KYUSHU UNIVERSITY

Public Relations Office, Kyushu University
744 Motooka Nishi-ku Fukuoka, 819-0395
TEL: +81-92-802-2130 FAX: +81-92-802-2139
E-MAIL: koho@jimu.kyushu-u.ac.jp
URL: https://www.kyushu-u.ac.jp

(2022/06/30)

Breaking AIs to make them better

Researchers investigate ways to make AIs more robust by studying patterns in their answers when faced with the unknown

Fukuoka, Japan—Today's artificial intelligence systems used for image recognition are incredibly powerful with massive potential for commercial applications. Nonetheless, current artificial neural networks—the deep learning algorithms that power image recognition—suffer one massive shortcoming: they are easily broken by images that are even slightly modified.

This lack of 'robustness' is a significant hurdle for researchers hoping to build better AIs. However, exactly why this phenomenon occurs, and the underlying mechanisms behind it, remain largely unknown.

Aiming to one day overcome these flaws, researchers at Kyushu University's Faculty of Information Science and Electrical Engineering have published in *PLOS ONE* a method called 'Raw Zero-Shot' that assesses how neural networks handle elements unknown to them. The results could help researchers identify common features that make AIs 'non-robust' and develop methods to rectify their problems.

"There is a range of real-world applications for image recognition neural networks, including self-driving cars and diagnostic tools in healthcare," explains Danilo Vasconcellos Vargas, who led the study. "However, no matter how well trained the AI, it can fail with even a slight change in an image."

In practice, image recognition AIs are 'trained' on many sample images before being asked to identify one. For example, if you want an AI to identify ducks, you would first train it on many pictures of ducks.

Nonetheless, even the best-trained AIs can be misled. In fact, researchers have found that an image can be manipulated such that—while it may appear unchanged to the human eye—an AI cannot accurately identify it. Even a single-pixel change in the image can cause confusion.

To better understand why this happens, the team began investigating different image recognition AIs with the hope of identifying patterns in how they behave when faced with samples that they had not been trained with, i.e., elements unknown to the AI.

"If you give an image to an AI, it will try to tell you what it is, no matter if that answer is correct or not. So, we took the twelve most common AIs today and applied a new method called 'Raw Zero-Shot Learning,'" continues Vargas. "Basically, we gave the AIs a series of images with no hints or training. Our hypothesis was that there would be correlations in how they answered.

They would be wrong, but wrong in the same way."

What they found was just that. In all cases, the image recognition AI would produce an answer, and the answers—while wrong—would be consistent, that is to say they would cluster together. The density of each cluster would indicate how the AI processed the unknown images based on its foundational knowledge of different images.

"If we understand what the AI was doing and what it learned when processing unknown images, we can use that same understanding to analyze why AIs break when faced with images with single-pixel changes or slight modifications," Vargas states. "Utilization of the knowledge we gained trying to solve one problem by applying it to a different but related problem is known as Transferability."

The team observed that Capsule Networks, also known as CapsNet, produced the densest clusters, giving it the best transferability amongst neural networks. They believe it might be because of the dynamical nature of CapsNet.

"While today's AIs are accurate, they lack the robustness for further utility. We need to understand what the problem is and why it's happening. In this work, we showed a possible strategy to study these issues," concludes Vargas. "Instead of focusing solely on accuracy, we must investigate ways to improve robustness and flexibility. Then we may be able to develop a true artificial intelligence."

###

For more information about this research, see "Transferability of features for neural networks links to adversarial attacks and defences," Shashank Kotyan, Moe Matsuki, and Danilo Vasconcellos Vargas, *PLOS ONE* (2022). https://doi.org/10.1371/journal.pone.0266060

About Kyushu University

Kyushu University is one of Japan's leading research-oriented institutes of higher education since its founding in 1911. Home to around 19,000 students and 8,000 faculty and staff, Kyushu U's world-class research centers cover a wide range of study areas and research fields, from the humanities and arts to engineering and medical sciences. Its multiple campuses—including the largest in Japan—are located around Fukuoka City, a coastal metropolis on the southwestern Japanese island of Kyushu that is frequently ranked among the world's most livable cities and historically known as a gateway to Asia.

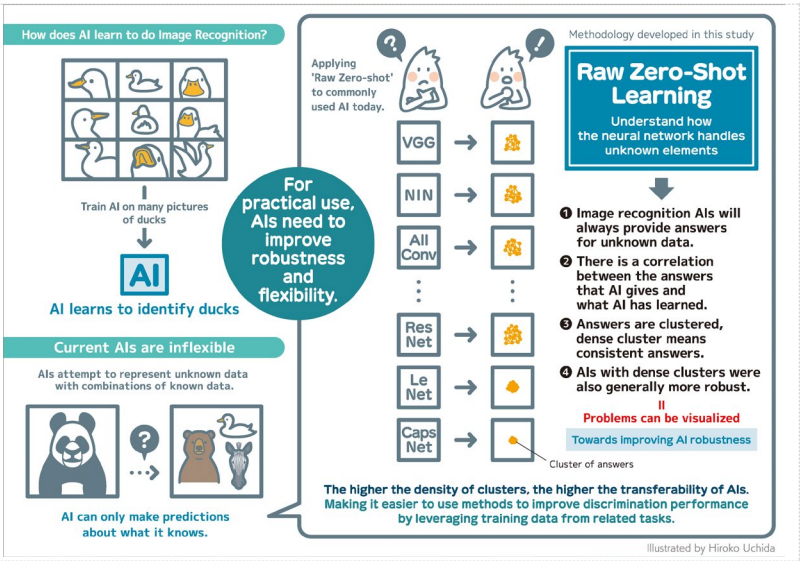

Fig. 1. Graphical abstract, Raw Zero-Shot. Image recognition AIs are powerful but inflexible and cannot recognize images unless they are trained on specific data. In Raw Zero-Shot Learning, researchers give these image recognition AIs a variety of data and observe the patterns in their answers. The research team hopes that this methodology can help improve the robustness of future AI. Illustrated by Hiroko Uchida

---

[Contact]
Danilo Vasconcellos Vargas, Associate Professor
Faculty of Information Science and Electrical Engineering, Department of Informatics
Tel: +81-92-802-3809
E-mail: vargas@inf.kyushu-u.ac.jp