Exploring Implicit and Explicit Relations with the Dual Relation-Aware Network for Image Captioning

Zhiwei Zha[1], Pengfei Zhou[1,2], and Cong Bai[1](✉)

[1] College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China
congbai@zjut.edu.cn

[2] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

Abstract. Recently, Transformer based architectures using object region features and graph convolutional networks using scene graphs have made significant progress in the image captioning task. However, previous works paid little attention to discovering the high-level semantic relations in visual space. Specifically, they typically neglected the problem of relation mismatching between sentences and images, which may result in generating a pale list of image objects. From the perspective of alignment, there are elements such as objects, attributes, and relations in a sentence, but in visual space, there are only objects and their attributes that can be directly detected. Previous works merely focused on aligning objects and attributes between sentences and images while ignoring the relations that just appeared in sentences but cannot be visually observed in images. In this paper, we introduce a novel dual relation-aware network (DRAN) for image captioning which composes of a dual-path relation encoder and an adaptive context relation decoder to alleviate this problem. Concretely, the dual-path relation encoder in DRAN learns to encode implicit relations and explicit relations between objects into relation-aware features. Then the contextual gated fusion module in the decoder fuses adaptively two types of relation-aware features to help the decoder generate semantically richer captions. Experimental results on the MSCOCO dataset demonstrate the superiority of DRAN in relation encoding and learning, which indicates that the proposed DRAN can capture more semantic relations and details. These conclusions are reflected by the best performance of SPICE score and also by the visual examples illustrated qualitatively.

Keywords: Image captioning · Implicit and explicit relations · Dual relation-aware network · Transformer · Scene graph

1 Introduction

As one of the core tasks in vision and language fields, image captioning is gaining increasing attention in recent years. It requires recognizing salient objects in

B. Þór Jónsson et al. (Eds.): MMM 2022, LNCS 13142, pp. 97–108, 2022.
https://doi.org/10.1007/978-3-030-98355-0_9

an image, identifying the appearance attributes of objects, understanding their interactions, and finally verbalizing them with natural language, thus making it a challenging task.

To solve the problem of vision and language tasks like image captioning, the key point is to find a way to bridge the giant semantic gap between visual modality and language modality. Firstly, we need to extract high-quality low-level appearance features in visual space, which can be done with object detection networks. Then we need to explore the high-level semantic information between objects and align them with the lingual words in semantic space.

Remarkable progress [2–4,20] has been made in recent years with the help of deep learning based encoder-decoder architecture. Attention and self-attention mechanism show great competitiveness in the latest image captioning framework. Anderson et al. [2] use visual region features detected with object detection networks to replace the features extracted by CNNs, which greatly improves the quality of visual appearance representation. Scene graph based [13,21,22,24] image captioning and Transformer based [5,7,18] image captioning have explored the structured semantic information in visual space and the interactions between objects, respectively. However, most previous works paid little attention to discovering high-level semantic relations in visual space. Specifically, they typically neglected the problem of relation mismatching between sentences and images. Take the image sentence pair in Fig. 1 as an example, “people and city street and snow and cars” can be visually observed from the image, but “skiing” and “clean snow off cars” are high-level semantic relations between objects which need to be discovered or inferred from the image content. This problem of relation mismatching hinders the precise alignment of visual and language semantics, ignoring which may result in generating a pale list of detected objects. Therefore, discovering and describing the relations between the objects correctly is of great importance to generate high-quality, semantically rich, and vivid captions. In this research, we pay more attention to discovering high-level semantic relations between objects, including implicit relations and explicit relations, with the proposed DRAN model which composes of a dual-path relation encoder and an adaptive contextual relation decoder to alleviate this problem.

Specifically, we use a Transformer-based encoder to learn implicit relations and use a GCN-based encoder to learn explicit relations. These two types of relations are softly encoded in the relation-aware features which are the output of relation encoders. To make better use of these encoded features, we devise a contextual gated fusion module in our decoder to fuse the implicit and explicit relation-aware visual features. We evaluate our model on the MSCOCO [12] dataset and the results show that our method achieves the best performance in terms of BLEU@4, METEOR, ROUGE-L, and SPICE compared with other methods. In addition, we conduct ablation studies on different variants of our model and demonstrate that our method captures more semantic relations and details, which are reflected in the SPICE score quantitatively and also reflected by the visual examples qualitatively.

Overall, the major contributions of this paper are summarized as follows:

- We propose a dual-path relation encoder to learn relations between objects and thus two types of relation-aware visual features are produced. Specifically, we introduce a Transformer-based encoder to learn implicit relations and introduce a GCN-based encoder to learn explicit relations.
- We design an adaptive context relation decoder to decode sentences from the encoded relation-aware features, in which a contextual gated fusion module is devised to fuse implicit and explicit relation-aware visual features.
- Experimental results on the MSCOCO dataset show that our method outperforms previous state-of-the-art models in most metrics. Results and examples indicate that our approach captures more semantic relations than other methods.

2 DRAN

In this section, we introduce a novel dual relation-aware network (DRAN) for image captioning, which composes of a dual-path relation encoder and an adaptive context relation decoder. It takes advantage of the high-level semantic features for modeling implicit and explicit relations between objects. The overall structure of our model is illustrated in Fig. 1.

2.1 Problem Formulation

We claim that the image captioning problem can be viewed as a weakly supervised multi-classification task while iterating through time. In time step t, the objective of classification can be formulated as:

$$\begin{aligned} w_0 &= <S>, \\ w_t &= F(I, w_0, w_1, ..., w_{t-1}), \end{aligned} \tag{1}$$

where I is the representation of the image, w_i is the word embedding of the i-th word, and $<S>$ is the starting token. F is the learned model whose function is to map from image representation and partially complete sentences to the next word.

To solve this problem with our proposed DRAN, we first use the object detector of Faster-RCNN with a backbone of ResNet-101 [6] to extract a set of object region features and use the scene graph detector of Motif-Net to extract an image scene graph. Then the extracted features are fed to our dual-path relation encoder to learn the implicit relation-aware and explicit relation-aware attended features. In the next step, the embedding of the partially complete sentence serves as the context to help our adaptive context relation decoder to fuse two types of relation-aware features adaptively. Based on the fused visual features, our decoder takes into account implicit relations and explicit relations in this way to generate the remaining tokens. We will introduce more details of the components of the DRAN in the following sections.

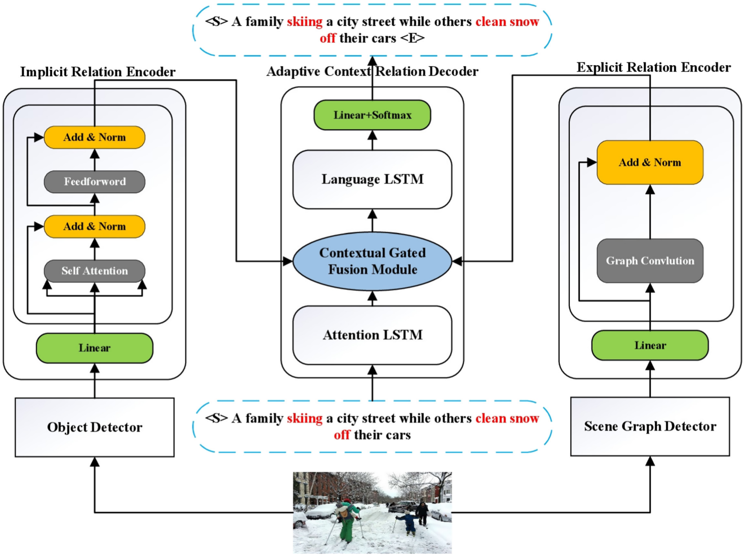

<S> A family skiing a city street while others clean snow off their cars <E>
Implicit Relation Encoder
Adaptive Context Relation Decoder
Explicit Relation Encoder
Linear+Softmax
Add & Norm
Feedforword
Add & Norm
Self Attention
Linear
Language LSTM
Contextual Gated Fusion Module
Attention LSTM
Add & Norm
Graph Convlution
Linear
Object Detector
<S> A family skiing a city street while others clean snow off their cars
Scene Graph Detector

Fig. 1. Overview of the proposed DRAN which composes of a dual-path relation encoder and an adaptive context relation decoder. The input of the implicit relation encoder (left) is the output of the object detector. The input of the explicit relation encoder (right) is the output of the scene graph detector. The word embedding of the partially complete sentence is fed to the Attention LSTM in the decoder to get a sentence representation. The outputs of the dual-path relation encoder and the sentence representation are then fed to the contextual gated fusion module in the adaptive context relation decoder (middle) to help the decoder to generate the next token.

2.2 Dual Path Relation Encoder

The dual-path relation encoder consists of an implicit relation encoder and an explicit relation encoder. We consider two types of relations for our model to learn, namely, implicit relations and explicit relations. Implicit relations are such kinds of relations that can be inferred from the appearance features of objects in images. For example, “next to” is a kind of positional relation that can be inferred from a close distance between two objects. And we believe that there are more implicit relations beyond positional relations that can be inferred from the appearance features of objects in images, such as actions or interactions of two objects. While explicit relations are relations detected by our scene graph detector and then directly encoded in word embedding of relation labels. To promote the diversity of relations, the explicit relation encoder is a complementary method to discover relations between objects.

Implicit Relation Encoder. Self-attention operation in Transformer can learn implicit relations and finally reflect in the attended version of object features, which makes Transformer a good implicit relation encoder. Since objects that are more similar in appearance tend to have stronger connections, we introduce the Transformer as our implicit encoder to discover the implicit relations between objects. We modify the Transformer encoder by applying a layer-wise residual connection between Transformer blocks, which makes our encoder network deeper, thus increasing the possibility of more abstract relations being learned. Specifically, we denote the extracted object features as X_{obj}, then the implicit encoder is given formally by:

$$\begin{aligned} X_Q &= X_K = X_V = X_{obj}, \\ \hat{X}_V &= Attention(W^Q X_Q, W^K X_K, W^V X_V) \\ V_{imp} &= X_{obj} + \hat{X}_V, \end{aligned} \tag{2}$$

where

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V,$$

and X_Q, X_K, X_V, W^Q, W^K, W^V, d_k are standard Transformer components. Multi-head self-attention, layer normalization, point-wise feed-forward layer, and another layer normalization operation in a Transformer block are omitted for the simplicity of description. Details about Transformer can be found in [18]. The output of the implicit relation encoder is the attended version of object features, in which the implicit relations are softly encoded.

Explicit Relation Encoder. Due to the ability to aggregate information from neighboring nodes, we introduce the graph convolutional operation in our explicit relation encoder to learn explicit relations between objects. The aggregated information from the neighborhood help refines the explicit relations. In this research, we extend the graph convolution so that it can perform convolution operations on graphs with edge features. Specifically, a node feature is updated with the sum of incoming edge features when the node is a subject and with the sum of outgoing edge features when a node is an object and its own feature. Correspondingly, an edge feature is updated with the sum of the features of its connected nodes and its own feature. Firstly, We denote the node feature matrix as $X_n^f = [x_n^f] \in \mathbb{R}^{d_f \times |N|}$ and the edge feature matrix as $X_e^f = [x_e^f] \in \mathbb{R}^{d_f \times |E|}$. Then aggregation and update operation is given formally by:

$$\begin{aligned} X_{e\_nbr}^f &= ReLU(W_{rs} X_e^f A_{rs}) + ReLU(W_{ro} X_e^f A_{ro}), \\ \hat{X}_n^f &= X_n^f + X_{e\_nbr}^f, \\ X_{n\_nbr}^f &= ReLU(W_{sr} X_n^f A_{sr}) + ReLU(W_{or} X_n^f A_{or}), \\ \hat{X}_e^f &= X_e^f + X_{n\_nbr}^f, \end{aligned} \tag{3}$$

where $W_{rs}, W_{ro}, W_{sr}, W_{or} \in \mathbb{R}^{d^f \times d^f}$ are learnable parameters that connect relation features with object or subject features and X_{e_nbr}, X_{n_nbr} are the aggregated neighboring features from edges and nodes, respectively. $A_{rs}, A_{ro} \in$

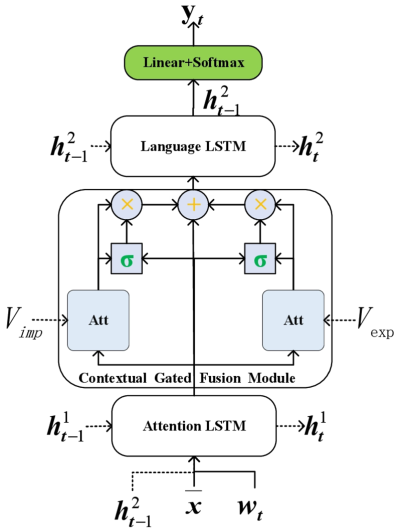


Fig. 2. Overview of the adaptive context relation decoder which composes of an attention LSTM a contextual gated fusion module and a language LSTM. The input of attention operation in contextual gated fusion module are the two types of relation-aware features from the dual-path relation encoder and the output of the attention LSTM.

$\mathbb{R}^{|E|\times|N|}$, $A_{sr}, A_{or} \in \mathbb{R}^{|N|\times|E|}$ are the normalized adjacency matrix between relations and subjects, between relations and objects, between subjects and relations and between objects and relations, respectively. $\hat{X}_n^f, \hat{X}_e^f$ are the updated node features and edge features, respectively. Then, a stack of graph convolution layers makes up our explicit relation encoder. The output of the explicit relation encoder is the updated and explicit relation-aware node features, which would be further used in our decoder.

2.3 Adaptive Context Relation Decoder

Based on the top-down attention LSTM [2], we design an adaptive context relation decoder which is composed of an attention-based LSTM, a contextual gated fusion module, and a language LSTM. The overall structure of our adaptive context relation decoder and the details of the contextual gated fusion module are shown in Fig. 2. We claim that our decoder has an adaptive functionality because, during the training process, our decoder learns to determine the fusion ratio of the implicit relation-aware features and the explicit relation-aware features to help the model understand and predict more reasonable relations between objects. When decoding the next word, the inputs of our decoder are the representation of the partially complete sentences and two types of aforementioned relation-aware features which are the outputs of the implicit encoder and explicit encoder, respectively.

Attention-Based LSTM. We denote the outputs of implicit relation encoder and explicit relation encoder as V_{imp}, V_{exp}, respectively. At time step t, the input of the Attention LSTM which is denoted as $LSTM_A$ is composed of the following four parts: the previous hidden state of the Attention LSTM, the previous output of the Language LSTM, the mean-pooled visual features $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} X_i$, which is a global representation of input image and the word embedding of the input token at time step t. Formally, the attention-based LSTM is given by:

$$\begin{aligned} h_t^1 &= LSTM_A([h_{t-1}^1, h_{t-1}^2, \bar{x}, w_t]), \\ \alpha_t^1 &= softmax(W_{a1}^T tanh(W_{vi}V_{imp} + W_{hi}h_t^1)), \\ \alpha_t^2 &= softmax(W_{a2}^T tanh(W_{ve}V_{exp} + W_{he}h_t^1)), \\ \hat{V}_{imp} &= \alpha_t^1 V_{imp}, \\ \hat{V}_{exp} &= \alpha_t^2 V_{exp}, \end{aligned} \tag{4}$$

where $h_{t-1}^1, h_{t-1}^2, \bar{x}, w_t$ are the previously mentioned four parts of the input of the top LSTM and $W_{a1}^T, W_{a2}^T, W_{vi}, W_{ve}, W_{hi}, W_{he}$ are all learned parameters.

Contextual Gated Fusion Module. We design a contextual gated fusion module for the decoder to adaptively fuse two types of learned relation-aware features. When predicting the next word, this kind of gating control allows the decoder to adaptively attend to more implicit relations or more explicit relations based on the contextual information. Formally, the contextual gated fusion module is given by:

$$\begin{aligned} \alpha_1 &= \sigma(W_h^1 h_t^1 + W_v^1 \hat{V}_{imp}), \\ \alpha_2 &= \sigma(W_h^2 h_t^1 + W_v^2 \hat{V}_{exp}), \\ \hat{V}_f &= \frac{\alpha_1 \times \hat{V}_{imp} + \alpha_2 \times \hat{V}_{exp}}{\sqrt{2}} \end{aligned} \tag{5}$$

where h_t^1 are the output of the Attention LSTM and $W_h^1, W_h^2, W_v^1, W_v^2$ are learned parameters. $\hat{V}_f$ is the fused visual features.

Language LSTM and Objective. The language LSTM and a generator layer which consists of a linear layer and a softmax activation generate a distribution of words. Then the distribution of the entire sentence is generated as the product of a series of conditional distributions at all time steps. Given a ground truth sequence $Y_{1:T}^*$ and an image captioning model with parameter θ, we optimize the cross-entropy loss:

$$Loss_{XE}(\theta) = -\sum_{t=1}^{T} log(p_\theta(Y_t^* | Y_{1:t-1}^*)). \tag{6}$$

3 Experiments

3.1 Datasets

All our experiments are conducted on the benchmark dataset of image captioning, MSCOCO [12], which contains 123,287 images. Each image has 5 different artificially labeled captions. We adopt the Karpathy splits [8] as [2,21,22,24] used for our offline testing. Specifically, 113,287 images are used for training, and 5,000 images are used for validation and testing respectively. For pre-processing, we convert all sentences to lower case, tokenize and discard words that occur less than 5 times, and trim each caption to a maximum of 16 words, resulting in the final vocabulary of 9,487 words.

3.2 Metrics

Standard automatic evaluation metrics including BLEU@N [14], METEOR [17], ROUGE-L [11], CIDEr [19], and SPICE [1] are used to evaluate the quality of generated captions. A higher value suggests a better performance for all those above-mentioned metrics.

3.3 Settings

To extract object features, we use the pre-trained Faster-RCNN [16] with ResNet-101 from [2]. To extract image scene graphs, we follow previous work [24], in which a Motif-Net [23] is trained on Visual Genome [10] with 1600/200 object/relation classes. We apply the detector for each image and 36/64 objects/triplets are kept in the scene graph. 2048D visual features and 300D GloVe [15] word embedding for labels of nodes and edges are projected into 1024D before feeding to encoders. For the Transformer encoder, we set the number of heads to 8 and d_{model} to 1024. The number of encoder layers is set to 4. For the GCN encoder, we set GCN_{dim} to 1024 and the number of layers to 4. For training, we use Adam [9] optimizer with an initial learning rate of 0.0005 and a mini-batch of 100 images. Beam search is adopted in the decoding phase with beam size 3. Testing results are reported on Karpathy split of the MSCOCO [12] dataset using cross-entropy loss optimization and without using reinforcement learning optimization.

3.4 Performance Comparison

Table 1 shows the performance comparison between the proposed DRAN and state-of-the-art methods in recent three years. The models we compared include Up-Down [2], GCN-LSTM [22], SGAE [21], Sub-GC [24], among which the last three also take advantage of image scene graphs like us. We present the results of our model with the concatenation fusion method and with the gated fusion method for fusing two types of relation-aware features.

Table 1. Comparison results on MSCOCO caption dataset using Karpathy split. B@1, B@4, M, R, C, and S represent BLEU-1, BLEU-4, METEOR, ROUGE-L, CIDEr, and SPICE scores respectively. The same below.

Method	B@1	B@4	M	R	C	S
Up-Down [2]	77.2	36.2	27.0	56.4	113.5	20.3
GCN-LSTM [22]	77.3	36.8	27.9	57.0	116.3	20.9
SGAE [21]	**77.6**	36.9	27.7	**57.2**	**116.7**	20.9
Sub-GC [24]	76.8	36.2	27.7	56.6	115.3	20.7
DRAN (w/concat)	76.8	36.6	28.0	57.1	115.5	21.0
DRAN (w/gated fusion)	76.9	**37.1**	**28.1**	**57.2**	116.1	**21.1**

As shown in Table 1, our DRAN model surpasses all other approaches in terms of BLEU@4, METEOR, and SPICE, and shares the best with [21] on ROUGE-L while achieving comparable performance on BLEU@1 and CIDEr compared to others. Note that two different fusion methods of our model achieve the first and second place in performance on SPICE respectively, which is a metric measuring how well caption models recover objects, attributes, and relations. It proves that our model captures more semantic relation information than other methods.

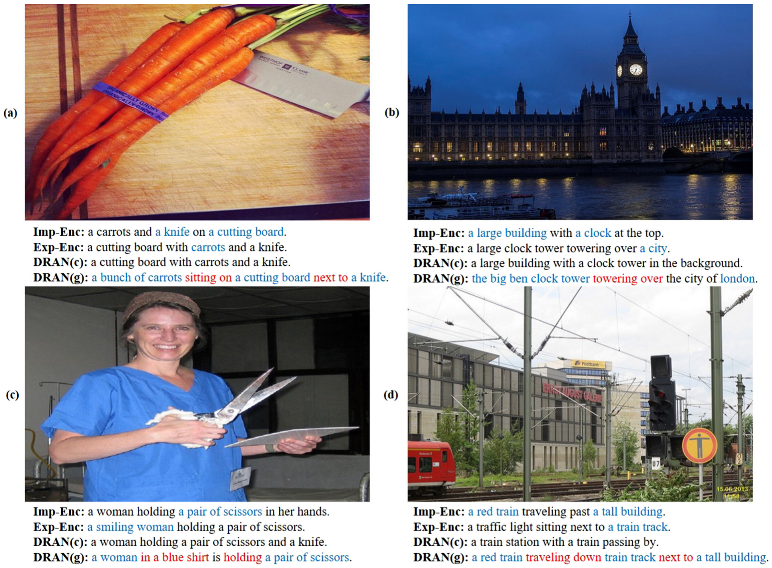
Table 2. Ablation study on different variants of our model

Model	B@1	B@4	M	R	C	S
Implicit relation encoder	76.4	36.3	27.8	56.7	114.6	20.9
Explicit relation encoder	76.2	36.4	27.7	56.6	114.0	20.8
DRAN (Concatenation)	76.8	36.6	28.0	57.1	115.5	21.0
DRAN (Gated fusion)	**76.9**	**37.1**	**28.1**	**57.2**	**116.1**	**21.1**

3.5 Ablation Study

We conduct several ablation studies to further verify the effectiveness of each component of our method. Following the previous finding, we set the layer number of the encoder to 4 and conduct the ablation study on different variants of the model, as follows:

- **Implicit Relation Encoder:** It only uses the implicit relation encoder, i.e., the Transformer-based encoder and the decoder of LSTM.
- **Explicit Relation Encoder:** It only uses the explicit relation encoder, i.e., the GCN-based encoder and the decoder of LSTM.



Imp-Enc: a carrots and a knife on a cutting board.
Exp-Enc: a cutting board with carrots and a knife.
DRAN(c): a cutting board with carrots and a knife.
DRAN(g): a bunch of carrots sitting on a cutting board next to a knife.

Imp-Enc: a large building with a clock at the top.
Exp-Enc: a large clock tower towering over a city.
DRAN(c): a large building with a clock tower in the background.
DRAN(g): the big ben clock tower towering over the city of london.

Imp-Enc: a woman holding a pair of scissors in her hands.
Exp-Enc: a smiling woman holding a pair of scissors.
DRAN(c): a woman holding a pair of scissors and a knife.
DRAN(g): a woman in a blue shirt is holding a pair of scissors.

Imp-Enc: a red train traveling past a tall building.
Exp-Enc: a traffic light sitting next to a train track.
DRAN(c): a train station with a train passing by.
DRAN(g): a red train traveling down train track next to a tall building.

Fig. 3. Examples of image captioning results generate by different variants of our model. The captured implicit and explicit relations are marked in red and the captured details from the two encoders are marked in blue. (Color figure online)

- **DRAN + Concatenation:** It uses simultaneously the dual-path relation encoder and the adaptive context relation decoder, with the concatenation fusion method to fuse two types of relation-aware visual features.
- **DRAN + Gated Fusion:** It uses simultaneously the dual-path relation encoder and the adaptive context relation decoder, but with the contextual gated fusion method to fuse two types of relation-aware visual features.

As shown in Table 2, simultaneous using dual-path relation encoders achieve better performance than using one of them alone on all the metrics, regardless of whether the fusion method is concatenation or contextual gated fusion. When comparing two different fusion methods, the contextual gated fusion is a better one to fuse the implicit and explicit relation-aware features, which proves that our gated fusion method can adaptively fuse more accurate semantic relation information to generate the better captions.

Several visual examples are illustrated in Fig. 3. We present the captioning results of the Implicit Relation Encoder model, the Explicit Relation Encoder model, the DRAN(concatenation) model, and the DRAN(gated fusion) model respectively under each image of examples. We denote these four models as **Imp-Enc**, **Exp-Enc**, **DRAN(c)**, **DRAN(g)**, respectively. As suggested by these examples, the proposed DRAN can generally capture more high-level semantic relation information and generate more accurate and semantically rich descriptions. For example, in Fig. 3, the single encoder model typically captures a list of detected objects while the dual-path encoder model capture semantic relations like “sitting on” and “next to” in (a), “towering over” in (b), “traveling down” and “next to” in (d). We infer that relations related to position or direction like “next to”, “on top of” are typically learned by the implicit relation encoder while relations composed of verbs are typically learned by the explicit relation encoder, both of which may occur in the process of dynamically fusing visual features for decoding sentences. In addition, we find that the dual-path encoder model with the gated fusion method can also capture and fuse richer and more precise details like attributes of an object from the two encoders respectively. For example, “a bunch of” in (a), “the big ben clock” and “london” in (b), “a red train” and “train track” in (d), “a blue shirt” in (c) and so on. The captured implicit and explicit relations are marked in red and the captured details from the two encoders are marked in blue in Fig. 3.

4 Conclusion

In this paper, we propose a novel image captioning model named DRAN which composes of a dual-path relation encoder and an adaptive context relation decoder. The problem of relation mismatching between sentences and images is alleviated by our relation-aware encoder-decoder framework. Specifically, the dual-path encoder learns to encode implicit relations and explicit relations into relation-aware features. Then the contextual gated fusion module we devised adaptively fuse these two types of features to help the decoder generate semantically richer captions. Experimental results on the MSCOCO dataset and visual examples we illustrated indicate that our method captures more semantic relations and details quantitatively and qualitatively.

Acknowledgement. This work is partially supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LR21F020002 and Natural Science Foundation of China under Grant No. U20A20196, 61976192. The source code of this work will be released on https://github.com/Zjut-MultimediaPlus.

References

1. Anderson, P., Fernando, B., Johnson, M., Gould, S.: SPICE: semantic propositional image caption evaluation. Adapt. Behav. **11**(4), 382–398 (2016)
2. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering (2017)
3. Bai, C., Huang, L., Chen, J.N., Pan, X., Chen, S.Y.: Optimization of deep convolutional neural network for large scale image classification. J. Softw. **29**, 1029–1038 (2018)
4. Bai, C., Zheng, A., Huang, Y., Pan, X., Chen, N.: Boosting convolutional image captioning with semantic content and visual relationship. Displays **70**, 102069 (2021)
5. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)

6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. IEEE (2016)
7. He, S., Liao, W., Tavakoli, H.R., Yang, M., Rosenhahn, B., Pugeault, N.: Image captioning through image transformer. arXiv (2020)
8. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 664–676 (2016)
9. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv e-prints (2014)
10. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Li, F.F.: Visual genome: connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis. **123**(1), 32–73 (2017)
11. Lin, C.Y.: ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out, Barcelona, Spain, pp. 74–81. Association for Computational Linguistics, July 2004
12. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
13. Liu, F., Liu, Y., Ren, X., He, X., Sun, X.: Aligning visual regions and textual concepts for semantic-grounded image representations. In: 33rd Conference on Neural Information Processing Systems (NeurIPS 2019) (2019)
14. Papineni, K.: BLEU: a method for automatic evaluation of MT. Research Report, Computer Science RC22176 (W0109–022) (2001)
15. Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation. In: Conference on Empirical Methods in Natural Language Processing (2014)
16. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks (2017)
17. Satanjeev, B.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: ACL-2005, pp. 228–231 (2005)
18. Vaswani, A., et al.: Attention is all you need. arXiv (2017)
19. Vedantam, R., Zitnick, C.L., Parikh, D.: CIDEr: consensus-based image description evaluation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
20. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: lessons learned from the 2015 MSCOCO image captioning challenge. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 652–663 (2016)
21. Yang, X., Tang, K., Zhang, H., Cai, J.: Auto-encoding scene graphs for image captioning (2018)
22. Yao, T., Pan, Y., Li, Y., Mei, T.: Exploring visual relationship for image captioning. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. LNCS, vol. 11218, pp. 711–727. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01264-9_42
23. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph parsing with global context (2017)
24. Zhong, Y., Wang, L., Chen, J., Yu, D., Li, Y.: Comprehensive image captioning via scene graph decomposition. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12359, pp. 211–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58568-6_13