.

3.2. Problem Statement

Image content description by using the neural networks and concepts of deep learning.

3.3. Input

A query image.

3.4. Dataset

MSCOCO, Flickr30k, and Flickr8k.

3.5. Output

Captioning of that query image, according to the learning of the model.

3.6. Feature Extraction

For a query image, firstly we extract its features based on its colour, texture, boundaries, objects, stuff, and position of things in it. This process of feature extraction is very crucial in our model, as it provides all basic required information about the image. This process is done with help of CNN models. As CNN contains certain specified number of layers and those layers are responsible for storing the features which are extracted from the images, then these features are passed forward to LSTM model so make a common representation for sentence generation. As shown in figure 2, an image is given to the model, and then model extracts all relevant features of that image such as:

- Low-level features, it involves around details of pixels level such as the colour of a pixel and edges and corners in images.
- Mid-level features, it involves between low-level and high-level features, it discuss any curves in the image of an object present in it.
- High-level features denote detection of the object in the image. It is hard to predict the exact object and scene of that object in that image, so image captioning is all about minimizing the gap between this low-level and high-level features methods.

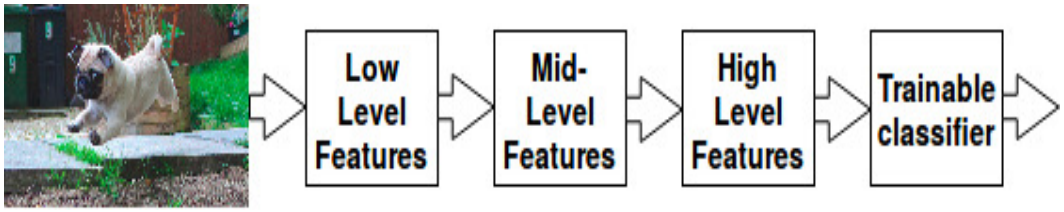

Figure 2. An image and extraction of its features by CNN

After obtaining all relevant features, CNN gives it to a trainable classifier. And from that to the common representation model. It is a neutral network which is fully trainable with the help of mathematical simulations like stochastic gradient descent.

The model takes an input image and provides the possible caption C from the available dictionary of 1-to-T words, such as:

$$C = C_1, C_2, \ldots, C_N, \text{ where } C_i \epsilon R^T \quad (1)$$

, where T equals to the available number of keywords of the vocabulary and N is the possible length of the sentence.

We use CNN to extract the image features as a set of feature vectors. Then it produces M vectors, which belongs to a part of the image and having an L-dimensional form, such as:

$$B = B_1, \ldots\ldots, B_M, \; B_i \epsilon R^L \quad (2)$$

As shown in figure 2, CNN extracts features from lower levels, so that each relevant region is completed for sentence generation. So that decoder, LSTM model could focus on only relevant portions of the image by using feature vectors, r(I) for an image I having a specified dimension.

$$CNN\_MODEL(I) = W^i r(I) + b \quad (3)$$

As per the measurement of accuracy of the features extracted by our model, we trained our model based on visualization parameters, which helps in examining of the different feature activations and their relation to features embedding. Also, our model worked on both image classification and the localization tasks. It is analysed that as the network grows, there is rise in the number of filters used. So, the accuracy is optimised by incoporating more number of filters.

3.7. Language Modeling

In our model, we have used n-gram model for language modelling, it means a statistical probability function based on conditional factors such as for N words =

$$P(a_i \mid a_{i-N+1}, \ldots, a_{i-1}) \quad (4)$$

, it means the possibility of the next word of the sequence is based on the previously occurred words in the sequence.

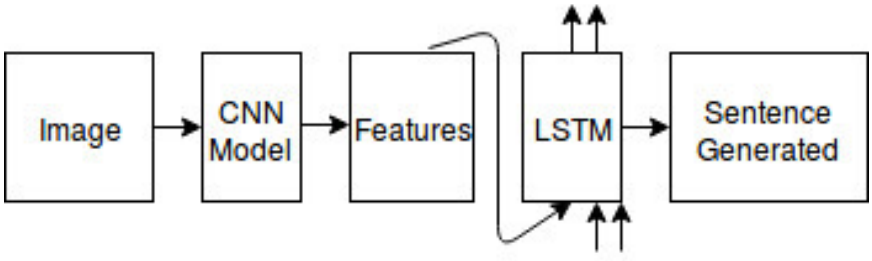

Figure 3. Basic Framework of our model.

3.8. LSTM

Long Short-Term Memory (LSTM) [15,16] acts as the decoder in our model when features are transferred to it, it uses the common representation of all gathered information and the based on it provide sentence. As shown in figure 3, the basic framework of our model is depicted, as shown an image is provided as an input to our model, then CNN used for feature extraction and then extracted features are given as input to the LSTM unit, which finally generates sentence, as shown in figure 3, which provides the basic framework of our model.

LSTM model generally consists of three important gates such as input gate, forget gate and output gate. And the main part of an LSTM model is its memory cell c, which keeps the whole information about the image features, previously generated words and track functionality of all three gates.

3.9. Words Representation

It is based on the size of the vocabulary of our model, like we taken image dimension as I_D = 4096, so word form will become of order:

$$T \; x \; I_D \tag{5}$$

4. DESIGN OF PROPOSED METHOD

In our model, image features are extracted by CNN, then LSTM model acts like a decoder of that features. As CNN_MODEL(I) is passed to LSTM model, as an input. It takes it, and further evaluate values of gates defined in its inner working system. And whole working information of the system is stored in memory cells, c. These gates units are trained to learn when to open and close permit of access to information to memory cells. Three gates used as whether the current cell value is to forget(forget gate f), input gate (i) is to read input and output gate (o) to whether output the new cell value.

LSTM model computes on the basis of the memory cell information and previously calculated words of the sequence, such as:

$$P(S_t|I,S_0 \, , \, …., \, S_{t-1}) \tag{6}$$

, where I is an image and S is a possible sentence which depends on previously generated words.

These are the main equations which explains our model:

$$a_{t-1} = CNN\_MODEL(I) \tag{7}$$

$$a_t = W_e \, S_t \tag{8}$$

$$p_{t-1} \; = LSTM\_MODEL(a_t) \tag{9}$$

, where a_t means input to LSTM model, as CNN_MODEL(I) is initial input to LSTM model, and then after it works recursively and obtain one word of sentence at each time.

Updating of gates and cell values in a LSTM model as such:

$$i_t = \sigma \, (W_{ia} \, a_t + W_{ik} \, k_{t-1}) \tag{10}$$

$$f = \sigma \, (W_{fa} \, a_t + W_{fk} \, k_{t-1}) \tag{11}$$

$$o = \sigma\,(W_{oa}\,a_t + W_{ok}\,k_{t-1}) \tag{12}$$

$$c = f_t\,\Theta\,c_{t-1} + i_t\,\Theta\,h(W_{ca}\,a_t + W_{ck}\,k_{t-1}) \tag{13}$$

$$k_t = o_t\,\Theta\,c_t \tag{14}$$

$$p_{t+1} = Softmax(k_t) \tag{15}$$

where, Θ means multiplication of a gate value, Softmax() is used as for higher dimension balancing between the various stages of values. And the pair (k_t , c_t) is passed as the present form of hidden state to the upcoming hidden state. And k_t is given to Softmax, which provide a probability distribution.

As shown in figure 4, LSTM models work recursively after a word is found, and use that information to predict the next word of the sentence.

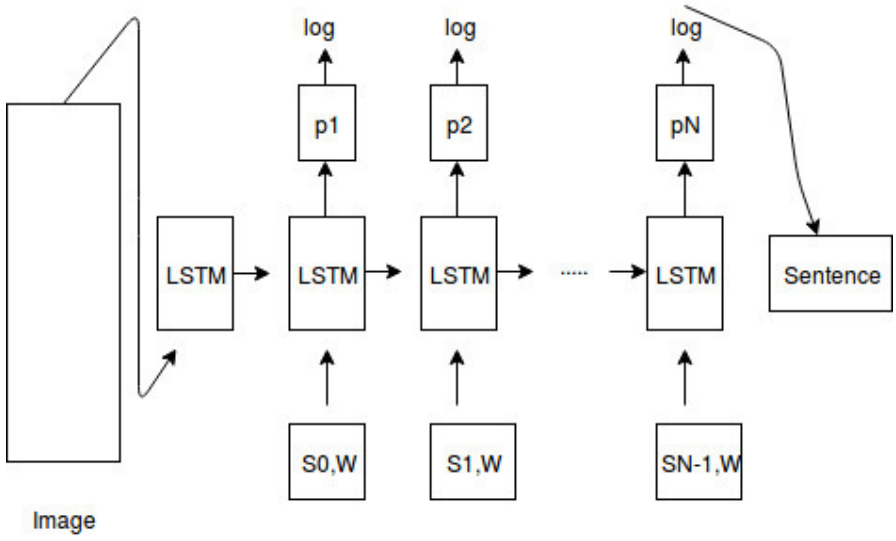

Figure 4. Working on LSTM model for sentence generation.

4.1. Sentence Generation

In our model, LSTM is used for sentence generation, the process of sentence generation involves certain basic steps, such as it starts from *"##START##*" or any other sentence generation reference words, which conveys that next word that will be generated will be the first word of our desired sentence. Our method calculates the probability distribution for the upcoming word, $P(S_t|I,S_0\ ,\ ....,\ S_{t-1})$. After that we use this distribution method and previously calculated words for the calculating probability of the next word. And cycle goes on until we encounter the last word of sequence and then after model produce output as end sign *"##END##"*. We use our model to calculate the probability of generating a sentence given an image. The sentence generation task is incorporated by using the perplexity of a sentence conditioned on the averaged image feature across the training set as the reference perplexity to normalize the original perplexity as discussed in [8].

For an example, in figure 5, an input image is given, our model starts predicting a word at a time and by using the previously calculated word, with the further calculation of probability distribution, it predicts the next word. Like first word predicted based on the image features is "man", then by using it and model parameters, it predicts the next word as "bench", and process

goes on until the model encounters the stop word. In this way, LSTM generates the most optimal sentence for a new input image.

5. IMPLEMENTATION AND RESULTS

Our model includes the novel combination of CNN and LSTM techniques with using deep learning approaches. We test our method on benchmark datasets like MSCOCO [12], Flickr30k [13] and Flickr8k [14]. The dataset MSCOCO contains around 80k images, and each image has at least 5 different captions of different lengths related to it. This contains images of almost everything such as sports, landscapes, portraits, persons, groups etc. Out of 80k images, we have taken 5k images for testing phase and check the implementation of our model on that testing dataset.

We used deep learning approach for implementation of our model, and it is implemented in python by using Keras [17] library, a high-level neural network for fast and accurate implementation which is run on Theano as backend.

5.1. Input

Human caption: A man lying on the bench and a dog sitting on the ground.

Figure 5. An input image

5.2. Output

Caption generated by our model: A woman sitting on a bench with a dog.

Figure 6. Output: A woman sitting on a bench with a dog.

5.3. Metric

We have used BLEU-4 metric instead of BLEU-1 metric, as it is a far better metric for measuring the efficiency of our model.

BLEU-4 metric score of our model: 20.84

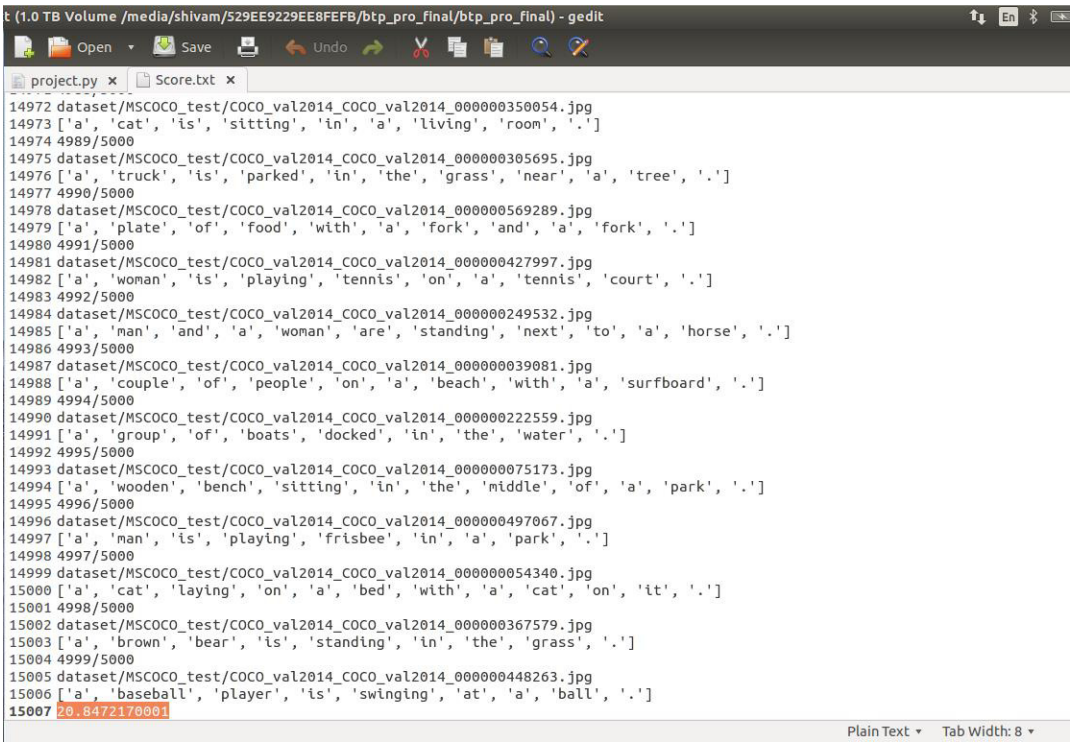

Figure 7. BLUE-4 Score: 20.84.

5.4. Parameters for Implementation

Our LSTM model is implemented by using 2 layers, as we have checked it with 1 layer too. But former method produces more optimal captions. We have observed that on increasing further number of layers, generated captions efficiency degraded. So, we concluded that a number of layers are 2, is the most optimal method for implementation. And weights are initialized uniformly from [-0.06, 0.06].

We have taken maximum caption length = 16.

Batch size = 200

Dimension of LSTM output = 512

Image dimension parameter = 4096

Word vector dimension = 300

5.5. Comparison

We have compared our model with state-of-art techniques [1, 2, 3, 4], and based on BLEU-4 metric.

Table 1. Comparison of models

Dataset	Model	BLEU-4 Score
MSCOCO	Random	4.6
MSCOCO	Nearest Neighbour[2]	9.9
MSCOCO	CNN & RNN[8]	19.5
MSCOCO	Karpathy[3]	20.4
MSCOCO	Human	20.51
MSCOCO	Our model	20.84

From Table 1, we can see that our model is far better and efficient than previous works done in the field of automatic image captioning.

6. CONCLUSIONS

Our novel method showed that it is an efficient and a robust system, and can produce the description of any unseen image, which is more specific or related to the content of that image. And, it also is shown that our model is much better than state-of--art models and others previous automated works. As the measure of the efficiency of our model, we calculated BLEU-4 metric which is around 20.84 for our model. Several experiments are performed on different datasets, which depicts the robustness of our method.

In future works, we can make the current model more fast and efficient by applying fast machine learning algorithms. Also, we can fine-tune features extracted by CNN to improve correctness of our model. Also, we can test our model on more number of testing the dataset for better results.

ACKNOWLEDGEMENTS

The authors would like to thank Department of Computer Science and Engineering, IIT-BHU, Varanasi for providing wonderful opportunity and complete facility for research works.

REFERENCES

[1] Pan, Jia-Yu, Hyung-Jeong Yang, Pinar Duygulu, and Christos Faloutsos. "Automatic image captioning." In Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on, vol. 3, pp. 1987-1990. IEEE, 2004.

[2] Devlin, Jacob, Saurabh Gupta, Ross Girshick, Margaret Mitchell, and C. Lawrence Zitnick. "Exploring nearest neighbor approaches for image captioning."arXiv preprint arXiv:1505.04467 (2015).

[3] Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128-3137. 2015.

[4] Kulkarni, Girish, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. "Babytalk: Understanding and generating simple image descriptions."IEEE Transactions on Pattern Analysis and Machine Intelligence 35, no. 12 (2013): 2891-2903.

[5] Ordonez, Vicente, Girish Kulkarni, and Tamara L. Berg. "Im2text: Describing images using 1 million captioned photographs." In Advances in Neural Information Processing Systems, pp. 1143-1151. 2011.

[6] Farhadi, Ali, Mohsen Hejrati, Mohammad Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. "Every picture tells a story: Generating sentences from images."Computer vision–ECCV 2010 (2010): 15-29.

[7] Elliott, Desmond, and Frank Keller. "Image Description using Visual Dependency Representations." In EMNLP, vol. 13, pp. 1292-1302. 2013.

[8] Mao, Junhua, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. "Explain images with multimodal recurrent neural networks."arXiv preprint arXiv:1410.1090 (2014).

[9] Pan, Jia-Yu, Hyung-Jeong Yang, Christos Faloutsos, and Pinar Duygulu. "Gcap: Graph-based automatic image captioning." In Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on, pp. 146-146. IEEE, 2004.

[10] Yao, Benjamin Z., Xiong Yang, Liang Lin, Mun Wai Lee, and Song-Chun Zhu. "I2t: Image parsing to text description."Proceedings of the IEEE 98, no. 8 (2010): 1485-1508.

[11] Kalchbrenner, Nal, and Phil Blunsom. "Recurrent Continuous Translation Models." In EMNLP, vol. 3, no. 39, p. 413. 2013.

[12] Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft coco: Common objects in context." In European Conference on Computer Vision, pp. 740-755. Springer International Publishing, 2014.

[13] Young, Peter, Alice Lai, Micah Hodosh, and Julia Hockenmaier. "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions."Transactions of the Association for Computational Linguistics 2 (2014): 67-78.

[14] Hodosh, Micah, Peter Young, and Julia Hockenmaier. "Framing image description as a ranking task: Data, models and evaluation metrics."Journal of Artificial Intelligence Research 47 (2013): 853-899.

[15] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory."Neural computation 9, no. 8 (1997): 1735-1780.

[16] https://deeplearning4j.org/lstm LSTM Documentation and Tutorial.

[17] https://keras.io – Keras Library Documentation.