Informatics Institute of Technology

In Collaboration With

University of Westminster, UK

University of Westminster, Coat of Arms

AURIX: Augmented Multimodal Knowledge Integration for Adaptive Zero-Shot Scene Understanding

A dissertation by

Malith Srineth Amarawickrama

W1867105 | 20210353

Supervised by

Mr. Guhanathan Poravi

April 2025

This Dissertation is submitted in partial fulfilment

of the requirements for the BSc(Hons) Software Engineering degree

at the University of Westminster.

ABSTRACT

Effectively understanding complex scenes in real-time remains a major challenge in computer vision, especially with unseen or dynamic elements. Traditional models struggle to generalize across environments and lack contextual enrichment from visual, textual, and knowledge-based data. Scalability and computational efficiency are also significant barriers, particularly for large-scale real-time applications. This research proposes an adaptive scene understanding model that integrates zero-shot learning with multimodal data like visual, text, and knowledge graphs as external knowledge to enhance performance and adaptability.

This project presents a multimodal framework for adaptive zero-shot scene understanding using visual, textual, and knowledge graph data. Embeddings from these modalities are mapped into a shared semantic space. When given an input image, visual features are extracted. These features are then processed by a zero-shot learning model, which uses shared embeddings to identify unseen objects based on semantic similarity. Finally, the zero-shot learning model's predictions are sent to a generative model like Flan-T5, which creates a caption describing the scene.

This research successfully designed, implemented, and tested the system, integrating multimodal data and external knowledge sources like ConceptNet. The system achieved a BLEU-4 score of 49.8%, CIDEr 112.4, and ROUGE-L 65.3%, demonstrating strong caption generation. ConceptNet integration improved contextual relevance with a 94% query success rate and 2.3 additional relevant concepts per caption. Further optimizations, including improved knowledge extraction, batch processing, and UI enhancements, will enhance efficiency and scalability, ensuring real-world applicability in multimodal AI.

Subject Descriptors

Computing methodologies → Artificial intelligence → Computer vision → Computer vision tasks → Scene understanding

Computing methodologies → Machine learning → Learning paradigms → Multi-task learning → Transfer learning

Keywords

Zero-Shot Learning, Multimodal, Scene Understanding, Computer Vision, Knowledge Graph