DHA Suffa University (DSU), Karachi Campus
Department of Computer Science

Automatic Image Description Generation

Submitted by

Muhammad Kashif (CS-172032)
Muhammad Ali Khan (CS-172058)

Supervisor

Tehniyat Mirza

Co - Supervisor

Conrad D' Silva

In partial fulfillment of the requirements for the degree of
Bachelor of Science
(2020)

Certificate of Approval

It is certified that the research work presented in this report, entitled Automatic Image Description was conducted by Muhammad Ali Khan and Muhammad Kashif under the supervision of Ms.Tehniyat Mirza.

No part of this report has been submitted anywhere else for any other degree.

This report is submitted to the Department of Computer Science in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science

at the

DHA Suffa University (DSU)

Karachi Campus, Pakistan

Team Lead Name: Muhammad Kashif Signature: _____

Member 1 Name: Muhammad Ali Khan Signature: _____

Supervisor Name: Tehniyat Mirza Signature: _____

Supervisor Name: Tehniyat Mirza Signature: _____

Evaluation Committee:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

_____

HOD, Computer Science Department, DSU, Karachi

Authors' Declaration

We declare that this project report was carried out in accordance with the rules and regulations of the DHA Suffa University (DSU). The work is original except where indicated by special references in the text and no part of the report has been submitted for any other degree. The report has not been presented to any other University for examination.

Dated:

Authors Signatures:

Muhammad Kashif
<Team Lead name>

Muhammad Ali Khan
<Member 1 name>

Plagiarism Undertaking

We, Muhammad Kashif and Muhammad Ali Khan, solemenly declare that the work presented in the CS Project Final Report titled Automatic Image Description Generation has been carried out solely by ourselves with no significant help from any other person except few of those which are duly acknowledged. I confirm that no portion of our report has been plagiarized and any material used in the report from other sources is properly referenced.

Dated:

Authors Signatures:

Muhammad Kashif
<Team Lead name>

Muhammad Ali Khan
<Member 1 name>

Acknowledgements

First of all, we want to thank DHA Suffa University's marvelous instructors for providing us with the right information. Under the supervision of these mentors. Our were successful in completing our voyage and completing this significant part of the previous journey, and as a result, had the opportunity to recognize and use our own potential.

In addition, we would like to thank our supervisor, Ms. Tehniyat Mirza, for all the time and effort she invested in this project. She believed our modest idea was worth a shot in order to build something legendary. In addition to guiding us and providing critical feedback, she was always by our side, helping us achieve our goal.

Our thanks also go to Mr. Conrad D'Silva who has always believed in us. His constant encouragement has kept us on target and has inspired us to keep striving for greater success.
It was a great pleasure to have him onboard as our Co - Supervisor. His availability to discuss consultations was always a plus, and he would return our calls with patient answers. Lastly, we would like to thank you for taking the time to go through all this with us, despite the effort we have put in over the years.

Document Information

Customer	DHA Suffa University (DSU)
Project Title	Automatic Image Description Generation
Document	CS Project Final Report
Document Version	1.0
Identifier	<Project Code>Final Report
Status	Final
Author(s)	Muhamamd Kashif, Muhammad Ali Khan
Approver(s)	Ms. Tehniyat Mirza
Issue Date	

Definition of Terms, Acronyms and Abbreviations

This section should provide the definitions of all terms, acronyms, and abbreviations required to interpret the terms used in the document properly.

Term	Description
ML	Machine learning is an artificial intelligence (AI) technology that allows computers to automatically learn and improve from experience without being explicitly designed. Machine learning is concerned with the creation of computer programmes that can access data and utilise it to understand on their own.
DL	Deep Learning is a type of machine learning and artificial intelligence (AI) that imitates the way humans gain certain types of knowledge. Deep learning is an important element of data science, which includes statistics and predictive modeling.
CV	Computer Vision is an artificial intelligence area that teaches computers to interpret and comprehend the visual environment. Machines can reliably recognise and categorise items utilizing digital photographs from cameras and movies and deep learning techniques — and then react to what they "see."

CNN	Convolutional neural networks are a type of neural network that excels at processing data with a grid-like layout, such as images. Each neuron has its own receptive field and is coupled to other neurons so that the full visual field is covered..
RNN	Artificial neural networks that use sequential data or time series data are known as recurrent neural networks (RNNs). The output of recurrent neural networks is affected by the prior elements in the sequence, regardless of what the input is.
BLSTM	Bidirectional Long-Short Term Memory (LSTM) with bidirectional capability is a form of LSTM that is often employed for sequence classification. In problems where all timesteps of the input sequence are known, bidirectional LSTMs train two instead of one LSTM on the input sequence. As a result, the network can learn the problem more quickly and thoroughly by using two variations of the input sequence. The first version is based on the original

	input sequence, and the second is based on a reverse replica of the original input sequence.
NLP	Computers can understand human languages using Natural Language Processing, which combines artificial intelligence with linguistics.
LSTM	Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems.

Abstract

To generate automatic descriptions from images, we must understand how humans describe images in order to analyze a given image to predict its most likely interpretation. As part of our study, we examined a number of existing natural language description generating models for images - how they work, and how they produce descriptions for unknown images.

On the basis of the current model, we developed a deep learning model for extracting image data with Convolutional Neural Networks (CNN) and Deep learning models utilize Long Short-Term Memory (LSTM) techniques to generate natural language descriptions from the image features.

Our deep learning model will be connected to our web application backend so users can submit images and generate descriptions with ease. The web interface will also allow users to upload images, and then the backend will take care of processing the image and generating a description that will be visible to the users.

Table of Contents

Contents

Chapter 1

INTRODUCTION

Description about the Project

Using ResNet and LSTM (Long Short Term Memory), the Picture Description Generator is created in this project. A ResNet model will be used to extract the image features and train it against the ImageNet dataset. Then the features will be fed into an LSTM model and used to describe the images. As a picture description generator, Flickr 8k dataset will be used. Other big datasets like Flickr 30K and MSCOCO are available, but training a network with those datasets might take weeks, so this modest dataset, Flickr8k, will be utilized.

Details about the Domain

Computer Vision and Deep Learning are both part of the project.

1. **Computer Vision:** An area of computer science known as Computer Vision deals with the study of digital image modeling and, in particular, identifying relations between objects in photos and videos. Computer Vision is applied in automated sectors such as cybernetics and surveillance.[8] Computing vision is concerned with allowing computers to perceive

visuals on a human level.

2. **Deep Learning** : The application of "Deep Learning" to unstructured data, such as photos, videos, text, and speech, has outperformed standard [9] Machine Learning models. Deep learning models use neurons that look like humans, referred to as perceptrons.

Relevant Background

The Image Description Generation process involves framing written descriptions for provided photographs. It is an important and fundamental task in Deep Learning. Image descriptions have a wide range of uses.

NVIDIA is developing an application using its picture description generating capabilities for use by individuals with reduced or no vision. Several approaches have been introduced to address the challenge. This is a well-known problem and one of the most important solutions presented by Andrej Karpathy, Director of AI, Tesla, in his Ph.D. [10] at Stanford.

Chapter 2

RELEVANT BACKGROUND & DEFINITIONS

Image description generation may be seen as an end-to-end Sequence to Sequence issue, since it converts images, which is regarded as a sequence of pixels to a sequence of words. For this aim, we need to process both the words or statements and the visuals. For the Language component, we utilize recurrent Neural Networks and for the Image part, we use Convolutional Neural Networks to acquire the feature vectors correspondingly.

Now, How does the notion work?

Say, we as humans are experiencing a situation as depicted below.

Fig 1. A picture of a dog sitting on a blue towel.

If we are prompted to describe it, maybe we may describe it as: “A puppy on a blue towel” or “ A brown dog playing with a green ball”. So, how are we doing this? While forming the description, we are seeing the image but at the same time, we are looking to generate a meaningful sequence of words. The first half is handled by CNNs while the second is handled by LSTMs.

If we can collect a decent dataset with photos and their related human descriptions, we can train networks to automatically label images. FLICKR 8K, FLICKR 30K, and MS-COCO are among the most utilized datasets for this purpose.

Definitions:

Page numbers 6 - 8 of this document list all definitions relevant to this project.

Chapter 3

LITERATURE REVIEW & RELATED WORK

Literature Review

Author	Year	Name	Technique
MOHD SANAD ZAKI RIZVI	2020	Learn Image Classification on 3 Datasets using Convolutional Neural Networks (CNN)	CNN
Victor Zhou	2019	Machine Learning for Beginners: An Introduction to Neural Networks	Feed Forward
Sagar Sharma	2017	Activation Functions in Neural Networks	Sigmoid tanh Softmax ReLU

Robert Di Pietro Gregory D.Hager	2020	Deep learning: RNNs and LSTM	RNN LSTM

Related Work

Much study has been dedicated to automated picture description generation recently. The study may be roughly grouped into three major groups comprising the template-based techniques, retrieval-based approaches, and new picture description generating approaches.

Template-based approaches allow for descriptions to be produced by utilizing templates that have an unfilled number of slots. To accomplish this, the blank spaces in the template are filled by identifying distinct objects, properties, and actions first. Li et al. extract the related sentences about identified objects, properties, and their connections [1] for filling the template slots. Farhadi et al. use a triplet of scene components to fill in template slots.[2] To fill in the gaps, Kulkarni et al. perform an inference based on conditional random fields.[3] Although templates can create grammatically accurate descriptions, descriptions cannot be changed once they have been created.

Using retrieval-based strategies, descriptions for an image are generated by choosing from the sentence pool or copying from visually similar images the sentences that are most semantically related. Using stacked auxiliary embedding, Gong et al. include millions of poorly annotated photos into picture descriptions.[4] According to Ordonez et al., the Flickr database contains millions of photos and their accompanying descriptions and could easily identify comparable photos in it. [5]

Sun et al. first cluster comparable phrases and pictures together based on their semantic similarity and visual similarity scores, and then obtain a description of the target image based on descriptions of similar images in the same cluster.[6] A rank-based approach to sentence-based picture description uses Hodosh et al. to come up with generic and syntactically accurate descriptions to solve the problem of sentence-based picture description.[7] Yet, the production of semantically precise, image-specific descriptions poses a challenge for them.

Chapter 4

METHODOLOGY

Software Engineering Methodology

In order to break this project into two sections, we must use the following technique:

1. **Machine Learning Lifecycle:** Data Scientists, as well as ML developers, use Machine Learning Lifecycle in their efforts to construct an AI. It is an iterative process used to build an AI to address specific challenges.
    - Baseline modeling without augmentation of raw data..
    - Performance evaluation
    - Input features have been expanded with custom data
    - New data will be used to retrain the model

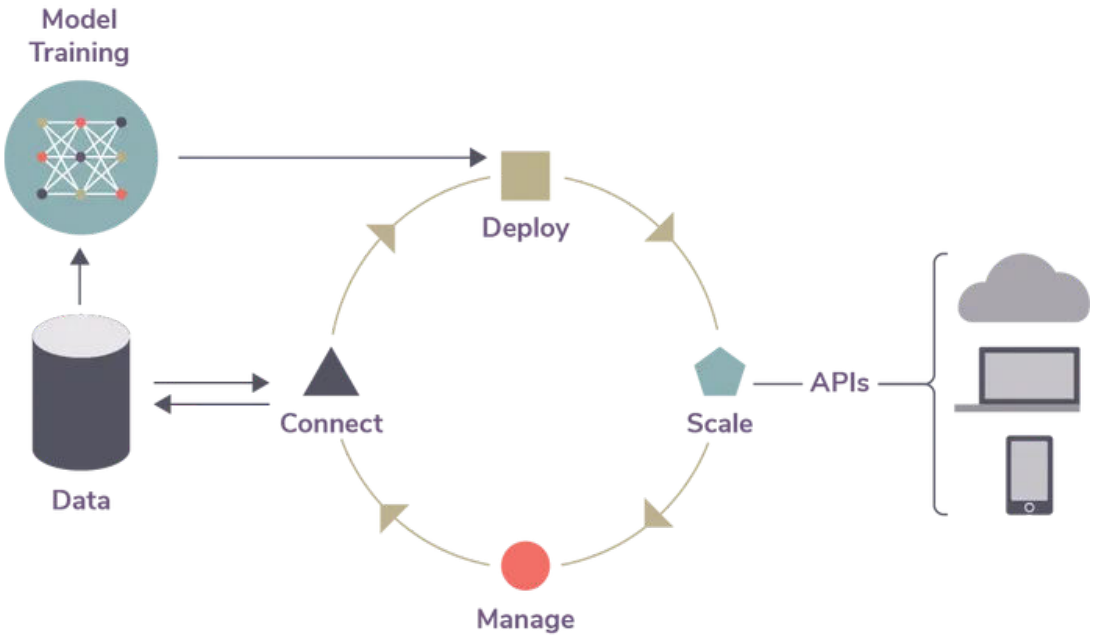


Fig 2. An image of machine learning lifecycle methodology

Advantages

- Progress always needs to be made
- Everything will be automated
- Identification of patterns and trends
- Several applications are available

2. A software project is incrementally modeled as follows:

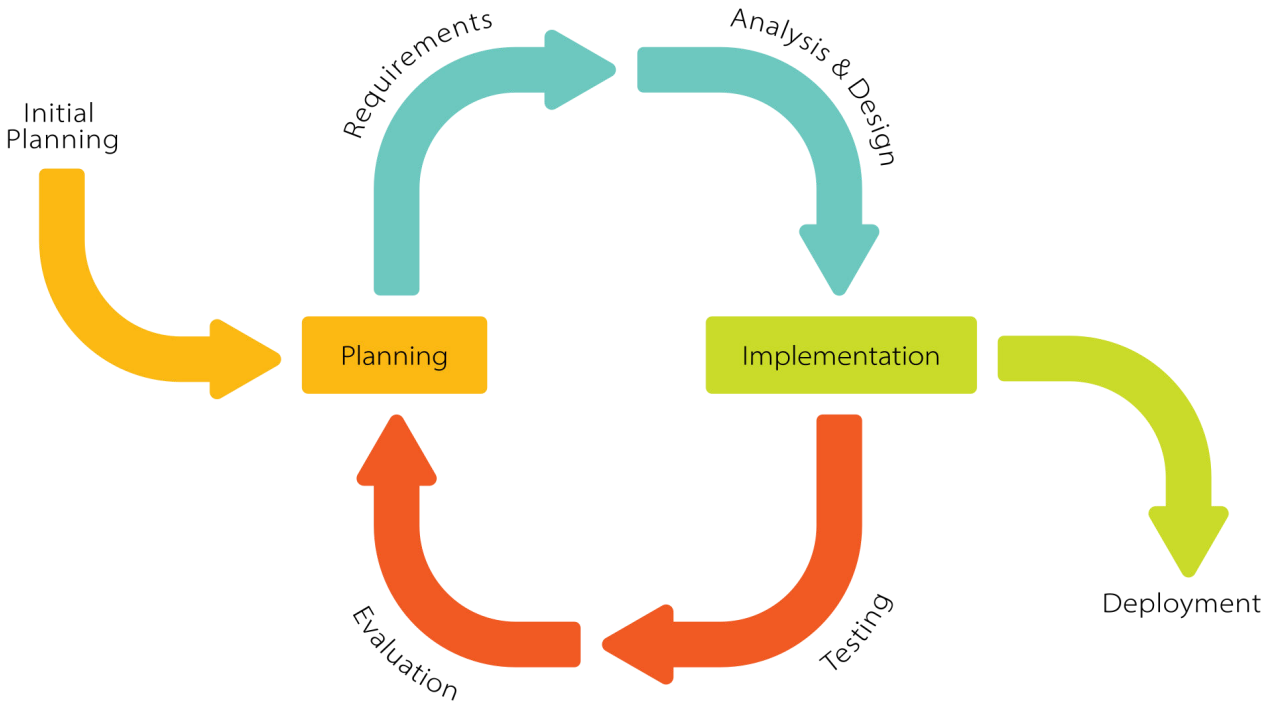


Fig 3. Incremental Software Methodology

- The addition of one feature at a time and deployment
- Using Flask to create APIs
- Development of web applications to input images to model.

Project Methodology

Gantt Chart

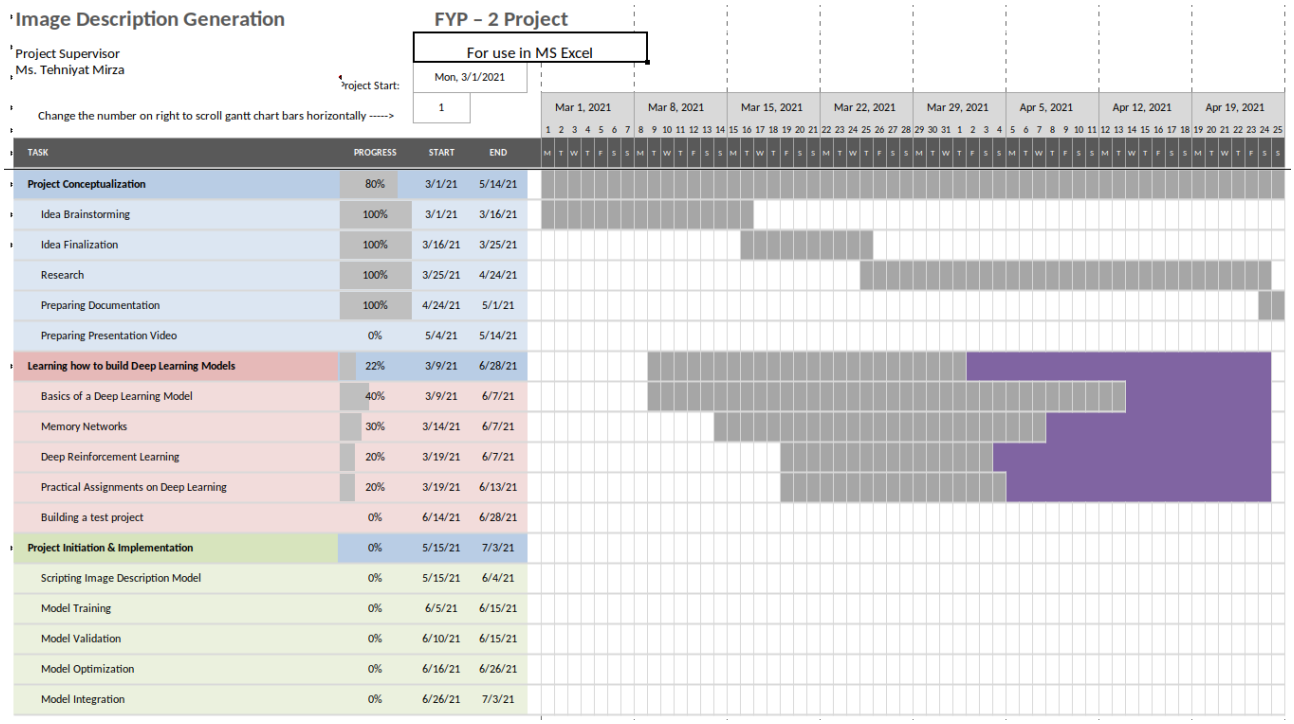


Fig 4. Gantt Chart showing timelines

Project Budgeting:

S#	Item	Description	Cost
1	**PC with GPU**	To train statistical models and process images using Deep Learning, a PC must be used to run libraries such as TensorFlow, Pytorch, and Keras on a GPU.	320000 PKR (Cost can be cut down if trained in FYP Labs)
2	**Server Deployment (Load Dependent)**	To put our DL-trained model to work making predictions online.	20000 - 30000 PKR / Month
3	**Maintenance Cost**	Retraining model using new data manipulation scripts.	15000 PKR / Month + Item # 1

Budgeted Cost - The Project's estimated cost

If FYP lab access is not available, include PC with GPU in your budget	320000 PKR (One time costs) 35000 - 45000 PKR/ Month
Budget without PC with GPU (FYP lab access)	35000 - 45000 PKR/ Month

Chapter 5

EXPERIMENTAL EVALUATIONS & RESULTS

Evaluation Testbed

Based on induction, deep learning employs diverse methods of evaluating performance, such as accuracy, precision, recall, and other criteria based on the data.

Each batch of training is processed by a single epoch representing a forward + reverse pass.

Training Performance

```
Epoch 86/100
189/189 [==============================] - 12s 66ms/step - loss: 0.5204 - accuracy: 0.8553
Epoch 87/100
189/189 [==============================] - 13s 67ms/step - loss: 0.5194 - accuracy: 0.8554
Epoch 88/100
189/189 [==============================] - 12s 65ms/step - loss: 0.5002 - accuracy: 0.8602
Epoch 89/100
189/189 [==============================] - 13s 66ms/step - loss: 0.5073 - accuracy: 0.8585
Epoch 90/100
189/189 [==============================] - 12s 65ms/step - loss: 0.4960 - accuracy: 0.8613
Epoch 91/100
189/189 [==============================] - 12s 65ms/step - loss: 0.4901 - accuracy: 0.8634
Epoch 92/100
189/189 [==============================] - 13s 68ms/step - loss: 0.4892 - accuracy: 0.8632
Epoch 93/100
189/189 [==============================] - 12s 66ms/step - loss: 0.4838 - accuracy: 0.8646
Epoch 94/100
189/189 [==============================] - 12s 65ms/step - loss: 0.4818 - accuracy: 0.8661
Epoch 95/100
189/189 [==============================] - 12s 66ms/step - loss: 0.4717 - accuracy: 0.8674
Epoch 96/100
189/189 [==============================] - 12s 66ms/step - loss: 0.4755 - accuracy: 0.8658
Epoch 97/100
189/189 [==============================] - 13s 66ms/step - loss: 0.4694 - accuracy: 0.8677
Epoch 98/100
189/189 [==============================] - 13s 67ms/step - loss: 0.4634 - accuracy: 0.8683
Epoch 99/100
189/189 [==============================] - 12s 65ms/step - loss: 0.4661 - accuracy: 0.8674
Epoch 100/100
189/189 [==============================] - 12s 66ms/step - loss: 0.4543 - accuracy: 0.8709
```

Fig 5. Picture of model accuracy and loss

Chapter 6

CONCLUSION AND DISCUSSION

Limitations and Future Work

- Developing a CNN Architecture, or just any Deep Learning and Computer Vision model in particular, would need a significant amount of computational capability, often provided by Graphics cards.
- While training, the input of each time step is the real description, and when generating, each word is based on the one it replaced; it may take a long time for a word to get close to the truth once it has not been generated well. Discrepancies like this are known as the "exposure bias" problem. As a result, we presume that the word was formed correctly and is based on previously generated terms.
- We are assuming that if the developers want to generate more specific image descriptions they will buy our subscription and train the model to generate more specific image descriptions

Reasons for Failure – If Any

- If the blurry image or the image contain distortions the model may generate the wrong image description in such circumstances.

REFERENCES

[1] A. Farhadi *et al.*, “Every picture tells a story: Generating sentences from images,” in *Computer Vision – ECCV 2010*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 15–29. Accessed: Feb. 27, 2022. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-15561-1_2

[2] S. L. ; G. K. ; T. L. B. ; A. C. B. ; Y. Choi, “Composing Simple Image Descriptions using Web-scale N-grams”.

[3] G. Kulkarni *et al.*, “BabyTalk: Understanding and Generating Simple Image Descriptions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2891–2903, Dec. 2013, doi: 10.1109/tpami.2012.162.

[4] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, “Improving Image-Sentence Embeddings Using Large Weakly Annotated Photo Collections,” in *Computer Vision – ECCV 2014*, Cham: Springer International Publishing, 2014, pp. 529–545. Accessed: Feb. 27, 2022. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-10593-2_35

[5] V. Ordonez, G. Kulkarni, and T. Berg, “Im2Text: Describing Images Using 1 Million Captioned Photographs,” *Advances in Neural Information Processing Systems*, vol. 24.

[6] C. Sun, C. Gan, and R. Nevatia, “ICCV 2015 Open Access Repository.” https://openaccess.thecvf.com/content_iccv_2015/html/Sun_Automatic_Concept_Discovery_ICCV_2015_paper.html

[7] M. Hodosh, P. Young, and J. Hockenmaier, “Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics,” *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, Aug. 2013, doi: 10.1613/jair.3994.

[8] IBM, “What is Computer Vision? ,” *IBM*. https://www.ibm.com/topics/computer-vision (accessed Feb. 28, 2022).

[9] J. Brownlee , “What is Deep Learning?,” *Machine Learning Mastery*, Aug. 15, 2019. https://machinelearningmastery.com/what-is-deep-learning/ (accessed Feb. 28, 2022).

[10] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” Jun. 2015. Accessed: Feb. 28, 2022. [Online]. Available: http://dx.doi.org/10.1109/cvpr.2015.7298932

[11] “CNN Image Classification,” *Analytics Vidhya*, Feb. 18, 2020. https://www.analyticsvidhya.com/blog/2020/02/learn-image-classification-cnn-convolutional-neural-networks-3-datasets/ (accessed Feb. 28, 2022).

[12] V. Zhou, “Machine Learning for Beginners: An Introduction to Neural Networks,” *Towards Data Science*, Dec. 20, 2019. Accessed: Feb. 28, 2022. [Online]. Available: https://towardsdatascience.com/machine-learning-for-beginners-an-introduction-to-neural-networks-d49f22d238f9

[13] S. SHARMA, “Activation Functions in Neural Networks,” *Towards Data Science*, Jul. 04, 2021. Accessed: Feb. 28, 2022. [Online]. Available: https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6

[14] “Deep learning: RNNs and LSTM,” *ScienceDirect*. https://www.sciencedirect.com/science/article/pii/B9780128161760000260 (accessed Feb. 28, 2022).

APPENDICES

List of Appendices

A0. Copy of Project Registration Form

A1 A. Project Proposal and Vision Document

A1 B. Copy of Proposal Evaluation Comments by Jury

A2. Research Documentation

A3. Other Technical Details

A4. Flyer & Poster Design

A5. Copy of Evaluation Comments

A6. Meetings' Minutes

A7. Project Progress

A8. Research Paper

A0. Copy of Project Registration Form

Image Description Generation

Project Title:	Image Description Generation		
Supervisor Name:	Ms. Tehniat Mirza		
Co-Supervisor(s) Name:	Mr. Conrad D' Silva		
Team Lead Name:	Muhammad Kashif	**Reg. No.**	CS172032
Team Member 2 Name:	Muhammad Ali Khan	**Reg. No.**	CS172058
Team Member 3 Name:	-	**Reg. No.**	-
Project Keywords (Comma separated terms describing your project domain)			
Deep Learning, Image Captioning, Image Description Generation			
Project Abstract			

The project's overall strategy is to analyze a given image's visual content to anticipate its most likely interpretation, and then construct a phrase that reflects that meaning. Generating automatic descriptions from images requires an understanding of how humans describe images [1]. We looked at many existing natural language descriptions generating models for images and how they work in order to produce natural language descriptions for unknown images. We created a deep learning model for generating natural language descriptions of images based on the current model. Convolutional Neural Networks (CNN) are used to extract image data, and Long Bidirectional Short-Term Memory (BLSTM) is applied to generate natural language descriptions from the image features in a deep learning model. For easy integration of this model in various websites and applications an API will be implemented that can be used by the developers of these systems. After a developer POST the image to our API our system will generate a description relevant to the image and a description will be sent back to the developer using our API.

_____ **Supervisor**	_____ **Team Lead**	_____ **Team Member 2**	_____ **Team Member 3**

A1 A. Project Proposal and Vision Document

DHA Suffa University

Department of Computer Science

Final Year Project

Automatic Image Description Generation

(R-1706)

Research Documentation

Submitted by

Muhammad Kashif (CS172032)

Muhammad Ali Khan (CS172058)

Supervisor(s)

Ms. Tehniat Mirza

Co - Supervisor(s)

Mr. Conrad Walter D' Silva

Document Sign off Sheet

Document Information

Project Title	Automatic Image Description Generation
Document	Project Proposal
Document Version	1.0
Identifier	FYP01-PP
Status	< Draft/ Final / etc..>
Author(s)	Muhammad Kashif, Muhammad Ali Khan
Approver(s)	Ms. Tehniat Mirza, Mr. Conrad Walter D'Silva
Issue Date	

Team Lead Name: Muhammad Kashif
Date:30/03.2021

Signature:

Member 1 Name: Muhammad Ali Khan
Date: 30/03.2021

Signature:

Supervisor Name: Ms. Tehniat Mirza
Date: _____

Signature: _____

Co-Supervisor Name: Mr. Conrad Walter D'Silva
Date: _____

Signature: _____

CS Project I Coordinators:

Name: _____
Signature: _____ Date: _____

Name: _____
Signature: _____ Date: _____

Revision History

Date	Version	Description	Author
13/04/2021	1.0	initial document	

Definition of Terms, Acronyms, and Abbreviations

[This section should provide the definitions of all terms, acronyms, and abbreviations required to interpret the terms used in the document properly.]

Term	Description
CNN	Convolutional Neural Network
BLSTM	Bidirectional Long Short-Term Memory (BLSTM)
CV	Computer Vision
DL	Deep Learning
NLP	Natural Language Processing
RNN	Recurrent Neural Network
ML	Machine Learning

Table of Contents

1 Introduction

1.1 Problem Statement

Problem Of	Almost 11 billion photographic images are taken every day. Classifying and organizing these images in such a way that we can locate a specific group of pictures or a unique picture quickly and easily is a challenging task. You can organize your images much more quickly and easily with the automatic generation of image descriptions. The system generates text to describe what is shown in the image. Consequently, users and search engines will be able to see more images, which is beneficial in many ways.
Effect	Tagging products automatically in online catalogs, as well as other areas of business and research. This will also be invaluable

	for the blinds and visually impaired community that will enable them to better manage everyday situations without assistance. Converting the scene into text will allow us to do this. Developers can use our API to generate image descriptions for their relevant domains

1.2 Product Position Statement:

For	Developers
Who	For developers who want to generate image descriptions automatically by just passing an image to our API.
The (Product Name)	Automatic Image Description Generation
That	Generates description of images which are passed to our API which help in making automated applications in their relevant

[redacted]	domains.

1.3 Project Motivation and Background

Automatic visual description creation is an intriguing subject in computer vision due to the rapid advancement of deep learning methods. It facilitates access to photo collections on social media and provides assistance to visually challenged persons. The motivation of the project was that there was not any API to generate automated image descriptions available. Earlier., the developers were writing the description manually. Now with our API development. The developer may easily generate the automated image description which will be time saving and relevant for their specific domain.

1.4 Objectives

Image description generation is a complicated task to do. Doing it multiple times for multiple images can be a time consuming task. Our objective is to automate this process by creating an API to which an image can be passed and it will return an image relevant description.

1.5 Literature Review and GAP Analysis

In the literature review we focused on the problem of how a system can automatically understand an image and create relationships between objects because the human brain can only generate a description of an image if he understands the relations of objects in an image.

1.6 GAP Analysis:

Paper	Algorithm	Year	Citations	Limitations
Simple convolutional neural network on image classification	CNN	2017	173	Limitations of doing image classification using CNN is that it only identifies the class of an image.
Application of Deep Learning for Object Detection	CNN	2018	112	Object Detection was one of the problems of Deep Learning which was solved by CNN and its variants. Limitations of this approach was that, to detect a object in image it returned coordinates of

				object on a screen
Understanding Deep Learning Techniques for Image Segmentation	CNN RCNN CRF RNN	2019	73	After the task of image segmentation is performed the objects are separated from the image and it returns a vector representation of the individual objects in the image.
Detecting Visual Relationship s with Deep Relational Networks	RCNN CRF	2017	313	-

To understand the relationship between objects we first need to assign a text label (Image descriptions) of each individual object of an image and then return their respective vector representation. This task is carried out by fusing two models together. We first use an encoder that returns a vector of an object in an image and a decoder takes the vector as an input, returning a label

for that object. The decoder model can be RNN and its variants. In Order to generate labels and not only a description we can use the LSTM or Bi Directional LSTM model.

1.7 Additional Papers Read Related To Our Project:

Paper	Algorithm	Year	Citations
Simple Image Description Generation via a linear based approach	CNN LSTM	2014	33
Show and Tell: A Neural Image Caption Generator	CNN RNN	2015	4583
Collective generation of natural image descriptions	ILP Formulation Surface Realization	2012	338
A hierarchical and regional deep learning architecture for image description generation	R-CNN RNN	2019	33
Mind's Eye: A Recurrent Visual Representation for Image Caption Generation	RNN LSTM	2015	481
See No Evil, Say	Extension of The	2014	-

No Evil: Description Generation from Densely Labeled Images	Midge Algorithm		

2 Project Vision

2.1 Business Case and SWOT Analysis

Business Case:

As technologies advance, the world will become increasingly dependent on them and is moving towards automation in which machine learning, deep learning etc are being used to solve such problems. Our system aims to solve the problem of developers who want to generate image descriptions for their projects.For the business perspective the API/Function will be highly utilized when the image description generation process gets automated and if developers want to generate a more specific/in-depth image descriptions, they can use a subscription through which the model can be trained according to their relevant domain. After that it can generate more specific image descriptions.

SWOT Analysis

- **Strengths**

    1. Image descriptions are meant to explain what the image is, in a way that provides a similar effect to seeing the image. Due to which our project focuses primarily on generating automated image descriptions by just calling an API and using it.
    2. There are few available API for this task available on the Internet which makes it a very good problem to tackle and solve for developers.

- **Weaknesses:**

    1. While training the model if the input image is blurred or pixels are damaged then the model might predict wrong predictions for that image.
    2. Descriptions generated for an image might contain grammatical errors.

- **Opportunities:**

    1. There are very few API available to choose from for solving this particular problem.

2. By creating an API we can solve the problem of generating automated image descriptions.

- **Threats:**

1. We fail in the system due to unreliable images from users.
2. Because of lack of maintenance, our system may lose its quality after deployment.

2.2 Stakeholder Summary

Name	Description	Responsibilities
Team Members	Members working on this project from analysis to deployment	• Division of tasks equally. • Keep track of the tasks that have been accomplished. • Completing the tasks within the timeframes set out. • Ascertaining that the system satisfies the criteria.
Developers	The people who will be importing and using our API.	• Using our API to generate automated image descriptions.

2.3 User Summary

Name	Description	Responsibilities
Web/App Developers	The people who will be importing and using our API.	● Using our API to generate automated image descriptions for their websites and applications
Data Science Analyst	The people who will be importing and using our API.	● Using our API to work with image descriptions for analyzing image description data.

2.4 Business Objectives and Success Criteria:

Our system will have success if these factors are met:

- **High Quality Photos:** This would mean that the results would be wrong too if the photos used to train the model aren't of the highest quality.
- **Integration:** The integration with the projects for helping the developers to automate the process of generating image descriptions

- **Time saving:** When the time savings are noticed, our system will be hailed.

2.5 Project Risks and Risk Mitigation Plan

Project Risks:

It is unlikely that there will be many risks, including

1. If the input image is blurred or not clear, It may result in wrong description or inaccurate navigation.
2. The image description is obtained by predicting the most likely nouns, verbs, scenes, and prepositions that make up the sentence.Sometimes the description may contain grammatical issues.

Risk Mitigation Plan:

- According to the above risk, We need to provide clear photos and the photos that'll be training the model need to be up to the mark for better results.

2.6 Assumptions and Dependencies

- The model is given high quality pictures.

- According to our assumption the developers will highly use our API for generating automated image descriptions for their project.
- We are assuming that if the developers want to generate more specific image descriptions they will buy our subscription and train the model to generate more specific image descriptions.

3 Project Scope

3.1 In Scope

- Detection of Objects.
- Generation of Descriptions.
- Conversion of scene into text..
- Achieve accuracy of upto 80-85%
- Creating an API which takes images as input.

3.2 Out of Scope

- Blurry image result in incorrect description
- Mobile/Web Applications

4 Proposed Methodology

4.1 SDLC Approach (Waterfall/Agile/Spiral)

Because this research project does not include big portions, such as model training and API implementation, the Waterfall methodology will now be employed. We will first complete one phase, then move onto the next. Such a method enables everyone to complete every step before going off to any subsequent round. As a result of this approach, we are able to complete each phase before moving on to the previous one, ultimately culminating in the integration of all information.

4.2 Team Role & responsibilities

Tehniat Mirza - Project Supervisor

Conrad D' Silva - Project Co-supervisor

Team Member	*Responsibilities*
Muhammad Kashif (Team Lead)	• Project Management • Project Proposal • Model Training • Back End Development • Project Integration

Muhammad Ali Khan	• Project Proposal • Model Training • Back End Development • Project Integration • Project Testing

4.3 Requirement Development Methodology

The first phase outlines the data collection methods as well as the project's data collection criteria. The techniques used for gathering information are by reading the research papers, articles etc. The other phase will be working on image description generation function by applying CNN, BLSTM and different tools which will convert the image scene into text. Another Phase will be the development by converting the above function into an API which will be used by developers in their projects to generate automated image descriptions.

4.4 High level Architecture / Design

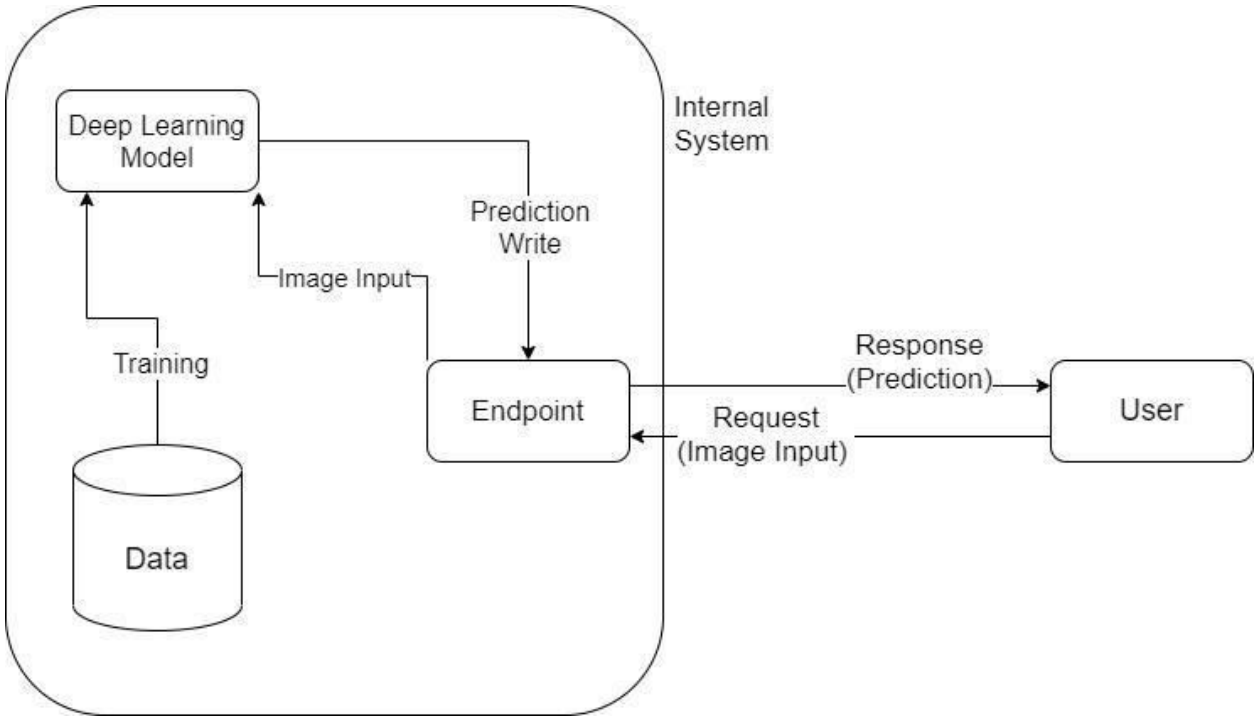
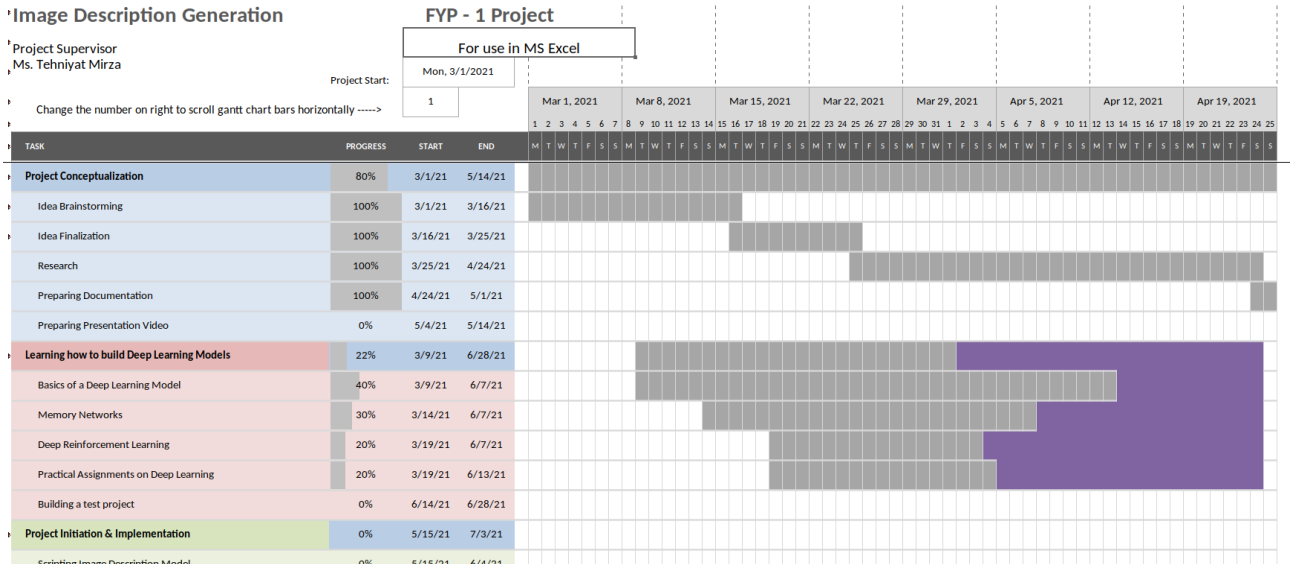


Fig 6. Data flow between model and users.

5. Project Planning

5.1 Gantt Chart



6 Project Requirements

6.1 Software tools requirements:

- Python (and it's Libraries)
- Jupyter Notebook / Google Colab
- Tensorflow/Keras
- Windows / Linux / another OS

6.2 Hardware requirements

- .A computer equipped with a GPU with enough compute power for training of a Deep Learning model.

7 Budget/Costing

7.1 Estimated Budgeted Cost of the Project

S.no	Item	Cost
1	Model Training	10000pkr/month
2	PC with GPU	320000 pkr
3	Server Cost (Load Dependent)	30k - 70k / Month
4	Maintenance Cost	15k pkr/ Month

8 Project Deliverables

8.1 Phase I - Alpha Prototype

Data collection and Research

8.2 Phase II - Beta Prototype

- Trained Computer Vision Model for testing

8.3 Phase III - Release Candidate

- Testing and tuning the model for possible errors for better optimization and speed.

8.4 Phase IV - Final Product

- Conversion of the model into an API
- Server Deployment

9 Proposed GUI (Prototype)

Since the output would be an API therefore there won't be GUI. Developers would be importing our API and using it in their projects. .

10 Meetings held with supervisor and/or client.

Meeting Number	Topic	Tasks	Date	Comments of Supervisor
1	Idea Discussion			
2	Explained Gradient Descent, Loss Function, Working of a DL Model and Back Propagation	Code a simple neural network	Tuesday March 9, 2021	completed
3	Explained Sequential/Parallel Models, activation function and how to evaluate a model	Code a simple neural network		

4	FYP Idea Discussion	Research on FYP Idea	Thursday March 25, 2021	
5	Explained Pooling, Filtering and CNN	Implement CNN		
6	Discussion on CNN Project			
7	Object Detection Explanation			

11 References

Chen, Xinlei, and C. Lawrence Zitnick. “Mind's Eye: A Recurrent Visual Representation for Image Caption Generation.” *Proceedings of the IEEE conference on computer vision and pattern recognition*, no. 2015, https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Chen_Minds_Eye_A_2015_CVPR_paper.html.

Dai, Bo, Yuqi Zhang, and Dahua Lin. “Detecting visual relationships with deep relational networks.” *Proceedings of the IEEE conference on computer vision and Pattern recognition*, 2017, https://openaccess.thecvf.com/content_cvpr_2017/html/Dai_Detecting_Visual_Relationships_CVPR_2017_paper.html.

dshahid380. “Convolutional Neural Network.” *Towards Data Science*, 15 February 2019, https://towardsdatascience.com/covolutional-neural-network-cb0883dd6529.

Evergreen. “Automatic Image Captioning Using Neural Networks.” *Evergreen.team*, Evergreen, 20 07 2020, https://evergreen.team/articles/automatic-image-captioning.html#:~:text=Automatic%20image%20captioning%20(tagging)%20allows,is%20shown%20on%20the%20picture.

Ghosh, S., Das, N., Das, I., & Maulik, U. “Understanding deep learning techniques for image segmentation.” *ACM Computing Surveys (CSUR) 52.4*, 2019, https://dl.acm.org/doi/pdf/10.1145/3329784.

JOUR, et al. “An Overview of Image Caption Generation Methods.” 2020. *Hindawi*, https://www.hindawi.com/journals/cin/2020/3062706/.

Kinghorn, Philip, Li Zhang, and Ling Shao. “A hierarchical and regional deep learning architecture for image description generation.” *Pattern Recognition Letters 119*, vol. 2019, https://www.sciencedirect.com/science/article/abs/pii/S0167865517303240.

Kuznetsova, Polina, et al. “Collective generation of natural image descriptions.” *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2012, https://www.aclweb.org/anthology/P12-1038.pdf.

Mark Yatskar Michel Galley Lucy Vanderwende Luke Zettlemoyer. “See No Evil, Say No Evil: Description Generation from Densely Labeled Images.” *Microsoft*, Microsoft, https://www.microsoft.com/en-us/research/publication/see-no-evil-say-no-evil-descripti o n-generation-from-densely-labeled-images/.

Oriol, et al, Vinyals. “Show and tell: A neural image caption generator.” *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Vinyals_Show_and_Tell_2015_CVPR_paper.html.

Pathak, Ajeet Ram, Manjusha Pandey, and Siddharth Rautaray. “Application of deep learning for object detection.” *Procedia computer science 132*, 2018.

Rémi, Lebret, et al. “Simple image description generator via a linear phrase-based approach.” *arXiv preprint arXiv:1412.8419*, no. 2014, https://arxiv.org/abs/1412.8419.

S R, S.; Idicula, S.M. “Dense Model for Automatic Image Description Generation with Game Theoretic Optimization.” 2019. *mdpi.com*, https://www.mdpi.com/2078-2489/10/11/354.

T. Guo, J. Dong, H. Li and Y. Gao. “Simple convolutional neural network on image classification.” *IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, no.721-724, doi: 10.1109/ICBDA.2017.8078730, 2017, https://ieeexplore.ieee.org/abstract/document/8078730.

A1 B.Copy of Proposal Evaluation Comments by Jury

Name	Code	Status	Changes	Suggestion
Jury Member 1	**R-1706**	ACCEPT (possibly with or without suggestions)	A literature review is missing recent papers.	
Jury Member 2	**R-1706**	ACCEPT (possibly with or without suggestions)	Please update your problem statement regarding blind or visually impaired persons. You will create the captions in text so it is not going to help blind or visually impaired persons unless you convert this text into sound/voice.	
Jury Member 3	**R-1706**	ACCEPT (possibly with or without suggestions)		
Jury Member 4	**R-1706**	ACCEPT (possibly with or without suggestions)		

A2 . Research Documentation

DHA Suffa University

Department of Computer Science

Final Year Project

Automatic Image Description Generation

(R-1706)

Research Documentation

Submitted by

Muhammad Kashif (CS172032)

Muhammad Ali Khan (CS172058)

Supervisor(s)

Ms. Tehniat Mirza

Co - Supervisor(s)

Mr. Conrad Walter D' Silva

Document Sign off Sheet

Document Information

Project Title	Automated Image Description Generation
Project Code	R - 1706
Document Name	Research Documentation
Document Version	<1.0>
Document Identifier	R - 1706
Document Status	< Draft/ Final / etc..>
Author(s)	Muhammad Kashif (CS172032) Muhammad Ali Khan (CS172058)
Approver(s)	Ms. Tehniat Mirza Mr. Conrad D'Silva
Issue Date	20/07/21

Name	Role	Signature	Date
Muhammad Kashif	Team Lead		19/July/2021
Muhammad Ali Khan	Team Member 2		19/July/2021
Ms. Tehniat Mirza	Supervisor		19/July/2021
Mr. Conrad Walter D' Silva	Co-Supervisor	Conrad	19/July/2021

Revision History

Date	Version	Description	Author
20/July/2021	1.0	Initial Document	Muhammad Kashif Muhammad Ali Khan

Definition of Terms, Acronyms, and Abbreviations

Term	Description
ML	Machine learning is an artificial intelligence (AI) technology that allows computers to automatically learn and improve from experience without being explicitly designed. Machine learning is concerned with the creation of computer programmes that can access data and utilise it to understand on their own.
DL	Deep Learning is a type of machine learning and artificial intelligence (AI) that imitates the way humans gain certain types of knowledge. Deep learning is an important element of data science, which includes statistics and predictive modeling.
CV	Computer Vision is an artificial intelligence area that teaches computers to interpret and comprehend the visual environment. Machines can reliably recognise and categorise items utilizing digital photographs from cameras and movies and deep learning techniques — and then react to what they "see."
CNN	Convolutional neural networks are a type of neural network that excels at processing data with a grid-like layout, such as images. Each neuron has its own receptive field and is coupled to other neurons so that the full visual field is covered.

RNN	Artificial neural networks that use sequential data or time series data are known as recurrent neural networks (RNNs). The output of recurrent neural networks is affected by the prior elements in the sequence, regardless of what the input is.
BLSTM	Bidirectional Long-Short Term Memory (LSTM) with bidirectional capability is a form of LSTM that is often employed for sequence classification. In problems where all timesteps of the input sequence are known, bidirectional LSTMs train two instead of one LSTM on the input sequence. As a result, the network can learn the problem more quickly and thoroughly by using two variations of the input sequence. The first version is based on the original input sequence, and the second is based on a reverse replica of the original input sequence.
NLP	Computers can understand human languages using Natural Language Processing, which combines artificial intelligence with linguistics.
LSTM	Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems. [4]

Table of Contents

1 Introduction

1.1 Purpose of Document

The purpose of this document is to consolidate as well as evaluate just about many of the various concepts that have emerged in order to describe the framework as well as its needs. In order to gain a better understanding of the project, we will forecast and organise how we anticipate such a study to be conducted, draft thoughts that might have been generated later, and document ideas that are being examined and might be deleted even as the solution progresses.

In summary, the goal of this study paper is to offer a thorough overview of our research, including its requirements and aims. This article describes the research's target audience, as well as the hardware and software requirements. It establishes our team's and the subject's impressions of something like the study as well as its capability.

1.2 Intended Audience

The said paper is intended primarily for DHA Suffa Campus's programmers, professional, and scholastic employees. The document provides developers and designers of the system a guideline to develop the Image Description Generation System the way it was envisioned. The API Development will be Implemented using Python and it’s framework, to understand the project implementation the reader should have some theoretical knowledge of Deep Learning, Machine Learning and Python programming language.

1.3 Document Convention

- Font Family - Times New Roman
- Line Spacing - 1.5

- Customization for the Primary Header - 18 Font Size
- Customization for the Secondary Header - 14 Font Size
- Customization for the Normal Paragraph - 12
- All Main Heading and Sub Headings are Bold

1.4 Project Overview

The project's overall strategy is to analyze a given image's visual content to anticipate its most likely interpretation, and then construct a phrase that reflects that meaning. Generating automatic descriptions from images requires an understanding of how humans describe images [1]. We looked at many existing natural language description generating models for images and how they work in order to produce natural language descriptions for unknown images. We created a deep learning model for generating natural language descriptions of images based on the current model. Convolutional Neural Networks (CNN) are used to extract image data, and Long Short-Term Memory (LSTM) is applied to generate natural language descriptions from the image features in a deep learning model. For easy integration of this model in various websites and applications an API will be implemented that can be used by the developers of these systems. After a developer POST the image to our API our system will generate a description relevant to the image and a description will be sent back to the developer using our API.

1.5 Scope

1.5.1 In Scope:

- Detection of Objects.
- Generation of Descriptions.

- Conversion of scene into text.
- Achieve accuracy upto 85%
- Creating an API which takes images as input generates descriptions relevant to the image.
- Predicting the most likely nouns, verbs, scenes, and prepositions that make up the sentence.
- Generating descriptions for different objects in the image
- Predicting the relation of things (such as person,animals etc) in images.

1.5.2 **Out of Scope**:

- Generating Descriptions for blurry images.
- Mobile Application.
- Websites
- Desktop Application
- API services on cloud.

2 Overall System Description

2.1 Problem Statement:

2.1.1 Problem

Almost 11 billion photographic images are taken every day. Classifying and organizing these images in such a way that we can locate a specific group of pictures or a unique picture quickly and easily is a challenging task. You can organize your images much more quickly and easily with the automatic generation of image descriptions. The system generates text to describe what is shown in the image. Consequently, users and search engines will be able to see more images, which is beneficial in many ways.

2.1.2 Effect

Automatic Product Tagging in online catalogs, and other areas of business and research. Now with our API development. The developer may easily generate the automatic image description which will be time saving and relevant for their specific domain.Transforming the scene into text will allow us to do this. Developers can use our API to generate image descriptions for their relevant domains.

2.1.3 Who

For developers who want to generate image descriptions automatically by just passing an image to our API. For example a developer integrates the API for an ecommerce website or a social media platform and after the end user uploads the image the API returns a description for the image which the developer can store in their database for other uses like auto filling forms etc.

2.1.4 That

This project (Image Description Generation) generates descriptions of images which are passed to our API which help in making automated applications in their relevant domains.

2.2 Project Objectives

- To create a system to automate the process of generating image descriptions.
- To learn about Deep Learning theoretically and practically as to apply its knowledge to build this project.
- To learn Deep Learning models like CNN, RNN, LSTM, AutoEncoders, BLSTM etc and find their advantages/disadvantages as compared to other Deep Learning Models.

- To apply the theoretical/practical knowledge of Computer Science to build this project as one of the requirements of Final Year Project.
- To learn about REST APIs and how to connect these APIs with a Deep Learning Model.
- To create APIs and to connect with Deep Learning Model to increase the reusability of the project.

2.3 StakeHolders And Effected Groups

2.3.1 Team Members

Description:

Developers, often known as product engineers, are members of a team that use their technical and computer language skills to create software. Experienced designers make sure the product is simple to use and enjoyable to use. User interviews, market research, and product design with end-users in mind are all things they do.

Responsibilities:

- To learn about Deep Learning theory and its practical implementation.
- To build a system to generate automatic image descriptions for a given image.
- To learn about Deep Learning models being used in the projects.
- To build REST APIs for the project
- To write the necessary documentation related to the project.
- To coordinate with Supervisor, Co-Supervisor and PMOs for project updates and guidance related to the project.
- Keep track of the tasks that have been accomplished.
- Completing the tasks within the timeframes set out.

- Ascertaining that the system satisfies the criteria.

2.3.2 Developers

Description:

A computer programmer, often known as a software developer, a programmer, or, more lately, a coder, is someone who develops computer software. A computer programmer might be an expert in a certain field of computing or a generalist who creates code for a variety of applications.

Responsibilities:

- To read and understand the documentation written by the developers of Automatic Image Description Generation System.
- To ensure the images passed to the Automatic Image Description Generation System are not blurry and are taken in sufficient lighting.
- To understand what type of parameters must be passed to the API by reading the documentation as to successfully generate image descriptions.
- To integrate our API within social media platforms, mobile applications, desktop applications etc which use images on their platform to generate image descriptions.

2.3.3 Supervisor / Co-Supervisor

Description:

An IT project supervisor oversees the process of planning, executing, and allocating tasks related to an organization's information technology (IT) ambitions and goals. Because almost every company relies on computer technology, IT project managers can work in a number of sectors.

Responsibilities:

- To understand the requirements of the project and the knowledge required to build it.
- To delegate tasks to team members required to understand and build the project.
- To provide necessary guidance to the team members which will help build,optimize and execute the project.
- To provide necessary guidance required to write documentation which is detailed in nature and easy to understand by designers,project engineers and developers of the project.

2.4 Operating Environment

The operations that should be completed in order to have a Jupyter Notebook. Therefore will require a gpu based computer or if not then operations will be performed on Google Colab. Google Colab is a free cloud service and now it supports free GPU! You can; improve your Python programming language coding skills. develop deep learning applications using popular libraries such as Keras, TensorFlow, PyTorch, and OpenCV

2.5 System Constraints

- Developing a Cnn Architecture, or just any Deep Learning and Computer Vision model in particular, would need a significant amount of computational capability, often provided by Graphics cards.
- If the blurry image or the image contain distortions the model may generate the wrong image description in such circumstances.

3 External Interface Requirements

3.1 Hardware Interfaces

There will be a requirement of a high computation GPU for our project. We might be using the FYP labs. GPU takes significantly less time to apply complex filters to photos and special effects to videos. That means less time waiting, more time creating. You'll also see improved performance when running these apps on higher-resolution screens, including 4K displays.

The developers will require laptops, PC’s etc to use the API in their development projects and to generate the image description automatically.

3.2 Software Interfaces

- Python (and it’s Libraries)
- Jupyter Notebook / Google Colab
- Windows / any other Operating System

4 System Functions / Functional Requirements

4.1 System Functions

A list of system functions is shown below as an example:

- Each function is described in one line in the Functions column.
- Categories indicate that the functionality for a proposed system has a certain status. Below is a list of the possible choices for the Category.

- Each column specifies characteristics of the system, while the Details and Constraints columns specify the conditions under which the characteristic applies.

Function Categories:

Function Category	Meaning
Evident	Should perform, and the user should be cognizant that it is performed.
Hidden	Should perform, but not be visible to users. This is true of many underlying technical services, such as save information in a persistent storage mechanism. Hidden functions are often missed during the requirements gathering process.
Frill	Optional; adding it does not significantly affect cost or other functions.

Ref #	Functions	Category	Details & Boundary Constraints
R1	Use API to take input from developers.	Evident	• Image must not be blurry • Image must be a png or jpg

			• Lightning conditions must be adequate
R2	Validate input images contain subjects.	Hidden	• Detect faces • Detect that there isn't text in image
R3	Preprocess image to feed to neural network.	Hidden	• Filtering Image to minimize computation.

System Attributes/ Nonfunctional Requirements:

Attribute	Details and Boundary Constraints	Category
When to Respond	This algorithm requires a couple of minutes to evaluate each imagery through its trained model.	Optional
Concurrent User Load	To be evaluated	Mandatory
Accuracy	This model generates a description with upto a 82%	Mandatory

	accuracy	

4.2 Use Cases

4.2.1 List Of Actors

- Software Engineers: Developers,Software Engineers project engineers integrate the API in software systems.
- Python Developers: Imports the trained model module to use its functions

4.2.2 List Of Use Cases

Unique ID	Use Cases	Description
U01	Software Developers integrate the API to their application.	Software Developers can integrate the API by generating an API Key. After they have integrated the API to their application they can POST an image to the API to generate a description of the given image.
U02	Generate description for social media image	Social media users upload an image. That image is used to generate description which can be stored in database

U03	Generate image description for python developers.	Python developers will import our model as a module and pass an image to its function. The function will return image description as output.

4.2.3 Use Case Diagram

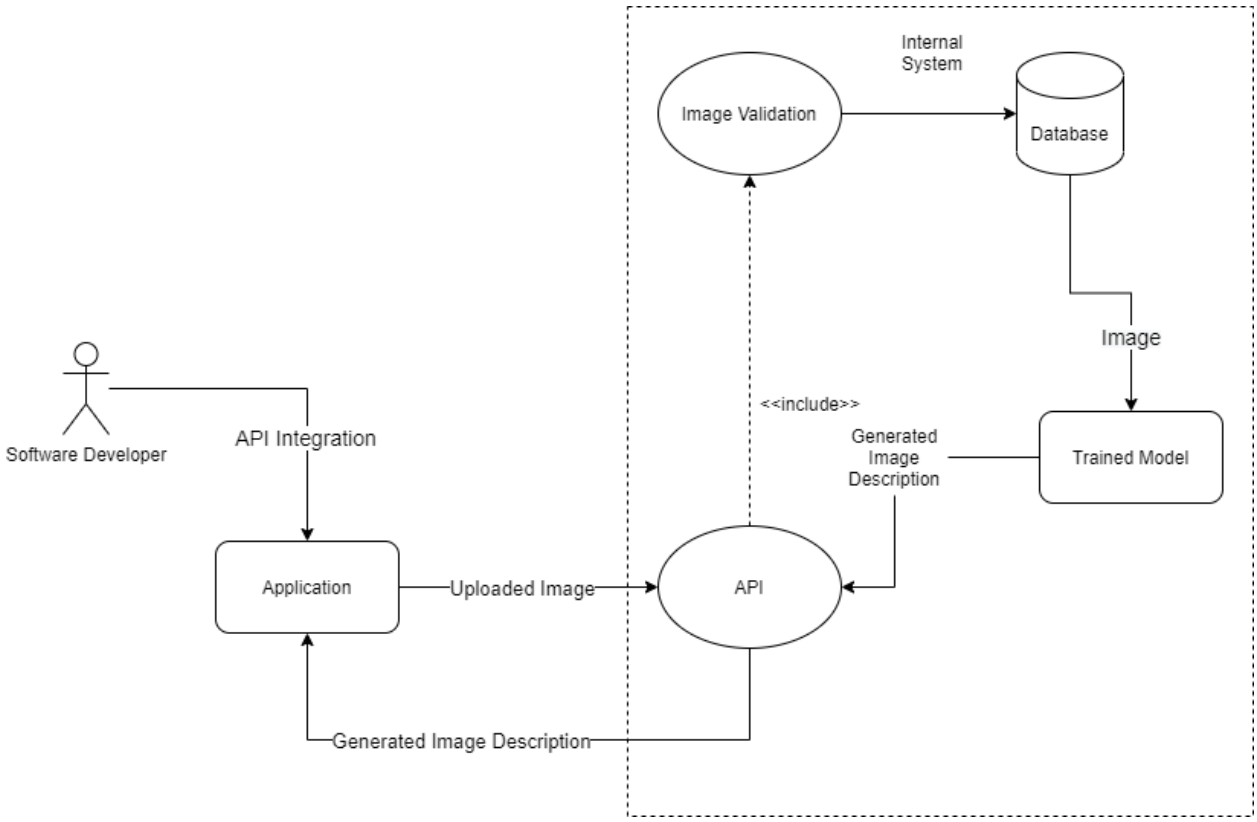


Fig 7. Data Flow between user and model in 1st use case

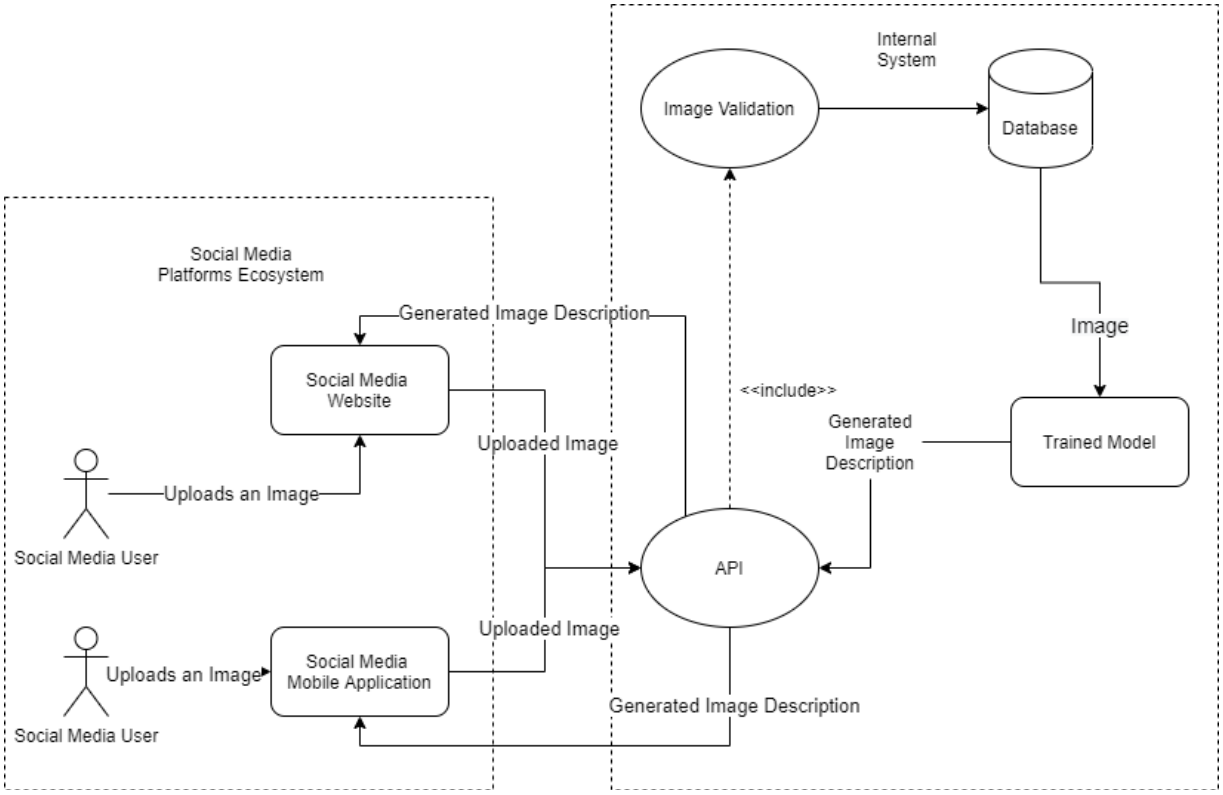


Fig 8. Data Flow between user and model in 2nd use case

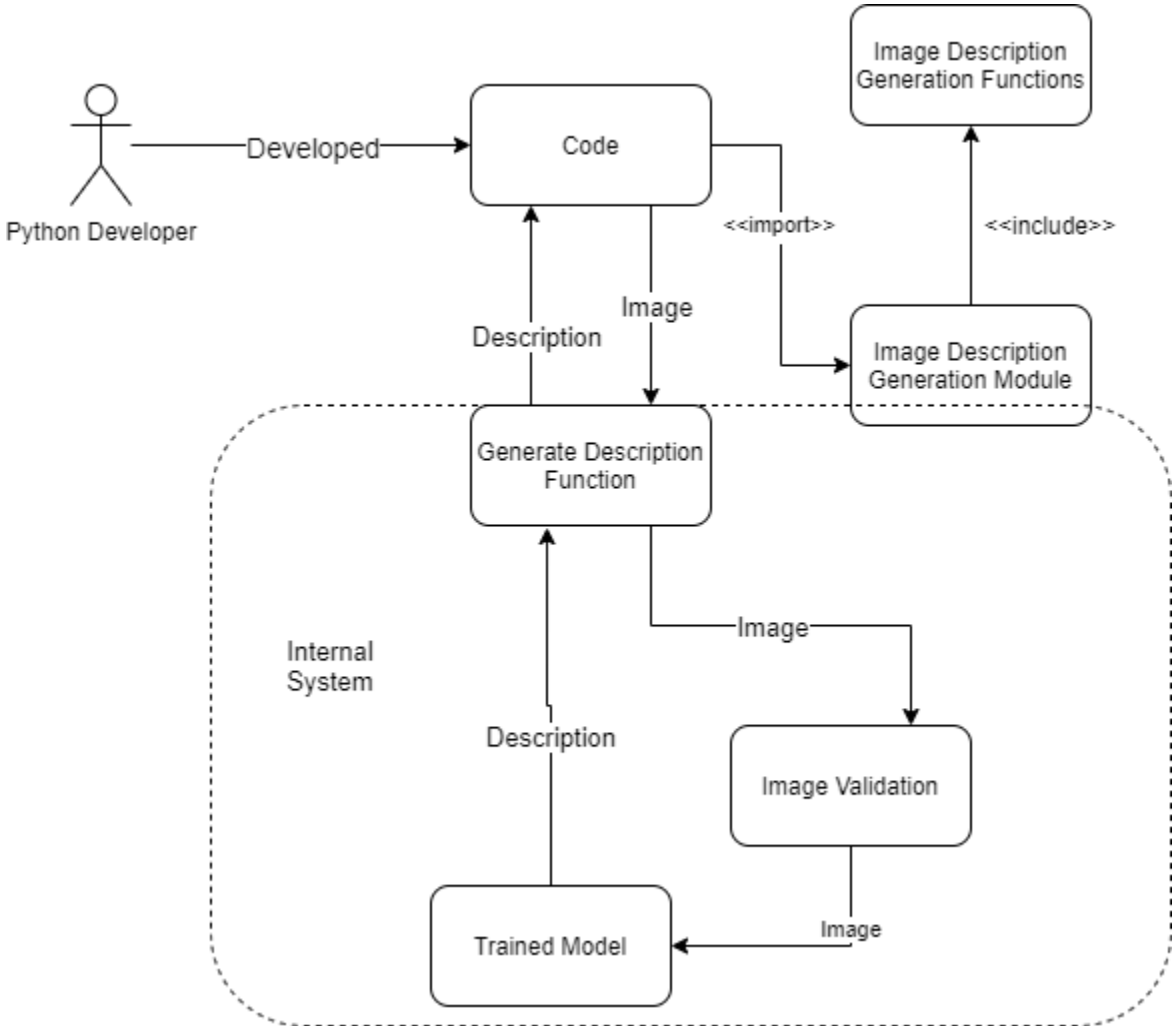


Fig 9. Data Flow between user and model in 3rd use case

4.2.4 Description Of Use Cases

Use Case 01 (U01):

Name: Integration of API (Development Phase of Applications)

Actors: Software Developer, Application, API

Purpose: To reduce the development time.

Description: A software developer develops an application and wants to generate image descriptions as one of the application requirements. The software developer integrates our API with the application and generates descriptions of images which helps in reducing development time.

Typical Course Of Events			
Actor Action		**System Response**	
01	This use case begins when a software developer generates/requests an API key.	02	A string consisting of alphabets,numbers and special characters are returned
03	The software developer integrates the API within the application being created		
04	An image is POST to the API.	05	HTTP code 201 is returned for successful POST.
06	Image is validated by the system. (Check that its not Blurry or Dark)		
07	After validation the image is stored on database		
08	The saved image on database is then sent		

	to the trained model		
09	Trained model generates the description for the image.		
		10	Image is sent back to the API which is given back to the application as output.

Alternative Course of Events:

Step 01: API Key is not generated due to invalid credentials

Step 06: Image is invalidated due to blurriness or low lighting

Use Case 02 (U02):

Name: Social Media Platforms Integration

Actors: Software Developer,Social Media Platform, API

Purpose: To generate data from the uploaded images.

Description: Thousands of images are uploaded on social media platforms online. Our API when integrated within the social media platforms can generate data such as images which can be used to categorize and catalog the images.

Typical Course Of Events

Actor Action		System Response	
01	This use case begins when the software developer generates/requests an API key.	02	A string consisting of alphabets,numbers and special characters are returned
03	The software developer integrates the API within the social media platform.		
04	The social media user uploads the image on social media platforms.		
05	The uploaded image is then POST to the API.	06	HTTP code 201 is returned for successful POST.
07	Image is validated by the system. (Check that its not Blurry or Dark)		
08	After validation the image is stored in database		
09	The image which is stored in the database is then sent to the trained model.		
10	Trained model generates the description for the image.		

		11	Image is sent back to the API which is then stored on the social media platform database.

Alternative Course of Events:

Step 01: API Key is not generated due to invalid credentials

Step 07: Image is invalidated due to blurriness or low lighting

Use Case 03 (U03):

Name: Import module in Python code.

Actors: Python Developer

Purpose: To generate image descriptions as a Python module.

Description: Python developers can import our code as a module inside their Python code. They can use its function to pass an image to its parameters and then generate the description relevant to it.

Typical Course Of Events			
Actor Action		**System Response**	
01	This use case begins when the Python developer imports the module inside their code.		

02	The developer then calls the function by passing the image as a parameter		
03	The image is then checked for validation (Check that its not Blurry or Dark)		
04	The image is then passed to a trained model and an image description is generated which is relevant to the image.	05	The image description is returned back from the function as a string output.

Alternative Course of Events:

Step 03: Image is invalidated due to blurriness or low lighting

5 Design Considerations

5.1 Assumptions and Dependencies

- The model is given high quality pictures.
- According to our assumption the developers will highly use our API for generating automated image descriptions for their project.
- While training, the input of each time step is the real description, and when generating, each word is based on the one it replaced; it may take a long time for a word to get close to the truth once it has not been generated well. Discrepancies like this are known as the "exposure bias" problem. As a result, we presume that the word was formed correctly and is based on previously generated terms.

- To successfully define a picture, it is necessary to recognise items in pictures, recognize scenarios, and properties of the items. The tags are therefore utilised to form semantically relevant phrases in order to generate paragraphs that explain the photos. Our research makes an attempt to complete the goal of producing a sentence that functions as an image description.
- We are assuming that if the developers want to generate more specific image descriptions they will buy our subscription and train the model to generate more specific image descriptions

5.2 Risks and Volatile Areas

Risk mitigation monitoring the Management Plan (RMMM):

Probability:

- High
- Medium
- Low

Effect:

- Catastrophic
- Severe
- Moderate
- Tolerable
- Insignificant

S.No	Types Of Risk	Probability	Effect
1.	**Software Risk:**		
1.1	Security Threats may arise due to Bugs and loopholes in the system.	MEDIUM	SEVERE
1.2	Images fed to the system are not upto the mark. In other words the images are blurred or the image taken in low light	MEDIUM	TOLERABLE
2.	**Hardware Risk:**		
2.1.	The server hosting the API goes down, due to which the developer will not be able to POST the image and will not be able to generate the description of the image	LOW	CATASTROPHIC
2.2	Power losses may result in the system going down due to which the description generation process will not be completed.	MEDIUM	SEVERE
2.3	Natural disasters such as fire,	LOW	TOLERABLE

	cyclone and floods also present risks to IT systems, data and infrastructure		
3.	**Technical Risk:**		
3.1	The descriptions are utilized to build statements that are linguistically relevant. words to produce sentences that explain the visuals. Suppose an image is blurred and the objects are not well detected. It will result in wrong descriptions and then the description paragraph will be wrong as well.	HIGH	SEVERE
3.2	Budget-related risks are financial risks that arise primarily as a result of budget overruns. Always handle the project's financial element as planned, however if the project's financial aspect is	LOW	MODERATE

	mismanaged, budget problems will arise, resulting in budget dangers.		

6 System Architecture

A system architecture is a conceptual model that defines a system's structure, behavior, and other aspects. A formal description and representation of a system arranged in a way that facilitates reasoning about the system's structures and behaviors is known as an architecture description. A system architecture can be made up of system components and subsystems that will collaborate to construct the overall system. There have been attempts to codify languages for describing system architecture, which are referred to as architecture description languages collectively (ADLs).[2]

6.1 *System Level Architecture*

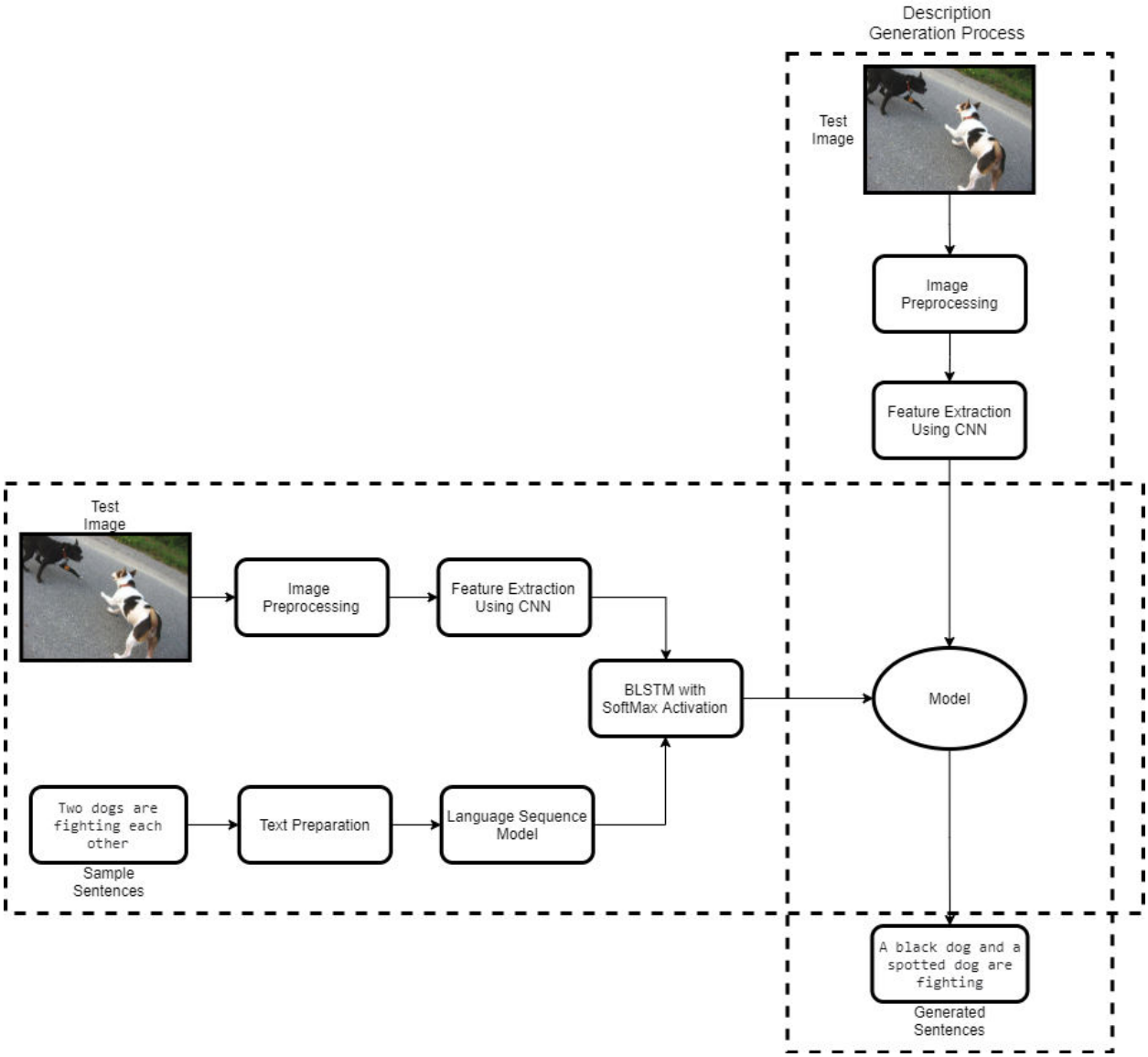


Fig 9. Image preprocessing and NLP process

For picture feature extraction, we used a pre-trained CNN, and for text feature extraction, we used a BLSTM. We utilised a decoder network with the SoftMax activation function to train the model. The training process and the natural language description generating process are the two aspects of the natural language description generation framework. We first pre-processed both the photos and the language descriptions for the training procedure. The features are subsequently extracted using

the CNN model on the preprocessed images. The language sequence model, which is paired with word embedding and the BLSTM model, is used to process the pre-processed phrases. The picture and language features are then concatenated into a single feature vector, which is then fed into the decoder model to train it using the SoftMax activation function. The image is the input to the natural language description creation process, and the sentence is the result. To extract the features, the input image is pre-processed, and the features are extracted using a CNN model. The retrieved image features are fed into a natural language description generating model, which generates the output sentence.

6.2 *Software Architecture*

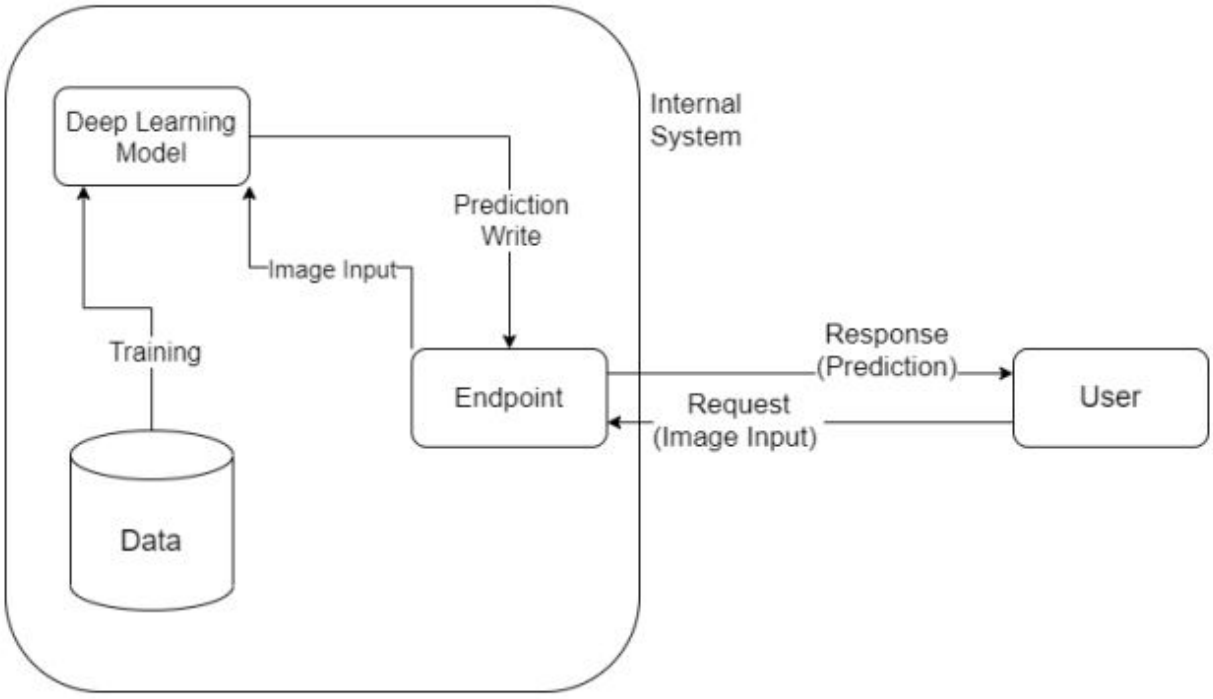


Fig 10. Data flow between user and model

7 Design Strategy

7.1 Future Enhancements

- The subscription will be provided for a training model for detailed domain specific image description generation.
- Credit based system for encouraging more users to use our API.
- Providing cloud based services to developers/end users to easy integration of our API
- The text generated can be converted into voice using a text to speech model.
- Collaboration with different cloud based services to introduce our compute engine for easy integrations.

7.2 System Reuse

Systems are built by combining components that have already been employed in other systems. Software engineering has traditionally concentrated on new development, but it is now understood that in order to produce better software faster and at a reduced cost, we need a design process based on systematic software reuse. Over the last ten years, there has been a significant shift toward reuse-based development.[6]

- Our Project (Image Description Generation) can be easily integrated within applications like (TapTapSee etc) for easily generating image descriptions.
- This project can be integrated within social media platforms like Facebook, Instagram, Twitter etc whenever a user uploads an image. This data can be stored in databases to perform data analysis.

7.3 User Interface Paradigms

Human users interact with a computer, website, or application through the user interface (UI). The objective of good user interface design is to make the user's experience simple and straightforward, requiring the least amount of work on the user's side to get the best possible result. Layers of interaction are used to build user interfaces that appeal to the senses (sight, touch, auditory and more). They include both input and output devices, such as keyboards, mice, trackpads, microphones, touch screens, fingerprint scanners, e-pens, and cameras.

- Our project aims to create an API which focuses on generating image descriptions efficiently and effectively. Our team decided to create a two screen mobile application which will be integrated with our API for testing and usability purposes. The application is a side product and is not the priority of our team.

7.4 Data Management (storage, distribution, persistence)

- Our team in the future aims to create a database which acts as a repository for our model. This repository would contain images required for training and testing. It would contain sample descriptions and predicted outcomes for testing,calculating errors and calculating the accuracy of the model.

8 Detailed System Design

8.1 Flowchart

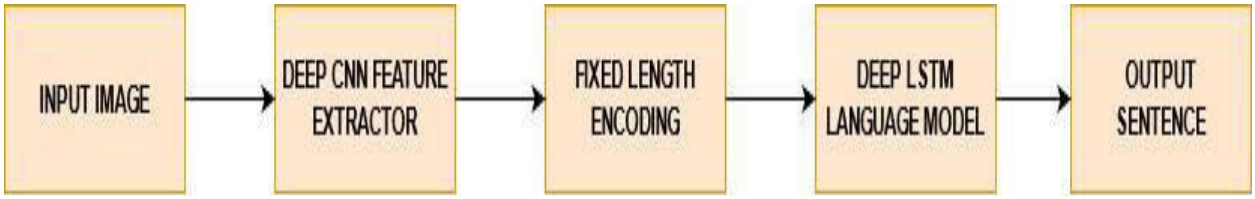


Fig 11. Data flow of neural network

8.2 Software Design

Software design is the process by which an agent produces a specification for a software artefact that is meant to achieve goals and is based on a set of primitive components and restrictions. "All the activity involved in conceiving, framing, implementing, commissioning, and eventually changing complex systems" or "the activity following requirements specification and preceding programming, as in a stylized software engineering process" are two definitions of software design. Typically, software design includes problem-solving and the development of a software solution. This entails both a low-level component and algorithm design as well as a higher-level architectural design.

8.2.1 Sequence Diagram

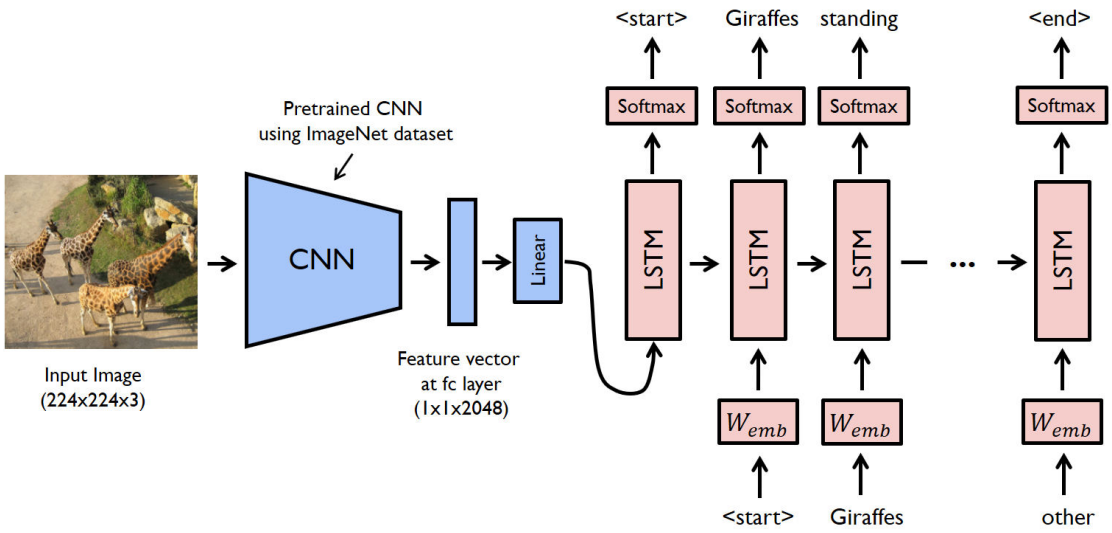


Fig 12. Data flow of neural network

8.2.2 State Diagram

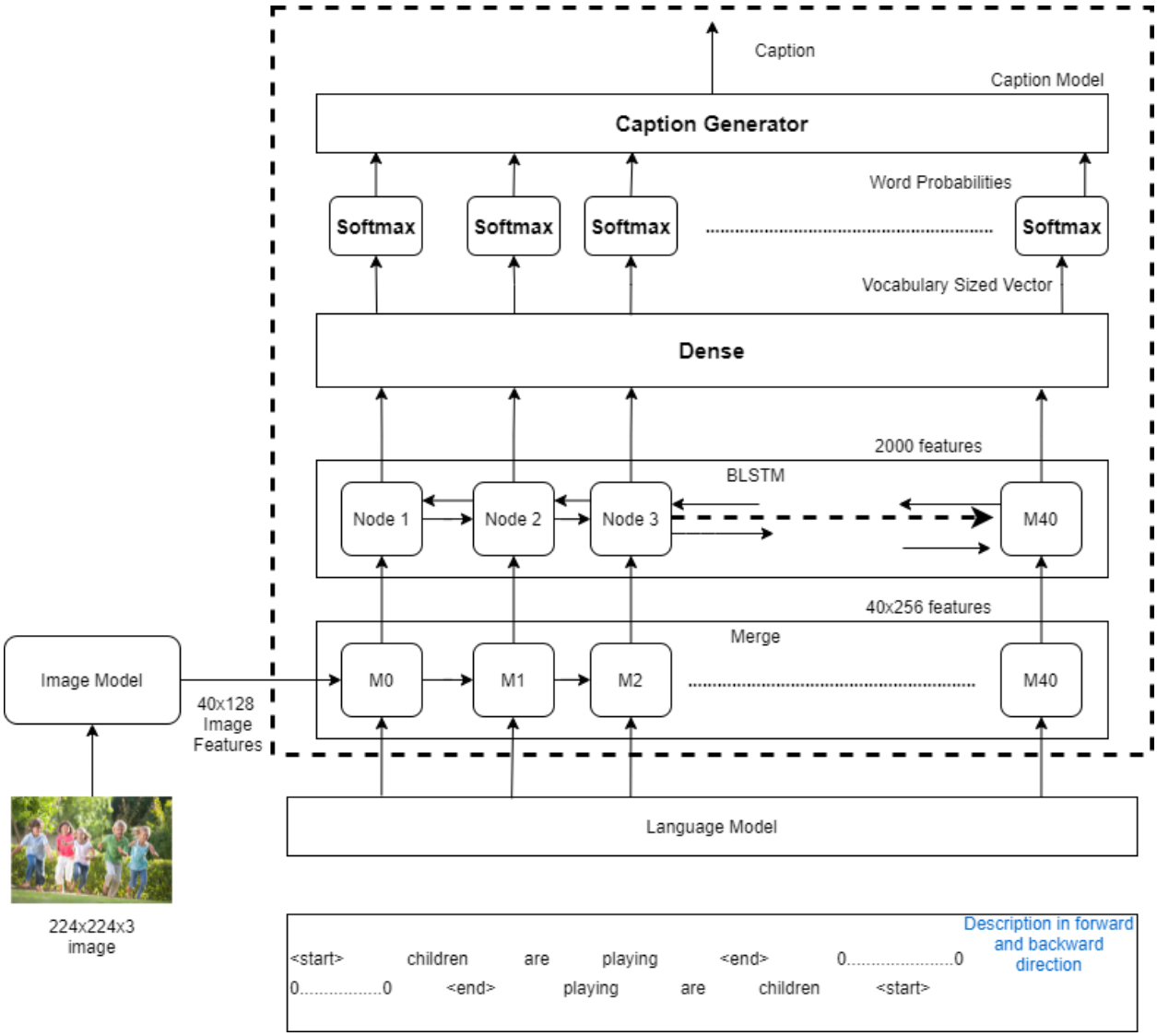


Fig 13. Data flow of neural network

9 GUI Design (For Testing Purposes)

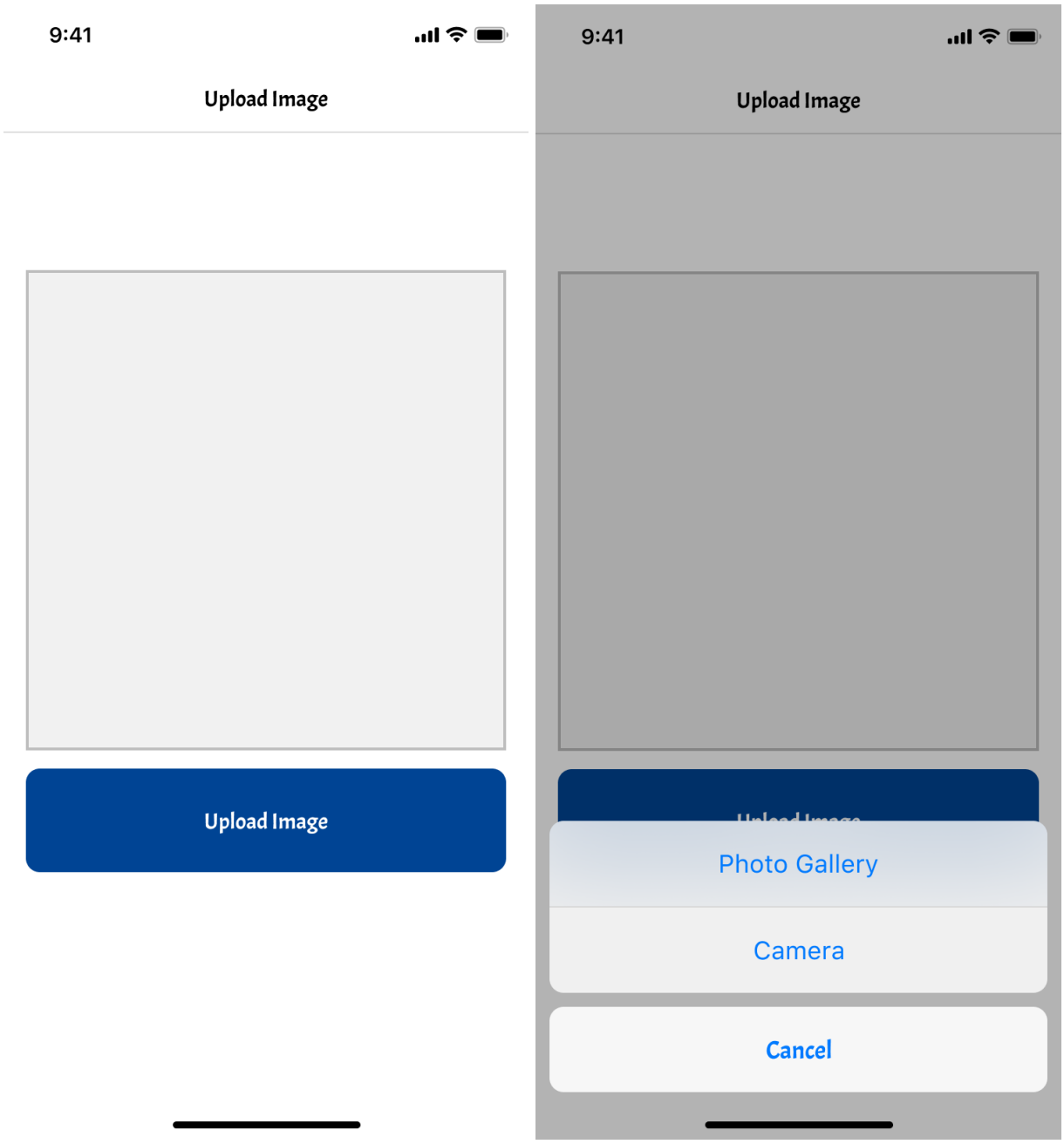


Fig 14. Mockup diagram of future mobile application

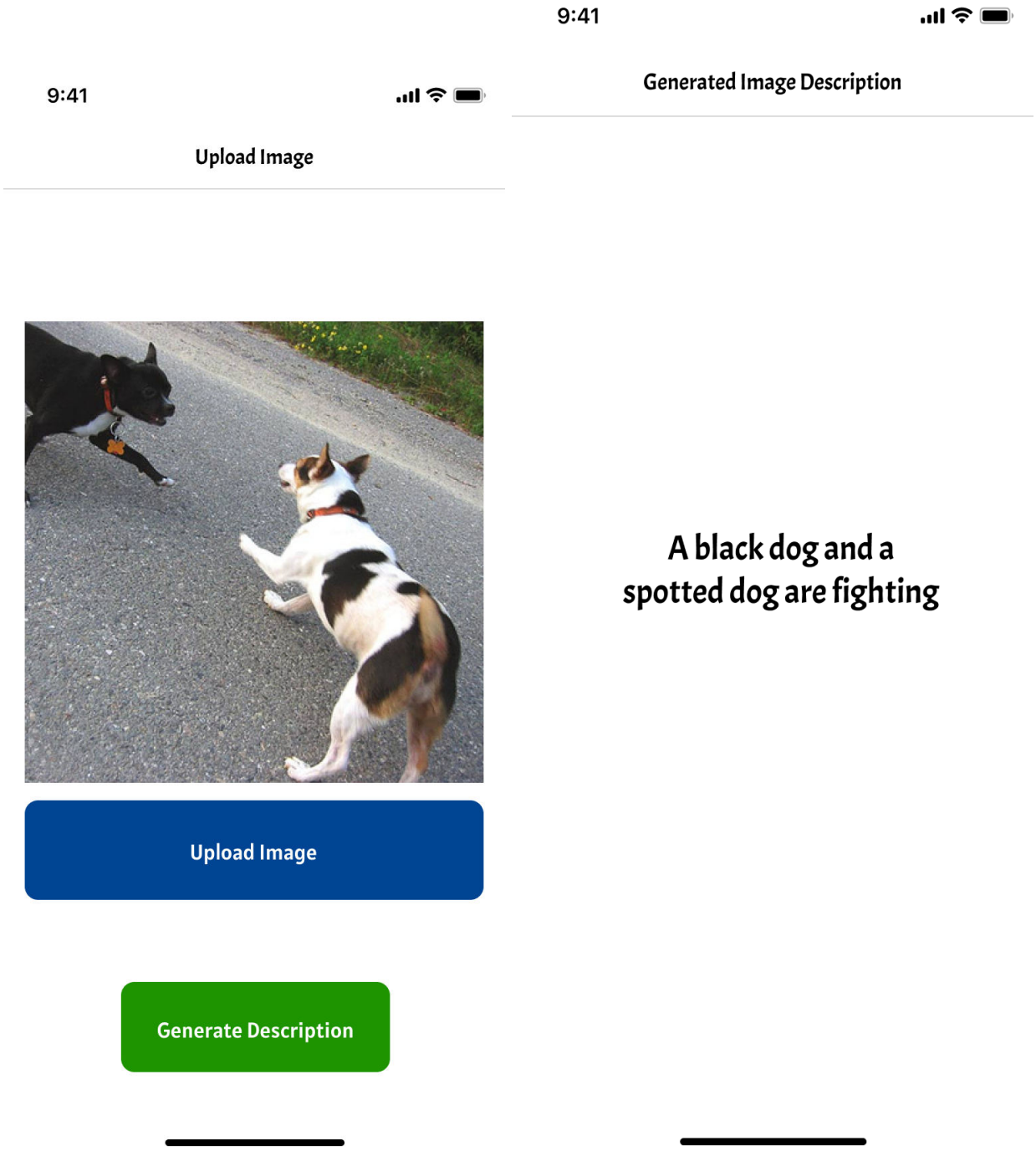


Fig 15. Mockup diagram of future mobile application

7. References

[1]R. Bernardi *et al.*, “Automatic description generation from images: A survey of models, datasets, and evaluation measures,” *Journal of Artificial Intelligence Research*, vol. 55, pp. 409–442, Feb. 2016, doi: 10.1613/jair.4900.

[2]Contributors to Wikimedia projects, “Systems architecture,” *Wikipedia*, 17-Jun-2021. [Online]. Available: https://en.wikipedia.org/wiki/Systems_architecture. [Accessed: 20-Jul-2021]

[3]Jason Brownlee, “How to Develop a Bidirectional LSTM For Sequence Classification in Python with Keras,” *Machine Learning Mastery*, 15-Jun-2017. [Online]. Available: https://machinelearningmastery.com/develop-bidirectional-lstm-sequence-classification-python-keras/. [Accessed: 20-Jul-2021]

[4]J. Brownlee, “A Gentle Introduction to Long Short-Term Memory Networks by the Experts,” *Machine Learning Mastery*, 23-May-2017. [Online]. Available: https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/. [Accessed: 20-Jul-2021]

[5]M. M. Aung, M. San, and P. P. Khaing, “Natural Language Description Generation for Image using Deep Learning Architecture,” *International Journal of Trend in Scientific Research and Development*, vol. Volume-3, no. Issue-5.

[6]CS.CCSU.edu, “CS 530 - Software Engineering class notes,” *CS 530 - Advanced Software Engineering Software Reuse*. [Online]. Available:

https://cs.ccsu.edu/~stan/classes/CS530/Notes18/15-SoftwareReuse.html. [Accessed: 20-Jul-2021]

[7]X. Chen and C. L. Zitnick, “Mind’s eye: A recurrent visual representation for image caption generation,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015 [Online]. Available: http://dx.doi.org/10.1109/cvpr.2015.7298856. [Accessed: 25-Jul-2021]

[8]B. Dai, Y. Zhang, and D. Lin, “Detecting Visual Relationships with Deep Relational Networks,” *arXiv.org*, 11-Apr-2017. [Online]. Available: https://arxiv.org/abs/1704.03114

[9]dshahid380, “Convolutional Neural Network,” *Towards Data Science*, 26-Feb-2019 [Online]. Available: https://towardsdatascience.com/covolutional-neural-network-cb0883dd6529. [Accessed: 25-Jul-2021]

[10]Evergreen, “Automatic Image Captioning,” *Evergreen*, 20-Jul-2020. [Online]. Available: https://evergreen.team/articles/automatic-image-captioning.html. [Accessed: 25-Jul-2021]

[11]S. Ghosh, N. Das, I. Das, and U. Maulik, “Understanding Deep Learning Techniques for Image Segmentation,” *ACM Computing Surveys*, vol. 52, no. 4, pp. 1–35, Sep. 2019, doi: 10.1145/3329784.

[12]H. Wang, Y. Zhang, and X. Yu, “An Overview of Image Caption Generation Methods,” *Computational Intelligence and Neuroscience*, vol. 2020, pp. 1–13, Jan. 2020, doi: 10.1155/2020/3062706.

[13]P. Kinghorn, L. Zhang, and L. Shao, “A hierarchical and regional deep learning architecture for image description generation,” *Pattern Recognition Letters*, vol. 119, pp. 77–85, Mar. 2019, doi: 10.1016/j.patrec.2017.09.013.

[14]P. K. ; V. O. ; A. B. ; T. B. ; Y. Choi, “Collective Generation of Natural Image Descriptions,” *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, vol. Volume 1: Long Papers, no. P12-1038, Jul. 2012.

[15]M. Yatskar, M. Galley, L. Vanderwende, and L. Zettlemoyer, “See No Evil, Say No Evil: Description Generation from Densely Labeled Images,” *Microsoft Research*. [Online]. Available: https://www.microsoft.com/en-us/research/publication/see-no-evil-say-no-evil-description-generation-from-densely-labeled-images/

[16]O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and Tell: A Neural Image Caption Generator,” *arXiv.org*, 17-Nov-2014. [Online]. Available: https://arxiv.org/abs/1411.4555

[17]A. R. Pathak, M. Pandey, and S. Rautaray, “Application of Deep Learning for Object Detection,” *Procedia Computer Science*, vol. 132, pp. 1706–1717, 2018, doi: 10.1016/j.procs.2018.05.144.

[18]R. Lebret, P. O. Pinheiro, and R. Collobert, “Simple Image Description Generator via a Linear Phrase-Based Approach,” *arXiv.org*, 29-Dec-2014. [Online]. Available: https://arxiv.org/abs/1412.8419

[19]S. S R and S. M. Idicula, “Dense Model for Automatic Image Description Generation with Game Theoretic Optimization,” *Information*, vol. 10, no. 11, p. 354, Nov. 2019, doi: 10.3390/info10110354.

[20]T. Guo, J. Dong, H. Li, and Y. Gao, “Simple convolutional neural network on image classification,” in *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)(*, 2017 [Online]. Available: http://dx.doi.org/10.1109/icbda.2017.8078730. [Accessed: 25-Jul-2021]

A3. Other Technical Details

Coding Standards

Standardizing the documentation is crucial as the project uses many different programming languages, paradigmatic frameworks that change the syntax of the programming languages, and accompanying files such as graph descriptions, JSONs, etc.

Structure	Python
Indentation	• It is preferable to use spaces or tabs instead of 4 spaces
Lines	• You can only have 79 characters per line • The top-level functions should have two spaces both above and below them • Use blank lines to group together relevant functions and variables • Create logical divisions using blank lines
Import	• Imports will always be on top. • The next step is to import specific modules.
Strings	• Characters should be enclosed in single quotes • Strings should be enclosed in double quotes • Strings should be broken using + • Instead of concatenating strings, use string formatting

Functions	• Use Camel Case
Assignments	• The variable name should be preceded and followed by a space
Definition indentation or curly braces	• Following the declaration, there should not be a blank line

Project Policy Document

Project Approval:

A few concepts were presented by the students, following which the supervisor reduced Image Description Generation to its most effective and productive elements. Automatic Image Description project defense was held and the created proposal was received by the Jury, which authorized this Final Year Project.

Project Funding

N/A

Risk Management

S.No	Types Of Risk	Probability	Effect
1.	**Software Risk:**		
1.1	Security Threats may arise due to	MEDIUM	SEVERE

	Bugs and loopholes in the system.		
1.2	Images fed to the system are not upto the mark. In other words the images are blurred or the image taken in low light	MEDIUM	TOLERABLE
1.3	The technology stack is becoming outdated.	MEDIUM	SEVERE
2.	**Hardware Risk:**		
2.1.	The server hosting the API goes down, due to which the developer will not be able to POST the image and will not be able to generate the description of the image	LOW	CATASTROPHIC
2.2	Power losses may result in the system going down due to which the description generation process will not be completed.	MEDIUM	SEVERE
2.3	Natural disasters such as fire, cyclone and floods also present	LOW	TOLERABLE

	risks to IT systems, data and infrastructure		
3.	**Technical Risk:**		
3.1	The descriptions are utilized to build statements that are linguistically relevant. words to produce sentences that explain the visuals. Suppose an image is blurred and the objects are not well detected. It will result in wrong descriptions and then the description paragraph will be wrong as well.	HIGH	SEVERE
3.2	Budget-related risks are financial risks that arise primarily as a result of budget overruns. Always handle the project's financial element as planned, however if the project's financial aspect is mismanaged, budget problems	LOW	MODERATE

	will arise, resulting in budget dangers.		
3.3	Key resources leaving project	LOW	CATASTROPHIC

Project Budgeting:

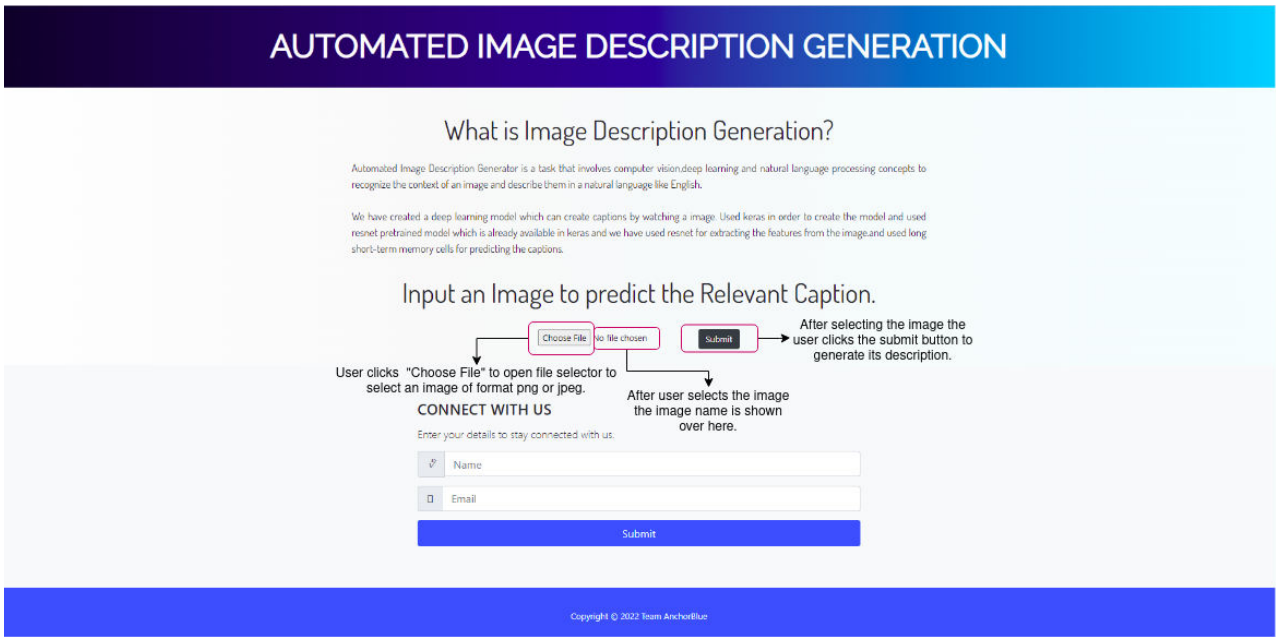
S#	Item	Description	Cost
1	**PC with GPU**	To train statistical models and process images using Deep Learning, a PC must be used to run libraries such as TensorFlow, Pytorch, and Keras on a GPU.	320000 PKR (Cost can be cut down if trained in FYP Labs)
2	**Server Deployment (Load Dependent)**	To put our DL-trained model to work making predictions online.	20000 - 30000 PKR / Month
3	**Maintenance Cost**	Retraining model	15000 PKR / Month +

		using new data manipulation scripts.	Item # 1

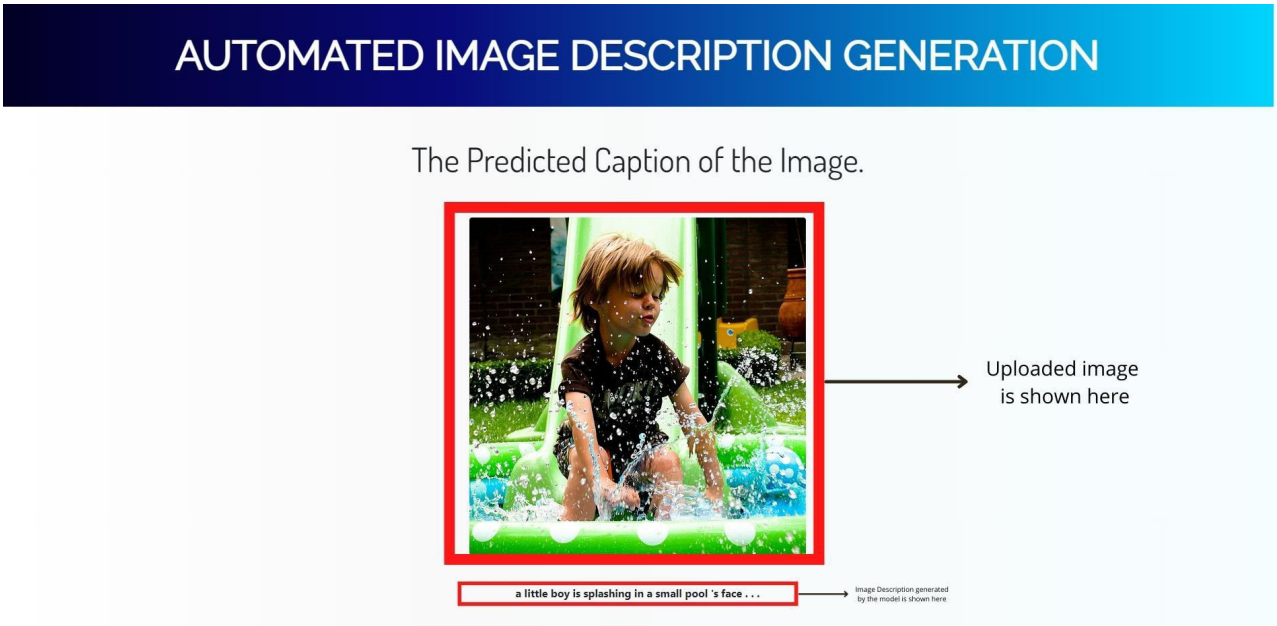
Budgeted Projected Cost - Estimated

Budget including PC with GPU (if no access to FYP lab)	320000 PKR (One time costs) 35000 - 45000 PKR/ Month
Budget without PC with GPU (FYP lab access)	35000 - 45000 PKR/ Month

User Manual



Selecting an image to provide input to the model.



Generated Image Description

A4. Flyer & Poster Design

Fall 2022
Final Year Project

Automatic Image Description Generation

Abstract

The project's overall strategy is to analyze a given image's visual content to anticipate its most likely interpretation and then construct a phrase that reflects that meaning. Generating automatic descriptions from images requires an understanding of how humans describe images. We looked at many existing natural language descriptions generating models for images and how they work in order to produce natural language descriptions for unknown images.

We created a deep learning model for generating natural language descriptions of images based on the current model. Convolutional Neural Networks (CNN) are used to extract image data, and Long Short-Term Memory (LSTM) is applied to generate natural language descriptions from the image features in a deep learning model. To easily use the model to submit images and generate descriptions against it we will create a web application and its back end will be connected to our deep learning model. Users will be able to upload images from the web interface and the back-end will process the image and generate a description against it which will be visible to the user.

Tools

Front-End

Back-End

Flask

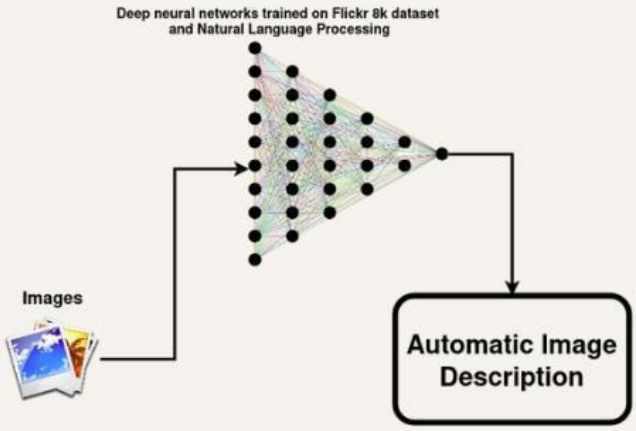
python

TensorFlow

OpenCV

kaggle

Architecture



Project Code

R - 1706

Supervisor

Ms. Tehniat Mirza

Co-Supervisor

Engr. Conrad D' Silva

Team Members

Muhammad Kashif (CS172032)

Muhammad Ali Khan (CS172058)

A5.

Copy of Evaluation Comments

Copy of Evaluation Comments by Supervisor for Project – I End Semester Evaluation

N/A

Copy of Evaluation Comments by Jury for Project – I End Semester Evaluation

Name	Code	Status	Changes	Suggestion
Jury Member 1	**R-1706**	ACCEPT (possibly with or without suggestions)	A literature review is missing recent papers.	
Jury Member 2	**R-1706**	ACCEPT (possibly with or without suggestions)	Please update your problem statement regarding blind or visually impaired persons. You will create the captions in text so it is not going to help blind or visually impaired persons unless you convert this text into sound/voice.	
Jury Member 3	**R-1706**	ACCEPT (possibly with or without suggestions)		
Jury Member 4	**R-1706**	ACCEPT (possibly with or without suggestions)		

Copy of Evaluation Comments by Supervisor for Project – II End Semester Evaluation

N/A

Copy of Evaluation Comments by Jury for Project – II End Semester Evaluation

N/A

A6. Meetings' Minutes

Meeting Number	Topic	Tasks	Date	comments of Supervisor
1	Idea Discussion			
2	Explained Gradient Descent, Loss Function, Working of a DL Model and Back Propagation	Code a simple neural network	Tuesday March 9, 2021	completed
3	Explained Sequential/Parallel Models, activation function and how to evaluate a model	Code a simple neural network		
4	FYP Idea Discussion	Research on FYP Idea	Thursday March 25, 2021	
5	Explained Pooling, Filtering and CNN	Implement CNN		
6	Discussion on CNN Project			
7	Object Detection Explanation			
8	Discussion on FYP II Documentation			
9	What are Pre Trained Networks (RESNET-50)	Implement RESNET - 50		
10	What is LSTM	Implement		

		LSTM		
11	Proofreading Documentation			
12	Panaflex Designing Discussion			

A7 . Project Progress

DHA Suffa University

Department of Computer Science

Final Year Project

Automatic Image Description Generation

(R-1706)

Research Documentation

Submitted by

Muhammad Kashif (CS172032)

Muhammad Ali Khan (CS172058)

Supervisor(s)

Ms. Tehniat Mirza

Co - Supervisor(s)

Mr. Conrad Walter D' Silva

Table of Contents

2. Minutes of Meeting

Meeting Number	Topic	Tasks	Date	comments of Supervisor
1	Idea Discussion			
2	Explained Gradient Descent, Loss Function, Working of a DL Model and Back Propagation	Code a simple neural network	Tuesday March 9, 2021	completed
3	Explained Sequential/Parallel Models, activation function and how to evaluate a model	Code a simple neural network		
4	FYP Idea Discussion	Research on FYP Idea	Thursday March 25, 2021	
5	Explained Pooling, Filtering and CNN	Implement CNN		
6	Discussion on CNN Project			
7	Object Detection Explanation			
8	Discussion on FYP II Documentation			
9	What are Pre Trained Networks (RESNET-50)	Implement RESNET - 50		
10	What is LSTM	Implement		

		LSTM		
11	Proofreading Documentation			
12	Panaflex Designing Discussion			

3. Document Finalization

- Research documentation complete
- Research Paper has been written completely and is created by up-to-date resources.

4. Project Plan Vs Actual

4.1. Hurdles

- Deep learning largely relies on computing power to test hypotheses, practise using existing deep neural networks, and draw conclusions. Because our idea is based on current advances in deep learning, we need a significant number of test benches.
- There were some problems when we were trying to train our model in FYP Lab and it needed some time to train on a large

number of images. Despite the powerful PCs in FYP lab we could only train our model during university hours.

4.2. Work done so far

- Due to the above circumstances , we concentrated our FYP-1 on:
    - Examining current journal articles, their flaws, and thought about how can we be able to solve them.
    - Learning about Deep Learning concepts required to build an Image Description Generation model.

A8 . Research Paper

DHA Suffa University

Department of Computer Science

Final Year Project

Automatic Image Description Generation

(R-1706)

Research Documentation

Submitted by

Muhammad Kashif (CS172032)

Muhammad Ali Khan (CS172058)

Supervisor(s)

Ms. Tehniat Mirza

Co - Supervisor(s)

Mr. Conrad Walter D' Silva

Automatic Image Description Generation

Muhammad Ali Khan
Computer Science Department
DHA Suffa University
Karachi, Pakistan
khanali198361@gmail.com

Muhammad Kashif
Computer Science Department
DHA Suffa University
Karachi, Pakistan
kashif.mmn1999@gmail.com

***Abstract*—The project's overall strategy is to analyze a given image's visual content to anticipate its most likely interpretation, and then construct a phrase that reflects that meaning. Generating automatic descriptions from images requires an understanding of how humans describe images. We looked at many existing natural language descriptions generating models for images and how they work in order to produce natural language descriptions for unknown images. We created a deep learning model for generating natural language descriptions of images based on the current model. Convolutional Neural Networks (CNN) are used to extract image data, and Long Bidirectional Short-Term Memory (LSTM) is applied to generate natural language descriptions from the image features in a deep learning model. To easily use the model to submit images and generate descriptions against it we will create a web application and its back end will be connected to our deep learning model. Users will be able to upload images from the web interface and the back-end will process the image and generate a description against it which will be visible to the user.**

***Index Terms*—deep-learning, image, description, RESNET-50, LSTM,CNN**

I. Introduction

With the rapid growth of digitization, there are a great number of photographs, accompanied with a lot of relevant writings . Automatic image description generation has lately gained substantial scholarly interest. The objective of automatic image descriptions is to generate properly formed English sentences to describe the content of an image automatically, which is of great impact in various domains such as virtual assistants, image indexing, recommendation in editing applications, and the help of the disabled. Although it is an easy process for a human to describe a picture, it becomes exceedingly tough for a computer to accomplish such a task. Image description generation does not only need to recognize the things present in an image but also capture how these objects connected to each other and their properties as well as the activities participating in. Moreover, the semantic information should be conveyed in a natural language, which needs a language model to be constructed based on the visual understanding.

Much research effort has been devoted to automatic image description generation, and it can be categorized into template-based image description generation, retrieval-based image description generation, and novel image description generation. Template-based image description generation first detects the objects/attributes/actions and then fills the blanks slots in a fixed template. Retrieval-based approaches first find the visually similar images with their descriptions from the training data-set, and then the image description is selected from similar images with descriptions. These methods are able to generate syntactically correct descriptions but are unable to generate image-specific and semantically correct descriptions. Differently, the novel image description generation approaches are to analyze the visual content of the image and then to generate image descriptions from the visual content using a language model. Compared to the first two categories, novel description generation can generate new descriptions for a given image that are semantically more accurate than previous approaches. Most of the works in this category rely on machine learning and deep learning, which is also the approach adopted in this paper. One common framework used in this category is the encoder-decoder framework for image description generation.

II. Background

A. Convolution Neural Network

A convolution neural network is an architecture of a neural network that has been used for image categorization and object recognition with considerable success. The most frequent design consists of three primary procedures that are performed numerous times. Firstly, a convolution is performed to the picture, and the weights of these convolution kernels or filters are not predefined but are learnt from the data. Next, the outcome of these convolutions is compressed into smaller matrices with the assistance of pooling. Pooling is the technique of collecting results from areas in a way similar to convolution. Pooling may be done in many methods such as max-pooling that preserves the maximum element in the pooling zone, min pooling or average pooling.

The size of the convolution filter, the pooling region, the number of filters to employ and the number of convolution and dense layers are hyper parameters and need to be tuned by exploring what works best for a problem. The number of dropout layers, that randomly enable a node to transmit its result farther, the frequency of dropout and their position is also crucial. These are the settings that need to be set appropriately for optimum outcomes.

There are several designs of Convolutions Neural networks that have been employed for diverse applications, and there is

current research in the subject. Three models were employed for the purpose of this study - VGG16, VGG19, and ResNet50

The VGG architecture was initially described in (Simonyan and Zisserman 2014) for image classification on the ImageNet data set. The design is noted for its simplicity as it employs only a 3 × 3 convolutional layers stacked on top of each other. The VGG19 model has more of these layers layered compared to the VGG16. The volume is lowered in each phase with the aid of max pooling. These convolutional layers are followed by two fully connected layers that flatten the output and a soft max layer that forecasts the probability of each of the items the model was trained for.

The Residual Networks (ResNet) described originally in (He et al. 2016) comprises of layered Residual blocks. The residual block contains a skip connection which makes it easier to understand the internal representation and so layering several residual blocks lets us go more in-depth and prevent the decreasing gradient problem.

B. LSTM Architecture

LSTM networks are an extension of recurrent neural networks (RNNs) principally created to handle instances when RNNs fail. Talking about RNN, it is a network that operates on the present input by taking into consideration the previous output (feedback) and keeping in its memory for a brief period of time (short-term memory) (short-term memory). Out of its multiple applications, the most popular ones are in the domains of voice processing, non- Markovian control, and music creation. Nevertheless, there are downsides to RNNs.

First, it fails to keep information for a longer period of time. At times, a reference to particular information saved quite a long time ago is necessary to anticipate the present result. But RNNs are totally incapable of handling such "long-term dependencies". Second, there is no finer control over which element of the context has to be brought forward and how much of the past needs to be 'forgotten'. Thus, Long Short-Term Memory (LSTM) was brought into the scene. It has been thus designed that the vanishing gradient problem is nearly totally removed, but the training model is kept unmodified. Long-time delays in particular issues are bridged utilising LSTMs where they also manage noise, distributed representations, and continuous data. LSTMs offer us with a vast variety of parameters like as learning rates, and input and output biases. Hence, no need for precise adjustments. Figure 1 displays the architecture of LSTM.

The figure 1 shows the LSTM Architecture which is explained in detail below.

1) Figure-A depicts what a simple LSTM network looks like. Only one layer of LSTM between an input and output layer has been presented here.
2) Figure-B illustrates Deep LSTM which incorporates a number of LSTM layers in between the input and output. The advantage is that the input values provided to the network not only transit through numerous LSTM layers but also propagate over time inside one LSTM cell. Hence, parameters are well spread throughout various layers. This leads in a complete process of inputs in each time step.
3) Figure-C illustrates LSTM with the Recurrent Projection layer where the recurrent connections are taken from the projection layer to the LSTM layer input. This architecture was developed to lower the high learning computational complexity (O(N)) for each time step) of the typical LSTM RNN.
4) Figure-D illustrates Deep LSTM with a Recurrent Projection Layer comprised of many LSTM layers where each layer has its own projection layer. The additional depth is particularly handy in the scenario where the memory capacity is too huge. Having additional depth inhibits over fitting in models as the inputs to the network need to travel via multiple nonlinear equations.

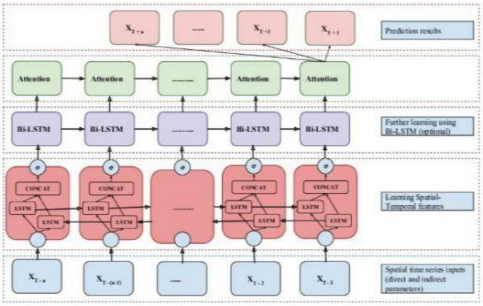


Fig. 1. LSTM Architecture

To conclusion, long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture utilised in the field of deep learning. Unlike normal feed forward neural networks, LSTM contains feedback connections. A standard LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values across arbitrary time periods and the three gates manage the flow of information into and out of the cell.

C. Residual Networks (ResNet-50)

ResNet is a sophisticated backbone model that is used extremely frequently in various computer vision tasks. It has been noticed that as we go putting on more layers to the neural network, it gets tough to train them and the accuracy begins saturating and eventually declines likewise. Res-Net is a network that is 50 layers thick. You may load a pretrained version of the network trained on more than a million photos from the ImageNet database. The pretrained network can categorise photos into 1000 item categories, such as keyboard, mouse, pencil, and numerous animals. As a consequence, the network has learnt rich feature representations for a wide range of pictures. The network has an image input size of 224-by-224. Residual Network is an unique sort of deep learning network that was designed to handle challenging issues by adding some additional layers in deep learning networks which

results in enhanced accuracy and performance. The process of training a model entails giving an algorithm (that is, the learning algorithm) with training data to learn from. The term model refers to the model artefact that is formed by the training process. The training data must contain the right response, which is known as a target or target characteristic. The learning algorithm detects patterns in the training data that translate the input data characteristics to the goal (the response that is to be predicted), and it returns a model that captures these patterns. This model may be used to make predictions on new data for which the target is unknown. ResNet Architecure Image data augmentation is a technique that may be used to artificially enlarge the size of a training dataset by producing changed copies of pictures in the dataset.

Training deep learning neural network models on additional data can result in more skillful models and the augmentation approaches can produce variants of the pictures that can increase the capacity of the fit models to generalise what they have learnt to new images. In our instance the quantity of photos is fewer. So, data augmentation will aid in increasing the amount of photos and give the images to the model in batches. The photographs in the batches are not repeated hence it also helps in preventing the overfitting of model. Keras ImageDataGenerator (only rotation range,zoom range and horizontal flip was used)generate training data from the numpy arrays in batches and process them with their labels. Training data was additionally scrambled during training, whereas validation data was utilised to determine the validation accuracy and validation loss during training. After data augmentation is accomplished, both training generator and testing generator is developed for generating the model. Keras- ImageDataGenerator provides a method named flow from directory which is used to bulid the generator. It accepts the target directory where the data is saved, a batch size which determines the number of photos provided each batch, class mode which specifies the kind of labels used and additionally shuffle and seed attributes can be enabled. ResNet-101 is a convolutional neural network that is 101 layers deep. Once all the pre-processing for the model is done, ResNet model is constructed using tensrflow keras package. Weights for the model is pretrained and utilised. Now more layers are added to produce much more precise findings. Since a highly deep network is employed, there is a very high risk for mistake. Now a completely linked layer is introduced. The fully connected layer encompasses flatten, dense and dropout layers. First layer is the flatten layer which is used to transfer the output of convolutional layer into the dense layer. Once it is sent into the dense layer, it transfers the scores of the convolutional layer into the right labels which triggers the function. In this situation SoftMax function is employed which turns the scores into probabilities that total to 1.

At the end, dropout layer is applied which would be utilised to prevent the neural network from overfitting during the training. This inhibits the network to learn duplicate information. Batch normalisation is a recently discovered approach which seeks to correctly initialising neural networks by deliberately

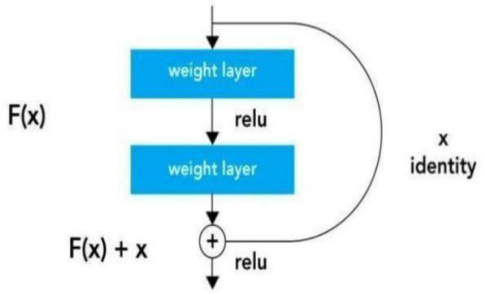


Fig. 2. ResNet Model

forcing the activations throughout a network to assume a unit Gaussian distribution at the beginning of the training. In practise, we put the Batch Normalization layers directly after thick layers. Networks that utilise Batch Normalization are substantially more resilient to faulty initialisation, because normalisation greatly limits the possibility of small number of outlying inputs to over-influence the training. It also tends to lessen the overfitting. Additionally, it may be understood as conducting pre-processing at each layer, but incorporated into the network itself. In our example, batch normalisation is used in the model to prevent arbitrarily big weights in the intermediate layers. As the batch normalisation normalises, the intermediate layers, aid to converge nicely. In our scenario early halting is used by monitoring the val loss till best fit model is obtained. Once, the best fit model is achieved, the keras ModelCheckpoint is used to auto store the best fit training model and the weights for our training data. Now the model being finalised. It is built using Adam as the optimizer which is an upgrade to RMSProp optimizer in which the running average of both the gradients and their magnitude is employed. Adam is considerably better than RMSProp because Adadelta learns too rapidly.

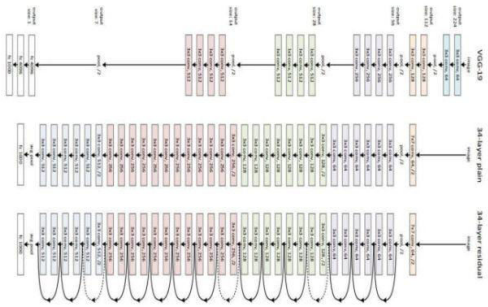


Fig. 3. ResNet Architecture

Now the model is trained and verified using Keras for generator which conducts the training of the generator. Once the best fit is attained the training will finish and the best model

and its weight are saved. Here ResNet comes into rescue and helps fix this problem. In this network, a technique called skip connections is employed as illustrated in the picture 3. The skip connection bypasses training from a few levels and links directly to the output. The advantage of introducing this type of skip connection is that if any layer damage the performance of architecture then it will be skipped by regularisation. Without employing this skip link, the input 'x' is multiplied by the weights of the layer followed by adding a bias term. This term runs via the activation function, f() and we receive our output as H(x) (x). This network employs a 34-layer simple network design inspired by VGG- 19 in which afterwards the shortcut connection is introduced. These shortcut connections subsequently change the design into residual network. Figure 3 displays the architecture of Resnet.

III. ResNet-LSTM Architecture for Image Description Generation

In this project, the picture description generator utilising ResNet and LSTM (Long Short Term Memory) is built. The image features will be taken from a ResNet model and trained on the image-net dataset and then fed the features into the LSTM model which will be responsible for creating the image descriptions. For the picture description generator, Flickr 8K dataset will be utilised. There are other other huge datasets like Flickr 30K and MSCOCO dataset but it might take weeks only to train a network thus we will be utilising a modest Flickr8k dataset. Figure 4 and 5 demonstrates how descriptions are created using Resnet and LSTM.

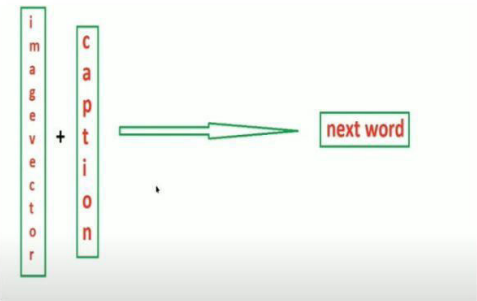


Fig. 4. How next word is predicted

A. Dataset

The FLickr8k data set is a collection of 8000 photos with five description each, aggregated in one location, and ready to use it for the evaluation of image description generation and image querying methodologies. The researchers suggest that effective outcomes may be attained while many descriptions are used with each image, to train the model.

B. Image Pre-processing

To produce image features we utilise pretrained weights of CNNs trained on ImageNet image classification dataset

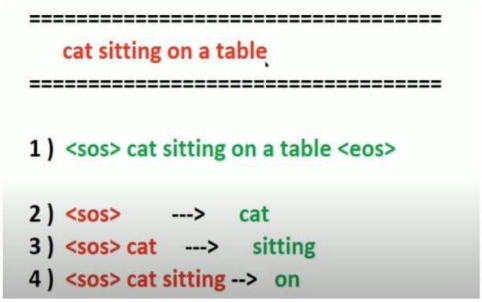


Fig. 5. How next word is predicted

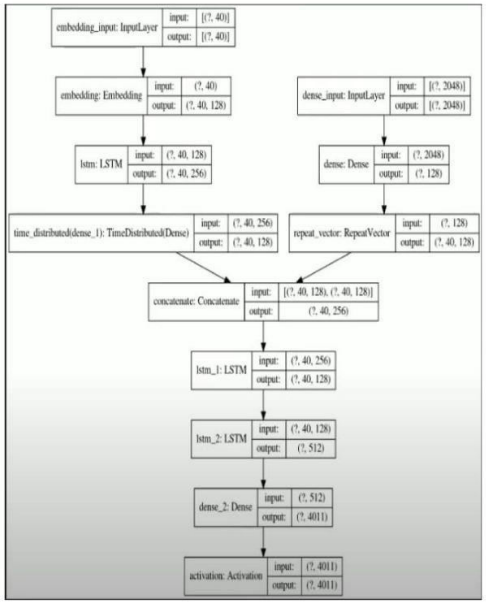


Fig. 6. Model Structure

(VGG16, VGG19, and ResNet50) and remove the last dense layers from the model. We preprocess pictures and produce image features using the by conducting a forward pass on the image on using these weights and store these features to a file.

C. Model

This model that has been employed for such research comprises of two independent input streams, one for the image features, and the other for the preprocessed input descriptions. The visual characteristics are transferred via a fully linked (dense) layer to produce a representation in a separate dimension. The input captions are routed via an embedding layer. These two input streams are then merged and provided as inputs to an LSTM layer. The picture is supplied as the starting

state to the LSTM while the caption embeddings are passed as the input to the LSTM. The architecture is seen in figure 6.

D. Training

The model is trained in DHA Suffa University FYP Labs . The DSU FYP Labs equipment features one NVIDIA 2070TI, with Ryzen 5 CPUs and 32 gigabytes of RAM. Training the model takes roughly 30 minutes in the DSU FYP Lab with total of 1500 photos. To train the model, for each picture and each of the input captions that were created during preprocessing, we feed the image features via the dense layer, and the pre-processed input captions to the embedding layer. We then utilise the picture as the starting state to the LSTM, along with the caption which is supplied as the input to the LSTM. The model generates a projected caption,

REFERENCES

[1] Donahue, J., Anne Hendricks, L., Guadarrama, M., Venugopalan, S., Saenko, K., and Darrell, T. (2015), 'Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 2625-2634.

[2] Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X. and Mitchell, M. (2015) 'Language models for image captioning: The quirks and what works', pp- 155.809.

[3] Farhadi, M. Hejrati, M. A. Sadeghi et al., (2010), 'Every picture tells a story: generating sentences from images, in Compute r Vision', pp. 15–29.

[4] Hodosh.M , P. Young, and J. Hockenmaier, (2013), 'Framing image description as a ranking task: Data, models and evaluation metrics, Journal of Artificial Intelligence Research', vol. 47, pp. 853–899.

[5] Ordonez. V, G. Kulkarni, and T. L. Berg, (2010), 'Im2Text: Describing images using 1 million captioned photographs, Advances in Neural Information Processing Systems', pp. 1143–1

[6] Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. J. Choi, (2011), 'Composing simple image descriptions using web-scale n-grams, in Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics', pp. 212- 1153

[7] Karpathy, A., and Fei-Fei, L.(2013), 'Deep visual-semantic alignments for generating image d

[8] Karpathy, A., and Fei-Fei, L.(2013), 'Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 3128-3137
escriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 3128-3137

Turnitin Originality Report

Processed on: 03-Mar-2022 12:06 PKT
ID: 1775384799
Word Count: 8625
Submitted: 1

FYP Report By Muhammad Ali Khan

Similarity Index	Similarity by Source	
16%	Internet Sources:	8%
	Publications:	5%
	Student Papers:	14%

1% match (student papers from 24-May-2021)
Submitted to University of Bradford on 2021-05-24

1% match (Internet from 24-Jan-2022)
https://www.techtarget.com/searchenterpriseai/definition/deep-learning-deep-neural-network

1% match (student papers from 26-Jul-2021)
Submitted to Westcliff University on 2021-07-26

1% match (student papers from 28-Oct-2021)
Submitted to National School of Business Management NSBM, Sri Lanka on 2021-10-28

1% match (student papers from 28-Nov-2021)
Submitted to Salalah College of Technology on 2021-11-28

1% match (student papers from 18-Jan-2022)
Submitted to Asia Pacific University College of Technology and Innovation (UCTI) on 2022-01-18

1% match (student papers from 09-Aug-2021)
Submitted to Queen Mary and Westfield College on 2021-08-09

1% match (student papers from 01-Jun-2021)
Submitted to Georgia State University on 2021-06-01

1% match (student papers from 04-Jun-2021)
Submitted to Nexford University on 2021-06-04

1% match (publications)
Alexandru-Ion Marinescu. "A Musical Similarity Metric based on Symbolic Aggregate Approximation", 2020 International Conference on Software, Telecommunications and Computer Networks (SoftCOM), 2020

1% match (student papers from 04-Jan-2022)
Submitted to ritindia on 2022-01-04

1% match (student papers from 31-Oct-2019)
Submitted to University of Computer Studies on 2019-10-31

< 1% match (student papers from 06-Dec-2021)
Submitted to Westcliff University on 2021-12-06

< 1% match (student papers from 26-Aug-2021)
Submitted to National School of Business Management NSBM, Sri Lanka on 2021-08-26

< 1% match (student papers from 05-Jul-2021)
Submitted to Tower Hamlets College on 2021-07-05

< 1% match (Internet from 12-Nov-2021)
https://www.mvorganizing.org/what-are-the-advantages-and-disadvantages-of-using-gpu/#:~:text=Why%20choose%20GPUs%20for%20Deep,GPU's%20memory%20bandw

< 1% match (student papers from 23-Jun-2020)
Submitted to K. J. Somaiya College of Engineering Vidyavihar, Mumbai on 2020-06-23

< 1% match (student papers from 04-Dec-2021)
Submitted to The Hong Kong Polytechnic University on 2021-12-04

< 1% match (Internet from 14-Feb-2021)
https://www.slideshare.net/nbalneg/business-risk-and-continuity-planning-109631969

< 1% match (student papers from 25-Feb-2022)
Submitted to Coventry University on 2022-02-25

< 1% match (student papers from 06-Oct-2021)
Submitted to Zambia Centre for Accountancy Studies on 2021-10-06

< 1% match (student papers from 02-Jun-2021)
Submitted to Middle East College of Information Technology on 2021-06-02

< 1% match (Internet from 29-Oct-2021)
https://link.springer.com/article/10.1007/s00530-021-00758-w?code=12b3477b-174f-41f9-b2c0-b3145bf83c80&error=cookies_not_supported

< 1% match (student papers from 29-Sep-2021)
Submitted to Midlands State University on 2021-09-29

< 1% match (Internet from 27-Sep-2021)
http://www.ijres.org/papers/Volume-9/Issue-8/Series-1/L09086471.pdf

< 1% match (student papers from 16-Nov-2017)
Submitted to Nazarbayev University on 2017-11-16

< 1% match (Internet from 29-Jan-2022)
https://journals.pen2print.org/index.php/ijr/article/download/20159/19752

< 1% match (publications)
"Machine Learning Paradigms", Springer Science and Business Media LLC, 2020

< 1% match (student papers from 14-Sep-2021)
Submitted to University of the Free State on 2021-09-14

< 1% match ()
Comput Intell Neurosci. 2020 Jan 9; 2020:3062706

< 1% match (Internet from 27-Oct-2021)
https://doku.pub/documents/loy-j-neural-network-projects-with-python-2019-el9re6ok1xly

< 1% match (publications)
Viet-Khoa Vo-Ho, Quoc-An Luong, Duy-Tam Nguyen, Mai-Khiem Tran, Minh-Triet Tran. "A Smart System for Text-Lifelog Generation from Wearable Cameras in Smart Environment Using Concept-Augmented Image Captioning with Modified Beam Search Strategy", Applied Sciences, 2019

< 1% match (Internet from 01-Feb-2022)
https://deepai.org/publication/diverse-and-styled-image-captioning-using-svd-based-

mixture-of-recurrent-experts

< 1% match (Internet from 24-Nov-2021)
https://www.hindawi.com/journals/cin/2020/3062706/

< 1% match ()
Bernardi, Raffaella, Cakici, Ruket et al. "Automatic Description Generation from Images:A Survey of Models, Datasets, and Evaluation Measures", 'AI Access Foundation', 2016

< 1% match (student papers from 09-Jul-2020)
Submitted to Universidad de Alcalá on 2020-07-09

< 1% match (student papers from 15-Jan-2022)
Submitted to University of Teesside on 2022-01-15

< 1% match (Internet from 08-Mar-2019)
https://victorzhou.com/

< 1% match (Internet from 24-Feb-2020)
https://www.analyticsvidhya.com/blog/

< 1% match (Internet from 23-Jan-2022)
http://www.arxiv-vanity.com/papers/1705.08759/

< 1% match (Internet from 04-Jan-2022)
https://sftlift.shopqueensknow.com/inpout32-dll-serial-port-examples-of-resumes/

< 1% match (publications)
Sreela S R, Sumam Mary Idicula. "Dense Model for Automatic Image Description Generation with Game Theoretic Optimization", Information, 2019

< 1% match (publications)
Sreela Sreekumaran Pillai Remadevi Amma, Sumam Mary Idicula. "Keyword Driven Image Description Generation System", Advances in Science, Technology and Engineering Systems Journal, 2020

< 1% match (Internet from 29-Nov-2021)
https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/

< 1% match (publications)
Sepp Hochreiter, Jürgen Schmidhuber. "Long Short-Term Memory", Neural Computation, 1997

< 1% match ()
Rjoob, Khaled, Bond, RR et al. "Reliable Deep Learning–Based Detection of Misplaced Chest Electrodes During Electrocardiogram Recording: Algorithm Development and Validation", 'JMIR Publications Inc.', 2021

< 1% match (Internet from 04-Mar-2021)
https://www.coursehero.com/file/73682012/CS561-RNNpdf/

< 1% match (Internet from 03-Nov-2021)
https://www.semanticscholar.org/paper/Image-Captioning-with-Semantic-Attention-You-Jin/bf55591e09b58ea9ce8d66110d6d3000ee804bdd

< 1% match (publications)
"Advanced Informatics for Computing Research", Springer Science and Business Media LLC, 2021