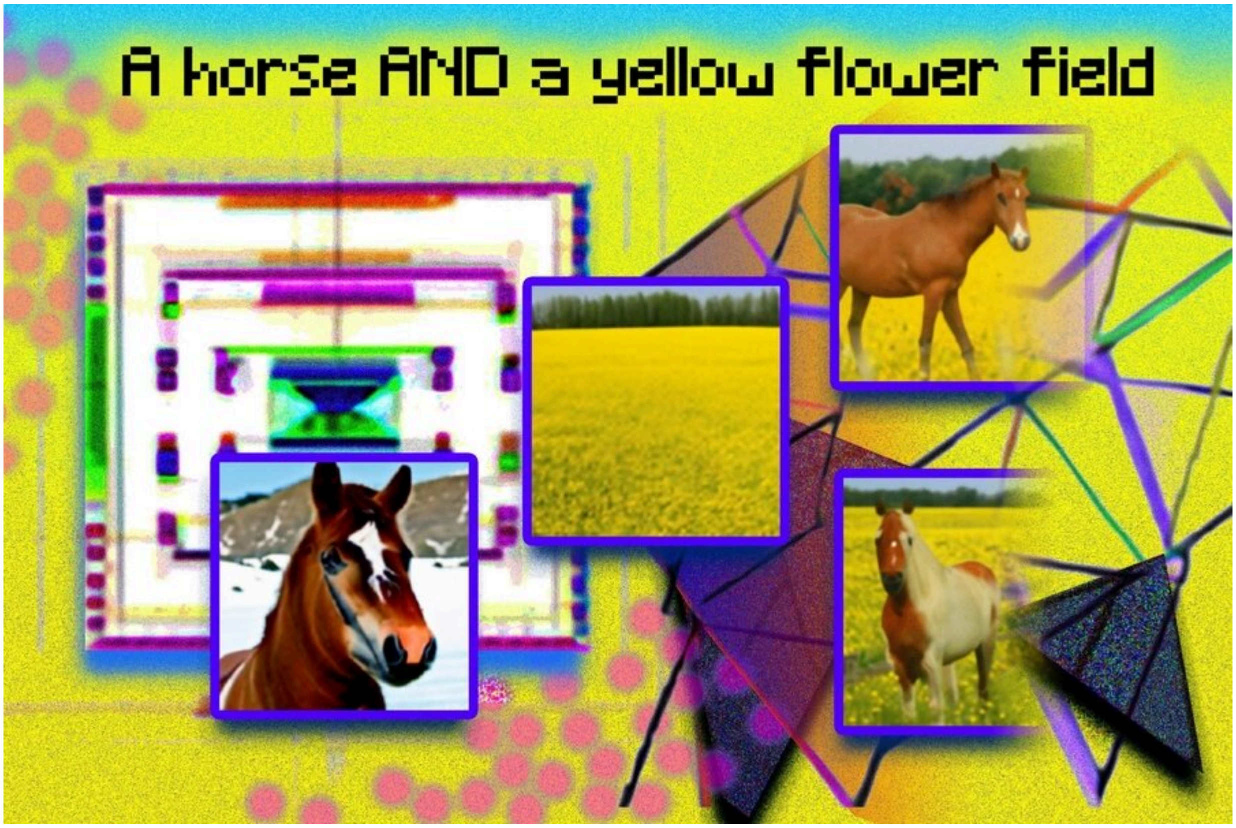
TechXplore

How AI image generators could help robots

October 27 2022, by Rachel Gordon



MIT doctoral student Yilun Du has been working on extending stable diffusion models, the technical backbone of generative art to other domains such as robotics. Credit: Jose-Luis Olivares/MIT and the researchers

AI image generators, which create fantastical sights at the intersection of dreams and reality, bubble up on every corner of the web. Their entertainment value is demonstrated by an ever-expanding treasure trove of whimsical and random images serving as indirect portals to the brains of human designers. A simple text prompt yields a nearly instantaneous image, satisfying our primitive brains, which are hardwired for instant gratification.

Although seemingly nascent, the field of AI-generated art can be traced back as far as the 1960s with early attempts using symbolic rule-based approaches to make technical images. While the progression of models that untangle and parse words has gained increasing sophistication, the explosion of generative art has sparked debate around copyright, disinformation, and biases, all mired in hype and controversy.

Yilun Du, a Ph.D. student in the Department of Electrical Engineering and Computer Science and affiliate of MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL), recently developed a new method that makes models like DALL-E 2 more creative and have better scene understanding. Here, Du describes how these models work, whether this technical infrastructure can be applied to other domains, and how we draw the line between AI and human creativity.

Q: AI-generated images use something called "stable diffusion" models to turn words into astounding images in just a few moments. But for every image used, there's usually a human behind it. So what's the the line between AI and human creativity? How do these models really work?

A: Imagine all of the images you could get on Google Search and their associated patterns. This is the diet these models are fed on. They're trained on all of these images and their captions to generate images similar to the billions of images it has seen on the internet.

Let's say a model has seen a lot of dog photos. It's trained so that when it gets a similar text input prompt like "dog," it's able to generate a photo that looks very similar to the many dog pictures already seen. Now, more methodologically, how this all works dates back to a very old class of models called "energy-based models," originating in the '70's or '80's.

In energy-based models, an energy landscape over images is constructed, which is used to simulate the physical dissipation to generate images. When you drop a dot of ink into water and it dissipates, for example, at the end, you just get this uniform texture. But if you try to reverse this process of dissipation, you gradually get the original ink dot in the water again.

Or let's say you have this very intricate block tower, and if you hit it with a ball, it collapses into a pile of blocks. This pile of blocks is then very disordered, and there's not really much structure to it. To resuscitate the tower, you can try to reverse this folding process to generate your original pile of blocks.

The way these generative models generate images is in a very similar manner, where, initially, you have this really nice image, where you start from this random noise, and you basically learn how to simulate the process of how to reverse this process of going from noise back to your original image, where you try to iteratively refine this image to make it more and more realistic.

In terms of what's the line between AI and human creativity, you can say that these models are really trained on the creativity of people. The internet has all types of paintings and images that people have already created in the past. These models are trained to recapitulate and generate the images that have been on the internet. As a result, these models are more like crystallizations of what people have spent creativity on for hundreds of years.

At the same time, because these models are trained on what humans have designed, they can generate very similar pieces of art to what humans have done in the past. They can find patterns in art that people have made, but it's much harder for these models to actually generate creative photos on their own.

If you try to enter a prompt like "abstract art" or "unique art" or the like, it doesn't really understand the creativity aspect of human art. The models are, rather, recapitulating what people have done in the past, so to speak, as opposed to generating fundamentally new and creative art.

Since these models are trained on vast swaths of images from the internet, a lot of these images are likely copyrighted. You don't exactly know what the model is retrieving when it's generating new images, so there's a big question of how you can even determine if the model is using copyrighted images. If the model depends, in some sense, on some copyrighted images, are then those new images copyrighted? That's another question to address.

Q: Do you believe images generated by diffusion models encode some sort of understanding about natural or physical worlds, either dynamically or geometrically? Are there efforts toward "teaching" image generators the basics of the universe that babies learn so early on?

A: Do they understand, in code, some grasp of natural and physical worlds? I think definitely. If you ask a model to generate a stable configuration of blocks, it definitely generates a block configuration that's stable. If you tell it, generate an unstable configuration of blocks, it does look very unstable. Or if you say "a tree next to a lake," it's roughly able to generate that.

In a sense, it seems like these models have captured a large aspect of common sense. But the issue that makes us, still, very far away from truly understanding the natural and physical world is that when you try to generate infrequent combinations of words that you or I in our working our minds can very easily imagine, these models cannot.

For example, if you say, "put a fork on top of a plate," that happens all the time. If you ask the model to generate this, it easily can. If you say, "put a plate on top of a fork," again, it's very easy for us to imagine what this would look like. But if you put this into any of these large models, you'll never get a plate on top of a fork. You instead get a fork on top of a plate, since the models are learning to recapitulate all the images it's been trained on. It can't really generalize that well to combinations of words it hasn't seen.

A fairly well-known example is an astronaut riding a horse, which the model can do with ease. But if you say a horse riding an astronaut, it still generates a person riding a horse. It seems like these models are capturing a lot of correlations in the datasets they're trained on, but they're not actually capturing the underlying causal mechanisms of the world.

Another example that's commonly used is if you get very complicated text descriptions like one object to the right of another one, the third object in the front, and a third or fourth one flying. It really is only able to satisfy maybe one or two of the objects. This could be partially because of the training data, as it's rare to have very complicated captions But it could also suggest that these models aren't very structured.

You can imagine that if you get very complicated natural language prompts, there's no manner in which the model can accurately represent all the component details.

Q: You recently came up with a new method that uses multiple models to create more complex images with better understanding for generative art. Are there potential applications of this framework outside of image or text domains?

A: We were really inspired by one of the limitations of these models. When you give these models very complicated scene descriptions, they aren't actually able to correctly generate images that match them.

One thought is, since it's a single model with a fixed computational graph, meaning you can only use a fixed amount of computation to generate an image, if you get an extremely complicated prompt, there's no way you can use more computational power to generate that image.

If I gave a human a description of a scene that was, say, 100 lines long versus a scene that's one line long, a human artist can spend much longer on the former. These models don't really have the sensibility to do this. We propose, then, that given very complicated prompts, you can actually compose many different independent models together and have each individual model represent a portion of the scene you want to describe.

We find that this enables our model to generate more complicated scenes, or those that more accurately generate different aspects of the scene together. In addition, this approach can be generally applied across a variety of different domains. While image generation is likely the most currently successful application, generative models have actually been seeing all types of applications in a variety of domains.

You can use them to generate different diverse robot behaviors, synthesize 3D shapes, enable better scene understanding, or design new materials. You could potentially compose multiple desired factors to generate the exact material you need for a particular application.

One thing we've been very interested in is robotics. In the same way that you can generate different images, you can also generate different robot trajectories (the path and schedule), and by composing different models together, you are able to generate trajectories with different combinations of skills. If I have natural language specifications of jumping versus avoiding an obstacle, you could also compose these models together, and then generate robot trajectories that can both jump and avoid an obstacle .

In a similar manner, if we want to design proteins, we can specify different functions or aspects—in an analogous manner to how we use language to specify the content of the images—with language-like descriptions, such as the type or functionality of the protein. We could then compose these together to generate new proteins that can potentially satisfy all of these given functions.

We've also explored using diffusion models on 3D shape generation, where you can use this approach to generate and design 3D assets. Normally, 3D asset design is a very complicated and laborious process. By composing different models together, it becomes much easier to generate shapes such as, "I want a 3D shape with four legs, with this style and height," potentially automating portions of 3D asset design.

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

Citation: How AI image generators could help robots (2022, October 27) retrieved 10 April 2024 from https://techxplore.com/news/2022-10-ai-image-robots.html