Relationship-Embedded Representation Learning for Grounding Referring Expressions

Sibei Yang, Guanbin Li, Yizhou Yu

Abstract—Grounding referring expressions in images aims to locate the object instance in an image described by a referring expression. It involves a joint understanding of natural language and image content and is essential for a range of visual tasks related to human-computer interaction. As a language-to-vision matching task, the core of this problem is to not only extract all the necessary information (i.e., objects and the relationships among them) in both the image and referring expressions, but also to to make full use of context information to achieve alignment of cross-modal semantic concepts in the extracted information. Unfortunately, existing work on grounding referring expressions fails to accurately extract multi-order relationships from the referring expressions and associate it with the object and its related context in the image. In this paper, we propose a Cross-Modal Relationship Extractor (CMRE) to adaptively highlight objects and relationships (spatial and semantic relations) related to the given expression, with a cross-modal attention mechanism, and represent the extracted information as language-guided visual relation graphs. In addition, we propose a Gated Graph Convolutional Network (GGCN) to compute multimodal semantic context by fusing information from different modes and propagating multimodal information in the structured relation graphs. Experimental results on three common benchmark datasets show that our Cross-Modal Relationship Inference Network, which consists of CMRE and GGCN, greatly surpass all existing state-of-the-art methods.

Index Terms—Referring Expressions, Cross-Modal Relationship Extractor, Gated Graph Convolutional Network.

1 Introduction

A Fundamental capability of AI for bridging humans and machines in the physical world is comprehending natural language utterances and their relationship with visual information. This capability is required by many challenging tasks, among which, grounding referring expressions [1], [2] is an essential one. The task of grounding referring expressions needs to locate the target visual object in an image by understanding the multimodal semantic concepts as well as relationships between referring natural language expressions (e.g. "the man with sun glasses", "the dog near a white car") and the image content.

Identifying an object proposal referred to by the expressions from a set of proposals in an image is a typical formulation of grounding referring expressions [3]. Recent methods adopt the combination of Convolutional Neural Networks (CNN) [4] and Long Short-Term Memory Neural Networks (LSTM) [5] to process the multimodal information in images and referring expressions. CNNs extract visual features of single objects, global visual contexts [2], [6] and pairwise visual differences [3], [7], [8], [9] while LSTMs encode global language contexts [2], [7], [8], [9], [10] and language features of decomposed phrases [3], [11], [12]. In addition, the cooperation between CNNs and LSTMs captures the context of object pairs [11], [12], [13]. However, existing work cannot extract all the required information (i.e. individual objects; first-order relationships or multi-order relationships) accurately from referring expressions and the captured contexts in such work also have discrepancies with the contexts described by referring expressions. In this paper, we refer to [12] and define the "context" as the objects as well as their attributes and relationships mentioned in the expression that help distinguish a referred object from other objects.

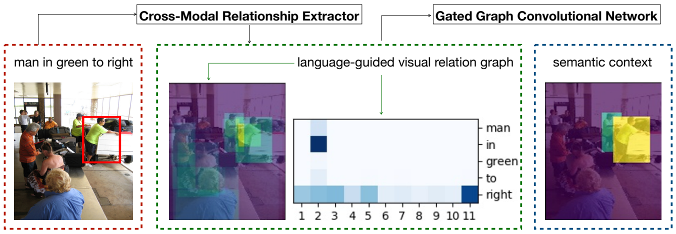

Cross-Modal Relationship Extractor
Gated Graph Convolutional Network
man in green to right
language-guided visual relation graph
semantic context
man
in
green
to
right
1 2 3 4 5 6 7 8 9 10 11

Fig. 1. Cross-Modal Relationship Inference Network. Given an expression and image, Cross-Modal Relationship Extractor constructs the language-guided visual relation graphs (spatial relation graph as an example, the attention scores of proposals and edges' types are visualized inside green dashed box). The Gated Graph Convolutional Network capture semantic context and computes the matching scores between context of proposals and context of expression (the matching scores of proposals are shown inside blue dashed box). Warmer color indicates higher scores of pixels and darker blue indicates higher scores of edges' types.

To solve the problem of grounding referring expressions, the accurate extraction of all required information (i.e. objects and the relationships among them in the image and referring expressions) is crucial for any given pair of expression and image. Because of the unpredictability and flexibility of an expression describing the scene in an image [2], the

This work was partially supported by the Hong Kong PhD Fellowship, the National Natural Science Foundation of China under Grant No.61976250 and No.61702565, the Science and Technology Program of Guangdong Province under Grant No. 2017B010116001 and the Fundamental Research Funds for the Central Universities under Grant No.18lgpy63. Corresponding authors are Guanbin Li and Yizhou Yu.

S. Yang and Y. Yu are with the Department of Computer Science, The University of Hong Kong, Hong Kong, and Y. Yu is also with Deepwise AI Lab (e-mail: sbyang9@hku.hk; yizhouy@acm.org).
G. Li is with the school of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China (e-mail: liguanbin@mail.sysu.edu.cn).

proposed model needs to extract the information adaptively. For example, if "The man holding a red balloon" is located in an image with two or more men, the nouns/noun phrases ("man" and "red balloon") and the relation word "holding" need to be extracted from the natural language expression; meanwhile, proposals for "man" and "red balloon" and the visual relationship ('holding") linking them together should be identified in the image. "The parking meter on the left of the man holding a red balloon" is a more complicated example, which involves an additional object "parking meter" and additional relational information "left". In this example, on one hand, there are three individual objects (i.e. "man", "red balloon" and "parking meter"), that need to be recognized in both the image and expression. Object proposals in the image can be either obtained with an object detector [14] or provided as part of the dataset [2], [8]. Nouns and noun phrases in the expression need to be extracted, and words in the same phrase should refer to the same object. Unfortunately, existing methods only consider individual words and softly parsed phrases [3], [11], [12], but words in the same softly parsed phrase cannot be constrained to the same object. On the other hand, the second-order relationship between the target and the "red balloon" via the "man" need to be inferred from either the detected direct semantic relationship "holding" or spatial relationship between object pairs "on the left of". Unfortunately, existing work either does not support relationship modeling or only considers first-order relationships among objects [11], [12], [13]. Theoretically, visual relation detectors [15], [16], [17] and natural language parsers can help achieve that goal by detecting relational information in the image and parsing grammatical relations among the words in the expression. However, existing visual relation detectors which focus only on the extraction of semantic relationship cannot deliver satisfactory and sufficient clues for highly unrestricted scene compositions [12], and existing language parsers have adverse effects on performance for grounding referring expressions due to their parsing errors [3], [12].

Moreover, the target object is distinguished from other objects on the basis of their contexts and the context of the expression [8], [12], [13]; therefore, accurate and consistent representation of contextual information in the referring expression and object proposals is essential. Nevertheless, existing methods for context modeling either cannot represent the context accurately or cannot achieve high-level consistency between both types of context mentioned above, and the reasons are given below. First, noisy information introduced by existing work on global language context modeling [2], [7], [8], [9], [10] and global visual context modeling [2], [6] makes it hard to align and match these two types of contexts. Second, pairwise visual differences computed in existing work [3], [7], [8], [9] can only represent instance-level visual differences among objects of the same category. Third, existing work on context modeling for object pairs [11], [12], [13] only considers first-order relationships instead of multi-order relationships (e.g., they directly extract the relationship between the pairs of (target, "man") and (target, "balloon") without considering the "man" is "holding the balloon" when extracting the relationship between the target "parking meter" and "the man"). In addition, multi-order relationships are actually structured information, which cannot be modeled by the context encoders adopted by existing work on grounding referring expressions.

Given the limitations of existing methods, our proposed end-to-end Cross-Modal Relationship Inference Network (CMRIN) aims to overcome the aforementioned difficulties. CMRIN consists of two modules, i.e., the Cross-Modal Relationship Extractor (CMRE) and the Gated Graph Convolutional Network (GGCN). An example is illustrated in Fig. 1. The CMRE extracts all the required information adaptively (i.e., nouns/noun phrases and relationship words from the expressions, and object proposals and their visual relationships from the image) for constructing a language-guided visual relation graph with cross-modal attention. First, CMRE constructs two scene graphs (a spatial relation graph as well as a semantic relation graph) for the image. Second, it extracts noun phrases in the expression using a constituency tree, meanwhile, it learns to classify the words in expression into four types and further assign the words/phrases to the vertices and edges in each scene graph. Finally, it constructs the language-guided visual relation graphs from the normalized attention distribution of words/phrases over vertices and edges of each scene graph. The GGCN fuses information from different modes and propagates the fused information in the language-guided visual relation graph to obtain semantic contexts referring to the expression by performing the following two steps. First, it fuses the contexts in the expression into the visual relation graph to form a multimodal relation graph, which includes the spatial/semantic relationships, visual information and language contexts; Second, gated graph convolutional operations are applied to the multimodal relation graph to obtain the semantic contexts. We have tested our proposed CMRIN on three common benchmark datasets, including RefCOCO [8], RefCOCO+ [8] and RefCOCOg [2], for grounding referring expressions. Experimental results show that our proposed network outperforms all other state-of-the-art methods.

In summary, this paper has the following contributions:

- Cross-Modal Relationship Extractor (CMRE) is proposed to convert the pair of input expression and image into a language-guided visual relation graph. For any given pair of expression and image, CMRE highlights objects as well as spatial and semantic relationships among them with a cross-modal attention mechanism by considering the words and phrases in the expression as guidance.
- Gated Graph Convolutional Network (GGCN) is proposed to capture multimodal semantic context with multi-order relationships. GGCN fuses information from different modes and propagates fused information in the language-guided visual relation graph.
- CMRE and GGCN are integrated into Cross-Modal Relationship Inference Network (CMRIN), which outperforms all existing state-of-the-art methods on grounding referring expressions using the ground-truth proposals. In addition, CMRIN shows its robustness using the detected proposals.

This paper is an extended version of [18], it provides a more complete introduction and analysis to the proposed cross-modal relationship inference network for referring expression comprehension, providing additional insights and relevant research introduction, effectiveness of framework components verification, network parameters analysis and more elaborated experimental comparison. Furthermore, we propose to add phrase parsing in the expression and apply it to enhance the representation of language-guided visual relation graphs, which helps to better align between the linguistic words and visual objects. Second, to complement spatial relation graph, we have also extracted semantic relations and use it as another guidance in edge gate computation for multi-order relationships inference in our proposed GGCN. Experimental results show that by introducing phrase decomposition on referring expression and semantic relationship modeling on images, it can bring different levels of performance improvement and make the algorithm more complete and robust.

2 Related Work

2.1 Grounding Referring Expressions

Grounding referring expression and referring expression generation [2] are dual tasks. The latter is to generate an unambiguous text expression for a target object in an image, and the former selects the corresponding object according to the image content referred by a text expression.

To address the problem of grounding referring expression, some previous work [2], [7], [8], [9], [10], [19] extracts visual object features from CNN and treat an expression as a whole to encode language feature through an LSTM. Among them, some methods [2], [8], [10] learn to maximize the posterior probability of the target object given the expression and the image, and the others [7], [9] model the joint probability of the target object and the expression directly. Specifically, MMI [2] applies the same CNN-LSTM network architecture for grounding referring expressions and referring expression generation respectively, and jointly optimize those two parts together. Speaker [8] improves MMI by taking more consideration between the comparisons on objects of the same type in the image, and it encodes visual appearance differences and relative spatial (i.e. location and size) differences between object and surrounding objects of the same object category. Speaker-Listener-Reinforcer [9] proposes an Reinforcer module to sample more discriminative expressions for helping the training of the Speaker. Attr [7] suggests that the attributes of objects help to distinguish the target object from other ones, and it learns the attributes of objects and encodes the features from the learned attributes and visual features. A-ATT [19] adopts joint attention mechanism on query, image and objects multiply round to obtain the communication among the three different types of information. However, all of the above methods independently encode the images and expressions without considering the interactions between them, and the learned monolithic representations in the two modes are not practical to the semantic-rich visual scenes and complex expressions.

Different from the methods above, Neg Bag [13] proposes to feed the concatenation of visual object representation, visual context representation and the word embedding to an LSTM model. Recent methods [3], [11], [12] learn to decompose an expression into different components and compute the language-vision matching scores of each module for objects. Specially, CMN [11] learns to parse the expression into a fixed form of subject-object-relationship; MAttNet [3] decomposes the expression into subject, location and relationship modules, and the module weights are computed for combining those three modules. VC [12] obtains the context-cue language features and referent-cue language features for both single objects and pairwise objects. However, all of the existing works are based on simple expression decomposition and match directly with the detected object features and the additionally computed relationship features [15], [16], [17], without considering the cross-modal alignment of multi-order relationship among objects and attributes, they are therefore arduous to adapt to the referring of objects in highly unrestricted scenes. Our Cross-Modal Relationship Extractor also learns to parse the expression, but we treat the parsed words as the guidance to highlight all the objects and their relationships described in the expression automatically to build the language-guided visual relation graphs which are further enhanced by a tailor-designed gated graph neural network for cross-modal multi-order context reasoning and alignment.

2.2 Context Modeling

Context modeling has been applied in many visual recognition tasks, e.g., object detection [20], [21] and semantic segmentation [22], [23]. For example, ION [20] uses four directional Recurrent Neural Networks (RNNs) to compute the context features on feature maps from four spatial directions. Ren *et al.* [21] propose Recurrent Rolling Convolution architecture to gradually aggregate context among the feature maps with different resolutions. Context Encoding Module [23] encodes the global semantic context by learning an inherent codebook which is a set of visual centers. Recently, Structure Inference Network [24] formulates the context modeling task as a graph structure inference problem [25], [26], [27], and it obtains the scene context by applying RNN to proposals in image.

As contextual information helps to distinguish the target from other objects, previous work on grounding referring expressions has also attempted to captured the context. For example, some early works [2], [6] propose to encode the entire image as a visual context, but that global contextual information usually cannot accurately match with the local context described by the expression. Other works [3], [7], [8], [9] capture the visual difference between the objects belonging to the same category in an image, but the visual difference of the object's appearance is often insufficient to distinguish the target from other objects. In fact, the visual difference between the context including appearance and relationship is essential, e.g., "Man holding a balloon", the necessary information to locate the "man" is not only the appearance of the "man" but the "holding" relation with the "balloon". There are also some works [11], [12], [13] which model the context from the context of object pairs, but they only consider the context with the first-order relationship between the objects. Inspired by Graph

Convolutional Network [26] for classification, our Gated Graph Convolutional Network flexibly capture the context referring to the expression by message passing, and the context with multi-order relationships can be captured.

2.3 Vision-Language

The combination of language and vision has been extensively studied in the last few years due to its significance for building AI systems. Besides grounding referring expressions, image/video captioning [28], [29] and visual question answering [30] are two popular and fundamental tasks.

Image caption is to generate image-relevant textual descriptions for given images. Early approaches [28], [31] extract visual concepts (i.e., objects and attributes) from images and format the sentences from those visual concepts and templates. Recently, some work [32], [33] starts to encode the image as visual representations (e.g., single visual representation for the whole image [32] and a set of visual representations for different sub-regions of the image [33]) by applying CNN, and then decode the visual representations into language descriptions through LSTM. The attention mechanism is adopted to attend the most relevant part of visual information [33] of it with the already generated text [34] in every time step of LSTM. Some of the recent approaches [35], [36] use the additional information (e.g. phrases and semantic words) extracted from image or text to help generate high quality sentences. There are also some works which focus on the description of a specified object in an image, a.k.a referring expression generation [2], which is a dual problem of visual grounding. It can be applied as an auxiliary component to referring expression comprehension to enhance the performance of cross-modal matching by computing the semantic distance between the generated statement and the given expression [2]. However, as the description of the object is varied and involves complex context information and relationships with other objects, the performance improvement for referring expression comprehension of unrestricted complex scenes is limited.

Visual question answering is to correctly infer the answer for a given pair of image and textual question. Most of the existing work [37], [38], [39], [40], [41], [42], [43] extracts the visual features from the image through CNN and encodes the question to language representation by passing the question into LSTM. And then, the answer is predicted by cooperation between those two types of representations. The cooperation is implemented by approaches, like learning a common embedding space for visual and language representations [37], [38], [39], or attending the most discriminate regions of image by applying different attention mechanisms on both representations [40], [41], [42], [43], or using both of them together. Besides the direct prediction of the answers, interpreting the reasoning procedure is important as well. The reasoning procedures are modeled from three different perspectives (i.e. relation-based modeling [44], [45], attention-based modeling [39], [46] and module-based modeling [47], [48]). Although visual question answering and referring expression comprehension have different problem definitions and solving goals, visual grounding is the key to endowing VQA with interpretability, which helps to ground their answers to relevant regions in the image [48], [49]. On the other hand, cross-modal feature fusion and semantics reasoning are equally effective and important for both issues. The study of the two problems can be integrated and learned from each other [39], [50].

2.4 Graph Neural Networks

Graph Neural Networks (GNNs) which are widely used to model the relational dependencies among elements of a graph through message passing [26], [51], [52], have been successfully applied to various context-aware visual tasks, e.g., semi-supervised classification [26], zero-shot recognition [53] and object detection [24].

Graph-structured representations and GNNs have also been introduced to the tasks of language and vision understanding. The methods in [54], [55], [56], [57] for VQA and image captioning represent an image as a graph structure where the vertices represent visual regions of an image and the edges are relationships among them, and then capture the visual context of each region node by GNN propagation. Specifically, [54] and [57] encode the contextual features at vertices by using the graph networks based on the recurrent unit [58] and the graph convolutional network (GCN) respectively. Their graph networks operate in the modes of vision and language independently. Different with them, our graph network performs on the top of multi-modal graph to learn the language-guided contexts at vertices. The recent works, [55] and [56], also obtain the convolved graph representations over the language-conditioned graphs: the former identifies neighbors for a vertex as its K most similar vertices and update feature at the vertex as sum of the learned features of its neighbors weighted by the learned weighting factors in each convolution layer, and the latter considers relationships between any pairs of vertices and aggregates the relational features for vertices using max pooling operator. Different with the above methods, we define gates of vertices and edges to implement the different influences of neighbors and relationships, and the gates are learned globally. To the best of our knowledge, we are the first to incorporate the graph convolutional networks in referring expressions comprehension for multi-order relationships representation learning.

3 Cross-Modal Relationship Inference Network

Our proposed Cross-Modal Relationship Inference Network (CMRIN) relies on relationships among objects and context captured in the multimodal relation graph to choose the target object proposal in the input image referred to by the input expression. First, CMRIN constructs a language-guided visual relation graph using the Cross-Modal Relationship Extractor. Second, it captures multimodal context from the relation graph based on the Gated Graph Convolutional Network. Finally, a matching score is computed for each object proposal according to its multimodal context and the context of the input expression. The overall architecture of our CMRIN for grounding referring expressions is illustrated in Fig. 2. In the rest of this section, we elaborate all the modules in this network.

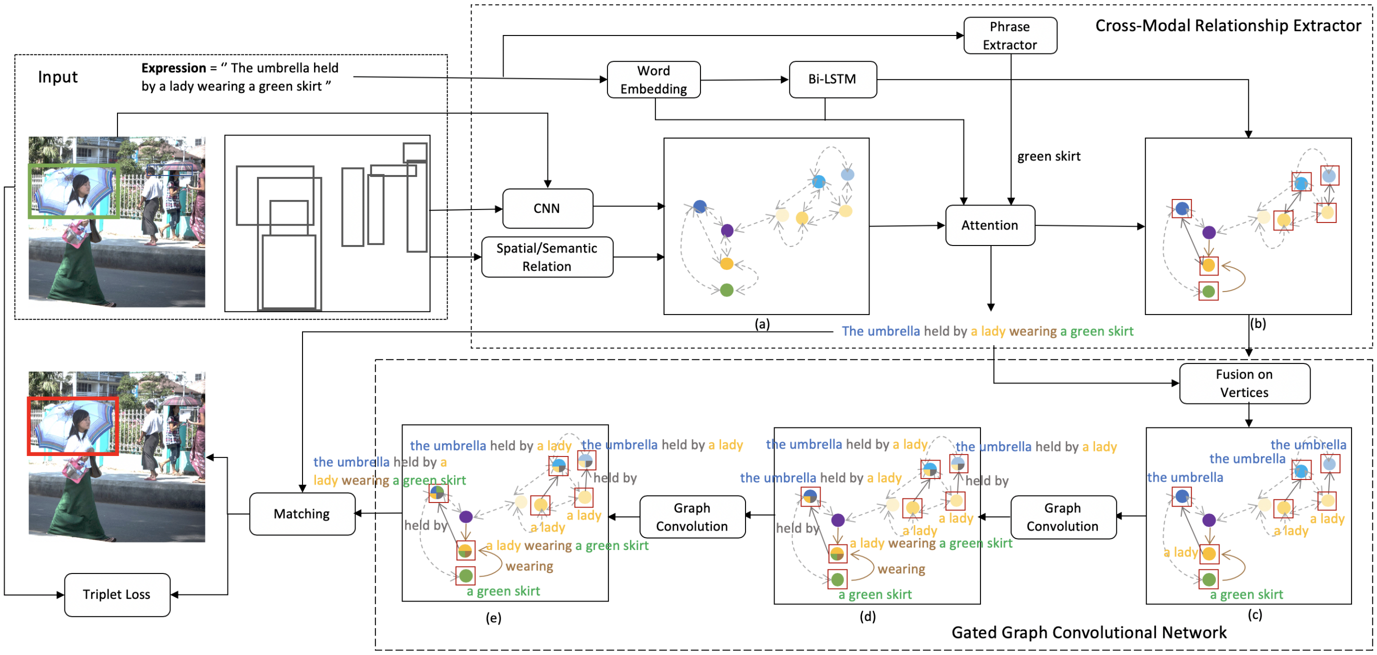

Cross-Modal Relationship Extractor
Input
Expression = " The umbrella held by a lady wearing a green skirt "
Phrase Extractor
Word Embedding
Bi-LSTM
CNN
Spatial/Semantic Relation
green skirt
Attention
(a)
(b)
The umbrella held by a lady wearing a green skirt
Fusion on Vertices
the umbrella
the umbrella held by a lady
held by
a lady
a lady wearing a green skirt
wearing
a green skirt
Graph Convolution
Graph Convolution
the umbrella held by a lady wearing a green skirt
Matching
Triplet Loss
(c)
(d)
(e)
Gated Graph Convolutional Network

Fig. 2. An overview of our Cross-Modal Relationship Inference Network for grounding referring expressions (better view in color). We use color to represent semantics, i.e. yellow denotes “person”, green denotes “green shirt”, blue denotes “umbrella”, purple means “white T-shirt”, brown means “wearing” and dark grey refers to “held by”. It includes a Cross-Modal Relationship Extractor (CMRE) and a Gated Graph Convolutional Network (GGCN). First, CMRE constructs (a) a spatial relation graph from the visual features of object proposals and spatial relationships between proposals. Second, CMRE parses the expression into a constituency tree and extracts the valid noun phrases. Third, CMRE highlights the vertices (red bounding boxes) and edges (solid lines) to generate (b) a language-guided visual relation graph using cross-modal attention between words/phrases in the referring expression and the spatial relation graph’s vertices and edges. Fourth, GGCN fuses the context of words into the language-guided visual relation graph to obtain (c) a multimodal (language, visual and spatial information) relation graph. Fifth, GGCN captures (d) the multimodal semantic context with first-order relationships by performing gated graph convolutional operations in the relation graph. By performing gated graph convolutional operations multiple iterations, (e) semantic context with multi-order relationships can be computed. Finally, CMRIN calculates the matching scores between semantic context of proposals and the global context of the referring expression. The triplet loss with online hard negative mining is adopted during training and the proposal with the highest matching score is chosen.

3.1 Cross-Modal Relationship Extractor

The Cross-Modal Relationship Extractor (CMRE) adaptively constructs the language-guided visual relation graph according to each given pair of image and expression using a cross-modal attention mechanism. Our CMRE considers both the word level and the phrase level. At the word level, it softly classify the words in the expression into four types (i.e., entity, relation, absolute location, and unnecessary words) according to the context of the words. At the phrase level, it extracts noun phrases, which are directly taken as entity phrases. Meanwhile, the context of the entire expression can be computed from the context of each individual word. In addition, a spatial relation graph of the image is constructed by linking object proposals in the image according to their size and locations and a semantic relation graph is constructed by an off-the-shelf object relationship detector [59]. Next, CMRE generates the language-guided visual relation graph by highlighting the vertices and edges of the relation graphs. Highlighting is implemented as computing cross-modal attention between the words/phrases in the expression and the vertices and edges in the relation graphs.

3.1.1 Relation Graph Construction

Exploring spatial relations and semantic relations among object proposals within an image is necessary for grounding referring expressions, because they are frequently occurs in referring expressions. Thus, we construct two different graphs by exploring two different types of relationships, i.e., spatial relation graph and semantic relation graph.

For spatial relation graph, we obtain the spatial relationship between each pair of object proposals according to their size and locations, which bears resemblance to the approach in [60]. For a given image I with K object proposals (bounding boxes), $O = \{o_i\}_{i=1}^{K}$, the location of each proposal o_i is denoted as $loc_i = (x_i, y_i, w_i, h_i)$, where (x_i, y_i) are the normalized coordinates of the center of proposal o_i, and w_i and h_i are the normalized width and height. The spatial feature $\mathbf{p}_i$ is defined as $\mathbf{p}_i = [x_i, y_i, w_i, h_i, w_i h_i]$. For any pair of proposals o_i and o_j, the spatial relationship r_{ij} between them is defined as follows. We compute the relative distance d_{ij}, relative angle θ_{ij} (i.e. the angle between the horizontal axis and vector $(x_i - x_j, y_i - y_j)$) and Intersection over Union u_{ij} between them. If o_i includes o_j, r_{ij} is set to “inside”; if o_i is covered by o_j, r_{ij} is set to “cover”; if none of the above two cases is true and u_{ij} is larger than 0.5, r_{ij} is set to “overlap”; otherwise, when the ratio between d_{ij} and the diagonal length of the image is larger than 0.5, r_{ij} is set to “no relationship”. In the rest of the cases, r_{ij} is assigned to one of the following spatial relationships, “right”, “top right”, “top”, “top left”, “left”, “bottom left”, “bottom” and “bottom right”, according to the relative angle

θ_{ij}. The details are shown in Fig. 3.

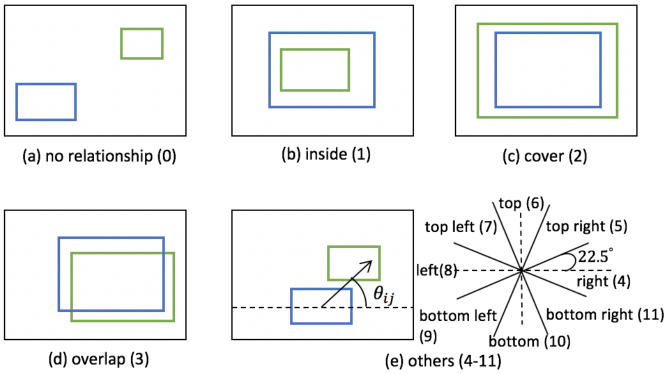


Fig. 3. All types of spatial relationships between proposal o_i (green box) and proposal o_j (blue box). The number following the relationship is the label.

The directed spatial relation graph $G^s = (V, E, \mathbf{X}^s)$ is constructed from the set of object proposals O and the set of pairwise relationships $R = \{r_{ij}\}_{i,j=1}^{K}$, where $V = \{v_i\}_{i=1}^{K}$ is the set of vertices and vertex v_i corresponds to proposal o_i; $E = \{e_{ij}\}_{i,j=1}^{K}$ is the set of edges and e_{ij} is the index label of relationship r_{ij}; $\mathbf{X}^s = \{\mathbf{x}_i^s\}_{i=1}^{K}$ is the set of features at vertices and $\mathbf{x}_i^s \in \mathbb{R}^{D_x}$ is the visual feature of proposal o_i, where D_x is the dimension of visual feature. $\mathbf{x}_i^s$ is extracted using a pretrained CNN model. A valid index label of E ranges from 1 to $N_e = 11$ (the label of "no relationship" is 0).

Similar to the spatial relation graph $G^s = (V, E, \mathbf{X}^s)$, the semantic relation graph $G^{sem} = (V, \mathbf{E}^{sem}, \mathbf{X}^s)$ shares the same sets of vertices and features at vertices as G^s, but instead the set of edges $\mathbf{E}^{sem}$ is extracted by a pretrained object relationship detector [59].

The spatial relation graph G^s and semantic relation graph G^{sem}, which are constructed from the image, involves the visual features of the proposals as well as their spatial relationships or semantic relationships. They are further transformed into the language-guided visual relation graph based on the guidance from the expression, which will be detailed in Section 3.1.4. To simplify the description and focus on the pipeline design of the proposed method, we adopt the G^s as the example for remaining part in Section 3. And the detail implementation for the semantic branch will be described in Section 4.4.5.

3.1.2 Phrase

Parsing the phrases in the expression is paramount as it helps to accurately highlight the vertices of graph G^s referred to by the expression. For example, if the expression is "the umbrella held by a lady wearing a green skirt", it is necessary to recognize the noun phrase (i.e. "green skirt"), and the words in this phrase refer to the same vertex. We only extract noun phrases in this paper since they are most relevant to the objects in the image. Specifically, Our CMRE follows the three steps below to extract the noun phrases. First, it parses the given expression into a constituency tree, which breaks the expression into sub-phrases. Second, it locates candidate noun phrases (i.e. "the umbrella", "a lady" and "a green skirt") from the leaves to the root. On each path from the leaves to the root, it extracts the first noun phrase and ignores the other noun phrases. Third, it eliminates determiners and words indicating absolute location in the extracted noun phrase candidates. A candidate phrase is valid if the number of remaining words is at least two. Thus, "green shirt" is a valid noun phrase. For a given expression $L = \{l_t\}_{t=1}^{T}$ (T is the number of words), we denote the set of extracted noun phrases as $Q = \{q_m\}_{m=1}^{M}$, where M is the number of phrases.

3.1.3 Language Representation

Inspired by the attention weighted sum of word vectors over different modules in [3], [11], [12], our CMRE defines attention distributions of words/phrases over the vertices and edges of the spatial relation graph G^s. In addition, different words in a referring expression may play different roles. For referring expressions, words can usually be classified into four types (i.e entity, relation, absolute location and unnecessary words), and the type for noun phrases is entity. By parsing the expression into different types of words and distributing words/phrases over the vertices and edges of graph G^s, the language embedding of every vertex and edge can be captured, and the global language context can also be obtained.

Given an expression $L = \{l_t\}_{t=1}^{T}$, CMRE first learns a D_f-dimensional embedding for each word, $\mathbf{F}^l = \{\mathbf{f}_t^l \in R^{D_f}\}_{t=1}^{T}$, and then applies a bi-directional LSTM [61] to encode the context of words. The context of word l_t is the concatenation of its forward and backward hidden vectors, denoted as $\mathbf{h}_t^l \in \mathbb{R}^{D_h}$. The weight $\mathbf{m}_t$ of each type (i.e. entity, relation, absolute location and unnecessary word) for word l_t is defined as follows.

$$\mathbf{m}_t = \text{softmax}(\mathbf{W}_{l1}\sigma(\mathbf{W}_{l0}\mathbf{h}_t^l + \mathbf{b}_{l0}) + \mathbf{b}_{l1}), \tag{1}$$

where $\mathbf{W}_{l0} \in \mathbb{R}^{D_{l0} \times D_h}$, $\mathbf{b}_{l0} \in \mathbb{R}^{D_{l0} \times 1}$, $\mathbf{W}_{l1} \in \mathbb{R}^{4 \times D_{l0}}$ and $\mathbf{b}_{l1} \in \mathbb{R}^{4 \times 1}$ are learnable parameters, D_{l0} and D_h are hyper-parameters and σ is the activation function. The feature vector of a phrase is computed as the mean embedding feature (context) of words appearing in the phrase. The set of features for all phrases in the expression is denoted as $\mathbf{F}^q = \{\mathbf{f}_m^q\}_{m=1}^{M}$ (contextual embeddings $\mathbf{H}^q = \{\mathbf{h}_m^q \in \mathbb{R}^{D_h}\}_{m=1}^{M}$).

Next, CMRIN computes the language context of every vertex in graph G^s from both words and phrases. When words are considered, on the basis of the word embedding $\mathbf{F}^l = \{\mathbf{f}_t^l\}_{t=1}^{T}$ and the entity weights of words $\{\mathbf{m}_t^{(0)}\}_{t=1}^{T}$, a weighted normalized attention distribution over the vertices of graph G^s is defined as follows.

$$\begin{aligned}\alpha_{t,i}^l &= \mathbf{W}_n^l[\tanh(\mathbf{W}_v^l\mathbf{x}_i^s + \mathbf{W}_f^l\mathbf{f}_t^l)],\\ \lambda_{t,i}^l &= \mathbf{m}_t^{(0)}\frac{\exp(\alpha_{t,i}^l)}{\sum_i^K \exp(\alpha_{t,i}^l)},\end{aligned} \tag{2}$$

where $\mathbf{W}_n^l \in \mathbb{R}^{1 \times D_n}$, $\mathbf{W}_v^l \in \mathbb{R}^{D_n \times D_x}$ and $\mathbf{W}_f^l \in \mathbb{R}^{D_n \times D_h}$ are transformation matrices and D_n is hyper-parameter. $\lambda_{t,i}$ is the weighted normalized attention, indicating the probability that word l_t refers to vertex v_i. Likewise, CMRIN computes $\alpha_{m,i}^q$ for the phrases $Q = \{q\}_i^M$ on the basis of their features $\mathbf{F}^q = \{\mathbf{f}^q_m\}_{m=1}^M$, and the normalized distribution over the vertices are computed as follows.

$$\lambda^q_{m,i} = \frac{\exp(\alpha^q_{m,i})}{\sum_i^K \exp(\alpha^q_{m,i})}, \tag{3}$$

The language context $\mathbf{c}_i$ at vertex v_i is computed by aggregating all attention weighted word contexts and phrase contexts.

$$\mathbf{h}_i = \frac{\sum_{t=1}^T \lambda^l_{t,i}\mathbf{h}^l_t + \sum_{m=1}^M \lambda^q_{m,i}\mathbf{h}^q_m}{\sum_{t=1}^T \lambda^l_{t,i} + \sum_{m=1}^M \lambda^q_{m,i}} \tag{4}$$

Then, the global language context $\mathbf{h}_g$ of graph G^s is calculated as follows.

$$\mathbf{h}_g = \sum_{t=0}^T (\mathbf{m}^{(0)}_t + \mathbf{m}^{(1)}_t + \mathbf{m}^{(2)}_t)\mathbf{h}^l_t \tag{5}$$

where the entity weight, relation weight and absolute location weight are the first three elements of $\mathbf{m}_t$. CMRIN computes the global context only from word contexts because phrases are only used for improving the accuracy of vertex highlighting in the relation graphs.

3.1.4 Language-Guided Visual Relation Graph

Different object proposals and different relationships between proposals do not have equal contributions in solving grounding referring expressions. The proposals and relationships mentioned in the referring expression should be given more attention. Our CMRE highlights the vertices and edges of the spatial relation graph G^s, that have connections with the referring expression, to generate the language-guided visual relation graph G^v. The highlighting operation is implemented by designing a gate for each vertex and edge in graph G^s.

The gate p^v_i for vertex v_i is defined as the sum over the weighted probabilities that individual words and phrases in the expression refer to vertex v_i,

$$p^v_i = \sum_{t=1}^T \lambda^l_{t,i} + \sum_{m=1}^M \lambda^q_{m,i} \tag{6}$$

Each edge has its own type and the gates for edges are formulated as the gates for edges' types. The weighted normalized distribution of words over the edges of graph G^s is defined as follows.

$$\mathbf{w}^e_t = \text{softmax}(\mathbf{W}_{e1}\sigma(\mathbf{W}_{e0}\mathbf{h}^l_t + \mathbf{b}_{e0}) + \mathbf{b}_{e1})\mathbf{m}^{(1)}_t, \tag{7}$$

where $\mathbf{W}_{e0} \in \mathbb{R}^{D_{e0}\times D_h}$, $\mathbf{b}_{e0} \in \mathbb{R}^{D_{e0}\times 1}$, $\mathbf{W}_{e1} \in \mathbb{R}^{N_e\times D_{e0}}$ and $\mathbf{b}_{e1} \in \mathbb{R}^{N_e\times 1}$ are learnable parameters, and D_{e0} is hyper-parameter. $w^e_{t,j}$ is the j-th element of $\mathbf{w}^e_t$, which is the weighted probability of word l_t referring to edge type j. And the gate p^e_j for edges with type $j \in \{1, 2, ..N^e\}$ is the sum over all the weighted probabilities that individual words in the expression refer to edge type j,

$$p^e_j = \sum_{t=1}^T w^e_{t,j}. \tag{8}$$

The language-guided visual relation graph is defined as $G^v = (V, E, \mathbf{X}, P^v, P^e)$, where $P^v = \{p^v_i\}_{i=1}^K$, and $P^e = \{p^e_j\}_{j=1}^{N_e}$.

3.2 Multimodal Context Modeling

Our proposed Gated Graph Convolutional Network (GGCN) further fuses the language context into the language-guided visual relation graph to generate multimodal relation graph G^m, and computes a multimodal semantic context for every vertex by performing gated graph convolutional operations on the graph G^m.

3.2.1 Language-Vision Feature

As suggested by visual relationships detection [15], [17], the spatial locations together with the appearance features of objects are the key indicators of visual relationship, and the categories of objects is highly predictive of relationship. Our GGCN fuses the language context of vertices into the language-guided visual relation graph G^v (G^v encodes the spatial relationships and appearance features of proposals) to generate multimodal relation graph G^m, which forms the basis for computing the semantic context of vertices.

We define feature $\mathbf{x}^m_i$ at vertex v_i in G^m to be the concatenation of the visual feature $\mathbf{x}^s_i$ at vertex v_i in the language-guided visual relation graph and the language context $\mathbf{h}_i$ at vertex v_i, i.e. $\mathbf{x}^m_i = [\mathbf{x}^s_i, \mathbf{h}_i]$. The multimodal graph is defined as $G^m = (V, E, \mathbf{X}^m, P^v, P^e)$, where $\mathbf{X}^m = \{\mathbf{x}^m_i\}_{i=1}^K$.

3.2.2 Semantic Context Modeling

Multi-order relationships may exist in referring expressions. We obtain semantic context representing multi-order relationships through message passing. On one hand, semantic features are obtained by learning to fuse the spatial relations, visual features and language features. On the other hand, context representing multi-order relationships is computed by propagating pairwise context in graph G^m.

Inspired by Graph Convolutional Network (GCN) for classification [26], [53], our GGCN adopts graph convolutional operations in multimodal relation graph G^m for computing semantic context. Different from GCN operating in unweighted graphs, GGCN operates in weighted directed graphs with extra gate operations. The n-th gated graph convolution operation at vertex v_i in graph $G^m = (V, E, \mathbf{X}^m, P^v, P^e)$ is defined as follows.

$$\begin{aligned}
\overrightarrow{\mathbf{x}_i}^{(n)} &= \sum_{e_{i,j}>0} p^e_{e_{i,j}}(\overrightarrow{\mathbf{W}}^{(n)}\hat{\mathbf{x}}^{(n-1)}_j p^v_j + \mathbf{b}^{(n)}_{e_{i,j}}), \\
\overleftarrow{\mathbf{x}_i}^{(n)} &= \sum_{e_{j,i}>0} p^e_{e_{j,i}}(\overleftarrow{\mathbf{W}}^{(n)}\hat{\mathbf{x}}^{(n-1)}_j p^v_j + \mathbf{b}^{(n)}_{e_{j,i}}), \\
\tilde{\mathbf{x}}^{(n)}_i &= \widetilde{\mathbf{W}}^{(n)}\hat{\mathbf{x}}^{(n-1)}_i + \widetilde{\mathbf{b}}^{(n)}, \\
\hat{\mathbf{x}}^{(n)}_i &= \sigma(\overrightarrow{\mathbf{x}_i}^{(n)} + \overleftarrow{\mathbf{x}_i}^{(n)} + \tilde{\mathbf{x}}^{(n)}_i),
\end{aligned} \tag{9}$$

where $\hat{\mathbf{x}}^{(0)}_i = \mathbf{x}^m_i$, $\overrightarrow{\mathbf{W}}^{(n)}, \overleftarrow{\mathbf{W}}^{(n)}, \widetilde{\mathbf{W}}^{(n)} \in \mathbb{R}^{D_e\times(D_x+D_h)}$ $\{\mathbf{b}^{(n)}_j\}_{j=1}^{N_e}, \widetilde{\mathbf{b}}^{(n)} \in \mathbb{R}^{D_e\times 1}$ are learnable parameters, and D_e is hyper-parameter. $\overrightarrow{\mathbf{x}_i}^{(n)}$ and $\overleftarrow{\mathbf{x}_i}^{(n)}$ are encoded features for out- and in- relationships respectively. $\tilde{\mathbf{x}}^{(n)}_i$ is the updated feature for itself. The final encoded feature $\hat{\mathbf{x}}^{(n)}_i$ is the sum of the above three features and σ is the activation function. By performing the gated graph convolution operation multiple iterations (N), semantic context representing multi-order relationships among vertices can be computed. Such semantic context are denoted as $\mathbf{X}^c = \{\mathbf{x}^c_i = \hat{\mathbf{x}}^{(N)}_i\}_{i=1}^K$.

Finally, for each vertex v_i, we concatenate its encoded spatial feature $\mathbf{p}_i$ mentioned before and its language-guided semantic context $\mathbf{x}_i^c$ to obtain the multimodal context $\mathbf{x}_i = [\mathbf{W}_p\mathbf{p}_i, \mathbf{x}_i^c]$, where $\mathbf{W}_p \in \mathbb{R}^{D_p \times 5}$ and D_p is hyper-parameter.

3.3 Loss Function

The matching score between proposal o_i and expression L is defined as follows,

$$s_i = \text{L2Norm}(\mathbf{W}_{s0}\mathbf{x}_i) \odot \text{L2Norm}(\mathbf{W}_{s1}\mathbf{h}_g), \tag{10}$$

where $\mathbf{W}_{s0} \in \mathbb{R}^{D_s \times (D_p + D_x)}$ and $\mathbf{W}_{s0} \in \mathbb{R}^{D_s \times D_h}$ are transformation matrices, and D_s is hyper-parameter.

Inspired by the deep metric learning algorithm for face recognition in [62], we adopt the triplet loss with online hard negative mining to train our CMRIN model. The triplet loss is defined as

$$loss = \max(s_{neg} + \Delta - s_{gt}, 0), \tag{11}$$

where s_{gt} and s_{neg} are the matching scores of the ground-truth proposal and the negative proposal respectively. The negative proposal is randomly chosen from the set of online hard negative proposals, $\{o_j | s_j + \Delta - s_{gt} > 0\}$, where Δ is the margin. During testing, we predict the target object by choosing the object proposal with the highest matching score.

4 Experiments

4.1 Datasets

We have evaluated our CMRIN on three commonly used benchmark datasets for referring expression comprehension (i.e., RefCOCO [8], RefCOCO+ [8] and RefCOCOg [2]).

In RefCOCO, there are 50,000 target objects, collected from 19,994 images in MSCOCO [63], and 142,210 referring expressions, collected from an interactive game interface [1]. RefCOCO is split into train, validation, test A, and test B, which has 120,624, 10,834, 5,657 and 5,095 expression-target pairs, respectively. Test A includes images of multiple people while test B contains images with multiple other objects.

RefCOCO+ has 49,856 target objects collected from 19,992 images in MSCOCO, and 141,564 expressions collected from an interactive game interface. Different from RefCOCO, RefCOCO+ does not contain descriptions of absolute location in the expressions. It is split into train, validation, test A, and test B, which has 120,191, 10,758, 5,726 and 4,889 expression-target pairs, respectively.

RefCOCOg includes 49,822 target objects from 25,799 images in MSCOCO, and 95,010 long referring expressions collected in a non-interactive setting. RefCOCOg [13] has 80,512, 4,896 and 9,602 expression-target pairs for training, validation, and testing, respectively.

4.2 Evaluation and Implementation

The Precision@1 metric (the fraction of correct predictions) is used for measuring the performance of a method for grounding referring expressions. A prediction is considered to be a true positive when the Intersection over Union between the ground-truth proposal and the top predicted proposal for a referring expression is larger than 0.5.

For a given dataset, we count the number of occurrences of each word in the training set. If a word appears more than five times, we add it to the vocabulary. Each word in the expression is initially an one-hot vector, which is further converted into a word embedding. We parse the expression into a constituency tree by Stanford CoreNLP toolkit [64]. Annotated regions of object instances are provided in RefCOCO, RefCOCO+ and RefCOCOg. The target objects in the three datasets belong to the 80 object categories in MSCOCO, but the referring expressions may include objects beyond the 80 categories. In order to make the scope of target objects consistent with referring expressions, it is necessary to recognize objects in expressions, even when they are not in the 80 categories.

Inspired by the Bottom-Up Attention Model in [65] for image captioning and visual question answering, we train ResNet-101 based Faster R-CNN [14], [66] over selected 1,460 object categories in the Visual Genome dataset [67], excluding the images in the training, validation and testing sets of RefCOCO, RefCOCO+ and RefCOCOg. We combine the detected objects and the ground-truth objects provided by MSCOCO to form the final set of objects in the images. We extract the visual features of objects as the 2,048-dimensional output from the pool5 layer of the ResNet-101 based Faster R-CNN model. Since some previous methods use VGG-16 as the feature extractor, we also extract the 4,096-dimensional output from the fc7 layer of VGG-16 for fair comparison. We set the mini-batch size to 64. The Adam optimizer [68] is adopted to update network parameters with the learning rate set to 0.0005 initially and reduced to 0.0001 after 5 epochs. Margin Δ is set to 0.1.

4.3 Comparison with the State of the Art

We compare the performance of our proposed CMRIN against the state-of-the-art methods, including MMI [2], Neg Bag [13], CG [10], Attr [7], CMN [11], Speaker [8], Listener [9], VC [12], A-ATT [19] and MAttNet [3].

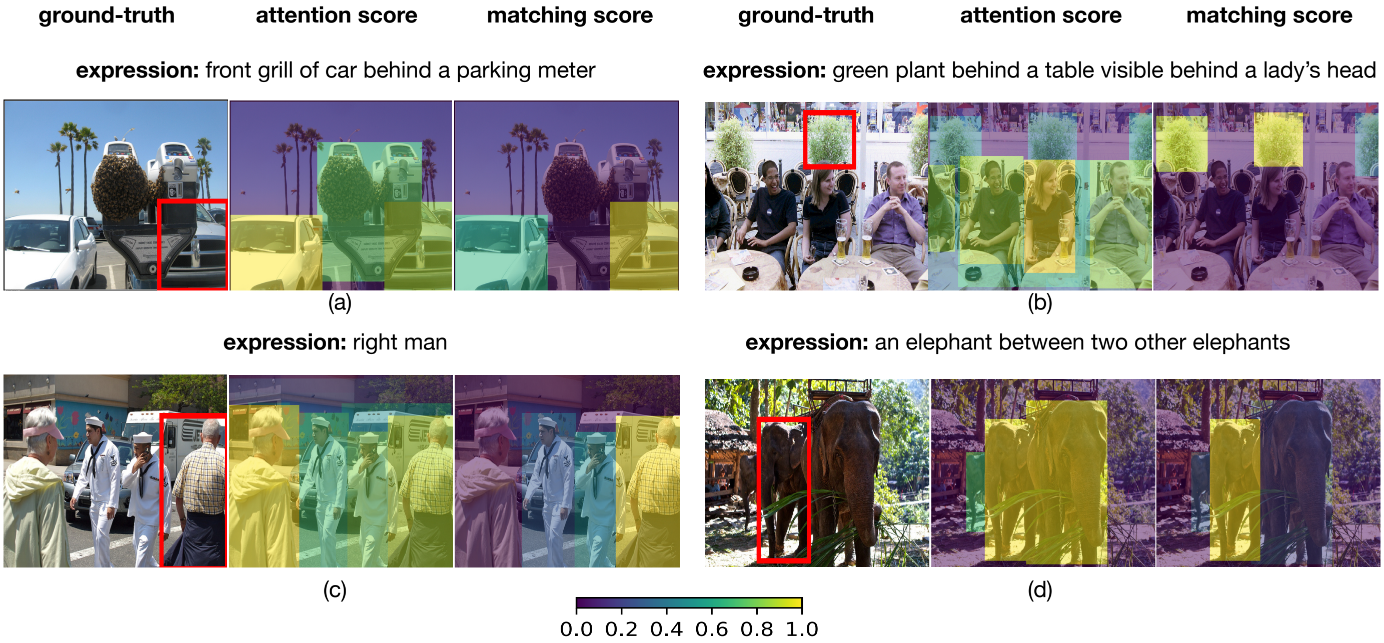
4.3.1 *Quantitative Evaluation*

Table 1 shows quantitative evaluation results on RefCOCO, RefCOCO+ and RefCOCOg datasets. Our proposed CMRIN consistently outperforms existing methods across all the datasets by a large margin, which indicates that our CMRIN performs well on datasets with different characteristics. Note that as semantic relationship computation introduces additional annotation data, in order to make a fair comparison with the state-of-the-art methods, we report the performance of CMRIN when using only spatial relation graph. We will explore the effectiveness of the further introduction of semantic relation graph during ablation study.

Specially, CMRIN improves the average Precision@1 over validation and testing sets achieved by the existing best-performing algorithm by 1.80%, 5.17% and 3.14% respectively on the RefCOCO, RefCOCO+ and RefCOCOg datasets when VGG-16 is used as the backbone network. Our CMRIN significantly improves the precision in the person category (test A of RefCOCO and RefCOCO+), which indicates that casting appearance attributes (e.g. shirt,

			RefCOCO			RefCOCO+			RefCOCOg	
		feature	val	testA	testB	val	testA	testB	val	test
1	MMI [2]	vgg16	-	71.72	71.09	-	58.42	51.23	-	-
2	Neg Bag [13]	vgg16	76.90	75.60	78.00	-	-	-	-	68.40
3	CG [10]	vgg16	-	74.04	73.43	-	60.26	55.03	-	-
4	Attr [7]	vgg16	-	78.85	78.07	-	61.47	57.22	-	-
5	CMN [11]	vgg16	-	75.94	79.57	-	59.29	59.34	-	-
6	Speaker [8]	vgg16	76.18	74.39	77.30	58.94	61.29	56.24	-	-
7	Listener [9]	vgg16	77.48	76.58	78.94	60.50	61.39	58.11	69.93	69.03
8	Speaker+Listener+Reinforcer [9]	vgg16	79.56	78.95	80.22	62.26	64.60	59.62	71.65	71.92
9	VC [12]	vgg16	-	78.98	82.39	-	62.56	62.90	-	-
10	A-ATT [19]	vgg16	81.27	81.17	80.01	65.56	68.76	60.63	-	-
11	MAttNet [3]	vgg16	80.94	79.99	82.30	63.07	65.04	61.77	73.04	72.79
12	Ours CMRIN	vgg16	83.57	83.97	82.69	71.57	75.60	65.56	75.65	76.45
13	MAttNet [3]	resnet101	85.65	85.26	84.57	71.01	75.13	66.17	78.10	78.12
14	Ours CMRIN	resnet101	86.43	87.36	85.52	75.63	80.37	**69.58**	79.72	80.85
15	Ours CMRIN + Semantic Branch	resnet101	**86.59**	**88.17**	**85.59**	**76.38**	**81.44**	68.81	**80.76**	**81.71**

TABLE 1
Comparison with the state-of-the-art methods on RefCOCO, RefCOCO+ and RefCOCOg. The two best performing methods using VGG-16 are marked in red and blue. The best performing method using ResNet is marked in bold.


ground-truth
attention score
matching score
ground-truth
attention score
matching score
expression: front grill of car behind a parking meter
expression: green plant behind a table visible behind a lady's head
expression: right man
expression: an elephant between two other elephants
0.0
0.2
0.4
0.6
0.8
1.0

(a) (b) (c) (d)

Fig. 4. Qualitative results showing initial attention score (gate) maps and final matching score maps. We compute the score of a pixel as the maximum score of proposals covering it, and normalize the score maps to 0 to 1. Warmer color indicates higher score.

glasses and shoes) of a person as external relationships between person and appearance attributes can effectively distinguish the target person from other persons. After we switch to the visual features extracted by ResNet-101 based Faster R-CNN, the Precision@1 of our CMRIN is further improved by another ∼4.40%. It improves the average Precision@1 over validation and testing sets achieved by MAttNet [3] by 1.62%, 5.03% and 3.13% respectively on the three datasets. Note that our CMRIN only uses the 2048-dimensional features from pool5 while MattNet uses the feature maps generated from the last convolutional layers of both the third and fourth stages.

4.3.2 Qualitative Evaluation

Visualizations of some samples along with their attention maps and matching scores are shown in Fig. 4. They are generated from our CMRIN using ResNet-101 based Faster R-CNN features.

Without relationship modeling, our CMRIN can identify the proposals appearing in the given expression (second columns), and it achieves this goal on the basis of mentioned objects in the given sentence (e.g. the parking meter in Fig. 4(a) and the elephant in full view in Fig. 4(d) have higher attention scores). After fusing information from different modes and propagating multimodal information in the structured relation graph, it is capable of learning semantic context and locating target proposals (third

		RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
1	global langcxt+vis instance	79.05	81.47	77.86	63.85	69.82	57.80	70.78	71.26
2	global langcxt+global viscxt(2)	82.61	83.22	82.36	67.75	73.21	63.06	74.29	75.23
3	weighted langcxt+guided viscxt(2)	86.02	86.21	84.51	73.59	78.62	68.01	77.14	78.29
4	weighted langcxt+guided viscxt(1)+fusion	85.89	87.27	84.61	74.28	79.24	69.16	79.41	79.38
5	weighted langcxt+guided viscxt(3)+fusion	86.20	87.24	84.91	75.26	80.06	69.52	79.55	80.55
6	weighted langcxt+guided viscxt(2)+fusion	**86.43**	**87.36**	**85.52**	**75.63**	**80.37**	**69.58**	**79.72**	**80.85**

TABLE 2
Ablation studies on variants of network architecture of our proposed CMRIN on RefCOCO, RefCOCO+ and RefCOCOg. The number following the "viscxt" refers to the number of gated graph convolutional layers used in the model.

		RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
1	softmax loss	85.43	86.81	84.47	74.23	79.74	67.78	78.94	79.37
2	triplet loss (0.2; 1; random hard)	86.25	87.50	84.99	75.15	80.11	68.91	79.96	80.49
3	triplet loss (0.5; 1; random hard)	85.84	86.87	84.69	74.52	79.99	68.32	79.04	79.97
4	triplet loss (0.1; 2; random hard)	86.32	87.54	84.97	75.12	80.75	69.13	79.39	81.30
5	triplet loss (0.1; 1; hardest)	86.38	87.68	85.06	74.32	79.69	68.30	79.68	80.38
6	triplet loss (0.1; 1; random semi-hard)	86.35	87.36	84.87	75.16	80.39	69.07	79.41	80.55
7	triplet loss (0.1; 1; random hard)	86.43	87.36	85.52	75.63	80.37	69.58	79.72	80.85

TABLE 3
Comparison of different schemes for training our proposed CMRIN on RefCOCO, RefCOCO+ and RefCOCOg. The contents in parentheses following the "triplet loss" represent the margin value, number of negative proposals and the sampling strategy respectively. The two best performing models are marked in red and blue. CMRIN is robust to different loss settings and consistently outperforms existing state-of-the-art models on all the three benchmark datasets.

columns) even when the target objects do not attract the most attention at the beginning. It is worth noting that our CMRIN learns semantic relationships ("behind") for pairs of proposals with different spatial relationships ("bottom right" between "car" and "parking meter" in Fig. 4(a); "top" between "green plant" and "lady's head" in Fig. 4(b)), which indicates that CMRIN is able to infer semantic relationships from the initial spatial relationships. In addition, CMRIN learns the context for target "elephant" (Fig. 4(d)) from "two other elephants" by considering the relations from multiple elephants together. Moreover, multi-order relationships are learned through propagation in CMRIN, e.g., the relationships ("right" in Fig. 4(c)) between object pairs are propagated gradually to the target proposal (most "right man").

Fig. 5 demonstrates more qualitative results from our proposed CMRIN. In order to better visualize the pixels covered by multiple proposals generated from the same object (e.g., the pixels covered by proposal "man" and proposal "shirt weared by the man"), we compute the score of a pixel in attention score maps as the sum of the scores of all covering proposals. And in order to distinguish different objects, we set the score of a pixel in matching score maps to be the maximum among scores of all covering objects. We exclude negative scores and normalize the range of each score map to $[0, 1]$.

4.4 Ablation Study

We evaluate the proposed CMRIN in five different aspects: 1) the effectiveness of the two modules in our proposed network architecture, i.e., CMRE and GGCN modules; 2) the impact of different training schemes on the performance of CMRIN; 3) the necessity of phrases; 4) the impact of variants of spatial relation graphs used in CMRIN; 5) we explore the effectiveness of incorporating semantic relation graph and detail its implementation. In the following experiments, features computed using ResNet-101 based Faster R-CNN are adopted.

4.4.1 Variants of Network Architecture

Our proposed CMRIN includes CMRE and GGCN modules. To demonstrate the effectiveness and necessity of each module and further compare each module against its variants, we have trained five additional models for comparison. The results are shown in Table 2.

As a baseline (row 1), we use the concatenation of instance-level visual features of objects and the location features as the visual features, and use the last hidden state of the LSTM based expression encoder as the language feature, and then compute the matching scores between the visual features and the language feature. In comparison, a simple variant (row 2) that relies on a global visual context, which is computed by applying graph convolutional operations to the spatial relation graph, already outperforms the baseline. This demonstrates the significance of visual context. Another variant (row 3) with a visual context computed in the language-guided visual relation graph outperforms the above two versions. It captures the context by considering cross-modal information. By fusing the context of words into the language-guided visual relation graph, the semantic context can be captured by applying gated graph convolutional operations (row 6, the final version of CMRIN). Finally, we explore the number of gated graph convolutional layers used in CMRIN. The 1-layer CMRIN (row 4) performs worse than the 2-layer CMRIN because it only captures context with first-order relationships. The 3-layer CMRIN (row 5) does not further improve the performance. One possible reason is that third-order relationships merely occur in the expressions.

4.4.2 Different Training Schemes

In this section, we evaluate the impact of different loss function designs on the performance of the proposed CMRIN. As shown in Table 3, CMRIN is robust with respect to different loss function settings (i.e., the softmax loss and triplet loss with different parameter settings) and consistently outperforms existing state-of-the-art models on all the three commonly used benchmark datasets (i.e. RefCOCO, RefCOCO+ and RefCOCOg).

Specifically, we compare different loss functions, hyperparameters and sampling strategies in the triplet loss for the training of CMRIN. We optimize the proposed CMRIN with the softmax loss (row 1), which is commonly adopted by existing works [8], [10], [11], [12]. The performance of CMRIN using the softmax loss performs worse than that of CMRIN using the triplet loss (row 7) because the matching score between the context of an expression and the context of a proposal is not always exactly zero or one. For example, in the image associated with expression "the umbrella held by a lady wearing a green skirt", there are three umbrellas held by three different ladies and only one of them wears a green skirt. The context of two umbrellas held by the ladies without wearing a green skirt partially matches the context ("the umbrella held by a lady") of the expression. In addition, we explore the effects of different margins (i.e. 0.1, 0.2 and 0.5) in the triplet loss. CMRIN trained using the triplet loss with a 0.5 margin achieves worse performance (row 3) than that with other margins (row 2 and 7) over all the three datasets. Moreover, the performance of CMRIN using the triplet loss with a 0.5 margin fluctuates during training. The models trained by the triplet loss with margin 0.1 and 0.2 have similar performance. Meanwhile, noting that sampling strategies for the triplet loss are essential in face recognition [62], [69], we also sample triplets using different online sampling strategies, including random sampling with one hard negative proposal (row 7), random sampling with two hard negative proposals (row 4), hardest negative mining (row 5) and random semi-hard negative mining (row 6; semi-hard negative proposals can be hard and some of their matching scores are smaller than the matching score of the ground-truth proposal). CMRINs using the triplet loss with one or two negative proposals have similar performance. Their differences in average precision over the three testing sets (RefCOCO, RefCOCO+ and RefCOCOg) are -0.19%, -0.04% and 0.06%, respectively. CMRINs trained using the triplet loss with three different definitions of negative proposals have similar performance except the triplet loss with hardest negative mining on the RefCOCO+ dataset. Their differences in precision over the validation sets and testing sets are within ±0.65%, which demonstrates the robustness of our proposed CMRIN with respect to different sampling strategies.

We report the performance of CMRIN (row 7: triplet loss with a margin of 0.1, one negative proposal and random hard negative mining) as the final version of our algorithm, which is chosen according to its performance on the validation sets.

4.4.3 Necessity of Phrases

	RefCOCO		RefCOCO+		RefCOCOg
	testA	testB	testA	testB	test
w/o phrase	86.95	84.40	79.57	68.23	79.71
implicit phrase	87.63	84.73	80.93	68.99	80.66
CMRIN	87.36	85.52	80.37	69.58	80.85

TABLE 4
Comparison of different phrase designs on RefCOCO, RefCOCO+ and RefCOCOg.

We discuss the necessity of phrases in this section and the results are shown in Table 4. The performance of the variant using words to highlight the vertices of the spatial relation graph (row 1) is worse than that of the final version using both words and phrases (row 3), which demonstrates the effectiveness of phrases in improving the accuracy of vertex highlighting. It is worth noting that we implicitly considered the word context (phrase-level information) in our conference version (row 2) by using the contextual features of words to attend the vertices instead of using the word embeddings. However, the contextual features of words introduce the global noise of the expressions, which increases the difficulty of learning the correspondence between words and vertices. The performance of CMRIN with implicit phrases is worse than that of it with explicit phrases in the object category (i.e., test B), because the visual contents of object category is sensitive to contextual noise. In addition, explicit use of phrases can help align between the linguistic words and visual objects.

4.4.4 Variants of Spatial Relation Graph

We conduct experiments for CMRINs with different spatial relation graphs to evaluate the effects of different designs for spatial relation graphs, and those designs come from two perspectives.

On one hand, we adopt three types of designs for edges, i.e, "type(11)", "type(7)" and "soft". Specifically, "type(11)" is the 11 types of edges introduced in Section 3.1.1; the "type(7)" is a coarse-grained version of "type(11)" and its 7 types of edges are "inside", "cover", "overlap", "right"', "top", "left" and "bottom"; the "soft" is a fine-grained version of "type(11)" and it directly encodes the edges as relative location representations [3] by calculating the offsets and area ratios between objects. As shown in Table 5, the performance of CMRIN with "type(7)" (row 1) is slightly worse than that of it with "type(11)" (row 7), because the design of "type(7)" is coarse than the design of "type(11)". The CMRIN with "soft" (row 2) and "type(11)" (row 3) have similar performance on RefCOCO and RefCOCOg datasets, but the performance of latter is better than that of the former on RefCOCO+ dataset, which indicates that "type(11)" is fine enough to capture spatial relationships. In addition, "type(11)" is more memory- and computation-efficient than "soft".

On the other hand, we evaluate different conditions for connecting between objects, i.e, "edge num", "axis dis" and "center dis". In particular, the "edge num(5)" constraints

		RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
1	type(7) + center dis(0.5)	86.50	87.56	84.26	75.18	80.58	69.01	79.58	80.37
2	soft + edge num(5)	86.94	88.12	84.46	75.12	80.06	68.62	80.43	80.80
3	type(11) + edge num(5)	86.58	88.39	84.47	75.29	81.63	68.64	80.11	81.07
4	type(11) + axis dis(0.15)	86.46	87.91	85.14	76.01	81.02	68.99	80.45	81.21
5	type(11) + center dis(0.3)	86.67	88.21	84.53	75.46	80.49	69.26	80.29	80.35
6	type(11) + center dis(0.7)	86.57	87.71	84.14	75.21	79.74	69.16	79.68	80.20
7	type(11) + center dis(0.5)	86.43	87.36	85.52	75.63	80.37	69.58	79.72	80.85

TABLE 5
Ablation studies on variants of spatial relation graph of our proposed CMRIN on RefCOCO, RefCOCO+ and RefCOCOg. The variant is denoted as design of edge type with condition of existence for edges. The numbers in parentheses following the "type", "dis" and "num" represent the number of types of edges, the threshold of normalized distance and the maximum number of edges at each vertices respectively. The two best performing models are marked in red and blue. CMRIN consistently outperforms existing state-of-the-art models on all the three benchmark datasets.

the maximum out-degrees of each vertices to 5 and a vertex is connected to its 5 nearest nodes based on the distances between their normalized center coordinates (i.e., center distances) [3]; the "axis dis(0.15)" connects each pair of objects as long as the relative distances between them in axes are smaller than 15% of the length and width of the image respectively [70]; the "center dis(threshold)" creates a edge for each pairs of objects whose center distance is smaller than the threshold. As shown in Table 5, CMRIN with "center dis(0.7)" (row 6) has relative lower precision than that with other conditions (row 3, 4, 5 and 7), because the "center dis(0.7)" covers several redundant edges which introduces noisy information. The CMRIN with remaining conditions have similar performance, and their differences in average precision over the validation and testing sets on RefCOCO, RefCOCO+ and RefCOCOg datasets are within ±0.07%, ±0.27% and ±0.55%, respectively.

4.4.5 *Semantic Relation Graph Branch*

It is intuitive to encode the semantic relationships among objects, in this section, we explore the effectiveness and detail the implementation of semantic relation graph branch.

	RefCOCO		RefCOCO+		RefCOCOg
	testA	testB	testA	testB	test
spatial	87.36	85.52	80.37	69.58	80.85
semantic	87.63	84.00	80.00	68.64	80.64
spatial+semantic	88.17	85.59	81.44	68.81	81.71

TABLE 6
Experimental results of spatial/semantic relation graph branch on RefCOCO, RefCOCO+ and RefCOCOg.

Effectiveness. We compare the CMRINs with single spatial relation graph branch, with single semantic relation graph branch and joint model including both branches. And the results are shown in the Table 6. The performance of single semantic branch is worse than that of single spatial branch, because the object relationship detector cannot recognize the semantic relations completely accurately in the highly unrestricted scenes. Moreover, the spatial relation graph branch also implicitly captures the semantic relationships as described in Section 3.2.2. In addition, the model including both branches achieves the best precision, which indicate the possibility of cooperation between the spatial and semantic relationship representations.

Implementation. To implement the branch using semantic relationship graph, we first use a visual relationship detector (the relationship detection part of [59]) trained on the Visual Genome datasets excluding the images in RefCOCO, RefCOCO+ and RefCOCOg datasets to extract the semantic relationships among objects. And, for each predicted edge, we compute its features as the probability-weighted embedding of the relationship categories, and the probabilities to relationship categories are predicted by the visual relationship detector. Since we represent each edge as a type in spatial branch and represent each edge as a feature in semantic branch, the implementation for semantic branch has some minor differences with that of spatial branch. In particular, 1) we attend the language representations over the features of edges instead of over the types of edges; 2) We learn the bias vectors for edges by using fully connected layers to encode the edge features in each gated graph convolutional layers instead of learning the bias vectors for types of edges.

For model including both branches, the semantic branch and spatial branch share the same language representations and the vertex attention computation, but have their individual attention learning between language representations and visual edges, gated graph convolutional layers and matching computations. The final score of the two-branched model is the mean of matching scores from those two branches. The whole model is end-to-end trained by the triplet loss with online hard negative mining.

4.5 Grounding Referring Expressions with Detected Object Proposals

We have also evaluated the performance of the proposed CMRIN for grounding referring expressions using automatically detected objects in the three datasets. The detected objects are provided by [3], and they were detected with a pretrained Faster R-CNN in COCO's training images with the images in the validation and testing sets of RefCOCO, RefCOCO+ and RefCOCOg excluded. The results are shown in Table 7. The proposed CMRIN outperforms existing state-of-the-art models, which demonstrates the robustness of CMRIN with respect to object detection results. Specifically, CMRIN improves the average precision in the person category achieved with the existing best-performing method by

3.80%, and it improves the Precision@1 on RefCOCO+'s test A and test B by 5.50% and 1.27%, respectively.

	RefCOCO		RefCOCO+		RefCOCOg
	testA	testB	testA	testB	test
MMI [2]	64.90	54.51	54.03	42.81	-
Neg Bag [13]	58.60	56.40	-	-	49.50
CG [10]	67.94	55.18	57.05	43.33	-
Attr [7]	72.08	57.29	57.97	46.20	-
CMN [11]	71.03	65.77	54.32	47.76	-
Speaker [8]	67.64	55.16	55.81	43.43	-
S+L+R [9]	73.71	64.96	60.74	48.80	59.63
VC [12]	73.33	67.44	58.40	53.18	-
MAttNet [3]	80.43	**69.28**	70.26	56.00	67.01
Ours CMRIN	**82.53**	68.58	**75.76**	**57.27**	**67.38**

TABLE 7
Comparison with the state-of-the-art methods on RefCOCO, RefCOCO+ and RefCOCOg when using detected proposals. The best performing method is marked in bold.

4.6 Effectiveness on multi-order relationships subsets

To evaluate the effectiveness of our method on multi-order relationships alone, we identify the subset of expressions with indirect references in the RefCOCOg's test set. There are 2,507 expressions with multiple verbs/location words and Precision@1 on this subset is 81.07% while the number of remaining expression is 7,095 and Precision@1 on them is 80.22%. In contrast, the Precision@1 of MattNet [3] (existing best method) is 76.31% and 79.30% respectively on these two subsets. This result demonstrates that our method can handle indirect references equally well as other simpler cases.

5 Conclusions

In this paper, we focus on the task of referring expression comprehension in images, and demonstrate that a feasible solution for this task needs to not only extract all the necessary information in both the image and referring expressions, but also compute and represent multimodal contexts for the extracted information. In order to overcome the challenges, we propose an end-to-end Cross-Modal Relationship Inference Network (CMRIN), which consists of a Cross-Modal Relationship Extractor (CMRE) and a Gated Graph Convolutional Network (GGCN). CMRE extracts all the required information adaptively for constructing language-guided visual relation graphs with cross-modal attention. GGCN fuses information from different modes and propagates the fused information in the language-guided relation graphs to obtain multi-order semantic contexts. Experimental results on three commonly used benchmark datasets show that our proposed method outperforms all existing state-of-the-art methods.

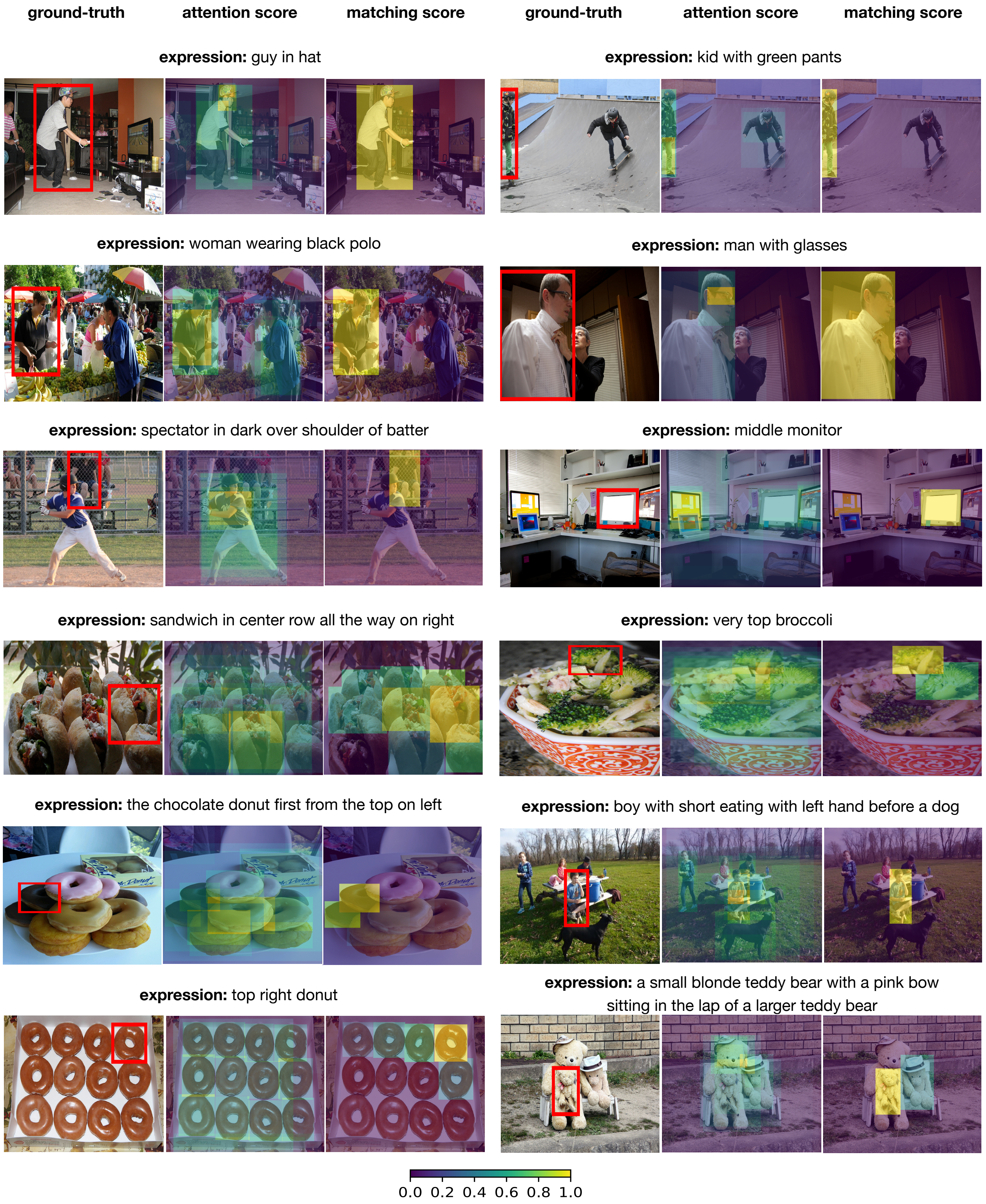

ground-truth
attention score
matching score
ground-truth
attention score
matching score
expression: guy in hat
expression: kid with green pants
expression: woman wearing black polo
expression: man with glasses
expression: spectator in dark over shoulder of batter
expression: middle monitor
expression: sandwich in center row all the way on right
expression: very top broccoli
expression: the chocolate donut first from the top on left
expression: boy with short eating with left hand before a dog
expression: top right donut
expression: a small blonde teddy bear with a pink bow sitting in the lap of a larger teddy bear
0.0
0.2
0.4
0.6
0.8
1.0

Fig. 5. Qualitative results showing initial attention score (gate) maps and final matching score maps. Warmer color indicates higher score.

References

[1] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "Referitgame: Referring to objects in photographs of natural scenes," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 787–798.

[2] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 11–20.

[3] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg, "Mattnet: Modular attention network for referring expression comprehension," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele, "Grounding of textual phrases in images by reconstruction," in *European Conference on Computer Vision*. Springer, 2016, pp. 817–834.

[7] J. Liu, L. Wang, and M.-H. Yang, "Referring expression generation and comprehension via attributes," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[8] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *European Conference on Computer Vision*. Springer, 2016, pp. 69–85.

[9] L. Yu, H. Tan, M. Bansal, and T. L. Berg, "A joint speakerlistener-reinforcer model for referring expressions," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, vol. 2, 2017.

[10] R. Luo and G. Shakhnarovich, "Comprehension-guided referring expressions," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, vol. 2, 2017.

[11] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko, "Modeling relationships in referential expressions with compositional modular networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE, 2017, pp. 4418–4427.

[12] H. Zhang, Y. Niu, and S.-F. Chang, "Grounding referring expressions in images by variational context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4158–4166.

[13] V. K. Nagaraja, V. I. Morariu, and L. S. Davis, "Modeling context between objects for referring expression understanding," in *European Conference on Computer Vision*. Springer, 2016, pp. 792–807.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[15] B. Dai, Y. Zhang, and D. Lin, "Detecting visual relationships with deep relational networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 3298–3308.

[16] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *European Conference on Computer Vision*. Springer, 2016, pp. 852–869.

[17] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5831–5840.

[18] S. Yang, G. Li, and Y. Yu, "Cross-modal relationship inference for grounding referring expressions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4145–4154.

[19] C. Deng, Q. Wu, Q. Wu, F. Hu, F. Lyu, and M. Tan, "Visual grounding via accumulated attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018.

[20] S. Bell, C. Lawrence Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2874–2883.

[21] J. S. J. Ren, X. Chen, J. Liu, W. Sun, J. Pang, Q. Yan, Y. Tai, and L. Xu, "Accurate single stage detector using recurrent rolling convolution," 2017, pp. 752–760.

[22] L. Chen, G. Papandreou, I. Kokkinos, K. P. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[23] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[24] Y. Liu, R. Wang, S. Shan, and X. Chen, "Structure inference net: Object detection using scene-level context and instance-level relationships," 2018.

[25] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: Deep learning on spatio-temporal graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5308–5317.

[26] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *International Conference on Learning Representations*, 2017.

[27] K. Marino, R. Salakhutdinov, and A. Gupta, "The more you know: Using knowledge graphs for image classification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[28] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. A. Forsyth, "Every picture tells a story: generating sentences from images," in *European conference on computer vision*, 2010, pp. 15–29.

[29] X. Wu, G. Li, Q. Cao, Q. Ji, and L. Lin, "Interpretable video captioning via trajectory structured localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018, pp. 6829–6837.

[30] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," *The IEEE International Conference on Computer Vision (ICCV)*, pp. 2425–2433, 2015.

[31] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Babytalk: Understanding and generating simple image descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2891–2903, Dec. 2013. [Online]. Available: doi.ieeecomputersociety.org/10.1109/TPAMI.2012.162

[32] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," 2015, pp. 3156–3164.

[33] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," 2015, pp. 2048–2057.

[34] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," 2017.

[35] H. Ling and S. Fidler, "Teaching machines to describe images with natural language feedback," 2017, pp. 5068–5078.

[36] B. Dai, S. Fidler, and D. Lin, "A neural compositional paradigm for image captioning," 2018.

[37] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," *The IEEE International Conference on Computer Vision (ICCV)*, pp. 1–9, 2015.

[38] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, "Are you talking to a machine? dataset and methods for multilingual image question answering," *Advances in neural information processing systems*, pp. 2296–2304, 2015.

[39] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *conference on empirical methods in natural language processing (EMNLP)*, pp. 457–468, 2016.

[40] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," *Advances in neural information processing systems*, pp. 289–297, 2016.

[41] I. Schwartz, A. G. Schwing, and T. Hazan, "High-order attention models for visual question answering," *Advances in neural information processing systems*, pp. 3664–3674, 2017.

[42] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," *IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 4613–4621, 2016.

[43] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola, "Stacked attention networks for image question answering," *IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 21–29, 2016.

[44] A. Santoro, D. Raposo, D. G. T. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. P. Lillicrap, "A simple neural network module for relational reasoning," *Advances in neural information processing systems*, pp. 4967–4976, 2017.

[45] C. Wu, J. Liu, X. Wang, and X. Dong, "Chain of reasoning for visual question answering," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 273–283.

[46] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," *The IEEE International Conference on Computer Vision (ICCV)*, pp. 1839–1848, 2017.

[47] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural module networks," *IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 39–48, 2016.

[48] Q. Cao, X. Liang, B. Li, G. Li, and L. Lin, "Visual question reasoning on general dependency tree," *IEEE conference on computer vision and pattern recognition (CVPR)*, 2018.

[49] Y. Zhang, J. C. Niebles, and A. Soto, "Interpretable visual question answering by visual grounding from attention supervision mining," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 349–357.

[50] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7w: Grounded question answering in images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4995–5004.

[51] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner *et al.*, "Relational inductive biases, deep learning, and graph networks," *arXiv preprint arXiv:1806.01261*, 2018.

[52] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *International Conference on Learning Representations, 2018*, 2018.

[53] X. Wang, Y. Ye, and A. Gupta, "Zero-shot recognition via semantic embeddings and knowledge graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6857–6866.

[54] D. Teney, L. Liu, and A. van den Hengel, "Graph-structured representations for visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1–9.

[55] W. Norcliffe-Brown, S. Vafeias, and S. Parisot, "Learning conditioned graph structures for interpretable visual question answering," in *Advances in Neural Information Processing Systems*, 2018, pp. 8334–8343.

[56] R. Cadene, H. Ben-Younes, M. Cord, and N. Thome, "Murel: Multimodal relational reasoning for visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1989–1998.

[57] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 685–10 694.

[58] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[59] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph r-cnn for scene graph generation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 670–685.

[60] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," 2018.

[61] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[62] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2015, pp. 815–823.

[63] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[64] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Association for Computational Linguistics (ACL) System Demonstrations*, 2014, pp. 55–60.

[65] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018.

[66] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 770–778.

[67] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[68] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 2015.

[69] R. Manmatha, C. Wu, A. J. Smola, and P. Krahenbuhl, "Sampling matters in deep embedding learning," *The IEEE International Conference on Computer Vision (ICCV)*, pp. 2859–2867, 2017.

[70] D. A. Hudson and C. D. Manning, "Learning by abstraction: The neural state machine," *arXiv preprint arXiv:1907.03950*, 2019.