

Pixels to Graphs by Associative Embedding

Credit: Alejandro Newell, Jia Deng

University of Michigan, Ann Arbor

Presenter: Fuwen Tan

<https://qdata.github.io/deep2Read>

Scene graph

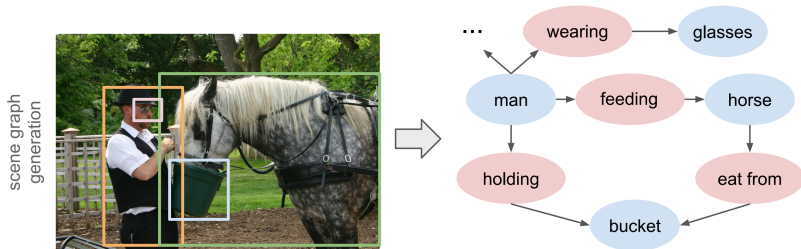


Figure: A visually-grounded scene graph that captures the objects in the image (blue nodes) and their pairwise relationships (red nodes). Credit: IMP [5]

- **SGGen**: Detect and classify all objects and determine the relationships between them;
- **SGCIs**: Ground-truth object boxes are provided, classify them and determine their relationships;
- **PredCIs**: Boxes and classes are provided for all objects, predict their relationships.

- **Knowledge base for downstream applications:**
 - Image retrieval [2]
 - Image synthesis [1]
 - VQA: <https://cs.stanford.edu/people/dorarad/gqa/>
- **Data compression (potentially)**

An end-to-end approach

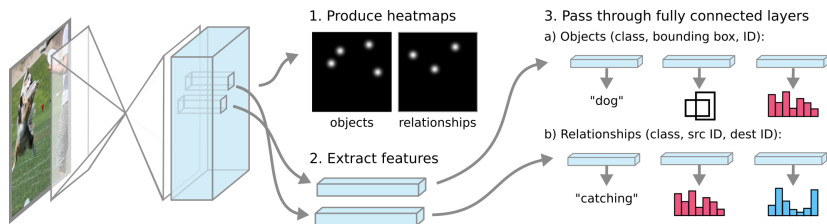


Figure: A network is trained to produce two heatmaps that activate at the predicted locations of objects and relationships. Feature vectors are extracted from the pixel locations of top activations and fed through fully connected networks to predict object and relationship properties. Embeddings produced at this step serve as IDs allowing detections to refer to each other.

How to associate the detected objects with the detected edges

$$L_{pull} = \frac{1}{\sum_{i=1}^n K_i} \sum_{i=1}^n \sum_{k=1}^{K_i} (h_i - h'_{ik})^2$$

$$L_{push} = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \max(0, m - \|h_i - h_j\|)$$

$h_i \in \mathbb{R}^d$: embedding vector for vertex i

$h'_{ik} \in \mathbb{R}^d$: embedding vector for edge (i, k)

- Visual Genome [3]:
 - Object category: 150
 - Edge category: 50
- Evaluation metric: a tuple (subject, predicate, object) is correct if
 - Object localization: IoU > 0.5 ;
 - Objects and relationships are correctly classified.

Result

	SGGen (no RPN)		SGGen (w/ RPN)		SGCls		PredCls	
	R@50	R@100	R@50	R@100	R@50	R@100	R@50	R@100
Lu et al. [4]	–	–	0.3	0.5	11.8	14.1	27.9	35.0
Xu et al. [5]	–	–	3.4	4.2	21.7	24.4	44.8	53.0
Pixel2Graph	6.7	7.8	9.7	11.3	26.5	30.0	68.0	75.2
Pixel2Graph (03/2018)	15.5	18.8	–	–	35.7	38.4	82.0	86.4

Table: Results on Visual Genome. Updated numbers (*bottom row*) are the result of longer training with more efficient code.

Take-home messages

- Main contribution: end-to-end;
- Scene graph generation is still a very challenging problem.

 Justin Johnson, Agrim Gupta, and Li Fei-Fei.


Image generation from scene graphs.

In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

 Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei.

Image retrieval using scene graphs.

In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

 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.

 Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei.

Visual relationship detection with language priors.

In *European Conference on Computer Vision (ECCV)*, 2016. 



Danfei Xu, Yuke Zhu, Christopher Choy, and Li Fei-Fei.

Scene graph generation by iterative message passing.

In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*,
2017.