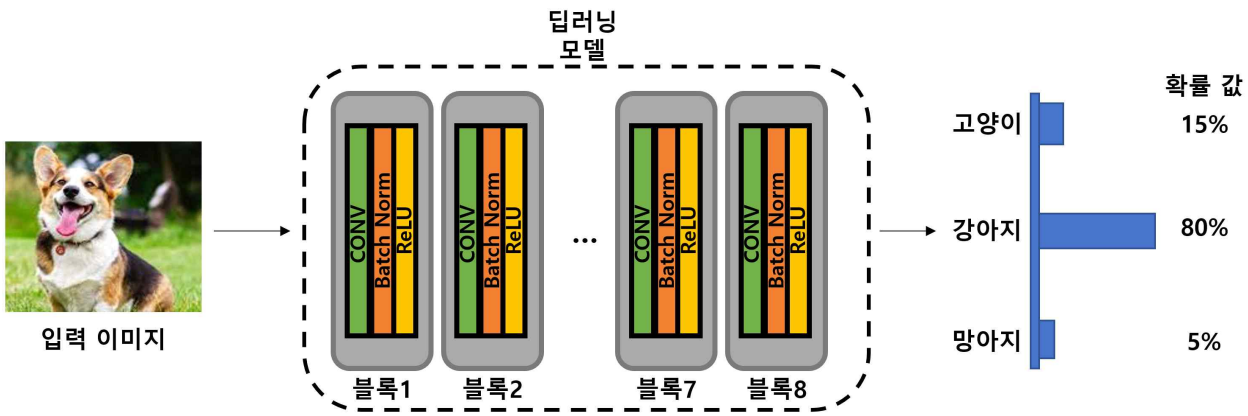
모르는 것도 안다고 우기던 AI,
“모르는 건 모른다”인지할 수 있게 된다

- 미학습 데이터 식별하는 기술 개발... 자율주행차, 의료진단AI 안전성 개선 기대돼
- 지스트 이규빈 교수팀, 컴퓨터비전 분야 세계 1위 <CVPR 학회>서 6월 18일 발표 예정

[사진] (앞줄 왼쪽부터) 융합기술학제학부 이규빈 교수, 유연국 박사과정생
(뒷줄 왼쪽부터) 이성주 박사과정생, 신성호 박사과정생

인공지능(AI) 기술은 2016년 알파고 등장 이후 급속도로 발전해 실생활에 폭넓게 활용되고 있다. 오늘날 이용되는 대부분의 AI는 주어진 후보 중 **정답이 없으면 가장 비슷한 답을 찾도록 설계**됐다.

특히 **딥러닝 모델(Deep learning, 심층학습)**은 이미지 인식 능력이 탁월해 컴퓨터비전 분야에서 다양하게 활용되고 있으나, **답을 몰라도 가장 유사한 값을 정답으로 잘못 인식**한다는 단점이 있다. 이 경우 **자율주행 차량이 장애물을 잘못 인식**하는 등 심각한 문제를 일으킬 수 있어 이를 보완할 AI모델의 필요성이 제기되고 있다.



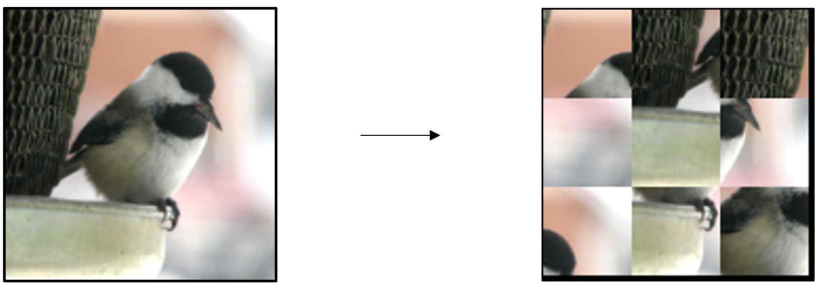
[그림1] 입력된 이미지(강아지)에 대해서 확률 값이 계산되는 딥러닝 모델 구조 도식. 딥러닝 모델은 여러 층(Layer)로 구성된 블록으로 이루어져 있다. 해당 그림에서 학습된 카테고리(즉, 분포 내 입력)는 고양이, 강아지, 망아지이다. AI는 답을 몰라도 기존에 학습한 카테고리에서 가장 유사한 값을 정답으로 잘못 인식하게 된다.

지스트(광주과학기술원, 총장직무대행 박래길) 융합기술학제학부 이규빈 교수 연구팀은 **학습한 적 없는 '모르는 데이터'를 구별해 내는 AI 기술을 개발**했다.

AI모델은 여러 블록으로 구성되어 있는데, 각 블록은 똑같은 작업을 수행한다. 컨베이어 벨트에 재료(데이터)가 들어오고, 여러 사람(블록)이 분업하여 순서대로 물건을 완성하는 것과 같다. 연구팀은 이 중 '모르는 데이터' 탐지에 적합한 블록을 찾아내기 위해 **직소 퍼즐을 이용**했으며, **블록의 활성도*를 기준으로 모르는 데이터를 탐지하는 방법**을 제안했다.

* **활성도**: 블록은 입력된 이미지에 대해 특징 맵을 출력하는데 그 특징 맵의 크기를 뜻한다. 모르는 데이터에 대해서는 크기(활성도)가 작아지고 아는 데이터에 대해서는 커진다.

연구팀은 **모르는 데이터의 예시**로써 **이미지를 직소 퍼즐처럼 잘게 쪼갠 뒤 무작위로 섞어서 입력**했다. 실제 이미지와 유사하지만 정답은 아닌 데이터를 입력한 후 활성도에 따라 모르는 데이터 탐지에 적합한 블록을 찾기 위해서다.



기존 이미지 → **기존 이미지로부터 만들어진 직소 퍼즐**

[그림2] 기존 이미지와 직소 퍼즐 예시. 본 연구에서는 직소 퍼즐 이미지를 일종의 '모르는 데이터' 입력(분포 외 입력)의 예시로 사용했다. 직소 퍼즐에는 기존 이미지에 있던 물체의 정보가 파괴되기 때문이다.

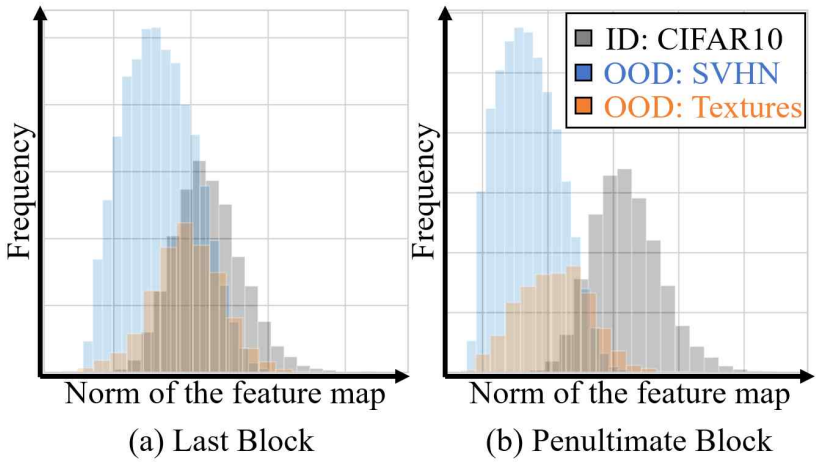
기존 연구에서는 가장 많은 데이터를 학습한 마지막 블록을 사용했으나, 연구팀은 마지막 블록이 과도한 학습으로 인해 모르는 데이터도 아는 데이터로 착각하는 경향이 있다는 점을 밝혀냈다.

연구팀은 **모르는 데이터(직소 퍼즐)에는 낮은 활성도**를, **아는 데이터에는 높은 활성도를 보이는 블록**이 모르는 데이터 탐지에 가장 적합한 것으로 보고, 직소 퍼즐에 대한 활성도 대비 학습된 이미지에 대한 활성도가 가장 높은 블록을 선택했다.

이 방식으로 기존에 사용하던 **첫 번째 벤치마크*에서는 5.8%, 두 번째 벤치마크에**

서는 6.8% 향상된 탐지 결과를 얻어 **현재까지 가장 높은 수준의 성능이 달성**됐다.

* **벤치마크**: 연구 결과의 공정한 성능 비교를 위해 동일한 데이터셋으로 평가 환경을 구성해둔 것을 뜻한다. 첫 번째는 CIFAR10 벤치마크, 두 번째는 ImageNet 벤치마크.



[그림3] CIFAR10 데이터셋으로 학습된 모델에서 마지막 블록(a)와 그 이전 블록(b)에서의 분포 내 입력(회색, 아는 데이터) 및 분포 외 입력(파랑, 주황)이 들어왔을 때의 활성화 정도를 비교한 히스토그램. '모르는 데이터(OOD)' 탐지에 적합한 블록일수록 (b)와 같이 '아는 데이터(ID)'에 대한 활성도는 크며, '모르는 데이터'에 대한 활성도는 작아야 함.

이번 연구성과로 딥러닝 모델의 메타인지*가 가능해지면 **지능을 증강하는 형태의 AI 모델**도 개발할 수 있게 된다. 또, **자율주행, 의료 진단 등 안전이나 생명과 직결되는 민감한 분야**에서 유용하게 이용할 수 있을 것으로 기대된다.

자율주행차 운행 중 동물을 사람으로 잘못 인식해 급정거하거나 학습한 적 없는 피부병을 기존에 학습한 피부병 중 가장 유사한 질환으로 오진하는 것과 같은 문제를 방지할 수 있다.

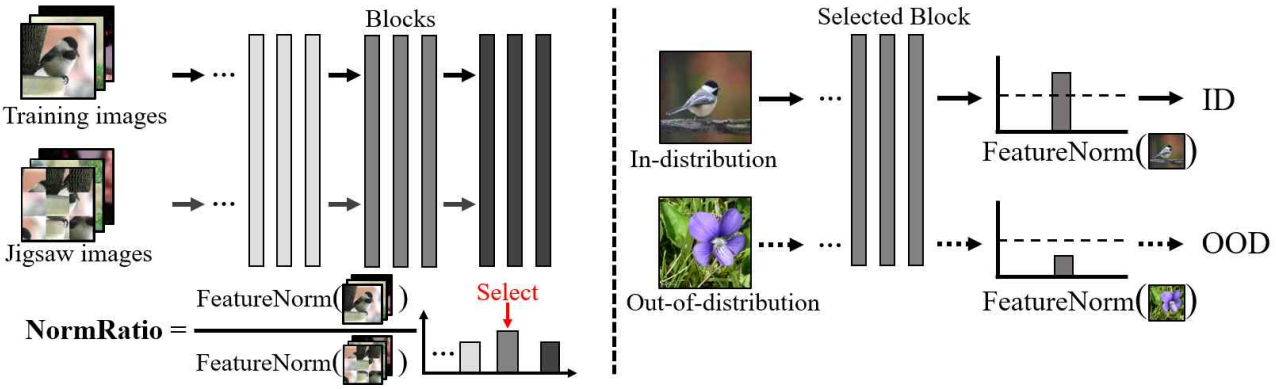
* **메타인지**: 자신의 생각에 대해 판단하는 능력. 본 연구에서는 아는 것을 안다, 모르는 것을 모른다라고 판단할 수 있는 능력을 뜻한다.

이규빈 교수는 "이번 연구성과를 발전시키면 딥러닝 모델이 **인식된 결과를 스스로 인지하는 메타인지 능력**을 얻을 수 있다"며 "모르는 것을 아는 것으로 잘못 인식해 발생할 수 있는 막대한 피해를 방지할 수 있을 뿐만 아니라, 지능 증강과 같은 다양한 기술로 응용될 것이라고 기대한다"고 밝혔다.

이 교수가 지도하고 유연국 박사과정생이 신성호 박사과정생, 이성주 박사과정생, 전창현 석사와 함께 진행한 이번 연구는 **과학기술정보통신부의 클라우드 로봇복합 인공지능 핵심기술개발사업**, **불확실성을 자각하고 성장하는 에이전트 기술개발사업**의 지원을 받아 수행됐다.

이번 연구 성과는 컴퓨터비전 분야에서 **세계 최고 수준의 학회인 <컴퓨터비전과**

패턴인식 학술대회(CVPR, Computer Vision and Pattern Recognition Conference)>에서 오는 **6월 18일 발표**될 예정이다. 연구에 사용된 코드는 깃허브에서 오픈소스로 이용할 수 있다. (https://github.com/gist-ailab/block-selection-for-OOD-detection)



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

[그림4] 이번 연구에서 제안된 '모르는 데이터' 탐지 방법의 개요도. 제안된 방법은 먼저, 활성도의 비율(학습 이미지 / 직소 퍼즐 이미지)을 모든 블록에 대해 계산하고, 가장 해당 값이 큰 블록을 선택한다 (a) 그 후, 적합한 블록의 활성도를 기반으로 아는 데이터(ID) 인지 모르는 데이터(OOD)인지를 판단함(b).

논문의 주요 정보

1. 논문명, 저자정보

- 학회명: IEEE / CVF Computer Vision and Pattern Recognition Conference(CVPR)

  * 세계 최고 컴퓨터 비전 학회(2022년 기준 h5-index=389 in Computer Vision and Pattern Recognition, Top 1)이며, 인공지능/컴퓨터비전 분야 최우수 학술대회, 한국정보과학회 기준 최우수 학술대회(S급)

- 논문명: Block Selection Method for Using Feature Norm in Out-of-distribution Detection

- 저자 정보: 유연국 박사과정생(제1저자, 융합기술학제학부), 신성호 박사과정생, 이성주 박사과정생, 전창현 석사, 이규빈 교수(교신저자, 융합기술학제학부)