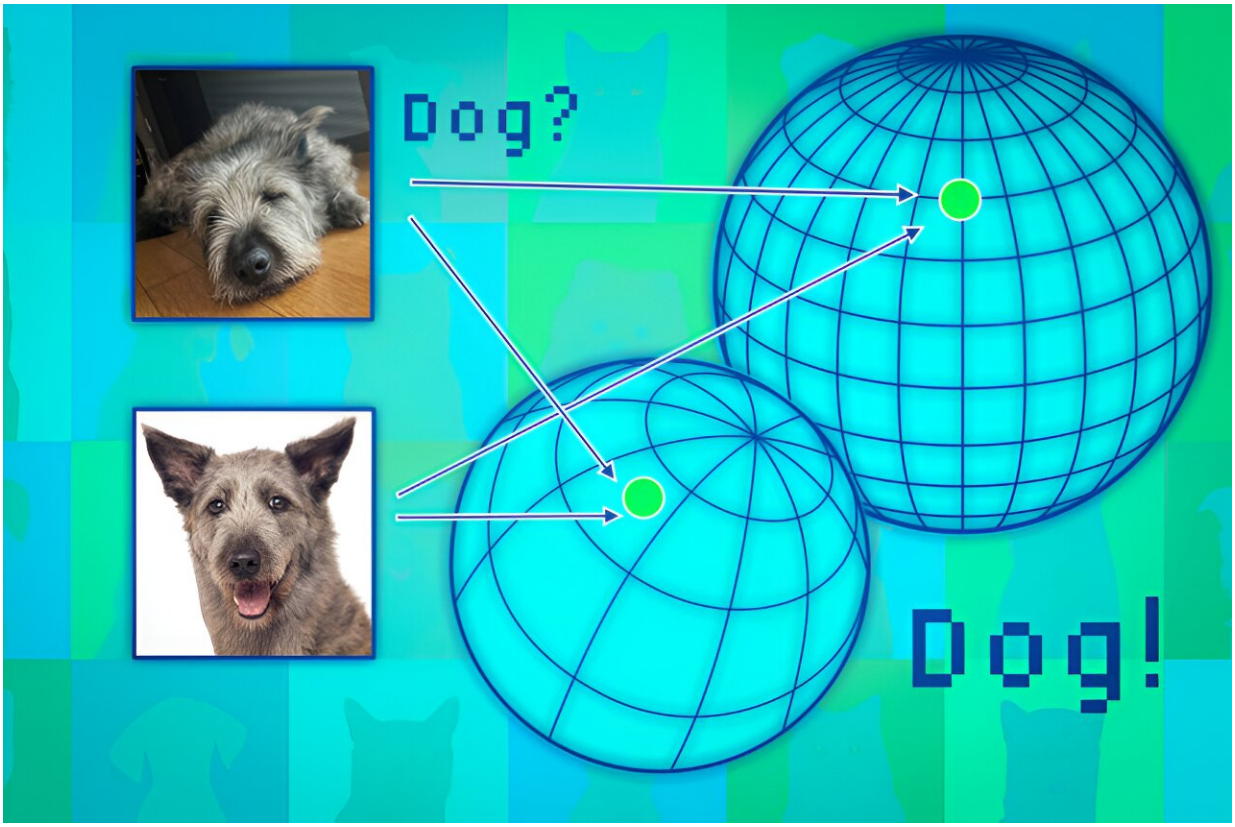
New technique to assess a general-purpose AI model's reliability before it's deployed

July 16 2024, by Adam Zewe



To estimate the reliability of massive deep-learning models called foundation models, MIT researchers developed a technique to assess the consistency of representations an ensemble of similar models learn about the same test data point. Credit: Massachusetts Institute of Technology

Foundation models are massive deep-learning models that have been

pretrained on an enormous amount of general-purpose, unlabeled data. They can be applied to a variety of tasks, like generating images or answering customer questions.

But these models, which serve as the backbone for powerful artificial intelligence tools like ChatGPT and DALL-E, can offer up incorrect or misleading information. In a safety-critical situation, such as a pedestrian approaching a self-driving car, these mistakes could have serious consequences.

To help prevent such mistakes, researchers from MIT and the MIT-IBM Watson AI Lab developed a technique to estimate the reliability of foundation models before they are deployed to a specific task.

They do this by training a set of foundation models that are slightly different from one another. Then they use their algorithm to assess the consistency of the representations that each model learns about the same test data point. If the representations are consistent, it means the model is reliable.

When they compared their technique to state-of-the-art baseline methods, it was better at capturing the reliability of foundation models on a variety of classification tasks.

Someone could use this technique to decide if a model should be applied in a certain setting, without the need to test it on a real-world dataset. This could be especially useful when datasets may not be accessible due to privacy concerns, like in health care settings. In addition, the technique could be used to rank models based on reliability scores, enabling a user to select the best one for their task.

"All models can be wrong, but models that know when they are wrong are more useful. The problem of quantifying uncertainty or reliability

gets harder for these foundation models because their abstract representations are difficult to compare. Our method allows you to quantify how reliable a representation model is for any given input data," says senior author Navid Azizan, the Esther and Harold E. Edgerton Assistant Professor in the MIT Department of Mechanical Engineering and the Institute for Data, Systems, and Society (IDSS), and a member of the Laboratory for Information and Decision Systems (LIDS).

He is joined on a paper about the work by lead author Young-Jin Park, a LIDS graduate student; Hao Wang, a research scientist at the MIT-IBM Watson AI Lab; and Shervin Ardeshir, a senior research scientist at Netflix. The paper will be presented at the Conference on Uncertainty in Artificial Intelligence (UAI 2024), held July 15–19 in Barcelona, and is available on the *arXiv* preprint server.

Counting the consensus

Traditional machine-learning models are trained to perform a specific task. These models typically make a concrete prediction based on an input. For instance, the model might tell you whether a certain image contains a cat or a dog. In this case, assessing reliability could simply be a matter of looking at the final prediction to see if the model is right.

But foundation models are different. The model is pretrained using general data, in a setting where its creators don't know all downstream tasks it will be applied to. Users adapt it to their specific tasks after it has already been trained.

Unlike traditional machine-learning models, foundation models don't give concrete outputs like "cat" or "dog" labels. Instead, they generate an abstract representation based on an input data point.

To assess the reliability of a foundation model, the researchers used an

ensemble approach by training several models which share many properties but are slightly different from one another.

"Our idea is like counting the consensus. If all those foundation models are giving consistent representations for any data in our dataset, then we can say this model is reliable," Park says.

But they ran into a problem: How could they compare abstract representations?

"These models just output a vector, comprised of some numbers, so we can't compare them easily," he adds.

They solved this problem using an idea called neighborhood consistency.

For their approach, the researchers prepare a set of reliable reference points to test on the ensemble of models. Then, for each model, they investigate the reference points located near that model's representation of the test point.

By looking at the consistency of neighboring points, they can estimate the reliability of the models.

Aligning the representations

Foundation models map data points in what is known as a representation space. One way to think about this space is as a sphere. Each model maps similar data points to the same part of its sphere, so images of cats go in one place and images of dogs go in another.

But each model would map animals differently in its own sphere, so while cats may be grouped near the South Pole of one sphere, another model could map cats somewhere in the Northern Hemisphere.

The researchers use the neighboring points like anchors to align those spheres so they can make the representations comparable. If a data point's neighbors are consistent across multiple representations, then one should be confident about the reliability of the model's output for that point.

When they tested this approach on a wide range of classification tasks, they found that it was much more consistent than baselines. Plus, it wasn't tripped up by challenging test points that caused other methods to fail.

Moreover, their approach can be used to assess reliability for any input data, so one could evaluate how well a model works for a particular type of individual, such as a patient with certain characteristics.

"Even if the models all have average performance overall, from an individual point of view, you'd prefer the one that works best for that individual," Wang says.

However, one limitation comes from the fact that they must train an ensemble of large foundation models, which is computationally expensive. In the future, they plan to find more efficient ways to build multiple models, perhaps by using small perturbations of a single model.

More information: Young-Jin Park et al, Quantifying Representation Reliability in Self-Supervised Learning Models, *arXiv* (2023). DOI: 10.48550/arxiv.2306.00206

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