Project Documentation:

VISUAL SEARCH ENGINE :

Visual search refers to the process of retrieving information based on an image input rather than text. Visual search enables users to upload an image—such as a photo or sketch—and retrieve semantically similar results. By using models like CLIP, the system understands the image content and compares it to other data (images, text, etc.) in a shared embedding space, allowing for accurate and intuitive cross-modal retrieval.

Prepared by :
Ankit Sreebastava
Chirantan Mondal
Aritra Paul

Mentor :
Dr. Shivnath Ghosh

1. Overview

The Multimodal Map Search System is an AI-driven search engine that enables users to query map images using either natural language descriptions or example images. It utilizes OpenAI's CLIP (Contrastive Language–Image Pretraining) model, which can interpret and embed both text and images into the same latent vector space. By leveraging CLIP, this project performs semantic similarity matching between a user's input and a large repository of pre-indexed map images.

The system is deployed via a Gradio-based interface, making it easy to use and interact with. It is optimized for performance by precomputing embeddings and utilizing cosine similarity for efficient retrieval.

2. Objectives

- To develop a robust search engine capable of understanding and processing both textual and visual queries.
- To integrate OpenAI's CLIP model for effective multimodal search capabilities.
- To create a lightweight, scalable solution with real-time inference for end-users.
- To design a clean and intuitive user interface using Gradio.
- To demonstrate the potential of CLIP in geospatial and cartographic retrieval systems.

3. Technologies Used

Technology	Purpose
Python	Core programming language
PyTorch	Deep learning framework used for CLIP and tensor computations
Transformers (HuggingFace)	Provides access to the pretrained `CLIPModel` and `CLIPProcessor`
Gradio	Building interactive user interfaces quickly
Pillow (PIL)	Image processing and manipulation
Requests	To fetch external image content via HTTP

Torch Tensors	Efficient storage and computation of precomputed embeddings
Cosine Similarity	Metric used for comparing embedding vectors

4. System Workflow

Model Initialization

- Load the CLIP model and processor from HuggingFace.
- Load precomputed image embeddings, image indices, and normalized tensors from `.pt` files.

Text-based Search

1. User enters a natural language query.
2. Query is preprocessed and passed through CLIP to extract text embeddings.
3. Text embeddings are normalized and compared against the image embeddings using cosine similarity.
4. Top-5 most similar map images are retrieved.

Image-based Search

1. User uploads an image.
2. The image is processed and passed through CLIP to get image embeddings.
3. These embeddings are similarly compared against the precomputed image dataset.
4. Results are ranked, and the top 5 images are displayed.

Unified Gradio Interface

- Combines both input methods into a single, user-friendly UI.
- Users can input either text or an image.
- Outputs are presented in a gallery format.

5. Key Features

- **Multimodal Search**: Accepts both images and natural language as input.
- **Top-K Retrieval**: Returns top 5 relevant map images based on similarity score.
- **Cosine Similarity Matching**: For fast and accurate nearest neighbor search.
- **Precomputed Embeddings**: Reduces inference time significantly.
- **Interactive UI**: Simple and minimalistic frontend for testing and demo purposes.
- **Error Handling**: Handles invalid inputs gracefully.

6. Advantages

Accuracy

- Leverages CLIP's ability to understand context and semantics rather than just keywords or pixel patterns.

Performance

- Embeddings are precomputed and normalized, resulting in low-latency retrieval.

Flexibility

- Works across different modalities without additional training or conversion logic.

User-Friendly

- **Gradio UI** allows users to interact with the system effortlessly without needing any technical expertise.

Reusable & Extensible

- Embedding infrastructure allows for easy updates, dataset expansion, or domain shifts.

7. Risk Assessment

Data Privacy Risk

- User-uploaded images may contain sensitive or personal information.

Model Bias & Misinterpretation

- CLIP may produce biased or inaccurate results due to its pre-trained nature on general data.

Performance Bottlenecks

- Large embedding datasets may cause latency or memory issues during similarity computation.

Network Dependency

- External image URLs used in retrieval may be broken, slow, or unavailable.

Scalability Constraints

- System may struggle with high user load or expanding map datasets.

8. Applications

- **Digital Atlas Search Engines**
  For educational tools and geographical learning platforms.

- **Navigation & Location Intelligence**
  Search and compare routes or city structures through examples or queries.

- **Historical Map Archives**
  Discover old maps similar to sketches or textual records.

- **Urban Planning & Research**
  Analyze and retrieve area blueprints or plans via visual inputs.

- **Geographic Education Platforms**
  Helps in building interactive geography tools for learning.

9. Future Enhancements

Area	Description
Live Geotag Integration	Link map images with GPS coordinates for interactive applications
Natural Language Reasoning	Allow more complex questions like “Where is this mountain located?”
Batch Upload & Comparison	Enable comparison of multiple query images simultaneously
Multilingual Support	Expand text search to support other languages using translation pipelines
Retraining with Custom Maps	Fine-tune CLIP or use domain-specific versions for better accuracy in specialized datasets
Cloud Deployment	Host the system on platforms like Hugging Face Spaces or AWS for global access
Analytics Dashboard	Track search statistics, usage trends, and model accuracy over time

10. Conclusion

This project showcases the power of multimodal AI by combining vision and language understanding into a unified search system. By leveraging CLIP, a state-of-the-art model by OpenAI, we enable natural and intuitive map exploration through both text and images. With its extensibility, speed, and user-friendly interface, this solution stands as a strong foundation for intelligent geospatial search engines and interactive cartographic applications of the future.

11. Roles

Coding Role (Handled by Developer)

Ankit Sreebastava has written the core Python code that powers the system. They handle loading the CLIP model, writing functions to process text and image inputs, compute cosine similarity, and retrieve the top results. They also manage how the model interacts with the precomputed embeddings and ensure everything runs efficiently. Their code connects the backend logic with the frontend interface.

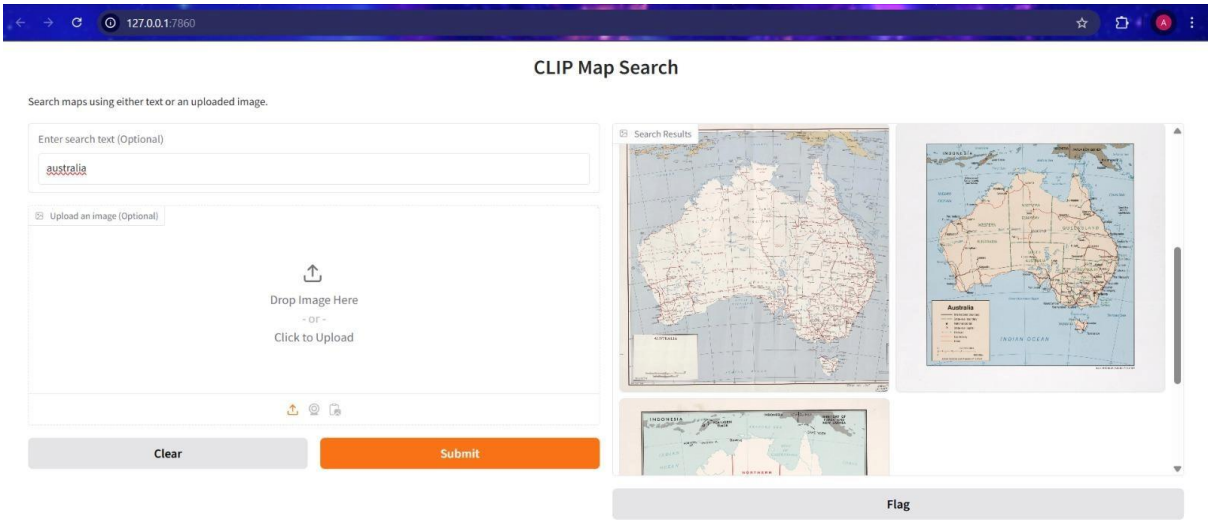
Data Embedding Role (Handled by ML/Backend Person)

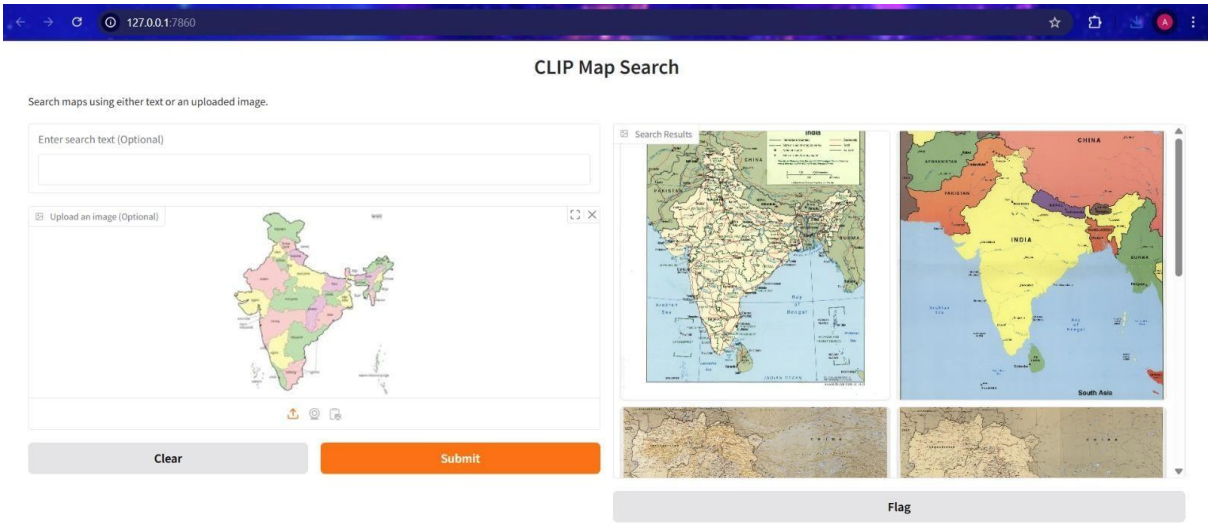
Chirantan Mondal is responsible for generating vector representations (embeddings) of all map images using the CLIP model. They process each image through CLIP to extract its semantic features, normalize them, and store them for fast comparison. These precomputed embeddings allow the system to quickly find similar images when a user enters a text or image query. Their work is essential for enabling fast and accurate similarity-based search.

Dependency Management Role

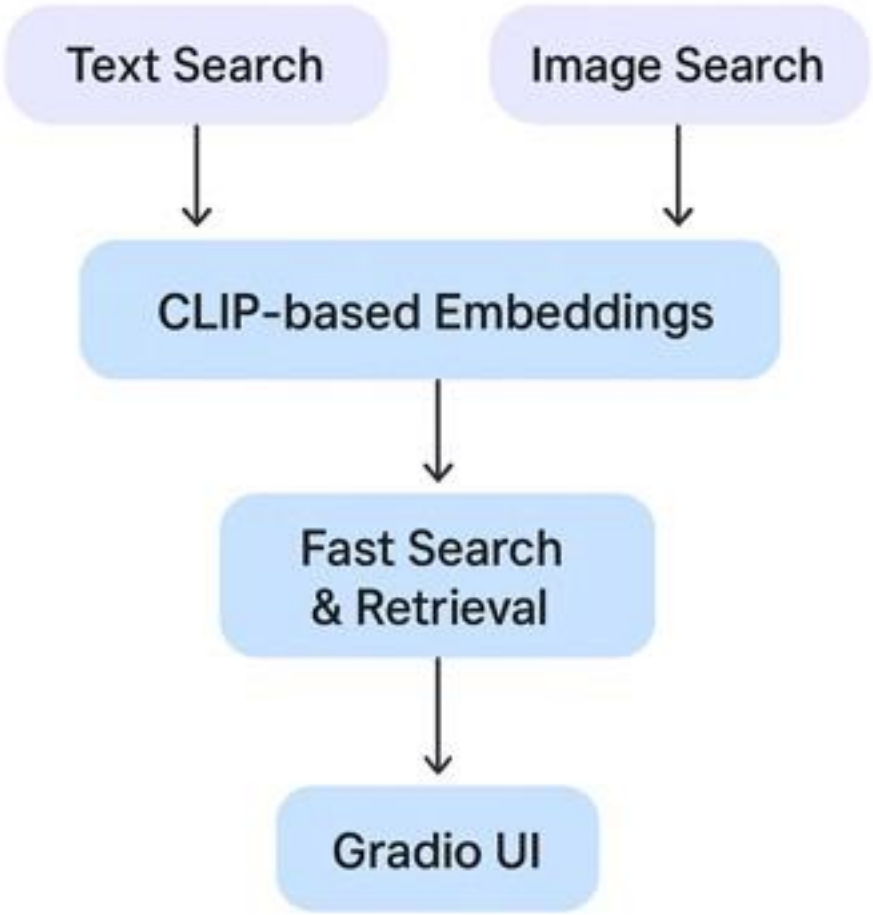
Aritra Paul has ensured that all the necessary libraries and tools (like `transformers`, `torch`, `gradio`, etc.) are correctly installed and compatible. They manage the `requirements.txt` file or environment setup, handle version conflicts, and make sure the project runs smoothly across different systems. Their work is crucial for setting up and maintaining the development environment.

12.Outputs


127.0.0.1:7860
CLIP Map Search
Search maps using either text or an uploaded image.
Enter search text (Optional)
australia
Upload an image (Optional)
Drop Image Here
- or -
Click to Upload
Clear
Submit
Search Results
Australia
INDIAN OCEAN
Flag
Use via API · Built with Gradio · Settings


127.0.0.1:7860
CLIP Map Search
Search maps using either text or an uploaded image.
Enter search text (Optional)
Upload an image (Optional)
Clear
Submit
Search Results
INDIA
CHINA
PAKISTAN
South Asia
Flag
Use via API · Built with Gradio · Settings

Custom Project Workflow Diagram



This flowchart illustrates the end-to-end process of the Multimodal-Map-Search system.