COSE474-2024F: Final Project Proposal
Multimodal Image Retrieval with CLIP

Author: Hania Akhlaqi

1. Introduction

Image retrieval, the task of finding relevant images from a large database, has been a fundamental problem in computer vision for decades. Traditional methods rely on keywords or image features, often failing to capture the full context or semantics of an image.

In this project a multimodal image retrieval system named CLIP (Contrastive Language-Image Pre-training) that bridges the gap between visual content and textual descriptions is studied. The project aims to explore and compare multiple techniques such as concatenation, element-wise multiplication, maybe attention for combining image and text features and evaluate the model performance in both object-based and contextual image searches. This provides a more comprehensive analysis of different approaches.

2. Problem definition & chanllenges

Current image retrieval methods face limitations: Keyword-based searches might miss relevant images with the desired object but different descriptions. In addition, solely reliance on visual features struggles with similar objects displayed in different contexts.

This project explores beyond these limitations by leveraging CLIP, which learns a joint representation for images and text. The challenge lies in effectively combining image and text features to capture both object presence and the surrounding scene.

While previous work has focused on either on object-based or contextual retrieval this project plan to compare the performance of different combination techniques on both object-based and conceptual datasets, providing a clear evaluation of their effectiveness in various scenarios.

3. Related Works

Several studies explore multimodal image retrieval:

One study utilize text-guided object localization to improve visual search accuracy.

Some studies explored ”Attention mechanisms” technique which allow the model to focus on specific parts of the image and text that are most relevant to the task. This can help to improve the effectiveness of the combination.

Another study investigates ”Fusion networks”. They are neural network architectures designed to combine multiple modalities. They can use techniques like gating mechanisms or feature fusion layers to integrate visual and text information.

Another common technique is Graph-based models. They represent the image and text as nodes in a graph and use graph operations to combine the features. This approach can capture complex relationships between the modalities.

Researchers have also explored more specialized methods for specific tasks such as Visual-Semantic Embedding. This approach learns a joint embedding space for visual and text features, allowing for direct comparison and retrieval. Hierarchical Fusion is another method that combines features at different levels of abstraction, such as low-level visual features and high-level semantic features.

These approaches highlight the benefits of combining visual and textual modalities. The concatenation and elemnet-wise multiplication planned to be used in this project are also common combination techniques. However, directly comparing the techniques to address object-based and contextual retrieval simultaneously by using different datasets and offering clear evaluation is unique with this project.

.

4. Datasets

We will utilize two benchmark datasets:

Flickr30K: This dataset contains images with corresponding user-generated captions, ideal for learning the relationship between visuals and text.
Conceptual Captions: This dataset focuses on providing descriptive captions for images, aiding in extracting contextual information. Using these diverse datasets ensures proper training of the model for both object and contextual aspects of image retrieval.

5. State-of-the-art methods and baselines

CLIP serves as the foundation for our system. We will explore different techniques for combining image and text features:

Concatenation: Simple concatenation creates a joint representation, but might not capture the complex interaction between modalities.
Element-wise Multiplication: This method multiplies corresponding elements of image and text features, potentially amplifying relevant information.

Question:

- How to compare the performance?
- How to handle the requirement for high computational power?

6. Schedule & Roles (if you have a teammate)

Write a brief schedule/timeline of your project.

References