



Relational Attention with Textual Enhanced Transformer for Image Captioning

Lifei Song¹(✉), Yiwen Shi², Xinyu Xiao³, Chunxia Zhang⁴, and Shiming Xiang³

¹ School of Artificial Intelligence, University of Chinese Academy of Sciences,
Beijing 100049, China
songlifei2018@ia.ac.cn

² Beijing City University Intelligent Electronic Manufacturing Research Center,
Beijing 101309, China

³ National Laboratory of Pattern recognition, Institute of Automation of Chinese
Academy of Sciences, Beijing 100190, China
{xinyu.xiao,smxiang}@nlpr.ia.ac.cn

⁴ School of Computer Science and Technology, Beijing Institute of Technology,
Beijing 100081, China
cxzhang@bit.edu.cn

Abstract. Image captioning has attracted extensive research interests in recent years, which aims to generate a natural language description of an image. However, many approaches focus only on individual target object information without exploring the relationship between objects and the surrounding. It will greatly affect the performance of captioning models. In order to solve the above issue, we propose a relation model to incorporate relational information between objects from different levels into the captioning model, including low-level box proposals and high-level region features. Moreover, Transformer-based architectures have shown great success in image captioning, where image regions are encoded and then attended into attention vectors to guide the caption generation. However, the attention vectors only contain image-level information without considering the textual information, which fails to expand the capability of captioning in both visual and textual domains. In this paper, we introduce a Textual Enhanced Transformer (TET) to enable addition of textual information into Transformer. There are two modules in TET: text-guided Transformer and self-attention Transformer. The two modules perform semantic and visual attention to guide the decoder to generate high-quality captions. We extensively evaluate model on MS COCO dataset and it achieves 128.7 CIDEr-D score on Karpathy split and 126.3 CIDEr-D (c40) score on official online evaluation server.

This research was supported by the National Key Research and Development Program of China under Grant No. 2020AAA0104903, and the National Natural Science Foundation of China under Grants 62072039, 62076242, and 61976208.

Keywords: Relational information · Attention · Textual enhanced transformer

1 Introduction

Image captioning [1–3] aims to automatically describe an image with natural language. It is a task at the intersection connecting computer vision and nature language processing. It is particularly challenging because it requires to simultaneously capture information of target objects as well as their relationships. Inspired by the development of neural machine translation [4], existing research on image captioning usually employs an encoder-decoder architecture with a Convolutional Neural Network (CNN) used for image feature extraction and a Recurrent Neural Network (RNN) for caption generation. The attention mechanism [5–7] also plays a crucial role in image captioning task, instead of transferring an entire image to a single representation, visual attention allows to focus on features relevant for captions.

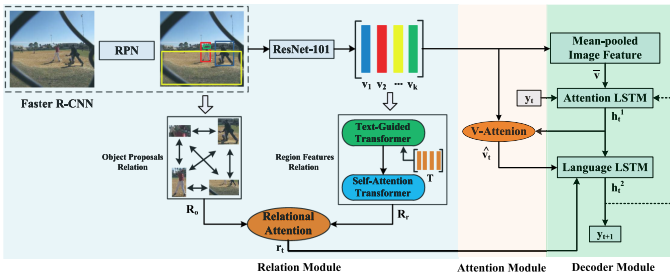


Fig. 1. The overall architecture of RA-TET. Our model consists of three modules. (1) In relation module, object proposals and region features are detected, then explores relational matrix R_o and R_r based on low-level and high-level information. (2) In attention module, V-attention is designed to obtain attended visual features \hat{v}_t based on image features V . (3) In LSTM decoder, decoding word by word y_t based on attended visual features and attended relational features so as to generate the final caption.

Despite impressive successes, there are still limitations in the current captioning frameworks. In fact, understanding inter relationships between objects facilitate a major component in visual understanding. Current methods focus on obtaining information of each object individually, but overlook existing relationships with other objects or the environment. Although some methods [8, 9] based on exploring the connections between objects have been proposed, but these methods only employ Graph Convolutional Networks (GCN) [10] to integrate object relationships into image encoder, completely replace the previous image input, and the performance of the result is limited. Considering that the structure of CNN is hierarchical and features are propagated through the network layer by layer, more efficient relational features can be derived from the

hierarchical information of the network. As shown in Fig. 1, a relational attention network is proposed in our paper, aiming at extracting relevant relations from two levels of information (low-level object proposals, high-level region features). For object proposals, location and size information of each object proposal (generated by a Region Proposal Network, RPN) are extracted to generate low-level features. For region features, the TET encourages our model to explore the relations between image regions and detected texts to generate high-level features. Afterwards, low-level and high-level features processed by a relational attention network to obtain attended relational features. Finally, the attended relational features combined with attended visual features which is obtained by visual attention(V-Attention) model to decoder model to guide the final caption generation.

In summary, our contributions are as follows:

- To address the issue of relation information missing, a novel relation model is proposed to explore relations from two levels of image information, *i.e.* low-level object proposals and high-level image region features.
- To expand the capability of complex multi-modal reasoning in image captioning, TET is designed to model interactions between image regions and detected texts.
- Experiments show that the proposed method can achieve competitive performance in MS COCO dataset compared with other state-of-the-art models, e.g. 128.7 CIDEr-D score on Karpathy split and 126.3 CIDEr-D (c40) score on official online evaluation server.

2 Related Work

2.1 Relationship Exploration

Recently, a few works have attempted to utilize relations between objects from images to improve the performance of image captioning. For example, Kim et al. [8] introduced a relational captioning, which aims to generate multiple captions about relational information between objects. To explore the relationship between objects and semantic attributes, Yao et al. [11] applied two kinds of graph convolution neural networks in the encoding stage. Moreover, Yang et al. [9] leveraged an image scene graph and a semantic scene graph which incorporate the language inductive bias into the image captioning framework. Wang et al. [28] proposed a hierarchical attention network that enables attention to be calculated on pyramidal hierarchy of features. Our method explores relations based on low-level and high-level image information, adding more relationship information to improve the performance of the model.

2.2 Transformer Architecture

The Transformer architecture is introduced in [12], which is based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. It

has advantages of low computational complexity, parallel computing and better learning about long-range distance dependence. For example, Li et al. [13] developed EnTangled Attention (ETA) that enables the transformer to exploit semantic and visual information simultaneously. In addition, Simao et al. [14] proposed a Object Relation Transformer, which incorporates information about the spatial relationship between detected objects through geometric attention. Our ETE seeks to model textual and visual information based on the input of detected texts and image features, which could be more comprehensive for caption generation.

3 The Proposed Approach

As illustrated in Fig. 1, our model consists of three modules: a Relation Module, an Attention Module, and a Decoder Module. In relation module, Faster R-CNN [15] with ResNet-101 [16] is used for object detection and feature extraction. Specially, RPN is leveraged to generate m object bounding boxes (non-maximum suppression with an intersection-over-union threshold is used to select the top box proposals), the location and size information of each object proposal can be obtained. Then, we take outputs of the *pool5* layer from ResNet-101 as image region features $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^m$, where $\mathbf{v}_i \in \mathbb{R}^{d_v}$ denotes the d_v -dimensional feature vector. Moreover, two relation modules are designed to extract relational information based on the above two sources of information. Finally, additional relational attention networks are proposed to extract attended relational features. In attention module, V-Attention is proposed to predict attended visual features based on image region features. In decoder module, two-layer LSTMs decoder is designed to generate the caption, which is based on attended relational features and attended visual features.

3.1 Relation Module

Object Proposals Relation. Each box proposal predicted by RPN is described by its location and size information: the center coordinates of bounding box $c_i = (x_i, y_i)$, width w_i and height h_i , with $i = 1, 2, \dots, m$. For two object proposals i and j , the relational vector o_{ij} can be defined as:

$$\mathbf{o}_{ij} = \left[\frac{|x_i - x_j|}{H}, \frac{|y_i - y_j|}{W}, \frac{h_i}{h_j}, \frac{w_i}{w_j} \right], \quad (1)$$

where W and H are the width and height of the image. By stacking these vectors together, a low-level feature relational matrix $\mathbf{R}_o \in \mathbb{R}^{m(m-1) \times 4}$ of object proposals can be created.

Region Features Relation. To process high-level region features, we employ TET to convert fixed number of image features into the unified relational feature representation. The Transformer module includes self-attention and multi-head attention. For self-attention, inputs consist of queries $\mathbf{Q} = (\mathbf{q}_1, \dots, \mathbf{q}_m)$,

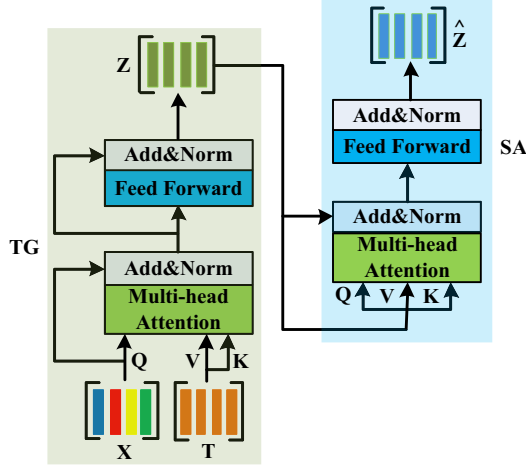


Fig. 2. The structure of TET. Two basic units with different types of inputs. Firstly, the TG unit takes the input of image features \mathbf{V} and text features \mathbf{T} , and outputs the features \mathbf{Z} . Then, the SA unit takes the input of attended features \mathbf{Z} and outputs the attended features $\hat{\mathbf{Z}}$.

keys $\mathbf{K} = (\mathbf{k}_1, \dots, \mathbf{k}_m)$ and values $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m)$, $\mathbf{q}_i, \mathbf{k}_i, \mathbf{v}_i \in \mathbb{R}^{d_v}$. The dot products of the query is computed with all keys, divided each by $\sqrt{d_v}$. Finally, a softmax function is applied to obtain the weights, the process can be seen as follows:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Softmax\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}. \quad (2)$$

To extend the capacity of exploring subspaces, the multi-head attention which consists of h parallel scaled dot-product attentions is adopted. The inputs include queries, keys, and values which are projected into h subspaces, and the attention is performed in the subspaces separately. Then, h heads are concatenated and linearly projected to the feature set:

$$MultiHead(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Concat(\mathbf{H}_1, \dots, \mathbf{H}_h)\mathbf{W}^O, \quad (3)$$

$$\mathbf{H}_i = Attention(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V), \quad (4)$$

where $\mathbf{W}^O \in \mathbb{R}^{d_v \times d_v}$ denotes the linear transformation. $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V \in \mathbb{R}^{\frac{d_v}{h} \times d_v}$ are independent head projection matrices, $i = 1, 2, \dots, h$. In addition to attention sub-layers, each of attention layer contains a point-wise feed-forward network (\mathcal{FFN}), which consists of two linear transformations with a ReLU activation.

$$\mathcal{FFN}(\mathbf{x}) = max(0, \mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2, \quad (5)$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_v \times m}$, $\mathbf{W}_2 \in \mathbb{R}^{m \times d_v}$, \mathbf{b}_1 and \mathbf{b}_2 are weights and biases of two fully connected layers.

Different from the original Transformer model, the ETE model is guided by detected texts, in order to associate text features with image features. As illustrated in Fig. 2, there are two units in our ETE: text-guided Transformer (TG) and self-attention Transformer (SA). In the TG unit, the input of image features $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m)$ is transformed into queries matrix \mathbf{Q}_v , and text features $\mathbf{T} = \{\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_n\}$ which are detected by the method [17], is transformed into keys matrix \mathbf{K}_t and values matrix \mathbf{V}_t . The calculation is as follows:

$$\mathbf{Q}_v, \mathbf{K}_t, \mathbf{V}_t = \text{Linear}(\mathbf{V}, \mathbf{T}, \mathbf{T}). \quad (6)$$

Then, with focusing on relevant image regions through detected texts, multi-head attention is applied. And the output is computed by residual connection, which is followed by layer normalization:

$$\mathbf{F}_{t \rightarrow v} = \mathcal{LN}(\mathbf{V} + \text{MultiHead}(\mathbf{Q}_v, \mathbf{K}_t, \mathbf{V}_t)), \quad (7)$$

where $\mathbf{F}_{t \rightarrow v}$ is the output features of images guided by text features, \mathcal{LN} represents layer normalization. Then, \mathcal{LN} layer and residual connection are used to obtain the attended image features $\tilde{\mathbf{V}}$:

$$\mathbf{F}'_{t \rightarrow v} = \mathcal{FFN}(\mathbf{F}_{t \rightarrow v}), \quad (8)$$

$$\tilde{\mathbf{V}} = \mathcal{LN}(\mathbf{F}_{t \rightarrow v} + \mathbf{F}'_{t \rightarrow v}). \quad (9)$$

In the SA unit, given the attended image features $\tilde{\mathbf{V}}$ guided by text features, then transformed into queries matrix $\mathbf{Q}_{\tilde{v}}$, keys matrix $\mathbf{K}_{\tilde{v}}$ and values matrix $\mathbf{V}_{\tilde{v}}$ through linear layers:

$$\mathbf{Q}_{\tilde{v}}, \mathbf{K}_{\tilde{v}}, \mathbf{V}_{\tilde{v}} = \text{Linear}(\tilde{\mathbf{V}}, \tilde{\mathbf{V}}, \tilde{\mathbf{V}}). \quad (10)$$

Specially, residual connection followed by layer normalization is applied to the multi-head attention as follows:

$$\mathbf{F}_{\tilde{v} \rightarrow \tilde{v}} = \mathcal{LN}(\tilde{\mathbf{V}} + \text{MultiHead}(\mathbf{Q}_{\tilde{v}}, \mathbf{K}_{\tilde{v}}, \mathbf{V}_{\tilde{v}})). \quad (11)$$

Then, with the optimization mentioned above, the relational matrix \mathbf{R}_r of image regions are obtained as follows:

$$\mathbf{F}'_{\tilde{v} \rightarrow \tilde{v}} = \mathcal{FFN}(\mathbf{F}_{\tilde{v} \rightarrow \tilde{v}}), \quad (12)$$

$$\mathbf{R}_r = \mathcal{LN}(\mathbf{F}_{\tilde{v} \rightarrow \tilde{v}} + \mathbf{F}'_{\tilde{v} \rightarrow \tilde{v}}), \quad (13)$$

the final output is $\mathbf{R}_r \in \mathbb{R}^{m \times d_v}$.

Based on two kinds of relational features, we adopt relational attention networks Att_{obj} and Att_{reg} to transform relational feature matrix outputs into two attended relational feature vectors:

$$\hat{\mathbf{r}}_{ot} = Att_{obj}(\mathbf{R}_o, \mathbf{h}_t^1), \quad (14)$$

$$\hat{\mathbf{r}}_{rt} = Att_{reg}(\mathbf{R}_r, \mathbf{h}_t^1), \quad (15)$$

where \mathbf{h}_t^1 represents the hidden state of attention LSTM, the calculation of Att_{obj} and Att_{reg} is consistent with the following calculation process of V-Attention, which is shown as the Eqs. (17) and (18).

After obtaining two attended relational feature vectors, the final attended relational feature vector can be concatenated as:

$$\mathbf{r}_t = Concat(\hat{\mathbf{r}}_{ot}, \hat{\mathbf{r}}_{rt}), \quad (16)$$

the final output \mathbf{r}_t will be sent to language LSTM as input.

3.2 Attention Module

The V-Attention, which is widely used in other attention methods, could focus on the image features that are most relevant to words at the current time step. Specially, given the image region features \mathbf{v}_i and hidden state \mathbf{h}_t^1 of attention LSTM, a single-layer neural network followed by a softmax layer is applied as V-Attention to obtain attention weights α_t :

$$\mathbf{a}_{it} = \omega_{\mathbf{h}}^T \tanh(\mathbf{W}_v \mathbf{v}_i + \mathbf{W}_h \mathbf{h}_t^1), \quad (17)$$

$$\alpha_t = softmax(\mathbf{a}_t), \quad (18)$$

where $\mathbf{W}_v \in \mathbb{R}^{H \times d_v}$, $\mathbf{W}_h \in \mathbb{R}^{H \times d_h}$, and $\omega_{\mathbf{h}} \in \mathbb{R}^H$ are parameters to be learned. Based on the weight distribution, attended image region feature $\hat{\mathbf{v}}_t$ can be calculated by weighted summing at the current time step t :

$$\hat{\mathbf{v}}_t = \sum_{i=1}^m \alpha_{it} \mathbf{v}_i. \quad (19)$$

3.3 Decoder Module

Based on the final attended relational feature vector \mathbf{r}_t and attended image feature vector $\hat{\mathbf{v}}_t$, decoder module uses a two-layer LSTM decoder, namely attention LSTM and language LSTM, to guide the process of generating captions sequentially. The input vector of attention LSTM at each time step \mathbf{x}_t^1 consists of mean-pooled image feature $\bar{\mathbf{v}} = \frac{1}{m} \sum_{i=1}^m \mathbf{v}_i$, the encoding of the previously generated word \mathbf{y}_{t-1} , and the previous output \mathbf{h}_{t-1}^2 of language LSTM:

$$\mathbf{x}_t^1 = [\bar{\mathbf{v}}, \mathbf{W}_{e1} \mathbf{y}_{t-1}, \mathbf{h}_{t-1}^2], \quad (20)$$

$$\mathbf{h}_t^1 = LSTM_{att}[\mathbf{x}_t^1, \mathbf{h}_{t-1}^1], \quad (21)$$

where \mathbf{W}_{e1} is a word embedding matrix for a vocabulary Σ .

Then, the output \mathbf{h}_t^1 of attention LSTM, attended relational features \mathbf{r}_t and attended image features $\hat{\mathbf{v}}_{ct}$ are used as input to the language LSTM, given by:

$$\mathbf{x}_t^2 = [\hat{\mathbf{v}}_t, \mathbf{h}_t^1, \mathbf{r}_t], \quad (22)$$

$$\mathbf{h}_t^2 = LSTM_{lan}[\mathbf{x}_t^2, \mathbf{h}_{t-1}^2]. \quad (23)$$

We model hidden state \mathbf{h}_t^2 of language LSTM to compute the conditional probabilities on the vocabulary:

$$p(y_t) = softmax(\mathbf{W}_h \mathbf{h}_t^2 + \mathbf{b}_h), \quad (24)$$

where $\mathbf{W}_h \in \mathbb{R}^{d_v \times d_h}$ is the weight parameters to be learnt and \mathbf{b}_h is bias. d_v is the size of whole vocabulary.

3.4 Training and Objectives

We first train our hierarchical relation attention captioning model by optimizing the Cross Entropy Loss:

$$L_{XE}(\theta) = - \sum_{t=1}^T \log(p_\theta(y_t^* | y_{1:t-1}^*)), \quad (25)$$

where $y_{1:T}^*$ denotes the ground truth word sequence.

Then we directly optimize the non-differentiable metrics with self-critical sequence training [18]:

$$L_{RL}(\theta) = -E_{y_{1:T} \sim p_\theta} [r(y_{1:T})], \quad (26)$$

where the reward r represents the score of CIDEr-D [19]. The gradients can be approximated:

$$\nabla_\theta L_{RL}(\theta) \approx -(r(y_{1:T}^s) - r(\hat{y}_{1:T})) \nabla_\theta \log p_\theta(y_{1:T}^s), \quad (27)$$

where $y_{1:T}^s$ means its a result sampled from probability distribution, while $\hat{y}_{1:T}$ indicates a result of greedy decoding.

4 Experiments

4.1 Datasets and Evaluation Metrics

MS COCO. We evaluate our proposed method on the popular MS COCO dataset [20]. MS COCO dataset contains 123,287 images, including 82,783 for training and 40,504 for validation. Each image has five human-annotated captions. For offline evaluation, we use the Karpathy split [21] for the performance comparison, where 5,000 images are used for validation, 5,000 images for testing, and the rest for training. We convert all sentences to lower case, drop the words that occur less than 5 times, and trim each caption to a maximum of 16 words, which results in a vocabulary of 10,369 words.

Evaluation Metrics. To evaluate caption quality, we use the standard automatic evaluation metrics, namely BLEU [22], METEOR [23], ROUGE-L, CIDEr-D [19], and SPICE [24], which are denoted as $B@n(n=1,2,3,4)$, MT, R-L, C-D and SP for short respectively.

4.2 Implementation Details

The size of input images is 224×224 . The dimension of the region vector d_v and the word embedding are respectively 2048 and 1024, the hidden size d_h of both two LSTM layers is set to 2048, and the dimension of attention layers H is set to 512. For each image, we set the IoU thresholds for box proposal suppression to 0.7. For the training process, we train our model under cross-entropy loss for 20 epochs, ADAM optimizer is used with mini-batch size of 64, the learning rate starts from 0.01 and after 10 epochs decays by the factor of 0.8 at every five epoch. Then we use self-critical sequence training (SCST) [18] to optimize the CIDEr-D score with Reinforcement Learning for another 20 epochs with an initial learning rate of $1e-5$, and annealed by 0.5 when the CIDEr-D score on the validation split has not improved.

4.3 Ablation Study

To show the effectiveness of different levels of relational strategies used in our framework, we conduct experiments to compare the models leveraging different image relational information, including low-level object proposals and high-level region features. The results are shown in Table 1. In the first row, we remove the attention operation and instead use two fully connected layers to get relational embedding vector. We notice that the use of region features get better performance with respect to object proposals with an improvement of 2.7% in terms of the C-D metric. The combination of the two relational information further improves the score by 2.4% compared with region features, hence demonstrating the effectiveness of using levels of relation information in our model.

To illustrate the effect of our relational attention strategy, we further carry out another ablation study and the results are exhibited in the lower column of Table 1. Compared with the previous methods that don't use relational attention, our relational attention networks all achieve an improvement in terms of the C-D metric. This is due to the ability of the relation attention mechanism to further extract object-related information on the basis of the original relational information.

We also compare the Textual Enhanced Transformer with the Original Transformer which includes no text information in Table 2. We observe a significant increase in performance when using TET, which leads to 128.7 on C-D. This is due to the semantic information added to the Transformer, and improves the language generalization ability of the model.

Table 1. The performance of ablation experiments on relational information with relational attention or not. The sign – means that we remove relational information from the model.

Attention	Relation	B@4	MT	R-L	C-D	SP
-	-	34.3	22.9	53.1	122.1	18.8
	Object Proposals	35.2	23.1	53.8	122.8	19.2
	Region Features	37.3	27.4	56.3	125.5	21.1
	Combination	38.0	27.7	57.8	127.9	21.8
Relational Attention	Object Proposals	35.9	24.0	54.2	123.9	20.1
	Region Features	37.9	27.5	57.3	126.7	21.5
	Combination	38.4	28.3	58.6	128.1	22.2

Table 2. The performance of ablation experiments on Original Transformer and Textual Enhanced Transformer

Transformer	B@4	MT	R-L	C-D	SP
Original transformer	38.4	28.3	58.6	128.1	22.2
Textual enhanced transformer	38.6	28.9	58.6	128.7	22.2

Specifically, we compared our methods with other state-of-the-art approaches, including SCST [18], Up-down [3], GCN-LSTM [11], SGAE [9], CNM [27] and ETA [13]. The experiment results on MS COCO dataset are shown in Table 3.

From the experiment results, we can observe that our RA-TET model performed significantly better than most captioning models. For instance, compared with GCN-LSTM [11] which is also the relation caption model similar to our model, our model improves the scores by a wide margin on all evaluation metrics, due to the use of relation information. Moreover, compared with CNM [27] which uses the Original Transformer, our method also has experimental result improvement due to the use of ETE.

We also submitted our RA-TET optimized with C-D score to online COCO testing server and evaluated the performance on official testing set. Table 4 shows the performance on official testing image set with 5 and 40 reference captions. Compared to other performing methods, our proposed model achieves superior performances across all the evaluation metrics on both c5 and c40 testing sets.

Figure 3 provides examples on generated captions. Compared with GCN[11], we can find that our model can have more accurate description.

4.4 Comparison with State-of-the-Art

Table 3. Performance of our model and other state-of-the-art methods on MS-COCO Karpathy test split.

Method	Cross-Entropy Loss					Self-Critical Loss				
	B@4	MT	R-L	C-D	SP	B@4	MT	R-L	C-D	SP
Adaptive [27]	33.2	26.6	–	108.5	–	–	–	–	–	–
SCST [18]	30.0	25.9	53.4	99.4	–	34.2	26.7	55.7	114.0	–
Up-Down [3]	36.2	27.0	56.4	113.5	20.3	36.3	27.7	56.9	120.1	21.4
GCN-LSTM [11]	36.8	27.9	57.0	116.3	20.9	38.2	28.5	58.3	127.6	22.0
CNM [28]	37.1	27.9	57.3	116.6	20.8	–	–	–	–	–
ETA [13]	37.8	28.4	57.4	119.3	21.6	39.9	28.9	59.0	127.6	22.6
Our Method	37.0	28.0	57.2	117.3	21.1	38.6	28.9	58.6	128.7	22.2

Table 4. Comparison of various methods on the online MS-COCO test server.

Method	B@1		B@2		B@4		MT		R-L		C-D	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Adaptive [27]	74.8	92.0	58.4	84.5	33.6	63.7	26.4	35.9	55.0	70.5	104.2	105.9
SCST [18]	78.1	93.7	61.9	86.0	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.0
Up-down [3]	80.2	95.2	64.0	88.8	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
GCN-LSTM [11]	–	–	65.5	89.3	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
CNM [28]	–	–	–	–	37.9	68.4	28.1	36.9	58.3	72.9	123.0	125.3
ETA [13]	81.2	95.0	65.5	89.0	38.9	70.2	28.6	38.0	58.6	73.9	122.1	124.4
Our Method	80.7	94.8	65.4	89.0	38.8	69.7	28.6	37.5	58.6	73.5	125.5	126.3




Image	Captions
	<p>Ground Truth: This is a bathroom with a jacuzzi, shower, sink, and toilet.</p> <p>GCN-LSTM: A bathroom with toilet, tub and a stand up shower.</p> <p>RA-TET: A bathroom with a jacuzzi, shower, sink, and toilet.</p>
	<p>Ground Truth: A baby is holding onto a piece of bread.</p> <p>GCN-LSTM: A little girl is holding a piece of bread.</p> <p>RA-TET: A little girl is holding a piece of bread with her right hand.</p>
	<p>Ground Truth: Five horses are roaming in a snowy field.</p> <p>GCN-LSTM: A group of horses stand together in a snowy field.</p> <p>RA-TET: A group of horses are roaming in a snowy field.</p>

Fig. 3. The example of generated captions of RA-TET.

5 Conclusion

In this paper, we introduce a novel captioning model, namely Relational Attention with Textual Enhanced Transformer. Our model accounts for the relationship between objects within an image. There are two relation modules that could learn dependencies from two levels of image information, including low-level object proposals and high-level region features. Moreover, a Textual Enhance Transformer is designed to infer attended information in both textual and visual domains to guide the caption generation. Extensive experiments conducted on the MS COCO dataset demonstrate the effectiveness of our framework compared to state-of-the-art approaches.

References

1. Mao, J., Xu, W., Yang, Y., Wang, J., Yuille, A.L.: Deep captioning with multimodal recurrent neural networks (m-rnn). In: ICLR (2015)
2. Jia, X., Gavves, E., Fernando, B., Tuytelaars, T.: Guiding the long-short term memory model for image caption generation. In: ICCV (2015)
3. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR (2018)
4. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence learning with neural networks. NIPS (2014)
5. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.J.: Stacked attention networks for image question answering. In: CVPR (2016)
6. Jiasen, L., Xiong, C., Parikh, D.: and Richard Socher. Adaptive attention via a visual sentinel for image captioning. CVPR, Knowing when to look (2017)
7. Chen, L., et al.: SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning. In: CVPR (2017)
8. Kim, D.-J., Choi, J., Tae-Hyun, O., Kweon, I.S.: Dense relational captioning: Triple-stream networks for relationship-based captioning. In: CVPR (2019)
9. Yang, X., Tang, K., Zhang, H., Cai, J.: Auto-encoding scene graphs for image captioning. In: CVPR (2019)
10. Hong, D., Gao, L., Yao, J., Zhang, B., Plaza, A., Chanussot, J.: Graph convolutional networks for hyperspectral image classification. CoRR, vol. abs/2008.02457 (2020)
11. Yao, T., Pan, Y., Li, Y., Mei, T.: Exploring visual relationship for image captioning. In: ECCV (2018)
12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. NIPS (2017)
13. Li, G., Zhu, L., Liu, P., Yang, Y.: Entangled transformer for image captioning. In: ICCV (2019)
14. Herdade, S., Kappeler, A., Boakye, K., Soares, J.: Image captioning, Transforming objects into words. NIPS (2019)
15. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. NIPS (2015)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
17. Fang, H., et al.: From captions to visual concepts and back. In: CVPR (2015)

18. Steven, J.: Rennie, E.M., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: CVPR (2017)
19. Vedantam, R., Zitnick, C.L., Cider, D.P.: Consensus-based image description evaluation. In: CVPR (2015)
20. Chen, X., et al.: Microsoft COCO captions: Data collection and evaluation server. vol. abs/1504.00325 (2015)
21. Karpathy, A., Li, F.-F.: Deep visual-semantic alignments for generating image descriptions. In: CVPR (2015)
22. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. ACL (2002)
23. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. ACL (2005)
24. Anderson, P., Fernando, B., Johnson, M., Gould, S.: SPICE: semantic propositional image caption evaluation. In: ECCV (2016)
25. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: CVPR (2015)
26. Huang, L., Wang, W., Chen, J., Wei, X.: Attention on attention for image captioning. In: ICCV (2019)
27. Yang, X., Zhang, H., Cai, J.: Learning to Collocate Neural Modules for Image Captioning. In: ICCV (2019)
28. Wang, W., Chen, Z., Hu, H.: Hierarchical Attention Network for Image Captioning, AAAI (2019)