Teaching computers to describe images as people would

April 15 2016

The dog was ready to go.

Let's say you're scrolling through your favorite social media app and you

come across a series of pictures of a man in a tuxedo and a woman in a long white dress.

An automated image captioning system might describe that scene as "a picture of a man and a woman," or maybe even "a bride and a groom." But a person might look at the pictures and think, "Wow, my friends got married! They look so happy. What a beautiful wedding."

As image captioning tools get increasingly good at correctly recognizing the objects in an image, a group of researchers is taking the technology one step further. They are working on a system that can automatically describe a series of images in the same kind of way that a human would, by focusing not just on the items in the picture but also what's happening and how it might make a person feel.

"Captioning is about taking concrete objects and putting them together in a literal description," said Margaret Mitchell, a Microsoft researcher who is leading the research project. "What I've been calling visual storytelling is about inferring conceptual and abstract ideas from those concrete objects."

For example, while another image captioning system might describe an image as "a group of people dancing," the visual storytelling system would instead say "We had a ton of fun dancing." And while another captioning system might say, "This is a picture of a float in a parade," this system would instead say "Some of the floats were very colorful."

He had a great time on the hike.

The research project, which relies on a new Microsoft Sequential Image Narrative Dataset, doesn't just stop at one picture. Instead, it takes a series of pictures about the same event and strings together several sentences describing what's going on. The work will be presented in June at the annual meeting of the North American Chapter of the Association for Computational Linguistics.

'Ready for the next step'

The researchers say visual storytelling could eventually be helpful for people who are sharing a number of pictures on social media and want a tool that will help them build a narrative about those pictures. It also could potentially be used to provide richer descriptive tools for people who are blind or visually impaired.

And was very happy to be in the field.

"In image captioning, there are a lot of things we can do reasonably well, and that means we are ready for the next step," said Ting-Hao (Kenneth)

Huang, a Ph.D. candidate at Carnegie Mellon University who worked on the project as part of a summer internship at Microsoft Research. "I think the computer can generate a reasonably simple story, like what we see in a children's book."

Huang was the first author on a paper about the work, along with another summer intern from Johns Hopkins University, Francis Ferraro.

'Translating' from images to sentences

His mom was so proud of him.

The fields of computer vision and natural language processing have made significant advances in the past few years. That's thanks in part to the more widespread use of a machine learning methodology called deep neural networks. These methods have helped researchers get much more accurate results for pattern recognition tasks like speech recognition and identifying objects in photos.

To build the visual storytelling system, the researchers used the deep neural networks to create a "sequence to sequence" machine learning system that is similar to the kind other computer scientists have used for automated language translation. In this case, however, instead of translating from, say, French to English, the researchers were training the system to translate from images to sentences.

For a machine learning system to work, it needs a training set of data that it can learn from. To build the visual storytelling system's training set, the researchers hired crowdsourced workers to write sentences describing various scenes. To account for variations in how people described the scenes, the tool was trained to prefer language in which there was consensus, and to create sentences based on that common ground.

It was a beautiful day for him.

The team also created a separate test set, so they could compare the machine's descriptions with how a human described the scene.

Then, they fed the system new images and asked it to create sentences based on the knowledge it had from the training set.

The research is still in the early stages, and the researchers admit there's significant progress to be made. Still, the researchers say these most recent advances represent another milestone in the fast-moving effort to

use machine learning and other methods from the broader field of artificial intelligence for valuable applications. The new work on visual storytelling brings artificial intelligence a step closer to interpreting the world in the complex, nuanced ways that humans do.

"A picture is worth 1,000 words. It's not just worth three tags," Mitchell said.

Still, the researchers caution that this system – and other cutting-edge research projects like it – are still far from reaching a human level of cognition.

"We're really all scratching the surface," said Nasrin Mostafazadeh, a Ph.D. candidate at the University of Rochester who worked on the project as an intern at Microsoft Research. "It's not that we're doing it, really, in the way that humans do it. It's just that we're trying to."

Provided by Microsoft