Zero-Shot Image Retrieval Using Vision-Language Models

Dorna Dehghani
Department of Computer Science
University of British Columbia
EECE 570 Project Proposal
drna@cs.ubc.ca

Abstract

Zero-shot image retrieval aims to find relevant images based on textual queries without relying on labeled training data. In this work, I explore the use of text-to-image generation combined with VLMs for retrieval. The approach involves generating reference images from textual prompts, computing embeddings using pre-trained models, and retrieving similar images from an unlabeled dataset. This study aims to assess the feasibility of generative approaches in zero-shot retrieval settings.

1 Introduction

Recent advancements in Vision-Language Models (VLMs) have enabled significant progress in text-to-image generation and zero-shot image retrieval. Traditional image retrieval methods rely on annotated datasets with explicit labels, which are often unavailable in real-world applications. This project explores zero-shot image retrieval, where a text query (e.g., "dog playing with a red ball") is used to find relevant images from an unlabeled dataset. Instead of relying on manual annotations, I leverage a text-to-image model to generate representative query images and then perform an embedding-based similarity search on an existing image dataset. The goal is to assess whether this generative approach enhances retrieval performance in a zero-shot setting without requiring labeled training data.

2 Related Works

2.1 Text-to-Image Generation

Text-to-image models are beneficial in generating representative images from textual descriptions. Notable models include:

- **DALL.E [7]:** A transformer-based model trained on large-scale text-image pairs.
- **Stable Difusion [9]:** A latent diffusion model capable of generating high-resolution images efficiently.
- **Imagen [10]:** A high-fidelity text-to-image generation model trained on extensive datasets.
- **Parti [13]:** An autoregressive model for text-to-image synthesis.
- **Kandinsky [8]**: A diffusion-based text-to-image model designed for high-quality generation with improved efficiency.

These models will be explored as potential candidates for generating query images.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

2.2 Image Similarity and Retrieval

To compare generated images with dataset images, robust embedding models are necessary:

- **CLIP [6]:** Maps both text and images into a common representation space.
- **DINO [2]:** A self-supervised model for learning high-level semantic representations.

These approaches provide embedding representations that can be used for cosine similarity-based retrieval.

2.3 Zero-Shot Image Retrieval

Unlike conventional retrieval systems that require labeled data, zero-shot retrieval relies on different learning methods:

- **CLIP [6]:** Performs retrieval by leveraging image-text alignment.
- **Pic2Word [11]:** Mapping images to words using weakly labeled image-caption pairs and unlabeled datasets to improve zero-shot image retrieval.

By leveraging pre-trained embeddings, this project aims to achieve retrieval without requiring additional supervision.

3 Experiment Plan

3.1 Dataset Selection

The dataset should be lightweight, and preferably with a limited number of objects per image. Possible candidates include:

- **Pascal VOC [3]:** A small-scale dataset (11K images) with object segmentation.
- **COCO [5]:** A large-scale object detection, segmentation, and captioning dataset. It should be filtered for this work, as it has more than 330K images (e.g. selecting images containing at most three objects.)
- **Open Images [4]:** A dataset of 9M images annotated with image-level labels, object bounding boxes, object segmentation masks, visual relationships, and localized narratives. It also needs to be filtered based on image complexity and size.

The final dataset selection will be determined based on computational feasibility and task relevance.

3.2 Proposed Pipeline

The proposed system will consist of the following steps:

1. **Text-to-Image Generation:** Generating representative images based on user queries using models such as Stable Diffusion v1.5 [9], Kandinsky 2.1 [8], or commercial alternatives (GPT [1] or Grok [12]) if required.
2. **Embedding Extraction:** Computing representations using CLIP ViT-B/32 [6] or alternative self-supervised models (e.g., DINOv2 [2]).
3. **Similarity Computation:** Measuring cosine similarity between generated query embeddings and dataset image embeddings.
4. **Retrieval and Ranking:** Returning the top-k most similar dataset images.

3.3 Evaluation and Potential Challenges

3.3.1 Evaluation Metrics

Due to the lack of labeled ground-truth data, evaluation can be conducted through:

- **Human Inspection:** Assessing the relevance of retrieved images to the query.
- **Recall@k and Precision@k:** (if feasible) Evaluating retrieval ranking performance. This will need a partially labeled dataset, which can be choosing a labeled dataset, or manually labeling a part of unlabeled data.

3.3.2 Expected Outcomes

If successful, this system will enable text-based image retrieval without requiring labeled datasets. Expected outcomes include:

- Accurate retrieval of relevant images based on textual descriptions.
- Scalability of retrieval across different datasets.
- Potential applications in digital archives, medical imaging, and remote sensing.

3.3.3 Challenges

- Text-to-image models may not always generate perfect representations of the query.
- Embedding-based similarity may not fully capture semantic meanings, leading to imperfect retrieval.
- Hardware constraints may limit the choice of models, requiring trade-offs between efficiency and quality.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[3] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338, 2010.

[4] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020.

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[7] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[8] Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. *arXiv preprint arXiv:2310.03502*, 2023.

[9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[10] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[11] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19305–19314, 2023.

[12] xAI. Grok: The AI Model by xAI, 2024. URL `https://x.ai`. Accessed: Feb. 25, 2025.

[13] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.