Hila Levi, Guy Heller, Dan Levi, Ethan Fetaya

Object-Centric Open-Vocabulary Image Retrieval with Aggregated Features

Task & Motivation

Task: open-vocabulary object-centric image-retrieval, i.e., efficiently locating images that contain a specific object query.

Alternatives: existing open vocabulary methods
• CLIP – global embedding – Low accuracy on complex scenes with relatively small objects
• OwlViT – dense & local embeddings – not scalable

Challenges:
• Combine scalability with object-level processing
• Keep high retrieval accuracy, for both common and rare categories

Applications: include mining rare examples and ad-hoc performance analysis

Contributions

• We visit the task of object-centric open-vocabulary image retrieval and introduce DenseCLIP, which uses CLIP’s local features, keeping its original zero-shot properties.
• We present Cluster-CLIP which enables scalability via a compact representation.
• We show the effectiveness of our approaches by achieving significantly better results compared with a global feature (CLIP) on three datasets: COCO, LVIS, and nuImages, increasing retrieval accuracy by up to 15 points.
• We integrate Cluster-CLIP into a retrieval framework, showcasing its scalability and presenting empirical evidence of its efficacy through plausible results.

Method

Preliminaries, CLIP:
• Produces a single global embedding per image.
• Suboptimal retrieval accuracy on complex scenes and small objects.

Dense-CLIP:
• Creates dense patch embeddings.
• Keeps CLIP vision-language association.
• Increases retrieval performance.
• Impairs potential scalability.

Cluster-CLIP:
• Aggregates Dense-CLIP’s dense embeddings into sparse representatives with distinct local semantics. Produces compact representation.
• Empirically evaluated with a variety of clustering methods.
• Scalable. Increased retrieval performance.

Results

Datasets: COCO, LVIS and nuImages
Metric: mAP@50
Evaluation Protocol:
Step 1: generating & storing images embeddings
Step 2: ranking by similarity to dataset’s categories

Dense-CLIP outperforms
CLIP, on all backbones
OwlViT, on the challenging LVIS and rare categories benchmarks

Cluster-CLIP
Surpasses Dense-CLIP retrieval rates using ~50 representatives per image.