Paying attention to words not just images leads to better image captions

March 17 2016

A team of University and Adobe researchers is outperforming other approaches to creating computer-generated image captions in an international competition. The key to their winning approach? Thinking about words - what they mean and how they fit in a sentence structure - just as much as thinking about the image itself.

The Rochester/Adobe model mixes the two approaches that are often used in image captioning: the "top-down" approach, which starts from the "gist" of the image and then converts it into words, and the "bottom-up" approach, which first assigns words to different aspects of the image and then combines them together to form a sentence.

The Rochester/Adobe model is currently beating Google, Microsoft, Baidu/UCLA, Stanford University, University of California Berkeley, University of Toronto/Montreal, and others to top the leaderboard in an image captioning competition run by Microsoft, called the Microsoft COCO Image Captioning Challenge. While the winner of the year-long competition is still to be determined, the Rochester "Attention" system - or ATT on the leaderboard - has been leading the field since last November.

Other groups have also tried to combine these two methods by having a feedback mechanism that allows a system to improve on what just one of the approaches would be able to do. However, several systems that tried to blend these two approaches focused on "visual attention," which tries to take into account which parts of an image are visually more important to describe the image better.

Google caption: “A close-up of a plate of food on a table.” Rochester ATT caption: “A table topped with a cake with candles on it.”

The Rochester/Adobe system focuses on what the researchers describe as "semantic attention." In a paper accepted by the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), entitled "Image Captioning with Semantic Attention," computer science professor Jiebo Luo and his colleagues define semantic attention as "the ability to provide a detailed, coherent description of semantically important objects that are needed exactly when they are needed."

"To describe an image you need to decide what to pay more attention to," said Luo. "It is not only about what is in the center of the image or a bigger object, it's also about coming up with a way of deciding on the importance of specific words."

Google caption: “A baby is eating a piece of paper.” Rochester ATT caption: “A baby with a toothbrush in its mouth.”

For example, take an image that shows a table and seated people. The table might be at the center of the image but a better caption might be "a group of people sitting around a table" instead of "a table with people seated." Both are correct, but the former one also tries to take into account what might be of interest to readers and viewers.

Computer image captioning brings together two key areas in artificial intelligence: computer vision and natural language processing. For the computer vision side, researchers train their systems on a massive dataset of images, so they learn to identify objects in images. Language models can then be used to put these words together. For the algorithm that Luo and his team used in their system, they also trained their system on many texts. The objective was not only to understand sentence structure but also the meanings of individual words, what words often get used together with these words, and what words might be semantically more

important.

Google caption: “A white plate with a variety of food.” Rochester ATT caption: “A plate with a sandwich and french fries.”

A closely related paper on video captioning by Luo, graduate student Yuncheng Li, and their Yahoo Research colleagues Yale Song, Liangliang Cao, Joel Tetreault, andLarry Goldberg. "TGIF: A New Dataset and Benchmark on Animated GIF Description," will also be featured as a "Spotlight" presentation at CVPR.

Provided by University of Rochester

Citation: Paying attention to words not just images leads to better image captions (2016, March 17) retrieved 19 April 2024 from https://phys.org/news/2016-03-attention-words-images-image-

captions.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.