Transferrable Referring Expression Grounding with Concept Transfer and Context Inheritance

Xuejing Liu[1,2], Liang Li[1,*], Shuhui Wang[1], Zheng-Jun Zha[3], Dechao Meng[1,2], Qingming Huang[2,1]
[1]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, CAS, Beijing, China
[2]School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, China
[3]School of Information Science and Technology, University of Science and Technology of China, Hefei, China
{xuejing.liu,liang.li,dechao.meng}@vipl.ict.ac.cn,wangshuhui@ict.ac.cn,zhazj@ustc.edu.cn,{qmhuang}@ucas.ac.cn

ABSTRACT

Referring Expression Grounding (REG) aims at localizing a particular object in an image according to a language expression. Recent REG methods have achieved promising performance, but most of them are constrained to limited object categories due to the scale of current REG datasets. In this paper, we explore REG in a new scenario, where the REG model can ground novel objects out of REG training data. With this motivation, we propose a Concept-Context Disentangled network (CCD) which transfers concepts from auxiliary classification data with new categories meanwhile inherits context from REG data to ground new objects. Specially, we design a subject encoder to learn a cross-modal common semantic space, which can bridge the semantic and domain gap between auxiliary classification data and REG data. This common space guarantees CCD can transfer and recognize novel categories. Further, we learn the correspondence between image proposal and referring expression upon location and relationship. Benefiting from the disentangled structure, the context is relatively independent of the subject, so it can be better inherited from the REG training data. Finally, a language attention is learned to adaptively assign different importance to subject and context for grounding target objects. Experiments on four REG datasets show our method outperforms the compared approach on the new-category test datasets.

CCS CONCEPTS

• **Software and its engineering → Visual languages**.

KEYWORDS

Transferrable REG, Concept Transfer, Context Inheritance

ACM Reference Format:
Xuejing Liu[1,2], Liang Li[1,*], Shuhui Wang[1], Zheng-Jun Zha[3], Dechao Meng[1,2], Qingming Huang[2,1]. 2020. Transferrable Referring Expression Grounding with Concept Transfer and Context Inheritance. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413677

*Corresponding author.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413677

1 INTRODUCTION

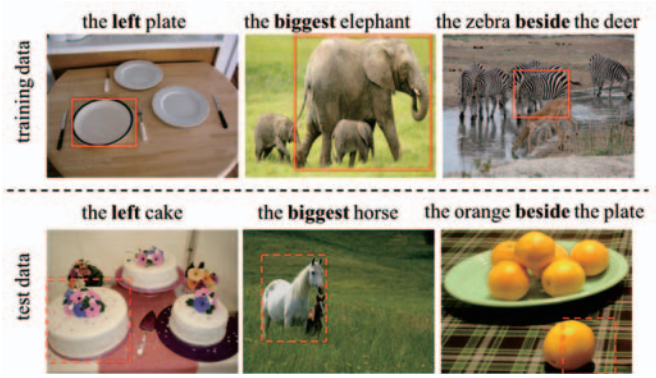

Figure 1: Some training and test examples in transferrable referring expression grounding (REG) for new categories. In testing data, the categories (e.g. cake, horse, orange) are out of REG training data while the dominant contexts (e.g. left, biggest, beside) exist in the REG training data.

Referring Expression Grounding (REG), also known as referring expression comprehension, is to ground a target object in an image according to linguistic expression. As a classical visual-language task, REG are attracting much attention in recent years [21–23, 32]. It is a fundamental problem for many visual-language applications, such as VQA [13, 29], interactive QA [5], and visual navigation [10, 14], etc.

However, most methods are constrained to limited categories due to the scale of current REG datasets [15, 23]. This constraint severely hinders the application of REG in the real world. Moreover, labeling proposals and referring expressions of new categories for REG is very expensive. Thus we introduce the **transferrable REG**, which aims at grounding the new objects out of the REG training data. In the training data of our settings, we have the grounding region of referring expression for very limited categories. In the testing data, the referring expressions are about new categories, but their dominant contexts appear in the training data. Some examples of training and testing data are shown in Fig. 1.

There are two main challenges in transferrable REG dealing with new categories. The first one is to *recognize new objects*. This needs to construct the correspondence between their visual appearances and categories, but such correspondence cannot be acquired in REG training data. We notice that the visual appearances of thousands of categories are learned in the image classification task [3, 17].

Unfortunately, they cannot be directly used in REG due to the domain gap. In this paper, we aim to transfer this correspondence in classification data to endow the REG model with the ability of recognizing new categories.

The second challenge is to *perceive category-independent context*. REG needs to understand context to distinguish the referent from other objects, usually of the same category. Traditional REG methods [11, 12, 24, 30, 31, 33, 34] model the category-dependent context of the referent by exploiting the relationship between the referent and its related objects under supervision. In transferrable REG for new categories, we need to learn the category-independent context (i.e., "on the right of") that does not rely on the representation of the target object. Because of lacking the ground truth for grounding new objects in the training data, the context can only be learned from existing REG training data. In our work, we attempt to model category-independent context to assist grounding a referred new object.

To address the above problems, we design the Concept-Context Disentangled network (CCD) which enables *concept transfer* from auxiliary classification data (images from new categories) and *context inheritance* from REG data to ground new objects. CCD is a modular network including subject representation and context modeling, which can disentangle the category-independent context for grounding.

For the subject representation, we introduce a subject encoder to learn the cross-modal common semantic space, which can bridge the semantic and domain gap between the auxiliary classification data and REG data to recognize new objects. To bridge the domain gap, we encode the categories of auxiliary data and the referring expressions from REG data into the same language space with Glove [25], and encode the images from both datasets into the same visual space with pre-trained Faster R-CNN. To bridge the semantic gap, we use the subject encoder to learn the correspondence between the visual and language space.

For the context modeling, we learn the mapping score between visual context feature and attentive language context feature upon location and relationship with other objects. The category-independent context can be inherited to perceive the context of new objects. Finally, a language attention is learned to adaptively assign different importance to subject and context. Jointly with the recognition of new objects, our CCD can perceive category-independent context to ground referred new object, especially when multiple objects of the same category situate together.

At the training stage, we introduce a multi-task learning mechanism in CCD which jointly trains the classification and REG models in an end-to-end manner. We evaluate our network under two scenarios, i.e., intra-dataset and inter-dataset, on four REG datasets: RefCOCO, RefCOCO+, RefCOCOg, and RefClef. Extensive experiments show that our method can outperform strong competitors on the new-category test set.

The contribution of this paper are summarized as follows:

(1) We introduce the transferrable referring expression grounding dealing with grouding new categories. It is more practical and more challenging in recognizing new objects and perceiving category-independent context.
(2) We propose a concept-context disentangled network to transfer new concept and inherit category-independent context for the transferrable REG for new categories.
(3) We evaluate our CCD on four datasets under two kinds of scenarios: intra-dataset and inter-dataset. Experimental results show CCD can get better performance in grounding new objects.

2 RELATED WORK

2.1 Referring Expression Grounding

To distinguish a particular object in an image, a referring expression should be unambiguous. Thus the expression usually consists of target category and context (e.g., attributes, location, and relation of the object [23]). How to make the best use of these cues becomes the key challenge for referring expression grounding.

Yu *et al.* [33] utilized visual comparison to other objects within an image to improve the performance. Nagaraja *et al.* [24] learned context through multiple instance learning to understand referring expressions. Hu *et al.* [12] presented a modular network identifying entities and relationships to analyze referential expressions. Liu *et al.* [20] found attribute learning significantly improved the performance of referring expression generation and comprehension. Zhang *et al.* [34] proposed a variational Bayesian method to model complex context in referring expression grounding. Yu *et al.* [32] designed a modular attention network to dynamically learn the scores upon three modules: subject, location and relationship. Yang *et al.* [30] introduced a language-guided visual relation graph to compute multi-modal semantic contexts. The above studies show the importance of context for REG. However, they modeled the category-dependent context of the referent by exploiting relationship between the referent and its related objects under supervision. In our settings, for lacking the ground truth on novel objects, we attempt to perceive the category independent context from REG training data to ground novel objects.

2.2 Novel Objects in Visual-Language Task

The novel object problem, i.e., categories not contained in the training data, has attracted much attention in recent years [26]. Many related works on visual and language tasks arise these days, such as object retrieval, image caption, language grounding, which enhance the generalization of model working in practical applications.

Guadarrama *et al.* [7] combined category and instance representation to handle the problem of open-vocabulary object retrieval. Anderson *et al.* [2] used constrained beam search to force the inclusion of open-vocabulary words in image caption. Agrawal *et al.* [1] presented the first large-scale benchmark to promote the open-vocabulary image caption. Li *et al.* [18] expanded the vocabulary via pointing mechanism for image caption. Hinami and Satoh [9] proposed Query-Adaptive R-CNN to address the open-vocabulary object retrieval and localization. Most related to our work, Sadhu *et al.* [28] grounded the novel, "unseen" nouns using a single-stage model which combined the detector network and grounding system. Their work focused on grounding the queries in which category is discriminative enough. In comparison, our work focuses on understanding the contexts and distinguishing the novel objects from

other same-category objects, which is also the key challenge of REG task.

3 METHOD

REG aims to ground a particular object described by the linguistic query in an image where multiple objects of the same category situated together. This problem can be formulated as follows. Given an image I, a linguistic query q, and a set of region proposals $\{r_i\}_{i=1}^{N}$, we aim at selecting the best-matched region r^* according to the query.

In our introduced transferrable REG, we aim at grounding the novel objects out of REG training data. To address this problem, we design a Concept-Context Disentangled network (CCD) which can transfer concept as well as inherit context in the REG model. We adopt the modular design of MAttNet [32] and make some changes as the backbone.

3.1 Overall Network Structure

The overall network structure of our proposed CCD is shown in Fig. 2. The model is disentangled into two modules: subject representation and context modeling. In the subject representation, we use a multi-task learning mechanism by jointly training classification and REG tasks. We learn a cross-modal common semantic space to bridge the domain and semantic gap between the classification and REG datasets. Specifically, we encode the images from the two datasets into the same visual space with pre-trained Faster R-CNN and encode the categories from classification data and the referring expressions from REG data into the same language space with Glove. With these operations, we respectively bridge the visual and language domain gap between the two datasets. Then, we design a subject encoder to learn the mapping between the visual space and the language space to bridge the semantic gap of two modalities. The cross-modal common semantic space enables our CCD to transfer the concepts from classification data to the REG model.

In the context modeling, we learn the correspondence between context visual features and attentive context language features. Benefiting from the concept-context disentangled structure, the CCD can perceive category-independent context. So we can inherit the learned context from REG training data when coming across novel objects. Finally, a language attention is learned to adaptively apply different importance to the two modules. It also denotes the ratio between the transferred concept and inherited context.

3.2 Concept Transfer

In this subsection, we introduce the concept transfer from classification data to REG data in subject representation. The detailed structure is shown in Fig. 3.

3.2.1 Language Feature Encoding. To bridge the language domain gap, we can encode the categories from classification data and referring expressions from REG data into the same language space. For categories in classification data, we directly encode them into word embeding c_t with Glove [25]. Golve is a kind of global vector for word representation. It has been widely used in visual and language tasks to improve the understanding for language. If the category has multiple words $\{w_n\}_{n=1}^{N}$, we use the average of word embedings as the final representation.

$$c_t = \text{Mean}\,(w_n)_{n=1}^{N} \tag{1}$$

For referring expressions of REG data, we first encode each word in query $q = \{w_t\}_{t=1}^{T}$ into word embedding e_t with Glove. Then the word embedding e_t is fed into an MLP (multi-layer perceptions) and a bi-directional LSTM. We concatenate the hidden vectors in both directions as the final representation $h_t = [\overrightarrow{h}_t, \overleftarrow{h}_t]$. Based on the representation h_t, we calculate attention score for each word in the query with attention mechanism, and this helps learn better subject representations. The final language representation for subject is calculated as follows:

$$q_s = \sum_t \text{softmax}_t\,(\text{fc}\,(h_t))\,e_t\,. \tag{2}$$

3.2.2 Subject Encoder. First, to bridge the visual domain gap, we encode the images from classification data and REG data into the same visual space. Pre-trained Faster R-CNN [27] based on ResNet [8] is used as the backbone network to extract the visual features for the images from both the classification and REG datasets. Second, to bridge the semantic gap, we design the subject encoder to construct the correspondence between the language space and visual space. To get better visual representation for the subject, the above extracted features are first fed into the fully-connected layer with ReLU.

$$\widetilde{r_i^s} = \text{ReLU}\left(fc\left(r_i^s\right)\right) \tag{3}$$

r_i^s denotes the visual features of a proposal. Then two branches are introduced to learn the attributes and transfer novel concepts respectively.

This attribute classification branch is to better preserve the attribute information in the visual features of the subject. Only the REG data go through this branch. The attribute label is extracted through an external language parser [15] according to [32]. We use the binary cross-entropy loss for the multi-label attribute classification:

$$Loss_{att} = f_{\text{BCE}}\left(y_i^{att}, fc\left(r_i^s\right)\right), \tag{4}$$

where y_i^{att} is the attribute label.

The category classification branch is to transfer novel categories from classification data to our model. Both classification data and REG data go through this branch. $\widetilde{r_i^s}$ is fed into an MLP followed by a fully-connected layer.

$$\overline{r_i^s} = fc\left(\text{MLP}\left(\widetilde{\text{r}_\text{i}^\text{s}}\right)\right) \tag{5}$$

Finally, subject matching functions as the loss to construct the cross-modal correspondence between language and visual space. This guarantees that the concept can be transferred from classification data to REG model. In detail, we use Mean Squared Error (MSE) criterion to minimize the distance between the visual features and category embeddings.

$$Loss_{cat} = \text{MSE}(c_t, \overline{r_i^s}), \tag{6}$$

where c_t is the category embedding.

For the REG data, we calculate the cosine distance between the language representation q_s and visual representation $\overline{r_i^s}$ as the

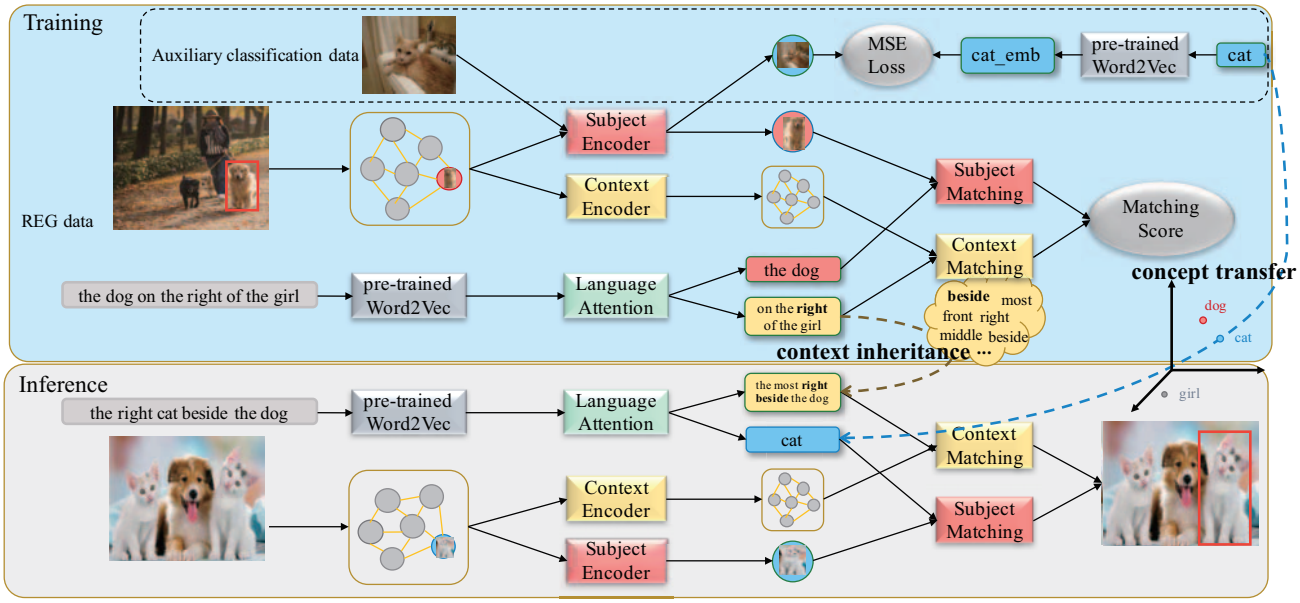

Figure 2: The network structure of our CCD. In the training stage, we jointly train classification and REG. The input data source is from two datasets, i.e., REG data and auxiliary classification data with unseen categories. In the subject representation, we bridge the semantic and domain gap between the two datasets to transfer concepts from classification data to REG data. In the context modeling, we perceive the category-independent context to better inherit context from REG training data to ground novel objects. The language attention is learned to adaptively apply different importance to the two modules.

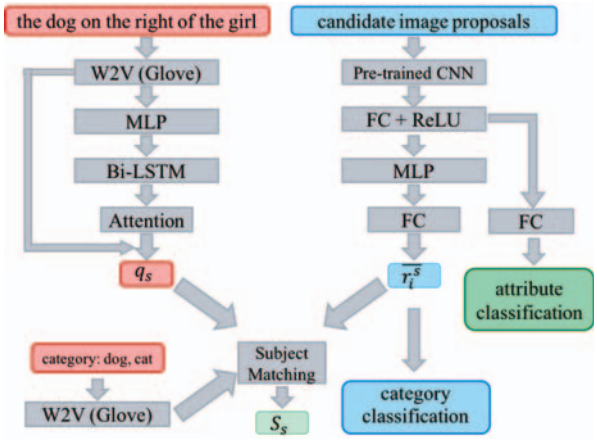

Figure 3: The detailed pipeline of subject representation. The bottom left branch shows the language feature encoding for auxiliary classification data. "dog" is the category of this referring expression. We also feed novel categories such as "cat" to the classification branch. The top left branch presents the language feature encoding for referring expressions of REG data. The right longer branch is the visual feature encoding for images from both datasets. The shorter one represents the attribute classification.

matching score between proposals and referring expressions upon subject.

$$\text{Score}_s = \text{COS}(q_s, \overline{r_i^s}), \tag{7}$$

3.3 Context Inheritance

The context information is crucial for REG to distinguish the target object from other same-category objects. Because of lacking the ground truth for grounding novel objects in the training data, here we learn to perceive category-independent context for novel objects. Our CCD is a concept-context disentangled structure, which can learn the category-independent context. The context modeling models two kinds of information, namely the location and relationship with other objects of the target object.

3.3.1 Language Feature Encoding. Similar as the subject feature encoding (Sec. 3.2.1), the language encoding for the context is as follows.

$$q_x = \sum_t \text{softmax}_t\left(\text{fc}\left(h_t\right)\right) e_t \quad x \in (l, r) \tag{8}$$

3.3.2 Context Encoder. The visual feature encoding for context models the location and relationship features. Location feature r_i^l consists of absolute position and relative locations with other objects of the same category in the image. Relationship feature r_i^r represents the relationship between the target proposal with its surrounding objects. We choose one with the maximum response to the query from 5 surrounding proposals as the relationship feature.

The location and relationship features are fed into a fully-connected layer to rescale to same dimension.

$$\widetilde{r_i^x} = \text{fc}\left(r_i^x\right) \quad x \in (l, r) \tag{9}$$

Then, we calculate the semantic similarity between the language and visual features as the matching score for context. The features are first fed into a two layer perceptron with batchnormalization layer.

$$\begin{aligned} \overline{r_i^x} &= \text{BN}\left(W_2 \phi_{\text{ReLU}}\left(W_1[\widetilde{r_i^x}]\right)\right), \quad x \in (l, r) \\ q_x &= \text{BN}\left(W_2 \phi_{\text{ReLU}}\left(W_1[q_x]\right)\right), \quad x \in (l, r) \end{aligned} \tag{10}$$

Finally, we calculate the cosine distance between the language representation and visual representation as the location and relationship matching scores, which represents the correspondence between proposals and referring expressions.

$$\text{Score}_x = \text{COS}(q_x, \overline{r_i^x}), \quad x \in (l, r) \tag{11}$$

3.4 End-to-End Joint Training and Inference

The overall matching score is the linear combination of the subject and context score. A language attention is introduced to learn their weights based on the referring expression features.

$$\begin{aligned} w_x &= \text{softmax}_w\left(\text{fc}\left([h_0, h_T]\right)\right), \qquad x \in (s, l, r) \\ S_t &= \sum_x w_x \text{Score}_x, \qquad x \in (s, l, r) \end{aligned} \tag{12}$$

For novel objects at inference, the weights measure the importance of subject and context.

A combined hinge loss is calculated to evaluate the grounding result as follows.

$$\begin{aligned} Loss_{reg} = \sum_i \Big[& \lambda_1 \max\left(0, \Delta + S_t^{ij} - S_t^{ii}\right) \\ & + \lambda_2 \max\left(0, \Delta + S_t^{ji} - S_t^{ii}\right) \Big] \end{aligned} \tag{13}$$

S_t^{ij} and S_t^{ji} denote the final matching score of negative pairs (r_i, q_j) and (r_j, q_i) respectively. S_t^{ii} is the matching score for the positive pair (r_i, q_i).

We use a multi-task learning mechanism to jointly train classification and REG tasks in an end-to-end manner. The network learns the concept information from both classification and REG data, and learns the context information from REG data. The cross-modal common semantic space and the concept-context disentangled structure guarantee the concept transfer and context inheritance for novel objects in the inference. The final loss is the combination of the attribute classification, category classification, and REG hinge loss:

$$L = Loss_{att} + Loss_{cat} + Loss_{reg}. \tag{14}$$

In the inference, we calculate the matching score between each candidate proposal and the query. We select the proposal with the maximum matching score as the best-matched region.

4 EXPERIMENTS

4.1 Datasets

We evaluate our method under two kinds of scenarios with novel objects, i.e., intra-dataset and inter-dataset, on four datasets: RefCOCO, RefCOCO+, RefCOCOg and RefCLEF. In the intra-dataset scenario, we randomly select 60 categories from RefCOCO, RefCOCO+, and RefCOCOg as their novel-object categories. In the inter-dataset scenario, the categories in the REG training data come from RefCOCO while the categories in the new-category test set come from RefClef. For the selected novel objects, we discard their referring expression and remain their corresponding categories as classification dataset. We do not use classification datasets such as ImageNet [3] because we do not have a corresponding REG test set for these classification data.

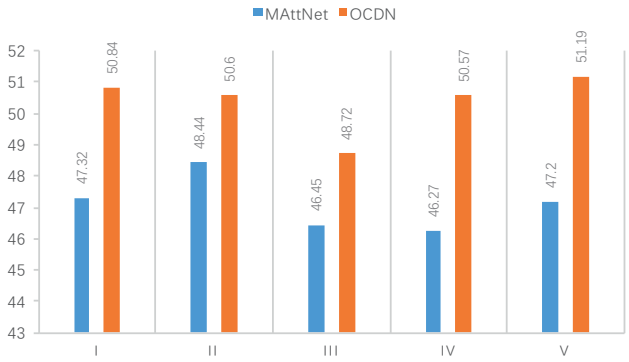

Figure 4: The results of different split on RefCOCO+ dataset. We randomly select different 60 categories as new-category REG test set.

RefCOCO [33]: It is also called UNC RefExp, which contains 142,209 queries for 50,000 objects with 80 categories in 19,994 images from MSCOCO [19]. We split the dataset into train, validation, Test A, Test B, and Test N. Test A contains multiple people; Test B contains multiple objects; Test N contains the novel objects.

RefCOCO+ [33]: It has 141,564 queries for 49,856 referents with 80 categories in 19,992 images from MSCOCO [19]. Different from RefCOCO, the queries in this dataset are disallowed to use locations to describe the referents. This dataset is also split into train, validation, Test A, Test B, and Test N respectively.

RefCOCOg [23]: It is also called Google Refexp. It has 95,010 queries for 49,822 objects with 80 categories in 25,799 images from MSCOCO [19]. It has longer queries containing appearance and location to describe the referents. The images are split into train, validation, Test S and Test N. Test S contains the seen objects while Test N contains the novel objects.

RefCLEF [15]: It is also called ReferIt. It contains 20,000 annotated images with 238 categories from IAPR TC-12 dataset [6] and SAIAPR-12 dataset [4] . The dataset includes some ambiguous queries, such as anywhere. It also has some mistakenly annotated image regions. We split the images into the train and validation set. The maximum length of all the queries is 19 words.

It is worth noting that each image from these four datasets contains at least 2 objects of the same object category.

4.2 Training Settings

4.2.1 Implementation details. The network is trained through Adam [16] algorithm with an initial learning rate of 4e-4, which is dropped

Table 1: Accuracy (IoU > 0.5) of REG on RefCOCO, RefCOCO+ and RefCOCOg dataset. The val, testA, testB or testS are seen-category test sets and the testN is new-category test set.

Methods	RefCOCO				RefCOCO+				RefCOCOg		
	val	testA	testB	testN	val	testA	testB	testN	val	testS	testN
MAttNet[32]	82.75	83.71	78.99	74.18	68.19	70.43	49.71	46.27	73.96	73.12	51.33
CCD	80.38	82.06	75.55	76.94	64.68	67.37	50.00	50.57	69.72	69.68	53.18

Table 2: Accuracy of classification on RefCOCO, RefCOCO+ and RefCOCOg dataset.

Methods	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val	test
CCD	79.98	82.76	75.69	77.34	80.28	73.04	79.18	78.60

by 10 after every 4,000 iterations. The training iterations are up to 12,000 with a batch size of 15 images. We can get better performance if the number of training iterations is larger. Here we use 12,000 as training iterations for time efficiency. Each image has an indefinite number of annotated queries. The rectified linear unit (ReLU) is used as the non-linear activation function. Batch normalization operations are used in our framework. ResNet is our main feature extractor for RoI visual features.

4.2.2 Metrics. The Intersection over Union (IoU) between the selected region and the ground-truth are calculated to evaluate the grounding performance. If the IoU score is greater than 0.5, the predicted region is considered as the right grounding. To evaluate the classification, we calculate the semantic distance between all categories and our results. The semantic distance is in the form of the euclidean metric. We use the closest category as the predicted category.

4.3 Experiments on RefCOCO, RefCOCO+ and RefCOCOg Datasets

4.3.1 Quantitative Results. We verify our method under intra-dataset scenario on the RefCOCO, RefCOCO+ and RefCOCOg datasets. We split the images of 80 classes in these datasets into two different sets. One set includes images of 20 classes (including human), and we further split the images into train, validation, and test sets to evaluate the performance of the seen-category set. One set includes the images of the other 60 classes, and this set is considered as the new-category test set. We use the variant of MAttNet [32] as our compared method. We delete the subject visual attention in MAttNet to keep consistent with our approach.

Table 1 shows the comparison between MAttNet and our method on RefCOCO, RefCOCO+ and RefCOCOg. In the seen-category test set (val, testA, testB or testS), the performance of our method slightly dropped. This may be because our method focuses more on the concept (category information) in subject representation and ignores the discriminative attribute information. In the new-category set, our method can improve the accuracy by 2.76%, 4.30% and 1.85% than MAttNet on three datasets respectively. This benefits from the learned cross-modal common semantic space in the subject representation of CCD, which can bridge the domain and semantic gap between classification data and REG data, and transfer the novel concepts from classification data to recognize them.

To better evaluate the stability of our method, we randomly choose 5 sets of data (each set has 60 classes) as a new-category test. Our method outperforms MAttNet on all splits. Fig. 4 shows the results of different splits on RefCOCO+ datasets. We can observe that our CCD outperforms MAttNet with a margin of 2.16~4.3%. The results also prove the robustness of our method.

Table 2 shows the classification results of our CCD on the three datasets. The split of the datasets for classification follows the original setting without splitting the new-category test set. The results show our CCD can get good performance on classification, and this indicates that our model can recognize most transferred concepts to improve the results of novel object grounding. Noting that we do not have classification results for the compared method as it lacks the classification branch.

4.3.2 Qualitative Results. We show qualitative results upon recognizing novel objects and perceiving category-independent context under the intra-dataset scenario on the three datasets. For each dataset, the first row shows the results of MAttNet[32]. The second row shows the results of our CCD.

Recognize novel objects. The left three columns of Fig. 5 show the qualitative results about recognizing novel objects. Compared with MAttNet, our CCD can better ground the novel objects in the new-category cases. This proves our method has learned the correspondence between novel category and its visual appearance, which benefits from the learned cross-modal common semantic space in CCD. This space helps to bridge the domain gap between two different datasets and transfer concepts from auxiliary classification data.

Perceive category-independent context. The right three columns of Fig. 5 show the comparison of qualitative results between MAttNet and CCD about perceiving category-independent context under scenarios where multiple objects of the same novel category situated together. The qualitative results show our CCD can distinguish the target object from other same-category objects. This indicates CCD can not only recognize novel objects but also successfully inherit learned context from the limited REG training data.

4.4 Experiments on RefCLEF Dataset

4.4.1 Quantitative Results. In this subsection, we evaluate our method on RefCLEF dataset in inter-dataset scenario. We use the

Figure 5: The comparison of qualitative results between MAttNet [32] and CCD under intra-dataset scenario. The left three columns show the results about recognizing novel objects. The right three columns are the visualization about perceiving the category-independent context. The denotations of the bounding box colors are as follows. Solid red: ground truth; dashed blue: predicted proposal. The referring expression is shown above corresponding images.

Table 3: Accuracy (IoU > 0.5) on RefCLEF dataset. The sets of val, testA and testB in RefCOCO are seen-category test sets and the val in RegCLEF is new-category test set.

Methods	RefCOCO			RefCLEF
	val	testA	testB	val
MAttNet[32]	81.91	82.52	82.28	20.47
CCD	80.95	80.70	80.57	55.03

training data of 80 categories in RefCOCO dataset as REG training data; the training data in RefClef as auxiliary classification data; the validation data in RefClef as new-category test set.

Table 3 shows the comparison between MAttNet and our CCD. The left three columns are the results of seen-category test sets of RefCOCO. The accuracy of MAttNet and our CCD are beyond 80.00% and ours has a slight decrement. The right column shows the results of new-category test data. We can observe that the accuracy of CCD is much higher (+34.56%) than MAttNet. This results from that CCD learns a cross-modal common semantic space to bridge the domain gap between two different datasets and transfers novel concepts from the classification data to the REG data, even when the datasets follow different data distribution. Besides, we evaluate the classification result (36.83%) on RefCLEF dataset. The performance of REG is much better (+18.2%) than classification on val set. This

Figure 6: The comparison of qualitative results between MAttNet [32] and CCD on RefCLEF under inter-dataset scenario. We specially show some results of the same image with different queries, indicating CCD can inherit the context learned from other REG datasets.

means our CCD still works to ground novel objects even when CCD cannot recognize the novel objects. This phenomenon indicates the concept-context disentangled structure enables CCD to perceive category-independent context.

4.4.2 Qualitative Results. Fig. 6 shows qualitative results upon recognizing novel objects and perceiving category-independent context under inter-dataset scenario on the RefCLEF dataset. The left three columns show the ability of our network for recognizing novel objects while the right three columns for perceiving category-independent context. The promotion under inter-dataset scenario is more obvious than under intra-dataset scenario. We show some examples of the same image with different queries. The queries have different context to ground different novel objects. Our method succeeds in distinguishing the category-independent context. The results show CCD can inherit the context learned from RefCOCO dataset to ground novel objects in RefCLEF dataset.

5 CONCLUSION

This paper focuses on the study of REG for novel objects. This task brings two new challenges: recognizing novel objects and perceiving category-independent context. To address this problem, we design transferrable REG with concept transfer and context inheritance. Specifically, we introduce a concept-context disentangled network (CCD) that enables concept transferring from auxiliary classification data and context inheriting from REG data to ground novel objects. As a modular network, CCD includes subject representation and context modeling. In subject representation, we learn the cross-modal common semantic space to bridge the semantic and domain gap between auxiliary classification data and REG data to recognize novel objects. In the context modeling, benefiting from the disentangled structure, the context is relatively independent of the subject. Thus we can learn the correspondence between image proposal and referring expression upon location and relationship. Benefiting from the disentangled structure, the context is relatively independent of the subject, so it can be better inherited from the REG training data. Finally, a language attention is learned to adaptively assign different importance to subject and context for grounding the referred novel objects. We use a multi-task learning mechanism to jointly training classification and REG tasks. Experiments on four datasets under intra- and inter-dataset scenarios show our method can better ground novel objects with concept transfer and context inheritance, even when multiple objects of the same category situated together.

ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China: 61771457, 61732007, 61772494, 61672497, 61622211, 61836002, 61472389, 61620106009 and U1636214, in part by Key Research Program of Frontier Sciences, CAS: QYZDJ-SSW-SYS013, and Fundamental Research Funds for the Central Universities under Grant WK2100100030.

REFERENCES

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2018. nocaps: novel object captioning at scale. *CoRR* abs/1812.08658 (2018).

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2017. Guided Open Vocabulary Image Captioning with Constrained Beam Search. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. 936–945.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. 248–255.

[4] Hugo Jair Escalante, Carlos A. Hernández, Jesús A. González, Aurelio López-López, Manuel Montes-y-Gómez, Eduardo F. Morales, Luis Enrique Sucar, Luis Villaseñor Pineda, and Michael Grubinger. 2010. The segmented and annotated IAPR TC-12 benchmark. *Computer Vision and Image Understanding* 114, 4 (2010), 419–428.

[5] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. 2018. IQA: Visual Question Answering in Interactive Environments. In *CVPR*. IEEE Computer Society, 4089–4098.

[6] Michael Grubinger, Paul Clough, Henning MÃijller, and Thomas Deselaers. 2006. The IAPR TC12 Benchmark: A New Evaluation Resource for Visual Information Systems. *Workshop Ontoimage* (10 2006).

[7] Sergio Guadarrama, Erik Rodner, Kate Saenko, Ning Zhang, Ryan Farrell, Jeff Donahue, and Trevor Darrell. 2014. Open-vocabulary Object Retrieval. In *Robotics: Science and Systems*.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. IEEE Computer Society, 770–778.

[9] Ryota Hinami and Shin'ichi Satoh. 2018. Discriminative Learning of Open-Vocabulary Object Retrieval and Localization by Negative Phrase Augmentation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*. 2605–2615.

[10] Ronghang Hu, Daniel Fried, Anna Rohrbach, Dan Klein, Trevor Darrell, and Kate Saenko. 2019. Are You Looking? Grounding to Multiple Modalities in Vision-and-Language Navigation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. 6551–6557.

[11] Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. 2019. Language-Conditioned Graph Networks for Relational Reasoning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[12] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2017. Modeling Relationships in Referential Expressions with Compositional Modular Networks. In *CVPR*. IEEE Computer Society, 4418–4427.

[13] Pingping Huang, Jianhui Huang, Yuqing Guo, Min Qiao, and Yong Zhu. 2019. Multi-grained Attention with Object-level Grounding for Visual Question Answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. 3595–3600.

[14] Vihan Jain, Gabriel Magalhães, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. 2019. Stay on the Path: Instruction Fidelity in Vision-and-Language Navigation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. 1862–1872.

[15] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *EMNLP*. ACL, 787–798.

[16] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014).

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.

[18] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. 2019. Pointing Novel Objects in Image Captioning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. 12497–12506.

[19] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *ECCV (5) (Lecture Notes in Computer Science, Vol. 8693)*. Springer, 740–755.

[20] Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. 2017. Referring Expression Generation and Comprehension via Attributes. In *ICCV*. IEEE Computer Society, 4866–4874.

[21] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. 2019. Improving Referring Expression Grounding With Cross-Modal Attention-Guided Erasing. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. 1950–1959.

[22] Ruotian Luo and Gregory Shakhnarovich. 2017. Comprehension-Guided Referring Expressions. In *CVPR*. IEEE Computer Society, 3125–3134.

[23] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and Comprehension of Unambiguous Object Descriptions. In *CVPR*. IEEE Computer Society, 11–20.

[24] Varun K. Nagaraja, Vlad I. Morariu, and Larry S. Davis. 2016. Modeling Context Between Objects for Referring Expression Understanding. In *ECCV (4) (Lecture Notes in Computer Science, Vol. 9908)*. Springer, 792–807.

[25] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. 1532–1543.

[26] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: Better, Faster, Stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 6517–6525.

[27] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*. 91–99.

[28] Arka Sadhu, Kan Chen, and Ram Nevatia. 2019. Zero-Shot Grounding of Objects from Natural Language Queries. *CoRR* abs/1908.07129 (2019).

[29] Jialin Wu, Zeyuan Hu, and Raymond J. Mooney. 2019. Generating Question Relevant Captions to Aid Visual Question Answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. 3585–3594.

[30] Sibei Yang, Guanbin Li, and Yizhou Yu. 2019. Cross-Modal Relationship Inference for Grounding Referring Expressions. *CoRR* abs/1906.04464 (2019).

[31] Sibei Yang, Guanbin Li, and Yizhou Yu. 2019. Dynamic Graph Attention for Referring Expression Comprehension. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[32] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. 2018. MAttNet: Modular Attention Network for Referring Expression Comprehension. In *CVPR*. IEEE Computer Society, 1307–1315.

[33] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling Context in Referring Expressions. In *ECCV (2) (Lecture Notes in Computer Science, Vol. 9906)*. Springer, 69–85.

[34] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. 2018. Grounding Referring Expressions in Images by Variational Context. In *CVPR*. IEEE Computer Society, 4158–4166.