"AI News That's Fit to Print"

The AI Newsletter

Bimonthly

Late Edition

K-water연구원 AI연구센터에서 정기적으로 발간하는 뉴스레터입니다. AI 분야의 뉴스 뿐 아니라 코딩실습과 기술정보 등 다양한 내용과 난이도로 담아냈습니다.

AI Newsletter No.10A

AI LAB, MONDAY, JUL 22, 2024

AI News

What's NEW in AI

블랙박스 해독! 설명 가능한 인공지능, XAI

엔스로픽 자사 LLM 대상 연구결과 발표

XAI란 설명 가능한 인공지능(eXplainable AI)의 약자로 인공지능(AI) 모델의 결정을 사람이 이해할 수 있게 만드는 기술과 방법론을 의미합니다. 예를 들어 기존 AI 시스템이 강아지 사진을 분류할 때는 강아지인지 여부만을 판별하지만 XAI는 털이나 얼굴의 모양 등 그렇게 판단하게 된 근거까지 제시하는 방식입니다. 일반적인 AI 모델, 특히 딥러닝은 블랙박스 모델(Black Box Model)이라고 하는데, 이는 내부 구조나 작동 방식이 복잡하여 모델이 어떻게 특정 결정을 내리는지 사람이 이해하기 어렵다는 것을 의미합니다. XAI는 이런 블랙박스를 열어보고 그 안을 설명할 수 있도록 하는 것이 목표입니다. AI가 내린 결정의 이유를 알 수 있다면 시스템에 대한 신뢰도가 높아지는 것은 물론, 결정에 대한 책임도 명확해 지며, AI의 오류나 편향을 발견하고 이를 개선하는 데에도 도움이 됩니다. 최근 오픈AI의 대항마 앤스로픽이 그간 알기 어려웠던 거대언어모델(LLM)이 작동하는 원리의 실마리를 찾았다는 내용을 담은 'LLM의 마인드 매핑'이란 제목의 연구결과를 발표했습니다. 앤스로픽은 '딕셔너리 러닝(dictionary learning)'이란 기법을 통해 자사 LLM인 클로드 소넷 내부에서 수백만 개의 '특징'(feature)을 추출해 개념화한 지도를 만들었습니다. 예를들어 'Golden Bridge'라는 단어를 언급하면 LLM 내부에서 여러나라 언어로 된 단어, 이를테면 한국어 '현수교' '샌프란시스코' 등의 연관 단어나 단어 일부가 활성화되었습니다. AI 모델 내부 작동방식이 인간처럼 유사한 개념을 통해 해당 단어의 의미를 추론하는 것과 비슷한 면을 보여준다는 것입니다. 전문가들은 이번 연구를 발전시켜 AI 내부 프로세스를 완벽히 이해할 수 있게 된다면 그동안 생성형 AI의 치명적 약점으로 꼽혀온 환각(Hallucination)과 편향(Bias) 문제를 해결할 수 있을 것으로 기대하고 있습니다. XAI 기술의 발전을 통해 그동안 블랙박스 모델의 한계로 지적되었던 AI모델의 투명성(Transparency)과 설명가능성(Explainability), 책임성(Accountability), 탈 전문화(Deskill) 등의 문제를 해결할 수 있게 되기를 바랍니다.

Cluade 3.5 Sonnet 개발사인 앤스로픽의 CEO 다리오 아모데이. 앤스로픽은 최근 거대언어모델(LLM)이 작동하는 원리를 규명한 연구결과를 공개했다. (사진: Claude-AI.Net)

Golden Bridge 언급에 AI 내부에서 연간 단어들이 활성화된 모습(출처: 앤스로픽)

Meanwhile, in K-water

2024 실용인공지능컨퍼런스 참가

인공지능 분야의 최신 연구 동향과 기술을 공유하는 연례 학술대회인 2024년도 실용인공지능컨퍼런스(AAiCON: Applied Artificial Intelligence Conference)가 6월 27일부터 28일까지 대전 DCC에서 개최되었습니다. AI연구센터에서는 27일 김성훈 센터장이 "시계열 데이터 및 빅데이터" 세션에서 좌장을 맡아 진행했으며, 28일에는 이승한 연구원이 "LDA 기반 토픽모델링을 활용한 물산업 분야 기술 트렌드 분석 연구"라는 주제의 논문을 발표했습니다.

서울물연구원 K-water연구원 견학

수처리 분야의 효율성과 안전성을 높이고자 AI 기술 도입이 활발한 가운데 국내 수처리 대표 연구기관인 K-water연구원은 수처리와 AI의 기술융합을 선도하고 있습니다. 7월 4일 서울시 산하 서울물연구원이 Kwater연구원의 앞선 수처리 AI기술을 벤치마킹하고 협력 방안을 모색하기 위해 K-water연구원을 찾아 상하수도연구소와 AI연구센터를 견학하고 기술교류회를 가졌습니다.

K-water연구원은 AI정수장, 수요예측, 수도 동파예측 저수지 수질예측, 교육 및 홍보 등 K-water의 물분야 AI기술 도입 현황과 성과를 소개하였고, 서울물연구원은 도시오염 문제, 정수공정 최적화, 관망관리, 원격검침 등에서 AI기술 도입 가능성 등 관심사항을 논의하였습니다. 두 기관은 수처리 및 AI 분야에서 기술 및 정보교류 등 긴밀한 협력을 이어가기로 했습니다.

AI Reviews

물순환시스템의 물리적 이해를 위한 XAI 기술의 활용

수자원환경연구소 강태호 선임연구원

블랙박스 모델이 의사결정에 미치는 영향

딥러닝 등 AI기술은 데이터로부터 학습된 내용과 출력결과의 이유를 사람이 이해하는 것이 불가능한 블랙박스 모형입니다. 이러한 블랙박스 특성은 의료계에서 수술 등 치료를 결정하는데 딥러닝 AI기술에 의존하기 어려운 이유이기도 합니다. 의사결정의 영향이 중대한 결과를 가져올수록 블랙박스 딥러닝의 기술에 대한 의문이 높은 것이 사실입니다. 최근 그 대안으로 점차 관심을 받는 것이 설명 가능한 학습 및 의사결정 결과를 줄 수 있는 XAI 기술입니다. 대표적인 XAI 기술로는 결정트리(Decision tree)와 사례기반추론(Case-based reasoning) 등으로, 데이터로부터 학습된 AI 구조를 사람이 이해하는 것이 가능하며, 결과적으로 AI가 내린 의사결정에 대한 이유를 설명하는 것까지도 가능합니다. 결정트리의 경우 단계마다의 조건에 의해 학습된 결과가 의사결정에 이르게 되는 과정을 사람이 이해할 수 있습니다. 예를 들어, 코로나 판별에 XAI를 적용할 경우 일반 감기와의 증상 차이를 데이터를 통해 학습하고 어떤 조건(예: 열이 39도 이상 여부)에 의해 구별할 수 있는지 설명할 수 있게 됩니다.

물순환시스템 이해를 위한 XAI

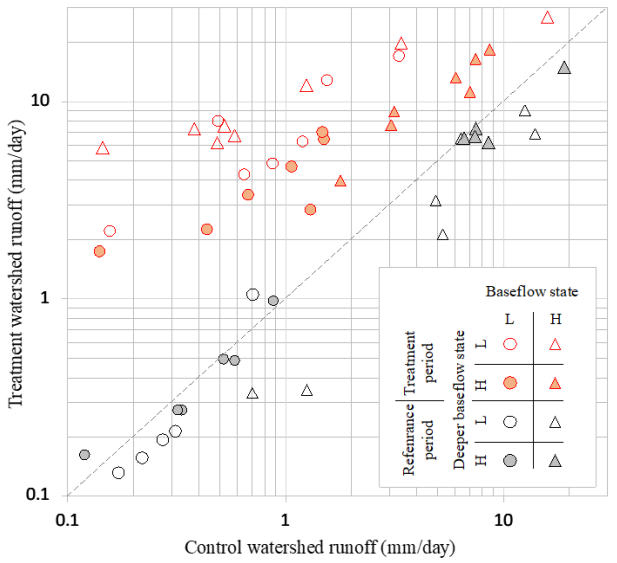
수문기상 분야에서는 유역의 강우-유출 관계 등의 데이터 학습을 통해 물순환시스템의 특성을 설명하는 방안으로 최근 사례기반추론 XAI 기술을 적용한 연구(Kang et al. 2024)가 있습니다. 유역 물순환 과정은 관측이 어려운 땅속 흐름부터 대기로의 증발산 등 복잡하고 광범위하므로, 그 물리적 현상을 명백하게 이해하기 어려운 것이 현실입니다. 일례로 숲이 홍수에 미치는 영향관계에 대해서만도 1900년대 이후 전세계 수백곳의 실제 유역에서

(2면에서 계속)

Paper Review

(1면에 이어서)

실험이 수행되었습니다. 하지만 실험유역 결과는 매우 불확실하여 그 결과가 받아들여지지 못하고 수백년간 논쟁이 이어지고 있습니다. Kang and Sharma (2024)는 불확실성이 있는 실험유역 데이터에 대해 사례기반추론 XAI 기술을 적용하여 사람이 이해할 수 있는 정보(숲과 홍수의 관계)를 얻을 수 있는지 연구하였습니다. 연구결과, XAI를 활용하면 기존 수문학에서 숲과 홍수 관계에 대해 개념적으로 추측한 이론들이 증명 가능하다는 것이 나타났습니다. 연구결과에 따르면 지표면 아래 지하수의 발달된 정도에 따라 숲이 홍수량을 크게 감소시킬 수 있으며, 나뭇잎의 강우 차단(Interception) 효과에 따라 추가적인 홍수저감 효과가 있음을 설명할 수 있었습니다.



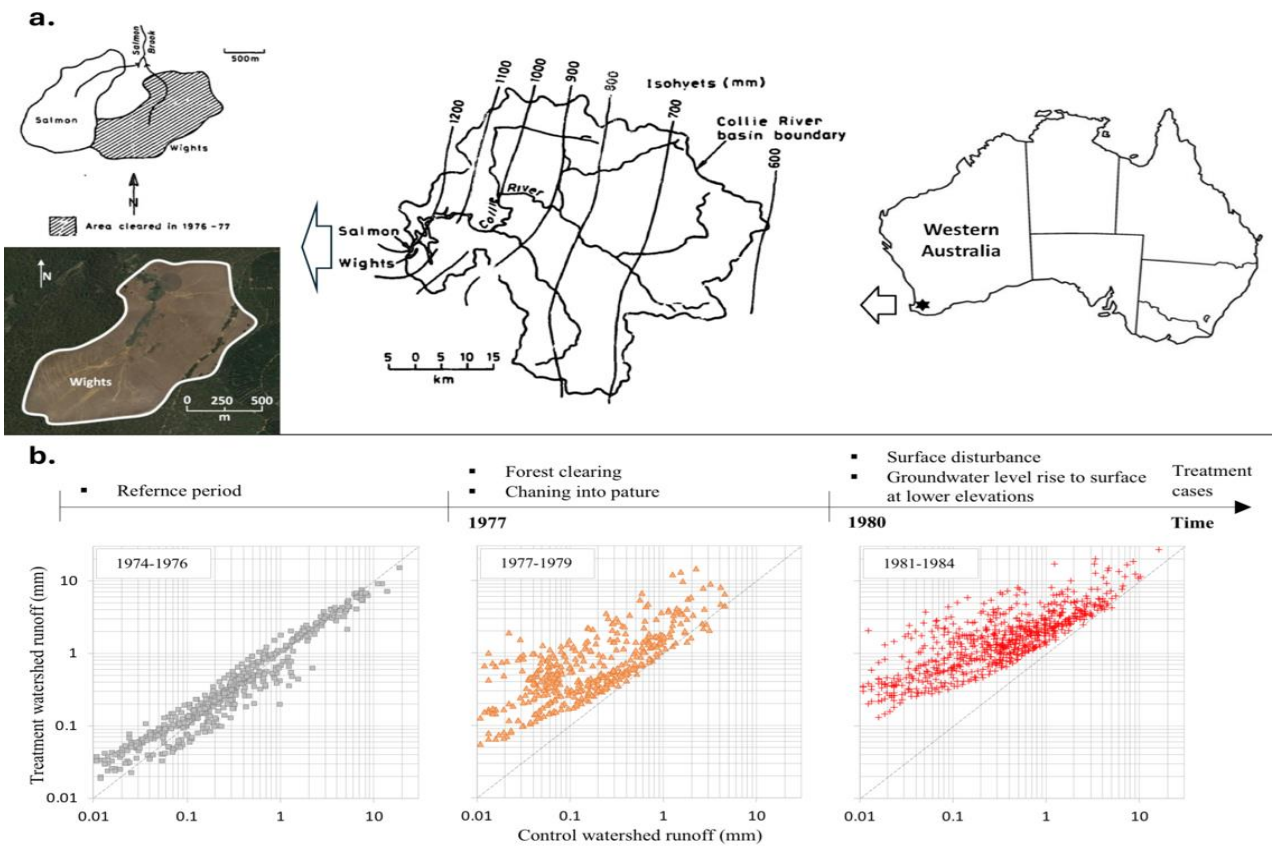
호주의 실험유역에서 20mm 이상 일강우에 대한 사례기반추론 분석 결과

본 논문의 결과는 호주의 실험유역 한곳에 대한 것이지만, 향후 다양한 실험유역에서 XAI기술을 적용하여 분석한다면 보다 일반화된 결론을 얻을 수 있을 것으로 제시하고 있습니다.

이상기후 대응을 위한 XAI

점차 심해지는 이상기후에 대응하기 위해서는 우선 이상기후의 특성과 홍수, 가뭄 등 물재해에 대한 영향을 이해하는 것이 중요할 것입니다. 위성과 수많은 지상관측소의 빅데이터는 이상기후에 대해 상세한 정보를 담고 있으므로, XAI 기술은 이의 해석에 큰 가능성을 제공하고 있는 것으로 보입니다. 이러한 가능성에 따라 수자원환경연구소 물순환연구팀에서는 2024년 기획연구를 통해 이상기후 진단을 위한 XAI 기술 적용 연구를 준비하고 있습니다.

Kang, T. H. and A. Sharma (2024). State-controlled hypothesis test of paired-watershed experiments. *Journal of Hydrology*, 2024.6.28. (To be published)



홍수에 대한 숲의 역할을 이해하기 위해 만들어진 호주의 실험유역(paired-watershed experiment)(上). 실험 결과, 숲과 홍수의 관계가 매우 불확실한 것으로 나타났다(下).