2581-4575
International Journal For Recent Developments in Science & Technology
Crossref
A Peer Reviewed Research Journal

AN APPROACH FOR GENERATING IMAGE CAPTION USING PYTHON 3

[1]S.Fowzia Sultana, [2]B.Ravi Teja, [3]G.Maneesh Yadav, [4]SMD.Ateeq Adnan, [5]N.Sai Achyuth
[1] Assistant Professor [2,3,4,5]B.Tech Scholar,
[1,2,3,4,5]Department Of Electronics And Communications Engineering
[1,2,3,4,5]G. Pullaiah College of Engineering and Technology, Nandikotkur Rd, near Venkayapalle, Pasupula Village, Kurnool, Andhra Pradesh 518002, India.

Abstract:
Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. In this paper, we 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. The model is trained to maximize the likelihood of the target description sentence given the training image. Experiments on several datasets show the accuracy of the model and the fluency of the language it learns solely from image descriptions. Our model is often quite accurate, which we verify both qualitatively and quantitatively. For instance, while the current state-of-the-art BLEU-1 score (the higher the better) on the Pascal dataset is 25, our approach yields 59, to be compared to human performance around 69. We also show BLEU-1 score improvements on Flickr30k, from 56 to 66, and on SBU, from 19 to 28. Lastly, on the newly released COCO dataset, we achieve a BLEU-4 of 27.7, which is the current state-of-the-art.
Key words: IMAGE, CAPTION, CNN, RNN, LSTM, ENCODER, DECODER, IDLE.

1. Introduction

Image Caption Generator is the one which generates a sentence which explains the input image. We must first comprehend the significance of this challenge in real-world scenarios. Let us consider few scenarios in which a solution to this problem could be extremely beneficial. Browsing in search engines with the help of an image. First the caption is generated through Image Caption Generator. Later searching is performed with the help of caption. Searching can be done efficiently. By generating captions visually challenged people can understand the image without others helps. First the caption is generated from image and later the caption can be converted into audio format. By listening the audio, visually challenged person can understand image. CCTV cameras are used for monitoring, but if we can provide useful captions in addition to watching, we can trigger warnings as soon as dangerous conduct is detected. This helps to minimize crime or accidents. Every day, we are exposed to a significant number of images from a variety of sources, including the internet, news articles, schematics in documents, and advertisements. These resources provide visuals that visitors must interpret for themselves. Majority of photos do not contain a description, yet humans can make sense of them without them. However, if people want automated image captions from the machine, the system must be able to understand and interpret them. In this Project, we are going to implement a encoder-decoder architecture in which the encoder is a pre-trained models like VGG16, ResNet50, InceptionV3 and Mobile Net. These will extract the features from the image. Later these are transferred to a special RNN which is LSTM (Long Short Term Memory) which generates caption word by word based on the input features and the caption generated up to that moment. We choose ResNet50 (Residual Networks) because to reduce vanishing gradient problem and LSTM to get captions efficiently.

2. Literature Review

Image caption generator is a popular research area of Artificial Intelligence that deals with image understanding and a language description for that image. Generating well-formed sentences requires both syntactic and semantic understanding of the language. Being able to describe the content of an image using accurately formed sentences is a very challenging task, but it could also have a great

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

impact, by visually impaired people better understanding the content of images. This task is significantly harder in comparison to the image classification or object recognition tasks that have been well researched. The biggest challenge is most definitely being able to create a description that must capture not only the objects contained in an image, but also express how these objects relate to each other.

Existing System:

Md. ZakirHossainet. al(2021) proposed "Text to Image Synthesis for Improved Image Captioning" which explained a Generative Adversarial Network (GAN) based text to image generator to generate synthetic images. In this, attention-based image captioning was used[1]. ChunleiWuet. al(2020) proposed "Hierarchical Attention-Based Fusion for Image Caption With Multi-Grained Rewards" in which a Hierarchical Attention Fusion (HAF) model is presented as a baseline for image caption based on RL, where multi-level feature maps of Resnet are integrated with hierarchical attention. Revaluation network (REN) is exploited for revaluating CIDEr score by assigning different weights for each word according to the importance of each word in a generating caption[2]. SongtaoDinget.al(2019) proposed "Image caption generation with high-level image features" which Introduce the theory of attention in psychology to image captioning and use to filter image features. Combine low-level information with high-level features to detect attention regions of an image.LSTM variant model is not only affected by long-term information, but also by the rules of attention [3]. N. Komal Kumar et. al(2019) proposed "Detection and recognition of objects in image caption generator system" which detected, recognized and generated worthwhile captions for a given image using deep learning. Regional Object Detector (RODe) is used for the detection, recognition and generating captions. The proposed method focuses on deep learning to further improve upon the existing image caption generator system[4]. Philip King horn et. al(2018) proposed "A Region-based Image Caption Generator with Refined Descriptions" a region based deep learning approach to generate the caption. It employs a regional object detector and RNN based attribute prediction. It also embeds encoder decoder based description of sentence[5]. Ali Farhadiet. al(2018) proposed "Every Picture Tells a Story: Generating Sentences from Images" which describe a system that can compute a score linking an image to a sentence. This score can be used to attach a descriptive sentence to a given image, or to obtain images that illustrate a given sentence. The score is obtained by comparing an estimate of meaning obtained from the image to one obtained from the sentence[6]. Aneja, J et. al(2018) proposed "Convolution Image Captioning" in which the LSTM is used to generate the final sentences. The resent has played a major role in identification of the image features. Vanishing gradient problem is removed using skip connections is resnet. Due to LSTM the sentences are generated effectively [7]. Marc Tanti et. al(2017) proposed "What is the role of recurrent neural networks (rnns) in an image caption generator?" which explains how the recurrent neural networks are important. In neural image captioning systems, a recurrent neural network (RNN) is typically viewed as the primary `generation' component. This view suggests that the RNN should only be used to encode linguistic features and that only the final representation should be `merged' with the image features at a later stage[8]. Karim F et. al(2017) proposed "LSTM fully convolution networks for time series classification" we understood about LSTM, a special kind of RNN. The main advantage of using LSTM in language description is more. It contains a special kind of architecture which contains input gate, forget gate, output gate. Because of these gates LSTM is efficient [9]. Sasha Target. al(2016) proposed "Resnet in resnet" which explains about the Residual Networks. Resnet50 is a special kind of CNN which performs very better than CNN. Resnet50 contains skip connections which removes the vanishing gradient problem. Because of this the weights on nodes were changed effectively. The features are recognized better when compared to other CNN [10]. OriolVinyalset. al(2015) proposed "A Neural Image Caption Generator" based on deep recurrent architecture. The features of the image are extracted by the CNN. The LSTM plays a major role in generating a sentence which describes accurately [11]

3. Proposed Method

Hardware Description:

Processor : Intel i5
Ram : 8GB Hard Disk
Space : 50GB

Software Description:

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

Operating System: Windows7/8/10 or Ubuntu
Front-end Design: HTML, CSS
Back-end Design: Python3
Tool: IDLE, Kaggle
Designing UML Diagrams: Rational Rose

IDLE:

IDLE (Integrated Development and Learning Environment) is an integrated development environment for Python, which has been bundled with the default implementation of the language. It is completely written in Python and the Tinker GUI toolkit. Its main features are:

1. Multi-window text editor with syntax highlighting, auto-completion, smart indent and other.
2. Python shell with syntax highlighting.
3. Integrated debugger with stepping, persistent breakpoints, and call stack visibility.

Installation:

Download Python 3.9

To start, go to python.org/downloads and then click on the button to download the latest version of Python

The modules that should be installed are listed below:

1. Tensor Flow

2.Keras

3.NumPy

4.Flask

5. Pandas

6. OpenCV

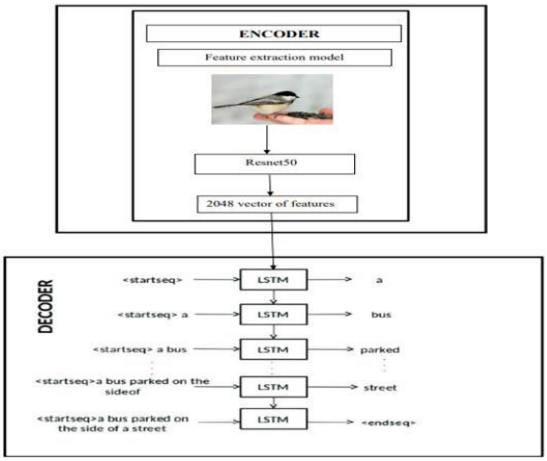

Fig 1: Block Diagram

As shown in the block diagram it is Encoder-Decoder Architecture, where image characteristics are extracted using pre-existing Resnet50 model. Later the sentence is generated word by word by LSTM. This model focuses dynamically on the different parts of the image during the generation of the output sequences.

The following steps are taken by a typical approach for this class:

1.Information on the image is obtained from a CNN on the basis of the entire scene.

2. The word generation phase produces words based on the above step.

3. Captions are updated dynamically until the end state of language generation model.

Working:

An image is sent to our model to generate a sentence which explains the image. To get the features of the image, it is provided as input to the pre-existing Resnet50 model. It provides a vector of 2048 values. These are the features of the image. This vector should be provided as input to the LSTM. Along with this vector a sentence vector which contains initially tag is provided. As LSTM is the time distributed layer, at each iteration the next word is predicted and appended to the pre-existing caption. Later the updated caption along with the image features vector are provided as input to LSTM to get the next word.

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

This process continues until the tag is generated or the maximum length is reached.

UML Diagrams:

Class Diagram:

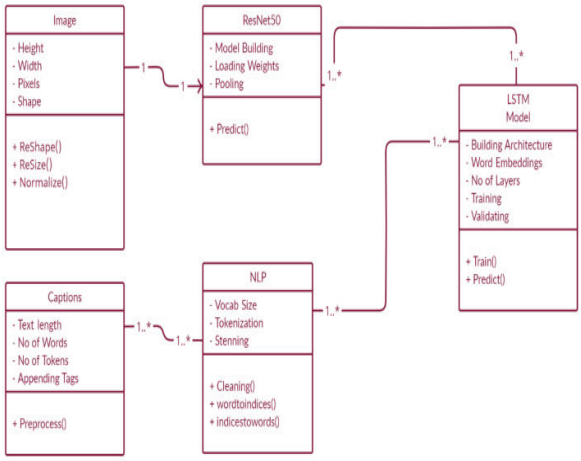

Fig 2: Class Diagram

Sequence Diagram:

A sequence diagram simply shows how items interact in a sequential order. A sequence diagram can also be referred to as an event diagram or an event scenario. Sequence diagrams show how and in what order the components of a system function. Businesspeople and software developers often use these diagrams to document and understand requirements for new and existing systems.

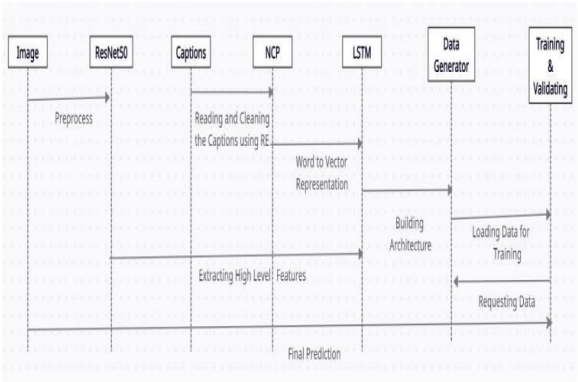

Fig 3: Sequence Diagram

Use Case Diagram:

The purpose of a use case diagram is to capture the dynamic aspect of a system. Use case diagrams are used to gather the requirements of a system including internal and external influences. These requirements are mostly design requirements. As a result, use cases are generated and actors are identified when a system is studied to gather its functionality.

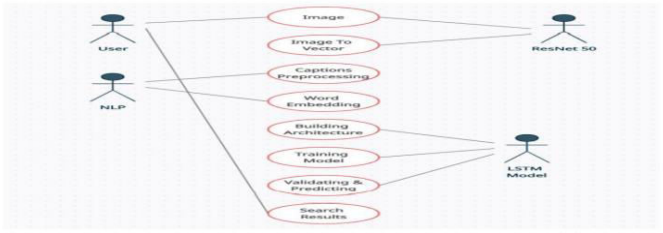

Fig 4: Case Diagram

State Chart Diagram:

State chart diagram itself clarifies the purpose of the diagram and other details. It describes different states of a component in a system. The states are specific to a component/object of a system. A State chart diagram describes a state machine. A state machine is a machine that defines multiple states of an entity and controls these states through external or internal events.

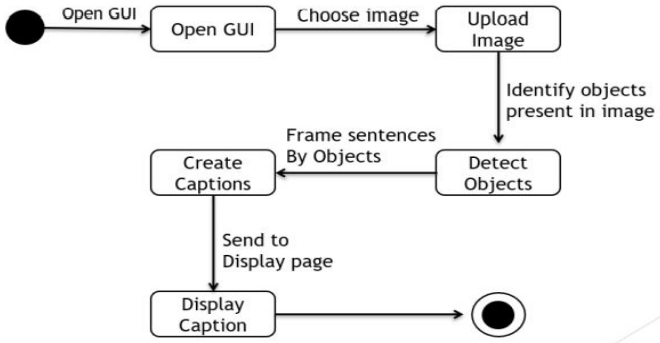

Fig 5: State Chart Diagram

Activity Diagram:

Activity diagram is another important diagram in UML to describe the dynamic aspects of the system. Activity diagram is basically a flowchart to represent the flow from one activity to another activity. The activity can be described as an operation of the system. The control flow is drawn from one operation to another. This flow can be sequential, branched, or

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

concurrent. Activity diagrams deal with all type of flow control by using different elements such as fork, join, etc. The basic purposes of activity diagrams is it captures the dynamic behavior of the system. Activity diagram is used to show message flow from one activity to another.

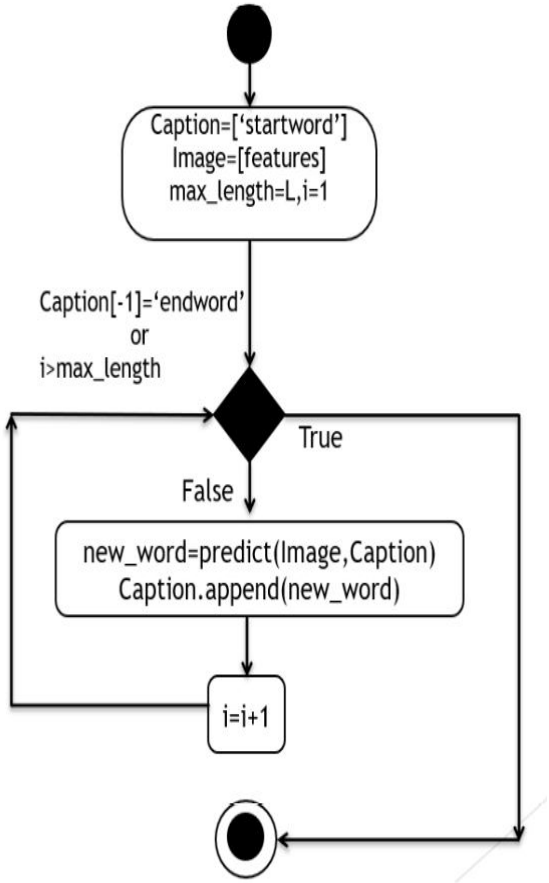

Fig 6: Activity Diagram

Flow Chart:

A flowchart illustrates the individual steps of a process in a sequence order. It is a generic tool that may be used for a wide range of purposes and can be used to describe a number of processes, including manufacturing, administrative and service processes, and project plans.

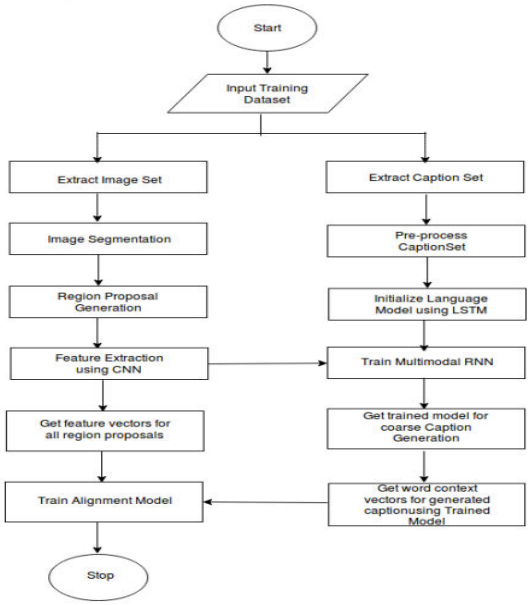

Fig 7: Flow Chart

MODULE DESIGN AND ORGANIZATION:

Encoder:

The encoder is used to encode the given data to extract the insights from it. In our project the encoder is used to extract the high-level features from the images. The encoder we are using here is ResNet50 as shown in the figure 4.6. ResNet50, or Residual Networks, is a well-known neural network that is utilized as the backbone for many computer vision tasks. In 2015, this model was the winner of the ImageNet challenge. The fundamental breakthrough with ResNet was, it allowed us to successfully train extraordinarily deep neural networks with 150+ layers. Due to the problem of vanishing gradients, training very deep neural networks was difficult before ResNet. ResNet is a sophisticated backbone model that is utilized in a wide range of computer vision applications. To add the output from an earlier layer to a later layer, ResNet uses skip connections. This helps in resolving the vanishing gradient issue.

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

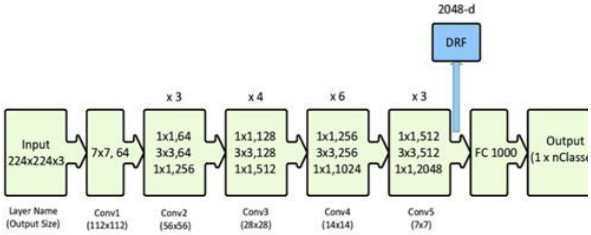

Fig 8: Encoder

The ResNet-50 model is divided into five stages, each with its own convolution and identity block. There are 3 convolution layers in each convolution block as shown in Fig 4.8, and 3 convolution layers in each identity block. There are around 23 million trainable parameters in the ResNet-50. There was a small change made for the ResNet50 and above that previously, shortcut connections skipped two layers, but now they skip three layers, and 1 * 1 convolution layers were added, which we will go over in detail with the ResNet50 Architecture. Using skip-connections or residual connections, you can bypass the training of a few layers. This is what the image above shows. Indeed, if you look closely, we can learn an identity function directly by relying on skip connections. This is the reason why identity shortcut connections are also known as skip connections.

Decoder:

Long Short-Term Memory networks – usually just called "LSTMs" – are a special kind of RNN, capable of learning long-term dependencies. They work tremendously well on a large variety of problems and are now widely used. LSTMs are specifically developed to prevent the problem of long-term dependency. They do not have to work hard to remember knowledge for lengthy periods of time, it's nearly second nature to them. All recurrent neural networks are made up of a series of repeated neural network modules. This repeating module in ordinary RNNs will have a relatively simple structure, such as a single tanh layer.

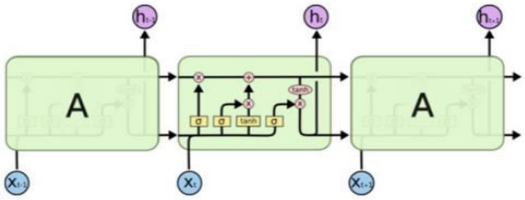

Fig 9: LSTM Architecture

4. Experimental Results

Fig 10: Front End Image

Predicted Caption

a black and white bird eating seeds out of someone 's hand over seeds . .

Predicted Caption

a long-necked bird standing on rocks by river rapids . . .

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

Predicted Caption

two dogs are playing on each other across a grassy lawn leaving some gras

5. Conclusion

As a result, we have successfully created the GUI of our project, which will accept the image from the user and display the caption. This contains two modules that have been fully implemented, Feature Extraction and Caption Generation. The model we developed has a high level of accuracy. Hence from the above proposed method we have presented one single joint model for automatic image captioning based on ResNet50 and LSTM. The proposed model was designed with one encoder-decoder architecture. We adopted ResNet50, a convolutional neural network, as the encoder to encode an image into a compact representation as the graphical features. After that, a language model LSTM was selected as the decoder to generate the description sentence. The whole model is fully trainable by using the stochastic gradient descent that makes the training process easier. The experimental evaluations indicate that the proposed model can generate good captions for images automatically.

6. References

[1]. M. Z. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga and M. Bennamoun, "Text to Image Synthesis for Improved Image Captioning," in IEEE Access, vol. 9, April 2021, pp. 64918-64928.

[2]. C. Wu, S. Yuan, H. Cao, Y. Wei and L. Wang, "Hierarchical Attention-Based Fusion for Image Caption With Multi-Grained Rewards," in IEEE Access, vol. 8, March 2020, pp. 57943-57951.

[3]. Ding, S., Qu, S., Xi, Y., Sangaiah, A. K., & Wan, S. (2020). Image caption generation with high-level image features. Pattern Recognition Letters, 123, 89-95.

[4]. Kumar, N. K., Vigneswari, D., Mohan, A., Laxman, K., & Yuvaraj, J. Detection and recognition of objects in image caption generator system: A deep learning approach. In 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS),March 2019, (pp. 107-109). IEEE.

[5]. Kinghorn, P., Zhang, L., & Shao, L. A region-based image caption generator with refined descriptions. Neurocomputing, Volume 272, 10 January 2018, 272, 416-424.

[6]. Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2018, September). Every picture tells a story: Generating sentences from images. In European conference on computer vision (pp. 15-29). Springer, Berlin, Heidelberg.

[7]. Aneja, J., Deshpande, A., & Schwing, A. G. (2018). Convolutional image captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5561-5570).

[8]. Tanti, M., Gatt, A., & Camilleri, K. P. What is the role of recurrent neural networks (rnns) in an image caption generator?.,Aug 2017 arXiv preprint arXiv:1708.02043.

[9]. Karim, F., Majumdar, S., Darabi, H., & Chen, S.(2017) LSTM fully convolutional networks for time series classification. IEEE Access, 6, 1662-1669.

[10]. Targ, S., Almeida, D., & Lyman, K.(2016) Resnet in resnet: Generalizing residual architectures. arXiv preprint arXiv:1603.08029.

[11]. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition,2015, (pp. 3156-3164).