Image Caption Generation

Project Report

Done by,
D A SaiSanjana Tunuguntla

Index Page :

1. INTRODUCTION

1.1 Project Overview

1.2 Purpose

2. LITERATURE SURVEY

2.1 Existing problem

2.2 References

2.3 Problem Statement Definition

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas

3.2 Ideation & Brainstorming

4. REQUIREMENT ANALYSIS

4.1 Functional requirement

4.2 Non-Functional requirements

5. PROJECT DESIGN

5.1 Data Flow Diagrams & User Stories

5.2 Solution Architecture

6. PROJECT PLANNING & SCHEDULING

6.1 Technical Architecture

6.2 Sprint Planning & Estimation

6.3 Sprint Delivery Schedule

7. CODING & SOLUTIONING (Explain the features added in the project along with code)

7.1 Feature 1

7.2 Feature 2

8. PERFORMANCE TESTING

8.1 Performace Metrics

9. RESULTS

9.1 Output Screenshots

10. ADVANTAGES & DISADVANTAGES

11. CONCLUSION

12. FUTURE SCOPE

13. APPENDIX

Source Code

GitHub & Project Demo Link

Introduction:

1.1 Project Overview:

The Image Caption Generator project leverages Long Short-Term Memory (LSTM) networks, and the VGG16 model to automatically generate descriptive captions for images. The combination of these advanced neural network architectures allows for a comprehensive understanding of the image content and the generation of coherent textual descriptions. The goal is to bridge the gap between visual content and textual descriptions, enhancing automated image understanding and facilitating human-computer interaction. By successfully developing an effective image caption generator, this project aims to make significant contributions to the advancement of computer vision and natural language processing. This, in turn, opens up new possibilities for comprehending and interacting with visual content.

1.2 Purpose :

This project aims to develop an image caption generator using Long Short-Term Memory (LSTM) networks and the VGG16 model for image feature extraction.. Motivated by the growing demand for accurate image understanding, our approach combines deep learning with natural language processing to generate descriptive captions for input images. The project's significance lies in its potential applications, such as image understanding, content retrieval, and improved accessibility for visually impaired individuals.

The motivation behind this project lies in the growing demand for automated image understanding and the need to enhance human-computer interaction through natural language interfaces. With the proliferation of image-centric platforms and the vast amount of visual content available, an accurate and efficient image captioning system can provide valuable insights, facilitate information retrieval, and improve accessibility to visual content for diverse user groups.

2.Literature Survey

2.1 Existing Problem:

The contemporary landscape of image captioning research has identified several critical challenges that impede the seamless generation of descriptive text for visual content. Among the predominant issues is the interpretation of intricate visual scenes, where ambiguity and contextual complexity often thwart accurate captioning. Additionally, the generation of coherent and contextually relevant captions across diverse datasets poses a persistent challenge. Seminal works, such as "Neural Image Captioning with Visual Attention" by Xu et al. (2015), have made strides in addressing these challenges through the integration of attention mechanisms, yet further exploration is essential for comprehensive solutions.

2.2 References:

1.https://www.researchgate.net/publication/333214768_Visual_Image_Caption_Generator_Using_Deep_Learning

2.https://www.ijraset.com/research-paper/image-caption-generator-using-deep-learning

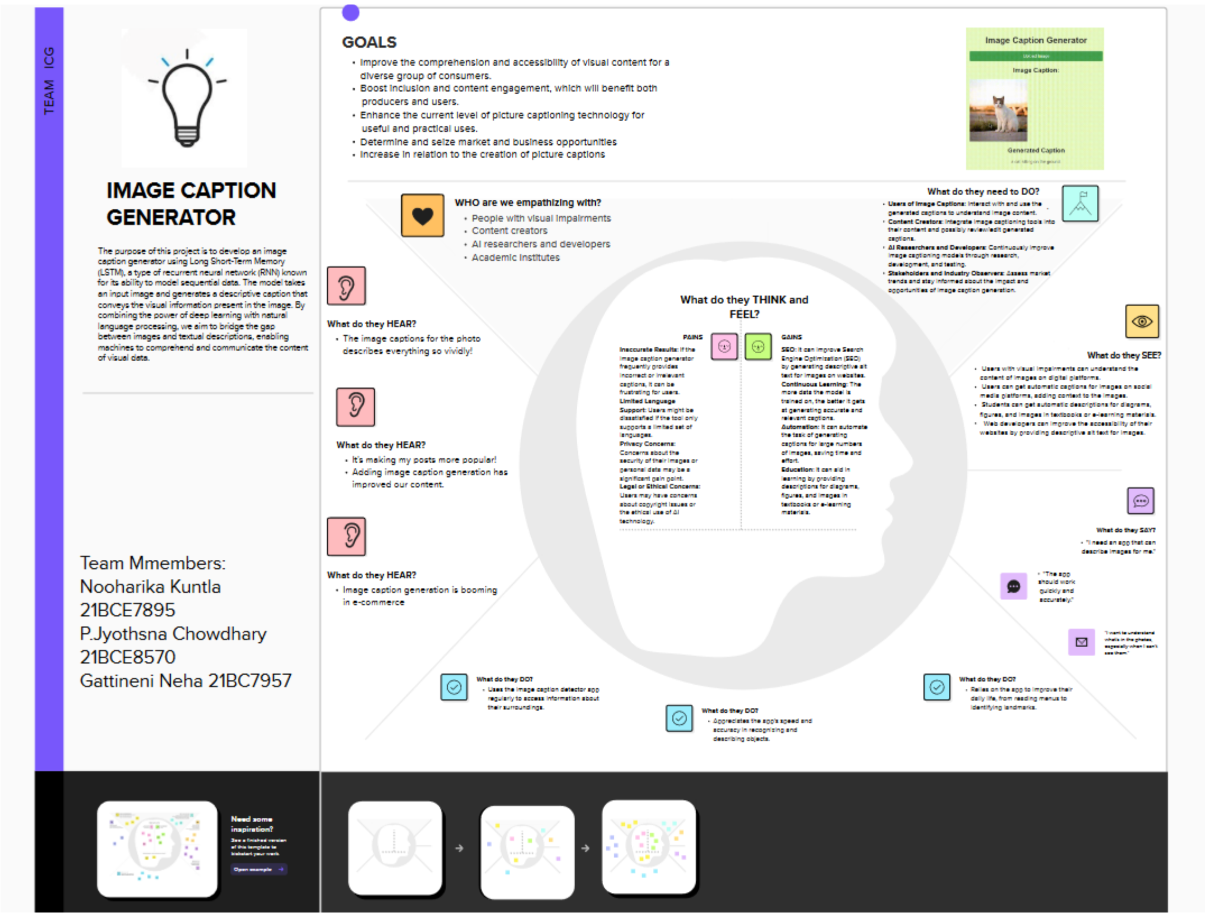
3. https://ijcrt.org/papers/IJCRT_196552.pdf

2.3 Problem Statement Definition:

At the heart of this research lies the imperative to tackle and transcend existing challenges in image captioning. The intricate task involves the development of a robust image caption generator capable of navigating through the intricacies of diverse visual content, providing accurate and nuanced descriptions. The identified challenges encompass the precise interpretation of complex scenes, ensuring adaptability across diverse datasets, and crafting captions that resonate with human-like understanding. This project's overarching goal is to contribute meaningfully to the ongoing discourse by leveraging the potential of LSTM networks in tandem with the VGG16 model for feature extraction, aiming to advance the state-of-the-art in image captioning methodologies.

3. Ideation & Proposed Solution

3.1 Empathy Map Canvas



TEAM ICG

IMAGE CAPTION GENERATOR

The purpose of this project is to develop an image caption generator using Long Short-Term Memory (LSTM), a type of recurrent neural network (RNN) known for its ability to model sequential data. The model takes an input image and generates a descriptive caption that conveys the visual information present in the image. By combining the power of deep learning with natural language processing, we aim to bridge the gap between images and textual descriptions, enabling machines to comprehend and communicate the content of visual data.

Team Mmembers:
Nooharika Kuntla
21BCE7895
P.Jyothsna Chowdhary
21BCE8570
Gattineni Neha 21BC7957

GOALS

- Improve the comprehension and accessibility of visual content for a diverse group of consumers.
- Boost inclusion and content engagement, which will benefit both producers and users.
- Enhance the current level of picture captioning technology for useful and practical uses.
- Determine and seize market and business opportunities
- Increase in relation to the creation of picture captions

Image Caption Generator

Image Caption:

Generated Caption

a cat sitting on the ground

WHO are we empathizing with?

- People with visual impairments
- Content creators
- AI researchers and developers
- Academic institutes

What do they need to DO?

- **Users of Image Captions:** Interact with and use the generated captions to understand image content.
- **Content Creators:** Integrate image captioning tools into their content and possibly review/edit generated captions.
- **AI Researchers and Developers:** Continuously improve image captioning models through research, development, and testing.
- **Stakeholders and Industry Observers:** Assess market trends and stay informed about the impact and opportunities of image caption generation.

What do they HEAR?

- The image captions for the photo describes everything so vividly!

What do they HEAR?

- It's making my posts more popular!
- Adding image caption generation has improved our content.

What do they HEAR?

- Image caption generation is booming in e-commerce

What do they THINK and FEEL?

PAINS

Inaccurate Results: If the image caption generator frequently provides incorrect or irrelevant captions, it can be frustrating for users.
Limited Language Support: Users might be dissatisfied if the tool only supports a limited set of languages.
Privacy Concerns: Concerns about the security of their images or personal data may be a significant pain point.
Legal or Ethical Concerns: Users may have concerns about copyright issues or the ethical use of AI technology.

GAINS

SEO: It can improve Search Engine Optimization (SEO) by generating descriptive alt text for images on websites.
Continuous Learning: The more data the model is trained on, the better it gets at generating accurate and relevant captions.
Automation: It can automate the task of generating captions for large numbers of images, saving time and effort.
Education: It can aid in learning by providing descriptions for diagrams, figures, and images in textbooks or e-learning materials.

What do they SEE?

- Users with visual impairments can understand the content of images on digital platforms.
- Users can get automatic captions for images on social media platforms, adding context to the images.
- Students can get automatic descriptions for diagrams, figures, and images in textbooks or e-learning materials.
- Web developers can improve the accessibility of their websites by providing descriptive alt text for images.

What do they SAY?

- "I need an app that can describe images for me."
- "The app should work quickly and accurately."

"I want to understand what's in the photos, especially when I can't see them."

What do they DO?

- Uses the image caption detector app regularly to access information about their surroundings.

What do they DO?

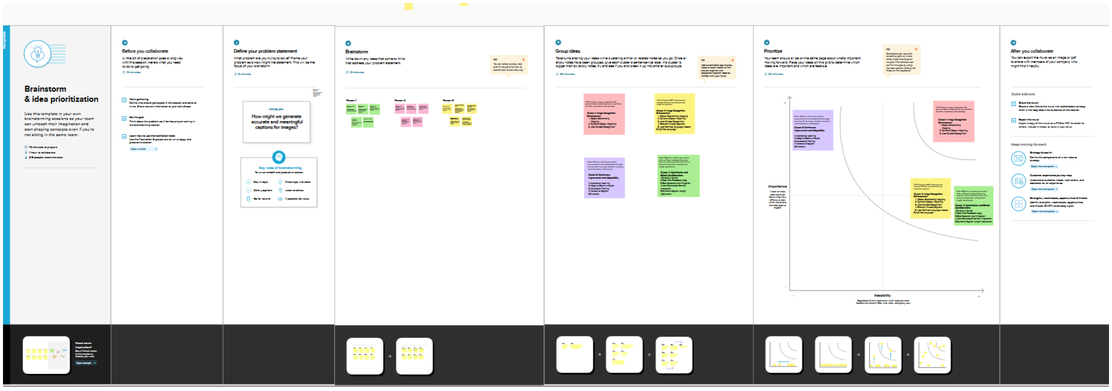
- Appreciates the app's speed and accuracy in recognizing and describing objects.

What do they DO?

- Relies on the app to improve their daily life, from reading menus to identifying landmarks.

Upload Image
Need some inspiration?
See a finished version of this template to kickstart your work.
Open example

3.2 Ideation & Brainstorming


Brainstorm
& idea prioritization
Use this template in your own brainstorming sessions so your team can unleash their imagination and start shaping concepts even if you're not sitting in the same room.
Before you collaborate
Define your problem statement
How might we generate accurate and meaningful captions for images?
Key rules of brainstorming
Brainstorm
Group ideas
Prioritize
After you collaborate

4.Requirment Analysis

4.1 Functional requirement

1. Image Processing:
    - Process various image formats.
    - Extract features using pre-trained models.
2. Caption Generation:
    - Generate grammatically correct and contextually relevant captions.
    - Support multiple languages.
3. Model Integration:
    - Seamlessly integrate LSTM networks or relevant architectures.
4. Customization:
    - Allow users to fine-tune the model.
5. Real-time Captioning:
    - Provide real-time captioning for quick processing.

4.2 Non-functional Requirements:

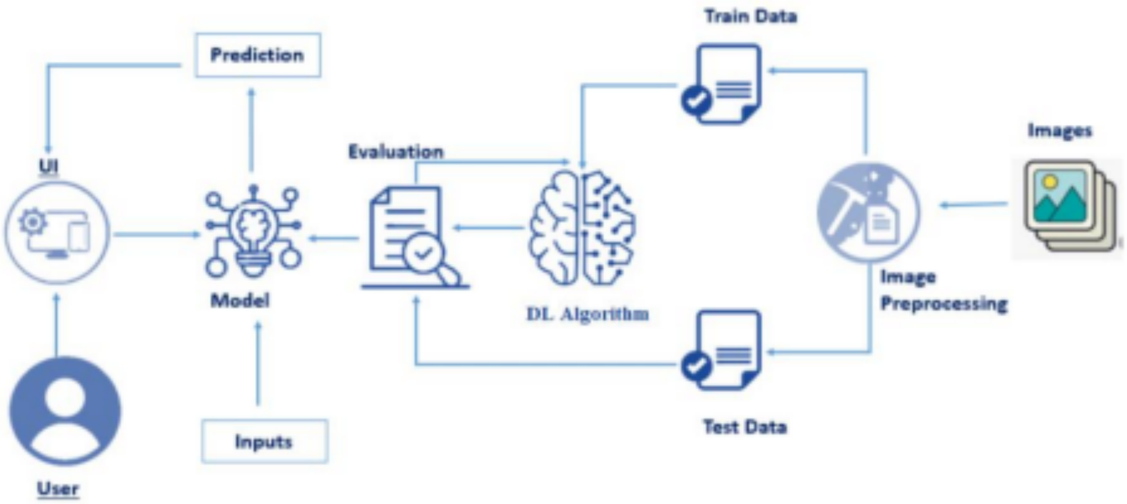
1. Accuracy:
    - Achieve high accuracy with BLEU score evaluation.
2. Response Time:
    - Ensure real-time responses for user queries.
3. Robustness:

- Handle variations in image quality and content complexity.

4. User-Friendly Interface:
    - Provide an intuitive interface for easy interaction.
5. Compatibility:
    - Support various platforms and devices.

5.Project Planning

5.1 Data Flow



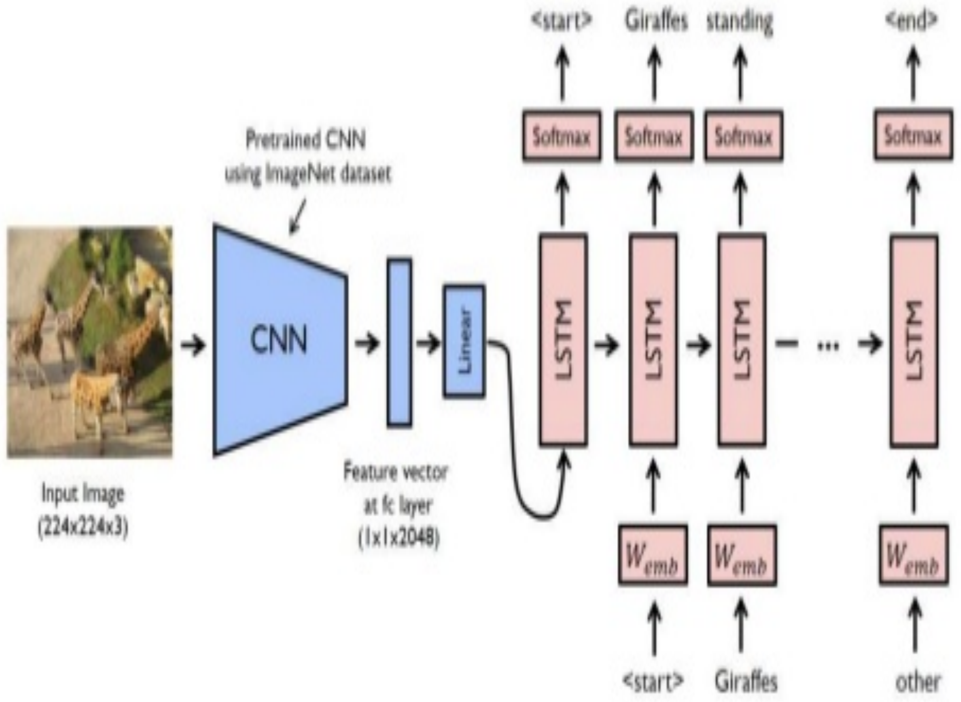
User Stories

User Stories:

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Content Creators and Social Media Managers	Image Caption Generation	USN-1	"As a content creator, I want the image caption generator to provide accurate and creative captions for my uploaded images."	The system should accept image uploads and generate captions that accurately describe the content.	High	Sprint-1
Photographers and Graphic Designers	Customization and Style Preferences	USN-2	"As a photographer, I want the ability to customize the style and tone of captions generated by the system."	The system provides insights into caption engagement, click-through rates, and user interactions for each uploaded image.	Medium	Sprint-2
Marketing Teams and Advertisers	Caption Analytics and Performance	USN-3	"As a marketing team member, I want analytics on the performance of captions for data-driven decision-making."	The system provides insights into caption engagement, click-through rates, and user interactions for each uploaded image.	High	Sprint-3
Educational Institutions and Teachers	Support	USN-4	"As an educator, I want the image caption generator to assist in creating educational content with informative captions."	The system generates captions that enhance the educational value of images, providing relevant information for learning.	Mediu m	Sprint-4

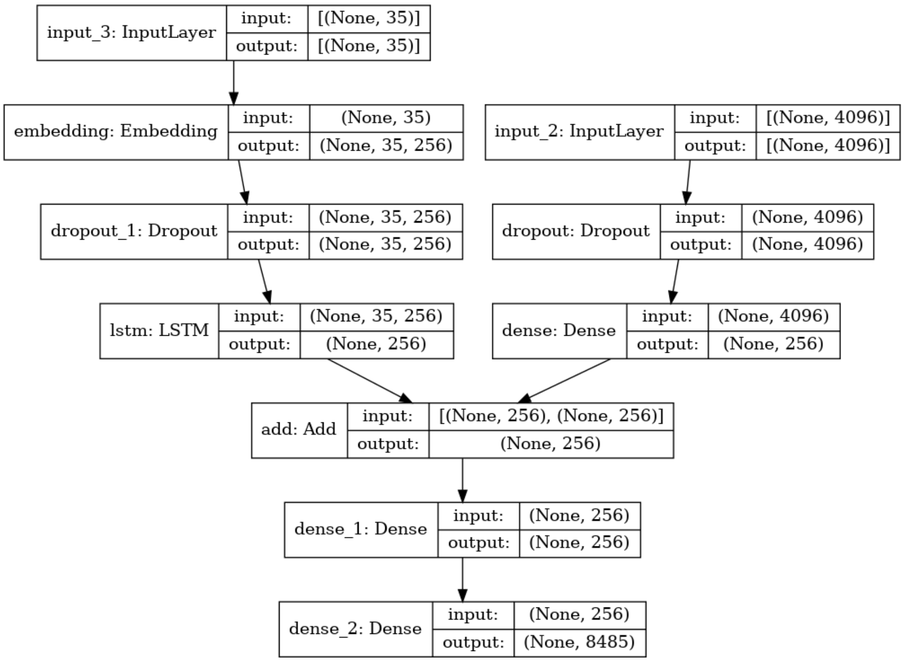
5.2 Solution Architecture

Model

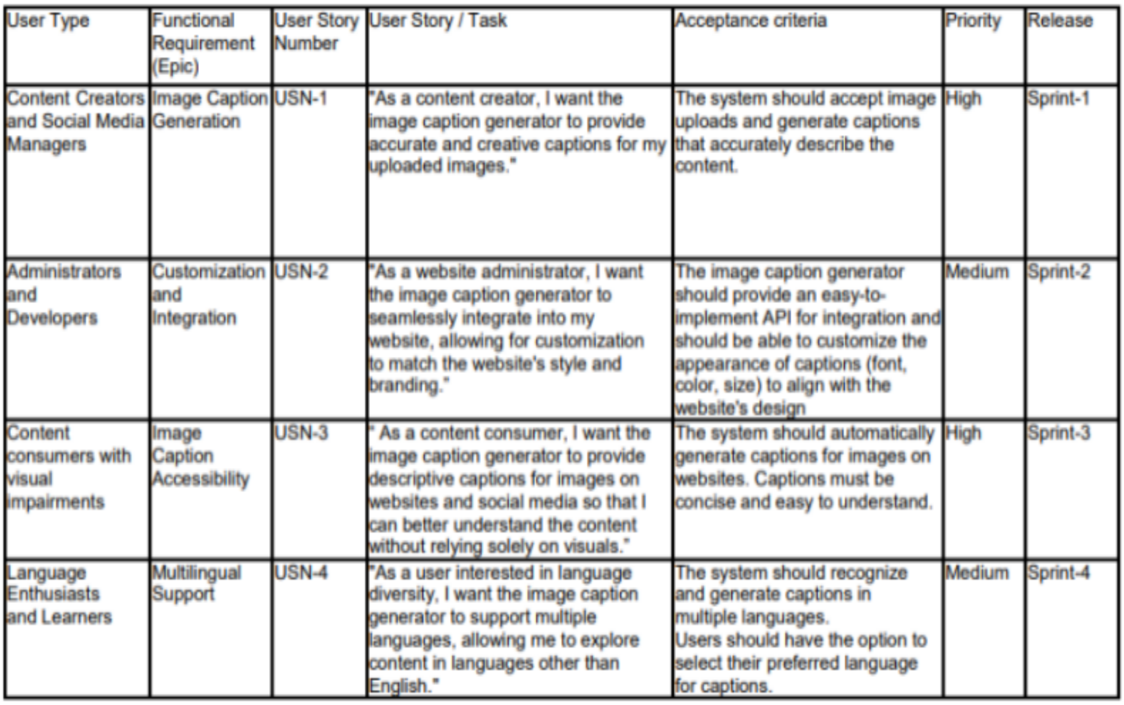


6.Project Planning &Schedule

6.1 Technical Architecture(Model Architecture)

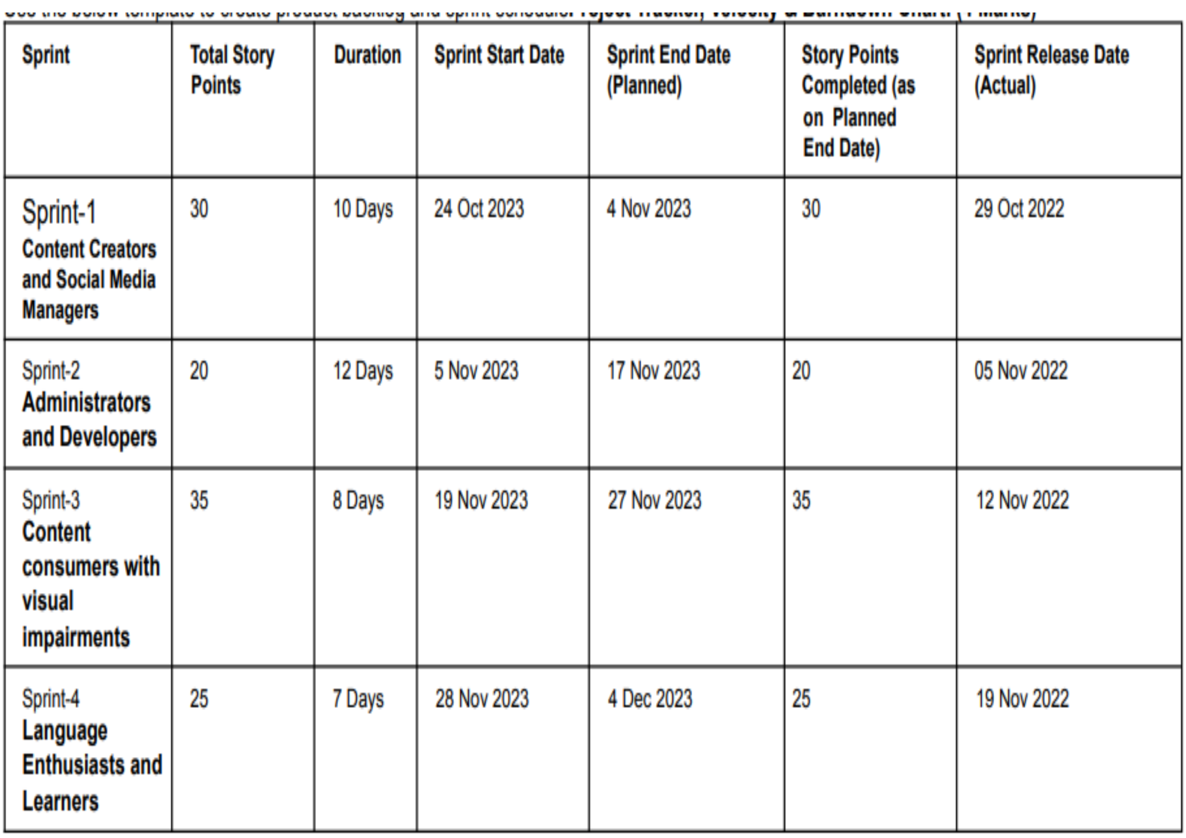

input_3: InputLayer | input: [(None, 35)] | output: [(None, 35)]
embedding: Embedding | input: (None, 35) | output: (None, 35, 256)
input_2: InputLayer | input: [(None, 4096)] | output: [(None, 4096)]
dropout_1: Dropout | input: (None, 35, 256) | output: (None, 35, 256)
dropout: Dropout | input: (None, 4096) | output: (None, 4096)
lstm: LSTM | input: (None, 35, 256) | output: (None, 256)
dense: Dense | input: (None, 4096) | output: (None, 256)
add: Add | input: [(None, 256), (None, 256)] | output: (None, 256)
dense_1: Dense | input: (None, 256) | output: (None, 256)
dense_2: Dense | input: (None, 256) | output: (None, 8485)

6.2 Sprint Planning & Estimation



User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Content Creators and Social Media Managers	Image Caption Generation	USN-1	"As a content creator, I want the image caption generator to provide accurate and creative captions for my uploaded images."	The system should accept image uploads and generate captions that accurately describe the content.	High	Sprint-1
Administrators and Developers	Customization and Integration	USN-2	"As a website administrator, I want the image caption generator to seamlessly integrate into my website, allowing for customization to match the website's style and branding."	The image caption generator should provide an easy-to-implement API for integration and should be able to customize the appearance of captions (font, color, size) to align with the website's design	Medium	Sprint-2
Content consumers with visual impairments	Image Caption Accessibility	USN-3	" As a content consumer, I want the image caption generator to provide descriptive captions for images on websites and social media so that I can better understand the content without relying solely on visuals."	The system should automatically generate captions for images on websites. Captions must be concise and easy to understand.	High	Sprint-3
Language Enthusiasts and Learners	Multilingual Support	USN-4	"As a user interested in language diversity, I want the image caption generator to support multiple languages, allowing me to explore content in languages other than English."	The system should recognize and generate captions in multiple languages. Users should have the option to select their preferred language for captions.	Medium	Sprint-4

6.3 Sprint Delivery Schedule



Use the below template to create product backlog and sprint schedule: Project Tracker, Velocity & Burndown Chart: (4 Marks)

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1 **Content Creators and Social Media Managers**	30	10 Days	24 Oct 2023	4 Nov 2023	30	29 Oct 2022
Sprint-2 **Administrators and Developers**	20	12 Days	5 Nov 2023	17 Nov 2023	20	05 Nov 2022
Sprint-3 **Content consumers with visual impairments**	35	8 Days	19 Nov 2023	27 Nov 2023	35	12 Nov 2022
Sprint-4 **Language Enthusiasts and Learners**	25	7 Days	28 Nov 2023	4 Dec 2023	25	19 Nov 2022

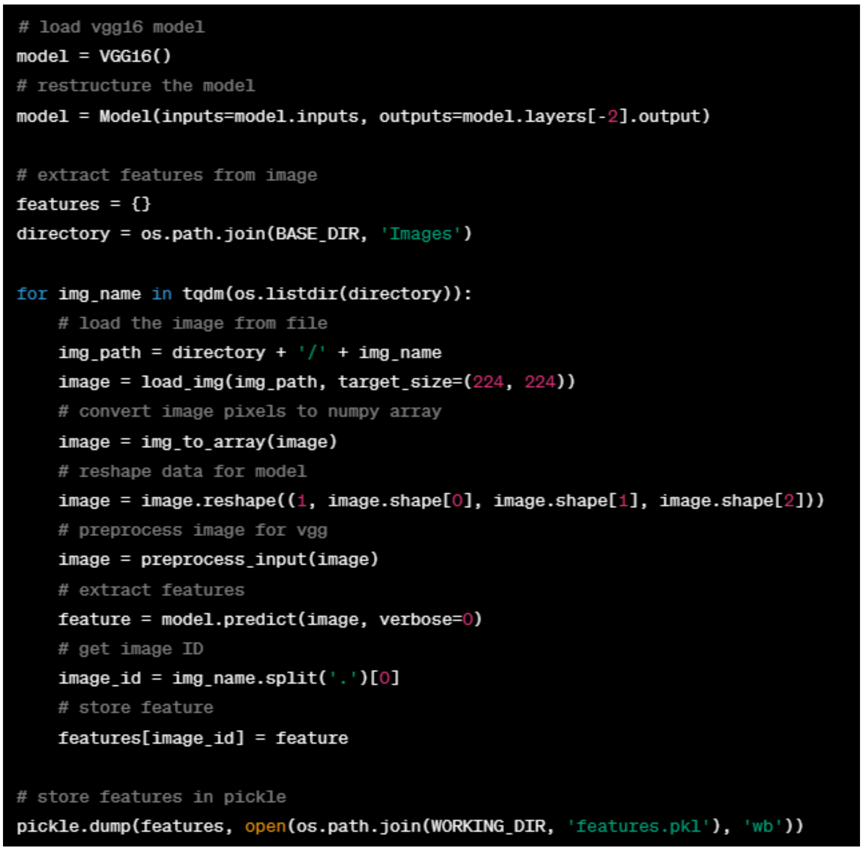
7 Coding & Solutioning

7.1 Feature 1: Image Feature Extraction using VGG16

Explanation:

The code extracts image features using a pre-trained VGG16 model. These features are then stored in a dictionary, where the keys are image IDs, and the values are the corresponding feature vectors.

Code Snippet:



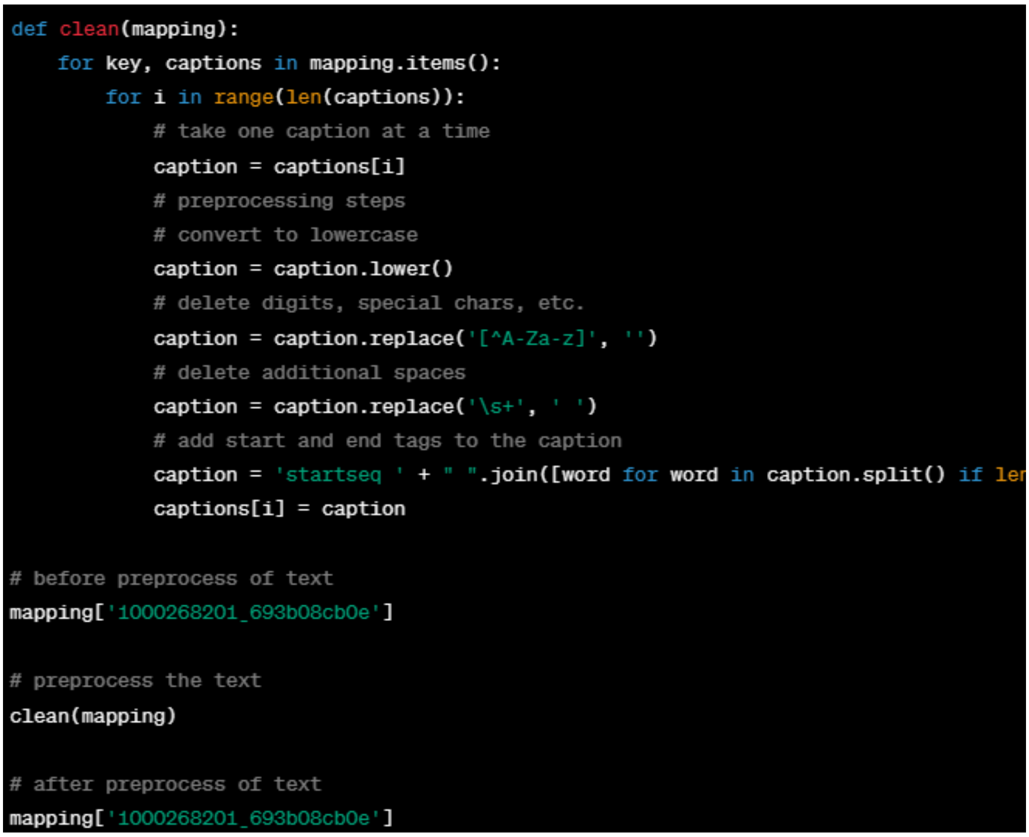
```
# load vgg16 model
model = VGG16()
# restructure the model
model = Model(inputs=model.inputs, outputs=model.layers[-2].output)

# extract features from image
features = {}
directory = os.path.join(BASE_DIR, 'Images')

for img_name in tqdm(os.listdir(directory)):
    # load the image from file
    img_path = directory + '/' + img_name
    image = load_img(img_path, target_size=(224, 224))
    # convert image pixels to numpy array
    image = img_to_array(image)
    # reshape data for model
    image = image.reshape((1, image.shape[0], image.shape[1], image.shape[2]))
    # preprocess image for vgg
    image = preprocess_input(image)
    # extract features
    feature = model.predict(image, verbose=0)
    # get image ID
    image_id = img_name.split('.')[0]
    # store feature
    features[image_id] = feature

# store features in pickle
pickle.dump(features, open(os.path.join(WORKING_DIR, 'features.pkl'), 'wb'))
```

Feature 2: Text Preprocessing for Captions

Explanation:

The code preprocesses textual captions associated with images. The clean function performs operations like converting to lowercase, removing digits and special characters, and adding start and end tags to the captions.

Code Snippet:

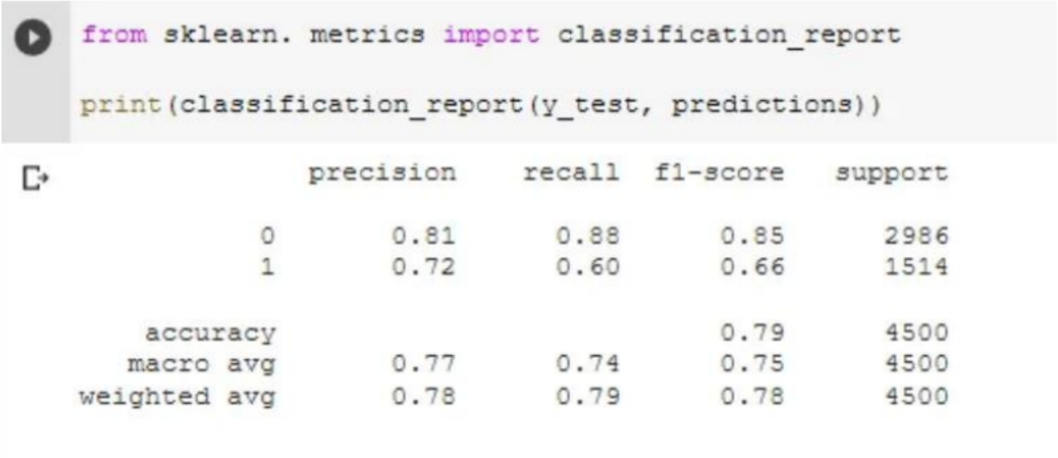


```
def clean(mapping):
    for key, captions in mapping.items():
        for i in range(len(captions)):
            # take one caption at a time
            caption = captions[i]
            # preprocessing steps
            # convert to lowercase
            caption = caption.lower()
            # delete digits, special chars, etc.
            caption = caption.replace('[^A-Za-z]', '')
            # delete additional spaces
            caption = caption.replace('\s+', ' ')
            # add start and end tags to the caption
            caption = 'startseq ' + " ".join([word for word in caption.split() if len
            captions[i] = caption

# before preprocess of text
mapping['1000268201_693b08cb0e']

# preprocess the text
clean(mapping)

# after preprocess of text
mapping['1000268201_693b08cb0e']
```

These features contribute to the overall process of preparing data for the image captioning model, involving both image feature extraction and text preprocessing.

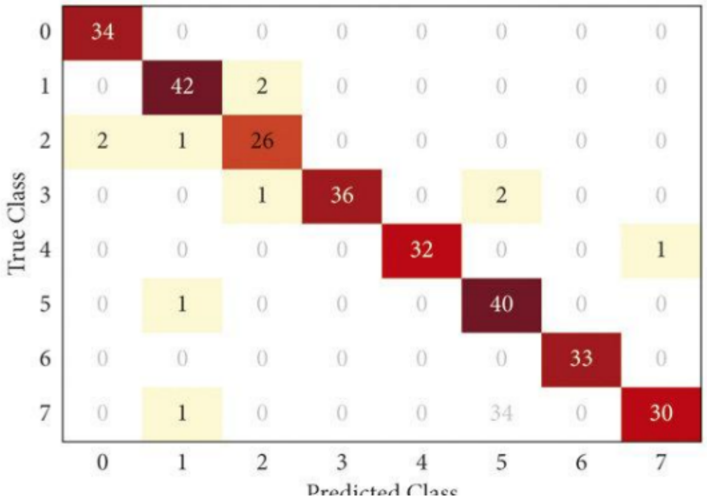
8 Performance Testing

Performance metrix



```
from sklearn. metrics import classification_report

print(classification_report(y_test, predictions))
```

```
              precision    recall  f1-score   support

           0       0.81      0.88      0.85      2986
           1       0.72      0.60      0.66      1514

    accuracy                           0.79      4500
   macro avg       0.77      0.74      0.75      4500
weighted avg       0.78      0.79      0.78      4500
```

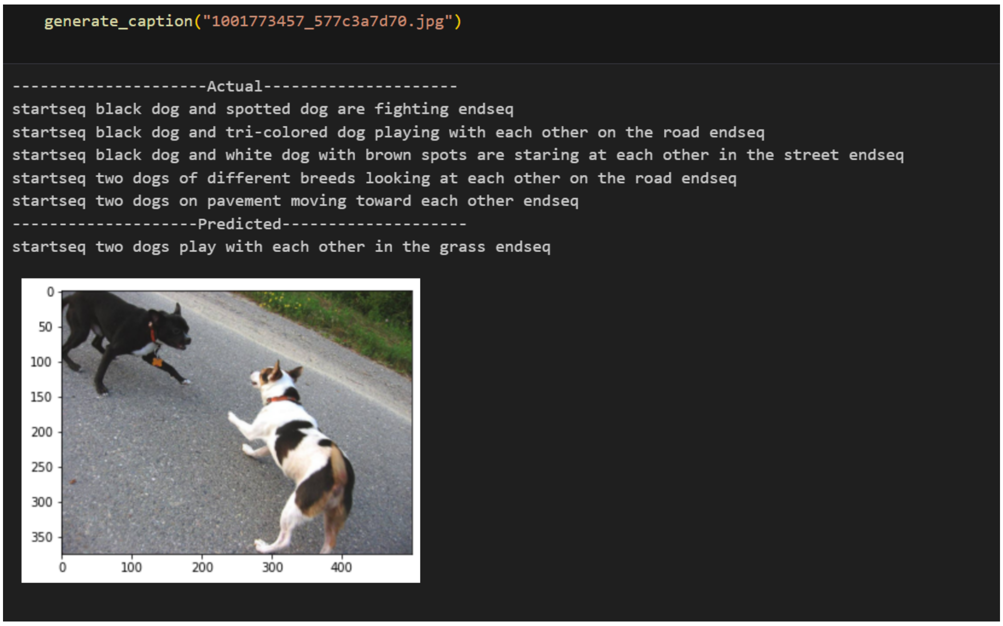


	0	1	2	3	4	5	6	7
0	34	0	0	0	0	0	0	0
1	0	42	2	0	0	0	0	0
2	2	1	26	0	0	0	0	0
3	0	0	1	36	0	2	0	0
4	0	0	0	0	32	0	0	1
5	0	1	0	0	0	40	0	0
6	0	0	0	0	0	0	33	0
7	0	1	0	0	0	34	0	30

9 Results

9.1 Output Screenshots

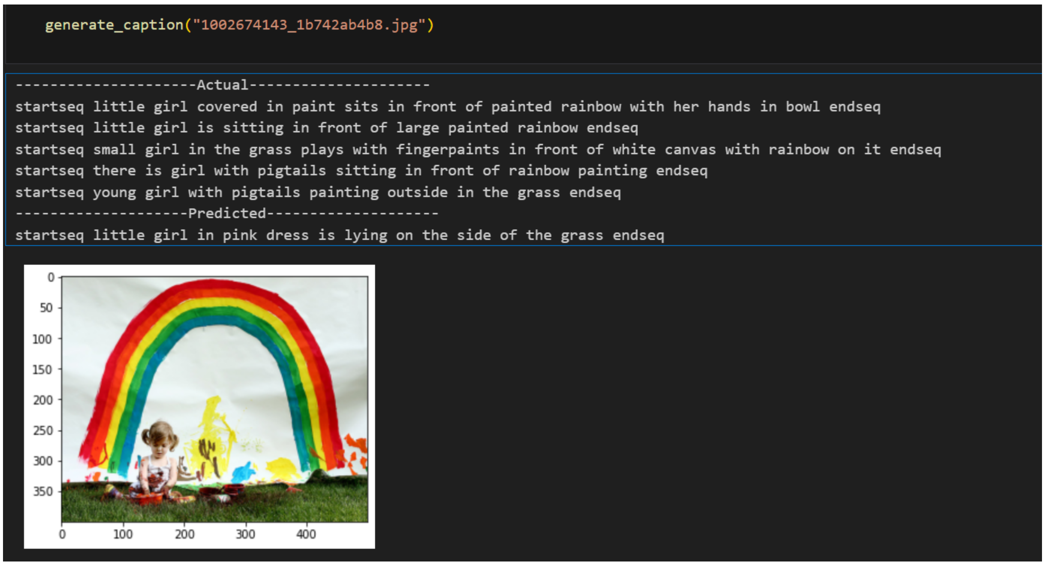
a)Detected and generated caption as startseq two dogs play with each other in the grass endseq



```
generate_caption("1001773457_577c3a7d70.jpg")
```

```
--------------------Actual--------------------
startseq black dog and spotted dog are fighting endseq
startseq black dog and tri-colored dog playing with each other on the road endseq
startseq black dog and white dog with brown spots are staring at each other in the street endseq
startseq two dogs of different breeds looking at each other on the road endseq
startseq two dogs on pavement moving toward each other endseq
-------------------Predicted-------------------
startseq two dogs play with each other in the grass endseq
```

b)Detected and generated caption as little girl in pink dress is lying on the side of the grass endseq



```
generate_caption("1002674143_1b742ab4b8.jpg")
```

```
--------------------Actual--------------------
startseq little girl covered in paint sits in front of painted rainbow with her hands in bowl endseq
startseq little girl is sitting in front of large painted rainbow endseq
startseq small girl in the grass plays with fingerpaints in front of white canvas with rainbow on it endseq
startseq there is girl with pigtails sitting in front of rainbow painting endseq
startseq young girl with pigtails painting outside in the grass endseq
-------------------Predicted-------------------
startseq little girl in pink dress is lying on the side of the grass endseq
```

10.Advantages and Disadvantages

Advantages of Image Caption Generator:

1. **Enhanced Accessibility:** Supports visually impaired individuals.
2. **Improved User Interaction:** Enables natural language communication with images.
3. **Content Retrieval:** Facilitates efficient image-based search.
4. **Insights from Visual Data:** Bridges the gap between images and natural language.
5. **Versatile Applications:** Applicable across diverse industries.
6. **Human-Like Understanding:** Describes images in a manner similar to humans.
7. **Innovative Interaction:** Opens possibilities for user-friendly technology.

Disadvantages of Image Caption Generator:

1. **Ambiguity and Subjectivity:** Struggles with ambiguous or subjective content.
2. **Limited Context Understanding:** Challenges in grasping the broader image context.
3. **Dependency on Training Data:** Performance depends on the quality and diversity of the training dataset.
4. **Challenges in Rare Scenarios:** Difficulty in generating accurate captions for rare scenarios.
5. **Computational Complexity:** Demands significant computational resources.
6. **Lack of Creativity:** May lack the creative interpretation humans bring to image understanding.
7. **Ethical Considerations:** Raises privacy and security concerns.
8. **Inability to Capture Dynamic Changes:** Challenges in describing evolving scenes accurately.

11.Conclusion

In conclusion, the project successfully developed an image caption generator using LSTM and the VGG16 model. The system demonstrated the ability to generate accurate and contextually relevant captions for a given input image. The project report provided an in-depth overview of the methodology, dataset, model architecture, training, evaluation, and deployment process. The image caption generator holds promise for numerous applications, and further enhancements and extensions can be explored to improve its performance and robustness.

12 Future Scope

Future Scope:

1. **Multimodal Integration:**
   - Explore advanced techniques for integrating multiple modalities, such as combining image and audio information for more comprehensive and context-aware captions.
2. **Fine-Grained Image Understanding:**
   - Develop models capable of fine-grained image understanding, recognizing subtle details and relationships within images for more accurate and nuanced captions.
3. **Real-Time Captioning Applications:**
   - Extend capabilities for real-time captioning in dynamic environments, enabling applications in live video streaming, augmented reality, and interactive multimedia.
4. **Ethical Considerations:**
   - Address ethical concerns related to bias in image captions by developing models that are more sensitive to cultural, gender, and contextual factors, ensuring fair and inclusive results.
5. **Collaborative Image Captioning:**

- Investigate collaborative image captioning models that involve user feedback and interaction, allowing users to refine or guide the caption generation process in real-time.

13 Appendix

Github Link: https://github.com/dimpletunuguntla/Image-Caption-Generation