IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Image Caption Generator

[1]M.V.S.Pravallika, [2]K.Yamuna, [3]N.Padma, [4]K.Pooja, [5]Mrs. Ch.Usha

[1]Student, [2]Student, [3]Student, [4]Student, [5]Assistant Professor

[1]Compter Science and Engineering,

[1]Vignan's Institute of Engineering for Women, Visakhapatnam, India

Abstract: The main aim of the Image Caption Generator is to generate a language description of the given image. It deals with image understanding and a language description for that image. Experimental results show that the embedded image caption generator generates captions for the selected images.

Index Terms **– Image Captioning, CNN, LSTM.**

I. INTRODUCTION

Image caption generation has emerged as a challenging and important research area following advances in statistical language modelling and image recognition. We will build a working model of the image caption generator by using CNN (Convolutional Neural Networks) and LSTM (Long short term memory) units. We use a deep convolutional neural network to generate a vectorized representation of an image that we then feed into a Long-Short-Term Memory (LSTM) network, which then generates captions. Image Captioning is the process of generating a textual description for given images. It has been a very important and fundamental task in the Deep Learning domain.

II. LITERATURE SURVEY

2.1 EXISTING SYSTEM

The existing system is a multimodal Recurrent Neural Network (m-RNN) model for generating novel image captions. It directly models the probability distribution of generating a word given previous words and an image. Image captions are generated by sampling from this distribution. The model consists of two sub-networks: a deep recurrent neural network for sentences and a deep convolutional network for images. These two sub-networks interact with each other in a multimodal layer to form the whole m-RNN model.

2.2 PROPOSED SYSTEM

This model will generate captions from a fixed vocabulary that describe the contents of images in the Flickr8k Dataset. Encoder-Decoder model is used in this experiment. Here the encoder model will combine both the encoded form of the image and the encoded form of the text caption and feed to the decoder. This model will treat CNN as the 'image model' and the RNN/LSTM as the 'language model' to encode the text sequences of varying length. The vectors resulting from both the encodings are then merged and processed by a dense layer to make a final prediction.

III. PROPOSED ALGORITHMS

3.1 Working

This model works on Encoder- Decoder principle using InceptionV3 model

Step1: Select an input image
Step2: Detecting the objects from the image
Step3: Processing the sequence from the text obtained from training dataset using glove embeddings.
Step4: Extract the feature vector from the image using the InceptionV3 model.
Step5: Decoding the output using softmax by concatenating the above two layers.
Step6: Training the model
Step7: Generating Image captions using greedy search and beam search algorithms.

3.2 Greedy Search Algorithm

While generating captions, greedy search algorithm selects one best candidate as an input sequence for each time step. Choosing just one best candidate might be suitable for the current time step, but when we construct the full sentence, it may be a sub-optimal choice.

3.3 Beam Search algorithm

The beam search algorithm selects multiple alternatives for an input sequence at each time step based on conditional probability. The number of multiple alternatives depends on a parameter called Beam Width B**.** At each time step, the beam search selects B number of best alternatives with the highest probability as the most likely possible choices for the time step.

3.4 Model

The model used in this experiment is InceptionV3. Inception v3 is a convolutional neural network for assisting in image analysis and object detection, and got its start as a module for Googlenet. It is the third edition of Google's Inception Convolutional Neural Network, originally introduced during the ImageNet Recognition Challenge.

3.5 Dataset

The dataset used for training this model is Flcker8k which contains 8000 images.

IV. EXPERIMENT AND RESULT

The test set for this evaluation experiment image randomly selected from the test_images folder in the system. Jupyter Notebook software platform is use to perform the experiment. The PC for experiment is equipped with an Intel i3 2.2GHz Personal laptop and 8GB RAM.

The proposed scheme is tested using ordinarily image processing. From the simulation of the experiment results, we can draw to the conclusion that this method is robust to many kinds of testing images.

Figure1. Interface

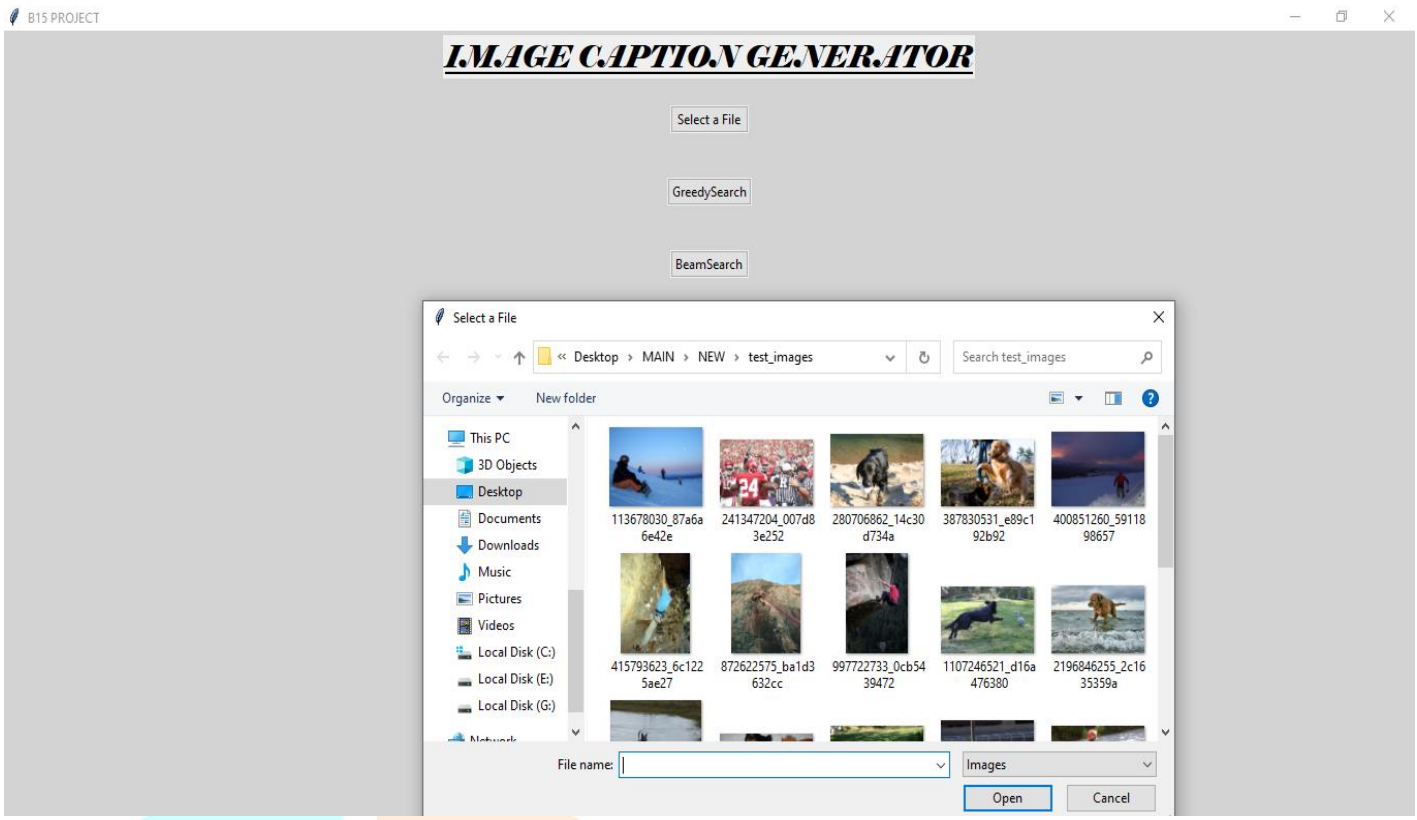

Figure2. Selecting a File

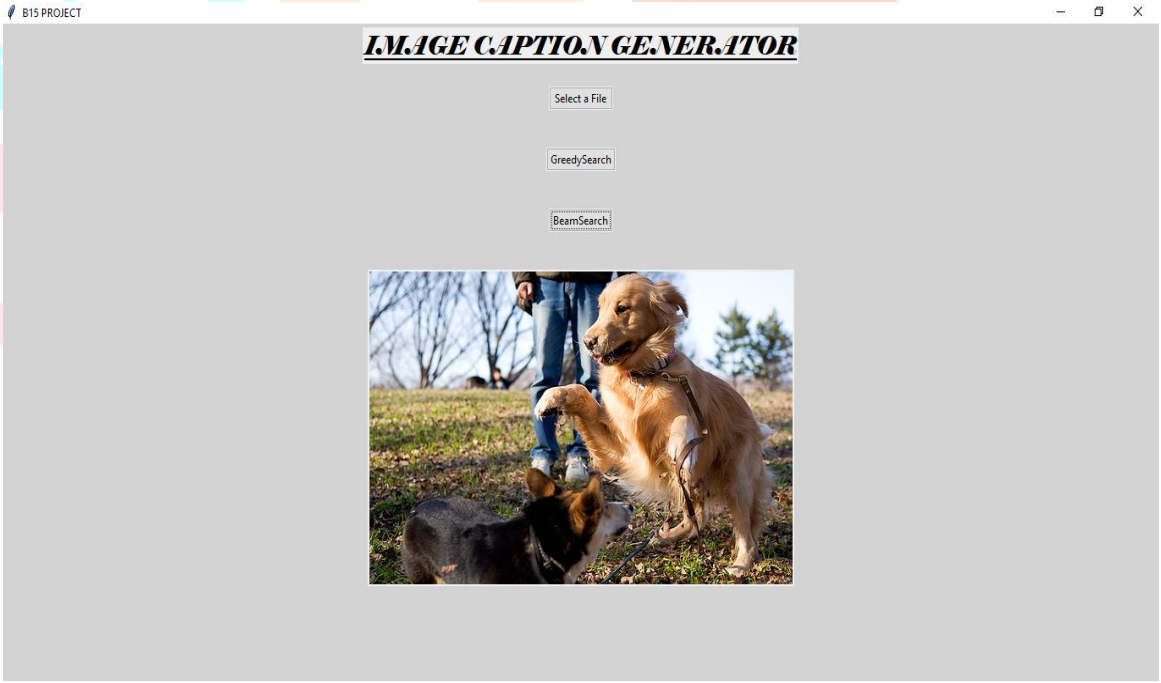

Figure3. Input Image

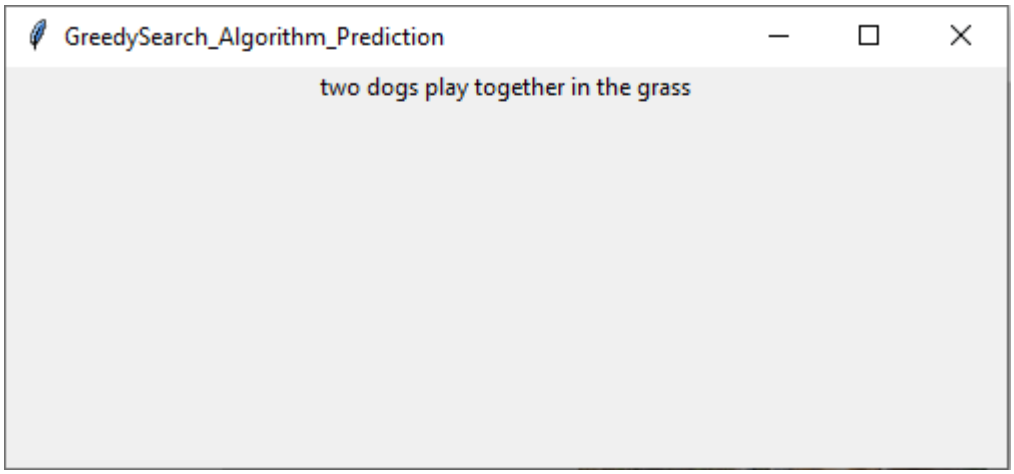

Figure4. Greedy Search Algorithm Output

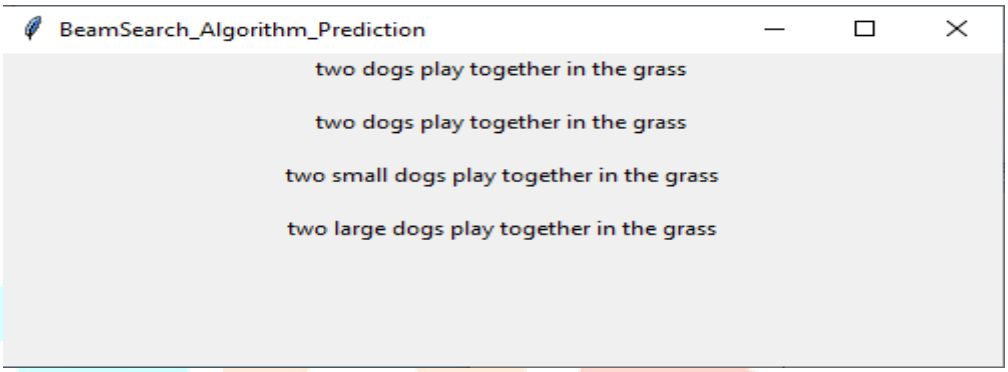

Figure5. Beam Search Algorithm Output

V. CONCLUSION

Image captioning has many advantages in almost every complex area of Artificial Intelligence. The main use case of our model is to help visually impaired to understand the environment and made them easy to act according to the environment. As, this is a complex task to do, with the help of pre trained models and powerful deep learning frameworks like Tensorflow and Keras, we made it possible. This is completely a Deep Learning project, which makes use of multiple Neural Networks like Convolutional Neural Network and Long Short Term Memory to detect objects and captioning the images. To create an interface to our model, we have used Tkinter.

REFERENCES

[1] K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares,H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation.In EMNLP, 2014.

[2] D. Elliott and F. Keller. Image description using visual dependency representations. In EMNLP,2013.

[3] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young,C. Rashtchian, J. Hockenmaier, and D. Forsyth.Every picture tells a story: Generating sentences from images. In ECCV, 2010.

[4] R. Gerber and H.-H. Nagel. Knowledge representation for the generation of quantified natural language descriptions of vehicle traffic in image sequences. In ICIP. IEEE, 1996.

[5] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In ECCV, 2014.

[6] S. Hochreiter and J. Schmidhuber. Long short-term memory.Neural Computation, 9(8), 1997.

[7] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. JAIR, 47, 2013.

[8] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In arXiv:1502.03167, 2015.

[9] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. NIPS,2014.

[10] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. In arXiv:1411.2539, 2014.

[11] R. Kiros and R. Z. R. Salakhutdinov. Multimodal neural language models. In NIPS Deep Learning Workshop, 2013.

[12] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In CVPR, 2011.

[13] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi. Collective generation of natural image descriptions. In ACL, 2012.

[14] P. Kuznetsova, V. Ordonez, T. Berg, and Y. Choi. Treetalk:Composition and compression of trees for image descriptions.ACL, 2(10), 2014.

[15] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. Composing simple image descriptionsusing web-scale n-grams. In Conference on Computational Natural Language Learning,2011