Dynamic Gated Graph Neural Networks for Scene Graph Generation

Mahmoud Khademi and Oliver Schulte

Department of Computing Science
Simon Fraser University
mkhademi@sfu.ca, oschulte@sfu.ca

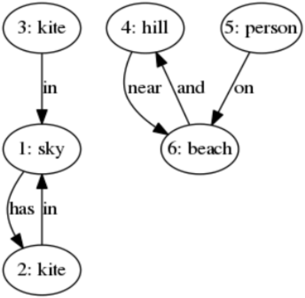
Presented by Rui Zeng

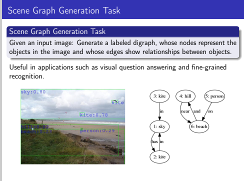
Scene Graph Generation Task

Scene Graph Generation Task

Given an input image: Generate a labeled digraph, whose nodes represent the objects in the image and whose edges show relationships between objects.

Useful in applications such as visual question answering and fine-grained recognition.



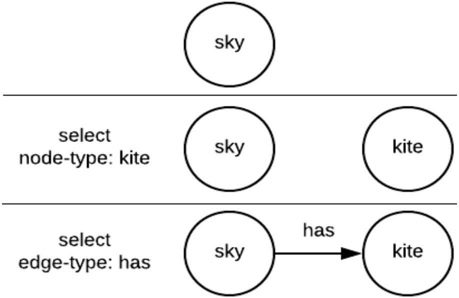


- A scene graph provides *scene understanding.*

D-GGNN: Reinforcement Learning for Scene Graph Generation

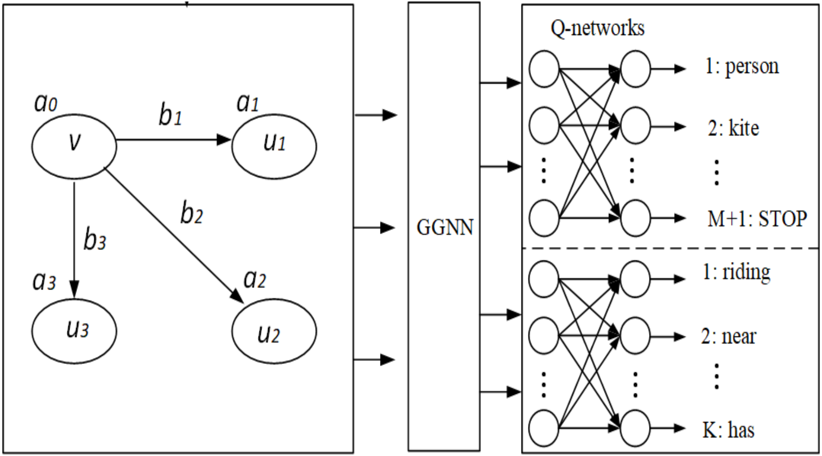
A scene graph generation algorithm needs to exploit visual contextual information.

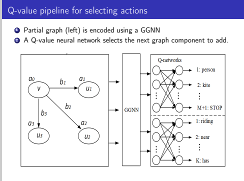
- State = (encoding of) partial graph
- Action = expands current graph
- Reward = agreement with ground truth



Q-value pipeline for selecting actions

1. Partial graph (left) is encoded using a GGNN
2. A Q-value neural network selects the next graph component to add.





- We use a standard object detector (Tensorflow API). The object detector produces objectness confidence scores. The graph construction starts with th highest-scoring object.
- Given an image and its candidate object bounding-boxes, we simultaneousl build the graph and assign node-types and edge-types to the nodes and edge in a Deep Q-Learning framework Mnih et al. [2013, 2015].
- The next slide explains more details.

Q-value network for selecting actions

State = Encoded Graph

- Feature vectors for each node v:
  - ResNet feature vector $\hat{\mathbf{x}}_v$
  - Node embedding $\mathbf{h}_v$, computed by GGNN Li et al. [2015].
    - Captures link information.
- Node feature vectors are combined using a soft attention mechanism that represents how important node v is for the next decision.

A Q-function takes as input a state s and an action a and outputs expected future reward $Q(s, a)$.

- Implemented by deep neural network
- Trained by temporal difference learning

Q-value network for selecting actions

State = Encoded Graph

- Feature vectors for each node v:
  - ResNet feature vector $\hat{\mathbf{x}}_v$
  - Node embedding $\mathbf{h}_v$, computed by GGNN Li et al. [2015].
    - Captures link information.
- Node feature vectors are combined using a soft attention mechanism that represents how important node v is for the next decision.

A Q-function takes as input a state s and an action a and outputs expected future reward $Q(s, a)$.

- Implemented by deep neural network
- Trained by temporal difference learning

- GGNN stands for Gated Graph Neural Net

Dataset

- The Visual Genome (VG) dataset 1.4 [Krishna et al., 2016] contains $108,077$ images. Annotations provide subject-predicate-object triples.
  - e.g. man-throwing-frisbee
- 5,000 images for hyperparameter validation, 5,000 for testing.
- Preprocessing:
  - `VG1.4-a` uses the most frequent 150 object categories and 50 predicates Xu et al. [2017].
  - `VG1.4-b` uses the most frequent 1750 object categories and 347 predicates.

Dataset

- The Visual Genome (VG) dataset 1.4 [Krishna et al., 2016] contains 108, 077 images. Annotations provide subject-predicate-object triples.
  - e.g. man-throwing-frisbee
- 5,000 images for hyperparameter validation, 5,000 for testing.
- Preprocessing:
  - `VG1.4-a` uses the most frequent 150 object categories and 50 predicates Xu et al. [2017].
  - `VG1.4-b` uses the most frequent 1750 object categories and 347 predicates.

- Predicate = edge type.

Metrics (`VG1.4-a`)

The goal is to find ground truth relationship triplets (subject-predicate-object). Different input information = different tasks.

- Predicate classification (PRED-CLS): location and object categories are given.
- Scene graph classification (SG-CLS) task: location of objects are given.
- Scene graph generation (SG-GEN) task: only the image is given.
- Relationship phrase detection (REL-PHRASE-DET) and Relationship detection (REL-DET) are similar to SG-GEN, applied on VG1.4-b Liang et al. [2017].
- Metric is Top-K recall (`Rec@K`): the number of the ground-truth-triples hit in the top-K predictions in an image.

Metrics (VG1.4-a)

The goal is to find ground truth relationship triplets (subject-predicate-object). Different input information = different tasks.

- Predicate classification (PRED-CLS): location and object categories are given.
- Scene graph classification (SG-CLS) task: location of objects are given.
- Scene graph generation (SG-GEN) task: only the image is given.
- Relationship phrase detection (REL-PHRASE-DET) and Relationship detection (REL-DET) are similar to SG-GEN, applied on VG1.4-b Liang et al. [2017].
- Metric is Top-K recall (Rec@K): the number of the ground-truth-triples hit in the top-K predictions in an image.

- Predictions are ranked by the product of the objectness confidence scores an the Q-values of the selected predicates.
- We are following the evaluation methodology in previous papers.

Experimental Results

Model	PRED-CLS		SG-CLS		SG-GEN	
	R@50	R@100	R@50	R@100	R@50	R@100
Lu et al. [2016]	27.88	35.04	11.79	14.11	00.32	00.47
Xu et al. [2017]	44.75	53.08	21.72	24.38	03.44	04.24
D-GGNN (ours)	**46.85**	**55.63**	**23.80**	**26.78**	**06.36**	**07.54**

Table: VG1.4-a results for scene graph generation (SG-GEN). D-GGNN finds twice as many triplets as the previous state-of-the-art.

Model	REL-PHRASE-DET		REL-DET	
	R@100	R@50	R@100	R@50
CNN+RPN Simonyan and Zisserman [2014]	01.39	01.34	01.22	01.18
Faster R-CNN Ren et al. [2015]	02.25	02.19	-	-
CNN+TRPN Ren et al. [2015]	02.52	02.44	02.37	02.23
Lu et al. [2016]	10.23	09.55	07.96	06.01
VRL Liang et al. [2017]	16.09	14.36	13.34	12.57
D-GGNN (ours)	**18.21**	**15.78**	**14.85**	**14.22**

Table: On VG1.4-b results on variants of the scene graph generation task. D-GGNN shows an improvement over the most recent baseline, and almost double for the older methods.

Experimental Results

Model	PRED-CLS		SG-CLS		SG-GEN	
	R@50	R@100	R@50	R@100	R@50	R@100
Lu et al. [2016]	27.88	35.04	11.79	14.11	00.32	00.47
Xu et al. [2017]	44.75	53.08	21.72	24.38	03.44	04.24
D-GGNN (ours)	**46.85**	**55.63**	**23.80**	**26.78**	**06.36**	**07.54**

Table: VG1.4-a results for scene graph generation (SG-GEN). D-GGNN finds twice as many triplets as the previous state-of-the-art.

Model	REL-PHRASE-DET		REL-DET	
	R@100	R@50	R@100	R@50
CNN+RPN Simonyan and Zisserman [2014]	01.39	01.34	01.22	01.18
Faster R-CNN Ren et al. [2015]	02.25	02.19	-	-
CNN+TRPN Ren et al. [2015]	02.52	02.44	02.37	02.23
Lu et al. [2016]	10.23	09.55	07.96	06.01
VRL Liang et al. [2017]	16.09	14.36	13.34	12.57
D-GGNN (ours)	**18.21**	**15.78**	**14.85**	**14.22**

Table: On VG1.4-b results on variants of the scene graph generation task. D-GGNN shows an improvement over the most recent baseline, and almost double for the older methods.

- For the VG1.4a dataset, the average number of ground-truth triplets in the images is 7.1.
- Results for VRL Liang 2017 are not available for VG1.4-a.

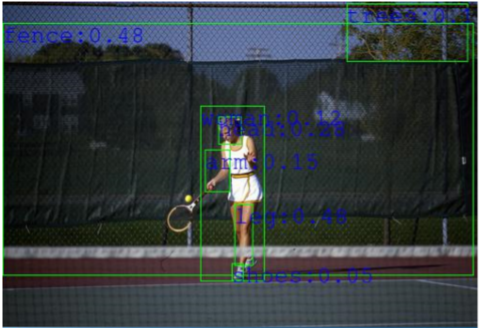
Conclusion

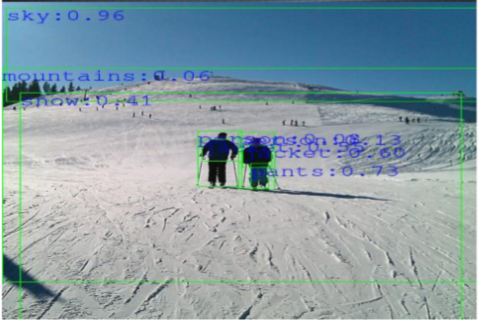
- Scene graph generation is an important part of scene understanding.
- We utilized a deep Reinforcement learning framework to sequentially generate a scene graph for an input image.
- New idea: entire partial graph is encoded as state information for RL.
  - A Gated Graph Neural Network computes node embeddings that capture relational information.
- We presented a generative deep architecture for graph-structured information from data sources (e.g. image, videos, text, program).
- Future Work: Evaluate in more applications, e.g. Visual Question Answering.
- We have a couple more scene graphs to show.

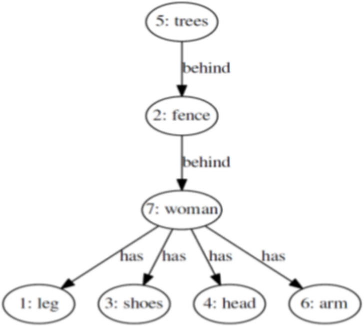
Conclusion

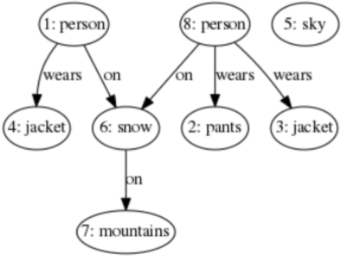
- Scene graph generation is an important part of scene understanding.
- We utilized a deep Reinforcement learning framework to sequentially generate a scene graph for an input image.
- New idea: entire partial graph is encoded as state information for RL.
  - A Gated Graph Neural Network computes node embeddings that capture relational information.
- We presented a generative deep architecture for graph-structured information from data sources (e.g. image, videos, text, program).
- Future Work: Evaluate in more applications, e.g. Visual Question Answering.
- We have a couple more scene graphs to show.

These are optional, it would be nice to leave the audience with pictures.

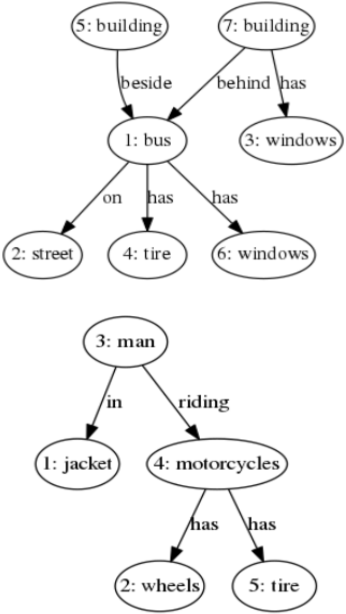

fence:0.48
arm:0.15
leg:0.48
shoes:0.05


sky:0.96
pants:0.73


5: trees
behind
2: fence
behind
7: woman
has
has
has
has
1: leg
3: shoes
4: head
6: arm


1: person
8: person
5: sky
wears
on
on
wears
wears
4: jacket
6: snow
2: pants
3: jacket
on
7: mountains


building:0.02
windows:0.04
windows:0.01
iPod
tire:0.28
man:0.75
wheels:0.02


5: building
7: building
beside
behind
has
1: bus
3: windows
on
has
has
2: street
4: tire
6: windows
3: man
in
riding
1: jacket
4: motorcycles
has
has
2: wheels
5: tire

References

R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*, 2016.

Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*, 2015.

X. Liang, L. Lee, and E. P. Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4408–4417. IEEE, 2017.

C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pages 852–869. Springer, 2016.

V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 2017.