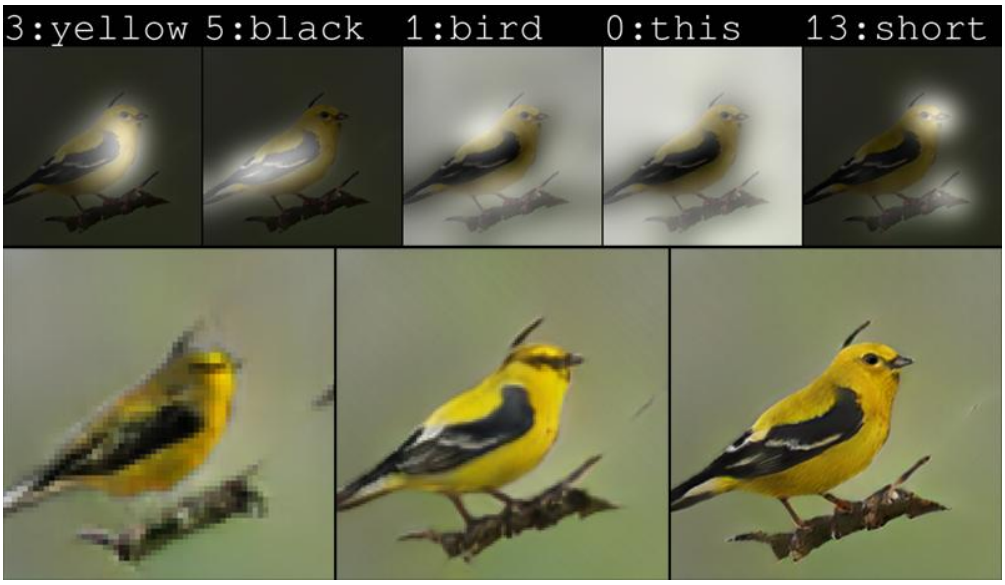
Microsoft researchers build a bot that draws what you tell it to

January 18, 2018 | John Roach



If you're handed a note that asks you to draw a picture of a bird with a yellow body, black wings and a short beak, chances are you'll start with a rough outline of a bird, then glance back at the note, see the yellow part and reach for a yellow pen to fill in the body, read the note again and reach for a black pen to draw the wings and, after a final check, shorten the beak and define it with a reflective glint. Then, for good measure, you might sketch a tree branch where the bird rests.

Now, there's a bot that can do that, too.

The new artificial intelligence technology under development in Microsoft's research labs is programmed to pay close attention to individual words when generating images from caption-like text descriptions. This deliberate focus produced a nearly three-fold boost in image quality compared to the previous state-of-the-art technique for text-to-image generation, according to results on an industry standard test reported in a research paper posted on arXiv.org.

The technology, which the researchers simply call the drawing bot, can generate images of everything from ordinary pastoral scenes, such as grazing livestock, to the absurd, such as a floating double-decker bus. Each image contains details that are absent from the text descriptions, indicating that this artificial intelligence contains an artificial imagination.

"If you go to Bing and you search for a bird, you get a bird picture. But here, the pictures are created by the computer, pixel by pixel, from scratch," said Xiaodong He, a principal researcher and research manager in the Deep Learning Technology Center at Microsoft's research lab in Redmond, Washington. "These birds may not exist in the real world — they are just an aspect of our computer's imagination of birds."

The drawing bot closes a research circle around the intersection of computer vision and natural language processing that He and colleagues have explored for the past half-decade. They started

with technology that automatically writes photo captions – the CaptionBot – and then moved to a technology that answers questions humans ask about images, such as the location or attributes of objects, which can be especially helpful for blind people.

These research efforts require training machine learning models to identify objects, interpret actions and converse in natural language.

“Now we want to use the text to generate the image,” said Qiuyuan Huang, a postdoctoral researcher in He’s group and a paper co-author. “So, it is a cycle.”

Image generation is a more challenging task than image captioning, added Pengchuan Zhang, an associate researcher on the team, because the process requires the drawing bot to imagine details that are not contained in the caption. “That means you need your machine learning algorithms running your artificial intelligence to imagine some missing parts of the images,” he said.

Attentive image generation

At the core of Microsoft’s drawing bot is a technology known as a Generative Adversarial Network, or GAN. The network consists of two machine learning models, one that generates images from text descriptions and another, known as a discriminator, that uses text descriptions to judge the authenticity of generated images. The generator attempts to get fake pictures past the discriminator; the discriminator never wants to be fooled. Working together, the discriminator pushes the generator toward perfection.

Microsoft’s drawing bot was trained on datasets that contain paired images and captions, which allow the models to learn how to match words to the visual representation of those words. The GAN, for example, learns to generate an image of a bird when a caption says bird and, likewise, learns what a picture of a bird should look like. “That is a fundamental reason why we believe a machine can learn,” said He.

GANs work well when generating images from simple text descriptions such as a blue bird or an evergreen tree, but the quality stagnates with more complex text descriptions such as a bird with a green crown, yellow wings and a red belly. That’s because the entire sentence serves as a single input to the generator. The detailed information of the description is lost. As a result, the generated image is a blurry greenish-yellowish-reddish bird instead a close, sharp match with the description.

As humans draw, we repeatedly refer to the text and pay close attention to the words that describe the region of the image we are drawing. To capture this human trait, the researchers created what they call an attentional GAN, or AttnGAN, that mathematically represents the human concept of attention. It does this by breaking up the input text into individual words and matching those words to specific regions of the image.

“Attention is a human concept; we use math to make attention computational,” explained He.

The model also learns what humans call commonsense from the training data, and it pulls on this learned notion to fill in details of images that are left to the imagination. For example, since many images of birds in the training data show birds sitting on tree branches, the AttnGAN usually draws birds sitting on branches unless the text specifies otherwise.

"From the data, the machine learning algorithm learns this commonsense where the bird should belong," said Zhang. As a test, the team fed the drawing bot captions for absurd images, such as "a red double-decker bus is floating on a lake." It generated a blurry, drippy image that resembles both a boat with two decks and a double-decker bus on a lake surrounded by mountains. The image suggests the bot had an internal struggle between knowing that boats float on lakes and the text specification of bus.

"We can control what we describe and see how the machine reacts," explained He. "We can poke and test what the machine learned. The machine has some background learned commonsense, but it can still follow what you ask and maybe, sometimes, it seems a bit ridiculous."

Practical applications

Text-to-image generation technology could find practical applications acting as a sort of sketch assistant to painters and interior designers, or as a tool for voice-activated photo refinement. With more computing power, He imagines the technology could generate animated films based on screenplays, augmenting the work that animated filmmakers do by removing some of the manual labor involved.

For now, the technology is imperfect. Close examination of images almost always reveals flaws, such as birds with blue beaks instead of black and fruit stands with mutant bananas. These flaws are a clear indication that a computer, not a human, created the images. Nevertheless, the quality of the AttnGAN images are a nearly three-fold improvement over the previous best-in-class GAN and serve as a milestone on the road toward a generic, human-like intelligence that augments human capabilities, according to He.

"For AI and humans to live in the same world, they have to have a way to interact with each other," explained He. "And language and vision are the two most important modalities for humans and machines to interact with each other."

In addition to Xiaodong He, Pengchuan Zhang and Qiuyuan Huang at Microsoft, collaborators include former Microsoft interns Tao Xu from Lehigh University and Zhe Gan from Duke University; and Han Zhang from Rutgers University and Xiaolei Huang from Lehigh University.

Related

- Read the research paper describing the AttnGAN
- Learn more about Microsoft's AI research efforts in Vision and Language Intelligence
- Check out the CaptionBot and Seeing AI
- Decades of computer vision research, one 'Swiss Army knife'

John Roach writes about Microsoft research and innovation. Follow him on Twitter.