Deep Learning approach for Image Caption Generation in Nepali Language

M.Sc. Project Mid Defense

Presented By:

PITAMBAR KHANAL(078MSIISE013)

Supervisor:

Er. Bibek Ropakheti

Department of Electronics and Computer Engineering

Institute of Engineering, Thapathali Campus

Presentation Outline

- Motivation
- Background
- Problem Statement
- Objectives of Project
- Scope of Project
- Originality of Project
- Potential Applications
- Literature Review
- Methodology
- Results
- Discussion and Analysis
- Remaining Tasks
- Tentative Timeline (Gantt Chart)
- References

Motivation

- A crucial task that pertains to both the field of computer vision and natural language processing.
- It's not novel to create image captions, but it can be difficult to translate them into Nepali Language.

Fig-1: Motivational Image

Background[1]

- Whenever an image appears in front of us, the brain is capable of labeling it based on its background and features.
- How can a machine process an image and label it with a highly relevant and accurate caption?
- Image caption generaton is a task that involves computer vision and natural language processing.

Background[2]

- The concept of CNN,RNN and LSTM have been used to generate image captioning in existing research work.
- The major datasets used were Google and Kaggle Dataset.
- Translating the caption of images into Nepali Language is new.
- The ability of a machine to mimic human abilities to describe images with Nepali caption

Problem Statement

- Generating the captions in the English new but what about the case of Nepali Language?
- Domain specific datasets are not available
- Researchers worked on various data-set to generate the captions for any images but have they done in Nepali data-set?

Objectives of Project

1. To develop a model that is able to generate captions for image in Nepali language.

Scope of Project

- This project work will be helpful for researcher in the field of generating image captions in Nepali Language
- Can be used to enhance the practical applications of computer vision and NLP concepts
- The model will not work on cross language text descriptions of languages
- The model can’t generate captions from videos
- The model will be able only to generate captions of limited domain images

Originality of Project

- Image captioning in English language has already been done
- **The length of caption will be improved by this project compared to previous work**
- A large annotated data-set of Nepali image captions will be compiled to facilitate further research in this domain

Potential Applications

The potential applications areas of this project work are listed as follows:

1. Advertising
2. News and Journalism
3. Social Media
4. E-Learning
5. Computer Vision and Natural Language Processing

Literature Review[2]

Paper	Year	Authors	Methodology	Results	Weakness	Strengths
Image Caption Generating Deep Learning Model	2021	Aishwarya Maroju , Sneha Sri Doma , Lahari Chandarlapati	ResNet-LSTM Model	ResNet is having better performance and accuracy compared to traditional CNN ,VGG	They were unable to generate exact caption	Provides the detail concepts of ResNet 50 and LSTM
Image Captioning Generator System With Caption To Speech Conversion Mechanism	2021	Shubham Rawale, Megha Ghotkar, Krishna Sonavane, Paras Surve, Shraddha Khonde, Deepali Patil	1. RNN and LSTM 2. gTTS engine (text to speech conversion)	The model provides results in the form of speech	This work is not sufficient for multiple domain	1. The use of g TTS engine to translate image captions into speech 2. The use of evaluation metric BLEU

Literature Review[2]

Paper	Year	Authors	Methodology	Results	Weakness	Strengths
Image Caption Generation Using A Deep Architecture	2020	Ansar Hani, Najiba Tagougui, Monji Kherallah	CNN as encoder and attention module(RNN) as decoder	The obtained results is better than previous results	This model was unable to use the BELU metirc	Provides the concept of attention based encoder-decoder model
Image Captioning Based on Deep Neural Networks	2018	Shuang Liu , Liang Bai,Yanli Hu and Haoran Wang	1. CNN-RNN based framework 2. CNN-CNN based framework 3. Reinforcement based framework	The CNN-RNN and Reinforcement based methods can get better performance than the CNN-CNN based framework	1. Inconsistent objects during training and testing 2.Cross-Language text description of images	1. Provides comparative results 2. Provides relevant datasets

Literature Review[3]

Paper	Year	Authors	Methodology	Results	Weakness	Strengths
An Automatic Approach for Translating Simple Images into Text Descriptions and Speech for Visually Impaired People	2015	Mrunmaye e Patil, Ramesh Kagalkar	Canny edge detection algorithm	Provides 100% accurate results for horse and Dinosaurs as compared to human and nature scene images	This work is not sufficient to make dynamic system	It can handle a variety of input photos and convert them to text and audio.

Methodology[1]

The steps to build model are arranged as shown in block diagram :

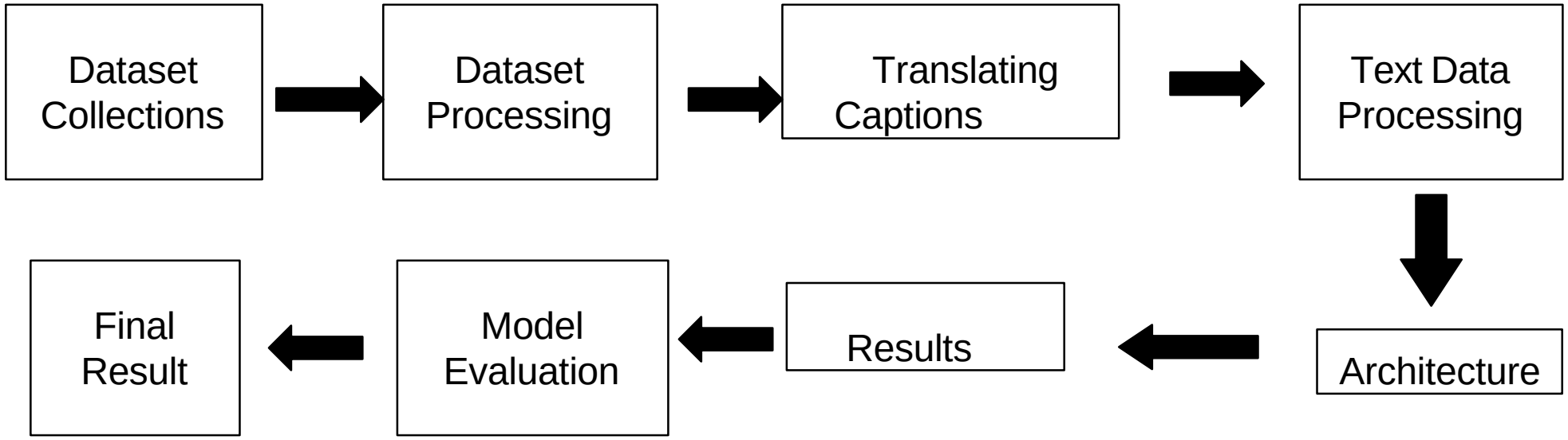


Fig -2: Working Flow Diagram

Methodology[2]

1. Data Collection:

- The primary resources of datasets are google or kaggle dataset
- The used dataset to achieve project's objective is Flickr 8k.

Images	Captions
6k(Training)	6*5=3000
1k for testing and validation	

Methodology[3]

2. Dataset Processing

Preprocess images

- Resize images and normalize pixel values

Preprocess captions

❖ Clean the text by removing punctuation, converting to lowercase, and removing non-alphanumeric characters.
❖ Tokenize the captions into words or sub words.

Translation into Nepali

- Manual and automated translation

Tokenization

- Tokenize the translated Nepali captions into words or sub words

Methodology[4]

3. Text Data Processing

- Tokenizing the captions has been done to simplify the representation of the captions as numerical sequences
- Building a vocabulary involves creating a list of all the unique words in the captions
- Encoding the captions involves converting the captions into numerical sequences that can be input to a model

Methodology[5]

4. Architecture/Model Selection

1. Encoder-Decoder Architecture
2. Attention-based Architecture
3. Generative Adversarial Network (GAN)
4. Transformer-based Architecture

Among all these proposed architecture, **transformer based architecture** is finalized.

Methodology[6]

Transformer-based Architecture

- Ability to handle complex dependencies in both image understanding and natural language generation

Image Feature Extraction

- CNN is used to extract features from the image

Transformer Architecture

Encoder: The CNN-extracted features are input to this encoder to capture more complex representations

Decoder: The transformer decoder generates the caption

Methodology[7]

Attention Mechanism

- To learn which parts of the image are relevant for generating the next word in the caption

Sequence Generation in Nepali

- The decoder is trained on a dataset of image-caption pairs where the captions are in Nepali

Methodology[8]

The steps involved include:

1. Tokenization
2. Embedding
3. Positional encoding
4. Training

Caption Generation

- The image is passed through CNN to extract features.
- Features are fed into transformer decoder
- Greedy decoding can be used to improve quality of generated captions

Methodology[7]

5. Model Evaluation

The model evaluating parameters are listed as follows:

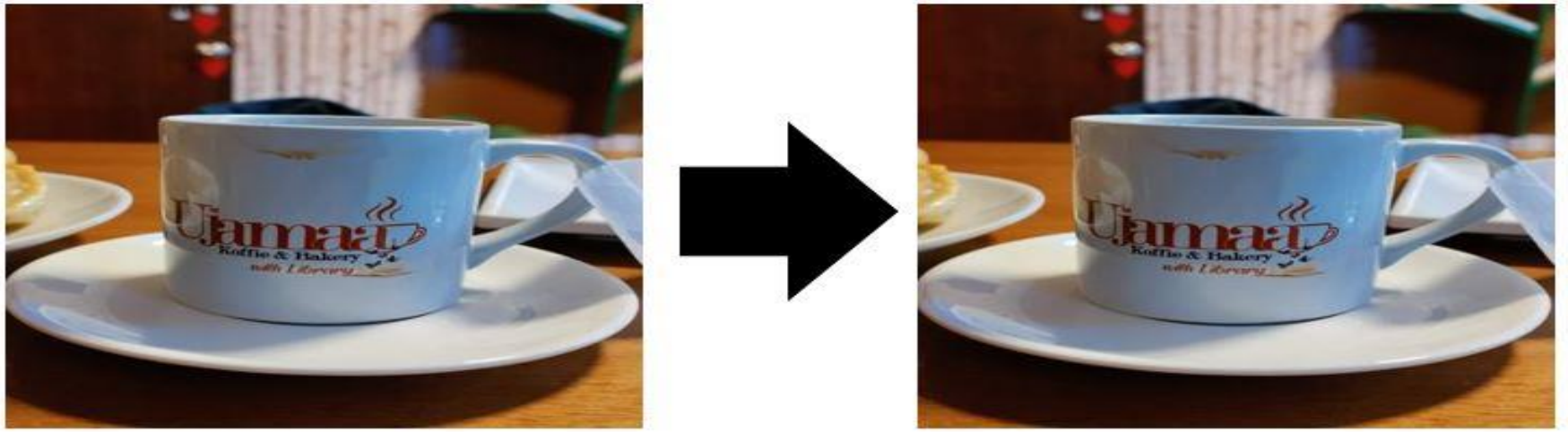
- BLEU((Bilingual Evaluation Understudy)) Score
  - The degree of overlap between the generated captions and a set of reference captions.
- CIDEr(Consensus-based Image Description Evaluation) Score
  - The CIDEr score is a metric that measures the degree of overlap between the generated captions and a set of reference captions.

Methodology[8]

- ROUGE(Recall-Oriented Understudy for Gisting Evaluation) Score
  - The degree of overlapping by focusing on the recall of important words and phrases.
- METEOR(Metric for Evaluation of Translation with Explicit ORdering) score
  - The degree of overlap between the generated captions and a set of reference captions.
  - By taking into account the alignment between the words in the generated and reference captions
- Human Evaluation
  - This can provide a more subjective and holistic evaluation of the model's performance.

Results[1]

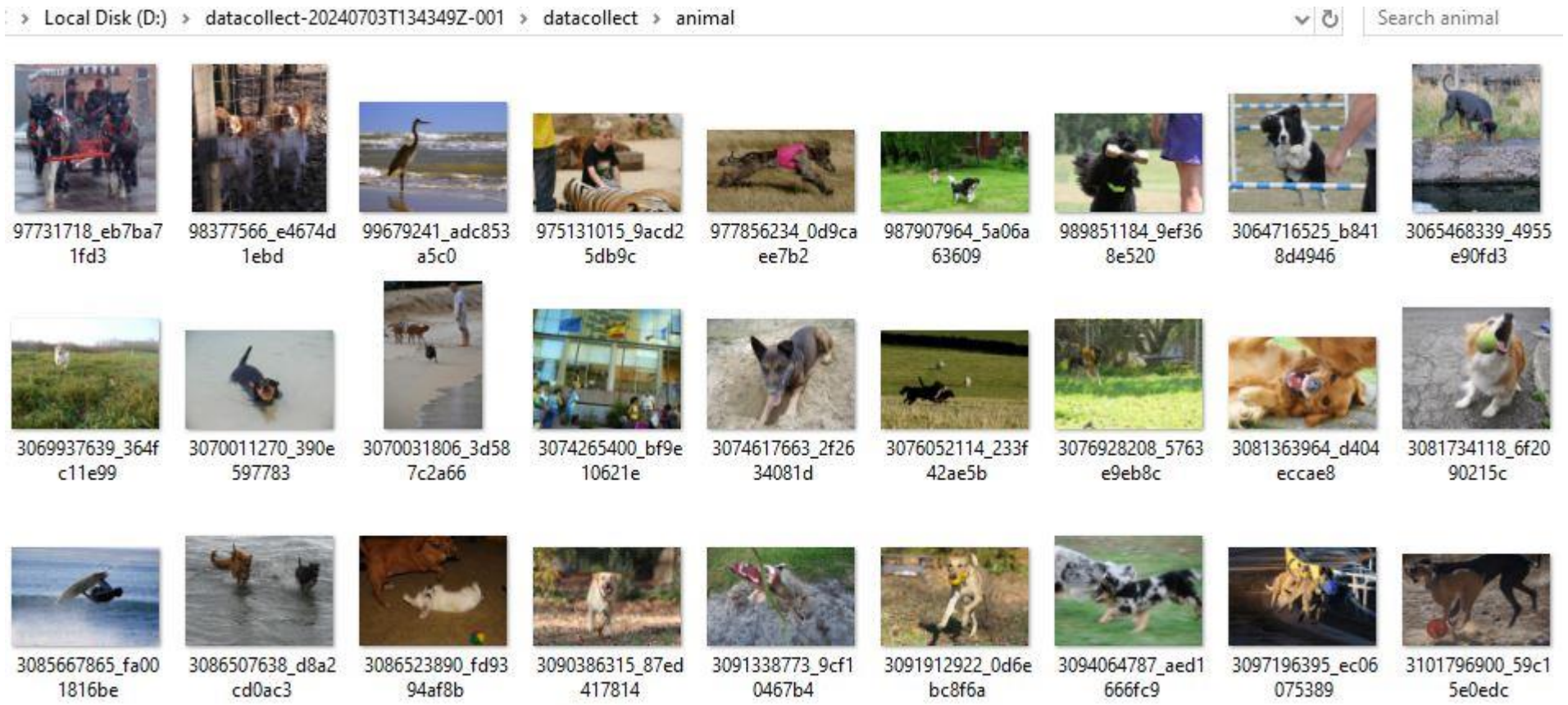
- The model will be capable to generate captions for input images.
- Accurate image captions will be generated.



Input Image without Caption

Excepted Output Image with Caption
True Caption: प्लेटमाथी सेतो कप छ।
Expected Caption: प्लेटमा सेतो कप।

Results[2]


Local Disk (D:) › datacollect-20240703T134349Z-001 › datacollect › animal
Search animal
97731718_eb7ba71fd3
98377566_e4674d1ebd
99679241_adc853a5c0
975131015_9acd25db9c
977856234_0d9caee7b2
987907964_5a06a63609
989851184_9ef368e520
3064716525_b8418d4946
3065468339_4955e90fd3
3069937639_364fc11e99
3070011270_390e597783
3070031806_3d587c2a66
3074265400_bf9e10621e
3074617663_2f2634081d
3076052114_233f42ae5b
3076928208_5763e9eb8c
3081363964_d404eccae8
3081734118_6f2090215c
3085667865_fa001816be
3086507638_d8a2cd0ac3
3086523890_fd9394af8b
3090386315_87ed417814
3091338773_9cf10467b4
3091912922_0d6ebc8f6a
3094064787_aed1666fc9
3097196395_ec06075389
3101796900_59c15e0edc

Results[3]

animal	7/3/2024 7:32 PM	File folder
object	7/3/2024 7:32 PM	File folder
outdoor	7/3/2024 7:32 PM	File folder
person	7/3/2024 7:32 PM	File folder
sports	7/3/2024 7:32 PM	File folder


3132006797_0482
2b5866


3132832452_c354
c6396c


3135504530_0f41
30d8f8


3135826945_f7c7
41e5b7 - Copy

Results[4]

```
PATH = "/kaggle/input/standford-paragraph-nepali-dataset/stanford_images/"
# PATH = "/kaggle/working/NumpyFiles/"
all_img_name_vector = []

for annot in data["Image_name"]:
    full_image_path = PATH + str(annot)+'.jpg'

    all_img_name_vector.append(full_image_path)
all_img_name_vector[:10]
```

```
21]: ['/kaggle/input/standford-paragraph-nepali-dataset/stanford_images/2356347.jpg',
 '/kaggle/input/standford-paragraph-nepali-dataset/stanford_images/2317429.jpg',
 '/kaggle/input/standford-paragraph-nepali-dataset/stanford_images/2414610.jpg',
 '/kaggle/input/standford-paragraph-nepali-dataset/stanford_images/2365091.jpg',
 '/kaggle/input/standford-paragraph-nepali-dataset/stanford_images/2383120.jpg',
 '/kaggle/input/standford-paragraph-nepali-dataset/stanford_images/2333990.jpg',
 '/kaggle/input/standford-paragraph-nepali-dataset/stanford_images/2388203.jpg',
 '/kaggle/input/standford-paragraph-nepali-dataset/stanford_images/2338364.jpg',
 '/kaggle/input/standford-paragraph-nepali-dataset/stanford_images/2410301.jpg',
 '/kaggle/input/standford-paragraph-nepali-dataset/stanford_images/2404368.jpg']
```

transformer_captioning+mlfl... Draft saved

File Edit View Run Add-ons Help

Run All Code Draft Session (13m)

```
# Specify the path to the Stanford Paragraph Captioning dataset file
dir_Stanford_text = '/kaggle/input/standford-paragraph-nepali-dataset/Caption_final.csv'

# Read the Stanford Paragraph Captioning dataset
data = pd.read_csv(dir_Stanford_text, delimiter=',')

# Remove unwanted rows if needed
# data = data[data['train'] == 'TRUE']  # Adjust as needed
print(data.columns)

# Reorganize the columns
data = data[['Image_name','Paragraph']]

# Display the first few rows of the DataFrame
data.head()
```

```
Index(['Image_name', 'Paragraph', 'train', 'test', 'val'], dtype='object')
```

[6]:

	Image_name	Paragraph
0	2356347	यसको अगाडि झ्यालहरूमा बारहरू भएको ठूलो भवन। भव...
1	2317429	एउटा सेतो गोलो प्लेट एउटा टेबलमा छ जसमा प्लास...
2	2414610	नीलो टेनिस पोशाकमा एउटी महिला हरियो टेनिस कोर...

Results[5]

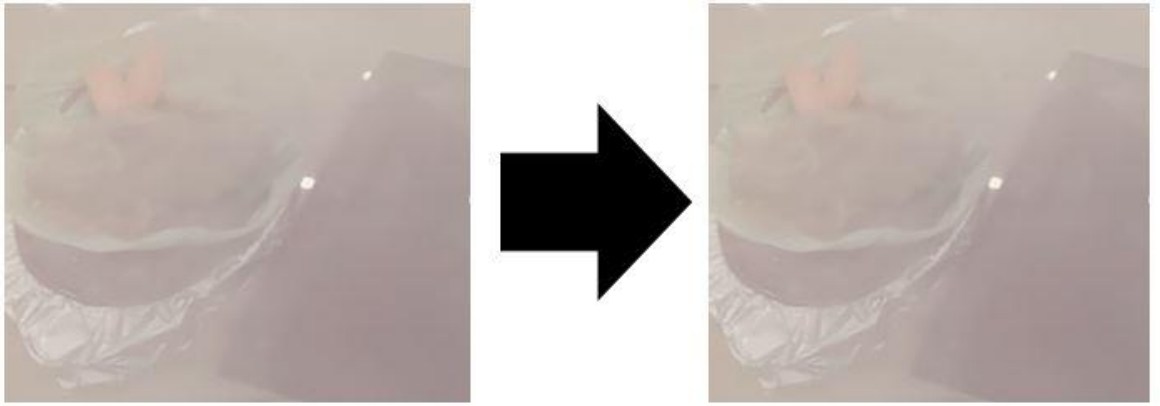
	Image	English Captions	Translated Nepali Captions
2	1000268201_693b08cb0e.jpg	A child in a pink dress is climbing up a set of stairs in an entry way .	गुलाबी पोशाकमा एक बच्चा एक प्रविष्टि तरीकामा सीढीको सेट चढ्दै छ।
3	1000268201_693b08cb0e.jpg	A girl going into a wooden building .	एक केटी काठको भवनमा जाँदै।
4	1000268201_693b08cb0e.jpg	A little girl climbing into a wooden playhouse .	एक सानो केटी काठको प्लेहाउस मा चढाई।
5	1000268201_693b08cb0e.jpg	A little girl climbing the stairs to her playhouse .	एक सानो केटी उनको खेतमा सीढीहरु मा चढाई।
6	1000268201_693b08cb0e.jpg	A little girl in a pink dress going into a wooden cabin .	गुलाबी पोशाकमा एउटी सानी केटी काठको केबकमा जाँदै थियो।
7	1001773457_577c3a7d70.jpg	A black dog and a spotted dog are fighting	कालो कुकुर र स्पट गरिएको कुकुरले लडाई गर्दैछ
8	1001773457_577c3a7d70.jpg	A black dog and a tri-colored dog playing with each other on the roa	एक कालो कुकुर र एक ट्राई-रंगको कुकुर सडकमा एक अर्कासँग खेल्दै।
9	1001773457_577c3a7d70.jpg	A black dog and a white dog with brown spots are staring at each ot	एक कालो कुकुर र खैरो दागहरू सडकमा एक अर्को एक अर्कामा घुम्दै छन्।
10	1001773457_577c3a7d70.jpg	Two dogs of different breeds looking at each other on the road .	सडकमा एक अर्कालाई हेर्दा विभिन्न जातका दुई कुकुरहरू।
11	1001773457_577c3a7d70.jpg	Two dogs on pavement moving toward each other .	फुटपाथमा दुई कुकुरहरू एक अर्कामा सर्छन्।

Results[6]

Excepted Output Image with Caption
True Caption: युवायुवती समुन्द्रको किनारामा।
Expected Caption: युवायुवती हिड्दै छन् ।

Excepted Output Image with Caption
True Caption: क्या राम्रो रातो फूल।
Expected Caption: रातो फूल।

Types of Dataset	Exact Results	Predicted Results
Dataset-1	100%	Not sure
Dataset-2	100%	Not sure



Input Image without Caption

Excepted Output Image with Caption
True Caption: पोखरी।
Expected Caption: खैरो सेतो।

Discussion and Analysis

- ✓ Domain classification for images
- ✓ Google translated Nepali captions are not accurate
- ✓ The following points justify the chosen model for this project.
    - Effective feature extraction
    - Handling language complexity
    - Transfer learning
    - Modularity and flexibility
- ✓ Insufficient and noisy data
- ✓ Model Training Issues : Overfitting and Underfitting

Remaining Tasks

- Dataset finalization(Images and Nepali Captions)
- Model building, training and evaluation
- Result interpretation

Tentative Timeline (Gantt Chart)

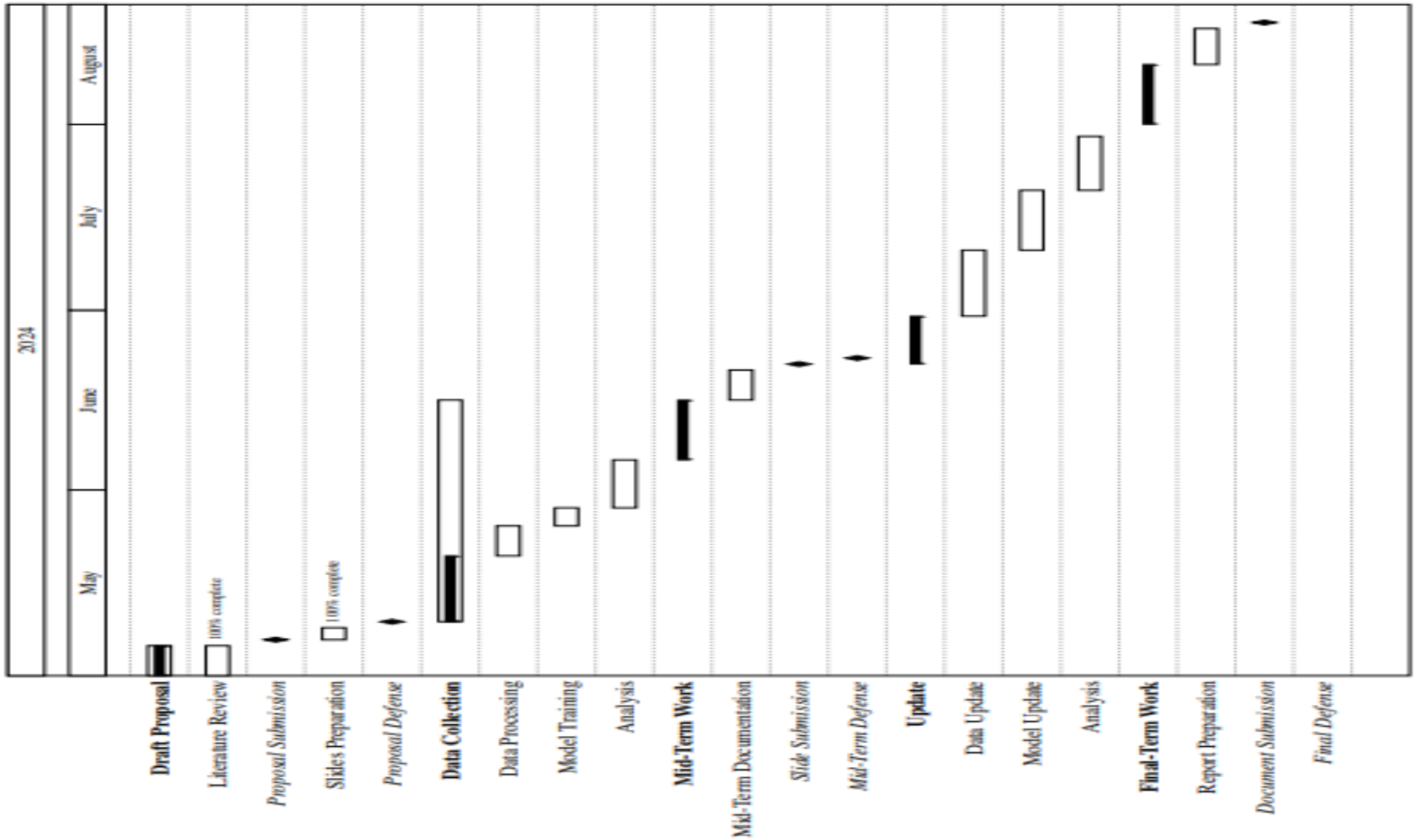


Figure A.7: Gantt Chart for Project Timeline

References[1]

1 Shuang Liu, Liang Bai,a, Yanli Hu and Haoran Wang.” Image Captioning Based on Deep Neural Networks” College of Systems Engineering, National University of Defense Technology, 410073 Changsha, China

2 Krizhevsky, Alex, I. Sutskever, and G. E. Hinton. "ImageNet classification with deep convolutional neural networks." International Conference on Neural Information Processing Systems Curran Associates Inc. 1097-1105. (2012)

3 Girshick, Ross, et al. "Region-based Convolutional Networks for Accurate Object Detection and Segmentation." IEEE Transactions on Pattern Analysis & Machine Intelligence 38.1:142-158. (2015)

4 Lei Ke , Wenjie Pei , Ruiyu Li , Xiaoyong Shen , Yu-Wing Tai” Reflective Decoding Network for Image Captioning”

References[2]

5 Li-Jia Li, Richard Socher, and Li Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In CVPR, 2009.

6 A. Hani, N. Tagougui and M. Kherallah, "Image Caption Generation Using A Deep Architecture," 2019 International Arab Conference on Information Technology (ACIT), 2019, pp. 246-251, doi: 10.1109/ACIT47987.2019.8990998

7 Aishwarya Maroju , Sneha Sri Doma , Lahari Chandarlapati," Image Caption Generating Deep Learning Model" Department Of Electronics and Computer Engineering J.N.T.U, Hyderabad , Sreenidhi Institute of Science And Technology, Ghatkesar, Yamnampet, Hyderabad,India

8 Zhongliang Yang, Yu-Jin Zhang, Sadaqat ur Rehman, Yongfeng Huang, Image Captioning with Object Detection and Localization, [Online] Available: https://arxiv.org/ftp/arxiv/papers/1706/1706.02430.p df

References[3]

9 Shubham Rawale, Megha Ghotkar, Krishna Sonavane, Paras Surve, Shraddha Khonde, Deepali Patil,'' IMAGE CAPTIONING GENERATOR SYSTEM WITH CAPTION TO SPEECH CONVERSION MECHANISM".

10 An Overview of Image Caption Generation Methods, Hindawi Computational Intelligence and Neuroscience Volume 2020, Article ID 3062706, 13 pages https://doi.org/10.1155/2020/3062706

11 M.D. Zeiler,R. Fergus, "Visualizing and understanding convolutional networks," In European conference on computer vision, Springer, Cham, pp. 818-833,2014

Thank You!!