GENBA SOPANRAO MOZE COLLEGE OF ENGINEERING
BALEWADI, PUNE-411045.

DEPARTMENT OF INFORMATION TECHNOLOGY

A 1st PHASE PROJECT REPORT ON
"Automatic Image Captioning"

Submitted in the partial fulfillment of the requirement for BE in Information Technology in 2022-23

Submitted by

Name	PRN:
Anand Dhakane	72024841J
Nihal Kazi	72024871L
Omkar Mangnale	72024886J
Dattatray Jadhav	72024856G

UNDER THE GUIDANCE OF

Prof. Priyanka Mane

GENBA SOPANRAO MOZE COLLEGE OF ENGINEERING BALEWADI, PUNE-411045.

DEPARTMENT OF INFORMATION TECHNOLOGY ENGINEERING

CERTIFICATE

This is to certify that the Project entitled

“Automatic Image Captioning"

Submitted by

Name	PRN
Anand Dhakane	72024841J
Nihal Kazi	72024871L
Omkar Mangnale	72024886J
Dattatray Jadhav	72024856G

It is a beneficial work carried out by them under the guidance of Priyanka Mane and is approved for the partial fulfillment of the requirement of SPPU for the award of BE in IT Engineering.

Date:…./…../……

Prof Priyanka Mane	**Prof. Sana Shaikh**	**Dr. Ratnarajkumar Jambi**
(Project Guide)	**(H.O.D)**	**(Principal)**

ACKNOWLEDGMENT

This is a great pleasure & immense satisfaction to express our deepest sense of Gratitude & thanks to everyone who have directly or indirectly helped in Completing Project Stage 1 successfully. It gives us great pleasure in presenting the project report on:

“**Automatic Image Captioning ”**

We would like to take this opportunity to thank our guide **Prof. Priyanka Mane** for giving all the help and guidance that we needed. We are really grateful to them for her kind support. Her valuable suggestions were very helpful.We would also like to express our gratitude towards our parents for their kind cooperation and encouragement which helped us in completion of this project and We also appreciate the guidance given by our honorable Principal **Dr.Ratnarajkumar Jambi** sir & our beloved HOD, **Prof.Sana Shaikh**.

Last but not the least, many thanks to the Project Co-ordinator, **Prof. Kaveri Kari** who has invested her full effort in guiding the team for achieving the goal. We are also grateful to her dispensable support and suggestions.

Regards,

CONTENTS

CHAPTER NO.	TITLE	PAGE. NO.
1. 1.1 1.2 1.3	Introduction Background Information Project Objective Report Layout	6
2.	Literature Survey	8
3. 3.1	Methodology System Architecture	10
4 4.1 4.2 4.3 4.4 4.5 4.6 4.7 4.8 4.9	Proposed System Architecture Convolutional Neural Network ReLU (Rectified Linear Unit) Pooling Layer Recurrent Neural Network LSTM (Long Short Term Memory) Hardware Requirements Software Requirements Dataset Sources	15
5 5.1 5.2 5.3 5.4	Implementation Detail Languages and Libraries Setup Used Model Training Implementation Web Site Implementation	20
6	Expected Result and Analysis	25
7	Conclusion and future scope	28

INTRODUCTION

CHAPTER 1

INTRODUCTION

1.1 Background Information

Image Captioning is the system of producing a textual description for given images.It has been an extremely important and basic endeavor in the Deep Learning space. Picture subtitling has a major amount of use. NVIDIA is the usage of picture captioning applied sciences to create a software to assist human beings who have low or no eyesight.

Picture inscribing can be considered as a start to finish Sequence to Sequence issue, as it changes over pictures, which is considered as a grouping of pixels to a succession of words. For this reason, we need to methodology each language or explanations and the pictures. For the Language part, we utilize intermittent Neural Networks and for the Image part, we use Convolutional Neural Networks to individually accomplish the capacity vectors.

Before moving to further chapters let's understand about digital imagers and their advantages.An image is a visual representation of an object. It can be anything from paintings, sculptures, photos etc. The images are in existence for a very long time now. As computers cannot understand images, it became necessary to develop special methods to represent images in computers. These days, they are addressed as a succession of 0s and 1s in PC.

1.2 Motivation

As per WHO, nearly a billion people have some disability and some 280 million have visual impairments. Many of them use some devices like screen readers which help them to understand digital text.

As of 2022, about 60% of the world has internet access. This increased internet usage has disrupted the market as many content creators are shifting to internet and web. This poses a challenge to provide a similar level of service to visually impaired people and also a challenge to maintain large amounts of images and their text.

1.3 Project Objective

We are going to develop a web application that helps us in identifying the action of an image. We aim at creating a Neural Network Model to analyze the images and create captions using transformer models.

1.4 Report Layout

Literature Review describes all the previous works done in this field.

Methodology describes the architecture of the project.

Implementation Details describes the tools that are used and the process that needs to be followed in each mode.

Results show the final outputs that are done in the course of this project.

End indicates the impediments and future extent of this undertaking.

LITERATURE SURVEY

CHAPTER 2

LITERATURE SURVEY

One of the most striking notices is the ImageNet project, where they publicly supported a huge number of named pictures and prepared models for the last ten years to perceive objects in the picture. Beginning around 2010, the yearly ImageNet Large Scale Visual Recognition Challenge (ILSVRC) holds a contention consistently, to vie for most elevated precision on different visual acknowledgment undertakings. Presently the profound CNN[2] networks have more exactness than people in acknowledgment. Anyway Captioning pictures could be a lot of testing task, since it includes object acknowledgment and tracking down connections among them. This has been unthinkable as of not long ago, attributable to gigantic improvement in computational power[13]. Despite the fact that there are different scientists taking care of a similar issue, there are two groups that stood apart with their calculations. One from Google, and the other from Stanford University. Google delivered a paper "Sharing time: A Neural Image Caption Generator" in 2014 [6]. Their model is prepared to expand the probability of the objective portrayal sentence, given the picture. The model is prepared on different datasets like Flickr30K, SBU, MSCOCO and has accomplished human level execution in creating subtitles. At the point when Google originally delivered a paper in 2014, the framework utilized the "Commencement V1 " picture characterization model which accomplished 89.6% exactness. The most recent delivery in 2016 utilized the "Commencement V3 " model, which accomplishes 93.9% precision[3]. Before Google, picture subtitling was conceivable utilizing the DistBelief software system. Then Google delivered TensorFlow execution, which utilizes GPU power and contrasted with before executions, the preparation time is decreased by a variable of 4. The other group that accomplished well in taking care of the issue is from Stanford University - Fei Li and Andrej Karpathy. Their paper "Profound Visual-Semantic Alignments for Generating Image Descriptions" which was delivered in 2015 [7], uses pictures and depictions to find out about multi-modular correspondences among language and visual information. They've utilized RNN and CNN to accomplish the errand. Their execution is clustered. It utilizes Torch library, which runs on GPU and upholds CNN finetuning, which sped up by immense component

METHODOLOGY

CHAPTER 3

METHODOLOGY

The methodology is a relevant structure for research. We have multiple sections that cover the Architecture, Working, and Tools Used in the project.

3.1 System Architecture

The accompany addresses the system architecture and the essential working of Web Application of Image caption system.

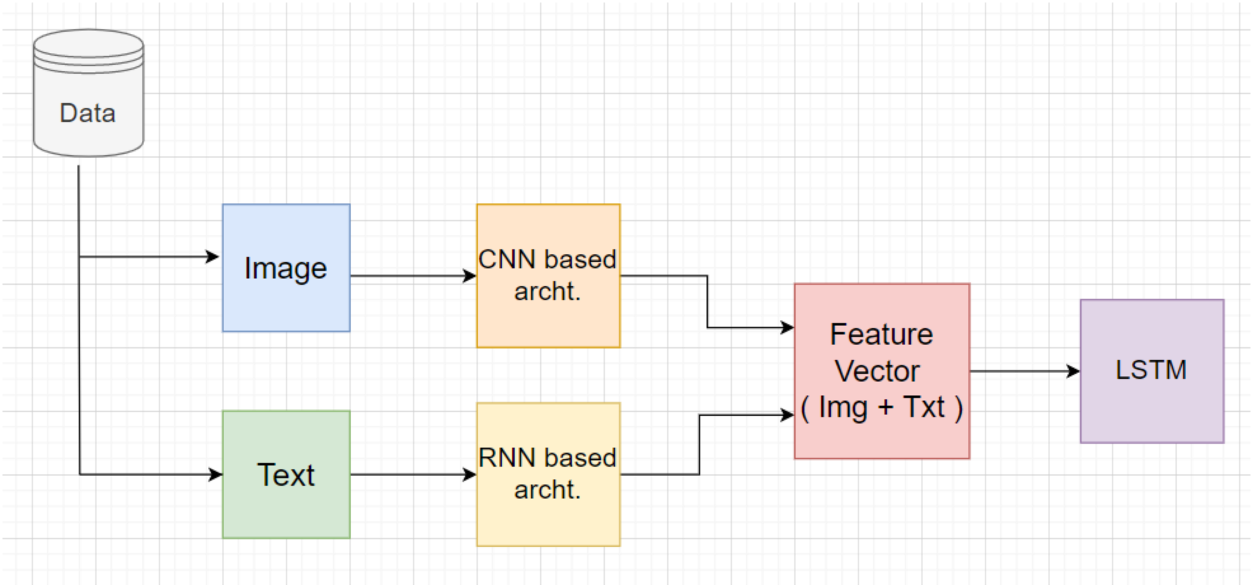
The framework is intended to give subtitles of a picture.

1. **RNN**: Recurrent Neural Networks (RNNs) are perhaps the most pervasive engineering in light of the capacity to deal with variable-length texts. They are networks with different circles in them, permitting data to proceed.

2. **LSTM**: LSTMs are unequivocally intended to avoid the drawn out reliance issue. Recalling records for an extended time frame is practically their default conduct, at this point not something they battle to learn. Long non super durable memory (LSTM) is a counterfeit repetitive brain local area (RNN) structure utilized in the circle of profound learning. Not at all like chic feedforward brain organizations, LSTM has remarks associations. LSTM networks are appropriate to ordering, handling, and making forecasts upheld measurement measurements on account that there could likewise be slacks of obscure period between essential events in a very measurement.

3. **Transformers:** The figure given below depicts transformers and which is also called a sequence-to-sequence architecture. Sequence-to-Sequence architecture[4] is a neural network which changes a specified succession of components, like grouping words in a sentence, into another grouping. These models are admissible

for interpretation, in which the grouping in words from one language is changed to a series of different words in some other dialect.



Above figure proposes the System Architecture of our task that arrangements with the text and pictures.

The walkthrough of the Architecture is as follows:

- The Text Data is cleaned and pre-processed. This Bag Of Words or Embedding Matrix is created to passed through RNN architecture which converts text into its feature vector.

- The image data is pre-processed and size is adjusted to 248*248 to be input into Resnet50 model. The Resnet50 model outputs a feature vector describing the image mathematically.

- Feature vectors created in above steps are concatenated and passed through LSTM model, which train on data and for a particular image learns suitable text.

PROPOSED SYSTEM

CHAPTER 4

PROPOSED SYSTEM

The system developed is a caption generation system. It contains components of pre-processing, caption generation and a caption decoder. During training and testing phase, images are read from the disk. This image is converted from JPEG/PNG to array representation. The image is resized to an array of shap 224X224X3. The image pre-processor normalizes the image to array of shape 224X224. The image is passed through a Renet50 model, which generates a compressed gist of information contained within the image.

For training phase only, text inputs are preprocessed as well. The text is changed to lower case and each sentence is appended with ‘startofseq’ and ‘endofseq’. After this each word is broken down into words and frequency of each word is stored temporarily in form of dictionary. A new dictionary with an indexing of every word is generated which is used to replace words in a sentence with their index. Then a mapping of every image to its index-positioned sentences is maintained.

Vectors obtained for image (in testing) or image and text in training is passed through an architecture of LSTM which generates text caption for the image.

4.1 Architecture

Engineering of the model basically include 2 kinds of brain organizations, RNN and CNN. RNN is Recurrent brain Network and CNN is Convolutional Neural Network. Insights concerning them are examined as follows:

4.2 Convolutional Neural Network:

Neural Networks are utilized for Image Recognition. However, the problem with simple Neural Networks is, in the event that the photo is of huge pixels, the no.of.parameters for a

Neural people group increments. This makes Neural organizations progressive and consumes a great deal of computational power.

To overcome this problem, CNN[8] are used. The convolutional neural community is a distinctive kind of feed ahead neural network. Convolutional neural networks ingest and procedure pix as tensors, and tensors are matrices of numbers with extra dimensions.

There are three primary kinds of layers to fabricate CNN structures:

Convolutional layer

Pooling layer

Fully-connected layer

The completely associated layer is very much like the standard brain organizations. The convolutional layer can be considered as playing out the convolution activity commonly on the past layer. The pooling layer can be however as downsampling by the limit of each block of the past layer. We stack these three layers to develop the full CNN engineering.

Convolutional Neural Networks have 2 main components.

Feature learning: We have convolution, ReLU,Pooling layer stages here. Edges,shades,lines,curves, in this Feature learning step are get extricated.

Classification: There is Fully Connected(FC) layer[9] in this stage. They will relegate a likelihood for the item on the picture being what the calculation predicts it is.

For our use-case, we are only interested in Feature learning component of CNN.

4.3 ReLU(Rectified Linear Unit):

An additional an activity known as ReLU[13] has been utilized after every single Convolution activity. Relu is a non-direct activity. ReLU is an issue savvy activity (applied per pixel) and replaces all awful pixel values in the limit map by using zero. The justification for ReLU is to introduce non-linearity in our ConvNet, when you consider that the greater part of this present reality measurements we would favor our ConvNet to look at would be non-straight.

4.4 Pooling layer:

In this segment the dimensionality of convlayer or trademark map gets diminished protecting the imperative data. from time to time this spatial pooling is furthermore known as Downsampling or subsampling. this pooling layers could likewise be Max pooling, Avg pooling, total pooling[14]. frequently we see Max pooling is utilized most.

4.5 Recurrent Neural Network:

RNN[1] represents Recurrent Neural Network. It is a sort of brain local area which conveys memory and good OK for consecutive information. RNN is utilized by utilizing Apples Siri and Googles Voice Search. We should discuss a few essential norms of RNN.

It is a speculation of feed-forward brain local area that has an inside memory. RNN repetitive in nature as it plays out the equivalent component for every single enter of data while the result of the present day enter depends upon on the past one calculation. In the wake of delivering the result, it is duplicated and despatched lower once more into the repetitive organization[5]. For going with a choice, it considers the current day enter and the result that it has found from the previous info.

The contrast among RNN[10] and feed forward brain network is that RNN can utilize inward memory to handle arrangement of data sources. It is reasonable for successive information where result of one info relies upon past conditions of info and result.

RNN has ability to retain past information. While chipping away at current information, it additionally thinks about what it has gained from past conditions of information and result. In this way, it ascertains its present status utilizing set of current info and the past state. Along these lines, the data burns through a circle.

4.6 LSTM (Long Short Term Memory):

LSTM[1,5] is a RNN structure that take note values over arbitrary intervals. It is used to classify, procedure and predict time sequence given time lags of unknown duration. Relative insensitivity to hole size offers an gain to LSTM over choice RNNs, hidden Markov fashions and different sequence gaining knowledge of methods.

The drawn out memory is regularly known as the telephone state. Each telephone has a recursive nature which allows in measurements from going before stretches to be put away in the LSTM cell[10]. Cell country is adjusted by utilizing the disregard entryway situated under the telephone realm and moreover alter via the enter regulation door.

The consider vector is by and large known as the disregard entryway. The result of the disregard entryway advises the cellphone realm which measurements to ignore through duplicating zero to a job in the framework. Assuming the result of the disregard door is 1, the measurements is saved in the mobilephone state.

The store vector is by and large alluded to as the enter entryway. These entryways conclude which information need to enter the versatile country/long haul memory. The important parts are the enactment highlights for each door. The enter entryway is a sigmoid trademark and have a shift of [0,1].

The focal point of consideration vector is regularly alluded to as the result entryway. The working reminiscence is normally known as the hidden state.

4.7 Hardware Requirements

16 GB RAM

NVIDIA GPU

M1 core

50 GB physical storage

Software Requirements

MACOS (X)

Python3

CUDA 9

Tensorflow

Keras

Numpy

Matplotlib

4.8 Dataset Sources

We picked our dataset from Kaggle. It is a web-based local area foundation of information researcher and AI lovers. It permits clients to team up on projects, find publicly released datasets, access GPU note pads and so forth.

Our dataset is Flickr images dataset which has about 8000 images. Each image is described in about 5 captions, which give different perspectives to an image. This is a standard dataset and is used by many researchers and developers to develop image captioning model.

IMPLEMENTATION DETAIL

CHAPTER 5

IMPLEMENTATION DETAIL

5.1 Languages And Libraries

- OpenCv[11]: OpenCv is a ML package library and associates ASCII text file laptop vision. It’s a library of programming functions chiefly geared toward period laptop vision. We have utilized this for the most part in pre-handling pictures.

- HTML: HTML or HyperText Markup Language is a markup language that permits web clients to make and design different pieces of a page like headers, tables and links using elements, tags, and attributes. It tends to be helped by innovations like Cascading Style Sheets (CSS) and prearranging dialects like JavaScript[14]. This would be based on the language for my website that I would create to deploy my three models of Sentiment Analysis.

- NumPy[9]: It is a open source Python library. NumPy works with Python objects called multi-dimensional arrays. Arrays are basically collections of values, and they have one or more dimensions. NumPy array data structure is also called *ndarray*, short for n-dimensional array. Datasets are usually built as matrices and it is much easier to open those with NumPy instead of working with lists. Numpy here is used for many processes. This library does all the numerical calculation in my undertaking.

- Tensorflow[2]: Tensorflow is an open source framework. It was initially designed to be a neural network library but with advancement it can perform much more functions. It is a machine learning library. It is the base library that we have used to create our model; this is the cover of Keras that helped in model designing and fitting the values.

5.2 Setup Used

- Flask[16]: It is a little net design written in Python. It's named a microframework. As a result of it doesn't need explicit tools or libraries. It's no information abstraction layer, type validation, or the other parts wherever pre-existing third-party libraries give standard functions. It has been used to create an interface between the website and models and also is responsible for returning the HTML pages accordingly to the output

- Kaggle Notebooks[17]: Kaggle Notebbok is a free Jupyter notebook climate that runs altogether in the cloud. In particular, it doesn't need an arrangement, and the notebooks that you make can be at the same time altered by your colleagues - how you vary reports in Google Docs. The free GPU of kaggle is used in this project for the training purpose of the Video and Audio modes for providing fast results.

- VS Code[19]: VS Code is a lightweight text editor, one of the best for coding in all most all languages. It provides you to code in any programming language for example Python, Java, C++, JavaScript, and more. Visual Studio Code is a source code editor, which helps businesses build and debug web applications running on Windows, Linux, and macOS. It is a source-*code* editor text editor program de

5.3 Model Training Implementation

The steps that we went through to train our model:

1. As our dataset is divided into 2 files of images and their corresponding text, we first worked on the image dataset.
2. We read pictures and went through the picture pre-handling step.
3. In picture pre-handling, pictures are resized to suitable elements of 224x224 and reshaped to 224x224x3. The third aspect addresses the RGB part.

4. Once pre-handled, picture is gone through Renet50 model to create an element vector of length 2048.
5. These vectors are stored against their image.
6. Next we pre-process our text dataset.
7. In this, we first read captions for an image which have been pre-processed to generate feature vectors.
8. Once read, captions are first converted to lowercase.
9. Then strings like ‘startofseq’ and ‘endofseq’ are added in start and end of captions respectively.
10. ‘start seq’ is added so that we can have a starting point when we use the deployed model.
11. ‘end of seq’ is added to show that prediction is over.
12. These captions are then broken into words and frequency of these words is counted.
13. These words/tokens are then indexed and a dictionary is created.
14. The tokens are replaced by their indices in captions.
15. Image feature vectors and the tokenized captions are then mapped together.
16. Next we create architecture for our model, which is divided into 2 parts, image-model and language-model.
17. These models are concatenated with each other and a Dense LSTM layer with softmax activation layer is appended as an output layer to the above architecture.
18. In this we use categorical_crossentropy to quantify misfortune, RMSprop as enhancer and exactness as metric.
19. We then passed our pre-processed data to our architecture and trained the model.

5.4 Web Site Implementation

The steps that we went through to develop the website:

1. We have used flask to develop our website as it is lightweigh.
2. On hitting our endpoint ‘http:127.0.0.1:5000/’, user getsredirected to a home page.
3. There user is asked to upload any image they want to see caption for.

4. Once uploaded, user hits predict button and a caption is generated for their image.
5. Under the hood, once predict button is pressed, image is passed as a parameter head of a user request to our server (localhost).
6. We capture the user image and pre-process it.
7. Image pre-processing happens in a similar manner as it happened during model training.
8. Once preprocessed, the image is passed through similar model architecture as that used in mode training.
9. The only difference in input is in textual data. During model training we passed entire captions, with 'startofseq' and 'endofseq' as prefix and suffix respectively. During user testing, we only pass 'startofseq' as text input.
10. 'startofseq' is used to create a sequence of textual predictions.
11. Once entire caption is predicted, it is returned to the user.
12. In model used to predict, we used trained model_vocabulary and model_weights in our architecture.

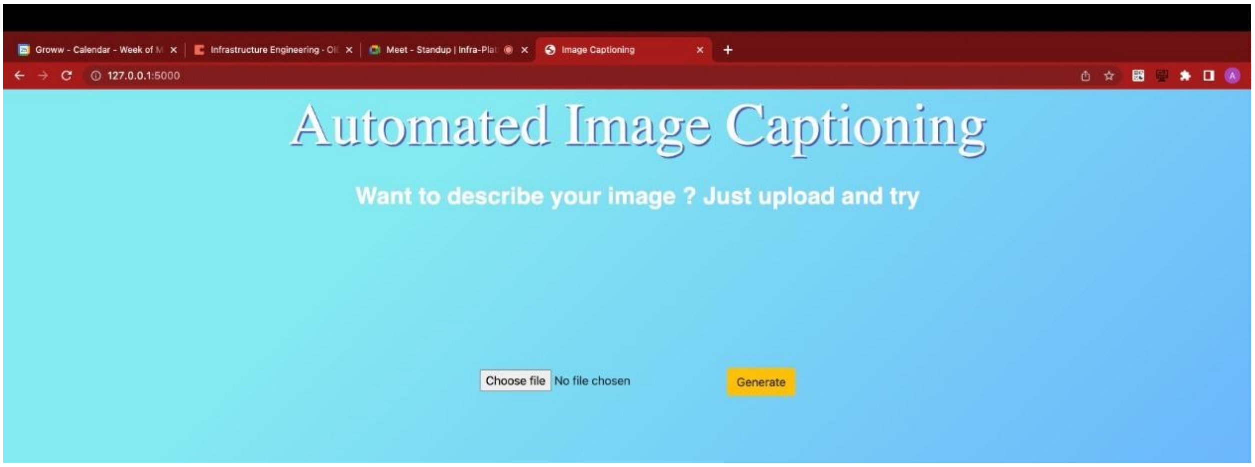
RESULT AND ANALYSIS

CHAPTER 6

EXPECTED RESULT AND ANALYSIS

The Web-App is deployed the models on a local server and ran it using Flask , results of models is shown in below :-

FLow of Web

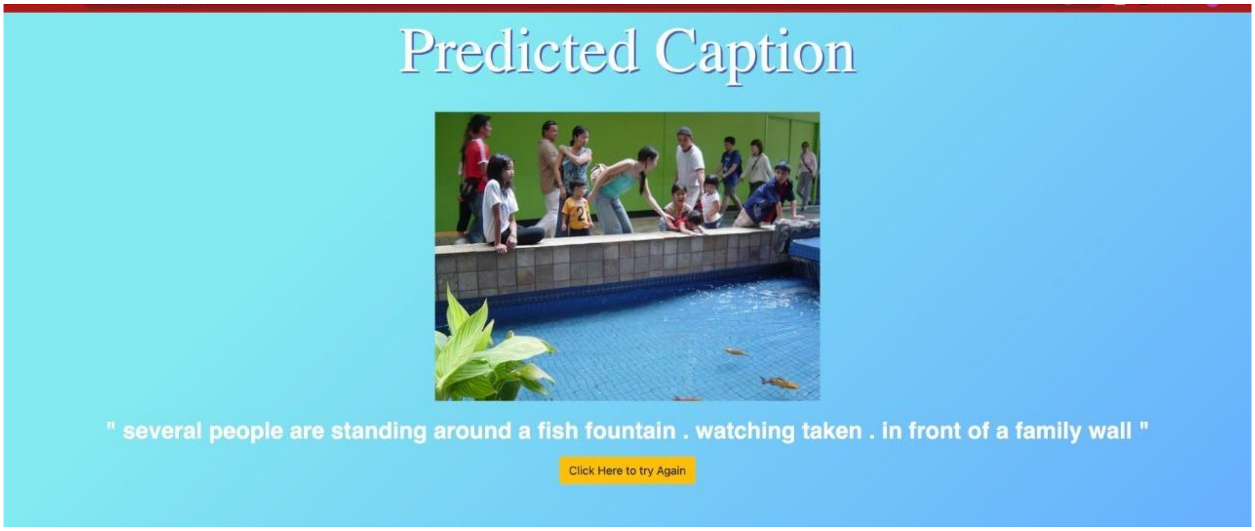


Automated Image Captioning

Want to describe your image ? Just upload and try

Groww - Calendar - Week of M
Infrastructure Engineering · Ol
Meet - Standup | Infra-Plat
Image Captioning
127.0.0.1:5000
Choose file
No file chosen
Generate

The generated sentence are shown in Fig 6.2 Generated sentences are " several people are standing around in a fish fountain " , while actual humans read as " several people are standing around fish fountain and watching them " .



Predicted Caption

" several people are standing around a fish fountain . watching taken . in front of a family wall "

shows image of captions generated for the image

CONCLUSION AND FUTURE SCOPES

CHAPTER 7

CONCLUSION AND FUTURE SCOPES

Image captioning is nonetheless a creating subject and many researches are nonetheless in progress. Recent work primarily based on deep studying methods has resulted in a leap forward in the accuracy of photograph captioning as it have breakdown the complicated fashions to easy structure. The textual content description of the picture can enhance the content-based photograph retrieval efficiency, the increasing utility scope of visible grasp in the fields of medicine, security, navy and different fields, which has a vast utility prospect. At the equal time, the hypothetical system and query procedures of photo inscribing can advance the improvement of the thought and utility of photograph comment and apparent question addressing (VQA), go media recovery, video subtitling and video exchange, which has vital instructive and reasonable programming esteem.

6.1 Use case and future Scopes

- Medical Application

  This model can be incorporated with a couple of sunglasses,cameras and listening devices, to assist the outwardly debilitated individual with getting the information on their environmental elements. One of the instances of this application is Horus Technology which in association with NVIDIA are chipping away at a similar task which is still in the improvement stage at the present time.

  This image is an example of Horus Technology[20]

- Intelligent monitoring permits the laptop to perceive and decide the behaviour of humans or automobiles in the captured scene and generate alarms beneath fantastic prerequisites to immediate the person to react to emergencies and forestall useless accidents.

One of such technologies is a developing phase where a palm-sized device is connected to a camera , giving a view of road alerting and helping drivers to reduce accidents . This technology is being developed by Intel IIIT Hyderabad and CPRI[21] .

- Campus Level Implementation

  This model can be integrated with cameras and alert systems such as messaging,sirens etc. to detect any absurd activities .

  Image cameras can detect ongoing robbery.

- Social Media, Platforms like facebook can surmise straightforwardly from the picture, where you are (ocean side, bistro and so forth), what you wear (variety) and all the more significantly the thing you're doing likewise (as it were) . This permits them to elevate promotions to the specific client of their advantage .

REFERENCES

1. Shuang Liu, Liang Bai,a, Yanli Hu and Haoran Wang. I*mage Captioning Based on Deep Neural Networks*. EITCE (2018)

   Available : Link

2. Lakshminarasimhan Srinivasan, Dinesh Sreekanthan,Amutha A.L. I2T: *Image Captioning - A Deep Learning Approach* (2018).

   Available : Link

3. Simao Herdade, Armin Kappeler, Kofi Boakye, Joao Soares.*Image Captioning: Transforming Objects into Words* . San Francisco, CA (2019).

   Available : Link

4. Aishwarya Maroju ,Sneha Sri Doma ,Lahari Chandarlapati , *Image Caption Generating Deep Learning Model ,J.N.T.U, Hyderabad , Sreenidhi Institute of Science And Technology* (2021).

   Available : Link

5. Zhongliang Yang, Yu-Jin Zhang, Sadaqat ur Rehman, Yongfeng Huang: *Image Captioning with Object Detection and Localization, Department of Electronic Engineering, Tsinghua University*, *Beijing 100084, China*

   Available : Link