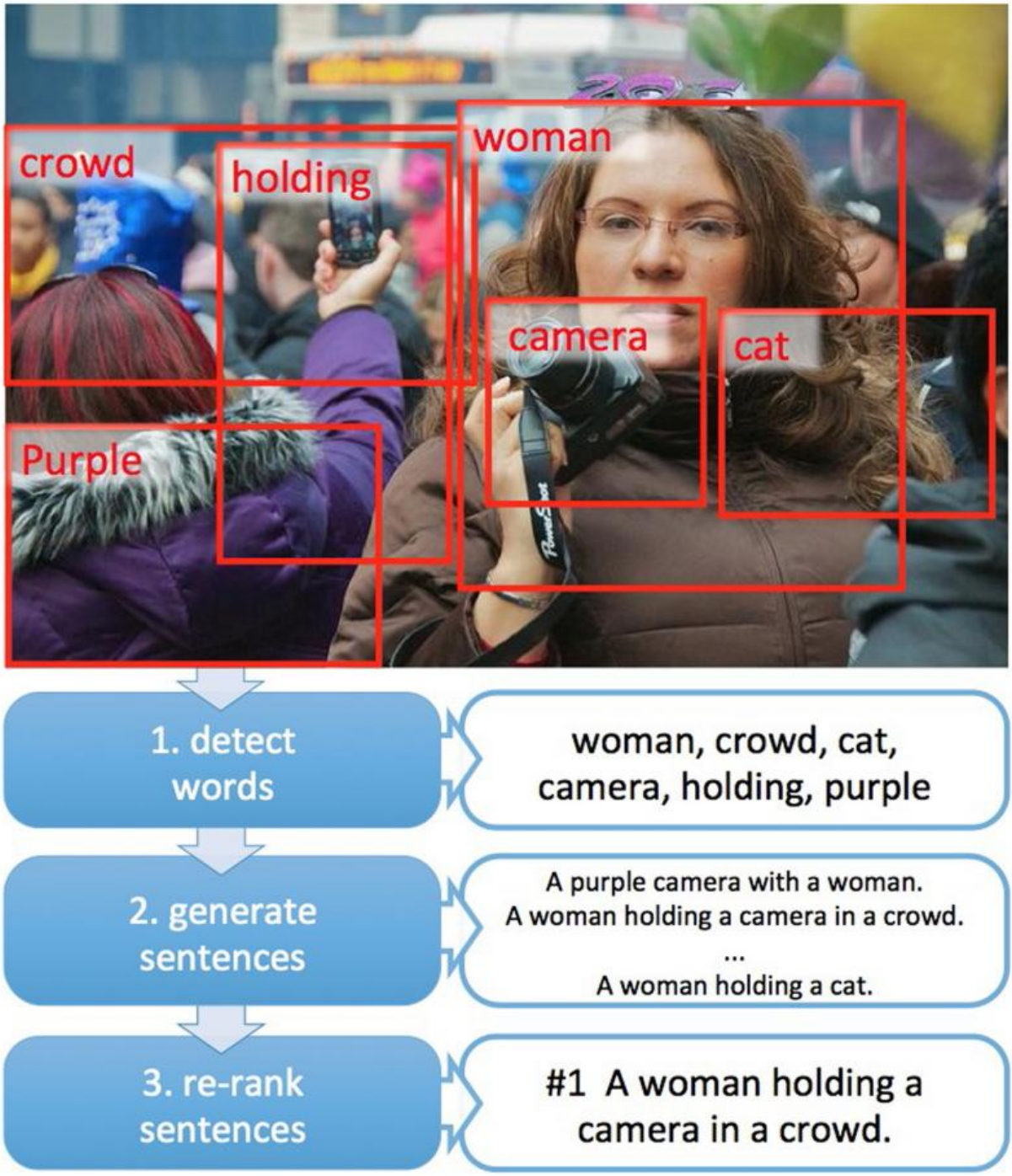
# Microsoft Research project can interpret, caption photos

May 29 2015


crowd
holding
woman
camera
cat
Purple
1. detect words
woman, crowd, cat, camera, holding, purple
2. generate sentences
A purple camera with a woman.
A woman holding a camera in a crowd.
...
A woman holding a cat.
3. re-rank sentences
#1 A woman holding a camera in a crowd.

If you're surfing the web and you come across a photo of the Mariners' Felix Hernandez on the pitchers' mound at Safeco Field, chances are you'll quickly interpret that you are looking at a picture of a baseball player on a field preparing to throw a pitch.

Now, there is technology that can do that, too.

Microsoft researchers are at the forefront of developing technology that can automatically identify the objects in a picture, interpret what is going on and write an accurate caption explaining it.

That's an important tool in and of itself. But the ability for a machine to correctly describe what's going on in a photo also has broader implications for helping Microsoft advance its overall work in the field of artificial intelligence, or the development of systems that can see, hear, speak and even understand.

"The machine has been trained to understand how a human understands the image," said Xiaodong He, a researcher with Microsoft Research's Deep Learning Technology Center and one of the people working on the project.

For example, when given a picture of a man sitting in front of a computer, the image captioning technology can accurately recognize that it should focus on describing the man in the foreground, not the image on the computer in the background. Because the man has facial hair, it also knows that it is a man, not a woman.

For decades, researchers have been tantalized by the possibility of creating systems that could accurately interpret and caption photos. But until a few years ago, most of the systems being developed just weren't getting it right, said Margaret Mitchell, a researcher in Microsoft Research's natural language processing group who also is working on the

technology.

That changed when researchers hit upon the idea of using neural networks, which are computing elements that are modeled loosely after the human brain, to connect vision to language. With that technology, the systems began to get it right more often, and error rates have been decreasing ever since.

"It's basically gone from not working to working because of neural networks," Mitchell said.

Automated image captioning still isn't perfect, but it has quickly become a hot research area, with experts from universities and corporate research labs vying for the best automated image captioning algorithm.

The latest competition to create the most informative and accurate captions, the MS COCO Captioning Challenge 2015, ends this Friday.

Throughout the competition, a leaderboard has been tracking how well the teams are doing using various technical measurements, and ranking them based on who is currently producing the best results. The top performers will have their results evaluated by human judges at the CVPR computer vision conference in early June.

The competitors are all using a dataset of images, called Microsoft COCO, which was developed by researchers from Microsoft and other research institutions. The challenge is to come up with the best algorithm that creates captions based on that dataset.

Microsoft's algorithm is trained to automatically write a caption using several steps.

First, it predicts the words that are likely to appear in a caption, using

what's called a convolutional neural network to recognize what's in the image.

The convolutional neural network is trained with many examples of images and captions, and automatically learns features such as color patches, shapes and other features. That's much like the way the human brain identifies objects.

Next, it uses a language model to take that set of words and create coherent possible captions.

"The critical thing is that the language model is generating text conditioned on the information in the image," said Geoffrey Zweig, who manages Microsoft Research's speech and dialog research group.

Finally, it deploys a checker that measures the overall semantic similarity between the caption and the image, to choose the best possible caption.

As the technology continues to improve, the researchers say they see vast possibilities for how these types of tools could be used to make significant gains in the field of artificial intelligence, in which computers are capable of intelligent behavior in an era of more personal computing.

"We want to connect vision to language because we want to have artificial intelligence tools," Mitchell said.

He, the deep learning researcher, said the technology could serve as a piece of the foundation for much more sophisticated AI tools, such as a universal augmented intelligence system that would constantly be with you, learning about you and the world around you and helping you when needed.

People have been looking forward to those types of capabilities for a long time.

"Now, we are optimistic to see them come true in the foreseeable future," He said.

Provided by Microsoft

Citation: Microsoft Research project can interpret, caption photos (2015, May 29) retrieved 20 September 2024 from https://phys.org/news/2015-05-microsoft-photos.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.