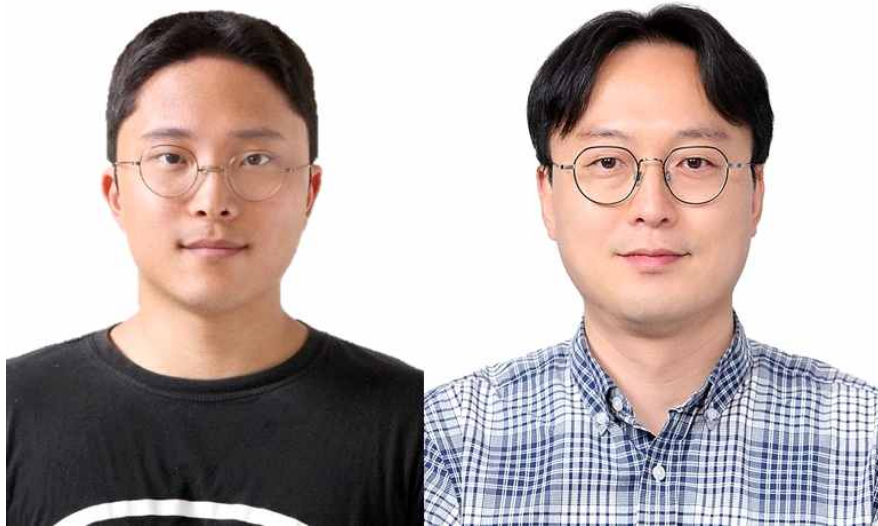


“자율주행차가 보행자를 놓친다면?”

GIST, AI의 시각 인식 오류 유도하는 새 기술 개발 취약점 진단 정밀도 높여

- AI학과 김승준 교수팀 AI가 학습한 정보를 활용해 혼동하기 쉬운 대상 찾아 인식 오류 유도하는 기술 'LSAA' 개발... 도심 주행 영상 실험서 사물 인식 성능 지표 68.35% → 0.41%로 하락
- 150종 사물배경 구분하는 AI에서도 취약점 확인... 로봇 분야 국제학술대회 'IROS 2026' 발표



▲ (왼쪽부터) GIST AI학과 김광빈 박사과정생, 김승준 교수

광주과학기술원(GIST·지스트, 총장 임기철)은 AI학과 김승준 교수 연구팀이 인공지능(AI)이 사물을 잘못 인식하도록 유도해 AI의 취약점을 찾아내는 새로운 ‘적대적 공격*’ 기법 ‘LSAA’를 개발했다고 밝혔다.

LSAA(Latent Space-Aware Adversarial Attack)는 AI가 학습 과정에서 형성한 내부 정보를 분석해 인식 오류를 일으키기 쉬운 대상을 찾고, 영상에 미세한 변화를 가해 AI가 사물을 잘못 인식하도록 유도하는 기술이다. 자율주행차와 로봇의 ‘눈’ 역할을 하는 시각 인식 AI의 취약점을 사전에 점검하는 데 활용될 것으로 기대된다.

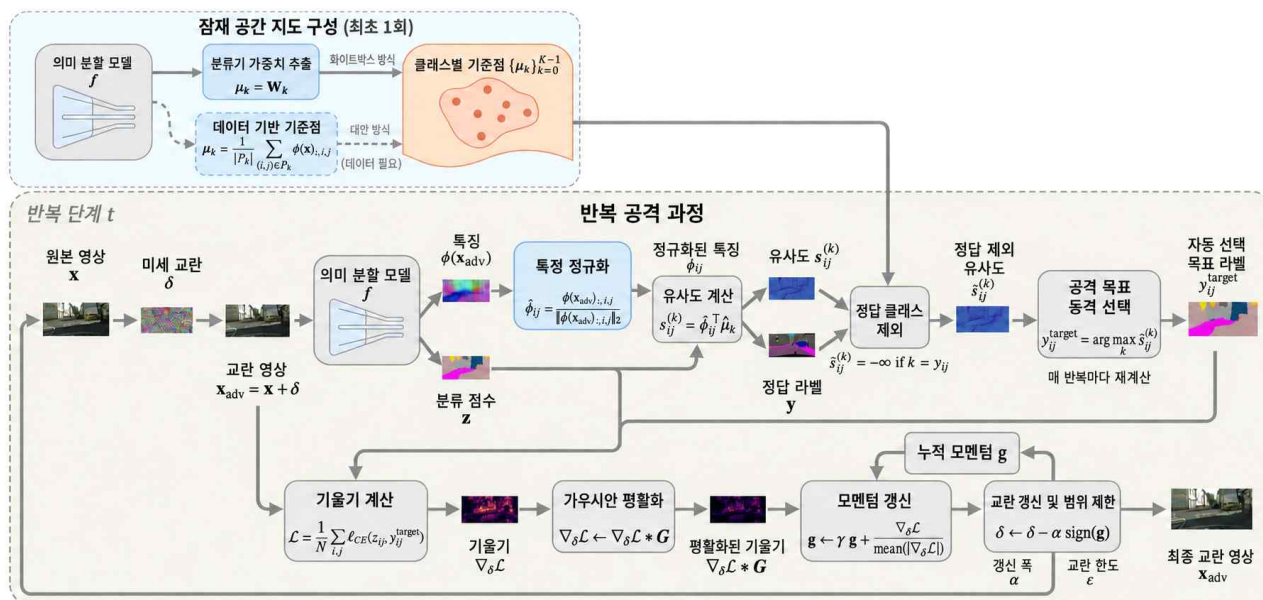
* **적대적 공격(Adversarial Attack)**: 사람의 눈으로는 알아차리기 어려울 만큼 영상 등 입력 데이터를 미세하게 바꿔 AI가 잘못 판단하도록 유도하는 기법이다. 이름은 ‘공격’이지만, AI가 어떤 조건에서 취약한지 미리 확인하고 안전성을 검증하는 ‘모의 해킹’ 도구로도 널리 쓰인다.

자율주행차는 카메라 영상에서 도로, 보행자, 차량 등을 식별해 주변 상황을 파악한다. 이때 영상의 각 화소(픽셀)가 어떤 사물이나 배경에 해당하는지 판별하는 AI 기술인 ‘의미 분할(Semantic Segmentation)’이 활용된다.

그러나 시각 인식 AI는 사람의 눈으로 구별하기 어려운 작은 영상 변화에도 사물을 잘못 인식할 수 있다. 예를 들어 도로 위 보행자를 배경의 일부로 잘못 인식하면 주변 상황을 제대로 파악하지 못해 주행 판단에 영향을 미칠 수 있다. 따라서 AI가 어떤 조건에서 인식 오류를 일으키는지 사전에 파악하는 것이 중요하다.

기존의 적대적 공격 기법은 영상에 변화를 가해 AI가 잘못 인식하도록 유도하거나, 사람이 미리 설정한 사물 간의 유사성이나 관계를 이용해 인식 오류를 일으키는 방식이었다. 그러나 사람이 비슷하다고 판단한 사물을 AI도 비슷하게 인식하는 것은 아니다. 또한 구분해야 할 사물의 종류가 많아질수록 사물 간 관계를 일일이 설정하기 어렵다는 한계가 있었다.

이에 연구팀은 AI가 학습 과정에서 사물의 특징을 수치로 표현하고 정리하는 내부 공간인 '잠재 공간(Latent Space)*'에 주목했다. 이 공간에서 특징이 비슷하게 표현되는 대상은 AI가 서로 혼동할 가능성이 있다는 점에 착안했다.



▲ LSAA를 이용한 AI 취약점 진단 과정. AI가 학습한 내부 정보(잠재 공간)를 분석해 영상의 각 화소(픽셀)에서 혼동하기 쉬운 대상을 찾고, 해당 대상으로 잘못 인식하도록 영상에 미세한 변화를 가한다. 이 과정에서 혼동하기 쉬운 대상을 계속 다시 선택해 인식 오류를 유도한다.

연구팀이 개발한 LSAA는 잠재 공간의 정보를 활용해 영상의 각 화소에서 정답을 제외하고 AI가 가장 혼동하기 쉬운 대상을 찾아낸 뒤, 해당 대상으로 잘못 인식하도록 영상에 미세한 변화를 가한다.

특히 영상에 변화를 가할 때마다 혼동하기 쉬운 대상을 다시 선택해, AI의 인식 상태 변화에 맞춰 공격 방향을 조정하도록 했다.

* 잠재 공간(Latent Space): AI가 입력 영상을 숫자 형태의 특징으로 바꿔 정리해 두는 내부 공간이다. 이 공간에서 서로 가까이 놓인 대상일수록 AI가 비슷하다고 여기며, 그만큼 헛갈리기도 쉽다.

연구진은 LSAA의 성능을 검증하기 위해 도심 주행 영상, 야간 주행 영상, 다양한 사물과 배경이 포함된 영상 등 3종의 데이터세트와 3종의 AI 모델을 이용해 기존 공격 기법과 비교 실험을 진행했다.

그 결과, 도심 주행 영상에서 원본 영상의 평균 사물 영역 일치도(mIoU)가 68.35%였던 AI 모델은 LSAA를 적용했을 때 0.41%를 기록했다. 같은 조건에서 기존 공격 기법을 적용했을 때의 2.04%보다 낮은 수치로, LSAA가 더 큰 인식 오류를 유도한 것으로 나타났다.

‘mIoU’는 AI가 예측한 사물 영역과 실제 정답 영역이 얼마나 일치하는지를 나타내는 지표다. 수치가 0%에 가까울수록 두 영역의 일치도가 낮다는 뜻으로, LSAA를 적용했을 때 AI의 사물 인식 결과가 실제 정답과 크게 달라졌음을 보여준다.

또한 혼동하기 쉬운 대상을 처음에 한 번만 선택했을 때 mIoU는 9.45%였지만, 영상 변화에 따라 대상을 계속 갱신했을 때는 0.41%까지 낮아졌다. 이는 AI의 인식 상태 변화에 맞춰 공격 대상을 갱신하는 방식이 인식 오류를 유도하는 데 효과적임을 뒷받침한다.



▲ LSAA 적용 전후 AI의 영상 인식 결과 비교. 도심 주행 영상에서 AI가 도로와 차량 등을 구분한 결과를 보여준다. 각 색상은 AI가 구분한 사물이나 배경의 종류를 나타내며, LSAA로 미세하게 변형한 영상을 입력하면 인식 결과가 크게 달라지는 것을 확인할 수 있다.

연구팀은 150종의 사물과 배경을 구분하는 데이터세트에서도 LSAA를 평가했다. 그 결과, 3종의 AI 모델 모두에서 mIoU가 0.6% 이하로 낮아졌다.

이는 사람이 사물 간 관계를 일일이 설정하지 않아도, 다양한 사물과 배경을 구분하는 AI 모델의 취약점을 효과적으로 찾아낼 수 있음을 보여준다.

김승준 교수는 “AI가 어떤 대상을 혼동하는지는 학습 과정에서 형성된 내부 정보에 이미 담겨 있다”며 “이번 연구는 이러한 정보를 활용해 시각 인식 AI의 취약점을 효과적으로 찾아낼 수 있음을 보여준 것으로, 향후 자율주행차와 로봇에 쓰이는 AI의 안전성과 신뢰성을 검증하는 데 기여할 것으로 기대한다”고 말했다.

GIST AI학과 김승준 교수가 지도하고 김광빈 박사과정생이 제1저자로 수행한 이번 연구는 과학기술정보통신부·한국연구재단과 국토교통부·국토교통과학기술진흥원의 지원을 받았다.

연구 결과는 2026년 9월 27일부터 10월 1일까지 미국 피츠버그에서 열린 국제학술대회 '2026 IEEE/RSJ 지능형 로봇 및 시스템 국제학술대회(IROS 2026)'에서 발표됐다.

논문 정보

○ 학회명, 저자 정보

- 학회명 : 2026 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2026), 2026년 9월 28일 발표 (미국 피츠버그, 9. 27.~10. 1. 개최)
- 논문명 : LSAA: Latent Space-Aware Adversarial Attack on Semantic Segmentation
- 저자 정보 : 김광빈(제1저자, GIST AI학과 박사과정생), 김승준(교신저자, GIST AI학과 교수)