Visual Genome

Ranjay Krishna et al.

IJCV, 2016

Christopher Painter (cpainter@seas.upenn.edu)
February 13th, 2019

Problem & Motivation

- Computer vision tasks typically focus on narrow sub-tasks of what humans actually do when observing an image.
  - Object classification (e.g. “this is a cat”, “this is a dog”)
  - Object detection (e.g. “this part of the image is a cat”)
  - Object generation (e.g. “a cat looks like this”)
- But humans do more than this

Problem & Motivation

- Example:

Problem and Motivation

- Another (low-resolution) Example:

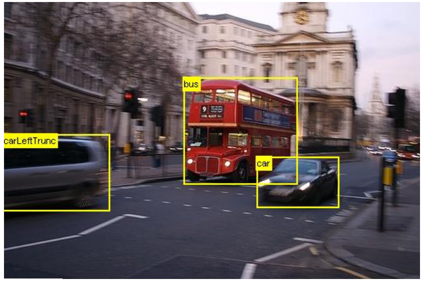
Problem & Motivation

- Existing image understanding datasets allow for the formulation of object classification, detection, and description tasks on salient elements of an image, but not comprehensive scene understanding
- Seeks to contribute three missing elements to the state of the art for image understanding datasets
  - Grounding of visual concepts to language
  - Complete set of descriptions and QAs
  - Formalized representation of image components

Contents:

Previous approaches

- Datasets have historically been made for a narrow task
  - *Caltech 101* was the first large-scale image dataset for image classification, with 101 categories and 15-30 examples in each
  - *Pascal VOC* (Everingham et al., 2010) shifted from object classification to object detection
  - *Imagenet* (Deng et al., 2009) crowdsourced a corpus of 14 million images for Wordnet synsets.



Previous approaches

- MS-COCO (Lin et al., 2014)
- VQA (Antol et al., 2015)

What color are her eyes?
What is the mustache made of?

Previous approaches

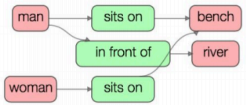
- “Deep Visual-Semantic Alignments for Generating Image Descriptions”, Karpathy et al., CVPR 2015

man in black shirt is playing guitar. | construction worker in orange safety vest is working on road. | two young girls are playing with lego toy. | boy is doing backflip on wakeboard.

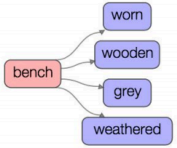
The Visual Genome is a Hybrid

- Why create this dataset? We won't be satisfied that an "Artificial Intelligence" understands an image until it can give us many descriptions of it ("a picture is worth a thousand words").
- How does it do this? It integrates elements of existing datasets and adds some new ones:
  - Places equal emphasis on object relationships and attributes as it does on objects themselves
  - Brings the utility of knowledge representation in NLP to image descriptions, formalizing image descriptions
  - Captures image "narratives" that are non-salient e.g. subplots occurring in the background of images

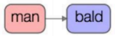
Scene Graphs



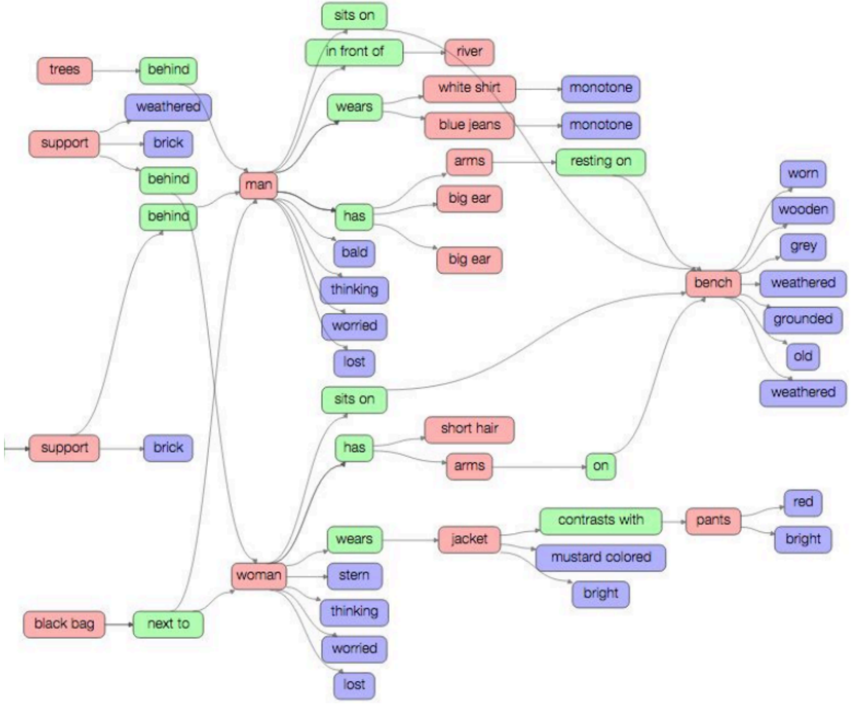
A man and a woman sit on a park bench along a river.



Park bench is made of gray weathered wood

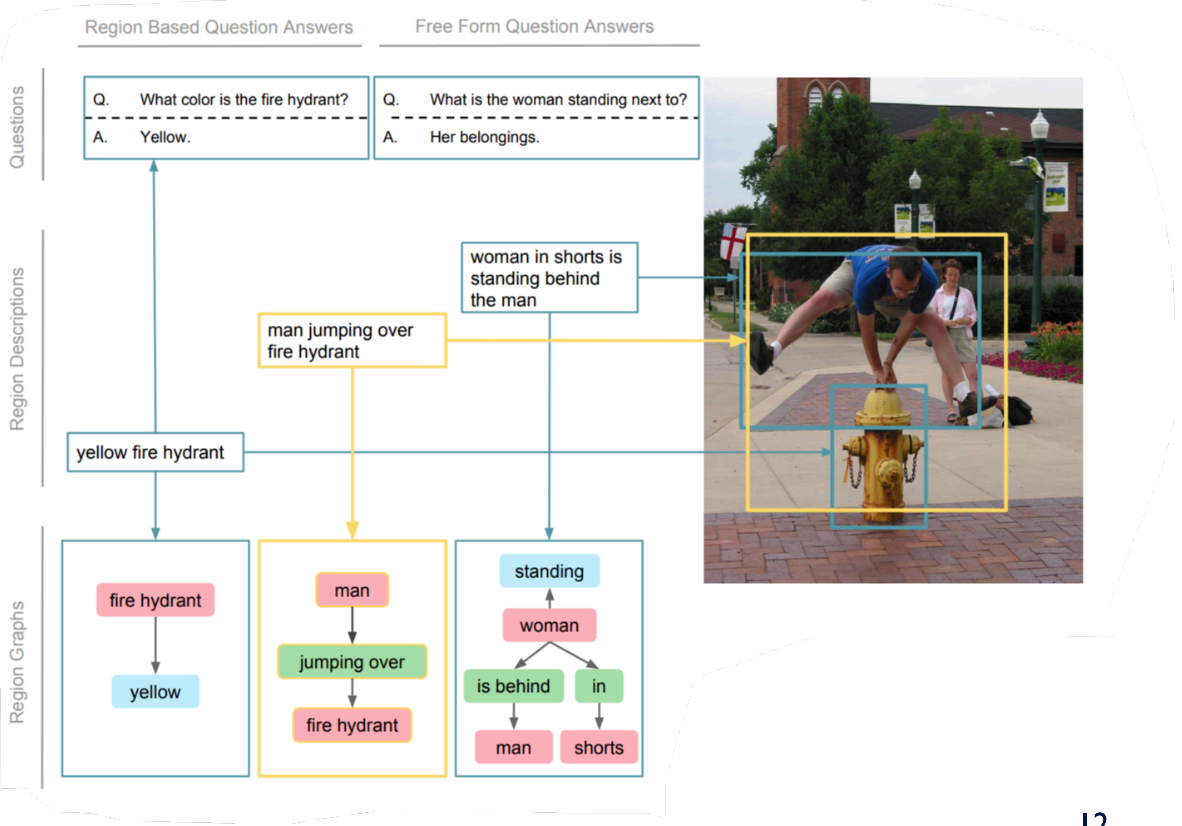


The man is almost bald



Visual Genome Data Representation

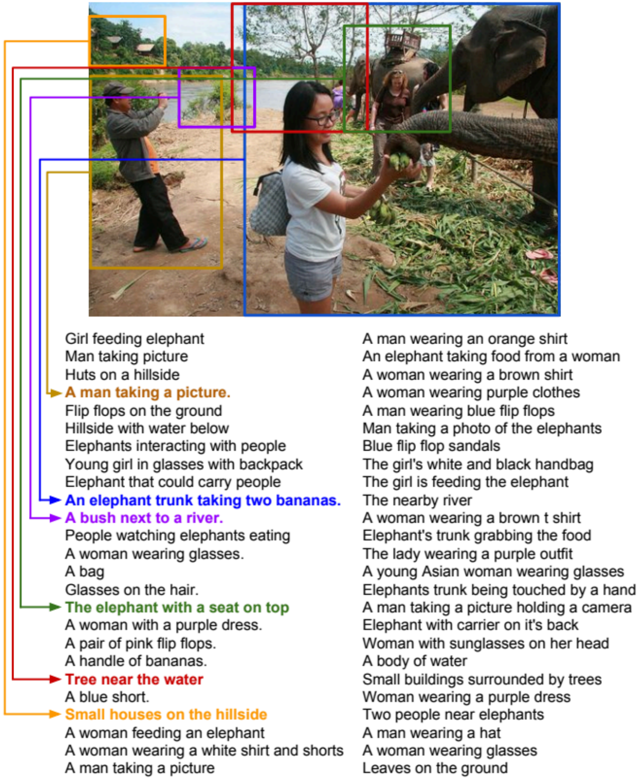
- What's in it?
  - Multiple region descriptions
  - Objects with bounding boxes
  - Object attributes
  - Object relationships
  - Region graphs
  - Scene graph (the union of all region graphs
  - QA Pairs



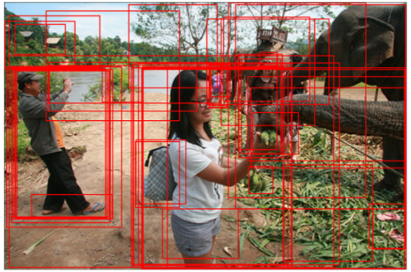
Preparation

- The Visual Genome was constructed through a series of Amazon Mechanical Turk tasks.
- Images prepared in order of: 1. Region descriptions 2. Objects 3. Attributes, Relationships, Region Graphs 4. Scene Graphs 5. Q&A
- Restrictions were placed on the Turks so as to improve dataset quality
  - Region bounding boxes were evaluated based on coverage, or whether they incorporated **at least** all of the region being described, while object bounding boxes also had to be **tight**
  - Descriptions were checked for dissimilarity to other image specific descriptions and globally common descriptions using BLEU scoring

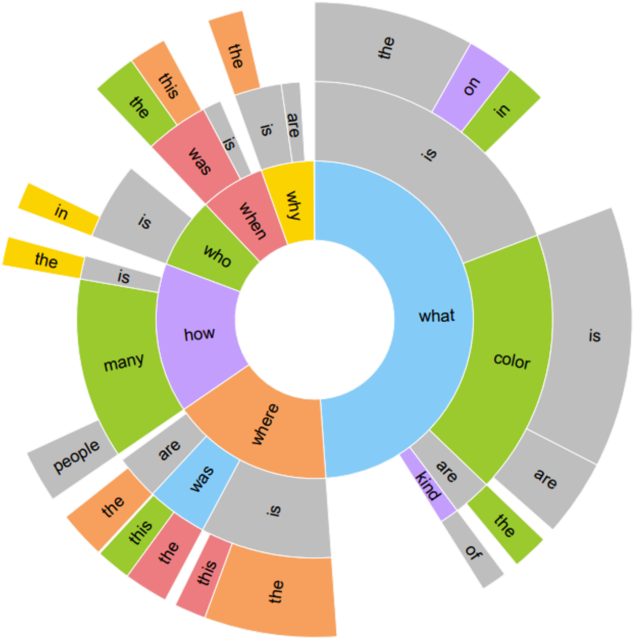
Result



- Regions have between 0 and 2 objects
- Images have 15 to 20 objects
- Each image has an average of 42 descriptions.
- Regions roughly correspond to relationships between objects
- Visual Genome descriptions sample more “image description clusters” than those of MS-COCO



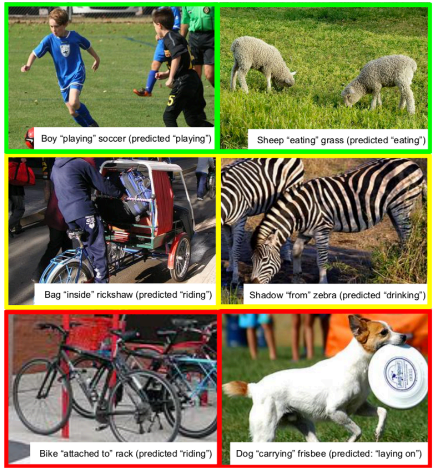
Dataset Properties- Questions and Answers

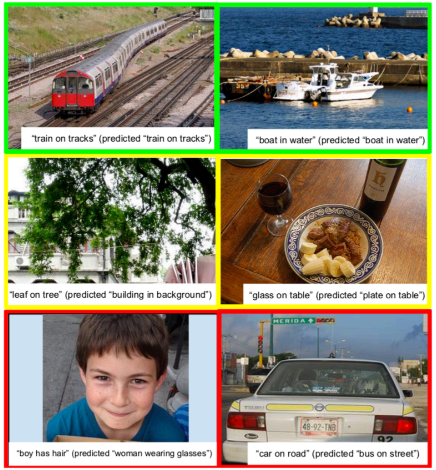


Experimental Results

- Struggles with “Why”, “Where”, and object-relationship association
- The dataset lends itself to innovation in:
  - Dense image captioning
  - Image understanding
  - Semantic image retrieval

	top-100	top-500	top-1000
What	0.420	0.602	0.672
Where	0.096	0.324	0.418
When	0.714	0.809	0.834
Who	0.355	0.493	0.605
Why	0.034	0.118	0.187
How	0.780	0.827	0.846
Overall	0.411	0.573	0.641





Penn Engineering

Takeaways

- Many of the most common image datasets do not lend themselves to questions of holistic image understanding.
- Datasets can make a massive difference in the type of questions that researchers can ask, and the efficiency with which they can ask them.
- Datasets that integrate different types of helpful data structure in one package are uniquely useful