RA-CLIP: Retrieval Augmented Contrastive Language-Image Pre-training

CVPR 2023

Chen-Wei Xie, Siyang Sun, Xiong Xiong, Yun Zheng, Deli Zhao, Jingren Zhou
Alibaba Group

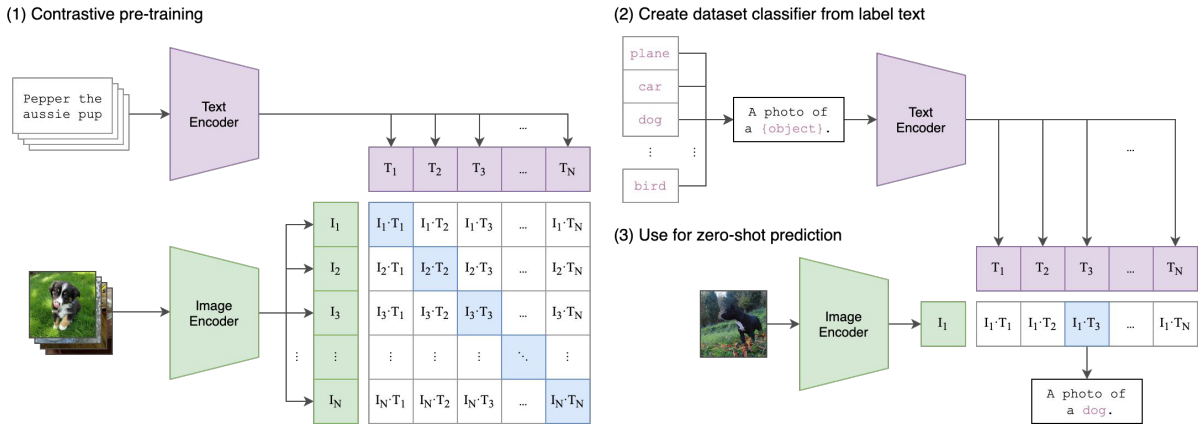
2024. 03. 15

발제자:
윤예준

연구 배경

- 전통적인 visual representation learning systems
  - 고정된 이미지 카테고리 데이터셋을 활용하여 학습
    → 한계점: 학습 외의 visual concepts가 들어오면 새로운 학습 데이터셋이 필요함

- 대안 방법: CLIP*
  - 대규모 이미지-텍스트 쌍을 학습하여 다양한 visual semantic concepts을 암기
    → 한계점: 대규모 이미지-텍스트 쌍과 다양한 visual semantic concepts를 암기할 수 있는 모델 필요함



  - 한계점을 완화하기 위한 연구로 DeCLIP*, SLIP*가 있지만 여전히 동일한 문제 존재

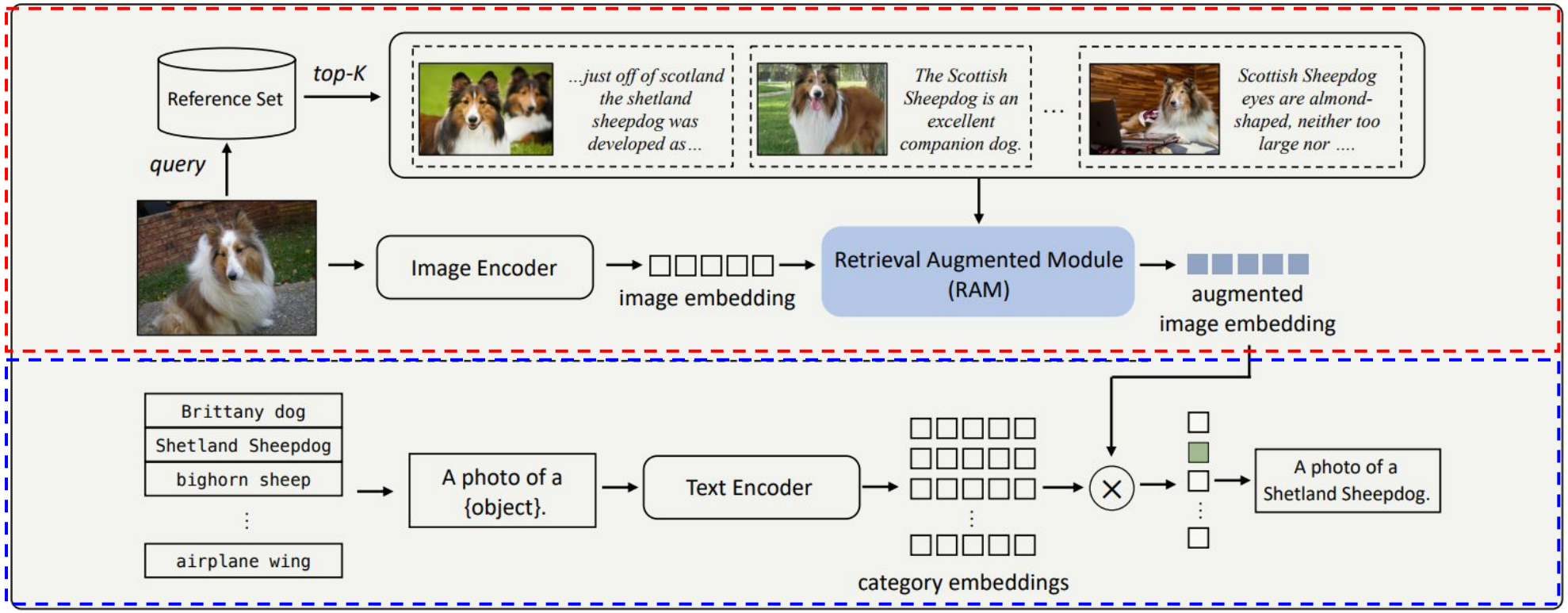
Radford, Alec, et al. “Learning Transferable Visual Models From Natural Language Supervision.” ICML, 2021
Yangguang Li, et al. “Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm.” ICLR, 2022
Norman Mu, et al. “Slip: Self-supervision meets language-image pre-training.” ECCV, 2022

연구 목표

- Retrieval Augmented을 활용하여 기존보다 효율적인 대조학습 이미지-텍스트 사전학습 프레임워크 제안
  - Reference Set(Cheat sheet)을 활용하여 image representation의 질을 높임

핵심 방법

visual recognition **task 적용**



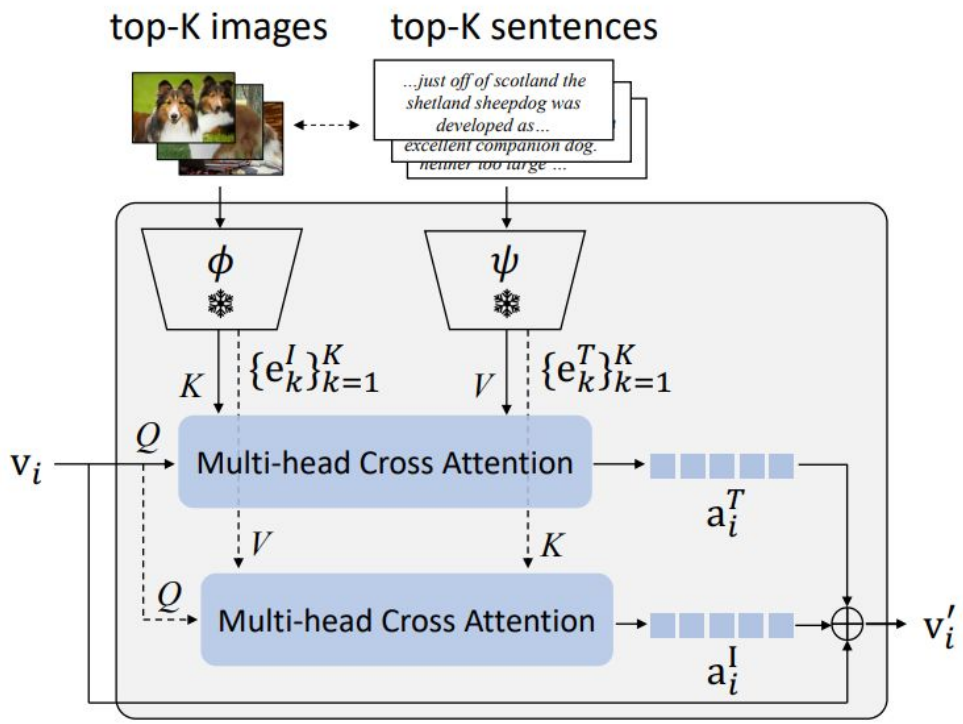
Overview of the proposed RA-CLIP

방법

- Reference Image-Text Retrieval
  - unsupervised pre-trained image encoder 사용
  - query, reference set 모두 같은 encoder를 이용하여 이미지 임베딩 추출 및 유사도 비교하여 Top-K image-text pair 추출 (FAISS 사용)

방법

- Retrieval Augmented Module (RAM)



$$\mathbf{e}_k^I = \phi(\mathbf{r}_k^I), \quad \mathbf{e}_k^T = \psi(\mathbf{r}_k^T). \tag{1}$$

$$\mathbf{a}_i^T = MultiheadAttn(\mathbf{v}_i, \{\mathbf{e}_k^I\}_{k=1}^K, \{\mathbf{e}_k^T\}_{k=1}^K). \tag{2}$$

$$\mathbf{a}_i^I = MultiheadAttn(\mathbf{v}_i, \{\mathbf{e}_k^T\}_{k=1}^K, \{\mathbf{e}_k^I\}_{k=1}^K). \tag{3}$$

$$\mathbf{v}_i' = \mathbf{v}_i + \mathbf{a}_i^T + \mathbf{a}_i^I, \tag{4}$$

Loss Function

$$\mathcal{L}_{v2t} = -log(\frac{exp(\sigma(\mathbf{t}_i, \mathbf{v}_i')/\tau)}{\sum_{j=1}^N exp(\sigma(\mathbf{t}_i, \mathbf{v}_j')/\tau)}), \tag{5}$$

$$\mathcal{L}_{t2v} = -log(\frac{exp(\sigma(\mathbf{v}_i', \mathbf{t}_i)/\tau)}{\sum_{j=1}^N exp(\sigma(\mathbf{v}_i', \mathbf{t}_j)/\tau)}), \tag{6}$$

$$\mathcal{L} = \mathcal{L}_{t2v} + \mathcal{L}_{v2t}. \tag{7}$$

구성 요소

- Data
  - Pre-training set
    - 13 million English subset of YFCC* dataset
  - Reference set
    - 1.6 million (about 1/10 amount of pre-train dataset)

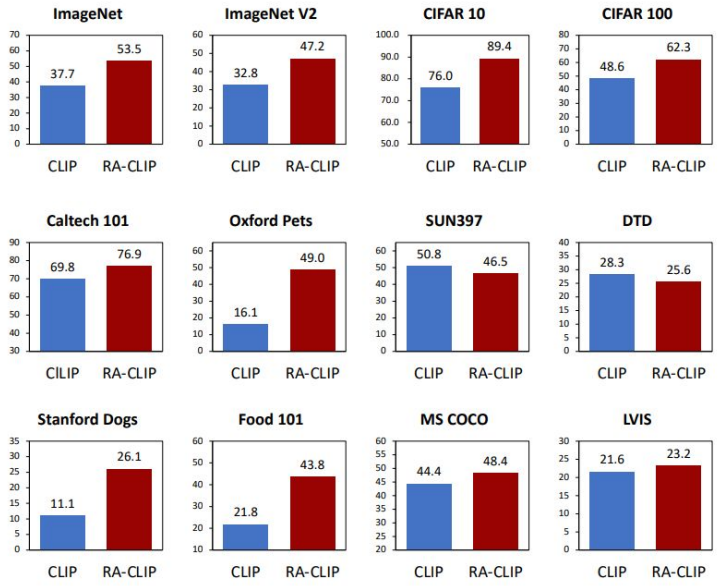
- Architecture
  - Image encoder: ViT-B/32
  - Text encoder: BERT-base
  - ϕ: DINO-S/8*
  - ψ: Sentence Transformer

- Optimization
  - Batch size: 4,096
  - Epoch: 32
  - Optimizer: LAMB
  - Learning Rate: 2.5e-3

Thomee, Bart, et al. "Yfcc100m: The new data in multimedia research." Communications of the ACM 59.2 (2016): 64-73.
Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." ICCV, 2021.

결과

- Visual recognition datasets 평가
  - 대부분 datasets에서 CLIP 등 다양한 기존 모델들보다 좋은 성능을 보임



Evaluation Type	Method	ImageNet	ImageNetV2	Pets	CIFAR10	CIFAR100	SUN397	Food101	Caltech101	DTD	Dogs	Avg.
Zero-shot	K-Lite [33]	45.3	–	–	–	–	–	–	–	–	–	–
	CLIP [29]	37.7	32.8	16.1	76.0	48.6	50.8	21.8	69.8	28.3	11.1	39.3
	MS-CLIP [39]	36.7	30.2	–	–	–	–	–	–	–	5.6	–
	SLIP [25]	38.3	33.3	28.3	72.2	45.3	45.1	44.7	65.9	21.8	11.8	40.7
	DeCLIP [24]	43.2	36.1	30.2	72.1	39.7	51.6	46.9	70.1	24.2	11.7	42.6
	RA-CLIP*	51.2	45.4	50.5	89.4	61.8	45.7	43.9	76.1	24.6	22.0	51.1
	RA-CLIP	53.5	47.2	49.0	89.4	62.3	46.5	43.8	76.9	25.6	26.1	**52.0**
Linear probe	CLIP [29]	63.5	51.3	69.8	91.7	74.1	64.7	69.1	84.9	66.5	50.5	68.6
	MS-CLIP [39]	68.1	49.8	62.1	87.2	66.7	71.7	76.0	81.6	69.4	46.1	67.9
	SLIP [25]	68.1	52.1	75.4	90.5	75.3	73.5	77.1	87.2	71.1	52.6	72.3
	DeCLIP [24]	69.2	53.1	76.5	88.6	71.6	75.9	79.3	88.0	69.1	49.9	72.1
	RA-CLIP*	73.3	62.3	88.2	94.9	78.4	60.7	65.5	86.8	65.5	75.5	75.1
	RA-CLIP	72.9	61.9	88.2	95.2	78.9	61.1	66.5	87.2	66.6	76.0	**75.5**

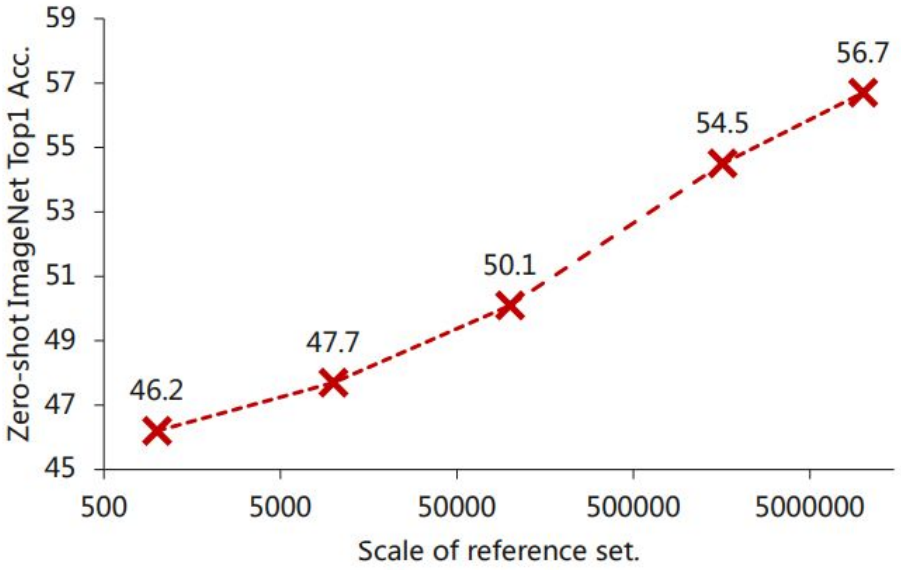
결과

- 1-3: 인코더
- 1 vs 4: Pretrain Dataset
- 4 vs 6: Method
- 5-7: Reference Dataset
- 6 vs 8: pre-trained text encoder

ID	Method	Init. of Image Enc.	Init. of Text Enc.	Pretrain Dataset	Reference Dataset	ϕ	ψ	ImageNet Top-1
1	CLIP	ViT rand.	BERT	YFCC	✗	✗	✗	37.7
2	CLIP	DINO-S	SentenceT	YFCC	✗	✗	✗	21.0
3	CLIP	ViT IN1K	BERT	YFCC	✗	✗	✗	46.1
4	CLIP	ViT rand.	BERT	YFCC+CC	✗	✗	✗	42.1
5	RA-CLIP	ViT rand.	BERT	YFCC	YFCC	SentenceT	DINO-S	53.5
6	RA-CLIP	ViT rand.	BERT	YFCC	CC	SentenceT	DINO-S	54.5
7	RA-CLIP	ViT rand.	BERT	YFCC	LAION	SentenceT	DINO-S	54.2
8	RA-CLIP	ViT rand.	BERT	YFCC	CC	Text Encoder	DINO-S	54.4

결과

- Different amounts of reference data
  - reference set의 scale이 커질 수록 더 좋은 표현을 얻을 수 있음을 보여줌

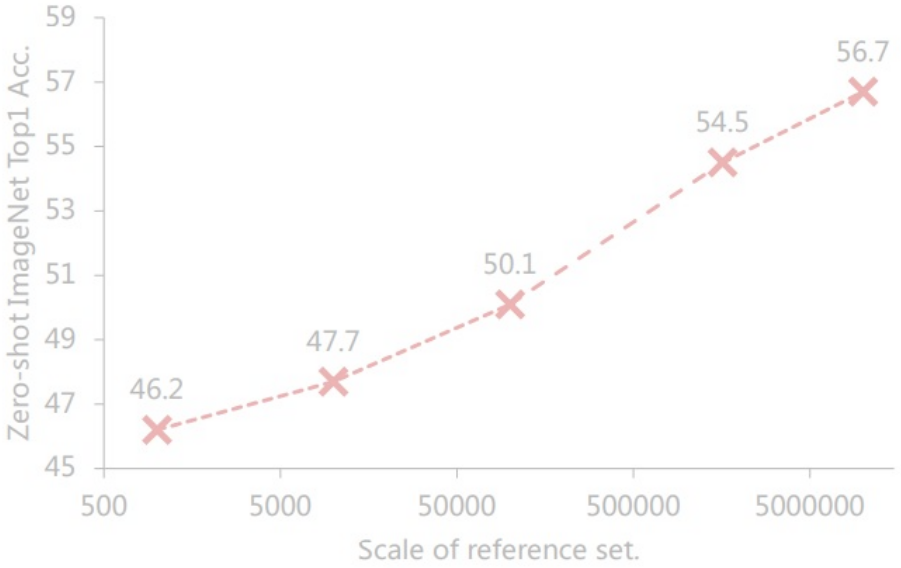


- Different hyper-parameters and design choices
  - Augment Text 적용시 성능 하락
    → text sentence는 덜 유익하고 검색된 이미지-텍스트 쌍이 더 다양하며 올바른 정보를 가져오지 못할 수 있다고 추측

Method	Augment Image	Augment Text	Fusion Type	K	ImageNet Top-1
RAM	✓	✗	$\mathbf{a}_i^T$	64	52.1
RAM	✓	✗	$\mathbf{a}_i^T + \mathbf{a}_i^I$	64	51.8
RAM	✓	✗	$\mathbf{a}_i^T + \mathbf{a}_i^I + \mathbf{v}_i$	64	54.5
RAM	✓	✗	$\mathbf{a}_i^T + \mathbf{a}_i^I + \mathbf{v}_i$	16	54.3
RAM	✓	✗	$\mathbf{a}_i^T + \mathbf{a}_i^I + \mathbf{v}_i$	128	53.9
RAM	✓	✓	$\mathbf{a}_i^T + \mathbf{a}_i^I + \mathbf{v}_i$	64	53.1

결과

- Different amounts of reference data
  - reference set의 scale이 커질 수록 더 좋은 표현을 얻을 수 있음을 보여줌

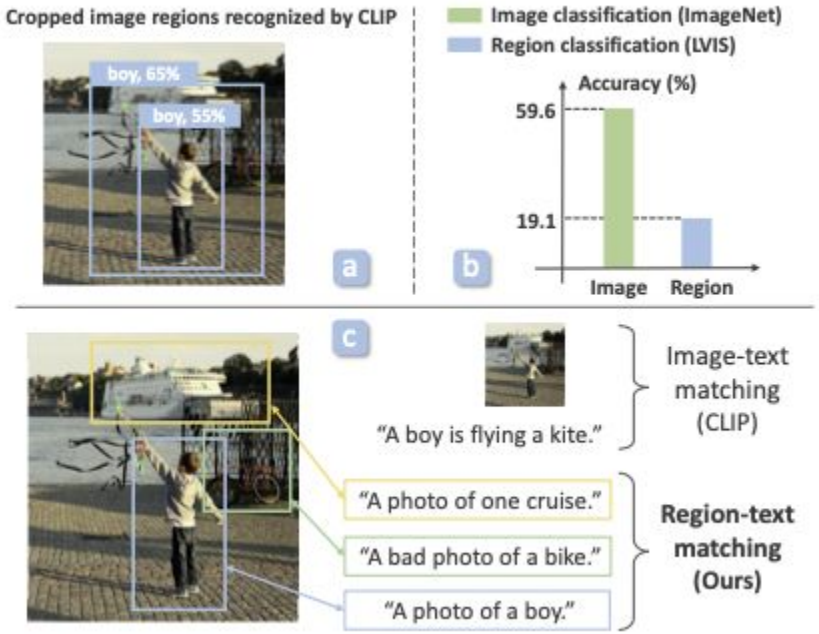


- Different hyper-parameters and design choices
  - Augment Text 적용시 성능 하락
    → text sentence는 덜 유익하고 검색된 이미지-텍스트 쌍이 더 다양하며 올바른 정보를 가져오지 못할 수 있다고 추측

Method	Augment Image	Augment Text	Fusion Type	K	ImageNet Top-1
RAM	✓	✗	$\mathbf{a}_i^T$	64	52.1
RAM	✓	✗	$\mathbf{a}_i^T + \mathbf{a}_i^I$	64	51.8
RAM	✓	✗	$\mathbf{a}_i^T + \mathbf{a}_i^I + \mathbf{v}_i$	64	54.5
RAM	✓	✗	$\mathbf{a}_i^T + \mathbf{a}_i^I + \mathbf{v}_i$	16	54.3
RAM	✓	✗	$\mathbf{a}_i^T + \mathbf{a}_i^I + \mathbf{v}_i$	128	53.9
RAM	✓	✓	$\mathbf{a}_i^T + \mathbf{a}_i^I + \mathbf{v}_i$	64	53.1

결과

- Zero-shot ROI classification
  - LVIS, COCO 데이터셋에 대하여 ROI classification 진행
  - Regin CLIP*보다 Small objects and medium objects에 대하여 더 잘하는 것을 확인



Method	LVIS				COCO			
	AP	APs	APm	APl	AP	APs	APm	APl
Regin CLIP	21.6	8.7	31.0	45.7	44.4	21.9	51.0	61.8
Region RA-CLIP	23.2	10.9	34.2	44.9	48.4	29.3	57.9	61.9

Zhong, Yiwu, et al. "Regionclip: Region-based language-image pretraining." CVPR, 2022.

결론

- Contrastive language-image pre-training을 위해 학습 데이터의 효율적인 활용 방법 제시
- Retrieval Augmented Contrastive Image-Language Pre-training 프레임워크 제안
- Visual recognition down-stream task에서 zero-shot, linear prob 분류 방법 모두 기존 모델들보다 뛰어난 성능을 보임

감사합니다.