# Image Captioning Generator

By
Sahith Damera
Prudhviraj Boddu

# CONTENTS

Here's what you'll find

# Decoding the Challenge

01

## PROBLEM & SOLUTION

Describing Images From Silent Images to Descriptive Narratives

02

## OUR PROCESS

Tokenizing and extracting image features with ResNet-18 and Inspection V3

03

## TARGET

Fine-Tuning and Adaptability

# Introduction

What do you see in the picture?

Image captioning involves generating descriptive text to describe the content of an image

In the era of AI, understanding images is crucial for machines to interact with the visual world. Image captioning bridges the gap between visual data and natural language

# UNDERSTANDING THE PROBLEM

## Generating Captions for Images

The central challenge lies in developing models that can effectively transform visual information into coherent and relevant textual descriptions. This requires overcoming complexities in understanding the context and semantics of images.

We are addressing the challenge by conducting a comparative analysis between ResNet-18 and Inception V3 to determine the superior solution

# Essential Building Blocks

## Data Preprocessing

For Tokenization

## Image Feature Extraction

Using ResNet-18

## Implementation

Using Transformer based Captioning model for text

## Deploying

Creating a Flask Application

# From Pixels to Text

## Data Preprocessing

This involves tokenization and vocabulary building. Here, we leverage the capabilities of ResNet-18 to extract rich image features and train a Transformer-based model for caption generation.

## Data Exploration

Exploring sample captions and associated images in Flickr 8k dataset. This step provides insights into the preprocessing techniques and the diversity of image content we are working with.

## Image Feature Extraction

The selection of ResNet-18 for image feature extraction is pivotal. Its deep architecture captures intricate visual patterns, enhancing our model's ability to understand and interpret visual content effectively.

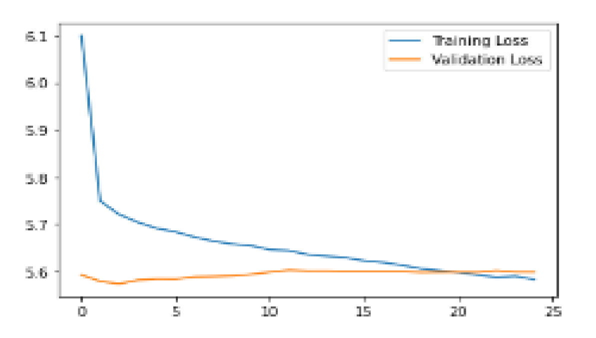
## Model Architecture

Our model adopts a Transformer-based architecture, a cutting-edge approach for handling sequential data. Incorporating positional encoding and attention mechanisms,

# Training Process

Training involves an iterative process, optimizing our model using the Adam optimizer and minimizing the Cross Entropy loss. Techniques to address challenges like sequence padding are employed, ensuring robust learning.

```
Epoch ->  20  Training Loss ->  5.597502708435059 Eval Loss ->  5.598044395446777
Epoch    21: reducing learning rate of group 0 to 2.6214e-06.
Epoch ->  21  Training Loss ->  5.591978073120117 Eval Loss ->  5.597818851470947
Epoch ->  22  Training Loss ->  5.586987018585205 Eval Loss ->  5.601054191589355
Epoch ->  23  Training Loss ->  5.589152336120605 Eval Loss ->  5.598156452178955
Epoch    24: reducing learning rate of group 0 to 2.0972e-06.
Epoch ->  24  Training Loss ->  5.582265377044678 Eval Loss ->  5.59806489944458
```



# Evaluation

The Model generating captions that closely align with ground truth annotations. Performance metrics, including BLEU scores, highlight the effectiveness of our approach. Visual comparisons further validate the quality of generated captions."

## Training

Phase-1

Learning the
Features of data

## Validation

BIEU Score

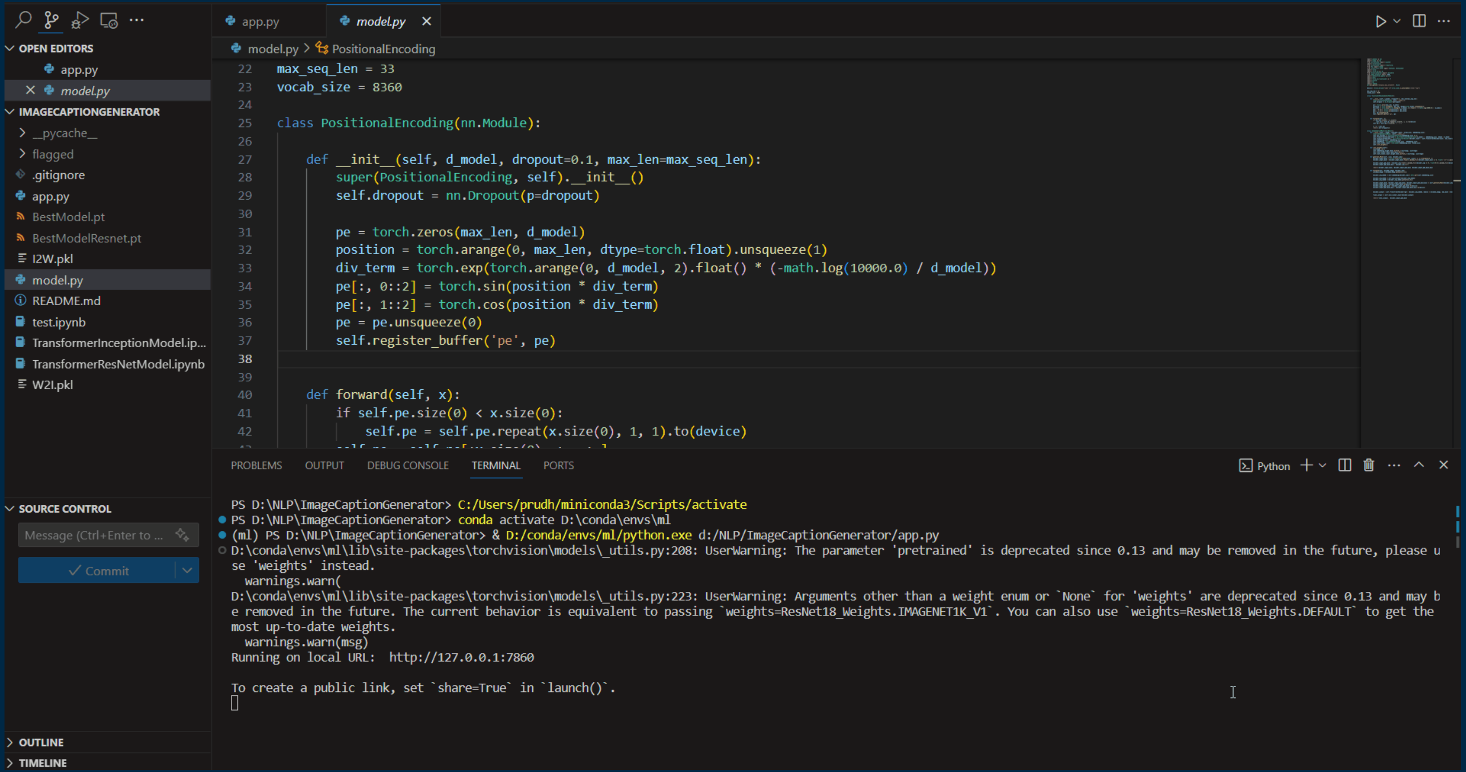
metric used for
evaluating the quality
of machine-generated
text

## Testing

Evaluation

Generated relevant
Caption’s for Image

# Implementation – Gradio App



# Future Work and Conclusion

In Future, we work could explore additional architectural enhancements or the utilization of diverse datasets to further elevate between different model's performance.

In conclusion, our work successfully integrates ResNet-18 features with a Transformer-based model for image captioning.

Do you have any questions?

sdame1@unh.newhaven.edu
pboddu@unh.newhaven.edu

# THANKS