Cops-Ref: A new Dataset and Task on Compositional Referring Expression Comprehension

Zhenfang Chen[1*] Peng Wang[2] Lin Ma[3] Kwan-Yee K. Wong[1] Qi Wu[4†]
[1]The University of Hong Kong [2]University of Wollongong
[3]Tencent AI Lab [4]Australian Centre for Robotic Vision, University of Adelaide
[1]{zfchen, kykwong}@cs.hku.hk [2]pengw@uow.edu.au
[3]forest.linma@gmail.com [4]qi.wu01@adelaide.edu.au

Figure 1: An example from the new Cops-Ref dataset for compositional referring expression comprehension. The task requires a model to identify a target object described by a compositional referring expression from a set of images including not only the target image but also some other images with varying distracting factors as well. The target/related/distracting regions are marked by green/yellow/blue boxes, respectively. More details about the reasoning tree can be seen in Sec. 3.1.

Abstract

Referring expression comprehension (REF) aims at identifying a particular object in a scene by a natural language expression. It requires joint reasoning over the textual and visual domains to solve the problem. Some popular referring expression datasets, however, fail to provide an ideal test bed for evaluating the reasoning ability of the models, mainly because 1) their expressions typically describe only some simple distinctive properties of the object and 2) their images contain limited distracting information. To bridge the gap, we propose a new dataset for visual reasoning in context of referring expression comprehension with two main features. First, we design a novel expression engine rendering various reasoning logics that can be flexibly combined with rich visual properties to generate expressions with varying compositionality. Second, to better exploit the full reasoning chain embodied in an expression, we propose a new test setting by adding additional distracting images containing objects sharing similar properties with the referent, thus minimising the success rate of reasoning-free cross-domain alignment. We evaluate several state-of-the-art REF models, but find none of them can achieve promising performance. A proposed modular hard mining strategy performs the best but still leaves substantial room for improvement. The dataset and code are available at: https://github.com/zfchenUnique/Cops-Ref.

1. Introduction

In recent years, computer vision tasks that require high-level reasoning have attracted substantial interest. Visual question answering (VQA) [14, 8] and visual dialog (VD) [5, 20] are typical examples of such a trend, where the system answers free-form questions based on an image by jointly reasoning over the textual and visual domains. A prerequisite to achieve this ultimate goal of artificial intelligence is the ability to ground the rich linguistic elements embodied in the language onto the visual content of the image. Referring expression comprehension (REF) is such a visual grounding task, which targets at identifying a particular object in a scene by an expression phased in natural language. A number of datasets [17, 29, 45] have been constructed for this task, and on top of which various mod-

*Work done while Zhenfang was visiting the University of Adelaide.
†Corresponding author.

els [27, 35, 44] have been developed.

Such popular datasets [17, 29, 45], however, cannot serve as ideal test beds to evaluate the reasoning ability of REF models. First, the expressions are typically simple and short, focusing mainly on some distinctive properties of the referent, such as object categories, attributes, or some simple relationships. For example, only some superficial reasoning is involved in expressions like '`the girl with glasses`' and '`the man sitting next to a table`'. Second, many images in the existing datasets contain only limited distracting information (*e.g.*, containing only two or three objects of the same category) and do not necessitate complex reasoning. For example, although we are given a complex expression '`The cat on the left that is sleeping and resting on the white towel.`' to localise the target cat in the example image shown in Fig. 1 (a), we can still have high chance to succeed even if we only use a simple expression '`The cat on the left`' as the query. Another non-negligible issue is dataset bias. As stated by Cirik *et al*. [4], a system that ignores the expression but uses only the image as input can still outperform random guess by a large margin. Recently, a synthetic dataset for referring expression, called CLEVR-Ref+ [26], is proposed to facilitate the diagnosis of visual reasoning. However, this dataset sacrifices visual realism and semantic richness by only describing some simple shapes and attributes.

To tackle the aforementioned problems, we propose a new challenging dataset for visual reasoning in context of referring expression comprehension. Our dataset is built on top of the real-world images in GQA [14] and therefore it pertains visual realism and semantic richness. The key novelty of our dataset lies in a new expression engine. We design six reasoning logics, namely, *and, or, order, same, not,* and *chain*, which can be flexibly combined with the rich visual information (*e.g.*, object categories, visual attributes, location information, and object interactions) to generate expressions with varying compositionality levels. Moreover, to overcome the sparse emergence of object categories and dataset bias, we design a new test setting by adding distracting images that contain objects sharing similar visual properties with the referent (*e.g.*, same object category and similar attributes). Along with the dataset, a new REF task named COmPoSitional Referring expression comprehension (Cops-Ref) is proposed, which requires a model to localise a region described by a flowery expression from a set of visually similar images. With the new dataset and task, the success rate of reasoning-free cross-domain alignment can be minimised.

We evaluate various state-of-the-art REF models on our proposed Cops-Ref dataset, but we find none of them can achieve a satisfactory performance. A modular hard-mining strategy is proposed to automatically mine hard negative examples embodying different visual properties. It achieves the best performance on the Cops-Ref task but still leaves much room for further improvement.

The contributions of this paper can be summarised as follows: 1) We introduce a new challenging task named Cops-Ref, which requires a model to localise the referent from a set of images with objects sharing similar visual properties; 2) We build a new dataset on top of real images, which pertains visual realism and semantic richness, and can complement the synthetic reasoning dataset to evaluate models' reasoning ability more rigorously; 3) We design a novel expression engine, which supports various reasoning logics that can be flexibly combined with rich visual stimuli to generate expressions with varying compositionality; 4) We conduct comprehensive evaluation on the REF models, among which the proposed modular hard mining strategy performs best but still leaves much room for improvement.

2. Related Work

Referring Expression Datasets. Toward tackling the REF task, many datasets [17, 29, 33, 45, 3] have been constructed by asking annotators to provide expressions describing regions of images. However, it is labor-intensive and hard to control the annotation quality, and most of the queries in the datasets can be easily solved by simply reasoning on object categories, attributes and shallow relations. Inspired by the synthetic dataset CLEVR [15] for VQA [43, 28], Liu *et al*. [26] built a synthetic REF dataset, called CLEVR-Ref+, by synthesising both images and expressions. However, it has been noticed in [14] that images in CLEVR, with only a handful of object classes, properties and spatial relations, are too simple for VQA. It is doubtful whether such synthetic images are representative enough to reflect the complexity of real-world images.

Recently, Hudson and Manning [14] proposed a new dataset GQA for VQA, which provides scene graph annotations for real-world images. By utilising the scene graph annotations and further data cleaning, we contribute a new dataset named Cops-Ref for referring expression, which contains not only region-expression pairs with complex reasoning chains but also visually similar distractors. It demands a much stronger reasoning ability to understand the whole expression and distinguish subtle visual differences in the images. Note that GQA also provides experiments localising related regions for questions but it is only regarded as a metric to evaluate the VQA task rather than targeting at the REF task. Neither expressions or distractors are considered in their setting.

Referring Expression Models. Referring expression [7, 12, 13, 19, 29, 30, 36, 40, 41, 42, 6] has attracted great attention. Karpathy and Fei-Fei [16] learned visual alignments between text and regions by multiple instance learning. Rohrbach *et al*. [35] localised a region by reconstructing the sentence using an attention mechanism. [45, 32, 46]

Index	Forms	Reasoning trees	Exemplar templates	Expression examples
1	chain	obj_0 (att_0) $\xrightarrow{\text{rel}_0}$ obj_1 (att_1) $\xrightarrow{\text{rel}_1}$ obj_2 (att_2)	The <att_0> <obj_0> that is <rel_0> the <att_1> <obj_1> that is <rel_1> the <att_2> <obj_2>.	The young girl that is touching the glazed donut that is on the round table.
2	and	obj_0 (att_0) $\begin{cases} \xrightarrow{\text{rel}_0} \text{obj}_1\ (\text{att}_1) \\ \xrightarrow{\text{rel}_1} \text{obj}_2\ (\text{att}_2) \end{cases}$	The <att_0> <obj_0> <rel_0> the <att_1> <obj_1> and <rel_1> the <att_2> <obj_2>.	The white fence near the building and behind the walking woman.
3	or	obj_0 (att_0) $\begin{cases} \xrightarrow{\text{rel}_0} \text{obj}_1\ (\text{att}_1) \\ \xrightarrow{\text{rel}_1} \text{obj}_2\ (\text{att}_2) \end{cases}$	The <att_0> <obj_0> <rel_0> the <att_1> <obj_1> or <rel_1> the <att_2> <obj_2>.	The green suitcase behind the black suitcase or near the yellow suitcase.
4	order	obj_0 (idx, dir, att_0)	The <idx> <obj_0> from the <dir> that is <att_0>.	The first glass from the left that is red.
5	same	$\text{obj}_0 \xrightarrow{\text{same cat}} \text{obj}_1$	The <obj_0> that has the same <cat> as the <obj_1>.	The bag that has the same color as the sweater.
6	not	obj_0 (not att_0)	The <obj_0> that is not <att_0>.	The apple that is not red.

Table 1: Examples of expression logic forms. Attributes of the objects are bounded with () and relations between objects are shown on $\rightarrow$. obj_0 denotes the target object, while $\text{obj}_{1,2}$ denote the related objects. $\text{att}_{0,1,2}$ and $\text{rel}_{0,1,2}$ denote the corresponding attributes and relations, respectively.

utilised context information to ground the expression. Yu *et al*. [44] and Liu *et al*. [25], respectively, used modular networks and neural module tree networks to match better structure semantics. Following [44], Wang *et al*. [39] and Liu *et al*. [27] increased the reasoning ability by watching neighbour regions and cross-modal attention-guided erasing. Different from these previous studies focusing on referring short expressions in a single image, we refer complex expressions in multiple similar images, which is more challenging and requires a stronger visual reasoning ability.

Text based Image Retrieval. Text based image retrieval returns relevant images from the gallery that is described by the text description [1, 9, 22, 23, 24, 33, 37, 38, 11, 2]. Different from text based image retrieval, Cops-Ref focuses on fine-grained region-level matching. The distracting regions in Cops-Ref are more semantically similar to the relevant region in the target image with only subtle differences. Such fine-grained and region-level similarity requires models with much stronger reasoning ability to ground the flowery expressions.

3. The Cops-Ref Dataset and Task

Previous natural image referring expression datasets [17, 29, 45] typically only require the ability to recognise objects, attributes and simple relations. Apart from such shallow ability, Cops-Ref also measures deeper reasoning ability like logic and relational inference. Compared with previous datasets, it has two main features, namely 1) flowery and compositional expressions which need complex reasoning ability to understand, and 2) a challenging test setting that includes controlled distractors with similar visual properties to the referent. Fig. 1 shows a typical example of our dataset. In the following subsections, we first introduce the construction of the dataset, including generating expressions (Sec. 3.1), discovering distractors (Sec. 3.2), and post-processing (Sec. 3.3). We then analyse the statistics of our dataset and formally define the task in Sec. 3.4 and Sec. 3.5.

3.1. The Expression Engine

The expression engine is the key to the construction of our dataset, responsible for generating grammatically correct, unambiguous and flowery expressions with various compositionality for each of the described regions. We propose to generate expressions from scene graphs based on some logic forms. Specifically, given a region to be described, we first choose a logic form from a predefined logic family and obtain a textual template for it. We then take the target object node in the scene graph as a root and expand it into a specific reasoning tree needed for the textual template. Finally, we fill the textual template with the content parsed from the reasoning tree and produce the expression. In the following paragraphs, we will describe the details of these three steps.

Expression logic forms. Expression logic forms summarise the abstract logics and provide specific structures for the expressions. Each of them is associated with several textual templates. Specifically, we define six types of expression logic forms, namely *chain*, *and*, *or*, *order*, *same*, and *not*. These high-level logic forms provide different contexts for the target object. Specifically, *chain*, *and* and *or* describe the relationship between the target object and other related objects. The *chain* form considers a sequence of related objects connected by a chain, while the *and* form indicates the target object must have some specific relations with two other objects and the *or* form only requires fulfilling one of the two relations. The *order* form provides relative spatial location between the target object and other objects of the same category. The *same* form shows that the target object shares the same attributes as the related object. The *not* form indicates a certain attribute or relation being absent in the target object. These basic logic forms can be further composed with each other and generate more complex and compositional expressions. The logic forms and their templates are shown in table 1.

Although these logic forms cannot fully reflect the com-

plexity of natural language, the basic logic units covered and their flexible compositions are sufficient to evaluate a model's reasoning ability. Moreover, experimental results show that knowledge learned from the Cops-Ref dataset can be directly applied to previous human-annotated datasets like refCOCO.

Reasoning tree parsing. While expression logic forms define the structures for the expressions, the dense scene graphs provide the corresponding semantic content. We use the scene graphs provided in [14, 21] to represent the internal semantic structures of images. Each node represents an object and edges between nodes represent the relations between them. Textual templates of different expression logic forms require different semantic content as input, which can be represented by reasoning trees with different structures extracted from the scene graph. Table 1 shows an instantiation of the reasoning tree for each of the forms, their corresponding textual templates and expression examples.

Specifically, for the *chain*, *and* and *or* forms, we simply parse the needed semantic reasoning trees from the scene graphs. For the *order* form, we sort all the regions that are of the same object category from left to right (vice versa) based on the coordinates of the centres of the bounding boxes. Since the order constraint is rather weak (*e.g.*, '`the left glass`' may also exist in the distracting images), we further add additional attributes and/or relations to make the expression unambiguous. Similarly, for the *not* form, we traverse the whole scene graph and collect the attributes/relations that present in all objects of the same category but not in the target object. For the *same* form, we find the attribute that only the target object and the related object have, and regard the category of that attribute as a relation between the two objects. The attribute categories used in the *same* form include colour, shape, material, gender and pattern.

Expression Generation. With the expression logic forms and the parsed reasoning trees ready, the expression engine can generate flexible expressions by filling the textual templates of the expression logic forms with the content from the reasoning tree. For example, given the *order* form and a textual template like `the <index> <object> from <direction>`, the expression engine can generate '`the first glass from the left`' for the reasoning tree, *glass (first, left)*. It can also generate more flowery expressions by adding more attributes or nodes to the reasoning tree. For example, it can produce '`the first glass from the left that is clear and to the left of the gravy`' by the expanded reasoning tree, *glass (first, left, clear)* $\xrightarrow{\textit{to the left of}}$ *gravy*.

3.2. Discovery of Distracting Images

Introducing distracting images in the test phase is another important feature of our proposed dataset. It provides more complex visual reasoning context and reduces dataset bias. The inclusion of distracting images guarantees that good performance can only be achieved by REF models that are able to reason over the complete expression and distinguish subtle visual differences. We define four types of distracting images, namely:

1. `DiffCat`: images that contain objects of different categories as the target object.
2. `Cat`: images that contain objects of the same category as the target object.
3. `Cat&attr`: images that contain objects of both the same category and attributes as the target object.
4. `Cat&cat`: images that contain all the objects in the reasoning tree but of different relations.

These distractors can be used to evaluate different aspects of REF models such as object recognition, attribute identification and relation extraction, etc. They force the models to fully understand the flowery and compositional expressions to achieve good performance. For each expression in the validation set and test set, we provide 3 distracting images under each distracting factor, apart from the image containing the ground-truth target. We simply discard those region-expression pairs for which we cannot find enough distractors. Fig. 1 shows an example of distracting images of different types for a given expression.

3.3. Post Processing and Balancing

We use synonyms parsed from wordNet [31] to further improve the diversity of the expressions. Besides, we remove expressions that target at classes that are hard to be bounded by a regular rectangle box (*e.g.*, '`sky`' and '`cloud`') and regions that are too small (*i.e.*, regions whose area is smaller than 1% of the whole image). We also notice that some of the scene graph annotations in GQA are incorrect or incomplete (*e.g.*, missing annotations for some objects/attributes/relations). They may make some regions in the distracting images also semantically match the expressions. To avoid such noise in the distractors, we manually check the expressions and images in the test set and discard these pairs with noise.

We also find some simple relations like '`to the left of`' being much more frequent than others in the scene graphs of GQA [14]. To address such bias issues, we adopt two strategies: 1) we sample relations for each node based on a probability that is directly proportional to the reciprocal of the frequency, downsampling most frequent relations and enriching diversity of expressions; 2) we abandon those expression-regions that only contain simple spatial relations.

3.4. Statistics of the Dataset

After the above processing and balancing, we have 148,712 expressions and 1,307,885 regions on 75,299 images, making our dataset the current largest real-world image dataset for referring expressions. The average length

(a) The most frequent object names (b) The most frequent attributes (c) The most frequent relations

Figure 2: The most frequent object names, attributes and relations of Cops-Ref. The size of words indicates frequency.

	Object Cat.	Att. Num.	Rel. Num.	Exp. length	Cand Num.	Cat Cand Num.
refCOCO	80[1]	-	-	3.5	10.6	4.9
refCOCOg	80	-	-	8.5	8.2	2.6
CLEVR-Ref+	3	12	5	22.4	-	-
Cops-Ref	508	601	299	14.4	262.5	20.3

Table 2: Statistic comparison of refCOCO [45], refCOCOg [29], ClEVR-Ref+ [26] and the Cops-Ref on number of object categories, number of attributes, number of relations, average length of expressions, average number of object candidates and average number of object candidates that are of the same categories for each expression.

of the expressions is 14.4 and the size of the vocabulary is 1,596. Since the scene graph annotation of the test set of GQA dataset is **not** publicly released, we use the validation set of GQA to construct our test set. A new validation set is split out from the training data of GQA to monitor the model training process. There are 119,603/16,524/12,586 expressions for training/validation/test, respectively.

Thanks to the dense annotations of the scene graphs, the proposed dataset contains fine-grained annotations for object categories, attributes and relations. The number of entry-level object categories, attributes and relations are 508, 601, and 299, respectively. We show the most frequent object names, attributes and relations in Fig. 2. We can see diverse object categories, with '`man`', '`building`' and '`tree`' being the most frequent object names. The most frequent attributes are colours (*e.g.* '`black`' and '`white`') and sizes (*e.g.* '`small`' and '`large`') while the most frequent relations are spatial relations like '`to the left/right of`'. We compare the statistics of the proposed Cops-Ref dataset with three widely-used referring dataset, refCOCO [45], refCOCOg [29] and CLEVR-Ref+ [26] in table 2[1]. As shown in table 2, the proposed dataset enjoys diverse object categories, attributes and relations. Moreover, it provides reasonably long expressions and much more candidate objects of same/different categories as the target object. The average length of our expressions is shorter than that of CLEVR-Ref+, but we find it is not necessary to use longer expressions to distinguish the target object in the real-world images even when distractors exist. More analysis about dataset bias and baseline results can be found in Sec. 5 and we provide more data examples and detailed statistics in the supplementary material.

[1]The definition of object categories between Cops-Ref and refCOCO are of different hierarchies. Cops-Ref does contain more diverse object categories like 'tree', 'shirt' and 'mat' which do not exist in refCOCO.

3.5. Task

Given a natural language referring expression and a set of similar images, the proposed Cops-Ref task requires a model to localise a target region described by the expression. Compared to the previous REF task [17, 29], the Cops-Ref demands a better understanding of longer and flowerier expressions, and the ability to distinguish the subtle differences of the distracting images. It requires REF models to have stronger reasoning ability for object detection, attribute recognition, and relation extraction. Formally, given N images and a query expression q, the Cops-Ref task identifies a target region r_{i^*,j^*} by

$$r_{i^*,j^*} = \underset{r_{i,j}, i\in[1,N], j\in[1,J_i]}{\arg\max} s(r_{i,j}|q), \tag{1}$$

where I_i denotes the i-th image, $r_{i,j}$ is the j-th region from I_i, J_i is the number of the regions in I_i, $s(r_{i,j}|q)$ denotes the matching score between $r_{i,j}$ and q.

Note that we do not use distracting images during training in our experimental setting because they are usually unavailable or hard-to-collect in the real world. Also, it is easier for us to follow the original training strategies in [22, 27, 44, 35] to re-train and evaluate the models.

4. Model

Although Cops-Ref is a new task that requires localising a region from a set of images instead of a single one, existing REF models can be directly applied to this new task by densely matching the query expression with each object in the image set and choosing the one with the highest matching score as the referring result.

MattNet [44] is a popular backbone model for solving the REF task because of its extraordinary capability in modeling different modules of the query expressions, including subject (sub), location (loc) and relationship (rel). Specifically, MattNet estimates the matching score between an expression q and the j-th region r_j by

$$s(r_j|q) = \sum_{md} w^{md} s(r_j|q^{md}), \tag{2}$$

where $md \in \{sub, loc, rel\}$, w^{md} is the learnt weight for the md module and q^{md} is the modular phrase embedding. More details about MattNet can be found in [44].

Given a positive pair (r_m, q_m), the whole model of MattNet is optimised by a ranking loss, given by

$$\begin{aligned}\mathcal{L}_{rank} = &\sum_m([\Delta - s(r_m|q_m) + s(r_m|q_n)]_+ + \\ &[\Delta - s(r_m|q_m) + s(r_o|q_m)]_+),\end{aligned} \tag{3}$$

where r_o and q_n are other random unaligned regions and expressions in the same image as r_m and q_m, Δ is a margin and $[x]_+ = max(x, 0)$. This loss function is suitable for the REF task and can successfully distinguish aligned region-expression pairs from unaligned ones within the same image. However, when it comes to the Cops-Ref task, it has the limitation on not being able to identify hard negative examples with similar visual properties in other images, because the training of MattNet does not consider hard negative regions and expressions in other images. To solve this problem, we propose a modular hard-mining training strategy based on MattNet.

Modular Hard-mining Strategy. To increase the ability of MattNet to distinguish hard negative regions in distracting images, we need to sample distracting regions/expressions in other images as negative training examples. However, since there are 119,603 expressions and 797,595 regions in the training set of Cops-Ref, how to mine hard negative regions and expressions effectively and efficiently becomes a challenge. To handle this challenge, we propose to use the similarity of modular phrase embedding q^{md} as a prior to sample the hard negative examples in other images, where $md \in \{sub, loc, rel\}$.

Specifically, for the m-th region-expression pair, we first extract its modular expression features $\{q_m^{md}\}$ and calculate their similarity with those of the n-th region-expression pair that has the same object category. We define the probability of sampling the n-th region-expression pair to be the negative pair by

$$\begin{aligned}s_{m,n}^{md} &= f(q_m^{md}, q_n^{md}), \\ p_{m,n}^{md} &= \frac{exp(s_{m,n}^{md})}{\sum_{n=1,m\neq n}^{n=N_C} exp(s_{m,n}^{md})},\end{aligned} \tag{4}$$

where f is a function for estimating the similarity between two expression features and N_C is the number of the region-expression pairs has the same object category as the m-th region-expression pair in the training set. For simplicity, We use cosine similarity as an instantiation of f. We mine hard distracting regions and expressions for each positive region-expression pair and send these distracting regions to a ranking loss as hard negative examples.

Formally, our modular hard-mining loss is

$$\begin{aligned}\mathcal{L}_{mine} = &\sum_m \sum_{md}([\Delta - s(r_m|q_m) + s(r_m|q_n^{md})]_+ + \\ &[\Delta - s(r_m|q_m) + s(r_n^{md}|q_m)]_+),\end{aligned} \tag{5}$$

where r_n^{md} and q_n^{md} are a region-expression pair sampled with $\{p_{m,n}^{md}\}_{n=1,m\neq n}^{N_C}$ as a prior.

Our total loss is $\mathcal{L} = \mathcal{L}_{rank} + \mathcal{L}_{mine}$, where $\mathcal{L}_{rank}$ targets at distinguishing distracting negative regions and expressions within the same image, and $\mathcal{L}_{mine}$ targets at distinguishing similar negative regions and expressions in other images.

Such a modular hard mining strategy is effective since it can mine hard negative region-expression pairs outside the image that contains the target region-expression pair. Besides, the mined regions have similar properties as the target, which demand stronger reasoning ability to distinguish. It is also efficient since it only requires the expressions as input without the need for loading images into memory. It enables the model to scan all the expressions in the training set in around 29 seconds with a naïve implementation. During training, we update the sample probability $p_{i,j}^{md}$ every 50 iterations. We distinguish the proposed hard mining model from the original MattNet by calling it **MattNet-Mine**.

5. Experiments

In this section, we conduct extensive experiments to analyse the Cops-Ref dataset and compare our proposed model with SOTA REF models. We first study the bias impact and transfer performance. We then compare the performance of the proposed MattNet-Mine with the baselines. We also provide extensive analysis, including "retrieve" + "REF" to handle the task, performance against logic forms and lengths of the expressions. We finally provide an ablation study on our mining strategy for distractors. We introduce experimental settings before we start.

5.1. Experimental Settings

Implementation Details. Following MattNet [44] and CM-Att-Erase [27], we extract visual features by res101-based Faster-RCNN [10, 34] pre-trained on COCO [24]. For each word in the sentence, we initialise it with an one-hot word embedding. We train all the models with Adam optimiser [18] until the accuracy of the validation set stops improving. We set the maximum time step for the text encoder to be 30. Expressions with words less than 30 are padded. For other settings for hyper-parameters, we keep them the same as the original MattNet to avoid cumbersome parameter fine-tuning. For the proposed MattNet-Mine, we first pre-train it by the ranking loss $\mathcal{L}_{rank}$ to obtain reasonable modular attentive phrase embeddings and then finetune the model with both $\mathcal{L}_{mine}$ and $\mathcal{L}_{rank}$. Following previous REF models like [27, 39, 44], we use ground-truth object bounding boxes as proposals. We consider it as a correct comprehension if the model successfully chooses the proposal pointed by the expression among all the proposals extracted from the similar image set.

Evaluation Settings. Table 3 shows different experiments settings. `Full` denotes the case when all the distractors are added while `WithoutDist` denotes no distractor is added. `DiffCat`, `Cat` and `Cat&attr`, respectively, represent

Method	Full	DiffCat	Cat	Cat&attr	Cat&cat	WithoutDist
Chance	0.4	1.7	1.8	1.9	1.7	6.6
GroundeR [35]	19.1	60.2	38.5	35.7	38.9	75.7
Deaf-GroundeR	2.2	7.7	7.9	8.0	8.0	27.1
Shuffle-GroundeR	13.1	41.8	28.6	27.2	27.6	58.5
Obj-Attr-GroundeR	15.2	53.1	32.6	29.6	32.7	68.8
MattNet-refCOCO	8.7	22.7	17.0	16.7	18.9	42.4
MattNet [44]	26.3	69.1	45.2	42.5	45.8	77.9
CM-Att-Erase [27]	28.0	**71.3**	47.1	43.4	48.4	**80.4**
SCAN [22]+MattNet	18.8	-	-	-	-	-
MattNet-Mine	**33.8**	70.5	**54.4**	**46.8**	**52.0**	78.4

Table 3: Results of baselines and state-of-the-art models on the Cops-Ref dataset. MattNet-refCOCO is trained on refCOCO.

the cases when certain type of distractors are added, including distracting images containing no object of the same category as the target object, images containing objects of the same category, images containing objects of the same category, and attributes and images containing all the objects in the reasoning tree but of different relations.

5.2. Dataset Analysis

Bias Analysis. Inspired by the bias analysis of Cirik *et al*. [4], we use similar ways to analyse the bias problem of Cops-Ref. To exclude the impact of specific models or mechanisms, we choose GroundeR [35], which is the simplest CNN-RNN baseline model for referring expression. We train several variants of GroundeR models, include deaf-GroundeR which masks out the language input of the GroundeR with an all-zero vector, shuffle-GroundeR which shuffles the order of the word sequence in the expression and Obj-Att-GroundeR which only keeps the nouns and adjectives of the text input.

The upper section of table 3 shows the results of the bias experiments. Deaf-GroundeR, an image only model achieves better performance than the "Chance" model, which selects a region from the images by chance. We observe that Deaf-GroundeR can filter out some irrelevant regions by providing higher matching scores for those regions whose categories like 'woman' and 'shirt' frequently appear in both the training set and test set. This indicates that the statistics bias problem in previous datasets like refCOCOg [29] also exists in our dataset. However, comparing the results of `WithoutDist` and `Full`, we see that the biased performance becomes much lower when distractors are added. Moreover, the bias problem in Cops-Ref is less significant than in refCOCOg. Deaf-GroundeR only achieves an accuracy of 2.2 in the `Full` case, while a similar "image only" model achieves an accuracy of 40.1 in [4].

Cirik *et al*. [4] also pointed out that shuffling the order of expressions and masking out other words that are not nouns or adjectives have minor effect on the performance of refCOCOg, resulting in only a relative drop of 4% and 3%, respectively. This suggests that a model does not need very strong reasoning ability for the whole sentence to handle the task. However, in Cops-Ref, comparing Shuffle-GroundeR and Obj-Att-GroundeR with GroundeR under the `Full` case, we observe a relative drop of 31% and 20%, respectively. It indicates that the syntactic structure and relations play more significant roles in Cops-Ref regarding performance improvement.

Transfer Performance. We directly apply a MattNet [44] model trained on refCOCO to our Cops-Ref and it only achieves an accuracy of 42.4 and 8.7 under the `WithoutDist` and `Full` cases, respectively. It shows our dataset and task are more complex and challenging. In contrast, a MattNet trained on Cops-Ref can achieve an accuracy of 56.5 and 64.5 on the testA and testB splits of refCOCO, which are about 65.7% and 76.4% of the performance of the original model trained on refCOCO, respectively. This demonstrates the realism of our synthetic expressions and that the knowledge learnt from Cops-Ref can be directly transferred to real datasets like refCOCO, while the reasoning ability gained from refCOCO cannot solve our Cops-Ref task.

5.3. Overall Evaluation

We evaluate the proposed Cops-Ref task with three baselines, namely GroundeR [35], MattNet [44] and CM-Att-Erase [27]. GroundeR is a simple CNN-RNN baseline. MattNet is one of the most popular REF models, and CM-Att-Erase was the best state-of-art in REF at the time of this submission. We densely ground the expressions on every image in the similar image set and choose the region with the highest score as the grounding result.

Performance of the REF Models. Table 3 reports the accuracy of all the baselines and the proposed MattNet-Mine. We have the following observations. (1) The performance gradually increases from GroundeR [35] to MattNet [44] and from MattNet [44] to CM-Att-Erase [27]. This is consistent with their performance on the refCOCO, refCOCO+ and refCOCOg [17, 29]. (2) The performance of these REF models decreases dramatically when distracting images containing the objects of the same object category are added, especially under the `Full` case. Among the 4 types of distractors, `DiffCat` affects the performance least while `Cat&Attr` affects most. This implies that existing

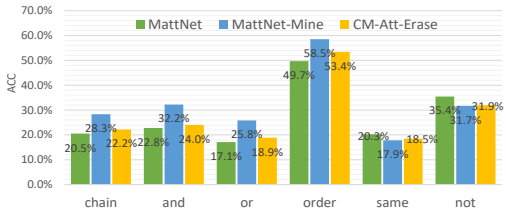

Figure 3: Accuracy of expressions of different logic form.

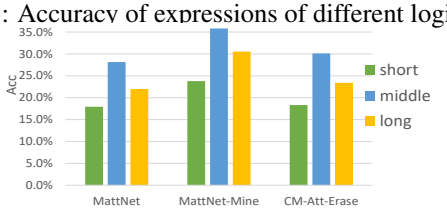

Figure 4: Accuracy of expressions of different lengths.

REF models strongly rely on object and attribute recognition to localise the target region. (3) Comparing with the original MattNet [44], our MattNet-Mine show improved performance under all cases, especially for the cases that contain fine-grained similar distractors. This demonstrates the effectiveness of the proposed hard mining strategy.

Performance of "Retrieve" + "REF" Strategy. We also evaluate another strategy to solve the problem in which we first use a text-based image retrieval model to select one image with the highest matching score and then ground the query expression in the selected image. We use SCAN (t2i) [22] as the retrieval model for its excellent performance, and we use MattNet to ground the expression in the returned image. We achieve an accuracy of 18.8 under the `Full` case. Compared with the other models in table 3, the "Retrieve"+"REF" strategy performs worse than densely referring the query expression in every image. We believe this may be caused by the fact that densely referring an expression in every image provides more fine-grained regional level matching than the retrieval model.

Performance of Different Logic Forms. We show the performance of expressions of each logic form in Fig. 3. We can see that while expressions of *chain*, *and*, *or* and *same* forms have similar accuracy, *order* and *not* forms have the best and second best accuracy, respectively. We believe the reasons are 1) the reasoning logic trees of *order* and *not* forms are simpler than other forms like *chain*, *and* and *or* (see table 1), and 2) *order* form has provided specific relative spatial location between the target object and the related objects of the same category within the same image.

Performance of Different Lengths of Expressions. We divide expressions into 3 kinds based on the number of the words in the expressions, namely short (less than 10 words), middle (10-20 words) and long (more than 20 words), and test them separately. As shown in Fig. 4, we find that expressions in the middle group have the best accuracy. We suspect that short expressions provide limited textual information for distinguishing distracting regions while long expressions usually contain complex logic or semantics that requires stronger reasoning ability.

Method	Full	DiffCat	Cat	Cat&attr	Cat&cat
MattNet	26.3	69.1	45.2	42.5	45.8
Random	27.6	**71.6**	47.4	43.5	47.3
Class-aware	32.2	70.3	53.2	46.1	51.4
Sentence-sim	32.3	70.4	53.6	46.4	51.2
Module-specific	**33.8**	70.5	**54.4**	**46.8**	**52.0**

Table 4: Ablation study of different hard mining strategies.

5.4. Ablation Study on Distractor Mining

We conduct an ablation study to investigate different hard negative mining strategies for the Cops-Ref task. Specifically, we have the following solutions by replacing the q_n^{md} and r_n^{md} in Eq. 5 with features from different regions and expressions. "Random" means using regions and expressions that are randomly-selected from the whole dataset regardless the object category. "Class-aware" means using random-selected regions and expressions that has the same object category as the target region. "Sentence-sim" means a region-expression pair that is sampled based on the similarity of global expression features. We define the global expression features as the average embedding of all the words in the expression. "Module-Specific" means the proposed modular specific hard mining strategy based on the similarity of the modular expression features.

Table 4 shows the ablation study results. Compared with the original MattNet, "Random" can provide an improvement under all cases. However, its improvement for the `Full` case is minor comparing with other mining strategy since it does not consider the similar distractors. "Class-awre" boosts the performance under the case where similar distractors are added, indicating the value of the distracting regions and expressions of the same category. "Sentence-sim" achieves only a comparable performance with "Class-aware", showing its inefficiency for hard negative mining. "Module-specific" achieves the best performance when similar distractors are added, showing its effectiveness to mine negative examples and distinguish similar distractors.

6. Conclusion

Expressions in existing referring expression datasets normally describe some simple distinguishing properties of the object which cannot fully evaluate a model's visual reasoning ability. In this paper, we proposed a new challenging dataset, named Cops-Ref, for referring expression comprehension. The new dataset covers various reasoning logics that can be flexibly combined with rich visual properties. Additionally, to better exploit the full reasoning chain embodied in the expression, we proposed a new test setting by adding some additional distracting images. This newly proposed dataset suffers less bias and we found existing state-of-the-art models fail to show promising results. We then proposed a modular based hard mining strategy that achieves the best performance but is still far from perfect. We wish this Cops-Ref dataset and task can attract more research attention and become a new benchmark in this area.

References

[1] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M Blei, and Michael I Jordan. Matching words and pictures. *Journal of machine learning research*, 2003. 3

[2] Zhenfang Chen, Lin Ma, Wenhan Luo, Peng Tang, and Kwan-Yee K Wong. Look closer to ground better: Weakly-supervised temporal grounding of sentence in video. *arXiv preprint arXiv:2001.09308*, 2020. 3

[3] Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee K Wong. Weakly-supervised spatio-temporally grounding natural sentence in video. In *Proc. 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 2019. 2

[4] Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. Visual referring expression recognition: What do systems actually learn? In *NAACL*, 2018. 2, 7

[5] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. In *CVPR*, 2017. 1

[6] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. Visual grounding via accumulated attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7746–7755, 2018. 2

[7] Pelin Dogan, Leonid Sigal, and Markus Gross. Neural sequential phrase grounding (seqground). In *CVPR*, 2019. 2

[8] Chuang Gan, Yandong Li, Haoxiang Li, Chen Sun, and Boqing Gong. Vqs: Linking segmentations to questions and answers for supervised attention in vqa and question-focused semantic segmentation. In *ICCV*, 2017. 1

[9] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision*, 2014. 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[11] Hexiang Hu, Ishan Misra, and Laurens van der Maaten. Binary Image Selection (BISON): Interpretable Evaluation of Visual Grounding. *arXiv preprint arXiv:1901.06595*, 2019. 3

[12] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *CVPR*, 2017. 2

[13] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *CVPR*, 2016. 2

[14] Drew A Hudson and Christopher D Manning. Gqa: a new dataset for compositional question answering over real-world images. In *CVPR*, 2019. 1, 2, 4

[15] 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 *CVPR*, 2017. 2

[16] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 2

[17] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 1, 2, 3, 5, 7

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[19] Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. What are you talking about? text-to-image coreference. In *CVPR*, 2014. 2

[20] Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Visual coreference resolution in visual dialog using neural module networks. In *ECCV*, 2018. 1

[21] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 2017. 4

[22] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, pages 201–216, 2018. 3, 5, 7, 8

[23] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. 2019. 3

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3, 6

[25] Daqing Liu, Hanwang Zhang, Zheng-Jun Zha, and Wu Feng. Learning to assemble neural module tree networks for visual grounding. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[26] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *CVPR*, 2019. 2, 5

[27] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. Improving referring expression grounding with cross-modal attention-guided erasing. In *CVPR*, 2019. 2, 3, 5, 6, 7

[28] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. In *ICLR*, 2019. 2

[29] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 1, 2, 3, 5, 7

[30] Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. A joint model of language and perception for grounded attribute learning. In *ICML*, 2012. 2

[31] George A. Miller. Wordnet: A lexical database for english. *COMMUNICATIONS OF THE ACM*, 38:39–41, 1995. 4

[32] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *ECCV*, 2016. 2

[33] 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 *ICCV*, 2015. 2, 3

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 6

[35] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, 2016. 2, 5, 7

[36] Mohammad Amin Sadeghi and Ali Farhadi. Recognition using visual phrases. In *CVPR*, 2011. 2

[37] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2014. 3

[38] Mingzhe Wang, Mahmoud Azab, Noriyuki Kojima, Rada Mihalcea, and Jia Deng. Structured matching for phrase localization. In *ECCV*, 2016. 3

[39] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *CVPR*, 2019. 3, 6

[40] Sibei Yang, Guanbin Li, and Yizhou Yu. Cross-modal relationship inference for grounding referring expressions. In *CVPR*, 2019. 2

[41] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *ICCV*, 2019. 2

[42] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *CVPR*, June 2019. 2

[43] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B Tenenbaum. Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. In *NIPS*, 2018. 2

[44] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, 2018. 2, 3, 5, 6, 7, 8

[45] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 1, 2, 3, 5

[46] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding referring expressions in images by variational context. In *CVPR*, 2018. 2