Say As You Wish: Fine-grained Control of Image Caption Generation with Abstract Scene Graphs

Shizhe Chen[1]*, Qin Jin[1]†, Peng Wang[2], Qi Wu[3]
[1]Renmin University of China, [2]Northwestern Polytechnical University
[3]Australian Centre for Robotic Vision, University of Adelaide
{cszhe1, qjin}@ruc.edu.cn, peng.wang@nwpu.edu.cn, qi.wu01@adelaide.edu.au

Abstract

*Humans are able to describe image contents with coarse to fine details as they wish. However, most image captioning models are intention-agnostic which can not generate diverse descriptions according to different user intentions initiatively. In this work, we propose the Abstract Scene Graph (ASG) structure to represent user intention in fine-grained level and control what and how detailed the generated description should be. The ASG is a directed graph consisting of three types of **abstract nodes** (object, attribute, relationship) grounded in the image without any concrete semantic labels. Thus it is easy to obtain either manually or automatically. From the ASG, we propose a novel ASG2Caption model, which is able to recognise user intentions and semantics in the graph, and therefore generate desired captions according to the graph structure. Our model achieves better controllability conditioning on ASGs than carefully designed baselines on both VisualGenome and MSCOCO datasets. It also significantly improves the caption diversity via automatically sampling diverse ASGs as control signals.*

1. Introduction

Image captioning is a complex problem since it requires a machine to complete several computer vision tasks, such as object recognition, scene classification, attributes and relationship detection, simultaneously, and then summarise them to a sentence. Thanks to the rapid development of deep learning [14, 15], recent image captioning models [3, 34, 43] have made substantial progress and even outperform humans in terms of several accuracy-based evaluation metrics [5, 30, 39].

However, most image captioning models are intention-agnostic and only passively generate image descriptions, which do not care about what contents users are interested in, and how detailed the description should be. On the contrary, we humans are able to describe image contents from coarse to fine details as we wish. For example, we can describe more discriminative details (such as the quantity and colour) of flowers in Figure 1 if we are asked to do so, but current systems totally fail to realise such user intention. What is worse, such passive caption generation can greatly hinder diversity and tend to generate mediocre descriptions [37, 41]. Despite achieving high accuracy, these descriptions mainly capture frequent descriptive patterns and cannot represent holistic image understanding, which is supposed to recognise different aspects in the image and thus be able to produce more diverse descriptions.

In order to address aforementioned limitations, few previous endeavours have proposed to actively control image

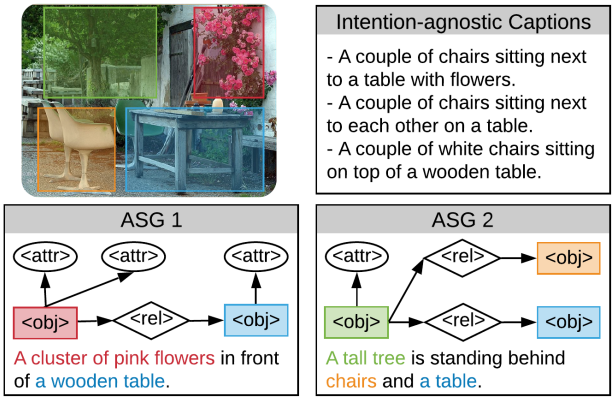


Figure 1: Although intention-agnostic captions can correctly describe image contents, they fail to realise what a user wants to describe and lack of diversity. Therefore, we propose Abstract Scene Graphs (ASG) to control the generation of user desired and diverse image captions in fine-grained level. The corresponding region, ASG node and generated phrase are labelled as the same colour.

*This work was performed when Shizhe Chen was visiting University of Adelaide.

†Qin Jin is the corresponding author.

captioning process. One type of works [10, 13, 27] focuses on controlling expressive styles of image descriptions such as factual, romantic, humorous styles *etc*., while the other type aims to control the description contents such as different image regions [17], objects [7, 50], and part-of-speech tags [9], so that the model is able to describe user interested contents in the image. However, all of the above works can only handle a coarse-grained control signal such as one-hot labels or a set of image regions, which are hard to realise user desired control at a fine-grained level, for instance describing various objects in different level of details as well as their relationships.

In this work, we propose a more fine-grained control signal, *Abstract Scene Graph* (ASG), to represent different intentions for controllable image caption generation. As shown in Figure 1, the ASG is a directed graph consisting of three types of abstract nodes grounded in the image, namely object, attribute and relationship, while no concrete semantic label is necessary for each node. Therefore, such graph structure is easy to obtain either manually or automatically since it does not require semantic recognition. More importantly, the ASG is capable of reflecting user's fine-grained intention on what to describe and how detailed to describe.

In order to generate captions respecting designated ASGs, we then propose an ASG2Caption model based on an encoder-decoder framework which tackles three main challenges in ASG controlled caption generation. Firstly, notice that our ASG only contains an abstract scene layout without any semantic labels, it is necessary to capture both intentions and semantics in the graph. Therefore, we propose a role-aware graph encoder to differentiate fine-grained intention roles of nodes and enhance each node with graph contexts to improve semantic representation. Secondly, the ASG not only controls what contents to describe via different nodes, but also implicitly decides the descriptive order via how nodes are connected. Our proposed decoder thus considers both content and structure of nodes for attention to generate desired content in graph flow order. Last but not least, it is important to fully cover information in ASG without missing or repetition. For this purpose, our model gradually updates the graph representation during decoding to keep tracking of graph access status.

Since there are no available datasets with ASG annotation, we automatically construct ASGs for training and evaluation on two widely used image captioning datasets, VisualGenome and MSCOCO. Extensive experiments demonstrate that our approach can achieve better controllability given designated ASGs than carefully designed baselines. Furthermore, our model is capable of generating more diverse captions based on automatically sampled ASGs to describe various aspects in the image.

The contributions of our work are three-fold:

- To the best of our knowledge, we are the first to propose fine-grained control of image caption generation with Abstract Scene Graph, which is able to control the level of details (such as, whether attributes, relationships between objects should be included) in the caption generation process.
- The proposed ASG2Caption model consists of a *role-aware graph encoder* and *language decoder for graphs* to automatically recognise abstract graph nodes and generate captions with intended contents and orders.
- We achieve state-of-the-art controllability given designated ASGs on two datasets. Our approach can also be easily extended to automatically generated ASGs, which is able to generate diverse image descriptions.

2. Related Work

2.1. Image Captioning

Image captioning [3, 11, 40, 42, 43] has achieved significant improvements based on neural encoder-decoder framework [38]. The Show-Tell model [40] employs convolutional neural networks (CNNs) [14] to encode image into fixed-length vector, and recurrent neural networks (RNNs) [15] as decoder to sequentially generate words. To capture fine-grained visual details, attentive image captioning models [3, 25, 43] are proposed to dynamically ground words with relevant image parts in generation. To reduce exposure bias and metric mismatching in sequential training [32], notable efforts are made to optimise non-differentiable metrics using reinforcement learning [24, 34]. To further boost accuracy, detected semantic concepts [11, 42, 47] are adopted in captioning framework. The visual concepts learned from large-scale external datasets also enable the model to generate captions with novel objects beyond paired image captioning datasets [1, 26]. A more structured representation over concepts, scene graph [18], is further explored [45, 46] in image captioning which can take advantage of detected objects and their relationships. In this work, instead of using a fully detected scene graph (which is already a challenging enough task [48, 49]) to improve captioning accuracy, we propose to employ abstract scene graph (ASG) as a control signal to generate desired and diverse image captions. The ASG is convenient to interact with human to control captioning in fine-grained level, and easier to be obtained automatically than fully detected scene graphs.

2.2. Controllable Image Caption Generation

Controllable text generation [16, 20] aims to generate sentences conditioning on designated control signals such as sentiment, styles, semantic *etc*., which is more interactive, interpretable and easier to generate diverse sentences. There are broadly two groups of control for image captioning, namely style control and content control. The style control researches [10, 13, 27, 28] aim to describe global image

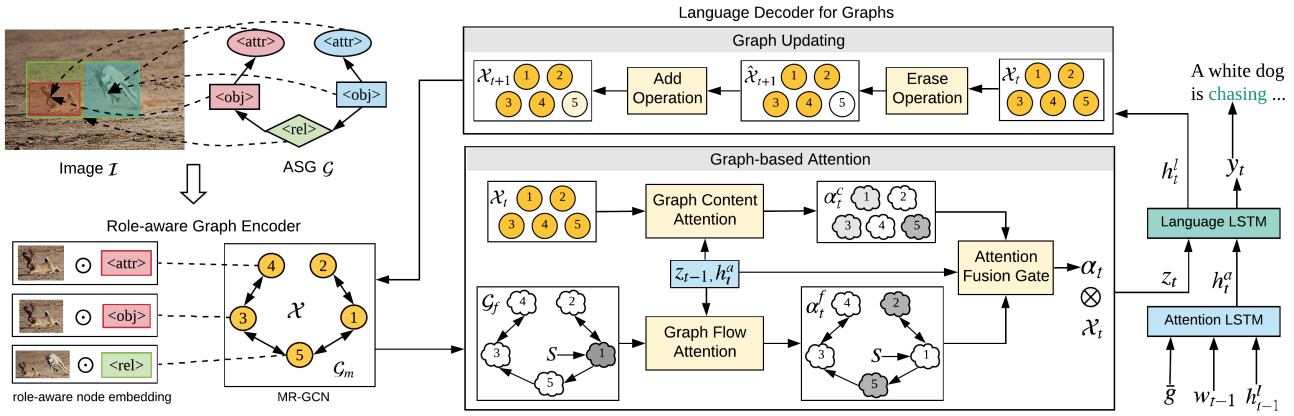


Figure 2: The proposed ASG2Caption model consists of a role-aware graph encoder and a language decoder for graphs. Given an image $\mathcal{I}$ and ASG $\mathcal{G}$, our encoder first initialises each node as role-aware embedding, and employs a multi-layer MR-GCN to encode graph contexts in $\mathcal{G}_m$. Then the decoder dynamically incorporates graph content and graph flow attentions for ASG-controlled captioning. After generating a word, we update the graph $\mathcal{X}_{t-1}$ into $\mathcal{X}_t$ to record graph access status.

content with different styles. The main challenge is lack of paired stylised texts for training. Therefore, recent works [10, 13, 27] mainly disentangle style codes from semantic contents so that unpaired style transfer can be applied.

The content control works [7, 17, 44, 50] instead aim to generate captions capturing different aspects in the image such as different regions, objects and so on, which are more relevant to holistic visual understanding. Johnson *et al*. [17] is the first to propose the dense captioning task, which detects and describes diverse regions in the image. Zheng *et al*. [50] constrain the model to involve a human concerned object. Cornia *et al*. [7] further control multiple objects and their orders in the generated description. Besides manipulating on object-level, Deshpande *et al*. [9] employ Part-of-Speech (POS) syntax to guide caption generation, which however mainly focus on improving diversity rather than POS control. Beyond single image, Park *et al*. [31] propose to only describe semantic differences between two images.

However, none of above works can control caption generation at more fine-grained level. For instance, whether (and how many) associative attributes should be used? Any other objects (and its associated relationships) should be included and what is the description order? In this paper, we propose to utilise fine-grained ASG to control designated structure of objects, attributes and relationships at the same time, and enable generating more diverse captions that reflect different intentions.

3. Abstract Scene Graph

In order to represent user intentions in fine-grained level, we first propose an Abstract Scene Graph (ASG) as the control signal for generating customised image captions. An ASG for image $\mathcal{I}$ is denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ are the sets of nodes and edges respectively. As illustrated in the top left of Figure 2, the nodes can be classified into three types according to their intention roles: object node o, attribute node a and relationship node r. The user intention is constructed into $\mathcal{G}$ as follows:

- add user interested object o_i to $\mathcal{G}$, where object o_i is grounded in $\mathcal{I}$ with a corresponding bounding box;
- if the user wants to know more descriptive details of o_i, add an attribute node $a_{i,l}$ to $\mathcal{G}$ and assign a directed edge from o_i to $a_{i,l}$. $|l|$ is the number of associative attributes since multiple $a_{i,l}$ for o_i are allowed;
- if the user wants to describe relationship between o_i and o_j, where o_i is the subject and o_j is the object, add relationship node $r_{i,j}$ to $\mathcal{G}$ and assign directed edges from o_i to $r_{i,j}$ and from $r_{i,j}$ to o_j respectively.

It is convenient for an user to construct the ASG $\mathcal{G}$, which represents the user's interests about objects, attributes and relationships in $\mathcal{I}$ in a fine-grained manner.

Besides obtaining $\mathcal{G}$ from users, it is also easier to generate ASGs automatically based on off-the-shelf object proposal networks and optionally a simple relationship classifier to tell whether two objects contain any relationship. Notice that our ASG is only a graph layout without any semantic labels, which means we do not rely on externally trained object/attribute/relationship detectors, but previous scene graph based image captioning models [46] need these well-trained detectors to provide a complete scene graph with labels and have to suffer the low detection accuracy.

The details of automatic ASG generation are provided in the supplementary material. In this way, diverse ASGs can be extracted to capture different aspects in the image and thus lead to diverse caption generation.

4. The ASG2Caption Model

Given an image $\mathcal{I}$ and a designated ASG $\mathcal{G}$, the goal is to generate a fluent sentence $y = \{y_1, \cdots, y_T\}$ that strictly aligns with $\mathcal{G}$ to satisfy user's intention. In this section, we present the proposed ASG2Caption model which is illustrated in Figure 2. We will describe the proposed encoder and decoder in Section 4.1 and 4.2 respectively, followed by its training and inference strategies in Section 4.3.

4.1. Role-aware Graph Encoder

The encoder is proposed to encode ASG $\mathcal{G}$ grounded in image $\mathcal{I}$ as a set of node embeddings $\mathcal{X} = \{x_1, \cdots, x_{|\mathcal{V}|}\}$. Firstly, x_i is supposed to reflect its intention role besides the visual appearance, which is especially important to differentiate object and connected attribute nodes since they are grounded in the same region. Secondly, since nodes are not isolated, contextual information from neighboured nodes is beneficial to recognise semantic meaning of the node. Therefore, we propose a role-aware graph encoder, which contains a *role-aware node embedding* to distinguish node intentions and a *multi-relational graph convolutional network (MR-GCN)* [35] for contextual encoding.

Role-aware Node Embedding. For the i-th node in $\mathcal{G}$, we firstly initialise it as a corresponding visual feature v_i. Specifically, the feature of object node is extracted from the grounded bounding box in the image; the feature of attribute node is the same as its connected object; and the feature of relationship node is extracted from the union bounding box of the two involved objects. Since visual features alone cannot distinguish intention roles of different nodes, we further enhance each node with role embedding to obtain a role-aware node embedding $x_i^{(0)}$ as follows:

$$x_i^{(0)} = \begin{cases} v_i \odot W_r[0], & \text{if } i \in o; \\ v_i \odot (W_r[1] + \text{pos}[i]), & \text{if } i \in a; \\ v_i \odot W_r[2], & \text{if } i \in r. \end{cases} \quad (1)$$

where $W_r \in \mathbb{R}^{3\times d}$ is the role embedding matrix, d is the feature dimension, $W_r[k]$ denotes the k-th row of W_r, and $\text{pos}[i]$ is a positional embedding to distinguish the order of different attribute nodes connected with the same object.

Multi-relational Graph Convolutional Network. Though the edge in ASG is uni-directional, the influence between connected nodes is mutual. Furthermore, since nodes are of different types, how the message passing from one type of node to another is different from its inverse direction. Therefore, we extend the original ASG with different bidirectional edges, which leads to a multi-relational graph $\mathcal{G}_m = \{\mathcal{V}, \mathcal{E}, \mathcal{R}\}$ for contextual encoding.

Specifically, there are six types of edges in $\mathcal{R}$ to capture mutual relations between neighboured nodes, which are: object to attribute, subject to relationship, relationship to object and their inverse directions respectively. We employ a MR-GCN to encode graph context in $\mathcal{G}_m$ as follows:

$$x_i^{(l+1)} = \sigma(W_0^{(l)} x_i^{(l)} + \sum_{\tilde{r}\in\mathcal{R}} \sum_{j\in\mathcal{N}_i^{\tilde{r}}} \frac{1}{|\mathcal{N}_i^{\tilde{r}}|} W_{\tilde{r}}^{(l)} x_j^{(l)}) \quad (2)$$

where $\mathcal{N}_i^{\tilde{r}}$ denotes neighbours of i-th node under relation $\tilde{r} \in \mathcal{R}$, σ is the ReLU activation function, and $W_*^{(l)}$ are parameters to be learned at l-th MR-GCN layer. Utilising one layer brings contexts from direct neighboured nodes for each node, while stacking multiple layers enables to encode broader contexts in the graph. We stack L layers and then the outputs of the final L-th layer are employed as our final node embeddings $\mathcal{X}$. We can also obtain a global graph embedding via taking an average of $\mathcal{X}$ as $\bar{g} = \frac{1}{|\mathcal{V}|}\sum_i x_i$. We fuse global graph embedding with global image representation as global encoded feature $\bar{v}$.

4.2. Language Decoder for Graphs

The decoder aims to convert the encoded $\mathcal{G}$ into an image caption. Unlike previous works that attend on a set of unrelated vectors [25, 43], our node embeddings $\mathcal{X}$ contain structured connections from $\mathcal{G}$, which reflects user designated order that should not be ignored. Furthermore, in order to fully satisfy user intention, it is important to express all the nodes in $\mathcal{G}$ without missing or repetition, while previous attention methods [25, 43] hardly consider accessed status of attended vectors. Therefore, in order to improve the graph to sentence quality, we propose a language decoder specifically for graphs, which includes a *graph-based attention mechanism* that considers both graph semantics and structures, and a *graph updating mechanism* that keeps a record of what has been described or not.

Overview of the Decoder. The decoder employs a two-layer LSTM structure [3], including an attention LSTM and a language LSTM. The attention LSTM takes the global encoded embedding $\bar{v}$, previous word embedding w_{t-1} and previous output from language LSTM h_{t-1}^l as input to compute an attentive query h_t^a:

$$h_t^a = \text{LSTM}([\bar{v}; w_{t-1}; h_{t-1}^l], h_{t-1}^a; \theta^a) \quad (3)$$

where $[;]$ is vector concatenation and θ^a are parameters.

We denote node embeddings at t-th step as $\mathcal{X}_t = \{x_{t,1}, \cdots, x_{t,|\mathcal{V}|}\}$ where $\mathcal{X}_1$ is the output of encoder $\mathcal{X}$. The h_t^a is used to retrieve a context vector z_t from $\mathcal{X}_t$ via the proposed graph-based attention mechanism. Then language LSTM is fed with z_t and h_t^a to generate word sequentially:

$$h_t^l = \text{LSTM}([z_t; h_t^a], h_{t-1}^l; \theta^l) \quad (4)$$

$$p(y_t|y_{<t}) = \text{softmax}(W_p h_t^l + b_p) \quad (5)$$

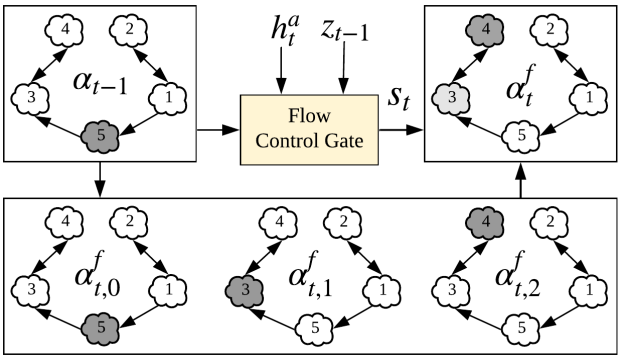


Figure 3: Graph flow attention employs graph flow order to select relevant nodes to generate next word.

where θ^l, W_p, b_p are parameters. After generating word y_t, we update node embeddings $\mathcal{X}_t$ into $\mathcal{X}_{t+1}$ via the proposed graph updating mechanism to record new graph access status. We will explain the graph-based attention and graph updating mechanisms in details in the following sections.

Graph-based Attention Mechanism. In order to take into account both semantic content and graph structure, we combine two types of attentions called *graph content attention* and *graph flow attention* respectively.

The graph content attention considers semantic relevancy between node embeddings $\mathcal{X}_t$ and the query h_t^a to compute an attention score vector $\boldsymbol{\alpha_t^c}$, which is:

$$\tilde{\alpha}_{t,i}^c = w_c^T \tanh(W_{xc} x_{t,i} + W_{hc} h_t^a) \tag{6}$$

$$\boldsymbol{\alpha_t^c} = \text{softmax}(\tilde{\boldsymbol{\alpha}}_{\boldsymbol{t}}^{\boldsymbol{c}}) \tag{7}$$

where W_{xc}, W_{hc}, w_c are parameters in content attention and we omit the bias term for simplicity. Since connections between nodes are ignored, the content attention is similar to teleport which can transfer from one node to another node in far distance in $\mathcal{G}$ at different decoding timesteps.

However, the structure of ASG implicitly reflects the user intended orders on caption generation. For example, if the current attended node is a relationship node, then the next node to be accessed is most likely to be the following object node according to the graph flow. Therefore, we further propose a graph flow attention to capture the graph structure. The flow graph $\mathcal{G}_f$ is illustrated in Figure 2, which is different from the original ASG in three ways. The first is that a start symbol S should be assigned and the second difference lies in the bidirectional connection between object node and attribute node since in general the order of objects and their attributes are not compulsive and should be decided by sentence fluency. Finally, a self-loop edge will be constructed for a node if there exists no output edge of the node, which ensures the attention on the graph doesn't vanish. Suppose M_f is the adjacent matrix of the flow graph $\mathcal{G}_f$, where the i-th row denotes the normalised in-degree of the i-th node. The graph flow attention transfers attention score vector in previous decoding step $\boldsymbol{\alpha_{t-1}}$ in three ways:

1) stay at the same node $\boldsymbol{\alpha_{t,0}^f} = \boldsymbol{\alpha_{t-1}}$. For example, the model might express one node with multiple words;
2) move one step $\boldsymbol{\alpha_{t,1}^f} = M_f \boldsymbol{\alpha_{t-1}}$, for instance transferring from a relationship node to its object node;
3) move two steps $\boldsymbol{\alpha_{t,2}^f} = (M_f)^2 \boldsymbol{\alpha_{t-1}}$ such as transferring from a relationship node to an attribute node.

The final flow attention is a soft interpolation of the three flow scores controlled by a dynamic gate as follows:

$$\boldsymbol{s_t} = \text{softmax}(W_s \sigma(W_{sh} h_t^a + W_{sz} z_{t-1})) \tag{8}$$

$$\boldsymbol{\alpha_t^f} = \sum_{k=0}^{2} s_{t,k} \boldsymbol{\alpha_{t,k}^f} \tag{9}$$

where W_s, W_{sh}, W_{sz} are parameters and $\boldsymbol{s_t} \in \mathbb{R}^3$. Figure 3 presents the process of graph flow attention.

Our graph-based attention dynamically fuses the graph content attention $\boldsymbol{\alpha_t^c}$ and the graph flow attention $\boldsymbol{\alpha_t^f}$ with learnable parameters w_g, W_{gh}, W_{gz}, which is:

$$\beta_t = \text{sigmoid}(w_g \sigma(W_{gh} h_t^a + W_{gz} z_{t-1})) \tag{10}$$

$$\boldsymbol{\alpha_t} = \beta_t \boldsymbol{\alpha_t^c} + (1 - \beta_t) \boldsymbol{\alpha_t^f} \tag{11}$$

Therefore, the context vector for predicting word at the t-th step is $z_t = \sum_{i=1}^{|\mathcal{V}|} \alpha_{t,i} x_{t,i}$, which is a weighted sum of graph node features.

Graph Updating Mechanism. We update the graph representation to keep a record of accessed status for different nodes in each decoding step. The attention score $\boldsymbol{\alpha_t}$ indicates accessed intensity of each node so that highly attended node is supposed to be updated more. However, when generating some non-visual words such as "the" and "of", though graph nodes are accessed, they are not expressed by the generated word and thus should not be updated. Therefore, we propose a visual sentinel gate as [25] to adaptively modify the attention intensity as follows:

$$\boldsymbol{u_t} = \text{sigmoid}(f_{vs}(h_t^l; \theta_{vs})) \boldsymbol{\alpha_t} \tag{12}$$

where we implement f_{vs} as a fully connected network parametrised by θ_{vs} which outputs a scalar to indicate whether attended node is expressed by the generated word.

The updating mechanism for each node is decomposed into two parts: an erase followed by an add operation inspired by NTM [12]. Firstly, the i-th graph node representation $x_{t,i}$ is erased according to its update intensity $u_{t,i}$ in a fine-grained way for each feature dimension:

$$e_{t,i} = \text{sigmoid}(f_{ers}([h_t^l; x_{t,i}]; \theta_{ers})) \tag{13}$$

$$\hat{x}_{t+1,i} = x_{t,i}(1 - u_{t,i} e_{t,i}) \tag{14}$$

Table 1: Statistics of VisualGenome and MSCOCO datasets for controllable image captioning with ASGs.

dataset	train		validation		test		#objs	#rels	#attrs	#words
	#imgs	#sents	#imgs	#sents	#imgs	#sents	per sent	per sent	per obj	per sent
VisualGenome	96,738	3,397,459	4,925	172,290	4,941	171,759	2.09	0.95	0.47	5.30
MSCOCO	112,742	475,117	4,970	20,851	4,979	20,825	2.93	1.56	0.51	10.28

Table 2: Comparison with carefully designed baselines for controllable image caption generation conditioning on ASGs.

Method	VisualGenome									MSCOCO								
	B4	M	R	C	S	G	G_o	G_a	G_r	B4	M	R	C	S	G	G_o	G_a	G_r
ST [40]	11.1	17.0	34.5	139.9	31.1	1.2	0.5	0.7	0.5	10.5	16.8	36.2	100.6	24.1	1.8	0.8	1.1	1.0
BUTD [3]	10.9	16.9	34.5	139.4	31.4	1.2	0.5	0.7	0.5	11.5	17.9	37.9	111.2	26.4	1.8	0.8	1.1	1.0
C-ST	12.8	19.0	37.6	157.6	36.6	1.1	0.4	0.7	0.4	14.4	20.1	41.4	135.6	32.9	1.6	0.6	1.0	0.8
C-BUTD	12.7	19.0	37.9	159.5	36.8	1.1	0.4	0.7	0.4	15.5	20.9	42.6	143.8	34.9	1.5	0.6	1.0	0.8
Ours	**17.6**	**22.1**	**44.7**	**202.4**	**40.6**	**0.7**	**0.3**	**0.3**	**0.3**	**23.0**	**24.5**	**50.1**	**204.2**	**42.1**	**0.7**	**0.4**	**0.3**	**0.3**

Therefore, a node can be set as zero if it is no longer need to be accessed. In case a node might need multiple access and track its status, we also employ an add update operation:

$$a_{t,i} = \sigma(f_{add}([h_t^l; x_{t,i}]; \theta_{add})) \quad (15)$$

$$x_{t+1,i} = \hat{x}_{t+1,i} + u_{t,i} a_{t,i} \quad (16)$$

where f_{ers} and f_{add} are fully connected networks with different parameters. In this way, we update the graph embeddings $\mathcal{X}_t$ into $\mathcal{X}_{t+1}$ for the next decoding step.

4.3. Training and Inference

We utilise the standard cross entropy loss to train our ASG2Caption model. The loss for a single pair $(\mathcal{I}, \mathcal{G}, y)$ is:

$$L = -\log \sum_{t=1}^{T} p(y_t | y_{<t}, \mathcal{G}, \mathcal{I}) \quad (17)$$

After training, our model can generate controllable image captions given the image and designated ASG obtained manually or automatically as described in Section 3.

5. Experiments

5.1. Datasets

We automatically construct triplets of (image $\mathcal{I}$, ASG $\mathcal{G}$, caption y) based on annotations of two widely used image captioning datasets, VisualGenome [21] and MSCOCO [23]. Table 1 presents statistics of the two datasets.

VisualGenome contains object annotations and dense regions descriptions. To obtain ASG for corresponding caption and region, we firstly use a Stanford sentence scene graph parser [36] to parse groundtruth region caption to a scene graph. We then ground objects from the parsed scene graph to object regions according to their locations and semantic labels. After aligning objects, we remove all the semantic labels from the scene graph, and only keep the graph layout and nodes type. More details are in the supplementary material. We follow the data split setting in [3].

MSCOCO dataset contains more than 120,000 images and each image is annotated with around five descriptions. We use the same way as VisualGenome to get ASGs for training. We adopt the 'Karpathy' splits setting [19]. As shown in Table 1, the ASGs in MSCOCO are more complex than those in VisualGenome dataset since they contain more relationships and the captions are longer.

5.2. Experimental Settings

Evaluation Metrics. We evaluate caption qualities in terms of two aspects, *controllability* and *diversity* respectively. To evaluate the controllability given ASG, we utilise ASG aligned with groundtruth image caption as control signal. The generated caption is evaluated against groundtruth via five automatic metrics including BLEU [30], METEOR [5], ROUGE [22], CIDEr [39] and SPICE [2]. Generally, those scores are higher if semantic recognition is correct and sentence structure aligns better with the ASG. We also propose a Graph Structure metric G based on SPICE [2] to purely evaluate whether the structure is faithful to ASG. It measures difference of numbers for (o), (o, a) and (o, r, o) pairs respectively between generated and groundtruch caption, where the lower the better. We also break down the overall score G for each type of pairs as G_o, G_a, G_r respectively. More details are in the supplementary material.

For the diversity measurement, we first sample the same number of image captions for each model, and evaluate the diversity of sampled captions using two types of metrics: 1) n-gram diversity (Div-n): a widely used metric [9, 4] which is the ratio of distinct n-grams to the total number of words

Table 3: Ablation study to demonstrate contributions from different proposed components. (role: role-aware node embedding; rgcn: MR-GCN; ctn: graph content attention; flow: graph flow attention; gupdt: graph updating; bs: beam search)

#	Enc		Dec				VisualGenome					MSCOCO				
	role	rgcn	ctn	flow	gupdt	bs	B4	M	R	C	S	B4	M	R	C	S
1							11.2	18.3	36.7	146.9	35.6	13.6	19.7	41.3	130.2	32.6
2			✓				10.7	18.2	36.9	146.3	35.5	14.5	20.4	42.2	135.7	34.6
3	✓		✓				14.2	20.5	40.9	176.9	38.1	18.2	22.5	44.9	166.9	37.8
4	✓	✓	✓				15.7	21.4	43.6	191.7	40.0	21.6	23.7	48.6	190.5	40.9
5	✓	✓	✓	✓			15.9	21.5	44.0	193.1	40.1	22.3	24.0	49.4	196.2	41.5
6	✓	✓	✓		✓		15.8	21.4	43.5	191.6	39.9	21.8	24.1	49.1	194.2	41.4
7	✓	✓	✓	✓	✓		16.1	21.6	44.1	194.4	40.1	22.6	24.4	50.0	199.8	41.8
8	✓	✓	✓	✓	✓	✓	**17.6**	**22.1**	**44.7**	**202.4**	**40.6**	**23.0**	**24.5**	**50.1**	**204.2**	**42.1**

in the best 5 sampled captions; 2) SelfCIDEr [41]: a recent metric to evaluate semantic diversity derived from latent semantic analysis and kernelised to use CIDEr similarity. The higher scores the more diverse captions are.

Implementation Details. We employ Faster-RCNN [33] pretrained on VisualGenome to extract visual features for grounded nodes in ASG and ResNet152 pretrained on ImageNet [8] to extract global image representations. For role-aware graph encoder, we set the feature dimension as 512 and L as 2. For language decoder, the word embedding and hidden size of LSTM layers are set to be 512. During training, the learning rate is 0.0001 with batch size of 128. In the inference phrase, we utilise beam search with beam size of 5 if not specified.

5.3. Evaluation on Controllability

We compare the proposed approach with two groups of carefully designed baselines. The first group contains traditional intention-agnostic image captioning models, including: 1) Show-Tell (ST) [40] which employs a pretrained Resnet101 as encoder to extract global image representation and an LSTM as decoder; and 2) state-of-the-art BottomUpTopDown (BUTD) model [3] which dynamically attends over relevant image regions when generating different words. The second group of models extend the above approaches for ASG-controlled image captioning. For the non-attentive model (C-ST), we fuse global graph embedding $\bar{g}$ with the original feature; while for the attentive model (C-BUTD), we make the model attend to graph nodes in ASG instead of all detected image regions.

Table 2 presents the comparison result. It is worth noting that controllable baselines outperform non-controllable baselines due to the awareness of control signal ASG. We can also see that baseline models are struggling to generate designated attributes compared to objects and relationships according to detailed graph structure metrics. Our proposed method significantly improves performance than compared

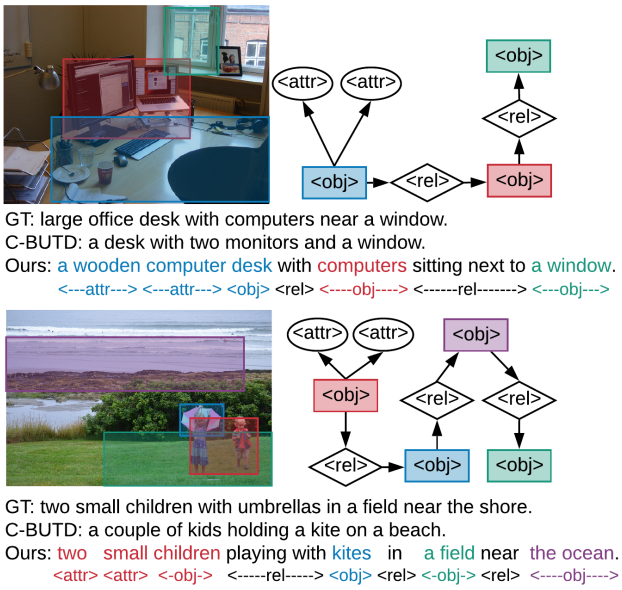


Figure 4: Examples on controllability given designated ASG for different captioning models.

approaches on all evaluation metrics in terms of both overall caption quality and alignment with graph structure. Especially for fine-grained attribute control, we reduce more than half of misalignment on VisualGenome (0.7 $\rightarrow$ 0.3) and MSCOCO (1.0 $\rightarrow$ 0.3) dataset. In Figure 4, we visualise some examples of our ASR2Caption model and the best baseline model C-BUTD. Our model is more effective to follow designated ASGs for caption generation than C-BUTD model. In the bottom image of Figure 4, though both models fail to recognise the correct concept "umbrella", our model still successfully aligns with the graph structure.

In order to demonstrate contributions from different components in our model, we provide an extensive ablation study in Table 3. We begin with baselines (Row 1 and 2) which are C-ST and C-BUTD model respectively. Then

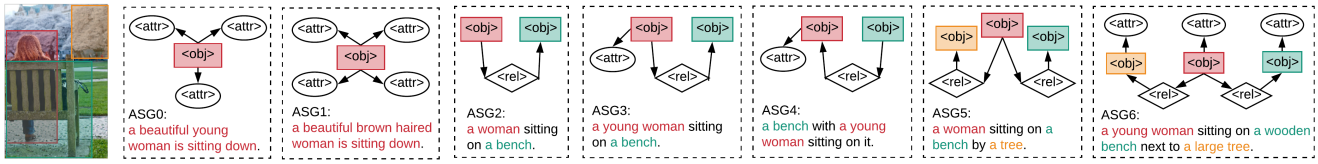


Figure 5: Generated image captions using user created ASGs for the leftmost image. Even subtle changes in the ASG represent different user intentions and lead to different descriptions. Best viewed in colour.

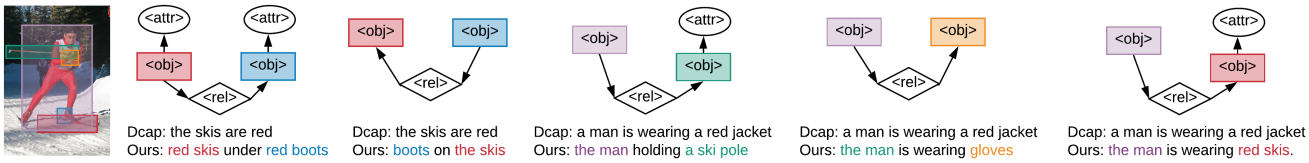


Figure 6: Examples for diverse image caption generation conditioning on sampled ASGs. Our generated captions are different from each other while the comparison baseline (dense-cap) generates repeated captions. Best viewed in colour.

in Row 3, we add the role-aware node embedding in the encoder and the performance is largely improved, which indicates that it is important to distinguish different intention roles in the graph. Comparing Row 4 against Row 3 where the MR-GCN is employed for contextual graph encoding, we see that graph context is beneficial for the graph node encoding. Row 5 and 6 enhance the decoder with graph flow attention and graph updating respectively. The graph flow attention shows complementarity with the graph content attention via capturing the structure information in the graph, and outperforms Row 4 on two datasets. However, the graph updating mechanism is more effective on MSCOCO dataset where the number of graph nodes are larger than on VisualGenome dataset. Since the graph updating module explicitly records the status of graph nodes, the effectiveness might be more apparent when generating longer sentences for larger graphs. In Row 7, we incorporate all the proposed components which obtains further gains. Finally, we apply beam search on the proposed model and achieves the best performance.

Besides ASGs corresponding to groundtruth captions, in Figure 5 we show an example of user created ASGs which represent different user intentions in a fine-grained level. For example, ASG0 and ASG1 care about different level of details about the woman, while ASG2 and ASG5 intends to know relationships between various number of objects. Subtle differences such as directions of edges also influence the captioning order as shown in ASG3 and ASG4. Even for large complex graphs like ASG6, our model still successfully generates desired image captions.

5.4. Evaluation on Diversity

The bonus of our ASG-controlled image captioning is the ability to generate diverse image descriptions that capture different aspects of the image at different level of details given diverse ASGs. We first automatically obtain a global ASG for the image (Section 3), and then sample subgraphs from the ASG. For simplicity, we randomly select connected subject-relationship-object nodes as subgraph and randomly add one attribute node to subject and object nodes. On VisualGenome dataset, we compare with dense image captioning approach which generates diverse captions to describe different image regions. For fair comparison, we employ the same regions as our sampled ASGs. On MSCOCO dataset, since there are only global image descriptions for images, we utilise beam search of BUTD model to produce diverse captions as baseline. We also compare with other state-of-the-art methods [4, 9] on MSCOCO dataset that strive for diversity.

Table 4: Comparison with state-of-the-art approaches for diverse image caption generation.

	Method	Div-1	Div-2	SelfCIDEr
Visual Genome	Region	0.41	0.43	0.47
	Ours	**0.54**	**0.63**	**0.75**
MS COCO	BS [4]	0.21	0.29	-
	POS [9]	0.24	0.35	-
	SeqCVAE [4]	0.25	0.54	-
	BUTD-BS	0.29	0.39	0.58
	Ours	**0.43**	**0.56**	**0.76**

As shown in Table 4, the generated captions of our approach are more diverse than compared methods especially on the SelfCider score [41] which focuses on semantic similarity. We illustrate an example image with different ASGs in Figure 6. The generated caption effectively respects the given ASG, and the diversity of ASGs leads to significant diverse image descriptions.

6. Conclusion

In this work, we focus on controllable image caption generation which actively considers user intentions to generate desired image descriptions. In order to provide a fine-grained control on what and how detailed to describe, we propose a novel control signal called Abstract Scene Graph (ASG), which is composed of three types of abstract nodes (object, attribute and relationship) grounded in the image without any semantic labels. An ASG2Caption model is then proposed with a role-aware graph encoder and a language decoder specifically for graphs to follow structures of the ASG for caption generation. Our model achieves state-of-the-art controllability conditioning on user desired ASGs on two datasets. It also significantly improves diversity of captions given automatically sampled ASGs.

References

[1] Harsh Agrawal, Karan Desai, Xinlei Chen, Rishabh Jain, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Proceedings of the European Conference on Computer Vision*, pages 382–398. Springer, 2016. 6, 12

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 1, 2, 4, 6, 7, 13

[4] Jyoti Aneja, Harsh Agrawal, Dhruv Batra, and Alexander Schwing. Sequential latent spaces for modeling the intention during diverse image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 6, 8

[5] Satanjeev Banerjee and Alon Lavie. 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*, pages 65–72, 2005. 1, 6

[6] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *ICCV*, pages 5561–5569, 2017. 11

[7] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2019. 2, 3

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 7

[9] Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander G. Schwing, and David Forsyth. Fast, diverse and accurate image captioning guided by part-of-speech. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2019. 2, 3, 6, 8

[10] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3137–3146, 2017. 2, 3

[11] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5630–5639, 2017. 2

[12] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014. 5

[13] Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Hanqing Lu. Mscap: Multi-style image captioning with unpaired stylized text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4204–4213, 2019. 2, 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1, 2

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 1, 2

[16] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1587–1596. JMLR.org, 2017. 2

[17] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574, 2016. 2, 3

[18] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3668–3678, 2015. 2

[19] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015. 6

[20] Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909*, 2019. 2

[21] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 6

[22] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014. 6

[24] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE international conference on computer vision*, pages 873–881, 2017. 2

[25] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 375–383, 2017. 2, 4, 5

[26] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7219–7228, 2018. 2

[27] Alexander Mathews, Lexing Xie, and Xuming He. Semstyle: Learning to generate stylised image captions using unaligned text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8591–8600, 2018. 2, 3

[28] Alexander Patrick Mathews, Lexing Xie, and Xuming He. Senticap: Generating image descriptions with sentiments. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016. 2

[29] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 11

[30] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics, 2002. 1, 6

[31] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *Proceedings of the IEEE international conference on computer vision*, October 2019. 3

[32] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *Proceedings of the International Conference on Learning Representations*, 2016. 2

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015. 7

[34] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024, 2017. 1, 2

[35] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607. Springer, 2018. 4

[36] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the Fourth Workshop on Vision and Language*, pages 70–80, 2015. 6, 11

[37] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4135–4144, 2017. 1

[38] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112, 2014. 2

[39] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015. 1, 6

[40] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015. 2, 6, 7

[41] Qingzhong Wang and Antoni B Chan. Describing like humans: on diversity in image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4195–4203, 2019. 1, 7, 8

[42] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton Van Den Hengel. What value do explicit high level concepts have in vision to language problems? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 203–212, 2016. 2

[43] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning*, pages 2048–2057, 2015. 1, 2, 4

[44] Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li. Dense captioning with joint inference and visual context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2193–2202, 2017. 3

[45] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10685–10694, 2019. 2

[46] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European Conference on Computer Vision*, pages 684–699, 2018. 2, 3

[47] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4651–4659, 2016. 2

[48] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global con-

text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5831–5840, 2018. 2

[49] Ji Zhang, Kevin J. Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[50] Yue Zheng, Yali Li, and Shengjin Wang. Intention oriented image captions with guiding objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2019. 2, 3

A. Automatic ASG Generation

Since the abstract scene graph does not require semantic labels, we could just utilize an off-the-shelf object proposal model to detect possible regions as object nodes. The attribute and relationship nodes then can be added arbitrarily on or between object nodes because we can always describe attributes of an object or find certain relationship between two objects in the image. However, not all relationships are meaningful and common to us. For example, "a dog is chasing a rabbit" is more common than "a dog is chasing a computer". Therefore, we can optionally employ a simple relationship classifier to tell whether two objects contain a meaningful relationship.

We train the relationship classifier with annotations in groundtruth ASGs. Instead of recognizing exact semantic labels which is rather challenging, we only predict three classes, with 0 for no relationship between two objects, 1 for subject-to-object relationship and 2 for object-to-subject relationship. Three types of features are utilized for the prediction. The first type is the global image appearance. The second type is the region visual features of the two objects respectively, and the third type is the feature for relative spatial location of the two objects. We balance the ratio of different classes as 2:1:1 during training.

For inference, we firstly detect bounding boxes of objects and apply SoftNMS [6] to reduce redundancy. Then we utilize the trained relationship classifier for each pair of objects. Two objects are considered to contain meaningful relationship if the probability of class 0 is below certain threshold (0.5 in our experiments) and the relationship of two objects are selected as class 1 or 2 according to the predicted probabilities. In this way, we build a global ASG which contains abstract object and relationship nodes.

B. ASG Dataset Construction

For the VisualGenome dataset, although there are grounded region scene graphs for each region description, we notice that these region graphs are noisy with missing objects, relationships and misaligned attributes. Therefore, we only utilize existing region scene graphs in VisualGenome as references to construct our ASGs. For the MSCOCO dataset, since there are no grounded scene graphs, we need to build grounded ASGs from scratch. The detailed steps of building an ASG $\mathcal{G}$ for image $\mathcal{I}$ and its image description y are as follows:

1. utilize Stanford scene graph parser [36] to parse description y to a scene graph, where there are both semantic label and node type for each node and connections between nodes.
2. collect candidate object bounding boxes and labels. For VisualGenome, we use the annotated object bounding boxes. For MSCOCO, we utilise an off-the-shelf object detector (Faster-RCNN pretrained on VisualGenome dataset) to detect objects.
3. ground objects in the parsed scene graph to candidate object bounding boxes in the image. For VisualGenome, we take into account both location overlap between candidate objects and the region and semantic similarity of labels based on WordNet [29] for grounding. For MSCOCO, we can only utilize the semantic similarity of labels for grounding.
4. remove noisy grounded scene graphs. If there are more than two objects in a scene graph without grounding, we remove the scene graph. For the remained scene graph, if an object cannot be grounded, we align the object with the region bounding box for VisualGenome and the global image for MSCOCO dataset.
5. remove all semantic labels of nodes and only keep the graph layout and nodes type as our ASG $\mathcal{G}$.

To be noted, since the two datasets are automatically constructed, there mainly exists two types of noises especially for MSCOCO dataset where no object grounding annotations are available. The two types of errors are sentence parsing error and object grounding error as shown in Figure 7. For example, in Figure 8 (a), the attribute "ornate" is

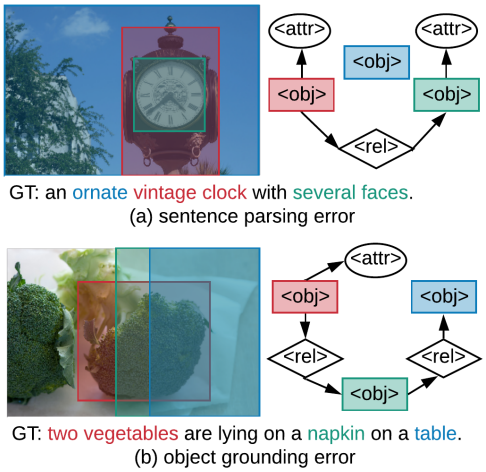
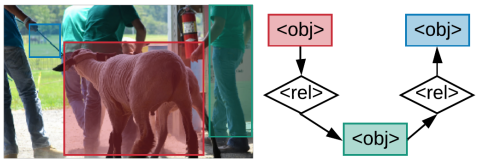
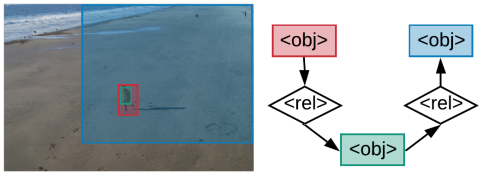


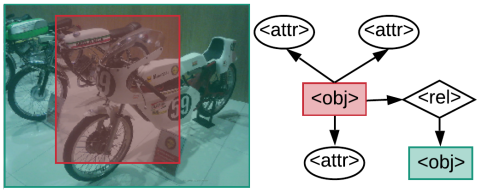
Figure 7: Two types of errors in the automatic dataset construction (examples from the testing set of MSCOCO).



GT: a sheep being pulled by a person with a rope.
Ours: a elephant is standing next to a man with a bat.
(a) object error



GT: a man in a wetsuit running on the beach.
Ours: a man in a wetsuit walking on the beach.
(b) relationship error



GT: there are many different dirt bikes on display.
Ours: a white and white motorcycle parked on display.
(c) attribute error

Figure 8: Three types of mistakes in our ASG2Caption model for controllable image caption generation (examples from the testing set of MSCOCO).

mistaken as an object by incorrect sentence parsing; in Figure 7 (b), the object "vegetables" is only grounded on one broccoli but not two of them in the image. However, since majority of the constructed pairs are correct, our model still can learn from the imperfect datasets.

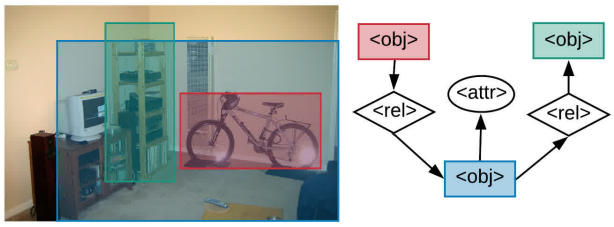
C. Graph Structure Metric

The proposed Graph Structure metric is based on SPICE metric [2]. The SPICE metric parses a sentence into three types of tuples (o), (o, a) and (o, r, o) and measures the semantic alignment of tuples between generated caption and groundtruth captions. However, our Graph Structure metric only cares about the structure alignment which reflects the structure control of ASG without considering the semantic correctness. For this purpose, we first calculate the numbers of the three types of tuples in the generated caption and groundtruth caption respectively. Then we employ the mean absolute error for each tuple type as the structure misalignment measure, which is G_o, G_a, G_r for measurement of (o), (o, a) and (o, r, o) respectively. The overall misalignment G is the average of errors of the three tuple types. The lower the score is, the better the structure alignment is.

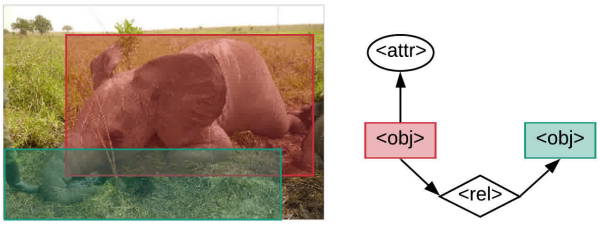
D. Additional Qualitative Results

Figure 9 presents additional examples on controllable image caption generation with designated ASGs. Figure 10 provides more examples on diverse image caption generation with sampled ASGs.

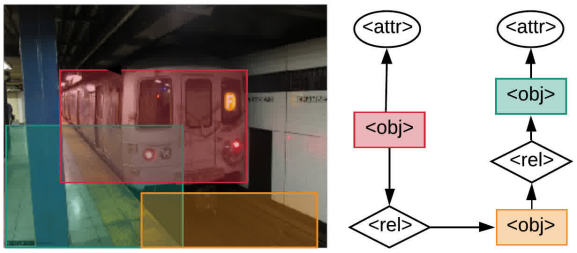
In Figure 8, we further present three main types of mistakes that our ASG2Caption model can make for controllable image caption generation, including object recognition error, relationship detection error and attribute generation error. The attribute generation error mostly occurs when multiple attributes are required, which can lead to generation of repeated or incorrect attributes.



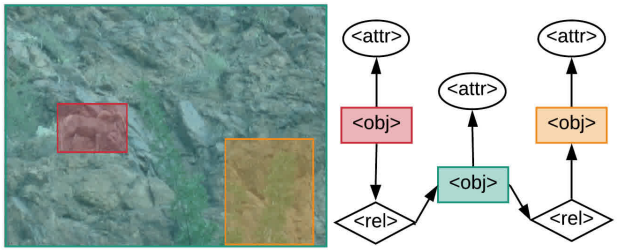
GT: a bicycle is in a living room next to some shelves.
C-BUTD: a bicycle and a bicycle in a room.
Ours: a bike parked in a living room next to a dresser.
<-obj-> <----rel----> <-attr-> <obj> <--rel--> <---obj--->



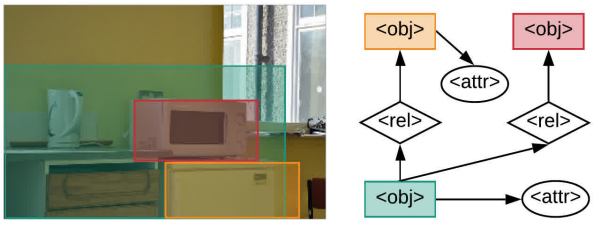
GT: a close up of a baby elephant laying in the grass.
C-BUTD: an elephant is laying down in the grass.
Ours: a large elephant is laying on the ground.
<attr> <---obj--> <-----rel-----> <----obj---->



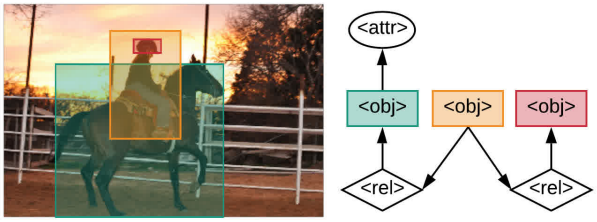
GT: a subway train sits on the track next to a loading platform.
C-BUTD: a train parked on a train platform next to a platform.
Ours: a white train on a track near a loading platform.
<attr> <obj> <rel> <obj> <rel> <--attr--> <--obj-->



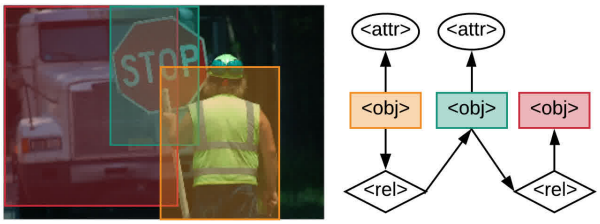
GT: a wild animal on a rocky hill with a few plants.
C-BUTD: a herd of animals standing on top of a lush green hillside.
Ours: a large animal is standing on a rocky hill with tall bushes.
<-attr-> <-obj-> <-------rel-------> <attr> <obj> <rel> <attr> <--obj-->



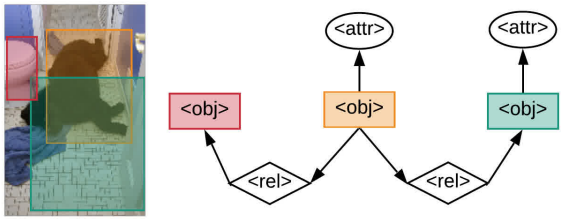
GT: a yellow kitchen contains a small refrigerator and microwave.
C-BUTD: a kitchen with a microwave a refrigerator and a refrigerator.
Ours: a small kitchen with a microwave and a white refrigerator.
<-attr-> <-obj--> <rel> <-----obj-----> <rel> <-attr-> <-----obj----->



GT: the woman wears a helmet as she rides a brown horse.
C-BUTD: a woman in a helmet riding a horse.
Ours: a woman in a hat is riding a brown horse.
<--obj--> <rel> <obj> <--rel--> <--attr--> <-obj->



GT: the street worker is holding a stop sign near a truck.
C-BUTD: a man is standing next to a stop sign.
Ours: a young man stands next to a stop sign in front of a truck.
<-attr-> <obj> <-------rel-------> <attr> <obj><----rel----> <-obj->



GT: the dog is laying on the bathroom floor next to the toilet.
C-BUTD: a cat sitting on the floor next to a toilet.
Ours: a cat laying on top of a bathroom floor next to a toilet.
<obj> <--------rel--------> <----attr----> <obj> <--rel--> <-obj->

Figure 9: Examples for controllable image caption generation conditioning on designated ASGs compared with captions from the groundtruth and the state-of-the-art model C-BUTD [3]. Best viewed in color.

Figure 10: Examples for diverse image caption generation conditioning on sampled ASGs. Best viewed in color.