04_XAI Evaluation

1. Evaluation of Explanations

What is an explanation in XAI?

XAI의 설명에 대한 평가 이전에, XAI의 설명에 대한 정의가 필요하다.

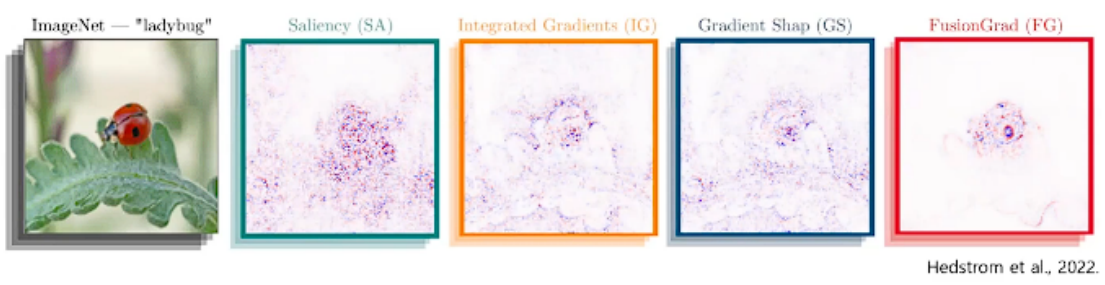
XAI의 설명이란 Agent, 예측 모델의 행동을 사람이 이해할 수 있게 하는 것을
XAI의 설명이라고 할 수 있다.

더 구체적으로 정의해보자면, 특정 결과에 대한 입력의 특징이나 관련성을 서술하기 위해
예측 모델 이외의 외부 알고리즘(XAI 알고리즘)으로부터 추가적인 정보를 생성하는 것을 설명이라고 정의할 수 있다.

Which explanation is better?

아래의 그림은 무당 벌레라는 클래스에 대한 샘플 이미지 각 픽셀의 특성과 관련성을 설명하기 위한 XAI 알고리즘의 결과이다.

아래의 4가지 XAI 알고리즘 중 어떤 것이 가장 적절할까?



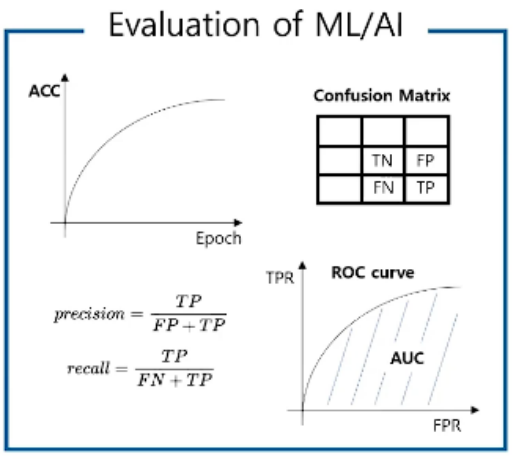
"Based on what grounds or criteria did you make your choice?"

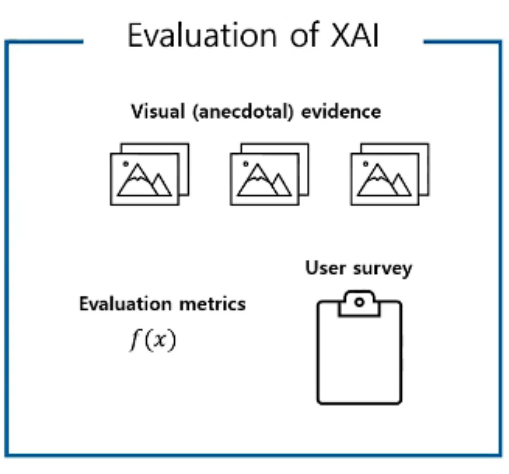
Evaluation of ML/AI and XAI

ML/AI 모델의 경우 정확도가 있고, 분류 모델의 경우 Confusion matrix, precision, recall, ROC curve 등등 일반적으로 표준화된 평가 지표가 존재한다.

그러나, XAI의 경우 특별히 합의된, 표준적인 평가 지표는 존재하지 않는다.
설명 가능 인공지능의 설명의 종류와 목적이 다양함에 따라 표준 평가 지표를 정의하기가 어렵기 때문이다.

일반적으로 사용되고 있는 XAI 평가 방법으로 시각적 사례 증거, 목적에 따른 평가 함수 선택,
응용 분야가 명확할 경우 설문을 통해서도 검증하기도 한다.



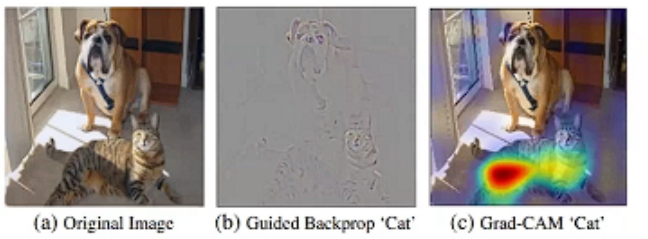


2. Types of Explanations

다양한 종류의 설명을 예시를 위주로 살펴보자.

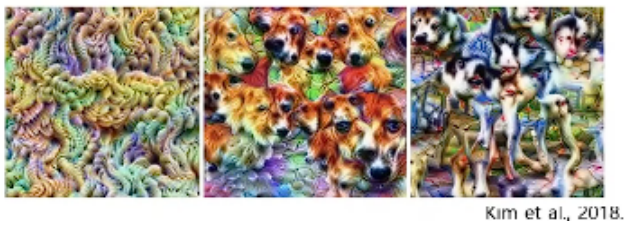
1. **Feature importance, Heatmap**

- Feature importance는 feature relevance score를 1차원 벡터 표현
- Heatmap은 2차원 이상이라고 생각하면 된다


(a) Original Image (b) Guided Backprop 'Cat' (c) Grad-CAM 'Cat'
Selvaraju et al., 2017

2. **Prototypes**

- 특정 샘플에 대해 컨셉과 같은 대표성이 있는 부분에 기반하여 설명하는 방식


Kim et al., 2018.

3. **Localization**

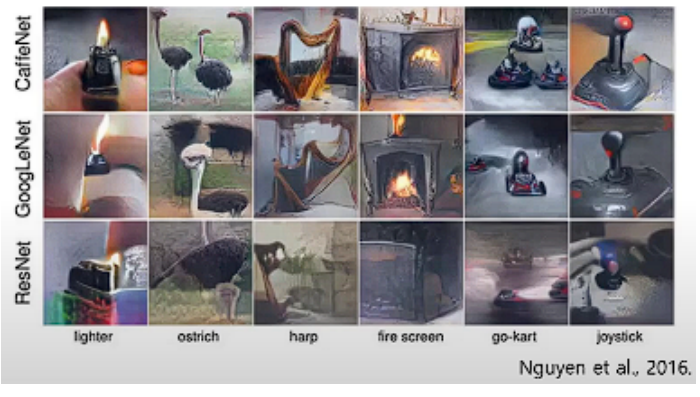
- binary한 feature importance로 생각하면 된다
- patch, segmentation, bounding bow 등등

Zhang et al., 2017

4. **Representation Synthesis**

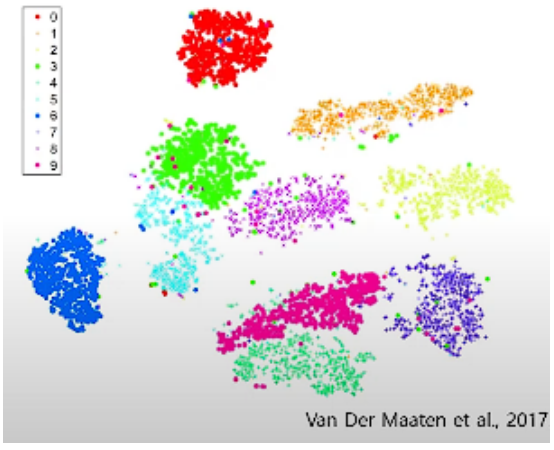
- 모델을 설명하기 위해서 이미지를 생성하여 시각화하는 방식
- Ex) Activation maximization : CNN의 특정 채널의 activation을 최대화 하는 패턴이 존재한다면, 그 패턴이 해당 채널이 학습한 feature의 패턴이라고 보고, 그 패턴을

찾기 위하여 noise input을 주어서 target 채널의 activation이 최대화 될 때까지
noise input을 업데이트 하는 방식



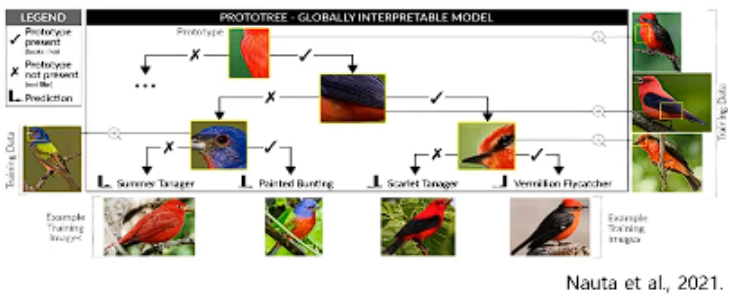
5. **Representation Visualization**

- 차트나 plot으로 예측 모델을 설명하는 방식
- PCA, t-SNE 등의 차원 축소를 통한 방식이 대표적
- 아래의 plot은 MNIST 데이터의 샘플 이미지를 차원 축소해서 나타낸 것, label 별로 군집화 되어있는 것을 알 수 있음

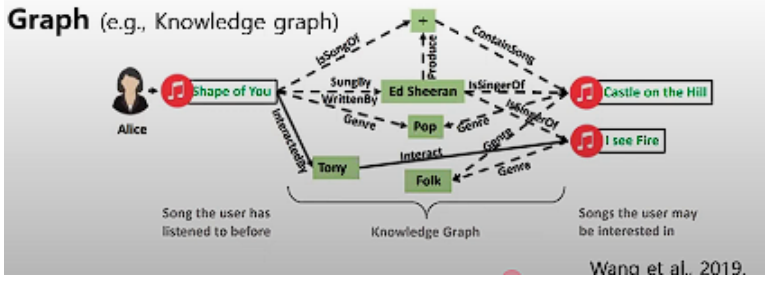


6. **Decision Tree**

- 이미지의 각 파트가 노드로 구성되어 조건문을 통과하도록 설계된 방식



7. **Graph**



8. **Text explanation**

Question:	What could people do that involves talking?
Choices:	**confession**, carnival, state park
CoS-E:	confession is the only vocal action.
Reason	people talk to each other
Rationale:	people talk to people
Question:	A child wants to play, what would they likely want?
Choices:	**play tag**, breathe, fall down
CoS-E:	A child to play tag
Reason	Children want to play tag, and they want to play tag with their friends.
Rationale:	Children want to play tag, what would they want to do?
Question:	They were getting ready for a really long hike, he put the food in his what?
Choices:	recycling center, house, **backpack**
CoS-E:	Backpacks are used on hikes
Reason	a backpack is a place to store food and supplies.
Rationale:	a backpack is used to carry food and supplies
Question:	You can do knitting to get the feeling of what?
Choices:	**relaxation**, your, arthritis
CoS-E:	Your are focusing on a repetitive task.
Reason	knitting is the only thing that is relaxing.
Rationale:	you can do knitting to get the feeling of what?

Rajani et al., 2019.

3. Properties of XAI evaluation

Faithfulness / Correctness

신뢰성/정확성, 설명이 예측 모델과 얼마나 정확하게 일치하는지 평가하는 기준

1. Incremental Deletion or Addition

- 특징을 하나씩 삭제하거나 추가하면서 예측 결과가 어떻게 변하는지 관찰한다.
- Ex) 특정 이미지 부분을 가려가면서 예측 결과가 바뀌는지 확인(Saliency map, Feature importance 활용)

2. Controlled Synthetic Data Check

- 특정 규칙을 따르는 합성 데이터를 생성하고, 모델이 그 규칙을 따르는지 평가한다.
- Ex) 모델이 특정 개념을 학습했다고 가정했을 때, 해당 개념을 포함한 합성 데이터를 넣어보고 예측이 기대와 일치하는지 확인

Complexity / Compactness

복잡성/간결성, 설명은 간결해야 하지만, 충분한 정보를 제공해야 한다

1. Size of Explanation

- 설명이 너무 길지 않고 적절한 크기를 유지해야 한다.
- Ex) Heat map, Decision tree, 텍스트 요약 등 사용

2. Counterfactual Compactness

- 예측을 변경하기 위해 입력에서 최소한으로 바꿔야 하는 부분을 확인한다.
- Ex) 텍스트 분류 모델에서 단어 하나만 바꿔도 결과가 바뀐다면, 그 단어가 중요한 요소

Completeness

완결성, 설명이 모델이 어떻게 동작하는지를 충분히 설명하는가?

1. Preservation or Deletion Check

- 중요한 특징만 유지한 상태에서 모델이 같은 결정을 내리는지 확인한다.
- Ex) 이미지에서 중요 부분만 남기고 예측 수행 → 같은 결과 나오면 해당 부분이 결정적으로 중요한 부분인 것이다.

Robustness / Continuity

강건성/연속성, 작은 변화가 있을 때 설명이 크게 변하지 않는가?

1. Stability for Slight Variations

- 유사한 입력에서 설명이 크게 달라지지 않는지 확인한다.
- Ex) 같은 물체가 약간 다른 각도에서 촬영된 경우, 히트맵이 크게 변하면 신뢰도가 낮다.

2. Connectedness

- 반사실적 예제(Counterfactual)가 기존 훈련 데이터와 유사한지 확인한다.
- Ex) 특정 특징이 추가되었을 때 결과가 바뀌더라도, 기존 데이터 분포 내에 존재하는지 확인.

Contrastivity

대비성(비교가능성), 설명이 비교 대상(다른 클래스, 상황 등)과 차이를 잘 보여주는가?

1. Target Discriminativeness

- 모델의 설명이 특정 목표를 예측하는 데 도움이 되는지 평가한다.
- Ex) 텍스트 분류에서 “이 기사가 스포츠 기사인 이유”와 “이 기사가 정치 기사인 이유”를 명확히 구별할 수 있는가?

2. Data Randomization Check

- 라벨을 랜덤하게 바꿔서 모델을 학습한 후, 기존 모델과 설명이 얼마나 다른지 비교한다.
- Ex) 훈련 데이터의 레이블을 섞었을 때도 같은 패턴의 설명이 나오면, 설명이 신뢰할 수 없다는 의미.

Covariate Complexity

공변량 복잡성, 설명에서 사용되는 특징이 이해 가능해야 한다.

1. Covariate Homogeneity

- 특정 특징이 항상 일정한 방식으로 해석되는지 평가한다.
- Ex) 히트맵이 특정 개념에 대해 항상 일관되게 반응하는지 확인.

Coherence

일관성, 논리적 정합성, 설명이 사용자의 기대와 논리적으로 일치하는가?

1. Alignment with Domain Knowledge

- 모델의 설명이 실제 전문가의 지식과 일치하는지 확인한다.
- Ex) AI 의사가 “폐암”을 진단할 때, 설명이 실제 의사의 판단과 유사한 방식으로 제시되는지 검토.

평가 기준	설명
Faithfulness / Correctness (신뢰성 & 정확성)	설명이 모델의 실제 동작과 얼마나 일치하는가?
Complexity / Compactness (복잡성 & 간결성)	설명이 너무 길거나 과도하지 않은가?
Completeness (완전성)	설명이 모델의 결정을 충분히 설명하는가?
Robustness / Continuity (강건성 & 연속성)	작은 입력 변화에 대해 설명이 안정적인가?
Contrastivity (대비성)	설명이 다른 클래스나 상황과 비교하여 차이를 명확히 보여주는가?
Covariate Complexity (공변량 복잡성)	설명에 사용된 특징이 사람에게 이해 가능한가?
Coherence (일관성)	설명이 사용자나 도메인 전문가의 기대와 일치하는가?

---

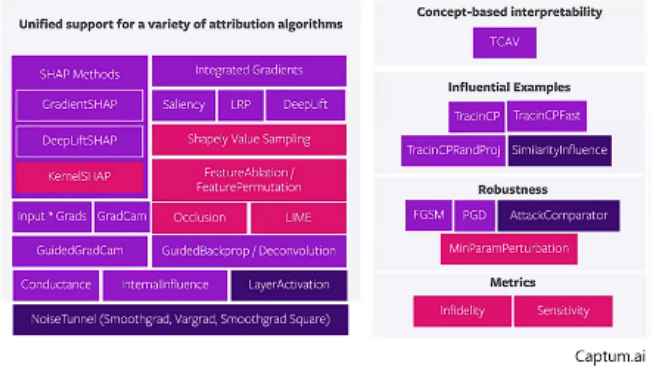
4. Tools for XAI evaluation

XAI 평가 메트릭을 지원하는 몇 가지 오픈소스 Toolkit에 대해 알아보자.

1. Captum

Captum

- An XAI package for *Pytorch*.
- Mainly supports attribution-based XAI methods.
- Supports two evaluation metrics.

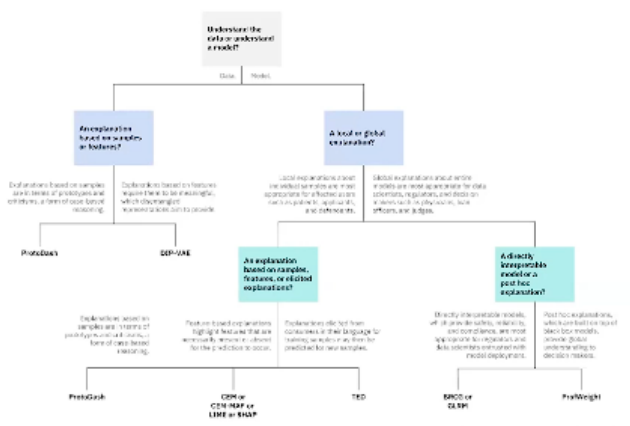


Captum.ai

2. AIX360

AIX360

- An XAI toolkit.
- Supports both local and global explanations for tabular, text, images, and time-series data.
- Supports two evaluation metrics.


IBM Research AIX 360

3. Quantus

Quantus

- A toolkit to evaluate XAI.
- Supports 35+ metrics in 6 categories for XAI evaluation.
- Supports images, time-series, and tabular data.
- Support *Pytorch* and *Tensorflow* models.