Parshvanath Charitable Trust's

A. P. SHAH INSTITUTE OF TECHNOLOGY

(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)

(Religious Jain Minority)

A MINI PROJECT REPORT

On

Image Captioning

Submitted in partial fulfillment of the
requirement of University of Mumbai for the
Course

In

Computer Engineering (VII SEM)

Submitted By

Ajay Arvind Nirmal (20102170)

Mrunal Mahendra Misale (20102110)

Suhas Keshava Murthy (20102099)

Subject Incharge

Prof. Merlin Priya Jacob

Parshvanath Charitable Trust's

A. P. SHAH INSTITUTE OF TECHNOLOGY

(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)

(Religious Jain Minority)

CERTIFICATE

This is to certify that the requirements for the project report entitled '**Image Captioning**' have been successfully completed by the following students:

Name	**Moodle Id**
Ajay Arvind Nirmal	20102170
Mrunal Mahendra Misale	20102110
Suhas Keshava Murthy	20102099

In partial fulfillment of the course **Machine Vision (CSDC 7011)** in Sem: VII of Mumbai University in the Department of Computer Engineering during academic year 2023-2024.

Prof. Merlin Priya
Jacob Sub-in-Charge

Parshvanath Charitable Trust's

A. P. SHAH INSTITUTE OF TECHNOLOGY

(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

PROJECT APPROVAL

The project entitled '**Image Captioning**' by **Ajay Arvind Nirmal, Mrunal Mahendra Misale** and **Suhas Keshava Murthy** is approved for the course of Machine Vision (CSDC 7011) in Sem: VII of Mumbai University in the Department of Computer Engineering.

Subject-in-Charge

Prof. Merlin Priya Jacob

Date:
Place: Thane

Table of Contents

Abstract

Our project contains image captioning to enhance visual content and enrich user experiences. We employ advanced neural networks for image captioning, enabling automatic text generation that describes the content of uploaded images. We integrate style transfer techniques, allowing users to transform image styles artistically. Neural style transfer merges image content with selected styles, yielding captivating visuals.

Our project unifies these models to enhance media in a comprehensive manner. Users can upload images for creative expression, and receive informative captions. This fusion enhances content aesthetics, accessibility, and comprehension, benefiting creators, educators, and diverse applications. By bridging computer vision, natural language processing, and artistic creativity, our multimodal deep learning project offers a versatile tool for elevating visual and textual content simultaneously.

Problem Definition

Natural language processing models have shown a large neural network to perform a variety of text generation tasks. It is also possible to generate visual images based on text prompts using transformers by establishing the relationship between the image caption pairs to create novel plausible images for a great variety of sentences which can produce a high level of abstraction in the generated images. We propose an image generation model harnessing the natural language processing capabilities of vision transformers and using Contrastive Language Image Pre-training (CLIP) in conjunction with Variable Auto-encoder Adversarial nets to generate novel images from scratch.

Introduction

Image captioning is a compelling application of deep learning, combining computer vision and natural language processing to generate descriptive textual descriptions for images. In this context, Convolutional Neural Networks (CNNs) and Transformers play pivotal roles in achieving impressive results.Convolutional Neural Networks (CNNs) excel at feature extraction from images. They process an image through multiple convolutional layers to identify various visual patterns, such as edges, textures, and shapes. This hierarchical feature representation allows CNNs to capture the essence of an image's content, making them a fundamental component in image captioning pipelines.

Transformers, on the other hand, have revolutionized natural language processing tasks by enabling efficient sequence modeling. They excel at capturing contextual relationships among words, making them highly suitable for generating coherent and contextually relevant image captions. Transformers incorporate self-attention mechanisms that help weigh the importance of different image regions when generating textual descriptions.In this project, we leverage the power of CNNs and Transformers in tandem, using CNNs for image feature extraction and Transformers for text generation. This combination enables our system to not only recognize objects and scenes within images but also produce meaningful and context-aware captions, enhancing the overall user experience and understanding of visual content.

Proposed System

To control the crime rate, we have come up with a system that is able to detect suspicious activity. With the help of OpenCV and NumPy, we are able to use the real time feed through the camera and thus create the region of interest through which the intrusion of a particular entity would be detected. The Region of interest would be created using Select ROI method using OpenCV. If an entity enters this restricted area, then an alarm would be set in order to alert the authorities. Other than ROI, object detection is also used, for object detection, we have used coco model v3 of MobileNetSSD which has 256 objects which would be used for object detection

SOFTWARE REQUIREMENTS

1) PYTHON- For app development.

2) MODULES USED-

OPENCV:

OpenCV (Open-Source Computer Vision Library) is the huge open-source library for computer vision, machine learning, and image processing and now it plays a major role in real-time operation which is very important in today's systems. By using it, one can process images and videos to identify objects, faces, or even handwriting of a human.

NUMPY:

NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python.

5.2 HARDWARE REQUIREMENTS

1) CPU: 32-Bit CPU or 64-Bit CPU (Intel/AMD architecture)

2) RAM: Minimum 4GB RAM

3) STORAGE: 1 GB Free Disk Space

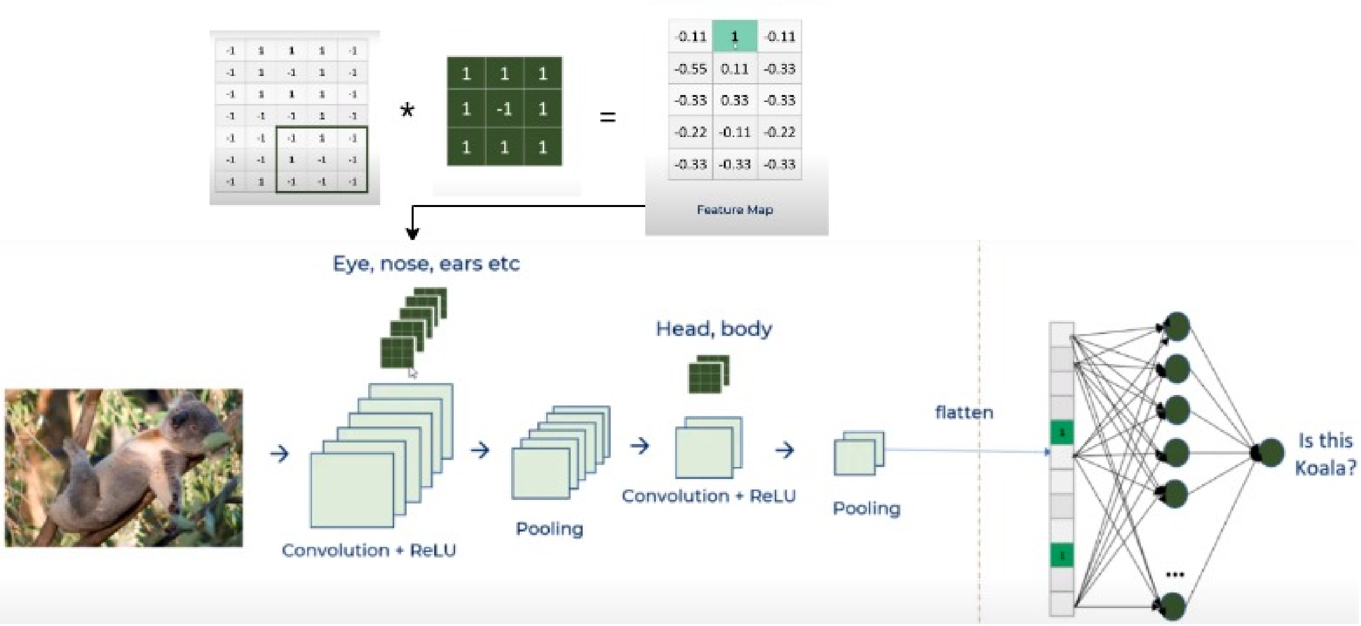
4) OS: Linux-Ubuntu, Windows 7-11

5) SURVEILLANCE CAMERA

6)LAPTOP /Desktop

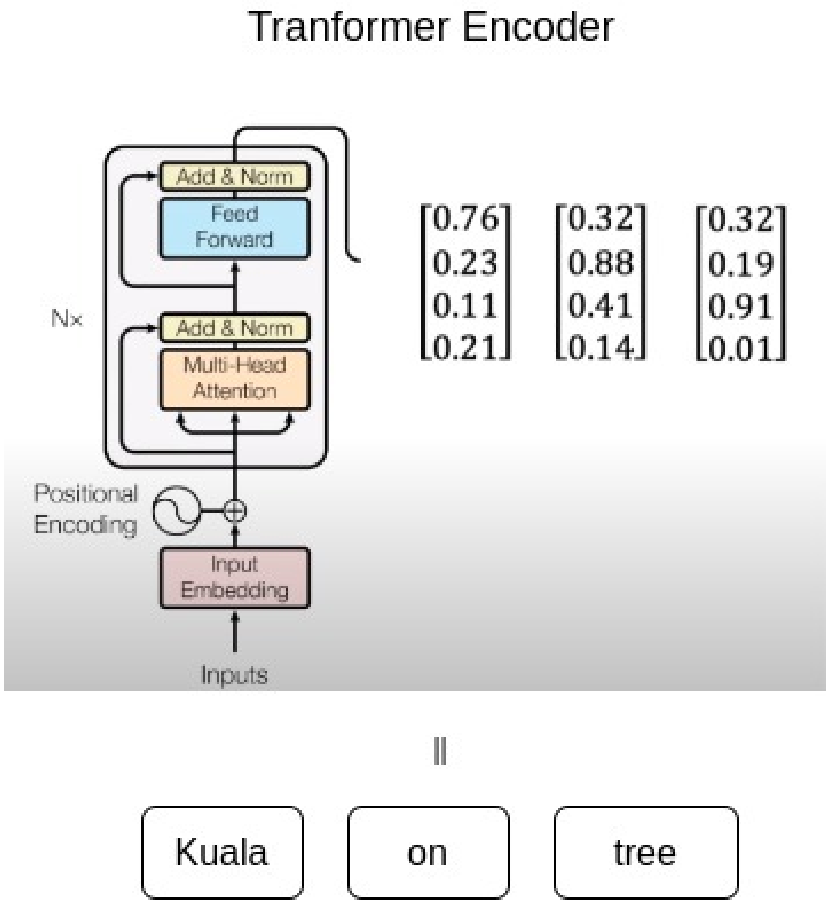
System Implementation

Module 1: Convolutional Neural Network



In the field of artificial intelligence and machine learning, CNN stands for Convolutional Neural Network. It's a type of deep learning algorithm designed for processing structured grid data, such as images and video. CNNs are particularly powerful for tasks like image recognition and computer vision.

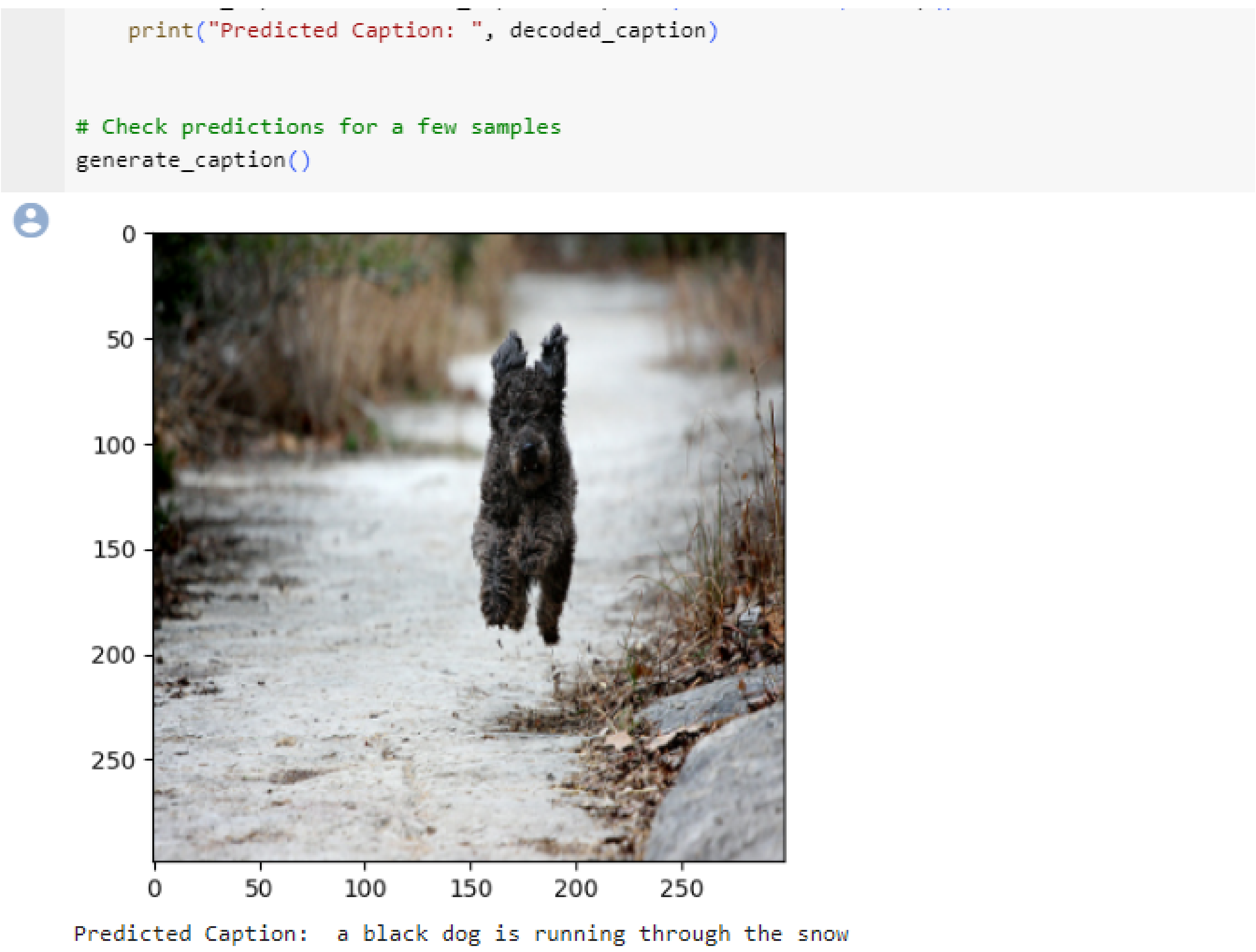
Module 2: Transformer Encoder and Decoder



Transformer encoding, in this context, is a crucial step because it allows the captioning model to understand and generate text descriptions based on the visual content of the image. By using pre-trained CNNs to extract features, the model leverages the knowledge learned from large image datasets, improving its ability to recognize and describe objects, scenes, and other visual elements in images accurately. This combination of computer vision and natural language processing enables the development of image captioning systems that can automatically generate human-readable descriptions for images.

The Transformer Decoder in image captioning effectively combines image information with textual generation, allowing the model to generate descriptive captions that are contextually grounded in the content of the input image. It has proven to be a powerful architecture for this task and has been used in various state-of-the-art image captioning models.

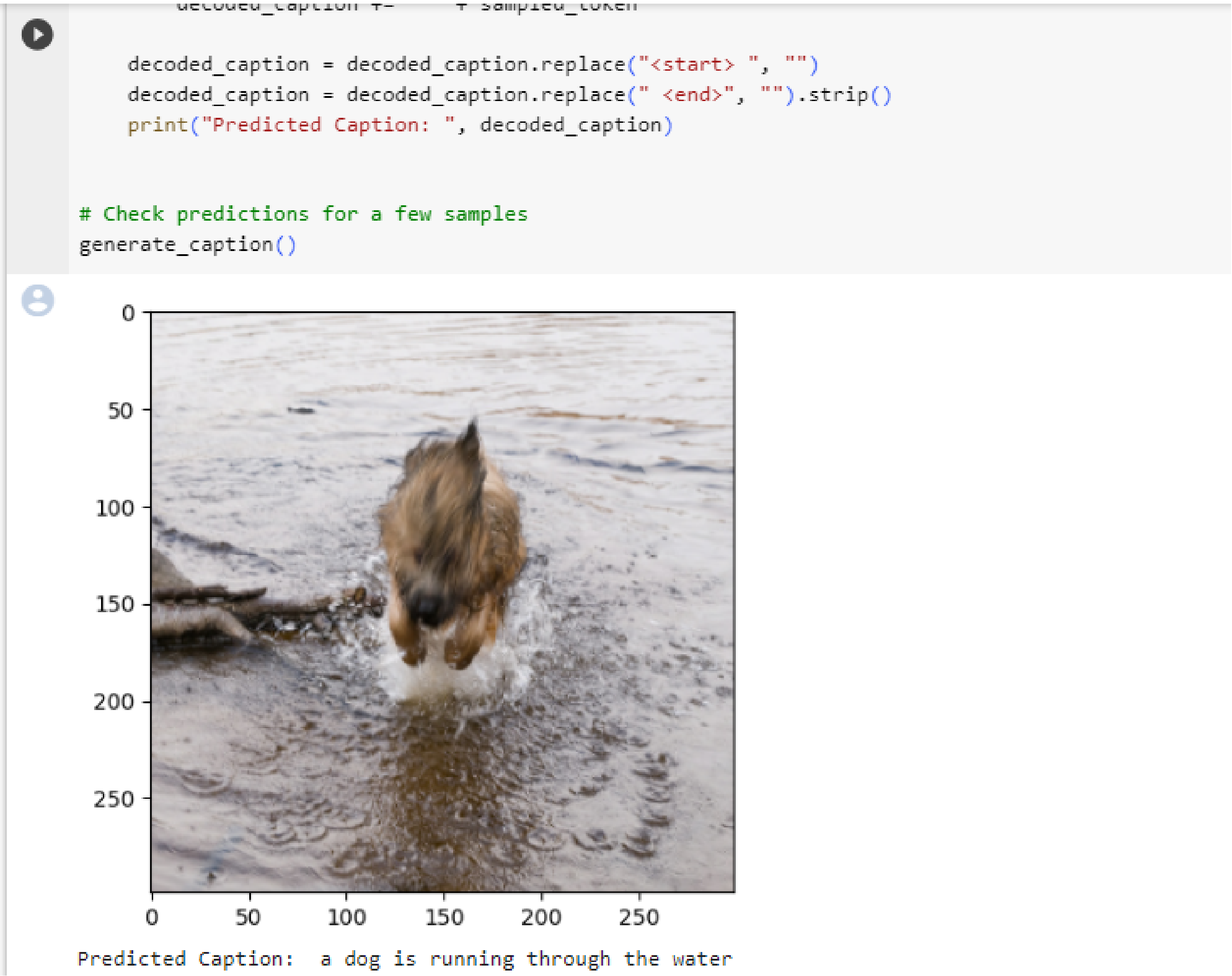
OUTPUT:



```
    print("Predicted Caption: ", decoded_caption)


# Check predictions for a few samples
generate_caption()
```

```
Predicted Caption:  a black dog is running through the snow
```

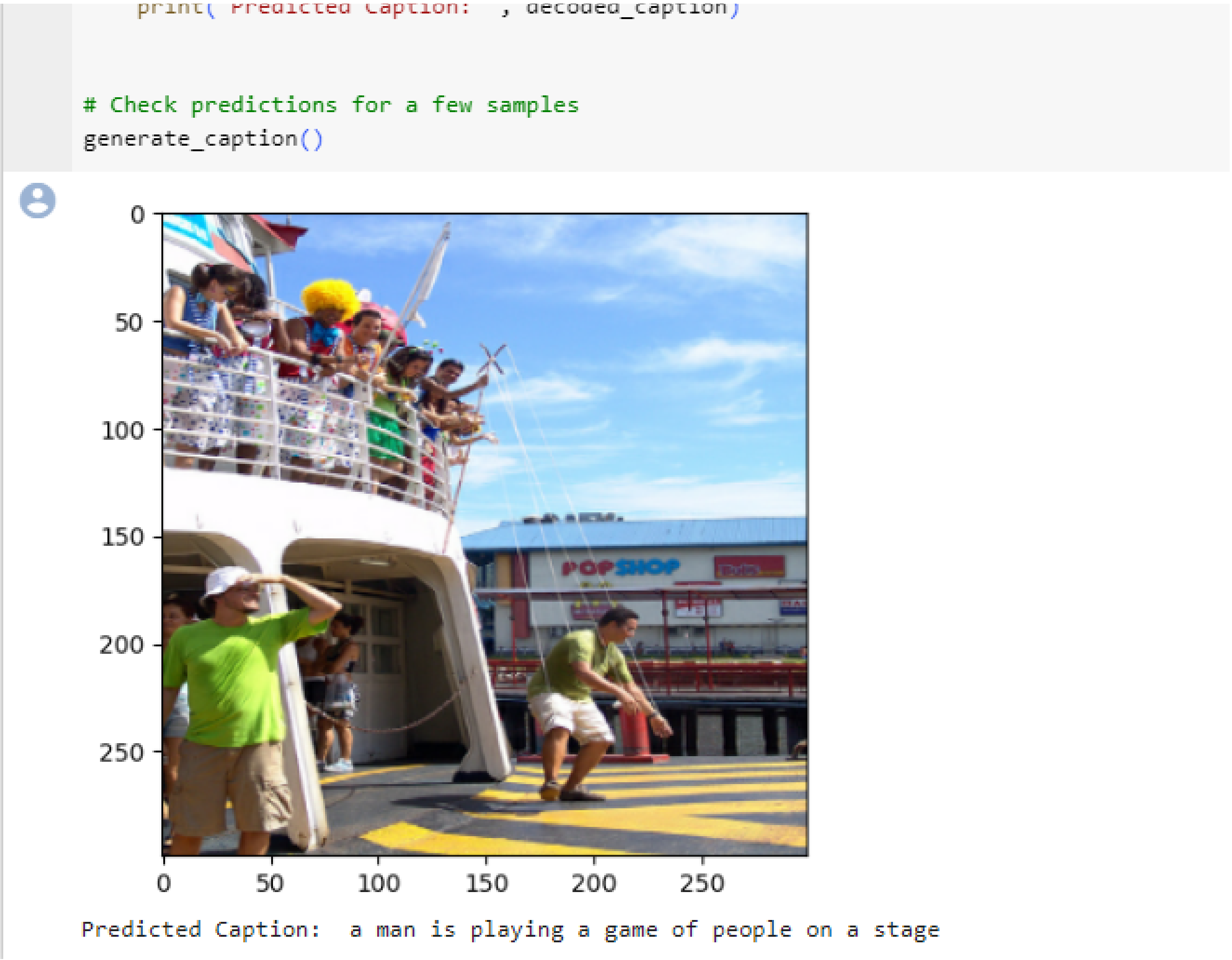


```
        decoded_caption += " " + sampled_token

    decoded_caption = decoded_caption.replace("<start> ", "")
    decoded_caption = decoded_caption.replace(" <end>", "").strip()
    print("Predicted Caption: ", decoded_caption)


# Check predictions for a few samples
generate_caption()
```

```
Predicted Caption:  a dog is running through the water
```



```
    print("Predicted Caption: ", decoded_caption)


# Check predictions for a few samples
generate_caption()
```

```
Predicted Caption:  a man is playing a game of people on a stage
```

References:

1. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H. (2016). “Generative adversarial text to image synthesis”. In ICML 2016. ↩
2. Andreas, J., Klein, D., and Levine, S. Learning with latent language. arXiv preprint arXiv:1711.00482, 2017.
3. Reed, S., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H. (2016). “Learning what and where to draw”. In NIPS 2016. ↩
4. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang X., Metaxas, D. (2016). “StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks”. In ICCY 2017. ↩
5. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D. (2017). “StackGAN++: realistic image synthesis with stacked generative adversarial networks”. In IEEE TPAMI 2018. ↩

Acknowledgement

We have great pleasure in presenting the mini project report on “Image Captioning ”. We take this opportunity to express our sincere thanks towards our guide Prof. Merlin Jacob Department of Computer Engineering, APSIT Thane for providing the technical guidelines and suggestions regarding line of work. We would like to express our gratitude for his constant encouragement, support and guidance through the development of the project.

We thank ma’am for her encouragement during the progress meeting and for providing guidelines to write this report.

We wish to express our deep gratitude towards all our colleagues of APSIT for their encouragement.

Student Names	**Student Id**
Ajay Arvind Nirmal	20102170
Mrunal Mahendra Misale	20102110
Suhas Keshava Murthy	20102099