설명 가능한 인공지능(eXplainable AI, XAI) 소개

(보안기술연구팀, 2018.3.23.)

1 개 요

□ 인공지능 기술(머신러닝 등)은 빅데이터 및 복잡한 알고리즘 등을 기반으로 사용자에게 의사결정, 추천, 예측 등의 정보를 제공하지만,

○ 일부 머신러닝 기술(딥러닝 등)은 알고리즘의 복잡성으로 인해 “블랙박스”라 불리며,

○ 도출한 최종 결과의 근거, 도출과정의 타당성 등을 제공하지 못하는 이슈가 존재

□ 한편, 금융, 보험, 의료 등의 분야에서 고객의 신뢰를 기반으로 개인 정보와 자산 등을 다루는 인공지능 시스템의 경우,

○ 공정성, 신뢰성, 정확성 등을 보장하기 위해 인공지능 시스템으로부터 생성된 결과의 도출 근거와 도출 과정의 타당성 등에 대한 확인이 필요

□ 이에 본 보고서에서는 사용자(개발자, 관리자 등)가 인공지능 시스템의 최종 결과를 이해하고,

○ 설명할 수 있도록 정보를 제공하는 설명 가능한 인공지능(eXplainable AI, XAI)에 대해 간략히 소개

2 XAI의 등장 배경 및 개념

□ **(등장 배경)** 인공지능 시스템 결과에 대한 사용자 및 사회의 수용과 신뢰가 우려되면서 XAI에 대한 관심이 높아짐

○ ‘70년대 인공지능 시스템인 전문가 시스템이 도출 결과를 전문가들에게 이해시키지 못하면서, 이후 설명 가능한 인공지능의 중요성이 인식되었고 일부 연구자들에 의해 연구

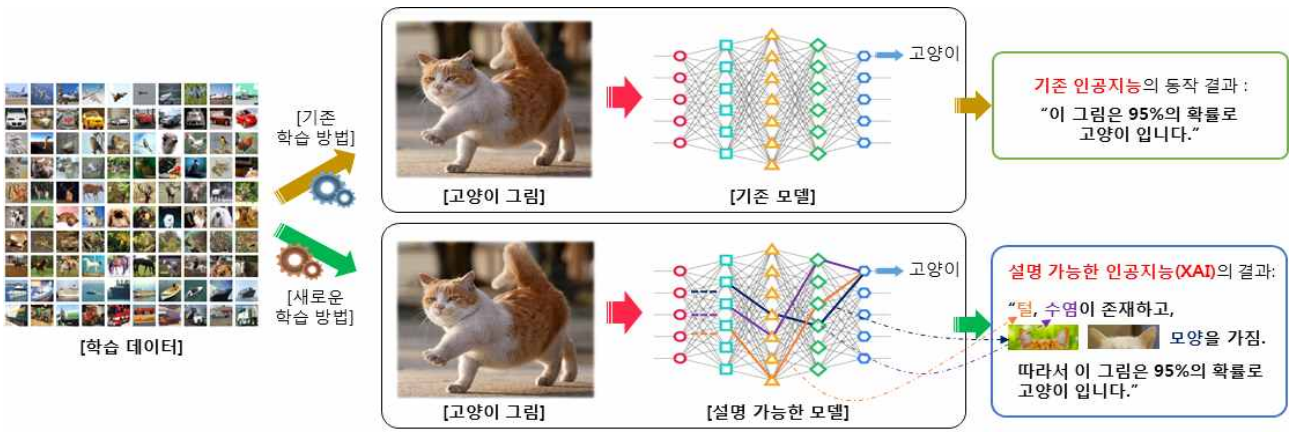
○ 최근 딥러닝이 전 세계적으로 확산되고, 다양한 분야에 도입되면서 설명 가능한 인공지능(XAI) 연구가 다시 주목

- 미(美) 국방성 산하 국방위고등연구계획국(DARPA)에서는 ‘17년부터 XAI 관련 프로젝트(XAI 학습 모델 개발 및 테스트)를 추진[1)]

□ **(개념)** XAI는 사용자가 인공지능 시스템의 동작과 최종 결과를 이해하고 올바르게 해석하여 결과물이 생성되는 과정을 설명 가능하도록 해주는 기술을 의미

○ 예를 들어 인공지능 시스템이 고양이 이미지를 분류할 경우, 기존 시스템은 입력된 이미지의 고양이 여부만을 도출하지만,

○ XAI는 고양이 여부를 도출하고, 이것의 근거(털, 수염 등)까지 사용자에게 제공(그림 1)



<그림 1. 설명 가능한 인공지능(XAI) 예시>

1) DARPA, Explainable Artificial Intelligence(XAI) DARPA-BAA-16-53, 2016.8.10.

3 XAI를 위한 기술적 접근방법

XAI를 위한 기술적 접근방법으로 **(가)기존 학습 모델 변형, (나)새로운 학습 모델 개발, (다)학습 모델 간 비교**에 기반을 둔 방법이 존재[2)3)]

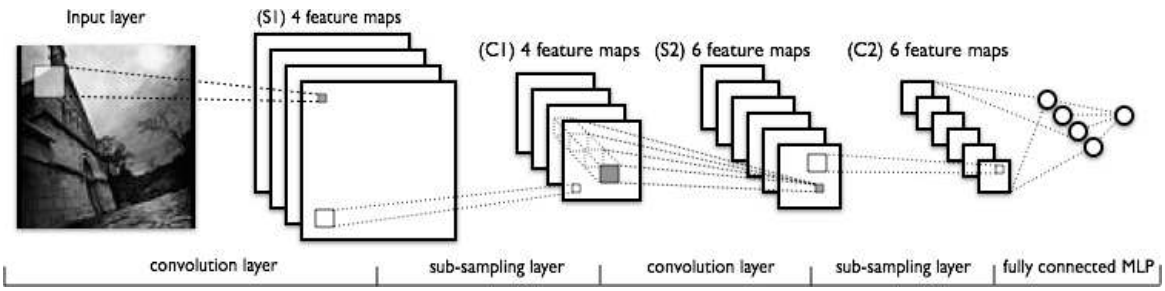
(가) 기존 학습 모델 변형

□ 기존 학습 모델 변형 방식은 기존 학습 모델에 역산 과정을 추가하거나 학습 모델을 수정

○ 예를 들어 합성곱 신경망의 결과 설명을 위해 역합성곱 신경망(Deconvolutional Network)을 추가한 방식은 다음과 같음

<설명 가능한 합성곱 신경망 예시>

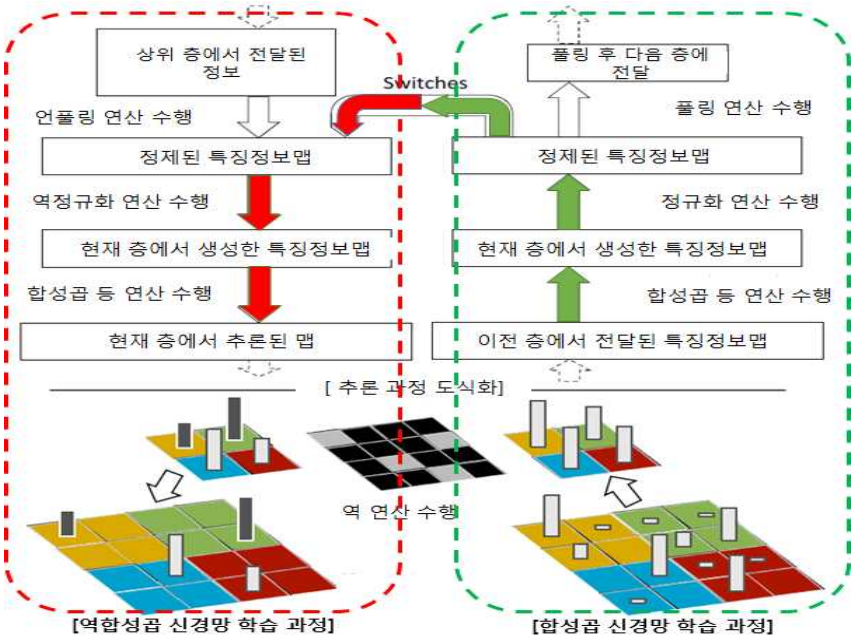
역합성곱 신경망(Deconvolutional Network)을 통한 학습 모델의 시각화 연구 사례[4)]
- **(개요)** 합성곱 신경망의 학습 과정을 역산하는 신경망을 구성함으로써 최종 결과에 영향을 미치는 요소들을 추론 및 시각화 - **(이미지 분류 예시)** 합성곱 신경망은 여러 개의 층(layer)으로 이루어져 있으며, 각 층은 이전 층에서 생성된 특징정보맵(feature maps)을 이용하여 새로운 특징정보맵을 생성(합성곱 연산)하는 층과 이를 축소(풀링 연산)시키는 층 등으로 구성 - 입력된 이미지는 각 층과 최종 분류를 위한 단계(완전연결층(fully connected MLP))를 거쳐 분류(그림2)



<그림 2. 합성곱 신경망 예시>

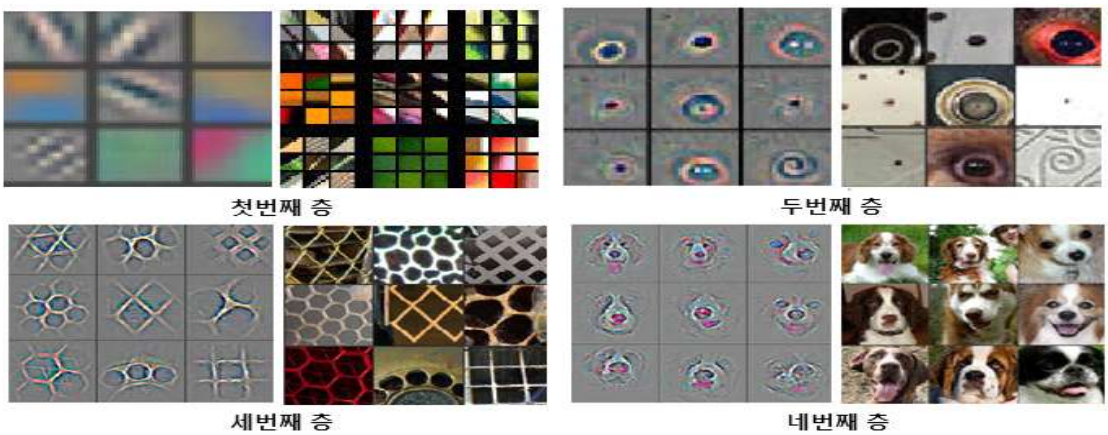
2) Biran, Cotton, Explanation and Justification in Machine Learning: A Survey, IJCAI, 2017
3) David Gunning, Explainable Artificial Intelligence(XAI), DARPA, 2016.

- 역합성곱 신경망은 이러한 학습 과정을 역으로 계산하면서 각 층에서 특징정보맵 생성시 영향을 미치는 요소를 추론하고 이를 시각화(영향을 미치는 요소란 이미지 분류에 영향을 미치게 되는 것을 의미하며, 이를 최종 분류 결과의 근거로 사용 가능)(그림3)



<그림 3. 역합성곱 신경망 구조 예시>

- 그림 4와 같이 역합성곱 과정을 거친 후, 특징정보맵에 영향을 미친 요소들을 시각화하고 실제 이미지와 매칭함으로써 이미지 분류 근거를 확인



<그림 4. 각 층의 특징정보맵 생성에 영향을 미치는 요인(왼쪽 회색)과
실제 이미지에 매칭된 부분(오른쪽 이미지)을 시각화한 예시>

4) Zeiler, M. D., & Fergus, R.Visualizing and understanding convolutional networks. In European conference on computer vision, 2014.

(나) 새로운 학습 모델 개발

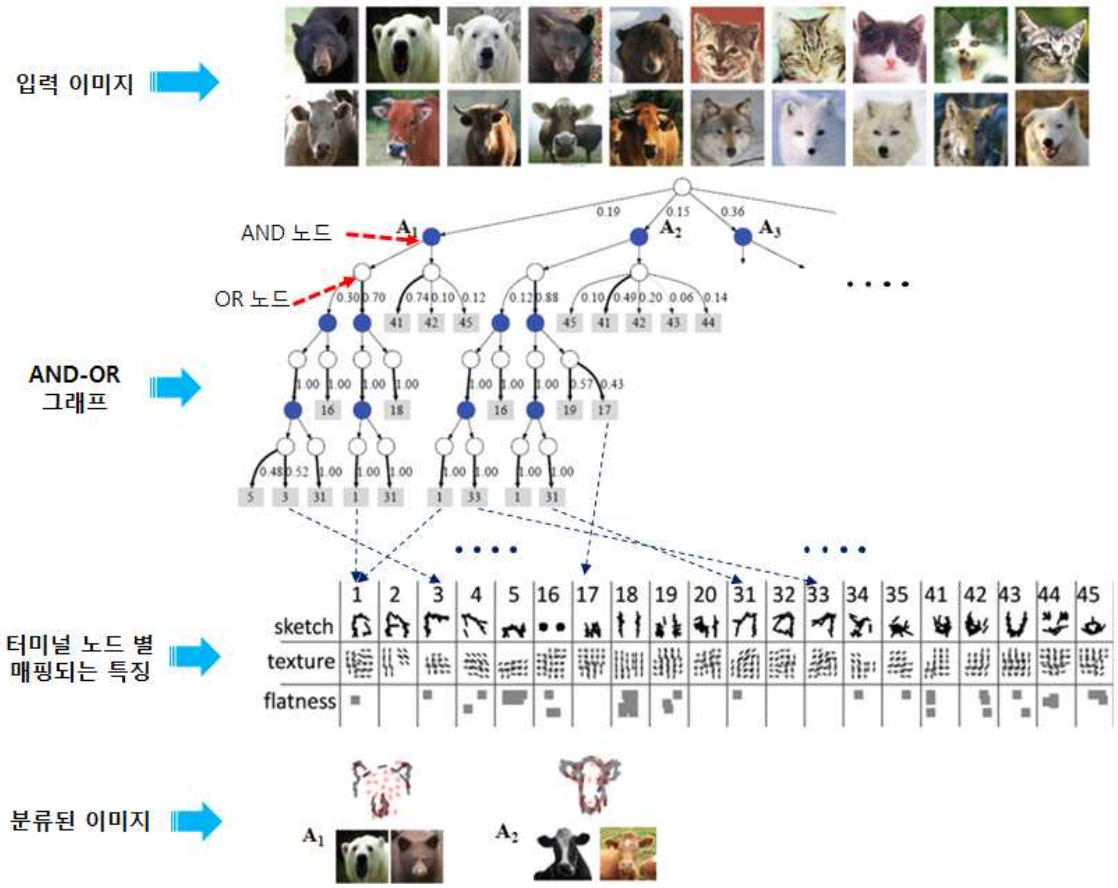
□ 새로운 학습 모델은 원인-결과와 같은 도출 과정이 표현 가능한 학습 모델을 새로이 만드는 것을 의미

○ 예를 들어 학습 모델 설명을 위해 개발된 확률적 그래프 기반 분류 모델은 다음과 같음

<설명을 위해 새로이 개발된 학습 모델 예시>

확률적 AND-OR 그래프 기반의 해석 가능한 분류 학습 모델 연구 사례[5]

- **(개요)** 입력(이미지, 텍스트 등) 데이터의 특징(경곗값, 텍스쳐, 색 등)을 관계 그래프로 생성하여 분류 결과에 연결된 노드로부터 분류 근거를 확인
- **(이미지 분류 예시)** 입력 이미지들과 이미지의 특징(스케치, 색, 텍스쳐, 주요 객체 위치 등)을 AND-OR 그래프로 표현(그림 5)



<그림 5. AND-OR 그래프를 이용한 이미지 분류>

= AND 노드는 입력 이미지에 표현된 객체를 의미하며, 그래프의 상위로 갈수록 큰 개념의 객체(곰, 공작새 등)를 표현하고, 하위로 갈수록 세분화된 객체(귀, 꼬리, 발 등)를 표현

= OR 노드는 AND 노드를 연결하는 노드로, 의미적 관계를 표현(예: '공작새 AND 노드'의 하위 노드로 '날개 AND 노드'가 있을 때, OR 노드가 이들을 연결하고, 이에 대한 전이 확률을 표시)

- AND-OR 그래프를 통해 터미널 노드에 도달하기까지 거친 노드들을 분석할 수 있으며, 분류된 이미지와 매핑되는 특징(스케치, 색, 텍스쳐, 주요 객체 위치 등)을 식별하여 설명 가능

(다) 학습 모델 간 비교

□ 학습 모델 간 비교 방법은 설명하려는 학습 모델에 대한 세부 지식 없이 설명 가능한 타 모델과의 비교로 최종 결과를 설명하는 것을 의미

○ 예를 들어 이미지 분류 모델의 설명을 위해 설명 가능한 학습 모델과 픽셀을 비교한 분류 모델은 다음과 같음

<학습 모델 간 비교로 학습 결과의 설명 예시>

학습 모델 비교를 통한 범용적 분류 모델 연구 사례[6]
- **(개요)** 설명 가능한 다른 분류 모델과의 상호 대조 및 추론으로 타깃 분류 모델(설명이 필요한 모델)의 최종 결과를 설명하는 기술 - **(이미지 분류 예시)** 설명 가능한 분류 모델과 타깃 분류 모델의 결과를 서로 비교하여 설명 가능한 분류 모델의 근거를 타깃 분류 모델에 적용 = 설명 가능한 분류 모델이 그림 6(a)을 전자 기타로 분류할 때, 설명 가능한 분류 모델에 사용된 픽셀(전자기타의 넥(neck) 부분 값)을 타깃 분류 모델에 적용하여 유사한 결과가 도출되면 설명 가능한 분류 모델의 근거를 타깃 분류 모델의 결과 도출 근거로 활용 가능

5) Si, Z., & Zhu, S. C. Learning and-or templates for object recognition and detection. IEEE transactions on pattern analysis and machine intelligence, 2013.


(a) 입력 이미지 (b) 전자 기타

<그림 6. 원본 이미지와 전자 기타 이미지>

- 동 연구에서는 이미지가 하나의 클래스(예: 전자 기타 등)로 분류되는 것을 설명하는데 약 10분 정도가 소요된다고 언급.
  상당히 오랜 시간이 필요하지만 분류 모델 설명을 위한 범용적 모델 진단 기술로 특정 알고리즘에 한정되지 않고 사용 가능한 것에서 의미가 있음

4 결론 및 시사점

□ 인공지능 기술이 다가올 미래에 핵심기술로 인식되고 있지만, 일각에서는 인공지능과 같이 빅데이터를 활용하는 기술로 인해 발생 가능한 사회의 차별, 불평등 등을 우려[7]

□ 이러한 우려 속에서, XAI는 다양한 분야(금융, 보험 등)의 인공지능 시스템이 사용자와 고객으로부터 신뢰를 얻고, 사회적 수용을 위한 공감대 형성 방안이 될 것으로 예상

○ 앞서 살펴본 바와 같이, XAI를 위한 기술적 접근방법으로 **1) 기존 학습 모델을 설명 가능하도록 변형**하거나, **2)새로운 학습 모델**이 개발될 수 있으며, **3)학습 모델 간의 비교**를 통한 방법이 존재

6) Ribeiro, M. T., Singh, S., & Guestrin, C. Why should i trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ACM, 2016.

7) Thelisson, E., Padh, K., & Celis, L. E. Regulatory Mechanisms and Algorithms towards Trust in AI/ML., IJCAI, 2017.

□ 또한 XAI를 활용함으로써 **1)인공지능 시스템의 성능 향상, 2)통찰력 습득, 3)법적 책임 및 준수 확인 등의 효과가 기대**됨[8)9)]

○ (성능 향상) 학습 모델의 편향 등 시스템의 성능저하 요인을 파악하고, 동일한 목적과 결과를 갖는 학습 모델 간 비교로 적합한 학습 모델을 도출함으로써 성능 향상 가능

○ (통찰력 습득) 학습 과정 중 빅데이터로부터 다양한 패턴을 추출·분석하여 드러나지 않았던 법칙, 전략 등을 도출 가능

- (예시) 인공지능 바둑 기사 알파고의 수(手)를 바둑 전문가들이 연구하여 새로운 전략 도출

○ (법적 책임 및 준수 확인) 인공지능 시스템의 잘못된 결과로 분쟁 발생시 원인파악이 가능하고, GDPR(유럽연합(EU) 개인정보보호 규정)과 같은 규정 준수 여부 검증 등이 가능

□ 인공지능 시스템을 개발 및 운영하는 담당자들은 아직 연구단계에 있는 XAI 관련 기술 등을 지속적으로 모니터링하는 것이 필요

8) Fox, Maria, Derek Long, and Daniele Magazzeni., Explainable Planning, IJCAI, 2017.
9) Samek, W., Wiegand, T., & Muller, K. R., Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models, 2017.