Scene Graph Generation by Iterative Message Passing

Danfei Xu[1], Yuke Zhu[1], Christopher Choy[2], Li Fei-Fei[1]

[1]Department of Computer Science, [2]Department of Electrical Engineering, Stanford University

Motivation

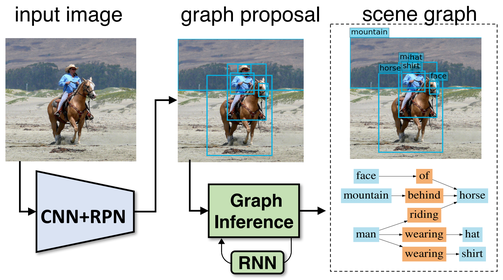
Existing methods:

- Focus on **isolated** predictions of objects and relationships while ignoring visual contexts.

Key insights:

- *Scene graphs* contain rich semantic and contextual information.
- Can be used to improve predictions and resolve ambiguities caused by lack of context.

Overview



- Our model generates a scene graph directly from an input image.
- The model **iteratively** refines graph predictions by passing visual context information along the topological structure of a scene graph using standard RNNs [3].
- The final model achieved state-of-the-art performance on the Visual Genome [4] and the NYU Depth V2 dataset [5].

Key components

Scene graph representation

- **Nodes**: object locations and categories
- **Edges**: pair-wise relationships between objects

Graph Inference using Recurrent Neural Network

- Approximate mean field inference using standard RNNs (GRU) [3]
- Each node and edge have their own hidden states (w/ shared weights)
- Iterative inference by passing hidden state as messages
- Node i: hidden state h_i, visual state f_i^v
- Edge $i \to j$ hidden state $h_{i\to j}$, visual feature $f_{i\to j}^e$

$$Q(x|I,B_I) = \prod_{i=1}^{n} Q(x_i^{cls}, x_i^{bbox}|h_i) Q(h_i|f_i^v) \prod_{i=1}^{n} Q(x_{i\to j}\,|h_{i\to j}) Q(h_{i\to j}|f_{i\to j}^e)$$

Primal-dual graph update and Message Pooling

We can exploit the **bipartite structure** of a scene graph:

- Neighbors of the edge GRUs are node GRUs, and vice versa.
- Passing messages along this structure forms two disjoint sub-graphs that are the **dual graph** to each other.

Aggregate incoming messages of a node / edge adaptively:

- Learn weight factors for each incoming message and fuse the messages using a **weighted sum**.

Primal-dual graph update with adaptive message pooling:

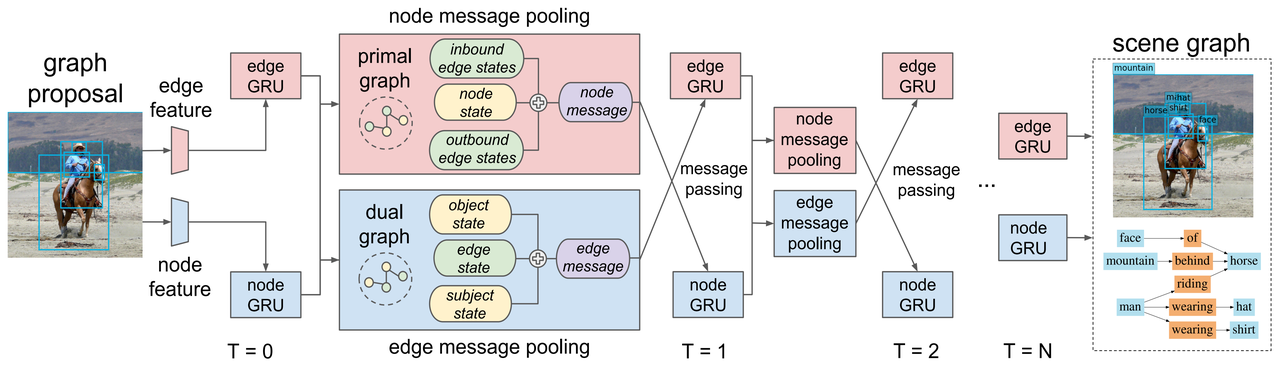
$$\textbf{Primal}: m_i = \sum_{j:i\to j} \sigma(\mathbf{v}_1^T[h_i, h_{i\to j}])h_{i\to j} + \sum_{j:j\to i} \sigma(\mathbf{v}_2^T[h_i, h_{j\to i}])h_{j\to i}$$

$$\textbf{Dual}: \quad m_{i\to j} = \sigma(\mathbf{w}_1^T[h_i, h_{i\to j}])h_i + \sigma(\mathbf{w}_2^T[h_j, h_{i\to j}])h_j$$

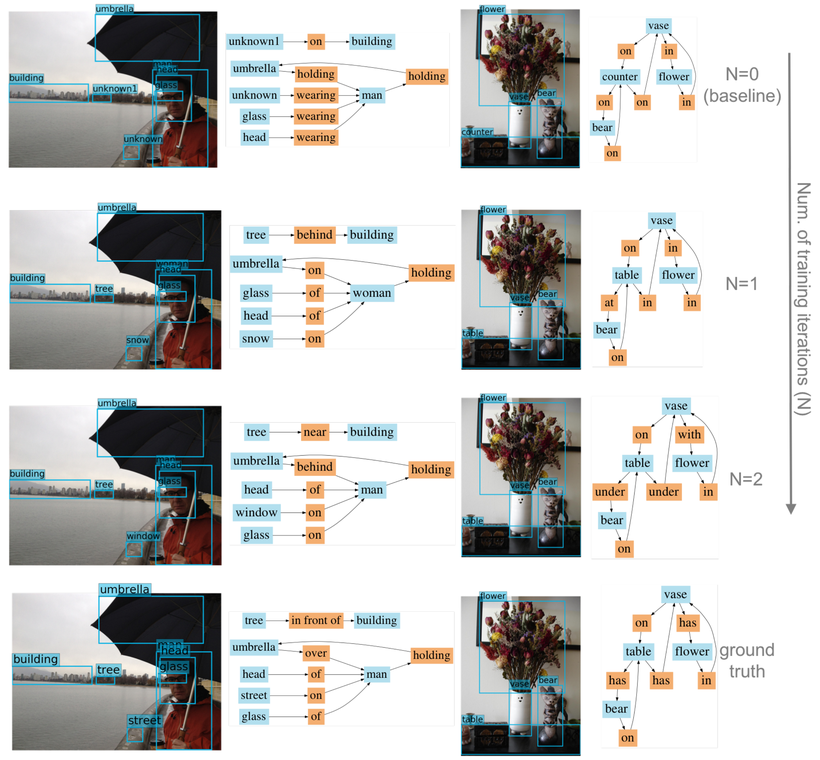
Model architecture

Inference procedure:

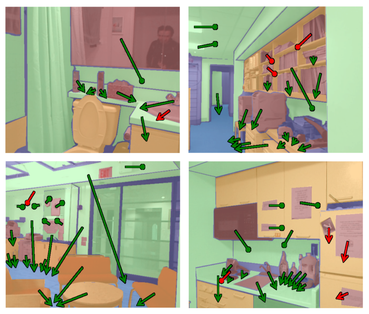
1. Proposes a candidate graph using a Regional Proposal Network (RPN)
2. Extract visual features of edges and nodes and update initial GRU hidden states with the visual features (T=0)
3. Aggregate hidden state messages using our adaptive message pooling method
4. (T=1) Update hidden states with the aggregated messages
5. (T=N) Use final step hidden states to infer object locations, categories, and pair-wise relationships.



Qualitative analysis of iterative inference



Evaluation on NYU Depth V2 [5]

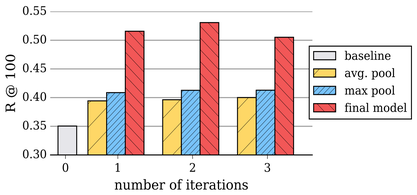


	Support Accuracy		PredCls	
	t-ag	t-aw	R@50	R@100
Silberman *et al.* [28]	75.9	72.6	-	-
Liao *et al.* [24]	88.4	82.1	-	-
Baseline [26]	87.7	85.3	34.1	50.3
Final model (ours)	**91.2**	**89.0**	**41.8**	**55.5**

Sample support relation predictions using our model. →: support from below, ➞: support from behind. Red arrows are incorrect predictions. We also color code structure classes: ground is in blue, structure is in green, furniture is in yellow, prop is in red. Purple indicates missing class.

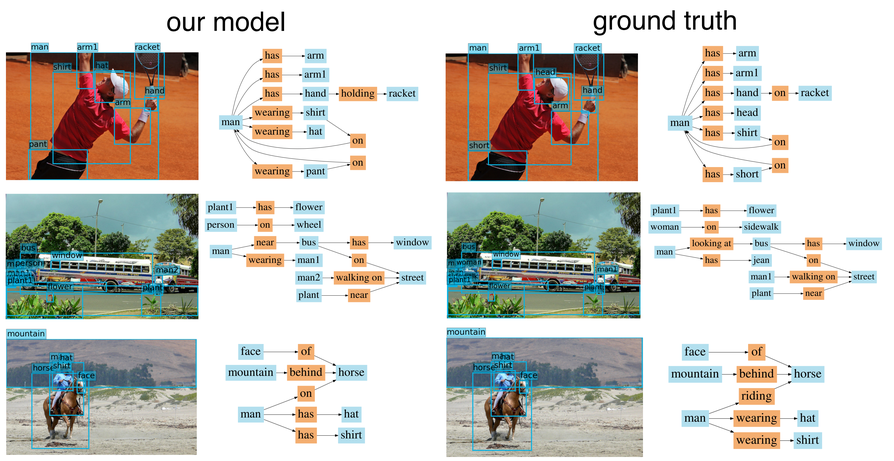
Evaluation on Visual Genome [4]

		[2]	avg. pool	max pool	final
PredCls	R@50	27.88	32.39	34.33	**44.75**
	R@100	35.04	39.63	41.99	**53.08**
SGCls	R@50	11.79	15.65	16.31	**21.72**
	R@100	14.11	18.27	18.70	**24.38**
SGGen	R@50	0.32	2.70	3.03	**3.44**
	R@100	0.47	3.42	3.71	**4.24**

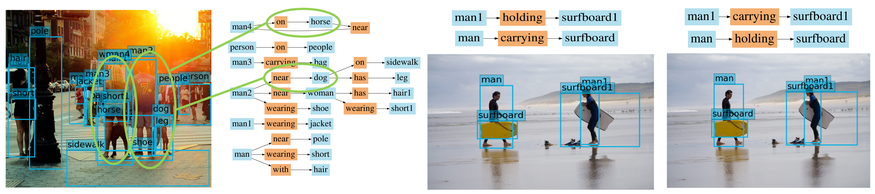


Comparison of our model (final) and Lu et al., ECCV'16 [2] and two other baseline architectures.

Sample Predictions



Failure modes



References

[1] J. Johnson, R. Krishna, M. Stark, L. J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei. Image retrieval using scene graphs. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[2] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In ECCV, 2016.

[3] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr. Conditional random fields as recurrent neural networks. In International Confer- ence on Computer Vision (ICCV), 2015.

[4] R. Krishna, Y. Zhu, O. Groth, J.Johnson, K.Hata, J.Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bern- stein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In arXiv, 2016.

[5] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In ECCV, 2012.