Adaptive Reconstruction Network for Weakly Supervised Referring Expression Grounding

Xuejing Liu[1,2], Liang Li[*1], Shuhui Wang[1], Zheng-Jun Zha[3], Dechao Meng[1,2], and Qingming Huang[2,1,4]

[1]Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]University of Science and Technology of China, Hefei, China
[4]Peng Cheng Laboratory, Shenzhen, China
xuejing.liu, liang.li, dechao.meng @vipl.ict.ac.cn, wangshuhui@ict.ac.cn, zhazj@ustc.edu.cn, qmhuang@ucas.ac.cn

Abstract

Weakly supervised referring expression grounding aims at localizing the referential object in an image according to the linguistic query, where the mapping between the referential object and query is unknown in the training stage. To address this problem, we propose a novel end-to-end adaptive reconstruction network (ARN). It builds the correspondence between image region proposal and query in an adaptive manner: adaptive grounding and collaborative reconstruction. Specifically, we first extract the subject, location and context features to represent the proposals and the query respectively. Then, we design the adaptive grounding module to compute the matching score between each proposal and query by a hierarchical attention model. Finally, based on attention score and proposal features, we reconstruct the input query with a collaborative loss of language reconstruction loss, adaptive reconstruction loss, and attribute classification loss. This adaptive mechanism helps our model to alleviate the variance of different referring expressions. Experiments on four large-scale datasets show ARN outperforms existing state-of-the-art methods by a large margin. Qualitative results demonstrate that the proposed ARN can better handle the situation where multiple objects of a particular category situated together[1].

1. Introduction

Referring expression grounding (REG), also known as phrase localization, has been a surge of interest in both computer vision and natural language processing [24, 14, 29, 41, 39, 19, 38, 36]. Given a query (referring expression) in natural language and an image, REG is to find the corresponding location of the referential object. REG can be widely used in interactive applications, such as robotic navigation [31, 1], visual Q&A [6, 10], or photo editing [5].

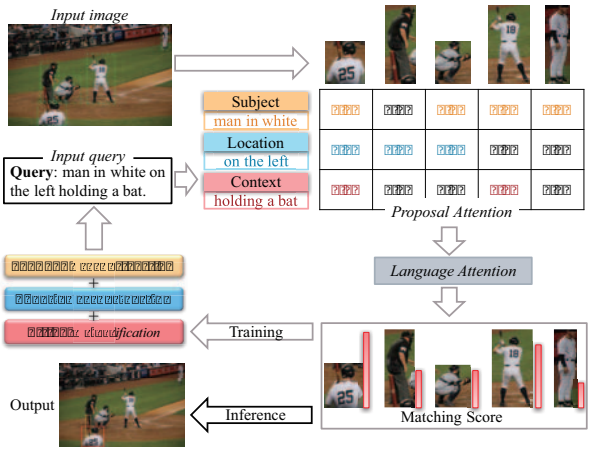


Figure 1. The proposed adaptive reconstruction network (ARN). Given a query and an image with region proposals, ARN localizes the referential object through adaptive grounding and collaborative reconstruction.

Training the REG model in a supervised manner requires expensive annotated data that explicitly draw the connection between the input query and its corresponding object proposal in the image. Besides, limited to the training data, supervised REG models can only handle the grounding with certain categories, which cannot meet the demand for real-world applications. Here we focus on weakly su-

*Corresponding author.
[1]Code is available at `https://github.com/GingL/ARN`

pervised REG task, where only the image-query pairs are used for training without the mapping information between the query and the object proposal.

Previous weakly supervised methods [29, 2, 44] learn to ground by reconstructing the input query. Xiao *et al.* [37] generate attention mask to localize linguistic query based on image-phrase pairs and language structure. Zhao *et al.* [45] try to find the location of the referential object by searching over the entire image. The above methods only exploit the visual appearance features of proposals during grounding and reconstruction. However, they ignore the discriminative information of location and context from the referential object, and cannot distinguish a specific object where multiple objects of a particular category situated together. As shown in Fig. 1 , given the query of "man in white on the left holding a bat", apart from the subject ("man in white"), location ("on the left") and context ("holding a bat") play an essential part in distinguishing the referential object.

Recently Yu *et al*. [41] find that people tend to use different syntax structures when referring to an object, and this brings the variance of different referring expressions. Taking Fig. 1 as an example, if the query is only "man in black", it can be grounded using subject features only. Similarly, "man on the far right" concentrates more on location features, and "man on the right of the man in black" should focus more on context features. Therefore, the grounding is triggered based on what features are present in the referring expression.

In light of these observations, we propose a novel end-to-end weakly supervised REG method, coined Adaptive Reconstruction Network (ARN). It learns the mapping between image region proposal and query upon the subject, location and context information in an adaptive manner. Fig. 1 shows the pipeline of ARN, that consists of two modules: adaptive grounding, and collaborative reconstruction.

Adaptive Grounding. First, we extract the subject, location, and context features of both the query and each region proposal in an image. Specifically, for the query, we introduce a recurrent net to parse it into these three features. For a proposal, we extract its visual appearance feature as the subject feature by Faster R-CNN [28]. Moreover, the location feature of the proposal consists of absolute position and relative locations with other proposals of the same category in the image. Furthermore, the context feature of the proposal is represented by concatenating the visual and relative location features of its surrounding proposals. Second, we propose the adaptive grounding module to compute the matching score between each proposal and query by a hierarchical attention model. The first attention helps generate attention scores upon subject, location, and context for each proposal respectively. The second one further learns the attention score of the above three components based on the syntax structure of the query. This module can alleviate the variance of different referring expressions.

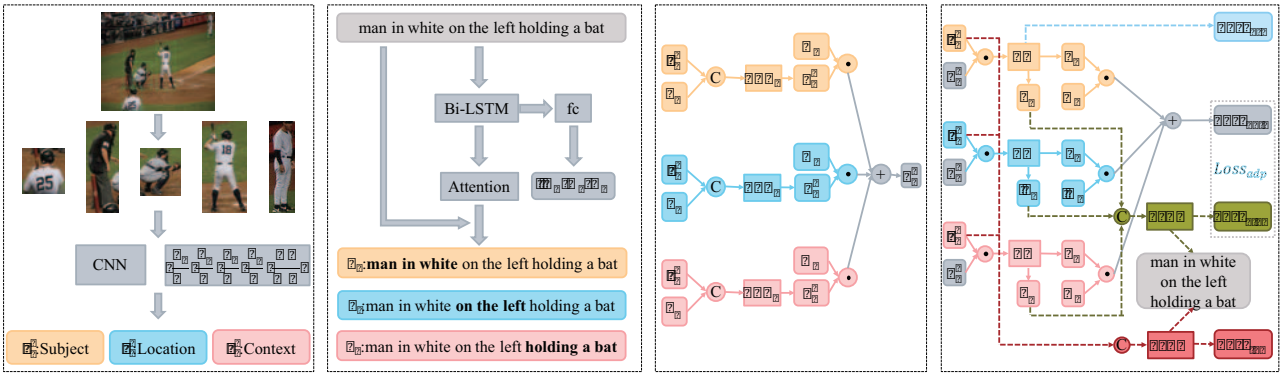
Collaborative Reconstruction. We design a collaborative loss to better formulate the measurement of weakly supervised grounding. The loss function derives from the following three parts. *Language reconstruction* directly reconstructs the input query based on the attentive proposal features. *Adaptive reconstruction* reconstructs attentive hidden features of subject, location and context respectively. *Attribute classification* leverages the attribute information of candidate proposal upon the subject to improve the grounding ability.

Both modules of ARN can be trained in an end-to-end manner. At the inference stage, ARN only utilizes the adaptive grounding to localize the referent without reconstruction. To summary, the main contribution of this paper is three-fold:

- We propose an end-to-end adaptive reconstruction network that models the mapping between input query and image upon subject, location and context features. ARN adaptively grounds the candidate proposals by hierarchical attention, which could alleviate the variance of different referring expressions.
- We design a collaborative reconstruction module to reconstruct the input query based on the matching score and proposal features. A collaborative loss of language reconstruction, adaptive reconstruction, and attribute classification are formulated for the measurement of adaptive grounding.
- Comparison experiments on the RefCLEF and three MS-COCO datasets show that the proposed ARN achieves state-of-the-art results in the weakly supervised REG task.

2. Related Work

Referring Expression Grounding (REG). REG [15, 25, 9, 24, 42, 43, 4, 22] is also known as referring expression comprehension or phrase localization, which is the inverse task of referring expression generation. REG aims to localize the corresponding object described by a free-form natural language query in an image. Given an image I, a query q and a set of region proposals $r_i\ _{i=1}^{N}$, REG selects the best-matched region r^* according to the query. Most REG methods can be roughly divided into two kinds. One is CNN-LSTM based encoder-decoder structure to model $P(q\,I, r)$ [24, 42, 27, 14, 23, 18, 30]. The other is the joint vision-language embedding framework to model $P(q, r)$. During training, the supervision is object proposal and referring expression pairs (r_i, q_i) [29, 33, 21, 3, 41, 17, 34]. Recently, MattNet [41] adopts subject, location and relation features on supervised REG



(a) Visual feature encoding. (b) Language feature encoding. (c) Adaptive grounding. (d) Collaborative Reconstruction.

Figure 2. The network architecture of the proposed ARN. It consists of feature encoding (Section 3.1), adaptive grounding (Section 3.2) and collaborative reconstruction (Section 3.3). Collaborative reconstruction module contains three losses: attribute classification loss, adaptive reconstruction loss (including adaptive language reconstruction and adaptive visual reconstruction loss) and language reconstruction loss. Visual features are pre-extracted from external networks. Language feature encoding, adaptive grounding and collaborative reconstruction are trained as an end-to-end network. The reconstruction module is not needed during inference. ATT: attention layer. $\oplus$: plus operation. $\odot$: element-wise vector multiplication. C: vector concatenation.

and gets state-of-the-art results. The above features prove to be effective in the grounding task, which are also used as the original feature representation in our method. But we design the collaborative reconstruction to bridge the gap between supervised and weakly supervised learning, achieving impressive results on weakly supervised REG.

Weakly Supervised Referring Expression Grounding. Weakly supervised REG only has image-level correspondence, and there is no mapping between image regions and referring expressions. To solve this problem, Rohrbach *et al.* [29] propose a framework which learns to ground by reconstructing the given referring expression through attention mechanism. Based on this framework, Chen *et al.* [2] design knowledge aided consistency network, which reconstructs both the input query and proposal's information. Xiao *et al.* [37] ground arbitrary linguistic phrase in the form of spatial attention mask and propose a network with discriminative and structural loss. Different from selecting the optimal region from a set of region proposals, Zhao *et al.* [45] propose multi-scale anchored transformer network, which can search the entire spatial feature map by taking region proposals as anchors to get more accurate location. Zhang *et al.* [44] propose a variational Bayesian method to exploit the relationship between the referent and context.

3. Methodology

We propose an adaptive reconstruction network (ARN) to ground the target proposal described by the query in weakly supervised scenario, where the training data do not have the region-query correspondence. This problem can be formulated as follows. Given an image I, a 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 without knowing any (q, r_i) pair. ARN chooses the most probable proposal through adaptive grounding, then reconstructs its corresponding query with a collaborative loss. The whole network architecture is shown in Fig. 2.

3.1. Feature Encoding

3.1.1 RoI Features

For each object proposal r_i, the subject, location and context features are extracted as shown in Fig. 2(a).

Subject feature is extracted as visual appearance features of proposals. We run the forward propagation of Faster R-CNN based on ResNet [13] for each image, and crop its C3 and C4 features as the subject feature $r_s^i = f_{CNN}(r_i)$. The C3 features represent lower-level features such as colors and shapes while C4 features contain higher-level representations.

Location feature consists of absolute position and relative locations with other objects of the same category in the image. Following [42, 43, 41], the absolute location feature of each proposal is decoded as a 5-dim vector $r_l^i = \frac{x_{tl}}{W}, \frac{y_{tl}}{H}, \frac{x_{br}}{W}, \frac{y_{br}}{H}, \frac{w \cdot h}{W \cdot H}$, denoting the top-left, bottom-right position and relative area of the proposal to the whole image. The relative location feature indicates the relative location information between the proposal and 5 surrounding proposals of the same category. For each surrounding proposal, we calculate its offset and area ratio to the candidate: $\delta r_l^{ij} = \frac{[\triangle x_{tl}]_{ij}}{w_i}, \frac{[\triangle y_{tl}]_{ij}}{h_i}, \frac{[\triangle x_{br}]_{ij}}{w_i}, \frac{[\triangle y_{br}]_{ij}}{h_i}, \frac{w_j h_j}{w_i h_i}$.

Finally, we concatenate the above absolute and relative location feature into the location feature of the proposal, which is a 30-dim vector: $r_l^i = \left[r_l^i; \delta r_l^i \right]$.

Context feature represents the relationship between the candidate proposal and environment. Following [41], we choose 5 surrounding proposals as the relative ones for each proposal. The feature of each proposal is composed of C4 feature $v_{ij} = f_{CNN}(r_j)$ and its relative location feature. The relative location feature is encoded as follows: $\delta m_{ij} = \left[\frac{[\triangle x_{tl}]_{ij}}{w_i}, \frac{[\triangle y_{tl}]_{ij}}{h_i}, \frac{[\triangle x_{br}]_{ij}}{w_i}, \frac{[\triangle y_{br}]_{ij}}{h_i}, \frac{w_j h_j}{w_i h_i} \right]$. The context feature is $r_c^{ij} = [v_{ij}; \delta m_{ij}]$. From above 5 proposals, we choose the one with the maximum response to the query as the final relative object, denoted as r_c^i .

3.1.2 Referring Expression Features

Corresponding to RoI features, the query features are also separated into **subject** q_s, **location** q_l and **context** q_c through attention mechanism, as shown in Fig. 2(b). Given an query $q = \left\{ w_t \right\}_{t=1}^{T}$, first each word in q is one-hot encoded and mapped into a word embedding e_t. Then the word embedding e_t is fed into a bi-directional LSTM. The final representation $h_t = [\overrightarrow{h}_t, \overleftarrow{h}_t]$ is the concatenation of the hidden vectors in both directions. Words are attended in each query for better representation of subject, location and context through attention mechanism. Take subject feature q_s as an example, its final hidden representation is calculated as follows:

$$\begin{aligned} m_t &= \mathrm{fc}\left(h_t\right), \\ \alpha_t &= \mathrm{softmax}_t\left(m_t\right), \\ q_s &= \sum_t \alpha_t e_t. \end{aligned} \tag{1}$$

Location feature q_l and context feature q_c can be obtained using the same mechanism. Besides, three different weights upon subject, location and context are calculated from the hidden state vector of the bi-directional LSTM.

$$[w_s, w_l, w_c] = \mathrm{softmax}_w\left(\mathrm{fc}\left([h_0, h_T]\right)\right) \tag{2}$$

3.2. Adaptive Grounding

Based on the subject, location and context features of both the proposal and query, ARN localizes the query through a hierarchical attention model. The first attention is the **proposal attention**, which calculates the matching score between the proposals and query upon subject, location and context respectively. The second attention is **language attention**, which assigns different weights to subject, location and context based on the query to alleviate variance in queries.

Detailed process can be seen in Fig. 2(c), r_s^i, r_l^i and r_c^i are the visual features extracted from the region proposals in the image through CNN. q_s, q_l and q_c represent the language feature extracted from the query through bi-directional LSTM. Taking the subject as an example, r_s^i and q_s are first concatenated into one vector. Then the vector is fed into the proposal attention, which is a two layer perceptron, to get the corresponding matching score. The biases are omitted in Eq. (3).

$$\overline{s}_x^i = f_{ATT}\left(q_x, r_x^i\right) = W_2 \phi_{\mathrm{ReLU}}\left(W_1 [q_x, r_x^i]\right), x \in (s, l, c) \tag{3}$$

We normalize the scores using softmax.

$$s_x^i = \mathrm{softmax}_i\left(\overline{s}_x^i\right), x \in (s, l, c) \tag{4}$$

The total score is calculated based on language attention, which is the linear combination of the three sub-score. The final score represents the probability of region i matching query q considering subject, location and context. The weights are calculated based on the query.

$$S_t^i = w_s s_s^i + w_l s_l^i + w_c s_c^i \tag{5}$$

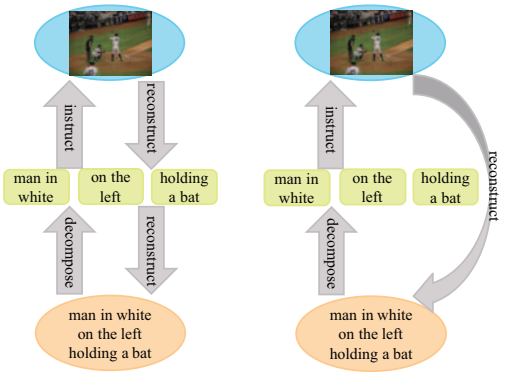


Figure 3. The sketch map of (a) Adaptive reconstruction and (b) Language reconstruction.

3.3. Collaborative Reconstruction

Because there are no mapping data between the query and the proposal of the image in the weak supervised training stage, the collaborative reconstruction is used to formulate the measurement of the grounding. The collaborative loss is designed with three losses, as shown in Fig. 2(d). Adaptive reconstruction reconstructs attentive hidden features of subject, location and context respectively. Language reconstruction directly reconstructs the input query based on the attentive features of proposals. Attribute classification takes advantage of attribute information of the referential proposal.

3.3.1 Adaptive Reconstruction Loss

Adaptive reconstruction loss utilizes three hidden vectors (subject, location, context) to bridge the gap between the

input query and proposals, as Fig. 3(a) shows. This loss consists of two sub-losses, adaptive visual reconstruction loss and adaptive language reconstruction loss. This is inspired by the idea that reconstructing different linguistic query using corresponding features of proposals can better handle the variance among different expressions in the datasets. The adaptive visual reconstruction loss is to reconstruct query features q_s, q_l and q_c using features of proposals r_s^i, r_l^i and r_c^i. We first compute a weighted sum over the different visual features and the matching scores.

$$\tilde{v}_s = \sum_{i=1}^{N} S_t^i \tilde{r}_s^i, \quad \tilde{v}_l = \sum_{i=1}^{N} S_t^i \tilde{r}_l^i, \quad \tilde{v}_c = \sum_{i=1}^{N} S_t^i \tilde{r}_c^i \tag{6}$$

The aggregation of the proposal features from the attentive proposals are then fed into a fully connected layer to get the same dimension with the language features.

$$v_s = FC(\tilde{v}_s), \quad v_l = FC(\tilde{v}_l), \quad v_c = FC(\tilde{v}_c) \tag{7}$$

Then we use the attentive proposal features v_s, v_l and v_c to reconstruct the language features q_s, q_l and q_c extracted from the original query. We use Mean Squared Error (MSE) criterion to minimize the distance between the proposal features and language features.

$$L_x = \mathrm{MSE}(v_x, q_x), \quad x \quad (s, l, c) \tag{8}$$

The final adaptive visual reconstruction loss is the weighted sum of the subject reconstruction loss, location reconstruction loss and context reconstruction loss. The weights are calculated based on the query, as subsection 3.1.2 shows.

$$Loss_{avis} = w_s L_s + w_l L_l + w_c L_c \tag{9}$$

However, the language feature extraction network is trained together with the grounding and reconstruction network. To reach convergence as soon as possible, network parameters might be set to zero roughly so that the network can not learn the correspondence between the visual modality and language modality. To avoid this circumstance, we add an adaptive language reconstruction loss, which utilizes the language features q_s, q_l and q_c to reconstruct the original query. First we concatenate q_s , q_l and q_c, then feed it into a one-layer perceptron.

$$f_{alan} = \phi_{\mathrm{ReLU}} \left(W_l([q_s, q_l, q_c]) + b_l \right) \tag{10}$$

Based on the fused language features f_{lan}, we reconstruct the input query through LSTM. This is inspired by the query generation methods [7, 32]. The language features f_{lan} are fed into a one-layer LSTM only at the first time step.

$$P(q \; f_{alan}) = f_{\mathrm{LSTM}} \left(f_{alan} \right) \tag{11}$$

The language reconstruction network aims to maximize the likelihood of the ground-truth query $\hat{q}$ generated by LSTM, as Eq. (12) shows. B is the batch size.

$$Loss_{alan} = \frac{1}{B} \sum_{b=1}^{B} \log(P(\hat{q} \; f_{alan})) \tag{12}$$

The final adaptive reconstruction loss is the weighted sum of the language reconstruction loss and visual reconstruction loss. α and β is the hyper-parameters defining the proportion of the two losses. In this adaptive reconstruction loss, both of the language and visual reconstruction loss are indispensable.

$$Loss_{adp} = \alpha Loss_{avis} + \beta Loss_{alan} \tag{13}$$

3.3.2 Language Reconstruction Loss

The second reconstruction loss is to directly reconstruct the input query based on the attentive proposal features, as Fig. 3(b) shows. First, the concatenation of the original proposal features r_s^i, r_l^i and r_c^i are fed into a one-layer perceptron.

$$r_{vis}^i = \phi_{\mathrm{ReLU}} \;\; W_v(\; r_s^i, r_l^i, r_c^i \;) + b_v \tag{14}$$

Then we calculate the weighted sum of the proposal features according to the total score.

$$f_{vis} = \sum_{i=1}^{N} S_t^i r_{vis}^i \tag{15}$$

Based on the fused proposal features, query are generated through LSTM.

$$P(q \; f_{vis}) = f_{\mathrm{LSTM}} \left(f_{vis} \right) \tag{16}$$

We use the same language reconstruction loss as Eq. 12.

$$Loss_{lan} = \frac{1}{B} \sum_{b=1}^{B} \log(P(\hat{q} \; f_{vis})) \tag{17}$$

Compared to the adaptive reconstruction, the language reconstruction reconstructs the input query directly, so it will not lose any useful language information during training.

3.3.3 Attribute Classi cation Loss

As mentioned in previous methods [40, 35, 41], attribute information is important on distinguish object of the same category. Here, we add an attribute classification branch in our model. The attribute label is extracted through an external language parser [15] according to [41]. Subject feature r_s^i of proposal is used for attribute classification. As each query has multiple attribute labels, we use the binary cross-entropy loss for the multi-label classification.

$$Loss_{att} = f_{\mathrm{BCE}}\left(y_{ij}, p_{ij}\right) \tag{18}$$

We use the reciprocal of the frequency that attribute labels appears as weights in this loss to ease unbalanced data.

3.4. Training and Inference

The referring expression feature extraction network, the grounding network and the reconstruction network are trained with end-to-end strategy. During training, only query with attribute words goes through attribute classification branch. At inference, the reconstruction module is not needed anymore. We feed the image and query into the network, and get the most related proposal whose final score is the maximal in the grounding module.

$$j = \arg\max_{i} f\left(p, r_i\right) \tag{19}$$

The final collaborative reconstruction loss is:

$$Loss = Loss_{adp} + \gamma Loss_{lan} + \lambda Loss_{att} \tag{20}$$

4. Experiments

4.1. Datasets

We evaluate our method on four popular benchmarks of referring expression grounding.

RefCOCO [42]. The dataset contains 142,209 queries for 50,000 objects in 19,994 images from MSCOCO [20]. The dataset is split into train, validation, Test A, and Test B, which has 16,994, 1,500, 750 and 750 images, respectively. Test A contains multiple people while Test B contains multiple objects. Each image contains at least 2 objects of the same object category.

RefCOCO+ [42]. It has 141,564 queries for 49,856 referents in 19,992 images from MSCOCO [20]. Different from RefCOCO, the queries in this dataset are disallowed to use locations to describe the referents. The split is 16,992, 1,500, 750 and 750 images for train, validation, Test A, and Test B respectively. Each image contains 2 or more objects of the same object category in this dataset.

RefCOCOg [24]. It has 95,010 queries for 49,822 objects in 25,799 images from MSCOCO [20]. It has longer queries containing appearance and location to describe the referents. The split is 21,149 and 4,650 images for training and validation. There is no open test split for RefCOCOg. Images were selected to contain between 2 and 4 objects of the same category.

RefCLEF [15]. It contains 20,000 annotated images from IAPR TC-12 dataset [11] and SAIAPR-12 dataset [8]. The dataset includes some ambiguous queries, such as anywhere. It also has some mistakenly annotated image regions. The dataset is split into 9,000, 1,000 and 10,000 images for training, validation and test for fair comparison with [29]. 100 bounding box proposals [14] are provided for each image using Edge Boxes [46]. Images contain between 2 and 4 objects of the same object category. The maximum length of all the queries is 19 words.

4.2. Experimental Setup

4.2.1 Implementation details

The proposed ARN is trained through Adam [16] algorithm with an initial learning rate 4e-4, which is dropped by 10 after every 8,000 iterations. The training iterations are up to 30,000 with a batch size of a single image. Each image has an indefinite number of annotated queries. ResNet is our main feature extractor for RoI visual features. We adopt EdgeBoxe [46] to generate 100 region proposals for RefCLEF dataset for fair comparison with [29, 2]. Besides, we also show the performance based on detected objects from Faster R-CNN. It is worth noting that we do not extract the context features for RefCLEF dataset. As there are 100 region proposals in each image of the dataset, it is not reasonable to choose one from 5 surrounding proposals as context of the candidate proposal.

4.2.2 Metrics

The Intersection over Union (IoU) between the selected region and the ground-truth are calculated to evaluate the network performance. If the IoU score is greater than 0.5, the predicted region is considered as the right grounding.

4.3. Results

4.3.1 RefCOCO Datasets

Performance Analysis: Table 1 reports the results on RefCOCO, RefCOCO+ and RefCOCOg datasets. We compared the proposed ARN with the only published unsupervised results on these datasets [44]. We can have the following findings. First, adaptive reconstruction performs better on testA, which contains multiple people. Language reconstruction achieves better performance on testB, which contains multiple other objects. Second, the collaborative loss can get second best on all the test, indicating it can better handle different kinds of datasets. We also show the results using detected object proposals from Faster R-CNN. The performance drops due to detection error.

Ablastion Study: Table 2 reports the results on RefCOCO datasets with different settings. α, β, γ, λ denoted the weights on $Loss_{avis}$, $Loss_{alan}$, $Loss_{lan}$, $Loss_{att}$, respectively. The proportion is based on the order of magnitude of different losses. We find that when $Loss_{lan}$ accounts for a more significant part in the collaborative loss, the performance on testA will drop greatly. While when the propor-

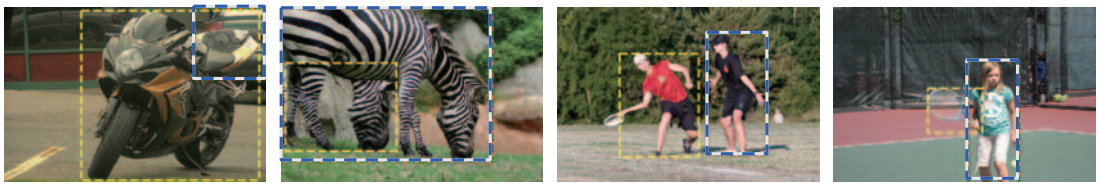
Table 1. Accuracy (IoU $>$ 0.5) on RefCOCO dataset. **Bond**: best result. Red: second best result. Blue: best result of VC.

Methods	Settings	RefCOCO			RefCOCO+			RefCOCOg
		val	testA	testB	val	testA	testB	val
VC	w/o reg	-	13.59	21.65	-	18.79	24.14	25.14
VC	-	-	17.34	20.98	-	23.24	24.91	33.79
VC	w/o α	-	33.29	30.13	-	34.60	31.58	30.26
VC (det)	w/o reg	-	17.14	22.30	-	19.74	24.05	28.14
VC (det)	-	-	20.91	21.77	-	25.79	25.54	33.66
VC (det)	w/o α	-	32.68	27.22	-	34.68	28.10	29.65
ARN	$L_{adp} + L_{att}$	33.07	**36.43**	29.09	33.53	**36.40**	29.23	33.19
ARN	$L_{lan} + L_{att}$	**38.05**	35.27	**36.47**	34.51	34.40	**36.12**	**39.62**
ARN	$L_{lan} + L_{adp}$	33.60	35.65	31.48	34.40	35.54	32.60	34.50
ARN (det)	$L_{lan} + L_{adp}$	31.58	35.50	28.32	31.73	34.23	29.35	32.60
ARN	$L_{lan} + L_{adp} + L_{att}$	34.26	36.01	33.07	**34.53**	36.01	33.75	34.66
ARN (det)	$L_{lan} + L_{adp} + L_{att}$	32.17	35.35	30.28	32.78	34.35	32.13	33.09

Table 2. Albation study on RefCOCO dataset.

	Settings				RefCOCO			RefCOCO+			RefCOCOg
	α	β	γ	λ	val	testA	testB	val	testA	testB	val
case 1	1	1	1	0	32.92	36.40	29.26	33.06	36.34	29.60	33.08
case 2	0.01	1	1	1	34.32	36.24	33.05	35.60	36.92	33.09	34.44
case 3	0.01	1	5	1	34.26	36.01	33.07	34.53	36.01	33.75	34.66
case 4	0.01	1	10	1	34.18	35.83	32.29	32.39	33.39	32.89	34.24
case 5	0.01	1	15	1	29.09	27.13	33.09	29.97	27.98	33.99	34.94
case 6	0.01	1	20	1	29.87	27.86	33.05	29.20	25.57	35.28	34.60

animal far left | cow on right | giraffe on right | tallest giraffe



a car parked behind a motorcycle | zebra on the right of the other zebra | a woman wearing black next to man in a red shirt | a girl in blue shirt playing tennis

middle donut | the biggest clock on the right | biggest lamb | bag to the far right

Figure 4. Qualitative results on MSCOCO datasets. The denotations of the bounding box colors are as follows. Solid white: ground truth; dashed blue: predicted proposal; dashed yellow: context ground.

Table 3. Accuracy (IoU > 0.5) on RefCLEF dataset.

Method	IoU
LRCN [7]	8.59
Caffe-7K [12]	10.38
GroundeR [29]	10.70
MATN [45]	13.61
VC [44]	14.11
VC w/o α [44]	14.50
KAC Net [2]	15.83
ARN ($loss_{lan}$)	21.86
ARN ($loss_{lan} + loss_{adp}$)	25.35
ARN ($loss_{lan} + loss_{adp} + loss_{att}$)	**26.19**

Table 4. Ablation study on RefCLEF dataset.

	α	β	γ	λ	val
case 1	0.001	1	10	0	24.14
case 2	0.01	1	10	0	21.83
case 3	0.001	10	10	0	22.55
case 4	0.001	1	1	0	22.34
case 5	0.001	1	1	1	25.35
case 6	0.001	1	10	1	24.34
case 7	0.001	1	20	1	24.76
case 8	0.001	1	30	1	26.19
case 9	0.001	1	40	1	25.53
case 10	0	0	1	0	21.86

tion of $Loss_{adp}$ is bigger, the results in testB will be a disaster. After the parameter search, we find that the settings in case5 get good result on all the datasets.

4.3.2 RefCLEF Dataset

Performance Analysis: We compare our adaptive reconstruction network (ARN) with state-of-the-art supervised referring expression grounding methods. Table 3 reports the results on RefCLEF dataset. We can see that ARN outperforms state-of-the-art result by 10.36%. We can have the following observations. First, with only language reconstruction loss, our method outperforms state-of-the-art result by 6.03%, which indicates our proposed adaptive grounding module taking effect. Second, adding our proposed adaptive reconstruction module, the performance achieves another 3.49% increase compared to with language reconstruction loss only. Third, the attribute classification loss also helps localization, the performance increase by 0.84% compared to previous result.

Ablation Study: We study the benefits of each loss module by running ablation experiments. Table 4 reports the results on RefCLEF dataset with different loss proportion. α, β, γ, $lambda$ denoted the weights on $Loss_{avis}$, $Loss_{alan}$, $Loss_{lan}$, $Loss_{att}$, respectively. The adaptive visual reconstruction loss is first set as 0.001 based on the order of magnitude. We can have the follow ablation experiment. We change the proportion of $Loss_{avis}$ and $Loss_{alan}$ in case 2 and case 3 compared to case 1, respectively. We find the result is better when α is 0.001, due to the order of magnitude in $Loss_{avis}$. The comparison of case 1 and case 6, case 4 and case 5 show that attribute classification loss can improve the grounding results. case 6, case 7 ,case 8 and case 9 show that when the proportion of $Loss_{lan}$ is 30, the performance of the network will be better. However, when we only use the $Loss_{lan}$ in case 10, the results are not as good as the combination of $Loss_{adp}$ and $Loss_{lan}$.

4.3.3 Qualitative Results

Fig. 4 shows qualitative example predictions on RefCOCO, RefCOCO+ and RefCOCOg datasets. The query is shown below corresponding images. The ground truth, grounding proposal and context proposal are denoted as solid white, dashed blue and dashed yellow, respectively. The first row shows the result based on different query in the same image. The proposed ARN can handle the location information correctly. The second row shows some examples with context information. ARN correctly grounds both the referential object and context object. The third row shows some difficult examples where multiple objects of the same category exist. It shows that the proposed ARN can help to ground in the hard cases which contain multiple objects of the same category.

5. Conclusion

To address the weakly supervised referring expression grounding problem, we propose a novel end-to-end adaptive reconstruction network. The ARN models the mapping between image proposal and query upon the subject, location and context information through adaptive grounding and collaborative reconstruction. Specially, a hierarchical attention model is designed to adaptively ground the query on the proposal with proposal attention and language attention. This model is trained by minimizing a collaborative reconstruction loss, which consists of language reconstruction loss, adaptive reconstruction loss and attribute classification loss. Experiments demonstrate that the proposed method provides a significant improvement in performance on RefCLEF, RefCOCO, RefCOCO+ and RefCOCOg datasets.

Acknowledgements. 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] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian D. Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, pages 3674–3683. IEEE Computer Society, 2018.

[2] Kan Chen, Jiyang Gao, and Ram Nevatia. Knowledge aided consistency for weakly supervised phrase grounding. In *CVPR*, pages 4042–4050. IEEE Computer Society, 2018.

[3] Kan Chen, Rama Kovvuri, and Ram Nevatia. Query-guided regression network with context policy for phrase grounding. In *ICCV*, pages 824–832. IEEE Computer Society, 2017.

[4] Xinpeng Chen, Lin Ma, Jingyuan Chen, Zequn Jie, Wei Liu, and Jiebo Luo. Real-time referring expression comprehension by single-stage grounding network. *CoRR*, abs/1812.03426, 2018.

[5] Ming-Ming Cheng, Shuai Zheng, Wen-Yan Lin, Vibhav Vineet, Paul Sturgess, Nigel Crook, Niloy J. Mitra, and Philip H. S. Torr. Imagespirit: Verbal guided image parsing. *ACM Trans. Graph.*, 34(1):3:1–3:11, 2014.

[6] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *CVPR*, pages 1–10. IEEE Computer Society, 2018.

[7] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, pages 2625–2634. IEEE Computer Society, 2015.

[8] 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. The segmented and annotated IAPR TC-12 benchmark. *Computer Vision and Image Understanding*, 114(4):419–428, 2010.

[9] Nicholas FitzGerald, Yoav Artzi, and Luke S. Zettlemoyer. Learning distributions over logical forms for referring expression generation. In *EMNLP*, pages 1914–1925. ACL, 2013.

[10] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. IQA: visual question answering in interactive environments. In *CVPR*, pages 4089–4098. IEEE Computer Society, 2018.

[11] Michael Grubinger, Paul Clough, Henning Mller, and Thomas Deselaers. The iapr tc12 benchmark: A new evaluation resource for visual information systems. *Workshop Ontoimage*, 10 2006.

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

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016.

[14] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *CVPR*, pages 4555–4564. IEEE Computer Society, 2016.

[15] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798. ACL, 2014.

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[17] Liang Li, Shuqiang Jiang, and Qingming Huang. Learning hierarchical semantic description via mixed-norm regularization for image understanding. *IEEE Trans. Multimedia*, 14(5):1401–1413, 2012.

[18] Liang Li, Shuqiang Jiang, Zheng-Jun Zha, Zhipeng Wu, and Qingming Huang. Partial-duplicate image retrieval via saliency-guided visual matching. *IEEE MultiMedia*, 20(3):13–23, 2013.

[19] Liang Li, Shuhui Wang, Shuqiang Jiang, and Qingming Huang. Attentive recurrent neural network for weak-supervised multi-label image classification. In *ACM Multimedia*, pages 1092–1100, 2018.

[20] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV (5)*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014.

[21] Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. Referring expression generation and comprehension via attributes. In *ICCV*, pages 4866–4874. IEEE Computer Society, 2017.

[22] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Li Su, and Qingming Huang. Knowledge-guided pairwise reconstruction network for weakly supervised referring expression grounding. In *ACM Multimedia*, 2019.

[23] Ruotian Luo and Gregory Shakhnarovich. Comprehension-guided referring expressions. In *CVPR*, pages 3125–3134. IEEE Computer Society, 2017.

[24] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20. IEEE Computer Society, 2016.

[25] Margaret Mitchell, Kees van Deemter, and Ehud Reiter. Generating expressions that refer to visible objects. In *HLT-NAACL*, pages 1174–1184. The Association for Computational Linguistics, 2013.

[26] Guido F. Montúfar, Razvan Pascanu, KyungHyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *NIPS*, pages 2924–2932, 2014.

[27] Varun K. Nagaraja, Vlad I. Morariu, and Larry S. Davis. Modeling context between objects for referring expression understanding. In *ECCV (4)*, volume 9908 of *Lecture Notes in Computer Science*, pages 792–807. Springer, 2016.

[28] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[29] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV (1)*, volume 9905 of *Lecture Notes in Computer Science*, pages 817–834. Springer, 2016.

[30] Guoli Song, Shuhui Wang, Qingming Huang, and Qi Tian. Multimodal similarity gaussian process latent variable model. *IEEE Trans. Image Processing*, 26(9):4168–4181, 2017.

[31] Jesse Thomason, Jivko Sinapov, and Raymond J. Mooney. Guiding interaction behaviors for multi-modal grounded language learning. In *RoboNLP@ACL*, pages 20–24. Association for Computational Linguistics, 2017.

[32] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164. IEEE Computer Society, 2015.

[33] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, pages 5005–5013. IEEE Computer Society, 2016.

[34] Shuhui Wang, Yangyu Chen, Junbao Zhuo, Qingming Huang, and Qi Tian. Joint global and co-attentive representation learning for image-sentence retrieval. In *ACM Multimedia*, pages 1398–1406, 2018.

[35] Qi Wu, Chunhua Shen, Peng Wang, Anthony R. Dick, and Anton van den Hengel. Image captioning and visual question answering based on attributes and external knowledge. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1367–1381, 2018.

[36] Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. Online asymmetric metric learning with multi-layer similarity aggregation for cross-modal retrieval. *IEEE Trans. Image Processing*, 28(9):4299–4312, 2019.

[37] Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *CVPR*, pages 5253–5262. IEEE Computer Society, 2017.

[38] Shijie Yang, Liang Li, Shuhui Wang, Dechao Meng, Qingming Huang, and Qi Tian. Srn: Structured stochastic recurrent network for linguistic video prediction. In *ACM Multimedia*, 2019.

[39] Shijie Yang, Liang Li, Shuhui Wang, Weigang Zhang, Qingming Huang, and Qi Tian. Skeletonnet: A hybrid network with a skeleton-embedding process for multi-view image representation learning. *IEEE Transactions on Multimedia*, 2019.

[40] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *ICCV*, pages 4904–4912. IEEE Computer Society, 2017.

[41] 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*, pages 1307–1315. IEEE Computer Society, 2018.

[42] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *ECCV (2)*, volume 9906 of *Lecture Notes in Computer Science*, pages 69–85. Springer, 2016.

[43] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L. Berg. A joint speaker-listener-reinforcer model for referring expressions. In *CVPR*, pages 3521–3529. IEEE Computer Society, 2017.

[44] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding referring expressions in images by variational context. In *CVPR*, pages 4158–4166. IEEE Computer Society, 2018.

[45] Fang Zhao, Jianshu Li, Jian Zhao, and Jiashi Feng. Weakly supervised phrase localization with multi-scale anchored transformer network. In *CVPR*, pages 5696–5705. IEEE Computer Society, 2018.

[46] C. Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *ECCV (5)*, volume 8693 of *Lecture Notes in Computer Science*, pages 391–405. Springer, 2014.