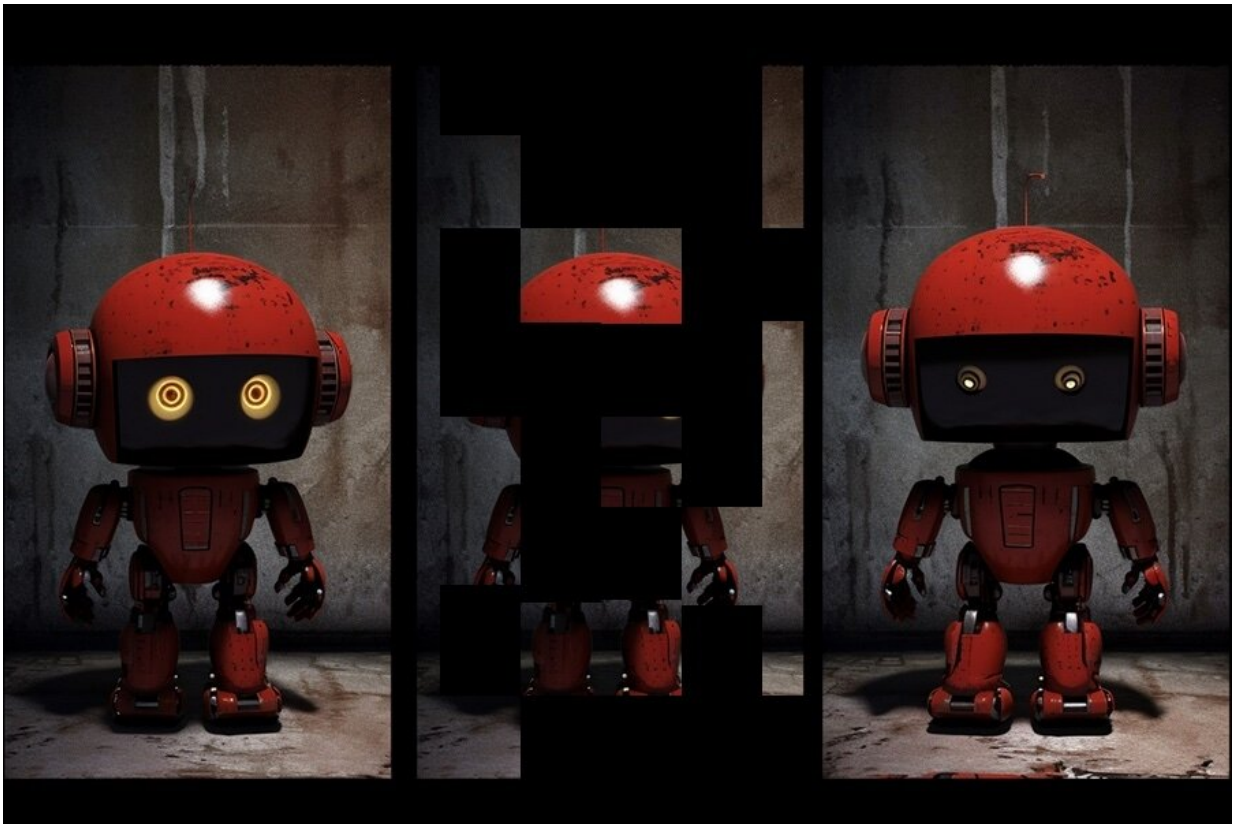
Computer vision system marries image recognition and generation

June 28 2023, by Rachel Gordon



A unified vision system known as Msked Generative Encoder (MAGE), developed by researchers at MIT and Google, could be useful for many things, like finding and classifying objects in an image, learning from just a few examples, generating images with specific conditions such as text or class, editing existing images, and more. Credit: Alex Shipps/MIT CSAIL via Midjourney

Computers possess two remarkable capabilities with respect to images: They can both identify them and generate them anew. Historically, these functions have stood separate, akin to the disparate acts of a chef who is good at creating dishes (generation), and a connoisseur who is good at tasting dishes (recognition).

Yet, one can't help but wonder: What would it take to orchestrate a harmonious union between these two distinctive capacities? Both chef and connoisseur share a common understanding in the taste of the food. Similarly, a unified vision system requires a deep understanding of the visual world.

Now, researchers in MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) have trained a system to infer the missing parts of an image, a task that requires deep comprehension of the image's content. In successfully filling in the blanks, the system, known as the Masked Generative Encoder (MAGE), achieves two goals at the same time: accurately identifying images and creating new ones with striking resemblance to reality.

This dual-purpose system enables myriad potential applications, like object identification and classification within images, swift learning from minimal examples, the creation of images under specific conditions like text or class, and enhancing existing images.

Unlike other techniques, MAGE doesn't work with raw pixels. Instead, it converts images into what's called "semantic tokens," which are compact, yet abstracted, versions of an image section. Think of these tokens as mini jigsaw puzzle pieces, each representing a 16x16 patch of the original image. Just as words form sentences, these tokens create an abstracted version of an image that can be used for complex processing tasks, while preserving the information in the original image. Such a tokenization step can be trained within a self-supervised framework,

allowing it to pre-train on large image datasets without labels.

Now, the magic begins when MAGE uses "masked token modeling." It randomly hides some of these tokens, creating an incomplete puzzle, and then trains a neural network to fill in the gaps. This way, it learns to both understand the patterns in an image (image recognition) and generate new ones (image generation).

"One remarkable part of MAGE is its variable masking strategy during pre-training, allowing it to train for either task, image generation or recognition, within the same system," says Tianhong Li, a Ph.D. student in electrical engineering and computer science at MIT, a CSAIL affiliate, and the lead author on a paper about the research. "MAGE's ability to work in the 'token space' rather than 'pixel space' results in clear, detailed, and high-quality image generation, as well as semantically rich image representations. This could hopefully pave the way for advanced and integrated computer vision models."

Apart from its ability to generate realistic images from scratch, MAGE also allows for conditional image generation. Users can specify certain criteria for the images they want MAGE to generate, and the tool will cook up the appropriate image. It's also capable of image editing tasks, such as removing elements from an image while maintaining a realistic appearance.

Recognition tasks are another strong suit for MAGE. With its ability to pre-train on large unlabeled datasets, it can classify images using only the learned representations. Moreover, it excels at few-shot learning, achieving impressive results on large image datasets like ImageNet with only a handful of labeled examples.

The validation of MAGE's performance has been impressive. On one hand, it set new records in generating new images, outperforming

previous models with a significant improvement. On the other hand, MAGE topped in recognition tasks, achieving an 80.9 percent accuracy in linear probing and a 71.9 percent 10-shot accuracy on ImageNet (this means it correctly identified images in 71.9 percent of cases where it had only 10 labeled examples from each class).

Despite its strengths, the research team acknowledges that MAGE is a work in progress. The process of converting images into tokens inevitably leads to some loss of information. They are keen to explore ways to compress images without losing important details in future work. The team also intends to test MAGE on larger datasets. Future exploration might include training MAGE on larger unlabeled datasets, potentially leading to even better performance.

"It has been a long dream to achieve image generation and image recognition in one single system. MAGE is a groundbreaking research which successfully harnesses the synergy of these two tasks and achieves the state-of-the-art of them in one single system," says Huisheng Wang, senior staff software engineer of humans and interactions in the Research and Machine Intelligence division at Google, who was not involved in the work. "This innovative system has wide-ranging applications, and has the potential to inspire many future works in the field of computer vision."

The findings are published on the *arXiv* preprint server.

More information: Tianhong Li et al, MAGE: MAsked Generative Encoder to Unify Representation Learning and Image Synthesis, *arXiv* (2022). DOI: 10.48550/arxiv.2211.09117