# IMAGE CAPTION GENERATOR USING DEEP LEARNING

-CATHERINE MARIANA PHILIPS

# IMAGE CAPTION GENERATOR USING DEEP LEARNING

## Table of Contents

## 1) ABSTRACT

In this project, the concept of Convolutional Neural Network is used to generate a caption that describes a particular image. The process of verbally explaining an image's content is known as image captioning. Natural language processing and computer vision come together in this project. This picture captioning project uses an encoder-decoder framework, which entails encoding an input image into an intermediary representation of its contents before decoding it into a descriptive text sequence. This is an application of CNN that was accomplished with the help of the Keras package in Python.

## 2) INTRODUCTION

When talking about the term "image captioning", we generally think it means labelling an image with a short and brief explanation of what is happening in the said image within a sentence. Consider this; a person looks at any sort of image or picture, and the brain has the capacity to comprehend what the picture is about and will be able to explain it well. However, a computer does not have the capacity to easily interpret images. Recent developments in Computer Vision (OpenCV) have made it possible for a computer to comprehend images. In Artificial Intelligence, image caption generator is an up-and-coming application which involves the techniques of deep learning, convolutional neural networks (CNN), and natural language processing (NLP). Image understanding and language description are key parts of this popular application of AI. Image Caption generator is one of the most popular and up and coming project in the field of Deep Learning. In this project, deep learning models namely VGG-16 will be used to generate feature vectors. For the purpose of generating captions for an image, an LSTM-based model is created. To predict the caption, it makes use of feature vectors generated by the VGG network. One of the important motivations for this project is

## 3) LITERATURE REVIEW

With the emergence of the Artificial Neural Network, the discipline of machine learning has undergone a drastic change recently. CNNs are typically employed to tackle challenging image recognition problems and, thanks to their accurate yet straightforward architecture, provide a streamlined way to get started with ANNs.[1] *(O'Shea, Keiron & Nash, Ryan. (2015). An Introduction to Convolutional Neural Networks. ArXiv e-prints.)*

The merging of computer vision and natural language processing has received a lot of interest recently due to the rise of deep learning. This field is represented by picture captioning, which trains a computer to perceive an image's visual information using one or more phrases. The ability to process the state, the properties, and the relationship between these objects is also necessary for the meaningful description generating process of high-level picture semantics. Despite the fact that image captioning is a challenging and intricate endeavour, numerous academics have made substantial advancements. Three deep neural network-based image captioning techniques—CNN-RNN based, CNN-CNN based, and reinforcement-based framework—are explored.[2] *(Liu, Shuang & Bai, Liang & Hu, Yanli & Wang, Haoran. (2018). Image Captioning Based on Deep Neural Networks. MATEC*

*Web of Conferences. 232. 01052. 10.1051/matecconf/201823201052.)*

Thanks to improvements in deep learning methods, the availability of big datasets, and computational power, we can now build models that can produce captions for images. We carefully investigate a deep neural network-based method for creating image captions. The method may result in a sentence that summarises the information contained in the image given as input. We focus on three components of the method: convolutional neural networks (CNN), recurrent neural networks (RNN), and phrase production.[3] *(Tawde, Kanishk. (2022). PYTHON BASED PROJECT TO BUILD IMAGE CAPTION GENERATOR WITH CNN & LSTM.)*

After receiving an image as input, the primary goal of the image caption generator is to produce an English sentence or caption. Getting appropriate and insightful captions for the image is a difficulty with this. A deep learning model receives an image as input, and a caption in English is generated by labelling the input image's components using two optimization strategies, such as beam search and greedy search. The features of the image are learned using a pre-trained CNN model (the VGG16 model), and the caption for the provided image is generated using an LSTM model.[4] *(Yeshasvi, Mogula & Thankaraj, Subetha. (2022). Image Caption Generator Using Machine Learning and Deep Neural Networks.)*

The functionality of the image caption generator is that it involves various concepts of computer vision to find the picture and to translate it to English. The frustrating part in this is to understand the context and produce an English caption which makes sense. The abilities of the deep learning architectures namely VGG16 and ResNe50 are compared. LSTM model is used to generate appropriate captions. BLEU score is used to compare the performance of the two deep learning architectures. An important application of an image caption generator is the explain the image to the visually impaired people.[5] *(Neha, V. & Nikhila, B. & Deepika, K. & Thankaraj, Subetha. (2022). A Comparative Analysis on Image Caption Generator Using Deep Learning Architecture—ResNet and VGG16)*

Picture captioning is the process of creating a description for an image. Recognizing the significant things, their characteristics, and the connections between the objects in an image are necessary. It produces phrases that are both semantically and syntactically sound. In this study, deep learning model that uses computer vision and machine translation to describe images and produce captions are implemented. This study sought to identify the various objects present in an image, understand their associations, and generate captions. The suggested experiment was demonstrated using the Flickr8k dataset, Python3 as the programming language, and a machine learning (ML) method called Transfer Learning with the use of the Xception model. The functions and structures of the various Neural networks involved was also covered in detail. A key component of computer vision and natural language processing is the creation of image captions. They'll be able to effortlessly automate the task of an image interpreter. Not to mention that it has a huge potential for assisting those who are visually challenged. [6] (*Megha J*

*Panicker, Vikas Upadhayay, Gunjan Sethi, Vrinda Mathur(2021), "Image Caption Generator" , Volume-10 Issue-3,*

One of the most robust dynamic classifiers now in use is the Long Short-Term Memory Recurrent Neural Network (LSTM-RNN). To obtain a sense of how the network functions, as well as the associated learning techniques, there is decent documentation available. This research focused on the early, groundbreaking publications to give additional light on how LSTM-RNNs emerged and why they function so admirably. [7] *K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink and J. Schmidhuber,(2017) "LSTM: A Search Space Odyssey,," IEEE Transactions on Neural Networks and Learning Systems, vol. 28*

# 4) METHODOLOGY

## 4.1) PYTHON

One of the most important decisions that has to be taken is what IDE is to be used. Unlike regular software initiatives, AI programmes are unique. The distinctions are in the software platform, the knowledge needed and the requirement for in-depth analysis. Programming language that is reliable, adaptable, and equipped with appropriate tools are to be selected. All of these features are provided by Python, which is why Python was selected for this project. Python aids developers in productivity regarding the software they are creating, from design through deployment and operation[8]. The advantages that make Python the best choice for projects based on machine learning and AI include flexibility, platform freedom, access to excellent libraries and frameworks for ML, and a large community. These increase the language's general appeal. Python is also simple and consistent [8]. The IDE used in this project is Google Colab.

## 4.2) ABOUT THE DATASET

Digital media like images and videos can be uploaded, arranged, and shared via the media platform Flickr. The dataset used here is made up of around 8000 images and is called Flickr 8K dataset [9]. 8,000 photos with five different captions, each describing the key elements and actions in the image, make up this new collection for caption-based image description and search. The pictures were hand-picked to represent a diversity of events and circumstances from six different Flickr groups, and they usually don't feature any famous individuals or places.

## 4.3) KERAS

Google created the high-level Keras deep learning API to implement neural networks. It is used to make the implementation of neural networks simple and is developed in Python. Various backend neural network computation is supported.Because it offers a high level of abstraction python frontend and the choice of many back-ends for computation, Keras is comparatively simple to learn and use[10]

There are built-in functions for many popular artificial intelligence techniques in Keras.

Because Keras has built-in modules for all neural network operations, it can be used to do deep learning quickly. Keras is incorporated in TensorFlow. Tensorflow also enables custom computation including

tensors, giving complete control over the application [10].

Keras is a user-friendly package which has different pre-trained models[21].

One of the cons of keras is that it is difficult to debug errors and log them[21]

## 4.4) VGG16

One of the top computer vision models to date is the CNN (Convolutional Neural Network) model known as VGG16.

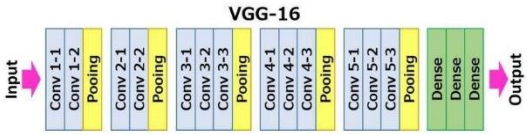


*Figure 1: VGG16 Architecture*

The above image is a VGG16 architecture The 16 in VGG16 stands for 16 weighted layers. Thirteen convolutional layers, five Max Pooling layers, three Dense layers, and a total of 21 layers make up VGG16, but only sixteen of them are weight layers, also known as learnable parameters layers. [11].

Throughout the whole architecture, the convolution and max pool layers are uniformly ordered. There are 64 filters in the Conv-1 Layer, 128 filters in Conv-2, 256 filters in Conv-3, and 512 filters in Conv-4 and Conv-5[11].

The most distinctive feature of VGG16 is that it prioritised convolution layers of a 3x3 filter with stride 1 rather than a large number of hyper-parameters and consistently employed the same padding and maxpool layer of a 2x2 filter with stride 2[11].

One of the challenges faced while training a VGG16 neural network is that it is very slow and takes a lot of time to train.

## 4.5) LSTM

Neural networks of the Long Short-Term Memory (LSTM) type can learn order dependence in sequence prediction issues.

This behaviour is essential for solving complicated problems in areas like speech recognition and language processing, among others.

A challenging area of deep learning is LSTMs. Understanding LSTMs and how concepts like bidirectional and sequence-to-sequence relate to the field can be challenging [12].

The memory unit is the only component of the LSTM architecture. There are four feedforward neural networks in the LSTM unit. There are two layers in each of these neural networks: the input layer and the output layer. Input neurons are linked to all output neurons in each of these neural networks.

The LSTM unit thus is made up of four completely linked layers [13].

The image below shows the hidden layers of LSTM.

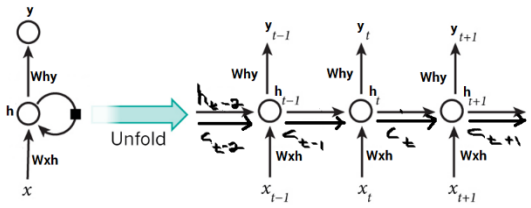


*Figure 2:Hidden Layers of LSTM (Credits: https://www.geeksforgeeks.org/understanding-of-lstm-networks/)*

Each LSTM cell has two outputs (h t and C t) and three inputs ($h_{t-1}$, $C_{t-1}$, and $x_t$). At time t, h t denotes the hidden state, C t, the

cell state or memory, and x t, the current input or data point. Two inputs—h $_{t-1}$ and x t—make up the first sigmoid layer, where h t-1 is the hidden state of the preceding cell. The forget gate gets its name because its output determines how much information from the previous cell will be included. The result is an integer in the range [0, 1] that is multiplied (point-wise) by the cell state from the preceding iteration, $C_{t-1}$[20].

One of the drawbacks of using LSTM in the real world is that, they need a lot of resources and time. Technically speaking, they require a high memory bandwidth because each cell contains linear layers, which the system typically is unable to provide. Thus, LSTMs become relatively inefficient in terms of hardware. This was one of the major drawback faced in this project as training the LSTM model took a minimum of 5 hours[20].

## 4.6) PROCEDURE

This project follows a systematic method as follows.

- Since Google Colab is used, the drive is first mounted.
- All the required packages are then imported.
- From each photo, features are extracted and saved as a pickle file.
- The document that contains the list of captions are then loaded into the memory.
- Each image has 5 descriptions and these descriptions are extracted.
- While using strings, they have to be pre-processed. The captions(descriptions) are tokenized, and converted into lowercase. Then punctuations and hanging words are removed.
- The clean captions are then saved into a separate file.
- Now, the clean captions are loaded into the memory.
- A function is created to load the features extracted from the pictures into the memory.
- Now, the training dataset is loaded and the features of these images are loaded using the function created.
- The clean captions are converted into a list.
- Tokenization is done on the list of captions.
- Input images, sequences, and caption sequences for a picture are built.
- The model which is used for captioning is to be defined.
- The model defining function will consist of a model that extracts the features, and decoder.
- A data generator function that generates random images is defined.
- The model is trained and each epoch are run manually and saved. The model will run for 20 epoch.
- Since this is a CNN model, the model training can take a very long time.
- After training is complete, a function is defined to map numbers to a word.
- A function to generate captions is defined.
- The model is loaded.
- A test image is given as an input and an appropriate caption is generated.

## 5)RESULTS AND DISCUSSION

The model was successfully trained for 20 epochs. A new image which was not present in the dataset was considered as the new test input and fed into the model. The image shown below was used as the test data.

*Figure 3: The new test data*

The output obtained by the model is as follows

```
young boy is wearing red shirt and holding his arm wide closed
```

*Figure 4:Output obtained*

When a normal person sees this image, they would say that it is a young boy with a red t-shirt. The model also gave a very similar output.

Further developments can be made to this project. Other neural networks such as Inception can be used to generate the image caption.

Further the different neural networks can be compared using BLEU score or CIDEr evaluation metric.

The measure BLEU (BiLingual Evaluation Understudy) is used to evaluate machine-translated text automatically. The resemblance of the machine-translated text to a collection of excellent reference texts is evaluated by the BLEU score, which ranges from zero to one[14].

Consensus-based The Image Description Evaluation (CIDEr) attempts to address the issue of the weak association between prior metrics and human judgement by comparing the similarity of a generated sentence to a set of ground-truth sentences written by people. However, CIDEr exhibits greater agreement with consensus as judged by humans[15].

## 6)CONCLUSION

The objective of this study was to find out if neural networks alone might be used to develop a robust image caption generator. The images from the Flickr8K dataset were pre processed before being given as an input to the VGG16 model. The captions were also cleaned (i.e tokenized, made into lower case, punctuations removed) before being used in the LSTM model. A whole new image was given as an input to be tested and an appropriate caption was generated. One of the main drawbacks faced while working on this project was that each time, the model training took a minimum of 5 hours to execute 20 epochs.

Further, different Neural networks can be used to create a caption generator and the accuracy can be compared using either BLEU or CIDEr evaluation metrics.

Future applications that can be built on top of this includes a web app using either Flask or Django to deploy this. The app can have the feature of text-to-speech i.e the generated caption can be read out as an audio. This will be very useful for the visually impaired people as this can help them identify a lot of things in their day-to-day life.

## 7)REFERENCES AND BIBLIOGRAPHY


[1](O'Shea, Keiron & Nash, Ryan. (2015). An Introduction to Convolutional Neural Networks. ArXiv e-prints.)

.[2] (Liu, Shuang & Bai, Liang & Hu, Yanli & Wang, Haoran. (2018). Image Captioning Based on Deep Neural Networks. MATEC Web of Conferences. 232. 01052. 10.1051/matecconf/201823201052.)

.[3] (Tawde, Kanishk. (2022). PYTHON BASED PROJECT TO BUILD IMAGE CAPTION GENERATOR WITH CNN & LSTM.)

[4](Yeshasvi, Mogula & Thankaraj, Subetha. (2022). Image Caption Generator Using Machine Learning and Deep Neural Networks.)

[5] (Neha, V. & Nikhila, B. & Deepika, K. & Thankaraj, Subetha. (2022). A Comparative Analysis on Image Caption Generator Using Deep Learning Architecture—ResNet and VGG16)

[6](Megha J Panicker, Vikas Upadhayay, Gunjan Sethi, Vrinda Mathur(2021), "Image Caption Generator" , Volume-10 Issue-3,

[7] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink and J. Schmidhuber,(2017) "LSTM: A Search Space Odyssey,," IEEE Transactions on Neural Networks and Learning Systems, vol. 28

[8] Angela Beklemysheva¸ "Why Use Python for AI and Machine Learning?", https://steelkiwi.com/blog/python-for-ai-and-machine-learning/

[9] Anon, "Flickr 8K dataset", Adityajn105, Kaggle, https://www.kaggle.com/datasets/adityajn105/flickr8k

[10] Simplilearn, (2022)"What Is Keras: The Best Introductory Guide To Keras", https://www.simplilearn.com/tutorials/deep-learning-tutorial/what-is-keras#:~:text=Keras%20is%20a%20high%2Dlevel,multiple%20backend%20neural%20network%20computation.

[11] Rohini G, (2021), "Everything you need to know about VGG16", https://medium.com/@mygreatlearning/everything-you-need-to-know-about-vgg16-7315defb5918#:~:text=What%20is%20VGG16%20used%20for,to%20use%20with%20transfer%20learning.

[12] Jason Brownlee,2017, "A Gentle Introduction to Long Short-Term Memory Networks by the Experts", https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/

[13] Ottavio Calzone, 2022, "An Intuitive Explanation of LSTM", https://medium.com/@ottaviocalzone/an-intuitive-explanation-of-lstm-a035eb6ab42c

[14] Anon, "Evaluation Metrics", https://cloud.google.com/translate/automl/docs/evaluate#:~:text=BLEU%20(BiLingual%20Evaluation%20Understudy)%20is,of%20high%20quality%20reference%20translations.

[15] Wei Di, Anurag Bhardwaj, Jianing Wei, 2018, "Deep Learning Essentials", Packt Publishing, ISBN: 9781785880360

[16] Chetan Sagathiya, 2021, "Image Caption Generator", https://github.com/Chetan-Sagathiya/Image-Caption-Generator

[17] AHMEDGAMAL12,2022, "Image_Caption_Generator",https://www.

kaggle.com/code/ahmedgamal12/image-caption-generator

[18] DAS KOUSHIK,2022, "ImageCaptioning", https://www.kaggle.com/code/daskoushik/imagecaptioning#Training-Model

[19] Masum Ahmed EeSha, 2021, "Image Caption Generator with CNN & LSTM", https://www.youtube.com/watch?v=yWAhC95n5RM&t=7s

[20] aditianu1998, 2021, “Understanding of LSTM networks”, https://www.geeksforgeeks.org/understanding-of-lstm-networks/

[21] Anon, “Python Keras Advantages and Disadvantages”, https://data-flair.training/blogs/python-keras-advantages-and-limitations/