Object Relation Attention for Image Paragraph Captioning

Li-Chuan Yang,[1] Chih-Yuan Yang,[1,2] and Jane Yung-jen Hsu[1,2]
[1]Computer Science and Information Engineering, National Taiwan University
[2]NTU IoX Center, National Taiwan University
{r07922100,yangchihyuan,yjhsu}@ntu.edu.tw

Abstract

Image paragraph captioning aims to automatically generate a paragraph from a given image. It is an extension of image captioning in terms of generating multiple sentences instead of a single one, and it is more challenging because paragraphs are longer, more informative, and more linguistically complicated. Because a paragraph consists of several sentences, an effective image paragraph captioning method should generate consistent sentences rather than contradictory ones. It is still an open question how to achieve this goal, and for it we propose a method to incorporate objects' spatial coherence into a language-generating model. For every two overlapping objects, the proposed method concatenates their raw visual features to create two directional pair features and learns weights optimizing those pair features as relation-aware object features for a language-generating model. Experimental results show that the proposed network extracts effective object features for image paragraph captioning and achieves promising performance against existing methods.

Introduction

With computer vision and natural language processing advancements, describing visual content using natural languages has achieved impressive results in recent years (Chen and Zitnick 2014; Donahue et al. 2015; Karpathy and Fei-Fei 2015; Mao et al. 2014; Vinyals et al. 2015; Xu et al. 2015; Luo et al. 2018; Anderson et al. 2018). Among studies integrating computer vision and natural language processing, image captioning aims to generate a single sentence from a given image. Due to the limited length of a single sentence expressed in a natural way, it is hard to describe rich details depicted by an image. To address this problem, Krause *et al* (Krause et al. 2017) compile a dataset containing thousands of images and their corresponding paragraphs in 5 to 8 descriptive sentences. This is the only dataset designed for image paragraph captioning to the best of our knowledge, which aims to generate multiple sentences as an integrated description of an image.

Since image paragraph captioning is extended from image captioning, a straight-forward extension for image paragraph captioning is to generate sentences after sentences to form a paragraph. However, image captioning approaches often generate similar sentences, which are highly different from human-written diverse sentences in a paragraph. To address this problem, Krause *et al*. propose a hierarchical recurrent network (Krause et al. 2017). Liang *et al*. introduce adversarial training for paragraph generation (Liang et al. 2017). Chatterjee *et al*. propose coherence vectors and global topic vectors to augment the hierarchical structure (Chatterjee and Schwing 2018). Melas-Kyriazi *et al* (Melas-Kyriazi, Rush, and Han 2018). propose repetition penalty and apply it to two existing image captioning methods, bottom-up and top-down attention (Anderson et al. 2018) and SCST (Self-Critical Sequence Training) (Rennie et al. 2017), and achieve state-of-the-art performance.

However, since image paragraph captioning is a new research problem, there are many unsolved challenges such as generating coherent and diverse sentences with details. All existing methods use Faster R-CNN (Ren et al. 2015) to extract individual object features and use the remaining language module to learn the relations among those objects from the training paragraphs. It is questionable how effective this approach will be because most training sentences consist of object and relation terms. And generating paragraphs requires more detailed relation information. Thus without strong relation information at the object feature level, the language module may learn to generate sentences incorrectly describing the input image as shown in Figure 5.

Based on this observation, we propose a set of two networks to encode relations into object features. One network learns to calculate the importance of object pairs with respect to the corresponding words occurring in training sentences. The other network learns to modify raw object features to encode the learned pair importance. We design a model to train the two networks together. We use spatial coherence as our object relations because this information is easy to obtain by simply checking whether objects' bounding boxes overlap. Existing object detector algorithms have shown promising performance in finding objects in images. We do not use object relations in other types such as object categories or semantic connections because we aim to evaluate how well our model can encode relation features by giving pairs of subjects in terms of their visual features. To encode relation features, we are inspired by the concept of graph convolution networks (Bruna et al. 2013; Niepert, Ahmed, and Kutzkov 2016) which generate a node's con-

volution output by calculating all of its connected nodes. In our case, a node equals a detected object in the given image and an edge between equals the condition that two objects' bounding boxes overlap. The proposed relation network can be viewed as an instantiation of a one-layer graph convolution network which can be trained in an end-to-end manner, encode and attend object relations automatically without manually labeled relation data. By applying the proposed network to a state-of-the-art method (Melas-Kyriazi, Rush, and Han 2018), we generate better qualitative and quantitative image captioning paragraphs.

Related Work

Image Captioning. Image captioning is a topic across computer vision and natural language processing, requires both image comprehension and text generation. Early developed approaches use template-based methods to generate sentences by filling object information into templates (Yang et al. 2011; Yao et al. 2010; Kulkarni et al. 2013). Several recently developed methods working in deep encoder-decoder structures use CNN (Convolutional Neural Networks) and RNN (Recurrent Neural Networks) as their encoders and decoders (Cho et al. 2014; Vinyals et al. 2015; Donahue et al. 2015). In order to locate highlight objects, several methods are proposed to use attention mechanisms to extract latent region features for text generation and generate improved caption sentences (Xu et al. 2015; Lu et al. 2017; Anderson et al. 2018; Huang et al. 2019). A reinforcement-learning-based algorithm is proposed to use language metrics instead of cross entropy to train a language generation model. Although language metrics are non-differentiable, their model is still trainable because they treat metric indices as a part of their reward (Rennie et al. 2017). Models utilizing high-level visual knowledge or scene understanding (Yao et al. 2017; You et al. 2016; Wu et al. 2018; Yang et al. 2019; Yao et al. 2018) are also proposed to improve captioning performance.

Image Paragraph Captioning. Image paragraph captioning is an extension of image captioning. A paragraph consists of several sentences and delivers more information than a single caption sentence. To address this problem how to evaluate automatically generated paragraphs describing images, a dataset is proposed along with a method (Krause et al. 2017) using a hierarchical structure to extract topic vectors by LSTM and to decode every sentence according to its topic vector. GAN and conditional GAN are applied to generate realistic paragraphs (Liang et al. 2017; Dai et al. 2017). To improve the sentence coherency in generated paragraphs, global vectors are proposed to bring information between topics (Chatterjee and Schwing 2018). To avoid repeated sentences and phrases which reduce sentence diversity, Melas-Kyriazi *et al.* propose a method to decrease probability of word occurrence if some words form repeated trigrams (Melas-Kyriazi, Rush, and Han 2018). To improve coherence, convolutional auto-encoder is applied on hierarchical structure to learn coherent topics (Wang et al. 2019).

Neural Networks on Relationship. There are two types of existing research topics related to our method in terms of relationship calculation: visual relation detection (VRD) and graph convolution networks (GCN). While VRD aims to model relationships between objects on an image, GCN deals with the relationship among nodes in a graph. In our case, GCN is the more major component.

Regarding GCN, an early method dealing with GCN uses an RNN structure to handle arbitrary graphs (Scarselli et al. 2009). Many following papers report the application of GCN for their problems (Henaff, Bruna, and LeCun 2015; Bruna et al. 2013; Duvenaud et al. 2015). The one highly related to our work is (Johnson, Gupta, and Fei-Fei 2018), which applies graph convolution on the relations between objects in scene graphs to extract information for image generation.

Approach

We observe that on many images, objects close to each other have certain relations and are often described in image captions. We exploit this property to develop a network for automatic image paragraph captioning. Fig. 1 shows the proposed architecture, and we explain its components in the following subsections.

Relations between Objects. Given an image I, we use an existing object detection algorithm Faster R-CNN to find objects $O = \{o_1, \ldots, o_k\}$ on I. Let $V = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}, \boldsymbol{v}_i \in \mathbb{R}^{2048}$ be their visual feature vectors, and $B = \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k\}, \boldsymbol{b}_i \in \mathbb{R}^4$ be their bounding boxes, where k is variable depending on the content of I. We exploit the relations among those objects in a primitive manner by only considering object pairs

$$p_{i,j} = \{o_i, o_j\}. \tag{1}$$

Because we observe that many captioning sentences describe major objects and their surrounding minor objects, we propose to reduce the pair set and take spatial coherence into account by retaining $p_{i,j}$ only if $\boldsymbol{b}_i$ overlaps with $\boldsymbol{b}_j$. For an object pair, we create two pair feature vectors by concatenating their visual features in two directions $\boldsymbol{w}_{i,j} = \boldsymbol{v}_i \oplus \boldsymbol{v}_j$ and $\boldsymbol{w}_{j,i} = \boldsymbol{v}_j \oplus \boldsymbol{v}_i$. This asymmetric concatenation design is motivated by our observation that many descriptive sentences in the Stanford dataset are formed as two visual objects with a verb if adjectives, adverbs, and tenses are ignored. The order of the two objects and the verb is asymmetric, for example, a man wears a shirt rather than a shirt wears a man.

Relation Encoding Network. To learn the relations among two objects and a verb in the visual feature stage, we propose a relation encoding network, which encodes relations in pair features $\{\boldsymbol{w}\}$ to generate new feature vectors $\{\boldsymbol{w}'\}$. We use bottleneck-layer architecture for graph convolution, proved to be efficient in the supplementary material, to merge feature vectors of two related objects into a single feature vector containing information of the two related objects. The network contains two one-dimension layers of sizes 2048 and 4096. The first layer has a ReLU and a Batch Normalization layer but the second has neither. We split its generated pair feature vectors into two equal-length intermediate object feature vectors $\boldsymbol{u}^i_{i,j}$ and $\boldsymbol{u}^j_{i,j}$ where $\boldsymbol{w}'_{i,j} = \boldsymbol{u}^i_{i,j} \oplus \boldsymbol{u}^j_{i,j}$.

Relation Attention Network. We use another fully-connected neural network to learn attention weights of object pairs in training paragraph captions. The network contains two one-dimension layers of size 1024 and 1. Same as

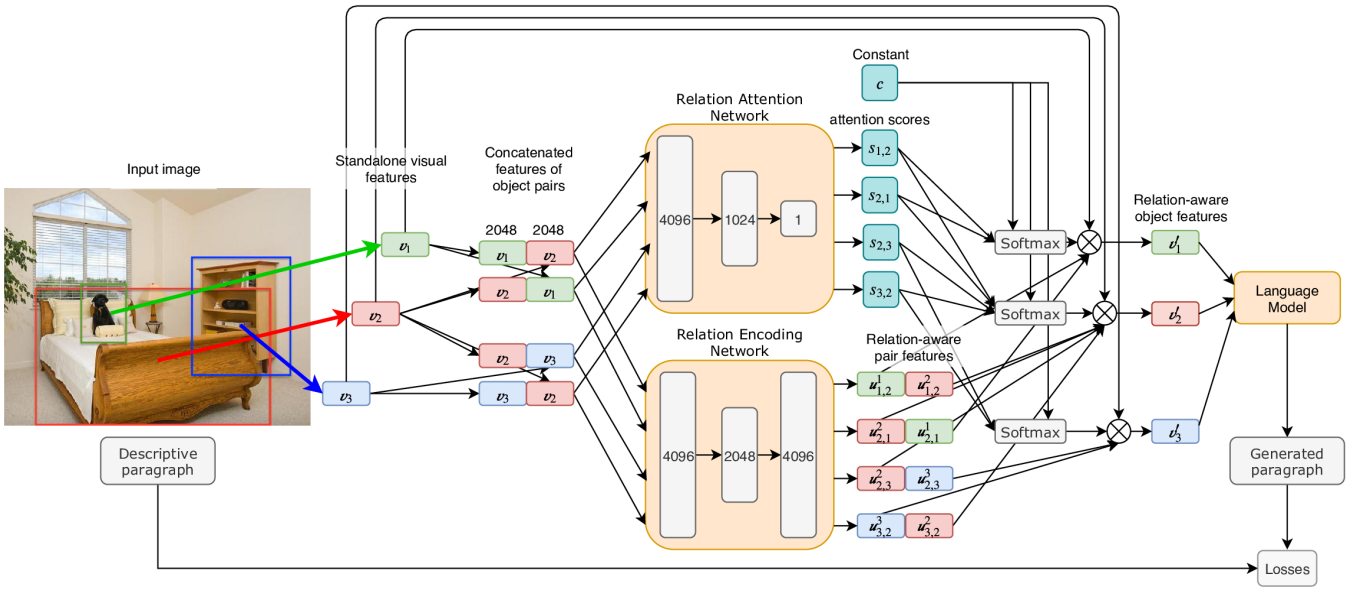


Figure 1: Flow chart of the proposed method. Given features from Faster R-CNN, we attend and encode relations into detected features asymmetrically to make language model easier to generate coherent paragraphs.

Relation Encoding Network, the first layer has a ReLU and a Batch Normalization layer but the second has neither. Denote f_A as a function of Relation Attention Network, and attention scores are generated by

$$s_{i,j} = f_A(\boldsymbol{w}_{i,j}) \tag{2}$$

We use $\{s_{i,j}\}$ to indicate the attention scores of object pairs mentioned in paragraphs describing images.

Feature Fusing. In our model, an image object o_i will generate multiple feature vectors including the raw one $\boldsymbol{v}_i$, and the relation-aware intermediate ones $\boldsymbol{u}^i_{i,x}$ and $\boldsymbol{u}^i_{x,i}$ where x is any number between 1 and k, $x \neq i$. We fuse those feature vectors by weight averaging

$$\boldsymbol{v}'_i = \alpha_0 \boldsymbol{v}_i + \sum_x \alpha_{i,x} \boldsymbol{u}^i_{i,x} + \sum_x \alpha_{x,i} \boldsymbol{u}^i_{x,i} \tag{3}$$

where weights α_0, $\alpha_{i,x}$ and $\alpha_{x,i}$ are calculated through a softmax function

$$\alpha_0, \{\alpha_{i,x}, \alpha_{x,i}\} = \text{softmax}(c, \{s_{i,x}, s_{x,i}\}) \text{ for } x \neq i. \tag{4}$$

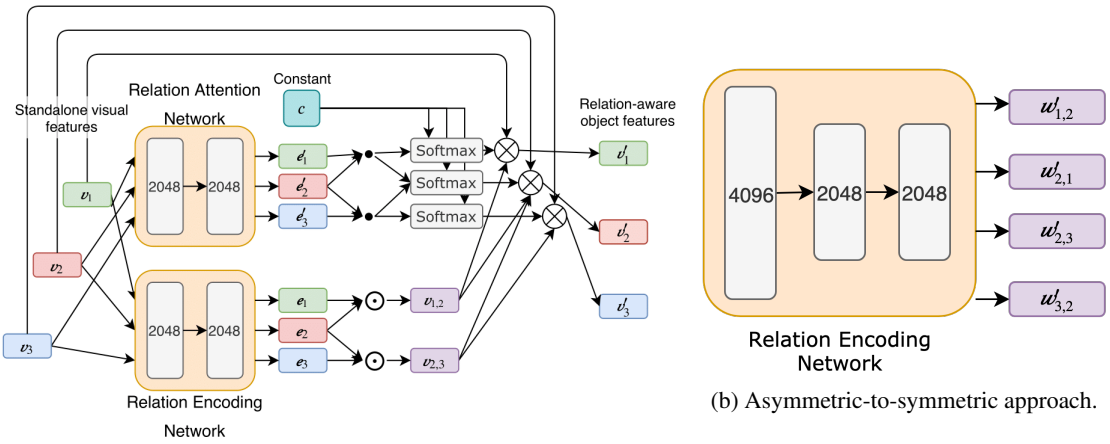
The constant c is a hyperparameter to balance the importance of raw object features $\boldsymbol{v}_i$ among related-aware ones and we tune the value empirically. We use the new object features $V' = \{\boldsymbol{v}'_1, \ldots, \boldsymbol{v}'_k\}, \boldsymbol{v}'_i \in \mathbb{R}^{2048}$ to train our paragraph generating module.

Simplified Symmetric and Asymmetric-to-Symmetric Approaches. To validate the effectiveness of the proposed asymmetric concatenation and its following steps of relation networks and feature fusing, we create two simplified networks as shown in Fig. 2 and report their numerical evaluation in the experiment section. In the first symmetric approach, we generate attention score vectors $\{\boldsymbol{e}'_i\}$ through a simplified relation attention network. We generate relation-aware intermediate feature vectors $\{\boldsymbol{e}_i\}$ through a simplified relation encoding network. We apply inner product on $\{\boldsymbol{e}'_i\}$ for overlapping objects to generate attention scores, which are used in softmax functions to compute weights. We create symmetric relation-encoded feature vectors $\{\boldsymbol{v}_{i,j}\}$ by applying element-wise product to $\{\boldsymbol{e}_i\}$ for overlapping objects. We generate relation-aware objects features $\{\boldsymbol{v}'_i\}$ by weight averaging $\{\boldsymbol{v}_{i,j}\}$. For the second asymmetric-to-symmetric (A2S) approach, we modify the relation encoding network by reducing the size of the second layer from 4096 to 2048. Thus, the output feature vectors $\{\boldsymbol{w}'_{i,j}\}$ no longer need to be split for feature fusing. We fuse those output feature vectors in the same way of Eq. 3 by

$$\boldsymbol{v}'_i = \alpha_0 \boldsymbol{v}_i + \sum_x \alpha_{i,x} \boldsymbol{w}'_{i,x} + \alpha_{x,i} \boldsymbol{w}'_{x,i} \tag{5}$$

Self-Critical Sequence Training. Cross entropy is widely used to measure language models at a word-based level. However, a trained language model which generates low cross-entropy loss may not generate natural and fluent sentences. To address this problem, Rennie *et al.* propose SCST (Self-Critical Sequence Training) (Rennie et al. 2017), a sequence-based training procedure using a policy gradient approach to train language models to generate high indices under non-differential captioning metrics. SCST puts language models in a reinforcement-learning framework and sets rewards as metric indices. To train the proposed method, we also adopt SCST under two different rewards. The first is pure CIDEr, for making a fair comparison with an existing method (Melas-Kyriazi, Rush, and Han 2018). The second is mixed CIDEr, METEOR, and BLEU-4, and we find it generates well-balanced indices.

Paragraph Generating. Our paragraph generating module is composed of a Top-Down Attention LSTM module and



(a) Symmetric approach.

(b) Asymmetric-to-symmetric approach.

Figure 2: Two various approaches to learn relations. Both sub-figures show the modified parts of the proposed relation network of Fig. 1. The unchanged parts such as the input image and language components are omitted for clarity.

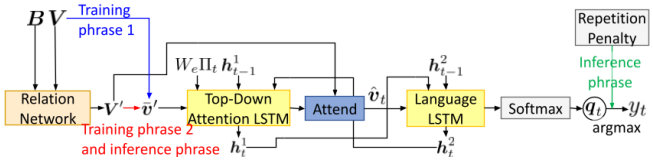


Figure 3: Paragraph generating module. Two training phases are in color blue and red.

a Language LSTM module, as illustrated in Fig. 3. The architecture is first used in (Anderson et al. 2018), and our changes are the input feature modified by the proposed relation network and the output probabilities modified by repetition penalty (Melas-Kyriazi, Rush, and Han 2018). For every timestamp t, a triplet—composed of an object feature vector $\bar{v}$, previous word embedding feature $W_e \Pi_t$, and hidden state h^2_{t-1} generated from the Language LSTM—is fed into the Top-Down Attention LSTM to get an attended feature vector $\hat{v}_t$. Then the Language LSTM outputs a word y_t based on the attended feature vector $\hat{v}_t$ and a hidden state h^1_t from the Top-Down Attention LSTM. The Language LSTM stops when it outputs an EOS (End of Sentence) signal.

We train the paragraph generating module in two stages. During the first 30 epochs, we use cross-entropy to calculate the loss against training paragraphs at a word-based level and use the mean of V to train our model to make training robust. Thereafter, we apply SCST to train our model and use the relation-encoded feature set V' instead of V.

Repetition Penalty. The bottom-up and top-down attention model trained under the SCST approach performs well for image captioning. However, this method keeps generating repeated sentences when it comes to paragraph captioning. Melas-Kyriazi *et al*. find that the problem is caused by the greedy search of the reinforcement-learning approach, which prefers to generate non-diverse sentences and results in poor policy gradient (Melas-Kyriazi, Rush, and Han 2018). To address this problem, they propose an inference-phrase probability-modifying approach, named repetition penalty, to adjust the selection of output words to reduce repeated tri-gram in output sentences. We also adopt repetition penalty in our method to suppress repeated sentences.

Experiments

Dataset and Metrics. We evaluate the proposed method on the Stanford paragraph dataset (Krause et al. 2017), which contains 19551 image/paragraph pairs, split into training/validation/test sets containing 14575/2487/2489 pairs, respectively. We use 6 metrics CIDEr (Vedantam, Zitnick, and Parikh 2015), METEOR (Banerjee and Lavie 2005), BLEU-1, BLEU-2, BLEU-3, and BLEU-4 (Papineni et al. 2002) as the literature (Krause et al. 2017; Liang et al. 2017; Chatterjee and Schwing 2018; Melas-Kyriazi, Rush, and Han 2018)

BLEU-n (Papineni et al. 2002) are invented to evaluate machine translation systems, and they calculate the accuracy of predicted sentences with n-grams. METEOR (Banerjee and Lavie 2005) is another metric designed for machine translation. It aligns words first using synonymy matching, calculates the precision and recall of unigrams, and sets a larger weight for recall over precision. CIDEr (Vedantam, Zitnick, and Parikh 2015) is designed for image captioning. It uses TF-IDF (Term Frequency Inverse Document Frequency) to re-weight different n-grams since key words like verbs and nouns carry more semantic information than prepositions.

Details of Model Training. We use a publicly available Faster R-CNN implementation (Anderson et al. 2018) pre-trained on the ImageNet (Deng et al. 2009) and Visual Genome datasets (Krishna et al. 2016).

To train our models, we use the Adam optimizer with a learning rate initialized as 5×10^{-4} and decaying 20% every two epochs. We manually set the attention hyperparameter c as 2 because we find the proposed method converges

Method	C	M	B-1	B-2	B-3	B-4
Krause *et al* (Krause et al. 2017) (Template)	12.15	14.31	37.47	21.02	12.30	7.38
Krause *et al* (Krause et al. 2017) (Flat)	11.14	13.54	37.30	21.70	13.07	8.07
Krause *et al* (Krause et al. 2017) (Hierarchical)	13.52	15.95	41.90	24.11	14.23	8.69
Liang *et al* (Liang et al. 2017)	16.87	17.12	41.99	24.86	14.89	9.03
Chatterjee *et al* (Chatterjee and Schwing 2018)	19.95	17.81	42.12	25.18	14.74	9.05
Chatterjee *et al* (Chatterjee and Schwing 2018) (GAN)	18.05	17.21	42.04	24.96	14.53	8.95
Chatterjee *et al* (Chatterjee and Schwing 2018) (VAE)	20.93	<u>18.62</u>	42.38	25.52	15.15	9.43
Transformer (Vaswani et al. 2017)	25.91	15.44	37.08	22.32	13.50	8.07
Melas-Kyriazi *et al* (Melas-Kyriazi, Rush, and Han 2018) (w/o r.p. C)	13.77	13.63	29.67	16.45	9.74	5.88
Melas-Kyriazi *et al* (Melas-Kyriazi, Rush, and Han 2018) (w/ r.p. C)	30.63	17.86	43.54	27.44	17.33	10.58
Melas-Kyriazi *et al* (Melas-Kyriazi, Rush, and Han 2018) (w/ r.p. C+M+B4)	30.37	17.80	43.69	27.55	17.41	10.66
Wang *et al* (Wang et al. 2019)	25.15	**18.82**	-	-	-	9.67
Relation All Pairs Asym. (C+M+B4)	30.88	17.23	42.87	27.24	17.42	10.68
Relation Overlapping Sym. (C)	32.98	17.67	43.22	27.64	17.56	10.79
Relation Overlapping Sym. (C+M+B4)	32.32	17.78	43.97	28.09	17.83	10.95
Relation Overlapping A2S. (C)	32.20	17.63	43.25	27.50	17.47	10.71
Relation Overlapping A2S. (C+M+B4)	32.70	17.64	<u>44.02</u>	<u>28.36</u>	<u>17.95</u>	<u>11.01</u>
Relation Overlapping Asym. (C)	**33.38**	17.82	43.76	28.08	17.88	10.95
Relation Overlapping Asym. (C+M+B4)	<u>33.12</u>	17.97	**44.55**	**28.54**	**18.19**	**11.18**

Table 1: Numerical evaluation. All methods are tested on the same dataset and split (Krause et al. 2017). Runner-up scores are underlined. Note that (Wang et al. 2019) does not report B-1, B-2, and B-3 scores.

well and performs stably when the value is between 1 to 3. We do experiments under two SCST configurations. One uses CIDEr only, for making a fair comparison with a state-of-the-art method. Another uses a mixed reward, calculated from scores of CIDEr, METEOR, BLEU-4 with weights 1, 0.5, and 0.5, respectively. As shown in Table 1, the first configuration achieves the highest CIDEr score 33.38, and the second one generates an overall improvement over the state-of-the-art method (Melas-Kyriazi, Rush, and Han 2018).

We train our model on a machine equipped with a 3.7GHz 12-core CPU and an NVidia GPU GTX 1080Ti. We set the training batch size as 10. The configuration of overlapping objects and asymmetric features consumes 2.3 GB GPU memory and takes 16 hours to run 80 epochs, including the first 30 cross-entropy epochs and the following 50 SCST epochs.

Numerical Evaluation. Table 1 shows the proposed method's numerical performance on the test set compared with existing methods under 6 widely used metrics. The methods exploiting object relations generate high scores. The configuration using asymmetric features of overlapping objects generates scores outperforming a state-of-the-art method (Melas-Kyriazi, Rush, and Han 2018) under all 6 metrics. The table also shows that both asymmetric concatenation and asymmetric splitting contribute to higher scores.

The Transformer (Vaswani et al. 2017) is designed for image captioning rather than image paragraph captioning, but it can be applied to image paragraph captioning with minor modification. Its relation learning and feature fusing belong to the all pair symmetric approach, which is similar to our simplified configuration. Thus we also report its performance for comparison.

The proposed method generates higher METEOR scores but slightly. METEOR evaluates machine translating systems and considers the precision and recall of unigram appearance, which hardly reflects accurate relationships since contradictory sentences can also get high scores(we illustrate in supplementary material). In contrast, BLEU (Papineni et al. 2002) and CIDEr (Vedantam, Zitnick, and Parikh 2015) consider n-grams and penalize repeated terms, and they better handle paragraph relationships and coherency.

Overlapping vs. All Pairs. When settings remained the same, experiments in Table 1 show that using only overlapping objects generates higher scores under all 6 metrics. Additionally, in the Stanford dataset with Faster-RCNN detector, overlapping pairs only take 40% of all pairs. Therefore, using only overlapping objects is effective and efficient.

Ablation Studies. We conduct ablation studies to investigate the influence of each component of the proposed method and show the results in Table 2, in which our model is trained by SCST on CIDEr. It shows that the proposed relation network improves the performance of image paragraph captioning under the 6 metrics in multiple training combinations.

Image Captioning. Because the proposed relation network is also applicable to image captioning, we conduct experiments by replacing original object features used by an existing method (Anderson et al. 2018) with ours. We use all the same parameters to train the model and evaluate the replacement on the same MS COCO (Lin et al. 2014) c5 dataset. Quantitative comparisons are reported in Table 3, which shows that the proposed relation-aware features help image captioning. In particular the proposed features generate a higher CIDEr index because CIDEr is the state-of-the-art image captioning metric.

Qualitative Comparison. To illustrate the relations learned by the proposed method, we show the top important pairs and their attention scores in Fig. 4, 5, and 6. The top atten-

Loss	Repetition penalty	Relation Encoding	Relation Attention	C	M	B-1	B-2	B-3	B-4
XE*				12.89	13.66	32.78	19.00	11.40	6.89
XE†		✓		13.11	13.30	33.01	19.21	11.67	6.94
XE		✓	✓	16.27	13.58	34.97	20.17	12.21	7.46
XE*	✓			22.68	15.17	35.68	22.40	14.04	8.70
XE†	✓	✓		22.79	15.21	35.60	22.44	14.07	8.71
XE	✓	✓	✓	22.85	15.43	37.50	23.34	14.63	9.00
SCST*				13.77	13.63	29.67	16.45	9.74	5.88
SCST†		✓		14.27	13.28	30.33	17.97	10.35	6.06
SCST		✓	✓	14.88	13.44	32.84	18.30	10.67	6.21
SCST*	✓			30.63	**17.86**	43.54	27.44	17.33	10.58
SCST†	✓	✓		31.80	17.42	43.67	27.53	17.62	10.73
SCST	✓	✓	✓	**33.38**	17.82	**43.76**	**28.08**	**17.88**	**10.95**

Table 2: Ablation study of the proposed method. *For the four rows, we show indices reported in (Melas-Kyriazi, Rush, and Han 2018) because their settings exactly match ours. †We disable the relation attention network by setting all of its output attention scores as the same constant c.

Used features	CIDEr	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
Anderson *et al* (Anderson et al. 2018)	117.9	**27.6**	**80.2**	64.1	49.1	36.9	57.1
Proposed	**122.0**	**27.6**	**80.2**	**64.3**	**49.4**	**37.1**	**57.8**

Table 3: Improvement of relation-encoded features on top of an existing method (Anderson et al. 2018) on the test split of the MS COCO c5 dataset.

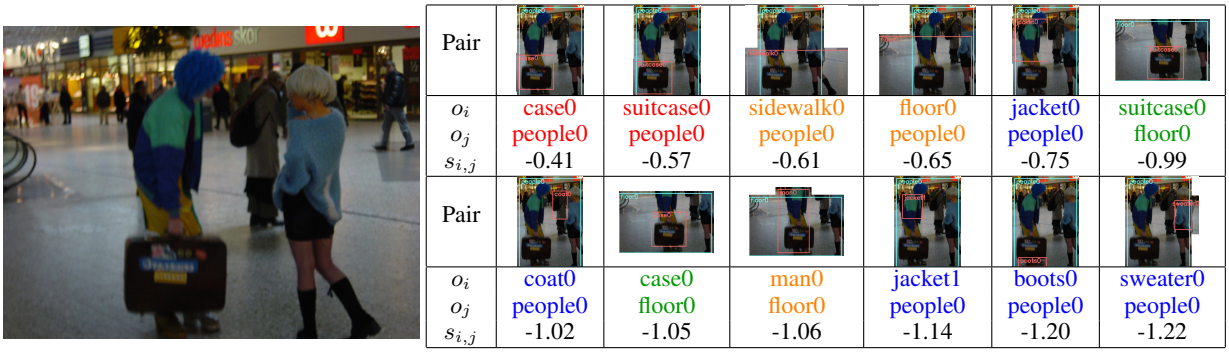
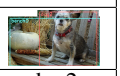


Figure 4: Street. The proposed method effectively recognizes important object relation patterns learned from training data. The 12 pairs with large attention weights (s values) indicate several important relations found on this image.

tion scores in Fig. 4 show that our method effectively learns object relation patterns, e.g. clothing objects are associated with people instead of background objects, and people as well as suitcases are associated with floors. Fig. 5 shows that our method pays most attention to the dog-and-bench pair and generates two sentences expressing this relation, i.e. "A dog is sitting on top of a wooden bench" and "The dog is wearing a black collar on the top of the bench". While the current state-of-the-art method suffers from accurate relation information and generates contradictory sentences, i.e. "The dog is standing in front of the bench", "The bench is on the front of the dog". Refer to the reference paragraph, our generated paragraph gets a higher CIDEr index. Fig. 6 shows that our relation-aware features directly help the language model generate better sentences. The four generated sentences—"A giraffe is standing in the enclosure", "The giraffe is spotted", "The giraffe has long horns on it", and "The is a brick wall behind the giraffe"—are highly similar to the ones in the reference paragraph because of the highly attended object pairs. In contrast, the compared method exploits no object relations and generates no sentences similar to the ones in the reference paragraph.

Conclusion

Image paragraph captioning is a newly developed research topic and there are still many problems unsolved. In this paper, we propose Relation Network to learn the attention importance of pairs of objects on images in respect to paragraphs depicting those images, and the network also encodes relation information into object features without relation labeled data. Experiments show that the proposed relation-aware features generate favorable descriptive paragraphs and their numerical scores under 6 widely used metrics are higher than the ones generated by existing methods.

Pair						
o_i o_j $s_{i,j}$	wall0 bench0 -1.77	plant0 bench0 -1.87	wall1 squash0 -1.88	bench0 squash0 -1.90	dog0 bench0 -1.96	bench0 wall1 -1.99
Pair						
o_i o_j $s_{i,j}$	collar0 mouth0 -2.00	head0 bench0 -2.02	dog0 bench0 -2.02	corn0 bench0 -2.03	hay0 bench0 -2.05	dog2 bench0 -2.09

The proposed method	Melas-Kyriazi *et al* (Melas-Kyriazi, Rush, and Han 2018)	Reference paragraph
A dog is sitting on top of a wooden bench. The dog is brown and white. **The dog is** wearing a black collar **on the top of the bench**. The dog has a black nose on the dog. The is a wooden on the bench. There is a tree on the ground. The bench is make of wood. The wood is made of wood.	A dog is standing on the ground. **The dog is sitting on the bench**. The dog is black. The dog has a black collar. **The dog is standing in front of the bench. The bench is on the front of the dog.** The is a white on the side of the dog. The dog is on the top of the fence. The fence is on the other side of the bench.	There is a wooden bench in front of a tan building. There is a bush next to the bench with green leaves on it. There is a white pumpkin on the bench. There is a colorful ear of corn in front of the pumpkin. There is a dog sitting on top of the bench. The dog is white and gray and has a red collar around its neck.

Method	CIDEr	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Melas-Kyriazi *et al* (Melas-Kyriazi, Rush, and Han 2018)	117.69	**23.24**	49.30	35.11	24.03	15.62
Proposed	**184.55**	22.05	**55.78**	**40.79**	**32.00**	**25.21**

Figure 5: Dog. The proposed method generates a better paragraph than the one generated by (Melas-Kyriazi, Rush, and Han 2018) in terms of message consistency. We highlighted sentences in bold to compare the similar messages delivered by both methods. And the three sentences generated by (Melas-Kyriazi, Rush, and Han 2018) are contradictory against each other.

Pair						
o_i o_j $s_{i,j}$	giraffe0 wall0 -1.90	giraffe0 nostril0 -2.28	giraffe0 mouth0 -2.44	head0 mouth0 -2.49	giraffe1 nostril0 -2.53	giraffe1 mouth0 -2.55
Pair						
o_i o_j $s_{i,j}$	head0 nostril0 -2.57	neck0 spot0 -2.58	neck0 spot1 -2.65	head1 mouth0 -2.65	head1 nostril0 -2.66	giraffe0 horn0 -2.67

The proposed method	Melas-Kyriazi *et al* (Melas-Kyriazi, Rush, and Han 2018)	Reference paragraph
A giraffe is standing in the enclosure. The giraffe is spotted. The giraffe has a long neck. The giraffes are brown. **The giraffe is looking at the giraffe.** The giraffes are black. **The giraffes have long necks.** The ears are black. The horns are black and white. The giraffe has long horns on it. The is a brick wall behind the giraffe.	A giraffe is standing in a giraffe. The giraffe is standing on the side of the giraffe. The giraffe is a brown. The giraffe has a black mane. The is standing on the top of the giraffe. The giraffes is standing in front of the fence. The fence is standing. The giraffe is standing in the front of the building.	An adult giraffe is licking a baby giraffe. The giraffes are in an enclosure made of grey bricks. The adult giraffe has a long black tongue. The giraffe has brown sports on the neck. The eye of the giraffe is big and black. The left ears of the giraffe is long and pointy. On top of the head there are two horns. The baby giraffe has two thin horns.

Method	CIDEr	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Melas-Kyriazi *et al* (Melas-Kyriazi, Rush, and Han 2018)	77.79	19.45	48.77	32.67	23.31	0.00
Proposed	**141.19**	**25.91**	**57.63**	**35.58**	**23.63**	**14.14**

Figure 6: Giraffe. The paragraph generated by the proposed method is easier to read rather than the ones generated by (Melas-Kyriazi, Rush, and Han 2018) for this image. We highlight generated sentences with colors if they have similar reference sentences, in bold if they are high-quality sentences.

Acknowledgements

We appreciate the open-source implementations of image captioning and repetition penalty by Ruotian Luo and Luke Melas-Kyriazi. This research was supported in part by the Ministry of Science and Technology of Taiwan (MOST 108-2633-E-002-001), National Taiwan University (NTU-108L104039), Intel Corporation, Delta Electronics and Compal Electronics.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *CVPR*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Bruna, J.; Zaremba, W.; Szlam, A.; and LeCun, Y. 2013. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203* .

Chatterjee, M.; and Schwing, A. G. 2018. Diverse and Coherent Paragraph Generation from Images. In *ECCV*.

Chen, X.; and Zitnick, C. L. 2014. Learning a recurrent visual representation for image caption generation. *arXiv preprint arXiv:1411.5654* .

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* .

Dai, B.; Fidler, S.; Urtasun, R.; and Lin, D. 2017. Towards diverse and natural image descriptions via a conditional GAN. In *ICCV*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*.

Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*.

Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; and Adams, R. P. 2015. Convolutional networks on graphs for learning molecular fingerprints. In *NeurIPS*.

Henaff, M.; Bruna, J.; and LeCun, Y. 2015. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163* .

Huang, L.; Wang, W.; Chen, J.; and Wei, X.-Y. 2019. Attention on Attention for Image Captioning. In *ICCV*.

Johnson, J.; Gupta, A.; and Fei-Fei, L. 2018. Image Generation from Scene Graphs. In *CVPR*.

Karpathy, A.; and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Krause, J.; Johnson, J.; Krishna, R.; and Fei-Fei, L. 2017. A Hierarchical Approach for Generating Descriptive Image Paragraphs. In *CVPR*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; Bernstein, M.; and Fei-Fei, L. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *arXiv preprint arXiv:1602.07332* .

Kulkarni, G.; Premraj, V.; Ordonez, V.; Dhar, S.; Li, S.; Choi, Y.; Berg, A. C.; and Berg, T. L. 2013. Babytalk: Understanding and generating simple image descriptions. *PAMI* 35(12): 2891–2903.

Liang, X.; Hu, Z.; Zhang, H.; Gan, C.; and Xing, E. P. 2017. Recurrent Topic-Transition GAN for Visual Paragraph Generation. In *ICCV*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Lu, J.; Xiong, C.; Parikh, D.; and Socher, R. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*.

Luo, R.; Price, B.; Cohen, S.; and Shakhnarovich, G. 2018. Discriminability objective for training descriptive captions. In *CVPR*.

Mao, J.; Xu, W.; Yang, Y.; Wang, J.; Huang, Z.; and Yuille, A. 2014. Deep captioning with multimodal recurrent neural networks (m-RNN). *arXiv preprint arXiv:1412.6632* .

Melas-Kyriazi, L.; Rush, A.; and Han, G. 2018. Training for diversity in image paragraph captioning. In *EMNLP*.

Niepert, M.; Ahmed, M.; and Kutzkov, K. 2016. Learning convolutional neural networks for graphs. In *ICML*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical sequence training for image captioning. In *CVPR*.

Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2009. The graph neural network model. *IEEE Transactions on Neural Networks* 20(1): 61–80.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. CIDEr: Consensus-based Image Description Evaluation. In *CVPR*.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *CVPR*.

Wang, J.; Pan, Y.; Yao, T.; Tang, J.; and Mei, T. 2019. Convolutional auto-encoding of sentence topics for image paragraph generation. In *IJCAI*, 940–946. AAAI Press.

Wu, Q.; Shen, C.; Wang, P.; Dick, A.; and van den Hengel, A. 2018. Image captioning and visual question answering based on attributes and external knowledge. *PAMI* 40(6): 1367–1381.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

Yang, X.; Tang, K.; Zhang, H.; and Cai, J. 2019. Auto-encoding scene graphs for image captioning. In *CVPR*.

Yang, Y.; Teo, C. L.; Daumé III, H.; and Aloimonos, Y. 2011. Corpus-guided sentence generation of natural images. In *EMNLP*.

Yao, B. Z.; Yang, X.; Lin, L.; Lee, M. W.; and Zhu, S.-C. 2010. I2T: Image parsing to text description. *Proceedings of the IEEE* 98(8): 1485–1508.

Yao, T.; Pan, Y.; Li, Y.; and Mei, T. 2018. Exploring visual relationship for image captioning. In *ECCV*.

Yao, T.; Pan, Y.; Li, Y.; Qiu, Z.; and Mei, T. 2017. Boosting image captioning with attributes. In *ICCV*.

You, Q.; Jin, H.; Wang, Z.; Fang, C.; and Luo, J. 2016. Image captioning with semantic attention. In *CVPR*.