MDETR - Modulated Detection for End-to-End Multi-Modal Understanding

Aishwarya Kamath[1] Mannat Singh[2] Yann LeCun[123] Gabriel Synnaeve[2] Ishan Misra[2]
Nicolas Carion[3]
[1]NYU Center for Data Science [2]Facebook AI Research [3]NYU Courant Institute
{aish, yann.lecun, nc2794}@nyu.edu, {mannatsingh,imisra,gab}@fb.com

Abstract

Multi-modal reasoning systems rely on a pre-trained object detector to extract regions of interest from the image. However, this crucial module is typically used as a black box, trained independently of the downstream task and on a fixed vocabulary of objects and attributes. This makes it challenging for such systems to capture the long tail of visual concepts expressed in free form text. In this paper we propose MDETR, an end-to-end modulated *detector that detects objects in an image conditioned on a raw text query, like a caption or a question. We use a transformer-based architecture to reason jointly over text and image by fusing the two modalities at an early stage of the model. We pre-train the network on 1.3M text-image pairs, mined from pre-existing multi-modal datasets having explicit alignment between phrases in text and objects in the image. We then fine-tune on several downstream tasks such as phrase grounding, referring expression comprehension and segmentation, achieving state-of-the-art results on popular benchmarks. We also investigate the utility of our model as an object detector on a given label set when fine-tuned in a few-shot setting. We show that our pre-training approach provides a way to handle the long tail of object categories which have very few labelled instances. Our approach can be easily extended for visual question answering, achieving competitive performance on GQA and CLEVR. The code and models are available at* https://github.com/ashkamath/mdetr.

1. Introduction

Object detection forms an integral component of most state-of-the-art multi-modal understanding systems [6, 24], typically used as a black-box to detect a fixed vocabulary of concepts in an image followed by multi-modal alignment. This "pipelined" approach limits co-training with other modalities as context and restricts the downstream model to only have access to the detected objects and not the whole image. In addition, the detection system is usually frozen, which prevents further refinement of the model's perceptive capability. In the vision-language setting, it implies restricting the vocabulary of the resulting system to the categories and attributes of the detector, and is often a bottleneck for performance on these tasks [65]. As a result, such a system cannot recognize novel combinations of concepts expressed in free-form text.

Figure 1: Output of MDETR for the query "A pink elephant". The colors are not segmentation masks but the real colors of the pixels. The model has never seen a pink nor a blue elephant in training.

A recent line of work [59, 39, 12] considers the problem of text-conditioned object detection. These methods extend mainstream one-stage and two-stage detection architectures to achieve this goal. However, to the best of our knowledge, it has not been demonstrated that such detectors can improve performance on downstream tasks that require reasoning over the detected objects, such as visual question answering (VQA). We believe this is because these detectors are not end-to-end differentiable and thus cannot be trained in synergy with downstream tasks.

Our method, MDETR, is an end-to-end *modulated* detector based on the recent DETR [2] detection framework, and performs object detection in conjunction with natural language understanding, enabling truly end-to-end multi-modal reasoning. MDETR relies solely on text and aligned boxes as a form of supervision for concepts in an image. Thus, unlike current detection methods, MDETR detects nuanced concepts from free-form text, and generalizes to unseen combinations of categories and attributes. We showcase such a combination as well as modulated detection in

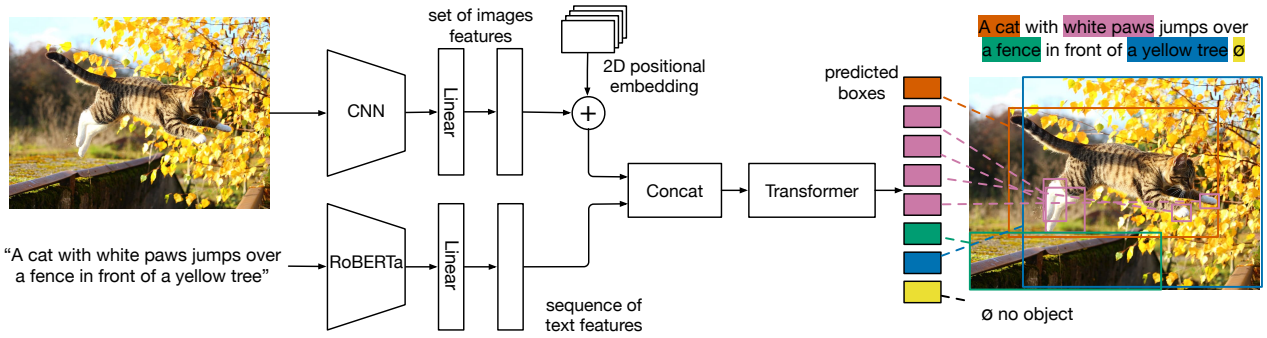

Figure 2: MDETR uses a convolutional backbone to extract visual features, and a language model such as RoBERTa to extract text features. The features of both modalities are projected to a shared embedding space, concatenated and fed to a transformer encoder-decoder that predicts the bounding boxes of the objects and their grounding in text.

Fig. 1. By design, our predictions are grounded in text, which is a key requirement for visual reasoning [58]. When pre-trained using a dataset of 200,000 images and aligned text with box annotations, we achieve best reported results on the Flickr30k dataset for phrase grounding, RefCOCO/+/g datasets for referring expression comprehension, and referring expression segmentation on PhraseCut, as well as competitive performance on the GQA and CLEVR benchmarks for visual question answering.

Our contributions are as follows:

- We introduce an end-to-end text-modulated detection system derived from the DETR detector.
- We demonstrate that the modulated detection approach can be applied seamlessly to solve tasks such as phrase grounding and referring expression comprehension, setting new state of the art performance on both these tasks using datasets having synthetic as well as real images.
- We show that good modulated detection performance naturally translates to downstream task performance, for instance achieving competitive performance on visual question answering, referring expression segmentation, and on few-shot long-tailed object detection.

2. Method

In this section we first briefly summarize the object detection pipeline [2] based on which we build our model in §2.1 and then describe how we extend it for modulated detection in §2.2.

2.1. Background

DETR Our approach to modulated detection builds on the DETR system [2], which we briefly review here. We refer the readers to the original paper for additional details. DETR is an end-to-end detection model composed of a backbone (typically a convolutional residual network [11]), followed by a Transformer Encoder-Decoder [51].

The DETR encoder operates on 2D flattened image features from the backbone and applies a series of transformer layers. The decoder takes as input a set of N learned embeddings called *object queries*, that can be viewed as slots that the model needs to fill with detected objects. All the object queries are fed in parallel to the decoder, which uses cross-attention layers to look at the encoded image and predicts the output embeddings for each of the queries. The final representation of each object query is independently decoded into box coordinates and class labels using a shared feed-forward layer. The number of object queries acts as a de facto upper-bound on the number of objects the model can detect simultaneously. It has to be set to a sufficiently large upper-bound on the number of objects one may expect to encounter in a given image. Since the actual number of objects in a particular image may be less than the number of queries N, an extra class label corresponding to "no object" is used, denoted by $\varnothing$. The model is trained to output this class for every query that doesn't correspond to an object.

DETR is trained using a Hungarian matching loss, where a bipartite matching is computed between the N proposed objects and the ground-truth objects. Each matched object is supervised using the corresponding target as ground-truth, while the un-matched objects are supervised to predict the "no object" label $\varnothing$. The classification head is supervised using standard cross-entropy, while the bounding box head is supervised using a combination of absolute error (L1 loss) and Generalized IoU [42].

2.2. MDETR

2.2.1 Architecture

We depict the architecture for MDETR in Fig. 2. As in DETR, the image is encoded by a convolutional backbone and flattened. In order to conserve the spatial information, 2-D positional embeddings are added to this flattened vec-

tor. We encode the text using a pre-trained transformer language model to produce a sequence of hidden vectors of same size as the input. We then apply a modality dependent linear projection to both the image and text features to project them into a shared embedding space. These feature vectors are then concatenated on the sequence dimension to yield a single sequence of image and text features. This sequence is fed to a joint transformer encoder termed as the *cross encoder*. Following DETR, we apply a transformer decoder on the object queries while cross attending to the final hidden state of the cross encoder. The decoder's output is used for predicting the actual boxes.

2.2.2 Training

We present the two additional loss functions used by MDETR, which encourage alignment between the image and the text. Both of these use the same source of annotations: free form text with aligned bounding boxes. The first loss function that we term as the *soft token prediction* loss is a non parametric alignment loss. The second, termed as the *text-query contrastive alignment* is a parametric loss function enforcing similarity between aligned object queries and tokens.

Soft token prediction For modulated detection, unlike in the standard detection setting, we are not interested in predicting a categorical class for each detected object. Instead, we predict the span of tokens from the original text that refers to each matched object. Concretely, we first set the maximum number of tokens for any given sentence to be L = 256. For each predicted box that is matched to a ground truth box using the bi-partite matching, the model is trained to predict a uniform distribution over all *token positions* that correspond to the object. Fig. 2 shows an example where the box for cat is trained to predict a uniform distribution over the first two words. In Fig. **??**, we show a simplified visualization of the loss for this example, in terms of a distribution over words for each box, but in practice we use token spans after tokenization using a BPE scheme [44]. Any query that is not matched to a target is trained to predict the "no object" label $\varnothing$. Note that several words in the text could correspond to the same object in the image, and conversely several objects could correspond to the same text. For example, "a couple" referred to by two boxes in the image, could further be referred to individually in the same caption. By designing the loss function in this way, our model is able to learn about co-referenced objects from the same referring expression.

Contrastive alignment While the soft token prediction uses *positional* information to align the objects to text, the contrastive alignment loss enforces alignment between the *embedded representations* of the object at the output of the decoder, and the text representation at the output of the cross encoder. This additional contrastive alignment loss ensures that the embeddings of a (visual) object and its corresponding (text) token are closer in the feature space compared to embeddings of unrelated tokens. This constraint is stronger than the soft token prediction loss as it directly operates on the representations and is not solely based on positional information. More concretely, consider the maximum number of tokens to be L and maximum number of objects to be N. Let T_i^+ be the set of tokens that a given object o_i should be aligned to, and O_i^+ be the set of objects to be aligned with a given token t_i.
The contrastive loss for all objects, inspired by InfoNCE [34] is normalized by number of positive tokens for each object and can be written as follows:

$$l_o = \sum_{i=0}^{N-1} \frac{1}{|T_i^+|} \sum_{j \in T_i^+} -\log \left(\frac{\exp(o_i^\top t_j / \tau)}{\sum_{k=0}^{L-1} \exp(o_i^\top t_k / \tau)} \right) \quad (1)$$

where τ is a temperature parameter that we set to 0.07 following literature [56, 41]. By symmetry, the contrastive loss for all tokens, normalized by the number of positive objects for each token is given by:

$$l_t = \sum_{i=0}^{L-1} \frac{1}{|O_i^+|} \sum_{j \in O_i^+} -\log \left(\frac{\exp(t_i^\top o_j / \tau)}{\sum_{k=0}^{N-1} \exp(t_i^\top o_k / \tau)} \right) \quad (2)$$

We take the average of these two loss functions as our contrastive alignment loss.

Combining all the losses In MDETR, a bipartite matching is used to find the best match between the predictions and the ground truth targets just as in DETR. The main difference is that there is no class label predicted for each object - instead predicting a uniform distribution over the relevant positions in the text that correspond to this object (soft token predictions), supervised using a soft cross entropy. The matching cost consists of this in addition to the L1 & GIoU loss between the prediction and the target box as in DETR. After matching, the total loss consists of the box prediction losses (L1 & GIoU), soft-token prediction loss, and the contrastive alignment loss.

3. Experiments

In this section we describe the data and training used for pre-training MDETR, and provide details and results on the tasks that we use to evaluate our approach. Results on the CLEVR dataset are reported in Table 1. For a discussion on the CLEVR results and further details on data preparation and training, please see Appendix **??**. Experimental details for pre-training and downstream tasks on natural images are detailed in §3.1 and §3.2.

3.1. Pre-training Modulated Detection

For pre-training, we focus on the task of modulated detection where the aim is to detect all the objects that are re-

Method	CLEVR	CLEVR-Hu		CoGenT		CLEVR-Ref+
	Overall	- FT	+ FT	TestA	TestB	Acc
MAttNet[62]	-	-	-	-	-	60.9
MGA-Net[66]	-	-	-	-	-	80.1
FiLM[36]	97.7	56.6	75.9	98.3	**78.8**	-
MAC [15]	98.9	57.4	81.5	-	-	-
NS-VQA[60]*	**99.8**	-	67.8	**99.8**	63.9	-
OCCAM [52]	99.4	-	-	-	-	-
MDETR	99.7	**59.9**	**81.7**	**99.8**	76.7	**100**

Table 1: Results on CLEVR-based datasets. We report accuracies on the test set of CLEVR. On CLEVR-Humans, we report accuracy on the test set before and after fine-tuning. On CoGenT, we report performance when the model is trained in condition A, without finetuning on condition B. On CLEVR-Ref+, we report the accuracy on the subset where the referred object is unique. *indicates method uses external program annotations. Further details in Appendix **??**.

ferred to in the aligned free form text. We create a combined dataset using images from the Flickr30k [40], MS COCO [26] and Visual Genome (VG) [20] datasets. Annotations from the referring expressions datasets, VG regions, Flickr entities and GQA train balanced set are used for training. An image may have several text annotations associated with it. Details on the datasets can be found in Appendix **??**.

Data combination For each image, we take all annotations from these datasets and combine the text that refers to the same image while ensuring that all images that are in the validation or testing set for all our downstream tasks are removed from our train set. The combination of sentences is done using a graph coloring algorithm which ensures that only phrases having boxes with GIoU ≤ 0.5 are combined, and that the total length of a combined sentence is less than 250 characters. In this way, we arrive at a dataset having 1.3M aligned image - text pairs. This combination step is important for two reasons: 1) data efficiency, by packing more information into a single training example and 2) it provides a better learning signal for our soft token prediction loss since the model has to learn to disambiguate between multiple occurrences of the same object category, as depicted in Fig 3. In the single sentence case, the soft token prediction task becomes trivial since it can always predict the root of the sentence without looking at the image. Experimentally, we find that such dense annotations translate to better grounding between text and image and subsequently to better downstream performance.

Model We use a pre-trained RoBERTa-base [27] as our text encoder, having 12 transformer encoder layers, each with hidden dimension of 768 and 12 heads in the multi-head attention. We use the implementation and weights from HuggingFace [54]. For the visual backbone, we explore two options. The first is a ResNet-101 [11] pretrained on ImageNet with frozen batchnorm layers, taken from

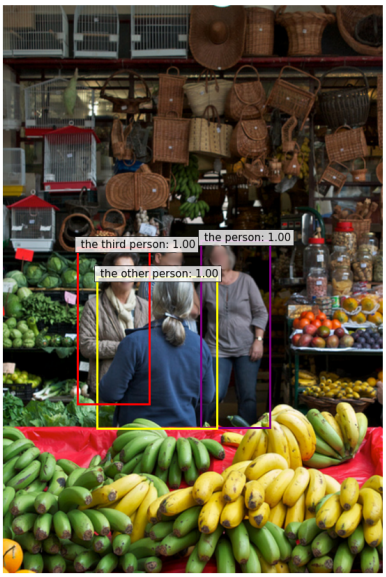

Figure 3: Our combination of annotations results in examples such as the following: "the person in the grey shirt with a watch on their wrist. the other person wearing a blue sweater. the third person in a gray coat and scarf." We show the predictions from our model for this caption. It is able to pay attention to all the objects in the image and then disambiguate between them based on the text. The model is trained to predict the root of the phrase as the positive token span, which as we can see in this figure, correctly refers to the three different people.

Torchvision. This is to be comparable with current literature in the space of multi-modal understanding, where the popular approach is to use the BUTD object detector with a Resnet-101 backbone from [1] trained on the VG dataset. In our work, we are not limited by the existence of pre-trained detectors, and inspired by its success in object detection [50], we choose to explore the EfficientNet family [49] for our backbone. We use a model which was trained on large amounts of unlabelled data in addition to ImageNet, using a pseudo-labelling technique called Noisy-Student [57]. We choose the EfficientNetB3, which achieves 84.1% top 1 accuracy on ImageNet with only 12M weights and EfficientB5 which achieves 86.1% using 30M weights. We use the implementation provided by the Timm library [53], and freeze the batchnorm layers. We pre-train our model for 40 epochs on 32 V100 gpus with an effective batch size of 64, which takes approximately a week to train. Training hyperparameters are detailed in Appendix **??**.

3.2. Downstream Tasks

We evaluate our method on 4 downstream tasks: referring expression comprehension and segmentation, visual question answering and phrase grounding. Training hyperprameters for all tasks can be found in Appendix **??**.

Phrase grounding Given one or more phrases, which

Method	Detection backbone	Pre-training image data	RefCOCO			RefCOCO+			RefCOCOg	
			val	testA	testB	val	testA	testB	val	test
MAttNet[62]	R101	None	76.65	81.14	69.99	65.33	71.62	56.02	66.58	67.27
ViLBERT[28]	R101	CC (3.3M)	-	-	-	72.34	78.52	62.61	-	-
VL-BERT_L [46]	R101	CC (3.3M)	-	-	-	72.59	78.57	62.30	-	-
UNITER_L[6]*	R101	CC, SBU, COCO, VG (4.6M)	81.41	87.04	74.17	75.90	81.45	66.70	74.86	75.77
VILLA_L[9]*	R101	CC, SBU, COCO, VG (4.6M)	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71
ERNIE-ViL_L[61]	R101	CC, SBU (4.3M)	-	-	-	75.95	82.07	66.88	-	-
MDETR	R101	COCO, VG, Flickr30k (200k)	**86.75**	**89.58**	**81.41**	**79.52**	**84.09**	**70.62**	**81.64**	**80.89**
MDETR	ENB3	COCO, VG, Flickr30k (200k)	**87.51**	**90.40**	**82.67**	**81.13**	**85.52**	**72.96**	**83.35**	**83.31**

Table 2: Accuracy results on referring expression comprehension. *As mentioned in UNITER [6], methods using box proposals from the BUTD detector [1] suffer from a test set leak, since the detector was trained on images including the validation and test set of the RE comprehension datasets. We report numbers for these methods from their papers using these "contaminated features" but we would like to stress that all of our pre-training excluded the images used in the val/test of any of the downstream datasets including for RE comprehension. CC refers to Conceptual Captions [45], VG to Visual Genome [20], SBU refers to the SBU Captions[35] and COCO to Micosoft COCO [26].

Method	Val			Test		
	R@1	R@5	R@10	R@1	R@5	R@10
ANY-BOX-PROTOCOL						
BAN [19]	-	-	-	69.7	84.2	86.4
VisualBert[22]	68.1	84.0	86.2	-	-	-
VisualBert†[22]	70.4	84.5	86.3	71.3	85.0	86.5
MDETR-R101	78.9	88.8	90.8	-	-	-
MDETR-R101†∗	**82.5**	**92.9**	**94.9**	**83.4**	**93.5**	**95.3**
MDETR-ENB3†∗	**82.9**	**93.2**	**95.2**	**84.0**	**93.8**	**95.6**
MDETR-ENB5†∗	**83.6**	**93.4**	**95.1**	**84.3**	**93.9**	**95.8**
MERGED-BOXES-PROTOCOL						
CITE [37]	-	-	-	61.9	-	-
FAOG [59]	-	-	-	68.7	-	-
SimNet-CCA [39]	-	-	-	71.9	-	-
DDPN [64]	72.8	-	-	73.5	-	-
MDETR-R101	79.0	86.7	88.6	-	-	-
MDETR-R101†∗	**82.3**	**91.8**	**93.7**	**83.8**	**92.7**	**94.4**

Table 3: Results on the phrase grounding task on Flickr30k entities dataset [40]. Models with † are pre-trained on COCO, models with ∗ are also pre-trained on VG and Flickr 30k. Our models (MDETR) use a RoBERTa text encoder while other models use RNNs, word2vec-based features, or BERT (comparable to RoBERTa) text encoders. All models use a ResNet101 backbone, except MDETR-ENB3 which uses EfficientNet-B3 and MDETR-ENB5 with an EfficientNet-B5.

may be inter-related, the task is to provide a set of bounding boxes for each phrase. We use the Flickr30k entities dataset for this task, with the train/val/test splits as provided by [40] and evaluate our performance in terms of Recall@k. For each sentence in the test set, we predict 100 bounding boxes and use the soft token alignment prediction to rank the boxes according to the score given to the token positions that correspond to the phrase. We evaluate under two protocols which we name ANY-BOX [22, 19] and MERGED-BOXES [38]. Please see Appendix **??** for a discussion on the two protocols. We compare our method to existing state-of-the-art results from two types of approaches - the text conditioned detection models [39, 59] and a transformer based vision-language pre-training model [22]. In the ANY-BOX setting, we obtain a 8.5 point boost over current state of the art on this task as measured in terms of Recall@1 on the validation set, without using any pre-training (no additional data). With pre-training, we further obtain a 12.1 point boost over the best model's performance on the test set, while using the same backbone.

Referring expression comprehension Given an image and a referring expression in plain text, the task is to localize the object being referred to by returning a bounding box around it. The approach taken by most prior work [62, 28, 6, 61] on this task has been to rank a set of pre-extracted bounding boxes associated with an image, that are obtained using a pre-trained object detector. In this paper, we solve a much harder task - we train our model to directly *predict* the bounding box, given a referring expression and the associated image. There are three established datasets for this task called RefCOCO, RefCOCO+ [63] and RefCOCOg [30]. Since during pre-training we annotate every object referred to within the text, there is a slight shift in the way the model is used in this task. For example, during pre-training, given the caption "The woman wearing a blue dress standing next to the rose bush.", MDETR would be trained to predict boxes for all referred objects such as the woman, the blue dress and the rose bush. However, for referring expressions, the task would be to only return *one* bounding box, which signifies the woman being referred to by the entire expression. For this reason, we finetune the

Method	Backbone	PhraseCut			
		M-IoU	Pr@0.5	Pr@0.7	Pr@0.9
RMI[3]	R101	21.1	22.0	11.6	1.5
HULANet[55]	R101	41.3	42.4	27.0	5.7
MDETR	R101	**53.1**	**56.1**	**38.9**	**11.9**
MDETR	ENB3	**53.7**	**57.5**	**39.9**	**11.9**

Table 4: Following [55], we report the mean intersection-over-union (IoU) of our masks with the ground-truth masks. We also report the precision Pr@I of our model, where success is marked when our proposed mask has an IoU with the ground-truth higher than the threshold I. With a comparable ResNet backbone, we observe consistent gains across all metrics over HULANet [55], the current state-of-the-art. The EfficientNet backbone further improves on those results.

model on the task specific dataset for 5 epochs. At inference time, we use the $\varnothing$ label to rank the 100 detected boxes. Let $P(\varnothing)$ be the probability assigned to the "no object" label, we rank by decreasing order of $1 - P(\varnothing)$. We report results in Table 2, showing large improvements over state-of-the-art across all datasets.

Referring expression segmentation Similarly to DETR, we show that our approach can be extended to perform segmentation by evaluating on the referring expression segmentation task of the recent PhraseCut [55] dataset which consists of images from VG, annotated with segmentation masks for each referring expression. These expressions comprise a wide vocabulary of objects, attributes and relations, making it a challenging benchmark. Contrary to other referring expression segmentation datasets, in PhraseCut the expression may refer to several objects. The model is expected to find all the corresponding instances. Our training occurs in two stages. In the first step, we take our pre-trained model after 40 epochs and fine-tune it for 10 epochs on this dataset, supervising the model to output correct boxes for the referred expressions. We use the box AP on the validation set for early stopping. In the second stage, following [2], we freeze the weights of the network, and only train a segmentation head for 35 epochs, with a learning rate drop at 25 epochs, supervised using a combination of the Dice/F1 loss[32] and the Focal loss [25]. At inference-time, we assign a confidence to each predicted box equal to $1 - P(\varnothing)$ where $P(\varnothing)$ is the probability assigned to the "no-object" token (see §2). We then filter the boxes with a confidence lower than 0.7. Finally, we merge the masks corresponding to each of these boxes into one binary mask corresponding to this referring expression. The results are collected in Table 4. Our model is able to produce clean masks for a wide variety of long tailed-concepts covered by PhraseCut. Example predictions from our model on this dataset are given in Appendix **??**.

Visual Question Answering We evaluate our hypothesis that modulated detection is a useful component for multi-modal reasoning by fine-tuning our pre-trained model on the GQA dataset. To train MDETR, we use the scene graph provided in GQA to obtain the alignment between question words and the boxes. Our model architecture is depicted in Fig 4. Object queries are learned embeddings input to the decoder, each of which can be used to a detect an object. Apart from the 100 queries that are used for detection, we use additional queries that specialize in the type of question as well as one that is used to predict the type of question, where the types are defined in the GQA annotations as REL, OBJ, GLOBAL, CAT and ATTR. We take our pre-trained model trained for 40 epochs on our combined dataset, and initialise these queries as well as the heads for each of them randomly, and fine-tune first for 125 epochs on the unbalanced *all* GQA split, followed by 10 epochs on the *balanced* split similar to what is done in prior work [24, 5]. During the first 125 epochs, we train the modulated detection losses along with the question answering, but put a weight on question answering loss that encourages the model to focus more on this task. For the balanced split fine-tuning, we only use the question answering loss. During inference, the type head predicts the type of question and the answer is taken from that head. Using our model with a Resnet-101 backbone, we not only outperform LXMERT [47] and VL-T5 [7] which use comparable amount of data, but also OSCAR [24] which uses magnitude more data in their pre-training. MDETR with the EfficientNet-B5 backbone is able to push performance even higher as reported in Table 5. The NSM model makes use of an external scene graph generation model, while the MMN model makes use of the scene graph and functional programs during training.

Method	Pre-training img data	Test-dev	Test-std
MoVie [33]	-	-	57.10
LXMERT[47]	VG, COCO (180k)	60.0	60.33
VL-T5 [7]	VG, COCO (180k)	-	60.80
MMN [5]	-	-	60.83
OSCAR [24]	VG, COCO, Flickr, SBU (4.3M)	61.58	61.62
NSM [16]	-	-	63.17
VinVL [65]	VG, COCO, Objects365, SBU Flickr30k, CC, VQA, OpenImagesV5 (5.65M)	65.05	64.65
MDETR-R101	VG, COCO, Flickr30k (200k)	62.48	61.99
MDETR-ENB5	VG, COCO, Flickr30k (200k)	62.95	62.45

Table 5: Visual question answering on the GQA dataset.

3.2.1 Few-shot transfer for long-tailed detection

Inspired by the success of CLIP [41], on zero-shot transfer for image classification, we explore the opportunity to construct a useful detector over a given label set from a pre-

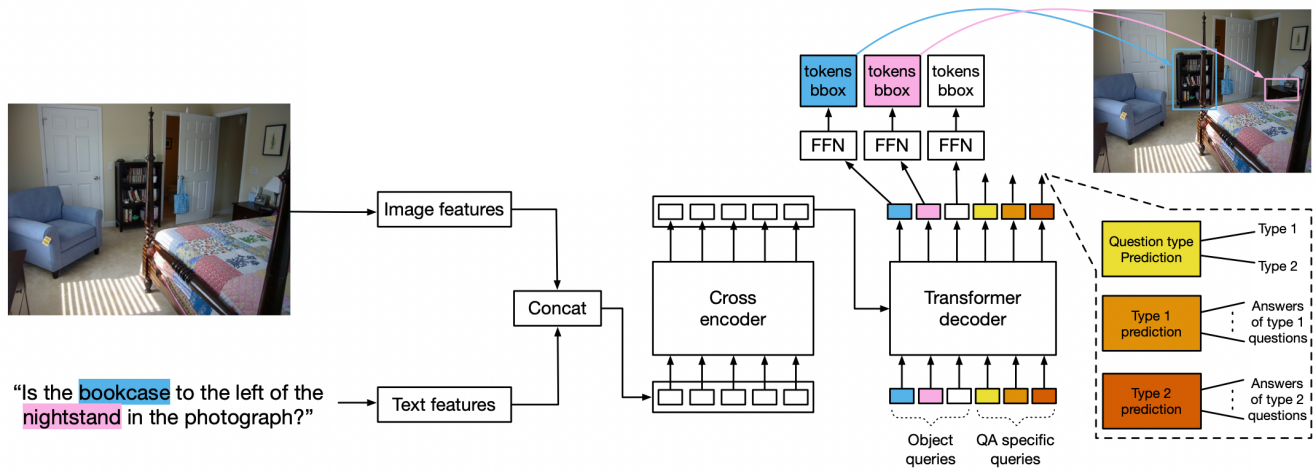

Figure 4: During MDETR pre-training, the model is trained to detect all objects mentioned in the question. To extend it for question answering, we provide QA specific queries in addition to the object queries as input to the transformer decoder. We use specialized heads for different question types.

Figure 5: MDETR provides interpretable predictions as seen here. For the question "What is on the table?", MDETR fine-tuned on GQA predicts boxes for key words in the question, and is able to provide the correct answer as "laptop". Image from COCO val set.

Method	Data	AP	AP50	AP_r	AP_c	AP_f
Mask R-CNN	100%	33.3	51.1	26.3	34.0	33.9
DETR	1%	4.2	7.0	1.9	1.1	7.3
DETR	10%	13.7	21.7	4.1	13.2	15.9
DETR	100%	17.8	27.5	3.2	12.9	24.8
MDETR	1%	16.7	25.8	11.2	14.6	19.5
MDETR	10%	24.2	38.0	20.9	24.9	24.3
MDETR	100%	22.5	35.2	7.4	22.7	25.0

Table 6: Box AP *fixed* results on LVIS-v1. Since the validation set of LVIS contains some training images from MSCOCO, we report results on the subset of 5k validation images that our model has never seen during training. We call this subset *minival*. All models use a Resnet 101 as backbone. Mask-RCNN can be regarded as a strong representative of the detection performance of current approaches on this dataset, using bells and whistles such as Repeat Factor Sampling (RFS) to address class imbalance. We use a vanilla DETR pretrained on MSCOCO as a few-shot transfer baseline, and show that our pre-training on natural text improves performance significantly, especially on rare categories.

trained MDETR model. Unlike CLIP, we do not ensure our pre-training dataset contains a balanced representation of all the target classes. By construction, our dataset has no training instances where there are zero boxes aligned to the text, biasing the model to always predict boxes for a given text. This prevents evaluating in a true zero-shot transfer setting, so we turn instead to a few-shot setting, where the model is trained on a fraction of the available labelled data. We conduct our experiments on the LVIS dataset [10], a detection dataset with a large vocabulary of 1.2k categories, with a long-tail that contains very few training samples, making it a challenging dataset for current approaches. Federated datasets often pose problems to standard detectors, and require developing specific loss functions [48]. However this property makes it well suited to train MDETR: for each positive category, we create a training instance composed of the image and a text version of the class name, and provide as annotations all the instances of this category. For each negative category, we provide the class name and an empty set of annotations. For inference on a given image, we query each possible class name, then merge the sets of boxes detected on each of the text prompts. This inference scheme costs about 10s/image on a GPU.

We fine-tune MDETR on three subsets of the LVIS train set, each containing respectively 1%, 10% and 100% of the images. We ensure a balanced sampling of the cate-

gories, such that our 1% set contains at least one positive and one negative examples from each category. We compare to two baselines: the first one is Mask-RCNN trained exclusively on the full training set of LVIS. The other is a DETR model pre-trained on MSCOCO then fine-tuned on the various subsets of the LVIS training set. Our results are shown in Table 6. Following recent recommendation [8] on AP evaluation in the context of large vocabulary, we report the box AP *fixed*, obtained by limiting the number of detections per category instead of per image. Even with as little as 1 example per class, MDETR leverages the text pre-training and outperforms a fully fine-tuned DETR on rare categories. We note however that under full fine-tuning on the whole training set, the performance on rare objects drops significantly from 20.9 AP with 10% data to 7.5 with 100%, likely due to the extreme class imbalance. We expect that common techniques such as Repeat Factor Sampling will improve the situation in future work.

4. Related work

The CLEVR dataset [17] is a popular vision-language benchmark for reasoning on objects, their relations, and the composition of such relations. A prominent line of work [18, 60, 31, 13] makes use of the functional programs annotations that are part of the CLEVR dataset. Such approaches tend to dominate on the question answering benchmark, but fail to generalize beyond synthetic data. Conversely, many approaches [36, 43, 52, 15] learn directly from images or pre-detected objects, with varying amounts of inductive bias tailored to the QA task. Our method can be seen as an in-between: while not explicitly using the program supervision, it is trained to detect objects that are required for performing intermediate reasoning steps.

Recent progress in multi-modal understanding has been mainly powered by pre-training large transformer models to learn generic multi-modal representations from enormous amounts of aligned image-text data [45], then fine-tuning them on downstream tasks. These methods can be divided into single stream [6, 24, 65, 22] and two-stream [47, 28, 29, 46] architectures depending on whether the text and images are processed by a single combined transformer or two separate transformers followed by some cross attention layers. For both these types, the prevalent approach is to extract visual and textual features independently and then use the attention mechanism of the transformers to learn an alignment between the two. While this approach has improved state of the art results on a wide variety of tasks such as image-text retrieval [65], phrase grounding [22], image captioning [24] and visual question answering [21], it leaves opportunity for a more tightly knit architecture, such as MDETR, in which information flows between the two modalities at an even earlier stage of the model. Some previous attempts at achieving this using modulated architectures such as [36] and [33] show improvements on counting tasks and visual question answering.

The visual features used by the current state-of-the-art models are extracted using an external pre-trained detector [1], which outputs regions that are noisy, often oversampled and ambiguous. [24] attempts to alleviate the problem of noisy image features by using tags as anchors between the text and images. This is still a weaker form of supervision than in MDETR where we have explicit alignment between words or phrases in text and the objects in the images. To alleviate the constraints implied by fixed vocabulary of concepts, [65] trains on a collection of much larger object detection datasets in pursuit of better coverage. [9] conduct adversarial training on top of existing high performing models pushing performance even higher. Other approaches [61] attempt to incorporate scene graph prediction as part of their pre-training to learn more robust representations. Some recent work also attempts to build multi-purpose multi-modal architectures that are able to tackle a variety of vision-language [7] as well as pure language tasks in a single architecture [14]. A separate line of work that attacks a similar problem to ours but with a much more task specialized model architectures are the single [59, 4, 23] and two stage [39, 12] referring expression segmentation and phrase detection models which are designed specifically for this task.

5. Conclusion

We presented MDETR, a fully differentiable modulated detector. We established its strong performance on multi-modal understanding tasks on a variety of datasets, and demonstrated its potential in other downstream applications such as few-shot detection and visual question answering. We hope that this work opens up new opportunities to develop fully integrated multi-modal architectures, without relying on black-box object detectors.

Acknowledgements

We would like to thank Kyunghyun Cho, Ethan Perez, Sergey Zagoruyko and Francisco Massa for helpful discussions and feedback at various points of this project. We would also like to thank Alex Kirillov and Ross Girshick for their help with the LVIS evaluations, Justin Johnson for test set evaluation on CLEVR, Bryan Plummer for discussions on best evaluation practices for phrase grounding and finally Runtao Liu and Chenxi Liu for their feedback on dataset construction and evaluation for CLEVR referring expressions.

Aishwarya Kamath was supported in part by AFOSR award FA9550-19-1-0343 and Nicolas Carion by a grant from NVIDIA.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 4, 5, 8

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 1, 2, 6

[3] D. Chen, S. Jia, Y. Lo, H. Chen, and T. Liu. See-through-text grouping for referring image segmentation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7453–7462, 2019. 6

[4] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. See-through-text grouping for referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7454–7463, 2019. 8

[5] Wenhu Chen, Zhe Gan, Linjie Li, Yu Cheng, William Wang, and Jingjing Liu. Meta module network for compositional visual reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 655–664, 2021. 6

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. 2019. 1, 5, 8

[7] Jaemin Cho, Jie Lei, H. Tan, and M. Bansal. Unifying vision-and-language tasks via text generation. *ArXiv*, abs/2102.02779, 2021. 6, 8

[8] Achal Dave, Piotr Dollár, Deva Ramanan, Alexander Kirillov, and Ross Girshick. Evaluating Large-Vocabulary Object Detectors: The Devil is in the Details. 8

[9] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *arXiv preprint arXiv:2006.06195*, 2020. 5, 8

[10] Agrim Gupta, Piotr Dollár, and Ross Girshick. LVIS: A Dataset for Large Vocabulary Instance Segmentation. 7

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 4

[12] Ryota Hinami and Shin'ichi Satoh. Discriminative learning of open-vocabulary object retrieval and localization by negative phrase augmentation. *arXiv preprint arXiv:1711.09509*, 2017. 1, 8

[13] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to Reason: End-to-End Module Networks for Visual Question Answering. *arXiv:1704.05526 [cs]*, September 2017. arXiv: 1704.05526. 8

[14] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. 2021. 8

[15] Drew A. Hudson and Christopher D. Manning. Compositional attention networks for machine reasoning. *ArXiv*, abs/1803.03067, 2018. 4, 8

[16] Drew A. Hudson and Christopher D. Manning. Learning by abstraction: The neural state machine. In *NeurIPS*, 2019. 6

[17] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017. 8

[18] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Inferring and Executing Programs for Visual Reasoning. *arXiv:1705.03633 [cs]*, May 2017. arXiv: 1705.03633. 8

[19] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *arXiv preprint arXiv:1805.07932*, 2018. 5

[20] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. 4, 5

[21] Chenliang Li, Ming Yan, Haiyang Xu, Fuli Luo, Wei Wang, Bin Bi, and Songfang Huang. Semvlp: Vision-language pre-training by aligning semantics at multiple levels, 2021. 8

[22] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 5, 8

[23] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2018. 8

[24] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 1, 6, 8

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988. 6

[26] Tsung-Yi Lin, M. Maire, Serge J. Belongie, James Hays, P. Perona, D. Ramanan, Piotr Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 4, 5

[27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 4

[28] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019. 5, 8

[29] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020. 8

[30] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and Comprehension of Unambiguous Object Descriptions. *arXiv:1511.02283 [cs]*, April 2016. arXiv: 1511.02283. 5

[31] David Mascharka, Philip Tran, Ryan Soklaski, and Arjun Majumdar. Transparency by Design: Closing the Gap Between Performance and Interpretability in Visual Reasoning. pages 4942–4950. 8

[32] F. Milletari, N. Navab, and S. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571, 2016. 6

[33] Duy-Kien Nguyen, Vedanuj Goswami, and Xinlei Chen. MoVie: Revisiting Modulated Convolutions for Visual Counting and Beyond. *arXiv:2004.11883 [cs]*, October 2020. arXiv: 2004.11883. 6, 8

[34] A. Oord, Y. Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. 3

[35] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24:1143–1151, 2011. 5

[36] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual Reasoning with a General Conditioning Layer. 4, 8

[37] Bryan A Plummer, Paige Kordas, M Hadi Kiapour, Shuai Zheng, Robinson Piramuthu, and Svetlana Lazebnik. Conditional image-text embedding networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 249–264, 2018. 5

[38] Bryan A. Plummer, Arun Mallya, Christopher M. Cervantes, Julia Hockenmaier, and Svetlana Lazebnik. Phrase Localization and Visual Relationship Detection with Comprehensive Image-Language Cues. *arXiv:1611.06641 [cs]*, August 2017. arXiv: 1611.06641. 5

[39] Bryan Allen Plummer, Kevin Shih, Yichen Li, Ke Xu, Svetlana Lazebnik, Stan Sclaroff, and Kate Saenko. Revisiting image-language networks for open-ended phrase detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 5, 8

[40] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 4, 5

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 3, 6

[42] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019. 2

[43] Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. *arXiv:1706.01427 [cs]*, June 2017. arXiv: 1706.01427. 8

[44] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015. 3

[45] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. 5, 8

[46] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 5, 8

[47] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 6, 8

[48] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 7

[49] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. 4

[50] Mingxing Tan, Ruoming Pang, and Quoc V. Le. EfficientDet: Scalable and Efficient Object Detection. 4

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 2

[52] Zhonghao Wang, Mo Yu, Kai Wang, Jinjun Xiong, Wen-mei Hwu, Mark Hasegawa-Johnson, and Humphrey Shi. Interpretable Visual Reasoning via Induced Symbolic Space. 4, 8

[53] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 4

[54] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. Transformers: State-of-the-art natural language processing. In *EMNLP*, 2020. 4

[55] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. PhraseCut: Language-based Image Segmentation in the Wild. *arXiv:2008.01187 [cs]*, August 2020. arXiv: 2008.01187. 6

[56] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance

discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 3

[57] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with Noisy Student improves ImageNet classification. 4

[58] Jianwei Yang, Jiayuan Mao, Jiajun Wu, Devi Parikh, David D. Cox, Joshua B. Tenenbaum, and Chuang Gan. Object-Centric Diagnosis of Visual Reasoning. 2

[59] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4683–4693, 2019. 1, 5, 8

[60] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. 4, 8

[61] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 2020. 5, 8

[62] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. MAttNet: Modular Attention Network for Referring Expression Comprehension. *arXiv:1801.08186 [cs]*, March 2018. arXiv: 1801.08186 version: 3. 4, 5

[63] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling Context in Referring Expressions. *arXiv:1608.00272 [cs]*, August 2016. arXiv: 1608.00272 version: 3. 5

[64] Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. Rethinking diversified and discriminative proposal generation for visual grounding. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018. 5

[65] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Making visual representations matter in vision-language models. *arXiv preprint arXiv:2101.00529*, 2021. 1, 6, 8

[66] Yihan Zheng, Zhiquan Wen, Mingkui Tan, Runhao Zeng, Qi Chen, Yaowei Wang, and Qi Wu. Modular graph attention network for complex visual relational reasoning. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020. 4