# Supplementary Material for "Say As You Wish: Fine-grained Control of Image Caption Generation with Abstract Scene Graph"

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

## 1. Automatic ASG Generation

Since the abstract scene graph does not require any semantic labels, we could simply utilise 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. Therefore, we optionally employ a binary relationship classifier to tell whether two objects contain a meaningful relationship.

**Training.** Generating ASGs does not need many annotations to train. For objects, since the ASG only utilizes object locations without semantics, we could employ off-the-shelf region proposal models for object detection, which is possible even with unsupervised training. For attributes, the attribute nodes can be automatically sampled for each object nodes without any recognition. For relationships, the ASGs only require the 'relation existence' binary classification between two objects instead of the more challenging relationship recognition. Such annotations can be automatically extracted from ground-truth captions, which do not need extra external visual relationship datasets.

We train the relationship classifier with annotations in groundtruth ASGs. Instead of recognising 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.

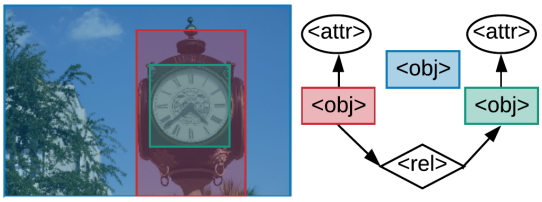
**Inference.** We firstly detect all object bounding boxes in the image and apply SoftNMS [1] to reduce redundancy. Then we utilize the pretrained relationship classifier for each pair of objects. Two objects are considered to contain meaningful relationship if the probability of class 0 is below certain 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 automatically build a full ASG which contains abstract object and relationship nodes.

When users control image captioning with desired ASG, they do not need to tediously generate the ASG from scratch. Users could simply select sub-graphs from the full ASG, or designate any objects as nodes and then let algorithms to automatically complete a ASG based on their preferences. Therefore, ASGs can provide users with more flexibility to control multiple objects, description details than previous controllable image caption works without much controlling burden.

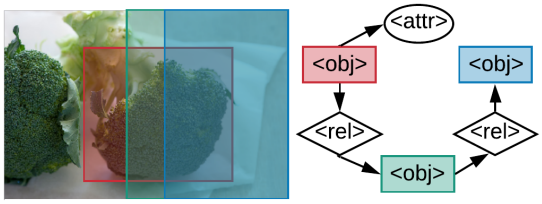
## 2. ASG Dataset Construction

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 region/image $\mathcal{I}$ and its region/image description $y$ are as follows:

1. utilize Stanford scene graph parser [2] 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




GT: an ornate vintage clock with several faces.
(a) sentence parsing error




GT: two vegetables are lying on a napkin on a table.
(b) object grounding error

Figure 1: Two types of errors in the automatic dataset construction (examples from the testing set of MSCOCO).

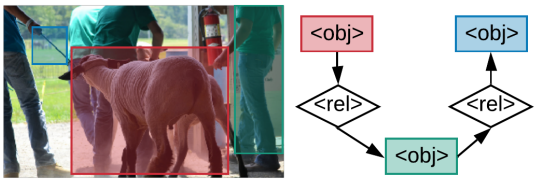
similarity of labels based on WordNet [3] for grounding. For MSCOCO, we can only utilize the semantic similarity of labels for grounding.

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 $\mathcal{G}$.

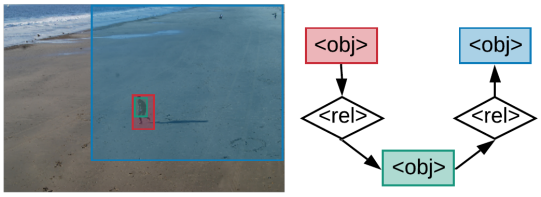
To be noted, since the two datasets are automatically constructed, there mainly exists two types of noises especially for MSCOCO dataset where no object grounding annotations are available. The two types of errors are sentence parsing error and object grounding error as shown in Figure 1. For example, in Figure 1 (a), the attribute "ornate" is mistaken as an object by incorrect sentence parsing; in Figure 1 (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.

## 3. Graph Structure Metric

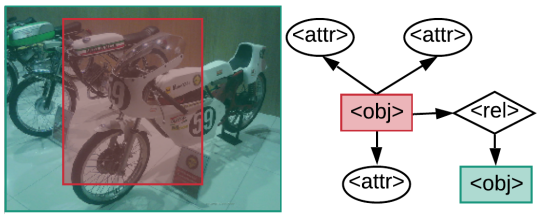
The proposed Graph Structure metric is based on SPICE metric [4]. The SPICE metric parses a sentence into three types of tuples $(o), (o, a)$ and $(o, r, o)$ and measures the semantic alignment of tuples between generated caption and groundtruth captions. However, our Graph Structure metric only cares about the structure alignment which reflects the structure control of ASG without considering the semantic




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

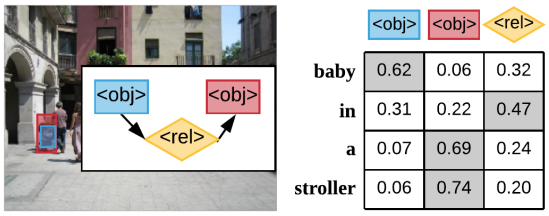
Figure 2: Three types of mistakes in our ASG2Caption model for controllable image caption generation (examples from the testing set of MSCOCO).

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.

## 4. Additional Qualitative Results

In Figure 2, we 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.

Figure 3 shows the learned graph attention over different nodes in the ASG, which demonstrate the effectiveness of




| | <obj> | <obj> | <rel> |
|---|---|---|---|
| **baby** | 0.62 | 0.06 | 0.32 |
| **in** | 0.31 | 0.22 | 0.47 |
| **a** | 0.07 | 0.69 | 0.24 |
| **stroller** | 0.06 | 0.74 | 0.20 |

Figure 3: Learned graph attention over nodes in ASG for each word in image caption generation.

our proposed graph-based attention mechanism.

Figure 4 presents additional examples on controllable image caption generation with designated ASGs. Figure 5 provides more examples on diverse image caption generation with sampled ASGs.

## References

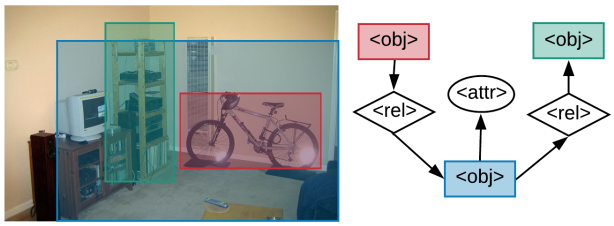
[1] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *ICCV*, pages 5561–5569, 2017. 1

[2] 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. 1

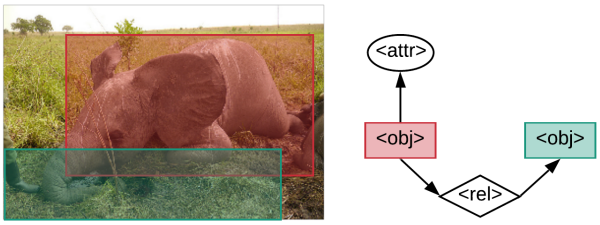
[3] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 2

[4] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, pages 382–398. Springer, 2016. 2

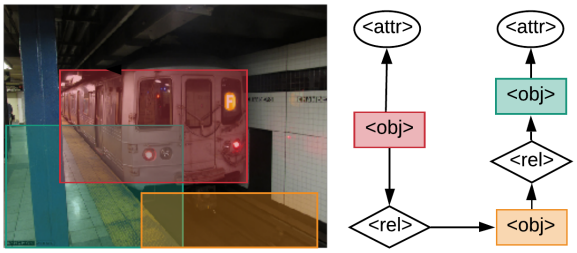
[5] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018. 4



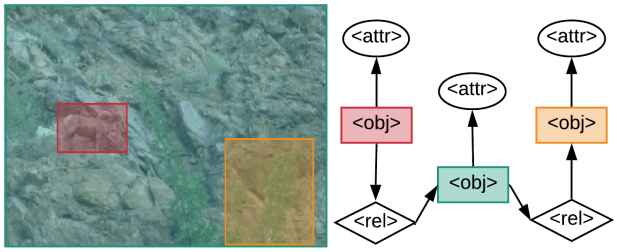

GT: a bicycle is in a living room next to some shelves.
C-BUTD: a bicycle and a bicycle in a room.
Ours: a bike parked in a living room next to a dresser.
<-obj-> <----rel----> <-attr-> <obj> <--rel--> <---obj--->



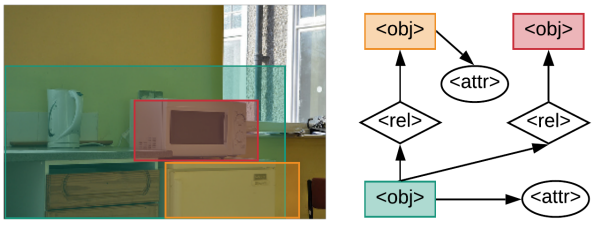

GT: a close up of a baby elephant laying in the grass.
C-BUTD: an elephant is laying down in the grass.
Ours: a large elephant is laying on the ground.
<attr> <---obj--> <-----rel-----> <----obj---->



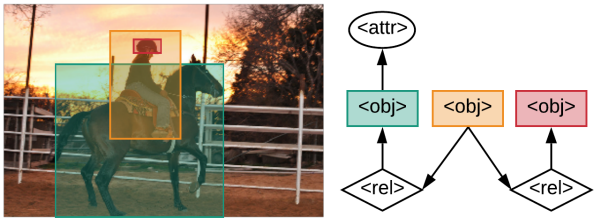

GT: a subway train sits on the track next to a loading platform.
C-BUTD: a train parked on a train platform next to a platform.
Ours: a white train on a track near a loading platform.
<attr> <obj> <rel> <obj> <rel> <--attr--> <--obj-->



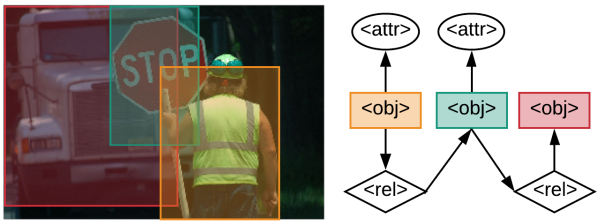

GT: a wild animal on a rocky hill with a few plants.
C-BUTD: a herd of animals standing on top of a lush green hillside.
Ours: a large animal is standing on a rocky hill with tall bushes.
<-attr-> <-obj-> <-------rel-------> <attr> <obj> <rel> <attr> <--obj-->



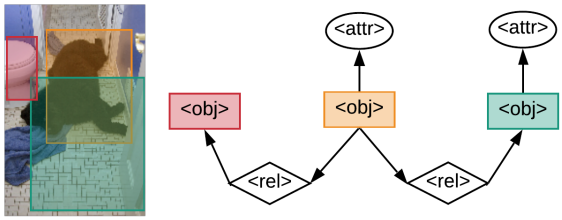

GT: a yellow kitchen contains a small refrigerator and microwave.
C-BUTD: a kitchen with a microwave a refrigerator and a refrigerator.
Ours: a small kitchen with a microwave and a white refrigerator.
<-attr-> <-obj--> <rel> <-----obj-----> <rel> <-attr-> <-----obj----->




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 4: Examples for controllable image caption generation conditioning on groundtruth ASGs compared with captions from the state-of-the-art model C-BUTD [5]. Best viewed in color.


Figure 5: Examples for diverse image caption generation conditioning on sampled ASGs. Best viewed in color.