Deep Variation-structured Reinforcement Learning (VRL) for Visual Relationship & Attribute Detection

Xiaodan Liang, Lisa Lee, Eric Xing
Machine Learning Department, Carnegie Mellon University

IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR July 21-26 2017 HONOLULU

Motivation

Computers still struggle to understand the interdependency of objects in the scene as a whole.

Existing methods ignore **global context cues** capturing the interactions among different object instances.

Relationship $(c, p, c') \in \mathcal{E}^{\mathcal{P}}$

girl (subject c) → feeding (predicate p) → elephant (object c')

Attribute modifier $(c, a) \in \mathcal{E}^{\mathcal{A}}$

elephant (subject c) → brown (attribute a)

Our Contributions

Use a novel **variation-structured traversal scheme** over the semantic action graph that dynamically constructs small action sets for each step.

Make **sequential decisions** using a deep RL framework, incorporating global context cues and semantic embeddings in the state vector.

Achieve **state-of-the-art** results on Visual Genome & Visual Relationship Detection datasets.

Method	Phr. R@100	Phr. R@50	Rel. R@100	Rel. R@50	Attr. R@100	Attr. R@50
Joint CNN+RPN	2.52	2.44	2.37	2.23	9.77	8.35
Lu et al. (2016)	10.23	9.55	7.96	6.01	-	-
Our VRL	**16.09**	**14.36**	**13.34**	**12.57**	**26.43**	**24.87**

Results for relationship phrase detection (Phr.), relationship detection (Rel.), and attribute detection (Attr.) on the Visual Genome dataset. R@N is short for Recall@N.

Methods

ε-greedy actions: Given the current **subject** and **object** instances, choose:

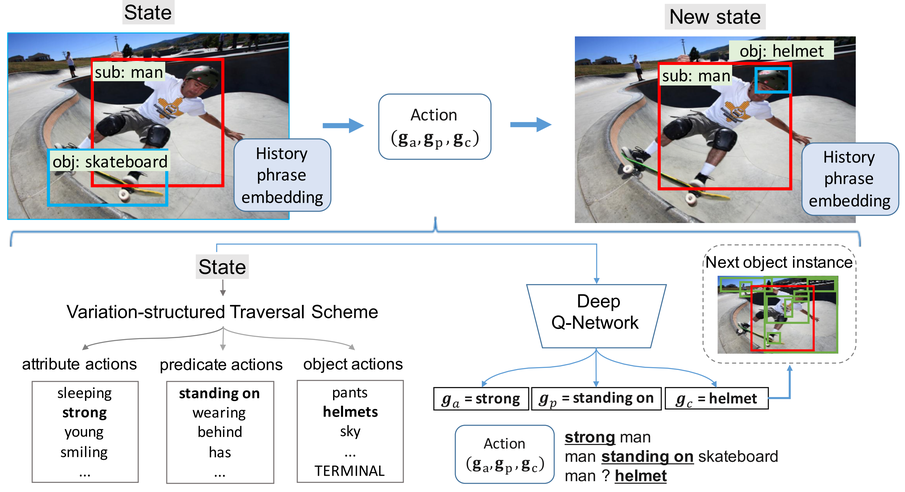
1. An **attribute** $\mathbf{g}_a$ describing the subject
2. A **predicate** $\mathbf{g}_p$ relating the subject and object
3. The **next object instance** $\mathbf{g}_c$.

Q-learning update given actions $(\mathbf{g}_a, \mathbf{g}_p, \mathbf{g}_c)$, state transition $\mathbf{f} \rightarrow \mathbf{f}'$, rewards $(\mathcal{R}_a, \mathcal{R}_p, \mathcal{R}_c)$:

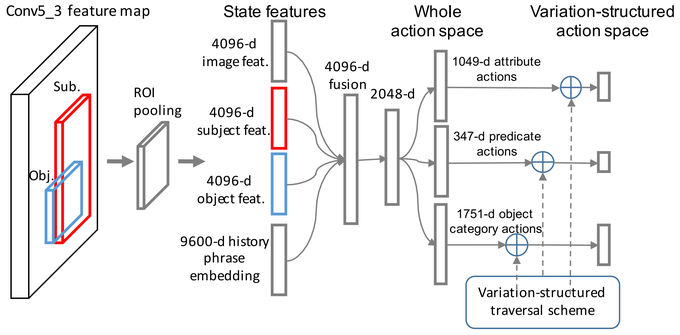
$$\theta_a^{(t+1)} = \theta_a^{(t)} + \alpha\left(\mathcal{R}_a + \lambda \max_{\mathbf{g}_{a'}} Q\left(\mathbf{f}', \mathbf{g}_{a'}; \theta_a^{(t)-}\right) - Q\left(\mathbf{f}, \mathbf{g}_a; \theta_a^{(t)}\right)\right)\nabla_{\theta_a^{(t)}} Q\left(\mathbf{f}, \mathbf{g}_a; \theta_a^{(t)}\right)$$

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

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



Network Architecture



Experimental Results

Comparison vs. VRD from Lu et al. (2016)

Example results generated by our VRL on the Visual Genome dataset.