과학기술정보통신부

IITP 정보통신기획평가원

KAIST

설명가능 인공지능 연구센터
NEWSLETTER

xaic

AI와 소통하는 방법

xaic
경기 성남시 분당구 성남대로331번길 8 킨스타워 18층 KAIST AI대학원 성남연구센터
xai.kaist.ac.kr

과학기술정보통신부 IITP 정보통신기획평가원 KAIST

AI가 설명을 한다고?

문제의 '답'만 알려주는 게 아니라
자세한 풀이 과정까지 설명해 주는 차세대 학습추론 인공지능
XAI(Explainable Artificial Intelligence)

About XAI

복잡한 인공지능 모델을 일반인들도 이해할 수 있게 해석을 제공해주는 시스템

AI와 함께할 가까운 미래
세상은 이미 변화하고 있다!

인공지능은 전 산업 분야에 걸쳐 혁신을 촉발하는
4차 지능 산업혁명의 핵심 기술
가까운 미래에 AI는 각 분야 전문가들의 의사결정을 지원하는
형태로 중요한 작업 분야에 투입됩니다.

xaic

"많은 사람들이 AI 때문에 세상이 바뀐다고 걱정합니다.
자칫 일자리를 뺏기거나 통제할 수 없을까봐 불안해합니다.
하지만 그런 막연한 두려움은 어쩌면 AI를 잘 모르기 때문일겁니다.
조금만 자세히 들여다보면 AI가 어떤 원리로 발전하고 있고,
어디까지가 한계인지를 알 수 있죠.
저희 연구센터가 개발하고 있는 XAI도 바로 그런 기술입니다.
AI가 사람들이 원치 않는 부분은 건드리지 않도록
우리가 원하는 방향으로 잘 컨트롤 할 수 있도록
투명하게 속을 들여다보는 기술,
AI가 좀 더 쉽고 편안하게 일반인에게 다가갈 수 있게 만드는
일종의 소통의 기술인 셈이죠.
AI를 너무 먼 미래나 다른 나라의 이야기로만 생각할 일이 아닙니다.
AI는 바로 지금,
우리가 주도적으로 바꿔 나가야 할 눈앞의 미래니까요."

Explainable Artificial Intelligence Center (XAIC)

설명가능 인공지능 연구센터를 소개합니다

국가혁신성장동력 프로젝트의 일환으로 2017년 7월
과학기술정보통신부와 정보통신기획평가원(IITP)의 후원으로 설립된 설명가능 인공지능 연구센터.
지난 2020년 1월 한국과학기술원에 새 터전을 마련하고
현재 KAIST AI대학원 성남연구센터에서 머신러닝 기술을 개발하고 있습니다.
추론·판단 근거를 논리적으로 설명할 수 있는 인공지능(XAI) 모델을 만드는 것이 목표입니다.

설명가능 인공지능 연구센터장
최재식 교수

AI가 실생활에 다양하게 응용되기 위해서는 의사결정과정이 투명하게 공개되고 검증되어야 합니다.
하지만 인공지능 모델이 복잡해질수록 그 추론 과정을 일일이 들여다보기는 쉽지 않죠.
때문에 인공 지능 내부의 작동원리와 판단의 근거를 분석하는 연구가 필요합니다.
설명가능 인공지능 모델은 AI가 산출해 낸 결과의 인과관계를 분석해 적절한 판단의 근거를 찾고,
해당 의사결정의 프로세스 및 결과를 사용자의 관점에서 설명해 주는 기술입니다.
쉽게 말해 사람이 AI가 내린 결정에 의문을 제기했을 때
왜, 어떻게 그런 결정을 내리게 됐는지를 정확하게 들여다볼 수 있는
현미경과 같은 도구를 만드는 것이 저희 연구의 목적입니다.

XAIC Partners

설명가능 인공지능 연구센터와 함께하는 협력기업사

고려대학교와 KAIST 기업 협력을 통해
경제적 가치와 사회적 가치를 함께 추구하며 발전해 나가는 설명가능 인공지능 연구센터.
설명가능 인공지능 연구센터의 비즈니스 파트너사와 함께 협업하고 동반성장하고 있습니다.

고려대학교 협력기업 신한은행

KAIST 협력기업 포스코 | KEB하나은행 | IBK기업은행 | 네이버
삼성전자(빅스비) | 분당서울대병원

XAIC INTERVIEW

설명가능 인공지능,
그리고 "표준화"의 필요성!

이재호 교수님(서울시립대학교)

서울시립대학교 인공지능학과 학과장과 사업단장을 맡고 계시고
설명가능 인공지능 연구센터에서 '설명가능한 인공지능 시스템 개발 지침'을 신규 표준화 과제를 진행하고 계시는
이재호 교수님에게 '설명가능 인공지능'과 표준화에 대해 들어보았습니다.

간단한 교수님 소개를 부탁드립니다.

현재 서울시립대학교 인공지능학과 학과장을 맡고 있는 이재호 교수입니다. 서울시립대학교 인공지능학과는 인공지능 분야 디지털혁신공유대학 사업에 참여하게 되어 사업단장도 맡고 있습니다.
80년대 후반에 인공지능을 분야 연구를 시작했으니 전통적 AI 연구자라고 할 수 있겠습니다.
현재는 지능 로봇과 자율 드론과 같은 지능형 시스템 개발 과제를 주로 수행하고 있습니다.

'설명가능한 인공지능 시스템 개발 지침'을 신규 표준화 과제(이하 'XAI 표준화')로 진행하고 계신다고 들었는데, 비전문가인 독자들도 이해할 수 있도록 '설명가능 인공지능'과 '표준화'에 대해 설명해 주실 수 있을까요?

인공지능 기술의 발달하며 활용 분야가 빠르게 확대됨에 따라 안전하게 사용하는 데 필수 요건인 인공지능 시스템의 신뢰성(Trustworthiness) 확보의 중요성이 강조되고 있습니다.
설명가능성은 기본적으로 인공지능 시스템이 "왜" 그런 결과를 만들어냈는지에 관한 질문에 답할 수 있는 능력으로 볼 수 있고, 좀 더 "표준적"으로는 "인공지능 시스템의 산출 결과에 영향을 미치는 중요한 요소를 사람이 이해할 수 있는 방식으로 표출하는 인공지능 시스템의 속성"으로 정의할 수 있습니다.
이는 모두 인공지능 시스템의 자율적 결정에 대한 신뢰성 확보에 필요한 요소입니다.

XAI 표준화 이전에도 '전자문서 호환성' 등 다른 표준화도 진행하셨던 걸로 아는데 표준화에 특별히 관심을 가지게 되신 계기가 있으신가요?

20여년 전 ISO/IEC JTC 1/SC 34에서 전자문서 관련 국제표준화 활동을 시작하였는데, 당시 온톨로지 분야 표준을 소개하는 글을 게재한 것이 계기가 되었습니다.
시장의 수요를 기반으로 하는 표준의 특성에 맞게 다른 기술 선진국들은 기업 중심으로 표준화 활동이 이루어지는 것에 비해. 특이하게도 우리나라는 교수의 참여가 많으며, 국내 기업에서 표준화 활동을 할 여건이 안되는 경우가 대부분인데 이를 대변하는 기술 외교관이라는 자부심으로, 현재까지도 교육정보, 사물인터넷, 그리고 가장 근래에는 인공지능 분야에서 활동하며 국제 표준을 만들어왔습니다.

XAI 표준화 과제는 어떻게 시작하게 되셨으며, 함께 참여중인 연구진이 있으신가요?

XAI 표준화 과제 ISO/IEC TS 6254는, 참여하고 있던 SC 42/WG 3 Trustworthiness에서 표준화 과제 개발 계획을 논의하는 AI Roadmap 그룹의 주요 개발 항목이었기에 개발 제안서를 제출하였고 개발 승인까지 상당한 노력이 필요하였습니다.
현재같이 참여하는 연구진은 없습니다만 제안 과정에서 KAIST 설명가능 인공지능연구센터의 최재식 센터장님과 연구진이 제공한 자료를 활용한 것이 큰 역할을 하였습니다. 앞으로도 센터에서 연구한 사례와 기술적 조치들이 표준 개발에 도움이 되리라 생각합니다.

표준화를 위해 필요한 작업(과정)들을 전반적으로 설명해 주실 수 있을까요?

표준화 기구별로 차이가 있습니다만 ISO의 경우 ISO/IEC 기술작업지침서에 따라서 일반적으로 제안부터 발행까지 예비, 제안, 준비, 위원회, 질의, 승인, 출판의 6단계로 구성됩니다.
이 중 제일 중요한 단계는 아마도 신규작업초안(NP)을 제안하는 예비 단계가 아닐까 생각합니다. 신규작업 초안의 제안자는 1차 작업초안을 제공하는데 다각적인 노력을 기울이거나 또는 최소한 이 작업초안에 대한 개요를 제공하고, 이를 참여국의 과반수 이상의 찬성과 5개국 이상의 적극적인 전문가 참여 표명이 필요합니다. 이 과정이야말로 기술적 요소 뿐만 아니라 외교적 협상 능력도 필요하기 때문에 가장 어려운 단계라고 생각합니다.

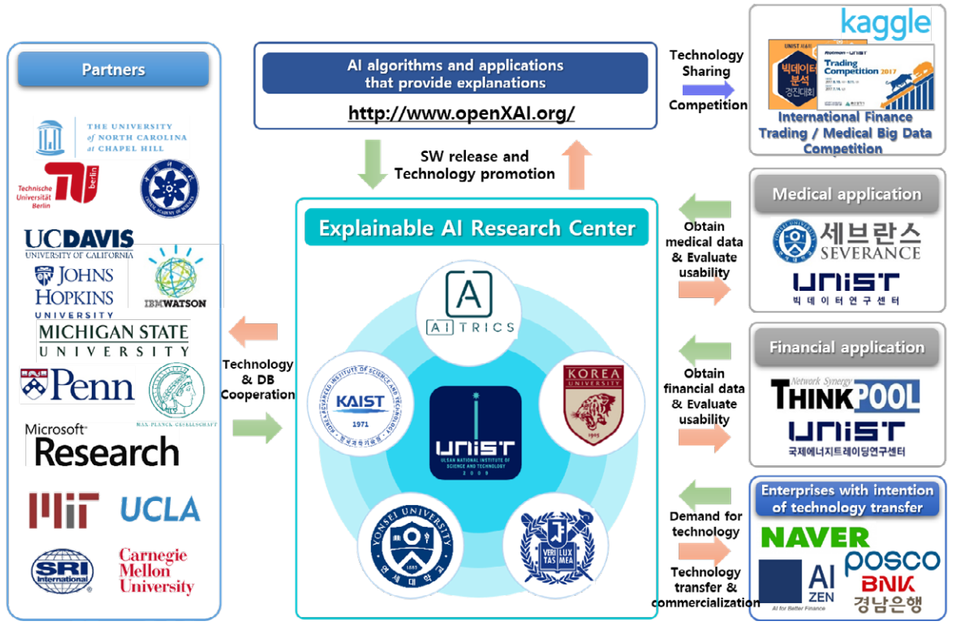
XAI 표준화가 제정되면 사회적으로 많이 활용될 수 있을 것 같습니다.
XAI 표준화는 왜 중요하고, 국제/한국 사회에 어떤 영향을 미칠까요?

국내외적으로 근래 "공정성"이 사회적으로 중요한 쟁점이 되고 있습니다. 인간을 대신한 인공지능의 자율적 결정과 행동에 점점 더 의존하게 되는 인공지능 사회가 현실이 되어 감에 따라 인공지능의 불공정한 의사결정으로 인한 사용자의 부당한 피해를 방지하거나 구제하고, 인공지능 시스템의 투명성을 전제로, 제조사나 서비스 제공자의 책임의 범위를 공정하게 규정하고 파악해야 합니다. 이러한 투명성의 핵심 요소인 설명가능성은 사용자의 신뢰를 증진하고, 인공지능 시스템의 안전한 활용도 촉진할 것입니다.
XAI 표준화는, 표준의 중요한 가치인 상호운용성(interoperability)을 기반으로 국내 외에 공통으로 적용할 수 있는 기술 기준과 공통 규정을 제공하여, 인공지능 시스템의 안전성을 확보하고, 통용되는 기준에 의한 인공지능 서비스 및 제품의 생산적인 활용에 이바지할 것입니다.

설명가능 인공지능이 아직 일반 인공지능처럼 범용적으로 활용되고 있지는 않은데, XAI 표준화가 특히 많이 활용될 것으로 예상하시는 분야가 있나요?

인공지능 시스템과 일반 소프트웨어 시스템의 차이점을 두드러지게 하는 특징은 인공지능 시스템의 "자율성" 이라 할 수 있습니다. 대표적인 예로 자율주행 자동차는 자율적 인식 및 판단으로 선택하여 수행한 행위를 검증하고 감시하여 책임과 보장의 범위를 정합니다.
그리고 이를 사용자가 신뢰할 수 있으려면 의사결정과 행위 선택의 이유에 대한 "설명"이 필수적으로 요구됩니다.
나아가 바이오헬스, 금융, 반도체, 지능형로봇, 에너지, 실감미디어 등의 첨단 분야를 포함한 사실상의 모든 산업 분야뿐만 아니라 생활 전반에서 인공지능의 확산과 같은 속도로 범용적으로 활용될 것으로 예상합니다.

설명가능 인공지능 연구센터를 만드는 사람들



KAIST를 주관기관으로 UNIST, 서울대학교, 고려대학교, AITRICS, 연세대학교, 연세의료원이 연구 참여

이런 연구가 진행되고 있어요

01

시각 및 언어적으로 설명이 가능한 인터페이스 X-CAD 개발

CT나 MRI 등의 의료 영상에서
전문적인 진단과 판단이 가능하도록 심층 학습모델 및 시각화 기술을 개발하고 있습니다.

02

**병렬 코퍼스가 존재하지 않는 새로운 도메인에서 번역모델을 학습하기 위해
각 언어의 새 도메인에 해당하는 단일 코퍼스를 이용한
비지도 학습 알고리즘을 개발하고 있습니다.**

03

연세의료원/AITRICS에서 개발한 'VitalCare'

환자들의 임상 데이터와 축적된 검사 수치 등을 통해
질병의 진행 과정, 위험도 등을 즉각적으로 분석하고 알려주는 모델을 개발하고 있습니다.

XAI센터의 주요 연구개발 기술

세부 연구과제 1

설명 가능한
인과관계
학습 추론 기술

해석 가능 인공지능, 인과관계 추론, 설명가능 인공지능 인터페이스 분야

딥러닝에서 입력한 각 데이터 간의 상관관계와 인과관계를 규명해 딥러닝 모델이 제시하는 예측 결과의 이유를 파악할 수 있게 하는 기술입니다.
또한, 인공지능 모델이 잘못 학습된 경우 사용자가 피드백을 주어 즉시 수정하게 하는 능동학습 기술을 적용해 상호작용을 통한 점진적 학습모델을 개발합니다.

최재식 교수, 김기응 교수, 신진우 교수, 이성환 교수, 뮬러 교수, 석흥일 교수

세부 연구과제 2

설명 가능한
원샷 / 제로샷
학습 모듈

소수샷 학습, 학습방법 학습, 상호작용 학습 분야

설명 가능한 원샷(One-shot) / 제로샷(Zero-shot) 학습 모듈은 적은 양의 데이터, 또는 학습 과정에서 경험하지 못한 새로운 데이터에 대해 클래스 예측이 가능한 학습 모듈을 의미합니다. 이는 데이터가 절대적으로 부족한 의료와 같은 분야에서 인공지능을 보다 제대로 활용하기 위해 필수적인 필수기술입니다.

한보형 교수, 황승원 교수, 오혜연 교수, 황성주 교수, 양은호 교수

세부 연구과제 3

산업분야에
실적용 가능한
인터페이스

설명가능 인공지능 및 소수샷 학습의 실세계 응용 분야

XAI센터에서는 인공지능의 의사결정에 대한 이유를 시각과 언어적 설명으로 제시하는 인터페이스를 개발합니다. 구체적으로 시계열 데이터를 설명하는 모델을 개발하고 그에 대한 설명을 자연어로 제시해 자동으로 보고서를 작성하는 기술, 심층 학습 특징의 시각적 해석이 가능한 입력 시각화 기술 등이 있습니다.
그 외 의료, 금융 등 XAI 과제의 최종 목표인 산업 분야에 실제 적용할 수 있도록 예측 가능한 모듈을 개발하고 실제 금융 거래에 참고할 수 있도록 근거자료를 자동으로 작성하는 기술을 개발하고 있습니다.

유진규 대표, 김광준 교수, 서병기 교수, 박현욱 교수, 노용만 교수

XAI센터의 대표연구

최재식 교수님 | 딥러닝을 기반으로 생성 모델의 생성 오류를 수정하는 인공지능 연구

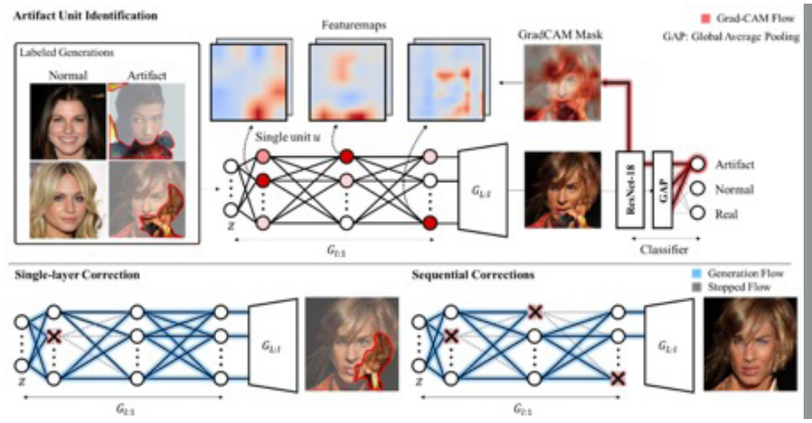


그림1. 개발 기술의 모식도.
(위) 결함이 존재하는 영역을 생성하는 단계
(아래) 검출된 내부 뉴런의 계층별 제거를 통한 생성 이미지내 결함 수리

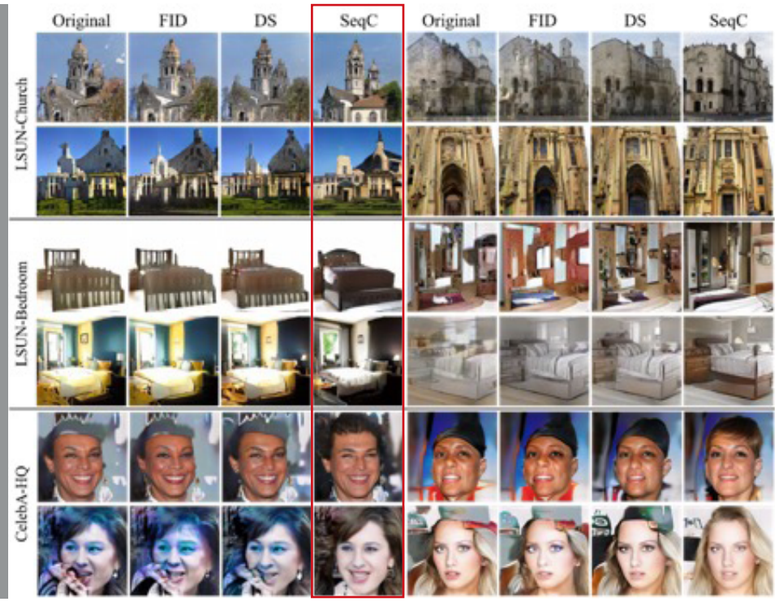
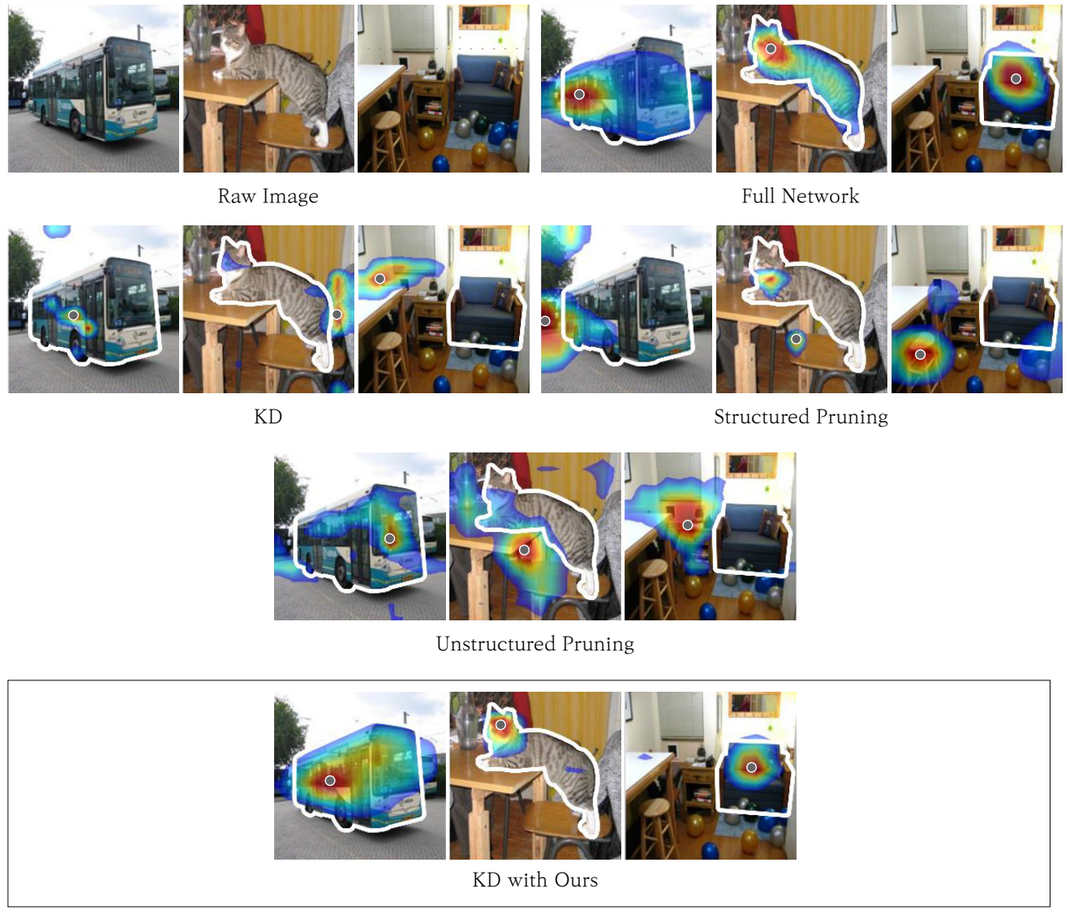


그림2. 여러가지 수리 기술에 대한 수리 성능 비교.
(FID) MIT에서 개발된 수리 기술
(DS) 개발 기술의 단일 계층 수리
(SeqC) 개발 기술의 계층별 수리

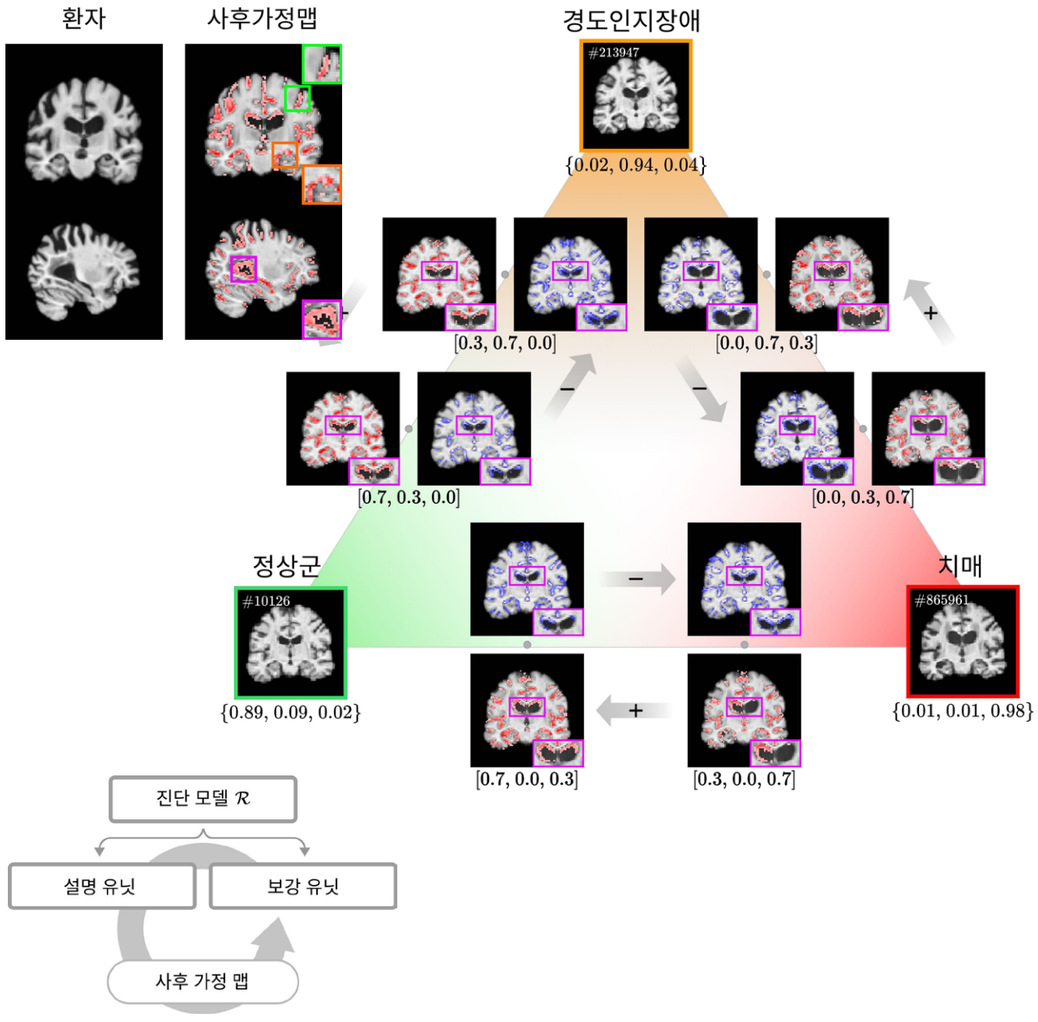
딥러닝을 기반으로 생성 모델 내부 뉴런이 생성 결과에 존재하는 물체[예) 나무, 선글라스]에 대한 정보를 표현하고 있는 기존 연구에서는 시각적 결함을 하나의 물체로 간주하고, 이를 유발하는 내부 뉴런을 검출하고 제거하여 결함을 수리합니다. 하지만 결함을 몇 개의 물체로 표현하는 것은 모델에서 발생할 수 있는 모든 결함을 포함할 수 없는 단점이 있어 'KAIST Breakthroughs'에 선정된 본 기술에서는 딥러닝 기반 분류기를 학습하여 시각적 결함에 대한 상위 수준(high-level) 특징을 이용하여 생성 모델 내부의 오류 유발 뉴런을 검출하고 제거하는 것을 목표로 합니다. 개발 기술은 (1) 정상 생성과 오류 생성을 구분할 수 있는 분류기를 학습하고, 설명가능 인공지능 기법을 적용하여, 생성 이미지의 오류 영역을 추출하는 과정과, (2) 오류 영역과 생성 모델 내부 뉴런 사이의 공유 영역(Intersection over Union)을 계산하여 오류 유발 뉴런 검출, (3) 검출된 뉴런의 계층별 제거를 통한 결함 수리로 구성되어 있어 개발된 생성 오류 수리 기술은 신경망 모델의 재학습을 요구하지 않으며 모델 구조에 대한 의존성이 적어, 다양한 딥러닝 기반 생성 신경망에 적용될 것으로 기대됩니다.

양은호 교수님 | 심층신경망 압축 시 발생하는 속성맵 손상으로부터 보호하는 학습 기술 연구 : 심층신경망 압축 시 신경망 매개변수 손실로 인한 설명가능성의 손상을 줄이는 학습 방법 개발



최근 발표되고 있는 심층신경망 모델들은 그 크기가 매우 거대하여 많은 컴퓨팅 장비를 요구하기 때문에 엣지 디바이스에서의 구동에 많은 제약을 받고 있어 신경망의 예측 성능을 최대한 보존하면서 신경망의 의 크기와 연산량을 줄이는 압축 알고리즘들이 사용되고 있습니다. 하지만 이러한 압축 방법들은 신경망의 예측 성능만을 보존하기 위해 설계되었기 때문에, 신경망을 압축할 시 해당 예측에 대한 설명은 손상의 위험에 노출되며 설명의 정확성을 보장할 수 없게 됩니다. 자료 그림에서 보여지는 바와 같이, 신경망 증류나 가지치기를 통해 압축된 신경망은 압축 전과 비교하였을 때 손상된 속성맵을 생성하는 것을 볼 수 있는데, 본 연구에서는 이러한 신경망 압축 손실 하에서의 설명력 손상을 줄이는 속성 매칭(Attribution Matching) 기술을 제안해 신경망 압축 과정 중 미세조정(fine-tuning) 단계에서 압축되는 신경망의 속성맵이 압축 전의 신경망(자신)을 따르도록 하는 전이 학습 방법을 통하여 압축 시 설명력의 손실을 최소화하였습니다. 개발된 기술은 다양한 압축 알고리즘 및 속성맵 알고리즘에 대하여 속성맵 일관성(integrity)을 보존함을 실험을 통해 확인하였고, 압축 신경망의 예측 성능에도 긍정적 영향을 미치는 것을 확인하였습니다. 이러한 방법을 통해 신경망의 설명력을 보존함으로써 압축된 신경망이 실제로 엣지 디바이스에 도입되었을 때에도 원본 신경망과 유사한 설명력을 보유할 수 있습니다.

석흥일 교수님 | 질병에 대한 사후 가정이 가능한 인공지능 연구

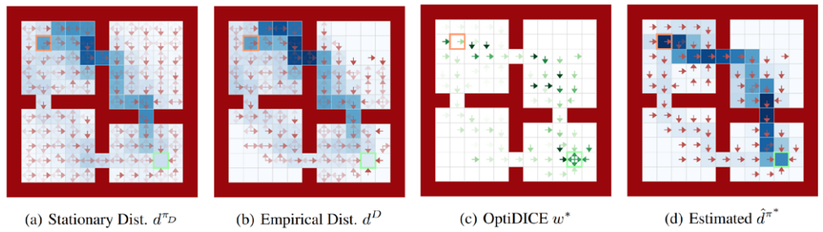


본 연구팀은 인간의 가장 고차원적인 사고 중 하나인 사후 가정 사고가 가능한 인공지능을 개발하였습니다. “사후 가정 사고(counterfactual reasoning)”는 이미 일어난 사건에 가상의 환경을 적용한 대안적 사건에 관한 생각하는 것입니다. 사후 가정 사고가 가능한 인공지능은 인간의 사고와 비슷한 기능을 가지기 때문에 기존의 설명가능한 인공지능보다 더욱 높은 차원의 설명성을 가지고 있습니다. 구체적으로, 본 연구팀은 질병에 대한 사후 가정 사고가 가능한 인공지능을 개발하였고, 이를 통해 좀 더 고차원적인 설명이 가능한 인공지능을 제시하였습니다.

본 연구는 사후 가정 사고를 활용해 기학습된 진단 모델의 성능을 향상 시킬 수 있고 다양한 구조의 심층 모델에 적용할 수 있는 일반화된 프레임워크를 제시하였습니다. 기존의 연구들은 진단 성능 향상 또는 설명성을 높이는 것을 추구하였지만, 본 연구는 진단 성능과 설명 가능성을 동시에 향상 시킬 수 있는 방법을 제안하고 있습니다.

참여연구실 대표연구 소개

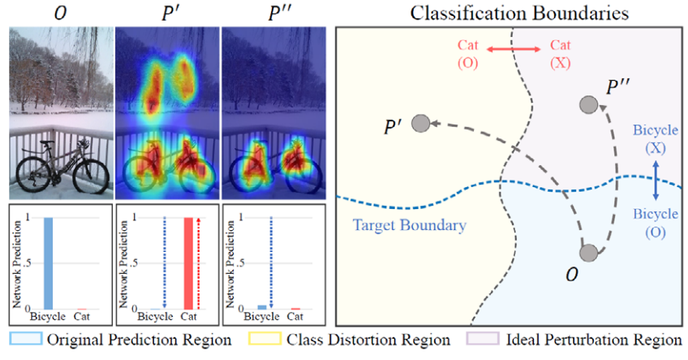
김기응 교수님 연구실



(a) Stationary Dist. d^{π_D} (b) Empirical Dist. d^D (c) OptiDICE w^* (d) Estimated $\hat{d}^{\pi^*}$

본 연구는 정적 분포 보정 추정(Stationary Distribution Correction Estimation)을 통한 오프라인 강화학습을 연구하고 있습니다.
환경과 상호작용이 없는 데이터 기반의 강화학습의 문제 해결을 위해 환경과의 상호작용 없이 사전에 수집된 데이터만을 통해 정책을 최적화하는 오프라인 강화학습 방법론을 고려하며 이는 에이전트(판단을 내리는 주체)가 환경과 직접 상호작용하는 것이 고비용/고위험인 다양한 현실문제에 유용합니다. 그러나 추가적인 데이터 수집이 제한된 상황에서 표준적인 과거에 배웠던 학습 내용과 현재 배우는 학습 내용이 다를 경우(Off-Policy) 데이터가 부족한 영역의 행동 가치가 과대추정되면서 정책 최적화에 실패하는 문제가 발생합니다. 본 연구에서는 행동가치 함수 최적화와 행동 정책 최적화, 이 2가지를 번갈아 수행하는 기존의 상태에 따라 행동하고 상태를 평가하는(Actor-Critic) 복잡한 방법론이 아닌, 최적 행동정책의 정적 분포를 직접적으로 최적화하는 새로운 배운 것만으로 수행하는(Offline) 강화학습 방법론을 제안하고 있습니다. 제안하는 방식은 근본적으로 가치 함수 과대 추정 문제를 겪지 않으며, 하나의 볼록 최소화 문제를 푸는 것으로 해결될 수 있어 안정적인 최적화가 가능한 장점이 있습니다.

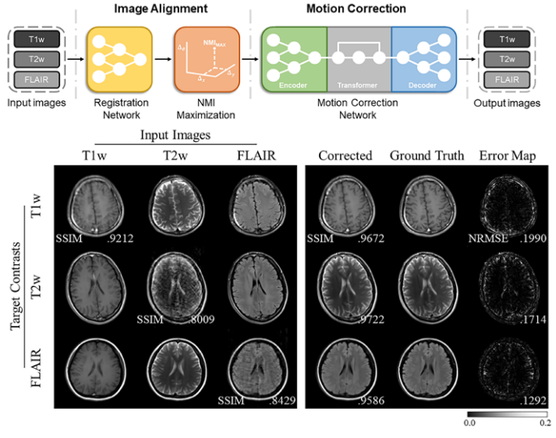
노용만 교수님 연구실



해당 연구에서는 블랙박스 딥러닝 모델의 의사 결정을 시각적으로 설명할 수 있는 변조 마스크 생성(Perturbation Mask Generation) 방법을 활용하고 있습니다. 변조 마스크 생성 방법은 학습된 모델의 관심 클래스 예측 점수를 최소화하는 마스크를 찾아 모델의 예측에 가장 많은 영향을 끼친 영역을 찾는 방법입니다. 기존 방법을 통해 변조 마스크를 생성하여 영상에 적용할 경우, 관심 외 클래스의 예측 점수가 왜곡되는 현상이 발생하는데, 본 연구는 해당 왜곡 현상을 모델 예측 점수 왜곡(Class Distortion) 이라 칭하고, 이 현상은 딥러닝 모델에 대한 시각적 설명의 강인함과 신뢰성을 하락 시키는 요인입니다. 본 연구는 (1) 마스크를 찾는 과정에서 관심 클래스의 예측 점수를 낮추며 관심 외 클래스의 예측 점수를 유지하는 새로운 마스크 최적화 및 생성 방법을 제안하며, (2) 관심 클래스의 영역을 가리는 정 방향 마스크 적용과 관심 클래스의 영역만을 남기는 역 방향 마스크 적용을 통한 교차 검증 방식(Cross-Checking) 의 마스크 최적화 방법을 제안합니다. 또한, 모델 예측 점수 왜곡(Class Distortion) 이라는 영상 변조 기반 시각적 설명 방법의 문제점을 제기함과 동시에 그 왜곡의 정도를 측정할 수 있는 새로운 평가지표(Class Distortion Score)를 제안합니다.

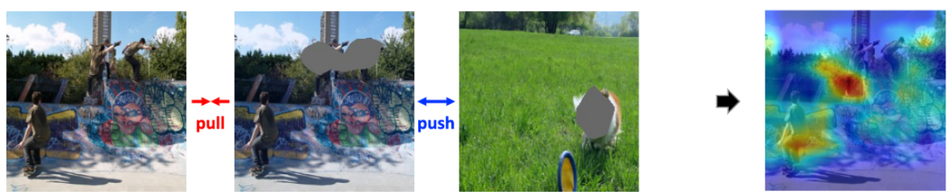
참여연구실
대표연구 소개

박현욱 교수님 연구실



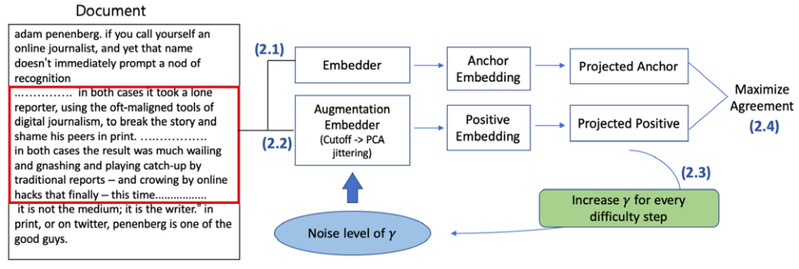
본 연구는 다중 대조도 MRI 뇌 영상에서 적은 데이터로 움직임 보정 기법의 문제 해결을 위해 뇌 MRI에서 환자의 움직임은 움직임 왜곡을 발생시킵니다. 기존의 방법은 주로 네이게이터 등 추가 신호를 획득하여 움직임 정보를 추정해낸 후, 추정한 움직임 정보를 바탕으로 보정하는 방식입니다. 최근 딥러닝 기술의 발전으로 뉴럴 네트워크를 통해 움직임 보정을 진행할 수 있는데, 기존의 단일 대조도 영상에서 진행하는 방법은 데이터 기반의 방식이지만, 데이터가 많지 않은 의료영상 분야에서는 부적합한 방법입니다. 본 연구는 뇌 MRI에서 다중 대조도 영상을 얻는 것에 착안하여 여러 영상을 입력 영상으로 서로를 참고하게 하여 데이터에 과한 의존을 피할 수 있게 합니다. 또한 왜곡이 없는 영상 데이터만 가지고 자기공명 물리학(MR Physics)에 의거한 움직임 시뮬레이션을 실행하여 적은 데이터로도 많은 데이터를 생성하여 과적합 문제를 해결하였습니다. 또한 합성곱 신경망(CNN)이 입력 영상의 오정렬에 민감하다는 것을 해결하기 위해 비지도학습 기반 정렬 네트워크를 이용하여 성능을 개선하였을 뿐만 아니라, 시뮬레이션을 통해 다양한 왜곡들을 학습시킨 결과 실제 자기공명 영상에도 적용하여 성능을 검증하였습니다.

신진우 교수님 연구실



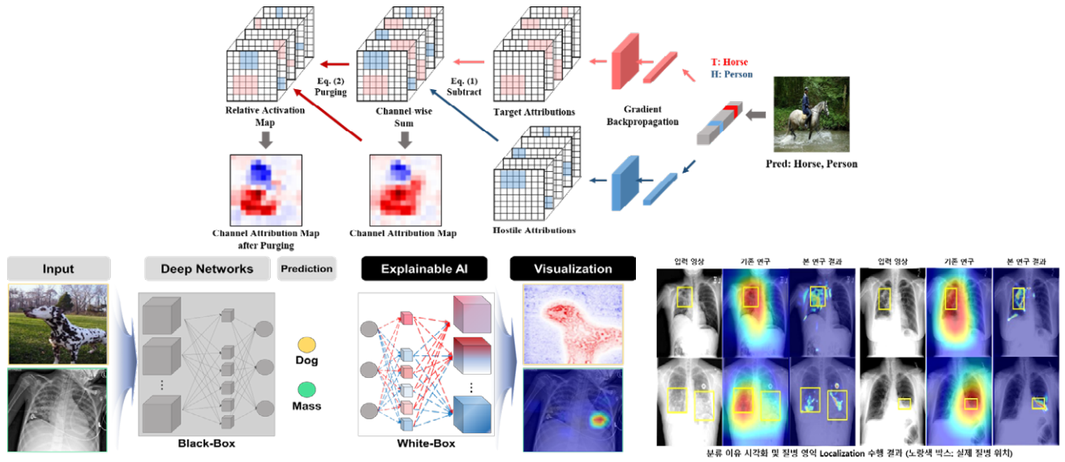
본 연구는 해석 가능한 대조 학습 기법 개발하기 위한 연구입니다. 분류 라벨 없이 학습한 표현 공간의 설명 가능성 증대시키기 위한 대조 학습 기반의 의사 결정 설명 기법을 통해, 분류 라벨 없이 학습한 표현 공간의 설명 가능성을 증대 시키는 연구로, 기존 입력의 특징을 시각화하는 기법(CAM)이 객체의 일부분을 놓치거나 여러 객체가 존재하는 경우 몇몇 객체를 놓치기도 하였으나 각 이미지에서 의사 결정의 이유가 될 수 있는 객체를 검출하는 CAM 기법을 제안하였습니다. 원래 이미지의 일부분을 가린 후, 원본 이미지와의 유사성은 높지만 배경만 있는 다른 이미지와의 유사성이 낮아지는 방향으로 대조 학습을 적용하여 의사 결정의 이유가 될 수 있는 영역을 검출하였습니다. 이와 같은 방법은 분류 라벨을 필요로 하지 않기 때문에, 대조 학습 기반으로 학습 된 표현 공간을 설명하기에 적합합니다. 분류 라벨을 필요로 하지 않는 점은 비지도 객체 검출에 응용되기 용이하며, 이미지 내에 여러 종류의 다른 객체가 존재하는 경우에도 성공적으로 객체를 검출할 수 있는 등 여러 적용 분야가 있을 것으로 기대됩니다.

오혜연 교수님 연구실



인간의 학습 과정을 모방한 대조 학습(Contrasitive Learning)과 커리큘럼 학습(Cirricculum Learning) 을 활용하여 효율적인 언어 모델 사전 훈련법을 제안하였습니다. (1) 새로운 데이터 증강 기법과 대조 학습을 적용하여 텍스트 표현을 학습하도록 하였으며, 텍스트 임베딩의 일부분(Cut-Off)과 프로젝션(Projection)을 통해 기존의 데이터와 유사한 증강 데이터 쌍을 생성하였습니다. 대조 학습을 적용하여 증강 데이터 쌍은 유사성을 높게, 그 외의 텍스트 쌍은 유사성을 낮게 평가하도록 모델을 학습하여 텍스트 특징 표현을 추출하였습니다. (2) 이 과정에서 커리큘럼 학습을 적용하였으며, 커리큘럼 학습은 난이도가 쉬운 단계부터 어려운 단계까지 단계적으로 학습을 시키는 방법으로, 데이터 어그멘테이션의 변형 정도를 높여가며 언어 모델을 훈련시킵니다. (3) 기존의 사전학습 방법보다 적은 메모리로 학습이 가능하며 높은 성능을 달성하였으며, 다양한 데이터에 적용 가능한 학습 효율이 높은 사전 학습을 제시하고 있습니다.

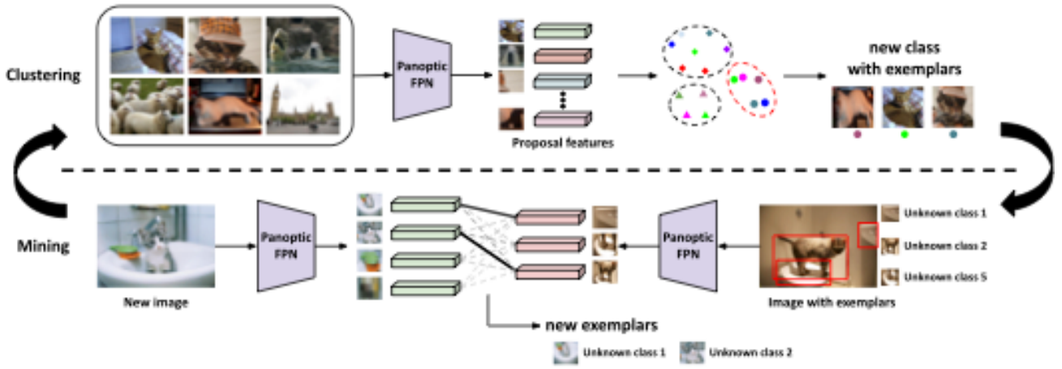
이성환 교수님 연구실



딥러닝의 의사결정 이유를 시각화하는 인공지능은 입력단에서 의사결정의 이유를 사람이 직관적으로 이해를 할 수 있게 시각화를 하는 기술입니다. 기존의 시각화 알고리즘은 정확한 객체의 위치를 시각화 하는 것이 어려웠으나, 본 연구에서는 좀 더 정밀하게 의사결정에 관련이 없는 부분과 연관성이 높은 부분을 명확하게 구분하여 시각화 할 수 있는 기술입니다. 더 나아가 다양한 객체가 존재하는 영상에서도 객체의 종류별로 시각화를 할 수 있는 기술도 제안하였습니다. 이는 기존의 시각화 결과에서 발생하는 객체간 설명이 힘든 한계를 극복하고 특정 객체에 중점적으로 시각화를 하는 기술입니다. VOC, COCO 등 다양한 데이터셋에서 실험이 진행하였고 화소 단위의 이미지 분류(Semantic Segmentation)에도 보조데이터로 활용될 수 있음을 보여 주었습니다. 알고리즘에 대해 간단히 설명하면, 해당 클래스에 대한 기울기(Gradient)와 분석하여 해당 객체에 대한 적대적인 관련성을 분석하여 상대적으로 영향력이 높은 뉴런에 높은 값을 부여 후 이 값을 시각적으로 보여주게 됩니다. 본 기술의 결과는 명확한 객체 간 관련성을 시각화하며 관련이 없는 배경 또한 분리가 가능하며 비모수적 Segmentation 연구에 활용 가능합니다.

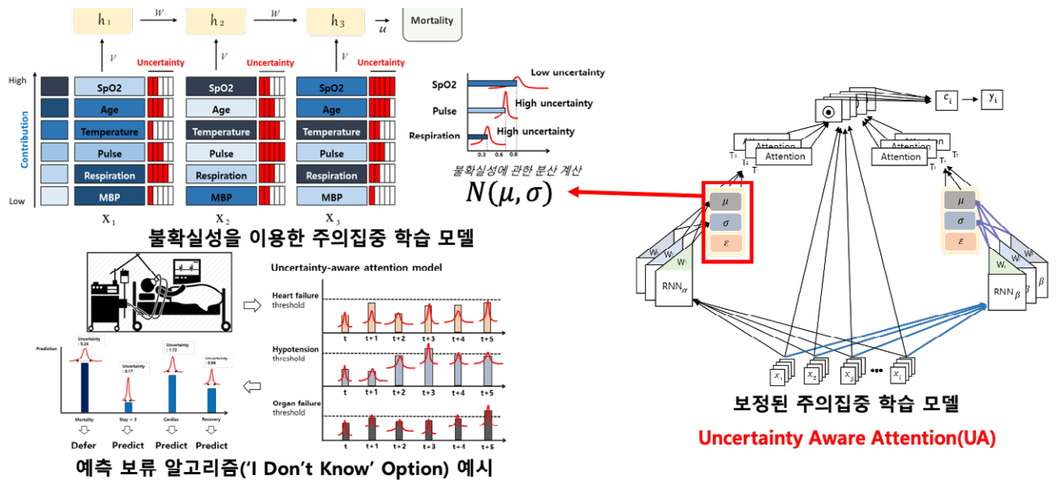
참여연구실
대표연구 소개

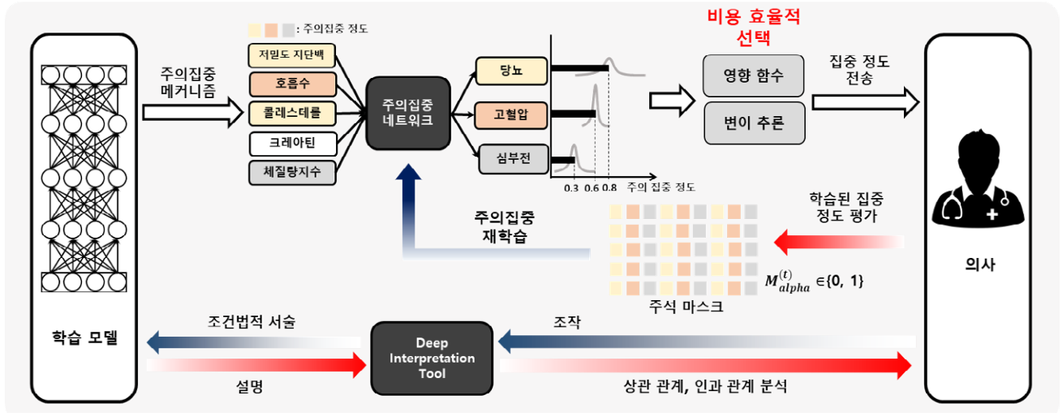
한보형 교수님 연구실



표본 기반 학습 동안 인지되지 않은 클래스에 대해서도 물체의 종류와 그 경계를 탐지하여 분할하는 문제 네트워크(Exemplar-Based Open-Set Panoptic Segmentation Network)를 위한 연구입니다. 오픈 셋 파노라마적 영상 분할을 위해 본 연구는 사전 학습된 클래스에 대해서만 영상 분할하는 기존 문제를 확장하여 학습동안 인지되지 않은 미지의 클래스에 대해서도 파노라마적 영상 분할하는 문제를 해결합니다. 본 연구팀은 해당 문제 상황의 실질적인 문제점들을 분석하고 기존 데이터셋을 변형한 새로운 벤치마크를 제시하여, 표본 이론에 기반한 새로운 오픈 셋 파노라마적 영상 분할 네트워크를 제안하고 있습니다. 제안된 네트워크는 군집 알고리즘을 통해 확인된 표본들을 기반으로 새로운 클래스를 발견하고 해당 표본들의 유사 정답으로써 활용됩니다. 각 클래스 데이터 크기는 관련 데이터 간 유사도를 기반으로한 표본의 지속적인 발견으로 점점 증가됩니다. 개발된 오픈 셋 파노라마적 영상 분할 기술은 기존의 탑다운 방식 영상 분할 모델들에 적용될 수 있는 포괄적인 기술이고 기존 연구보다 어렵고 현실적인 문제상황을 다루는 실용적인 기술입니다.

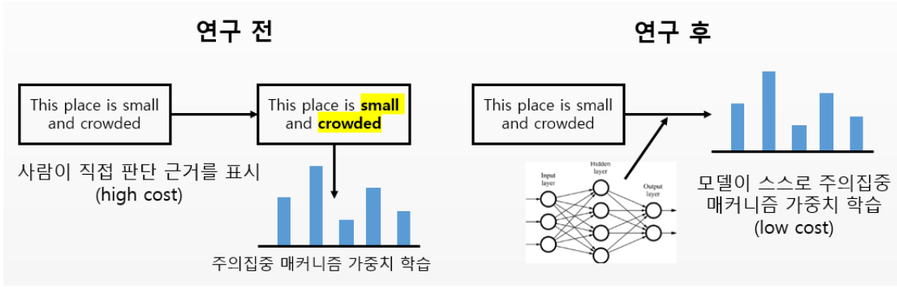
황성주 교수님 연구실





다양한 형태의 데이터가 실시간으로 들어오는 학습 환경에서 정확한 학습과 신뢰도 높은 판단 결과 해석을 해결하기 위해 실시간으로 들어오는 이미지, 텍스트, 시계열 등의 데이터를 효과적으로 통합하여 학습하는 알고리즘을 고안하여 의료, 금융 등 실제 AI 모델 활용 환경에서의 모델 사용성을 극대화한 연구입니다. 실제 산업 환경에서는 수치 형태의 시계열 데이터 외에도 이미지, 텍스트 등 다양한 형태의 데이터가 존재하는데 본 여러 형태의 데이터를 동시에 다루는(Multi-Modal) 설명가능 모델은 실시간으로 다양한 형태로 들어오는 데이터를 효과적으로 통합하고 정확한 판단이 가능할 뿐만 아니라 해당 설명가능 알고리즘을 바탕으로 멀티 모달 데이터를 통합한 모델 내부의 복합적인 해석이 가능합니다. 멀티 모달 설명가능 모델은 1) 불규칙적으로 들어오는 멀티모달 시계열 데이터를 통합하는 과정과, 2) 판단 결과를 해석하는 설명가능 기법을 활용하여 통합적인 모델 판단 설명가능성을 제공하는 과정을 포함합니다. 개발된 기술은 불규칙적으로 입력되는 시계열 데이터를 규칙적인 시간 간격으로 샘플링(Time-Interval Sampling)하는 데이터 정제 단계를 필요로 하지 않아 실제 산업 환경에서 비용 효율적으로 활용될 것으로 기대됩니다.

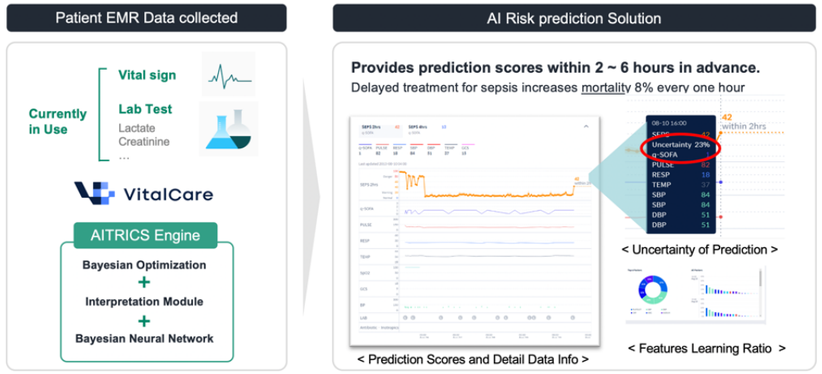
황승원 교수님 연구실



기존 모델들은 주의집중 매커니즘 학습 시 미리 생성한 정답 분포를 따르도록 학습을 진행하는데, 이 때 모든 데이터마다 판단 근거가 되는 단어를 사람이 직접 표시해야 하기 때문에 시간과 비용 부담이 크며, 문제나 상황에 따라 같은 데이터의 같은 단어라도 다른 중요도를 가질 수 있는데 이를 반영하기 어렵습니다. 본 연구에서는 사람이 생성한 지도 학습 데이터가 없거나 부족하더라도 모델이 스스로 가중치를 학습하는 방법을 제안합니다. 모델 내부에 주의집중 매커니즘에 사용되는 가중치를 예상하는 모듈을 추가하여, 학습 과정에서 해당 가중치를 스스로 생성하도록 하는데, 장점은 (1) 기존 모델이나 규칙에 기반한 데이터 레이블 생성 방법들에 비해 데이터 간의 인과관계를 잘 잡아내어 모델의 해석이 더 쉬워집니다. 따라서, 모델이 어떤 기준으로 데이터를 분류하였는지, 판단에 문제가 있었다면 어떤 점을 개선해야 하는지 결정하는 데 도움이 되며, (2) 데이터 변조에도 더 견고해지기 때문에 소수샷 환경에서도 학습이 가능합니다. 실제 환경에서는 충분한 학습 데이터가 주어지지 않는 경우가 많기 때문에 더 유리하다고 볼 수 있습니다.

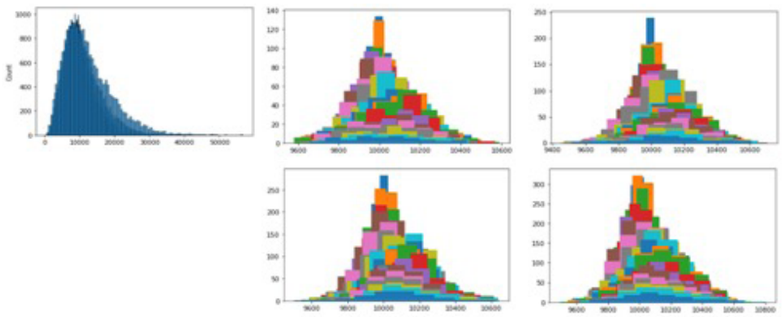
참여연구실
대표연구 소개

AITRICS



시계열 예측에서 입력, 예측의 불확실성을 고려하여 예측하는 모델·의료 문제 등, 안전이 중요한 문제에서 입력 노이즈, 인과관계의 애매함, 학습데이터의 부족, 미래예측에서의 근본적 불확실성 등을 정확히 모델링하여 예측의 신뢰도를 제공하는것은 중요한 문제입니다. 또한 모델에서 출력되는 예측 신뢰도 자체가 실제로 사람의 해석과 일치해야 합니다. 예로 모델이 0.5의 신뢰도(Confidence)를 가진다고 하면 실제 예측의 정확도도 50%가 되어야 하지만 많은 딥러닝 기반 예측모델들은 과신뢰화(Overconfident)되어, 실제 정확도가 낮은 경우도 높은 신뢰도를 출력하며, 이는 잘못된 예측으로 이어집니다. 따라서 다양한 소스에서의 불확실성을 고려하고, 신뢰도를 잘 보정하는것이 중요합니다. 이에 입력, 출력, 모델, 미래예측 등 모든 불확실성의 모델링을 통해 예측의 신뢰도를 향상시키고, 신뢰도 기반으로 예측이 힘든 경우 의료진에게 넘기도록 만들어 잘못 예측하는 경우를 최소화 하는 기술을 개발하며, 이는 의료 이외에 시계열 데이터 기반 다양한 의사결정의 신뢰성 제고에 기여할 것으로 기대됩니다.

UNIST

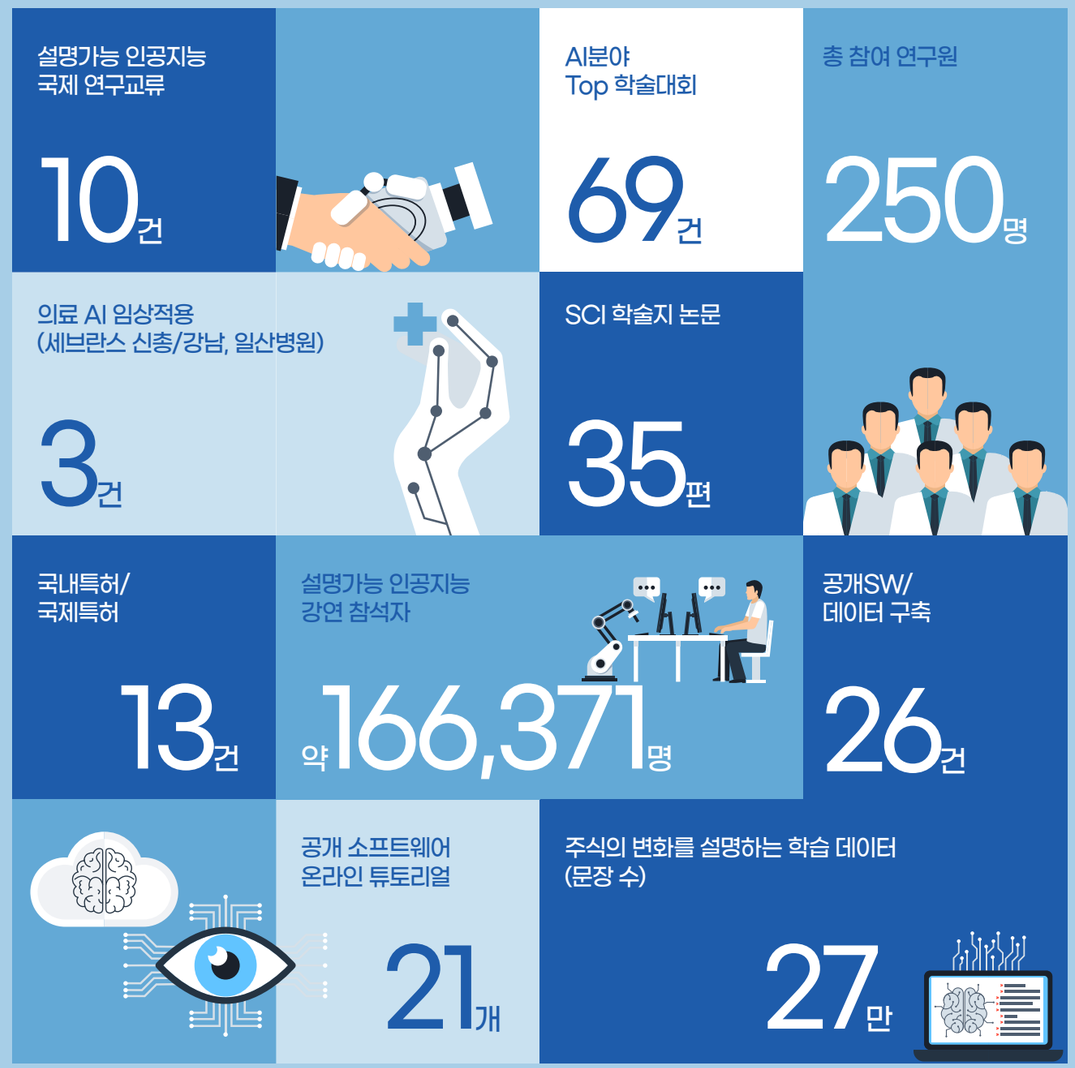


생성적 적대 신경망(GAN, Generative Adversarial Network)의 활용을 통한 정교한 VaR(Value-at-Risk) 리스크 측정 방법론 개발을 연구합니다.
정교한 리스크 측정 방법론 개발을 위해 금융시장에서 가장 널리 사용되는 리스크 측정 방법 중 하나인 VaR(Value-at-Risk) 모형은 금융 자산에 대한 포트폴리오의 위험의 정도를 측정하는 방법으로 해당 포트폴리오로 인한 가능한 분포를 구하여, 극단적인 경우를 리스크 척도로 표현한 것입니다. VaR 모형은 크게 역사적 VaR(Historical VaR) 모형과 모수적 VaR(Parametric VaR) 모형, 2가지로 나눌 수 있는데, 모수적 VaR 모형이 보다 간편하게 적용할 수 있는 방법이지만, 복잡한 포트폴리오 내 자산 간 상관관계를 추정하는 것이 어렵다는 단점이 있어, 역사적 VaR 모형이 더 실용적으로 활용되고 있습니다. 역사적 VaR 모형은 과거 실제 발생했던 시장의 움직임을 현재 똑같이 발생했다고 가정하여, 그 경우 가능한 손실 분포를 그리는 것인데, 과거 데이터가 한정적이라는 단점이 있습니다. 이에 생산적 적대 신경망 방법을 활용하여 실제 발생 가능한 시장의 움직임을 추가적으로 더 많이 생성하여 역사적 VaR 모형에 추가로 적용하면, 역사적 VaR 모형의 단점을 완화시켜 더욱 정교한 리스크 척도를 개발할 수 있을 것으로 기대됩니다.

세계가 주목하는
설명가능 인공지능 연구센터

한국 AI의 눈부신 선전

'설명가능 인공지능 연구센터'가 투자받은 연구 예산은 5년간 150억 수준이며 이는 이 분야 연구에 가장 적극적인 미국 방위고등연구계획국(DARPA)이 지원받는 예산(약 800억원)과 5배 정도 차이가 날 정도로 부족한 실정입니다.
하지만 열악한 환경 속에서도 XAIC는 세계적인 성과를 내놓고 있습니다.
성과와 숫자로 보는 지난 3년간 쏟아부은 땀과 열정의 결과입니다.



AI분야 Top-tier 논문 성과

No.	발생년도	학술지명	주저자	공동저자	논문명
1	2021	CVPR	황재동	오승욱, 이준영, 한보형	Exemplar-Based Open-Set Panoptic Segmentation Network
2	2021	CVPR	이승민	김동완, 한보형	CoSMo: Content-Style Modulation for Image Retrieval with Text Feedback
3	2021	CVPR	Ali Tousi, 정해동	최환일, 최재식	Automatic Correction of Internal Units in Generative Neural Network
4	2021	ICLR	이병준	이종민, 김기응	Representation Balancing Offline Model-Based Reinforcement Learning
5	2021	ICLR	장영수	서석인, 이종민, 김기응	Monte-Carlo Planning and Learning with Language Action Value Estimates
6	2021	ICLR	윤든솔, 홍성훈	이병준, 김기응	Winning the L2RPN Challenge: Power Grid Management via Semi-Markov Afterstate Actor-Critic
7	2021	ICLR	이종민, 전원석	이병준, Joelle Pineau, 김기응	A Roadmap to Never-Ending RL Workshop at ICLR 2021. OptiDICE: Offline Policy Optimization via Stationary Distribution Correction Estimation
8	2021	ICLR	윤태림	신수민, 황성주, 양은호	FEDMIX: Approximation of Mixup Under Mean Augmented Federated Learning
9	2021	ICLR	박세준	윤철희, 이재호, 신진우	Minimum Width for Universal Approximation
10	2021	ICML	이종민, 전원석	이병준, Joelle Pineau, 김기응	OptiDICE: Offline Policy Optimization via Stationary Distribution Correction Estimation
11	2021	ICML	유보선	이지우, 주장훈, 정세준, 김소연, 최재식	Conditional Temporal Neural Process with Covariance Loss
12	2021	AAAI	Anh Tong	최재식	Characterizing Deep Gaussian Processes via Nonlinear Recurrence Systems,
13	2021	AAAI	Anh Tong	Toan Tran, Hung Bui, 최재식	Learning Compositional Sparse Gaussian Processes with a Shrinkage Prior,
14	2021	AAAI	남우정	최재식, 이성환	Interpreting Deep Neural Networks with Relative Sectional Propagation by Analyzing Comparative Gradients and Hostile Activations
15	2021	EMNLP	예성현	김지선, 오혜연	Efficient Contrastive Learning via Novel Data Augmentation and Curriculum Learning
16	2021	EMNLP	조요한	유하늘, 오혜연	Knowledge-Enhanced Evidence Retrieval for Counterargument Generation
17	2021	EMNLP	박건도	김경만, 양은호	Distilling Linguistic Context for Language Model Compression
18	2021	ICCV	박재유	강민수, 한보형	Class-Incremental Continual Learning for Action Recogntion in Videos
19	2021	ICCV	이정현, 윤지훈	황성주, 양은호	Cluster-Promoting Quantization with Bit-Drop for Minimizing Network Quantization Loss
20	2021	IEEE CVPR	Ali Tousi	최재식, 정해동, 한지연, 최환일	Automatic Correction of Internal Units in Generative Neural Networks
21	2021	ICML	유보선	이지우, 주장훈, 정세준, 김소연	Learning Basis Functions of Deep Predictive Neural Networks with Global Spatio-Temporal Covariance Loss
22	2021	KDD	조소희, 장원준	최재식	Interpreting Deep Temporal Neural Networks by Selective Visualization of Internally Activated Nodes

No.	발생년도	학술지명	주저자	공동저자	논문명
23	2021	COLT	박세준	이재호, 윤철희, 신진우	Provable Memorization via Deep Neural Networks using Sub-linear Parameters
24	2021	IROS	김성운	최재식	Explaining the Decisions of Deep Policy Networks for Robotic Manipulations
25	2021	ICLR	박세준	윤철희, 이재호, 신진우	Minimum Width for Universal Approximation
26	2021	NeurIPS	황형주	김건형, 홍승훈, 김기응	Multi-View Representation Learning via Total Correlation Objective
27	2020	AAAI	장영수	-	Bayes-Adaptive Monte-Carlo Planning and Learning for Goal-Oriented Dialogues
28	2020	AAAI	이병준	-	Residual Neural Processes
29	2020	AAAI	이종민	-	Monte-Carlo Tree Search in Continuous Action Spaces with Value Gradients
30	2020	AAAI	강민수	문종환, 한보형	Towards Oracle Knowledge Distillation with Neural Architecture Search
31	2020	AAAI	Ruotian Luo	Ning Zhang, 한보형, Linjie Yang	Context-Aware Zero-Shot Recognition
32	2020	AAAI	전기영	정해동	An Efficient Explorative Sampling Considering the Generative Boundaries of Deep Generative Neural Networks
33	2020	CVPR	hong joo lee	-	Structure Boundary Preserving Segmentation for Medical Image with Ambiguous Boundary
34	2020	ICML	김건형	-	Variational Inference for Sequential Data with Future Likelihood Estimates
35	2020	ICML	이병준	이종민	Batch Reinforcement Learning with Hyperparameter Gradients
36	2020	ICML	안성수	-	Learning What to Defer for Maximum Independent Sets
37	2020	ICML	강민수	한보형	Operation-Aware Soft Channel Pruning using Differentiable Masks
38	2020	ICML	황도영	-	A benchmark study on reliable molecular supervised learning via Bayesian learning
39	2020	ECCV	최진영	한보형	Task-Aware Quantization Network for JPEG Image Compression
40	2020	ECCV	서성욱	서유민, 김동완, 김지호, 한종우, 한보형	Learning to Optimize Domain Specific Normalization for Domain Generalization
41	2020	NeurIPS	탁지훈	-	CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances
42	2020	NeurIPS	황형주	-	Variational Interaction Information Maximization for Cross-domain Disentanglement
43	2020	NeurIPS	이종민	-	Reinforcement Learning for Control with Multiple Frequencies
44	2020	NeurIPS	김서현	서유민, 김동완, 김지호, 한종우, 한보형	Rotation-Invariant Local-to-Global Representation Learning for 3D Point Cloud

2021년도 저널 논문

No.	발생년도	학술지명	주저자	공동저자	논문명
1	2021	Pattern Recognition	김건욱, 정홍규	이성환	Spatial Reasoning for Few-Shot Object Detection
2	2021	Neural Networks	서진우, 정홍규	이성환	Self-Augmentation: Generalizing Deep Networks to Unseen Classes for Few-Shot Learning
3	2021	Neural Networks	김정준, 이동규	Jialin Wu, 정홍규, 이성환	Visual Question Answering based on Local-Scene-Aware Referring Expression Generation
4	2021	Internet Research	Daejin Kim	Hyoung-Goo Kang, Kyounghun Bae, Seongmin Jeon	An artificial intelligence-enabled industry classification and its interpretation
5	2021	IEEE Transactions on Neural Networks and Learning Systems	윤지석	노명철, 석흥일	A Plug-in Method for Representation Factorization in Connectionist Models
6	2021	IEEE Transactions on Circuits and Systems for Video Technology	박민호	김학구, 이상민, 노용만	Robust Video Frame Interpolation With Exceptional Motion Map
7	2021	IEEE Transactions on Neural Networks and Learning Systems	Qin Xie	Peng Zhang, 유보선, 최재식	Semisupervised Training of Deep Generative Models for High-Dimensional Anomaly Detection
8	2021	Analytical Chemistry	김현빈, 이해님	최재식, 성지혜	Spatiotemporal Characterization of GPCR Activity and Function during Endosomal Trafficking Pathway
9	2021	IEEE Transactions on Circuits and Systems for Video Technology	이상민	김성엽, 김학구, 노용만	Assessing Individual VR Sickness through Deep Feature Fusion of VR Video and Physiological Response
10	2021	JMIR Medical Informatics	성민동, 한상철, 한창훈	이정모, 이자영, 유진규, 김영삼, 정경수	Event Prediction Model Considering Time and Input Error Using Electronic Medical Records in the Intensive Care Unit: A Retrospective Study
11	2021	Magnetic Resonance in Medicine	이원일	김병재, 박현욱	Quantification of intravoxel incoherent motion with optimized b-values using deep neural network
12	2021	Magnetic Resonance in Medicine	이종연	김병재, 박현욱	MC2-Net: motion correction network for multi-contrast brain MRI

기타 성과

대회 수상

No.	발생년도	구분	수행기관	주요내용
1	2021	외부 수상	KAIST	제33회 포스코 기술 컨퍼런스에서 최재식 교수 POSCO O&C(Open&Collaboration)상 수상

언론에 비친 설명가능 인공지능연구센터

[AIF2021] "AI에 차별화전략 중요… 제조에 설명가능 인공지능 역할 많아"

[AIF2021]"AI에 차별화전략 중요…제조에 설명가능 인공지능 역할 많아"

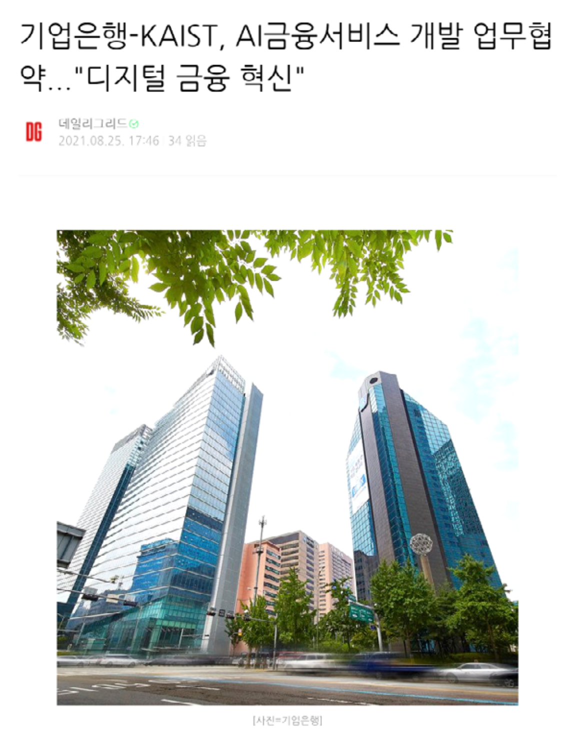
최재식 KAIST 센터장, AI기술의 제조 적용 효과 최대 92조원 예측

[AIF2021-인터뷰] "AI, 인간에게 판단 근거까지 제시해야"

[AIF2021-④/인터뷰] "AI, 인간에게 판단 근거까지 제시해야"

최재식 카이스트 AI대학원 성남연구센터 센터장 겸 교수

기업은행-KAIST, AI금융서비스 개발 업무협약…"디지털 금융 혁신"


기업은행-KAIST, AI금융서비스 개발 업무협약…"디지털 금융 혁신"

데일리그리드
2021.08.25. 17:46 34 읽음

[사진=기업은행]

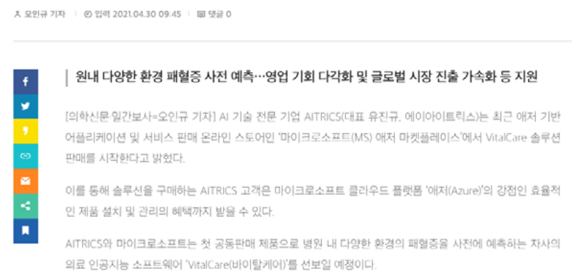
AITRICS, 서울아산병원 연구중심병원과 초연결 혁신융합기술 기반 공동연구 추진

AITRICS, 서울아산병원 연구중심병원과 초연결 혁신융합기술 기반 공동연구 추진

인공지능(AI) 기술 전문 기업 AITRICS(에이아이트릭스)가 서울아산병원 연구중심병원과 사람 중심 초연결 혁신융합기술 기반의 고위험 환자 안전확보를 위한 미래 의료환경 구축에 나선다.

▲ 연구중심병원 사무국 로고(위)와 AITRICS 로고(아래) /이미지 제공=AITRICS

AITRICS, '바이탈케어' MS 마켓플레이스서 판매 개시


오인규 기자 | 입력 2021.04.30 09:45 | 댓글 0

원내 다양한 환경 패혈증 사전 예측…영업 기회 다각화 및 글로벌 시장 진출 가속화 등 지원

[의학신문·일간보사=오인규 기자] AI 기술 전문 기업 AITRICS(대표 유진규, 에이아이트릭스)는 최근 애저 기반 어플리케이션 및 서비스 판매 온라인 스토어인 '마이크로소프트(MS) 애저 마켓플레이스'에서 VitalCare 솔루션 판매를 시작했다고 밝혔다.

이를 통해 솔루션을 구매하는 AITRICS 고객은 마이크로소프트 클라우드 플랫폼 '애저(Azure)'의 강점인 효율적인 제품 설치 및 관리의 혜택까지 받을 수 있다.

AITRICS와 마이크로소프트는 첫 공동판매 제품으로 병원 내 다양한 환경의 패혈증을 사전에 예측하는 자사의 의료 인공지능 소프트웨어 'VitalCare(바이탈케어)'를 선보일 예정이다.