Bachelor:

Enhancing Content-Based Image Retrieval (CBIR) via Semantic Clustering of VLM Embeddings

1 Introduction and Background

Content-Based Image Retrieval (CBIR) aims to find visually or semantically similar images from a large-scale database based on a query image. Traditional methods rely on global descriptors (e.g., CNN features), which often suffer from the "semantic gap"—the difference between low-level pixel data and high-level human concepts.

With the advent of Vision-Language Models (VLMs) like CLIP, we can now map images into a shared semantic space where distance represents conceptual similarity. However, as database sizes grow to millions of images, exhaustive nearest-neighbor search becomes computationally prohibitive. Clustering analysis offers a dual benefit in this context:

1. **Efficiency**: By partitioning the embedding space into clusters, we can implement an inverted file index (IVF) to accelerate retrieval.
2. **Precision**: Clustering can identify and group "semantic niches" within the dataset, allowing for more robust re-ranking of retrieval results.

This thesis investigates how the scaling of clustering hyperparameters affects the accuracy and speed of image retrieval tasks.

2 Research Question

How does the integration of density-based clustering (e.g., DBSCAN) into the VLM retrieval pipeline improve the trade-off between retrieval latency and Mean Average Precision (mAP)?

3 Tasks & Goals

- **Literature Review**: Study the fundamentals of VLM embeddings and vector quantization techniques for retrieval.
- **Feature Extraction**: Use a pre-trained VLM (e.g., CLIP or SigLIP) to generate semantic embeddings for a benchmark dataset (e.g., Flickr30k or MS-COCO).
- **Clustering Integration**: Implement a clustering-based indexing layer. Use DBSCAN or K-Means to group the database images based on their semantic vectors.
- **Hyperparameter Optimization**: Apply an AutoML approach (similar to HyperBand) to find the optimal clustering parameters (like ϵ and minPts) that maximize retrieval precision.
- **Evaluation**: Compare the performance of "Flat Search" vs. "Cluster-Pruned Search" in terms of retrieval time and mAP.
- **Discussion**: Analyze whether clustering helps in handling "out-of-distribution" queries or long-tail categories.

4 Expected Outcomes

- A functional Image Retrieval pipeline that utilizes clustering for efficient indexing.
- A comparative analysis showing the impact of clustering density on retrieval accuracy.
- A well-documented thesis and reproducible Python code (using `Faiss` or `scikit-learn`).

5 Requirements

- Solid programming experience in **Python**.
- Familiarity with **PyTorch** is beneficial.
- Basic understanding of **Unsupervised Learning** (Clustering).

Contact

If you are interested, please send your CV / self-introduction and transcripts to lan@dbs.ifi.lmu.de

References

- **[1] Radford, A., et al. (2021).** *Learning Transferable Visual Models from Natural Language Supervision.* In ICML. (The foundation for semantic image embeddings).
- [2] Ester, M., et al. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise.* In KDD.
- **[3] Johnson, J., et al. (2019).** *Billion-scale similarity search with GPUs.* IEEE Transactions on Big Data. (Crucial for image retrieval benchmarks).
- [4]Lokoč J, Andreadis S, Bailer W, Duane A, Gurrin C, Ma Z, Messina N, Nguyen TN, Peška L, Rossetto L, Sauter L. Interactive video retrieval in the age of effective joint embedding deep models: lessons from the 11th VBS. Multimedia Systems. 2023 Dec;29(6):3481-504.
- [5] Rossetto L, Gasser R, Lokoč J, Bailer W, Schoeffmann K, Muenzer B, Souček T, Nguyen PA, Bolettieri P, Leibetseder A, Vrochidis S. Interactive video retrieval in the age of deep learning–detailed evaluation of VBS 2019. IEEE transactions on multimedia. 2020 Mar 16;23:243-56.