A TEXT-GUIDED GRAPH STRUCTURE FOR IMAGE CAPTIONING

Depeng Wang, Zhenzhen Hu, Yuanen Zhou, Xueliang Liu, Le Wu, Richang Hong

Hefei University of Technology, Hefei, Anhui 230601, China
wdp18815514247@163.com, {huzhen.ice, y.e.zhou.hb, liuxueliang1982, lewu.ustc, hongrc.hfut}@gmail.com

ABSTRACT

Image captioning task requires a comprehensive understanding of visual content and has received a significant amount of attention. Recent studies have revealed that modelling relationships between visual objects imply a high-level semantic feature. However, most existing relationship modelling methods for image captioning heavily rely on the object detection results and handcrafted structured label to build the graph model. In this paper, we explore the relationships in a text-guided way via the descriptions from similar images to provide the context clues. We propose a novel framework named Text-Guided Graph (TGG) to employ image-related text to help build the relationship between objects in the image and incorporate the high-level graph information and captions associated with a certain image. Experiments conducted on the MS COCO dataset demonstrate the effectiveness of our text-guided graph model under various standard evaluation metrics.

Index Terms— Image captioning, Graph Convolutional Networks, Relationship

1. INTRODUCTION

Image captioning aims to automatically generate natural language description of an image [1]. As a new rising interdiscipline between computer vision and natural language processing, it is a high-level and complicated task requiring a comprehensive understanding of visual contents including a variety of entities as well as their relationships. This task attracts wide attention due to its value in practical applications, such as helping the impaired individuals to navigate their surroundings.

Inspired by the machine translation, the prominent pipeline of image captioning is to translate the image into a sentence via an encoder-decoder framework [1, 2]. This framework exploits a Convolutional Neural Network (CNN) as an encoder to extract image representations and utilizes a Recurrent Neural Network (RNN) as a decoder to generate the language sequences [3, 4]. Based on this framework, the significant upgrades are facilitated by the attention mechanism [5, 6, 7].

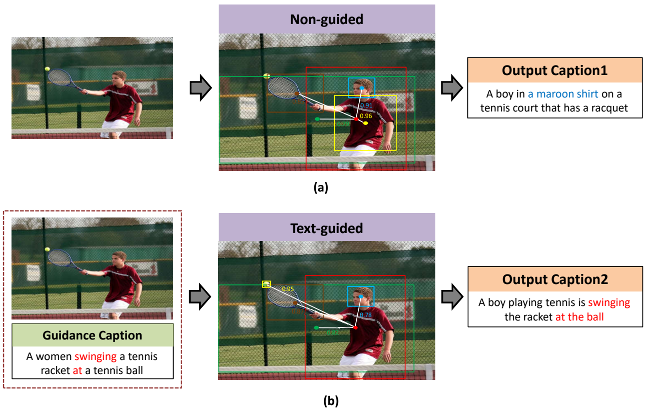

Non-guided
0.91
0.96
Output Caption1
A boy in a maroon shirt on a tennis court that has a racquet
(a)
Guidance Caption
A women swinging a tennis racket at a tennis ball
Text-guided
0.95
0.78
Output Caption2
A boy playing tennis is swinging the racket at the ball
(b)

Fig. 1. Text-guided Graph model illustration. In (a), the graph structure is directly constructed based on the object detector. In (b), the guidance caption provides the context clues to drive graph building to focus on key objects and relationships, resulting in more abundant and accurate image captions.

These methods only focus on the individual visual elements while ignoring the relationships between semantic entities. Vision and language researchers turn attention to the visual relationship discovery recently. The pioneering work from Yao *et al.* [2] extracted the internal information from the given image to learn relationships between objects. Yang *et al.* [8] utilized the scene graph to provide a deeper understanding of semantic relationships and shown the promising results. With the assistant from the additional consideration of relationships, the image representation has been boosted, so as to produce more abundant and accurate sentences. However, the internal information can only provide limited visual clues while the methods based on scene graph are heavily dependent on an external knowledge graph [9], which requires a lot of manual labelling.

Visually similar images may share similar descriptions which imply the reasoning knowledge to guide the target image caption generation. From this perspective, in this paper, we propose a novel captioning model, named Text-Guided Graph (TGG), to guide the image captioning process by exploring the external reasoning knowledge from the captioning dataset. Without a pre-defined knowledge base, the TGG model leverages the image-related text from the captioning dataset as the guidance caption to assist the relationship build-

Corresponding author: Zhenzhen Hu, This work is supported by the National Nature Science Foundation of China under the Grant No. 61802104 and 61976076.

978-1-7281-1485-9/20/$31.00 ©2020 IEEE

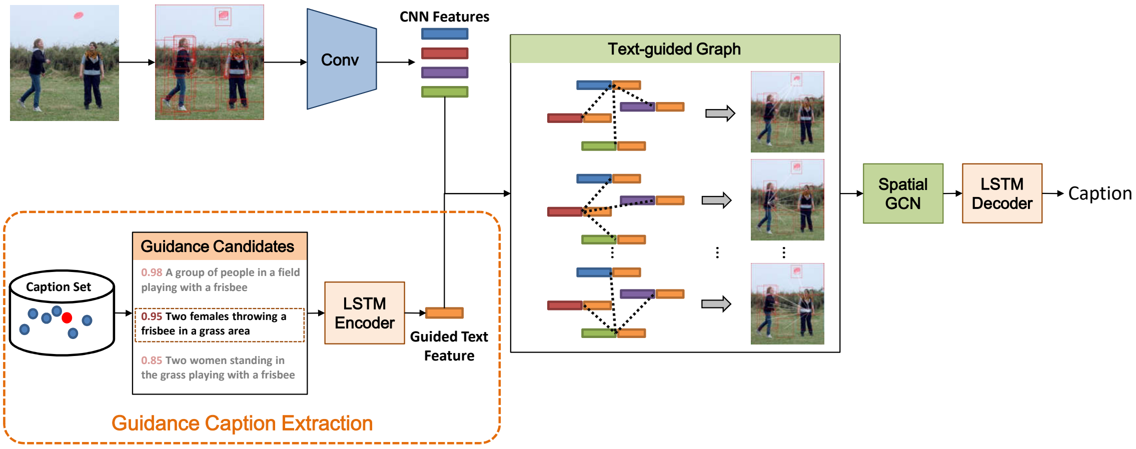

CNN Features
Conv
Text-guided Graph
Spatial GCN
LSTM Decoder
Caption
Caption Set
Guidance Candidates
0.98 A group of people in a field playing with a frisbee
0.95 Two females throwing a frisbee in a grass area
0.85 Two women standing in the grass playing with a frisbee
LSTM Encoder
Guided Text Feature
Guidance Caption Extraction

Fig. 2. Overall framework illustration of Text-Guided Graph model. We extract the region visual representations based on the CNN object detector. Among the image captioning dataset, we recall captions attached to the visually similar images and choose the top 3 of them according to the caption score as the guidance candidates. The guidance caption is randomly selected among the candidates and transformed into the vector by the LSTM encoder. Two modality features are concatenated as the graph nodes to construct the relational graph. For each node, we calculate the impact of every neighbouring feature. The relational representation vector is produced by adopting the means of spatial graph convolutions and decoded by LSTM to generate the caption.

ing between the objects. For example, as shown in Figure 1 (a), the caption captured the region of "a maroon shirt" and ignored the more important semantic clue "ball" without the guidance text, because the latter takes up a tiny place over the image. Comparatively, in Figure 1 (b), the guidance caption provides the related element of "tennis ball" and gives it a more significant relationship weight. The final output both capture the visual context and the relationship between the objects. Practically, we retrieve the visually similar image in the captioning dataset to extract the related sentences and select one as the guidance caption. The visual feature of detected objects and the representation of guidance caption are combined as the graph nodes, which drive to building the graph structure to learn cross-modality representations. Finally, an LSTM-based decoder translates the syncretic feature into a sentence.

The TGG model has two innovations. First, we introduce the sentences associated with a certain image to extend the diversity of input information. Besides, we not only pay attention to objects in the image but also focus on relationships constructed by the image-related text. Intuitively, the relationship between the objects in the sentences we generate is richer and more accurate.

The contributions of this paper are as follows:

- We propose a Text-Guided Graph model for image captioning generation, which employs the image-related text to help build the relationships between objects in the image and incorporates the high-level graph information and captions associated with a certain image. To our knowledge, our work is the first to utilize image-related text to build relationships between objects for image captioning.
- Different from the most existing knowledge-based vision-to-language frameworks, our model extracts the reasoning knowledge, i.e. the guidance captions, without a pre-defined knowledge base to reduce the amount of manual work.
- We demonstrate that captions produced by our model contain more and accurate relationship information and achieve comparable results against the state-of-the-art methods under various standard evaluation metrics on the MSCOCO dataset.

2. RELATED WORK

The successful application of deep neural networks in machine translation has promoted the solution of image captioning problem. In recent years, a large number of methods have been proposed based on the encoder-decoder framework [1, 10, 7, 11]. The well trained CNN model is used to encode images and an RNN is deployed to decode language sequences. SAT [6] is the first work to introduce the attention module into this framework with promising results and the follow-up works promoted this mechanism in various ways [12, 5, 13]. Anderson *et al.* [5] proposed to detect a set of salient image regions via bottom-up attention mechanism and then attend to the salient regions with top-down attention mechanism for sentence generation, resulting in striking performance improvement.

Although these methods achieved impressive successes, they tend to only focus on the individual visual elements while ignore the relationships between semantic entities which leads to generate relative rigid sentences. Most recent research finds that the relationships between semantic elements is an important property for the generated captions. Considering this point, Zellers *et al.* [14] and Yang *et al.* [8] used scene graph as the bridge to combine the objects, attributes and relationships together to generate more meaningful contexts. In this paper, we employ image-related sentences to help build the graph for image captioning.

3. METHOD

We propose a Text-Guided Graph (TGG) model to generate image descriptions via the extra consideration of relationships between objects and sequences related to the image. An overview of our architecture is illustrated in Figure 2. Based on the object detection results, we highlight the semantic areas crossing the image. In the guidance caption extraction module, we retrieve the visual similar image in the captioning dataset to extract the related sentences and select one as the guidance caption. An LSTM-based network is implemented to encode the caption to obtain the guidance caption vector. The visual feature of detected objects and the representation of guidance caption are combined as the graph nodes, which drive to build the graph structure to learn the cross-modality representations. Spatial graph convolution is performed to encode relational graph to result in relational representation vector, based on which the decoder is utilized to generate image descriptions.

3.1. Guidance Caption Extraction

The guided caption implies the reliable clues to guide the relationship generation, thus it is a crucial part of our model. In order to extract the appropriate guided caption from the dataset, we design a guidance caption extraction module. Mun *et al.* [15] indicated that visually similar images tend to have preference objects and meanings, thus guidance captions are extracted as follow. Given an target image, we first select the top 3 features according to confidence score from all object-level features extracted by Faster R-CNN [16]. Similar process for each image in the training set, then calculating visual similarity v between target image I_{tar} and each image I_{tra} in the training set and further filtrate top m images as caption set based on visual similarity:

$$v = \sum_{i=1}^{3} max(Cos(I_{tar_i}, I_{tra_j}), j \in \{1, 2, 3\}) \qquad (1)$$

where $Cos(.)$ denotes cosine similarity, I_{tar_i} and I_{tra_j} mean the i^{th} and j^{th} object-level feature of the target image and every image on training dataset respectively. Caption set $\{C_i\}, i = 1, ..., N$ consisted of the associated captions of m similar images is sorted according to the score s_i of each C_i, which is calculated by average similarity to all captions in $\{C_i\}$:

$$s_i = \frac{1}{N} \sum_{j=1}^{N} Sim(C_i, C_j) \qquad (2)$$

where $Sim(C_i, C_j)$ is the similarity between two captions C_i and C_j by TF-IDF [17].

We select top n captions from caption set according to caption scores as guidance candidates and randomly sample one of them as the guidance caption, as shown in Figure 3. The guidance caption is encoded by the pre-trained word embedding and a dynamic RNN to convert it to a single vector g.

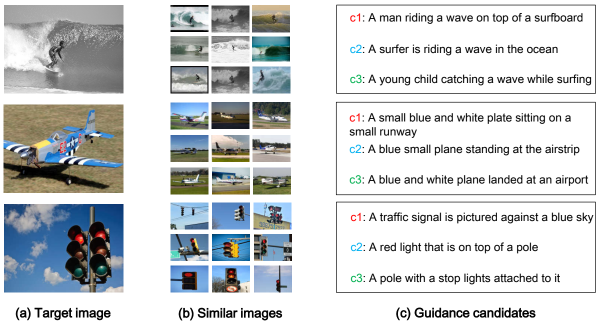

c1: A man riding a wave on top of a surfboard
c2: A surfer is riding a wave in the ocean
c3: A young child catching a wave while surfing
c1: A small blue and white plate sitting on a small runway
c2: A blue small plane standing at the airstrip
c3: A blue and white plane landed at an airport
c1: A traffic signal is pictured against a blue sky
c2: A red light that is on top of a pole
c3: A pole with a stop lights attached to it
(a) Target image
(b) Similar images
(c) Guidance candidates

Fig. 3. Examples of similar images and guidance candidates. Given a target image (a), we retrieve $m = 9$ nearest neighbor images (b) from the training data and extract the attached captions as the caption set. Then we choose top $n = 3$ sentences from the caption set according to the score as the guidance candidates (c).

3.2. Graph Builder

Inspired by [18], we first leverage Faster R-CNN [16] to produce fixed L object regions for every image and then treat every region as one vertex to construct relational graph $G = \{V, E, A\}$ conditioned on the guidance caption, where V denotes the group of detected object vertices consisting of bounding box coordinates and image features vectors, E means a collection of graph edges to learn and A represents the corresponding adjacency matrix. Our goal is to learn a specific-text adjacency matrix A in which each edge $(i, j, A_{i,j}) \in E$ denotes the strength of relationship between vertex i and vertex j based on guidance caption. This is done by modelling the similarity between detected object vertices and their association with the guidance caption. The specific operation is first to connect the guidance caption embedding g behind each of the L detected object vertices v_l, which we write as $[v_l|g]$ and then to obtain a fused feature e_l as:

$$e_l = F([v_l|g]), \qquad l = 1, 2, ..., L \qquad (3)$$

where $F : \mathcal{H}^{d_v+d_g} \rightarrow \mathcal{H}^{d_f}$ is a non-linear function and d_v, d_g, d_f denote the dimensions of the detected object feature vectors, guidance caption embedding and fused feature vectors respectively. Next, we combine all the fused feature e_l into one matrix $E_{L\times e_l}$, and further obtain a specific-text adjacency matrix A as : $A = EE^{\mathrm{T}}$. Besides, the strength of relationship between vertex i and vertex j is defined as : $A_{i,j} = e_i^{\mathrm{T}} e_j$.

The fully connected adjacency matrix A that does not consider any restrictions on the sparsity of graph result in a waste of computing resources. Besides, it also brings in a lot of unimportant information to affect the feature representation conditioned on the most relevant neighbours. Therefore, we adopt a ranking strategy to select the most relevant neighbours for graph nodes, which can save computing resources and focus more:

$$\mathcal{N}(i) = topk(\mathrm{a}_i) \tag{4}$$

where a_i is the i^{th} row of the adjacency matrix, and $topk$ function outputs the indices of the k largest values in a_i. This is to say, the neighbourhood system of a given node consists of nodes with which it has the strongest connections.

3.3. Spatial Graph Convolutions

Relational graph obtained from above section contains locations of objects in the image, solving the problem of relative position of objects in an image that is ignored by a lot of image captioning models. Inspired by [19, 18], we choose to employ a graph convolution approach that operates directly in the graphics domain and relies heavily on spatial relationships to process relational graph to obtain new feature vectors. In addition, a function $\mathbf{o}$ is defined to describe the relative spatial relationship, such as $\mathbf{o}(i, j)$ denotes the vertex j coordinate in the system that centred at the vertex i. And our model captures spatial relationships between detected objects in the image with the help of function $\mathbf{o}$.

A difficult point and focus in graph convolutions is to define a variable to reflect the impact of every neighboring features on a weighted sum of the neighbouring features. [19] introduces a method that using a group of R Gaussian kernels with learnt means and covariances to solve the above problem. Considering that the impact of each feature on a weighted sum of the neighbouring features is not only related to itself, but also to the relationship between them, we redefine a variable at kernel r for node i as :

$$\mathbf{f_r}(i) = \sum_{j \in \mathcal{N}(i)} w_r(\mathbf{o}(i,j))\mathbf{v}_j \alpha_{ij}, \quad r = 1, 2, ..., R \tag{5}$$

where $\mathcal{N}(i)$ denotes the neighbourhood of vertex i and $w_r(.)$ represents a kernel weight for the r^{th} Gaussian kernel. With $\alpha_{ij} = s(\mathrm{a}_i)_j$, where $s(.)_j$ means performing softmax operation and indexes the j^{th} element.

Finally, we obtain the output of spacial graph convolution at vertex i by concatenating the results of R kernels:

$$\mathbf{z}_i = \|_{r=1}^{R} \mathbf{G}_r \mathbf{f}_r(i) \tag{6}$$

where $\mathbf{G}_r \in \mathcal{H}^{\frac{d_z}{R} \times d_v}$ is a matrix of learnable weights for the n^{th} Gaussian kernel and d_z denotes the dimensionality of outputted graph convoluted features.

3.4. Decoder and Loss Function

For a triplet of input including an image I, a guidance caption g and a target caption c consisting of T words $(w_1, w_2, ..., w_T)$, we add two words w_0 (<BOS>) and w_{T+1} (<EOS>) at the begin and end of a caption c, respectively. Our decoder is formulated by

$$\mathbf{x}_{-1} = \mathbf{W}_z \mathbf{z} \tag{7}$$

$$\mathbf{x}_t = \mathbf{W}_e \mathbf{w}_t \tag{8}$$

$$\mathbf{h}_t = \mathrm{LSTM}(\mathbf{x}_t, \mathbf{h}_{t-1}) \tag{9}$$

$$\mathbf{p}_{t+1} = \mathrm{Softmax}(\mathbf{W}_h \mathbf{h}_t) \tag{10}$$

where the various $\mathbf{W}_z$, $\mathbf{W}_e$ and $\mathbf{W}_h$ matrices are learnt parameters for the context vector, the input word and the hidden state. At each time step t, the input word $\mathbf{w}_t$ is embeded to $\mathbf{x}_t$ and the current hidden state $\mathbf{h}_t$ is calculated by the word vector $\mathbf{x}_t$ and the previous hidden state $\mathbf{h}_{t-1}$. Next, $\mathbf{h}_t$ is fed to a Softmax to produce a probability distribution $\mathbf{p}_{t+1}$ over all words, with the model show the predicted output word for time step $t + 1$.

In the aspect of loss function, our image captioning model is trained to minimize the cross entropy as follows:

$$\begin{aligned} \mathcal{L} &= -\log \mathrm{P}(c|f_{graph}(I, g)) \\ &= -\log \mathrm{P}(w_1|w_0, f_{graph}(I, g)) \\ &+ \sum_{t=1}^{T} -\log \mathrm{P}(w_{t+1}|w_t, h_{t-1}) \end{aligned} \tag{11}$$

where f_{graph} is the proposed text-guided graph model and computed only once at the beginning. Besides, h_{t-1} is the previous hidden state of LSTM, w_0(<BOS>) and w_{T+1}(<EOS>) denotes the begin and end of a sentence, respectively. Compared with previous approaches, construction of our relational graph is driven by text features obtained from the guidance caption in addtion to visual feature of the image.

4. EXPERIMENTS

We evaluate our TGG model on the MS-COCO captioning dataset [23], which contains 123,287 images and each image is annotated with 5 sentences. Since the annotations of official test set are not public, we use the same splits provided by [3], which take 113,287 images for training, 5,000 images for validation and 5,000 images for testing. Moreover, we convert all annotations in the training set to lowercase and drop any word that has count less than five, resulting in a vocabulary of size 9,487 words.

4.1. Experiment Setup

We extract the region proposals by the CNN-based object detector [16]. For each image, we fix the number of object regions as 36 and the dimension of visual feature is 2052 (e.g. 2048-dimensional feature vectors attached with 4-dimensional absolute spatial information). We set $m = 10$ and $n = 3$ in the guidance caption extraction section and the guidance caption is encoded by a dynamic LSTM with a hidden state size of 512.

Additionally, the non-linear function F(see Eq. 3) fuses image feature vectors ($d_v = 2052$) and caption feature vectors ($d_g = 512$) into 512 dimension and we select the top 16 indexes of each row of the adjacency matrix as neighbours in graph builder module. During the stage of graph convolution, we employ two spatial graph convolution layers with dimensions 2048 and 1024 respectively, both layers have 8 Gaussian kernels. Besides, we adopt dropout to prevent overfitting and the Adam optimizer [24] with a learning rate of 0.0005 which we decrease 0.2 times every 3 epochs until it drop to 0.0001 after the 10^{th} epoch during training.

Table 1. Results of different caption models on MSCOCO dataset. All values are reported as percentage(%). * stands for more complicated decoder used in up-down model. (-) indicates that the metric is not provided.

Caption models	MSCOCO dataset						
	B-1	B-2	B-3	B-4	M	R	C
NICs [6]	66.6	46.1	32.9	24.6	-	-	-
ATT [20]	70.9	53.7	40.2	30.4	24.3	-	-
SGC [21]	67.1	48.8	34.3	23.9	21.8	48.8	73.3
Att-RegionCNN+LSTM [22]	74.0	56.0	42.0	31.0	26.0	-	94.0
ATT-kCC [15]	74.9	58.1	43.7	32.6	25.7	-	102.4
Up-down [5]*	77.2	60.0	47.3	36.2	27.0	56.4	113.5
GC	67.1	48.7	34.5	24.4	22.5	-	79.3
Graph	73.5	55.1	40.9	30.9	25.5	54.6	94.8
Graph*	77.9	60.5	47.6	36.5	27.2	56.8	114.4
Text-Guided Graph(TGG)	75.0	56.1	41.7	31.4	25.8	55.7	95.9
Text-Guided Graph(TGG)*	**78.8**	**61.3**	**48.2**	**36.7**	**27.8**	**57.5**	**115.8**

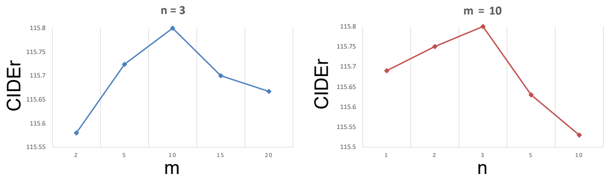


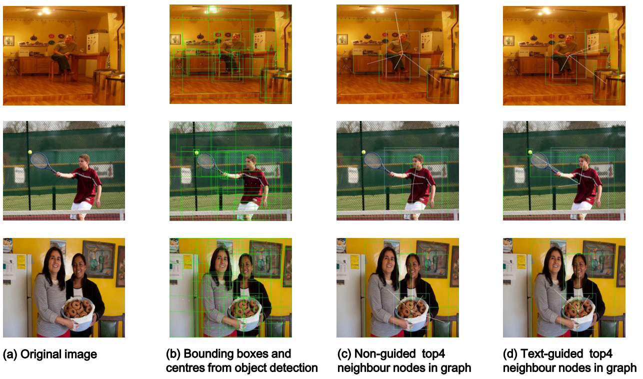
Fig. 4. The evaluation of CIDEr performance of hyper parameters n and m based on the MSCOCO.

In testing, we apply the beam search of size 5 to generate captions and evaluate our model on the common metrics for image captioning, BLEU, METEOR, CIDEr and ROUGE-L [25]. The performance on all metrics is computed by using MS-COCO caption evaluation tool [26].

4.2. Quantitative Analysis

To clarify the effect of the quality of guidance captions, we illustrate the performance curves over the evaluation metric CIDEr with hyper parameters m (the number different nearest neighbor images) and n (the number of sentences) in Figure 4. As shown in the figure, we can see that the performance curves are generally like "$\wedge$" shape. The best performance is achieved when $m = 10$ and $n = 3$. This proves that it is reasonable to exploit the appropriate guidance caption for boosting image captioning.

We evaluate the TGG model with ablative studies and compare the performances with several state-of-the-art algorithms of image captioning, such as NICs [6], ATT [20], SGC [21], Att-RegionCNN-LSTM [22], ATT-kCC [15] and Up-down [5]. The results are shown in Table 1. The bottom part reports the performance of baselines. The first baseline (GC) outputs the guidance captions as the final result of image captioning. The second (Graph) directly builds graph structure based on the detected objects of given image and adopts the means of graph convolutions to learn the image representation as the input of decoder. Based on the second method, our TGG model adds guidance captions to help construct graph structure, as show in the last two lines in Table 1. We can see that our Graph* model outperforms the Up-down model that also using object detection algorithm and attention decoder but without a view of relationship between objects, and our TGG model also surpasses the Att-RegionCNN-LSTM [22] that using the detected objects and additional text information, but not considering relationship between objects in the image. From the results we can see that the learnt relationship between objects can boost the image captioning.

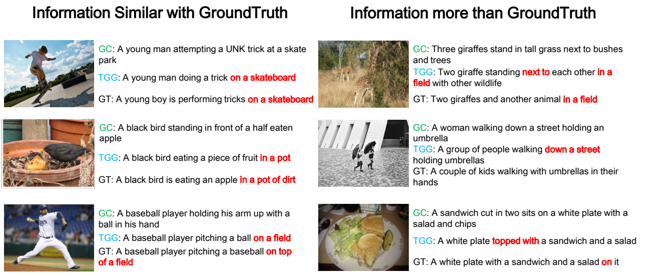


(a) Original image
(b) Bounding boxes and centres from object detection
(c) Non-guided top4 neighbour nodes in graph
(d) Text-guided top4 neighbour nodes in graph

Fig. 5. Examples of the non/text guided graph structures.

4.3. Qualitative Results

We compare the difference between the graph structures that with and without the guidance captions as shown in Figure 5. The visualization examples show that with the guidance caption, the relationship of objects pay more attention to the relative graph nodes, which is proving that guidance captions play an important role in the construction of the graph. In Figure 6, we show the results of image captioning, from which we can find that the sentences created by our method are more informative. In the first column, the relationships between objects



Information Similar with GroundTruth

GC: A young man attempting a UNK trick at a skate park

TGG: A young man doing a trick **on a skateboard**

GT: A young boy is performing tricks **on a skateboard**

GC: A black bird standing in front of a half eaten apple

TGG: A black bird eating a piece of fruit **in a pot**

GT: A black bird is eating an apple **in a pot of dirt**

GC: A baseball player holding his arm up with a ball in his hand

TGG: A baseball player pitching a ball **on a field**

GT: A baseball player pitching a baseball **on top of a field**

Information more than GroundTruth

GC: Three giraffes stand in tall grass next to bushes and trees

TGG: Two giraffe standing **next to** each other **in a field** with other wildlife

GT: Two giraffes and another animal **in a field**

GC: A woman walking down a street holding an umbrella

TGG: A group of people walking **down a street** holding umbrellas

GT: A couple of kids walking with umbrellas in their hands

GC: A sandwich cut in two sits on a white plate with a salad and chips

TGG: A white plate **topped with** a sandwich and a salad

GT: A white plate with a sandwich and a salad **on** it

Fig. 6. Examples of qualitative results. Each example attached with three captions: the guidance caption (top), the generated caption from our TGG model (middle) and ground truth (bottom).

in the sentences generated by TGG model are very similar to the words in the ground truth. Furthermore, we show more rich relationship information than the ground truth in the second column, which further demonstrate the effectiveness of our approach.

5. CONCLUSION

In this paper, we propose a Text-Guided Graph model to explore the impact of visual relationships on image captioning. We build the relationship between objects in a text-guided way via the context clues provided by the descriptions from similar images. Furthermore, we incorporate the high-level graph information and captions associated with a certain image to extend the diversity of information for boosting image captioning. Experimental results demonstrate our method achieves good performance on common evaluation metrics and intuitively generates more abundant and accurate captions.

6. REFERENCES

[1] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015.

[2] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei, "Boosting image captioning with attributes," in *ICCV*, 2017.

[3] Andrej Karpathy and Li Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015.

[4] Feng Liu, Tao Xiang, Timothy M Hospedales, Wankou Yang, and Changyin Sun, "Semantic regularisation for recurrent image annotation," in *CVPR*, 2017.

[5] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR*, 2018.

[6] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *IMCL*, 2015.

[7] Yuanen Zhou, Zhenzhen Hu, Ye Zhac, Xueliang Liu, and Richang Hong, "Enhanced text-guided attention model for image captioning," in *BigMM*, 2018.

[8] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai, "Auto-encoding scene graphs for image captioning," in *CVPR*, 2019.

[9] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, 2017.

[10] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge," *PAMI*, 2016.

[11] Jicheng Wang, Yuanen Zhou, Zhenzhen Hu, Xu Zhang, and Meng Wang, "Sequential image encoding for vision-to-language problems," *Multimedia Tools and Applications*, pp. 1–12, 2019.

[12] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *CVPR*, 2017, pp. 5659–5667.

[13] Daqing Liu, Zheng-Jun Zha, Hanwang Zhang, Yongdong Zhang, and Feng Wu, "Context-aware visual policy network for sequence-level image captioning," in *ACM MM*, 2018, pp. 1416–1424.

[14] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi, "Neural motifs: Scene graph parsing with global context," in *CVPR*, 2018.

[15] Jonghwan Mun, Minsu Cho, and Bohyung Han, "Text-guided attention model for image captioning," in *AAAI*, 2017.

[16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.

[17] Juan Ramos et al., "Using tf-idf to determine word relevance in document queries," in *ICML*, 2003.

[18] Will Norcliffe-Brown, Stathis Vafeias, and Sarah Parisot, "Learning conditioned graph structures for interpretable visual question answering," in *NIPS*, 2018.

[19] Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M Bronstein, "Geometric deep learning on graphs and manifolds using mixture model cnns," in *CVPR*, 2017.

[20] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo, "Image captioning with semantic attention," in *CVPR*, 2016.

[21] Ning Xu, An-An Liu, Jing Liu, Weizhi Nie, and Yuting Su, "Scene graph captioner: Image captioning based on structural visual representation," *JVCIR*, 2019.

[22] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton van den Hengel, "Image captioning and visual question answering based on attributes and external knowledge," *PAMI*, 2017.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

[24] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] C Lin, "Recall-oriented understudy for gisting evaluation (rouge)," *Retrieved August*, 2005.

[26] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.