

Image Caption Generator

[1]Shubham Kohli, [2]Himanshu Rawat [3]Priyanshu Joshi, [4]Madhav,
Mr. Devender Banga

^{1,2,3,4} Student, Department of Information Technology, Dr. Akhilesh Das Gupta Institute of Technology and Management, Delhi, India.

Assistant Professor, Department of Information Technology, Dr. Akhilesh Das Gupta Institute of Technology and Management, Delhi, India.

Date of Submission: 10-12-2020

Date of Acceptance: 25-12-2020

ABSTRACT: Image captioning models usually follow an associate encoder-decoder design that uses abstract image property vectors as input to the encoder. One of the foremost algorithms uses a feature vector extracted from the region proposals obtained from the associate object detector during this work, we have a tendency to introduce the Object Relation electrical device that builds upon this approach by expressly incorporating info about the spatial relationship between input detected objects through geometric attention. Quantitative and qualitative results demonstrate the importance of such geometric attention for image captioning, leading to enhancements on all common captioning metrics on the glint 8K dataset.

I. INTRODUCTION

Computer innovation in the image processing area has made remarkable progress in the past few years, like image classification and object detection. With the advancement of image classification and object detection, the process of generating textual description from an image based on the objects and actions in the image which is known as Image Captioning becomes feasible. Some applications like titles to news pictures, descriptions of medical pictures, text-based image retrieval, the knowledge required for blind users, human-robot interaction in image captioning have vital theoretical and sensible analysis worth. Therefore, image captioning becomes a more complicated task but the meaningful task in the age of artificial intelligence. On the natural language processing side aspect,

additional refined successive models, like attention-based perennial neural networks, have equally resulted in additional correct caption generation. Impressed by neural artificial intelligence, most standard image captioning systems utilize an encoder-decoder framework,

during which an input image is encoded into an intermediate illustration of the knowledge contained among the image, and afterward decoded into a descriptive text sequence.

This secret writing will carry with it one feature vector output of a CNN or multiple visual options obtained from totally different regions among the image. In the latter case, the regions are often uniformly sampled, or target-hunting by associate degree object detector that has been shown to yield improved performance. Whereas these detection primarily based encoders represent the progressive, at the moment they are doing not utilize data concerning the spatial relationships between the detected objects, like relative position and size. This data will typically be vital to understanding the content at intervals a picture, however, and is employed by humans once reasoning concerning the physical world. Relative position, as an example, will aid in identifying “a woman riding a horse” from “a woman standing beside a horse”. Similarly, relative size will facilitate differentiate between “a lady enjoying the guitar” and “a lady enjoying the ukulele”. Incorporating spatial relationships has been shown to boost the performance of object detection itself, as incontestable in. moreover, in AI encoders, point relationships are typically encoded, especially within the case of the electrical device, attention based encoder design.



Can you write a caption?

Well, a number of you would possibly say “A white dog in an exceedingly grass like area” and some might say “White dog with brown spots”. Definitely, all of those captions are unit relevant for this image and there could some others also.

However, the purpose we would like to form is; it’s really easy for America, as folks, to only have a look at an image associate degree describe it in an acceptable language. Regardless of such challenges, the matter has accomplished vital enhancements over the past few years. Image captioning algorithms are unit generally divided into 3 categories. The primary cluster tackles the matter via the retrieval-based strategies, which initially retrieves the highest matching pictures, so transfers their descriptions because of the captions of the question. These strategies will manufacture grammatically correct sentences however cannot modify the captions consistent with the new image. The second cluster generally uses template-based strategies to get descriptions with predefined grammar rules and slit sentences into many elements. These strategies initially make the most of many classifiers to acknowledge the objects, furthermore as their attributes and relationships in a picture, so use a rigid sentence example to create a whole sentence. Although it will generate a brand new sentence, these strategies either cannot specific the visual context properly or generate versatile and significant sentences. With the intensive application of deep learning, most up-to-date works make up the third cluster referred to as neural network-based ways galvanized by machine learning’s encoder-decoder design, recent years most image captioning ways use a Convolutional Neural Network (CNN) because of the encoder and a repeated Neural.

Network (RNN) because the decoder, particularly Long remembering (LSTM) to get captions, with the target to maximize the chance of a sentence given the visual options of a picture. Some ways square measure victimization CNN because the decoder and therefore the reinforcement learning because the decision-

making network. In step with these completely different coding and decipherment ways, during this paper, we tend to divide the image captioning ways with neural networks into 3 categories: CNN-RNN based, CNN-CNN based, and reinforcement-based framework for image captioning. Within the next half, we are going to point out their main concepts.

II. CONVOLUTIONAL NEURAL NETWORK [CNN]

Convolutional neural network (ConvNets or CNNs) is a class of neural networks commonly used in the field computer vision. Regularized versions of multilayer perceptrons, CNN is shared-weights architecture therefore sometimes called shift invariant artificial neural networks. In place of general matrix multiplication CNN uses convolution i.e. operation on two functions that produces a third function that expresses how the shape of one is modified by the other in a LTI system.

II.1. CNN Architecture


CNN architecture is inspired by functionality of the visual cortex and designed to mimic the connectivity pattern of neurons within the human brain. The neurons within a CNN are split into a 3-D structure, with each set of neurons analyzing a small region or feature of the image. In other words, each group of neurons specializes in identifying one part in the image. CNNs use the predictions from the layers to produce a final output that presents a vector of probabilities to represent the likelihood that a specific feature belongs to a certain class or group.

II.2. CNN Layers

1. Convolution layer –

When working on a Convolutional neural network, the input is a tensor with shape (number of images) x (dimensions of images). Then a tensor is passed to Convolution Layer which creates a feature map to predict the class probabilities for each feature by applying a filter that scans the whole image, a few pixels at a time.

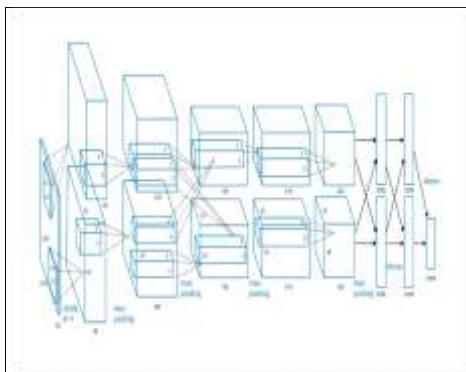
Some common feature maps:

Operation	Filter	Convolved Image
Identity	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	
Edge detection	$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}$	
	$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$	
	$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$	
Sharpen	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$	
Box blur (normalization)	$\frac{1}{9} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$	
Gaussian blur (approximation)	$\frac{1}{16} \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$	

2. Pooling layer (down sampling) - Pooling layers reduce the dimensions of the data by scaling down the amount of information the convolutional layer generated for each feature and maintains the most essential information
3. Fully connected input layer - Output from the previous layers are flattened to turn them into a single vector that can be used as an input for the next layer.
4. Fully connected layer - Fully connected layer is a multi-layer perceptron neural network (MLP) where a flattened matrix goes to classify the images.

Popular Convolutional Neural Network Architectures:

- **AlexNet (2012) -**
 AlexNet was not the first fast GPU-implementation of a CNN to win an image recognition contest. One of the most significant differences between AlexNet and other object detection algorithms is the use of ReLU for the non-linear part instead of Sigmoid function or Tanh like traditional neural networks. AlexNet leverages ReLU as activation to make their algorithm faster.



- **GoogleNet (2014) -**
 The GoogleNet model was inspired by LetNet, which is also named Inception V1, was made by a team at Google. It's architecture consists of a 22 layer deep CNN used a module based on inception module, which uses batch normalization, RMSprop and reduces the number of parameters from 60 million(in AlexNet) to only 4 million.

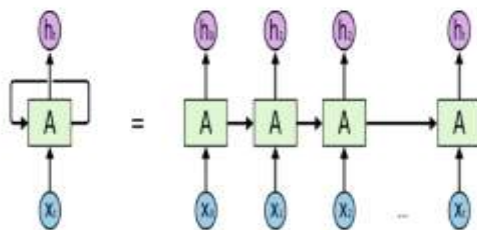
type	patch size	output size	depth	#1x1	#3x3	#5x5	#7x7	#9x9	pool	params	ops
conv2d	7x7	112x112x64	1							1.7M	30M
maxpool	3x3	56x56x64	0								
conv2d	3x3	56x56x128	2	64	64					1.1M	30M
maxpool	3x3	28x28x128	0								
inception (3)		28x28x256	2	64	96	128	16	32	32	1.9M	120M
inception (3)		28x28x480	2	128	128	160	32	96	64	3.8M	300M
maxpool	3x3	14x14x480	0								
inception (4)		14x14x512	2	128	96	208	16	48	64	3.6M	70M
inception (4)		14x14x512	2	160	112	224	32	64	64	4.7M	80M
inception (4)		14x14x512	2	128	128	256	32	64	64	4.8M	100M
inception (4)		14x14x512	2	112	144	288	32	64	64	5.8M	110M
inception (4)		14x14x512	2	128	160	320	32	128	128	6.8M	130M
maxpool	3x3	7x7x512	0								
inception (5)		7x7x512	2	128	160	320	32	128	128	10.7M	50M
inception (5)		7x7x512	2	192	192	384	48	128	128	13.6M	70M
avgpool	7x7	1x1x512	0								
dropout (50)		1x1x512	0								
linear		1x1x1000	1							1000K	1M
softmax		1x1x1000	0								

- **VGGNet (2014) -**
 VGGNet, consisting of 16 convolutional layers, reduces the parameters in the convolution layers and improves on training time. Similar to Alexnet but makes the improvement by replacing large kernel-sized filters of 11 and 5 in starting layers with multiple fixed sized 3X3 kernel-sized filters one after another.

Layer	Feature Map	Size	Kernel Size	Stride	Activation	
Input	Image	1	224 x 224 x 3	-	-	
1	2 X Convolution	64	224 x 224 x 64	3x3	1	relu
	Max Pooling	64	112 x 112 x 64	3x3	2	relu
3	2 X Convolution	128	112 x 112 x 128	3x3	1	relu
	Max Pooling	128	56 x 56 x 128	3x3	2	relu
5	2 X Convolution	256	56 x 56 x 256	3x3	1	relu
	Max Pooling	256	28 x 28 x 256	3x3	2	relu
7	3 X Convolution	512	28 x 28 x 512	3x3	1	relu
	Max Pooling	512	14 x 14 x 512	3x3	2	relu
10	3 X Convolution	512	14 x 14 x 512	3x3	1	relu
	Max Pooling	512	7 x 7 x 512	3x3	2	relu
13	FC	-	25088	-	-	relu
14	FC	-	4096	-	-	relu
15	FC	-	4096	-	-	relu
Output	FC	-	1000	-	-	Softmax

III. RECURRENT NEURAL NETWORK [RNN]

Recurrent Neural Network is an inference network of feed-forward neural network that has an Inner memory. RNN is repeated in nature because it performs an equivalent operation for each input data whereas the output of the present input depends on the past one computation. Once the output is manufactured, it is traced and sent into the Recurrent network. For call, it considers the current input and additionally the output that has been learned from the previous input. As compared with feed-forward neural networks, RNNs uses their inner state memory to process series of inputs. This makes them applicable to tasks like non-segmental, connected handwriting recognition, or speech recognition. Other than RNN in all Neural Network, all the inputs are independent of each other.



An unrolled recurrent neural network.

Initially, it takes the X (0) from the series of input then it outputs h (0) that along with X (1) is

the following step input. So, Inputs for the next step is h (0) and X (1). Similarly, h (1) from the following is that the input with X (2) for the following step and then on. This way, it keeps notes of context while processing. The formula is:

$$h_t = f(h_{t-1}, x_t)$$

Applying Activation method:

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$$

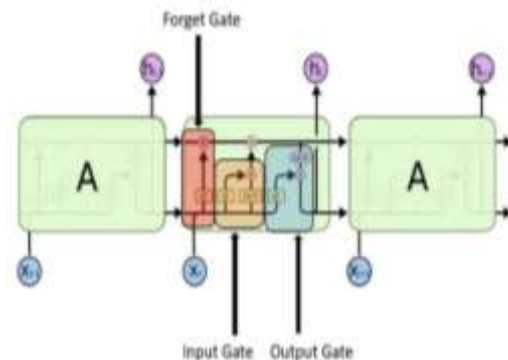
Here 'W' is weight, h is the single hidden vector, also Whh is the weight at previous hidden state, Whx is the weight at current input state, tanh is the Activation function in the formula.

Output:

$$y_t = W_{hy}h_t$$

Long Short Term Memory (LSTM):

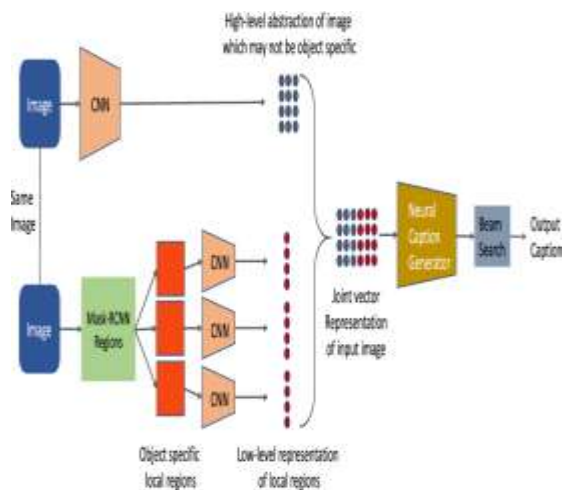
Long Short-Term Memory (LSTM) networks are an upgraded version of RNN, which makes it less complex to Memoize past data in memory. The vanishing gradient drawback of RNN is settled here .Long Short-Term Memory is used to classify ways and predict statistics given time lags of unknown length. It trains the model by backward-propagation. So, it has 3 gates units :



1. Input gate - finds out which value from input should be used to change the memory. The Sigmoid function checks which values to let pass from 0, 1 and tan-h function gives priority to the values which are passed checking their level of importance between -1 and 1 .
2. Forget gate - finds out what details to be ignored from the block. It is checked by the sigmoid function. It looks at the last state (ht-1) and the content input (X-t) and outputs a

number in range (0,1] for each number in the cell state C_{t-1} .

3. Output gate - input and the memory of the block is used to check the output. The Sigmoid function checks which values to let pass from 0, 1 and tan-h function gives priority to the values which are passed checking their level of importance between -1 and 1 and multiplied with an output of Sigmoid function.



IV. CNN –RNN FRAMEWORK

CNNs and RNNs are neural networks that can perform classification of image and text inputs. Although CNNs and RNNs, structured differently and applied for different purposes, they can process some of the same input types, creating an opportunity to combine the two network types for increased effectiveness. The combination can help in situations where the input to be classified is visually complex with some additional temporal characteristics then CNN alone would be unable to process.

When Convolutional and Recurrent neural networks are combined, they are sometimes referred to as a CRNN, inputs are first sent to and processed by CNN layers whose outputs are then

sent to RNN layers. CNN-LSTM architectures are particularly more promising than lower-level RNN architecture types as they facilitate analysis of inputs over longer periods.

Currently, these hybrid networks are being used in applications like gesture recognition, emotion detection, video identification and DNA sequence prediction. Most of the time, a captioning model is a combination of two separate architecture that is

Convolutional Neural Networks & Recurrent Neural Networks and in this case LSTM, which is a special kind of RNN, in order to maintain the information for a longer period of time. Our image caption generator model is a merge of these architectures and is called a CNN-RNN (CRNN) model.

- CNN extracts features from the image. We will use the pre-trained model
- LSTM will take information from CNN to generate a description of the image.

For the structure of the model, we will be using the Keras Model from Functional API. It will consist of three major parts:

- **Feature Extractor** – Extracted features from the image have a size of 2048, with the help of dense layers, it will be reduced to the dimensions of 256 nodes.
- **Sequence Processor** – Textual input will be handled by embedding layer, followed by the LSTM layer
- **Decoder** – By merging the output from the above two layers, we will process the dense layer to make the final prediction. The final layer will contain nodes equal to our vocabulary size.

Visual representation of the final model is given below –



V. DISCUSSION

V. 1. MAJOR CHALLENGES

Automatic image captioning remains challenging despite the recent spectacular progress in neural image captioning, though the application of neural network technology has opened up a brand new state of affairs for image captioning research, there are few issues that have not been solved.

i. Composibility and naturalness

Captioning systems suffer from a lack of compositionality and naturalness as they typically generate captions in an exceedingly ordered manner, i.e., next generated word depends on each previous word and the image feature. This could oftentimes result in syntactically correct, however semantically immaterial language structures, in addition to an absence of diversity in the generated captions.

ii. Generalization

The second challenge is that the dataset bias impacting current captioning systems. The

trained models overfit to objects that co-occur in a common context, that results in a drag such systems struggle to generalize to scenes where identical objects seem in unseen contexts.

iii. Evaluation and turing test

Using automatic metrics, although helpful remains dissatisfactory since they are doing not much to take the image under consideration. In several cases, their scoring remains inadequate and typically even misleading — particularly once evaluating various and descriptive captions. Human analysis remains a gold commonplace in the evaluation of captioning systems.

V. 2. BENIFITS

i. Human Computer Interaction

With the advancements of science and technology and the want for the event of human life, robots are utilized in additional industries. With safe and economical driving, it's additionally attainable to perform operations like automatic parking. Liberating the driver's eyes and hands will greatly facilitate people's lives and reduce safety

accidents. Auto-pilot robots can intelligently avoid obstacles, change lanes, and pedestrians based on the road conditions according to the surrounding driving environment they observe. If the machine wants to do the work better, it must interact with humans better. The machine can tell humans what it sees, and humans then perform appropriate processing based on machine feedback. To complete these tasks, we need to rely on automatic generation of image descriptions.

VI. CONCLUSION

So we've got enforced our CNN-RNN model by building a picture caption generator. Few key points to notice is that our model depends on the data, so, it cannot predict the words that which is out of its vocabulary. we tend to use a little dataset consisting of 8000+ pictures. For production-level models, we need to train on datasets larger than a 1 Lakh+ images which may turn out higher accuracy models.

Output (Based on Above Model):



CAPTION: a girl climbing the rock face

REFERENCES

- [1]. P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic propositional image caption evaluation. In European Conference on Computer Vision, pages 382–398. Springer, 2016.
- [2]. H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. From captions to visual concepts and back. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1473–1482, 2015.
- [3]. C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out, 2004.
- [4]. S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7008–7024, 2017.
- [5]. R. Vedantam, C. Lawrence Zitnick, and D. Parikh. CIDEr: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4566–4575, 2015.
- [6]. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3156–3164, 2015.
- [7]. T. Yao, Y. Pan, Y. Li, and T. Mei. Exploring visual relationship for image captioning. In European Conference on Computer Vision, pages 684–699, 2018.
- [8]. Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4651–4659, 2016.
- [9]. J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2625–2634, 2015.