Notes of "Visual Genome"

Jihang Li

January 21, 2020

Contents

Abstract

To achieve success at cognitive tasks, models need to understand the interactions and relationships between objects in an image.

The Visual Genome (VG) dataset collects dense annotations of objects, attributes, and relationships to help modeling such as identifying objects and relationships *riding(man, carriage)* when asked "what vehicle is the person riding?". VG contains over 108K images where each image has an average of 35 objects, 26 attributes, and 21 pairwise relationships between objects. The objects, attributes, relationships, and noun phrases in region descriptions and questions answer pairs to WordNet synsets.

3 Related Work

3.2 Image Descriptions

Most work related to describing images can be divided into two categories[1]:

- Retrieval of human-generated captions: methods in this category use similarity metrics between image features from predefined models to retrieve similar sentences.
- Generation of novel captions: methods in this category common theme has been to use recurrent neural networks to produce novel captions.

One drawback of these approaches is their attention to describing only the most salient aspect of the image.

VG pushes toward a more complete understanding of an image by collecting a dataset in which we capture not just scene-level descriptions but also myriad of low-level descriptions, the "grammar" of the scene.

3.3 Objects

Images in early datasets are iconic and do not capture the settings in which these objects usually co-occur. To remedy this problem, MS-COCO annotated real-world scenes that capture object contexts. However, MS-COCO was unable to describe all the objects in its images, since they annotated only 80 object categories.

VG aims at collecting annotations for **all visual elements** that occur in images, increasing the number of distinct categories to 33887.

3.4 Attributes

Even if we haven't seen an object before, attributes allow us to infer something about it; for example, "yellow and brown spotted with long neck" likely refers to a giraffe.

VG uses a generalized formulation (Johnson *et al.*, 2015) and extends it such that attributes are not image-specific binaries but rather object-specific for each object in a real-world scene. The types of attributes are also extended to include size, pose, state (*e.g.* "transparent"), emotion, and many more.

3.5 Relationships

Relationships have already shown their utility in improving visual cognitive tasks (Antol *et al.*, 2014, Yang *et al.*, 2012).

Relationships in a structured representation with objects have been defined as a graph structure called a **scene graph**, where the nodes are objects with attributes and edges are relationships between objects.

[1]For detailed related papers please read the original paper.

3.6 Question Answering

In previous datasets, most questions concentrated on simple recognition-based questions about the salient objects, and answers were often extremely short. VG aims to capture the details of the images with diverse question types and long answers. These questions should cover a wide range of visual tasks from basic perception to complex reasoning. The QA dataset of 1.7 million QAs is also larger than any currently existing dataset.

3.7 Knowledge Representation

A knowledge representation of the visual world is capable of tackling an array of vision tasks, from action recognition to general question answering.

VG attempts to learn common-sense relationships from images. VG scene graphs can also be considered a **dense** knowledge representations for images.

4 Crowdsourcing Strategies

4.2 Region Descriptions

Crowd workers naturally start with the most salient part of the image and then move to describing other parts of the image one by one. Inspired by this finding, we focused our attention towards collecting a dataset of region descriptions that is diverse in content.

4.7 Verification

Verification is conducted using two separate strategies[2]:

- majority voting
- rapid judgments

7 Future Applications and Directions

Image Understanding While there is a surge of image captioning (Kiros *et al.*, 2014) and question answering (Antol *et al.*, 2015) models, there has been little work on creating more comprehensive evaluation metrics to measure how well there models are performing. Such models are usually evaluated using BLEU, CIDEr, or METEOR and other similar metrics that **do not effectively** measure how well these models understand the images (Chen *et al.*, 2015).

Completing the Set of Annotations While VG is the most densely annotated visual dataset for cognitive image understanding, it is still not complete. **In most images, it is not feasible to collect an exhaustive set of attributes and relationships for every object or pair of objects.** This raises two new research questions:

[2]Both require manual verification.

1. Develop new evaluation metrics that do not penalize models due to a lack of a complete set of annotations.
2. Need to design new interfaces and workflows that incentivize humans to annotate visual common sense.