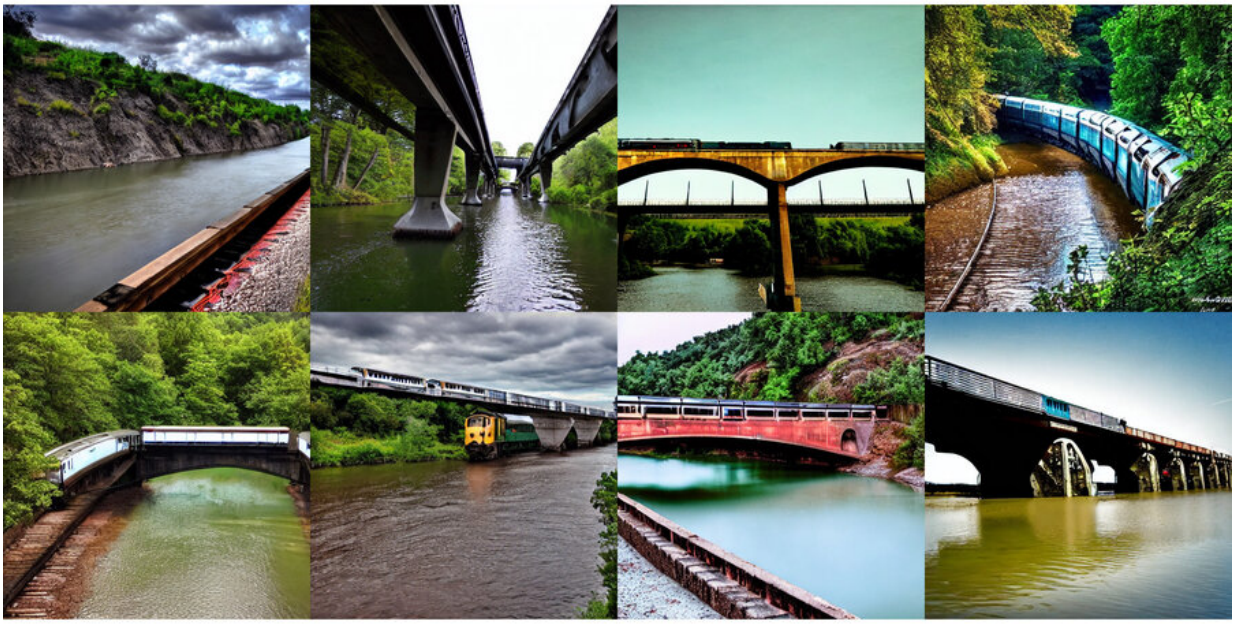
AI system makes image generator models like DALL-E 2 more creative

September 13 2022, by Rachel Gordon



This array of generated images, showing "a train on a bridge" and "a river under the bridge," was generated using a new method developed by MIT researchers. Credit: Massachusetts Institute of Technology

The internet had a collective feel-good moment with the introduction of

DALL-E, an artificial intelligence-based image generator inspired by artist Salvador Dali and the lovable robot WALL-E that uses natural language to produce whatever mysterious and beautiful image your heart desires. Seeing typed-out inputs like "smiling gopher holding an ice cream cone" instantly spring to life clearly resonated with the world.

Getting said smiling gopher and attributes to pop up on your screen is not a small task. DALL-E 2 uses something called a diffusion model, where it tries to encode the entire text into one description to generate an image. But once the text has a lot of more details, it's hard for a single description to capture it all. Moreover, while they're highly flexible, they sometimes struggle to understand the composition of certain concepts, like confusing the attributes or relations between different objects.

To generate more complex images with better understanding, scientists from MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) structured the typical model from a different angle: they added a series of models together, where they all cooperate to generate desired images capturing multiple different aspects as requested by the input text or labels. To create an image with two components, say, described by two sentences of description, each model would tackle a particular component of the image.

This array of generated images, showing “a river leading into mountains" and "red trees on the side,” was generated using a new method developed by MIT researchers. Credit: Massachusetts Institute of Technology

The seemingly magical models behind image generation work by suggesting a series of iterative refinement steps to get to the desired image. It starts with a "bad" picture and then gradually refines it until it becomes the selected image. By composing multiple models together, they jointly refine the appearance at each step, so the result is an image that exhibits all the attributes of each model. By having multiple models cooperate, you can get much more creative combinations in the generated images.

Take, for example, a red truck and a green house. The model will

confuse the concepts of red truck and green house when these sentences get very complicated. A typical generator like DALL-E 2 might make a green truck and a red house, so it'll swap these colors around. The team's approach can handle this type of binding of attributes with objects, and especially when there are multiple sets of things, it can handle each object more accurately.

"The model can effectively model object positions and relational descriptions, which is challenging for existing image-generation models. For example, put an object and a cube in a certain position and a sphere in another. DALL-E 2 is good at generating natural images but has difficulty understanding object relations sometimes," says MIT CSAIL Ph.D. student and co-lead author Shuang Li, "Beyond art and creativity, perhaps we could use our model for teaching. If you want to tell a child to put a cube on top of a sphere, and if we say this in language, it might be hard for them to understand. But our model can generate the image and show them."

Researchers were able to create some surprising, surreal imagery with the text, “a dog" and "the sky.” On the left appear a dog and clouds separately, labeled “dog” and “sky” underneath, and on the right appear two images of cloud-like dogs with the label, “dog AND sky,” underneath. Credit: Massachusetts Institute of Technology

Making Dali proud

Composable Diffusion—the team's model—uses diffusion models alongside compositional operators to combine text descriptions without further training. The team's approach more accurately captures text details than the original diffusion model, which directly encodes the words as a single long sentence. For example, given "a pink sky" AND "a blue mountain in the horizon" AND "cherry blossoms in front of the

mountain," the team's model was able to produce that image exactly, whereas the original diffusion model made the sky blue and everything in front of the mountains pink.

"The fact that our model is composable means that you can learn different portions of the model, one at a time. You can first learn an object on top of another, then learn an object to the right of another, and then learn something left of another," says co-lead author and MIT CSAIL Ph.D. student Yilun Du. "Since we can compose these together, you can imagine that our system enables us to incrementally learn language, relations, or knowledge, which we think is a pretty interesting direction for future work."

This photo illustration was created using generated images from an MIT system called Composable Diffusion, and arranged in Photoshop. Phrases like

“diffusion model” and “network” were used to generate the pink dots and geometric, angular images. The phrase “a horse AND a yellow flower field” is included at the top of the image. Generated images of a horse and yellow field appear on the left, and the combined imagery of a horse in a yellow flower field appear on the right. Credit: Massachusetts Institute of Technology

While it showed prowess in generating complex, photorealistic images, it still faced challenges since the model was trained on a much smaller dataset than those like DALL-E 2, so there were some objects it simply couldn't capture.

Now that Composable Diffusion can work on top of generative models, such as DALL-E 2, the scientists want to explore continual learning as a potential next step. Given that more is usually added to object relations, they want to see if diffusion models can start to "learn" without forgetting previously learned knowledge—to a place where the model can produce images with both the previous and new knowledge.

"This research proposes a new method for composing concepts in text-to-image generation not by concatenating them to form a prompt, but rather by computing scores with respect to each concept and composing them using conjunction and negation operators," says Mark Chen, co-creator of DALL-E 2 and research scientist at OpenAI. "This is a nice idea that leverages the energy-based interpretation of diffusion models so that old ideas around compositionality using energy-based models can be applied. The approach is also able to make use of classifier-free guidance, and it is surprising to see that it outperforms the GLIDE baseline on various compositional benchmarks and can qualitatively produce very different types of image generations."

"Humans can compose scenes including different elements in a myriad

of ways, but this task is challenging for computers," says Bryan Russel, research scientist at Adobe Systems. "This work proposes an elegant formulation that explicitly composes a set of diffusion models to generate an image given a complex natural language prompt."

More information: Yilun Du, Shuang Li, Igor Mordatch, Compositional Visual Generation and Inference with Energy Based Models. arXiv:2004.06030v3 [cs.CV], arxiv.org/abs/2004.06030

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology