AI 신뢰성을 위한 XAI 기술 동향

심혜진[1,†], 최창우[1], 김호원[1,2,‡]
[1]부산대학교 정보컴퓨터공학부
[2]스마트엠투엠
simhyejini17@gmail.com, changwoo7463@gmail.com, howonkim@gmail.com

XAI Technology Trends for AI Reliability

Hye-Jin Sim[1,†], Chang-Woo Choi[1], Ho-Won Kim[1,2,‡]
[1]School Computer of Computer Science and Engineering, Pu-san National University
[2]SmartM2M

요 약

4차 산업 시대가 도래하며 인공지능이 비약적으로 발전함에 따라, 인공지능은 다양한 산업 분야에 도입되어 업무의 효율성을 높이고 인류 발전에 중요한 역할을 하고 있다. 그러나 사회 전반에 걸쳐 인공지능의 역할이 커질수록 인공지능의 오판단, 오작동으로 인한 문제 또한 크게 작용한다. 따라서 인공지능 모델의 판단, 행동에 대한 신뢰성을 확보하기 위해 XAI 기술의 중요성이 크게 대두되었다. 본 논문에서는 이러한 XAI 기술에 대한 동향을 조사, 분석한다.

1. 서론

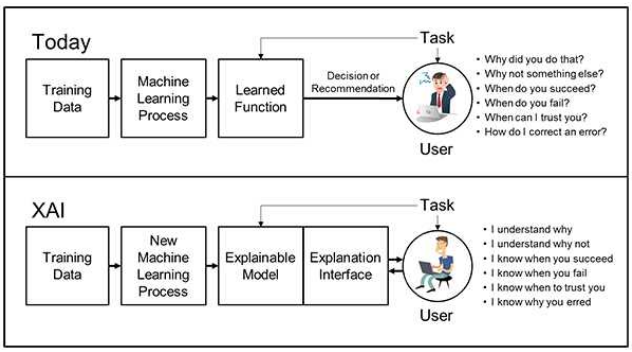
현대 사회에는 분야를 막론한 다양한 산업 분야에서 인공지능 기술을 개발하고 적용하고 있다. 인공지능이 인간의 일상과 더불어 여러 분야에 깊이 개입되어 있고, 이로 인해 다양한 업무에 효율성과 생산성 향상, 일상생활에서의 편리성 제공 등 인공지능이 이제는 인류 사회에서 완전히 배제될 수 없는 위치로 자리하게 되었다. 그러나 인공지능이 인류 사회의 모든 분야에 중요하게 작용하고 있음과 동시에 인공지능의 잘못된 판단으로 인해 발생하는 문제에 대한 우려 또한 커지고 있다. 인공지능에 대한 부정적 시각의 가장 큰 이유는 인공지능의 블랙박스(Black-Box) 성향으로 인해 인공지능의 잘못된 판단에 대한 근거를 인간이 알 수 없었기 때문이다. 이 문제를 해결하기 위해 설명 가능한 인공지능(eXplainable Artificial Intelligence, XAI) 이 등장하고, 인공지능의 블랙박스 성향을 분석·파악하여 설명 가능성을 제공할 수 있도록 하는 XAI 기술이 크게 중요시되었다[1]. 설명 가능한 인공지능(eXplainable Artificial Intelligence, XAI)이란 인공지능의 학습 결괏값에 대한 이유를 인간이 이해할 수 있도록 설명을 제공하는 것이다. 인공지능의 신뢰성 확보를 위한 다양한 접근 방법 중 특히 XAI가 중요한 이유는 판단을 내린 이유와 더불어 그 과정을 제시함으로써, XAI를 통해 인공지능의 오판단·오작동에 대한 원인 식별이 가능하기 때문이다[1]. 따라서 본 논문에서는 인공지능의 작용 과정에 대한 신뢰성 확보를 위해 사용되는 XAI 기술과 관련된 연구 동향과 활용 사례, 향후 연구 방향에 대해 살펴보고자 한다.

2. 본론

2.1 등장 배경

인공지능이 발전하면서 심층신경망(Deep Neural Network)과 같은 불투명한 의사결정 시스템이 등장했고, 이러한 심층신경망(DNN)과 같은 딥러닝 모델은 효율적인 학습 알고리즘과 방대한 매개변수 공간의 조합에서 학습을 수행한다[2]. 이 공간은 수백 개의 layer와 수백만 개의 parameter로 구성되어 있어서 DNN 기반의 인공지능 모델을 복잡한 블랙박스 모델로 간주하는 것이다[2]. 이러한 블랙박스 성향을 보이는 인공지능 모델이 사용됨에 따라 투명성에 대한 요구가 증가했고, 모델 메커니즘에 대한 이해를 높이기 위해 인공지능을 설계할 때 '해석 가능성'을 추가로 적용하는 XAI가 등장하게 되었다.

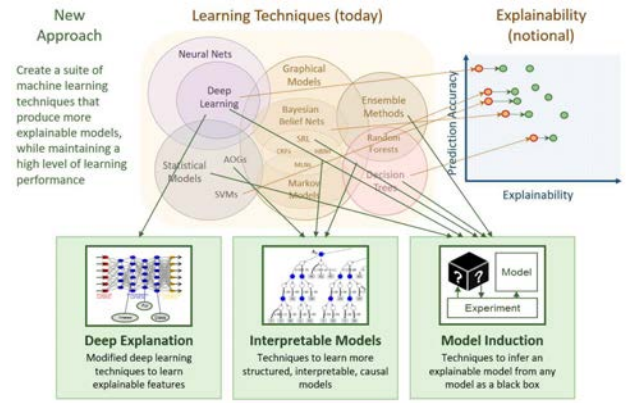
(그림 1)과 같이 인공지능이 잘못된 결과를 제시했을 시, 왜 이러한 결과가 도출되었는지 작동 방식에 대한 설명 가능성을 제공하여 메커니즘의 결함을 수정할 수 있도록 하는 것이 XAI 이다.



(그림 1) XAI Concept
출처: 방위고등연구계획국(Defense Advanced Research Projects Agency, DARPA).

2.2 연구 동향

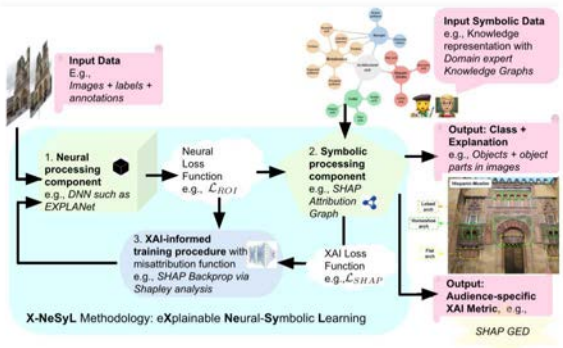
인공지능 분야에서 판단 근거의 불확실성을 해소하기 위한 AI 신뢰성 증대 요구가 점차 강해지면서 XAI 연구가 활발히 진행되고 있다. 방위고등연구계획국(Defense Advanced Research Projects Agency, DARPA)은 사용자가 인공지능 메커니즘을 더 잘 이해하고, 신뢰하며 효과적으로 관리할 수 있도록 2015년 XAI 프로그램 개발을 공식화했으며, 2017년부터 4년간의 XAI 연구 프로그램을 수행하였다[3]. (그림 2)는 DARPA가 제시한 XAI 연구 프로그램의 세 가지 기술적 접근 방법이다.



(그림 2) DARPA가 제안한 XAI 기술 연구의 세 가지 기술적 접근 방법
출처: DARPA-BAA-16-53, Proposers Day Slides

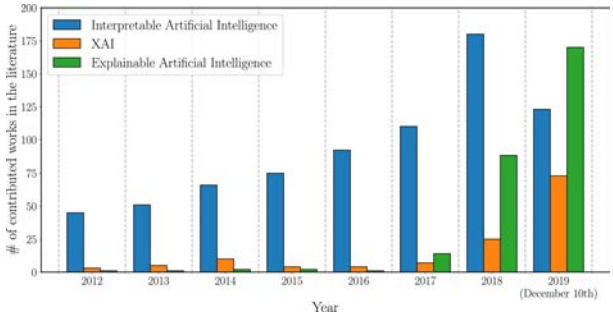
DARPA가 제시한 세 가지 접근 방법을 분석하자면, 첫 번째는 Deep Explanation으로 더 많은 설명 가능한 기능, 설명 가능한 표현 또는 설명 생성 기능을 학습하는 하이브리드 딥러닝 기술개발이다[4]. 이는 인공지능이 최종적으로 결과를 도출하기까지 어떤 과정을 거치는지에 대한 정보를 알 수 있는 발판이 된다[5]. 두 번째는 Interpretable Models이다. 기존 모델보다 구조화되고, 해석이 가능하거나 인과 관계가 있는 모델을 학습하는 대체 기계 학습 기술을 개발한다[4]. 즉, 의사결정 트리와 같은 설명력이 높은 단순한 학습방법과 연계하여 일치성을 찾고 이에 빗대어 결과 도출 과정을 해석할 수 있는 모델을 개발한다는 의미이다[5]. 마지막으로 Model Induction으로 주어진 기계학습 모델을 블랙박스로 실험하여 대략적이고 설명 가능한 모델을 추론하는 기술을 개발하는 것이다[4].

최근에 나온 EXplainable Neural-Symbolic Learning (이하 X-NeSyL)은 사전 지식을 통합하는 동시에 수학 방정식이나 DSL(Domain-Specific Languages)과 같이 해석 가능한 출력을 생성하는 학습 방법론이다[6]. X-NeSyL은 딥러닝 모델, 특히 CNN(Convolution Neural Network) 분류 모델의 성능과 설명 가능성을 모두 향상하도록 설계되었다[6]. 해당 논문에서의 X-NeSyL 방법론은 세 가지 주요 구성 요소로 구성되는데, 첫 번째는 상징적 표현(Symbolic representation)을 처리하기 위한 symbolic processing component, 두 번째는 표현을 학습하기 위한 eXplainable Part-based classifying network 아키텍처(EXPLANet), 세 번째는 신경망의 표현을 지식 그래프(KG)의 상징적인 표현과 정렬하기 위한 SHAP-backprop을 제안한다[6].



(그림 3) eXplainable Neural-Symbolic learning

최근 우리나라도 과학기술정보통신부를 중심으로 XAI 연구센터를 설립하여 세계적 수준의 XAI 연구를 진행하고 있다[7]. 이처럼 4차 산업에서는 AI 메커니즘의 작동 방식과 판단 과정에 대한 이해가 보안 적인 관점에서 더욱 중요해진다는 의미를 담고 있다[7]. (그림 3)은 세계 각국에서 XAI 기술 연구가 2017년부터 급격하게 증가하고 있음을 보여준다[2].



(그림 3) XAI 분야를 참조하는 총 출판물 수의 추이

2.3 적용 사례

최근 산업 분야에서의 XAI 기술 도입이 두드러지게 나타나고 있는데, 특히 사용자의 신뢰가 바탕이 되어야 하는 산업 분야들에 XAI의 중요성이 훨씬 부각 되고 있다. 대표적으로 자율주행차, 금융, 의료 등이 있다.

사용자의 자율주행차에 대한 불신은 꽤 높은 편이다. 자율주행차를 타면서 발생하는 여러 사고나 돌발 상황에 대한 우려가 크기 때문이다. 금융 분야에서는 사용자의 신용정보를 관리하는데 인공지능 기술이 사용되고 있다. 하지만 사용자는 개개인의 신용에 대한 변화가 생겼을 경우 왜 이런 변화가 생겼는지에 대한 정보를 원한다[9]. 게다가 의료 인공지능에서는 진단에 대한 근거가 절대적으로 중요한 만큼 상용화를 위해 XAI 기술이 필수적으로 자리잡는 추세이다[8]. 이러한 이유로 여러 분야에서 XAI 기술 도입을 위한 개발이 활발하게 진행되고 있다.

3. 결론

인공지능을 적용한 기술은 다양한 산업 분야와 인간의 일상생활에 깊게 관여되어 있을뿐더러 그 영향력을 더 넓혀나가고 있지만, 인공지능의 신뢰성에 대한 문제는 꾸준히 제기되었다. XAI는 진화된 인공지능의 하나의 형태라고 볼 수 있다. 인공지능에 설명 능력을 더함으로써 인간의 인공지능에 대한 신뢰성 향상은 물론 편향을 줄일 수 있다[1]. 현재 XAI에 대한 보편적이고 확실한 솔루션은 없는 상태이다[3]. XAI의 기초 기술개발을 넘어 인간과 완전한 상호작용을 할 수 있는 XAI 개발을 위해서는 심리학, 인적 요소 등 여러 분야에 걸친 협업이 필요하다는 것이 전문가들의 견해이다[3]. 앞으로도 전반적인 산업 분야를 막론하고 인공지능이 적용된 어느 곳이든, XAI에 대한 요구가 증가할 것으로 생각된다. 그러므로 XAI는 향후 몇 년간 활발한 연구 분야가 될 것으로 기대한다.

Acknowledgment

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획 평가원의 지원을 받아 수행된 연구임(IITP-2022-0-01201)

참고문헌

[1] IRS Global, “[ICT/정보통신] 설명 가능한 인공지능(XAI)이란?”, May, 2022.

[2] A.B Arrieta, Natalia Díaz-Rodríguez, and J.D Ser, “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible A”, Information Fusion, vol. 58, pp. 82-115, June. 2020.

[3] David Gunning, Eric Vorm and Yunyan Wang, “DARPA’s Explainable AI (XAI) program: A retrospective”, Nov. 2021.

[4] DARPA, “Explainable Artificial Intelligence(XAI)”, Broad Agency Announcement, DARPA-BAA-16-53, Aug. 2016.

[5] 정호근, “이상탐지를 위한 XAI 기법에 관한 연구”, 석사학위논문, 건국대학교, 2019년 8월.

[6] Natalia Díaz-Rodríguez, A. Lamas, J. Sanchez, G. Franchia, I. Donadellob, S. Tabikf, D. Filliata, P. Cruze, R. Montes, F. Herrera, “EXplainable Neural-Symbolic Learning (X-NeSyL) methodology to fusedeep learning representations with expert knowledge graphs: theMonuMAI cultural heritage use case”, Information Fusion, vol. 79, pp. 58-83, Mar. 2022.

[7] 김원정, “‘AI는 왜 그런 판단을 내렸을까’…신뢰성 요구 높아져”, 산업일보, 2021년 12월.

[8] 김남국, KHIDI 디지털 헬스케어 리포트 “의료 인공지능의 신뢰성과 안전성”, 한국보건산업진흥원, 2021년 10월.

[9] 천예은, 김세빈, 이자윤, “설명 가능한 AI 기술을 활용한 신용평가 모형에 대한 연구”, 한국데이터정보과학회지, pp. 238-295, 2021년 1월.