X-DETR: A Versatile Architecture for Instance-wise Vision-Language Tasks

Zhaowei Cai, Gukyeong Kwon, Avinash Ravichandran, Erhan Bas, Zhuowen Tu, Rahul Bhotika, Stefano Soatto
AWS AI Labs

ECCV TEL AVIV 2022

I. Introduction

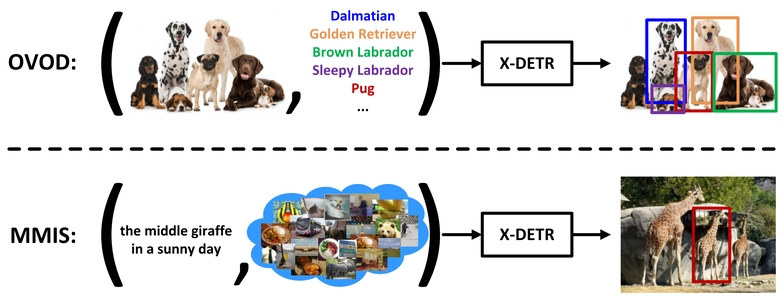

Figure 1: Open-vocabulary object detection (OVOD) and multi-modal instance search (MMIS).

- Motivation
  - Although vision-language (V+L) understanding, e.g., CLIP, has achieved very exciting results on image-level tasks, such as open-vocabulary classification and image-text retrieval, how to develop a system for instance-wise localization based V+L tasks is not clear, e.g., open-vocabulary object detection (OVOD) and multi-modal instance search (MMIS).
  - The straightforward solution, i.e., using CLIP in the framework of R-CNN, denoted as R-CLIP, is 1) very slow due to the repeated computations, and 2) sub-optimal for object-level tasks since CLIP is optimized for image-level tasks.
- Contributions
  - We propose a simple yet effective architecture, X-DETR, which is end-to-end optimized for various instance-wise V+L tasks, such as OVOD, MMIS, phrase grounding, and referring expression. It also shows better transferring capacity on downstream detection tasks than other detectors.
  - We have empirically shown that the CLIP-style of vision-language alignment, i.e., simple dot-product, can achieve good results with fast speeds for instance-wise V+L tasks, and the expensive cross-modality attention may not be necessary.
  - We have shown that X-DETR is capable of using different weak supervisions, which are helpful to expand the knowledge coverage of the model.
- Code Available
  - https://github.com/amazon-research/cross-modal-detr

II. X-DETR

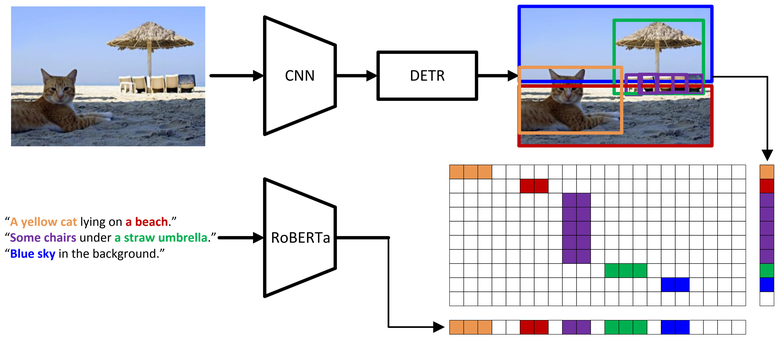

Figure 2: X-DETR architecture overview.

- Architecture
  - X-DETR has three major components: an object detector (DETR), a language encoder (RoBERTa), and vision-language alignment. The vision and language streams are independent until the end and they are aligned using an efficient dot-product operation.
- Object Detection
  - Deformable DETR is used as the stand-alone detection component, where the detection results are conditioned on only the image input. Decoupling the vision and language streams makes the detection results independent of the queries.
  - It is class-agnostic detection: classifying a hypothesis to foreground or background. In total, three losses come from detection, a binary cross-entropy, a generalized IoU and L1 regression loss.
- Vision and Language Alignment
  - Object-phrase: object is aligned with the textual phrase, and InfoNCE loss is used for optimization.
  - Object-sentence: object is aligned with the full sentence, and standard contrastive leaning is used.
  - Image-caption: the global image is aligned with the global caption, using contrastive loss similar to CLIP.
- Training Data
  - Object-language: Flickr30k-entities, RefCOCO/RefCOCO+/RefCOCOg, VG, GQA
  - Object detection: COCO, OpenImages
  - Image-caption: Flickr30k, COCO Captioning, Conceptual Captions, Localized Narratives
  - Pseudo-labeled: Localized Narratives

III. Experimental Results

- Open-vocabulary Object Detection (OVOD)

Method	Data	Train Time	Test Time	AP	AP50	AP_r	AP_c	AP_f
R-CLIP	0%	-	5s	12.7	19.3	17.0	16.0	9.0
R-CLIP+	0%	-	10.6s	13.7	20.6	18.5	17.3	9.6
MDETR [19]	0%	-	5s	6.4	9.1	1.9	3.6	9.8
X-DETR (ours)	0%	-	0.05s	16.4	24.4	9.6	15.2	18.8
DETR [5]	1%	0.5h	0.05s	4.2	7.0	1.9	1.1	7.3
MDETR [19]	1%	11h	5s	16.7	25.8	11.2	14.6	19.5
X-DETR (ours)	1%	1h	0.05s	22.8	35.0	17.6	22.0	24.4
DETR [5]	10%	3h	0.05s	13.7	21.7	4.1	13.2	15.9
MDETR [19]	10%	108h	5s	24.2	38.0	20.9	24.9	24.3
X-DETR (ours)	10%	5.2h	0.05s	29.5	44.7	29.4	30.6	28.6
Mask R-CNN [13]	100%	16h	0.1s	33.3	51.1	26.3	34.0	33.9
DETR [5]	100%	35h	0.05s	17.8	27.5	3.2	12.9	24.8
MDETR [19]	100%	1080h	5s	22.5	35.2	7.4	22.7	25.0
X-DETR (ours)	100%	45h	0.05s	34.0	49.0	24.7	34.6	35.1

  - X-DETR achieves 16.4 AP on LVIS detection of 1.2K categories at ∼20 frames per second without using any LVIS annotation during training.
  - X-DETR also shows better transferring ability.
- Multi-modal Instance Search (MMIS)

Method	FT	time	RefCOCO val			RefCOCO+ val			RefCOCOg val		
			R@5	R@10	R@30	R@5	R@10	R@30	R@5	R@10	R@30
R-CLIP	✗	∼0.19ms	5.6	8.1	14.8	7.3	10.2	17.3	21.7	29.4	42.9
R-CLIP+	✗	∼0.19ms	5.0	7.1	12.8	6.3	9.0	14.7	20.0	27.3	40.6
12-in-1 [33]	✗	∼3.5s	1.0	2.1	5.8	0.9	1.8	5.4	2.7	5.4	12.9
MDETR [19]	✗	∼25s	1.3	2.5	6.6	1.1	2.2	5.4	1.5	2.8	7.5
X-DETR (ours)	✗	∼0.15ms	21.5	30.8	47.8	14.8	22.1	37.7	23.4	33.2	52.0
UNITER [7]	✓	∼1.4s	8.1	14.3	28.9	13.5	21.0	36.0	14.5	22.1	37.7
MDETR [19]	✓	∼25s	2.0	3.7	9.0	2.5	4.4	10.9	3.5	5.9	15.3
X-DETR (ours)	✓	∼0.15ms	29.9	40.7	59.6	23.7	33.5	53.8	40.0	53.4	72.5

  - MMIS is a task to retrieve the most similar object region from a large-scale (millions or billions) database given a free-form language query.
  - X-DETR achieves the best results in all three datasets, at very fast speeds.