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

Zhaowei Cai, Gukyeong Kwon, Avinash Ravichandran, Erhan Bas,
Zhuowen Tu, Rahul Bhotika, and Stefano Soatto

AWS AI Labs
{zhaoweic,gukyeong,ravinash,erhanbas,ztu,bhotikar,soattos}@amazon.com

Abstract. In this paper, we study the challenging instance-wise vision-language tasks, where the free-form language is required to align with the objects instead of the whole image. To address these tasks, we propose X-DETR, whose architecture has three major components: an object detector, a language encoder, and vision-language alignment. The vision and language streams are independent until the end and they are aligned using an efficient dot-product operation. The whole network is trained end-to-end, such that the detector is optimized for the vision-language tasks instead of an off-the-shelf component. To overcome the limited size of paired object-language annotations, we leverage other weak types of supervision to expand the knowledge coverage. This simple yet effective architecture of X-DETR shows good accuracy and fast speeds for multiple instance-wise vision-language tasks, e.g., 16.4 AP on LVIS detection of 1.2K categories at ~20 frames per second without using any LVIS annotation during training. The code is available at https://github.com/amazon-research/cross-modal-detr.

Keywords: Vision-language, object detection, multi-modality, retrieval

1 Introduction

Vision-language (V+L) understanding has achieved promising progresses in the past a few years [7,8,18,24,26,32,33,42,49]. [18,42] have shown that strong vision-language alignment can be enabled by a simple dot-product between vision and language representations, with the help of large-scale image-caption pairs (hundred of millions to billions). Although they have 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 still unknown, e.g., open-vocabulary object detection (OVOD) and multi-modal instance search (MMIS). OVOD detects any object categories defined by free-form language descriptions without finetuning (see Fig. 1 top), where the size of categories could span from dozens to thousands. On the other hand, MMIS retrieves the most similar object region from a database given a free-form language query (see Fig. 1 bottom), where the database size could be millions or billions for a commercial search engine.

One straightforward solution is to use a R-CNN framework [10] with a pretrained object proposal detector, e.g., [4,5,46,61], and a pretrained V+L model, e.g., CLIP [42], denoted as R-CLIP, similar to [11]. Then the pipeline is to 1) detect object proposals,
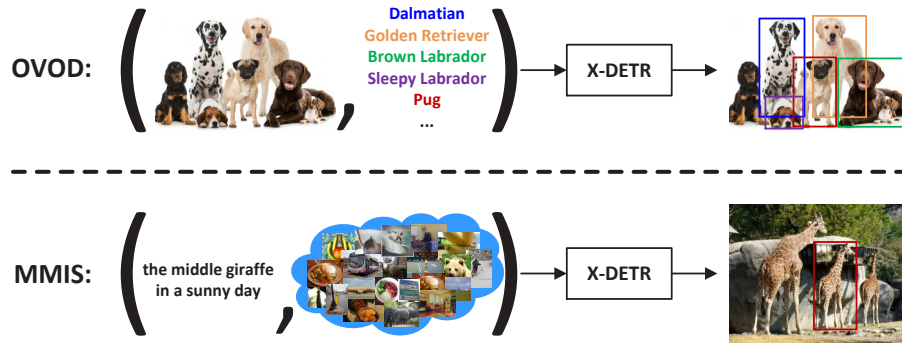
Fig. 1. The illustration for open-vocabulary object detection (OVOD) and multi-modal instance search (MMIS).

and 2) crop image regions of proposals, and 3) forward the cropped image regions through the V+L model for final vision-language alignment. However, this framework has its limitations. First, it is very slow because feature extraction is repeated in over-lapped image regions. Speed is the key for applications of instance-wise V+L tasks. For example, OVOD usually requires real-time speeds (e.g., 25 frames per second), and MMIS requires instant retrieval results (e.g., in at most a few seconds) from a database of millions or billions instances. Second, the image-level representation of CLIP is sub-optimal for the instance-wise V+L tasks. For example, the cropped regional representation lacks global context, which is required in some tasks. As shown in the example of locating "the middle giraffe in a sunny day" in Fig. 1 (bottom), the conditions of "middle" and "in a sunny day" require to know the relations with other giraffes and the global image context, respectively.

To resolve these issues, we propose an efficient and effective architecture for various instance-wise V+L tasks, denoted as X-DETR (cross-modal DETR). It has three major components: a visual object detector (transformer based DETR [5]), a language encoder (a RoBERTa model [31]) and alignment between the visual instance and the language description. This model is trained end-to-end, optimizing all components simultaneously. Hence, the detector is adapted to the instance-wise V+L tasks instead of as an off-the-shelf component as in the previous V+L efforts [7, 26, 32]. Furthermore, motivated by the success of CLIP [42], we keep the visual and language streams independent as much as possible, and align them together by a simple dot-product operation at the very end, instead of using an expensive joint-modality transformer as in [7, 19, 26, 32]. In this sense, X-DETR can be seen as a detection counterpart of CLIP, but it overcomes the two issues of R-CLIP framework. First, thanks to the architecture design of X-DETR, the representations of multiple instances and language queries can be obtained by a single feed-forward pass. For example, X-DETR can run 20 fps for OVOD on LVIS of ∼1.2K categories [12], and for MMIS, it can retrieve the object in seconds from one million of instances given a query. Second, each instance feature representation of DETR encodes the information from global image and thus can be more

accurate for tasks requiring global context. For example, X-DETR is much better than R-CLIP on RefCOCO/RefCOCO+ datasets [57] requiring context information.

A cornerstone of the success of CLIP is the large-scale training data, ∼400 million image-caption pairs. The paired image-caption annotation is relatively easy to collect, e.g., by crawling from the internet. However, paired object-language data usually requires human annotation and thus is very expensive. As a result, only a few datasets have this kind of annotations, i.e., Flickr30k entities [39] for grounding, RefCOCO/RefCOCO+/RefCOCOg [34,57] for referring expression comprehension (REC), VG [22] for dense captioning and GQA [17] for visual question answering (VQA). In fact, the union of them is of relatively small size, with only ∼90K unique images for training, which is not enough to learn universal instance-wise V+L representations. To deal with this challenge, our framework resorts to other types of weak supervisions than the paired object-language annotation, including image-caption pairs, object bounding boxes and pseudo-labels, from datasets including COCO [28], OpenImages [21], CC [48], and LocNar [41]. This expanded combination of full and weak annotations provides broader knowledge coverage for X-DETR to learn universal representation, as will be seen in our experiments. Our contributions can be summarized as follows:

- 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.

2 Related Work

Vision-Language Learning is a popular interdisciplinary research topic [7, 8, 24, 26, 32, 33, 42, 49]. Early efforts focused on a single specific task, e.g., [35,53] on image captioning, [1,55] on visual question answering (VQA) [2], [23,36] on image-text retrieval and [16, 56] on REC [20], etc. Recent efforts have focused on the joint pretraining of two modalities [7, 24, 26, 32, 33, 49], aiming for a multi-task model that can work on multiple downstream tasks simultaneously. Although these methods do not necessarily work on detection related tasks, most of them use an off-the-shelf detector, e.g., [1,59], for more accurate visual feature representation. However, the system is not end-to-end and it is not guaranteed that the detector is optimized for the following V+L tasks. Differently, X-DETR, a versatile architecture for multiple instance-wise V+L tasks, has all components optimized end-to-end.

Uni-modal Object Detection has achieved great progresses in recent years, including the pioneering two-stage object detectors [4, 10, 27, 46], efficient one-stage detectors [30, 45], and the very recent transformer-based object detectors [5, 61]. However, these uni-modal detectors are constrained to predefined categories, e.g., COCO/Objects365/OpenImages [21, 28, 47] of 80/365/601 classes, and they are unable to detect

any categories beyond the predefined ones. In natural images, the objects are orders more diverse than the predefined categories (usually 10∼1,000) in the sense of categories, attributes (e.g., colors, materials, etc.), geometric location (e.g., "man on the right"), relation to the environment ("man sitting in a couch"), etc. The traditional unimodal object detectors are incapable of dealing with these problems.

Multi-modal Object Detection tries to detect objects with free-form language description. Previous work [15,40,54] extended traditional uni-model detection framework to accommodate the language inputs. Recently, MDETR [19], a modulated detector for multi-modal understanding, achieved state-of-the-art results on multiple datasets. However, it adopts the expensive joint-modality transformer to model the alignment between language and vision, which prevents it from being used in practical applications like OVOD and MMIS. For example, it takes 5 seconds per image for OVOD and hours for MMIS. OVOD [15,25,40,58,60] usually leverages language models as it is impossible to pre-define all open-vocabulary categories. It is related to zero-shot object detection (ZSOD) [3,43], where the model is trained on the seen categories and evaluated on unseen ones. X-DETR falls into the category of OVOD because it is trained on large V+L data. X-DETR shares some similarities with some concurrent OVOD works [11,25,60], but it 1) also works on other practical tasks, e.g., MMIS; 2) leverages more diverse weak supervisions; 3) has fast speeds for both OVOD and MMIS tasks.

Cross-modal Retrieval Cross-modal (image and text) retrieval has a long research history [44]. This task is also popular in recent V+L learning efforts [36,51]. Many X-DETR tasks are related to cross-modal retrieval, but they are instance-wise. For example, OVOD [15,40,58] and MMIS [15,29,40] are to retrieve the most similar bounding boxes across the full dataset given the free-form query, and phrase grounding [39] and REC [34,57] are to retrieve the most similar bounding box in an image given a query. X-DETR aims to tackle them in an efficient and effective manner.

3   Instance-wise Cross-modality Network

3.1   Overall Architecture

Figure 2 gives an overview of X-DETR architecture with its three components: a visual object detector D, a text encoder ψ and an alignment between visual instance and textual description h. The method takes an image I and a language query y as inputs and outputs the detected object o and its alignment score with the language query y. The object detector D is used to generate the instance $o = D(I)$ by processing the input image I, and the language encoder ψ is to encode the tokenized text input y into embeddings $\psi(y)$ which can be mapped to the joint space with the visual instance embeddings o. Vision-language alignment h is the key component in V+L models, aligning the visual instance o and language description y in a joint feature space, such that $h(o, \psi(y))$ is higher (lower) for paired (unpaired) visual instances and language descriptions. For example, as shown in Fig. 2, the alignment will pull the same concepts together, e.g., the detected cat (yellow bounding box) and the language description "a yellow cat", but push different concepts away, e.g., the cat and other language descriptions like "a beach", "some chairs", "blue sky", etc. In addition to the object-language alignment,
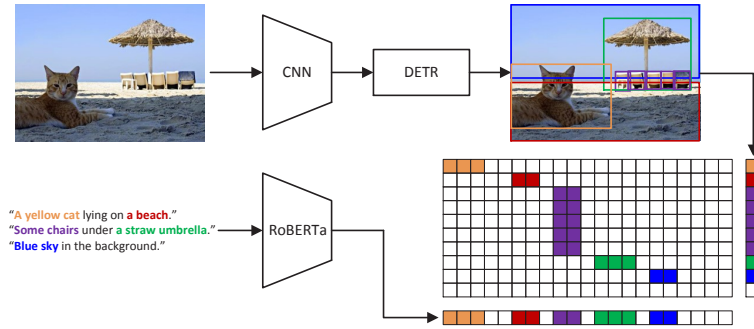
Fig. 2. Overview of our X-DETR architecture. The same color of text and object means they are aligned concepts. The matrix is the similarity matrix between text tokens and instance hypotheses, where the colored score items should be maximized but the rest should be minimized.

X-DETR also has an image-language alignment component to leverage the weak data of image-caption pairs to learn broader knowledge coverage. The details of each component are discussed next.

3.2   Object Detection

There are many choices for the object detection component [4,5,27,30,45,61], and we are not constrained to any specific one. We chose the transformer based framework of DETR [5], because 1) it is simpler without heuristics compared with the other popular one or two-stage frameworks [4,27,30,45], and 2) each detected instance encodes information from global image due to the attention mechanism of transformers. In DETR, a standard CNN (e.g., ResNet [14]) is applied to an image I to extract convolutional feature maps. They are then flattened into a sequence of features passed through a encoder-detector transformer architecture [50]. To detect objects, the detector also takes object queries as input, and outputs the decoded queries as the detection results o. The detection architecture of X-DETR is shown in Fig. 3 (a). Due to the slow convergence of the original DETR, we resort to Deformable DETR [61], which has faster convergence speed and better accuracy. Please refer to [5,61] for more details.

   The detector in X-DETR is a stand-alone component, where the detection results are conditioned on only the image input. This is different from previous multi-modal object detection frameworks [19,54], where the detection results depend on both image and language inputs. Decoupling the vision and language streams makes the detection results independent of the queries. This is closer to a human detection system, where salient objects can be detected before being asked to detect something of interest.

3.3   Object-Language Alignment

A common and powerful strategy for image-level vision-language alignment is to leverage the interaction between two modalities by a joint-modality transformer [7,26,32], MA, as follows:

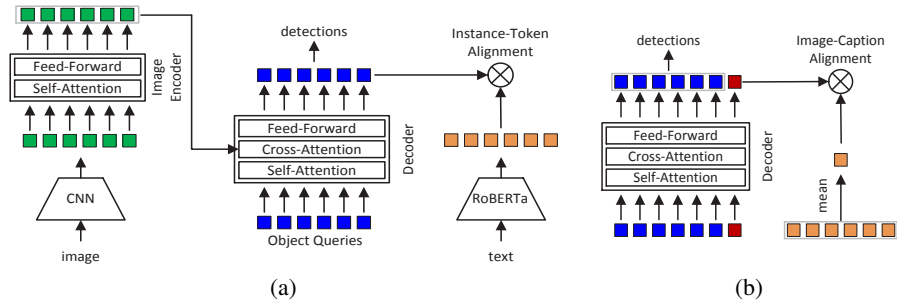$$h(I,y) = MA(\phi(I), \psi(y)), \tag{1}$$

Fig. 3. (a) is the architecture of X-DETR. (b) illustrates how the encoded image-level query (red square) is aligned with the caption (yellow square).

where ϕ is the image encoder, e.g., CNN. The joint-modality transformers could be either self-attention on concatenated vision and language features, e.g., [7,26], or cross-attention between vision and language streams, e.g., [32]. Although more cross-modal interaction could lead to stronger cross-modal representations, the computation of MA needs to be repeated if any vision or language input is changed. This is impractical for tasks such as OVOD and MMIS. For example, MA needs to be repeated for 1.2K times for LVIS detection, and one million times for MMIS of database image size of one million, although $\phi(I)$ and $\psi(y)$ do not need to be recomputed.

Motivated by CLIP [42], where strong cross-modal alignment can be enabled by a simple and efficient dot-product operation, we adopt a similar approach, i.e., the vision-language alignment h is a linear mapping of the two modality streams into a common feature space followed by a dot-product operation between them,

$$h(o, y) = f(o) \odot g(\psi(y)), \qquad (2)$$

where f and g are linear mapping for instance representation o and text representation $\psi(y)$, respectively, and $\odot$ is a dot-product operation. Since o is generated by the stand-alone object detector D, (2) can be formulated as

$$h(o, y) = f(D(I)) \odot g(\psi(y)). \qquad (3)$$

Since the two modalities are coupled at the very end via an efficient dot-product, no additional computation needs to be repeated. This is a key difference with methods that use joint-modality transformers [7, 19, 26, 32] and enables X-DETR to be used in practical tasks such as real-time OVOD and instant MMIS without sacrificing accuracy.

3.4   Image-Language Alignment

As shown in Fig. 3 (b), in addition to the queries used for object detection, an additional query is added to the query list as image query, which encodes image-level representation instead of instance-level representation like the other object queries. All queries are forwarded to the decoder with no difference. The language representation is the mean of all token representations in the sentence. Similar to CLIP, the alignment is also a simple dot-product between the encoded image query and the caption representation.

4   Training

In this section, we describe our multi-task loss design involving losses from object detection and vision-language alignment, and different types of data we used. More details about the losses can be found in the supplementary.

4.1   Class-agnostic Object Detection

In DETR [5], object detection is a set-to-set prediction, where a set of queries is firstly mapped to a set of ground truth objects during training. Then the matched object query is learned to regress to the corresponding ground truth, with a classification loss (cross-entropy) and bounding box regression loss (generalized IoU and L1 loss). Different from the general object detection of [5,61], X-DETR does not use the class information of each object. Instead, it is class-agnostic detection: classifying a hypothesis to foreground or background. This design is similar to [11]. In total, three losses come from detection, a binary cross-entropy, a generalized IoU and L1 regression loss.

4.2   Vision-Language Alignment

We have explored different levels of vision-language alignment, including object-phrase, object-sentence and image-caption alignment. The diverse levels of alignments allow us to expand the training data, which is one of our key contributions.

Object-Phrase Alignment In the phrase grounding dataset of Flickr30k entities [39], the ground truth is a pair of sentence phrases and objects. For example in Fig. 2, given the sentence "Some chairs under a straw umbrella.", the phrase queries are "Some chairs" and "a straw umbrella", with associated one or a few bounding boxes. We used the contrastive loss of InfoNCE [37] to optimize for the object-phrase alignment. The similarity of every potential object-token pair is computed, as shown in similarity matrix of Fig. 2, and contrastive loss is applied for each row (object-token alignment) and column (token-object alignment) of this matrix. Note that although this loss does not directly optimize the phrase-object alignment, it achieves similar results.

Object-Sentence Alignment is a special case of object-phrase alignment, where the length of the text is the full sentence, with data from REC datasets [34,57]. Contrastive loss is applied between the object query and whole sentence embedding, which is averaged from token embeddings.

Image-Caption Alignment leverages the large-scale weak data of image-caption pairs. Similar to CLIP, the loss is a cross-modality contrastive loss between the encoded image queries and the captions.

4.3   Training Efficiency

In a typical image, there are multiple objects associated multiple language descriptions. The most efficient way for training is to use all object-language pairs at a single forward-backward pass. However, this is problematic for the models with cross-modality interaction MA, e.g., [7, 19, 25, 26], due to the reasons discussed in Section 3.3. As a compromise, they usually merge all independent text queries into a paragraph as a single

query, and then computes MA only once. However, this violates the independence assumption of queries and queries can see each other during training, due to the property of transformer. At inference, only a single query is provided each time. On the contrary, X-DETR does not have these problems, due to the design of fully independent vision and language streams. It can take queries as many as possible at high efficiency while not violating the independence assumption, thus better performances.

4.4   Training Data

The most preferred data to learn instance-wise cross-modal representation is the paired object-language annotations, but they are very expensive and limited. X-DETR is capable of leveraging other weak supervision to cover broader knowledge.

Object-Language Data The paired object-language data comes from Flickr30k entities [39] for grounding, RefCOCO/RefCOCO+/RefCOCOg [34, 57] for REC, VG [22] for dense captioning and GQA [17] for VQA. We used the *mixed* dataset of them following MDETR [19]. Please refer to [19] for more details.

Object Detection Data Since X-DETR has a stand-alone object detector, it can leverage data with detection annotation only. COCO [28] bounding box annotations are used because many images of the *mixed* dataset are from COCO. But category information is not used, since the detection in X-DETR is class-agnostic as described in Section 3.2. Instead of having an image with detection annotations only, we add COCO objects into the existing images of the *mixed* dataset, since many of them are sparsely annotated. For example, usually only 3-5 objects are annotated per image in REC datasets.

Image-Caption Data Many datasets have image-caption annotations, e.g., Flickr30 [39], COCO Captioning [6], CC [48], SBU [38], etc. Since the *mixed* dataset already includes Flickr30k and COCO images, we added their captions into the *mixed* dataset. Beyond that, we also use the large-scale CC and Localized Narratives (LocNar) [41] (only the subset of OpenImages [21]). Note that we do not use the weak localization annotations of LocNar, i.e., the pointer tracks along narratives, because we found they are quite noisy and have no much benefit.

Pseudo-Labeled Data We also used the pseudo labeled data generated by our X-DETR model on LocNar. Given an image and its corresponding caption, at first we use Spacy[1] to extract the noun phrases which are possible objects in the corresponding image. Then we treat the pseudo-labeling as a phrase grounding task, retrieving the bounding box that is most aligned with the noun phrase. This pseudo-labeled dataset is then used with no difference to the other object-language pairs. In addition, the Open-Images object annotations were also added to LocNar similar to COCO.

5   Experiments

In this section, we are going to show how X-DETR performs on the challenging and practical OVOD and MMIS tasks, and how it can be generalized well on the other simpler tasks, such as phrase grounding and REC.

---

[1] https://spacy.io/

5.1   Implementation Details

In X-DETR, we used RoBERTa-base model as the text encoder (with implementation and pretrained model from HuggingFace [52]), ResNet as the vision backbone, and Deformable DETR as the object detector. The model has been trained for 10 epochs, w.r.t. the *mixed* dataset, with class-agnostic detection, object-phrase alignment, and image-caption alignment losses introduced in Section 4. The object-sentence alignment loss is not used during pretraining but used in some of the downstream tasks finetuning. In each mini-batch, the batch size for fully/pseudo/weakly-annotated data is 4/2/4 for a single GPU. The image is resized such that the minimum of width and height is 600, which is smaller than the standard practice of 800 in object detection [13, 27], for training cost reduction. See the supplementary for more training details. We report the results with and without finetuning. X-DETR without finetuning is a multi-task architecture for various instance-wise V+L tasks. All the compared algorithms use ResNet-101 (except R-CLIP). We do not use any prompt engineering as in CLIP [42], e.g., adding a prefix of "a photo of" to the language query.

In the baseline of R-CLIP with the pipeline introduced in Section 1, the proposal detector is a binary Faster R-CNN detector [27] of ResNet-50, trained on the COCO *train2017* with binary (object/non-object) annotations. Note the binary detector is not a RPN network [46]. The V+L model is a pretrained ViT-B-32 CLIP model [42]. For each image, the top 300 detection results are selected as the proposals. Since the proposals are usually tight bounding boxes, covering few context region. To include some context, we expand the proposal region by 50% on each side of the bounding box, following [11]. When using the context, the CLIP feature representations from the original and expanded regions are averaged before $L2$ normalization. This is denoted as R-CLIP+.

To evaluate the similarity between a free-form language query and an object is straightforward for R-CLIP and X-DETR. The whole sentence is forwarded through the language encoder to get feature embeddings for the input tokens. The feature embeddings corresponding to the phrase or the full sentence (excluding start and end tokens) are then averaged and $L2$-normalized as the phrase/sentence feature representation, which is then dot-producted with the $L2$-normalized object feature representation.

5.2   Open-vocabulary Object Detection (OVOD)

First, X-DETR is evaluated on OVOD task [15, 40, 58]. LVIS [12], consisting of ∼1.2K categories, is used for OVOD evaluation. Two settings are evaluated: 1) without finetuning and 2) finetuning with different amount (i.e., 1%/10%/100%) of LVIS annotations. The former is for OVOD and the latter is to show the transferring ability of X-DETR to other downstream detection datasets. For finetuning, the model is finetuned for 50 epochs, with the object detection and object-sentence alignment losses, learning rate dropped at 40th epoch and an image resolution of 800. For X-DETR and R-CLIP, the final score of an instance is the product between the objectness score and the classification probabilities over the categories of interest.

The results are shown in Table 1. For OVOD, R-CLIP achieves 12.7 AP even when the CLIP model is trained with image-caption annotations, and including some context information gives additional gains of 1 AP, but with 2 times slower speeds. X-DETR

Table 1. OVOD detection results on LVIS-v1 (box AP). We followed [19] to evaluate on the 5k *minival* subset. The subscript "r/c/f" of AP is for rare/common/frequent categories of LVIS. "Train Time" is the finetuning time on LVIS.

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

outperforms R-CLIP by 3.7 points and is about 100 times faster, without using any LVIS annotations. When compared with MDETR, the recent state-of-the-art localization based V+L model, X-DETR is 10 points better. Note that X-DETR is even close to the fully-supervised vanilla DETR baseline (16.4 v.s. 17.8). These experiments support that the X-DETR can serve as an effective open-vocabulary object detector.

To evaluate the transferring ability, Mask R-CNN [13] trained on full LVIS data, using repeat factor sampling (RFS) for class imbalance, is regarded as a strong baseline, and a vanilla DETR pretrained on COCO as a transferring baseline, following [19]. The backbone for both of them is ResNet-101. When finetuning, X-DETR still has very strong improvements over MDETR, 6.1/5.3/11.5 points for 1%/10%/100% data. Using more data leads to worse results in MDETR (100% v.s. 10% data), showing that MDETR does not leverage detection data very well. But X-DETR has increasing gains with more detection data, whose result using 100% data is better than the strong Mask R-CNN baseline (34.0 v.s. 33.3). Note that the X-DETR finetuning is straightforward and uses no strategy for the category imbalance issue.

In addition to OVID results on LVIS, we test X-DETR on COCO (80 classes) [28], OpenImages-v6 (601 classes) [21] and Objects365-v1 (365 classes) [47], in Table 2. Although our pretraining uses the images from COCO and OpenImages, no category information is used. X-DETR achieves relatively good results on COCO (26.5 AP), but COCO only has 80 classes, and it is possible that the pretraining data has covered those categories information in the free-form language descriptions. Since Ojbects365 and OpenImages are much more challenging, the numbers of X-DETR are relatively lower, but are still much better than MDETR. R-CLIP has close results as X-DETR on COCO

Table 2. OVOD results on other datasets. "FT" means being finetuned on the target dataset.

Method	FT	COCO		Objects365		OpenImages	
		AP	AP50	AP	AP50	AP	AP50
Faster R-CNN [46]	✓	38.1	58.9	19.7	30.6	20.6	30.8
R-CLIP	✗	22.8	34.6	5.9	9.2	15.2	23.1
R-CLIP+	✗	24.4	36.8	6.5	10.1	16.0	24.1
MDETR [19]	✗	3.0	3.9	0.5	0.7	0.4	0.5
X-DETR (ours)	✗	26.5	38.9	5.7	8.6	4.8	6.7

Table 3. Multi-modal instance search results. Time is evaluated per query on RefCOCO (1,500 images) after feature indexing.

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

and Objects365, but much better on OpenImages (∼10 points gain). These have shown that OVOD is still a very challenging task and requires more research efforts.

5.3   Multi-modal Instance Search (MMIS)

Next, X-DETR is evaluated on MMIS [15,29,40]. This 1) is a practical problem, like the commercial search engines (Google, Bing, etc.), 2) is challenging, as there are tons of false positives due to the large-scale image database, and 3) requires high efficiency, i.e., instant retrieval results for millions/billions of images for commercial search engines. However, collecting MMIS datasets is very expensive, because it needs to annotate all bounding boxes in the full database for any free-form language query. This is a main reason that there is no publicly available dataset for this task yet[2].

To evaluate on this task, we converted the referred expression comprehension (REC) datasets (RefCOCO/RefCOCO+/RefCOCOg [20, 34, 57]), because REC is a special (simpler) case of MMIS, which is to search the result in a single image. We changed the evaluation protocol of REC, such that the retrieval operates on the full database instead of a single image. However, only one bounding box is associated with a given

---

[2] A recent work [29] discusses this, but no data is released yet.

Table 4. Phrase grounding results on Flickr30k entities validation.

Method	FT	R@1	R@5	R@10	Method	FT	R@1	R@5	R@10
R-CLIP	✗	21.9	48.4	60.0	VisualBert [24]	✓	68.1	84.0	86.2
MDETR [19]	✗	**82.5**	92.9	94.9	VisualBert† [24]	✓	70.4	84.5	86.3
X-DETR (ours)	✗	81.4	**93.6**	**95.6**	X-DETR (ours)	✓	81.8	**93.6**	95.5

query in REC datasets, which is not exclusive across the full database. For example, multiple bounding boxes from different images could correspond to a single query, especially when the query is somewhat general, e.g., "left man". Therefore, we used a loose evaluation metric, the recall @ the top {5, 10, 30} bounding boxes. The cross-modal similarity scores are used to rank and retrieve the objects for each query. The results shown in Table 3, and some MMIS examples are shown in the supplementary.

R-CLIP does not perform well for RefCOCO/RefCOCO+, because these two datasets require more context information, e.g., "left/right", which is missing in R-CLIP. Even when including the context region, it does not improve. For RefCOCOg, which focuses more on general language description, which CLIP was mainly trained on, R-CLIP shows much better results. On the contrary, X-DETR is better in all three datasets, especially on RefCOCO/RefCOCO+. When compared with 12-in-1 [33], UNITER [7] and MDETR [19], which use cross-modality transformers to model vision-language alignment and achieved very good results on REC tasks, X-DETR outperforms them by a large margin with and without finetuning. These observations are consistent with [40]: a good model for REC is not necessarily a good model for the MMIS task.

5.4 Phrase Grounding and Referring Expression Comprehension

Phrase grounding and referring expression comprehension (REC) are reduced versions of MMIS, which retrieve the targets most similar to the given query in a single image. They are simpler tasks than MMIS, which assumes that the object referred by the query definitely exists in the image.

Flickr30k entities dataset [39] is used for phrase grounding evaluation, with the train/val/test splits of [39] and Recall@{1, 5, 10} as the evaluation metrics. For finetuning, the pretrained X-DETR model was finetuned for 3 epochs on the target dataset, with the same losses as pretraining. The results are shown in Table 4. R-CLIP performs poorly on this task. Compared with MDETR [19], X-DETR only has a small gap for R@1 but better performance at R@5 and R@10, considering X-DETR does not use the much stronger transformer as the joint-modality modeling. Finetuning also helps X-DETR, but does not help MDETR as mentioned in [19].

RefCOCO, RefCOCO+ and RefCOCOg [20, 34, 57] are used for REC evaluation with Recall@{1, 5, 10} as the evaluation metrics. For finetuning, the pretrained model was finetuned for 4 epochs on the union of all three dataset (excluding all images in all three validation sets), with the object detection and object-sentence alignment losses introduced in Section 4. The results are shown in Table 5. Similar to grounding in Table 4, R-CLIP does not work very well for REC task. X-DETR outperforms UNITER$_L$ [7]

Table 5. Comparison with state-of-the-art on REC datasets.

Method	FT	RefCOCO val			RefCOCO+ val			RefCOCOg val		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
R-CLIP	✗	21.6	52.2	68.5	24.7	57.4	72.8	36.4	72.3	86.4
R-CLIP+	✗	17.3	45.6	62.0	19.8	48.6	65.4	32.4	68.0	82.7
MDETR [19]	✗	72.4	92.2	94.7	58.3	86.3	90.5	55.9	87.0	91.8
X-DETR (ours)	✗	78.7	95.4	97.6	63.5	92.5	96.2	60.4	91.8	95.6
MAttNet [56]	✓	76.7	-	-	65.3	-	-	66.6	-	
UNITER$_L$ [7]*	✓	81.4	-	-	75.9	-	-	74.9	-	
VILLA$_L$ [9]*	✓	82.4	-	-	76.2	-	-	76.2	-	
MDETR [19]	✓	**86.8**	96.0	97.2	**79.5**	96.2	97.5	**81.6**	95.5	96.8
X-DETR (ours)	✓	86.2	**97.8**	**98.9**	77.0	**97.1**	**98.6**	80.4	**96.8**	**97.9**

and VILLA$_L$ [9], on all three datasets. When compared with the current state-of-the-art MDETR [19] on the finetuning setting, X-DETR is slightly worse at R@1 but better at R@5 and R@10. When the model is not finetuned, X-DETR achieves much better results than MDETR. These results have shown that X-DETR can be generalized well for simpler tasks such as phrase grounding and REC, in addition to the challenging tasks of OVOD and MMIS.

5.5   Speed Comparisons

X-DETR is an efficient architecture for both training and inference. When compared with R-CLIP, X-DETR is about 100 times faster for OVOD due to the slow R-CNN pipeline. For MMIS, X-DETR has close retrieval time with R-CLIP, since they both use the simple dot-product as the vision-language alignment and vision/language features are fully indexable. However, the indexing speed of X-DETR is about 100 times faster than R-CLIP due to the R-CNN pipeline. When compared with the state-of-the-art localization based V+L work of MDETR, X-DETR is a few times faster during training, since it can process an image with all of its queries simultaneously but MDETR needs to process the queries one by one due to the joint-modality transformer design, as discussed in Section 4.3. For example, finetuning on 10% LVIS, MDETR needs 108 hours for 150 epochs, but X-DETR only needs 5.2 hours for 50 epochs with improved results (see Table 1). At inference, for OVOD/MMIS, X-DETR is about 100/100,000 times faster than MDETR (see Table 1 and 3). Note that we have already indexed the indexable features, e.g., the ones before the joint-modality transformers for MDETR. The training (inference) speeds are reported on 8 (1) A100 GPUs.

5.6   Ablation on Pretraining Data

We ablated the effect of the pretraining data in Table 6 (with ResNet-50 for efficiency purpose). The rows are for experiments using data of 1) the original *mixed* of MDETR

Table 6. Ablation studies. "*" means MMIS experiments. Finetuning on LVIS is on 10% data.

Pre-training data	FT	LVIS AP	LVIS AP@r	RefCOCOg* R@5	RefCOCOg* R@10	Flickr R@1	RefCOCO R@1
mixed	✗	11.5	6.0	16.2	24.6	79.7	63.9
*mixed**	✗	13.5	5.2	20.4	29.8	79.9	75.4
*mixed**+boxes	✗	14.7	5.1	22.4	32.4	79.9	78.0
+CC	✗	15.9	7.9	22.2	31.8	80.1	76.9
+CC+LocNar	✗	15.7	6.8	21.0	29.8	81.3	78.3
mixed	✓	20.0	13.4	34.3	45.8	80.1	83.0
*mixed**	✓	21.7	18.6	36.2	48.9	80.6	83.9
*mixed**+boxes	✓	22.8	14.8	36.5	49.4	80.3	84.1
+CC	✓	23.3	16.6	37.5	49.4	80.9	84.5
+CC+LocNar	✓	25.6	17.3	37.5	50.3	81.8	85.4

[19] ("*mixed*"), where all queries of an image are merged into a single paragraph, violating the independence assumption of the queries; 2) *mixed* with independent queries ("*mixed**"), by splitting the merged paragraph into independent queries; 3) adding COCO objects without category ("*mixed**+boxes"); 4) adding CC [48] weakly labeled image-caption data ("+CC"); 5) adding LocNar [41] pseudo-labeled and OpenImage [21] bounding box data ("+CC+LocNar"). It can be found that the original *mixed* dataset has a mismatch with the inference data, which leads to inferior results especially when the model is not finetuned. Independent queries of *mixed** can have significant accuracy boosts in multiple tasks. Adding COCO object data ("*mixed**+boxes") also has nontrivial improvements, showing the advantage of X-DETR to leverage detection-only data. When using CC data ("+CC"), the results of OVOD on LVIS are improved because more concepts are covered by more data, but the results on RefCOCO are decreased. When adding more LocNar pseudo-labeled data with OpenImage objects ("+CC+LocNar"), the results on Flickr30k and RefCOCO are improved over "+CC", but decreased on MMIS. The observations on finetuning settings are consistent. These ablation studies have shown that it is not enough to just use the fully annotated object-phrase annotations, and leveraging other weaker types of supervision could be helpful for instance-wise V+L learning.

6   Conclusion

In this paper, we propose a simple yet effective architecture for instance-wise vision-language tasks, which uses dot-product to align vision and language. It has shown that the expensive joint-modality transformer may not be necessary for those V+L tasks and the weak annotated data can be a big help to improve the model performances. The proposed X-DETR has shown benefits in terms of accuracy and speed, when compared to the previous V+L language state-of-the-art, on the practical and challenging tasks such as open-vocabulary detection and multi-modal instance search.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR. pp. 6077–6086 (2018)
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: ICCV. pp. 2425–2433 (2015)
3. Bansal, A., Sikka, K., Sharma, G., Chellappa, R., Divakaran, A.: Zero-shot object detection. In: ECCV. pp. 384–400 (2018)
4. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: CVPR. pp. 6154–6162 (2018)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229. Springer (2020)
6. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
7. Chen, Y., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: UNITER: universal image-text representation learning. In: ECCV. pp. 104–120. Springer (2020)
8. Desai, K., Johnson, J.: Virtex: Learning visual representations from textual annotations. In: CVPR. pp. 11162–11173 (2021)
9. Gan, Z., Chen, Y.C., Li, L., Zhu, C., Cheng, Y., Liu, J.: Large-scale adversarial training for vision-and-language representation learning. arXiv preprint arXiv:2006.06195 (2020)
10. Girshick, R.B., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR. pp. 580–587. IEEE Computer Society (2014)
11. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921 (2021)
12. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: CVPR. pp. 5356–5364 (2019)
13. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV. pp. 2961–2969 (2017)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
15. Hinami, R., Satoh, S.: Discriminative learning of open-vocabulary object retrieval and localization by negative phrase augmentation. arXiv preprint arXiv:1711.09509 (2017)
16. Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., Darrell, T.: Natural language object retrieval. In: CVPR. pp. 4555–4564 (2016)
17. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: CVPR. pp. 6700–6709 (2019)
18. Jia, C., Yang, Y., Xia, Y., Chen, Y., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: Meila, M., Zhang, T. (eds.) ICML. vol. 139, pp. 4904–4916. PMLR (2021)
19. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding. In: ICCV. pp. 1780–1790 (2021)
20. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: EMNLP. pp. 787–798 (2014)
21. Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Veit, A., et al.: Openimages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from https://github. com/openimages **2**(3), 18 (2017)

22. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis. **123**(1), 32–73 (2017)

23. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: ECCV. pp. 201–216 (2018)

24. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)

25. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: CVPR. pp. 10965–10975 (2022)

26. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., Gao, J.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) ECCV. pp. 121–137. Springer (2020)

27. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR. pp. 2117–2125 (2017)

28. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Fleet, D.J., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV. vol. 8693, pp. 740–755. Springer (2014)

29. Liu, S., Lin, K., Wang, L., Yuan, J., Liu, Z.: Ovis: Open-vocabulary visual instance search via visual-semantic aligned representation learning. arXiv preprint arXiv:2108.03704 (2021)

30. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: ECCV. pp. 21–37. Springer (2016)

31. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

32. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: NeurIPS. pp. 13–23 (2019)

33. Lu, J., Goswami, V., Rohrbach, M., Parikh, D., Lee, S.: 12-in-1: Multi-task vision and language representation learning. In: CVPR. pp. 10437–10446 (2020)

34. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: CVPR. pp. 11–20 (2016)

35. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632 (2014)

36. Miech, A., Alayrac, J.B., Laptev, I., Sivic, J., Zisserman, A.: Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In: CVPR. pp. 9826–9836 (2021)

37. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)

38. Ordonez, V., Kulkarni, G., Berg, T.L.: Im2text: Describing images using 1 million captioned photographs. In: NeurIPS. pp. 1143–1151 (2011)

39. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. Int. J. Comput. Vis. **123**(1), 74–93 (2017)

40. Plummer, B.A., Shih, K., Li, Y., Xu, K., Lazebnik, S., Sclaroff, S., Saenko, K.: Revisiting image-language networks for open-ended phrase detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)

41. Pont-Tuset, J., Uijlings, J.R.R., Changpinyo, S., Soricut, R., Ferrari, V.: Connecting vision and language with localized narratives. In: ECCV. vol. 12350, pp. 647–664. Springer (2020)

42. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML. vol. 139, pp. 8748–8763. PMLR (2021)

43. Rahman, S., Khan, S., Porikli, F.: Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In: ACCV. pp. 547–563. Springer (2018)
44. Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G.R., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieval. In: ACMMM. pp. 251–260 (2010)
45. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR. pp. 779–788 (2016)
46. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. NeurIPS **28**, 91–99 (2015)
47. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: ICCV. pp. 8430–8439 (2019)
48. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL. pp. 2556–2565 (2018)
49. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490 (2019)
50. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS. pp. 5998–6008 (2017)
51. Wang, L., Li, Y., Huang, J., Lazebnik, S.: Learning two-branch neural networks for image-text matching tasks. IEEE Trans. Pattern Anal. Mach. Intell. **41**(2), 394–407 (2018)
52. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771 (2019)
53. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML. pp. 2048–2057. PMLR (2015)
54. Yang, Z., Gong, B., Wang, L., Huang, W., Yu, D., Luo, J.: A fast and accurate one-stage approach to visual grounding. In: ICCV. pp. 4683–4693 (2019)
55. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: CVPR. pp. 21–29 (2016)
56. Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L.: Mattnet: Modular attention network for referring expression comprehension. In: CVPR. pp. 1307–1315 (2018)
57. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: ECCV. vol. 9906, pp. 69–85. Springer (2016)
58. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: CVPR. pp. 14393–14402 (2021)
59. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: Vinvl: Revisiting visual representations in vision-language models. In: CVPR. pp. 5579–5588 (2021)
60. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: CVPR. pp. 16793–16803 (2022)
61. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. In: ICLR (2021)