

# Show and Tell: A Neural Visual Story-Teller

Linton Pereira

Final Year, Computer Engineering  
St. Francis Institute of Technology  
Mumbai, India  
pereira.linton@student.sfit.ac.in

Melron Pinto

Final Year, Computer Engineering  
St. Francis Institute of Technology  
Mumbai, India  
m.godzzzz@student.sfit.ac.in

Kinshuk Shah

Final Year, Computer Engineering  
St. Francis Institute of Technology  
Mumbai, India  
kinshukshah@student.sfit.ac.in

Shamsuddin Khan

Asst. Professor Computer  
Engineering  
St. Francis Institute of Technology  
Mumbai, India  
shamsuddinkhan@sfit.ac.in

**Abstract**—Automatically narrating the idea of an image is an elemental problem in artificial intelligence that links computer vision and natural language processing. RNN, CNN and NLP are powerful techniques that help us to work toward generation of captions and further generating a story. This model will describe complex images using technologies like Recurrent Neural Network (RNN) and Natural Language Processing (NLP). RNN consists of memory: it feeds its inputs back into itself, so it can remember past information. RNNs are used for sequential data. But they do have some flaws in remembering information. This is why CNN is used for generating captions as they are good at identifying patterns.

**Keywords**— *Recursive Neural Network (RNN), Convolution Neural Network (CNN), Natural Language Processing (NLP), Long Short-Term Memory (LSTM).*

## I. INTRODUCTION

This model is an extension of a single image captioning model to a sequence to sequence model that produces a story-outline from a sequence of five images. The context vector is derived from a series of images by using an encoder LSTM. The LSTM decoders use valuable information from the image sequence that is present in the context vector. Recursive Neural Network (RNN), Convolution Neural Network (CNN) and Natural Language Processing (NLP) are the most important and valuable algorithms that help us to convert images to captions. Many deep learning algorithms such as CNN helps us to visualize and get the meaning of an image. CNN algorithms take various input parameters which would mainly consist of pictures of any material or things like dogs, cats, humans, etc. Computer only recognizes the image as a sequence of pixels and that truly depends upon the picture resolution. Based on the resolution of the image, it will generate Height(H), Width(W) and Dimensions(D) as  $H \times W \times D$ . Convolution helps to extract features from an input image. Convolution learns the input image by its learning parameter and considering the small squares of the input image. RNNs are used for sequential data. Since they have hidden states to remember old information, they can understand all the words present in a sentence, video frames, pixels in an image. In the old approach of the neural network the input and output images are independent of each other but

now as our model deals with the connected images it is necessary to store the previous history and words. This problem is solved by RNN which includes a Hidden Layer. The main and the most important thing in RNN is the hidden layer which helps us to store some amount of information for a particular amount of time. It uses the same parameters for each input as it performs the same task on all the inputs or hidden layers to produce the output. Natural Language processing helps us to understand each text generated and to process and give valuable insights of the data by analyzing it. LSTM which is mainly called as Long short-term memory is an important part of deep learning which helps us to store the information generated and later helps to give more accurate results by knowing the past data. Unlike the standard feedforward neural networks, LSTM has feedback connections. It can process both images as well as videos at the same time. Recurrent Neural Networks suffer from short-term memory. The long sequence takes a large amount of time to process the data. So, it is useful to store and get the information previously. LSTM was created as a solution to short-term memory.

## II. LITERATURE SURVEY

A baseline approach was proposed in another paper which consisted of a sequence to sequence model. In this paper the encoder takes a sequence of images as input and the decoder takes the last state of the encoder as its first state to generate the story.[2]

Another paper presented a multi-task model that performed album summarization and story generation. The model achieved quality outcome on the VIST datasets yet, some of the descriptions were confusing and the output that was generated was unclear. This is because in this paper they did not use the approach of LSTM which our model is using.[3]

The task of describing images with sentences has been explored. Several approaches generate image captions based on fixed templates that are filled based on the content of the image or generative grammars, but this approach limits the variety of possible outputs. Most closely related [4]

developed a log bilinear model that can generate full sentence descriptions for images, but their model uses a fixed window context while our Recurrent Neural Network (RNN) model works with LSTM which gives better accuracy and each time the data is stored and used, the output of our model gets more closer to human thinking. [5]

Since analysts started working on object detection in images, it seems clear that just providing the names of the objects seen in the pictures is not sufficient, unlike describing what the image wants to tell us. Hence, it became necessary to give more detail and thus came the need for further classification of those images. As opposed to object-bounding box localization and semantic pixel level segmentation, we will use CNN to perform the major tasks, namely image recognition and image classification. Finally, instead of evaluating instance detection with segmentation mask we will use RNN to generate generic phases which will be used to generate and improve the generated story. [8]

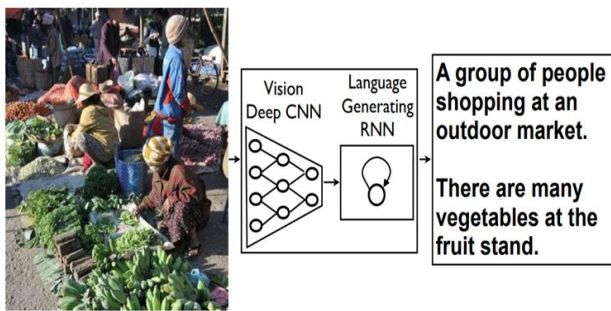


Figure 1: Basic model of Image Description

This model expands the single image captioning model to a sequence to sequence model that will produce a story-outline based on the series of five input images given. This add-on is dependent on LSTM to determine a context vector from the series of images. This context vector contains helpful data from the sequence of images that is used by the LSTM decoders that will lead to a story-outline.

### III. CHALLENGES IDENTIFIED

#### A. Describing a complex image

Describing a simple image is not that difficult as it contains a single digit or a single object. Whereas, complex images contain many objects, they don't have a standard background. Due to the diversity present in complex image; our approach is to identify different objects and then create tags based on those images. After identifying the tags, a sentence most suitable to those tags needs to be generated. However, to have a better accuracy in identifying objects and generating accurate information as that of a human mind, requires a large dataset.

#### B. Creating a story outline of the description of the images

Combining the description of five images to one story is not just appending the description but making sense of the available data (description). Each image should have an inter-connected story which can give the final story outline with a sequence or the description of the story part by part.

### IV. PROBLEM DEFINITION

Describing the contents of an image was a challenge in itself, here we will be generating caption as well as passing that captions into LSTMs to make generate a story line. Since caption generation has already been established, we have taken this a step further by generating a storyline which will be a mixture of all the five images that have been described and the connectors which will be generated based on the probability generated by our system.

### V. PROPOSED SYSTEM METHODOLOGY

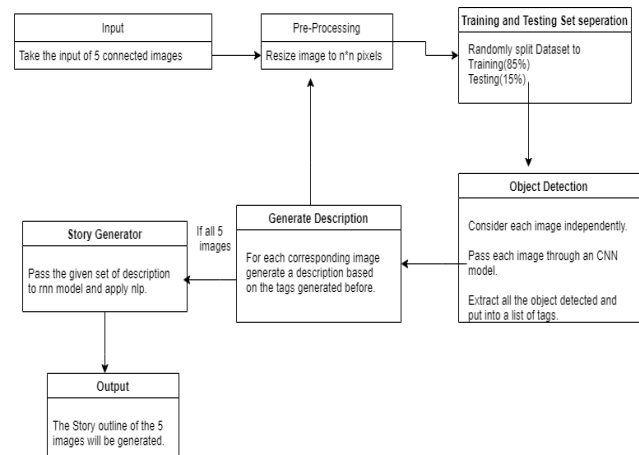


Figure 2: Flow Chart of Methodology

#### A. Data

This model will be trained using the Stories of Images-in-Sequence (SIS) dataset. Mostly, the story generated from SIS will be a five-sentence text where each sentence corresponds to one of five images in the sequence. The dataset contains 81,743 distinct photos in 20,211 sequences, lined up to descriptive and story language.

#### B. Recurrent Neural Network (RNN)

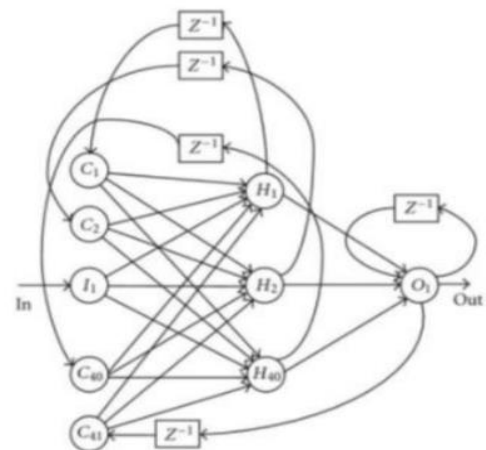


Figure 3: RNN model

RNN is a type of neural network where the output of the previous steps is again fed to the current step. The important

feature of RNN is that it includes a hidden state which helps the model to remember the information for a short time span. The information that has been generated is stored in the memory of RNN. Similar parameters are considered for all inputs as they carry out similar tasks on all the inputs or hidden layers to produce an output. Unlike other neural networks, this reduces the complexity of parameters. RNN transforms the independent activation into dependent activation by providing the biases to all the layers and similar weights. This will help reduce the complexity of increasing parameters and by remembering all the previous results by giving each output as an input to the next hidden layer. Therefore, these three layers can be joined together into a single recurrent layer, such that the weights and bias of all the hidden layers is the same. Figure 3 shows the basic Architecture of Recurrent Neural Network.

### C. Long-Short Term Memory (LSTM)

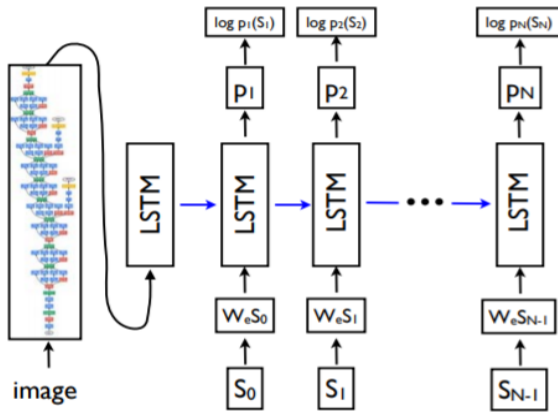


Figure 4: Working of LSTM

LSTM is an artificial RNN architecture used in the field of deep learning. LSTM has feedback connections unlike standard feed forward neural networks. It can process single data points (such as images) as well as entire sequences of data (such as speech or video). Recurrent Neural Networks suffer from short-term memory. If these networks are given a long sequence, they will have a hard time carrying information from earlier time steps to later ones. So, for example if we are trying to process a paragraph of text to do some kind of predictions, in this case RNN may leave out some useful information. LSTM was created as a solution to short-term memory. The gates help us to regulate the flow of data. These gates can learn which data in a sequence of data is important to keep or throw away. This can help to make predictions by passing the long chain of sequence. Almost all state-of-the-art outcomes based on RNN are accomplished with these two networks.

### D. Process

Our approach will consist of 2 major steps: Firstly, we will describe all the five images independently. To describe these images, we will implement Convolution Neural Network. CNN is one of the main categories to perform tasks like image recognition and image classification. The CNN which acts as an encoder is first pre-trained for image classification tasks and later the last hidden layer will act as an input to RNN decoder which will generate the sequences (see Figure 1). Since the images obtained will not be simple but complex images, we will identify tags and using those tags try to describe the image. Then we'll feed this description obtained

from these five images to our LSTM which will convert the description to a story-outline.

## VI. MODEL

Our model extends the image description model which consists of an architecture of encoder and decoder. As shown in Figure 5 we can observe that the given CNN acts as encoder and LSTM acts as decoder. The given images are passed through the encoder which will generate the description of the image then it is passed to decoder which helps us to know the actual content of the image and after evaluating it generates the description word by word. Below we have described how we extended this model for the visual storytelling task.

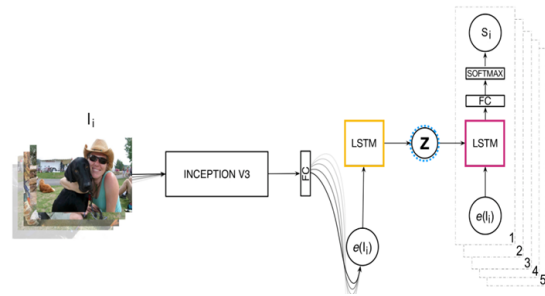


Figure 5: Working of our model

### A. Encoder

The initial component of our model is an RNN, precisely an LSTM network which sums-up the sequence of images. The network takes an image as input from the sequence at every timestep  $t$ . When  $t=5$ , the LSTM will have encoded the 5 images and will provide the sequence's context through its last hidden state.

### B. Decoder

The second LSTM network of our model is a decoder that uses the information obtained from the encoder to generate the description's. The last hidden state of the encoder is initialized to the first hidden state of the decoder. The first input to the decoder is the image for which the descriptions are being generated. Using this similar strategy, we provide the decoder the context of the whole sequence as well as the current picture content (i.e. global and local information) to generate the corresponding description which will contribute to our overall story-outline. For each image in we have got independent decoders. All the decoder's first hidden state is being initialized to the last hidden state of the encoder and takes the corresponding image plant as its first input. The first image is described in an order of words by the first decoder, second image by second decoder and so on. A particular language model for each location is well read by each decoder which corresponds to the order.

For example, the first decoder will gain an understanding of opening sentences while the last decoder will learn about closing sentences. For each image in the order, we obtain its portrayal  $\{e(I_1), e(I_2), e(I_3), e(I_4), e(I_5)\}$  by making use of Inception v3. The encoder takes one picture at every timestep  $t$ . At  $t=5$ , we get the context vector through  $\{Z=h(t)e\}$ . This vector is used to store each decoder's hidden state while the first input to each decoder is its corresponding image place.

Each decoder leads to a sequence of texts {p1, p2, p3,..pn} for each image in series. The final story generated is the integration of the output of 5 decoders.

## VII. PERFORMANCE EVALUATION PARAMETER

As seen in Figure 1 we observe that a particular picture can be interpreted in many different ways and this may vary from person to person. Our model might also give a different explanation for a given picture. We will evaluate our performance of our model by comparing the output of our model with the story outline obtained by a human.

## VIII. RESULT

As seen in Table1 we observe that a particular picture can be interpreted in many different ways and this may vary from person to person. Table I shows some sample stories generated by our model from the public test set of the Visual Storytelling. The description of the images may contain some repetitive ideas but the sentences are grammatically correct. Sometimes it may happen that the stories generated does not relate to the actual content of the image. The limitations of our model completely depends upon how the humans interpret it. Currently, we trained up to 50,000 images. As it's difficult to evaluate the accuracy of our model we used meteor metrics which are used for machine translation evaluation. Our model obtained a correlation of 0.3331 in the test set of 1000.

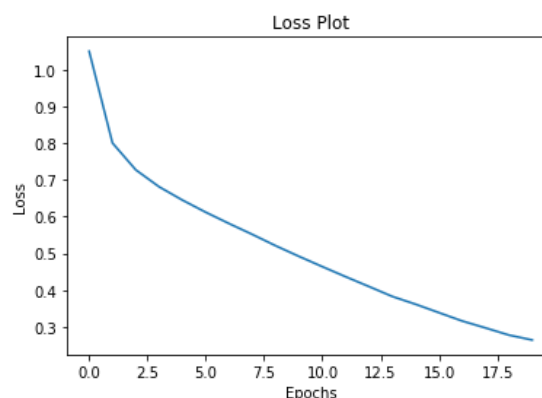


Figure 6: Train Loss Function.

Figure 6 displays the decrease in loss with respect to each epoch.






## CONCLUSION & FUTURE WORK

In this paper, we tackled a new problem of generating a story-outline from a sequence of images which was far more complex than describing independent images. We exploited context aware similarity to compute similarities between the image descriptions. This approach can be further improved by adding a larger dataset for a specific application. This technology has the capacity to make our world more approachable and understandable for us humans. This model could be taken a step further by using it in live commentary. Based on a live video narration can be generated. This system can also be used in image indexing. For example, we can describe the image we are looking for and the model will search for the image in the database or a directory. This model can also be used to help the visually impaired by capturing live images to guide them as they walk on the streets by narrating to them about the environment. It can also make children more creative as children can use this system by selecting five images and can try to write their own story before the model generates its own.

Table I. Sample stories generated by our model, compared to generated by humans.

Images					
Computer Evaluation	The dog was ready to go.	He had a great time on the hike.	And was very happy to be in the field.	His mom was so proud of him.	It was a beautiful day for him.
Human Evaluation	The dog is going on a search.	The dog spotted something in the grass.	The dog is waiting for his master.	The Dog's master is patting the dog.	It was a happy day for him.



<b>Images</b>					
<b>Computer Evaluation</b>	Today was graduation day.	The students were excited.	My parents were so happy.	He was very happy to be graduating.	Everyone was proud of him.
<b>Human Evaluation</b>	Today was graduation day and he was extremely happy.	However, he was nervous about what the future would bring.	His parents assured him that he would do well in life. That helped a little.	Of course, when Benny the squirrel gave him life advice his whole demeanor turned happily.	He is now ready for life, after the first big chapter ending high school.

#### REFERENCES

- [1] Cesc C Park & Gunhee Kim, "Expressing an image stream with a sequence of natural sentences. Advances in Neural Information Processing Systems", 2015, pp. 73-81.
- [2] Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, & Dhruv Batra, "Visual storytelling", In 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2016.
- [3] Licheng Yu, Mohit Bansal, & Tamara Berg, "Hierarchically-attentive rnn for album summarization and storytelling", In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 966-971.
- [4] R. Kiros, R. S. Zemel, and R. Salakhutdinov. Multimodal neural language models. ICML, 2014.
- [5] 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.
- [6] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 664-676, 1 April 2017.
- [7] K. Subramanian, D. Stallard, R. Prasad, S. Saleem and P. Natarajan, "Semantic translation error rate for evaluating translation systems," 2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU), Kyoto, 2007, pp. 390-395.
- [8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick, "Microsoft coco: Common objects in context", In Computer Vision - ECCV, 2014, pp. 740-755
- [9] Y. Cui, G. Yang, A. Veit, X. Huang and S. Belongie, "Learning to Evaluate Image Captioning," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 5804-5812.
- [10] N. K. Kumar, D. Vigneswari, A. Mohan, K. Laxman and J. Yuvaraj, "Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Coimbatore, India, 2019, pp. 107-109.
- [11] A. Poghosyan and H. Sarukhanyan, "Short-term memory with read-only unit in neural image caption generator," 2017 Computer Science and Information Technologies (CSIT), Yerevan, 2017, pp. 162-167.