KYUSHU UNIVERSITY

Public Relations Initiative, Kyushu University
744 Motooka Nishi-ku Fukuoka, 819-0395
TEL: +81-92-802-2443 FAX: +81-92-802-2139
E-MAIL: sysintlkh@jimu.kyushu-u.ac.jp
URL: https://www.kyushu-u.ac.jp

PRESS RELEASE (2024/12/12)

Unlocking the 'black box': scientists reveal AI's hidden thoughts

Researchers introduce a new method to assess how deep neural networks interpret information, ensuring its reliability and robustness for real-world applications.

Fukuoka, Japan – Deep neural networks are a type of artificial intelligence (AI) that imitate how human brains process information, but understanding how these networks "think" has long been a challenge. Now, researchers at Kyushu University have developed a new method to understand how deep neural networks interpret information and sort it into groups. Published in *IEEE Transactions on Neural Networks and Learning Systems,* the study addresses the important need to ensure AI systems are accurate and robust and can meet the standards required for safe use.

Deep neural networks process information in many layers, similarly to humans solving a puzzle step by step. The first layer, known as the input layer, brings in the raw data. The subsequent layers, called hidden layers, analyze the information. Early hidden layers focus on basic features, such as detecting edges or textures—like examining individual puzzle pieces. Deeper hidden layers combine these features to recognize more complex patterns, such as identifying a cat or a dog—similar to connecting puzzle pieces to reveal the bigger picture.

"However, these hidden layers are like a locked black box: we see the input and output, but what is happening inside is not clear," says Danilo Vasconcellos Vargas, Associate Professor from the Faculty of Information Science and Electrical Engineering at Kyushu University. "This lack of transparency becomes a serious problem when AI makes mistakes, sometimes triggered by something as small as changing a single pixel. AI might seem smart, but understanding how it comes to its decision is key to ensuring it's trustworthy."

Currently, methods for visualizing how AI organizes information rely on simplifying high-dimensional data into 2D or 3D representations. These methods let researchers observe how AI categorizes data points—for example, grouping images of cats close to other cats while separating them from dogs. However, this simplification comes with critical limitations.

"When we simplify high-dimensional information into fewer dimensions, it's like flattening a 3D object into 2D—we lose important details and fail to see the whole picture. Additionally, this method of visualizing how the data is grouped makes it difficult to compare between different neural networks or data classes," explains Vargas.

In this study, the researchers developed a new method, called the k* distribution method, that more clearly visualizes and assesses how well deep neural networks categorize related items together.

The model works by assigning each inputted data point a "k* value" which indicates the distance to the nearest unrelated data point. A high k* value means the data point is well-separated (e.g., a cat far from any dogs), while a low k* value suggests potential overlap (e.g., a dog closer to a cat than other cats). When looking at all the data points within a class, such

as cats, this approach produces a distribution of k* values that provides a detailed picture of how the data is organized.

"Our method retains the higher dimensional space, so no information is lost. It's the first and only model that can give an accurate view of the 'local neighborhood' around each data point," emphasizes Vargas.

Using their method, the researchers revealed that deep neural networks sort data into clustered, fractured, or overlapping arrangements. In a clustered arrangement, similar items (e.g., cats) are grouped closely together, while unrelated items (e.g., dogs) are clearly separated, meaning the AI is able to sort the data well. Fractured arrangements, however, indicate that similar items are scattered across a wide space, while overlapping distributions occur when unrelated items are in the same space, with both arrangements making classification errors more likely.

Vargas compares this to a warehouse system: "In a well-organized warehouse, similar items are stored together, making retrieval easy and efficient. If items are intermixed, they become harder to find, increasing the risk of selecting the wrong item."

AI is increasingly used in critical systems like autonomous vehicles and medical diagnostics, were accuracy and reliability is essential. The k* distribution method helps researchers, and even lawmakers, evaluate how AI organizes and classifies information, pinpointing potential weaknesses or errors. This not only supports the legalization processes needed to safely integrate AI into daily life but also offers valuable insights into how AI "thinks". By identifying the root causes of errors, researchers can refine AI systems to make them not only accurate but also robust—capable of handling blurry or incomplete data and adapting to unexpected conditions.

"Our ultimate goal is to create AI systems that maintain precision and reliability, even when faced with the challenges of real-world scenarios," concludes Vargas.

Written by Science Communicator Intern, Negar Khalili

###

For more information about this research, see "k* Distribution: Evaluating the Latent Space of Deep Neural Networks Using Local Neighborhood Analysis," Shashank Kotyan; Tatsuya Ueda; Danilo Vasconcellos Vargas, *IEEE Transactions on Neural Networks and Learning Systems*, https://doi.org/10.1109/TNNLS.2024.3446509

About Kyushu University

Founded in 1911, Kyushu University is one of Japan's leading research-oriented institutes of higher education, consistently ranking as one of the top ten Japanese universities in the Times Higher Education World University Rankings and the QS World Rankings. The university is one of the seven national universities in Japan, located in Fukuoka, on the island of Kyushu—the most southwestern of Japan's four main islands with a population and land size slightly larger than Belgium. Kyushu U's multiple campuses—home to around 19,000 students and 8000 faculty and staff—are located around Fukuoka City, a coastal metropolis that is frequently ranked among the world's most livable cities and historically known as Japan's gateway to Asia. Through its VISION 2030, Kyushu U will "drive social change with integrative knowledge." By fusing the spectrum of knowledge, from the humanities and arts to engineering and medical sciences, Kyushu U will strengthen its research in the key areas of decarbonization, medicine

and health, and environment and food, to tackle society's most pressing issues.

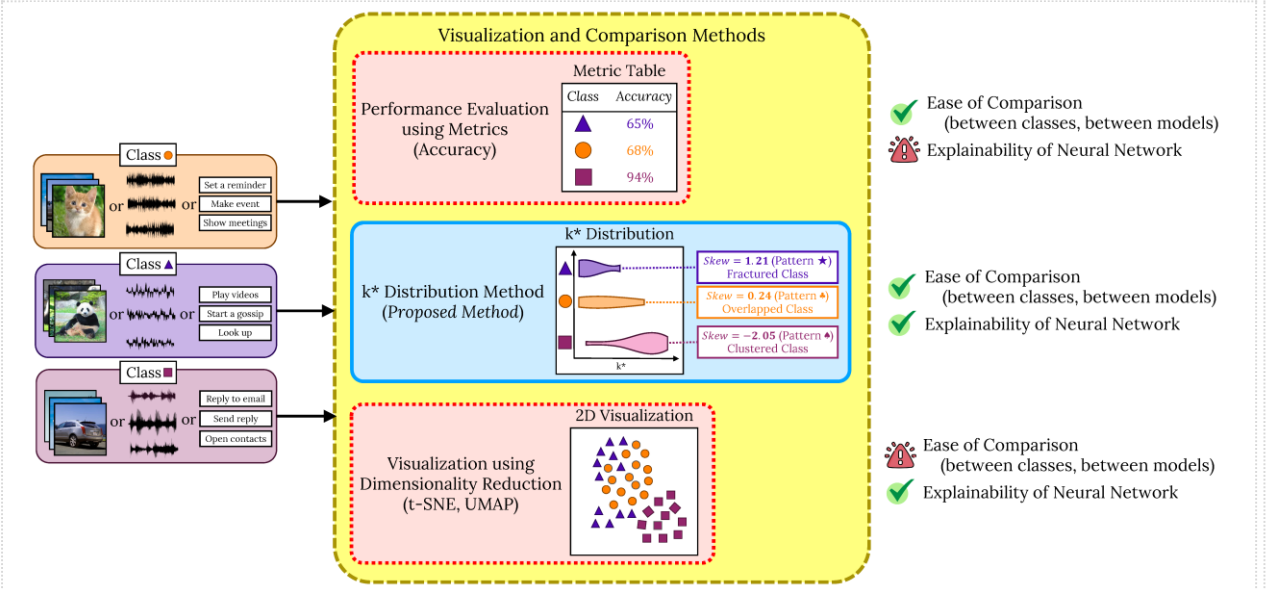

Fig. 1. The k* distribution method allows clear visualization and evaluation of how a neural network interprets data.

[Contact]
Danilo Vasconcellos Vargas, Associate Professor
Department of Informatics, Faculty of Information Science and Electrical Engineering
Tel: +81-92-802-3599
E-mail: vargas@inf.kyushu-u.ac.jp