Final Project

Video Retrieval System

Mahdi Khoursha, Morteza Hajiabadi

Information retrieval systems hold a special place in today's world, given the vast amount of information published across resources like the internet and the importance of time efficiency and accuracy in their functionality. These systems sift through extensive data to find records that best match the user's needs. High accuracy in such systems helps users achieve the best results in the shortest possible time, enhancing their user experience.

One of the most significant media formats with which users interact frequently is videos. The high volume of video content published on virtual platforms means users often forget the content quickly. However, they may remember vague details or only parts of certain scenes. Therefore, a system is needed that can, based on a scene description provided by the user, find and suggest the relevant video containing that scene.

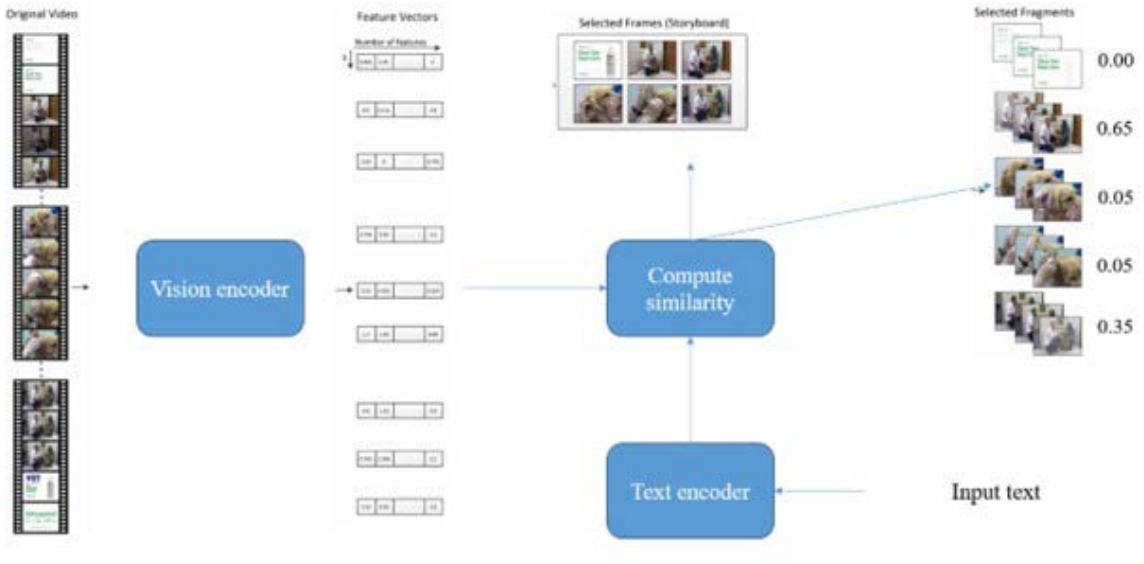


Figure 1: An overview of the proposed system.

This project involves creating a Video Retrieval System that allows users to search for specific video sequences by providing textual descriptions. The system should support Persian language input and return the top 5 relevant videos based on a given description.

This project is divided into three main phases:

1. **Research and Data Collection**
2. **Model Training and System Implementation**
3. **Final Phase and Presentation**

1 Research Phase and Data Collection

In this phase, you will gather and prepare the data needed to build and train the model. This includes:

1.1 Video Dataset Creation

Curate a dataset of videos from which you will retrieve specific scenes based on Persian descriptions. We also recommend that you use scenes from **soccer** games instead of scenes from movies and series. In other words, instead of retrieving scenes from movies and series, retrieve the events of soccer games, which can be an interesting topic if you can get the necessary data.

If no pre-existing annotations are available, manually create or obtain textual descriptions for key frames or sequences within the videos, ensuring each description accurately represents the associated visual content.

1.2 Scene Description Collection in Persian

Since this project requires input in Persian, ensure all textual descriptions are translated or generated in Persian. If only English descriptions are available, use machine translation tools (e.g., Google Translate) to translate descriptions and validate their accuracy.

Additional resources, such as DistilUSE for evaluating similarity, may assist in verifying translation quality by measuring alignment between original English and translated Persian captions.

1.3 Frame Extraction

Extract frames from the videos at specific intervals, or based on scene changes. Each frame (or set of frames) should be associated with a corresponding Persian description to facilitate training.

1.4 Data Preprocessing

Preprocess and clean the data, including text normalization for Persian (tokenization, removing stop words, etc.)

Format the data for compatibility with your chosen model, ensuring that frames and descriptions are correctly aligned for training.

(a) Godfather Hand Kiss

(b) Hand of God

Figure 2: Example of scene and event

2 Model Training and System Implementation

With your data prepared, the next phase involves model selection, training, and implementation of the retrieval system.

2.1 Model Choice

We recommend using the CLIP model (Contrastive Language-Image Pre-training) for this project. CLIP aligns text and image embeddings, which allows for efficient similarity matching between descriptions and frames.

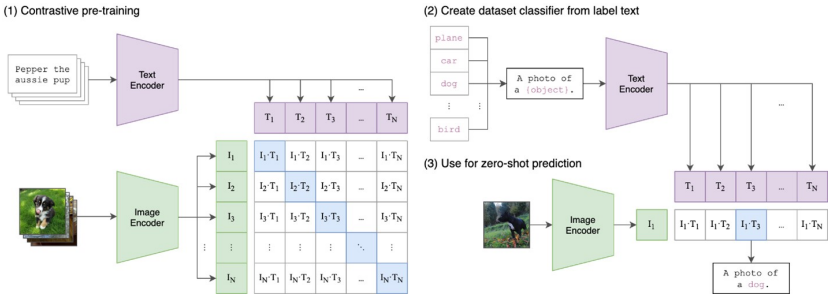


Figure 3: CLIP model architecture.

You may choose alternative models if they support the task requirements and demonstrate effectiveness with text-to-image matching.

2.2 Model Training and Fine-Tuning

Fine-tune your model using the collected Persian-language scene descriptions. Experiment with hyperparameters (e.g., learning rate, batch size) to optimize performance.

Encode each video frame into a vector representation, which captures the essence of each scene and allows for similarity comparisons with text descriptions.

2.3 System Workflow

Encode each frame and store it in a vector database (e.g., Qdrant) for quick retrieval.

During inference, the text description is converted into a vector embedding. This vector is used to retrieve matching frames from the database.

The challenges in this project for which you need to provide suitable solutions:

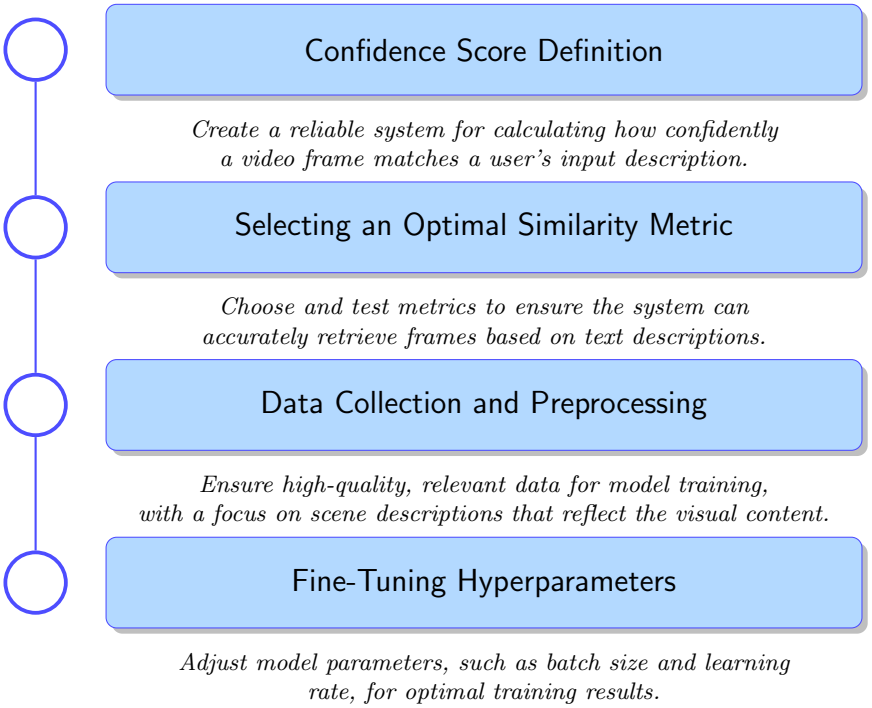


Figure 4: The challenges in this project

3 Final Phase and Presentation

In this final phase, you will present your project, demonstrating your model's performance and explaining your design choices.

3.1 Project Presentation

3.1.1 System Demonstration

Showcase the system's ability to return relevant video sequences based on a user's text description.

Explain how your model processes the description, retrieves matches, and ranks results by relevance.

3.1.2 Reporting Metrics

Present evaluation metrics such as AUC and F1-score to indicate model performance in retrieving relevant scenes.

Discuss the precision and relevance of the top 5 retrieved sequences based on the model's confidence scores.

3.1.3 Summary of Challenges and Solutions

Describe the main challenges faced during data collection, model training, and deployment. Explain the solutions implemented to address these challenges.

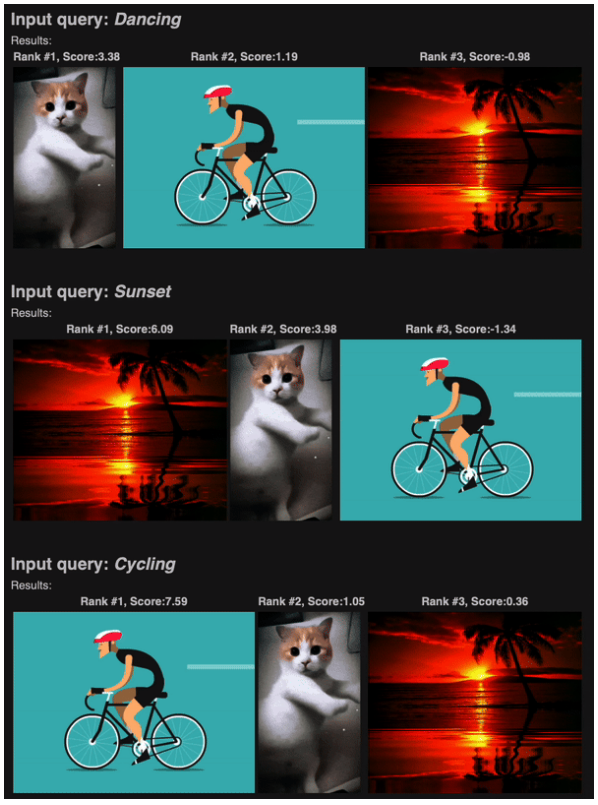


Figure 5: Example of desired system(output is video).

4 Bonus Opportunities (Optional)

- **System Deployment** : Implement a simple API or interface for querying your model and viewing results.

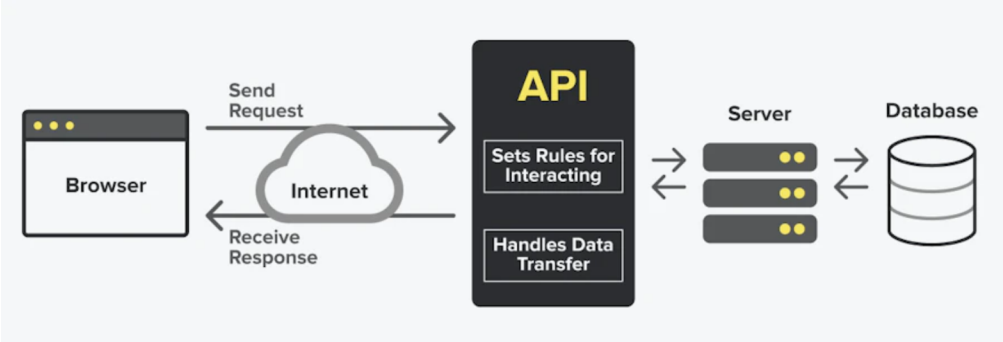


Figure 6: API.

- **MLOps Workflow**: Use MLOps tools to automate model training, monitoring, and deployment.

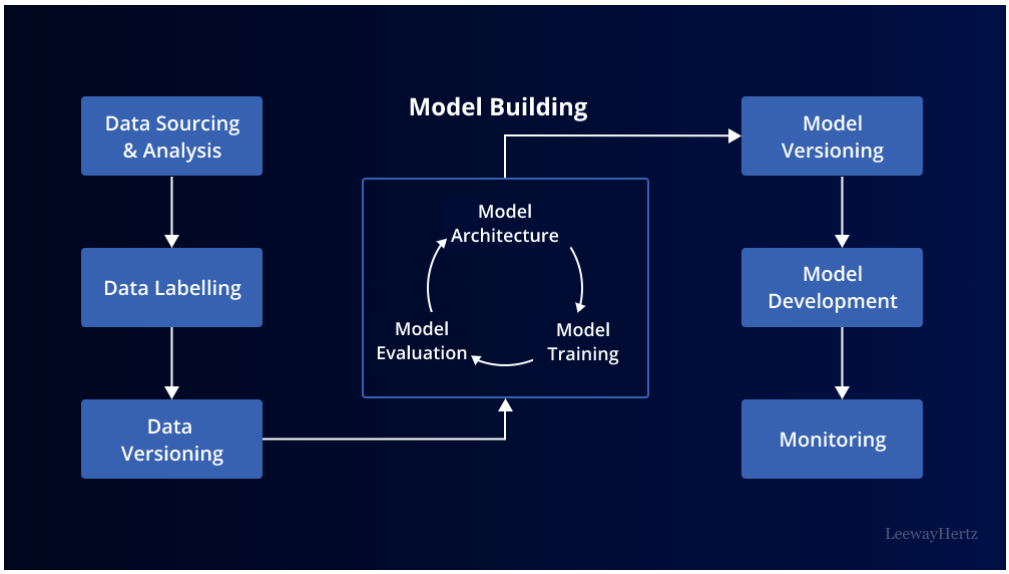


Figure 7: MLOps pipeline.

- **Advanced UI**: Create a polished, user-friendly interface for interacting with your system.

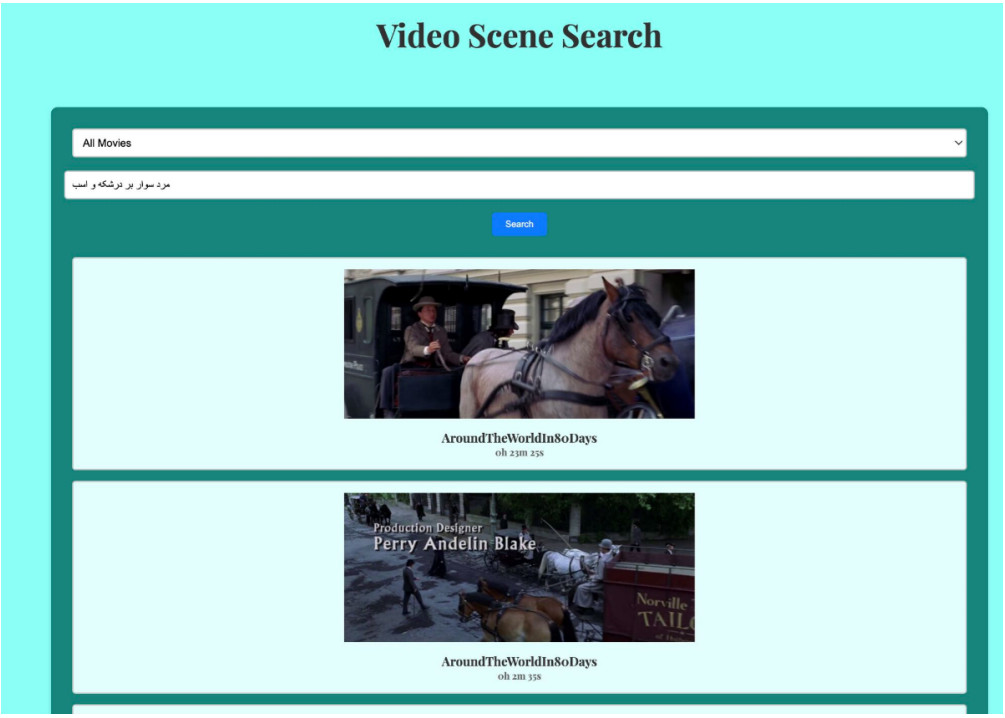
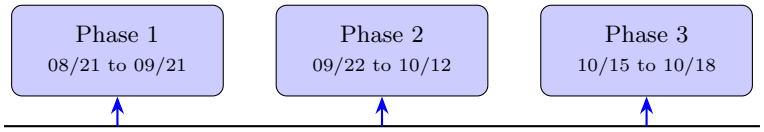


Figure 8: UI.

5 Timeline



Timeline of Phases

Additional descriptions of the expected output for each phase will be published prior to the commencement of that phase's timeline.