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)

DEPARTMENT OF COMPUTER ENGINEERING

Image Captioning

Ajay Arvind Nirmal	Roll No.: 119	Division: B
Suhas Keshava Murthy	Roll No.: 112	Division: B
Mrunal Mahendra Misale	Roll No.: 106	Division: B

Under the Guidance of:
Prof. Merlin Priya Jacob

Problem statement

This is a model which is responsible for generating captions that describes the images that have been provided to the model as an input by the user.

Objective

- The goal of creating a frame-by-frame caption generator using CNN and Transformer Transformer is to automatically generate descriptive captions standalone images.
- The model should understand the visual content of each frame and generate coherent and meaningful textual descriptions.
- The model should be flexible enough to handle both images generating captions for single frames and frames sequentially, considering temporal context.
- A systematic approach and deep understanding of computer vision and natural language processing techniques are required to achieve this objective.

Scope and Features

Image Captioning:

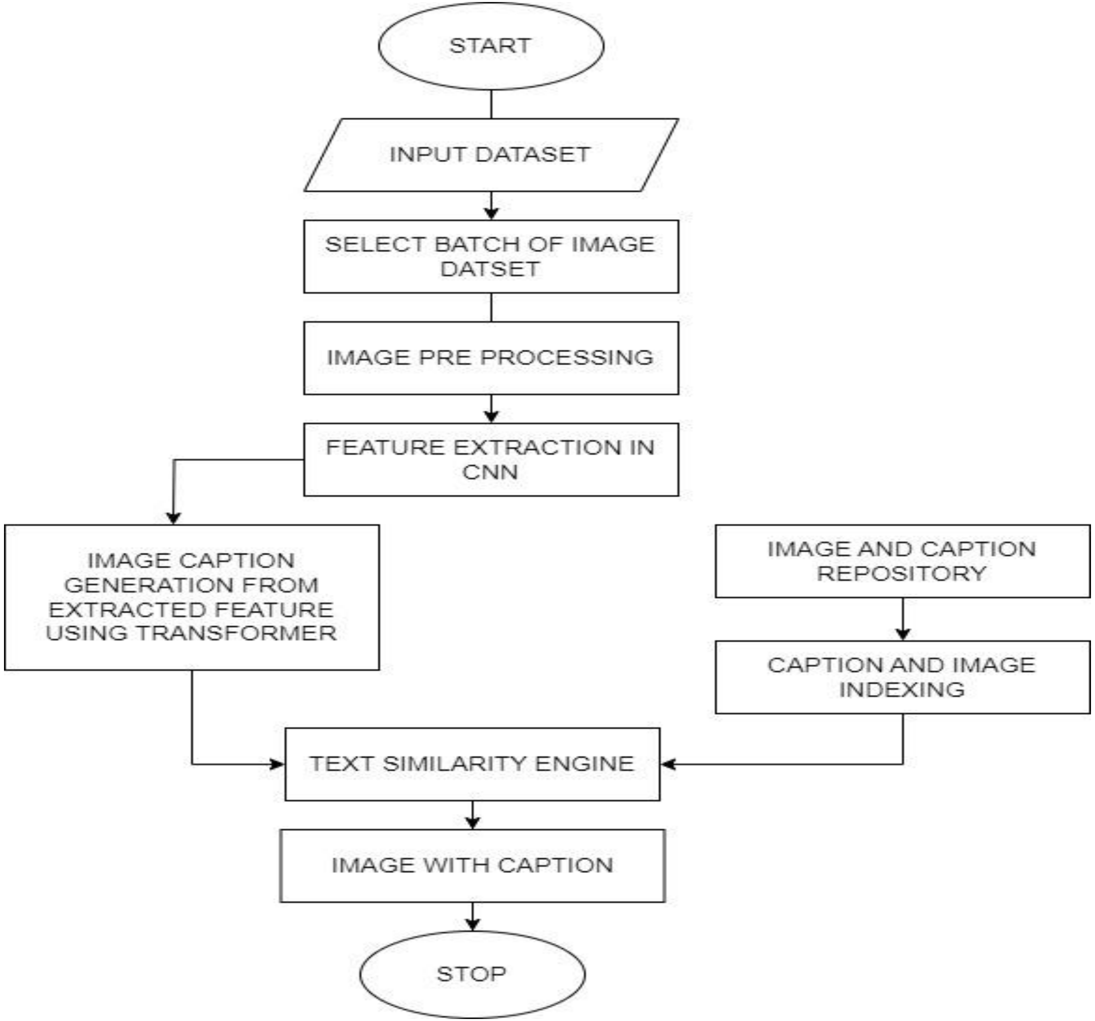
- It describes the image entered by the user.
- It uses a CNN to extract features from the image.
- It uses a Transformer to generate the caption based on those features.
- It is trained on a dataset of 8,000 images, each with five different captions.

Image Captioning

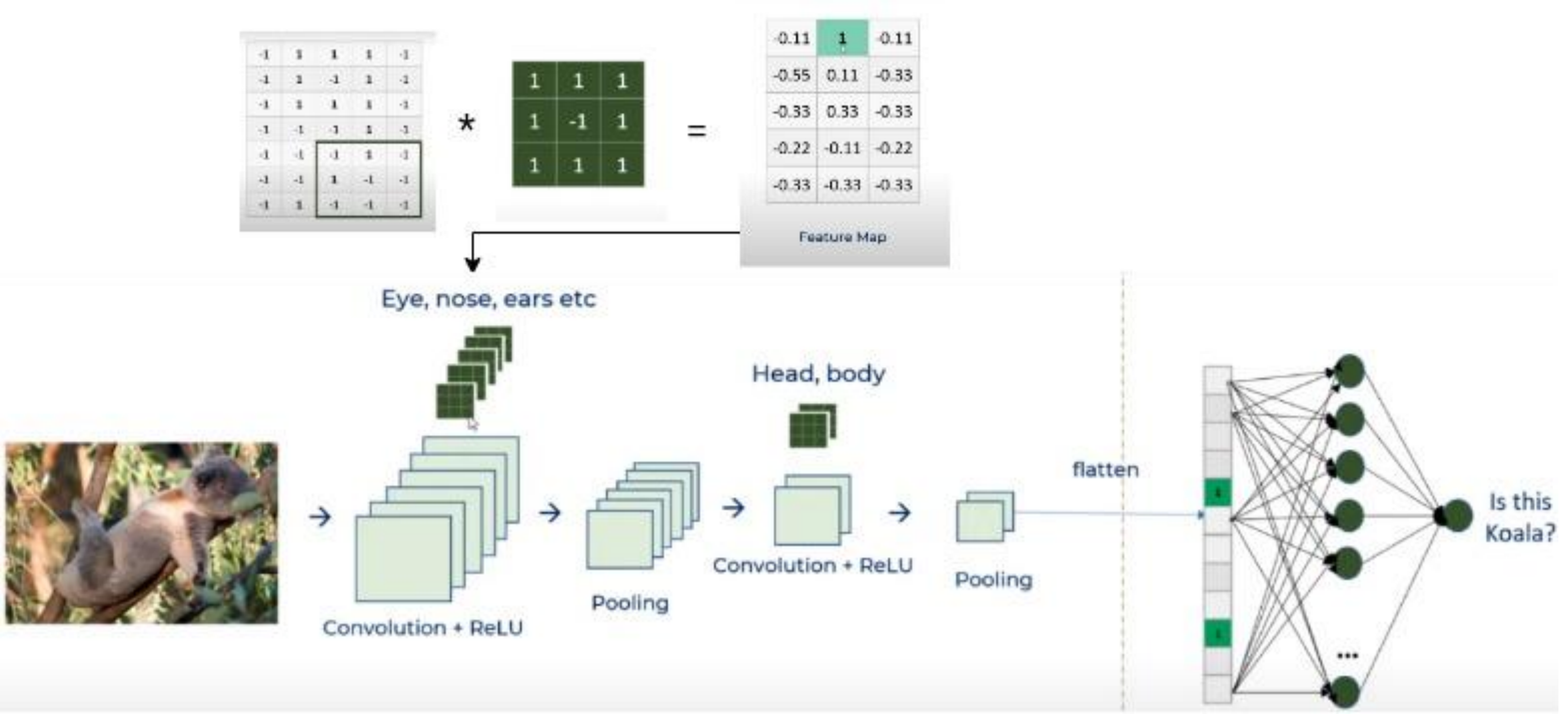
Name	Conclusion	Findings
Chetan Amritkar, Vaishali Jabade“Image Caption Generation using Deep Learning Technique”, Department of EnTC.	Artificial Intelligence (AI) generates image content using computer vision and Natural Language Processing (NLP). Regenerative neural models, including CNN and RNN, extract features and generate sentences, ensuring accuracy and smoothness.	CNN, Ensuring Accuracy and Smoothness
Ansar Hani, Najiba Tagougui, Monji Kherallah“Image Caption Generation Using A Deep Architecture” .	This paper presents a model combining computer vision, natural language processing, and machine learning to generate natural language captions using convolutional neural networks and attention mechanisms.	About Computer Vision, Natural Language processing and Machine Learning.

Name	Conclusion	Findings
Suganthe Ravi Chandaran, Shanthi Natesan, "Image Captioning Using Deep Learning Techniques for Partially Impaired People," Computer Science and Engineering Department, Kongu Engineering College (KEC), Perundurai, Erode, India.	Image captioning uses encoding and decoding structures, but current models struggle with gradient explosion. A new model, YoLOv5 and Bidirectional LSTM, addresses this issue. Tested on Flickr8k, it outperforms other methods and achieves a 0.7 BLEU score.	Encoding - Decoding, Flickr8k dataset.
Quan Sun1, Qiying Yu, “Generative Pretraining in Multimodality”.	Emu is a Transformer-based multimodal foundation model that generates images and texts in multimodal contexts. It uses a one-model-for-all autoregressive training process, enabling exploration of diverse pretraining data sources and demonstrating excellent performance in zero-shot tasks.	

Flowchart


START
INPUT DATASET
SELECT BATCH OF IMAGE DATSET
IMAGE PRE PROCESSING
FEATURE EXTRACTION IN CNN
IMAGE CAPTION GENERATION FROM EXTRACTED FEATURE USING TRANSFORMER
IMAGE AND CAPTION REPOSITORY
CAPTION AND IMAGE INDEXING
TEXT SIMILARITY ENGINE
IMAGE WITH CAPTION
STOP

CNN



-1	1	1	1	-1
-1	1	-1	1	-1
-1	1	1	1	-1
-1	-1	-1	1	-1
-1	-1	-1	1	-1
-1	-1	1	-1	-1
-1	1	-1	-1	-1

*

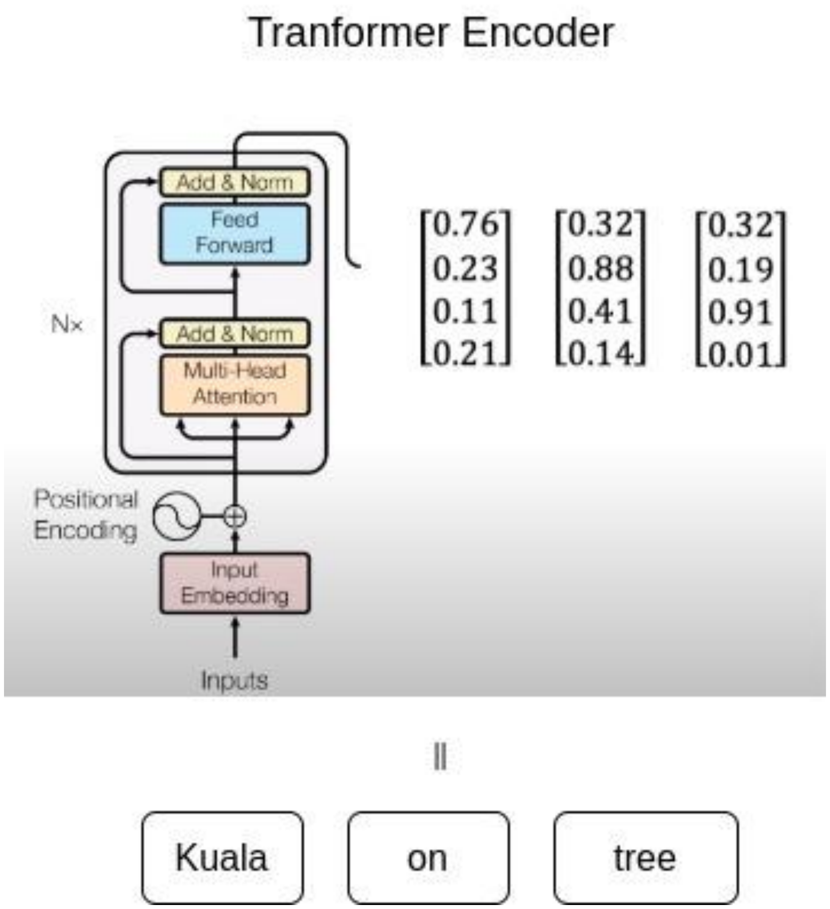
1	1	1
1	-1	1
1	1	1

=

-0.11	1	-0.11
-0.55	0.11	-0.33
-0.33	0.33	-0.33
-0.22	-0.11	-0.22
-0.33	-0.33	-0.33

Feature Map

Transformer Encoder



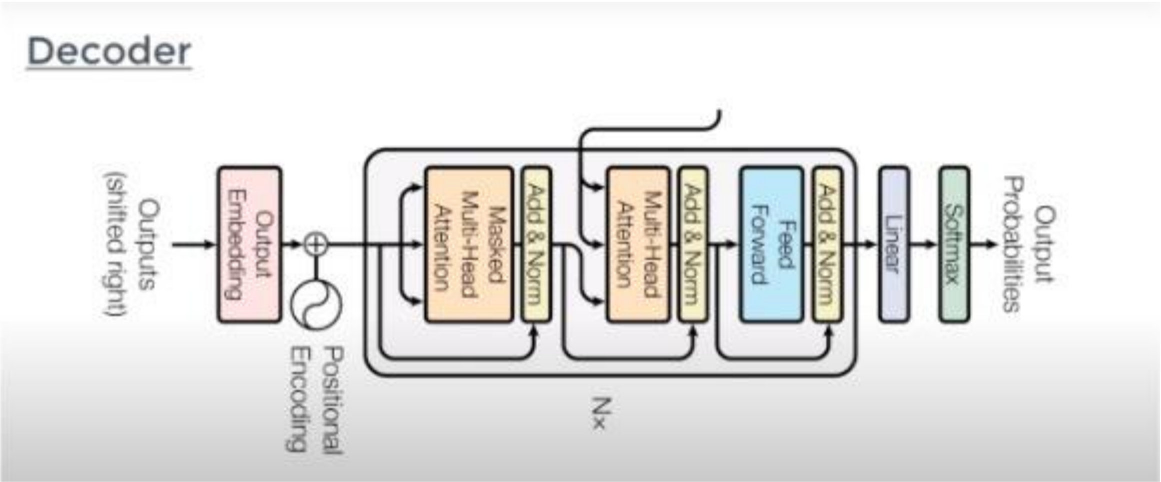
$$\begin{bmatrix}0.76\\0.23\\0.11\\0.21\end{bmatrix} \quad \begin{bmatrix}0.32\\0.88\\0.41\\0.14\end{bmatrix} \quad \begin{bmatrix}0.32\\0.19\\0.91\\0.01\end{bmatrix}$$

Transformer Decoder

Kuala =

on =

tree =



Kuala on a tree

Thank You !