17th International Conference on Machine Vision Applications (MVA)
Fully Online, July 25–27, 2021.

Saliency based Subject Selection for Diverse Image Captioning

Quoc-An Luong
Department of Informatics, SOKENDAI, Japan
lqan@nii.ac.jp

Duc Minh Vo
The University of Tokyo, Japan
vmduc@nlab.ci.i.u-tokyo.ac.jp

Akihiro Sugimoto
National Institute of Informatics, Japan
sugimoto@nii.ac.jp

Abstract

Image captioning has drawn more and more attention because of its practical usefulness in many multimedia applications. Multiple criteria such as accuracy, detail or diversity exist to evaluate the quality of generated captions. Among them, diversity is the most difficult because for a given image, its multiple captions should be generated while retaining their accuracy. We approach to diverse image captioning by explicitly selecting objects in an image one by one as a subject in generating captions. Our method has three main steps: (1) After generating scene graph of a given image, we first give selection priority to the nodes (namely, subjects) in the scene graph based on the size and visual saliency of objects. (2) With a selected subject, we prune a portion of the scene graph structure that is irrelevant to the subject to have subject-oriented scene graph for accurate captioning. (3) We convert the subject-oriented scene graph into its more sentence-friendly abstract meaning representation (AMR) to generate the caption whose the subject is the selected root. In this way, we can generate captions whose subjects are different from each other, achieving diversity. Our proposed method achieves comparable results with other methods in both diversity and accuracy.

1 Introduction

Image captioning is a challenging joint problem of computer vision and natural language processing, and important in both academic research and real-world applications. Concerning accuracy-based evaluation, recent efforts [2, 8, 12, 13] show remarkable results which are sometimes even better than humans to some extent. We humans, however, possess outstanding capability of flexibly describing a given image with different levels of details and subjects as what we wish. In the step towards humans' ability, generating diverse captions is a crucial requirement for image captioning.

Some methods [6, 7, 9, 10, 23, 25] are able to generate diverse captions with multiple levels of details via, for instance, part of speech [10], set of regions [7], and abstract scene graph [6]. Considering the diversity of image captions, besides their details, how to select different subjects to describe is a key ingredient toward a better mimic humans' level. This obvious observation comes from the fact that we tend to be attracted by visually salient objects in a given image, resulting in subject-oriented descriptions. Visual saliency can thus be an important clue as the subject for generating diverse captions. Nonetheless, yet such diverse subject-oriented image captioning has not been well explored in all of the aforementioned methods.

Early work [2, 12, 26] follows the encoder–decoder framework where the encoder projects an image into latent variables and then the decoder converts the latent variables into a sentence. The latent space, however, works as a black-box intermediate representation, resulting in its inability of controlling the diversity of captions, in particular subjects. The graph representation, on the other hand, is able to represent objects and their relations, opening up a high probability of subject selection to describe. As a result, scene graph [15] is widely used in image captioning to improve in both accuracy and diversity [6, 28, 29]. However, the scene graph structure does not have its root-node, failing to provide a clear subject for the caption. Meanwhile, in natural language processing, abstract meaning representation (AMR) [4] is used to represent the abstract meaning of sentence. AMR has acyclic graph structure with its root-node where the root-node acts as the main concept of the sentence. This characteristic of AMR is suitable for subject-oriented captioning.

We propose a method for selecting subjects one by one from scene graph based on visual saliency, and extract a portion of the scene graph most relevant to the selected subject to convert it into AMR for generating captions. Our main contributions are as follows:

- We propose to select different subjects to describe an image based on visual saliency for generating diverse captions.
- We propose to use only a portion of scene graph relevant to the subject and convert it to AMR for better caption generation. To our best knowledge, this is the first work to bring AMR to the computer vision community.

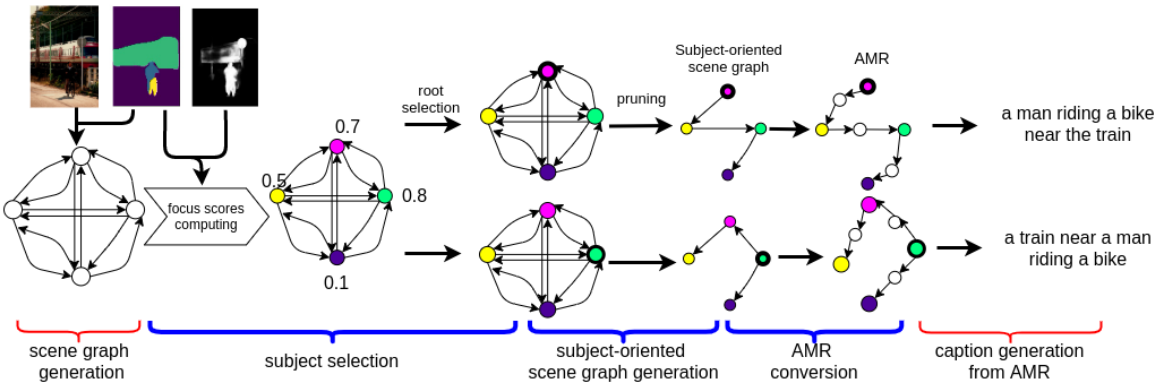

Figure 1: Overall pipeline of our method.

2 Proposed method

Our method consists of three main steps: (1) after the pre-process of scene graph generation, we define subject selection priority based on size and saliency of an object, and select nodes (i.e., subjects) in scene graph one by one depending on the priority. (2) With respect to each selected subject, we prune the scene graph by eliminating irrelevant nodes and edges, producing subject-oriented scene graph. (3) Finally, we convert the subject-oriented scene graph to AMR before generating caption from the AMR.

2.1 Pre-process: scene graph from image

We use Mask-RCNN [14] to detect and segment objects. We then combine the detected regions with our four predefined background regions (top-half, bottom-half, left-half, right-half) to have proposed regions. Next, we apply [27] to the proposed regions to generate scene graph.

The scene graph contains the set of detected objects $O = \{o_1, o_2, o_3, ..., o_n\}$ and the set of directed edges E in the form of (o_i, r_{ij}, o_j) connecting pairs of objects in O. We remark that the number of detected objects and edges varies depending on a given image. Object o_i and edge r_{ij} are assigned with its object confidence and relation confidence, respectively. To ensure the detected objects are relevant to the image, we use a threshold ϵ for the object confidence to keep only objects with high confidence and edges connecting them.

2.2 Subject selection priority based on saliency

The scene graph represents the full visual information obtained in an image. However, humans tend to select an object in an image as a subject and describe it with its relevant information. In principle, any object can be a candidate as a subject, but it will be unnatural. This is because humans tend to focus on visually appealing objects such as "human", "dog" or "car" and describe the image using them as a subject while small and decorative details such as "hat" or "bag" are not often selected as a subject. Inspired by this observation, we select objects one by one as a subject from scene graph based on visual saliency.

We define the selection priority for an object using the focus score as below. We employ two factors for the focus score: object size and visual saliency. The size of an object can be directly obtained as the segmentation mask by Mask-RCNN [14]. Visual saliency of an object is, on the other hand, computed using salient object detection [18]. The focus score $foc(o_i)$ for object o_i is then computed as follows:

$$size(o_i) = \sigma\left(\frac{\sum_{x=1}^{W}\sum_{y=1}^{H} seg(o_i)_{xy}}{W \cdot H}\right), \quad (1)$$

$$sal(o_i) = \frac{\sum_{x=1}^{W}\sum_{y=1}^{H} seg(o_i)_{xy} \cdot mask(o_i)_{xy}}{\sum_{x=1}^{W}\sum_{y=1}^{H} seg(o_i)_{xy}}, \quad (2)$$

$$foc(o_i) = \sqrt{sal(o_i) \cdot size(o_i)}, \quad (3)$$

where $seg(o_i)$ is the segmentation mask of object o_i, $mask(o_i)$ is the saliency mask of o_i, and σ is a normalization function defined as $\sigma(x) = \frac{1}{1+\exp(-(20x-0.2))}$. W and H are the image width and height. As the number of objects and their sizes and saliency values vary from image to image, some small objects may fall into top priority if there are too few objects in an image. Therefore, to avoid such cases, we filter out any object whose focus score is below the average $\frac{\sum_{i=1}^{n} foc(o_i)}{n}$ of focus scores over all the objects. We then select at most top k objects one by one as a subject based on the focus score.

2.3 Subject-oriented scene graph generation

For each selected object as a subject, we generate its subject-oriented scene graph. In order to generate more accurate captions, we had better not use information irrelevant to the subject. To this end, we prune the scene graph so that the pruned scene graph, called *subject-oriented scene graph*, represents only relevant information to the selected object.

The reasonable conditions on the most confident sub-graph of the scene graph with respect to the selected object are (1) the selected object should be the root-node of the sub-graph, (2) the nodes of the sub-graph should be reached from the root-node through the edges with high confidence, and (3) the sub-graph should have the tree structure. The maximum spanning tree rooted with the selected object (node) of the scene graph satisfies these conditions. We thus use the Edmond algorithm [21], an optimal spanning arborescence algorithm, to obtain the subject-oriented scene graph. With at most k top selected objects, we obtain at most k subject-oriented scene graphs, each of which is used to generate a caption. Since the subject of generated captions are different from each other, we are able to achieve diverse captions.

2.4 Conversion to AMR

Considering the structural similarity between scene graph and AMR, we define a set of conversion protocols that transform a scene graph structure to its corresponding AMR structure. After analysing the set of relation labels, we observe that the relation labels can be divided into three groups: (1) *action* (such as "ride" or "hold"), (2) *location* (such as "near" or "behind") and (3) *combination* (such as "walk on" or "hanging from"). We thus introduce a conversion protocol to each group:

- *action*:
  $(o_1, r_a, o_2) \rightarrow (o_1, \text{ARG0-of}, r_a), (r_a, \text{ARG1}, o_2)$.
- *location*:
  $(o_1, r_l, o_2) \rightarrow (o_1, \text{location}, r_l), (r_l, \text{op1}, o_2)$.
- *combination*:
  $(o_1, r_c, o_2) \rightarrow (o_1, \text{ARG0-of}, r_c), (r_c, \text{location}, o_2)$.

o_1, and o_2 are subject and object of the *subject-object* relation. r_a, r_l, r_c are the relation in each group. ARG0-of, ARG1, location, op1 are the semantic relations defined in AMR [4]. Applying the conversion protocols to each relation triplet (o_1, r, o_2) in the scene graph, we obtain its corresponding AMR structure.

2.5 Caption generation

We employ [20] to generate the caption from an AMR. This is because [20] is one of the state-of-the-arts and achieves high performance on the standard LDC2015E86 AMR-to-text benchmark [1]. [20] has the encoder–decoder architecture which encodes the graph structure of an AMR and decodes it to the caption.

3 Experiments

3.1 Experimental setup

We evaluate our method on the COCO dataset [17]. We empirically set the threshold $\epsilon = 0.2$ when generating scene graphs from images. For each image, we generate captions by selecting at most $k = 3$ different subjects in scene graph.

To evaluate the accuracy of generated captions, we use BLEU [19], METEOR [5], ROUGE-L [16], CIDEr [22], and SPICE [3] denoted by B, M, R, C, and S, respectively. We compute the average and the maximum of those scores of generated captions for each image and then report their averages over the dataset. We also employ self-CIDEr [24] to evaluate the diversity of generated captions. We follow standard implementations when computing all the evaluation metrics.

3.2 Comparison with other methods

We first evaluated the accuracy of generated captions (Table 1). For a fair comparison, we solely compare our method with an unsupervised image captioning method (denoted by unsupv) [11]. This is because other methods apply supervised learning to fit the distribution of ground-truth dataset while our method does not; therefore, comparison with supervised learning methods is not fair.

Table 1 reveals that the average accuracy obtained by our method are lower than that by unsupv [11]. The lower performance comes from the design of our method. Specifically, due to generating multiple captions for each given image, some of them do not match the ground-truth captions. When comparing our maximum accuracy scores against the scores by unsupv [11], however, we observe that our method achieves comparable scores on the non-n-gram metrics (i.e., METEOR, ROUGE-L, and SPICE) while poorly performs on the n-gram metrics (i.e., BLEU and CIDEr). This is because our method strictly relies on the vocabulary of the scene graph detection and thus the number of vocabularies in our method is limited to that of scene graph detection. More precisely, the vocabulary set of scene graph detection [27] consists of 200 words which is extremely smaller than that of COCO dataset [17] (10010 words), leading to our generated captions mismatch with the ground-truth COCO dataset.

We also evaluated the diversity of generated captions through comparing with ASG [6] and Len-Ctrl [9]. Table 2 shows self-CIDEr scores obtained by our and compared methods. We see that even without any involving learning process, our method is on a par with other (learning-based) methods. It is worth noting that our method significantly suffers from the lack of diversity in vocabulary. We thus conclude that our method is capable of generating diverse captions thanks to our subject selection.

Figure 2 illustrates examples of our generated captions, showing diversity and a variety of subjects of the captions. We see that images with more objects provide more captions since we can select more subjects. We also see that the generated captions preserve all the objects connected to the subject. However, if the scene graph has no out-going edge from a selected subject, the subject-oriented scene graph becomes disconnected and the caption results in just the subject.

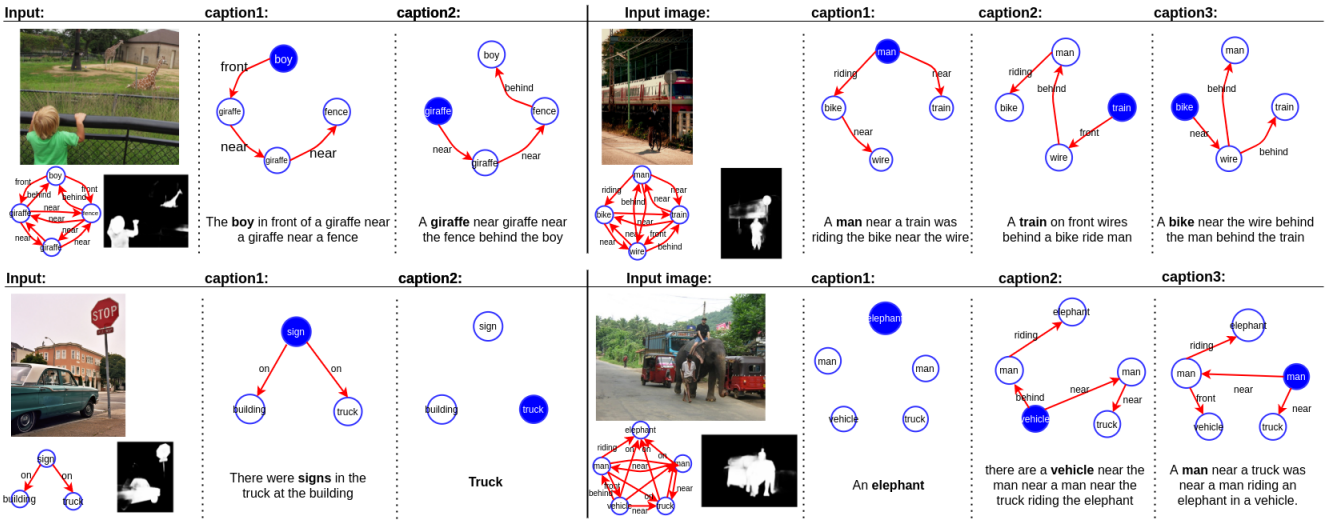

Figure 2: Examples of generated captions by our method on the COCO dataset. Blue circles mean selected nodes as subjects. Images with few saliency objects (left) provide fewer captions while images with more saliency objects (right) do more captions.

Table 1: Comparison of accuracy with unsupervised method [11].

	B1	B4	M	R	C	S
unsupv [11]	**41.0**	**5.6**	**12.4**	**28.7**	**28.6**	8.1
Ours (avg)	17.9	0.1	8.4	18.5	13.4	6.1
Ours (max)	25.4	0.2	10.7	23.6	18.1	**8.4**

Table 2: Comparison of diversity with other methods.

	self-CIDEr
ASG [6]	0.84
Len-Ctrl [9]	0.76
Human baseline	**0.90**
Ours	0.77

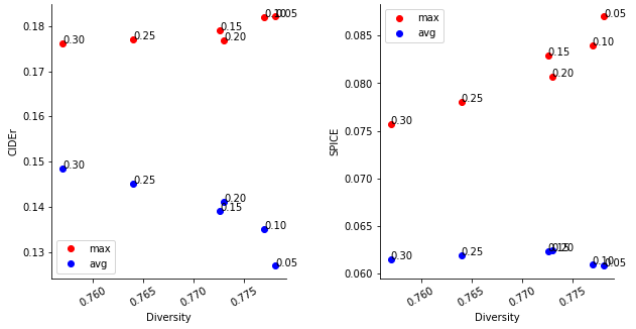

Figure 3: Accuracy and diversity under different object confidence thresholds.

3.3 Impact of threshold ϵ on diversity and accuracy

It is clear that incorporating more accurately detected objects into scene graph leads more accurate captions yet in less diversity while incorporating less accurately detected objects directs captions into the opposite direction. We therefore set up experiments to figure out the impact of object confidence score on the accuracy–diversity trade–off. To this end, we change the threshold ϵ by 0.05 from 0.05 to 0.3 when generating scene graph.

Figure 3 shows the behaviors of accuracy (on CIDEr and SPICE metrics) and diversity (self-CIDEr metric) as the threshold changes. As can be seen in Fig. 3, the lower the threshold is the more diverse the generated captions are. This is reasonable as we have more objects in the scene graph to be described in the captions. Meanwhile, with the threshold decreasing, the average accuracy decreases as more false detection is allowed into the captions. However, a low threshold also generate more captions where an accurate caption is potentially included in them. This leads to the increase of maximum accuracy even for lower thresholds.

4 Conclusion

We proposed to select subjects to describe one by one based on saliency for diverse image captioning. We also proposed to use only a portion of scene graph relevant to the selected subject for better captioning with the help of AMR. Different from other methods using the whole scene graph for captioning, our approach brings gains in both diversity and accuracy. Indeed, even without involving any learning process, our method achieves comparable results in both accuracy and diversity metrics against our compared methods. Moreover, our method is able to specify subjects to describe, which brings us to comprehensive captioning.

References

[1] *Abstract Meaning Representation benchmark*. https://amr.isi.edu/download.html.

[2] Peter Anderson et al. "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering". In: *CVPR*. 2018.

[3] Peter Anderson et al. "SPICE: Semantic Propositional Image Caption Evaluation". In: *ECCV*. 2016.

[4] Laura Banarescu et al. "Abstract Meaning Representation for Sembanking". In: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. 2013.

[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*. 2005.

[6] Shizhe Chen et al. "Say As You Wish: Fine-Grained Control of Image Caption Generation With Abstract Scene Graphs". In: *CVPR*. 2020.

[7] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions". In: *CVPR*. 2019.

[8] Marcella Cornia et al. "Meshed-Memory Transformer for Image Captioning". In: *CVPR*. 2020.

[9] Chaorui Deng et al. "Length-Controllable Image Captioning". In: *ECCV*. 2020.

[10] Aditya Deshpande et al. "Fast, Diverse and Accurate Image Captioning Guided by Part-Of-Speech". In: *CVPR*. 2019.

[11] Yang Feng et al. "Unsupervised Image Captioning". In: *CVPR*. 2019.

[12] Junlong Gao et al. "Self-Critical N-Step Training for Image Captioning". In: *CVPR*. 2019.

[13] Longteng Guo et al. "Normalized and Geometry-Aware Self-Attention Network for Image Captioning". In: *CVPR*. 2020.

[14] K. He et al. "Mask R-CNN". In: *ICCV*. 2017.

[15] Justin Johnson et al. "Image Retrieval Using Scene Graphs". In: *CVPR*. 2015.

[16] Chin-Yew Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. 2004.

[17] Tsung-Yi Lin et al. "Microsoft COCO: Common Objects in Context". In: *ECCV*. Ed. by David Fleet et al. Springer International Publishing, 2014.

[18] Jiang-Jiang Liu et al. "A Simple Pooling-Based Design for Real-Time Salient Object Detection". In: *CVPR*. 2019.

[19] Kishore Papineni et al. "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *ACL*. 2002.

[20] Linfeng Song et al. "A Graph-to-Sequence Model for AMR-to-Text Generation". In: *ACL*. 2018.

[21] Robert Endre Tarjan. "Finding optimum branchings." In: *Networks* (1977).

[22] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. "CIDEr: Consensus-Based Image Description Evaluation". In: *CVPR*. 2015.

[23] Jiuniu Wang et al. "Compare and Reweight: Distinctive Image Captioning Using Similar Images Sets". In: *ECCV*. Ed. by Andrea Vedaldi et al. 2020.

[24] Qingzhong Wang and Antoni B. Chan. "Describing Like Humans: On Diversity in Image Captioning". In: *CVPR*. 2019.

[25] Zeyu Wang et al. "Towards Unique and Informative Captioning of Images". In: *ECCV*. Ed. by Andrea Vedaldi et al. 2020.

[26] Kelvin Xu et al. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". In: *ICML*. 2015.

[27] Jianwei Yang et al. "Graph R-CNN for Scene Graph Generation". In: *ECCV*. 2018.

[28] Xu Yang et al. "Auto-Encoding Scene Graphs for Image Captioning". In: *CVPR*. 2019.

[29] Yiwu Zhong et al. "Comprehensive Image Captioning via Scene Graph Decomposition". In: *ECCV*. 2020.