Improving Image Captioning with Better Use of Captions

Zhan Shi[*], **Xu Zhou**[†], **Xipeng Qiu**[†], **Xiaodan Zhu**[*]
[*]Ingenuity Labs Research Institute, Queen's University
[*]Department of Electrical and Computer Engineering, Queen's University
[†]School of Computer Science, Fudan University
{z.shi,xiaodan.zhu}@queensu.ca,{16210240095,xpqiu}@fudan.edu.cn

Abstract

Image captioning is a multimodal problem that has drawn extensive attention in both the natural language processing and computer vision community. In this paper, we present a novel image captioning architecture to better explore semantics available in captions and leverage that to enhance both image representation and caption generation. Our models first construct caption-guided visual relationship graphs that introduce beneficial inductive bias using weakly supervised multi-instance learning. The representation is then enhanced with neighbouring and contextual nodes with their textual and visual features. During generation, the model further incorporates visual relationships using multi-task learning for jointly predicting word and object/predicate tag sequences. We perform extensive experiments on the MSCOCO dataset, showing that the proposed framework significantly outperforms the baselines, resulting in the state-of-the-art performance under a wide range of evaluation metrics. The code of our paper has been made publicly available. [1]

1 Introduction

Automatically generating a short description for a given image, a problem known as image captioning (Chen et al., 2015), has drawn extensive attention in both the natural language processing and computer vision community. Inspired by the success of encoder-decoder frameworks with the attention mechanism, previous efforts on image captioning adopt variants of pre-trained convolution neural networks (CNN) as the image encoder and recurrent neural networks (RNN) with visual attention as the decoder (Lu et al., 2017; Anderson et al., 2018; Xu et al., 2015; Lu et al., 2018).

Many previous methods translate image representation into natural language sentences without explicitly investigating semantic cues from texts and images. To remedy that, some research has also explored to detect high-level semantic concepts presented in images to improve caption generation (Wu et al., 2016; Gan et al., 2017; You et al., 2016; Fang et al., 2015; Yao et al., 2017). It is believed by many that the inductive bias that leverages structured combination of concepts and visual relationships is of importance, which has led to better captioning models (Yao et al., 2018; Guo et al., 2019; Yang et al., 2019). These approaches obtain visual relationship graphs using models pre-trained from visual relationship detection (VRD) datasets, e.g., Visual Genome (Krishna et al., 2017), where the visual relationships capture semantics between pairs of localized *objects* connected by *predicates*, including spatial (e.g., *cake-on-desk*) and non-spatial semantic relationships (e.g., *man-eat-food*) (Lu et al., 2016).

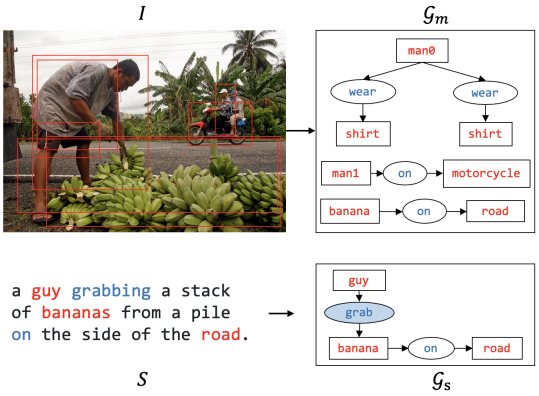

Figure 1: Visual relationship graphs from a pre-trained detection model (Yao et al., 2018) (upper) and from the ground-truth caption (bottom).

As in many other joint text-image modeling problems, it is crucial to obtain a good semantic representation in image captioning that bridges semantics in language and images. The existing approaches, however, have not yet adequately leveraged the semantics available in captions to con-

[1] https://github.com/Gitsamshi/WeakVRD-Captioning

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7454–7464

struct image representation and generate captions. As shown in Figure 1, although VRD detection models present a strong capacity in predicting salient objects and the most common predicates, they often ignore predicates vital for captioning (e.g., "grab" in this example). Exploring better models would still be highly desirable.

A major challenge for establishing a structural connection between captions and images is that the links between predicates and the corresponding object regions are often ambiguous: within the "image-level" label $(obj_1, pred, obj_2)$ extracted from captions, there may exist multiple object regions corresponding to obj_1 and obj_2. In this paper, we propose to use weakly supervised multi-instance learning to detect if a bag of object (region) pairs in an image contain certain predicates, e.g., predicates appearing in ground-truth captions here (or in other applications, they can be any given predicates under concerns). Based on that we can construct caption-guided visual relationship graphs.

Once the visual relationship graphs (VRG) are built, we propose to adapt graph convolution operations (Marcheggiani and Titov, 2017) to obtain representation for object nodes and predicate nodes. These nodes can be viewed as image representation units used for generation.

During generation, we further incorporate visual relationships—we propose multi-task learning for jointly predicting word and tag sequences, where each word in a caption could be assigned with a tag, i.e., *object*, *predicate*, or *none*, which takes as input the graph node features from the above visual relationship graphs. The motivation for predicting a tag in each step is to regularize which types of information should be taken into more consideration for generating words: predicate nodes features, object nodes features, or the current state of language decoder. We study different types of multi-task blocks in our models.

As a result, our models consist of three major components: constructing caption-guided visual relationship graphs (CGVRG) with weakly-supervised multi-instance learning, building context-aware CGVRG, and performing multi-task generation to regularize the network to take into account explicit predicate object/predicate constraints. We perform extensive experiments on the MSCOCO (Lin et al., 2014) image captioning dataset with both supervised and Reinforcement learning strategy (Rennie et al., 2017). The experiment results show that the proposed models significantly outperform the baselines and achieve the state-of-the-art performance under a wide range of evaluation metrics. The main contributions of our work are summarized as follows:

- We propose to construct caption-guided visual relationship graphs that introduce beneficial inductive bias by better bridging captions and images. The representation is further enhanced with neighbouring and contextual nodes with their textual and visual features.
- Unlike existing models, we propose multi-task learning to regularize the network to take into account explicit object/predicate constraints in the process of generation.
- The proposed framework achieves the state-of-the-art performance on the MSCOCO image captioning dataset. We provide detailed analyses on how this is attained.

2 Related Work

Image Captioning A prevalent paradigm of existing image captioning methods is based on the encoder-decoder framework which often utilizes a CNN-plus-RNN architecture for image encoding and text generation (Donahue et al., 2015; Vinyals et al., 2015; Karpathy and Fei-Fei, 2015). Soft or hard visual attention mechanism (Xu et al., 2015; Chen et al., 2017) has been incorporated to focus on the most relevant regions in each generation step. Furthermore, adaptive attention (Lu et al., 2017) has been developed to decide whether to rely on visual features or language model states in each decoding step. Recently, bottom-up attention techniques (Anderson et al., 2018; Lu et al., 2018) have also been proposed to find the most relevant regions based on bounding boxes.

There has been increasing work focusing on filling the gap between image representation and caption generation. Semantic concepts and attributes detected from images have been demonstrated to be effective in boosting image captioning when used in the encoder-decoder frameworks (Wu et al., 2016; You et al., 2016; Gan et al., 2017; Yao et al., 2017). Visual relationship (Lu et al., 2016) and scene graphs (Johnson et al., 2015) have been further employed for image encoder in a unimodal (Yao et al., 2018) or multi-modal (Yang et al., 2019; Guo et al., 2019) manner to improve the over-

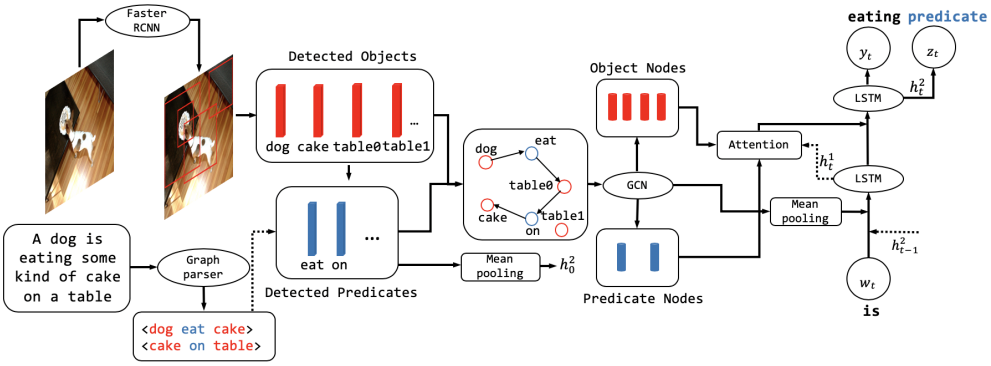

Figure 2: An overview of the proposed image captioning framework.

all performance via the graph convolutional mechanism (Marcheggiani and Titov, 2017). Besides, Kim et al. (2019) proposes a relationship-based captioning task to lead better understanding of images based on relationship. As discussed in introduction, we will further explore the relational semantics available in captions for both constructing image representation and generating caption.

Visual Relationship Detection Visual relations between objects in an image have attracted more studies recently. Conventional visual relation detection have dealt with ⟨*subject-predicate-object*⟩ triples, including spatial relation and other semantic relation. Lu et al. (2016) detect the triples by performing subject, object, and predicate classification separately. Li et al. (2017) attempt to encode more distinguishable visual features for visual relationships detection. Probabilistic output of object detection (Dai et al., 2017; Zhang et al., 2017) is also considered to reason about the visual relationships.

3 The Models

Given an image $\boldsymbol{I}$, the goal of image captioning is to generate a visually grounded natural language sentence. We learn our model by minimizing the cross-entropy loss with regard to the ground truth caption $\boldsymbol{S}^* = \{w_1^*, w_2^*, ..., w_T^*\}$:

$$L_{XE} = -\log p(\boldsymbol{S^*}|\boldsymbol{I}) \quad (1)$$

$$= -\sum_{t=1}^{T} \log p(w_t^*|w_{<t}^*, \boldsymbol{I}) \quad (2)$$

The model is further tuned with a Reinforcement Learning (RL) objective (Rennie et al., 2017) to maximize the reward of the generated sentence $\boldsymbol{S}$:

$$J_{RL} = E_{\boldsymbol{S}\sim p(\boldsymbol{S}|\boldsymbol{I})}(d(\boldsymbol{S}, \boldsymbol{S}^*)) \quad (3)$$

where d is a sentence-level scoring metric.

An overview of our image captioning framework is depicted in Figure 2, with the detail of the components described in the following sections.

3.1 Caption-Guided Visual Relationship Graph (CGVRG) with Weakly Supervised Learning

A general challenge of modeling $p(\boldsymbol{S}|\boldsymbol{I})$ is obtaining a better semantic representation in the multimodal setting to bridge captions and images. Our framework first focuses on constructing caption-guided visual relationship graphs (CGVRG).

3.1.1 Extracting Visual Relationship Triples and Detecting Objects

The process of constructing CGVRG first extracts relationship triples from captions using textual scene graph parser as described in (Schuster et al., 2015). Our framework employs Faster R-CNN (Ren et al., 2015) to recognize instances of objects and returns a set of image regions for objects: $V = \{v_1, v_2, \cdots, v_n\}$.

3.1.2 Constructing CGVRG

The main focus of CGVRG is constructing visual relationship graphs. As discussed in introduction, the existing approaches use pre-trained VRD (visual relationship detection) models, which often ignore key relationships needed for captioning. This gap can be even more prominent if the domain/data used to train image-captioning is farther from where VRD is pretrained. A major challenge to use predicate triples from captions to construct CGVRG is that, the links between predicates and the corresponding object regions are often ambiguous as discussed in introduction. To solve this problem, we use weakly supervised, multi-instance learning.

Obtaining Representation for Object Region Pairs For an image I with a list of salient object regions obtained in object detection $\{v_1, v_2, \cdots, v_n\}$, we have a set of region pairs $U = \{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_N\}$, where $N = n(n-1)$. As shown in Figure 3(b), the visual features of any two object regions and their union box will be collected to compute $p^{r_j}_{\mathbf{u}_n}$, the probability that a region pair $\mathbf{u}_n$ is associated with the predicate r_j, where $r_j \in R$ and $R = \{r_1, r_2, \cdots, r_M\}$ include frequent predicates obtained from the captions in training data. The feed-forward network of Figure 3(b) will be trained in weakly supervised training.

Weakly Supervised Multi-Instance Training As shown in Figure 3(c), during training, one object pair $t = (o_1, o_2)$, e.g., (*women*, *hat*), can correspond to multiple pairs of object regions: the four women-hat combinations between the two women and two hats. To make our description clearer, we refer to $t = (o_1, o_2)$ as an *object pair*, and the four women-hat pairs in the image as *object region pairs*. Accordingly, for a triple we extracted $t = (o_1, r, o_2), r \in R$, e.g., (*woman*, *in*, *hat*), the predicate r (i.e., *in*) can be associated with multiple *object region pairs* (here, (*w0*, *h0*), (*w0*, *h1*), (*w1*, *h0*), and (*w1*, *h1*)).

To predict predicates over object region pairs, we propose to use Multi-Instance Learning (Fang et al., 2015) as our weakly supervised learning approach. Multi-Instance Learning receives a set of labeled bags, each bag containing a set of instances. A bag would be labeled *negative* if all the instances in it are negative. On the other hand, a bag is labeled *positive* if there is at least one positive instance in the bag.

In our problem, an instance is a region pair. Therefore for a candidate predicate $r \in R$ (e.g., *in*), we use $\mathcal{N}_r$ to denote the object region pairs corresponding to predicate r. If r appears in the caption $\boldsymbol{S}$, $\mathcal{N}_r$ would be a positive bag. We use $\mathcal{N} \setminus \mathcal{N}_r$ to denote the negative bag for r. When r is not contained in the caption, the entire $\mathcal{N}$ would be the negative bag (the last row of Figure 3(c)). The probability of a bag b having the predicate r_j is measured with "noisy-OR":

$$p^{r_j}_b = 1 - \prod_{n \in b}(1 - p^{r_j}_{\mathbf{u}_n}) \tag{4}$$

where $p^{r_j}_{\mathbf{u}_n}$ has been introduced above. We adopt the cross-entropy loss on the basis of all predicate

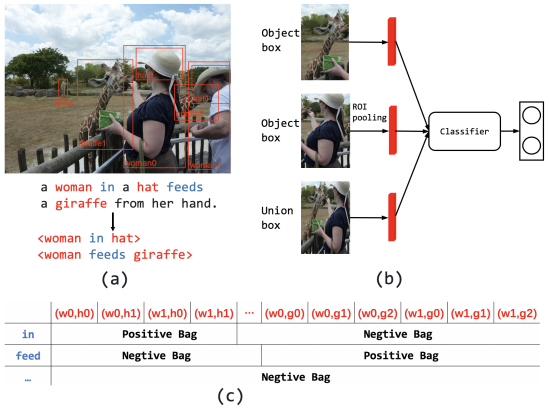

Figure 3: Subcomponents in constructing CGVRG: (a) detecting objects and extracting triples; (b) obtaining representation for object region pairs; (c) examples of positive and negative bags in multi-instance learning for predicate "in" and "feed", respectively. Here, w, h, and g denote *woman*, *hat*, and *giraffe*, respectively.

probabilities over bags, given an image $\boldsymbol{I}$ and caption $\boldsymbol{S}$:

$$L(\boldsymbol{I}) = -\sum_{j=1}^{M}\Big[\mathbb{1}_{(r_j \in \boldsymbol{S})}(\log p^{r_j}_{\mathcal{N}_{r_j}} + \log(1 - p^{r_j}_{\mathcal{N}\setminus\mathcal{N}_{r_j}})) + \mathbb{1}_{(r_j \notin \boldsymbol{S})}(\log(1 - p^{r_j}_{\mathcal{N}}))\Big] \tag{5}$$

where the indicator function $\mathbb{1}_{condition} = 1$ if the condition is true, otherwise $\mathbb{1}_{condition} = 0$.

Constructing the Graphs Once obtaining the trained module, we can build a CGVRG graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ for a given image $\boldsymbol{I}$, where the node set $\mathcal{V}$ includes two types of nodes: object nodes and predicate nodes. We denote o_i as the i^{th} object node and r_{ij} as a predicate node that connects o_i and o_j (refer to Figure 1 or the middle part of Figure 2). The edges in $\mathcal{E}$ are added based on triples; i.e., (o_i, r_{ij}, o_j) will assign two directed edges from node o_i to r_{ij} and from r_{ij} to o_j, respectively.

Note that due to the use of the proposed weakly supervised models, the acquired graphs can now contain predicates that exist in captions but not in the VRD models used in the previous work that does not explicitly consider predicates in captions. We will show in our experiments that this improves captioning quality.

3.2 Context-Aware CGVRG

We further enhance CGVRG in the context of both modalities, images and text, using graph convolution networks. We first integrate visual and textual features: the textual features for each node are

from a word embedding and the visual features are regional visual representations extracted via RoI pooling from Faster R-CNN. The specific features $\mathbf{g}_{o_i}, \mathbf{g}_{r_{ij}}$ for object o_i and predicate r_{ij} are shown as follows:

$$\mathbf{g}_{o_i} = \phi_o([\mathbf{g}^t_{o_i}; \mathbf{g}^v_{o_i}]) \quad (6)$$

$$\mathbf{g}_{r_{ij}} = \phi_r(\mathbf{g}^t_{r_{ij}}) \quad (7)$$

where ϕ_r and ϕ_o are feed-forward networks using ReLU activation; $\mathbf{g}^t_{o_i}, \mathbf{g}^t_{r_{ij}}$, and $\mathbf{g}^v_{o_i}$ denote textual features of o_i, r_{ij} and visual features of o_i, respectively.

We present the process of encoding $\mathcal{G}$ to produce a new set of context-aware representation $\mathcal{X}$. The representation of predicate r_{ij} and o_i are computed as follows:

$$\mathbf{x}_{r_{ij}} = f_r([\mathbf{g}_{o_i}; \mathbf{g}_{o_j}; \mathbf{g}_{r_{ij}}]) \quad (8)$$

$$\mathbf{x}_{o_i} = \frac{1}{N_i}\left[\sum_{r\in\mathcal{N}_{out}(o_i)} f_{out}([\mathbf{g}_{o_i}; \mathbf{g}_r]) + \sum_{r\in\mathcal{N}_{in}(o_i)} f_{in}([\mathbf{g}_{o_i}; \mathbf{g}_r])\right] \quad (9)$$

where f_r, f_{in}, f_{out} are feed-forward networks using ReLU activation. $\mathcal{N}_{in}$ and $\mathcal{N}_{out}$ denote the adjacent nodes with o_i as head and tail, respectively. N_i is the total number of adjacent nodes.

3.3 Multi-task Caption Generation

Unlike the existing image-captioning models, we further incorporate visual relationships into generation — we propose multi-task learning for jointly predicting word and tag sequences as each word in a caption will be assigned a tag, i.e., *object*, *predicate*, or *none*. The module takes as input the graph node features from the context-aware CGVRG. The output of the generation module is hence the sequence of words $\mathbf{y} = \{y_1, \cdots, y_T\}$ as well as the tags $\mathbf{z} = \{z_1, \cdots, z_T\}$. Two different approaches are leveraged to train the two tasks jointly.

The bottom LSTM is used to align a textual state to graph node representations:

$$\mathbf{h}^1_t = \text{LSTM}(\mathbf{h}^1_{t-1}, [\mathbf{h}^2_{t-1}; \overline{\mathbf{x}}; \mathbf{e}_{w_t}]) \quad (10)$$

where LSTM means one step of recurrent unit computation via LSTM; $\overline{\mathbf{x}}$ is the mean-pooled representation of all nodes in the graph; $\mathbf{h}^1_{t-1}$ and $\mathbf{h}^2_{t-1}$ denote hidden states of bottom and top LSTM in time step $t-1$, respectively; $\mathbf{e}$ is the word embedding table.

The state h^1_t is then used as a query to attend over graph node features $\{\mathbf{x}_o\}$ and $\{\mathbf{x}_r\}$ separately to get attended features $\hat{\mathbf{x}}^r_t$ and $\hat{\mathbf{x}}^o_t$:

$$\hat{\mathbf{x}}^r_t = \text{ATT}(\mathbf{h}^1_t, \{\mathbf{x}_r\}) \quad (11)$$

$$\hat{\mathbf{x}}^o_t = \text{ATT}(\mathbf{h}^1_t, \{\mathbf{x}_o\}) \quad (12)$$

where ATT is a soft-attention operation between a query and graph node features.

The top LSTM works as a language model decoder, in which the hidden state $\mathbf{h}^2_0$ is initialized with the mean-pooled semantic representation of all detected predicates $\{r\}$. In time step t, the input consists of the output from the bottom LSTM layer h^1_t and attended graph features $\hat{\mathbf{x}}^r_t$, $\hat{\mathbf{x}}^o_t$:

$$\mathbf{h}^2_t = \text{LSTM}(\mathbf{h}^2_{t-1}, [\mathbf{h}^1_t; \hat{\mathbf{x}}^o_t; \hat{\mathbf{x}}^r_t]) \quad (13)$$

3.3.1 Multi-task Learning

We propose two different blocks to perform the two tasks jointly, as shown in Figure 4. In each step, a multi-task learning block deals with task s_1 as predicting a tag z_t and task s_2 as predicting a word y_t. Specifically **MT-I** treats the two tasks independent of each other:

$$p(z_t|y_{<t}, \boldsymbol{I}) = \text{softmax}(f_z(\mathbf{h}^2_t)) \quad (14)$$

$$p(y_t|y_{<t}, \boldsymbol{I}) = \text{softmax}(f_y(\mathbf{h}^2_t)) \quad (15)$$

where f_z and f_y are feed-forward networks with ReLU activation. Inspired by the adaptive attention mechanism (Lu et al., 2017), **MT-II** further exploits the probability from $p(z_t|y_{<t}, \boldsymbol{I})$ to integrate the representation of current hidden state h^2_t and attended features from graph $\hat{\mathbf{x}}^r_t$, $\hat{\mathbf{x}}^o_t$:

$$p(y_t|y_{<t}, \boldsymbol{I}) = \text{softmax}(f_y(\hat{\mathbf{h}}^2_t)), \quad (16)$$

$$\hat{\mathbf{h}}^2_t = \mathbf{h}^2_t p_{na} + \hat{\mathbf{x}}^r_t p_r + \hat{\mathbf{x}}^o_t p_o \quad (17)$$

$$p(z_t|y_{<t}, \boldsymbol{I}) = \text{softmax}(f_z(\mathbf{h}^2_t)) \quad (18)$$

where p_{na}, p_r, p_o denote the probabilities of tag z_t being "none", "predicate", and "object", respectively. The multi-task loss function is as follows:

$$L_{MT}(\boldsymbol{I}) = -\sum_{t=1}^{T} \log p(y_t|y_{<t}, \boldsymbol{I}) + \gamma \log p(z_t|y_{<t}, \boldsymbol{I}) \quad (19)$$

where γ is the hyper-parameter to balance the two tasks.

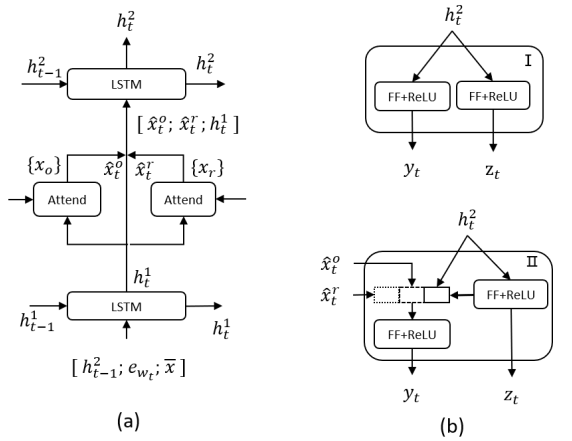

Figure 4: An overview of multi-task caption generation module. Subfigure (a) is a two-layer LSTM; Subfigure (b) depicts two different types of multi-task block.

3.4 Training and Inference

The overall training process can be broken down into two parts: the CGVRG detection module training period and the caption generator training period; the latter includes cross-entropy optimization and the CIDEr-D optimization. For CGVRG detection module training, the detection module is optimized with the multi-instance learning loss in Equation 5. For caption generator training, the model is first optimized with the cross-entropy loss in Equation 19, and then we directly optimize the model with the expected sentence-level reward (CIDEr-D in this work) shown in Equation 3 by self critical sequence learning (Rennie et al., 2017).

In the inference stage, given an image, the CGVRG detection module obtains a graph upon them. The graph convolution network encodes graphs to obtain the context aware multi-modal representations. Then graph object/predicate node features are further provided to the multi-task caption generation module to generate sequences with beam search.

4 Experiments

4.1 Datasets and Experiment Setup

MSCOCO We perform extensive experiments on the MSCOCO benchmark (Lin et al., 2014). The Karpathy split (Karpathy and Fei-Fei, 2015) is adopted for our model selection and offline testing, which contains 113K training images, 5K validation images and 5K testing images. As for the online test server, the result is trained on the entire training and validation set (123K images). To evaluate the generated captions, we employ standard evaluation metrics: SPICE (Anderson et al., 2016), CIDEr-D (Vedantam et al., 2015), METEOR (Denkowski and Lavie, 2014), ROUGE-L (Lin, 2004), and BLEU (Papineni et al., 2002).

Visual Genome We use the Visual Genome (Krishna et al., 2017) dataset to pre-train our object detection model. The dataset includes 108K images. To pre-train the object detection model with Faster R-CNN, we strictly follow the setting in (Anderson et al., 2018), taking 98K/5K/5K for training, validation, and testing, respectively. The split is carefully selected to avoid contamination of the MSCOCO validation and testing sets, since nearly 51K Visual Genome images are also included in the MSCOCO dataset.

Implementation Details We use Faster R-CNN (Ren et al., 2015) to identify and localize instances of objects. The object detection phase consists of two modules. The first module proposes object regions using a deep CNN, i.e., ResNet-101 (He et al., 2016). The second module extracts feature maps using region-of-interest pooling for each box proposals. Practically, we take the final output of the ResNet-101 and perform non-maximum suppression for each object class with an IoU threshold. As a result, we obtain a set of image regions, $V = \{v_1, v_2, \cdots, v_n\}$, where $n \in [10, 100]$ varies with input images and confidence thresholds. Each region is represented as a 2,048-dimensional vector obtained from the pool5 layer after the RoI pooling. We then apply a feed-forward network with a 1000-dimensional output layer for predicates classification. The network of the same size is also used for feature projection (ϕ_o, ϕ_i) and GCN (f_r, f_{in}, f_{out}). In the decoder LSTM, the word embedding dimension is set to be 1,000 and the hidden unit dimension in the top-layer and bottom-layer LSTM is set to be 1,000 and 512, respectively. The trade-off parameter γ in multi-task learning is 0.15. The whole system is trained with the Adam optimizer. We set the initial learning rate to be 0.0005 and mini-batch size to be 100. The maximum number of training epochs is 30 for Cross-entropy and CIDEr-D optimization respectively. For sequence generation in the inference stage, we adopt the beam search strategy and set the beam size to be 3.

We construct object and predicate categories for VRD training. Similar to (Lu et al., 2018), we manually expand the original 80 object categories to

	Cross entropy						CIDEr-D optimization					
	B1	B4	ME	RG	CD	SP	B1	B4	ME	RG	CD	SP
SCST	-	31.3	26.0	54.3	101.3	-	-	33.3	26.3	55.3	111.4	-
LSTM-A	75.4	35.2	26.9	55.8	108.8	20.0	78.6	35.5	27.3	56.8	118.3	20.8
Up-Down (Baseline)	77.2	36.2	27.0	56.4	113.5	20.3	79.8	36.3	27.7	56.9	120.1	21.4
StackCap	76.2	35.2	26.5	-	109.1	-	78.6	36.1	27.4	-	120.4	-
CAVP	-	-	-	-	-	-	-	38.6	28.3	58.5	126.3	21.6
GCN-LSTM	77.3	36.8	27.9	57.0	116.3	20.9	80.5	38.2	28.5	58.3	127.6	22.0
VSUA	-	-	-	-	-	-	-	38.4	28.5	58.4	128.6	22.0
SGAE	77.6	36.9	27.7	57.2	116.7	20.9	80.8	38.4	28.4	58.6	127.8	22.1
This Work (MT-I)	**78.1**	**38.4**	**28.2**	**58.0**	**119.0**	21.1	**80.8**	**38.9**	**28.8**	**58.7**	**129.6**	22.3
This Work (MT-II)	77.9	38.0	28.1	57.6	117.8	**21.3**	80.5	38.6	28.7	58.4	128.7	**22.4**

Table 1: Single-model performances on the MSCOCO dataset (Karpathy split) in both cross-entropy and RL training period. B1, B4, ME, RG, CD, and SP denote BLEU-1, BLEU-4, METEOR, ROUGE, CIDEr-D and SPICE, respectively.

	B4		ME		RG		CD	
	c5	c40	c5	c40	c5	c40	c5	c40
GCN-LSTM*	**38.7**	69.7	28.5	37.6	58.5	73.4	**125.3**	126.5
VSUA	37.4	68.3	28.2	37.1	57.9	72.8	123.1	125.5
SGAE	38.5	69.7	28.2	37.2	58.6	73.6	123.8	126.5
Baseline	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
This Work	38.6	**70.1**	**28.6**	**37.8**	**58.8**	**74.5**	125.1	**126.7**

Table 2: The performance on COCO online test server of various methods that incorporate visual relationships. * denotes that their training batch size and epochs are far beyond average setting in (Anderson et al., 2018; Yang et al., 2019).

413 fine-grained categories by utilizing a list of caption tokens. For example, the object category *"person"* is expanded to a list of fine-grained categories [*"boy"*, *"man"*, $\cdots$]. Then for all extracted triples that have both objects appearing in the 413 category list, we select the 200 most frequent predicates as our predicate categories.

4.2 Quantitative Analysis

Model Comparison We compare our models with the following state-of-the-art models: (1) SCST (Rennie et al., 2017) employs an improved policy gradient algorithm by utilizing its own inference output to normalize the rewards; (2) LSTM-A (Yao et al., 2017) integrates the detected image attributes into the CNN-plus-RNN image captioning framework; (3) Up-Down (Anderson et al., 2018) uses both a bottom-up and top-down attention mechanism to focus more on salient object regions; (4) GCN-LSTM (Yao et al., 2018) leverages graph convolutional networks over the detected objects and relations; (5) CAVP (Liu et al., 2018) proposes a context-aware policy network by accounting for visual attentions as context for generation; (6) VSUA (Guo et al., 2019) exploits the alignment between words and different categories of graph nodes; (7) SAGE (Yang et al., 2019) utilizes an additional graph encoder to incorporate language inductive bias into the encoder-decoder framework.

Our baseline is built on Up-Down (Anderson et al., 2018). We propose two variants of final models using different multi-task blocks, namely MT-I and MT-II shown in Fig 4(b). We conduct extensive comparisons on the dataset with the above state-of-the-art techniques. We also perform detailed analysis to demonstrate the impact of different components of our framework.

Table 1 lists the results of various single models on the MSCOCO Karpathy split. Our model outperforms the baseline model significantly, with CIDEr-D scores being improved from 113.5 to 119.0 and 120.1 to 129.6 in the cross-entropy and CIDEr-D optimization period, respectively. In addition, the model with MT-II shows an advantage over that with MT-I on SPICE, which implies that the proposed adaptive visual attention mechanism works in multi-task block II.

Table 2 compares our model with three models that also incorporate VRG, plus the baseline model, on the MSCOCO online test server. Our model improves significantly from the baseline (from 120.5 to 126.7 in CIDEr-D) and has achieved the best results across all evaluation metrics on c40 (40 reference captions).

Figure 5 shows the effect of taking different weights γ in the multi-task loss item (Equation 19). The results indicate that the weight around 0.15 yields the best performance in both multi-task blocks. Meanwhile, Figure 6 shows the ablation analysis by removing the multi-task caption generation and graph convolution operation, respectively, to check the effect of these components. The results

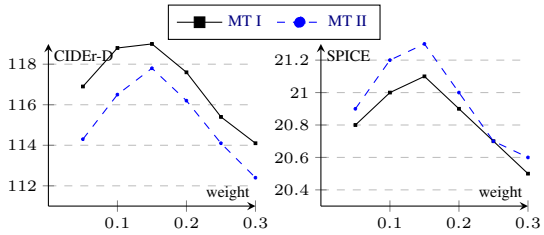

Figure 5: Test results (cross-entropy optimization) on various γ.

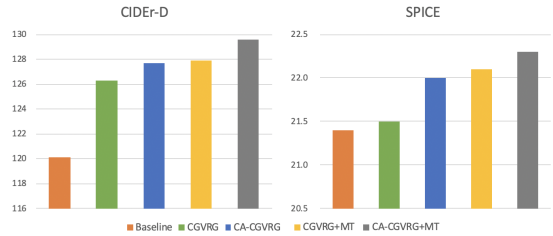

Figure 6: Ablation results (CIDEr-D optimization).

show that both the graph convolution operation and multi-task learning help improve the quality of the generated captions.

Note that the code of our paper has been made publicly available in the webpage provided in the abstract.

Human evaluation We performed human evaluation with three non-author human subjects, using a five-level Likert scale. For each image and each pair of systems in comparison (MT-I vs. Up-Down, MT-I vs. GCN-LSTM, and MT-I vs. SGAE), we show the captions generated by the two systems to the human subjects. We ask each subject if the first caption sentence is: significantly better (2), better (1), equal (0), worse (-1), or significantly worse (-2), compared to the second.

Following (Zhao et al., 2019), we obtain the subjects' ratings for fidelity (the first caption is superior in terms of making less mistakes?), informativeness (the first caption provides more informative and detailed description?), and fluency (the first caption is more fluent?). For each question asked for an image, we calculate the average of the three subjects' scores. For each pair of models in comparison, we randomly sampled 50 images from the Karpathy testset.

- MT-I vs. Up-Down: For fidelity, MT-I is better or significantly better on 44% images (where the average of the three human subjects' scores is larger than 0.5), equal to Up-Down on 46% images (the average is in range $[-0.5, 0.5]$), and worse or significantly worse on 10% images (average is less than -0.5). For informativeness, MT-I is better or significantly better on 60% images, equal on 34%, and worse or significantly worse on 6%. For fluency, the numbers are 18%, 72%, and 10%.

- MT-I vs. GCN-LSTM: For fidelity, MT-I is better or significantly better on 40% images, equal to GCN-LSTM on 52%, and worse or significantly worse on 8%. For informativeness, the numbers are 32%, 50%, and 18%, respectively. For fluency, the numbers are 12%, 76%, and 12%.

- MT-I vs. SGAE: For fidelity, MT-I is better or significantly better on 36% images, equal to SGAE on 56%, and worse or significantly worse on 8%. For informativeness, the numbers are 30%, 48%, and 22%, respectively. For fluency, the numbers are 6%, 90%, and 4%.

4.3 Qualitative Analysis

Figure 7 shows several specific examples, each including an image, a detected caption guided visual relationship graph, a ground truth sentence, a generated word sequence, and a learned visual relationship composition. We can see that the proposed model generates more accurate captions coherent to the visual relationship detected in the image. Consider the upper middle demo as an example; our model extracts a visual relationship graph covering the critical predicates "filled with" and "in front of" for understanding the image, thus producing a comprehensive description. In addition, we observe that the model generates the triple $(table, filled\ with, food)$, which is a new composition that has not appeared in the training set.

Figure 8 visualizes the effect of our tag sequence generation process. Specifically, we visualize the tag probabilities of the "object", "predicate", and "none" category in each generation step. Our model successfully learns to distinguish the correct category for each time step, which is in consistent with the tag of the predicted word. For example, for the generated words "flying over", the probability for the "predicate" category is the highest, which is also true for words like "bird" and "water".

5 Conclusions

This paper presents a novel image captioning architecture that constructs caption-guided visual relationship graphs to introduce beneficial inductive

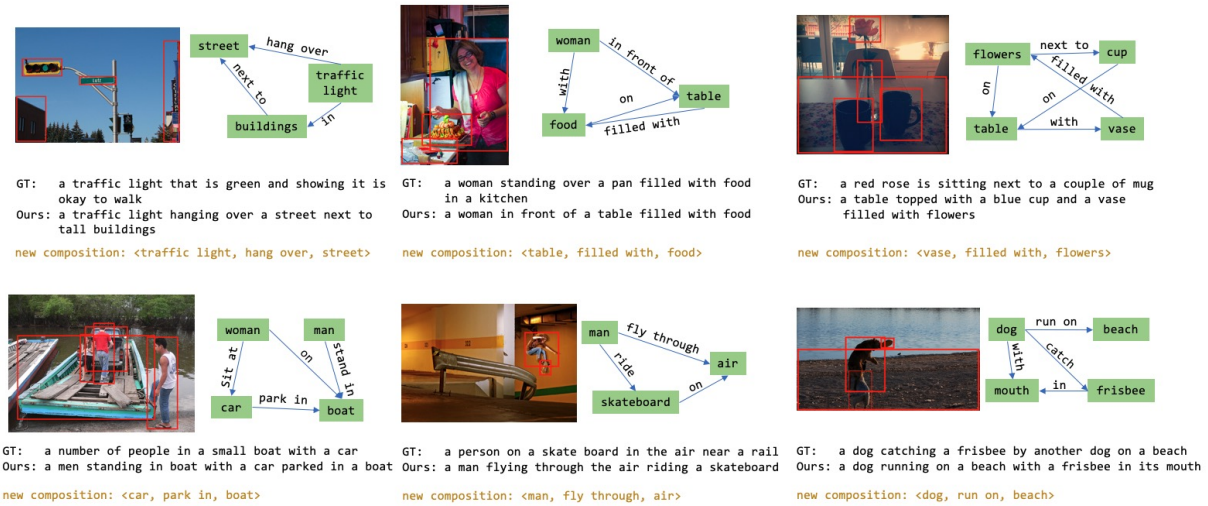

Figure 7: Several image captioning examples generated by our model.

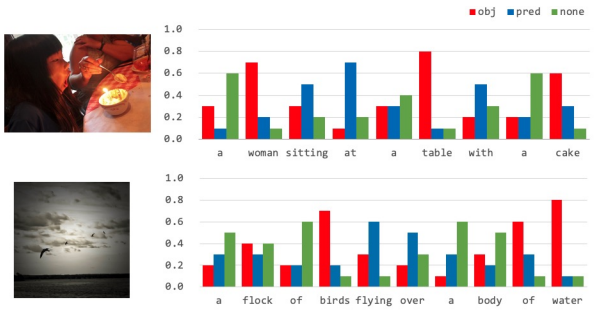

Figure 8: Examples of generated word and tag sequences.

bias to better utilize captions. The representation is further enhanced with text and visual features of neighbouring nodes. During generation, the network is regularized to take into account explicit object/predicate constraints with multi-task learning. Extensive experiments are performed on the MSCOCO dataset, showing that the proposed framework significantly outperforms the baselines, resulting in the state-of-the-art performance under various evaluation metrics. In the near future we plan to extend the proposed approach to several other language-vision modeling tasks.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments. This research of the first and last author is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *ECCV*, pages 382–398. Springer.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086.

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 *CVPR*, pages 5659–5667.

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

Bo Dai, Yuqi Zhang, and Dahua Lin. 2017. Detecting visual relationships with deep relational networks. In *CVPR*, pages 3076–3086.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, pages 2625–2634.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *CVPR*, pages 1473–1482.

Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. 2017. Semantic compositional networks for visual captioning. In *CVPR*, pages 5630–5639.

Longteng Guo, Jing Liu, Jinhui Tang, Jiangwei Li, Wei Luo, and Hanqing Lu. 2019. Aligning linguistic words and visual semantic units for image captioning. *arXiv*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*, pages 770–778.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *CVPR*, pages 3668–3678.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137.

Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. 2019. Dense relational captioning: Triple-stream networks for relationship-based captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6271–6280.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Yikang Li, Wanli Ouyang, Xiaogang Wang, and Xiao'ou Tang. 2017. Vip-cnn: Visual phrase guided convolutional neural network. In *CVPR*, pages 1347–1356.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer.

Daqing Liu, Zheng-Jun Zha, Hanwang Zhang, Yongdong Zhang, and Feng Wu. 2018. Context-aware visual policy network for sequence-level image captioning. *arXiv*.

Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016. Visual relationship detection with language priors. In *ECCV*, pages 852–869. Springer.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, pages 375–383.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural baby talk. In *CVPR*, pages 7219–7228.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. *arXiv*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *CVPR*, pages 7008–7024.

Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D. Manning. 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Workshop on Vision and Language (VL15)*, Lisbon, Portugal. ACL.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164.

Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton Van Den Hengel. 2016. What value do explicit high level concepts have in vision to language problems? In *CVPR*, pages 203–212.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057.

Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-encoding scene graphs for image captioning. In *CVPR*, pages 10685–10694.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *ECCV*, pages 684–699.

Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting image captioning with attributes. In *ICCV*, pages 4894–4902.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *CVPR*, pages 4651–4659.

Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. 2017. Visual translation embedding network for visual relation detection. In *CVPR*, pages 5532–5540.

Sanqiang Zhao, Piyush Sharma, Tomer Levinboim, and Radu Soricut. 2019. Informative image captioning with external sources of information. *arXiv preprint arXiv:1906.08876*.