SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

RAMAPURAM, CHENNAI – 600 089

FACULTY OF SCIENCE & HUMANITIES

(2023-2025 BATCH)

MINI PROJECT

Visual Analytics For Efficient Image -To -Text Prediction Based On Visually Aware Context Learning

PATEL SAI KRUPA

RA2332241020062

MCA-B

AGENDA

- ABSTRACT
- INTRODUCTION
- MODULES DESCRIPTION
- NECESSARY DIAGRAMS
- DATABASE DESIGN

ABSTRACT

- Captioning is process of assembling a description for an image.
- Our method solves this by refining attention to finer details.
- Introduced a dual-attention module designed to independently process features from distinct classes, thereby enhancing model's ability to discern complex scenes.
- Our methodology advances current practices by extracting feature vectors from segmented regions, enabling a more enhanced understanding of image components.
- In the realm of image-to-text prediction, efficiently bridging the gap between visual content and textual description remains a significant challenge. This paper introduces a novel approach termed Visual Analytics for Efficient Image-to-Text Prediction, which enhances predictive accuracy by leveraging Visually Aware Context
- Learning (VACL). VACL integrates visual analytics techniques with advanced contextual understanding to
- improve the alignment between images and their corresponding textual descriptions.

INTRODUCTION

1

Digital images are becoming increasingly ubiquitous in our current age of tech devices, such as smart phones, computers, home security systems, smart watches, etc., most of which contain some sort of camera.

2

Image captioning is research hotspot of computer vision technology, and it is one of the main tasks to realize the scene understanding.

3

Image captioning is based on detecting and recognizing objects, reasoning the relationship of the detected objects, and using natural language to describe the semantic content of the scene image.

4

Image and description text are two different representation manners, they are symmetric and unified in the semantic content of the same visual scene.

MODULES DESCRIPTION

➢ Region Level Feature:

- This module is foundational for understanding the spatial relationships within an image, a crucial aspect for accurate caption generation.

• It begins by employing target-detection techniques to identify objects, their attributes, and their interconnections within the image.

• These features were then used to guide the Long Short-Term Memory (LSTM) in generating the caption.

• Instead of directly extracting grid-level features, the target-detection technique is now used to extract region- level features of the image.

• The system initially identifies the correlation between the object, its attributes, and other objects in the image using target-detection methods.

➢ Domain Object Pre– Filtering :

- Input order of image objects and image region information in filter captioning model is modified by utilizing domain object dictionary during inference of image captions.

• Currently, to highlight objects displayed in domain object dictionary, image object tag is duplicated and included in repetitions based on value specified by number of repetitions parameters.

• Main objects that are removed due to addition of repeated objects are deleted in reverse order.

• By rearranging order of tags instead of replacing and deleting them, majority of image object tags are included in filter captioning model.

• This module effectively enhances interpretability and relevance of the generated captions, improving overall performance

➢ Text Generation :

• Text generation lies at the heart of the captioning task, and this module delves into the intricacies

generating accurate and diverse descriptions of visual content.

• Traditional text generation methods often rely on deterministic approaches like greedy search or beam search, which may produce repetitive or monotonous Captions.

• By introducing randomness into generation process, model can produce more diverse and creative captions that better capture enhances of visual content.

• This approach efficiently aligns text and picture information inside a single semantic space by using multi-model representation learning methods, which makes caption creation more coherent and contextually appropriate.

➢ Visually-Aware Context Network :

- The objective of the text semantic enhancement network is to extract improved semantic representations of text by capturing semantic associations, enhancing semantic features, and reducing the impact of modal features. The system comprises two primary elements: a text encoder and a semantic enhancement module. The text encoder transforms the input text sequence into a semantic space, producing an initial semantic representation.
- In order to improve the sentences, text enhancement techniques such as rule enhancement and semantic enhancements are utilized. The network model is used to encode the improved text sequence into a vector representation of a fixed length.
- The Multilayer Perceptron (MLP) is composed of several fully connected layers, which allows for the extraction of complex features and semantic information. By incorporating residual connectivity, the problems of gradient disappearance are mitigated and the expressiveness of feature representation is enhanced, resulting in an improvement in the quality of semantic representation.

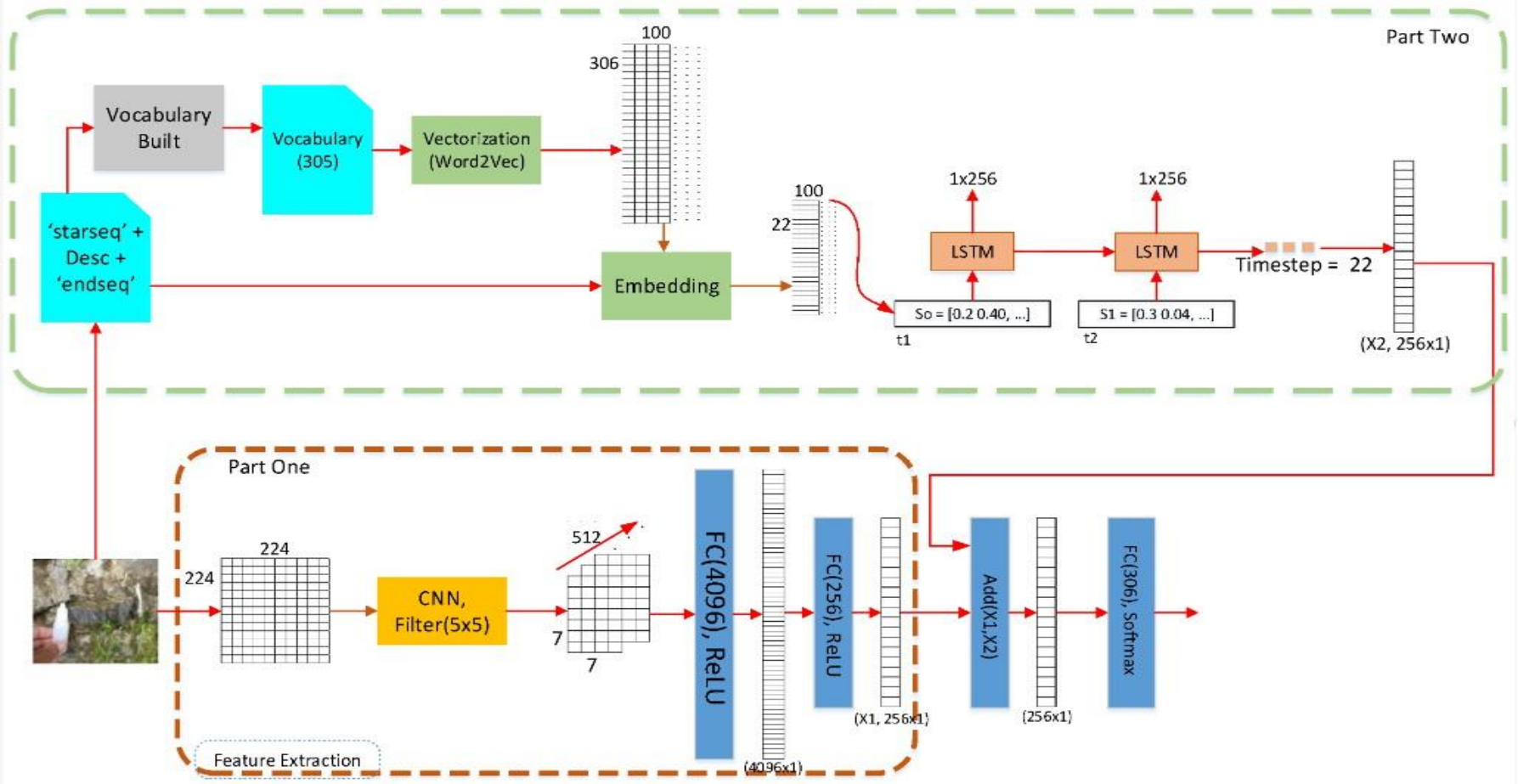
➢ Evaluation Metrics :

- In accordance with the prevailing convention in the literature, we conduct an assessment using BLEU-1 (B@1), BLEU-4 (B@4), METEOR (M), ROUGE-L (R-L), CIDEr, and SPICE. BLEU-1 (B@1) and BLEU-4 (B@4) evaluate the accuracy of 1-gram and up to 4-gram sequences, respectively. These metrics are commonly represented as percentages, ranging from 0 to 100%. The METEOR metric, which ranges from 0 to 1, offers a well-balanced evaluation of precision and recall by considering synonymy. (c) ROUGE-L is a dimensionless metric that measures the extent of overlap between the longest common subsequence.
- It has a range of 0 to 1. CIDEr is a metric that measures the agreement between human captions and generated descriptions. It is usually scored on a scale of 0 to 10 or higher. Finally, the SPICE metric evaluates the semantic propositional content on a scale of 0 to 1. In the academic literature, authors frequently multiply these values by 100 to convert them into percentages and present them as two-digit numbers, aiming to enhance intuitive comprehension. This practice ensures that the results are in accordance with a widely accepted convention that is familiar to many readers.

ARCHITECTURE DIAGRAM

➢ General Architecture:

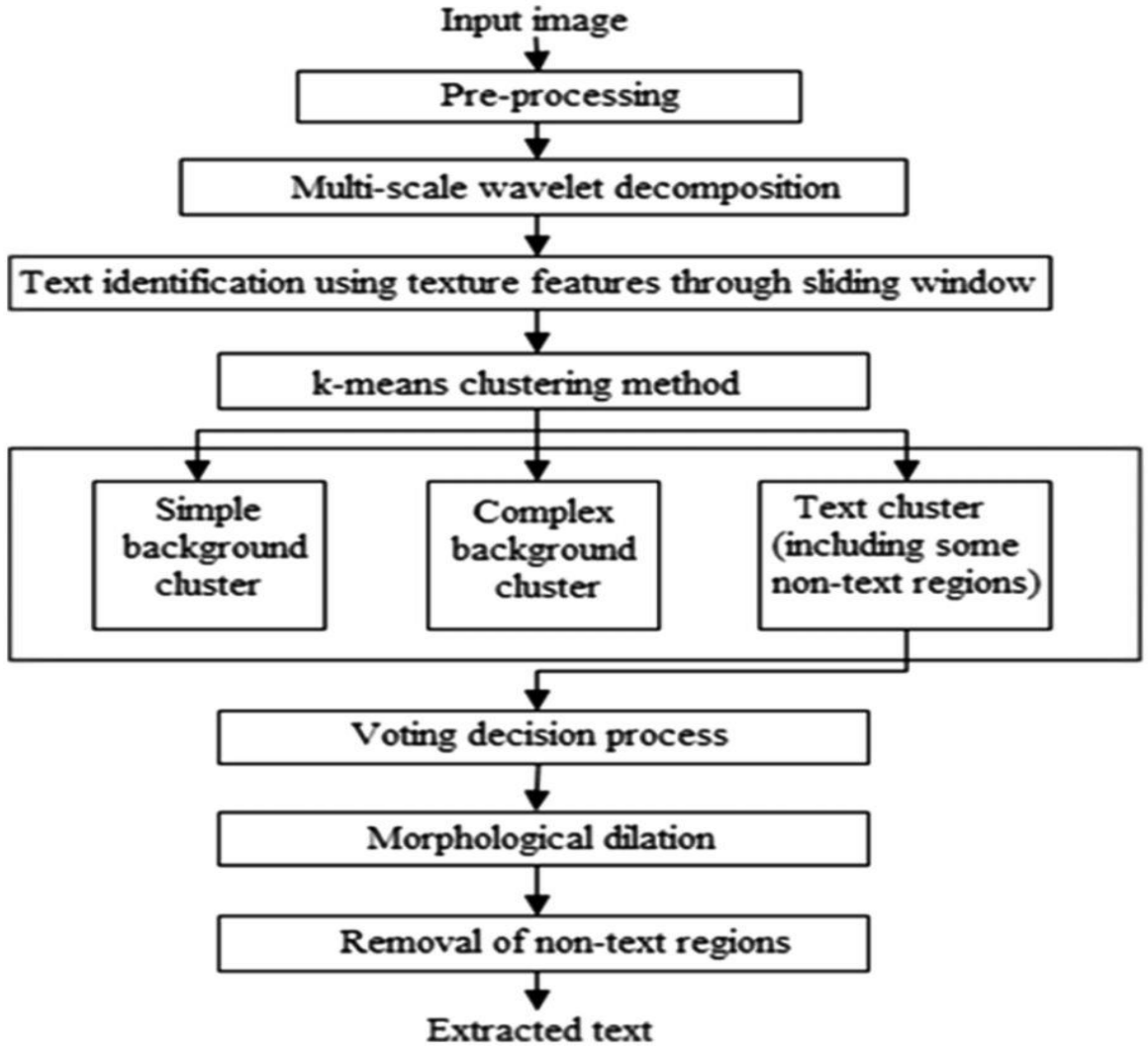
It represents the architecture diagram of the project. The system initially identifies the correlation between the object, its attributes, and the object in the image using target-detection techniques. Subsequently, it proceeds to create a graph structure. The Graph Convolutional Network (GCN) was employed to extract the structural features of the graph.



DESIGN PHASE

➢ **Data Flow Diagram :**

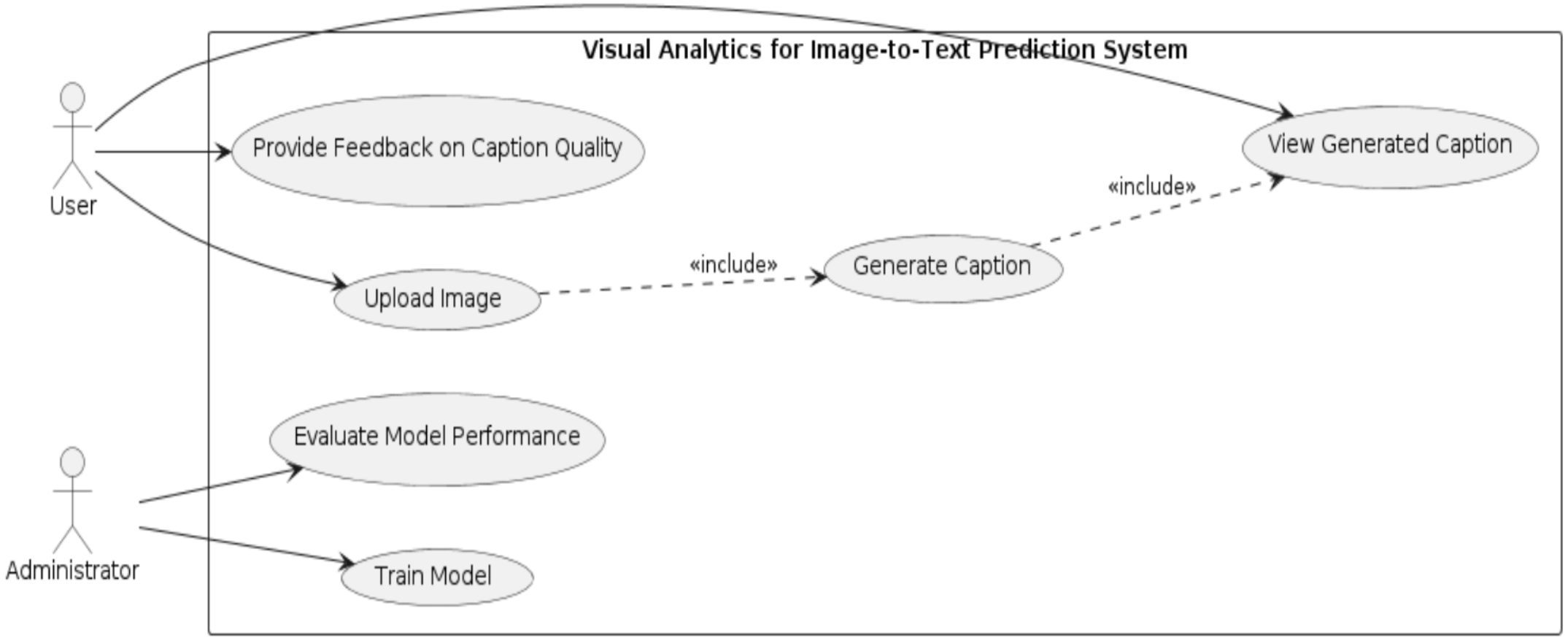
The process starts with a request from a client. This request is handled by the Client Handler, which is a Decision Maker component. The Client Handler checks if there are any available cloudlets (small cloud computing nodes) to offload the task



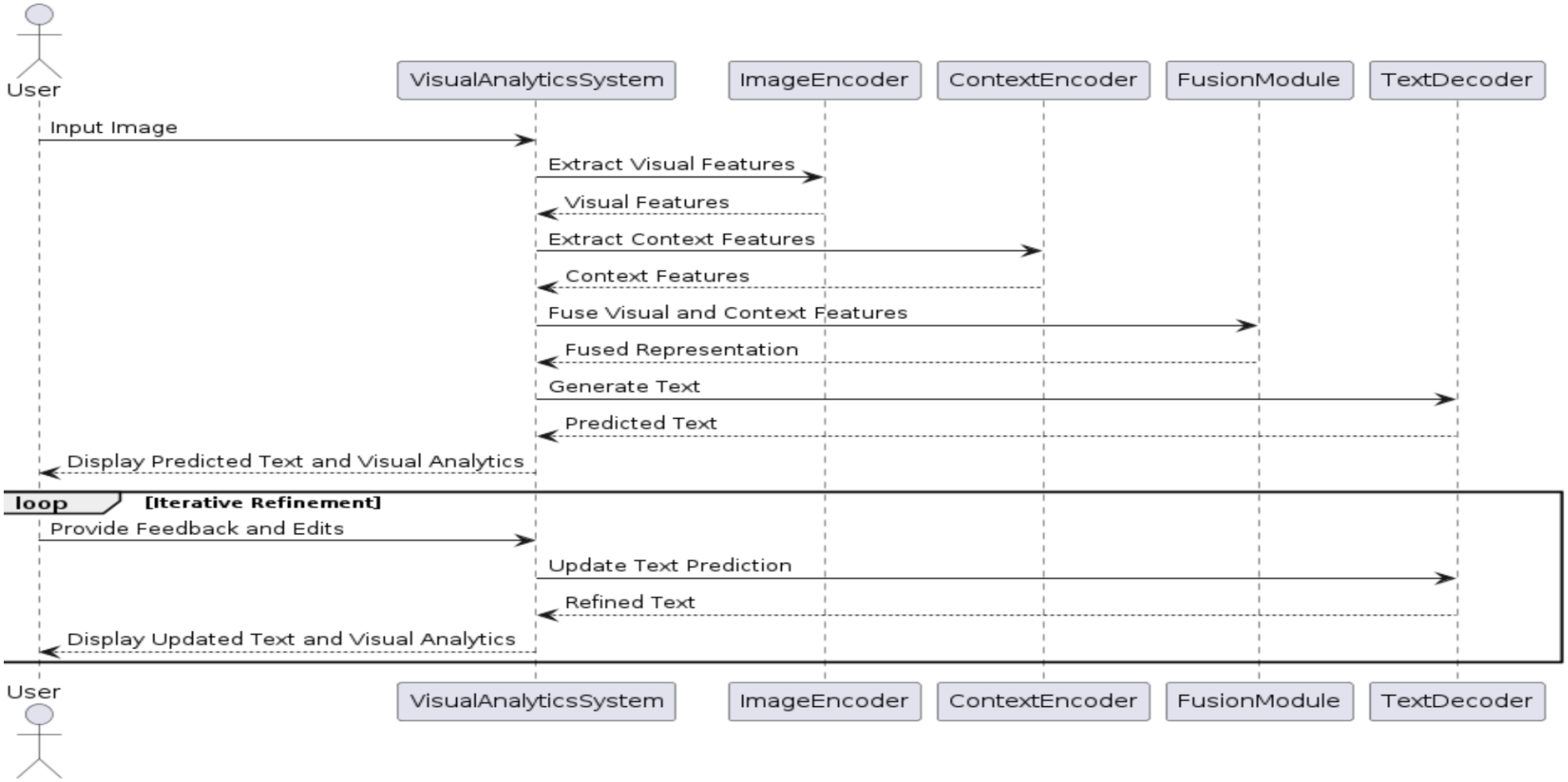
UML DIAGRAMS

- **Activity Diagram :**

➢ Use Case Diagram :



➢ Sequence Diagram :



THANK YOU 🙏