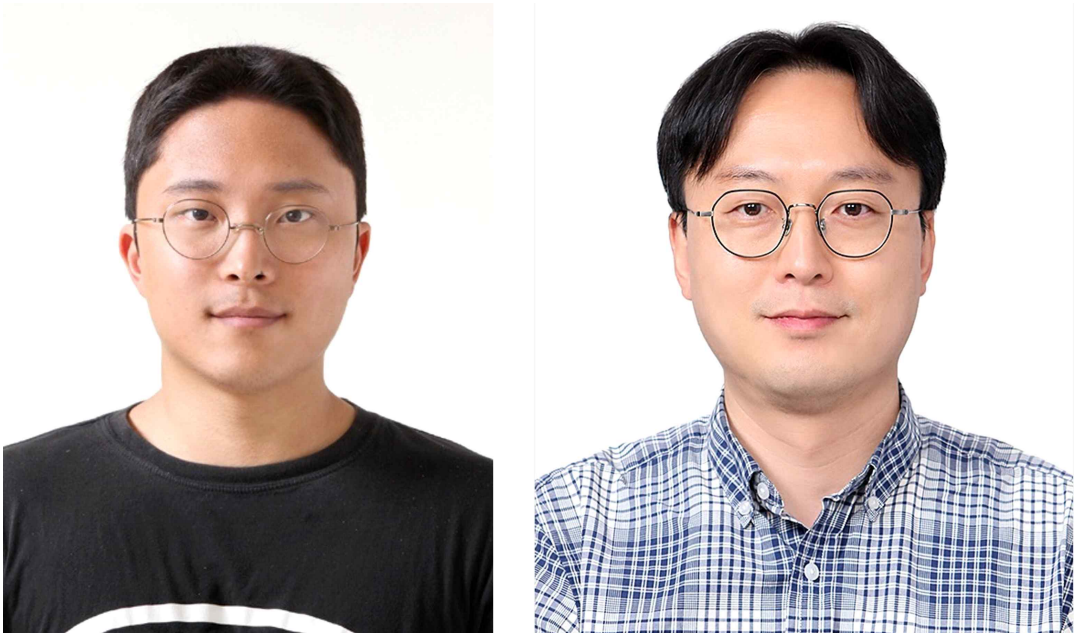
“보행자를 표지판으로? 자율주행차 맹점 찾아냈다”
GIST, 자율주행차 인식 시스템 취약점 드러내는 새 알고리즘 개발… 안전성 개선 활용 기대

- AI융합학과 김승준 교수 연구팀, 보행자·차선·신호등 등 객체 간 계층적 구조 반영한 새 공격 기법 제시… 인식률 95.3%→3.23% 급감, 기존보다 3.4배 높은 공격 성공률 확인
- 딥러닝 기반 시각 인식 모델이 미세한 데이터 변형(적대적 공격)에 취약한 점에 착안… 지능형 교통시스템(ITS), 스마트시티 보안, 로봇·국방 등 다양한 AI 분야 안전성 강화에 기여 기대
- 《IEEE Robotics and Automation Letters》 게재 및 ‘ICRA 2026’ 발표 예정



▲ (왼쪽부터) GIST AI융합학과 박사과정 김광빈 학생, 김승준 교수

광주과학기술원(GIST, 총장 임기철)은 **AI융합학과 김승준 교수 연구팀이 자율주행차에 활용되는 시각 인식 시스템의 보안 취약점을 진단할 수 있는 새로운 ‘적대적 공격*’ 알고리즘을 개발**했다고 밝혔다.

자율주행차는 도로 위 보행자, 차량, 차선, 신호등 등 다양한 객체를 실시간으로 정확하게 인식해야 안전하게 주행할 수 있다. 이를 가능하게 하는 핵심 기술이 바로 의미 분할 모델*이다.

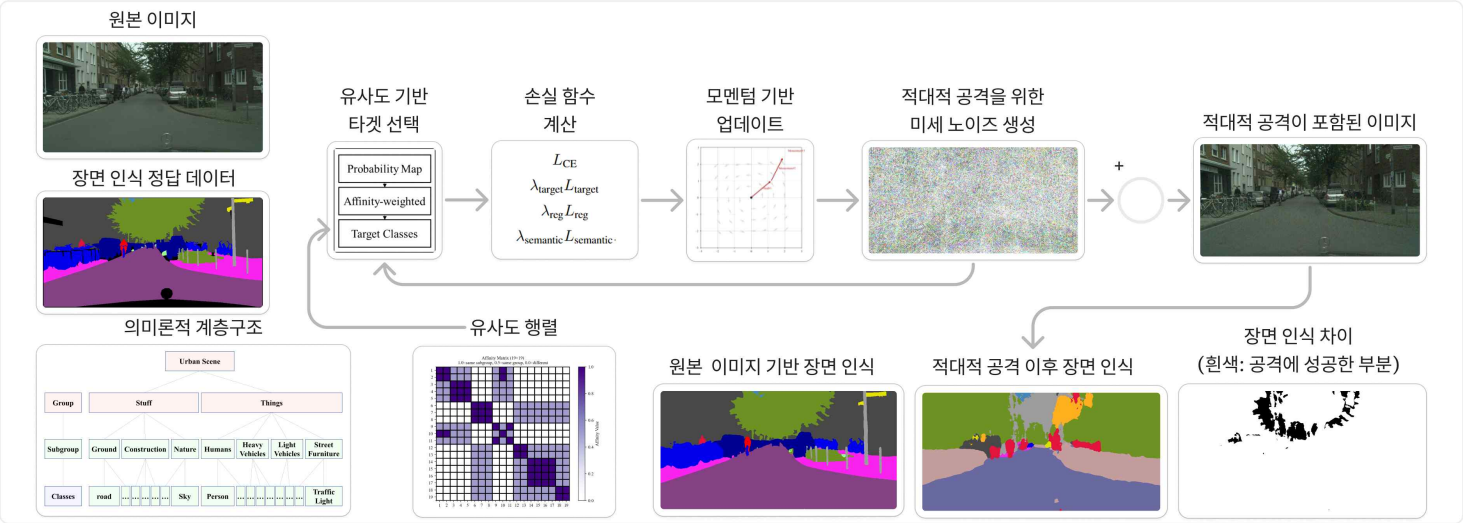
이번 연구는 **자율주행차가 의미 분할 모델에 가해질 수 있는 공격을 사전에 파악**해, **더욱 안전하고 신뢰성 높은 자율주행 시스템을 구축하는 데 기여**할 것으로 기대된다.

* **적대적 공격(Adversarial Attack)**: 인공지능 모델, 특히 딥러닝 기반 모델이 잘못된 판단을 내리도록 의도적으로 입력 데이터를 교묘히 조작하는 기법으로, 예를 들어 이미지의 픽셀 값을 인간이 눈치채기 힘들 정도로 미세하게 변형해도 모델은 전혀 다른 사물로 잘못 인식할 수 있어 보안 및 신뢰성 측면에서 큰 위협이 된다.

* **의미 분할(Semantic Segmentation) 모델:** 이미지 속의 모든 픽셀을 사물이나 배경 등 특정 의미 단위로 구분해주는 인공지능 기술로, 예를 들어 자율주행차가 도로 상황을 인식할 때 차선, 보행자, 차량, 신호등 등 각각의 객체를 픽셀 단위로 정확히 구분해 안전한 주행을 가능하게 한다.

최근 딥러닝 기반 모델은 적대적 공격이라 불리는 교묘한 데이터 변형에 취약한 것으로 알려졌다. 사람의 눈으로는 거의 구분되지 않는 미세한 변화에도 **딥러닝 기반 모델이 보행자를 도로 표지판으로 인식하거나 신호등을 무시하는 등의 오류가 발생할 수 있어 안전을 위협**한다.

이에 연구팀은 **의미 분할 모델의 구조적 특성에 주목해, 기존의 공격 방식보다 훨씬 정밀하게 취약점을 드러낼 수 있는 새로운 공격 기법을 고안**했다.



▲ **의미론적 계층 구조 기반 적대적 공격 기법 개요**. 의미론적 계층 구조를 활용한 타겟 선택을 통해 적대적 공격을 생성한다. 생성된 이미지는 육안으로는 원본과 구별되지 않지만, AI 모델에는 완전히 다른 장면으로 인식되어 높은 공격 성공률을 보인다.

자율주행차의 취약점 진단에 사용되는 기존의 적대적 공격이 단순히 무작위 오류를 유발하는데 그쳤다면, 이번 연구는 **보행자·차선·신호등처럼 서로 다른 객체들이 계층적으로 구분되는 구조를 반영해 실제 주행 상황에서 더 위협적인 오류 시나리오를 구현**했다.

실험 결과, 연구팀이 새롭게 개발한 기법은 주·야간 주행 환경에서 자율주행차의 시각 인식 모델을 **기존보다 3.4배 이상 교란시키는 등 월등히 높은 공격 성공률**을 보였다.

특히 자율주행 상황을 모사한 테스트에서는 **신호등·보행자·차선 등 핵심 객체의 인식률을 95.3%에서 3.23%로 급감시켜, 의미 분할 모델의 취약성을 실증**했다.

이번 연구는 단순히 강력한 공격 기법을 제시한 데 그치지 않고, 모델이 특히 취약한 의미 단위 계층을 규명했다는 점에서 의미가 크다. 이는 향후 **더 안전하고 신뢰성 높은 자율주행 인식 시스템을 설계하는 데 중요한 기초 자료로 활용**될 수 있다.

연구팀은 이번 성과가 자율주행 분야를 넘어, **지능형 교통 시스템(ITS), 스마트시티 보안, 로봇, 국방 등 다양한 인공지능(AI) 응용 분야의 안전성을 높이는 데 기여할 것으로 기대**하고 있다.

김승준 교수는 "이번 연구는 **자율주행차의 시각 인식 시스템이 가진 근본적인 취약점을 체계적으로 분석할 수 있는 도구를 제공**한다"며, "이론적으로 가장 취약한 시나리오를 **사전에 분석·진단하여 더욱 견고한 방어 전략을 수립**하고, **궁극적으로 자율주행차의 안전성과 신뢰성을 크게 향상**시킬 수 있을 것"이라고 밝혔다.

GIST AI융합학과 김승준 교수가 지도하고 김광빈 박사과정생이 제1저자로 수행한 이번 연구는 GIST-MIT 공동연구사업, 정보통신기획평가원 인공지능대학원지원사업, 한국연구재단 중견연구지원사업과 국토교통과학기술진흥원의 지원을 받았다.

연구 결과는 국제학술지 《IEEE 로보틱스 앤 오토메이션 레터스(Robotics and Automation Letters)》 2025년 8월호에 게재됐으며, 이에 따라 내년 6월 오스트리아 빈에서 개최될 예정인 세계 최고 권위의 로봇 학술대회 'ICRA(IEEE International Conference of Robotics and Automation)'에서 연구 결과를 발표할 예정이다.

한편 GIST는 이번 연구 성과가 학술적 의의와 함께 산업적 응용 가능성까지 고려한 것으로, 기술이전 관련 협의는 기술사업화센터(hgmoon@gist.ac.kr)를 통해 진행할 수 있다고 밝혔다.

논문의 주요 정보

1. 논문명, 저자정보

- 저널명 : IEEE Robotics and Automation Letters (Q1, IF: 5.3, 2025년 기준)에 게재되었으며, 해당 내용으로 IEEE International Conference of Robotics and Automation (ICRA), 로봇 분야 최고 학회에 2026년 6월에 발표할 예정
- 논문명 : Semantic Hierarchy-guided Adversarial Attack for Autonomous Driving
- 저자 정보 : 김광빈 (제1저자, GIST AI융합학과), 김승준 (교신저자, GIST AI융합학과)