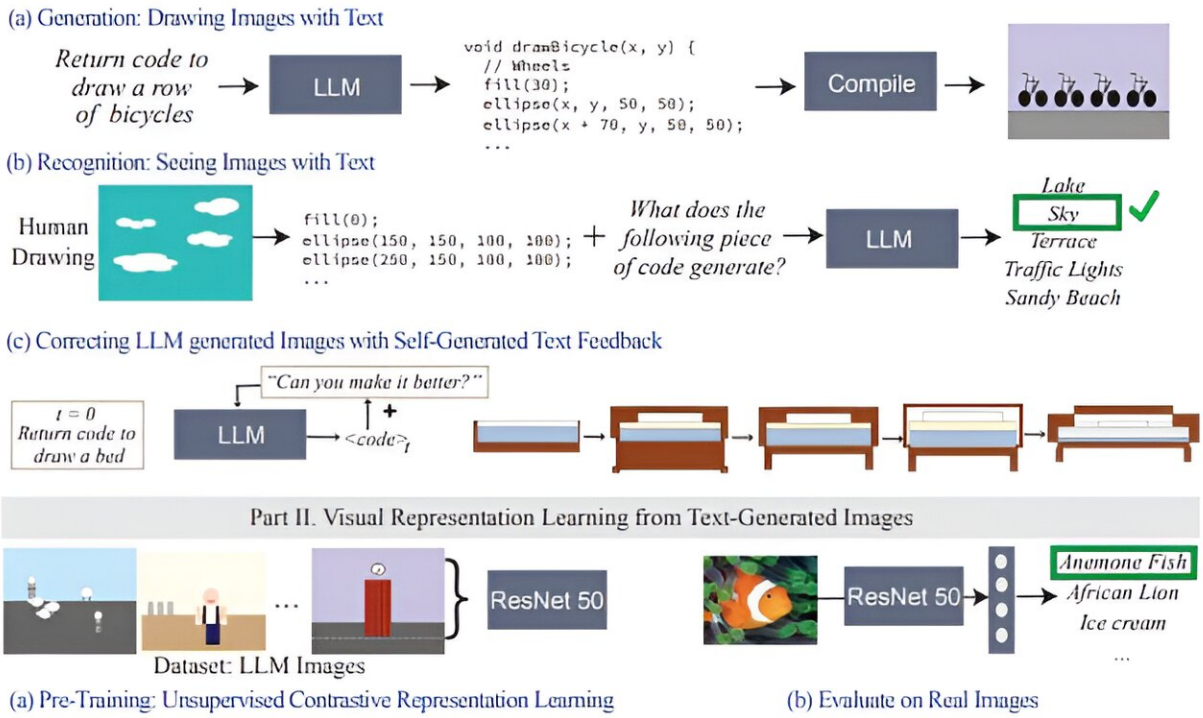
Using illustrations to train an image-free computer vision system to recognize real photos

June 17 2024, by Alex Shipps



Vision check-up for LLMs. I. Testing the visual knowledge of Language Models. We suggest a set of tests to check the vision abilities of language models, these include (a) the ability to write code that renders complex visual concepts (b) recognizing visual concepts from code (c) correcting rendering code with text-only self-feedback. II. We test whether LLMs can generate data to train a high-performance vision system that can be used to make semantic judgments on natural images. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2401.01862

You've likely heard that a picture is worth a thousand words, but can a large language model (LLM) get the picture if it's never seen images before?

As it turns out, language models that are trained purely on text have a solid understanding of the visual world. They can write image-rendering code to generate complex scenes with intriguing objects and compositions—and even when that knowledge is not used properly, LLMs can refine their images. Researchers from MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) observed this when prompting language models to self-correct their code for different images, where the systems improved on their simple clipart drawings with each query.

The visual knowledge of these language models is gained from how concepts like shapes and colors are described across the internet, whether in language or code. When given a direction like "draw a parrot in the jungle," users jog the LLM to consider what it's read in descriptions before.

To assess how much visual knowledge LLMs have, the CSAIL team constructed a "vision checkup" for LLMs: using their "Visual Aptitude Dataset," they tested the models' abilities to draw, recognize, and self-correct these concepts. Collecting each final draft of these illustrations, the researchers trained a computer vision system that identifies the content of real photos.

Their work is published on the *arXiv* preprint server.

"We essentially train a vision system without directly using any visual data," says Tamar Rott Shaham, co-lead author of the study and an MIT electrical engineering and computer science (EECS) postdoc at CSAIL. "Our team queried language models to write image-rendering codes to

generate data for us and then trained the vision system to evaluate natural images. We were inspired by the question of how visual concepts are represented through other mediums, like text. To express their visual knowledge, LLMs can use code as a common ground between text and vision."

To build this dataset, the researchers first queried the models to generate code for different shapes, objects, and scenes. Then, they compiled that code to render simple digital illustrations, like a row of bicycles, showing that LLMs understand spatial relations well enough to draw the two-wheelers in a horizontal row. As another example, the model generated a car-shaped cake, combining two random concepts. The language model also produced a glowing light bulb, indicating its ability to create visual effects.

"Our work shows that when you query an LLM (without multimodal pre-training) to create an image, it knows much more than it seems," says co-lead author, EECS Ph.D. student, and CSAIL member Pratyusha Sharma. "Let's say you asked it to draw a chair. The model knows other things about this piece of furniture that it may not have immediately rendered, so users can query the model to improve the visual it produces with each iteration. Surprisingly, the model can iteratively enrich the drawing by improving the rendering code to a significant extent."

The researchers gathered these illustrations, which were then used to train a computer vision system that can recognize objects within real photos (despite never having seen one before). With this synthetic, text-generated data as its only reference point, the system outperforms other procedurally generated image datasets that were trained with authentic photos.

The CSAIL team believes that combining the hidden visual knowledge of LLMs with the artistic capabilities of other AI tools like diffusion

models could also be beneficial. Systems like Midjourney sometimes lack the know-how to consistently tweak the finer details in an image, making it difficult for them to handle requests like reducing how many cars are pictured, or placing an object behind another. If an LLM sketched out the requested change for the diffusion model beforehand, the resulting edit could be more satisfactory.

The irony, as Rott Shaham and Sharma acknowledge, is that LLMs sometimes fail to recognize the same concepts that they can draw. This became clear when the models incorrectly identified human re-creations of images within the dataset. Such diverse representations of the visual world likely triggered the language models' misconceptions.

While the models struggled to perceive these abstract depictions, they demonstrated the creativity to draw the same concepts differently each time. When the researchers queried LLMs to draw concepts like strawberries and arcades multiple times, they produced pictures from diverse angles with varying shapes and colors, hinting that the models might have actual mental imagery of visual concepts (rather than reciting examples they saw before).

The CSAIL team believes this procedure could be a baseline for evaluating how well a generative AI model can train a computer vision system. Additionally, the researchers look to expand the tasks they challenge language models on. As for their recent study, the MIT group notes that they don't have access to the training set of the LLMs they used, making it challenging to further investigate the origin of their visual knowledge. In the future, they intend to explore training an even better vision model by letting the LLM work directly with it.

Sharma and Rott Shaham are joined on the paper by former CSAIL affiliate Stephanie Fu and EECS Ph.D. students Manel Baradad, Adrián Rodríguez-Muñoz, and Shivam Duggal, who are all CSAIL affiliates; as

well as MIT Associate Professor Phillip Isola and Professor Antonio Torralba.

They present their paper this week at the IEEE/CVF Computer Vision and Pattern Recognition Conference.

More information: Pratyusha Sharma et al, A Vision Check-up for Language Models, *arXiv* (2024). DOI: 10.48550/arxiv.2401.01862

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