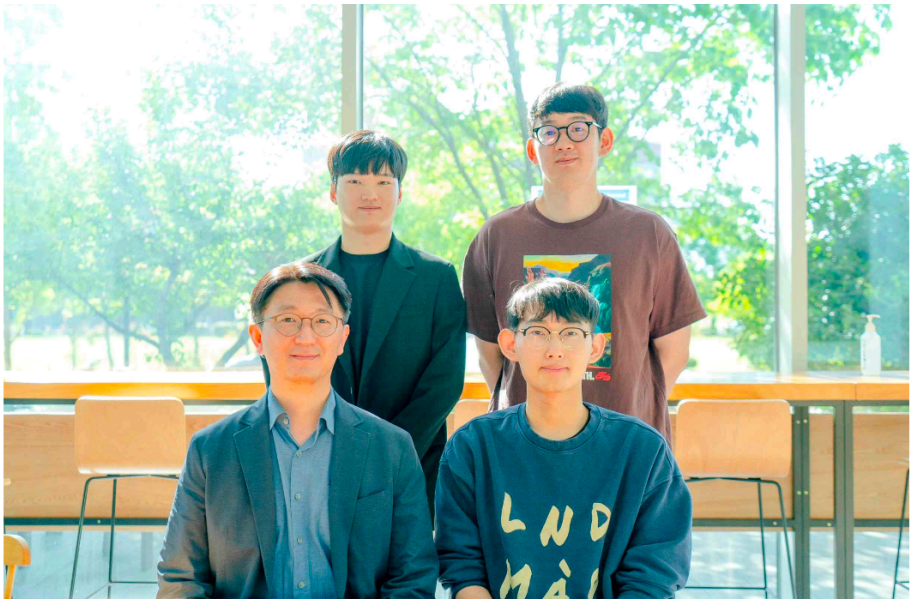
AI, which claimed to know what it didn't know, will be able to recognize "I don't know what I don't know"

- Development of technology to identify unlearned data... Improvement in safety of self-driving car and medical diagnosis AI is expected

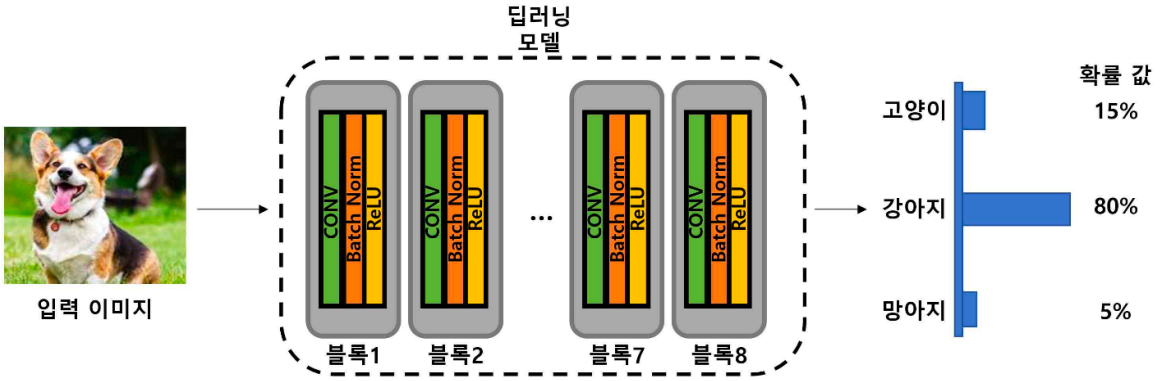
- GIST Professor Kyoobin Lee's team to present at the world's No. 1 computer vision conference <CVPR Conference> on June 18



[Photo] (front row from left) School of Integrated Technology Professor Kyoobin Lee and Ph.D. student Yeonguk Yu (back row from left), Ph.D. student Seongju Lee, and Ph.D. student Sungho Shin

Since the appearance of AlphaGo in 2016, artificial intelligence (AI) technology has developed rapidly and is widely used in real life. Most AIs used today are designed to find the most similar answer if there is no correct answer among given candidates.

In particular, the deep learning model is widely used in the field of computer vision because of its excellent image recognition ability. The disadvantage is that even if the answer is not known, the most similar value is mistakenly recognized as the correct answer. In this case, the self-driving vehicle can cause serious problems such as misrecognizing obstacles, so the need for an AI model to compensate for this is being raised.



[Figure 1] A diagram of the structure of a deep learning model in which probability values are calculated for an input image (dog). A deep learning model is composed of blocks composed of several

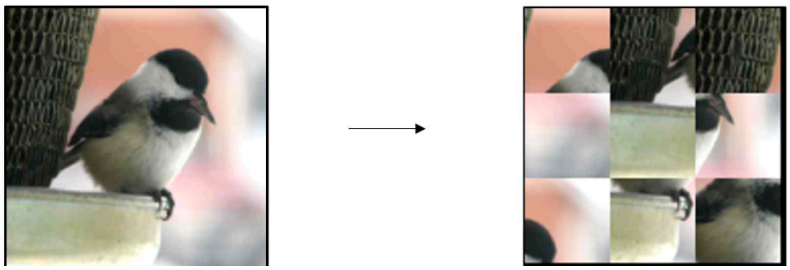
layers. The categories learned from that figure (i.e. inputs within the distribution) are cats, dogs, and foals. Even if AI does not know the answer, it incorrectly recognizes the most similar value in the previously learned category as the correct answer.

GIST (Gwangju Institute of Science and Technology, Acting President Raekil Park) School of Integrated Technology Professor Kyoobin Lee's research team developed an AI technology that distinguishes 'unknown data' that has not been learned.

An AI model is made up of several blocks, each performing the same task. It is like materials (data) coming in on a conveyor belt, and several people (blocks) dividing their labor to complete things in order. The research team used a jigsaw puzzle to find a block suitable for detecting 'unknown data', and proposed a method for detecting unknown data based on the activity* of the block.

* activity: A block outputs a feature map for an input image, which means the size of the feature map. The size (activity) decreases for unknown data and increases for known data.

As an example of unknown data, the research team split the image into small pieces like a jigsaw puzzle and randomly shuffled them into input. This is to find a block suitable for detecting unknown data according to the activity after inputting data similar to the actual image but not the correct answer.



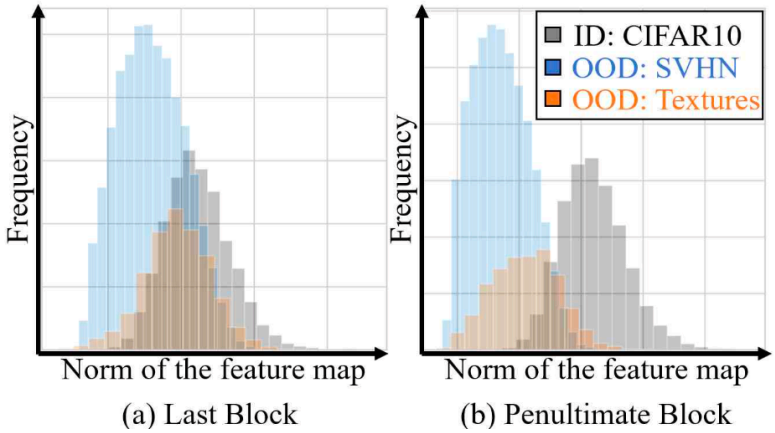
[Picture 2] An example of an existing image and jigsaw puzzle. In this study, a jigsaw puzzle image was used as an example of a kind of ‘unknown data’ input (non-distributed input). This is because the information of the object in the existing image is destroyed in the jigsaw puzzle.

Previous studies used the last block that learned the most data, but the research team found that the last block tends to mistake unknown data for known data due to excessive learning.

The research team reported that blocks with low activity for unknown data (jigsaw puzzle) and high activity for known data were most suitable for detecting unknown data. The block with the highest activity for the learned image was selected compared to the activity for the jigsaw puzzle.

In this way, detection results improved by 5.8% in the first benchmark* and 6.8% in the second benchmark, the highest level of performance to date.

* benchmark: It means that the evaluation environment is configured with the same dataset for fair performance comparison of research results. The first is the CIFAR10 benchmark, the second is the ImageNet benchmark.



[Figure 3] Activation when inputs within the distribution (grey, known data) and out-of-distribution inputs (blue, orange) are received in the last block (a) and the previous block (b) in the model trained with the CIFAR10 dataset Histogram comparing the degree. As shown in (b), the block suitable for detecting 'unknown data (OOD)' should have a high activity for 'known data (ID)' and a small activity for 'unknown data'.

If metacognition* of a deep learning model becomes possible with this research result, it will be possible to develop an AI model in the form of augmenting intelligence. In addition, it is expected to be useful in sensitive fields directly related to safety and life, such as autonomous driving and medical diagnosis.

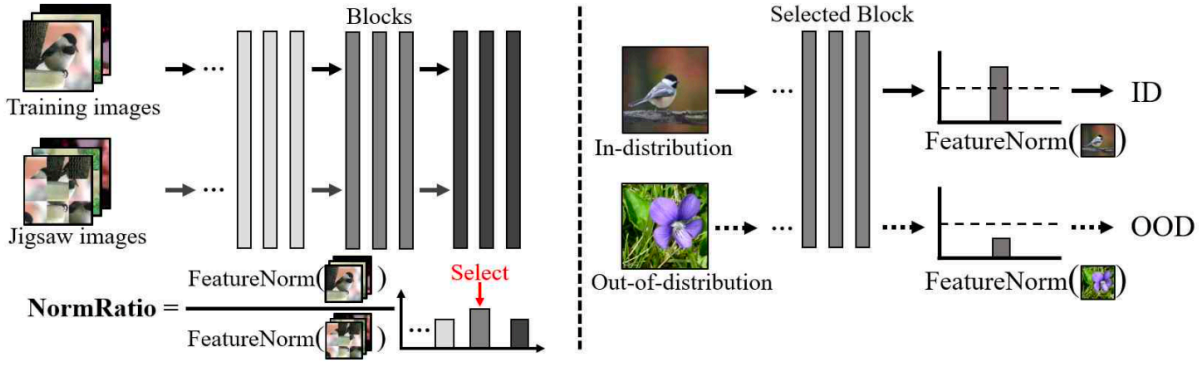
It can prevent problems such as mistakenly recognizing an animal as a person while driving an autonomous vehicle and making a sudden stop or misdiagnosing a skin disease that has never been learned as the most similar skin disease among previously learned skin diseases.

* metacognition: The ability to make judgments about one's thoughts. In this study, it means the ability to judge that you know what you know and don't know what you don't know.

Professor Kyoobin Lee said, "If the results of this research are developed, the deep learning model can acquire the metacognitive ability to recognize the recognized result on its own. It is expected that not only can it prevent enormous damage that can occur due to misrecognition as knowing what it does not know, but it will also be applied to various technologies such as intelligence augmentation."

The research was led by Professor Lee and conduced by Ph.D. student Yeonguk Yu, Ph.D. student Seongju Lee, Ph.D. student Sungho Shin, and master's student Changhyun Jun with the support of the Ministry of Science and ICT's Cloud Robot Complex Artificial Intelligence Core Technology Development Project and Agent Technology Development Project.

The results of this research will be presented on June 18th at the Computer Vision and Pattern Recognition Conference (CVPR), the world's top computer vision conference. The code used in the study is available as open source on GitHub. (https://github.com/gist-ailab/block-selection-for-OOD-detection)



(a) Block Selection using **NormRatio** (b) OOD detection with **FeatureNorm**

[Figure 4] Overview of the 'unknown data' detection method proposed in this study. The proposed method first calculates the ratio of activity (training image / jigsaw puzzle image) for all blocks. The block with the largest corresponding value is selected (a), and then, based on the activity of the suitable block, it is determined whether it is known data (ID) or unknown data (OOD) (b).