Microsoft unveils AI model that understands image content, solves visual puzzles

Enlarge / An AI-generated image of an electronic brain with an eyeball.
Ars Technica

reader comments

81 with

Share this story

On Monday, researchers from Microsoft introduced Kosmos-1, a multimodal model that can reportedly analyze images for content, solve visual puzzles, perform visual text recognition, pass visual IQ tests, and understand natural language instructions. The researchers believe multimodal AI—which integrates different modes of input such as text, audio, images, and video—is a key step to building artificial general intelligence (AGI) that can perform general tasks at the level of a human.

“Being a basic part of intelligence, multimodal perception is a necessity to achieve artificial general intelligence, in terms of knowledge acquisition and grounding to the real world,” the researchers write in their academic paper, “Language Is Not All You Need: Aligning Perception with Language Models.”

Visual examples from the Kosmos-1 paper show the model analyzing images and answering questions about them, reading text from an image, writing captions for images, and taking a visual IQ test with 22–26 percent accuracy (more on that below).

- A Microsoft-provided example of Kosmos-1 answering questions about images and websites.
- A Microsoft-provided example of "multimodal chain-of-thought prompting" for Kosmos-1.
- An example of Kosmos-1 doing visual question answering, provided by Microsoft.

While media buzz with news about large language models (LLM), some AI experts point to multimodal AI as a potential path toward general artificial intelligence, a hypothetical technology that will ostensibly be able to replace humans at any intellectual task (and any intellectual job). AGI is the stated goal of OpenAI, a key business partner of Microsoft in the AI space.

In this case, Kosmos-1 appears to be a pure Microsoft project without OpenAI's involvement. The researchers call their creation a "multimodal large language model" (MLLM) because its roots lie in natural language processing like a text-only LLM, such as ChatGPT. And it shows: For Kosmos-1 to accept image input, the researchers must first translate the image into a special series of tokens (basically text) that the LLM can understand. The Kosmos-1 paper describes this in more detail:

Advertisement

> *For input format, we flatten input as a sequence decorated with special tokens. Specifically, we use <g> and </g> to denote start- and end-of-sequence. The special tokens <image> and </image> indicate the beginning and end of encoded image embeddings. For example, "<g> document </g>" is a text input, and "<s> paragraph <image> Image Embedding </image> paragraph*

> *</s>” is an interleaved image-text input.*
>
> *… An embedding module is used to encode both text tokens and other input modalities into vectors. Then the embeddings are fed into the decoder. For input tokens, we use a lookup table to map them into embeddings. For the modalities of continuous signals (e.g., image, and audio), it is also feasible to represent inputs as discrete code and then regard them as “foreign languages”.*

Microsoft trained Kosmos-1 using data from the web, including excerpts from The Pile (an 800GB English text resource) and Common Crawl. After training, they evaluated Kosmos-1’s abilities on several tests, including language understanding, language generation, optical character recognition-free text classification, image captioning, visual question answering, web page question answering, and zero-shot image classification. In many of these tests, Kosmos-1 outperformed current state-of-the-art models, according to Microsoft.

Enlarge / An example of the Raven IQ test that Kosmos-1 was tasked with solving.

Of particular interest is Kosmos-1's performance on Raven's Progressive Reasoning, which measures visual IQ by presenting a sequence of shapes and asking the test taker to complete the sequence. To test Kosmos-1, the researchers fed a filled-out test, one at a time, with each option completed and asked if the answer was correct. Kosmos-1 could only correctly answer a question on the Raven test 22 percent of the time (26 percent with fine-tuning). This is by no means a slam dunk, and errors in the methodology could have affected the results, but Kosmos-1 beat random chance (17 percent) on the Raven IQ test.

Still, while Kosmos-1 represents early steps in the multimodal domain (an approach also being pursued by others), it's easy to imagine that future optimizations could bring even more significant results, allowing AI models to perceive any form of media and act on it, which will greatly enhance the

abilities of artificial assistants. In the future, the researchers say they’d like to scale up Kosmos-1 in model size and integrate speech capability as well.

Microsoft says it plans to make Kosmos-1 available to developers, though the GitHub page the paper cites has no obvious Kosmos-specific code upon this story’s publication.

https://arstechnica.com/?p=1920920