On the Role of Scene Graphs in Image Captioning

Dalin Wang **Daniel Beck** **Trevor Cohn**
School of Computing and Information Systems
The University of Melbourne
dalinw@student.unimelb.edu.au
{d.beck,t.cohn}@unimelb.edu.au

Abstract

Scene graphs represent semantic information in images, which can help image captioning system to produce more descriptive outputs versus using only the image as context. Recent captioning approaches rely on ad-hoc approaches to obtain graphs for images. However, those graphs introduce noise and it is unclear the effect of parser errors on captioning accuracy. In this work, we investigate to what extent scene graphs can help image captioning. Our results show that a state-of-the-art scene graph parser can boost performance almost as much as the ground truth graphs, showing that the bottleneck currently resides more on the captioning models than on the performance of the scene graph parser.

1 Introduction

The task of automatically recognizing and describing visual scenes in the real world, normally referred to as image captioning, is a long standing problem in computer vision and computational linguistics. Previously proposed methods based on deep neural networks have demonstrated convincing results in this task, (Xu et al., 2015; Lu et al., 2018; Anderson et al., 2018; Lu et al., 2017; Fu et al., 2017; Ren et al., 2017) yet they often produce dry and simplistic captions, which lack descriptive depth and omit key relations between objects in the scene. Incorporating complex visual relations knowledge between objects in the form of scene graphs has the potential to improve captioning systems beyond current limitations.

Scene graphs, such as the ones present in the Visual Genome dataset (Krishna et al., 2017), can be used to incorporate external knowledge into images. Because of the structured abstraction and greater semantic representation capacity than purely image features, they have the potential to improve image captioning, as well as other downstream tasks that rely on visual components. This has led to the development of many parsing algorithms for scene graphs (Li et al., 2018, 2017; Xu et al., 2017; Dai et al., 2017; Yu et al., 2017). Simultaneously, recent work also aimed at incorporating scene graphs into captioning systems, with promising results (Yao et al., 2018; Xu et al., 2019). However, these previous work still rely on ad-hoc scene graph parsers, raising the question of how captioning systems behave under potential parsing errors.

In this work, we aim at answering the following question: "to what degree scene graphs contribute to the performance of image captioning systems?". In order to answer this question we provide two contributions: 1) we investigate the performance of incorporating scene graphs generated by a state-of-the-art scene graph parser (Li et al., 2018) into a well-established image captioning framework (Anderson et al., 2018); and 2) we provide an upper bound on the performance by comparative experiments with *ground truth* graphs. Our results show that scene graphs can be used to boost performance of image captioning, and scene graphs generated by state-of-art scene graph parser, though still limited in the number of objects and relations categories, is not far below the ground-truth graphs, in terms of standard image captioning metrics.

2 Methods

Our architecture, inspired by Anderson et al. (2018) and shown in Figure 1, assumes an off-the-shelf scene graph parser. To improve performance, we also incorporate information from the original image through a set of region features obtained through an object detection model. Note we experiment with each set of features in isolation in

Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN), pages 29–34
Hong Kong, China, November 3, 2019.

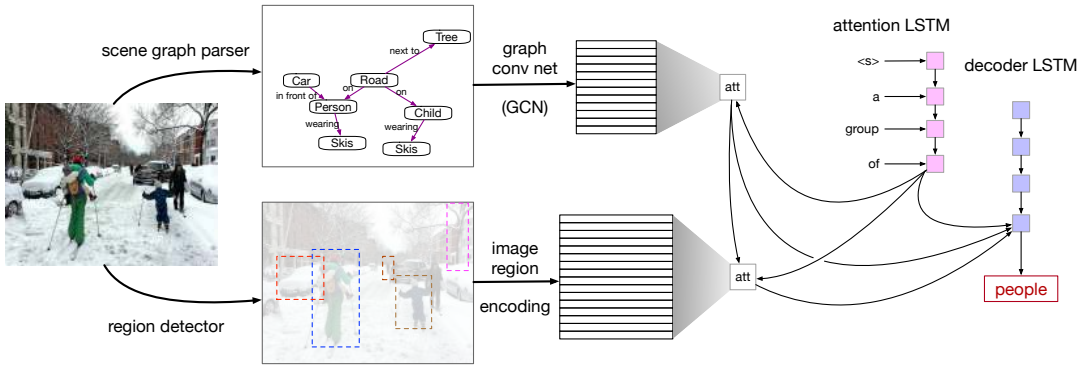

Figure 1: Overview of our architecture for image captioning.

Section 3.1. Given those inputs, our model consists a scene graph encoder, an LSTM-based attention module and another LSTM as the decoder.

2.1 Scene Graph Encoder

The scene graph is represented as a set of *node* embeddings which are then updated into contextual hidden vectors using a Graph Convolutional Network (Kipf and Welling, 2017, GCN). In particular, we employ the GCN version proposed by Marcheggiani and Titov (2017), who incorporate directions and edge labels. We treat each relation and object in the scene graph as nodes, which are then connected with five different types of edges.[1] Since we assume scene graphs are obtained by a parser, they may contain noise in the form of faulty or nugatory connections. To mitigate the influence of parsing errors, we allow edge-wise gating so the network learns to prune those connections. We refer to Marcheggiani and Titov (2017) for details of their GCN architecture.

2.2 Attention LSTM

The Attention LSTM keeps track of contextual information from the inputs and incorporates information from the decoder. At each time step t, the Attention LSTM takes in contextual information by concatenating the previous hidden state of the Decoder LSTM, the mean-pooled region-level image features, the mean-pooled scene graph node features from the GCN and the previous generated word representation: $\mathbf{x}_t^1 = [\mathbf{h}_{t-1}^2, \overline{\mathbf{v}}, \overline{\mathbf{f}}, W_e \mathbf{u}_t]$ where W_e is the word embedding matrix for vocabulary Σ and $\mathbf{u}_t$ is the one-hot encoding of the word at time step t. Given the hidden state of the Attention LSTM h_t^1, we generate cascaded attention features, first over scene graph features, and then we concatenate the attention weighted scene graph features with the hidden state of the Attention LSTM to attend over region-level image features. Here, we only show the second attention step over region-level image features as they are identical procedures except for the input:

$$b_{i,t} = \mathbf{w}_b^T \text{ReLU}(W_{fb}\mathbf{v}_i + W_{hb}[\mathbf{h}_t^1, \hat{\mathbf{f}}_t])$$

$$\beta_t = \text{softmax}(\mathbf{b}_t); \quad \hat{\mathbf{v}}_t = \sum_{i=1}^{N_v} \beta_{i,t}\mathbf{v}_i$$

where $\mathbf{w}_b^T \in \mathbb{R}^H, W_{fb} \in \mathbb{R}^{H \times D_f}, W_{hb} \in \mathbb{R}^{H \times H}$ are learnable weights. $\hat{\mathbf{v}}_t$ and $\hat{\mathbf{f}}_t$ are the attention weighted image features and scene graph features respectively.

2.3 Decoder LSTM

The inputs to the Decoder LSTM consist of the previous hidden state from the Attention LSTM layer, attention weighted scene graph node features, and attention weighted image features. $x_t^2 = [\mathbf{h}_t^1, \hat{\mathbf{f}}_t, \hat{\mathbf{v}}_t]$ Using the notation $y_{1:T}$ to refer to a sequence of words $(y_1, ..., y_T)$ at each time step t, the conditional distribution over possible output words is given by: $p(y_t|y_{1:t-1}) = \text{softmax}(W_p \mathbf{h}_t^2 + \mathbf{b}_p)$ where $W_p \in \mathbb{R}^{|\Sigma| \times H}$ and $\mathbf{b}_p \in \mathbb{R}^{|\Sigma|}$ are learned weights and biases.

2.4 Training and Inference

Given a target ground truth sequence $y_{1:T}^*$ and a captioning model with parameters θ, we minimize the standard cross entropy loss. At inference time, we use beam search with a beam size of 5 and apply length normalization (Wu et al., 2016).

[1] We use the following types: *subj* indicates the edge between a subject and predicate, *obj* denotes the edge between a predicate and an object, *subj'* and *obj'*, their corresponding reverse edges, and lastly, *self*, which denotes a self loop.

3 Experiments

Datasets **MS-COCO**, (Lin et al., 2014) is the most popular benchmark for image captioning, which contains 82,783 training images and 40,504 validation images, with five human-annotated descriptions per image. As the annotations of the official testing set are not publicly available, we follow the widely used Kaparthy split (Karpathy and Fei-Fei, 2017), and take 113,287 images for training, 5K for validation, and 5K for testing. We convert all the descriptions in training set to lower case and discard rare words which occur less than five times, resulting in a vocabulary with 10,201 unique words. For the oracle experiments, we take a subset of MS-COCO that intersects with Visual Genome (Krishna et al., 2017) to obtain the ground truth scene graphs. The resulting dataset (henceforth, *MS-COCO-GT*) contains 33,569 training, 2,108s validation, and 2,116 test images respectively.

Preprocessing The scene graphs are obtained by a state-of-the-art parser: a pre-trained Factorizable-Net trained on MSDN split (Li et al., 2017), which is a cleaner version of the Visual Genome[2] that consists of 150 object categories and 50 relationship categories. Notice that the number of object categories and relationships are much smaller than the actual number of objects and relationships in the Visual Genome dataset. All the predicted objects are associated with a set of bound box coordinates. The region-level image features[3] are obtained from Faster-RCNN (Ren et al., 2017), which is also trained on Visual Genome, using 1,600 object classes and 400 attributes classes.

Implementation Our models are trained with AdamMax optimizer (Kingma and Ba, 2015). We set the initial learning rate as 0.001 with a mini-batch size as 256. We set the maximum number of epochs to be 100 with early stopping mechanism.[4] During inference, we set the beam width to 5. Each word in the sentence is represented as a one-hot vector, and each word embedding is a 1,024-dimensional vector. For each image, we have $K = 36$ region features with bounding box coordinates from Faster-RCNN. Each region-level image feature is represented as a 2,048-dimensional vector, and we concatenate the bounding box coordinates to each of the region-level image features. The dimension of the hidden layer in each LSTM and GCN layer is set to 1,024. We use two GCN layers in all our experiments.

[2]The MSDN split might contain training instances that overlap with the Karpathy split

[3]These regions are different to those from the scene graph. To help the model learn to match regions, the inputs to attention include bounding box coordinates.

[4]We stop training if the CIDEr score does not improve for 10 epochs, and we reduce the learning by 20 percent if the CIDEr score does not improve for 5 epochs.

	B	M	R	C	S
No edge-wise gating					
I	34.1	26.5	55.5	108.0	19.9
G	22.8	20.6	46.7	66.3	13.5
I+G	34.2	26.5	55.7	108.2	20.1
With edge-wise gating					
G	22.9	21.1	47.5	70.7	14.0
I+G	**34.5**	**26.8**	**55.9**	**108.6**	**20.3**

Table 1: Results on the full MS-COCO dataset. "I", "G" and "I+G" correspond to models using image features only, scene graphs only and both, respectively. "B", "M", "R", "C" and "S" correspond to BLEU, METEOR, ROUGE, CIDEr and SPICE (higher is better).

Evaluation We employ standard automatic evaluation metrics including BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), ROUGE (Lin, 2004), CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016), and we use the coco-caption tool[5] to obtain the scores.

3.1 Quantitative Results and Analysis

Table 1 shows the performances of our models against baseline models whose architecture is based on Bottom-up Top-down Attention model (Anderson et al., 2018). Overall, our proposed model incorporating scene graph features achieves better results across all evaluation metrics, compared to image features only or graph features only. The results show that our model can learn to exploit the relational information in scene graphs and effectively integrate those with image features. Moreover, the results also demonstrate the effectiveness of edge-wise gating in pruning noisy scene graph features.

We also conduct experiments comparing Factorizable-Net generated scene graph with ground-truth scene graph, as shown in Table 2. As expected, the results show that the performance is

[5]https://github.com/tylin/coco-caption

GT: a cop riding a motorcycle next to a white van
Image: a police officer riding a motorcycle on a city street
Graph: a man riding on the back of a motorcycle down a street
I + G: a man riding a motorcycle down a city street in front of a white bus

GT: the baby is playing with the phone in the park
Image: a little girl is holding a cell phone
Graph: a woman sitting on a bench with a cell phone
I + G: a little girl is holding a cell phone in a field of grass in a park

Figure 2: Caption generation results on COCO dataset. All results are generated by models trained on the full version of Karpathy split, and all graph features are processed by GCN with edge-wise gating. 1) Ground Truth(GT) 2) Image features only(Image) 3) Graph features only(Graph) 4) Ours: Image features plus graph features (I + G)

	B	M	R	C	S
I	32.0	25.6	54.3	102.2	19.0
G (pred)	17.4	16.5	41.3	49.5	10.6
G (truth)	18.4	17.9	42.5	50.8	11.2
I+G (pred)	32.2	25.8	54.4	103.4	19.1
I+G (truth)	**32.5**	**26.1**	**54.8**	**105.2**	**19.5**

Table 2: Results on the MS-COCO-GT dataset. "G (pred)" refers to the parsed scene graphs from Factorizable-Net while "G (truth)" corresponds to the ground truth graphs obtained from Visual Genome.

better with ground-truth scene graph. Notably the SPICE score, which measures the semantic correlation between generated captions and ground truth captions, improved by 2.1%, since there are considerably more types of objects, relations and attributes present in the ground-truth scene graphs. Overall, the results show the potential of incorporating automatically generated scene graph features for the captioning system, and we argue with better scene graph parser trained on more objects, relations and attributes categories, the captioning system should provide additional improvements.

Compared to a recent image captioning paper[6] (Li and Jiang, 2019) using scene-graph features, our results are superior, demonstrating the effectiveness of our model. Moreover, compared to a state-of-art image captioning system (Yu et al., 2019),[7] our scores are inferior, as we do not apply scheduled sampling, reinforcement learning, transformer cell or ensemble predictions, which have all been proven to improve the scores significantly. However, our method of incorporating scene-graph features is orthogonal to the state-of-art methods.

3.2 Qualitative Results and Analysis

Figure 2 shows some generated captions by different approaches trained on the full Karpathy split of MS-COCO dataset. We can see that all approaches can produce sensible captions describing the image content. However, our approach of incorporating scene graph features and image features can generate more descriptive captions that more closely narrate the underlying relations in the image. In the first example, our model correctly predicts that the motercycle is in front of the white van while the image-only model misses this relational detail. On the other hand, purely graph features sometimes introduce noise. As shown in the second example, the graph-only model mistakes the little girl in a park as a woman on a bench, whereas the image features in our model helps disambiguate faulty graph features.

4 Conclusion

We have presented a novel image captioning framework that incorporates scene graph features extracted from state-of-art scene graph parser Factorizable-Net. Particularly, we investigate the problem of integrating relation-aware scene graph features encoded by Graph Convolution with region-level image features to boost image captioning performance. Extensive experiments conducted on MSCOCO image captioning dataset has shown the effectiveness of our method. In the future, we want to experiment with building an

[6]The Hierarchical Attention Model incorporating scene-graph features reports scores: Bleu4 33.8, METEOR 26.2, ROUGE 54.9, CIDEr 110.3, SPICE 19.8

[7]This transformer-based captioning system reports scores: Bleu4 40.4, METEOR 29.4, ROUGE 59.6, CIDEr 130.0.

end-to-end multi-task framework that jointly predicts visual relations and captions.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: semantic propositional image caption evaluation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, pages 382–398.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086.

Bo Dai, Yuqi Zhang, and Dahua Lin. 2017. Detecting visual relationships with deep relational networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3298–3308.

Kun Fu, Junqi Jin, Runpeng Cui, Fei Sha, and Changshui Zhang. 2017. Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(12):2321–2334.

Andrej Karpathy and Li Fei-Fei. 2017. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):664–676.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

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

Alon Lavie and Abhaya Agarwal. 2007. METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation, WMT@ACL 2007, Prague, Czech Republic, June 23, 2007*, pages 228–231.

Xiangyang Li and Shuqiang Jiang. 2019. Know more say less: Image captioning based on scene graphs. *IEEE Trans. Multimedia*, 21(8):2117–2130.

Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, and Xiaogang Wang. 2018. Factorizable net: An efficient subgraph-based framework for scene graph generation. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part I*, pages 346–363.

Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. 2017. Scene graph generation from objects, phrases and region captions. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1270–1279.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3242–3250.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural baby talk. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7219–7228.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1506–1515.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA.*, pages 311–318.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, ukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3097–3106.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 2048–2057.

Ning Xu, An-An Liu, Jing Liu, Weizhi Nie, and Yuting Su. 2019. Scene graph captioner: Image captioning based on structural visual representation. *J. Visual Communication and Image Representation*, 58:477–485.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*, pages 711–727.

Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. 2019. Multimodal transformer with multi-view visual representation for image captioning. *CoRR*, abs/1905.07841.

Ruichi Yu, Ang Li, Vlad I. Morariu, and Larry S. Davis. 2017. Visual relationship detection with internal and external linguistic knowledge distillation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1068–1076.