Paragraph Generation Network with Visual Relationship Detection

Wenbin Che
Harbin Institute of Technology
Harbin, China
chewenbin@hit.edu.cn

Xiaopeng Fan*
Harbin Institute of Technology
Harbin, China
fxp@hit.edu.cn

Ruiqin Xiong
Peking University
Beijing, China
rqxiong@pku.edu.cn

Debin Zhao
Harbin Institute of Technology
Harbin, China
dbzhao@hit.edu.cn

ABSTRACT

Paragraph generation of images is a new concept, aiming to produce multiple sentences to describe a given image. In this paper, we propose a paragraph generation network with introducing visual relationship detection. We first detect regions which may contain important visual objects and then predict their relationships. Paragraphs are produced based on object regions which have valid relationship with others. Compared with previous works which generate sentences based on region features, we explicitly explore and utilize visual relationships in order to improve final captions. The experimental results show that such strategy could improve paragraph generating performance from two aspects: more details about object relations are detected and more accurate sentences are obtained. Furthermore, our model is more robust to region detection fluctuation.

CCS CONCEPTS

• **Computing methodologies → Computer vision**; *Natural language generation*; • **Computer systems organization → Neural networks**;

KEYWORDS

image caption; object detection; paragraph generation; relationship detection

ACM Reference Format:
Wenbin Che, Xiaopeng Fan, Ruiqin Xiong, and Debin Zhao. 2018. Paragraph Generation Network with Visual Relationship Detection. In *2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea,* Jennifer B. Sartor, Theo D'Hondt, and Wolfgang De Meuter (Eds.). ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240508.3240695

1 INTRODUCTION

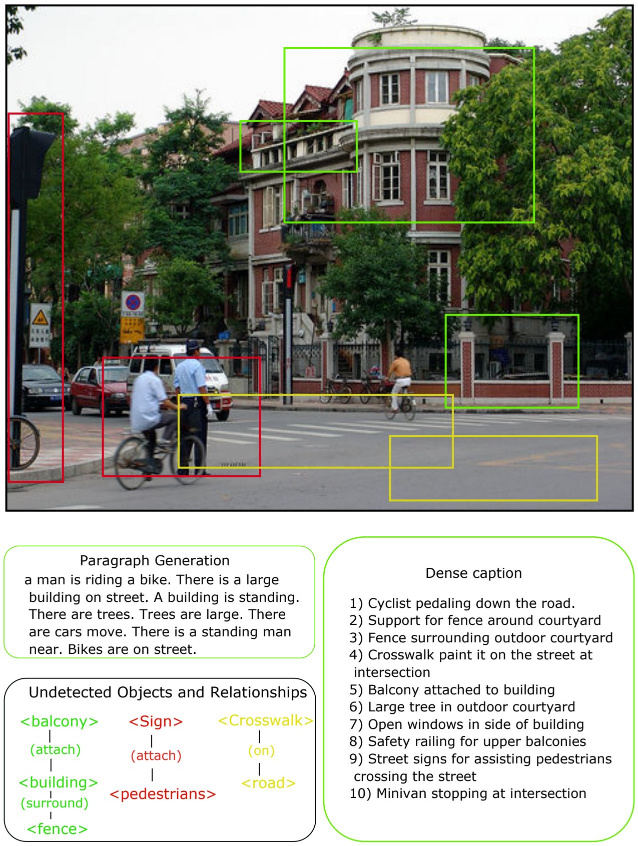
With the rapid development of classification and detection in computer vision(CV) tasks, more and more efforts are made to explore more contents of images. The most intuitive form to depict an image is sentences. Inspired by machine translation techniques and other visual understanding models, early literatures on image captions generate sentences by transforming visual space into text space. Generally speaking, early image caption models take an entire image as input and output a sentence to describe its content. Due to the great success of deep learning, it is possible to introduce deep models, such as Convolutional Neural Network(CNN) and Recurrent Neural Network(RNN), to CV and natural language processing(NLP) tasks. One of the most important advantages for CNN is that image information can be represented by feature maps. For this reason, many CV models use a CNN to extract visual features. As a result, deep learning based image caption models adopt "CNN+RNN" pattern for sentence generation. This pattern is proved to be effective when the given image does not contain substantial divergences. However, as images often contain rich visual contents, giving only one sentence for such description is either limited to the salient objects of the images or tend to broadly depict the entire visual scene [9]. In other words, only coarse depictions can be obtained by this way.

To overcome these limitations, dense captioning is proposed. Each caption is generated through two stages. First, the locations of important objects are detected through a Region Proposal Network(RPN). Then final captions are calculated based on visual features of the corresponding regions. In this respect, dense captioning can be regarded as a model combining two targets: object detection and caption generation. Compared with models which generate sentences for an entire image, dense captioning models provide more details of image regions. However, descriptions produced by dense captioning are not correlated as each sentence is generated only considering a region of the whole image. In order to make these dense captions form a cohesive whole describing the entire image, Krause in [11] proposed a novel captioning

*Corresponding Author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '18, October 22–26, 2018, Seoul, Republic of Korea
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5665-7/18/10...$15.00
https://doi.org/10.1145/3240508.3240695



Paragraph Generation

a man is riding a bike. There is a large building on street. A building is standing. There are trees. Trees are large. There are cars move. There is a standing man near. Bikes are on street.

Undetected Objects and Relationships

Dense caption

1) Cyclist pedaling down the road.
2) Support for fence around courtyard
3) Fence surrounding outdoor courtyard
4) Crosswalk paint it on the street at intersection
5) Balcony attached to building
6) Large tree in outdoor courtyard
7) Open windows in side of building
8) Safety railing for upper balconies
9) Street signs for assisting pedestrians crossing the street
10) Minivan stopping at intersection

Figure 1: The results of paragraph generation [11] and dense caption[9]. Paragraph captions generate more informative and complex results than dense caption, which is a sentence level description. Nevertheless, some details may be lost.

model which is designed to address the shortcomings of both image captioning and the recently-proposed dense image captioning by introducing the task of generating paragraphs for image description. Object regions are first detected and languages are reasoned through a hierarchical RNN, which consists of two modules: a sentence RNN and a word RNN. The sentence RNN is responsible for deciding the number of sentences and outputs a topic vector for the following sentence generation. Each sentence is then predicted through the word RNN. Different from dense captioning, hierarchical RNN takes relations between objects into consideration. It is implemented implicitly by the sentence RNN. However, this implementation may ignore some objects and their relationships, as shown in Fig.1.

To better predict captions which are related to multiple objects, we predict visual relationships of two detected objects and generate sentences using the corresponding relation feature. Inspired by [24] which aims to detect visual relations, we design a relation prediction module and train it to learn relationship features between two objects. In this way, a relation pair <subject, relationship, object> is obtained. The language is then reasoned with the predicted relation pair. Our contribution in this paper consists of two folds:

(1) Given detected regions, we design a discriminant network to decide whether a region is subject or object. Then we predict the relationship between them. We select relation pair achieving high scores for paragraph generation. This mechanism will make the model use more visual information than simply using detected regions from RPN. We take the union of subject and object boxes to represent visual relationship regions.
(2) Most previous works on paragraph generation and image caption produce the final sentence by converting image features to languages directly. In other words, RNN units which are responsible for word generation only take image features as input. Our method use both region features and their semantics to infer final paragraphs. Besides, language prior is also utilized in relationship detection.

2 RELATED WORKS

Generating captions for images is a challenging task, as it requires computers to deal with much more complicated semantics than low level computer vision tasks. Early works regard image captioning as a retrieval problem[5, 8]. With the success of deep learning in computer vision, various CNN models, such as VGG-net, Res-net etc., are proved to be powerful feature extractors and thus are employed in many image captioning works. As RNN based models are widely used in natural language processing like machine translation, they are also applied in language generation in image captioning. Therefore, Most early works follow the CNN+RNN pattern and only output coarse descriptions for an entire image[4, 15, 21].

To meet demands of exploring region details, attention mechanism[1, 14, 17] and dense captioning are proposed. Attention mechanism explores interactions of image regions and indicate where should be paid attention to. Nevertheless, as most of the attention based methods train their models on MS-COCO data set[19], which only consists of images and corresponding sentences and lack region information, it is hard to locate relevant parts accurately. To overcome the uncertainty of region localization, region information is added into Visual Genome data set[12]. Some dense captioning methods[9, 10, 22] are then proposed with the help of this data set. Karpathy and Fei-Fei[10] establish mapping between image regions and languages but do not generate caption for each region. Dense captioning is first introduced in [9] and it is obvious that using region features can give more complicated descriptions than global features. [22] tried to improve performance for dense captioning with the introduction of visual context. To make dense captions form a coherent whole, Jonathan[11] proposed a hierarchical approach for image paragraph generation.

Despite the great improvement of object detection, some CV tasks, such as semantic segmentation[6], action recognition[7] and some works concerned with mapping from images to language[23], require understanding relationships between objects. Relationship prediction is formalized as a task onto itself in [13]. Dai in [2] extended visual relation detection model by adding Statistical information. In [24], it is proved that visual relation can be detected given two related regions by a deep neural network. Motivated by [13] and [24], we extract relation features containing both visual and language information to aid paragraph generation. Visual relationship prediction in recent works takes two forms. One is to regard each visual phrase, *a triplet like <subject-relation-object>*, as a different category and the other is to recognize each component of the triplet individually. As the former suffers from the excessively large number of classes and we only care about the relation contents, we focus on the *relation* component. To make similar relationship to be close in the feature space, we regard each relation candidate as a class and produce a probability distribution over relation candidates and optimize the relationship prediction module by minimizing classification error.

3 METHODS

Overview The overview of the entire model is depicted in Fig.2. The model takes an image as input and outputs several sentences to describe the given image. In this section, we introduce our strategy of generating paragraph description for an input image. We split the whole process into three stages: relation pair detection, visual relationship prediction and caption generation. Compared with the method used by [11], the main advantage is that visual relations are explicitly considered. In [11], all regions of interest are detected first by a RPN. In order to aggregate these region features for describing the contents of image compactly, pooled vectors are computed through a projection and pooling process. The relation information is included implicitly in the pooled vectors.

3.1 Relation pair detection

The first step of our method is to locate regions which may contain important information of the given image. Obviously, this task is very similar to object detections. Works on on dense captioning extend detection models to a novel region proposal network(RPN). Regions with special characteristic can be effectively detected. The paragraph generation model in [11] demonstrated that these regions contain rich information that people may care about. Suppose the input image ,which is denoted by I, is of size $3 \times H \times W$. We adopt the region detector of [9]. Image features V are first extracted through a convolution neural network(CNN) trained from VGG-16 network:

$$V = f_{vgg}(I), V \in \Re^{H' \times W' \times C}$$

where $C = 512, H' = \lfloor \frac{H}{16} \rfloor, andW' = \lfloor \frac{W}{16} \rfloor$. Then we search for the regions of interest through a RPN. For each region, we compute three kinds of information: coordinate of bounding box, region scores and region features. We select regions with top B scores for next stage. We use a tensor v_i with shape of $X \times Y \times C$ to represent the $i-th$ region features.

From the results of [11], we find that visual relations not only exist within one sentence but also between two different sentences. This phenomenon explains the difference between paragraph description and sentence-level caption. For this reason, we design relationship detection module to output a more informative paragraph. We divided all object categories into three sets: Subject set Ω_{sbj}, Object set Ω_{obj} and unknown set. Motivated by [24] which designed a relationship proposal network to solve relationship prediction problems, we add a discriminant network which is responsible for deciding which set these regions belong to. The discriminant network is a multi-layered perceptron with a softmax layer to output probability distribution over these three sets. Among these N detected regions, we assume $K_{sbj}(K_{sbj} < N)$ of them are classified as subjects and $K_{obj}(K_{sbj} < N)$ are objects. Each combination of two sets of elements forms a relation pair $[x_s, x_o]_T$. Note that some regions may be judged in $\Omega_{sbj} \bigcap \Omega_{obj}$ and may also belong to neither one of them. Once we have subject and object regions, we construct relationship candidates by pairing them. The positive relation pairs should meet two conditions: Both subject and object overlap with ground truth are over a threshold $t(IoU > t)$.

3.2 Visual relation prediction

At the second stage, we attempt to explicitly explore the visual relation between subjects and objects. In order to infer relations more accurately, we design our algorithm based on two facts:

(1) Most relations are contained in visual information. Suppose the relationship is denoted by a phrase like *<subject, relationship, object>* in this paper. Given subject and object regions, it is easy for humans to recognize their relationships immediately. However, even for the most sophisticated model, it is still challenging to handle numerous kinds of circumstances. For this reason, we take subject, and object and their union as input for relationship prediction.

(2) Inspired by the success of [2] and [13], we observe that relationship are often decided by some language semantic correlations. For example, if we have known the subject-object pair is *<man,bicycle>*, the relation is supposed to be like *ride* or *on* and not likely to be *eat*. This phenomenon motivates us the introduction of attention mechanism.

Considering these two facts, we design our relationship prediction module. Fig.3 gives the detail. As relation information often exist in both subject and object region, we add union of these two regions to the visual input vector. Hence, the input of relationship prediction module is denoted by $[x_s, x_o, x_u]^T$. All elements are reshaped to column vectors of the same length $L_f = X \times Y \times C$. Suppose we have K_{sbj} subject boxes

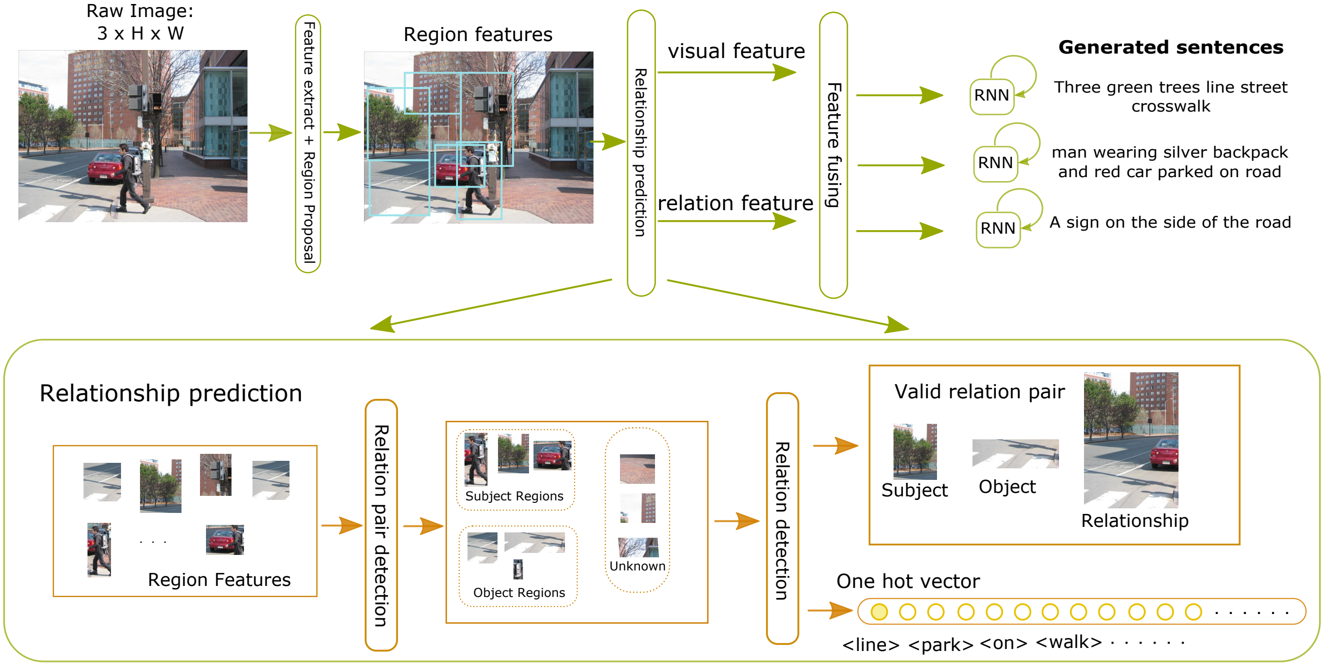


Figure 2: Overview of the paragraph generation model. The main contribution is the Relationship prediction module which generates relation feature for every detected object pair.

and K_{obj} object boxes, the relationship prediction module may generate at most $K_{sbj} \times K_{obj}$ valid relationships.

As discussed above, some relationships are more related to subjects than objects. Take <*man,ride,bicycle*> for example, '*ride*' is obviously more likely to be done by humans. There are also some relationships that are more closer to objects, such as <*girl,drink,water*>. In this example, the prediction of relationship *drink* largely depends on object *water*. Besides, some relations like location relations depend on visual information. In order to solve this issue, we adopt co-attention mechanism to fuse visual features and language features. Attention is often used to calculate expectation of features from CNN or RNN as it can help models focus on interested parts rather than the whole feature map. For this reason, many works on natural language processing and Visual Question Answering(VQA) adopt this strategy. In our case, we employ attention mechanism in order to select relevant features from $\{x_s, x_o, x_u\}$. We first transform them into a common space by the following equation:

$$\begin{aligned} H &= tanh(W_l[x_s, x_o, x_u]) \\ &= [tanh(W_l \cdot x_s), tanh(W_l \cdot x_o), tanh(W_l \cdot x_u)] \\ &= [h_s, h_o, h_u] \end{aligned} \tag{1}$$

where $W_l \in \Re^{L_e \times L_f}$ is a transformation matrix. H is a $L_e \times 3$ matrix where each column corresponds to that of $[x_s, x_o, x_u]$.

Then we calculate the weights of $\{x_s, x_o, x_u\}$ and obtain their weighted sum through:

$$\begin{aligned} [p_s, p_o, p_u]^T &= softmax(W_{lx}H) \\ \hat{x} &= p_s x_s + p_o x_o + p_u x_u \end{aligned} \tag{2}$$

where W_{lx} in is a matrix with shape $1 \times L_e$. Note that $[p_s, p_o, p_u]$ are all scalars and the sum of them is equal to 1.

With the weighted sum $\hat{x}$, we infer the relationship through a fully connected layer as

$$H^l = tanh(W_x \hat{x}) \tag{3a}$$

$$p = softmax(W_h H^l) \tag{3b}$$

where p represents probability distribution over relationship candidates. We assign all the invalid relationships a *unk* token.

3.3 Caption Generation

Similar to [11], our model also generates a paragraph consisting of several sentences. In [11], the length of paragraph and its sentence is decided by two recurrent networks: a sentence RNN and a word RNN. Sentence RNN is responsible to calculate the number of sentences that should be in the generated paragraph and output topic vectors for them. Then the word RNN takes each topic vector as input and generates words of the corresponding sentence. Different from this strategy, we use one-layer Long Short Term Memory(LSTM)

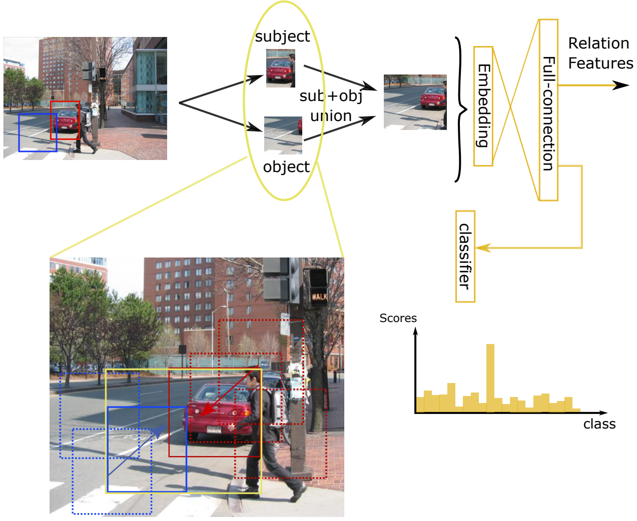


Figure 3: Relationship prediction module. The subject box(red car) is denoted by red squares and the object is indicated by a blue square. The dash squares represent boxes achieving lower scores than the solid ones.

network for sentence generation. The advantage of such architecture is that the training complexity can be reduced as training parameters of multi-layer RNN is challenging. The number of sentences is controlled by results of relationship prediction. Only the valid relationship shall generate sentence. We represent a valid predicted relationship by a one-hot vector y and embed it by a linear mapping:

$$\hat{y} = W_{embed} y$$

With the utilization of embedding, similar relationships are closer in the embedding space. For each valid relation, we feed visual and relationship features $\hat{x}, \hat{y}$ into LSTM units for word generation. Before we do this, it is necessary to fuse these two features into h_{xy} through following process:

$$h_{xy} = tanh(w_{xy}(h_x + h_y))$$

where $h_x = tanh(w_{hx}\hat{x})$ and $h_y = tanh(w_{hy}\hat{y})$.

As all the combinations of subjects and objects are considered, our paragraphs contain an average of 72.4 words and 5.9 sentences, compared with the results of 67.5 and 5.7 in previous work [11]. Following the formulation of [11], we set the hidden size of LSTM $H = 512$. The first and second inputs to the LSTM are h_{xy} and a special START token. At each time step, the hidden unit will predict a probability distribution over words in dictionary. Each sentence prediction is finished when the hidden unit outputs a special END token.

4 EXPERIMENTAL RESULTS

To demonstrate our model's advantage over [11], our results are evaluated by exploring the connection between visual relation and caption generation. Specifically, we first report the precision of relation prediction and then demonstrate our contribution by comparing paragraphs generated from our model and previous works. All the experiments are conducted on images from Visual Genome(VG) and MS-COCO.

4.1 Data set

Data sets containing images and their semantic information are necessary for works on image caption models. Many data sets, including Flickr30k, COCO and Visual Genome have been recently used for these models' training and testing. In order to demonstrate our model's effectiveness, we use the same dataset used in [11], which contains 19,551 images selected from Visual Genome and MS-COCO. Each image has been annotated with a paragraph description. Compared with combined sentence-level descriptions, these paragraph descriptions present even richer information as they removed redundancy tokens. We divide the data set into 14,575 training, 2,487 validation, and 2,489 testing images, following the same setup in [11]. Each image includes an average of 35 objects and 21 relationships.

4.2 Training Strategy

From a holistic point of view, our designed model consists of three sub-networks: object detection, relation detection and word generation module. Thereby, we train the entire model through three phases. At the first stage, we initialize the region detection network by copying parameters from the model in [9] and train it to be a network for object localization. We do not choose Faster RCNN[18], which is often adopted in relationship detection works, because dense caption model tends to cover more noteworthy visual elements. We want our model to detect objects that contain more relationship information for paragraph generation. The object locations in VG data set are set to be ground truth for this phase. Relation prediction module is composed of box discrimination and visual detection. These two sub-networks are trained alternately. Finally, the parameters of word generation are initialized from RNN units in [9] and fine-tuned at last. We map each sentence in paragraph to a relationship pair by matching words with <*subject,relation,object*> format. For this reason, only images in VG can be used at training stage.

4.3 Baselines

Dense caption. Dense caption model proposed in [9] outputs sentences based on regions detected by Region proposal networks. We demonstrate that using multiple regions outperforms using single region for sentence generation.

Hierarchical RNN. As the first work especially on dealing with paragraph generation problem, [11] put region features and their relations in a pooled vector and outputs sentences based on it. We compare the results with [11] as we need to demonstrate that generated paragraph would contain richer information if relation features are considered explicitly. Furthermore, our model's output presents more correlation of different visual elements.

Other Baselines. In order to reveal the connection between relation prediction and sentence generation, we design another baseline by replacing detected regions of our method with ground truth regions. Besides, we also report results when the relation prediction module performs poor.

4.4 Visual Relation Results

Before we validate the performance of final paragraph, we first evaluate model's ability of detecting visual relations and then explore to what extent visual detection results affect paragraph generation. Works on relation detection usually evaluate their models in two folds: the Recall rate of Subject and Object boxes(RoSO) and Recall rate of relationship detection. The RoSO of each image is defined as:

$$RoSO = \frac{\sum_{s \in \Omega_{sbj}, o \in \Omega_{obj}} I(s,o)}{K_{sbj} + K_{obj}} \tag{4a}$$

$$I(s,o) = \begin{cases} 1 & both\ IoU(s) > t\, and\, IoU(o) > t \\ 0 & otherwise \end{cases} \tag{4b}$$

Any <subject, object> box pair would be recognized as positive sample if both detected subject and object overlap with ground-truth are over threshold t. Suppose the total number of test images is $T(T = 2489$ in our case), the average RoSO($aRoSO$) is then supposed to be:

$$aRoSo = \frac{\sum_{i=1}^{T} RoSO_i}{T}$$

The $aRoSO$ is used to validate whether regions generated by Region Proposal Network is well classified into Ω_{sbj} and Ω_{obj}. As for evaluating relation prediction result, we calculate the recall rate of positive relation pairs. A positive pair sample needs to meet two conditions: (1)The subject and object boxes are correctly detected. (2) The correct relationship is predicted. Given the detected K_{sbj} subjects and K_{obj} objects for each image, we obtain $K_{sbj} \times K_{obj}$ pairs for relationship prediction. We select the top 499 items from the most frequent used relationships in data set as candidates. We also add a special candidate Unk to represent the case when some pairs are not related to each other at all.

For performance comparison, we select models focusing on relationship detection as baselines. Among them, Rel-PN[24] and DR-net[2] involve detecting subject and object boxes besides predicting relationships. We adopt the same setup and abbreviations for convenience. We show results with different IoUs and $K_{sbj} \times K_{obj}$ in Table 1. $IoU > t$ means both subject and object boxes overlap with ground-truth by at least t. $K_{sbj} \times K_{obj}$ means that we select N relationship pair proposals and $K_{sbj} = K_{obj} = \sqrt{N}$. We set the output number of object proposals as 100. From the results, it is obvious that the recall rates decrease when we set IoU a large value. However, if we enlarge the size of $K_{sbj} \times K_{obj}$, the recall rate can be improved even at high IoU. Table 1 presents detection results of subject and object boxes.

$IoU > 0.5$	$K_{sbj} \times K_{obj}$			
	32×32	45×45	54×54	71×71
Rel-PN	21.53	25.60	27.66	32.30
DR-net	25.38	27.13	28.94	30.88
Our method	26.33	28.45	29.14	32.53
$IoU > 0.6$	$K_{sbj} \times K_{obj}$			
	32×32	45×45	54×54	71×71
Rel-PN	17.44	20.71	21.93	23.46
DR-net	19.28	20.94	21.81	22.53
Our method	19.70	21.34	22.27	23.94
$IoU > 0.7$	$K_{sbj} \times K_{obj}$			
	32×32	45×45	54×54	71×71
Rel-PN	6.37	8.11	9.07	10.79
DR-net	6.98	8.05	8.87	9.98
Our method	7.34	9.04	10.23	11.64

Table 1: Average Recall rate of relation pair detection.

$IoU > 0.5$	Recall@50	Recall@100
Rel-PN	23.72	27.12
DR-net	22.87	25.97
Our method	24.93	28.25

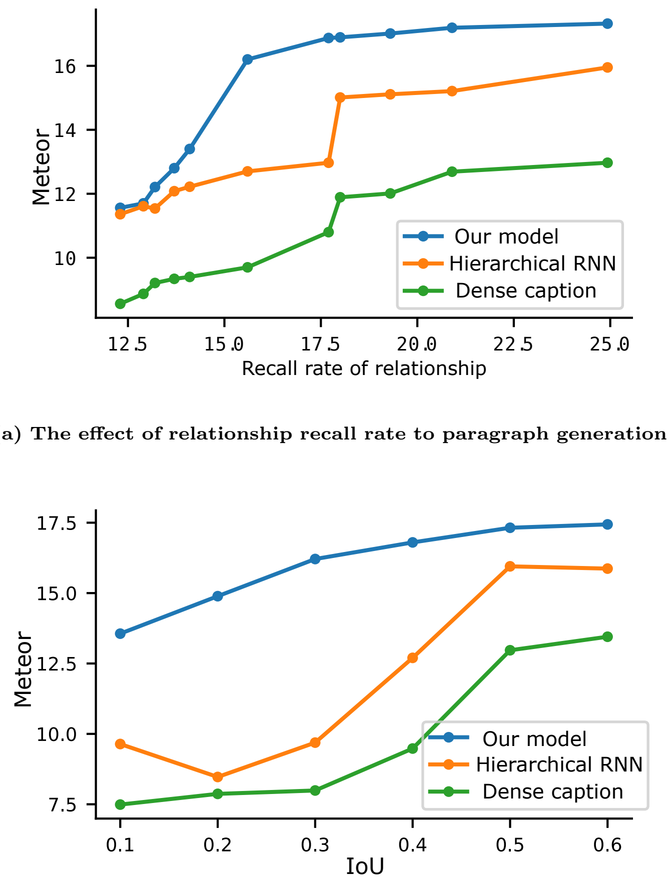
Table 2: The Recall rates of relationship prediction.

Rel-PN generates subject and object boxes with a proposal network and DR-net achieves this purpose by assigning labels to detected objects output by Faster RCNN[18]. Compared with these two methods, Our proposal network is trained from [9] which aims to detect regions containing more visual elements. For this reason, the detection results of our model tend to cover more visual relationship parts. We can see that with relationship considered, the accuracy of relation pair detection is improved as Rel-PN and our method performs better.

We also report the recall rate of relationship prediction in Table2. We set the IoU to be 0.5 and select 2000 proposals. Recall@K means that we select the top K relationship candidates for evaluation. The results of Table 2 demonstrate that our proposal network outperforms previous works on relationship detection, even though the box detection results of these three methods are approximate.

4.5 Paragraph generation Results

We present our results of paragraph generation in Table 3. We adopt six language metrics: CIDEr[20], METEOR[3], and BLEU-1,2,3,4[16] to evaluate our model, following the metric setup in [11]. For all baseline methods, the regions are from three sources: Ground truth regions(GT), generated by RPN of [9](RPN), Union of subject and object(S+O). For a fair comparison, they are all trained with ground truth regions and their descriptions. From the results of Dense caption and Hierarchical RNN, using S+O regions is better



(a) The effect of relationship recall rate to paragraph generation

(b) The effect of IoU to paragraph generation

Figure 4: The effect of relationship prediction to paragraph generation

for word generation than using RPN regions. This phenomenon demonstrates that $S + O$ regions contain more useful visual information than those generated from RPN. Both Dense caption method and Hierarchical RNN take all regions as input. Thus the paragraph contains large amounts of visual information. Therefore, they achieved competitive performance of our proposed method at these six metrics. We also present the effect of relationship detection accuracy to the final sentences. We extract corresponding features of $S + O$ from images and take them as input of dense caption and Hierarchical RNN. These results present that our model overcomes Hierarchical RNN and Dense caption under the same setup. The performance of our method(Gt) demonstrates that our model is robust to region localization error. Even some relationship regions are not localized well, our method still achieves promising results. Fig.4 illustrates the relations between relationship detection and word generation. In Fig.4(a), two baseline methods are more sensitive to accuracy of relationship detection. Fig.4(b) demonstrate that our method still performs well even at the low IoU thresholds. These results further demonstrate the robustness to region detection fluctuation.

4.6 Qualitative Results

In order to present the advantage of our design, we show the qualitative results in Fig.5. Compared with dense captioning and Hierarchical RNN, our method performs more similar to human beings. Dense caption method performs poorly because it simply describes regions' contents by a single sentence and these sentences can not be regarded as a 'paragraph'. Obviously, both two baseline works generate much redundant information. For example, one object appears in many sentences and some sentences actually express the same meaning. Another improvement is that our model's results present relationships among objects better. Some details that are not included in Ground truth are also detected. For example, our model detect the shadow and number of legs while the other two methods only detect the salient objects.

5 CONCLUSION

We proposed a paragraph generation network which considers the visual relation among objects. By explicitly predicting visual relationship of the detected objects, we obtain paragraphs containing richer and more accurate information than previous works. Furthermore, our proposed model is more robust to object detection errors as we use both visual and relation features to generate sentences.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous referees for their valuable comments and helpful suggestions. This work was supported in part by the National Science Foundation of China (NSFC) under grants 61472101 and 61631017, and the Major State Basic Research Development Program of China (973 Program 2015CB351804).

REFERENCES

[1] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6298–6306.

[2] Bo Dai, Yuqi Zhang, and Dahua Lin. 2017. Detecting Visual Relationships With Deep Relational Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*. 3298–3308.

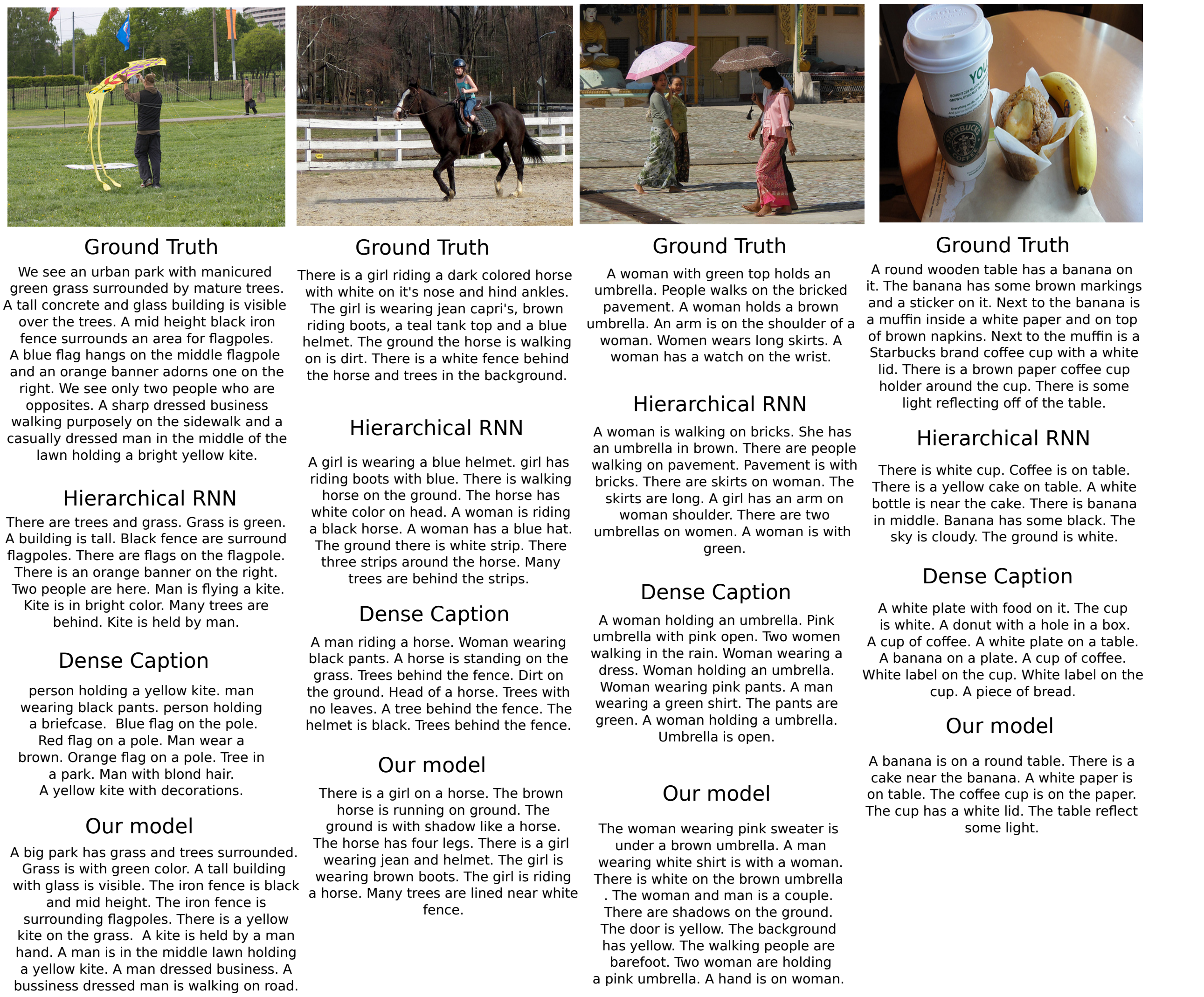
[3] Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

[4] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Computer Vision and Pattern Recognition*. 677–691.

[5] Andrea Frome, Gregory S Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marcaurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. *neural information processing systems* (2013), 2121–2129.

	METEOR	CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Dense caption(GT)	13.58	11.88	34.97	20.74	13.05	8.23
Dense caption(RPN)	12.82	11.06	34.04	19.95	12.20	7.71
Dense caption(S+O)	12.97	11.25	34.27	20.18	12.42	7.79
Hierarchical RNN(GT)	16.73	14.48	40.78	25.19	15.05	9.26
Hierarchical RNN(RPN)	15.95	13.52	41.90	24.11	14.23	8.69
Hierarchical RNN(S+O)	16.01	13.57	41.86	24.86	14.86	9.22
Our method(GT)	17.58	14.98	42.16	25.06	15.12	9.11
Our method	17.32	14.55	41.74	24.94	14.94	9.34
Humans	19.22	28.55	42.88	25.68	15.55	9.66

Table 3: Results of paragraph generation. We alternate the source of regions for each baseline method. The last row is human performance.



Ground Truth

We see an urban park with manicured green grass surrounded by mature trees. A tall concrete and glass building is visible over the trees. A mid height black iron fence surrounds an area for flagpoles. A blue flag hangs on the middle flagpole and an orange banner adorns one on the right. We see only two people who are opposites. A sharp dressed business walking purposely on the sidewalk and a casually dressed man in the middle of the lawn holding a bright yellow kite.

Hierarchical RNN

There are trees and grass. Grass is green. A building is tall. Black fence are surround flagpoles. There are flags on the flagpole. There is an orange banner on the right. Two people are here. Man is flying a kite. Kite is in bright color. Many trees are behind. Kite is held by man.

Dense Caption

person holding a yellow kite. man wearing black pants. person holding a briefcase. Blue flag on the pole. Red flag on a pole. Man wear a brown. Orange flag on a pole. Tree in a park. Man with blond hair. A yellow kite with decorations.

Our model

A big park has grass and trees surrounded. Grass is with green color. A tall building with glass is visible. The iron fence is black and mid height. The iron fence is surrounding flagpoles. There is a yellow kite on the grass. A kite is held by a man hand. A man is in the middle lawn holding a yellow kite. A man dressed business. A bussiness dressed man is walking on road.

Ground Truth

There is a girl riding a dark colored horse with white on it's nose and hind ankles. The girl is wearing jean capri's, brown riding boots, a teal tank top and a blue helmet. The ground the horse is walking on is dirt. There is a white fence behind the horse and trees in the background.

Hierarchical RNN

A girl is wearing a blue helmet. girl has riding boots with blue. There is walking horse on the ground. The horse has white color on head. A woman is riding a black horse. A woman has a blue hat. The ground there is white strip. There three strips around the horse. Many trees are behind the strips.

Dense Caption

A man riding a horse. Woman wearing black pants. A horse is standing on the grass. Trees behind the fence. Dirt on the ground. Head of a horse. Trees with no leaves. A tree behind the fence. The helmet is black. Trees behind the fence.

Our model

There is a girl on a horse. The brown horse is running on ground. The ground is with shadow like a horse. The horse has four legs. There is a girl wearing jean and helmet. The girl is wearing brown boots. The girl is riding a horse. Many trees are lined near white fence.

Ground Truth

A woman with green top holds an umbrella. People walks on the bricked pavement. A woman holds a brown umbrella. An arm is on the shoulder of a woman. Women wears long skirts. A woman has a watch on the wrist.

Hierarchical RNN

A woman is walking on bricks. She has an umbrella in brown. There are people walking on pavement. Pavement is with bricks. There are skirts on woman. The skirts are long. A girl has an arm on woman shoulder. There are two umbrellas on women. A woman is with green.

Dense Caption

A woman holding an umbrella. Pink umbrella with pink open. Two women walking in the rain. Woman wearing a dress. Woman holding an umbrella. Woman wearing pink pants. A man wearing a green shirt. The pants are green. A woman holding a umbrella. Umbrella is open.

Our model

The woman wearing pink sweater is under a brown umbrella. A man wearing white shirt is with a woman. There is white on the brown umbrella . The woman and man is a couple. There are shadows on the ground. The door is yellow. The background has yellow. The walking people are barefoot. Two woman are holding a pink umbrella. A hand is on woman.

Ground Truth

A round wooden table has a banana on it. The banana has some brown markings and a sticker on it. Next to the banana is a muffin inside a white paper and on top of brown napkins. Next to the muffin is a Starbucks brand coffee cup with a white lid. There is a brown paper coffee cup holder around the cup. There is some light reflecting off of the table.

Hierarchical RNN

There is white cup. Coffee is on table. There is a yellow cake on table. A white bottle is near the cake. There is banana in middle. Banana has some black. The sky is cloudy. The ground is white.

Dense Caption

A white plate with food on it. The cup is white. A donut with a hole in a box. A cup of coffee. A white plate on a table. A banana on a plate. A cup of coffee. White label on the cup. White label on the cup. A piece of bread.

Our model

A banana is on a round table. There is a cake near the banana. A white paper is on table. The coffee cup is on the paper. The cup has a white lid. The table reflect some light.

Figure 5: The qualitative results of our model. We mainly compare our model with two baseline methods: Hierarchical RNN[11] and dense caption[9].

[6] Stephen Gould, Jim Rodgers, David Cohen, Gal Elidan, and Daphne Koller. 2008. Multi-Class Segmentation with Relative Location Prior. *International Journal of Computer Vision* 80, 3 (2008), 300–316.

[7] Abhinav Gupta, Aniruddha Kembhavi, and Larry S. Davis. 2009. Observing Human-Object Interactions: Using Spatial and Functional Compatibility for Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 10 (2009), 1775–1789.

[8] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: data, models and evaluation metrics. *Journal of Artificial Intelligence Research* 47, 1 (2013), 853–899.

[9] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. DenseCap: Fully Convolutional Localization Networks for Dense Captioning. In *Computer Vision and Pattern Recognition*. 4565–4574.

[10] Andrej Karpathy and Fei Fei Li. 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 4 (2015), 664.

[11] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A Hierarchical Approach for Generating Descriptive Image Paragraphs. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3337–3345.

[12] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. https://arxiv.org/abs/1602.07332

[13] Cewu Lu, Ranjay Krishna, Michael S Bernstein, and Li Feifei. 2016. Visual Relationship Detection with Language Priors. *European Conference on Computer Vision* (2016), 852–869.

[14] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[15] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. 2014. Explain Images with Multimodal Recurrent Neural Networks. *Computer Science* (2014).

[16] Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. *Association for Computational Linguistics* (2002), 311–318.

[17] Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. 2017. Areas of Attention for Image Captioning. In *The IEEE International Conference on Computer Vision (ICCV)*. 1251–1259.

[18] Shaoqing Ren, Kaiming He, Ross B Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 6 (2017), 1137–1149.

[19] Lin Tsung-Yi, Maire Michael, Belongie Serge, Bourdev Lubomir, Girshick Ross, Hays James, Perona Pietro, Ramanan Deva, Zitnick C. Lawrence, and Dollar Piotr. 2015. Microsoft COCO: Common Objects in Context. https://arxiv.org/pdf/1405.0312.pdf

[20] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. *computer vision and pattern recognition* (2015), 4566–4575.

[21] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition*. 3156–3164.

[22] Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li. 2017. Dense Captioning With Joint Inference and Visual Context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1978–1987.

[23] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. 2017. Visual Translation Embedding Network for Visual Relation Detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3107–3115.

[24] Ji Zhang, Mohamed Elhoseiny, Scott Cohen, Walter Chang, and Ahmed Elgammal. 2017. Relationship Proposal Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5226–5234.