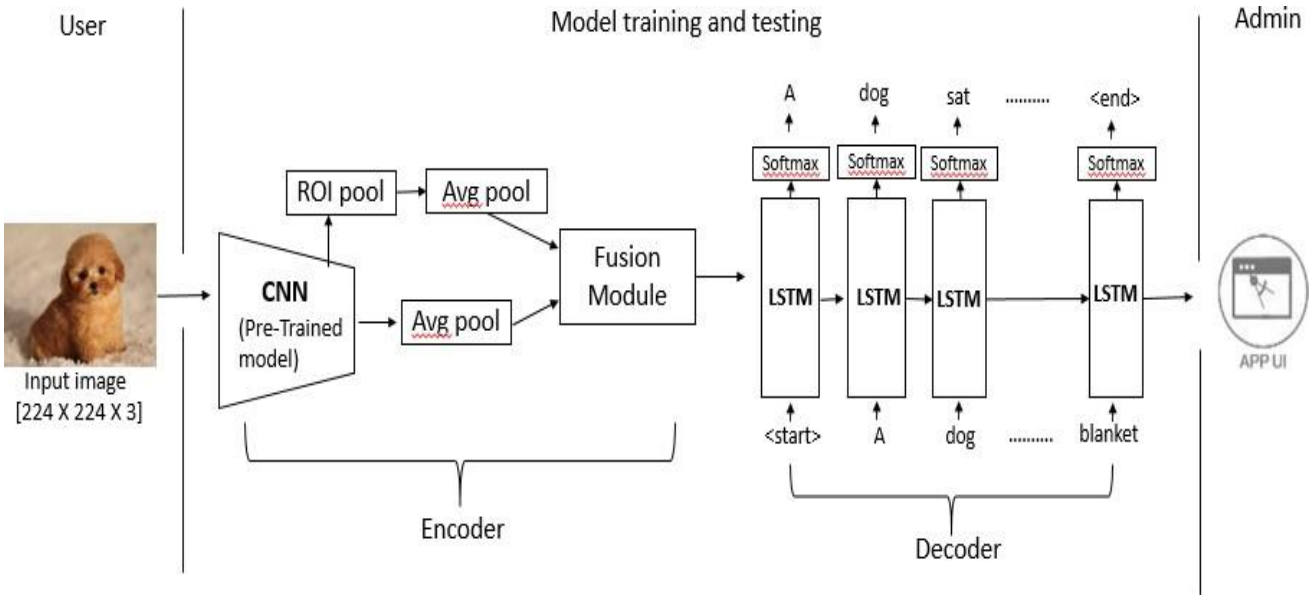
Project Design Phase-II Technology Stack
(Architecture & Stack)

Date	08th November 2023
Team ID	Team - 591718
Project Name	**Image caption Generation**
Maximum Marks	4 Marks

Technical Architecture:



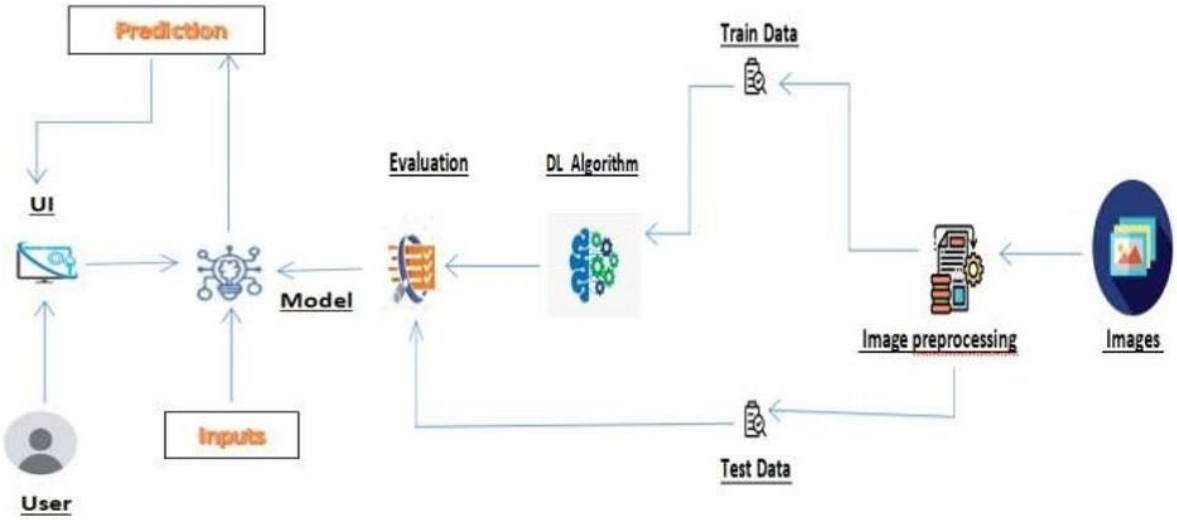


Table-1: Components & Technologies:

S.No	Component	Description	Technology
1.	Convolutional Neural Networks (CNN)	CNN are used for feature extraction from the input image. These networks are designed to identify and extract relevant visual features from the image, which serve as the foundation for generating captions.	VGG16, ResNet, and Inception
2.	Recurrent Neural Networks (RNN)	RNN are used for generating sequential data like natural language captions. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are commonly employed RNN architectures for image captioning. They take the visual features from the CNN as input and generate words one at a time.	TensorFlow, PyTorch, Keras,GPU,TPU
3.	Encoder-Decoder	The encoder, typically the CNN, encodes the image into a fixed-length feature vector, while the decoder, usually an RNN, generates the caption based on this encoded information.	TensorFlow, PyTorch, Keras,GPU,TPU

4.	Pretrained Word Embeddings	These embeddings help the model understand the relationships between words and provide a foundation for generating coherent and contextually accurate captions.	Word2Vec, GloVe, fastText
5.	Dataset	Image captioning models require a large dataset of images paired with human-generated captions for supervised training.	MS COCO, Flickr30k, and Pascal VOC
6.	Evaluation Metrics	They compare the generated captions to reference captions and measure their similarity.	BLEU (Bilingual Evaluation Understudy), METEOR, CIDER, and ROUGE
7.	Transfer Learning	models pretrained on large text and image datasets can be fine-tuned on specific captioning tasks.	fine-tuning pretrained models
8.	Reinforcement Learning	This involves providing rewards based on the quality of generated captions and adjusting the model's parameters accordingly.	TensorFlow, PyTorch, Keras,MDP,RLlib
9.	Machine Learning Model	Purpose of Machine Learning Model	Object detection Model, etc.
10.	API	Application Programming Interface	Microsoft Azure Computer Vision, Google Cloud Vision, IBM Watson Visual Recognition

Table-2: Application Characteristics:

S.No	Characteristics	Description
1.	Image Understanding	The application needs to comprehend and analyze the content of images, identifying objects, scenes, and other visual elements.
2.	Natural Language Generation	It should be capable of generating coherent, contextually relevant, and human-like textual descriptions based on the image content.
3.	Context Awareness	Understanding the context and relationships among objects in the image is crucial for generating meaningful captions.

4.	Multimodal Fusion	Some applications benefit from combining information from both text and images, requiring the fusion of these modalities to improve caption quality.
5.	Multilingual Support	Support for generating captions in multiple languages, depending on the target audience.
6.	Customization	The ability to fine-tune or customize the model for specific domains or to improve caption quality for a particular dataset.