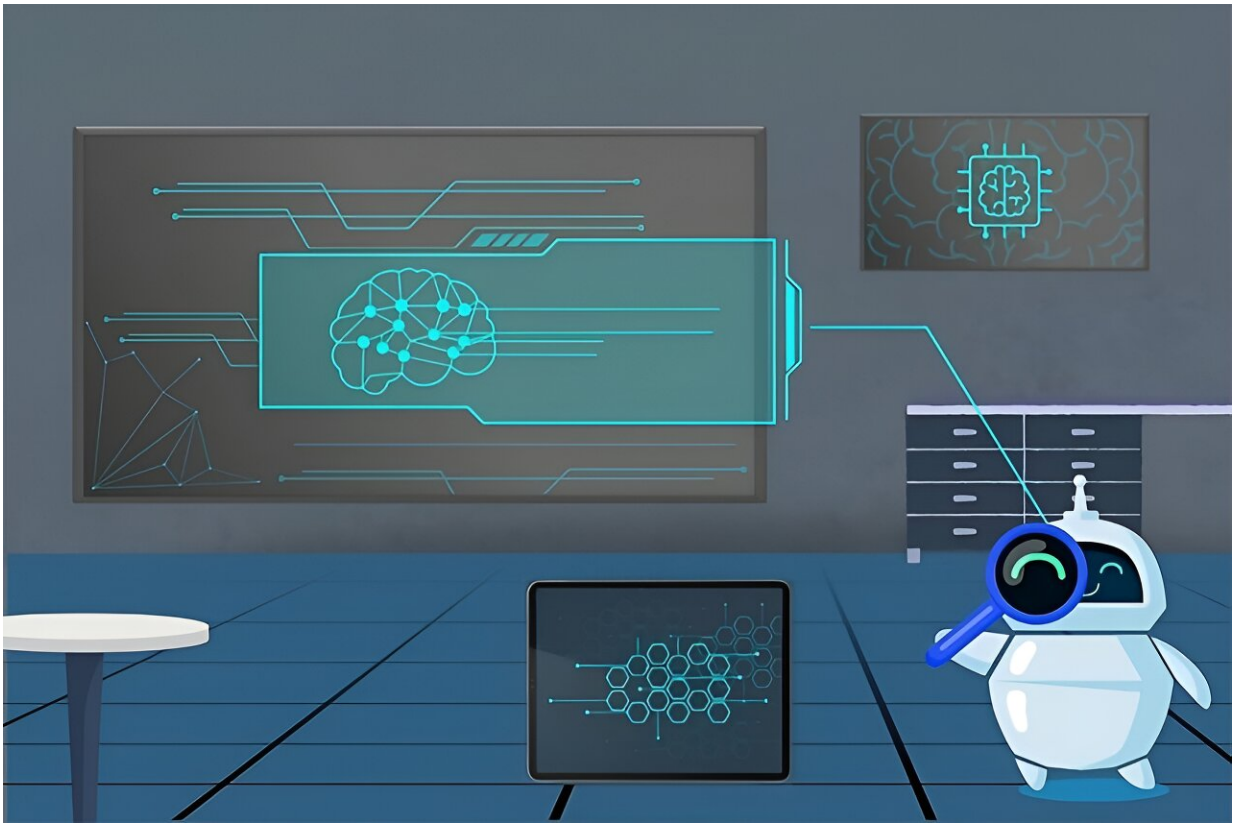
AI agents help explain other AI systems

January 3 2024, by Rachel Gordon



FIND is a new benchmark suite for evaluating automated interpretability methods in neural networks, featuring functions that mimic real-world network components and their complexities. It also presents a novel interactive method using automated interpretability agents, which employ pretrained language models to generate descriptions of function behavior, demonstrating the agent's ability to infer function structure while highlighting the need for further refinement in capturing local details. Credit: Alex Shipps / MIT CSAIL

Explaining the behavior of trained neural networks remains a compelling puzzle, especially as these models grow in size and sophistication. Like other scientific challenges throughout history, reverse-engineering how artificial intelligence systems work requires a substantial amount of experimentation: making hypotheses, intervening on behavior, and even dissecting large networks to examine individual neurons.

To date, most successful experiments have involved large amounts of human oversight. Explaining every computation inside models the size of GPT-4 and larger will almost certainly require more automation—perhaps even using AI models themselves.

Facilitating this timely endeavor, researchers from MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) have developed a novel approach that uses AI models to conduct experiments on other systems and explain their behavior. Their method uses agents built from pretrained language models to produce intuitive explanations of computations inside trained networks.

Central to this strategy is the "automated interpretability agent" (AIA), designed to mimic a scientist's experimental processes. Interpretability agents plan and perform tests on other computational systems, which can range in scale from individual neurons to entire models, in order to produce explanations of these systems in a variety of forms: language descriptions of what a system does and where it fails, and code that reproduces the system's behavior.

Unlike existing interpretability procedures that passively classify or summarize examples, the AIA actively participates in hypothesis formation, experimental testing, and iterative learning, thereby refining its understanding of other systems in real time.

Complementing the AIA method is the new "function interpretation and

description" (FIND) benchmark, a test bed of functions resembling computations inside trained networks, and accompanying descriptions of their behavior.

One key challenge in evaluating the quality of descriptions of real-world network components is that descriptions are only as good as their explanatory power: Researchers don't have access to ground-truth labels of units or descriptions of learned computations. FIND addresses this long-standing issue in the field by providing a reliable standard for evaluating interpretability procedures: explanations of functions (e.g., produced by an AIA) can be evaluated against function descriptions in the benchmark.

For example, FIND contains synthetic neurons designed to mimic the behavior of real neurons inside language models, some of which are selective for individual concepts such as "ground transportation." AIAs are given black-box access to synthetic neurons and design inputs (such as "tree," "happiness," and "car") to test a neuron's response. After noticing that a synthetic neuron produces higher response values for "car" than other inputs, an AIA might design more fine-grained tests to distinguish the neuron's selectivity for cars from other forms of transportation, such as planes and boats.

When the AIA produces a description such as "this neuron is selective for road transportation, and not air or sea travel," this description is evaluated against the ground-truth description of the synthetic neuron ("selective for ground transportation") in FIND. The benchmark can then be used to compare the capabilities of AIAs to other methods in the literature.

Sarah Schwettmann, Ph.D., co-lead author of a paper on the new work and a research scientist at CSAIL, emphasizes the advantages of this approach. The paper is available on the *arXiv* preprint server.

"The AIAs' capacity for autonomous hypothesis generation and testing may be able to surface behaviors that would otherwise be difficult for scientists to detect. It's remarkable that language models, when equipped with tools for probing other systems, are capable of this type of experimental design," says Schwettmann. "Clean, simple benchmarks with ground-truth answers have been a major driver of more general capabilities in language models, and we hope that FIND can play a similar role in interpretability research."

Automating interpretability

Large language models are still holding their status as the in-demand celebrities of the tech world. The recent advancements in LLMs have highlighted their ability to perform complex reasoning tasks across diverse domains. The team at CSAIL recognized that given these capabilities, language models may be able to serve as backbones of generalized agents for automated interpretability.

"Interpretability has historically been a very multifaceted field," says Schwettmann. "There is no one-size-fits-all approach; most procedures are very specific to individual questions we might have about a system, and to individual modalities like vision or language. Existing approaches to labeling individual neurons inside vision models have required training specialized models on human data, where these models perform only this single task.

"Interpretability agents built from language models could provide a general interface for explaining other systems—synthesizing results across experiments, integrating over different modalities, even discovering new experimental techniques at a very fundamental level."

As we enter a regime where the models doing the explaining are black boxes themselves, external evaluations of interpretability methods are

becoming increasingly vital. The team's new benchmark addresses this need with a suite of functions, with known structure, that are modeled after behaviors observed in the wild. The functions inside FIND span a diversity of domains, from mathematical reasoning to symbolic operations on strings to synthetic neurons built from word-level tasks.

The dataset of interactive functions is procedurally constructed; real-world complexity is introduced to simple functions by adding noise, composing functions, and simulating biases. This allows for comparison of interpretability methods in a setting that translates to real-world performance.

In addition to the dataset of functions, the researchers introduced an innovative evaluation protocol to assess the effectiveness of AIAs and existing automated interpretability methods. This protocol involves two approaches. For tasks that require replicating the function in code, the evaluation directly compares the AI-generated estimations and the original, ground-truth functions. The evaluation becomes more intricate for tasks involving natural language descriptions of functions.

In these cases, accurately gauging the quality of these descriptions requires an automated understanding of their semantic content. To tackle this challenge, the researchers developed a specialized "third-party" language model. This model is specifically trained to evaluate the accuracy and coherence of the natural language descriptions provided by the AI systems, and compares it to the ground-truth function behavior.

FIND enables evaluation revealing that we are still far from fully automating interpretability; although AIAs outperform existing interpretability approaches, they still fail to accurately describe almost half of the functions in the benchmark.

Tamar Rott Shaham, co-lead author of the study and a postdoc in

CSAIL, notes that "while this generation of AIAs is effective in describing high-level functionality, they still often overlook finer-grained details, particularly in function subdomains with noise or irregular behavior.

"This likely stems from insufficient sampling in these areas. One issue is that the AIAs' effectiveness may be hampered by their initial exploratory data. To counter this, we tried guiding the AIAs' exploration by initializing their search with specific, relevant inputs, which significantly enhanced interpretation accuracy." This approach combines new AIA methods with previous techniques using pre-computed examples for initiating the interpretation process.

The researchers are also developing a toolkit to augment the AIAs' ability to conduct more precise experiments on neural networks, both in black-box and white-box settings. This toolkit aims to equip AIAs with better tools for selecting inputs and refining hypothesis-testing capabilities for more nuanced and accurate neural network analysis.

The team is also tackling practical challenges in AI interpretability, focusing on determining the right questions to ask when analyzing models in real-world scenarios. Their goal is to develop automated interpretability procedures that could eventually help people audit systems—e.g., for autonomous driving or face recognition—to diagnose potential failure modes, hidden biases, or surprising behaviors before deployment.

Watching the watchers

The team envisions one day developing nearly autonomous AIAs that can audit other systems, with human scientists providing oversight and guidance. Advanced AIAs could develop new kinds of experiments and questions, potentially beyond human scientists' initial considerations.

The focus is on expanding AI interpretability to include more complex behaviors, such as entire neural circuits or subnetworks, and predicting inputs that might lead to undesired behaviors. This development represents a significant step forward in AI research, aiming to make AI systems more understandable and reliable.

"A good benchmark is a power tool for tackling difficult challenges," says Martin Wattenberg, computer science professor at Harvard University who was not involved in the study. "It's wonderful to see this sophisticated benchmark for interpretability, one of the most important challenges in machine learning today. I'm particularly impressed with the automated interpretability agent the authors created. It's a kind of interpretability jiu-jitsu, turning AI back on itself in order to help human understanding."

Schwettmann, Rott Shaham, and their colleagues presented their work at NeurIPS 2023 in December. Additional MIT co-authors, all affiliates of the CSAIL and the Department of Electrical Engineering and Computer Science (EECS), include graduate student Joanna Materzynska, undergraduate student Neil Chowdhury, Shuang Li, Ph.D., Assistant Professor Jacob Andreas, and Professor Antonio Torralba. Northeastern University Assistant Professor David Bau is an additional co-author.

More information: Sarah Schwettmann et al, FIND: A Function Description Benchmark for Evaluating Interpretability Methods, *arXiv* (2023). DOI: 10.48550/arxiv.2309.03886

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