Machine Intelligence Research
www.mi-research.net
DOI: 10.1007/s11633-022-1408-2

Comprehensive Relation Modelling for Image Paragraph Generation

Xianglu Zhu[1,2] Zhang Zhang[2,3] Wei Wang[2] Zilei Wang[1]

[1] Automation Department, University of Science and Technology of China, Hefei 230027, China

[2] Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

[3] University of Chinese Academy of Sciences, Beijing 100864, China

Abstract: Image paragraph generation aims to generate a long description composed of multiple sentences, which is different from traditional image captioning containing only one sentence. Most of previous methods are dedicated to extracting rich features from image regions, and ignore modelling the visual relationships. In this paper, we propose a novel method to generate a paragraph by modelling visual relationships comprehensively. First, we parse an image into a scene graph, where each node represents a specific object and each edge denotes the relationship between two objects. Second, we enrich the object features by implicitly encoding visual relationships through a graph convolutional network (GCN). We further explore high-order relations between different relation features using another graph convolutional network. In addition, we obtain the linguistic features by projecting the predicted object labels and their relationships into a semantic embedding space. With these features, we present an attention-based topic generation network to select relevant features and produce a set of topic vectors, which are then utilized to generate multiple sentences. We evaluate the proposed method on the Stanford image-paragraph dataset which is currently the only available dataset for image paragraph generation, and our method achieves competitive performance in comparison with other state-of-the-art (SOTA) methods.

Keywords: Image paragraph generation, visual relationship, scene graph, graph convolutional network (GCN), long short-term memory.

Citation: X. Zhu, Z. Zhang, W. Wang, Z. Wang. Comprehensive relation modelling for image paragraph generation. *Machine Intelligence Research*. http://doi.org/10.1007/s11633-022-1408-2

1 Introduction

Image description is a crucial task in the interdisciplinary field of computer vision and natural language processing, which is significant for video summarization and support of the blind. Recently, great successes have been achieved in image captioning with the development of deep learning. However, image captioning methods usually produce one sentence for a given image, missing rich visual contents. To overcome this shortcoming, Johnson et al.[1] proposed dense captioning to generate captions based on a set of detected salient regions. In this way, these generated captions can tell more details of an image, but they are still independent of each other, thus making up a set of cluttered and incoherent descriptions. For the purpose of addressing the weaknesses of both image captioning and dense captioning, Krause et al.[2] introduced a new task of generating a paragraph that provides a coherent natural language description for describing fine-grained details of an image. From the perspective of the generated descriptions for images, paragraph generation takes advantage of the previous tasks but avoids their shortcomings, which is a promising task of image description.

Image paragraph generation is a very challenging task that needs to generate multiple sentences with more than 60 words. A single recurrent neural network generally cannot work well for this kind of long sequence generation task due to its limited capability of long-term dependency modelling. Existing methods for image paragraph generation generally decompose a paragraph into several sentences, which are generated by a recurrent neural network (RNN) based on the corresponding topic vectors. Therefore, topic vectors are critical to the quality of paragraph generation. Previous methods[2–4] usually take image features as the input to a long short-term memory (LSTM), which produces a sequence of hidden states. Then, each hidden state is transformed to the corresponding topic vector via a linear projection or a shallow neural network. These methods ignore the modelling of visual relationships, which are very important for image understanding. Although a recent work[5] tries to

Research Article
Manuscript received on August 9, 2022; accepted on December 13, 2022
Recommended by Associate Editor Da-Cheng Tao
Colored figures are available in the online version at https://link.springer.com/journal/11633

model visual relationships to produce topic vectors, it oversimplifies visual relation modelling by just fusing object features.

Considering that a scene graph is a visually-grounded graphical structure of an image, where the nodes depict the object instances and the edges represent their pairwise relationships, it contains the key visual objects and their comprehensive relations for paragraph generation. As shown in Fig. 1, most of the visual objects (e.g., person, street, building and car) in the scene graph and their relations (e.g., cars driving on the street and man walking on the street) occur in the annotated paragraph and our generated paragraph. In particular, several indirect relations in the scene graph should also be taken into account for paragraph generation, such as people walking in front of the building.

In this paper, we propose a new method to generate paragraphs by modelling comprehensive visual relationships based on scene graphs. First, we build a scene graph through visual relationship detection, and obtain the visual features of the objects and their relations. Second, by leveraging the structure of the scene graph, we propose two graph convolutional networks (GCNs) to enrich the visual features with contextual cues by encoding visual and high-order relations. In particular, we also extract the linguistic features for the predicted labels of the detected objects and relationships in the scene graph. Third, the enriched features are selectively fed into an attention-based topic generation network, which produces a sequence of topic vectors. With the linguistic features, the topic vectors generate a paragraph via an RNN-based language model.

The main contributions of our work can be summarized as follows:

1) We propose a novel image paragraph generation method based on scene graphs, where the visual relations between objects as well as the high-order features between visual relations are specifically explored by GCNs to improve the quality of generated paragraphs.

2) We extract the linguistic features from the labels of the detected objects and their relationships as supplements to the visual features. The multimodal features are then fused via an attention-based topic generation network to generate topic vectors.

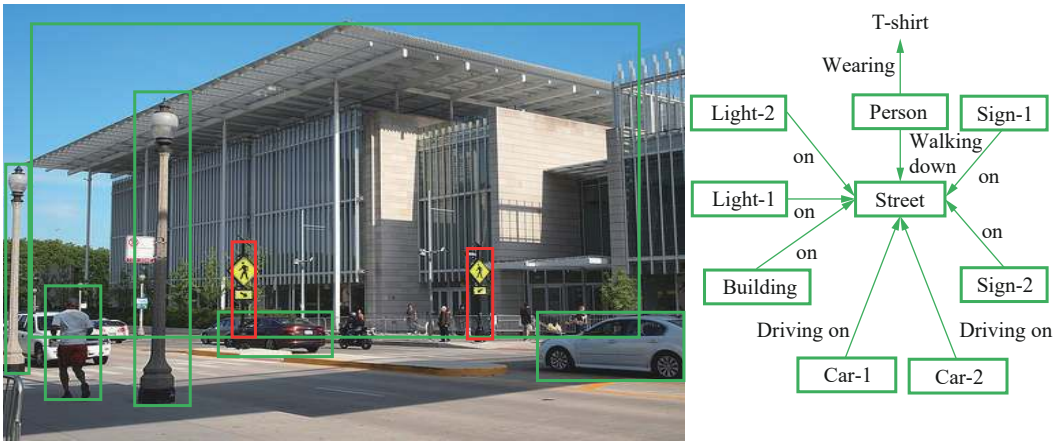
3) We perform extensive evaluations and ablation studies on the benchmark of image paragraph generation (i.e., the Stanford image-paragraph dataset), and the comparable results with the state-of-the-art (SOTA) methods verify the effectiveness of our model.

2 Related work

In this section, we briefly review the literature that is closely related to the proposed method, including scene graph parsing and image paragraph generation.

2.1 Scene graph parsing

Similar to visual relationship detection[6–9], scene graph parsing is a visual task aiming to understand a



(a)

A large building with bars on the windows in front of it. There arepeople walking in front of the building. There is a street in front of the building with many cars on it.

(b)

This picture is taken outside on a sunny day. There is a large building behind the street. There are cars driving on the street in front of the building. There is a man walking on the sidewalk. Tall lights are standing on the sidewalk.

(c)

Fig. 1 Examples of a scene graph and paragraphs: (a) An image and its scene graph; (b) Ground-truth of the paragraph; (c) Our generated paragraph. Objects and the visual relationships in the scene graph are beneficial to generate an informative paragraph.

visual scene by recognizing the objects and detecting their relationships. In a scene graph, an object is denoted as a graph node with a bounding box and an object category label, while a relationship is denoted as a directed edge between two objects with a relational predicate (i.e., subject → predicate → object). Instead of inferring each object and relationship in isolation, Xu et al.[10] addressed the problem of scene graph parsing by passing contextual information through a graph topology and iteratively improving the predictions with standard RNNs. To leverage the mutual connections of object detection, scene graph parsing and region captioning, Li et al.[11] introduced a novel framework to solve the three tasks jointly in an end-to-end manner and utilize their complementary information for mutual improvements. Observing that the relational predicate is closely related to the labels of the subject and object, Zellers et al.[12] proposed to predict the labels of subjects and objects using a bidirectional LSTM, which are then used to infer the relational predicates with the prior knowledge of recurring relationships in scene graphs. To improve the efficiency of generating scene graphs, Li et al.[13] constructed a shared representation for highly overlapped ⟨subject, object⟩ pairs and factorized the entire scene graph into subgraphs, thus significantly reducing the redundant computation and accelerating the inference speed. Woo et al.[14] designed a relational embedding module to model interdependence among entire objects, and Chen et al.[15] learned a routing mechanism to explore the relationships by propagating messages through a graph. Li et al.[16] proposed an unbiased graph neural network with adaptive message propagation to alleviate error propagation and achieve effective context modelling. Different from traditional methods using cross-entropy losses, Suhail et al.[17] introduced a novel energy-based learning framework for effectively incorporating the structure of scene graphs in the output space. In this work, we adopt [16] to generate scene graphs owing to its efficiency and performance.

2.2 Image paragraph generation

Generating an informative paragraph for an image is a challenging task, since it not only needs the image encoder to understand objects and relationships but also requests a robust natural language model to generate long descriptions. To solve this problem, Krause et al.[2] decomposed the input image into semantically meaningful regions of interest, and employed a hierarchical recurrent neural network (HRNN) to generate multiple sentences. The proposed HRNN consists of SentenceRNN and WordRNN, where SentenceRNN produces topic vectors and WordRNN generates the corresponding sentences. Liang et al.[3] built an adversarial framework containing a structured paragraph generator and two discriminators to drive model learning. The generator aims to constructed meaningful and coherent paragraphs, which are then fed into a sentence discriminator and a recurrent topic-transition discriminator. The two discriminators are used to measure the plausibility of paragraphs and the smoothness of semantic transition between different sentences. Chatterjee and Schwing[4] introduced coherence vectors to guarantee a gradual transition of contextual topics. In addition, to generate a set of diverse paragraphs, Chatterjee and Schwing[4] formulated paragraph generation into a variational autoencoder (VAE)[18] framework. Different from prior studies that generate the topic vectors with SentenceRNN, Che et al.[5] constructed topic vectors based on relation features, which are generated with the visual features of related pairwise objects, and Wang et al.[19] proposed generating topic vectors from region features through convolutional layers. Melas-Kyriazi et al.[20] applied self-critical training for this task and punished the repeated trigram when decoding into paragraphs. Zha et al.[21] proposed a context-aware visual policy network that explicitly considers the previous visual attentions as context when generating the current visual attention. To alleviate the repetitive and incomplete captioning problems, Xu et al.[22] designed an interactive key-value memory-augmented attention mechanism to track the information of each object. Yang et al.[23] used an image scene graph to incorporate rich semantic knowledge and hierarchical constraints into the model. Shi et al.[24] proposed a tree-structured decoder network to better organize visual clues holistically. Xu et al.[25] proposed a retrieval-enhanced adversarial framework to enhance image description generation with retrieved candidate captions. Guo et al.[26] designed a visual-textual coupling model to distill multilayer semantic topics of a given image, and used a language model to interpret the extracted image features and semantic topics into captions. In this work, to obtain more contextual representations, we employ the object-based GCN to encode visual relations and the relation-based GCN to encode high-order relationships. Then, we design an attention-based topic generation network to produce the topic vectors.

3 Our method

As shown in Fig. 2, an image is first input to the scene graph detection module which detects a scene graph and extracts the object and relation features. These two kinds of visual features are enriched with contextual cues by encoding visual and high-order relations via object-based and relation-based GCNs, respectively, which are then fed into an attention-based topic generation network to produce a set of topic vectors and attentive weights. Moreover, the labels of objects and visual relationships in the detected scene graph are embedded into the linguistic features, which are then integrated with attentive weights. At each time step t, the attended linguistic features are input to an RNN to generate a sentence with

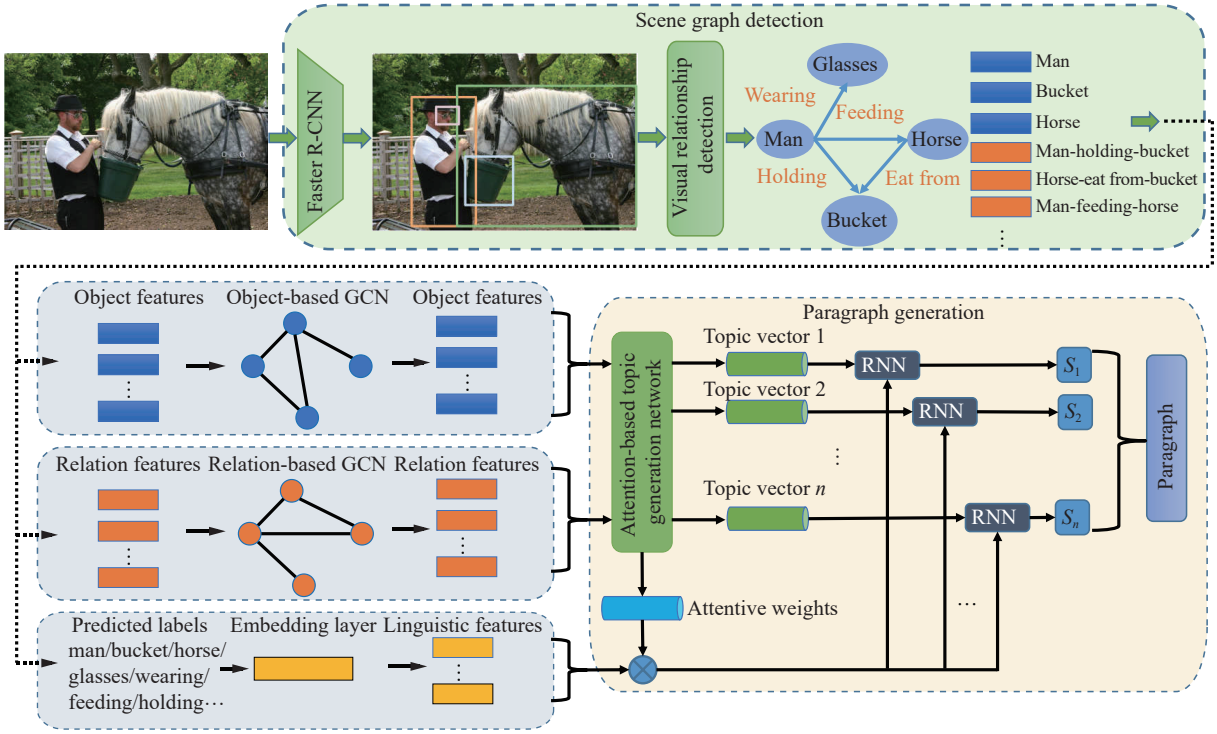


Fig. 2 Architecture of our method. Scene graph detection module parses an image into a scene graph and produces the visual features of the objects and their relationships, which are enriched through two GCNs. Enriched visual features are then fed into the paragraph generation module to generate a paragraph with the aid of the linguistic features.

the corresponding topic vector. After unrolling the attention-based topic generation network for n steps, we obtain all the sentences that are orderly concatenated into a complete paragraph. Details are presented as follows.

3.1 Scene graph parsing

We adopt the scene graph generator in [16] to parse a given image into a scene graph. First, the Faster R-CNN[27] is leveraged to produce a set of entity proposals. The detected entity proposals are fed into a multistage graph network to obtain a context-aware representation. The network adopts directed edges to model different information flows between entity and relationship proposals as a bipartite graph, and an adaptive message propagation strategy based on relation confidence estimation to reduce the noise in context modelling. Finally, the refined entity and predicate representations are used to predict their categories with linear classifiers.

During the process of constructing the scene graph for an image, plenty of contextual cues for understanding visual contents are available. For instance, the object feature for each proposal can be obtained by projecting the visual representation of the corresponding region from 4 096-d to 1 024-d. For a pair of related objects, we learn the relation feature of a triplet ⟨subject-predicate-object⟩ based on the features of the subject, the object and the bounding box enclosing the relationship. In addition to the aforementioned visual cues, the labels of the object categories and the relationships play the role of attributes and provide abundant textual information, which are projected into the linguistic feature via an embedding layer.

Furthermore, based on the structure of the scene graph, we build two GCNs (i.e., the object-based GCN and relation-based GCN) to enrich the visual representations with high-order information, which draws inspiration from [28]. The object-based GCN takes the object features as nodes and models the visual relations with the edges in the original scene graph. Then, we switch its nodes and edges to obtain the relation-based GCN, as shown in Fig. 3, where the nodes denote the visual relations and the edges model the high-order relations between objects. The details are shown in Sections 3.2 and 3.3.

3.2 Modelling visual relations between objects

Suppose that we detect N region proposals in an image I, and extract the set of object features $\mathcal{O} = \{o_i\}_{i=1}^{N}$ from the bounding boxes, where $o_i \in \mathbf{R}^{D_o}$ denotes the D_o-dimensional visual feature of the i-th object proposal. To implicitly encode visual relationships into the object features, we build a graph $\mathcal{G}_{obj} = (\mathcal{O}, \mathcal{E}_{obj})$ according to the structure of the generated scene graph, where $\mathcal{O}$ denotes the object features and $\mathcal{E}_{obj}$ denotes the set of semantic connections between the related objects. Based on the topology of the graph $\mathcal{G}_{obj}$, we employ a GCN to update

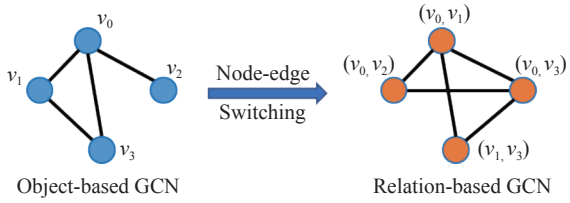


Fig. 3 Illustration of node-edge switching

the node representations via message passing. In this way, the object feature o_i can be iteratively improved by incorporating the neighboring object features. In particular, the iteration process of o_i is formulated as

$$o_i^t = \rho(W_s^o o_i^{t-1} + \sum_{o_j \in E_{o_i}} W o_j^{t-1} + b) \quad (1)$$

where ρ is the ReLU activation function. E_{o_i} denotes the set of neighbours of o_i. $W_s^o \in \mathbf{R}^{D_o \times D_o}$ and $W \in \mathbf{R}^{D_o \times D_o}$ are transformation matrices for the node itself and the neighbors, respectively. b represents the bias vector.

Furthermore, to take full advantage of the predicted labels of the visual relationships, we extend the transformation matrix W with a set of weights $\{W_k\}_{k=1}^K$ where K denotes the number of relational predicates. In this way, we can select the corresponding weight depending on the predicate of the relationship when we update o_i using the neighboring node o_j (e.g., W_1 for "on", W_2 for "eating"). In this way, o_i is enhanced to be a more contextual visual representation with the information of the relationships. However, there is a "long-tail" phenomenon in the frequency distribution of the relational predicates. For example, the predicates "on" and "has" appear frequently compared with "carrying" and "eating". Moreover, it costs much storage capacity to update the nodes with $\{W_k\}_{k=1}^K$. To mitigate the imbalance of the frequency distribution of the predicates and reduce the storage capacity of the iteration process for the nodes, we divide these predicates into four classes: Geometric (e.g., above, behind, under), Possessive (e.g., has, part of, wearing), Semantic (e.g., carrying, eating, using) and Misc (for, from, made of) based on their property following[12]. The updated frequency distribution of the predicates is presented in Table 1. Then, the transformation matrices $\{W_k\}_{k=1}^K$ can be updated as $\{W_{k'}\}_{k'=1}^4$, i.e., W_1 for Geometric, W_2 for Possessive, W_3 for Semantic, W_4 for Misc. Thus, the iteration process is updated as follows:

$$o_i^t = \rho(W_s^o o_i^{t-1} + \sum_{o_j \in E_{o_i}} g(o_i, o_j) W_{cls(o_i,o_j)} o_j^{t-1} + b_{cls(o_i,o_j)}) \quad (2)$$

$$g(o_i, o_j) = \sigma(\tilde{W}_{cls(o_i,o_j)}[o_i, o_j] + \tilde{b}_{cls(o_i,o_j)}) \quad (3)$$

where $cls(o_i, o_j)$ denotes the class of the predicate between the pairwise objects (o_i, o_j), and $[\cdot, \cdot]$ is the concatenating function. Moreover, when aggregating the representations of the neighboring nodes, we set an elementwise gate $g(o_i, o_j)$ to reduce redundant information as shown in (3), where σ is the logistic sigmoid function. $\tilde{W}_{cls(o_i,o_j)} \in \mathbf{R}^{2D_o \times D_o}$ is the transformation matrix and $\tilde{b}_{cls(o_i,o_j)} \in \mathbf{R}^{D_o}$ is the bias vector.

After all the nodes are updated according to (2) and (3), the object features are enriched by implicitly encoding the visual relationships between the objects, which are denoted as $\{o_i'\}_{i=1}^N$.

Table 1 Frequency distribution of predicates, which is drawn from [12]

Classes	Examples	# Predicates	#Instances
Geometric	Above, behind, under	15	228 K (50.0%)
Possessive	Has, part of, wearing	8	186 K (40.9%)
Semantic	Carrying, eating, using	24	39 K (8.7%)
Misc	For, from, made of	3	2 K (0.3%)

3.3 Modelling high-order relations

Considering that visual relations usually appear in paragraphs, we devise relation features based on scene graph representation. As illustrated in Fig. 4, for a relation instance ⟨woman-play-football⟩, the visual features of the subject box "woman" v_s, the object box "football" v_o and the union box v_u are first extracted from the corresponding regions, and then multiplied to obtain the relation feature r in an elementwise manner (i.e., $r = v_s \otimes v_o \otimes v_u$). In this way, we obtain the set of relation features $\mathcal{R} = \{r_i\}_{i=1}^M$ from an image I, where $r_i \in \mathbf{R}^{D_r}$ represents the D_r-dimensional relation feature and M denotes the number of related object pairs.

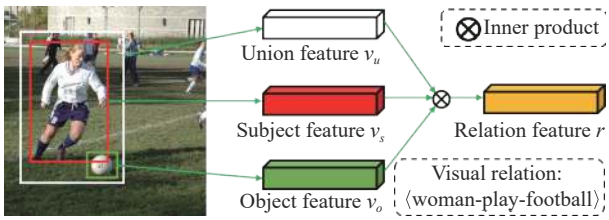


Fig. 4 Illustration of generating relation features

We observe that there are high-order relations between different visual relations. For example, as shown in Fig. 2, ⟨man-feeding-horse⟩ is related to ⟨man-holding-bucket⟩ because the reason that the man is carrying the bucket is to feed the horse. However, the high-order relations are not always meaningful. For example, as seen in Fig. 2, ⟨man-wearing-glasses⟩ and ⟨man-holding-bucket⟩ are related to the same object "man", but there is no semantic association between them. To model the implicit and complex high-order relations, we construct another graph $\mathcal{G}_{rel} = (\mathcal{R}, \mathcal{E}_{rel})$, where $\mathcal{R} = \{r_i\}_{i=1}^M$ denotes the nodes and $\mathcal{E}_{rel}$ denotes the edges connecting two relevant

relationships. Specifically, for two relations $r_1 = \langle s_1, p_1, o_1 \rangle$ and $r_2 = \langle s_2, p_2, o_2 \rangle$ (s, p, o denote subject, predicate, object, respectively), if the two sets $\{s_1, o_1\} \cap \{s_2, o_2\} \neq \emptyset$, we consider that there is a high-order relation between r_1 and r_2, and use an undirected edge to connect them. Thus, these visual relations can be enriched through the graph structure of $\mathcal{G}_{rel}$ by using a GCN, which is formulated as

$$r_i^t = \rho(W_s^r r_i^{t-1} + \sum_{r_j \in E_{r_i}} W^r r_j^{t-1} + b^r) \tag{4}$$

where E_{r_i} denotes the set of neighboring visual relations of r_i. $W_s^r \in \mathbf{R}^{D_r \times D_r}$ and $W^r \in \mathbf{R}^{D_r \times D_r}$ represent transformation matrices. b^r is the bias vector.

As discussed above, some relations connected by the edges are actually not relevant at the semantic level. To measure the relevance of the two connected relations, we design an edgewise gate unit to compute a scale factor based on the corresponding relation features. Similar to graph attention networks[29], we incorporate the gate into the GCN as follows:

$$r_i^t = \rho(W_s^r r_i^{t-1} + \sum_{r_j \in E_{r_i}} \alpha(r_i, r_j) W^r r_j^{t-1} + b^r) \tag{5}$$

$$\alpha(r_i, r_j) = \frac{relu(\beta([r_i, r_j]))}{\sum_{r_{j'} \in E_{r_i}} relu(\beta([r_i, r_{j'}]))} \tag{6}$$

where $\beta(\cdot)$ denotes a linear function and $[\cdot, \cdot]$ means the concatenating operation. $\alpha(r_i, r_j)$ is an edgewise gate to control the information flow from r_j to r_i. In this way, the model learns to focus on potentially important edges that contain meaningful high-order relations. Consequently, after updating all the nodes of $\mathcal{G}_{rel}$ with the relevant relations via the GCN as in (5) and (6), the enhanced relation features denoted as $\{r_i'\}_{i=1}^M$ are encoded with the implicit high-order contextual information between the visual relations.

3.4 Paragraph generation

Based on the enriched visual features $\{o_i'\}_{i=1}^N$ and $\{r_j'\}_{j=1}^M$, paragraphs are generated via the caption generation module. Drawing inspiration from [2–5], the caption generation module is constructed based on hierarchical RNN. Concretely, at each time step t, a high-level RNN (i.e., SentenceRNN) outputs two vectors: topic vector and sentence state. The topic vector is fed into a low-level RNN (i.e., WordRNN) to generate the corresponding sentence. The sentence state is used to decide whether to generate the next sentence or not. Therefore, the generation of topic vectors is crucial for the quality of the final paragraphs.

To produce informative topic vectors, we take the enriched visual features $\{o_i'\}_{i=1}^N$ and $\{r_j'\}_{j=1}^M$ as inputs to the SentenceRNN. Moreover, since the multiple sentences of a paragraph usually describe different visual contents of an image, the corresponding topic vectors are supposed to focus on specific regions of interest. To this end, an attention mechanism is employed to select the relevant features from $\{o_i'\}_{i=1}^N$ and $\{r_j'\}_{j=1}^M$. As shown in the Fig. 5, the topic generation network is composed of two LSTMs: attention-LSTM and topic-LSTM. Attention-LSTM is responsible for outputting the attentive weights for the enriched visual features, and topic-LSTM is in charge of producing topic vectors as well as predicting the state of generating sentences. Concretely, at each time step t, we concatenate the mean-pooled object and relation features as a vector $v = [\bar{o}, \bar{r}]$, where $\bar{o} = \frac{1}{N}\sum_{i=1}^N o_i'$ and $\bar{r} = \frac{1}{M}\sum_{j=1}^M r_j'$. Then, vector v is concatenated with the previous hidden state h_{t-1}^p of the topic-LSTM. Finally, the attention-LSTM takes the concatenated vector as input, and outputs the attentive weights for the object and relation features. In particular, the updating process of the attention-LSTM is formulated as follows:

$$h_t^a = f(W_1[v, h_{t-1}^p]) \tag{7}$$

where h_t^a is the hidden state of the attention-LSTM. W_1 is the matrix for transforming the concatenation $[v, h_{t-1}^p]$. $f(\cdot)$ is the updating function within the attention-LSTM

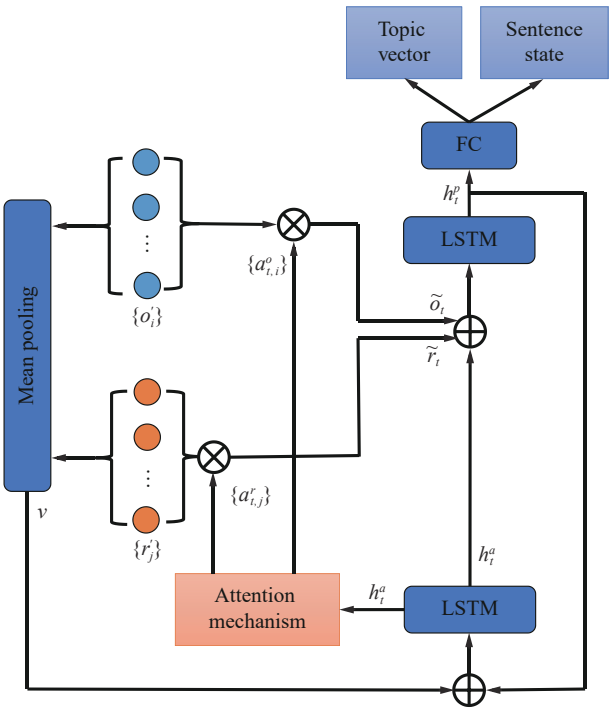


Fig. 5 Overview of the attention-based topic generation network. At each time step t, one of the LSTMs is used to output the attentive weights for the enriched object and relation features via an attention mechanism, while the other LSTM is responsible for predicting the topic vector and the sentence state with the attended features.

unit. Based on the output h_t^a, the attention distributions over the object and relation features are calculated as

$$a_{t,i}^o = W_a^o(tanh(W_v^o o_i' + W_h^o h_t^a)) \tag{8}$$

$$a_{t,i}^o = softmax(a_{t,i}^o) \tag{9}$$

$$a_{t,j}^r = W_a^r(tanh(W_v^r r_j' + W_h^r h_t^a)) \tag{10}$$

$$a_{t,j}^r = softmax(a_{t,j}^r) \tag{11}$$

where $a_{t,i}^o$ and $a_{t,j}^r$ denote the attentive weights of o_i' and r_j', respectively at time step t. W_a^o, W_a^r, W_v^o, W_v^r, W_h^o and W_h^r are the transformation matrices.

Depending on these two sets of normalized attention distributions $\{a_{t,i}^o\}_{i=1}^N$ and $\{a_{t,j}^r\}_{j=1}^M$, the object and relation features are aggregated into the attended visual features, which are calculated as $\tilde{o}_t = \sum_{i=1}^N a_{t,i}^o o_i'$ and $\tilde{r}_t = \sum_{j=1}^M a_{t,j}^r r_j'$. Furthermore, we concatenate the attended visual features and the output of the attention-LSTM as a vector, which is then projected as the input of the topic-LSTM. The updating procedure of topic-LSTM is thus formulated as follows:

$$h_t^p = g(W_2[\tilde{o}_t, \tilde{r}_t, h_t^a]) \tag{12}$$

where h_t^p is the output of the topic-LSTM, whose updating function is denoted as $g(\cdot)$. W_2 is a transformation matrix. Next, the hidden state h_t^p is utilized to predict a probability distribution p_t over the two states "CONTINUE" and "STOP" with a logistic classifier. When the probability of "CONTINUE" exceeds that of "STOP", WordRNN will continue to generate the next topic vector, and vice versa. Then, we reuse h_t^p to produce the topic vector T_t via two fully connected layers.

To generate a detailed paragraph relevant to semantic concepts of an image, we leverage the linguistic features to help word generation. As discussed in Section 3.1, the linguistic features are generated by encoding the labels of the object categories and the relationships via an embedding layer. Specifically, the label of the object category is first converted into a one-hot vector e_o and then embedded by a linear mapping as follows:

$$\hat{e}_o = W_{embed} e_o \tag{13}$$

where W_{embed} is an embedding matrix. Since the labels of visual relations usually contain multiple elements (e.g., ⟨man-drinking-water⟩), the corresponding linguistic feature is calculated by summarizing the embedding vectors of all the elements:

$$\hat{e}_r = \sum_{k=1}^K W_{embed} e_k \tag{14}$$

where K denotes the number of elements, and e_k is the k-th element in the label of a visual relation (e.g., for ⟨man-drinking-water⟩, e_1, e_2 and e_3 are "man", "drinking" and "water", respectively). We perform language attention on the embedding vectors to selectively aggregate the linguistic features. Since the labels are semantically related to the object or relation features, we thus reuse the corresponding attentive weights $\{a_{t,i}^o\}_{i=1}^N$ and $\{a_{t,j}^r\}_{j=1}^M$ to obtain the two integrated embedding vectors $\tilde{e}_t^o = \sum_{i=1}^N a_{t,i}^o \hat{e}_o$ and $\tilde{e}_t^r = \sum_{j=1}^M a_{t,j}^r \hat{e}_r$ at time step t of the topic-LSTM. Therefore, the topic vector T_t is updated as

$$T_t' = W_T T_t + W_e^o \tilde{e}_t^o + W_e^r \tilde{e}_t^r \tag{15}$$

where W_T, W_e^o and W_e^r are the transformation matrices.

Given the updated topic vector T_t', WordRNN is capable of generating the corresponding sentence. The first and second inputs to WordRNN are the topic vector T_t' and the ⟨START⟩ token, respectively, and subsequent inputs are learned embedding vectors for the words. During the updating procedure of WordRNN, each hidden state is used to predict a distribution over the words in the vocabulary. After all the words of every sentence have been generated, the sentences are concatenated to form an orderly paragraph.

3.5 Loss function

Consider pairs of training data (x, y), where x is an image and y is a ground-truth paragraph description for that image. Suppose that y consists of S sentences, with the i-th sentence having N_i words. In the training stage, we unroll SentenceRNN for S time steps and predict a set of distributions $\{p_i\}_{i=1}^S$ over the states of generating sentences (i.e., {CONTINUE, STOP}), which is then used to calculate a binary cross-entropy sentence-level loss. WordRNN produces the distribution $p_{i,j}$ for the j-th word of the i-th sentence, which is used to calculate a cross-entropy word-level loss:

$$l_{sentence} = \sum_{i=1}^S l_s(p_i, I[i = S]) \tag{16}$$

$$l_{word} = \sum_{i=1}^S \sum_{j=1}^{N_i} l_w(p_{ij}, y_{ij}) \tag{17}$$

where $l_s(\cdot)$ and $l_w(\cdot)$ are two cross-entropy functions. $I[\cdot]$ denotes the indicator function. y_{ij} is the j-th word of the i-th sentence.

To alleviate the mismatch problem between training and testing, we use a popular reinforcement learning (RL) loss[23], which has been proven to be efficient and effective training the paragraph generator:

$$l_{para} = -\mathrm{E}_{p_{ij}\sim y'}[r(y'; y)] \tag{18}$$

where r is the CIDEr[30] metric between the generated paragraph y' and ground-truth paragraph y.

Moreover, to encourage the model to pay attention to every object and relationship of the image when generating all the topic vectors, we devise two penalties to encourage $\sum_{i=1}^{S} a_{i,n}^{o} \approx 1$ and $\sum_{i=1}^{S} a_{i,m}^{r} \approx 1$ where $n = 1, 2, \cdots, N$ and $m = 1, 2, \cdots, M$:

$$l_{penalty} = \sum_{n=1}^{N}(1 - \sum_{i=1}^{S} a_{i,n}^{o})^2 + \sum_{m=1}^{M}(1 - \sum_{i=1}^{S} a_{i,m}^{r})^2. \tag{19}$$

We combine the word-level, sentence-level, paragraph-level losses and the penalties on the attentive weights as the final training loss of our model:

$$l(x, y) = l_{word} + \lambda_s l_{sentence} + \lambda_a l_{para} + \lambda_p l_{penalty} \tag{20}$$

where λ_s, λ_a and λ_p are the scale factors.

4 Experiments

In this section, we first introduce the dataset and evaluation metrics used in our experiments, then present extensive ablation studies on our model, and finally report our results and comparisons with other methods.

4.1 Dataset and evaluation metrics

Dataset: We conduct experiments on the Stanford image-paragraph dataset[2], which is the only generally acknowledged benchmark for the task of generating image paragraphs. The dataset consists of 19 551 images from MS COCO[31] and Visual Genome[32], where each image has been annotated with a human-labelled paragraph description containing 67.5 words on average. Following the experimental protocol of [2], we divide this dataset into 14 575 training, 2 487 validation, and 2 489 testing images.

Evaluation metrics. We adopt six widely used language generation metrics: BLEU-1, BLEU-2, BLEU-3, BLEU-4[33], METEOR[34], and CIDEr[30] to evaluate our model. BLEU is a popular metric for machine translation evaluation that computes an n-gram based precision for the candidate sentence with respect to the references. METEOR returns its judgment of the generated sentences by computing the F-measure based on matches, and CIDEr provides evaluation based on consensus.

4.2 Implementation details

In our experiments, we extract the visual features from the scene graph using the top $N = 50$ objects and $M = 20$ relations. The dimensions of the original object and relation features are 4 096, which are then transformed into 1 024 via a linear projection. The linguistic features have the same dimension as the word embedding vectors, which is set to 300. The embedding layer is initialized with global vectors (GloVe)[35], and then trained under our loss. Two LSTMs of the topic generation network have a single layer with 512 dimensions, while WordRNN adopts two LSTM layers. In addition, the weights of our linear layers are initialized using Kaiming initialization[36].

In the training stage, we train our network for 25 epochs using the Adam optimizer[37]. The initial learning rate is set to 0.001, and the batch size is set to 128. Instead of decaying the learning rate at regular intervals, we change it depending on the performance (e.g., the average of METEOR and CIDEr scores) on the validation set. In particular, the learning rate is decayed only when the performance stops improving for 5 epochs. According to the validation set performance, λ_s, λ_a and λ_p are set to 5, 1 and 1, respectively.

At the inference time, SentenceRNN keeps producing topic vectors until the stopping probability p(STOP) exceeds the continuing probability p(CONTINUE) or the number of sentences reaches the threshold $S_{\max}$, whichever comes first. Based on the produced topic vectors, WordRNN samples the words using beam search (beam size = 7) and stops when ⟨END⟩ token is met or after $N_{\max}$ words. In our experiments, we set the parameters $S_{\max}$ to 6 and $N_{\max}$ to 30.

4.3 Ablation studies

In this section, we perform extensive ablation studies on our method, including the components of our model, the GCN used to update the object features, the manner of generating the relation feature, the size of the beam search decoder and the hyperparameters of the loss function.

4.3.1 Contributions of model components

To demonstrate the effectiveness of each component in our model, we design a baseline "OBJ" that only uses the object features to generate paragraphs without the GCNs and the attention mechanism. The analysis is also performed on the benchmark, as shown in Table 2. We report the results of the baseline in the first row. Based on the baseline, we use the object-based GCN (denoted as GCN-O) to enrich the object features and add the relation features, which are referred to as "OBJ+GCN-O" and "OBJ+REL", respectively. The corresponding results show a certain amount of improvements over the performance of the baseline at all the six metrics, which demonstrates that GCN-O and the relation features are both beneficial to obtaining more contextual representations for generating paragraphs. Then, we use the relation-based GCN (denoted as GCN-R) to enrich the relation features, and the results ("OBJ+REL+GCN-R") show that GCN-R is also beneficial to our model compared to the previous row. Next, the "OBJ+REL+

Table 2 Ablations of our method

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
OBJ (baseline)	39.99	25.73	16.25	9.79	16.94	29.37
OBJ+GCN-O	41.79	26.84	17.03	10.32	17.45	29.24
OBJ+REL	42.18	27.02	17.14	10.46	17.60	29.47
OBJ+REL+GCN-R	42.87	27.43	17.32	10.55	17.71	29.75
OBJ+REL+GCNs	43.43	27.72	17.51	10.63	17.78	29.91
OBJ+REL+GCNs+ATT	43.57	27.95	17.67	10.75	**17.95**	30.33
+ Linguistic feature (Ours)	**43.81**	**27.98**	**17.69**	**10.78**	17.94	**31.64**

GCNs" model uses GCN-O and GCN-R to enrich the object and relation features, which further verifies the effectiveness of the GCNs. By introducing the attention-based topic generation network, the "OBJ+REL+GCNs+ATT" model obtains successive improvements on all six metrics compared to the previous model. Finally, the comparison between the "Ours" and "OBJ+REL+GCNs+ATT" models (especially 1.31% improvement on CIDEr) indicates that the linguistic features are helpful for generating more informative paragraphs.

4.3.2 Comparison of variant GCNs

As mentioned in Section 3.2, we employ a modified GCN to update the object features via message passing. Compared to the general GCN, the modified GCN used in our model adds an elementwise gate to reduce redundant information and multiple weights according to the classes of the predicates. To verify the effectiveness of these two components, we first remove the multiple weights from the GCN (i.e., all the nodes are transformed by using a common weight), and the corresponding model is denoted as "GCN-O w/o weights". We further remove the gate from the GCN, and the update rules of nodes can be formulated as (1). The corresponding model is denoted as "w/o gate".

We conduct experiments for these models and present the results in Table 3. From Table 3, we observe that the multiple weights bring 0.45%, 0.33%, 0.2%, 0.11%, 0.11% and 0.39% on all the metrics by comparing the results of "Ours" and "GCN-O w/o weights", and the results of "GCN-O w/o weights" increase 0.54%, 0.1%, 0.18% and 0.14% on BLEU-1, BLEU-2, METEOR, CIDEr compared to the results of "w/o gate". These comparison results show that the gate and multiple weights are beneficial to generating better paragraphs. In addition, we also remove the gate from GCN-R (the relation-based GCN) and the results ("GCN-R w/o gate") drop 0.36%, 0.27%, 0.13%, 0.07%, 0.15% and 0.43% on BLEU-{1, 2, 3, 4}, METEOR and CIDEr compared to "Ours", which further verifies the effectiveness of the gate for GCN.

4.3.3 Comparison with other relation features

There is no doubt that the relation features play an important role in image paragraph generation. As shown in Fig. 4, we denote the features of the subject, object, and union area as v_s, v_o and v_u, respectively. Then, we multiply them in an elementwise manner to obtain the relation feature, which is denoted as CRM-MUL. To verify the effectiveness of this method, we design two other methods of generating the relation feature for comparison. The first model uses the sum of v_s, v_o and v_u as the relation feature, which is denoted as CRM-SUM. The second model generates the relation feature by concatenating v_s, v_o and v_u, which is denoted as CRM-CC.

To perform ablation analysis on the different manners of generating relation features, we present the results of the three models (CRM-SUM, CRM-CC, CRM-MUL) in Table 4. It is obvious that our model CRM-MUL has achieved better performance than CRM-SUM and CRM-CC on the six language metrics. These comparison results demonstrate that multiplying the features of the subject, object and union area is an effective way to obtain the relation feature.

4.3.4 Sensitivity analysis of beam size

As mentioned in Section 4.2, the beam search decoder is adopted to generate image paragraphs at the inference time. To study the sensitivity of our model to beam size (i.e., the size of candidate sentences), we conduct experiments on the benchmark dataset by varying the beam size from 2 to 9. As presented in Fig. 6(a), the experimental result for CIDEr increases as the beam size increases from 2 to 7. Compared to the result of "beam

Table 3 Ablations of GCN

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
Ours	43.81	27.98	17.69	10.78	17.94	31.64
GCN-O w/o weights	43.36	27.65	17.49	10.67	17.83	31.25
w/o gate	42.82	27.55	17.49	10.65	17.65	31.11
GCN-R w/o gate	43.45	27.71	17.56	10.71	17.79	31.21

Table 4 Performance comparison with the different relation features

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
CRM-SUM	42.60	27.28	17.25	10.49	17.57	30.07
CRM-CC	43.05	27.56	17.44	10.59	18.15	30.97
CRM-MUL (Ours)	43.81	27.98	17.69	10.78	17.94	31.64

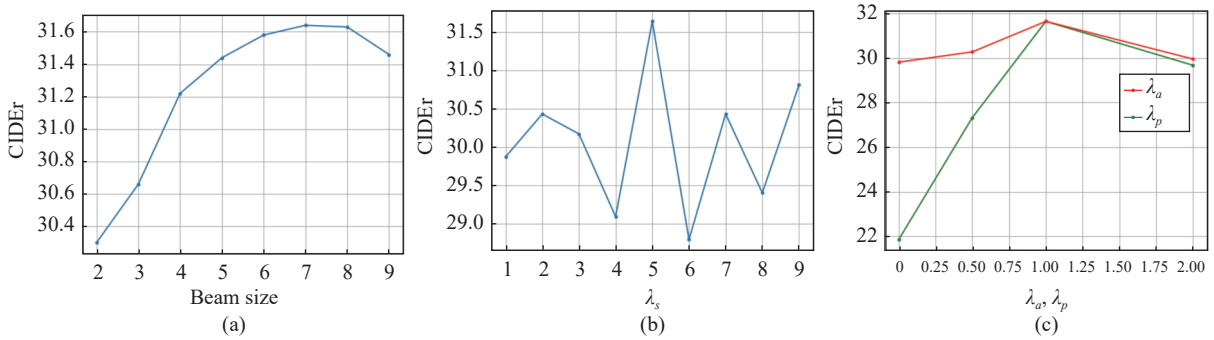


Fig. 6 Results with different beam sizes and hyperparameters of loss: (a) CIDEr with various beam size; (b) CIDEr with various λ_s; (c) CIDEr with various λ_a and λ_p.

size = 2", our model increases CIDEr by 1.34% when the beam size is set to 7. We can conclude that beam size has an effect on the performance of our model. In addition, from the figure, we find that the result degrades with a large beam size. This phenomenon has also appeared in other work[38].

4.3.5 Three hyperparameters of loss

We first study the value of λ_s by varying it from 1 to 9 at intervals of 1.0, as presented in Fig. 6(b). It is crucial to select a suitable λ_s because it directly affects the number of sentences composing the generated paragraph. From Fig. 6(b), we observe that our model achieves the best performance on CIDEr when λ_s is set to 5. Therefore, we select $\lambda_s = 5$ as the best choice.

We also perform sensitivity analyses on λ_a and λ_p used in the loss function. Concretely, we conduct experiments on the benchmark dataset by changing λ_a and λ_p as [0, 0.5, 1, 2], and report the results in Fig. 6(c). It is obvious that our model achieves the best results at $\lambda_a = 1$ and $\lambda_p = 1$. Furthermore, our model achieves better results at 0.5, 1 and 2 compared to the result at 0, which shows that l_{para} and $l_{penalty}$ of the loss function work well. In addition, the comparisons between the results at 1 and 2 indicate that the overweight l_{para} and $l_{penalty}$ adversely affect model training.

4.4 Comparison with the state-of-the-art (SOTA)

In this section, we show the quantitative results on the benchmark in Table 5 and some qualitative results in Fig. 7. As shown in Table 5, we evaluate our method on the benchmark, and compare it with several recent methods. Our approach shows comparable performance to the state-of-the-art methods on the scores of BLEU-{1, 2, 3, 4}. In particular, our model achieves a better CIDEr score than a majority of previous methods (e.g., a retrieval-enhanced adversarial training with dynamic memory-augmented attention for image paragraph captioning (RAMP)[25] and a visual-textual coupling model (VTCM)[26]), except hierarchical scene graph encoder-decoder (HSGED)[23]. The reasons why our model performs worse than HSGED are analysed as follows. First, HSGED integrates attribute information (such as color) into features, which requires extra labelled training data to train the model to detect the attributes of objects. Second, HSGED constructs a subgraph for each object, and incorporates such subgraphs into embeddings by a graph neural network (GNN). However, the subgraph-level embedding undoubtedly requires many computations, thus leading to a longer runtime. Third, to improve the quality of the generated paragraphs, HSGED adopts two reinforcement learning (RL) based losses (while we only use one RL based loss), which may increase the difficulty of model training. Therefore, compared to HSGED, our method does not need extra knowledge (attribute labels) and adopts GCNs to make the model more efficient.

As seen in Fig. 7, we also present some qualitative results. Our results are close to the ground-truth and give good descriptions for the visual contents of the images. We also list the corresponding paragraphs generated by the baseline, which is mentioned in Section 4.3.1. It is obvious that our model generates more comprehensive and richer paragraphs than the baseline. For example, given the first figure, our model describes the helmet while the baseline misses it. For the third figure, our model gives a detailed description of the color of the zebras and the distribution of the trees. However, in the corresponding paragraph produced by the baseline, the description of the zebras is very simple, and trees are mistaken as grass. It is worth noting that our model does not handle

Table 5 Comparisons with the SOTA models on the benchmark

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
Regions-hierarchical[2] (CVPR 2017)	41.9	24.11	14.23	8.69	15.95	13.52
RTT-GAN[3] (ICCV 2017)	42.06	25.35	14.92	9.21	18.39	20.36
DC with VAE[4] (ECCV 2018)	42.38	25.52	15.15	9.43	18.62	20.93
PG with VRD[5] (ACM MM 2018)	41.74	24.94	14.94	9.34	17.32	14.55
SCST with penalty[20] (EMNLP 2018)	43.54	27.44	17.33	10.58	17.86	30.63
CAVP[21] (TPAMI 2019)	42.01	25.86	15.33	9.26	16.83	21.12
IMAP[22] (COLING 2020)	44.45	27.93	17.14	10.29	17.36	24.07
HSGED[23] (ACM MM 2020)	44.51	28.69	18.28	11.26	18.33	36.02
S2TD[24] (ACM MM 2021)	44.47	27.38	16.87	10.17	17.64	24.33
RAMP[25] (Knowledge-based systems 2021)	45.27	28.60	17.63	10.48	17.49	23.22
VTCM-Transformer[26] (IJCV 2022)	40.93	25.51	15.94	9.96	16.88	26.15
Ours	43.81	27.98	17.69	10.78	17.94	31.64
Human[2]	42.88	25.68	15.55	9.66	19.22	28.55

Ground truth

In the middle of a pine tree forest, a skier flies in the sky doing a trick. His skis are black , yellow and red and are high up in the air. His poles are also yellow and black. The skier is wearing tan ski pants and a red and white ski jacket. His gloves, hat, goggles and backpack are all black. The white snow is falling around him.

Our model

A man is skiing down a hill. The ground is covered with snow. The person is wearing a red jacket and a black helmet. The snow is white and the sky is grey.

Baseline

A person is skiing down the snow. This is a ski slope. He is also wearing a red jacket. The snow is densely.

Ground truth

There are many people riding bikes on a street. One of the bikes has a brown box on the back of it. The people are stopped at two red lights on the road. Large white buildings surround the people. Part of a white sky can be seen in between the buildings. Also, a tall black pole is standing near two buildings.

Our model

This photo is taken outside on a cloudy day. There are people walking on the sidewalk in front of the building. There is a man in a white shirt standing next to the bicycle. There is a man walking down the sidewalk in front of the building. There are many people standing on the side of the street.

Baseline

People are walking along the street with a yellow light. There appears to be two men riding bikes. Many bikes are on the road.

Ground truth

Four zebras are standing in a field. The zebras are black and white. Some of the zebras' heads are down as they are grazing on grass .The grass on the field is a yellowish-green color and it is short. Behind the zebras are trees growing densely, then another field, where the grass is a darker green color, then another dense forest.

Our model

There are brown and white zebras standing in a field. The zebras are grazing on the grass on the left side of the image. There are a bunch of trees and small bushes in the back of the field that is green. There are a lot of trees in the background along with a forest in the distance.

Baseline

There are zebras in a field.The grass is green. The grass are very tall. The animals are grazing in a field.

Ground truth

There are many people sitting on a beach on a clear, blue sky day. There are three boats sitting in the water. One boat is yellow, while the other two are smaller and white. There are various different umbrellas shading the people from the sun. Behind the sandy beach there are many green, lively trees. Some people are riding in a boat. Some are sunbathing and some are standing in the area between the water and sand.

Our model

This is a picture of a beach. The water is light blue in color. There are people standing next to the water looking at the boats. There is a small boat in the water with people standing on it. The sky is blue and clear.

Baseline

A large sail boat is in the water They have white caps on the front of the boat. The sky above the boat is blue with a few white clouds.

Fig. 7 Qualitative results of our method. We mark more information in red compared to the baseline.

crowded scenes well. For example, the second example of Fig. 7 contains many persons, which are detected by the scene graph detection module. In this case, our model is attracted by these people and uses redundant sentences

to describe them. This causes the model to ignore the descriptions of other objects and relationships. Therefore, in future work, we need to further reduce or merge the similar proposals generated by the scene graph detection module in crowded environments. In addition, we also need a stronger attention mechanism to select salient objects to generate the most important descriptions of the picture.

5 Conclusions

In this work, we have proposed a novel paragraph generation network that comprehensively models visual and high-order relationships. To enrich the visual features extracted from the scene graph of an image, we build two graphs according to the structure of the scene graph, which aim to update the object and relation features with contextual information using GCNs. Then, these enhanced representations are selectively fed into the attention-based topic generation network to produce topic vectors, which are taken as input to the natural language model to generate multiple sentences composing the final paragraph. We empirically evaluate our model on the benchmark of image paragraph generation, which has achieved comparable performance with the state-of-the-art methods.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (Nos. 61721004, 61976214, 62076078 and 62176246).

Declarations of conflict of interest

The authors declared that they have no conflicts of interest to this work.

References

[1] J. Johnson, A. Karpathy, F. F. Li. DenseCap: Fully convolutional localization networks for dense captioning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 4565–4574, 2016. DOI: 10.1109/CVPR.2016.494.

[2] J. Krause, J. Johnson, R. Krishna, F. F. Li. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 3337–3345, 2017. DOI: 10.1109/CVPR.2017.356.

[3] X. D. Liang, Z. T. Hu, H. Zhang, C. Gan, E. P. Xing. Recurrent topic-transition GAN for visual paragraph generation. In *Proceedings of IEEE International Conference on Computer Vision*, Venice, Italy, pp. 3382–3391, 2017. DOI: 10.1109/ICCV.2017.364.

[4] M. Chatterjee, A. G. Schwing. Diverse and coherent paragraph generation from images. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 747–763, 2018. DOI: 10.1007/978-3-030-01216-8_45.

[5] W. B. Che, X. P. Fan, R. Q. Xiong, D. B. Zhao. Paragraph generation network with visual relationship detection. In *Proceedings of the 26th ACM International Conference on Multimedia*, Seoul, Republic of Korea, pp. 1435–1443, 2018. DOI: 10.1145/3240508.3240695.

[6] C. W. Lu, R. Krishna, M. Bernstein, F. F. Li. Visual relationship detection with language priors. In *Proceedings of the 14th European Conference on Computer Vision*, Springer, Amsterdam, The Netherlands, pp. 852–869, 2016. DOI: 10.1007/978-3-319-46448-0_51.

[7] Y. K. Li, W. L. Ouyang, X. G. Wang, X. O. Tang. ViP-CNN: Visual phrase guided convolutional neural network. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 7244–7253, 2017. DOI: 10.1109/CVPR.2017.766.

[8] B. Dai, Y. Q. Zhang, D. H. Lin. Detecting visual relationships with deep relational networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 3298–3308, 2017. DOI: 10.1109/CVPR.2017.352.

[9] Y. H. Zhu, S. Q. Jiang. Deep structured learning for visual relationship detection. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence and the 33th Innovative Applications of Artificial Intelligence Conference and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence*, Louisiana, USA, Article Number 934, 2018.

[10] D. F. Xu, Y. K. Zhu, C. B. Choy, F. F. Li. Scene graph generation by iterative message passing. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 3097–3106, 2017. DOI: 10.1109/CVPR.2017.330.

[11] Y. K. Li, W. L. Ouyang, B. L. Zhou, K. Wang, X. G. Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of IEEE International Conference on Computer Vision*, Venice, Italy, pp. 1270–1279, 2017. DOI: 10.1109/ICCV.2017.142.

[12] R. Zellers, M. Yatskar, S. Thomson, Y. Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 5831–5840, 2018. DOI: 10.1109/CVPR.2018.00611.

[13] Y. K. Li, W. L. Ouyang, B. L. Zhou, J. P. Shi, C. Zhang, X. G. Wang. Factorizable net: An efficient subgraph-based framework for scene graph generation. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 346–363, 2018. DOI: 10.1007/978-3-030-01246-5_21.

[14] S. Woo, D. Kim, D. Cho, I. S. Kweon. LinkNet: Relational embedding for scene graph. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Montreal, Canada, pp. 558–568, 2018.

[15] T. S. Chen, W. H. Yu, R. Q. Chen, L. Lin. Knowledge-embedded routing network for scene graph generation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 6156–6164, 2019. DOI: 10.1109/CVPR.2019.00632.

[16] R. J. Li, S. Y. Zhang, B. Wan, X. M. He. Bipartite graph network with adaptive message passing for unbiased scene graph generation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 11104–11114, 2021. DOI: 10.1109/CVPR46437.2021.01096.

[17] M. Suhail, A. Mittal, B. Siddiquie, C. Broaddus, J. Eledath, G. Medioni, L. Sigal. Energy-based learning for scene graph generation. In *Proceedings of IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, IEEE, Nashville, USA, pp. 13931–13940, 2021. DOI: 10.1109/CVPR46437.2021.01372.

[18] D. P. Kingma, M. Welling. Auto-encoding variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations*, Banff, Canada, 2014.

[19] J. Wang, Y. W. Pan, T. Yao, J. H. Tang, T. Mei. Convolutional auto-encoding of sentence topics for image paragraph generation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, Macao, China, pp. 940–946, 2019.

[20] L. Melas-Kyriazi, A. Rush, G. Han. Training for diversity in image paragraph captioning. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, ACL, Brussels, Belgium, pp. 757–761, 2018. DOI: 10.18653/v1/D18-1084.

[21] Z. J. Zha, D. Q. Liu, H. W. Zhang, Y. D. Zhang, F. Wu. Context-aware visual policy network for fine-grained image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 2, pp. 710–722, 2022. DOI: 10.1109/TPAMI.2019.2909864.

[22] C. P. Xu, Y. Li, C. M. Li, X. Ao, M. Yang, J. W. Tian. Interactive key-value memory-augmented attention for image paragraph captioning. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain, pp. 3132–3142, 2020. DOI: 10.18653/v1/2020.coling-main.279.

[23] X. Yang, C. Y. Gao, H. W. Zhang, J. F. Cai. Hierarchical scene graph encoder-decoder for image paragraph captioning. In *Proceedings of the 28th ACM International Conference on Multimedia*, Seattle, USA, pp. 4181–4189, 2020. DOI: 10.1145/3394171.3413859.

[24] Y. H. Shi, Y. Liu, F. X. Feng, R. F. Li, Z. Y. Ma, X. J. Wang. S2TD: A tree-structured decoder for image paragraph captioning. In *Proceedings of ACM Multimedia Asia*, Gold Coast, Australia, pp. 5, 2021. DOI: 10.1145/3469877.3490585.

[25] C. P. Xu, M. Yang, X. Ao, Y. Shen, R. F. Xu, J. W. Tian. Retrieval-enhanced adversarial training with dynamic memory-augmented attention for image paragraph captioning. *Knowledge-based Systems*, vol. 214, Article number 106730, 2021. DOI: 10.1016/j.knosys.2020.106730.

[26] D. D. Guo, R. Y. Lu, B. Chen, Z. Q. Zeng, M. Y. Zhou. Matching visual features to hierarchical semantic topics for image paragraph captioning. *International Journal of Computer Vision*, vol. 130, no. 8, pp. 1920–1937, 2022. DOI: 10.1007/s11263-022-01624-6.

[27] S. Q. Ren, K. M. He, R. Girshick, J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, Montreal, Canada, pp. 91–99, 2015.

[28] T. Yao, Y. W. Pan, Y. H. Li, T. Mei. Exploring visual relationship for image captioning. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 711–727, 2018. DOI: 10.1007/978-3-030-01264-9_42.

[29] P. Veličković, A. Casanova, P. Liò, G. Cucurull, A. Romero, Y. Bengio. Graph attention networks. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, Canada, 2018. DOI: 10.17863/CAM.48429.

[30] R. Vedantam, C. L. Zitnick, D. Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 4566–4575, 2015. DOI: 10.1109/CVPR.2015.7299087.

[31] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the 13th European Conference on Computer Vision*, Springer, Zurich, Switzerland, pp. 740–755, 2014. DOI: 10.1007/978-3-319-10602-1_48.

[32] R. Krishna, Y. K. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. J. Li, D. A. Shamma, M. S. Bernstein, F. F. Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017. DOI: 10.1007/s11263-016-0981-7.

[33] K. Papineni, S. Roukos, T. Ward, W. J. Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL, Philadelphia, USA, pp. 311–318, 2002. DOI: 10.3115/1073083.1073135.

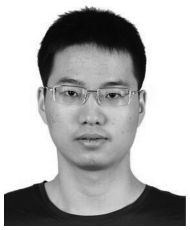
[34] M. Denkowski, A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the 9th Workshop on Statistical Machine Translation*, ACL, Baltimore, USA, pp. 376–380, 2014. DOI: 10.3115/v1/W14-3348.

[35] J. Pennington, R. Socher, C. Manning. GloVe: Global vectors for word representation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, ACL, Doha, Qatar, pp. 1532–1543, 2014. DOI: 10.3115/v1/D14-1162.

[36] K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of IEEE International Conference on Computer Vision*, Santiago, Chile, pp. 1026–1034, 2015. DOI: 10.1109/ICCV.2015.123.

[37] D. Kingma, J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, USA, 2015.

[38] E. Cohen, C. Beck. Empirical analysis of beam search performance degradation in neural sequence models. In *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, USA, pp. 1290–1299, 2019.

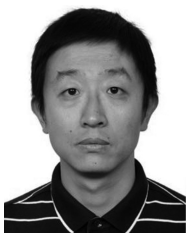


Xianglu Zhu received the B. Sc. degree in automation from University of Science and Technology of China (USTC), China in 2016. He is currently a Ph. D. degree candidate in automation at USTC, and is an intern at National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), China.

His research interests include image caption, pose estimation and deep learning.

E-mail: zxl9531@mail.ustc.edu.cn

ORCID iD: 0000-0001-6288-9237



Zhang Zhang received the B. Sc. degree in computer science and technology from Hebei University of Technology, China in 2002, and the Ph. D. degree in pattern recognition and intelligent systems from National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China in 2009. Currently, he is an associate professor at Na-

tional Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences (CASIA), China. He has published more than 40 research papers on computer vision and pattern recognition, including some highly ranked journals and conferences, e.g., IEEE TPAMI, IEEE TIP, CVPR, and ECCV.

His research interests include action and activity recognition, human attribute recognition, person re-identification, and large-scale person retrieval.

E-mail: zzhang@nlpr.ia.ac.cn (Corresponding author)
ORCID iD: 0000-0001-9425-3065

Wei Wang received the B. Eng. degree in automation from Wuhan University, China in 2005, and the Ph. D. degree in information science and engineering from University of Chinese Academy of Sciences, China in 2011. He is currently an associate professor in National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), China. He has published more than fifty papers in refereed international journals and conferences such as TPAMI, TIP, CVPR, ICCV and NeurIPS.

His research interests include computer vision and machine learning, particularly on the computational modelling of visual attention and memory, vision and language understanding.

E-mail: wangwei@bigai.ai

Zilei Wang received the B. Sc. and Ph. D. degrees in control science and engineering from University of Science and Technology of China (USTC), China in 2002 and 2007, respectively. He is currently an associate professor with Department of Automation, USTC, and the founding leader of the Vision and Multimedia Research Group. Before joining USTC as a faculty, he was a postdoctoral research fellow with National University of Singapore, Singapore.

His research interests include computer vision, multimedia and deep learning.

E-mail: zlwang@ustc.edu.cn