Motivation and Contribution

CLIP with its powerful ability to link images with text, is an excellent choice for image retrieval tasks. Users can input a text query and retrieve the most relevant image based on this image-text pairing. However, the model has a key limitation: its maximum token length of 77. This means that longer inputs are truncated, making it difficult for users to find exactly what they are looking for especially if their prompt isn't perfectly phrased.

In reality, people rarely recall things in fully formed-sentences. Instead, memories are often fragmented. To address this, I have combined longer descriptions with traditional prompts to ensure even less precise inputs still lead to accurate matches. Additionally, by incorporating hashtags as a tool, users can gradually refine their search, narrowing results even if they only recall partial details or key terms. This leads to an intuitive and flexible search system that aligns with how users naturally recall events and scenes.

Besides, there are some highlights of the system including:

- Immediate Feedback System: Real-time refinement of search results based on user reactions, allowing users to interact with the system dynamically.

- Aggregated Feedback System: Refinement based on historical reactions, minimizing the effects of outdated feedback.

- Relevant Lookup Feature: This enables users to initiate new searches based on any image in the results, creating a more interactive and personalized retrieval experience.

- VideoID and Timestamp Filters: Helps users find adjacent frames when searching for specific moments in video clips, improving the search experience for video-based content.

This research bridges computer vision, natural language processing, and human-computer interaction, contributing to the broader field of cognitive computing. This enhances the retrieval experience and contributes to more advanced human-computer interaction models, where systems adapt to human recall which is a fluid and imperfect nature.