설명가능한 AI (XAI) 기술 동향

Technology Trends in Explainable AI (XAI)

장종현 (J.H. Jang, jangjh@etri.re.kr) 워리어능력증강연구실 책임연구원
이소연 (S.Y. Lee, sylee@etri.re.kr) 워리어능력증강연구실 책임연구원/실장

ABSTRACT

Explainable artificial intelligence (XAI) has emerged as an essential technology in high-stakes domains requiring reliability and safety, including autonomous driving and defense systems. Conventional XAI research has largely focused on post-hoc explanation methods, including SHAP and LIME; however, such approaches exhibit fundamental limitations in terms of explanation fidelity and consistency. We analyze current technical trends in XAI with a primary focus on Kolmogorov–Arnold networks (KANs) as an intrinsically interpretable architecture designed to overcome these shortcomings. Unlike traditional neural networks, which assign fixed weights to nodes, a KAN introduces learnable univariate functions on network edges, enabling the direct interpretation of input–output relationships and simultaneously achieving both high expressiveness and structural transparency. A case study of applying KAN to a vehicle dynamics model further demonstrates that a hybrid XAI framework integrating KAN-based structural interpretability with post-hoc explanation methods, such as SHAP and LIME, offers a practically effective approach to robust and faithful explanations. Going forward, XAI is expected to evolve through the convergence of intrinsically interpretable architectures and post-hoc techniques, alongside the establishment of quantitative evaluation frameworks for explanations. Within this context, KANs are anticipated to serve as a pivotal building block for next-generation XAI systems.

KEYWORDS Explainability, Intrinc(Inherent) Interpretability, Post-hoc, XAI

I. 서론

AI 기술의 집합체인 자율주행 기술의 비약적인 발전과 함께 이동체의 기동 예측을 위한 AI 모델 학습 방법론이 크게 주목받고 있다. 그러나, AI 시스템의 의사결정 과정은 '블랙박스' 형태로 존재하여 사회적 수용성 측면에서 한계가 있으며, 이러한 한계를 극복하고자 설명가능한 AI(XAI) 기술이 자율주행 시스템의 안전성과 신뢰성을 입증하는 핵심 요소로 자리 잡고 있다.

* DOI: https://doi.org/10.22648/ETRI.2026.J.410408
* This work was supported by Korea Research Institute for defense Technology planning and advancement (KRIT) grant funded by the Korea government(DAPA(Defense Acquisition Program Administration)) [No. KRIT-CT-22-087, XR-based Combat Vehicle Training Technology.]

본고에서는 설명가능한 AI, 즉 XAI 기술의 최신 기술 동향을 관련 논문과 기술자료를 기반으로 분석하고자 한다. Ⅱ장에서는 XAI 분야에서 그동안 수행되어 온 연구내용을 살펴보고, Ⅲ장에서는 기존 XAI의 구조적 한계를 해결하기 위해 제안된 KAN 기반 설명성을 요약한다. Ⅳ장에서는 차량 동역학 모델에 적용되고 있는 XAI 연구에 대해 간략히 소개하고, Ⅴ장에서 결론을 맺는다.

II. XAI 주요 연구 동향

1. XAI 연구의 선구자-DARPA

민간 분야보다 훨씬 더 엄격한 안전성과 신뢰성이 필요한 국방 분야에서는 일찍이 설명가능한 AI에 대해 주목하였다. 미국의 방위고등연구계획국(DARPA: Defense Advanced Research Projects Agency)은 2015년 XAI 프로그램을 시작하여 2021년까지 AI 시스템의 투명성과 설명가능성을 높이는 것을 목표로 다양한 연구를 진행했다. 이 프로그램은 복잡한 AI 모델이 사용자에게 그 의사결정의 근거를 설명할 수 있도록 하는 것을 목표로 하였으며, 이를 통해 사용자들이 AI 시스템을 신뢰하고 효과적으로 관리할 수 있게 돕고자 했다.

DARPA의 XAI 프로그램은 크게 세 가지 기술 영역으로 구성되었다. 첫 번째 영역은 새로운 XAI 기계학습 및 설명 기법의 개발을 목표로 했으며, 이는 AI 모델이 생성하는 설명을 효과적으로 전달하기 위한 기법을 포함한다. 두 번째 영역에서는 설명의 심리학적 모델을 개발하여 사용자가 설명을 어떻게 받아들이고 이해하는지를 연구했다. 마지막으로, 해군 연구소가 담당한 평가 영역에서는 프로그램의 성능을 비교하고 효과를 측정하는 작업을 수행했다. 이 세 영역의 협력은 XAI 프로그램의 주요 목표 달성을 위한 통합적 접근을 가능하게 했다[1]. DARPA의 XAI 연구는 AI 모델의 투명성과 신뢰성을 증진시키기 위해 다학제적 접근을 시도한 중요한 연구 프로그램으로, 특히 AI가 제공하는 정보를 기반으로 임무 중심의 의사결정을 수행해야 하는 국방 도메인에 XAI가 반드시 수반되어야 함을 시사한다.

최근 XAI 연구는 다양한 산업과 응용 분야에서 빠르게 발전하고 있으며, 특히 자율주행, 의료, 국방 등 고위험 및 복잡한 임무 분야에서 큰 주목을 받고 있다. 다음 절에서는 XAI 연구 동향을 기술적 및 산업적 측면에서 정리하고자 한다.

2. 이동체의 자율주행 관련 XAI 연구

AI의 안전성과 신뢰성 확보가 중요한 이동체의 자율주행 분야에서 XAI 기술을 적용하기 위한 연구가 매우 활발히 이루어져 왔다.

자율주행의 AI 모델에 대한 설명성은 표 1과 같이 크게 네 가지 방식으로 연구됐다[2]. 시각적 설명은 자율주행 차량이 주변 환경을 얼마나 정확하게 감지하고 있는지 이해하고 모델의 예측 결과를 디버깅하며 신뢰성을 확보하는 것이 목적이다. 강화

표 1 자율주행 관련 XAI 연구

분류	기술적 접근방법
시각적 설명	• Class Activation Map(CAM) • Saliency Map • Guided Backpropagation • Layer-wise Relevance Propagation (LRP)
강화학습 및 모방학습 기반 설명	• Semantic Predictive Control(SPC) • 조감도 마스크 기반 Motion Planning
특징 중요도 기반 설명	• Decision Tree • SHAP • Random Forest & Hybrid
LLM 및 VLM 기반 설명	• Live Natural Language Explanations • VideoQA • RAG-driver • Graph of Thoughts

학습 기반 설명은 인지된 환경 상태가 어떻게 특정 행동으로 이어지는지 설명하는 것을 목적으로 하고, 모방학습 기반 설명은 전문가의 주행을 학습하는 과정에서 시각적 주의 집중이나 논리적 구조를 결합하여 설명력을 확보한다. 특징 중요도 기반 설명은 모델의 예측 결과에 각 입력 변수가 이바지한 정도를 수치화하거나 논리적 구조로 변환하여 제공하므로 계산적으로 투명하고 상대적으로 적은 계산 자원으로 설명할 수 있다는 장점을 갖고 있다. 마지막으로, 대규모 언어모델 및 시각-언어모델 기반의 설명 방식은 자율차량의 의사결정을 해석하고 교통 장면을 묘사하는 방식으로, 2022년 이후 급부상하는 설명 기술이다.

3. 사후 분석 vs 내재적 설명성

2025년도에 개최된 NeurIPS에서는 하버드 연구진이 XAI 기술에 대한 튜토리얼을 진행하였다. 해당 튜토리얼에서는 AI의 설명성 관련 연구의 발전 이력을 체계적으로 정리함과 동시에 '사후 분석적(Post-hoc) 설명성'과 '내재적인(Intrinsic) 설명성'의 두 갈래로 심층적인 분석을 정리하였다[3].

XAI 연구 방법론의 전환점이 되는 시점으로 DNN의 가능성이 확인된 2014년과 LLM이 대두된 2022년을 들 수 있다. 2014년 이전에는, 모델의 의사결정 과정을 이해하기 위해 별도의 사후 분석 기법을 적용할 필요 없이 모델의 구조 자체가 투명하게 설계된 모델, 즉 선형모델, 트리모델, 규칙 기반 모델 등이 적용되었다. 딥러닝이 도입되면서 다양한 사후 분석적 설명성, 즉 Post-hoc 방법이 개발되었다.

2022년에 LLM이 등장하면서 AI 연구뿐만 아니라 설명성 연구에도 큰 변화가 일어나서, 모델의 동작으로 모델 내부의 구성 요소인 뉴런(또는 Layer)

표 2 XAI 기술의 분류

2014년 이전	2014~2022년	2022년 이후	분류
• 선형모델 • 트리모델 • 규칙 기반 모델 • 프로토타입 기반 모델	피처 기여도 데이터 기여도	컴포넌트 기여도	사후 분석
	내재적으로 설명가능한 DNNs	내재적으로 설명가능한 LLMs	내재적 설명성

을 통해 내부 메커니즘을 역설계하여 해석하는 사후 분석적 방법과 LLM을 학습시킬 때 다양한 조건(Data, Architecture, Representations)을 손실함수 등에 반영함으로써 내재적 설명성을 확보하는 방법으로 진화하고 있다(표 2)[4].

III. KAN 기반 설명성

1. KAN 모델의 개발 배경

KAN은 기존 MLP의 구조적 한계와 XAI에 대한 요구를 배경으로 등장했다. MLP는 보편 근사 정리에 기반하여 고정된 활성화 함수와 선형 가중치를 사용하지만, 층이 깊어질수록 내부 작동 기제 해석이 어렵고 특정 비선형 패턴 학습에 과도한 파라미터가 요구되는 비효율성이 존재했다. 제안자는 1957년 콜모고로프-아르놀트(Kolmogorov-Arnold)가 증명한 표현 정리, 즉 복잡한 다변수 함수가 단변수 함수의 중첩과 덧셈으로 표현될 수 있다는 이론을, 현대적인 역전파 알고리즘과 결합하여 학습 가능한 네트워크 구조로 실현하고자 했다[5].

또한, 물리학, 로보틱스 등 도메인 지식이 중요한 과학적 머신러닝(SciML) 분야에서는 단순 예측을 넘어 물리적 법칙과의 연계 및 '법칙 발견'이 요구된다. KAN은 학습된 함수를 기호적 수식으로 변환하거나 시각화하기 쉬운 구조로, 이러한 필요성에 직접적으로 응답하는 아키텍처로 개발되었다[5].

2. KAN 개념

Kolmogorov-Arnold Networks(KAN)은 Kolmogorov-Arnold 표현 정리를 신경망 구조로 구현한 아키텍처로 제안된 모델이다. 기존 MLP의 '선형 가중치 + 고정 활성화 함수' 구조를 재해석한 모델이다. KAN의 핵심 차별점은 각 연결(엣지)에 학습 가능한 1변수 함수 Φ(x)를 배치한다는 점이다. 이러한 함수는 주로 B-spline 등으로 파라미터화되며, 결과적으로 '가중치 = 함수'라는 관점을 취한다. 이 구조는 다음과 같은 특징을 갖는다.

- 엣지 함수의 직접 시각화 가능
- 프루닝(Pruning)을 통한 구조 단순화
- 심볼릭 근사를 통한 수식 형태 변환 가능

이러한 특성 때문에 KAN은 구조 내재적 설명가능성을 갖는 모델로 분류된다. 다만, KAN이 항상 MLP보다 파라미터가 적거나 효율적이라고 일반화하는 것은 부정확하다. 그리드 해상도, 스플라인 차수, 레이어 폭에 따라 메모리 및 계산 비용이 증가할 수 있으며, 이를 해결하기 위한 효율화 연구가 병행되고 있다.

3. KAN 아키텍처

KAN은 레이어 간 연결마다 1변수 함수 $\Phi_{i,j}(\cdot)$를 정의하고, 레이어 출력은 이러한 함수들의 합 또는 합성으로 구성된다. 각 함수는 스플라인 기반 그리드를 통해 근사되며, 다음과 같은 기능이 가능하다.

- Grid Extension을 통한 정밀도 향상
- Pruning을 통한 구조 단순화

이와 같은 구조는 전통적인 노드 기반 활성화 구조와 구별되는 KAN의 핵심 설계 특징이다.

4. KAN 모델의 적용 사례

KAN은 엣지 함수의 시각화와 심볼릭 근사 특성을 활용하여 다양한 과학 · 공학 도메인에 적용되고 있다.

물리 정보 신경망(PINN)과의 결합에서는, 유체역학 등 편미분 방정식(PDE) 풀이에 KAN을 도입함으로써 복잡한 경계 조건과 비선형성을 정밀하게 모사한다. B-spline의 국소적 학습 특성은 불연속점이 있는 물리 현상 모델링에 특히 강점을 보인다.

심볼릭 회귀(Symbolic Regression) 분야에서는, KAN의 엣지 함수를 sin, exp 등 단순 함수로 근사함으로써 복잡한 데이터셋으로부터 $F=ma$와 같은 물리적 수식을 도출하는 연구가 진행 중이다. 이는 앞서 언급한 심볼릭 근사 기능의 실제 적용 사례에 해당한다.

지능형 제어 및 시뮬레이션 영역에서는, 6-DOF 매니퓰레이터나 Stewart Platform과 같은 복잡한 기구학적 시스템의 역기구학 계산에 적용되어, 실시간 시뮬레이션 환경에서의 제어 알고리즘 최적화에 활용되고 있다.

시계열 분석 및 신호 처리 분야에서는, 데이터의 특징을 스플라인 함수로 압축 표현하여 노이즈 제거 및 장기 의존성 파악에서 RNN 계열 모델의 대안으로 검토되고 있다.

5. KAN 기반 설명가능성

KAN의 설명가능성은 사후 기법이 아니라, 모델 구조 자체에서 비롯된다. MLP가 블랙박스적 가중치 집합이라면, KAN은 각 엣지의 함수 형태를 직접 확인할 수 있어 입력-출력 관계를 구조적으로 해석할 수 있다.

설명가능성 측면에서 KAN은 다음과 같은 특징

을 갖는다.

- **전역(Global) 설명 가능:** 엣지 함수 형태로 입력-출력 관계를 직접 표현
- 엣지 수준 기여도 분석 가능
- **심볼릭 근사 가능성 존재:** 희소화 이후 수식 근사(Symbolic Approximation)를 통해 구조적 해석 강화

설명가능성은 크게 사후 설명 방식과 구조 내재적 방식으로 구분된다. SHAP, LIME, Integrated Gradients(IG)는 사후 근사 기반 설명 방식이며, Decision Tree, GAM, KAN 등은 구조 자체가 해석 가능한 모델로 분류된다. 다만, 구조가 복잡해질 경우 해석이 오히려 어려워질 수 있으며, 설명의 충실도(Fidelity) 및 안정성에 대한 정량적 검증은 추가 연구가 필요하다. 또한, 일부 응용 사례에서 과도하게 보고되는 '성능 대폭 향상' 수치는 데이터셋 및 실험 환경에 따라 달라질 수 있으므로, 산업 적용 전에는 재현성 검증이 필요하다.

6. KAN 기술의 분류

KAN 연구는 효율성 개선, 표현력 확장, 그리고 도메인 특화 응용의 세 가지 계열로 나뉘어 진행된다. 표 3부터 표 5는 계열별 세부 기술에 관한 연구 방법을 요약한 것으로, 계열별 특성을 대표한다.

표 3 KAN 모델의 효율성 개선 계열

접근방법	연구방법
FastKAN	KAN의 3차 B-spline 기저를 Gaussian RBF로 근사하여, 스플라인 평가비용을 줄이는 빠른 구현방식[6]
MetaKANs	KAN의 파라미터(또는 가중치)를 '메타-러너'가 생성하도록 설계해, 학습 시 메모리 사용량과 학습비용을 줄이는 접근[7]
추론가속·경량화	CPU/엣지환경에서 spline 평가가 병목이 되는 문제를 겨냥해, 구간별 LUT(Look-Up Table) 기반 양자화 등 추론 최적화 연구 제안[8]

표 4 KAN 모델의 표현력 확장 계열

접근방법	연구방법
Wav-KAN	Wav-KAN은 웨이블릿 함수를 KAN 구조에 통합해 다중해상도(Multi-Resolution) 특성을 활용하는 변형임[9]
SineKAN	학습가능한 B-spline 그리드를 '재가중된 사인 함수 그리드'로 대체하여, 주기성 표현과 속도/모델 크기 측면의 개선 시도[10]
RBF 기반확장	De Boor 알고리즘이 필요한 spline 평가부담을 줄이기 위해, RBF 커널을 엣지함수로 사용하는 흐름[11]

표 5 도메인 특화 응용계열

접근방법	연구방법
TimeKAN	혼재된 주파수 성분을 분해(Cascaded Frequency Decomposition)한 뒤, 각 주파수 대역의 패턴을 KAN 기반블록(M-KAN)으로 학습하고 다시 혼합하는 구조를 제안[12]
TFKAN	시간 영역과 주파수 영역 특징을 병렬로 처리하는 Dual-Branch 구조를 사용해, 주기적 패턴(주파수)과 추세/국소변화(시간)를 동시에 학습하도록 설계[13]
강화학습 및 과학계산(PDE) 확장	연속함수 근사능력과 해석가능성(엣지함수시각화)을 바탕으로, 강화학습(World Model/정책 네트워크 구성 요소교체) 및 과학 계산(PDE/연산자 학습)에서 실험적 적용이 진행 중[14].

IV. XAI 분석 사례

현재, 국산 이동체에 장착된 IMU/GPS 센서 데이터를 기반으로 차량 동역학 AI 모델 개발 및 그에 대한 설명성을 분석하는 연구가 진행되고 있다. 즉, 차량 동역학 AI 모델에 대한 XAI 적용은 베이스라인 모델의 설명성을 분석하고, 동시에 타 탑승체 동역학 모델 개발 시에도 최소한의 데이터 수집으로 빠르게 모델을 개발하기 위한 방법론을 연구하는 것이 핵심이다. 표 6은 KAN, SHAP, LIME의 설명 가능성 기법을 설명 범위, 설명 방식, 수식 근사 가능 여부 및 적용 단계 측면에서 비교한 것이다.

표 6 설명가능성 방식 비교(KAN vs. SHAP vs. LIME)

구분	KAN	SHAP	LIME
설명 범위	전역(Global)	전역(Global)	지역(Local)
설명 방식	구조 내재적	사후(Post-hoc)	사후(Post-hoc)
수식 근사	가능	불가	불가
적용 단계	KAN 최종 계층	PINN/GRU 출력	이상 샘플 추적

1. 데이터 기반 이동체 동역학 모델 개발

데이터 기반 차량 동역학 모델은 물리적 특성과 운동 법칙에 기반하여 정확하게 수학 모델로 만들 수 있으나, 비선형 복잡성 및 고도의 전문 지식이 요구되는 등 한계가 있어, 최근 들어 실 주행 데이터를 기반으로 학습하여 동적 거동을 예측하는 데이터 기반 모델링의 개념을 도입하고 있다[15].

그 결과 종방향에서는 속도와 가속도 데이터에서 높은 정합성을 보였으며, 횡방향에서도 유사한 성능을 확인하였으므로, 데이터 기반 AI 기술 적용이 효과가 있음을 보였다. 향후에는 '차량 동역학 수집데이터의 전처리모델 → 설명성 → 잔차모델'의 순서로 연결되며, 물리 법칙 기반 예측과 데이터 기반 잔차 보정을 결합한 하이브리드 구조를 개발 중이다.

2. KAN 기반 XAI 분석 적용 예

KAN의 설명가능성은 사후 기법이 아니라 모델 구조 자체에서 비롯된다. MLP가 블랙박스와 같은 가중치 집합이라면, KAN은 각 엣지(Edge)의 함수 형태를 직접 확인할 수 있어 입력–출력 관계를 구조적으로 해석할 수 있다.

일부 응용에서 보고되는 '성능 대폭 향상' 수치는 데이터셋 및 실험 환경에 따라 달라질 수 있으며, 산업 적용 전 재현성 검증이 필요하다. 설명의 충실도 및 안정성에 대한 정량적 검증 또한 추가 연구가 필요하다.

그림 1은 KAN 개념을 적용한 피처의 영향력 분석 결과를 보여준다.

3. SHAP 기반 사후 설명

SHAP(SHapley Additive exPlanations)은 게임 이론의 샤플리 값(Shapley Value)을 기반으로 각 입력 특성의 예측 기여도를 전역적(Global) 수준에서 정량화한다. AI Dynamics 파이프라인에서는 PINN 및 GRU 출력 단계에 적용되어 GYRO_Z, ROLL, PITCH, ACC 예측에 영향을 미치는 주요 센서 채널(가속도, 각속도, 지형 특성 등)의 기여도를 순위화한다. Summary Plot, Waterfall, Bar, Force Plot 등 다양한 시각화를 통해 운용자에게 모델 동작 근거를 제공한다. 그림 2는 SHAP을 이용한 피처 중요도 및 각 피처가 예측에 미치는 영향 분포를 나타낸 결과이다.

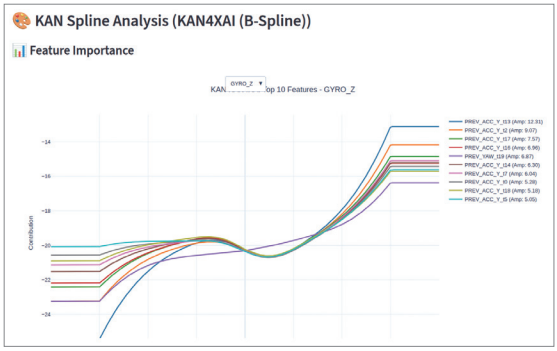

그림 1 KAN 적용 예

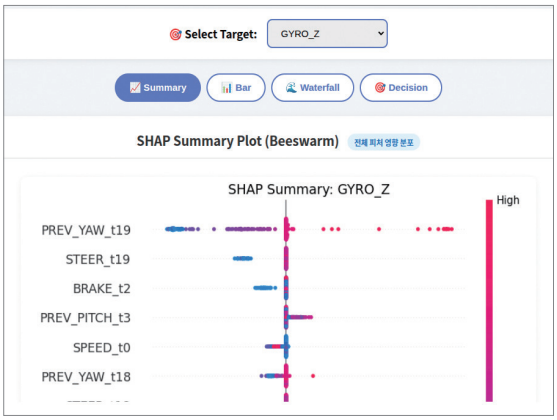

그림 2 SHAP 적용 예

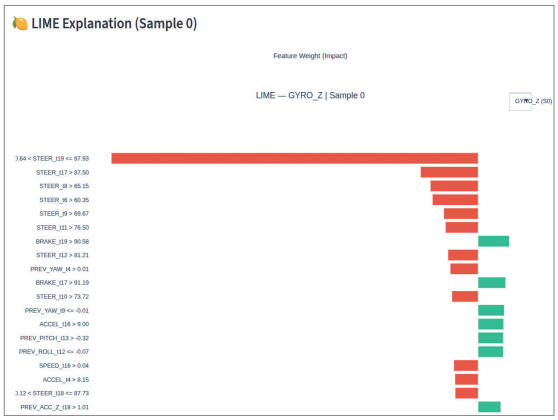

그림 3 LIME 분석 결과 예

4. LIME 기반 국소적 설명

LIME(Local Interpretable Model-agnostic Explanations)은 특정 입력 샘플 근방을 선형 모델로 근사하여 국소적(Local) 설명을 제공한다. 이는 KAN의 전역 설명과 상호 보완적으로 작동하며, 특이 기동 구간(급선회, 험지 주행 등)에서 예측이 특정 방향으로 편향될 때 해당 샘플에 대한 원인 변수를 국소적으로 추적한다. SHAP이 전체 분포에 대한 기여도를 설명한다면, LIME은 개별 예측 사례에 대한 근거를 제시하여 모델 신뢰성 검증에 활용된다.

전체적인 XAI 분석 결과, 차량 동역학 AI 모델이 타겟 출력값인 '자이로/자세(롤, 피치)/가속도 채널(GYRO, ROLL, PITCH, ACC)'에 영향을 미치는 입력 피처들을 확인할 수 있었고, XAI(KAN)이 모델 해석 계층에 잘 동작되는 것을 확인하였다. 그림 3은 LIME을 이용한 피쳐 중요도(영향력) 분석 결과를 보여준다.

V. 결론

본고에서는 설명가능한 AI, 즉 XAI 기술의 최신 기술 동향을 관련 논문과 기술자료를 기반으로 정리하였다. AI 기술의 산업적 · 사회적 파급력이 높아지면서, 자연적으로 AI 모델에 대한 안전성과 신뢰성이 주요 쟁점이 되었고, XAI는 이를 뒷받침하는 도구로서 지속적으로 연구가 진행되고 있다. AI 기술이 적용되는 도메인 중에서 특히, 국방, 안전, 재난과 같이 임무 중요도가 높은 분야에서는 항상 다음과 같은 질문이 선행된다.

"How can we build robust, assured, and therefore trustworthy AI-based systems?"

향후 XAI 기술은 위 질문에 답할 수 있는 연구 방향을 정립하는 데 결정적 역할을 할 것으로 기대된다. 또한, XAI 기술은 구조 내재적 설명과 사후 설명 기법의 통합, 그리고 설명의 정량적 평가 체계 확립을 중심으로 발전할 것으로 예상하며, KAN은 이러한 흐름에서 중요한 역할을 할 것으로 기대한다.

용어해설

KAN(Kolmogorov-Arnold Networks) 1957년 콜모고로프-아르놀트 표현 정리를 신경망 구조로 구현하여, 기존 MLP의 고정된 활성화 함수 대신 각 연결(엣지)에 학습 가능한 1변수 함수를 배치함으로써 구조적 해석 가능성을 극대화한 차세대 신경망 아키텍처

XAI(Explainable AI) AI 모델의 복잡한 블랙박스와 같은 내부 작동 기제를 인간이 이해할 수 있는 형태로 설명하여 신뢰성과 투명성을 제공하는 기술로, KAN과 같이 구조 자체가 해석 가능한 방식과 SHAP/LIME과 같은 사후 설명 방식으로 구분됨

SHAP(SHapley Additive exPlanations) 협동 게임 이론의 샤플리 값을 기반으로 각 입력 특성이 전체 예측 결과에 미치는 기여도를 전역적(Global) 수준에서 정량화하여 순위와 분포를 시각화하는 사후 설명 기법

LIME(Local Interpretable Model-agnostic Explanations) 특정 입력 샘플 근방을 단순한 선형 모델로 근사하여, 개별 예측 사례나 특이 구간에서의 판단 근거를 국소적(Local)으로 추적하는 지역적 사후 설명 기술

참고문헌

[1] D. Gunning et al., "DARPA's explainable AI (XAI) program: A retrospective," Applied AI Letters, vol. 2, no. 4, 2021. doi: 10.1002/ail2.61

[2] S. Atakishiyev et al., "Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions," IEEE Access, vol. 12, 2024, pp. 101603-101625. doi: 10.1109/ACCESS.2024.3431437

[3] S. Zhang et al., "Towards unified attribution in explainable AI, data-centric AI, and mechanistic interpretability," arXiv preprint, 2025. doi: 10.48550/arXiv.2501.18887

[4] https://shichangzh.github.io/xaiTutorial/

[4] S. Zhang et al., "Explain AI models: Methods and opportunities in explainable AI, data-centric AI, and mechanistic interpretability," NeurIPS 2025 Tutorial. https://shichangzh.github.io/xaiTutorial/

[5] Z. Liu et al., "KAN: Kolmogorov-Arnold Networks," arXiv preprint, 2024. doi: 10.48550/arXiv.2404.19756

[6] Z. Li, "Kolmogorov-Arnold networks are radial basis function networks," arXiv preprint, 2024. doi: 10.48550/arXiv.2405.06721

[7] Z. Zhao et al., "Improving Memory Efficiency for Training KANs via Meta Learning," arXiv preprint, 2025. doi: 10.48550/arXiv.2506.07549

[8] O. Kuznetsov, "LUT-KAN: Segment-wise LUT Quantization for Fast KAN Inference," arXiv preprint, 2026. doi: 10.48550/arXiv:2601.03332

[9] Z. Bozorgasl and H. Chen, "Wav-KAN: Wavelet Kolmogorov-Arnold networks," arXiv preprint, 2024. doi: 10.48550/arXiv.2405.12832

[10] E.A.F. Reinhardt et al., "SineKAN: Kolmogorov-Arnold Networks Using Sinusoidal Activations," arXiv preprint, 2024. doi: 10.48550/arXiv.2407.04149

[11] S.T. Chiu, S.W. Cheung, U. Braga-Neto, C.S. Lee, and R.P. Li, "Free-RBF-KAN: Kolmogorov-Arnold networks with adaptive radial basis functions for efficient function learning," arXiv preprint, 2026. doi: 10.48550/arXiv.2601.07760

[12] S. Huang et al., "TimeKAN: KAN-based frequency decomposition learning architecture for long-term time series forecasting," arXiv preprint, 2025. doi: 10.48550/arXiv.2502.06910

[13] X. Kui et al., "TFKAN: Time-frequency KAN for long-term time series forecasting," arXiv preprint, 2025. doi: 10.48550/arXiv.2506.12696

[14] C. Shi and X. Luan, "KAN-Dreamer: Benchmarking Kolmogorov-Arnold networks as function approximators in world models," arXiv preprint, 2025. doi: 10.48550/arXiv.2512.07437

[15] 유재승, 이기범, "물리 정보 신경망을 통한 차량 동역학 모델 파라미터 추정," 자동차안전학회지, 제17권 제2호, 2025, pp. 30-35.