International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

Image Caption Generator

Parth Kotak
Department of Computer Engineering
Vidyalankar Institute of Technology
Mumbai,India

Prem Kotak
Department of Computer Engineering
Vidyalankar Institute of Technology
Mumbai, India

***Abstract*— Automatically creating the description or caption of an image using any natural language sentences is a very challenging task. It requires both methods from computer vision to understand the content of the image and a language model from the field of natural language processing to turn the understanding of the image into words in the right order. In addition to that we have discussed how this model can be implemented on web and will be accessible for end user as well. Our project aims to implement an Image caption generator that responds to the user to get the captions for a provided image. The ultimate purpose of Image caption generator is to make user's experience better by generating automated captions. We can use this in image indexing, for visually impaired persons, for social media, and several other natural language processing applications. Deep learning methods have demonstrated state-of-the-art results on caption generation problems. What is most impressive about these methods is a single end-to-end model can be defined to predict a caption, given a photo, instead of requiring sophisticated data preparation or a pipeline of specifically designed models. In this an Image caption generator, Basis on our provided image It will generate the caption from our trained model. The basic idea behind this is that users will get automated captions when we use or implement it on social media or on any applications.**

Keywords*— *Caption Generator , Machine Learning, Automated Captions, Convolutional Neural Network(CNN), Long Short Term Memory(LSTM)

I. INTRODUCTION

In the past few years, computer vision in the image processing area has made significant progress, like image classification and object detection. Benefiting from the advances of image classification and object detection, it becomes possible to automatically generate one or more sentences to understand the visual content of an image, which is the problem known as Image Captioning. Generating complete and natural image descriptions automatically has large potential effects, such as titles attached to news images, descriptions associated with medical images, text-based image retrieval, information accessed for blind users, human-robot interaction. These applications in image captioning have important theoretical and practical research value.Image captioning is a more complicated but meaningful task in the age of artificial intelligence. Given a new image, an image captioning algorithm should output a description about this image at a semantic level. In this an Image caption generator, basis on our provided or uploaded image file It will generate the caption from a trained model which is trained using algorithms and on a large dataset. The main idea behind this is that users will get automated captions when we use or implement it on social media or on any applications.

II. LITERATURE SURVEY

Caption generation is a challenging artificial intelligence problem where a textual description must be generated for a given photograph. It requires both methods from computer vision to understand the content of the image and a language model from the field of natural language processing to turn the understanding of the image into words in the right order. Recently, deep learning methods have achieved state-of-the-art results on examples of this problem.

Deep learning methods have demonstrated state-of-the-art results on caption generation problems. What is most impressive about these methods is a single end-to-end model can be defined to predict a caption, given a photo, instead of requiring sophisticated data preparation or a pipeline of specifically designed models.

RNN's have become very powerful. Especially for sequential data modelling. Andrej Karapathy has very nicely explained the use of RNN's in his blog The Unreasonable Effectiveness of Recurrent Neural Networks. There are basically four types of RNN's. Fig 1 demonstrates them.

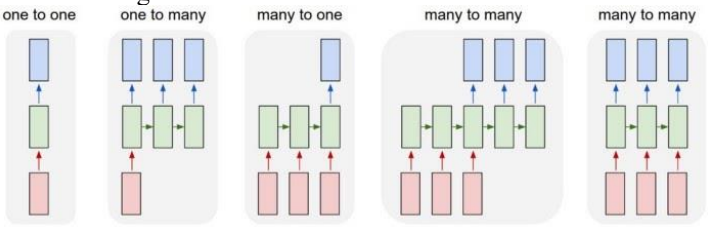


Figure1-Types of RNN's

A. Convolutional Neural Network(CNN)

Convolutional Neural Network (CNN) is a Deep Learning algorithm which takes in an input image and assigns importance (learnable weights and biases) to various aspects/objects in the image, which helps it differentiate one image from the other.

One of the most popular applications of this architecture is image classification. The neural network consists of several convolutional layers mixed with nonlinear and pooling layers. When the image is passed through one convolution layer, the output of the first layer becomes the input for the second layer. This process continues for all subsequent layers.

After a series of convolutional, nonlinear and pooling layers, it is necessary to attach a fully connected layer. This layer takes the output information from convolutional networks. Attaching a fully connected layer to the end of the network results in an N dimensional vector, where N is the number of classes from which the model selects the desired class.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

B. Long short-term Memory

Long Short-Term Memory (LSTM) networks are a type of Recurrent Neural Network (RNN) capable of learning order dependence in sequence prediction problems. This is most used in complex problems like Machine Translation, Speech Recognition, and many more.

The reason behind developing LSTM was, when we go deeper into a neural network if the gradients are very small or zero, then little to no training can take place, leading to poor predictive performance and this problem was encountered when training traditional RNNs. LSTM networks are well-suited for classifying, processing, and making predictions based on time series data since there can be lags of unknown duration between important events in a time series.

LSTM is way more effective and better compared to the traditional RNN as it overcomes the short term memory limitations of the RNN. LSTM can carry out relevant information throughout the processing of inputs and discards non-relevant information with a forget gate.

C. CNN LSTM Architecture

The CNN-LSTM architecture involves using CNN layers for feature extraction on input data combined with LSTMs to support sequence prediction. This model is specifically designed for sequence prediction problems with spatial inputs, like images or videos. They are widely used in Activity Recognition, Image Description, Video Description and many more.

The general architecture of the CNN-LSTM Model is shown in Fig-2:

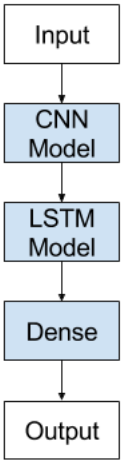


Figure2-General Architecture of CNN- LSTM Model

III. TECHNOLOGY USED

A. Python

Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library.

B. Jupyter Notebook

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

C. Google Colab

Colaboratory, or “Colab” for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education. More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing free access to computing resources including GPUs

D. Python Libraries

1) Pandas
2) Numpy
3) Matplotlib
4) Keras
5) Re
6) Nltk
7) String
8) Json
9) Pickle
10) TensorFlow

E. Flask

Flask is a micro web framework written in Python. It is classified as a micro framework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions.

IV. PROPOSED SYSTEM

The overall workflow can be divided into these main steps:

1. Read Captions File:

Reading the text and token flickr8k file , finding the length of the file and splitting it.

2. Data Cleaning:

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. ... If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct.

3. Loading Training Testing Data:

The process includes training Images File, testing it and creating a train description dictionary that adds starting and ending sequence

4. Data Preprocessing - Images :

Loading the image, preprocessing and encoding it and testing it.

5. Data Preprocessing - Captions :

Loading the captions ,appending the start and the end sequence , finding the maximum length of the caption

6. Data Preparation using Generator:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data and the combining of data sets to enrich data.

7. Word Embedding:

Converting words into Vectors(Embedding Layer Output)

8. Model Architecture:

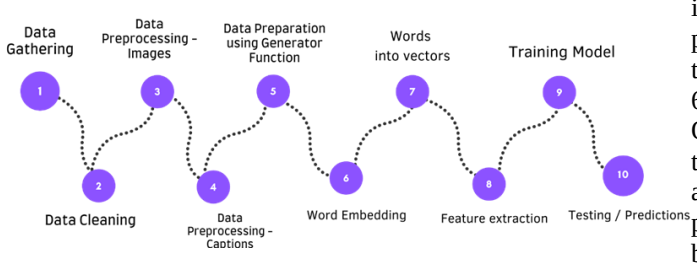
Making an image feature extractor model, partial caption sequence model and merging the two networks

9. Train Our Model:

A training model is a dataset that is used to train an ML algorithm. It consists of the sample output data and the corresponding sets of input data that have an influence on the output.

10. Predictions:

Prediction refers to the output of an algorithm after it has been trained on a historical dataset and applied to new data when forecasting the likelihood of a particular outcome here predicting Caption for a photo.



V. METHODOLOGY

Our project uses the Word to vector, Word embedding, SoftMax and ReLU Activation Functions for Training our model.

The System uses the following methodology:

1. Read Captions File
2. Data Cleaning
3. Loading Training Testing Data
4. Data Preprocessing - Images
5. Data Preprocessing - Captions
6. Data Preparation using Generator Function
7. Word Embedding
8. Model Architecture
9. Train Our Model
10. Predictions

VI. ANALYSIS

Iterative Development Process Model

Iterative development model aims to develop a system through building small portions of all the features, across all components. We build a system which helps to analyse ratings of customers on basis of which it recommends movies to the user.

The phases of iterative development are:

1. Planning:

As with most any development project, the first step is go through an initial planning stage to map out the specification documents, establish software or hardware requirements, and generally prepare for the upcoming stages of the cycle.

2. Requirements:

In this phase, requirements are gathered from customers and check by an analyst whether requirements will fulfil or not. Analyst checks that need will achieve within budget or not. After all of this, the software team skips to the next phase.

3. Design:

Once planning is complete, an analysis is performed to nail down the appropriate business logic, database models, and the like that will be required at this stage in the project .In the design phase, team design the software by the different diagrams like Data Flow diagram, activity diagram, class diagram, state transition diagram, etc.

4. Implementation:

With the planning and analysis out of the way, the actual implementation and coding process can now begin. All planning, specification, and design docs up to this point are coded and implemented into this initial iteration of the project.

5. Verification:

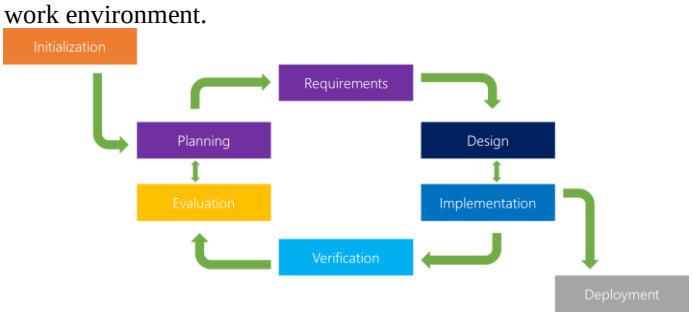
Once this current build iteration has been coded and implemented, the next step is to go through a series of testing procedures to identify and locate any potential bugs or issues that have cropped up.

6. Evaluation:

Once all prior stages have been completed, it is time for a thorough evaluation of development up to this stage. This allows the entire team, as well as clients or other outside parties, to examine where the project is at, where it needs to be, what can or should change, and so on.

7. Deployment:

After completing all the phases, software is deployed to its work environment.



VII. FEASIBILITY STUDY

Technology Considerations image captioning generation systems available in the market are dependent on the dataset to contain large of clusters of similar users and items.

- Product/ Service Market place Image Captioning system will impact client institutions in several ways. The following provides a high level explanation of how the organization, tools, processes, and roles and responsibilities will be affected as a result of the image captioning generation system
- Tools: The existing requirement for on site management systems will be eliminated completely with the availability of a cloud-based system.
- Processes: With the image captioning system comes more efficient and streamlined administrative processes.
- Hardware/Software: Clients can use this in image indexing, for visually impaired persons, for social media, and several other natural language processing applications.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

- Operational Feasibility: The project will be implemented in a way that it will allow the functioning of image captioning generation smoothly. It will provide a user-friendly user interface in a modular fashion.

VIII. COST ANALYSIS

Infrastructure services

These services include infrastructural components such as where the model is hosted, where data is stored and how the data is delivered. All these also need redundancies and load balancers for backup and security servers, which add both the cost and complexities.

Servers - Servers are where the app will be hosted. Some of the popular companies that provide hosting services are amazon aws, google gcp and azure.

We can use this to implement it globally over the internet. It has some the cost for training ML

GCP: $0.54 per hour

Azure: $9.99 per ML studio workspace per month $1 per studio experimentation hour

AWS: $0.42 per hour

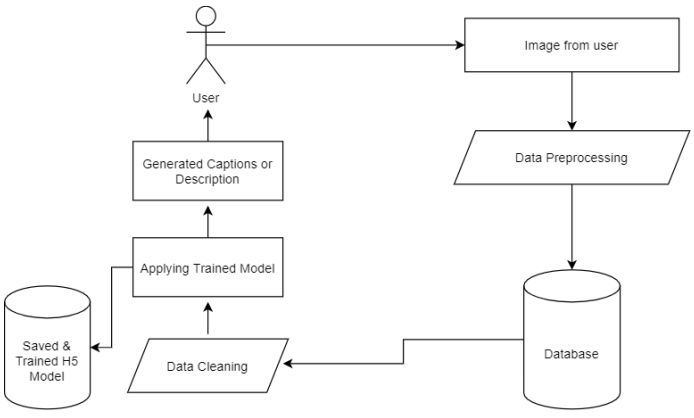
If we need to increase the number of GPUs that your model supports, we will have to pay $3.06 per Hour in AWS and GCP Charges hourly with different cost depends on the GPUs.

ML model training – For the training of the AI - agent we will be using our local system with the following system specifications:

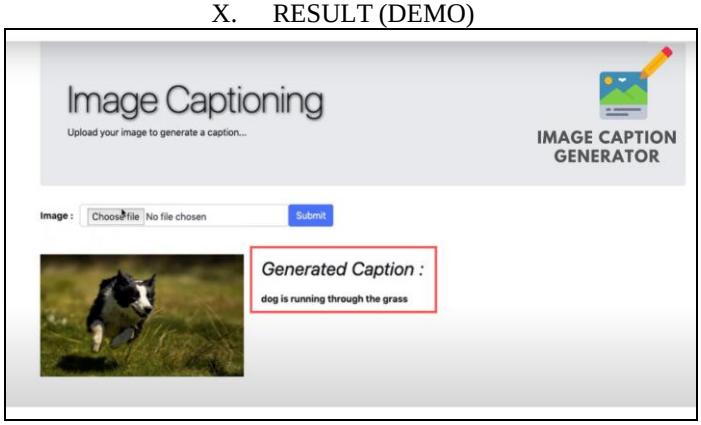
Component	Model	Cost
CPU	i7 8750h	₹ 28,991
GPU	GeForce gtx 1050 TI	₹ 26,850
RAM	16GB RAM	₹ 8,000

We are using the Jupyter Notebook and Google Colab for testing and checking the prediction of our model

IX. DESIGN



X. RESULT (DEMO)



XI. CONCLUSION

In this paper we have learned and designed a technique of Image Caption Generator which will respond to User with captions or description based on an image. The Image Based Model extracts features of an image and the Language based model translates the features and objects extracted by image based model to a natural sentence. Image based model uses CNN whereas Language Based model used LSTM .The workflow is Data gathering followed by Pre-processing, Training model and Prediction. The ultimate purpose of an Image caption generator is to improve the social media platforms as well as in image indexing and for visually impaired persons with automated generated captions or description

XII. REFERENCES

[1] https://towardsdatascience.com/a-guide-to-image-captioning-e9fd5517f350
[2] https://ieeexplore.ieee.org/document/8276124
[3] https://machinelearningmastery.com/develop-a-deep-learning-caption-generation-model-in-python/
[4] https://medium.com/@raman.shinde15/image-captioning-with-flickr8k-dataset-bleu-4bcba0b52926
[5] https://blog.clairvoyantsoft.com/image-caption-generator-535b8e9a66ac
[6] https://www.matec-conferences.org/articles/matecconf/abs/2018/91/matecconf_eitce2018_01052/matecconf_eitce2018_01052.html
[7] https://www.analyticsvidhya.com/blog/2018/04/solving-an-image-captioning-task-using-deep-learning/