Image Retrieval using Scene Graphs

Justin Johnson[1], Ranjay Krishna[1], Michael Stark[2], Li-Jia Li[3,4], David A. Shamma[3], Michael S. Bernstein[1], Li Fei-Fei[1]
[1]Stanford University, [2]Max Planck Institute for Informatics, [3]Yahoo Labs, [4]Snapchat

This paper develops a novel framework for semantic image retrieval based on the notion of a scene graph. Our scene graphs represent objects ("man", "boat"), attributes of objects ("boat is white") and relationships between objects ("man standing on boat"). We use these scene graphs as queries to retrieve semantically related images. To this end, we design a conditional random field model that reasons about possible groundings of scene graphs to test images. The likelihoods of these groundings are used as ranking scores for retrieval. We introduce a novel dataset of 5,000 human-generated scene graphs grounded to images and use this dataset to evaluate our method for image retrieval. In particular, we evaluate retrieval using full scene graphs and small scene subgraphs, and show that our method outperforms retrieval methods that use only objects or low-level image features. In addition, we show that our full model can be used to improve object localization compared to baseline methods.

1 Real-World Scene Graphs Dataset

To use scene graphs as queries for image retrieval, we need many examples of scene graphs grounded to images. To our knowledge no such dataset exists. To this end, we introduce a novel dataset of *real-world scene graphs*, which is freely available at the first author's website.

We selected 5,000 images from the intersection of the YFCC100m [3] and Microsoft COCO [2] datasets. For each of these images, we use Amazon's Mechanical Turk (AMT) to produce a human-generated scene graph. For each image, three workers write (object, attribute) and (object, relationship, object) tuples using an open vocabulary to describe the image, and draw bounding boxes for all objects. An example scene graph can be seen in Figure 1. Our full dataset of 5,000 images contains over 93,000 object instances, 110,000 instances of attributes, and 112,000 instances of relationships.

2 Model

We wish to use a scene graph as a query to retrieve images portraying scenes similar to the one described by the graph. To do so, we need to measure the agreement between a query scene graph and an unannotated test image. We assume that this agreement can be determined by examining the best possible grounding of the scene graph to the image.

To this end we construct a conditional random field (CRF) that models the distribution over all possible groundings. We perform maximum a posteriori (MAP) inference to find the most likely grounding; the likelihood of this MAP solution is taken as the score measuring the agreement between the scene graph and the image.

The unary potentials of our CRF measure the degree to which image regions agree with the known object classes and attributes of the objects of the scene graph; these are modeled by training R-CNN [1] detectors for all object classes and attributes in our dataset.

The binary potentials of our CRF measure the degree to which a pair of image regions express the known relationships between the objects of the scene graph.

3 Experiments

We perform image retrieval experiments using two types of scene graphs as queries. First, we use full ground-truth scene graphs as queries; this shows that our model can effectively make sense of extremely precise descriptions to retrieve images. Second, we jump to the other end of the query complexity spectrum and use extremely simple scene graphs as queries; this shows that our model is flexible enough to retrieve relevant images when presented with more open-ended and human-interpretable queries. In addition, we directly evaluate the groundings found by our model and show that our model is able to take advantage of scene context to improve object localization.

In all experiments we compare our full model (SG-obj-attr-rel) with ablated versions of our model that only consider objects (SG-obj) or only objects and attributes (SG-obj-attr). For the partial scene graph retrieval experiment we also compare with baselines based only on image features. The results of all experiments are shown in Table 1.

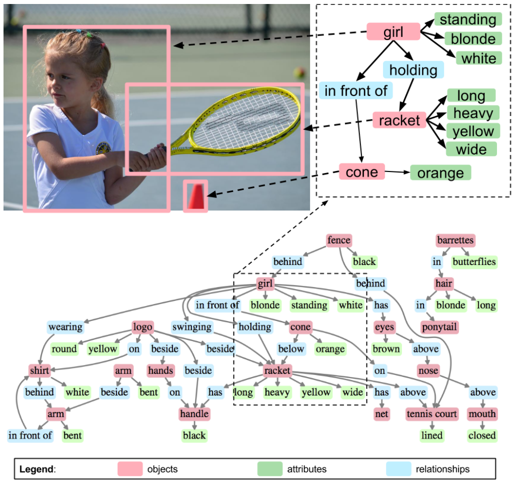

Figure 1: An example of a scene graph (bottom) and a grounding (top). The scene graph encodes objects ("girl"), attributes, ("girl is blonde"), and relationships ("girl holding racket"). The grounding associates each object of the scene graph to a region of an image. The image, scene graph, and grounding are drawn from our *real-world scene graphs* dataset.

		Rand	SIFT	GIST	CNN	SG-obj	SG-obj-attr	SG-obj-attr-rel
(a)	Med r	420	-	-	-	28	17.5	**14**
	R@1	0	-	-	-	0.113	0.127	**0.133**
	R@5	0.007	-	-	-	0.260	**0.340**	0.307
	R@10	0.027	-	-	-	0.347	0.420	**0.433**
(b)	Med r	94	64	57	36	17	12	**11**
	R@1	0	0	0.008	0.017	0.059	0.042	**0.109**
	R@5	0.034	0.084	0.101	0.050	0.269	0.294	**0.303**
	R@10	0.042	0.168	0.193	0.176	0.412	**0.479**	**0.479**
(c)	Med IoU	-	-	-	-	0.014	0.026	**0.067**
	R@0.1	-	-	-	-	0.435	0.447	**0.476**
	R@0.3	-	-	-	-	0.334	0.341	**0.357**
	R@0.5	-	-	-	-	0.234	0.234	**0.239**

Table 1: Quantitative results in entire scene retrieval (a), partial scene retrieval (b), and object localization (c).

[1] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition*, 2014.

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014.

[3] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, 2015.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.