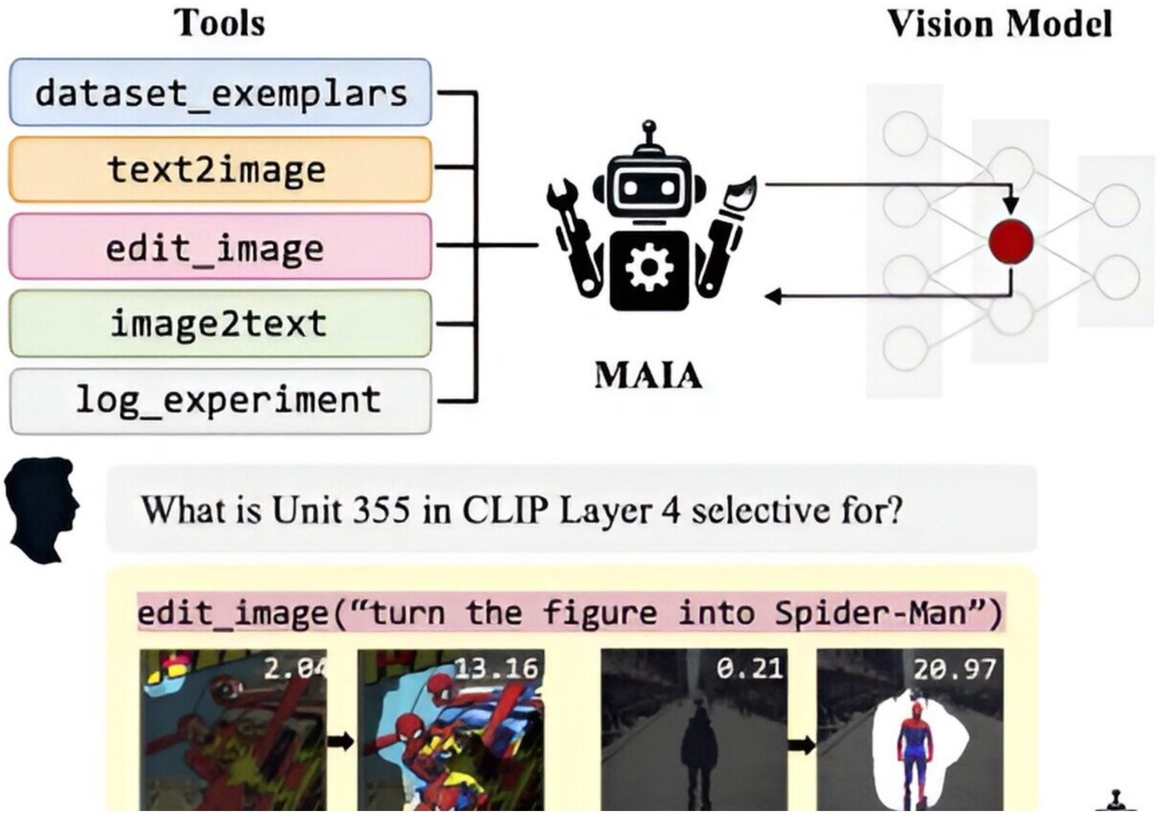
Multimodal agent can iteratively design experiments to better understand various components of AI systems

July 23 2024, by Rachel Gordon



MAIA framework. MAIA autonomously conducts experiments on other systems to explain their behavior. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2404.14394

As artificial intelligence models become increasingly prevalent and are integrated into diverse sectors like health care, finance, education, transportation, and entertainment, understanding how they work under the hood is critical. Interpreting the mechanisms underlying AI models enables us to audit them for safety and biases, with the potential to deepen our understanding of the science behind intelligence itself.

Imagine if we could directly investigate the human brain by manipulating each of its individual neurons to examine their roles in perceiving a particular object. While such an experiment would be prohibitively invasive in the human brain, it is more feasible in another type of neural network: one that is artificial. However, somewhat similar to the human brain, artificial models containing millions of neurons are too large and complex to study by hand, making interpretability at scale a very challenging task.

To address this, MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) researchers decided to take an automated approach to interpreting artificial vision models that evaluate different properties of images. They have developed "MAIA" (Multimodal Automated Interpretability Agent), a system that automates a variety of neural network interpretability tasks using a vision-language model backbone equipped with tools for experimenting on other AI systems.

The research is published on the *arXiv* preprint server.

"Our goal is to create an AI researcher that can conduct interpretability experiments autonomously. Existing automated interpretability methods merely label or visualize data in a one-shot process. On the other hand, MAIA can generate hypotheses, design experiments to test them, and refine its understanding through iterative analysis," says Tamar Rott Shaham, an MIT electrical engineering and computer science (EECS) postdoc at CSAIL and co-author on a new paper about the research.

"By combining a pre-trained vision-language model with a library of interpretability tools, our multimodal method can respond to user queries by composing and running targeted experiments on specific models, continuously refining its approach until it can provide a comprehensive answer."

The automated agent is demonstrated to tackle three key tasks: It labels individual components inside vision models and describes the visual concepts that activate them, it cleans up image classifiers by removing irrelevant features to make them more robust to new situations, and it hunts for hidden biases in AI systems to help uncover potential fairness issues in their outputs.

"But a key advantage of a system like MAIA is its flexibility," says Sarah Schwettmann, Ph.D., a research scientist at CSAIL and co-lead of the research. "We demonstrated MAIA's usefulness on a few specific tasks, but given that the system is built from a foundation model with broad reasoning capabilities, it can answer many different types of interpretability queries from users, and design experiments on the fly to investigate them."

Neuron by neuron

In one example task, a human user asks MAIA to describe the concepts that a particular neuron inside a vision model is responsible for detecting. To investigate this question, MAIA first uses a tool that retrieves "dataset exemplars" from the ImageNet dataset, which maximally activate the neuron. For this example neuron, those images show people in formal attire, and closeups of their chins and necks. MAIA makes various hypotheses for what drives the neuron's activity: facial expressions, chins, or neckties. MAIA then uses its tools to design experiments to test each hypothesis individually by generating and editing synthetic images—in one experiment, adding a bow tie to an image of a human face increases the neuron's response.

"This approach allows us to determine the specific cause of the neuron's activity, much like a real scientific experiment," says Rott Shaham.

MAIA's explanations of neuron behaviors are evaluated in two key ways. First, synthetic systems with known ground-truth behaviors are used to assess the accuracy of MAIA's interpretations. Second, for "real" neurons inside trained AI systems with no ground-truth descriptions, the authors design a new automated evaluation protocol that measures how well MAIA's descriptions predict neuron behavior on unseen data.

The CSAIL-led method outperformed baseline methods describing individual neurons in a variety of vision models such as ResNet, CLIP, and the vision transformer DINO. MAIA also performed well on the new dataset of synthetic neurons with known ground-truth descriptions. For both the real and synthetic systems, the descriptions were often on par with descriptions written by human experts.

How are descriptions of AI system components—like individual neurons—useful?

"Understanding and localizing behaviors inside large AI systems is a key part of auditing these systems for safety before they're deployed—in some of our experiments, we show how MAIA can be used to find neurons with unwanted behaviors and remove these behaviors from a model," says Schwettmann. "We're building toward a more resilient AI ecosystem where tools for understanding and monitoring AI systems keep pace with system scaling, enabling us to investigate and hopefully understand unforeseen challenges introduced by new models."

Peeking inside neural networks

The nascent field of interpretability is maturing into a distinct research area alongside the rise of "black box" machine learning models. How can researchers crack open these models and understand how they work?

Current methods for peeking inside tend to be limited either in scale or in the precision of the explanations they can produce. Moreover, existing methods tend to fit a particular model and a specific task. This caused the researchers to ask: How can we build a generic system to help users answer interpretability questions about AI models while combining the flexibility of human experimentation with the scalability of automated techniques?

One critical area they wanted this system to address was bias. To determine whether image classifiers displayed bias against particular subcategories of images, the team looked at the final layer of the classification stream (in a system designed to sort or label items, much like a machine that identifies whether a photo is of a dog, cat, or bird) and the probability scores of input images (confidence levels that the machine assigns to its guesses).

To understand potential biases in image classification, MAIA was asked to find a subset of images in specific classes (for example "labrador retriever") that were likely to be incorrectly labeled by the system. In this example, MAIA found that images of black labradors were likely to be misclassified, suggesting a bias in the model toward yellow-furred retrievers.

Since MAIA relies on external tools to design experiments, its performance is limited by the quality of those tools. But, as the quality of tools like image synthesis models improve, so will MAIA. MAIA also shows confirmation bias at times, where it sometimes incorrectly confirms its initial hypothesis. To mitigate this, the researchers built an image-to-text tool, which uses a different instance of the language model to summarize experimental results. Another failure mode is overfitting to a particular experiment, where the model sometimes makes premature conclusions based on minimal evidence.

"I think a natural next step for our lab is to move beyond artificial systems and apply similar experiments to human perception," says Rott Shaham. "Testing this has traditionally required manually designing and testing stimuli, which is labor-intensive. With our agent, we can scale up this process, designing and testing numerous stimuli simultaneously. This might also allow us to compare human visual perception with artificial systems."

"Understanding neural networks is difficult for humans because they have hundreds of thousands of neurons, each with complex behavior patterns. MAIA helps to bridge this by developing AI agents that can automatically analyze these neurons and report distilled findings back to humans in a digestible way," says Jacob Steinhardt, assistant professor at the University of California at Berkeley, who wasn't involved in the research. "Scaling these methods up could be one of the most important routes to understanding and safely overseeing AI systems."

More information: Tamar Rott Shaham et al, A Multimodal Automated Interpretability Agent, *arXiv* (2024). DOI: 10.48550/arxiv.2404.14394

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: Multimodal agent can iteratively design experiments to better understand various components of AI systems (2024, July 23) retrieved 23 July 2024 from https://techxplore.com/news/2024-07-multimodal-agent-iteratively-components-ai.html