Topic Scene Graph Generation by Attention Distillation from Caption

Wenbin Wang[1,2], Ruiping Wang[1,2,3], Xilin Chen[1,2]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
[3]Beijing Academy of Artificial Intelligence, Beijing, 100084, China
wenbin.wang@vipl.ict.ac.cn, {wangruiping, xlchen}@ict.ac.cn

Table 1. The details of the parameters settings.

Params.	Meanings	Values
d_v	dimension of the visual features $\boldsymbol{v}_i$	4,096
d_u	dimension of the union visual features $\boldsymbol{v}_{ij}$	512
d_l	dimension of the transformed features	1,024
d_h	dimension of the hidden states in UD model [1]	512
d_a	dimension of the attention embedding in UD model	1,024
d_e	dimension of the word embedding in UD model	1,024
d_{tr}	dimension of the input/output inside the Transformer [6]	512
d_s	dimension of the query/key of the Transformer	512
d_{sem}	dimension of the semantic embedding	200
T_R	the maximum length of the relational captions	18
T_C	the maximum length of the image caption	18

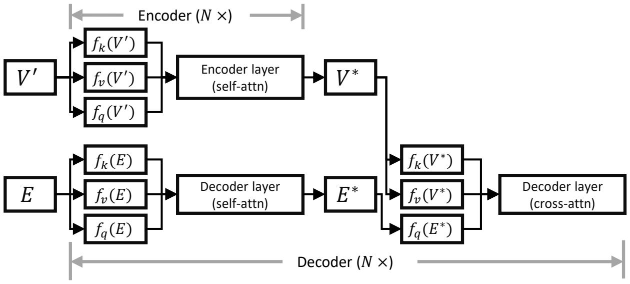


Figure 1. The details of the Transformer.

1. Implementation Details

The parameters settings mentioned in the main paper are shown in Table 1.

In Figure 1 we show the details of the Transformer. The Transformer has 6 encoder layers and 6 decoder layers, both of which have one attention head for simplicity. Only the attentions from the last decoder layer are collected. It is noted that when generating the relational captions, we still feed all of the object transformed features (*i.e.*, $\boldsymbol{V}' \in \mathbb{R}^{d_l \times n}$) into the encoder, but only select the subject and object features from the output of the encoder and their union feature to assemble the $\boldsymbol{V}^* \in \mathbb{R}^{d_{tr} \times 3}$ as the input of the decoder.

Table 2. Image captioning results. B1, B4, M, R, C, S denote the BLEU-1, BLEU-4, METEOR, ROUGE-L, CIDEr-D, and SPICE respectively. "-ICRC" denotes that the model is trained with image captions and relational captions.

Model	λ	B1	B4	M	R	C	S
UD [1]	-	69.8	29.6	25.0	52.3	94.1	18.0
UD-ICRC	0.1	**71.1**	**30.4**	**25.1**	**52.6**	**95.3**	**18.3**
	0.3	70.7	30.0	24.9	52.5	94.6	18.1
	0.7	70.5	30.1	25.0	52.4	94.8	18.2
	1.0	71.0	30.0	24.8	52.5	93.5	17.9
	3.0	70.0	29.2	24.3	51.8	91.4	17.5
	5.0	69.6	28.9	23.9	51.3	89.6	17.2
	7.0	69.0	28.3	23.6	50.9	87.4	16.8
	10.0	68.7	27.4	23.2	50.3	84.8	16.7
Transformer [6]	-	68.8	26.8	23.5	50.4	85.6	17.3
Transformer-ICRC	0.7	**70.3**	**28.6**	**24.4**	**51.7**	**91.5**	**18.0**

The scene graph training is divided into two stages. In the first stage we train the captioning module for 30 epochs with the Adam [3] optimizer and the Noam policy [6]. The initial learning rate is set as 5e-4 and the batch size is 12. In the second stage, the attention alignment module is trained with the SGD optimizer for 12 epochs during which the parameters of captioning module are frozen. The initial learning rate is set as 2e-2. We use the ground truth objects (bounding boxes and categories) for training and evaluation to eliminate the interference of the error from detector.

2. Experiments about Topic Scene Graph

In Table 2 and Table 3, we provide more results as the λ changes. From Table 2, it further proves the conclusion mentioned in the main paper that mixed training actually brings benefit to the image captioning, but as the λ increases, this benefit will slightly drop. When the λ is larger than 1, it will be harmful to the performance. In Table 3, as the λ increases, the UD-ICRC roughly performs better on the image level metrics (mAP, METEOR, and Img-Lv. Recall), and surpasses the UD-RC baseline when λ is greater than 0.7. When the λ is larger than 7.0, the performance begins to drop. As for the important relationship recall metrics, the performance reaches a peak when λ is 0.1.

Apart from using the sentence likelihoods for sorting the relationships, we additionally use the randomization strategy, *i.e.*, K relationships are randomly chosen for evaluation. The results are presented in Table 4. It is observed that using sentence likelihoods is even worse (for TriLSTM) or just slightly better (for UD-RC) than using the randomization strategy. It suggests that the sentence likelihood is not suitable for representing the importance of the relationships, which further demonstrates the value of our predicted importance scores.

We provide more experiment results about different configurations for topic scene graph based on the Up-Down [1] model in Table 5. It is observe that the max pooling function works well under any configuration of the input feature and the application of masking. The mean pooling only works when configured with SOU or SOUS input features and masking the non-noun words (index (9) and (17) in Table 5). It is interesting that max pooling achieves the best result configured with no masking while mean pooling is more compatible with masking. We think that the mean pooling function reduces the attention about the objects mentioned in the caption, and therefore masking the non-noun objects helps correct the attention on these objects again. As for the input feature, SOU brings the most significant improvement as it contains both the visual features of the subjects and objects, and the relative spatial configuration of two objects, which are useful information for estimating their importance. The semantic embedding brings less improvement when the attention alignment module is applied, while the improvement is more significant for the upper bound model. We think it is mainly because the attention is not so accurate and the categories of detected objects do not absolutely match with the objects mentioned in the captions. When it comes to the important relationships with ground truth labels, the semantic embedding is helpful for the model as a type of prior knowledge, *e.g.*, usually the object of the category *person* is important which catches attention and is mentioned in the caption.

We provide more qualitative results in Figure 2. In these examples, the words of the captions are correctly grounded to the objects, and the top 5 relationships are indeed highly relative to the major events of the images, while the top 5 relationships from motifs [7] are too trivial and not image-specific.

Some failure cases are shown in Figure 3. They can be roughly categorized into three types: (1) The image captioner fails to ground some words to the appropriate object regions. As a result, some unimportant or wrong relationships emerge in top 5, *e.g.*, the *wave* is wrongly grounded to the *sky* in the first example, resulting in the relationship *the sky is above the water* which is not mentioned in the caption, and in the fourth example, the grounded *bear* and the *chair* does not match because of the confusion when multiple instances of a category emerge, resulting in the wrong relationship *bear-0 is sitting on the chair-4*. (2) The detector fails to detect some objects. In the second example, the missing *zebra* on the right side makes the relationships about this zebra are lost. (3) The detector provides highly overlapped bounding boxes with similar semantics and the context of the image is complicated. As a result, the image captioner in the third example repeats the same pattern and the model produces some repeated relationships.

3. Topic Scene Graph for Retrieval

As the topic scene graph provides relationships relevant to the major events in an image, we further conduct the image retrieval experiment to prove it. We use the classic image-text matching model SCAN [4]. 1,000 images are randomly chosen from the test set, and their top 1 or 5 relationships are collected as query for retrieving correct target images. We use 3 different strategies for retrieval, (1) using the top 1 relationship (Top-1), (2) using the top 5 relationships (Top-5), and (3) using the single sentence by connecting the top 5 relationships (Top-5-CON). The recall (R@K, K is 1, 5, 10) and the median rank of the correctly retrieved images [2] are used as the metrics. We run through this process 3 times and report the average results. The results are shown in Table 6. It's shown that our baseline outperforms the TriLSTM, and the attention alignment module successfully makes the model surpass the baseline significantly. In addition, we provide more qualitative results in Figure 4. The retrieval results using the top 2 relationships for every image are demonstrated respectively. In these images, the major events can be decomposed into multiple relationships. If one directly use the original image or traditional scene graph to retrieve similar images, the results may be not the desired ones. The proposed topic scene graph provides fine-grained descriptions of major events and makes it possible to designate the target content to be retrieved. These results suggest that the top relationships given by the topic scene graph are indeed much more relevant to the major events. What's more, the topic scene graph makes the fine-grained controllable retrieval feasible.

4. Topic Scene Graph for Image Captioning

We evaluate the topic scene graph on another downstream task, image captioning. The model in [5] is reimplemented and receives different scene graphs as input: N: Neural-Motifs [7], T: topic SG, and R: randomly selected relationships (baseline). Top k (*e.g.*, k=2, 5) relationships are used following [5]. We train and evaluate on the subset

Table 3. Results of relational captioning (%). "-RC" denotes that the model is only trained with the relational captions. "-ICRC" denotes that the model is trained with image captions and relational captions. R-ns means Recall-ns. Img-Lv. Recall means the image level recall.

Model	λ	mAP	METEOR	Img-Lv. Recall	R@20	R-ns@20	R@50	R-ns@50	R@100	R-ns@100
TriLSTM [2]	-	3.80	30.21	72.72	1.31	3.20	3.93	9.58	8.42	20.88
UD-RC [1]	-	5.61	42.40	88.77	3.02	3.71	**10.46**	12.92	**22.97**	28.90
UD-ICRC	0.1	4.84	38.31	84.81	**3.45**	**4.43**	10.22	**13.99**	20.77	**29.00**
	0.3	5.14	40.36	86.93	3.39	4.18	9.87	12.88	21.57	27.99
	0.7	5.43	42.26	89.15	2.75	3.49	9.97	12.40	20.76	26.46
	1.0	5.41	42.75	89.52	2.31	2.90	8.09	10.20	19.97	25.56
	3.0	5.56	43.51	90.28	1.84	2.29	7.03	8.73	17.82	22.28
	5.0	5.61	43.60	90.44	1.74	2.23	6.81	8.57	17.74	22.43
	7.0	**5.62**	**43.74**	**90.49**	1.74	2.05	7.03	8.55	17.75	22.00
	10.0	5.54	43.62	90.43	1.69	2.03	7.16	9.05	17.17	21.82
Transformer-RC [6]		5.26	41.62	88.65	2.11	2.73	6.83	9.12	16.36	21.91
Transformer-ICRC	0.7	5.15	41.63	88.64	2.05	2.70	6.86	9.19	16.21	21.91

Table 4. Results (%) of relational captioning when using different sorting strategies (denoted by "Strtg."). the R denotes randomly choosing the K relationships, while the L denotes sorting the relationships according to the sentence likelihoods.

Model	Strtg.	R@20	R-ns@20	R@50	R-ns@50	R@100	R-ns@100
TriLSTM	R	**1.51**	**3.68**	**4.59**	**12.72**	**10.72**	**28.85**
	L	1.31	3.20	3.93	9.58	8.42	20.88
UD-RC	R	2.13	3.12	8.48	12.34	18.73	27.51
	L	**3.02**	**3.71**	**10.46**	**12.92**	**22.97**	**28.90**

of our dataset (naturally the subset of MSCOCO) with the train/val/test split (5,000/100/1,000). The results are shown in Table 7. It is shown that N is not much better than R, especially when using less input relationships, while T outperforms them. It suggests that the topic SG can provide important content more efficiently.

References

[1] 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 (CVPR)*, pages 6077–6086, 2018. 1, 2, 3

[2] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Dense relational captioning: Triple-stream networks for relationship-based captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6271–6280, 2019. 2, 3

[3] Jimmy Kingma, Diederik P.and Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 1

[4] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of European Conference on Computer Vision (ECCV)*, volume 11208, pages 201–216. Springer, 2018. 2

[5] Xiangyang Li and Shuqiang Jiang. Know more say less: Image captioning based on scene graphs. *IEEE Transactions on Multimedia (TMM)*, 21(8):2117–2130, 2019. 2

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5998–6008, 2017. 1, 3

[7] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5831–5840, 2018. 2, 4, 5, 6

Table 5. Results (%) comparison on discovering the important relationships. "Feat." denotes different input features. "P" denotes the pooling function. "Mask" denotes masking the non-noun words (✓) or not (✗).

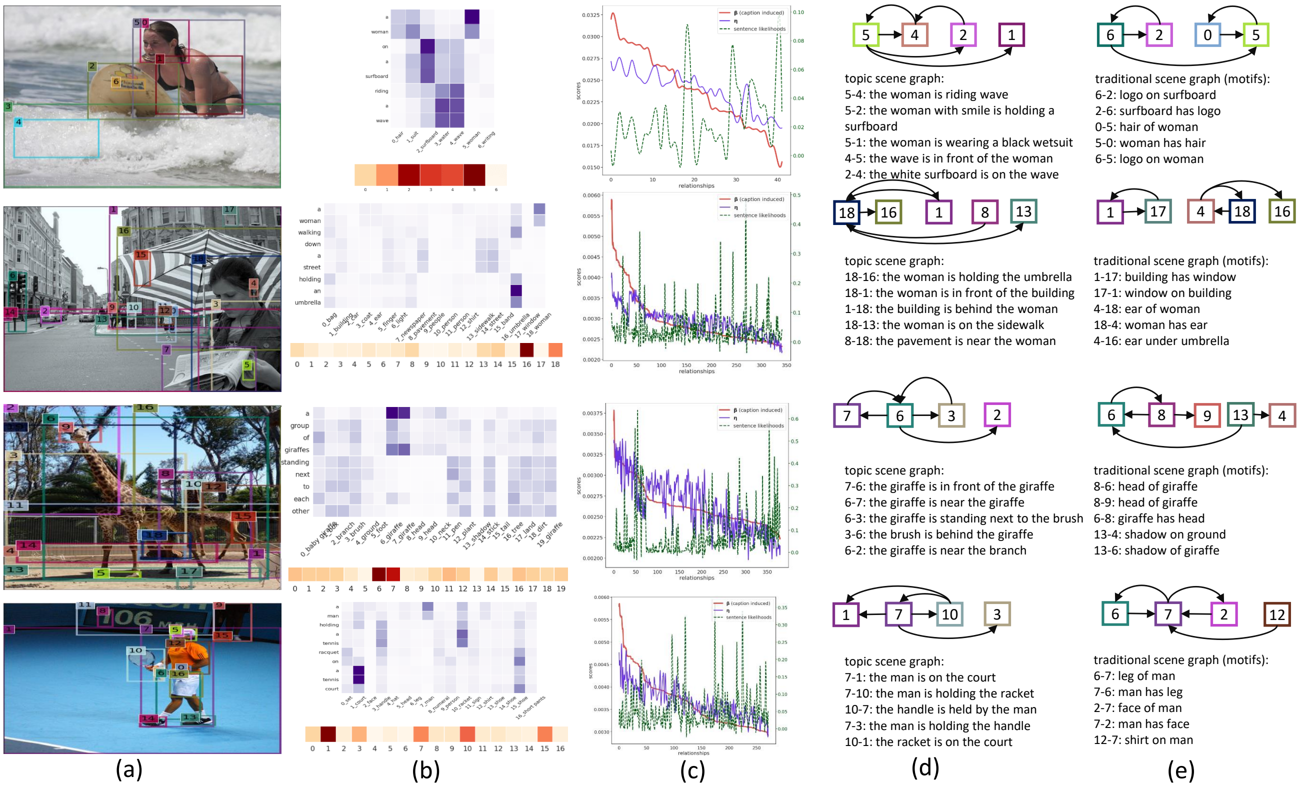
index	Model	Feat.	*P*	Mask	R@20	R-ns@20	R@50	R-ns@50	R@100	R-ns@100	mean
(1)	TriLSTM	-	-	-	1.31	3.20	3.93	9.58	8.42	20.88	7.89
(2)	UD-ICRC	-	-	-	2.75	3.49	9.97	12.40	20.76	26.46	12.64
(3)	UD-ICRC-attn	SO	MAX	✓	4.13	6.20	14.53	21.02	29.65	41.62	19.53
(4)		SO	MAX	✗	7.49	10.88	20.61	28.79	37.06	51.07	25.98
(5)		SO	MEAN	✓	1.55	2.59	7.09	10.70	18.02	26.18	11.02
(6)		SO	MEAN	✗	1.28	2.06	5.65	9.08	16.51	24.13	9.79
(7)		SOU	MAX	✓	11.35	16.20	22.02	30.53	33.84	47.20	26.86
(8)		SOU	MAX	✗	**15.71**	21.80	28.85	39.39	41.09	**55.73**	33.76
(9)		SOU	MEAN	✓	3.49	5.50	10.52	15.71	21.40	31.16	14.63
(10)		SOU	MEAN	✗	2.39	4.02	8.04	12.61	18.81	27.78	12.28
(11)		U	MAX	✓	5.15	7.56	13.47	19.22	25.79	36.73	17.99
(12)		U	MAX	✗	7.27	10.53	17.12	24.10	30.44	42.22	21.95
(13)		U	MEAN	✓	2.77	4.19	7.91	11.89	17.29	25.74	11.63
(14)		U	MEAN	✗	2.17	3.27	6.73	10.08	15.46	22.84	10.09
(15)		SOUS	MAX	✓	10.72	15.43	21.59	30.26	34.43	47.34	26.63
(16)		SOUS	MAX	✗	15.46	**21.81**	**29.55**	**40.72**	**41.14**	55.68	**34.06**
(17)		SOUS	MEAN	✓	2.74	4.53	8.76	13.71	19.27	28.05	12.84
(18)		SOUS	MEAN	✗	2.39	4.02	8.04	12.61	18.81	27.78	12.28
(19)	UD-ICRC-label	U	-	-	13.04	17.35	25.25	33.28	36.72	49.22	29.14
(20)		SO	-	-	30.14	38.86	41.45	53.95	51.55	67.70	47.28
(21)		SOU	-	-	32.17	41.38	43.57	56.68	53.65	70.81	49.71
(22)		SOUS	-	-	**34.39**	**45.13**	**46.03**	**60.97**	**54.60**	**72.44**	**52.26**

Table 6. The image retrieval results using different strategies: using top 1 relationship (Top-1), top 5 relationships (Top-5), and one sentence by connecting the top 5 relationships (Top-5-CON). We use the recall at K (R@K, higher is better) and the median rank of the target image (Med, lower is better).

Model	Top-1				Top-5				Top-5-CON			
	R@1	R@5	R@10	Med	R@1	R@5	R@10	Med	R@1	R@5	R@10	Med
TriLSTM	1.73	7.47	12.83	135.33	1.70	7.23	12.87	128.0	4.43	15.20	24.70	42.00
UD-ICRC	5.67	20.40	31.73	27.33	5.93	20.23	31.10	27.33	17.97	42.57	56.80	7.67
UD-ICRC-attn	**9.73**	**31.67**	**46.13**	**12.33**	**8.97**	**29.23**	**43.67**	**14.00**	**18.57**	**47.53**	**63.87**	**6.00**
UD-ICRC-label	17.77	49.17	67.37	5.67	13.47	39.47	55.63	8.67	29.27	63.50	79.87	3.00

Table 7. The image captioning performances (%) of different methods using top k relationships. N: Neural-Motifs [7], T: our topic SG, and R: randomly selected relationships (baseline).

Model	k	B1	B4	M	R	C	S
R	2	68.48	27.26	23.22	50.50	85.86	16.35
N		68.19	27.03	23.08	50.62	85.45	16.27
T		**68.93**	**28.44**	**23.75**	**51.39**	**90.31**	**17.02**
R	5	68.97	27.58	23.36	51.20	87.77	16.26
N		69.01	27.87	23.53	51.17	88.57	16.74
T		**69.41**	**28.76**	**23.79**	**51.58**	**91.63**	**17.20**



topic scene graph:
5-4: the woman is riding wave
5-2: the woman with smile is holding a surfboard
5-1: the woman is wearing a black wetsuit
4-5: the wave is in front of the woman
2-4: the white surfboard is on the wave

traditional scene graph (motifs):
6-2: logo on surfboard
2-6: surfboard has logo
0-5: hair of woman
5-0: woman has hair
6-5: logo on woman

topic scene graph:
18-16: the woman is holding the umbrella
18-1: the woman is in front of the building
1-18: the building is behind the woman
18-13: the woman is on the sidewalk
8-18: the pavement is near the woman

traditional scene graph (motifs):
1-17: building has window
17-1: window on building
4-18: ear of woman
18-4: woman has ear
4-16: ear under umbrella

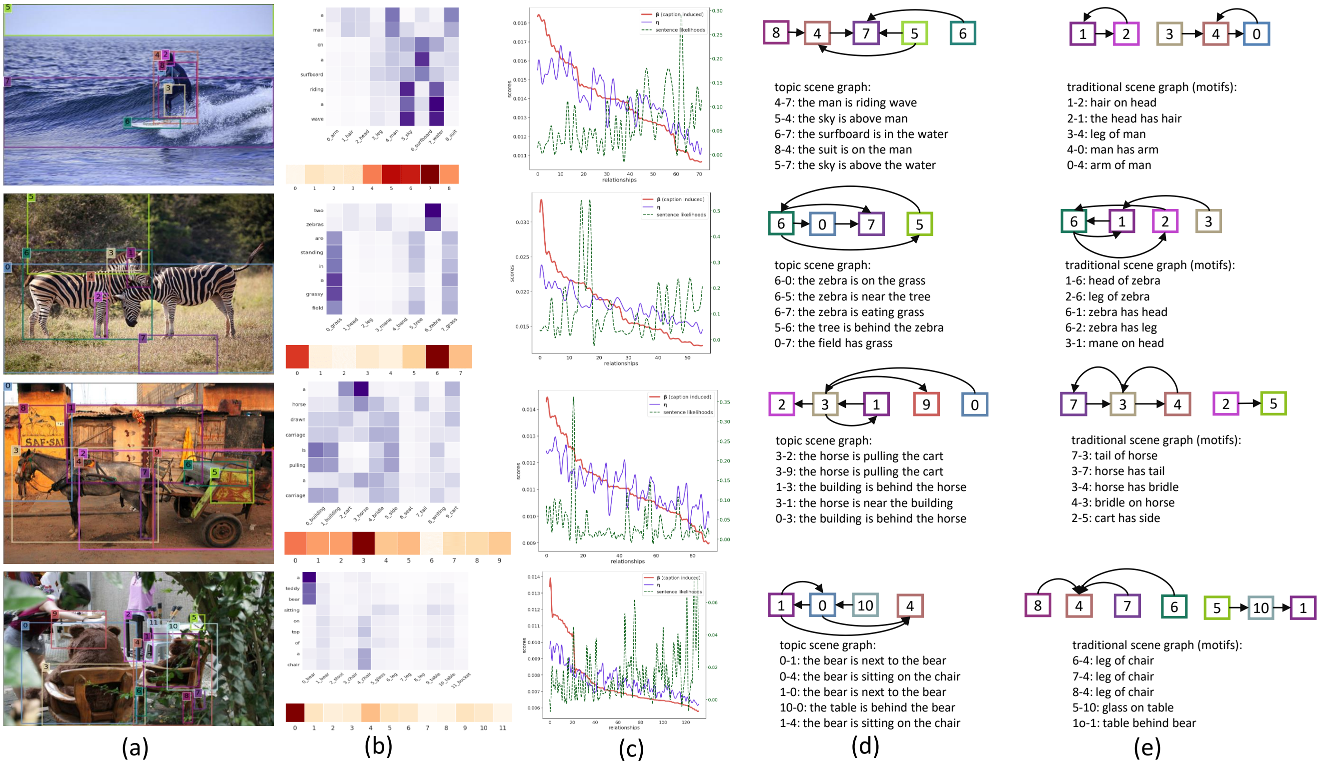
topic scene graph:
7-6: the giraffe is in front of the giraffe
6-7: the giraffe is near the giraffe
6-3: the giraffe is standing next to the brush
3-6: the brush is behind the giraffe
6-2: the giraffe is near the branch

traditional scene graph (motifs):
8-6: head of giraffe
8-9: head of giraffe
6-8: giraffe has head
13-4: shadow on ground
13-6: shadow of giraffe

topic scene graph:
7-1: the man is on the court
7-10: the man is holding the racket
10-7: the handle is held by the man
7-3: the man is holding the handle
10-1: the racket is on the court

traditional scene graph (motifs):
6-7: leg of man
7-6: man has leg
2-7: face of man
7-2: man has face
12-7: shirt on man

Figure 2. The qualitative results. (a) The bounding boxes of objects with their unique ids (number on the top left corner) are shown. (b) The attention about the objects during caption generation (the purple heat map) and the pooled attention (the reddish brown heat map) are visualized. Darker color indicates larger weights. Along the X-axis, the objects with their ids and categories correspond to the bounding boxes in (a). Along the Y-axis in the purple heat map, each word of the generated caption is shown. (c) Importance scores of the relationships are drawn. Along the X-axis, the relationships are sorted by the β scores in a descending order. All the lines are smoothed. (d-e) The scene graph from our methods and motifs [7] consisting of top 5 relationships are shown.



topic scene graph:
4-7: the man is riding wave
5-4: the sky is above man
6-7: the surfboard is in the water
8-4: the suit is on the man
5-7: the sky is above the water

traditional scene graph (motifs):
1-2: hair on head
2-1: the head has hair
3-4: leg of man
4-0: man has arm
0-4: arm of man

topic scene graph:
6-0: the zebra is on the grass
6-5: the zebra is near the tree
6-7: the zebra is eating grass
5-6: the tree is behind the zebra
0-7: the field has grass

traditional scene graph (motifs):
1-6: head of zebra
2-6: leg of zebra
6-1: zebra has head
6-2: zebra has leg
3-1: mane on head

topic scene graph:
3-2: the horse is pulling the cart
3-9: the horse is pulling the cart
1-3: the building is behind the horse
3-1: the horse is near the building
0-3: the building is behind the horse

traditional scene graph (motifs):
7-3: tail of horse
3-7: horse has tail
3-4: horse has bridle
4-3: bridle on horse
2-5: cart has side

topic scene graph:
0-1: the bear is next to the bear
0-4: the bear is sitting on the chair
1-0: the bear is next to the bear
10-0: the table is behind the bear
1-4: the bear is sitting on the chair

traditional scene graph (motifs):
6-4: leg of chair
7-4: leg of chair
8-4: leg of chair
5-10: glass on table
1o-1: table behind bear

Figure 3. The qualitative results. (a) The bounding boxes of objects with their unique ids (number on the top left corner) are shown. (b) The attention about the objects during caption generation (the purple heat map) and the pooled attention (the reddish brown heat map) are visualized. Darker color indicates larger weights. Along the X-axis, the objects with their ids and categories correspond to the bounding boxes in (a). Along the Y-axis in the purple heat map, each word of the generated caption is shown. (c) Importance scores of the relationships are drawn. Along the X-axis, the relationships are sorted by the β scores in a descending order. All the lines are smoothed. (d-e) The scene graph from our methods and motifs [7] consisting of top 5 relationships are shown.

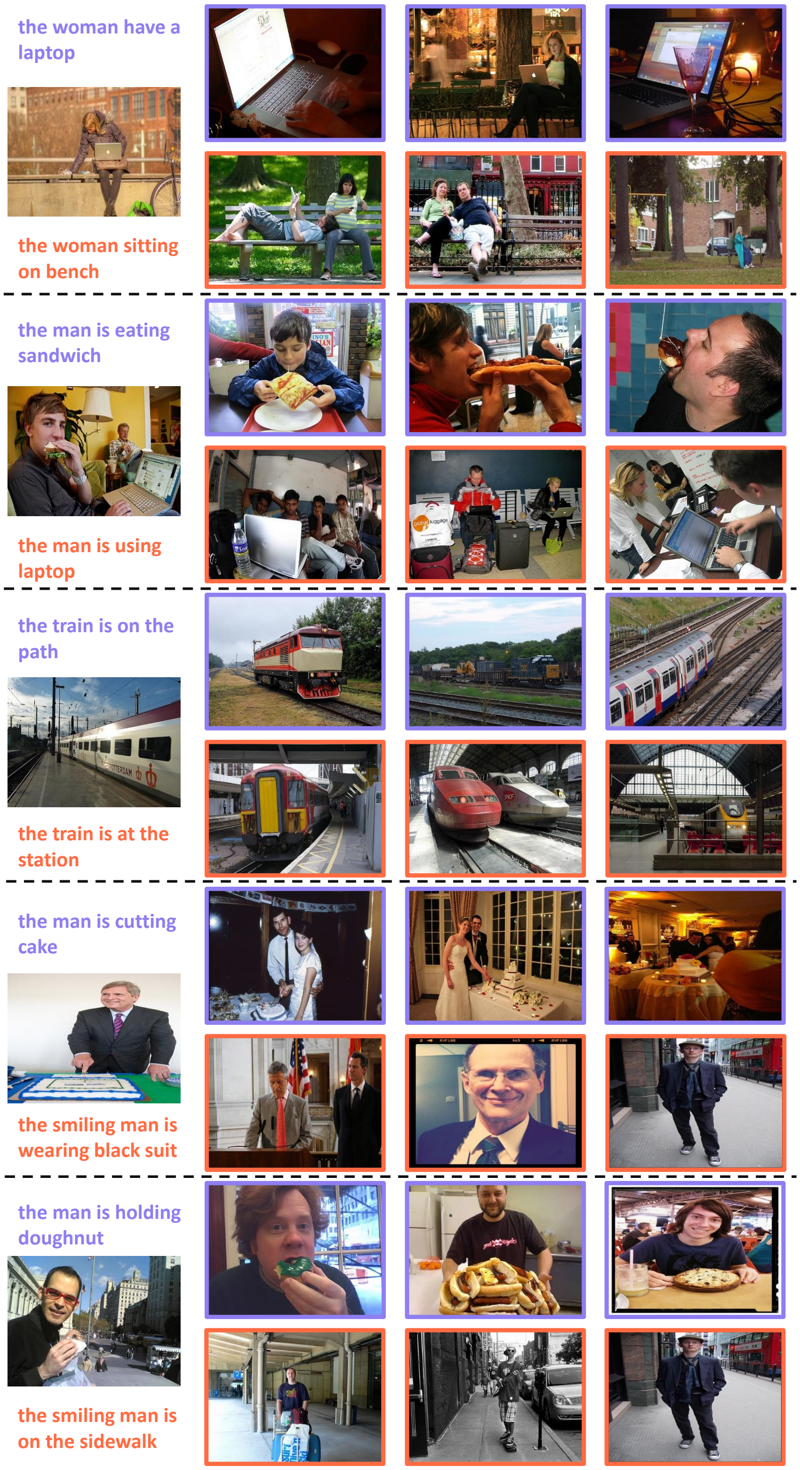

the woman have a laptop
the woman sitting on bench
the man is eating sandwich
the man is using laptop
the train is on the path
the train is at the station
the man is cutting cake
the smiling man is wearing black suit
the man is holding doughnut
the smiling man is on the sidewalk

Figure 4. Two important relationships given by topic scene graph of the left image are used to retrieve similar images respectively. The images with purple or orange boundaries correspond to the purple or orange relationships.