Image Caption Generator

Saurav Pandey[1], Sudhir Kumar[2], Shailesh Kumar Gupta[3], Ritik Shukla[4]

Department of Computer Science and Engineering, Babu Banarasi Das Institute of Technology and Management, Lucknow, India

ABSTRACT

Wouldn't be great if blind people can know what is going around them without depending on anyone, or we can know if something suspicious is going in our house or place.

These all things can be achieved by state of the art machine learning and deep learning techniques. Our machine learning model tries to solve this problem by providing caption for image. It takes image as input and output text describing it.

It is comparatively challenging task than any image recognition or face-recognition, classification task. We have used several exciting deep learning concepts and tools. We have used tensorflow, CNN for image processing and state of the art Transformers for text processing.

KEYWORDS : Deep learning, CNN, Transformers, Tensorflow

INTRODUCTION

Image captioning is very difficult task because of few reasons. First if in an image suppose more than one events are happening like there is football match going on and There are lots of people gathered so if our model predicts that there is lots of people gathered or football match is going on then both the results are correct but there is currently no way we can tell which context is more important. But research is going on this field. In image captioning image is passed to model and it generates respective descriptions about that image. We have used Convolutional Neural Network (CNN) for extracting images features and then this is passed to transformer to generate captions.

Image caption generation will help in various industry for detecting faulty products, surveillance etc.

LITERATURE REVIEW

Automatically generating captions of an image is a task very close to the heart of scene understanding—one of the primary goals of computer vision. Not only must caption generation models be powerful enough to solve the computer vision challenges of determining which objects are in an image, but they must also be capable of capturing and expressing their relationships in a natural language. For this reason, caption generation has long been viewed as a difficult problem. It is a very important challenge for machine learning algorithms, as it amounts to mimicking the remarkable human ability to compress huge amounts of salient visual information into descriptive language.

In **Show and Tell: A Neural Image Caption Generator[4]** paper author present a generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image.

First, they presented an end-to-end system for the problem. It is a neural net which

is fully trainable using stochastic gradient descent. Second, their model combines state of-art sub-networks for vision and language models. These can be pre-trained on larger corpora and thus can take advantage of additional data. Finally, it yields significantly better performance compared to state-of-the-art approaches;

In **Show Attend and Tell: Neural Image Caption Generation with Visual Attention [3]** paper author used Attention mechanism to improve the performance of model in this task.

Inspired by recent work in machine translation and object detection, we introduce an attention based model that automatically learns to describe the content of images.

They describe how we can train this model in a deterministic manner using standard backpropagation techniques and stochastically by maximizing a variational lower bound. They also show through visualization how the model is able to automatically learn to fix its gaze on salient objects while generating the corresponding words in the output sequence. They validate the use of attention with state-of-the-art performance on three benchmark datasets: Flickr8k, Flickr30k and MS COCO. Automatically generating captions of an image is a task very close to the heart of scene understanding - one of the primary goals of computer vision. Not only must caption generation models be powerful enough to solve the computer vision challenges of determining which objects are in an image, but they must also be capable of capturing and expressing their relationships in a natural language. For this reason, caption generation has long been viewed as a difficult problem. It is a very important challenge for machine learning algorithms, as it amounts to mimicking the remarkable human ability to compress huge amounts of salient visual infomation into descriptive language. Despite the challenging nature of this task, there has been a recent surge of research interest in attacking the image caption generation problem.

Aided by advances in training neural networks (Krizhevsky et al., 2012) and large clas-sification datasets (Russakovsky et al., 2014), recent work has significantly improved the quality of caption generation using a combination of convolutional neural networks (convnets) to obtain vectorial representation of images and recurrent neural networks to decode those representations into natural language sentences (One of the most curious facets of the human visual system is the presence of attention (Rensink, 2000; Corbetta &amp; Shulman, 2002).

Rather than compress an entire image into a static representation, attention allows for salient features to dynamically come to the forefront as needed.

Using representations (such as those from the top layer of a convnet) that distill information in image down to the most salient objects is one effective solution that has been widely adopted in previous work. However working with these features necessitates a powerful mechanism to steer the model to information important to the task at hand. In this paper, they describe approaches to caption generation that attempt to incorporate a form of attention with.

With **Attention is all you need [1]** paper Transformers comes into existent. Since it can be parallelize hence GPU can be used efficiently and it is also more accurate.

PROPOSED WORK

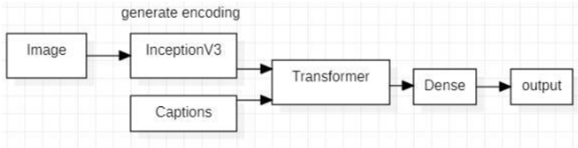


Figure 1: Flow of execution of model

1) Feature Extraction Using CNN
2) Using Transformer for decoding text

Image Captioning are performed using ENCODER-DECODER architecture.

Image Captioning basically comprises two parts first extracting features from images and then processing text along with image to train our machine learning model.

For feature extraction we are using inceptionV3 architecture which is trained on imagenet dataset. We remove last layer (prediction layer) from this architecture.

And store the encoding of every image in numpy format for letter processing with texts.

We are using Transformer architecture for text processing with images. Captions along with their encoded image passed to transformer architecture to train our model.

Since it is difficult to measure accuracy for text generation but there are methods like BLEU score which can help us getting whether predicted and real caption are similar or not.

Previously there were many models which were using RNN-LSTM architecture but problem was that RNN is slow because we have to wait for output of previous input but for transformer we can get benefits of parallelism.

We have previously worked with RNN but were not able to get good results, so we move to transformer architecture for text processing to achieve better

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 05 | May - 2022 Impact Factor: 7.185 ISSN: 2582-3930

results.
There is a slight difference between transformer architecture while training and testing.

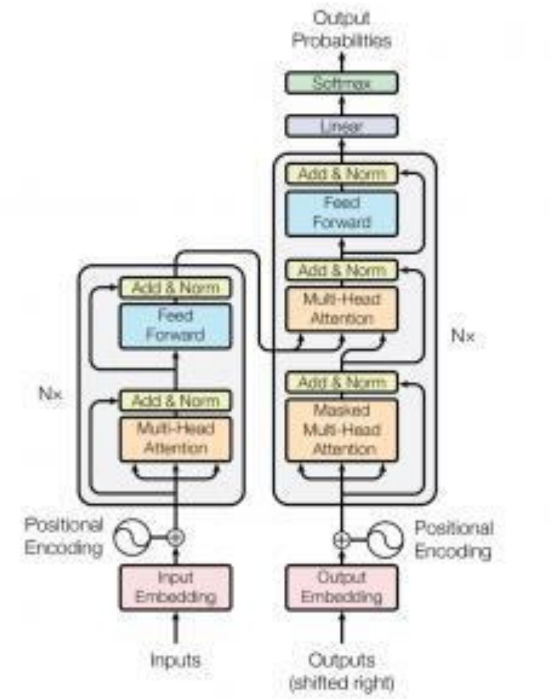
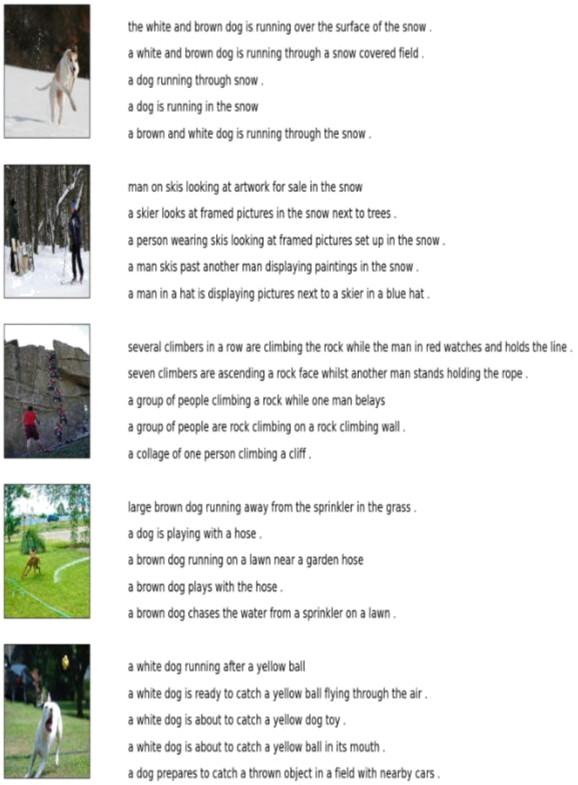


Figure 2: Transformer Architecture

DATASET

We have used flickr 8K dataset which consists of 8 thousands images and each image has 5 captions associated with it.
So there are 40 thousand training data.

We have divided dataset into 32 thousand training set and 8 thousand tests set (20%).



the white and brown dog is running over the surface of the snow .
a white and brown dog is running through a snow covered field .
a dog running through snow .
a dog is running in the snow
a brown and white dog is running through the snow .

man on skis looking at artwork for sale in the snow
a skier looks at framed pictures in the snow next to trees .
a person wearing skis looking at framed pictures set up in the snow .
a man skis past another man displaying paintings in the snow .
a man in a hat is displaying pictures next to a skier in a blue hat .

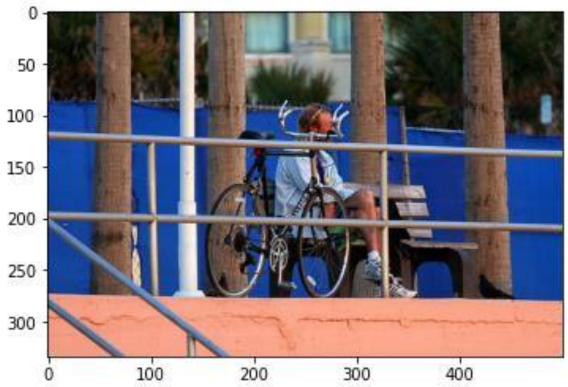
several climbers in a row are climbing the rock while the man in red watches and holds the line .
seven climbers are ascending a rock face whilst another man stands holding the rope .
a group of people climbing a rock while one man belays
a group of people are rock climbing on a rock climbing wall .
a collage of one person climbing a cliff .

large brown dog running away from the sprinkler in the grass .
a dog is playing with a hose .
a brown dog running on a lawn near a garden hose
a brown dog plays with the hose .
a brown dog chases the water from a sprinkler on a lawn .

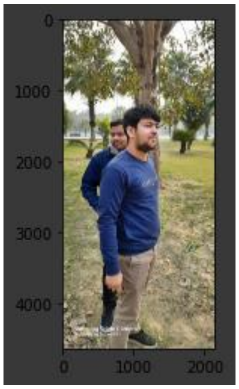
a white dog running after a yellow ball
a white dog is ready to catch a yellow ball flying through the air .
a white dog is about to catch a yellow dog toy .
a white dog is about to catch a yellow ball in its mouth .
a dog prepares to catch a thrown object in a field with nearby cars .

RESULTS

When we shifted our model from rnn-lstmarchitecture to transformer our results improved here are some samples of results.



```
Predicted Caption: man sitting on bench next to
bike parked at the view of bicycle
```



```
Predicted Caption: boy hiding behind tree in front
of tree in front of tree in front of trees
```

Conclusions

The need for AI in today's word is increasing since we want to automate every task which human can do manually. The recent pandemic has taught us that if we have new trained robots like in hospital and shops which can helps perform task without involving other person than it can prevent spread of some disease.

Our model can be used to monitor what patient is doing, or we can install camera with our model in various public places including shops, hospital, and traffic to monitor suspicious activities.

Future Scopes

With more data and training our model can improve performance and can be used in real time.

Our system can be used for blind people to get the event happening around them.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. CoRR, abs/1706.03762, 2017.

[2] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. CoRR, abs/1411.4555, 2014.

[3] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. CoRR, abs/1502.03044, 2015.

[4] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition. 3156–3164.

[5]***Sen He, Wentong Liao, Hamed R. Tavakoli, Michael Yang, Bodo Rosenhahn, Nicolas Pugeault***; Proceedings of the Asian Conference on Computer Vision (ACCV), 2020

[6]***Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, Lei Zhang***; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6077-60