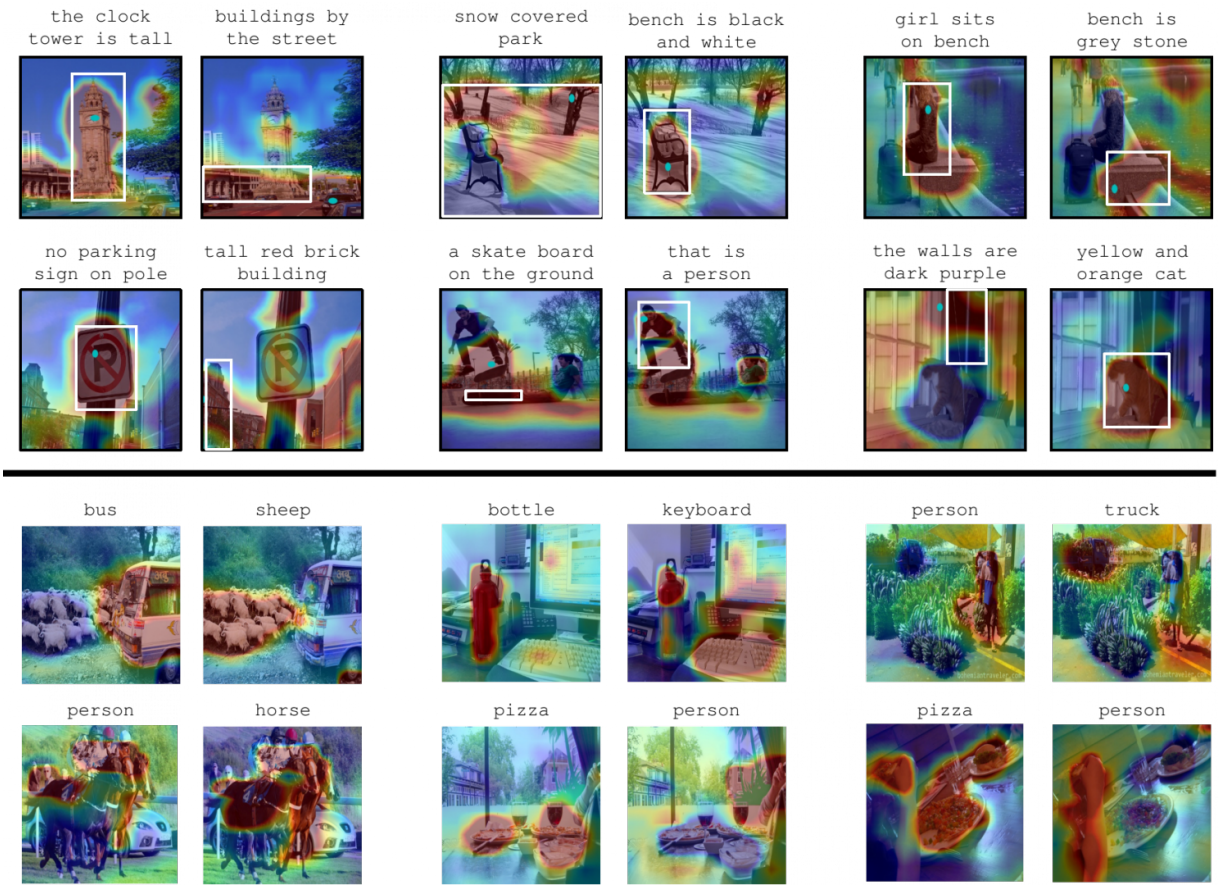
Computers using linguistic clues to deduce photo content

July 21 2017



Credit: Disney Research

Scientists at Disney Research and the University of California, Davis have found that the way a person describes the content of a photo can

provide important clues for computer vision programs to determine where various things appear in the image.

According to Leonid Sigal, a senior research scientist at Disney Research, it's not just the words, but the sentence structure of a caption that can help a computer determine where in an image a particular object or action is depicted. By parsing the sentence and applying deep learning techniques, the computer can use the hierarchy of the sentence to better understand spatial relationships and associate each phrase with the appropriate part of the image.

A neural network based on this approach potentially could automate the process of annotating images that subsequently can be used to train visual recognition programs. The researchers, including Fanyi Xiao and Yong Jae Lee of UC Davis, will present their findings at the IEEE Conference on Computer Vision and Pattern Recognition on July 22 in Honolulu.

"We've seen tremendous progress in the ability of computers to detect and categorize objects, to understand scenes and even to write basic captions, but these capabilities have been developed largely by training computer programs with huge numbers of images that have been carefully and laboriously labeled as to their content," said Markus Gross, vice president at Disney Research. "As computer vision applications tackle increasingly complex problems, creating these large training data sets has become a serious bottleneck."

Using just a little bit of labeled data to generate these large training sets has been a goal of researchers for years and the approach by the Disney and UC Davis scientists may be the first to leverage sentence structure in doing so.

The phrase "a grey cat staring at a hand with a donut," for instance,

suggests that a hand and a donut will appear together while "staring" suggests that the grey cat should be spatially disjointed from the hand with the donut.

Xiao said recognizing these constraints - natural language that indicates which things are together and which are apart - provides important context that enables the neural network to produce more accurate visual localizations for language inputs at all levels (words, phrase and sentence).

Different language inputs thus will provide different results for the same image. In a photo of a park, the phrase "girl sits on bench" results in the computer highlighting a girl sitting, while "bench is grey stone" highlights just the stone end of the bench, without highlighting the girl.

In testing this approach with existing visual data sets, the researchers showed their system produced more accurate localizations than baseline systems that do not consider the structure of natural language. "While mainstream weakly-supervised localization approaches have used image tags as the source of supervision, our work instead uses captions and is thus able to exploit the rich structure in language. We hope this work will inspire more research in this direction." said Yong Jae.

Combining creativity and innovation, this research continues Disney's rich legacy of leveraging technology to enhance the tools and systems of tomorrow.

More information: "Weakly-Supervised Visual Grounding of Phrases with Linguistic Structures-Paper" [PDF, 2.46 MB]

Provided by Disney Research