Graphr: Scene Graph Generation using Deep Variation-structured Reinforcement Learning

Apoorva Dornadula, Aarti Bagul

Stanford University

Overview

Problem: Generate scene graphs for images

What is a Scene Graph?: A scene graph is a graph structure with nodes as objects and edges as relationships. An object can have attributes.

Motivation for using Reinforcement Learning:

- Sequentially decide relationships/attributes depending on previous relationships/attributes assigned in an image
- Large and dynamic state and action space

Dataset

- We use the Visual Genome [2] (VG) dataset which provides images and its corresponding scene graph information.

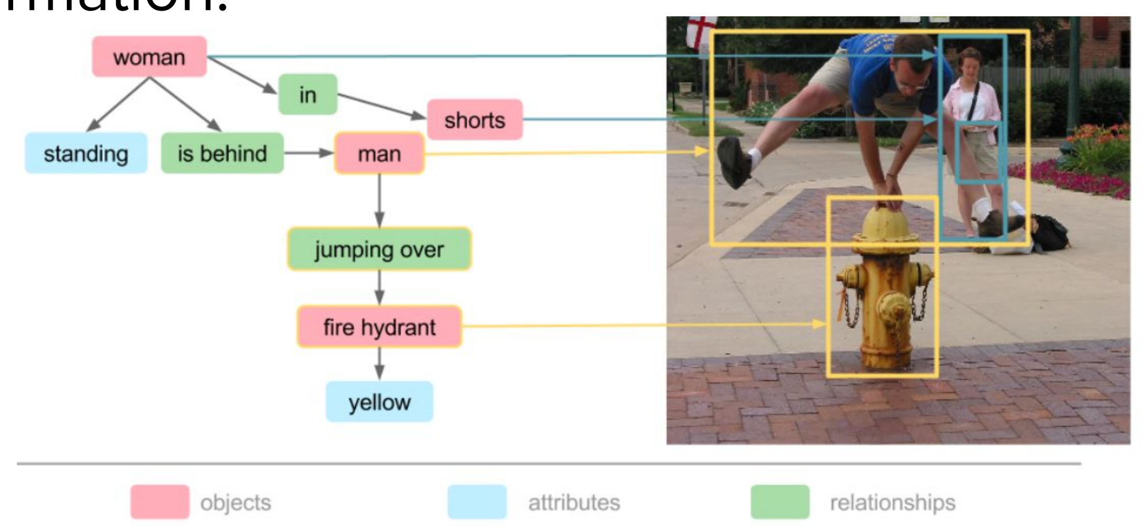

Fig. 1: An image from Visual Genome and its scene graph.

Baseline Model

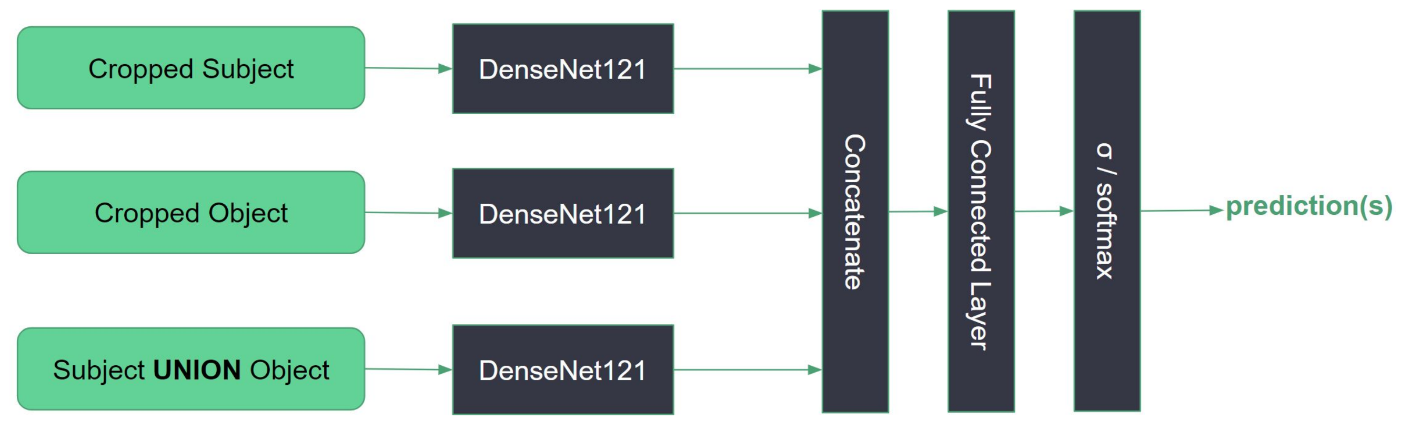

Fig. 2: Network architecture for our baseline model. We used this architecture to predict 1). whether or not an edge exists between a subject & object, 2). the relationship between a subject & object

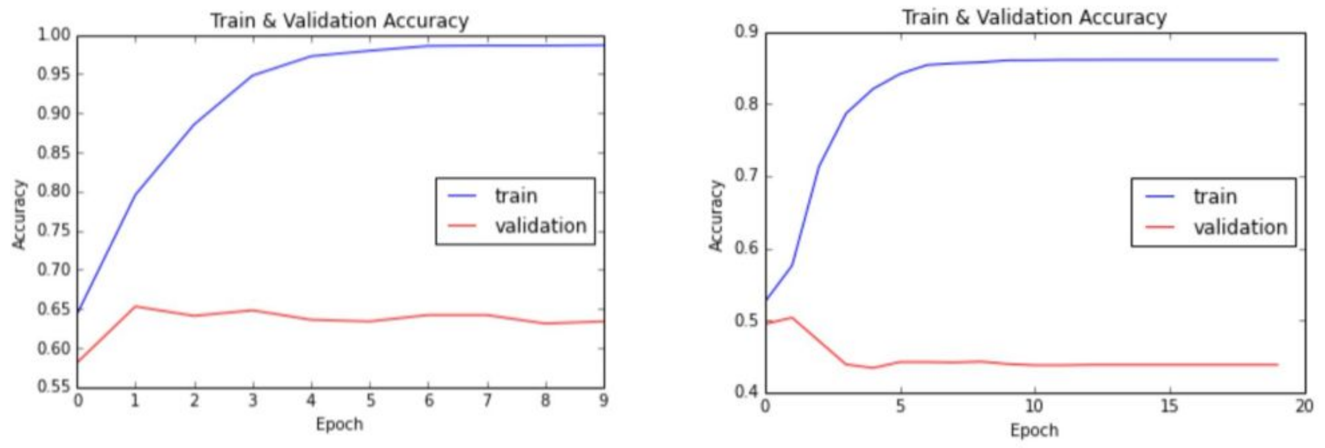

Fig. 3: Edge Existence Prediction (left), Edge Prediction (right)

Model - VRL (w/ experience replay & target network)

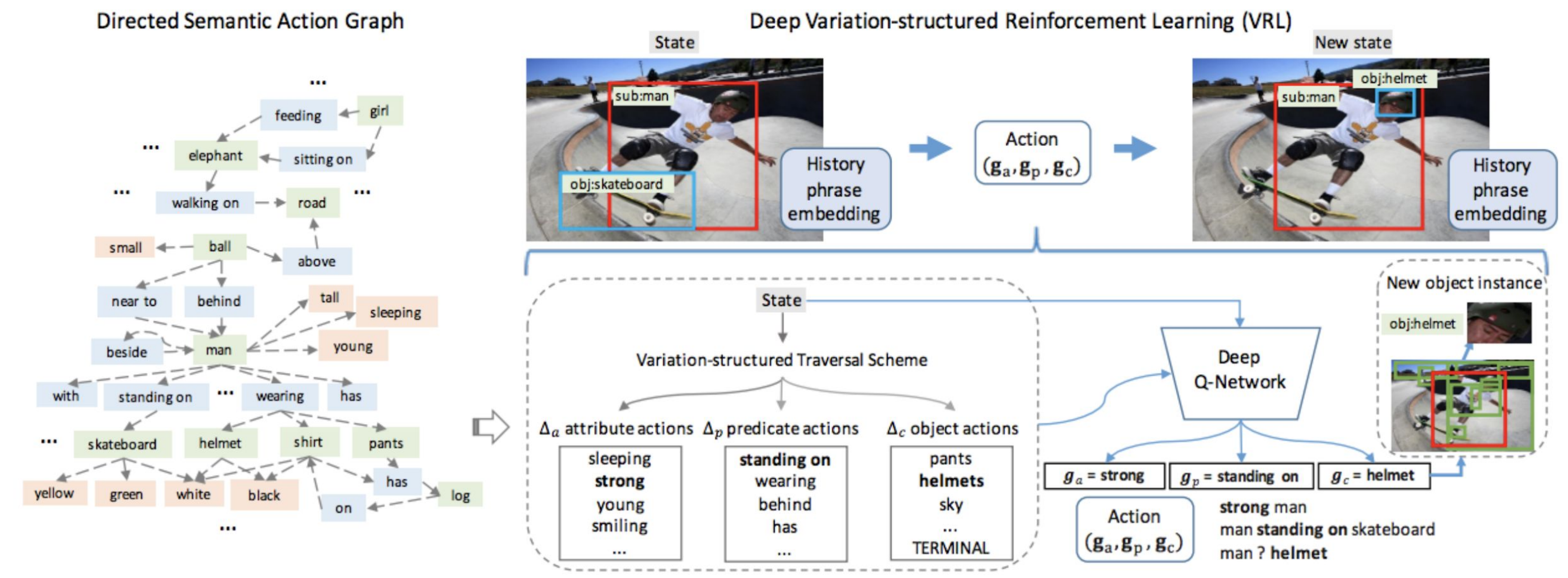

Fig 4: Overview of the VRL [1] framework

- **Directed Semantic Action Graph (SAG)**: a graph whose nodes are entities, relationships, or attributes. An edge can either connect an entity with an attribute (attribute edge) or connect two entities with a relationship (predicate edge). We created the graph using the VG dataset.
- **State space**: Image features, subject features, object features, history embedding
- **Action space**: SAG; Smaller adaptive action sets (Δ_a, Δ_p, Δ_a) are generated using a variation-structured traversal scheme.
- **Rewards**: correct attribute for subject predicted (+1), correct relationship between subject and object predicted (+1), next object proposal overlaps with an object not yet explored (+5), else -1
- **DQN**: There are 3 DQNs [6][7] used in this architecture to predict the following: attribute for a subject, relationship between the subject & object, and the next object to pair with the subject.
- **Why DQN?**: able to capture global dependencies, data efficient, make sequential predictions

Attribute DQN Parameters Update

$$\theta_a^{(t+1)} = \theta_a^{(t)} + \alpha(\mathcal{R}_a + \lambda \max_{\mathbf{g}_{a'}} Q(\mathbf{f}', \mathbf{g}_{a'}; \theta_a^{(t)-}) - Q(\mathbf{f}, \mathbf{g}_a; \theta_a^{(t)})) \nabla_{\theta_a^{(t)}} Q(\mathbf{f}, \mathbf{g}_a; \theta_a^{(t)}),$$

Predicate DQN Parameters Update

$$\theta_p^{(t+1)} = \theta_p^{(t)} + \alpha(\mathcal{R}_p + \lambda \max_{\mathbf{g}_{p'}} Q(\mathbf{f}', \mathbf{g}_{p'}; \theta_p^{(t)-}) - Q(\mathbf{f}, \mathbf{g}_p; \theta_p^{(t)})) \nabla_{\theta_p^{(t)}} Q(\mathbf{f}, \mathbf{g}_p; \theta_p^{(t)}),$$

Next Object DQN Parameters Update

$$\theta_c^{(t+1)} = \theta_c^{(t)} + \alpha(\mathcal{R}_c + \lambda \max_{\mathbf{g}_{c'}} Q(\mathbf{f}', \mathbf{g}_{c'}; \theta_c^{(t)-}) - Q(\mathbf{f}, \mathbf{g}_c; \theta_c^{(t)})) \nabla_{\theta_c^{(t)}} Q(\mathbf{f}, \mathbf{g}_c; \theta_c^{(t)}),$$

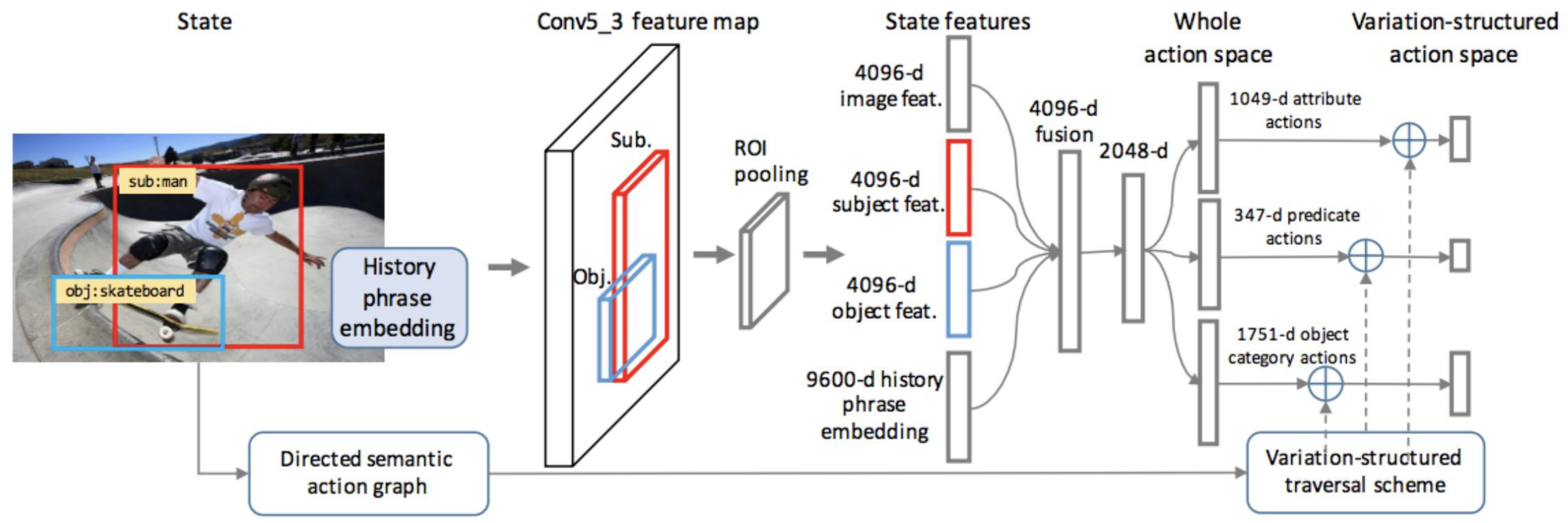

Fig 5: DQN network used to choose the relationship and attribute with respect to the current subject, as well as the next object

Results

Visualization:

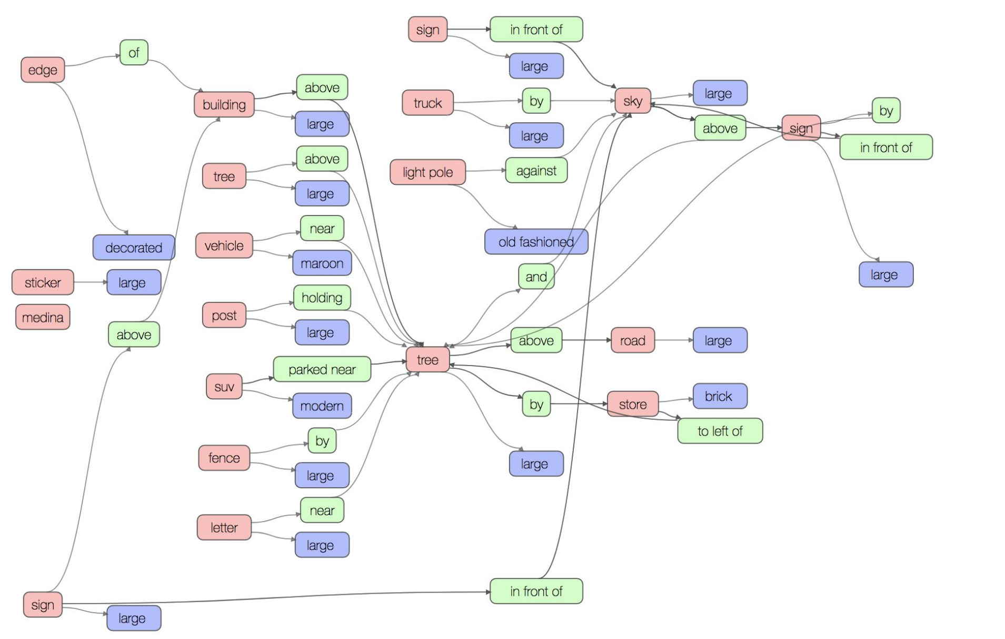

Fig. 6 VRL Generated Scene Graph

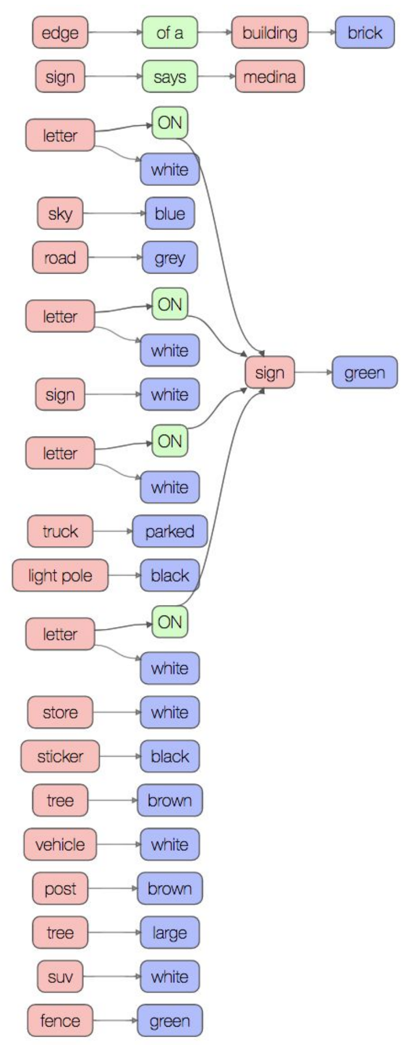

Fig. 7 Ground Truth Scene Graph

	Precision	Recall
Entities	1.0	1.0
Attributes	0.026	0.034
Relationships	0.0	0.0

Table 1: Results

Next Steps

- We can create a more comprehensive semantic action graph using natural language sentences
- This architecture can be generalized, to an unsupervised learning framework. This would allow us to learn from unlabeled images.
- We can adapt this architecture to include an active learning component. This would allow us to learn new objects and new predicates not in the initial dataset.

References

[1] X. Liang, L. Lee, and E. P. Xing. Deep variation structured reinforcement learning for visual relationship and attribute detection. In CVPR, 2017

[2] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016

[3] Huang, Gao and Liu, Zhuang and van der Maaten, Laurens and Weinberger, Kilian Q, Densely connected convolutional networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017

[4] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015

[5] Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proc. International Conference on Learning Representations.

[6] Mnih, Volodymyr, Kavukcuoglu, Koray, Silver, David, Rusu, Andrei A., Veness, Joel, Bellemare, Marc G., Graves, Alex, Riedmiller, Martin, Fidjeland, Andreas K., Ostrovski, Georg, Petersen, Stig, Beattie, Charles, Sadik,Amir, Antonoglou, Ioannis, King, Helen, Kumaran, Dharshan, Wierstra, Daan, Legg, Shane, and Hassabis, Demis. Human-level control through deep reinforcement learning. Nature, 518(7540):529–533, 02 2015

[7] Van Hasselt, Hado, Guez, Arthur, and Silver, David. Deep reinforcement learning with double q-learning. arXiv preprint arXiv:1509.06461, 2015.