Design Image- Based Retrieval-Augmented Generation (RAG) System Using CLIP and Flan T5 Base Model on Flickr30K

Wulan Nuri Rahmawati[1], Yunita Sangadji[2], Nurlisa Safi[3], Annisaa Salsabila Shafiyyah Fitriyani[4]
[1, 2, 3, 4] Informatics Enginering, University of Muhammadiyah Malang

INTRODUCTION

- RAG combines retrieval and generation for text and images
- Most existing RAG systems focus on text, with limited exploration on images
- Extending RAG to images enables smarter AI capable of understanding and describing visual content
- An image-based RAG system using CLIP and FLAN-T5 is evaluated on the Flickr30k dataset

CONTRIBUTION

- End-to-end image retrieval with text generation
- Integrates CLIP + FAISS for efficient and scalable image retrieval
- Generate relevant captions from retrieved images

DATASET

The **Flickr30k** dataset is a publicly available collection of images from Flickr, each annotated with 5 human-written captions.

Figure 1. Image Sample

DATA OVEVIEW

- Total images: ± 31,000 images
- Gallery set: seluruh dataset
- Query set: 100 images (randomly selected)
- Training set: not applicable
- The dataset offers rich visual diversity and semantic annotations, making it suitable for image retrieval, captioning, and multimodal learning

METHODOLOGY

BASELINE

- Image embeddings are extracted using CLIP (ViT-B/32)
- Similarity search is performed using FAISS.
- Retrieved captions are used as input for FLAN-T5 Base to generate text

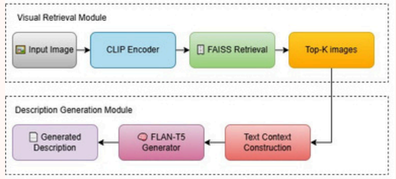

Figure 2. Pipeline for Image Retrieval and Generative Text

EVALUATION PROTOCOL

- Retrieval: Recall@5, Recall@10, mAP
- Generation: BLEU-4, ROUGE-L, CIDEr

RESULT ANALYSIS

QUANTITATIVE ANALYSIS

The **retrieval system** achieves high Recall@K and mAP, indicating that relevant images are consistently retrieved. This result demonstrates the effectiveness of CLIP-based embeddings and FAISS similarity search

Text generation scores are relatively lower due to the use of a pre-trained model without fine-tuning and the limitations of n-gram-based evaluation metrics.

→ Image Retrieval

Metric	Score
Recall@5	1.0
Recall@10	1.0
mAp	1.0

→ Generative Text

Metric	Score
BLUE-4	0.0051
ROUGE-L	0.0143
CIDEr	0.0068

QUALITATIVE ANALYSIS

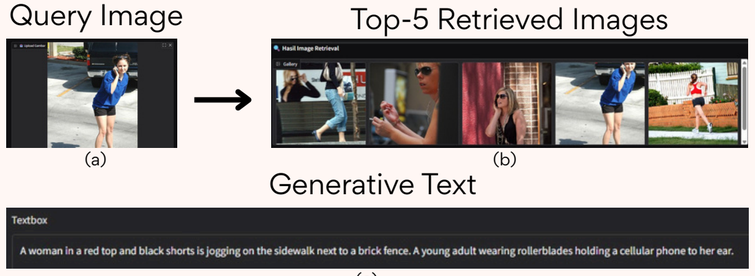

Figure 3. System Retrieval-Augmented Generation)

Qualitative analysis shows that the system successfully retrieves images with strong visual and semantic relevance. The generated descriptions capture the main objects and activities present in the images. Although automatic evaluation scores are relatively low, qualitative observation confirms that the generated text remains semantically coherent and contextually appropriate.

KEY FINDINGS

- CLIP effectively captures visual–semantic similarity between images
- FAISS enables fast and scalable image retrieval
- The retrieval stage achieves high Recall@K and mAP
- Retrieval-Augmented Generation improves contextual text descriptions
- Generated captions remain semantically coherent despite lower automatic metric scores