**Project Title**- “Image and Video Caption”

Abstract- Image and Video Caption is a system based on deep learning for generating captions for images and videos automatically with descriptions in natural human language. Due to the rapid proliferation of multimedia content, there is an increasing desire for tools to achieve a level of accessibility, comprehension of the content, and indexing with respect to it. This project thus seeks to bridge the domain of visual data to natural language processing by building a commendable captioning system. The system builds on an advanced computer vision solution and a natural language generation model to make sense of the visual content and provide human-readable descriptions. On very large datasets containing paired images/videos and descriptive texts, the system generates captions after being trained. Apart from recognizing objects and actions in images, the system accounts for some levels of contextual information for the generation of semantically meaningful captions.

A major challenge remains the modelling of spatial and temporal relationships in videos without losing fluency and contextual coherence in the captions. Our approach aims to address these by leveraging advanced models grounded in Convolutional Neural Networks (CNNs) for feature extraction, Long Short-Term Memory (LSTM) networks for sequence modelling, and a Transformer-based architecture for context-rich caption generation.

* **Project Objectives-** The project aims to design an automated system that would be capable of generating captions in natural language for images and videos. It aims at higher accuracy and contextual relevance while describing visual content. This would improve accessibility to multimedia data and would enable efficient retrieval and understanding of content within it.
* **Methodology-**
* Data collection and preprocessing
* Feature extraction using CNNs
* Sequence modeling with LSTM networks
* Transformer-based caption generation
* Evaluation using BLEU, METEOR, and CIDEr metrics
* Key findings-
* The transformer models improved the accuracy of the captions.
* There was a greater understanding of the context of the given video captions.
* They also maintained spatial and temporal dependencies effectively.

Step wise solution approach-

* Step 1:Data Acquisition and Pre-Processing
* Collect a large-scale dataset of images, videos, and their associated captions.
* Clean and pre-process the data, including the resizing of images, extraction of video frames, and tokenization of text captions.
* Step 2:Feature Extraction: With the aid of some pre-trained Convolutional Neural Networks (e.g., ResNet, Inception), visual features are extracted from images and video frames.
* Step 3: Caption Generation: The caption generation process is also called sequence modelling by LSTM networks with attention mechanism.
* Step 4: Integration of Temporal Features for Video Captioning: Using 3D-CNNs or Temporal Attention Network techniques to model temporal dependencies across video frames.
* Step 5: Model Training and Evaluation: Train your captioning model on image/video-caption pairs. Evaluate your captions using quantitative metrics such as BLEU, METEOR, and CIDEr.

Reference:

https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=4934726

Team:

1. Umangi Nigam-E22CSEU0526

2. Tushar Swarnkar-E22CSEU0436