Image retrieval using scene graphs

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

Ganga Meghanath
EE15B025

Stanford University, Max Planck Institute for Informatics, Yahoo Labs, Snapchat

The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3668-3678
(Cited by 212)

April 27, 2019

Schedule

1 Problem Description

2 Requirements
- I/O
- Dataset

3 Analytics

4 Results

5 Criticism

Problem Description

Problem Statement

- Retrieving images by describing their contents (objects, structured relationships and attributes)
- Utilize the structured nature of the query; perfect recognition of detailed semantics

- **Main challenges** :
  - Interactions between objects in a scene can be highly complex
  - Assumption of a closed universe (all classes are known beforehand) : does not hold

Aim : Semantic image retrieval using a scene graph

- **Scene Graph** explicitly models
    - objects
    - attributes of objects
    - relationships b/w objects
- **Idea** : Perform semantic image retrieval using scene graphs as queries
    - Scene graph used to represent the detailed semantics of a scene
    - Design Conditional Random Field model (CRF model of visual scenes) grounding scene graphs to images
    - The likelihoods - used as ranking scores for retrieval

Requirements

Input - Output

- **Input Query** : Scene Graph
- **Output** : Ranked retrieved images (in decreasing order of relevance)

Real-World Scene Graphs Dataset

- Introduced a novel dataset of 5,000 human-generated scene graphs grounded to images; used to evaluate the method for image retrieval.
- The dataset consists of :
  - 5,000 images
  - Over 93,000 object instances
  - 110,000 instances of attributes
  - 112,000 instances of relationships
- Per image statistics :
  - Mean of 13.8 objects
  - Mean of 18.9 attributes
  - Mean of 21.9 relationships

Real-World Scene Graphs Dataset

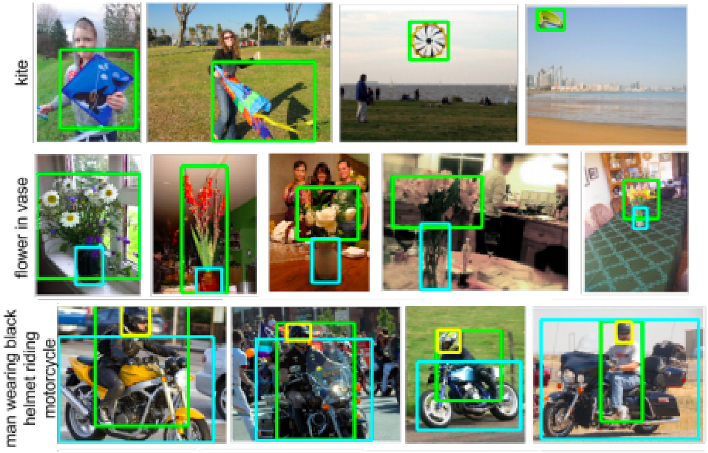

Figure: [1]Examples of scene sub-graphs of increasing complexity (top to bottom) from the Real-World Scene Graphs dataset, with attributes & different objects.

Analytics

Scene Graph and Grounding

- **Scene Graph** :
  - A data structure that describes the contents of a scene
  - Given a set of object classes C, a set of attribute types A, and a set of relationship types R, a scene graph $\implies$ a tuple $G = (O, E)$
    - $O = \{o_1, ..., o_n\}$ is a set of objects
    - $E \subseteq O \times R \times O$ is a set of edges
    - Each object, $o_i = (c_i, A_i)$
    - $c_i \in C$ is the class of the object
    - $A_i A$ are the attributes of the object
- **Grounding a scene graph in an image**
  - A scene graph can be grounded to an image by associating each object instance of the scene graph to a region in an image.
  - Given a scene graph and an image, there are many possible ways of grounding the scene graph to the image.
  - A grounding of a scene graph $G = (O, E)$ is a map $\gamma : O \to B$ where an image is represented by a set of candidate bounding boxes B.

Depiction : Scene Graph and Grounding

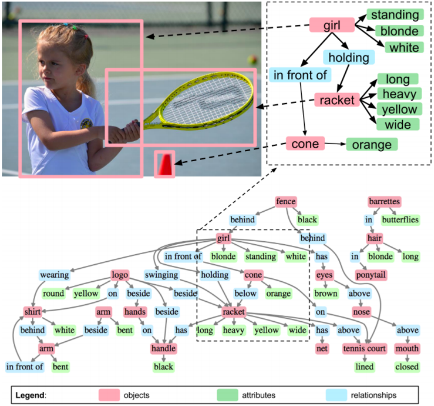

Figure: [1]An example of a scene graph (bottom) and a grounding (top). The scene graph encodes objects (“girl”), attributes, (“girl is blonde”), and relationships (“girl holding racket”). The grounding associates each object of the scene graph to a region of an image.

Image Retrieval by Scene Graph Grounding

- Use a scene graph as a query to retrieve images portraying scenes similar to the one described by the graph
- Measure the agreement between a query scene graph and an unannotated test image
- Conditional Random Field (CRF) models the distribution over all possible groundings
- MAP (Maximum a posteriori) estimate used to find the most likely grounding (measure of agreement between the scene graph and the image)

Image Retrieval by Scene Graph Grounding

- Use GOP(Geodesic Object Proposals)[2] for generating candidate boxes for images (provides the best trade-off between object recall and number of regions per image).
- **CRF formulation** :
  Distribution over possible groundings,

$$P(\gamma|G,B) = \prod_{o\in O} P(\gamma_o|o) \prod_{(o,r,o')\in E} P(\gamma_o,\gamma_{o'}|o,r,o')$$

$$\gamma^\star = \underset{\gamma}{argmax} \prod_{o\in O} P(o|\gamma_o) \prod_{(o,r,o')\in E} P(\gamma_o,\gamma_{o'}|o,r,o')$$

where $P(\gamma_o|o) = P(o|\gamma_o)\frac{P(\gamma_o)}{P(o)}$ by Bayes rule
(Uniform prior $\implies$ $P(\gamma_o), P(o)$ are constants)

Image Retrieval by Scene Graph Grounding

- **Unary potentials** :
  - $P(o|\gamma_o)$ models how well the appearance of the box γ_o agrees with the known object class and attributes of the object o.
  - $$P(o|\gamma_o) = P(c|\gamma_o) \prod_{a \in A} P(a|\gamma_o)$$
  - R-CNN is used to train the detectors for each of the $|C| = 266$ and $|A| = 145$ object classes and attribute types.
  - Platt scaling is used to convert the SVM classification scores for each object class and attribute into probabilities.

Image Retrieval by Scene Graph Grounding

- **Binary potentials** :
  - $P(\gamma_o, \gamma_{o'} | o, r, o')$ models how well the pair of bounding boxes γ_o, $\gamma_{o'}$ express the tuple (o, r, o')
  - $$f(\gamma_o, \gamma_{o'}) = \Big((x - x')/w, (y - y')/h, w'/w, h'/h\Big)$$

    where $\gamma_o = (x, y, w, h)$ and $\gamma_{o'} = (x', y', w', h')$ are the coordinates of the bounding boxes in the image
  - Train a Gaussian mixture model (GMM) to model $P(f(\gamma_o, \gamma_{o'})|c, r, c')$ where c and c' are classes of objects o and o'.
  - Fewer than 30 instances of (c, r, c') in the training data *implies* use object agnostic model $P(f(\gamma_o, \gamma_{o'})|r)$
  - Use Platt scaling to convert the value of the GMM density function evaluated at $f(\gamma_o, \gamma_{o'})$ to a probability $P(\gamma_o, \gamma_{o'}|o, r, o')$

Methods used in Experiments

- **SG-obj-attr-rel**: Model uses unary object and attribute potentials and binary relationship potentials.
- **SG-obj-attr**: Model uses only object and attribute potentials.
- **SG-obj**: Model uses only object potentials (object class potentials are rescaled R-CNN detection scores).
- **CNN**: L2 distance between the last layer features extracted using the reference model.
- **GIST**[4]: L2 distance between the GIST descriptors of the query image and each test image.
- **SIFT**[3]
- **Random**: A random permutation of the test images.

Results

Results

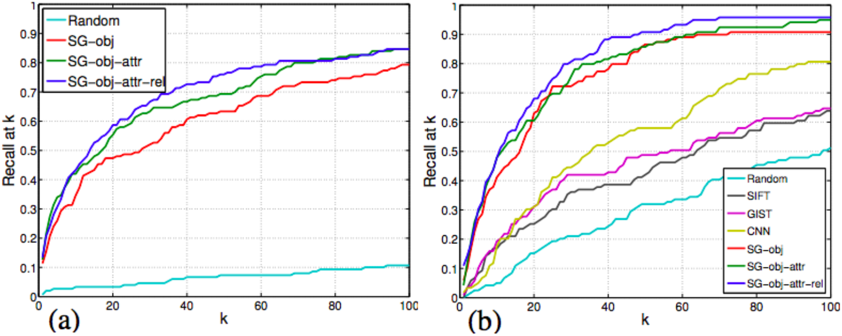

Figure: [1](a) Retrieval performance for entire scenes and (b) for partial scenes.

Results

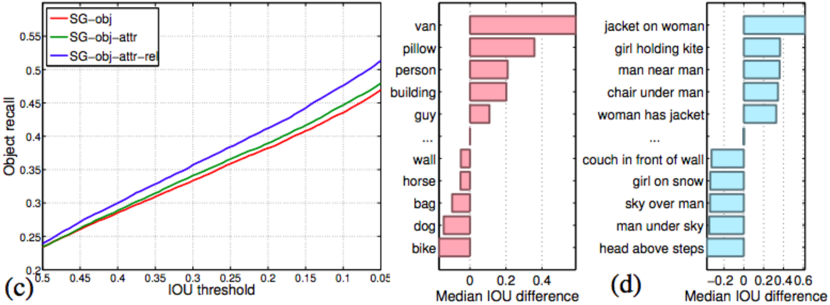

Figure: [1](c) Object localization performance for entire scenes. (d) Increase in localization performance of our full model SG-obj-attr-rel vs SG-obj for individual objects (left) and objects participating in a relation (right). In (d), positive values indicate the SG-obj-attr-rel performs better than SG-obj

Results

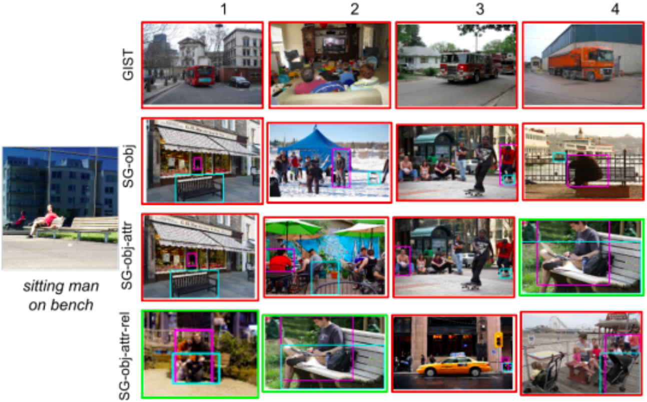

Figure: [1]Top-4 retrieval results returned by different methods using a partial scene graph queries

Results

- Evaluation of image retrieval using full scene graphs and small scene subgraphs, shows that this outperforms retrieval methods that use only objects or low-level image features
- The full model can be used to improve object localization compared to baseline methods.
- Context provided by SG-obj-attr-rel helps to localize rare objects and objects with large appearance variations.
- SG-obj-attr-rel gives large gains for tuples with well-defined spatial constraints.

Criticism

Novelty

- Dataset : 5,000 human-annotated scene graphs grounded to images that use an open-world vocabulary to describe images in greater detail
- CRF model : semantic image retrieval using scene graph queries. Model outperforms baseline models that reason only about objects, and simple content-based image retrieval methods based on low-level visual features

Pros

- Replacing textual queries with scene graphs allows our queries to describe the semantics of the desired image in precise detail without relying on unstructured text.
- Using paragraph description for image retrieval (instead of scene graphs) would require the need to resolve co-references in the text, perform relationship extraction to convert the unstructured text into structured tuples, and ground the entities of the tuples into regions of the image described by the text.
- Can model multiple modes of interaction between pairs of objects while traditional CRF models are more restricted, and encode a fixed relation given two nodes.
- **Dataset** : contains significantly more labeled object instances per image than existing datasets, and also provides annotated attributes and relationships for individual object instances.

Cons

- Requires the creation of a scene graph (Converting a description into a scene graph for query).
- Tuples encoding less well-defined spatial relationships may suffer in performance as the model penalizes valid configurations not seen at training time.
- Creation of training data is a cumbersome process (For each image : produce scene graph (using Amazon's Mechanical Turk - AMT), write (object, attribute) and (object, relationship, object) tuples using an open vocabulary to describe the image, draw bounding boxes for all objects, correct and verify the bounding boxes).

References

- Justin Johnson et al. “Image Retrieval Using Scene Graphs”. In: **The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. 2015.
- Philipp Krähenbühl and Vladlen Koltun. “Geodesic object proposals”. In: **European conference on computer vision**. Springer. 2014, pp. 725–739.
- David G Lowe. “Distinctive image features from scale-invariant keypoints”. In: **International journal of computer vision** 60.2 (2004), pp. 91–110.
- Aude Oliva and Antonio Torralba. “Modeling the shape of the scene: A holistic representation of the spatial envelope”. In: **International journal of computer vision** 42.3 (2001), pp. 145–175.

THANK YOU :)