MUTATT: VISUAL-TEXTUAL MUTUAL GUIDANCE FOR REFERRING EXPRESSION COMPREHENSION

Shuai Wang, Fan Lyu*, Wei Feng*, Song Wang*†*

*College of Intelligence and Computing, Tianjin University, China
†Department of Computer Science and Engineering, University of South Carolina, US
{wangshuai201909, fanlyu}@tju.edu.cn, wfeng@ieee.org, songwang@cec.sc.edu

ABSTRACT

Referring expression comprehension (REC) aims to localize a text-related region in a given image by a referring expression in natural language. Existing methods focus on how to build convincing visual and language representations independently, which may significantly isolate visual and language information. In this paper, we argue that for REC the referring expression and the target region are semantically correlated and subject, location and relationship consistency exist between vision and language. On top of this, we propose a novel approach called MutAtt to construct mutual guidance between vision and language, which treats vision and language equally thus yields compact information matching. Specifically, for each module of subject, location and relationship, MutAtt builds two kinds of attention-based mutual guidance strategies. One strategy is to generate vision-guided language embedding for the sake of matching relevant visual features. The other reversely generates language-guided visual features to match relevant language embedding. This mutual guidance strategy can effectively enforce the vision-language consistency in three modules. Experiments on three popular REC datasets demonstrate that the proposed approach outperforms the current state-of-the-art methods.

Index Terms— Referring expression comprehension, vision-language matching, mutual guidance

1. INTRODUCTION

Referring expression comprehension (REC), also known as visual grounding, aims at finding the text-related object in a given image according to the description of referring expressions. As a vision-language problem, REC has widespread applications in real-world scenarios, e.g., in an autopilot system, we need to localize the exact location in images or videos from text expressions like "park the car on the right side". Although much progress has been made in REC, grounding referring expressions remains challenging because it requires a comprehensive understanding of complex language semantics and various types of visual information simultaneously.

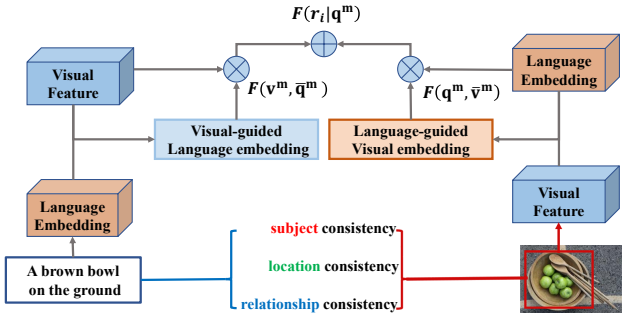

Fig. 1. Framework of the proposed MutAtt. We assume there exist three kinds of consistency between referring expression and the target region proposal. MutAtt builds mutual attention-based guidance strategy between visual and language information, which consists of visual-guided language embedding and language-guided visual embedding.

Researches on REC can be categorized into generative methods and discriminative methods. Generative methods, originated from image captioning [1, 2], generate description for each localized region in searching by maximum posteriori probability [3, 4, 5]. However, generative methods overrelies on the local region captioning model, which cannot describe the relative location and relationships with other objects. Discriminative methods try to learn the joint vision-language matching score and select object by ranking all scores [6, 7, 8, 9], which has become the most common ways in REC. Existing discriminative methods always focus on extracting more powerful visual and language features. Generally, these methods use Convolutional Neural Networks to encode the visual features for each candidate region, and use Recurrent Neural Networks to encode the referring expression [6, 10]. Compositional modular networks [7, 8] decompose the referring expression into three parts: subject, location and relationship, and design three visual feature representations to achieve fine-grained matching. Variational context [9] exploits the reciprocal relation between the referent

This work was supported by the National Natural Science Foundation of China (Nos. U1803264, 61672376, 61671325).

978-1-7281-1331-9/20/$31.00 ©2020 IEEE

and context to solve the problem of complex context modeling in referring expression comprehension. Nevertheless, the previous discriminative methods focus on how to build convincing visual and language representations independently, where the referring expressions are always only treated as unrequited queries. This may significantly isolate visual and language information, thus hinders the effective matching between vision and language, especially when the scene or expression are complex.

In our view, *REC can only work based on the hypothesis that the referring expression and the target region represent the same semantics*, including **subject consistency**, **location consistency** and **relationship consistency**. By considering these three kinds of consistency, REC model can achieve more compact vision and language combination and more accurate prediction. Based on this hypothesis, we design an innovative mutual attention-based guidance method MutAtt in the perspective of vision-language matching by enforcing these three consistencies. Specifically, to ensure effective cross modal consistency, we first treat REC as a vision-language matching problem in order to make visual and language information equal. MutAtt provides two strategies to achieve the above hypothesis as shown in Fig. 1. One strategy uses visual features to guide the language and then matches the guided language features with visual features. While improving the consistency of cross-modal information, it will make the model focus on vision over language. The other strategy uses language embedding to guide vision and then match the generated visual features with language embedding. This allows us to balance the weight of vision and language information while further improving cross-model consistency. We apply this approach to subject, location and relationship modules, which significantly enhance the three kinds of consistency while maintaining vision and language equality. We conduct experiments on three popular REC datasets to verify the advantages of the proposed method, and the experimental results show the superiority of the proposed MutAtt.

2. RELATED WORK

Referring expression comprehension. Existing REC methods generally fall into two categories: generative model and discriminative model. Many generative models[3, 4, 5] use the encoder-decoder structure to localize the region that can generate the sentence with maximum posteriori probability. Discriminative model [7, 8, 9] tends to use various feature vectors to represent the expression and the image region, and then measures the similarity of them to select the region with the highest scores. Earlier methods[6] separately encode the entire related expression and the entire image feature , which ignores the complex structures in the language as well in the image. Later works in [7, 8] overcomes this limitation through decomposing the expression into sub-components and computing the vision-language matching scores of each module. The method in [9] lowers the requirement of joint grounding and reasoning to a holistic association score between the sentence and region features. In addition, recent work [8, 11] uses the attention mechanism to make the model focus on more critical information and achieves significant effectiveness. However, the previous discriminative methods focus on how to build convincing visual and language representations independently, and never consider the information consistency between vision and language. These methods only regard referring expression as a complementary query and overemphasize the importance of visual information. In contrast, we propose to enhance the vision-language consistency by cross-modal attention-based mutual guided matching.

Vision-language matching. Vision-language matching has been studied for years, the key challenge of which is measuring the similarity between vision and language embeddings. The most popular vision-language matching methods [12, 13, 14] rely on relatively similar procedures: extract discriminative visual and language features and measure as accurately as possible the distance between the two representations. The work in [15, 16] adopts CNN and Skip-Gram or LSTM to extract feature representations for both modals. Then a ranking loss is used to force the model to get closer to the matched vision-language pair and away for the unmatched pair. [17] the learning of cross-view feature embedding is further improved by incorporating generative objectives. Through region relationship reasoning and global semantic reasoning, image representation can be enhanced [18] to align with the corresponding text caption better. In this paper, we treat the fusion of visual and language features as a kind of cross-modal matching. By enhancing the vision-language consistency, we make vision and language play an equally important role in REC. In this way, the proposed MutAtt can discover more discriminative joint visual-textual representations.

3. METHOD

3.1. Problem formulation and background

Given an image I with a set of regions of interest $\mathcal{R} = \{r_i\}$ tagged by people or detection algorithm and referring expression $\mathcal{E} = \{u_t\}_{t=1}^{T}$, where u_t means the t-th word in sentence, the purpose of REC is to find the target region r^* best matching $\mathcal{E}$. The effective solution is to match the visual features of each candidate region and the language embedding of expression, and select the region with the highest score. We follow the modular design of MAttNet [8] as our backbone for its capability to handle subject, location and relationship information in referring expressions. MAttNet decomposes the expression embedding into three modular components, i.e. $\{\mathbf{q}^{\text{subj}}, \mathbf{q}^{\text{loc}}, \mathbf{q}^{\text{rel}}\}$, via a language attention network, and designs three visual models to encode the corresponding vi-

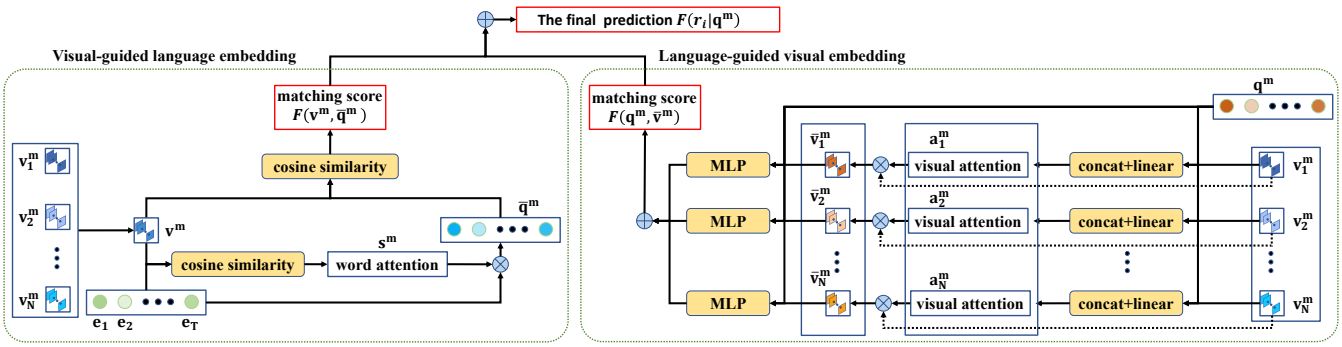

Fig. 2. Illustration of MutAtt. $\{\mathbf{v}_n^m\}_{n=1}^N$ represent visual feature of region proposal, $\{\mathbf{e}_t\}_{t=1}^T$ represent word embedding of sentence and $\mathbf{q}^m$ represent phrase embedding of sentence. The left part shows visual-guided language embedding, where we compute word attention to guide the generation process of language embedding and match it with the visual feature by cosine similarity. The right part shows language-guided visual embedding, where we compute attention on visual feature guided by language embedding and match them by MLPs. Finally, we combine the matching result of two parts for a final score.

sual feature $\mathbf{v}^m$, where $m \in (\text{subj}, \text{loc}, \text{rel})$. In this paper, we introduce a mutual attention-based guidance approach called MutAtt to improve vision-language consistency, including vision-guided language embedding and language-guided visual feature, as shown in Fig. 2. As we treat the REC problem as a matching problem, we only consider one region r (not specific) in $\mathcal{R}$ as the visual input, while in the inference, the region with largest matching score will be selected.

3.2. Mutual attention-based guidance

3.2.1. Visual-guided language embedding

We first design to use visual feature help the formation of language embedding through matching vision and language from *word-level* to *sentence-level* for each module $m \in \{\text{subj}, \text{loc}, \text{rel}\}$. To be specific, we compute the cosine similarity vector $\mathbf{s}^m$ between word embedding $\{\mathbf{e}_t\}_{t=1}^T$ and visual feature $\{\mathbf{v}_n^m\}_{n=1}^N$ of region proposal r, which can be computed as

$$\mathbf{s}_t^m = \frac{(\mathbf{v}^m)^\top \mathbf{e}_t}{\|\mathbf{v}^m\| \, \|\mathbf{e}_t\|}, t \in [1, T], \tag{1}$$

where $\mathbf{v}^m$ is the average pooled visual feature of $\{\mathbf{v}_n^m\}_{n=1}^N$ and can be obtained by

$$\mathbf{v}^m = \frac{1}{N} \sum_{n=1}^{N} \mathbf{v}_n^m, \tag{2}$$

where N represents the number of visual elements in different module for the candidate region. In Eq. (1), $\mathbf{s}_t^m$ represents the attention from visual feature of module m to the t-th word embedding. By this word-level similarity, we compute the fine-grained similarity between each visual and language element pair, which can significantly compose of the visual-guided language embedding. Thus, we use the similarity as the weight of each word embedding to generate visual-guided sentence-level embedding as follows:

$$\bar{\mathbf{q}}^m = \sum_{t=1}^{T} \text{softmax}(\lambda^m \mathbf{s}_t^m) \cdot \mathbf{e}_t, \tag{3}$$

where λ^m is the word-level language attention obtained from language attention network in MAttNet [8], which helps form the language embedding corresponding to different visual modules. Under the guidance of word-level vision-language similarities, the sentence-level embedding can be enhanced by visual feature.

After that, we further calculate the score of visual feature $\mathbf{v}^m$ and visual-guided language embedding $\bar{\mathbf{q}}^m$ by the cosine similarity through matching vision and language in sentence level:

$$F(\mathbf{v}^m, \bar{\mathbf{q}}^m) = \frac{(\mathbf{v}^m)^\top \bar{\mathbf{q}}^m}{\|\mathbf{v}^m\| \, \|\bar{\mathbf{q}}^m\|}. \tag{4}$$

Note that, we propose to match vision and language information from word-level to sentence-level, which can guarantee the multi-scale vision-language matching. If the region and referring expression never match, the score would be small by this two level matching method, which could help omit failed predictions. To ensure that vision and language information have equal importance in the matching process and further improve vision-language consistency, in the following we also construct a language-guided visual embedding.

3.2.2. Language-guided visual embedding

In our framework, we assume that the language and vision play equal role. Thus, after using visual information to guide language embedding, we also build the reverse guided embedding, i.e., the language-guided visual embedding. Given the visual feature $\{\mathbf{v}_n^m\}_{n=1}^N$ of region proposal r and the corresponding language embedding $\mathbf{q}^m$ of referring expression $\mathcal{E}$, we first compute language-guided visual attention on subject,

location and relationship modules.

$$\mathbf{h}_n = \tanh(\mathbf{W}_1^m[\mathbf{v}_n^m, \mathbf{q}^m] + \mathbf{b}^m), \tag{5}$$

$$\mathbf{a}_n^m = \text{softmax}(\mathbf{W}_2^m \mathbf{h}_n), \tag{6}$$

where $[\cdot]$ is the concatenation operation, $\mathbf{W}_1^m, \mathbf{b}^m, \mathbf{W}_2^m$ are model parameters and $\mathbf{a}_n^m$ represents the attention from language embedding $\mathbf{q}^m$ to the n-th visual element of region proposal r. After that, we generate a more discriminative language-guided visual feature by

$$\bar{\mathbf{v}}^m = \sum_{n=1}^{N} \mathbf{a}_n^m \cdot \mathbf{v}_n^m \tag{7}$$

$$F(\mathbf{q}^m, \bar{\mathbf{v}}^m) = \text{MLP}(\mathbf{q}^m, \bar{\mathbf{v}}^m). \tag{8}$$

In Eq. (8), we use MLP structure to calculate the score between language-guided visual feature and the language embedding. Each MLP is composed of two fully connected layers with ReLU activation, which help transform cross modal information into a common embedding space. With language-guided visual embedding, we guarantee the consistency of visual and language information and prevent the model from paying too much attention to one of them. Note, the language-guided visual embedding is similar to the common attention using in other REC methods, as they only consider the simple visual language fusion. The drawback of these methods is that the language is treated as the complementary query, without considering their mutual guidance.

3.3. Matching result and loss function

We combine the proposed MutAtt in subject, location and relationship modules. The overall matching score for the region proposal and expression is:

$$F(r_i|\mathcal{E}) = \omega_{\text{subj}} F(r_i|\mathbf{q}^{\text{subj}}) + \omega_{\text{loc}} F(r_i|\mathbf{q}^{\text{loc}}) + \omega_{\text{rel}} F(r_i|\mathbf{q}^{\text{rel}}), \tag{9}$$

$$F(r_i|\mathbf{q}^m) = F(\mathbf{v}^m, \bar{\mathbf{q}}^m) + F(\mathbf{q}^m, \bar{\mathbf{v}}^m), \tag{10}$$

where $(\omega_{\text{subj}}, \omega_{\text{loc}}, \omega_{\text{rel}})$ represent the weights of subject module, location module and relationship module obtained from language attention network in MAttNet.

For positive candidate object and query pair $(\mathcal{R}_i, \mathcal{E}_i)$ and negative pairs $(\mathcal{R}_i, \mathcal{E}_j)$, $(\mathcal{R}_j, \mathcal{E}_i)$, the ranking loss is minimized during training:

$$\begin{aligned} \text{Loss} = \sum_i ([k - F(\mathcal{R}_i, \mathcal{E}_i) + F(\mathcal{R}_i, \mathcal{E}_j)]_+ \\ + [k - F(\mathcal{R}_i, \mathcal{E}_i) + F(\mathcal{R}_j, \mathcal{E}_i)]_+), \end{aligned} \tag{11}$$

where $[x]_+ = \max(x, 0)$, and k is the margin for the loss.

4. EXPERIMENTS

4.1. Dataset and implementation details

Dataset. We use three popular datasets for the evaluation, i.e., RefCOCO, RefCOCO+ and RefCOCOg [5, 6]. RefCOCO has 50,000 target objects collected from 19,994 images. RefCOCO+ has 49,856 target objects collected from 19,992 images. These two datasets are split into four parts of "train", "val", "testA" and "testB". RefCOCOg includes 49,822 target objects from 25,799 images, which are split into three parts of "train", "val" and "test".

Visual feature representation. We use faster R-CNN with ResNet101 as backbone to extract subject features, location features and relationship features for each region proposal and follow [8] to construct modular visual network. For the subject network, we feed the whole image into faster R-CNN and extract 7×7 features maps from the last convolutional output of 3rd-stage and the last convolutional output of 4-th stage to represent subject features. For the location network, we represent location features of candidate object by encoding position and relative area as $\mathbf{l}_i = [\frac{x_{tl}}{W}, \frac{y_{tl}}{H}, \frac{x_{br}}{W}, \frac{y_{br}}{H}, \frac{w \cdot h}{W \cdot H}]$, and encoding relative location offsets and relative areas of up-to-five surrounding same-category objects as $\delta\mathbf{l}_{ij} = [\frac{[\Delta x_{tl}]_{ij}}{w_i}, \frac{[\Delta y_{tl}]_{ij}}{h_i}, \frac{[\Delta x_{br}]_{ij}}{w_i}, \frac{[\Delta y_{br}]_{ij}}{h_i}, \frac{w_j \cdot h_j}{W \cdot H}]$. For the relationship network, we first find up-to-five surrounding objects, then extract their average-pooled visual features and encode their relative position offsets and relative areas to represent relationship features of context objects. For the visual features $\{\mathbf{v}_n^m\}_{n=1}^N$ mentioned in Sec. 3.2, $n = (49, 1, 5)$ when $m = (\text{subj}, \text{loc}, \text{rel})$ respectively.

Training setting. The training batch size is 15, which means in each training iteration we feed 15 images and the referring expressions associated with these images to the network. Adam is used as the training optimizer, with initial learning rate 0.0004, which decays by a factor of 10 every 8,000 iterations. We implement MutAtt based on PyTorch.

Evaluation setting. Following previous work [19, 20], we take the region proposals from human annotated (gt) and detection methods (det). For gt, the evaluation favors the region with the highest matching score to be the same as the ground-truth region. For det, the evaluation favors the intersection over union between the region with highest matching score and ground-truth region to be greater than 0.5.

4.2. Results

Comparisons with State-of-The-Art. We provide a comparison of our method with other SOTA methods in Table. 1, including the results of using two settings on three datasets. As can be seen, MutAtt shows the advantage of the proposed approach. On the ground-truth setting, MutAtt is significantly better than the previous method on the RefCOCO dataset, and performs similarly to the previous method on the RefCOCO+ and RefCOCOg datasets. On more important detec-

Table 1. Comparison with state-of-the-art REC approaches on ground-truth and automatically detected regions. It can be seen that our method has significantly improved performance compared with other methods, and is superior to SOTA in most metrics.

Method	Box	RefCOCO			RefCOCO+			RefCOCOg		
		val	testA	testB	val	testA	testB	val*	val	test
visdif+MMI [6]	gt	-	73.98	76.59	-	59.17	55.62	64.02	-	-
Speaker/visdif [6]	gt	76.18	74.39	77.30	58.64	61.29	56.24	59.40	-	-
S-L-R [3]	gt	79.56	78.65	80.22	62.26	64.60	59.62	72.63	71.65	71.92
VC [9]	gt	-	78.98	82.39	-	62.56	62.90	73.98	-	-
Attr [21]	gt	-	78.05	78.07	-	61.47	57.22	69.83	-	-
Accu-Att [22]	gt	81.27	81.17	80.01	65.56	68.76	60.63	73.18	-	-
PLAN [23]	gt	81.67	80.81	81.32	64.18	66.31	61.46	69.47	-	-
Multi-hop Film [24]	gt	84.9	87.4	83.1	73.8	**78.7**	65.8	71.5	-	-
MattNet [8]	gt	85.65	85.26	84.57	71.01	75.13	66.17	-	78.10	78.12
NMT_{REE} [25]	gt	85.65	85.63	85.08	72.84	75.74	67.62	78.03	78.57	78.21
LGRANS [19]	gt	82.0	81.2	84.0	66.6	67.6	65.5	-	75.4	74.7
DGA [20]	gt	86.34	86.64	84.79	73.56	78.31	**68.15**	-	80.21	**80.26**
MutAtt	gt	**86.58**	**87.20**	**85.38**	**73.69**	76.30	67.74	-	**80.37**	79.24
S-L-R [3]	det	69.48	73.71	64.96	55.71	60.74	48.80	-	60.21	59.63
PLAN [23]	det	-	75.31	65.52	-	61.34	50.86	58.03	-	-
MattNet [8]	det	76.40	80.43	69.28	64.93	70.26	56.00	-	66.67	67.01
LGRANS [19]	det	-	76.6	66.4	-	64.0	53.4	62.5	-	-
DGA [20]	det	-	78.42	65.53	-	69.07	51.99	-	-	63.28
MutAtt	det	**78.35**	**82.52**	**71.50**	**67.90**	**72.60**	**58.60**	-	**68.67**	**69.03**

Table 2. Ablation studies on RefCOCOg dataset.

		val	test
1	MutAtt:subj+loc+rel	77.96	77.14
2	MutAtt:subj(V→L)+loc+rel	79.33	78.53
3	MutAtt:subj(V→L+L→V)+loc+rel	80.00	79.34
4	MutAtt:subj(V⇌L)+loc(V⇌L)+rel	80.35	79.03
5	MutAtt:subj(V⇌L)+loc(V⇌L)+rel(V⇌L)	80.37	79.24

tion settings, we use the features of res101-frcn and compare with other methods. MutAtt outperforms the state-of-the-art on various split sets of the three datasets. It demonstrates that MutAtt can ensure the equality of vision and language in matching and improve the vision-language consistency on subject, location and relationship modules.

Ablation Study. We perform ablation study to verify the reliability of visual and language mutual guidance on each module. In the ablation study, we give the evaluation results of ground-truth setting on the RefCOCOg dataset. The results are shown in Table. 2. Row1 shows the result without mutual guidance. Row2∼3 show the results of adding visual guidance and language guidance to subject module. The results show that both visual guidance and language guidance improve the comprehension of the model and prove the effectiveness of our method. Row∼5 show the results of the same method applied to location module and relationship module. We can see that the help for the improvement of model comprehension gradually decreases. The reason for this phenomenon is that of the three module weights generated by the language attention network, the subject module has the highest weight, the relationship module has the lowest weight to be less than 0.1 in most cases.

4.3. Visualization

We visualize the attention of image and the weights of expressions in Fig. 3. The first column is the comprehension result of our approach and the third column is the comprehension result of MAttNet. From the first set of examples, it is obvious that our method is superior to MAttNet in terms of visual attention, language embedding and comprehensive understanding. With the guidance of "a brown bowl on the ground", the focus area of the model is moved from the edge of the "bowl" to the main body of the "bowl". Correspondingly, with the help of the guidance of visual features, the model improves the understanding of the relationship between "bowl" and "ground" in "a brown bowl on the ground", and encodes the "ground" as a related object rather than a target object.

5. CONCLUSION

In this paper, we proposed a mutual attention-based guidance method (MutAtt) for the task of REC. MutAtt contains two key components for vision-language matching: *visual-guided language embedding* and *language-guided visual embedding*. By combining two matching processes, we maintains vision and language equality. So MutAtt can learn more discriminative visual feature and language embedding while guarantee vision-language consistency during the matching process in three sub-components, which are beneficial to matching on cross-modal information. Experiments on three REC datasets with two settings show that MutAtt outperforms other methods on most evaluation metrics, which demonstrates the effectiveness of MutAtt.

6. REFERENCES

[1] Jiuxiang Gu, Jianfei Cai, Gang Wang, and Tsuhan Chen, "Stack-captioning: Coarse-to-fine learning for image captioning," in *AAAI*, 2018.

[2] Xu Yang, Hanwang Zhang, and Jianfei Cai, "Learning to col-

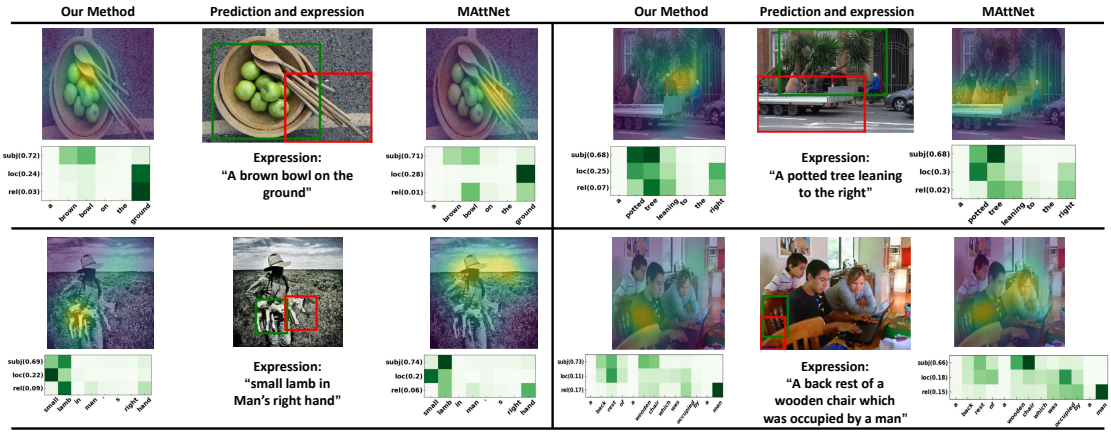

Fig. 3. Visualization comparisons between MutAtt and MAttNet of visual attention and language attention of each word on three modules. Green box is the prediction result of MutAtt and red box is the prediction result of MAttNet. We can see that MutAtt can adaptively capture the weight of each word, and accurately focus on the objects described in the language.

locate neural modules for image captioning," *arXiv preprint arXiv:1904.08608*, 2019.

[3] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg, "A joint speaker-listener-reinforcer model for referring expressions," in *CVPR*, 2017.

[4] Ruotian Luo and Gregory Shakhnarovich, "Comprehension-guided referring expressions," in *CVPR*, 2017.

[5] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy, "Generation and comprehension of unambiguous object descriptions," in *CVPR*, 2016.

[6] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg, "Modeling context in referring expressions," in *ECCV*, 2016.

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

[8] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg, "Mattnet: Modular attention network for referring expression comprehension," in *CVPR*, 2018.

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

[10] Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao, "Rethinking diversified and discriminative proposal generation for visual grounding," *arXiv preprint arXiv:1805.03508*, 2018.

[11] Fan Lyu, Qi Wu, Fuyuan Hu, Qingyao Wu, and Mingkui Tan, "Attend and imagine: Multi-label image classification with visual attention and recurrent neural networks," *TMM*, 2019.

[12] Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan, "Cascade attention network for person search: Both image and text-image similarity selection," *arXiv preprint arXiv:1809.08440*, 2018.

[13] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang, "Identity-aware textual-visual matching with latent co-attention," in *ICCV*, 2017.

[14] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang, "Person search with natural language description," in *CVPR*, 2017.

[15] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov, "Devise: A deep visual-semantic embedding model," in *NeurIPS*, 2013.

[16] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.

[17] Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *CVPR*, 2018.

[18] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu, "Visual semantic reasoning for image-text matching," in *ICCV*, 2019.

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

[20] Sibei Yang, Guanbin Li, and Yizhou Yu, "Dynamic graph attention for referring expression comprehension," in *ICCV*, 2019.

[21] Jingyu Liu, Liang Wang, and Ming-Hsuan Yang, "Referring expression generation and comprehension via attributes," in *ICCV*, 2017.

[22] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan, "Visual grounding via accumulated attention," in *CVPR*, 2018.

[23] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton van den Hengel, "Parallel attention: A unified framework for visual object discovery through dialogs and queries," in *CVPR*, 2018.

[24] Florian Strub, Mathieu Seurin, Ethan Perez, Harm De Vries, Jérémie Mary, Philippe Preux, and Aaron CourvilleOlivier Pietquin, "Visual reasoning with multi-hop feature modulation," in *ECCV*, 2018.

[25] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha, "Learning to assemble neural module tree networks for visual grounding," in *ICCV*, 2019, pp. 4673–4682.