**Explainable Artificial Intelligence (xAI) Approaches and Deep Meta-Learning Models**

[**https://www.intechopen.com/online-first/explainable-artificial-intelligence-xai-approaches-and-deep-meta-learning-models**](https://www.intechopen.com/online-first/explainable-artificial-intelligence-xai-approaches-and-deep-meta-learning-models)

**<0. Introduction>**

|  |  |
| --- | --- |
|  |  |
| XAI는 **Machine Learning, Human-Computer Interaction, End User Explanation**이 결합된 것이다. | 정확도가 높다고 알려진 알고리즘일수록 **설명 가능성(explainability)은 낮아진다.**   * **딥러닝**은 정확도가 매우 높지만 신경망으로 구성되어 있기 때문에 반대로 설명 가능성은 매우 낮다. * **Classification Rule**은 정확도가 매우 낮지만 반대로 설명 가능성은 매우 높다. |

**<1. Explainable Artificial Intelligence (xAI)>**

|  |
| --- |
| 현재의 일반적인 머신 러닝에서는 **머신 러닝 결과에 대한 근거, 언제 머신러닝을 믿을 수 있는지** 등을 알 수 없다.  Learned Function을 **Explainable Model + Explanation Interface**로 바꾸면 머신러닝 결과의 근거, 머신러닝을 언제 신뢰할 수 있는지 등을 알 수 있다. |

* 이 모델은 **이해할 수 있고 유용한 explanation dialog로 변환될 수 있는 인터페이스 기술과 결합**하는 것을 목표로 한다.

**<1. Explainable Artificial Intelligence (xAI)>**

일반적인 딥러닝과 달리 **Explanatory model과 Explanation interface**라는 요소가 있다.

|  |  |
| --- | --- |
| Explanatory model | 적용 가능한 **rule-based reasoning system**   * 입력 데이터와 머신러닝 출력값의 **cause-effect 관계**를 드러내는 구조 * 이 구조는 **내부 딥러닝 방법을 이용하여 rule을 학습**하고, 이를 통해 **원인을 탐색**하고 **새로운 상황에 대한 전략**을 개발한다. |
| Explanation interface | **User Interaction**의 한 부분   * **Voice digital assistant**의 question-answer 인터페이스와 유사 * 사용자의 **요구 사항을 평가하는 decoder unit**과 explanatory model로부터의 **답변을 가능하게 하는 encoder unit**으로 구성 |
| Knowledge Graph | **Entity(semantic technology)들과 relationship들**이 연결된 **Knowledge Graph (KG)**는 **understandability, reasoning mechanism** 등의 이슈에 대한 유용한 해결책을 제시한다.   * **입력 데이터, hidden layer, encoded features, 딥러닝 모델의 예측된 output**이 **Knowledge graph** 또는 **ontology의 concept/relationship**으로 pass된다. * **알고리즘의 내부 동작**은 딥러닝 요소(input features, hidden unit/layers, output, …)들의 **knowledge matching**으로 보다 이해하기 쉬워진다. |

* xAI 구조가 모델별로 서로 다르다고 해도 **Explanatory model, Explanation Interface, Deep Learning model은 서로 협동**하여 동작한다.

**<여러 가지 딥러닝 모델> [Figure 5]**

|  |  |
| --- | --- |
|  |  |
| AutoEncoder | CNN (Convolutional Nerual Networks) |
|  | |
| RNN (Recurrent Neural Network) | |

**<XAI 모델>**

|  |
| --- |
|  |
| **Explainable AI 모델을 위한 Semantic Knowledge Matching** **[Figure 6]** |
|  |
| **IBM에서 개발한 Explainable AI 모델: SENN (self-explaining neural network)** **[Figure 7]**  **Back-propagation을 이용하여 end-to-end로 학습**할 수 있으며, 입력값은 **interpretable basis feature들의 집합으로 변환**된다.   * Relevance score는 **input-dependent parametrizer**에 의해 생성된다. * 생성되는 예측은 **aggregation function에 의해 합성**된다. * 전체 모델은 **변수 가 있는 선형 함수 h(x)에 의해 독립적**으로 작동하며, **concept와 relevance에 대한 설명**을 robustness loss에 의해 생성한다.   + 모델링은 중요하기 때문에 **model은 higher-capacity architecture를 유지**할 수 있도록, concept가 raw input이 된다고 해도 **충분히 rich**해야 한다. |

**<5. Explainable meta-reinforcement learning (xMRL)>**

xMRL(Explainable meta-reinforcement learning)은 **설명 가능한 reinforcement learning model**이다.

|  |
| --- |
| **Deep reinforcement learning model:** 특정 상태에서 **어떤 행동을 취해야 reward가 가장 큰지**를 학습하여 **보상을 최대화**하는 것을 목표로 하는 강화학습 모델   * **로보틱스, 자율주행 자동차, 게임** 등에서 많이 이용 * **xAI를 deep reinforcement learning model에 적용**하려고 할 때, 모델의 **causal decision-making 능력**을 **agent의 입출력 간의 relation link**를 통해 얻을 수 있다. * **Inverse Reinforcement Learning 모델**을 이용하여 **reward derivation 메커니즘을 학습**할 수 있다.   + 이때 이전의 접근법과는 달리 **여러 환경에 적용할 수 있는 meta-cognitive 인공지능 모델**이 개발된다.   + 이것을 XAI 모델과 같이 생각하면 **이미 개발된 agent는 cause-effect relationship을 통한 자체 전략을 수립**할 수 있다.   + **Explainable meta-reinforcement learning agent**는 바둑, 체스와 같은 게임, 혹은 **전혀 새로운 게임도 학습**할 수 있다. 또한 XAI이므로 **왜 이런 행동을 했는지 설명**도 할 수 있을 것이다. |
| (a)는 **Reinforcement learning**이고 (b)는 **inverse reinforcement learning**이다. |