International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-07, Issue-12, MAR 2022

Image Captioning

[1]Shyamkrishna Menon, [2]Atharva Ranade, [3]Omkar Thavai, [4]Siddharth Nair, [5]Sofiya Mujawar

[1,2,3,4]Student, [5]Assistant Professor, A P Shah Institute of Technology, Thane, India,

[1]reachskmenon2017@gmail.com, [2]ranadeatharva2112@gmail.com,

[3]nair.siddharth01@rediffmail.com, [4]omkarthavai63@gmail.com, [5]ssmujawar@apsit.edu.in

Abstract--**This paper aims at generating automated captions by learning the contents of the image. At present images are annotated with human intervention and it becomes nearly impossible task for huge commercial databases. The image database is given as input to a deep neural network (Convolutional Neural Network (CNN)) encoder for generating "thought vector" which extracts the features and nuances out of our image and RNN (Recurrent Neural Network) decoder is used to translate the features and objects given by our image to obtain a sequential, meaningful description of the image [1] [16].**

Keywords* — *image caption, convolution neural network(CNN), long short term model(LSTM), recurrent neural network(RNN), VGG16, Flickr30k.

I. Introduction

A large amount of information is stored in an image. Every day, image data that are generated are enormous in amounts. These data are generated by social media, CCTV footage, etc. generating captions manually thus becomes a tedious task. Deep learning can be used to automatically annotate these images, thus replacing the manual annotations done. This will greatly reduce human error as well as the efforts by removing the need for human intervention. The generation of captions from images has various practical benefits, ranging from aiding the visually impaired, to enable the automatic, cost-saving labeling of the millions of images uploaded to the Internet every day, recommendations in editing applications, beneficial in virtual assistants, for indexing of images, for visually challenged people, for social media, and several other natural language processing applications [1]. Image captioning can also be used for educational purposes for teaching pre-primary children to make them aware of what all entities are present within a picture. The image captioning model can be used for the enhancement of products like Google Lens. Google Lens is used by users to identify objects and provide relative e-commerce links. With our project imbibed, Lens can also explain the scenario to a confused user. The field brings together state-of-the-art models in Natural Language Processing and Computer Vision, two of the major fields in Artificial Intelligence. There are many Natural Language Processing (NLP) applications right now, which extract insights/summary from a given text data or an essay etc. The same benefits can be obtained by people who would benefit from automated insights from images. One of the challenges is the availability of a large number of images with their associated text on the ever-expanding internet. Generating captions automatically from images is a complex task as it entails the model extracting features from the images and then forming a meaningful sentence from the available features [1]. Basically, the feature extraction is done by training a Convolutional Neural Network (CNN) with a huge number of images, and the correct weights are identified by multiple forward and backward iterations. With the help of RNN (Recurrent Neural Network) and the extracted features, a sentence is generated [1].

II. LITERATURE SURVEY

A. Deep Learning based Automatic Image Caption Generation [1]

The aim of the paper [1] is to generate captions to the image which is normally, manually annotated by data annotators. It first creates feature vectors with the help of CNN and later uses RNN for the creation of sentences with the help of features gained before. For the purpose of automated captioning, a pre-trained model called VGG16 model is being used. This model [1] makes use of a RNN which encodes the variable length input into a fixed dimensional vector and uses this representation to "decode" it to the desired output sentence [1] [4]. An encoder is a process of extracting vectors which describe contents of an image. A decoder reverses the process of encoding. Decoder process uses layers like tokenizer, embedding, GRU and dense layer. The paper also points few previous works done on image captioning. The paper [1] uses 2 approaches for obtaining image captioning with the same dataset i.e. MS-COCO, one without using Attention Model and one using Attention Model. Finally, the paper concludes with important points like different epochs used

DOI : 10.35291/2454-9150.2022.0100

for different models, deeper network constitutes to easier image captioning, etc.

B. Image Annotation via deep neural network [2]

The authors of this paper [2] have proposed a deep learning framework. A novel framework of multimodal deep learning where the CNNs with unlabeled data are utilized to pre-train the multimodal deep neural network to learn intermediate representations and provide a good initialization for the network then use backpropagation to optimize the distance metric functions on individual modality[1] [2]. NUS-WIDE dataset is being used in the paper. The proposed framework consist of a unified two-stage learning path where (i) learning to fune-tune the parameters of deep neural network with respect to each individual modality, and (ii) learning to find the optimal combination of diverse modalities simultaneously in a coherent process[2].

C. Automatic image annotation using DL representation[3]

In this paper [3], the authors propose a model for image annotation. They have used the Canonical Correlation Analysis (CCA) framework and have reported the results of all 3 variants of CCA i.e. linear CCA, kernel CCA and CCA with k-nearest neighbor (CCA-KNN) [3]. In the CNN based model (which is used for feature extraction and word embedding vectors for the representation of associated tags) the last layer of CaffeNet of the CNN based model is replaced with a projection layer to perform regression and the resulting network is trained for mapping images to semantically meaningful word embedding vectors. The advantage of this modeling is: firstly, it does not require dozens of handcrafted features and secondly, the approach is simpler to formulate than any other generative or discriminative models [3] [1]. Finally, the paper concludes by stating that CCA-KNN Model provides the best results.

D. Show and Tell: A Neural Image Caption Generator [4]

This paper [4] proposes a network of the same name. the model is used for the generation of captions from images using computer vision and machine translation. Here CNN is used for feature extraction and RNN helps in sentence formation [1] [4]. Pascal, MS-COCO [13], and Flickr30k [12] are some of the datasets that are being used. BLEU [10] scores are used as the evaluation metric.

E. An Empirical Study of Language CNN for Image Captioning [5]

In this paper [5], the authors have introduced a language CNN model which is suitable for statistical language modelling tasks and shows competitive performance in image captioning. The primary contribution lies in incorporating a language CNN, which is very powerful for text representation [5] [8] [9], is capable of capturing long-range dependencies in sequences, with RNNs for image captioning. The model yields comparable performance with the state-of-the-art approaches on Flickr30k [12] and MS COCO [13] which validate the proposal and analysis of the experiments conducted. Performance improvements are clearly observed when compared with other image captioning methods.

F. Image Captioning - A Deep Learning Approach [6]

This paper [6] proposes a fusion between CNN and LSTM. Here CNN is used for building vocabulary and LSTM is used for forming meaningful sequence of words obtained. The efficiency of the proposed model is checked using Flickr30k and Flickr8k datasets and also by utilizing Bleu [10] metric gives superior results in comparison to other state-of-the-art models.

III. EXPERIMENT SET-UP

We use Python as our programming language as it is a popular language when it comes to using deep learning approaches and image processing. We use Deep learning for training the model using Convolutional Neural Networks and Recurrent Neural Networks (deep learning model) to detect features from image and predict the captions respectively. There are few python libraries that we will be using. We use pandas for data manipulation and analysis, opencv for loading images, numpy for mathematical operations, Keras Framework(Using Tenserflow Backend) is used for building our model architecture for Image Captioning and also used for importing VGG-16 for Transfer Learning. All these are implemented in Jupyter Notebook [17] enabling Python 3 language.

IV. PROPOSED SYSTEM

1) Explanation of proposed system

a. First we will import Flickr30k [7] [11] dataset and process. Flicker datasets are used for image captioning. 30k stands for around 30,000 images of various instances.

b. We use VGG16 [14] model for image captioning. VGG16 is used for embedding of features within the image like identifying a person, thing, etc and LSTM [15] is used for encapsulating all features and describing it as a sentence.

c. We have considered our model with thresholds of both 0 (i.e. no threshold) and 10. A threshold is a frequency below which we do not consider a certain word. When the threshold is 10, it means that the frequency of words in the captions of the Flickr30k [7] dataset that are lesser than 10 are eliminated. Thresholds are kept for simplifying

DOI : 10.35291/2454-9150.2022.0100

the computation of the model by removing unimportant, less recurring words.

2) Block diagram

Fig 1 depicts the block diagram.

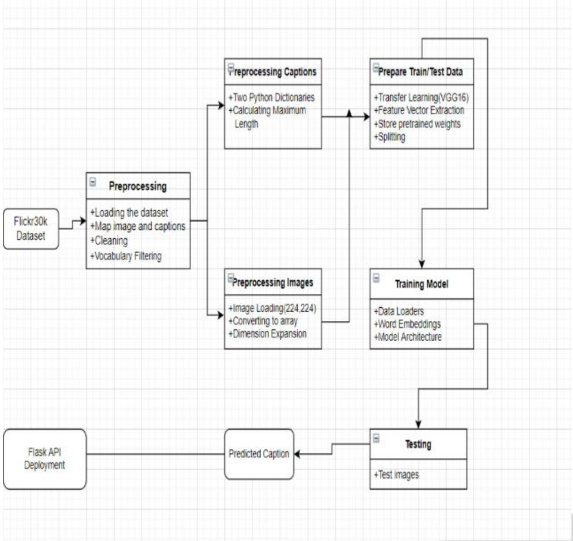


Fig. 1. Block diagram.

This figure represents the workflow of our project starting from loading the dataset to deploying it using Flask.

3) Model algorithm diagram

Fig 2 depicts the Model algorithm diagram

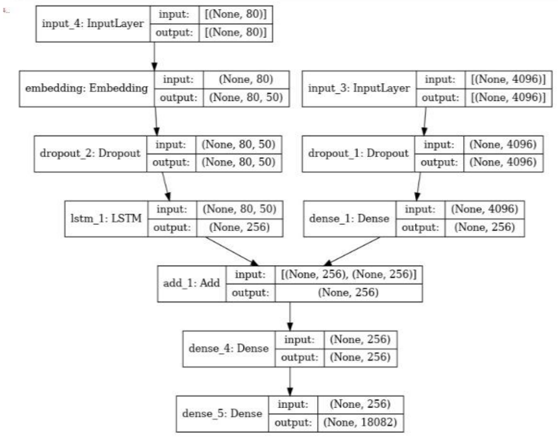


Fig. 2. Model algorithm diagram.

This figure shows the complete architecture of our model.

4) Pseudocode

a. Read dataset- read the Flickr30k [7] dataset.

b. Processing of data- firstly, create a dictionary of imageID and descriptions, then create a vocabulary and finally filter out words that are more frequent.

c. Transfer learning- use VGG16 [14] for getting vectors for every image.

d. Word embeddings- preprocess captions and put them into a fixed-length using glove.

e. Training- combine image and caption as input (obtained from steps c and d) and train the model.

f. Testing- obtain caption from learned weights during training of the model.

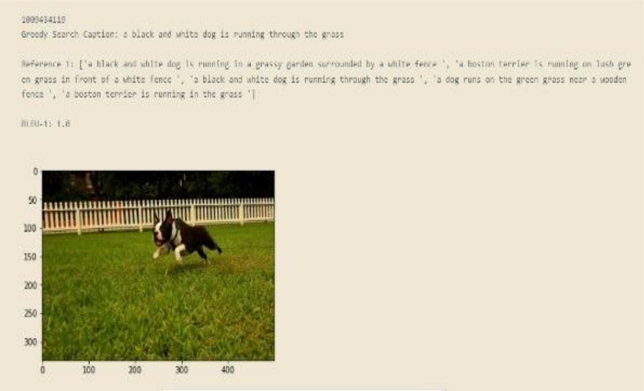
5) Expected output

a. Expected output would be apt sentence formation of the given input image.

b. With the help of different features learned, the model will provide words relevant to the image.

V. RESULTS

Fig 3 shows the final results of few images that are obtained by our paper.

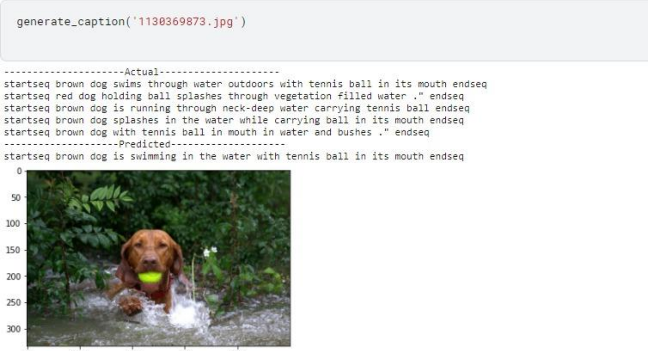


```
1000434115
Greedy Search Caption: a black and white dog is running through the grass

Reference 1: ['a black and white dog is running in a grassy garden surrounded by a white fence ', 'a boston terrier is running on lush gre
en grass in front of a white fence ', 'a black and white dog is running through the grass ', 'a dog runs on the green grass near a wooden
fence ', 'a boston terrier is running in the grass ']

BLEU-1: 1.0
```

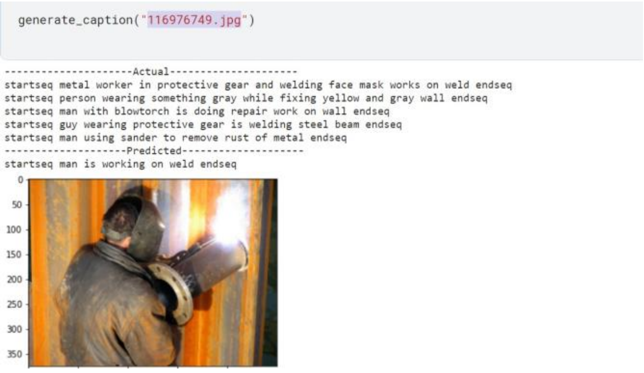
Fig. 3. a



```
generate_caption('1130369873.jpg')
```

```
--------------------Actual--------------------
startseq brown dog swims through water outdoors with tennis ball in its mouth endseq
startseq red dog holding ball splashes through vegetation filled water ." endseq
startseq brown dog is running through neck-deep water carrying tennis ball endseq
startseq brown dog splashes in the water while carrying ball in its mouth endseq
startseq brown dog with tennis ball in mouth in water and bushes ." endseq
-------------------Predicted-------------------
startseq brown dog is swimming in the water with tennis ball in its mouth endseq
```

Fig. 3. b

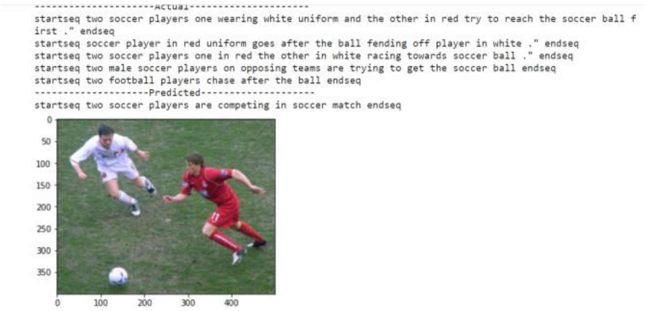


```
generate_caption("116976749.jpg")
```

```
--------------------Actual--------------------
startseq metal worker in protective gear and welding face mask works on weld endseq
startseq person wearing something gray while fixing yellow and gray wall endseq
startseq man with blowtorch is doing repair work on wall endseq
startseq guy wearing protective gear is welding steel beam endseq
startseq man using sander to remove rust of metal endseq
-------------------Predicted-------------------
startseq man is working on weld endseq
```

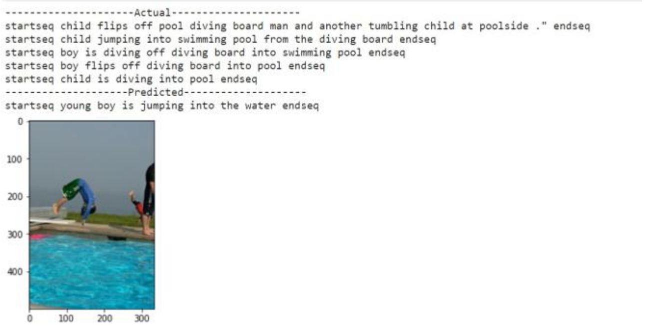
Fig. 3. c

DOI : 10.35291/2454-9150.2022.0100



```
--------------------Actual--------------------
startseq two soccer players one wearing white uniform and the other in red try to reach the soccer ball f
irst ." endseq
startseq soccer player in red uniform goes after the ball fending off player in white ." endseq
startseq two soccer players one in red the other in white racing towards soccer ball ." endseq
startseq two male soccer players on opposing teams are trying to get the soccer ball endseq
startseq two football players chase after the ball endseq
-------------------Predicted-------------------
startseq two soccer players are competing in soccer match endseq
```

Fig. 3. d



```
--------------------Actual--------------------
startseq child flips off pool diving board man and another tumbling child at poolside ." endseq
startseq child jumping into swimming pool from the diving board endseq
startseq boy is diving off diving board into swimming pool endseq
startseq boy flips off diving board into pool endseq
startseq child is diving into pool endseq
-------------------Predicted-------------------
startseq young boy is jumping into the water endseq
```

Fig. 3. e

Fig. 3. Few Screenshots of the Final Results obtained through our model.

As seen in figures 3. a,b,c,d, and e, our model has captioned images as per the features that are available in the image.

VI. Conclusion

Image captioning has become a booming topic and how leveraging deep learning concepts has eased the process of annotations. Our paper has used both CNN and RNN for the generation of captions to the inputted image. We have used the glove file for word embedding purposes. Along with the creation of vocabulary, we have considered our model with thresholds of both 0 (i.e. no threshold) and 10. The advantage of our model is that we are able to obtain caption to the input images with few relevant features included in the caption. Our findings show that certain features are not perfectly captured because those features have very little frequency in the dataset. For eg. The colors of T-shirts are captioned as either blue or black if any new color is given to the model. The disadvantage that we have observed is that the model requires hours to process and run to finally obtain image captions. Also, as the dataset consists of lesser images than those that are normally needed for image captioning, the results obtained for some images are not quite as expected. The main focus was to obtain relevant captions for the input image. In the future, the project can be improved to optimize the prediction of captions by training it on bigger datasets and using better computational resources like using GPUs. Usage of attention models can also benefit in the process of obtaining more relevant captions as the attention model will emphasize smaller details in the input image. As mentioned in the introduction, we can build an end-to-end application for visually impaired people who can benefit from our image captioning model by listening to the captions that our model predicts using text-to-speech for the predicted captions.

References

[1] Shahar Banu , Seemakousar B , Sanchita S M , Nivedita A, Arun Joshi, Rajeshwari S.G, 2021, varsha, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH TECHNOLOGY (IJERT) Volume 10, Issue 08 (August 2021)

[2] S. Chengjian, S. Zhu and Z. Shi, ”Image annotation via deep neural network,” 2015 14th IAPR International Conference on Machine Vision Applications (MVA), 2015, pp. 518-521, doi: 10.1109/MVA.2015.7153244.

[3] Venkatesh N.Murty et al, Automatic image annotation using DL representation, ICMR ’15: Proceedings of the 5th ACM on International Conference on Multimedia RetrievalJune 2015 Pages 603–606.

[4] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, ”Show and tell: A neural image caption generator,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3156-3164, doi: 10.1109/CVPR.2015.7298935.

[5] J. Gu, G. Wang, J. Cai and T. Chen, ”An Empirical Study of Language CNN for Image Captioning,” 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1231-1240, doi: 10.1109/ICCV.2017.138.

[6] Srinivasan, Lakshminarasimhan and Dinesh Sreekanthan. “Image Captioning-A Deep Learning Approach.” (2018).

[7] Kaggle dataset for flickr30k-https://www.kaggle.com/datasets/adityajn105/flickr30k?select=Images

[8] N. Kalchbrenner, E. Grefenstette, and P. Blunsom. A convolutional neural network for modelling sentences. ACL, 2014.

[9] M. Wang, Z. Lu, H. Li, W. Jiang, and Q. Liu. gen cnn: A convolutional architecture for word sequence prediction. ACL, 2015.

[10] BLEU: a Method for Automatic Evaluation of Machine Translation. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu IBM T. J. Watson Research Center Yorktown Heights, NY 10598, USA.

[11] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: data, models and evaluation metrics. Journal of Artificial Intelligence Research, 2013.

[12] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL, 2014.

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft coco: Common objects in context. arXiv preprint arXiv:1405.0312, 2014.

[14] Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[15] S. Hochreiter and J. Schmidhuber, ”Long short-term memory”, Neural Comput., vol. 9, no. 8, pp. 1735-1780, 1997.

[16] V. Kesavan, V. Muley and M. Kolhekar, ”Deep Learning based Automatic Image Caption Generation,” 2019 Global Conference for Advancement in Technology (GCAT), 2019, pp. 1-6, doi: 10.1109/GCAT47503.2019.8978293.

[17] https://python.engineering/getting-started-with-jupyter-notebook-python/

DOI : 10.35291/2454-9150.2022.0100