FACULTY OF TECHNOLOGY
Information & Communication Technology
Subject: Capstone Project

Name: Fenil Vadher

En Roll no: 92200133023

Subject: Capstone Project

Project Title: Multimodal Movie Script Search: Context-Aware Dialogue and Scene Retrieval

Introduction

Movies and web series contain vast multimodal information where **visual scenes and textual dialogues** are closely interlinked. Traditional search systems in multimedia archives often focus on **single modalities**, such as text-only search or image-only retrieval, which limits their ability to capture context and narrative flow. With the rise of **AI and multimodal learning**, there is an urgent need for intelligent systems capable of cross-modal retrieval that aligns **dialogues with scenes** and vice versa. This project proposes a **context-aware multimodal movie script search engine** that enables **dialogue-to-scene, scene-to-dialogue, and multimodal contextual queries**, bridging the gap between natural human queries and multimedia archives.

Problem Statement

Current multimedia search engines do not effectively capture the **semantic alignment between dialogues and scenes** in movies/web series. This limitation results in poor retrieval accuracy when users attempt to search for a **specific scene using dialogue** or vice versa. The lack of context-aware multimodal systems hinders applications in **content recommendation, video indexing, summarization, and intelligent media understanding**.

Objectives

1. **Develop a multimodal retrieval system** capable of bidirectional search between dialogues and scenes.
2. **Integrate contextual fusion** of dialogue and visual cues for improved retrieval accuracy.
3. **Evaluate system performance** using established metrics (BLEU, METEOR, ROUGE-L, CIDEr, CLIP-SIM, Precision, Recall).
4. **Design a scalable ICT solution** (Flask backend + React frontend) for real-world usability.
5. **Contribute a reusable dataset** (custom JSON dataset of movies with scene–dialogue pairs).

Relevance to ICT Domain

This project lies at the intersection of **Artificial Intelligence, Machine Learning, Computer Vision, and Natural Language Processing (NLP)**.:

- **Multimodal AI frameworks** for cross-modal search.
- **Information retrieval in ICT** to improve multimedia content management.
- **Software engineering practices** (Flask + React full stack system) for scalable deployment.
  The system aligns with current ICT trends in **AI-driven media analytics, content-based retrieval, and smart entertainment systems**.

Feasibility Analysis

a) Technical Feasibility

- **Tools/Frameworks:** HuggingFace pre-trained models (Vid2Seq, BLIP-2, GIT2, CLIP), Python (Flask), ReactJS, PyTorch, NumPy.
- **Dataset:** Custom JSON dataset + publicly available datasets (LSMDC, MSR-VTT).
- **Suitability:** Pre-trained models reduce training cost while ensuring high-quality embeddings.

b) Economic Feasibility

- Open-source models and frameworks → **zero licensing cost**.
- Storage & computation → possible via **Google Colab/Kaggle or low-cost GPU services**.
- Affordable within academic project constraints.

c) Ethical Considerations

- **Data Privacy:** Use only open datasets and avoid copyrighted raw media.
- **Bias:** Ensure diverse datasets (Hollywood + Bollywood) to avoid cultural bias.
- **Responsible Use:** Restrict system usage to **research and educational purposes**.

Market/User Needs Analysis

The system targets:

- **Film researchers** (narrative and script analysis).
- **Streaming platforms** (improving content search).
- **Educators/students** (media and film studies).
- **AI researchers** (benchmark for multimodal retrieval).

Supporting Sources:

1. Rohrbach et al., Movie Description Dataset (LSMDC), ICCV 2015.
2. Xu et al., MSR-VTT: A Large Video Description Dataset, CVPR 2016.
3. Radford et al., Learning Transferable Visual Models with CLIP, NeurIPS 2021.
4. Li et al., BLIP-2: Bootstrapping Language–Image Pretraining, ICML 2023.
5. Google DeepMind, Vid2Seq: Video-to-Text Models, 2023.

Literature Review (Brief)

- **Unimodal retrieval** approaches (TF-IDF, CNN-based) fail to capture multimodal context.
- **CLIP (Radford et al., 2021)** introduced joint vision–language embeddings, enabling cross-modal search.
- **Vid2Seq (Google, 2023)** advanced video-to-text generation, improving temporal understanding.
- **BLIP-2 (Li et al., 2023)** and **GIT2** further refined multimodal pretraining for retrieval and generation.
  However, none of these directly address **context-aware movie retrieval with custom datasets**, making our approach novel.

Conclusion

This project proposes a **novel multimodal framework** for intelligent movie understanding by integrating **scene–dialogue retrieval** with **contextual multimodal queries**. The system is technically feasible, economically sustainable, and ethically sound. By combining cutting-edge **AI/ML models, NLP, and computer vision**, it directly addresses pressing ICT challenges in multimedia retrieval and contributes both a **working prototype** and a **custom dataset** to the community.