Department of Computer Science
IIT Bombay

PROFESSOR IN CHARGE

Dr,. Sunita Sarawagi

STUDENTS

Sourabh Kale	(23M0783)
Akshay Patidar	(23M0792)
Ravi Patidar	(23M0796)
Ravikant Chandrawat	(23M0804)
Priyanshu Sharma	(23M0834)

Image Captioning using Transformers

INTRODUCTION

- Image captioning describes the challenge of converting visual content into textual descriptions, a key focus for AI systems.
- Recent years have seen increased interest in automatic image captioning, driven by the success of deep learning in language and image processing.
- Models commonly adopt a translational approach, combining a visual encoder with a linguistic decoder.
- Automatic translation faces complexities, particularly when words' meanings are context-dependent, exacerbated in cross-modal tasks like image-to-text translation.
- Attention mechanisms play a pivotal role, guiding models to focus during encoding and utilizing recurrent neural networks with attention during decoding.
- The transformer, designed for natural language processing, offers a comprehensive approach. It relates embedded words in sentences, training end-to-end without explicit auxiliary models for relation detection.
- Unlike traditional recurrent neural networks, transformers process input sentences/sequences in

THE PURPOSE

- Allows visually impaired individuals to access and comprehend visual content.
- Enables efficient image search based on textual queries.
- Enhances human-AI communication by providing meaningful textual context to images.
- Provides a bridge between visual and linguistic understanding, aiding in richer content comprehension.
- Supports educational tools by narrating visual content for better learning experiences.
- Boosts engagement by automatically generating descriptive captions for shared images.
- Essential for applications requiring integration of both visual and textual information.
- Drives advancements in assistive technologies, fostering inclusivity and accessibility.
- Improves user experience by providing meaningful context, and creating more immersive applications.

Problem Statement

In the realm of computer vision and natural language processing, the task of generating descriptive and contextually rich captions for images poses a compelling challenge. Existing image captioning methods, often reliant on traditional approaches, struggle to capture intricate relationships and contextual nuances within visual content. The emergence of transformer architectures in natural language processing has shown immense potential for parallel processing and holistic contextual understanding.

This project aims to address the limitations of conventional image captioning methods by harnessing the capabilities of transformers, paving the way for more accurate, context-aware, and efficient image description generation.

Dataset Exploration

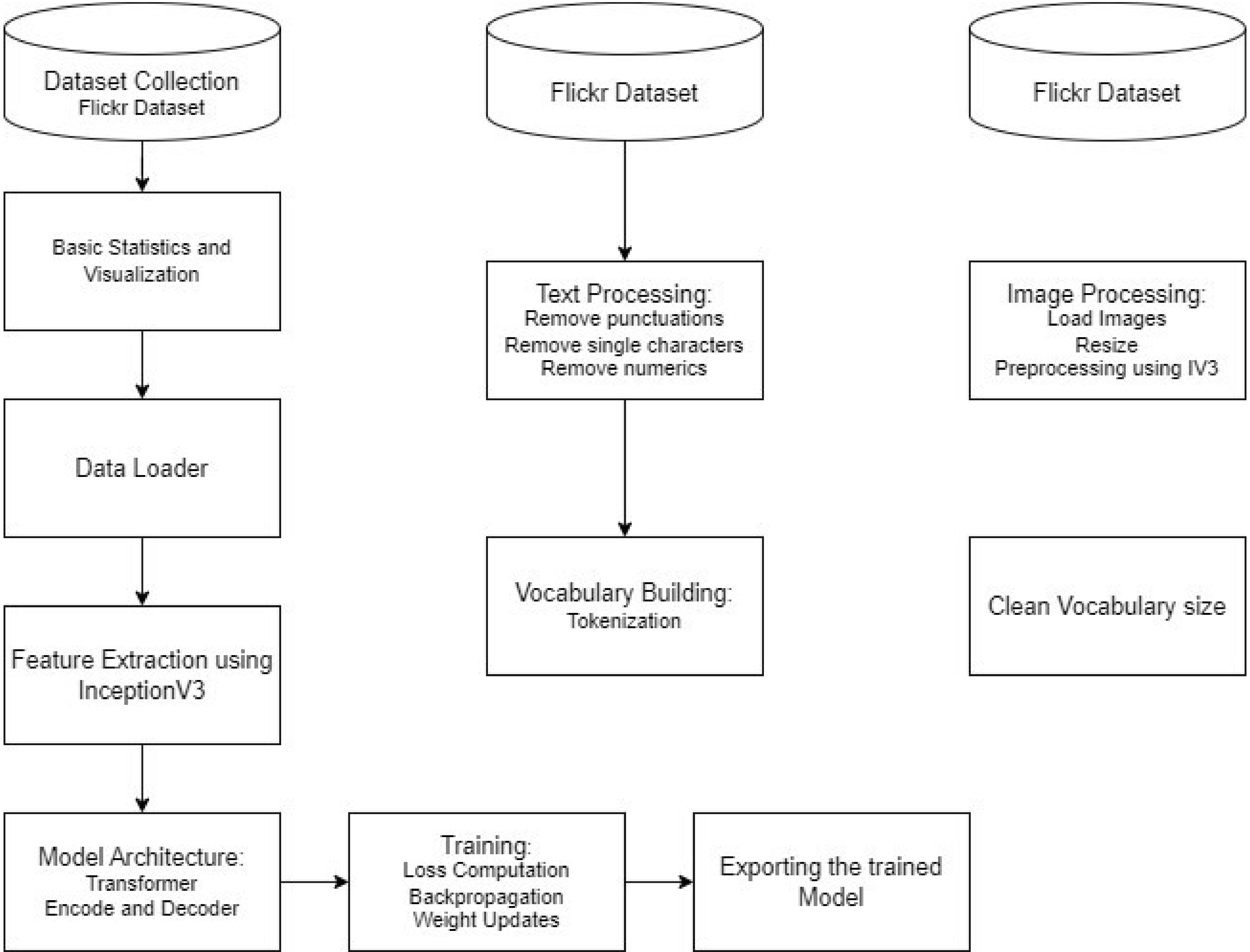
The Flickr8k dataset is a large corpus of images and captions that is widely used for image captioning research. It consists of 8,092 images and up to five captions for each image. The images were collected from Flickr and tend not to contain any well-known people or locations and cover a wide range of topics, including people, animals, objects, and landscapes.

Some key points about the Flickr8k dataset:

- Dataset size: 8,092 images
- Captions per image: Up to 5
- Image types: Wide range, including people, animals, objects, and landscapes
- Image resolution: Varies, but typically around 300x200 pixels
- Image format: JPEG
- Caption format: Plain text

We have seen it being used to train and evaluate a wide range of image captioning models, and it has contributed to significant advances in the field.

Training Workflow



Dataset Collection
Flickr Dataset

Basic Statistics and Visualization

Data Loader

Feature Extraction using InceptionV3

Model Architecture:
Transformer
Encode and Decoder

Training:
Loss Computation
Backpropagation
Weight Updates

Exporting the trained Model

Flickr Dataset

Text Processing:
Remove punctuations
Remove single characters
Remove numerics

Vocabulary Building:
Tokenization

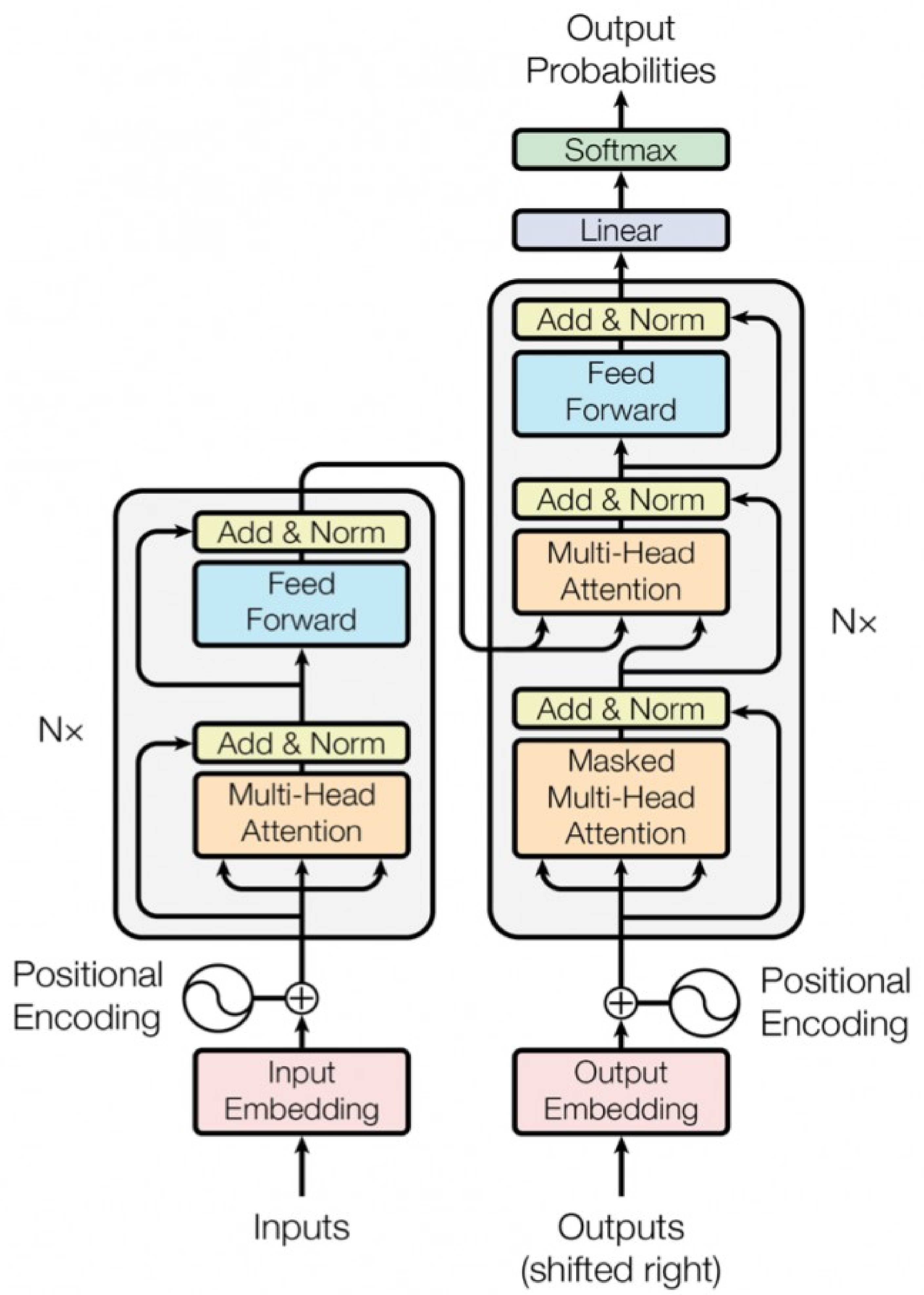
Flickr Dataset

Image Processing:
Load Images
Resize
Preprocessing using IV3

Clean Vocabulary size

Feature Extraction using InceptionV3

- InceptionV3 is employed to extract rich visual features from images in the dataset.
- : Reads and decodes image files, resizes them to a standard size (299x299), and preprocesses them using the InceptionV3 preprocessing function.
- We leverage the InceptionV3 model pre-trained on ImageNet to benefit from its learned representations.
- Using only the last layer so as to obtain a set of high-level features that the Inception V3 model has learned.
- It alllows the model to capture complex patterns and relationships in the input images.



Transformer Architecture

Thank you!