# Image Captioning using Deep Neural Architectures

Parth Shah
Department of Computer Engineering
C. G. Patel Institute of Technology
Uka Tarsadia University
Bardoli, India
parthpunita@yahoo.in

Vishvajit Bakrola
Department of Computer Engineering
C. G. Patel Institute of Technology
Uka Tarsadia University
Bardoli, India
vishvajt.bakrola@utu.ac.in

Supriya Pati
Department of Computer Engineering
C. G. Patel Institute of Technology
Uka Tarsadia University
Bardoli, India
supriya.pati@utu.ac.in

***Abstract*— Automatically creating the description of an image using any natural language sentences is a very challenging task. It requires expertise of both image processing as well as natural language processing. This paper discusses about different available models for image captioning task. We have discussed about how the advancement in the task of object recognition and machine translation has greatly improved the performance of image captioning model in recent years. In addition to that we have discussed how this model can be implemented. At the end, we have also evaluated the performance of model using standard evaluation matrices.**

***Index Terms*— Deep Learning, Deep Neural Network, Image Captioning, Object Recognition, Machine Translation, Natural Language Processing, Natural Language Generation.**

## I. INTRODUCTION

A single image contains large amount of information. Humans have ability to parse this large amount of information by single glance of it. Humans normally communicate through written or spoken languages. We can use natural languages for describing any image. Every individual can generate different caption for same image. If we can achieve same task with machine it will be greatly helpful for variety of tasks. However, generating captions for an image is very challenging task for today's machine. Generation of image caption with machine requires brief understanding of natural language processing and ability to identify and relate objects in an image. Some of the earlier approaches focuses on solving this challenge are based on hard-coded features and well defined syntax. This limits the type of sentences that can be generated by any given model. In order to overcome this limitation the main challenge is to make a model free of any hard-coded feature or sentence templates. Rules for forming models should be learned from the training data.

Another challenge is that there are large number of images available with their associated captions in the ever expanding internet. However, most of them are noisy and it cannot be directly used in image captioning model. Training an image captioning model requires dataset having collection of large number of properly annotated images generated by multiple persons.

In this paper, we have studied collection of different existing natural image captioning models and how they compose new caption for unknown images. We have also presented results of our implementation of this model along with comparison.

Section II of this paper describes related work in detail. Existing image captioning model named - Show & Tell is described in Section III. Section IV contain details about implementation environment and dataset. Results and comparison is presented in Section V. At the end this article in section VI we have provided our concluding remarks.

## II. RELATED WORK

With the intelligent reasoning capacity humans are capable to generate caption by combining objects and their relationship in an image. Creating captioning system that accurately generate captions like human is challenging task. Image can be described using more than one sentences but to efficiently train the image captioning model we require only single sentence that can be provided as a caption. This leads to a problem of text summarization in natural language processing.

There are mainly two different ways to perform the task of image captioning. These two types are basically retrieval based method and generative method. Most of work is done based on retrieval based method. One of the best model of retrieval based method is Im2Txt model [1]. It was proposed by Vicente Ordonez, Girish Kulkarni and Tamara L Berg. Their system is divided into mainly two part - Image matching and Caption generation.

First we will provide our input image to the model. Matching image will be retrieved from database that contain images and its appropriate caption. Once we find matching images we will compare extracted high level objects from original input image and matching images. Images will be then reranked based on matching content. Once it is reranked, caption of top-n ranked images will be returned. The main limitation of these retrieval based method is that it can only produce captions which are already present in database. It cannot generate novel captions.

This limitation of retrieval based method is solved in generative models. Using generative models we can create novel sentences. Generative models can be of two types either pipeline based model or end to end model. Pipeline type models uses two separate learning process, one for language modeling and other for image recognition. They first identify objects in image and provide the result of it to language modeling task. While in end-to-end models we combine both

978-1-5090-3294-5/17/$31.00 ©2017 IEEE

language modeling and image recognition models in single end to end model [2]. Both part of model learn at the same time in end-to-end system. They are typically created by combination of convolutional and recurrent neural networks.

Show & Tell model proposed by Vinyals et al. is of generative type end-to-end model. Show & Tell model uses recent advancement in image recognition and neural machine translation for image captioning task. It uses combination of Inception-v3 model and LSTM cells [3].

Here, Inception-v3 model will provides object recognition capability while LSTM cell provides language modeling capability [4][5].

## III. Show & Tell Model

Recurrent neural networks is generally used in neural machine translation [6]. They encodes the variable length inputs into a fixed dimensional vectors. Then it uses these vector representation to decode the desired output sequence [7][8]. Instead of using text as input to encoder, Show & Tell model uses image as an input. This image is then converted to word vector and then word vector is translated to caption using recurrent neural networks as decoder.

To achieve this, Show & Tell model is created by hybridizing two different models. It takes input as an image and provides it to Inception-v3 model. At the end of Inception-v3 model, single fully connected layer is added. This layer will transform output of Inception-v3 model into word embedding vector. We input this word embedding vector into series of LSTM cells. LSTM cell provides ability to store and retrieve sequential information through time. This helps to generate the sentences with keeping previous words in context.

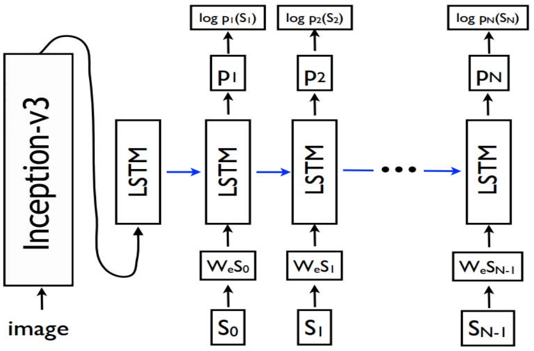


Fig. 1. Architecture of Show & Tell Model

Training of Show & Tell model can be divided into two part. First part is of training process where model learns its parameters. While second part is of testing process. In testing process we infer the captions and we compare and evaluate these machine generated caption with human generated captions.

### A. Training

During training phase we provide pair of input image and its appropriate captions to Show & Tell model. Inception-v3 part of model is trained to identify all possible objects in an image. While LSTM part of model is trained to predict every word in the sentence after it has seen image as well as all previous words. For any given caption we add two additional symbols as start word and stop word. Whenever stop word is encountered it stops generating sentence and it marks end of string. Loss function for model is calculated as,

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

where $I$ represent input image and $S$ represent generated caption. $N$ is length of generated sentence. $p_t$ and $S_t$ represent probability and predicted word at the time $t$ respectively. During the process of training we have tried to minimize this loss function.

### B. Inference

From various approaches to generate a caption from given image, Show & Tell model uses Beam Search to find suitable words to generate caption. If we keep beam size as K, it recursively consider K best words at each output of the word. At each step it will calculate joint probability of word with all previously generated words in sequence. It will keep producing the output until end of sentence marker is predicted. It will select sentence with best probability and outputs it as caption.

## IV. Implementation

For evaluation of image captioning model we have implemented Show & Tell model. Details about dataset, implementation tool and implementation environment is given as follows:

### A. Datasets

For task of image captioning there are several annotated images dataset are available. Most common of them are Pascal VOC dataset and MSCOCO Dataset. In this work MSCOCO image captioning dataset is used. MSCOCO is a dataset developed by Microsoft with the goal of achieving the state-of-the-art in object recognition and captioning task. This dataset contains collection of day-to-day activity with their related captions. First each object in image is labeled and after that description is added based on objects in an image. MSCOCO dataset contains image of around 91 objects types that can be easily recognizable by even a 4 year old kid. It contains around 2.5 million objects in 328K images. Dataset is created by using crowd sourcing by thousands of humans [9].

### B. Implementation Tool and Environment

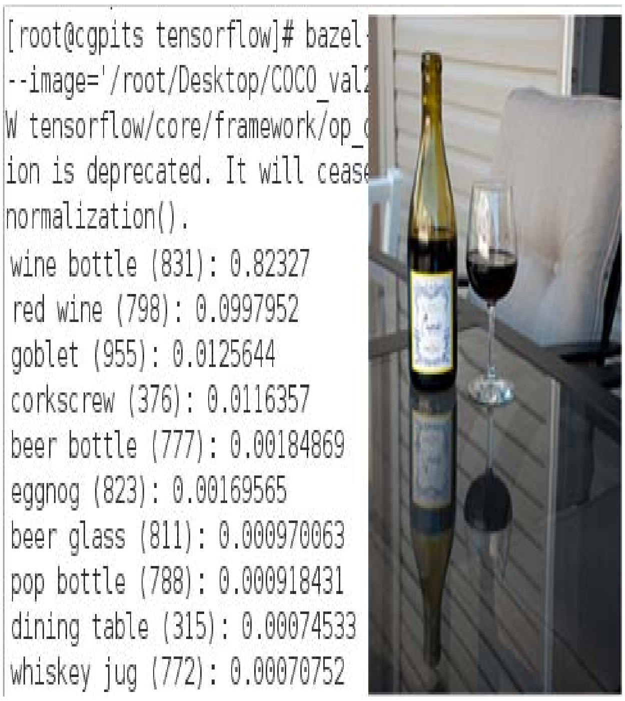
For the implementation of this experiment we have used machine with Intel Xeon E3 processor with 12 cores and 32GB RAM running CentOS 7. TensorFlow library is used for creating and training deep neural networks. TensorFlow is a deep learning library developed by Google [10]. It provides heterogeneous platform for execution of algorithms i.e. it can be run on low power devices like mobile as well as large scale distributed system containing thousands of GPUs. In order to define structure of our network TensorFlow uses graph definition. Once graph is defined it can be executed on any supported devices.

## V. Results and Comparison

### A. Results

By the implementation of the Show & Tell model we are able to generate moderately comparable captions with compared to human generated captions.

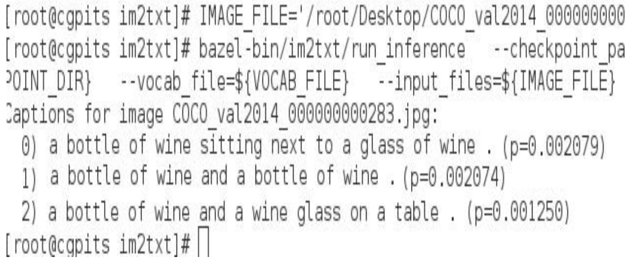
First of all model will identify all possible objects in image.



```
[root@cgpits tensorflow]# bazel
--image='/root/Desktop/COCO_val2
W tensorflow/core/framework/op_
ion is deprecated. It will cease
normalization().
wine bottle (831): 0.82327
red wine (798): 0.0997952
goblet (955): 0.0125644
corkscrew (376): 0.0116357
beer bottle (777): 0.00184869
eggnog (823): 0.00169565
beer glass (811): 0.000970063
pop bottle (788): 0.000918431
dining table (315): 0.00074533
whiskey jug (772): 0.00070752
```

Fig. 2. Generated word vector from Sample Image

As shown in Fig. 2 Inception-v3 model will assign probability of all possible objects in an image and convert image into word vector. This word vector is provided as input to LSTM cells which will then form sentence from this word vector as shown in Fig. 3 using beam search as described in previous section.



```
[root@cgpits im2txt]# IMAGE_FILE='/root/Desktop/COCO_val2014_00000000
[root@cgpits im2txt]# bazel-bin/im2txt/run_inference   --checkpoint_pa
POINT_DIR}   --vocab_file=${VOCAB_FILE}   --input_files=${IMAGE_FILE}
Captions for image COCO_val2014_000000000283.jpg:
  0) a bottle of wine sitting next to a glass of wine . (p=0.002079)
  1) a bottle of wine and a bottle of wine . (p=0.002074)
  2) a bottle of wine and a wine glass on a table . (p=0.001250)
[root@cgpits im2txt]# 
```

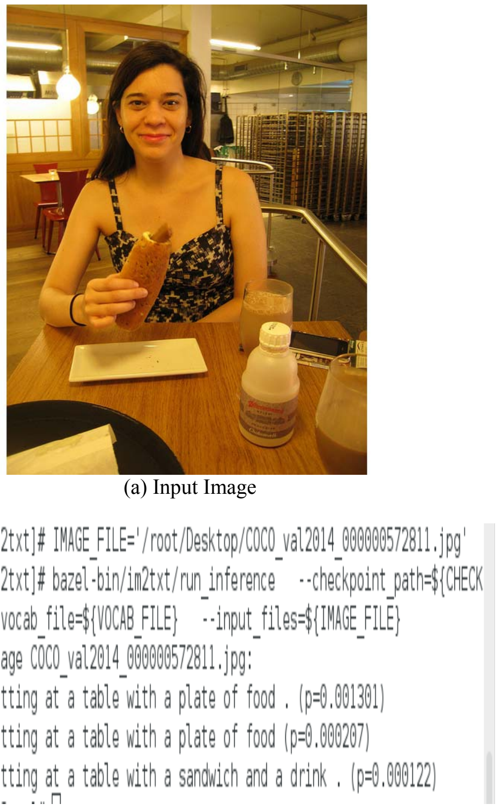
Fig. 3. Generated caption from word vector for Sample Image.

### B. Evaluation Matrices

Evaluation of any model that generates natural language sentence BLEU (Bilingual Evaluation Understudy) Score is used. It describes how natural sentence is compared to human generated sentence [11]. It is widely used to evaluate performance of Machine translation. Sentences are compared based on modified n-gram precision method for generating BLEU score [12] where precision is calculated using following equation:

$$p_n = \frac{\sum_{C\in\{Candidates\}} \sum_{ngram\in C} Count_{clip}(ngram)}{\sum_{C'\in\{Candidates\}} \sum_{ngram'\in C'} Count(ngram')} \quad (2)$$

In order to evaluate our model we have used image from validation dataset of MSCOCO Dataset. Some of the captions generated by Show & Tell model is shown as follows:



(a) Input Image

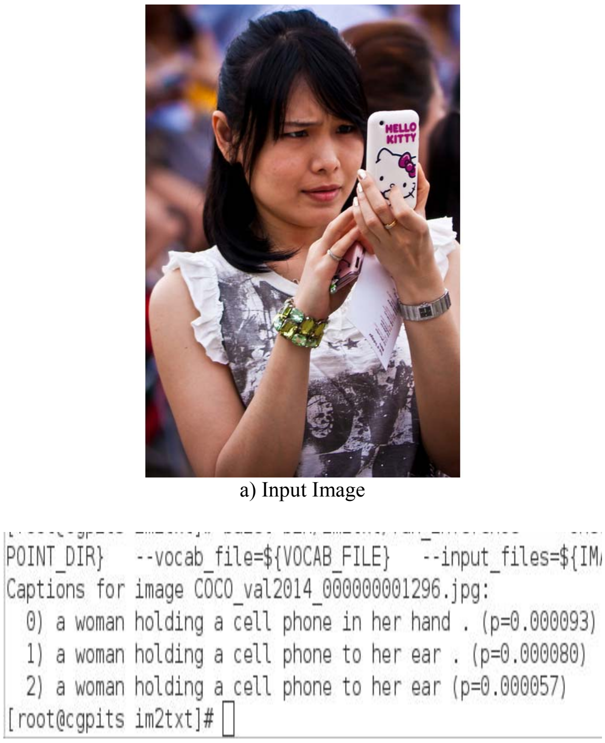
```
[root@cgpits im2txt]# IMAGE_FILE='/root/Desktop/COCO_val2014_000000572811.jpg'
[root@cgpits im2txt]# bazel-bin/im2txt/run_inference   --checkpoint_path=${CHECK
POINT_DIR}   --vocab_file=${VOCAB_FILE}   --input_files=${IMAGE_FILE}
Captions for image COCO_val2014_000000572811.jpg:
  0) a woman sitting at a table with a plate of food . (p=0.001301)
  1) a woman sitting at a table with a plate of food (p=0.000207)
  2) a woman sitting at a table with a sandwich and a drink . (p=0.000122)
```

(b) Generated Captions

Fig. 4. Experiment Result

As you can see in Fig. 4, generated sentence is “a woman sitting at a table with a plate of food.”, while actual human generated sentences are “The young woman is seated at the table for lunch, holding a hotdog.”, “a woman is eating a hotdog at a wooden table.”, “there is a woman holding food at a table.”, “a young woman holding a sandwich at a table.” and “a woman that is sitting down holding a hotdog.”. This result in BLEU score of 63 for this image.

Similarly in Fig. 5, generated sentence is “a woman holding a cell phone in her hand.” while actual human generated sentences are “a woman holding a Hello Kitty phone on her hands”, “a woman holds up her phone in front of her face”, “a woman in white shirt holding up a cellphone”, “a woman checking her cell phone with a hello kitty case” and “the asian girl is holding her miss kitty phone". This result in BLEU score of 77 for this image.



a) Input Image

```
POINT_DIR}   --vocab_file=${VOCAB_FILE}   --input_files=${IM
Captions for image COCO_val2014_000000001296.jpg:
  0) a woman holding a cell phone in her hand . (p=0.000093)
  1) a woman holding a cell phone to her ear . (p=0.000080)
  2) a woman holding a cell phone to her ear (p=0.000057)
[root@cgpits im2txt]# 
```

b) Generated Captions

Fig. 5. Experiment Result

While calculating BLEU score of all image in validation dataset we get average score of 65.5. Which shows that our generated sentence are very similar compared to human generated sentence.

## VI. Conclusion

We can conclude from our findings that we can combine recent advancement in Image Labeling and Automatic Machine Translation into an end-to-end hybrid neural network system. This system is capable to autonomously view an image and generate a reasonable description in natural language with better accuracy and naturalness.

## Acknowledgment

We would like to thank Department of Computer Engineering, C. G. Patel Institute of Technology for providing us computer resources as and when needed for training and implementing models presented in this paper here.

## References

[1] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2text: Describing images using 1 million captioned photographs," in Advances in Neural Information Processing Systems, pp. 1143–1151, 2011

[2] A. Karpathy and L. 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.

[3] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PP, no. 99, pp. 1–1, 2016.

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9, 2015. 28.

[5] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," arXiv preprint arXiv:1512.00567, 2015. 37.

[6] D. Britz, "Introduction to rnns." WILDML, http://www.wildml.com/, 2016. [Accessed 4-September-2016].

[7] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," CoRR, vol. abs/1609.08144, 2016.

[8] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in Advances in neural information processing systems, pp. 3104–3112, 2014.

[9] 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," in European Conference on Computer Vision, pp. 740–755, Springer, 2014.

[10] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," arXiv preprint arXiv:1603.04467, 2016.

[11] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in Proceedings of the 40th annual meeting on association for computational linguistics, pp. 311– 318, Association for Computational Linguistics, 2002.

[12] D. Jurafsky, Speech & language processing. Pearson Education India, 2000.