Object-Centric Open-Vocabulary Image Retrieval with Aggregated Features

Hila Levi, Guy Heller, Dan Levi, Ethan Fetaya

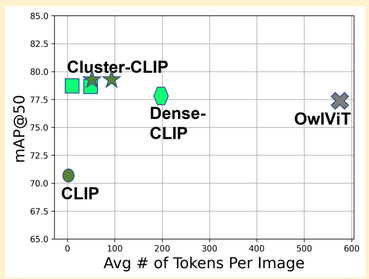
RESEARCH & DEVELOPMENT

Bar-Ilan University
אוניברסיטת בר־אילן

BMVC 2023

Task & Motivation

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



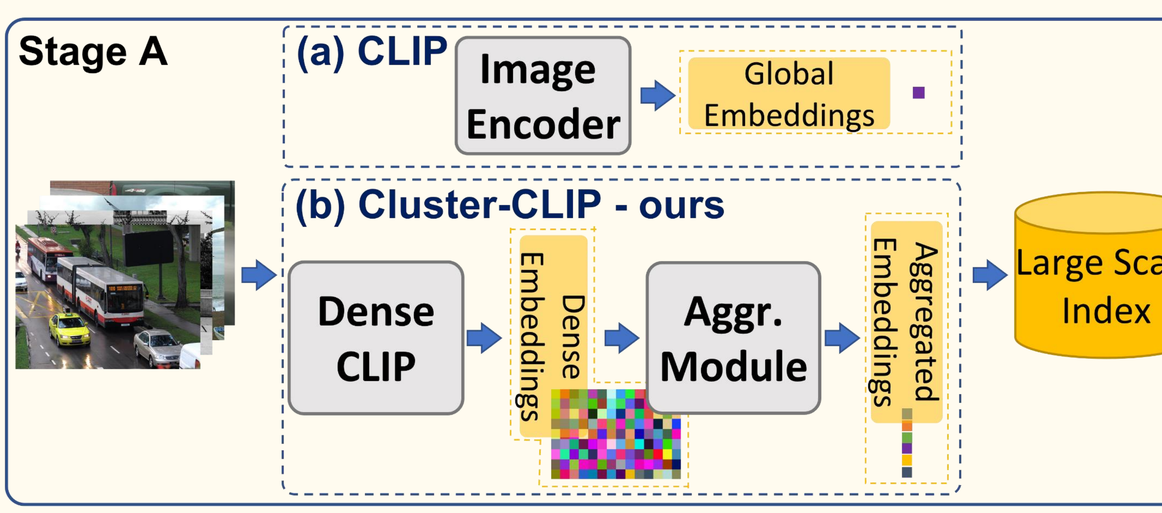
Alternatives: existing open vocabulary methods
- CLIP – global embedding – Low accuracy on complex scenes with relatively small objects
- OwlViT – dense & local embeddings – not scalable

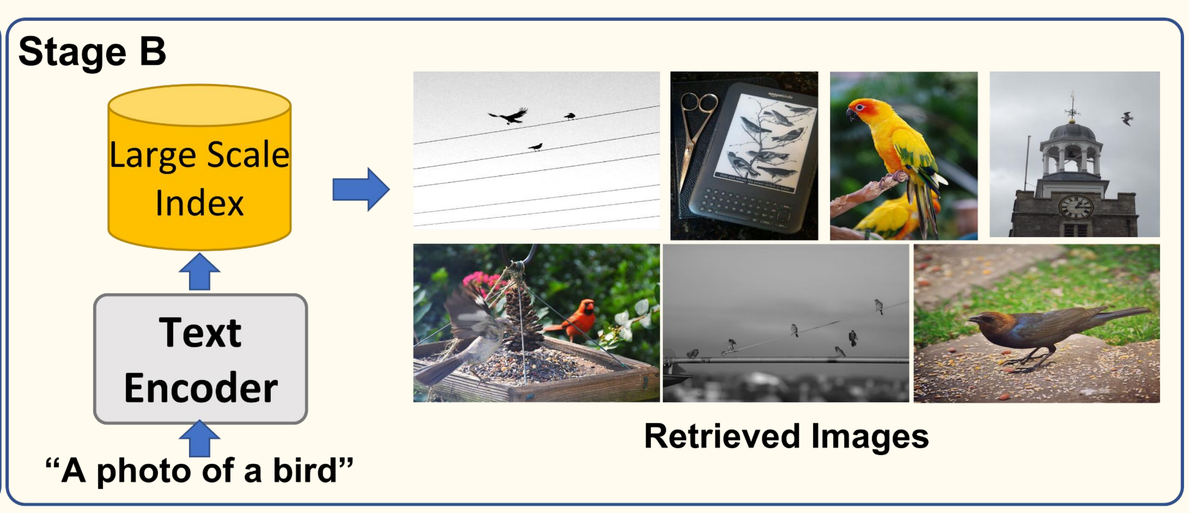
Challenges:
- Combine scalability with object-level processing
- Keep high retrieval accuracy, for both common and rare categories

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.

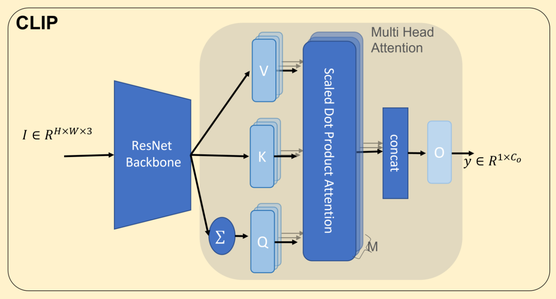




Method

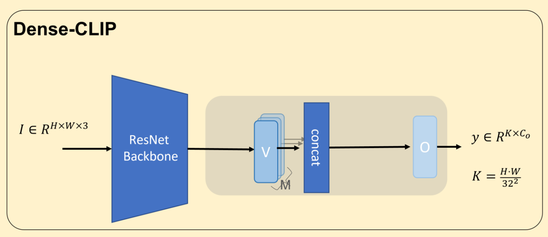
Preliminaries, CLIP:

- Produces a single global embedding per image.
- Suboptimal retrieval accuracy on complex scenes and small objects.



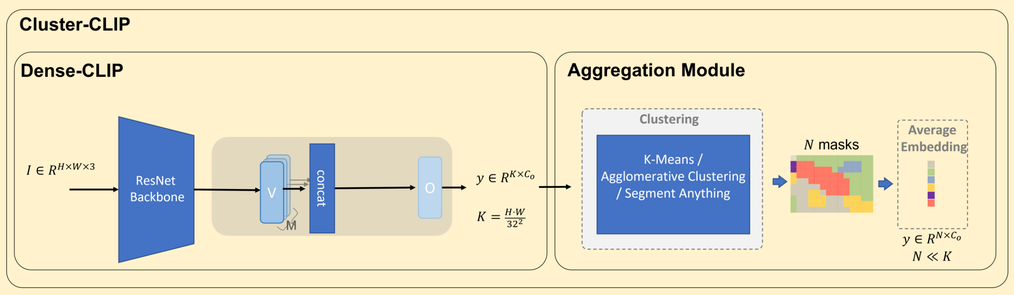
Dense-CLIP:

- 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.

Backbone	Res.	#rep.	COCO mAP@50	COCO mAP	COCO mAP@50$_{s-m}$	LVIS mAP@50	LVIS mAP	LVIS-rare mAP@50	LVIS-rare mAP
VSRN, Flicker30K	600	1	44.28	37.23	21.49	31.56	37.35	10.52	11.53
VSRN, COCO-Caption	600	1	70.29	52.33	30.99	42.09	45.60	20.46	21.36
PCME	224	7	69.69	57.98	29.98	47.38	51.06	27.85	28.58
CLIP, RN50	224	1	56.70	50.80	21.91	52.69	55.94	39.84	40.74
CLIP, RN50x4	288	1	64.58	56.39	29.39	57.85	60.35	43.37	44.28
CLIP, RN50x64	448	1	70.62	61.03	36.60	62.60	64.71	53.14	53.92
OwlViT, ViT-B/32	768	576	77.31	70.37	52.22	66.09	67.86	42.60	42.95
OwlViT, ViT-B/16	768	2304	74.96	65.91	47.02	61.28	63.39	34.75	35.82
OwlViT, ViT-L/14	840	3600	76.61	71.06	58.95	66.15	67.86	40.62	41.10
Dense-CLIP, RN50	512	256	58.83 (+2.13)	52.35 (+1.55)	32.58 (+10.67)	55.41 (+2.72)	57.46 (+1.52)	38.80 (-1.04)	39.86 (-0.88)
Dense-CLIP, RN50x4	512	256	69.61 (+5.03)	62.10 (+5.71)	41.18 (+11.79)	63.88 (+6.03)	65.88 (+5.53)	55.32 (+11.95)	56.40 (+12.12)
Dense-CLIP, RN50x64	448	196	77.78 (+7.16)	69.65 (+8.62)	51.47 (+14.87)	70.86 (+8.26)	71.80 (+7.09)	57.97 (+4.83)	58.72 (+4.80)
CLIP + Dense-CLIP, RN50	512	257	67.27 (+10.57)	59.33 (+8.53)	34.99 (+13.08)	61.32 (+8.63)	63.02 (+7.08)	46.96 (+7.12)	47.82 (+7.08)
CLIP + Dense-CLIP, RN50x4	512	257	66.21 (+1.63)	58.08 (+1.69)	29.80 (+0.41)	61.82 (+3.97)	64.12 (+3.77)	49.49 (+6.12)	50.35 (+6.07)
CLIP + Dense-CLIP, RN50x64	448	197	77.48 (+6.86)	69.45 (+8.42)	48.73 (+12.13)	72.24 (+9.64)	73.36 (+8.65)	61.68 (+8.54)	62.33 (+8.41)
Cluster-CLIP, K.M., RN50x64	448	10	78.67 (+8.05)	63.66 (+2.63)	46.69 (+10.09)	62.72 (+0.12)	64.11 (-0.6)	51.66 (-1.48)	52.45 (-1.47)
Cluster-CLIP, AG-T., RN50x64	448	10	76.72 (+6.10)	64.00 (+2.97)	45.03 (+8.43)	63.70 (+1.1)	65.16 (+0.45)	49.53 (-3.61)	50.38 (-3.54)
Cluster-CLIP, AG-F., RN50x64	448	50	78.35 (+7.73)	69.84 (+8.81)	51.95 (+15.35)	71.63 (+9.03)	72.60 (+7.89)	58.29 (+5.15)	59.07 (+5.15)
Cluster-CLIP, R.P., RN50x64	448	91	79.24 (+8.62)	69.43 (+8.4)	50.51 (+13.91)	70.74 (+8.14)	71.92 (+7.21)	58.28 (+5.14)	58.88 (+4.96)
CLIP + Cluster-CLIP, K.M, RN50x64	448	11	75.51 (+4.89)	64.60 (+3.57)	41.61 (+5.01)	66.39 (+3.79)	67.96 (+3.25)	57.02 (+3.88)	57.73 (+3.81)
CLIP + Cluster-CLIP, AG-T, RN50x64	448	11	75.42 (+4.8)	65.19 (+4.16)	43.22 (+6.62)	67.16 (+4.56)	68.64 (+3.93)	57.15 (+4.01)	57.85 (+3.93)
CLIP + Cluster-CLIP, AG-F, RN50x64	448	51	77.06 (+6.44)	69.03 (+8.00)	48.08 (+11.48)	71.79 (+9.19)	73.02 (+8.31)	60.92 (+7.78)	61.61 (+7.69)
CLIP + Cluster-CLIP, R.P., RN50x64	448	92	77.75 (+7.13)	68.59 (+7.56)	46.75 (+10.15)	71.26 (+8.66)	72.59 (+7.88)	60.98 (+7.84)	61.21 (+7.29)