From an Image to a Description

Margaret Mitchell
University of Aberdeen
Computing Science Department
Aberdeen, Scotland, U.K.
m.mitchell@abdn.ac.uk

Computer vision is noisy. For most images, correct recognition of the image content does not work well. Or perhaps it works too well: Computer vision can find hats and cat faces and people in an image that to a human observer is just a picture of a mountain.

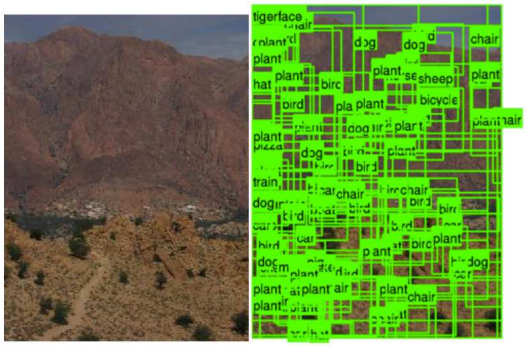

Figure 1: Output of running several object classifiers on a novel image.

This is not to say that computer vision does not work well when some constraints are in place. Object identification – when the system is told what kind of object to look for – achieves relatively high accuracy (Dalal and Triggs, 2005). Object segmentation – when there is not too much clutter or lighting variation – can also work relatively well (Friedland et al., 2005). Using data sets that traditionally inform the development of computer vision algorithms, such as the Pascal object recognition database (Everingham et al., 2010), may also achieve reasonable accuracy compared to novel images (Ponce et al., 2006).

However, in unconstrained, real-world situations, a computer vision system will tend not to make sense of the world. Until the state of the art improves, a system that links vision to language must rely on further semantic knowledge to constrain

Figure 2: Output of running an object classifier for a given object on a novel image.

what the vision system sees. If we can learn that cats, dogs, and sofas tend to be together; or that lions, cages, and sky tends to be together; then we can filter the output of a vision system to a smaller subset of likely objects, and generate language in an iterative procedure that checks what the vision system sees and what a semantic model expects until consensus is reached between both systems.

One way to learn how to talk about the visual world and what kinds of things we expect to see is by learning from examples of descriptive text, such as image captions obtained from Flickr (Flickr, 2011). From here, we can map language-based information to vision-based information. This affects several aspects of language generation:

1. Visual constraints: The visual world can roughly be characterized as consisting of objects (like people, sofas and dogs); "stuff" (like grass, water, and sky) (Kulkarni et al., 2011); spatial relations, the distance and placement of different objects and stuff relative to one another; poses/actions, estimated from the form and positions of given objects;

and features,[1] such as the color `red` and the material `wooden`. By representing recognized objects, stuff, poses/actions and features within an object's bounding box as sets of <ATTRIBUTE:`value`> pairs (e.g., <COLOR:`red`>), we can begin to connect visual output directly to language generation input. Initial work on connecting spatial relations and relative attributes such as size to generation suggests that for some properties, utilizing a *vector* of features instead of a single-featured `value` helps to generate further nuanced language for each property (Mitchell et al., 2011).

2. Syntactic constraints: A recognized object or stuff can be translated to a noun with syntactic constraints. Such constraints include whether or not the noun usually takes a determiner or not (corresponding to the count/mass noun distinction), and what kind of determiner it tends to take (corresponding to the given/new distinction). For example, "rice" is usually a mass noun, appearing without a determiner; "sky" usually appears with "the", and not "a", since "sky" is shared knowledge for any person viewing an image.

Beyond noun phrase constraints, the construction of verb phrases can leverage information available from action and pose detection. Accuracy on this task in novel images is still quite low, however, if we can detect a pose or an action based on likely poses/actions in our semantic model for a given subject, we can flesh out whether an appropriate verb is transitive or intransitive, and what kinds of complements it usually takes.

3. Syntactic-Semantic constraints: A given pair of objects/stuffs can be characterized at the language level as having two kinds of relationships: prepositional relationships and verbal relationships. A BOY can be *on* the TABLE (prepositional phrase); A BOY can also be *cleaning* the TABLE (verb phrase). The kinds of syntactic relationships expected between objects/stuffs is determined in part by their semantics. Boys can clean, but tables can't clean. A duck can be on the lake, but a lake can't be on a duck. Characterizing such semantic properties of objects as they are reflected in syntax can help couple the raw information about the placement and types of objects in a scene to what is happening in a scene.

4. Semantic constraints: One factor that influences what we describe in a scene is what is interesting about the scene. From a semantic standpoint, this can mean drawing from a knowledge base of prototypes to determine what properties are expected for objects/stuffs in the scene and what properties are unexpected. Those that are expected may not be mentioned unless they serve to distinguish an item from a similar confusable item, but those that are unexpected may tend to be mentioned.

More work is necessary to understand how to use language to temper the output of a vision system. But the sum output may be greater than its parts. With information about what different objects do, objects that tend to appear together, common values for attributes of objects, and the kinds of arguments different verbs take, we can begin to connect noisy vision to syntactically and semantically well-formed language structures.

Acknowledgements. Thanks to Jesse Dodge, Karl Stratos, Kota Yamaguchi, Xufeng Han, Alyssa Mensch, Amit Goyal, Hal Daumé III, Alexander Berg, Tamara Berg, and the Johns Hopkins Center for Language and Speech Processing.

References

N. Dalal and B. Triggs. 2005. Histograms of oriented gradients for human detections. *CVPR*.

M. Everingham, L. van Gool, and C. K. I. Williams, et al. 2010. The pascal visual object classes (voc) challenge. *Int. Journal of Computer Vision*, 88(2):303–338.

Flickr. 2011. http://www.flickr.com. Accessed 1.Sep.11.

G. Friedland, K. Jantz, and R. Rojas. 2005. SIOX: simple interactive object extraction in still images. *ISM*.

G. Kulkarni, V. Premraj, and S. Dhar, et al. 2011. Baby talk: Understanding and generating image descriptions. *CVPR*.

M. Mitchell, K. van Deemter, and E. Reiter. 2011. Applying machine learning to the choice of size modifiers. *PRE-CogSci*.

J. Ponce, T. L. Berg, and M. Everingham, et al. 2006. Dataset issues in object recognition. In J. Ponce, et al., editor, *Toward Category-Level Object Recognition*, volume 4170 of *LNCS*. Springer.

[1]The vision community does not usually make an attribute/value distinction, as is done in language work, e.g., "red" is called an *attribute* without a "color" specification. To avoid confusion, I refer to the vision-based "attributes" as features.