[DSBA] 1. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey

문서 유형	[XAI]
상태	진행 중
작성자	
출처	

1. Why do we need to “Explain” ?

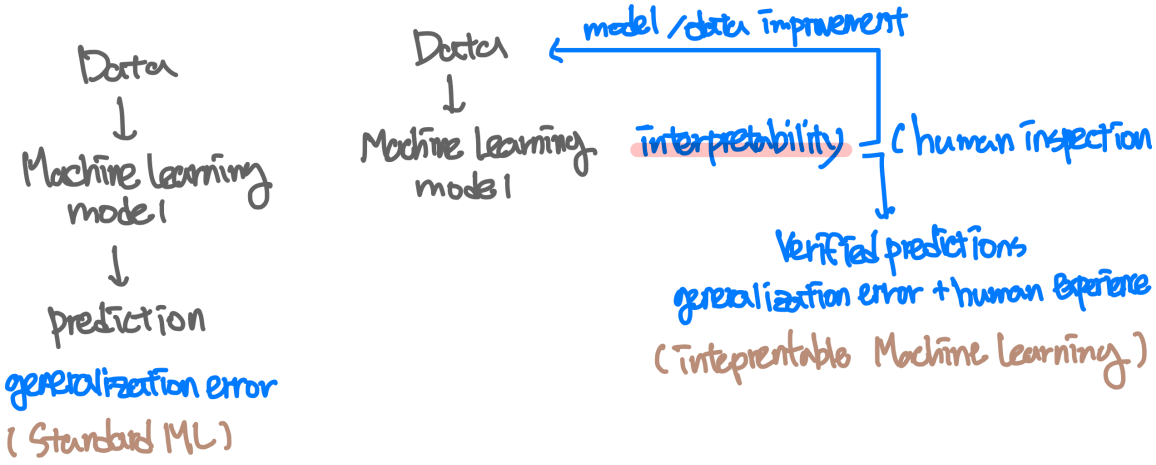
▼ XAI란 인간의 Explanation이 아닌 **AI가 Explanation을 도출하며, 사람이 AI의 동작과 최종 결과를 이해하고 올바르게 해석할 수 있고, 결과물이 생성되는 과정을 설명 가능하도록 함**

1) Improves transparency using human understandable justifications to decisions

2) Improves trust by improving confidence in decision making

3) Improves bias understanding and fairness by global model and behavior

4) **Improve models and verify the model**

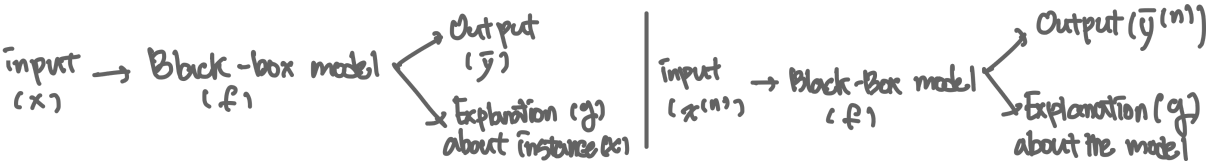


[그림 1] Diff. between General AI / XAI model architecture

2. Where is the XAI method forcusing on ?

2.1 Scope

- **local : individual data를 설명함** (→ $x \in X$에 대해 하나의 explanation map 생성)
- **Global : model 전체를 하나의 단위로 인식하여 설명함** (→ $x \in X$들의 group에 대해 하나의 explanation map 생성)



[그림 2] Diff. between model architecture (scope)

▼ **Local Explainable Artificial Intelligence (XAI)**

(1) Activation Maximization

CNN의 각 layer의 중요성을 알기 위해 **특정 hidden unit activation을 최대화하는 input 패턴을 탐지**

$x^* = argmax_{x_{s.t}||x||=p}\ z_{ij}\ (\theta,\ x)$

→ **특정 Input 패턴과 correlated된 layer filter activation maximization loss를 최적화 목적**

이를통해 input instance의 layer별 feature importance를 찾을 수 있음

(2) Saliency Map Visualization

각 output class에 해당하는 gradient를 픽셀별로 계산하여 각 픽셀의 importance 도출

$S_c(I) = W^T x + b$

→ **픽셀별로 gradient를 계산했을때 positive인 gradient 일수록 결과에 영향을 주는 픽셀로 가정함**

이를 통해 positive gradient를 갖는 픽셀로 saliency map을 구할수 있음

(3) LRP (Layer-wise Relevance Back-propagation)

output prediction을 back-propagation을 통해 분해하여 input data의 각 feature의 relevance score를 계산

$y_j = \sum_i w_{ij} x_i + b_j$, $z_j = f(y_j)$

$R_{i \leftarrow j} = R(z_j) \frac{x_j w_{ij}}{y_j + \epsilon \, sign(y_j)}$, $R(x) = \sum_j R_{i \leftarrow j}$

▼ **Global Explainable Artificial Intelligence (XAI)**

복잡한 Deep model을 linear counterparts로 축소하여 해석하기 쉬운 형태로 만듦
(Rule-base, tree-base model, ... etc.)

(1) Class model Visualization

Activation maximization을 global explanation으로 확장한 형태

$l' = \ argmax_l \, Sc(l) \ - \lambda ||l||_2^2$

(2) CAVs (Concept Activation Vectors)

positive concept (사람이 구분할수 있는 컨셉 : p_c) 과 negative를 구분하는 방법론

directional derivatives를 활용, class prediction의 sensitivity **(input의 특정 p_c로 가까워 질수록 드러나는 layer의 변화) 를 평가하는 방법 (TCAVs)**

$$S_{c,k,j}(x) = \lim_{\epsilon \to 0} \frac{h_{j,k}(z_j(x) + \epsilon v_i^c) - h_{j,k}(z_j(x))}{\epsilon} = \nabla h_{j,k}(z_j(x)) \cdot v_i^c$$

$$TCAV_{Q\,c,k,j} = \frac{|\{x \in X_k : S_{c,k,j}(x) > 0\}|}{|X_k|}$$

(3) SpRAy (Spectral Relevance Analysis)

군집화 한 후 relevant 한 입력값 도출하는 방법론으로 **LRP를 통해 각 입력값의 relevant score를 구한후 해당 score 기반으로 입력값들을 군집화** (spectral clustering)

(가장 relevant한 cluster를 찾기 위해 eigenvalue간의 차이를 기반으로 함)

2.2 Methodology

- **Gradient based (Back-propagation based) :** Back-propagation된 gradient를 통해 neuronal influence나 input / output의 relevance를 파악
- **Perturbation based** : 입력데이터의 feature 변화에 따른 input / output relevance를 파악

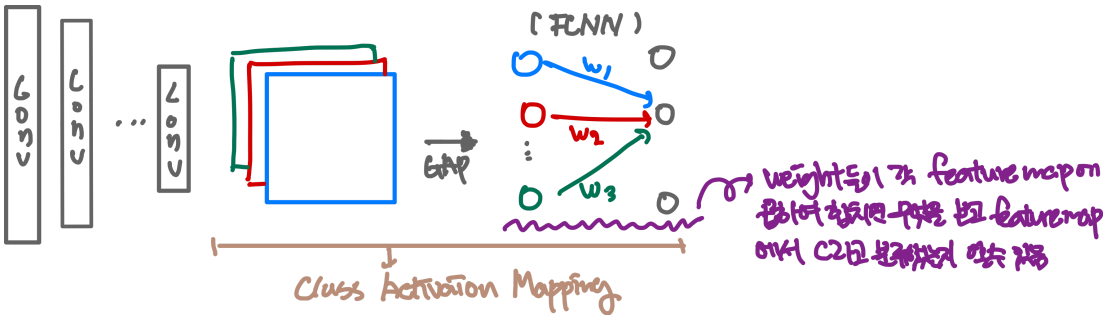
▼ **Gradient based Explainable Artificial Intelligence (XAI)**

(1) Saliency maps

gradients of the output over the input으로 인해 만들어짐 **즉, 이미지 안에 어떤 부분이 classification에 relevance한지 highlight**

(2) GradCAM (Gradient class Activation Mapping)

Convolution layer를 쌓고 FCNN 대신 GAP(Global Average Pooling)을 사용하여 위치정보 보전



[그림 3] GradCAM architecture

(3) Salient Relevance Maps

background layer와 foreground layer의 차이를 구하여 context aware한 information (canny-edge based detector를 Salient Relevance map에 추가)을 찾음

(즉, 픽셀별로 saliency value를 찾는데 인접 픽셀이 distinct하고, 서로다른 스케일의 pixel patch들과 다르면 salient)

(4) Attribute Maps

IG(Integrated Gradients)를 주장하며, **input data x가 model f에 가한 attribution은 gradients 들의 integral을 통해 얻을수 있다고 가정**함 또한, 도메인 지식을 학습단계에 추가하여 모델을 해석

$IG_j(x,\ x')$ = $(x_j$ - $x'_j) * \int_{\alpha=0}^{l} \frac{\partial F(x'+\alpha*(x-x'))}{\partial x_j}\ \partial\alpha$ (x : input data, x' : base-line data)

$$EG(x) = \int_{x'} (x_j - x'_j) \int_{\alpha=0}^{l} \frac{\delta F(x'+\alpha*(x-x'))}{\delta x_j} \delta\alpha) * P_{D(x)}(x')\delta x'$$

$$= E[(x_j - x'_j)\frac{\delta f(x'+\alpha*(x-x'))}{\delta x_j}]$$

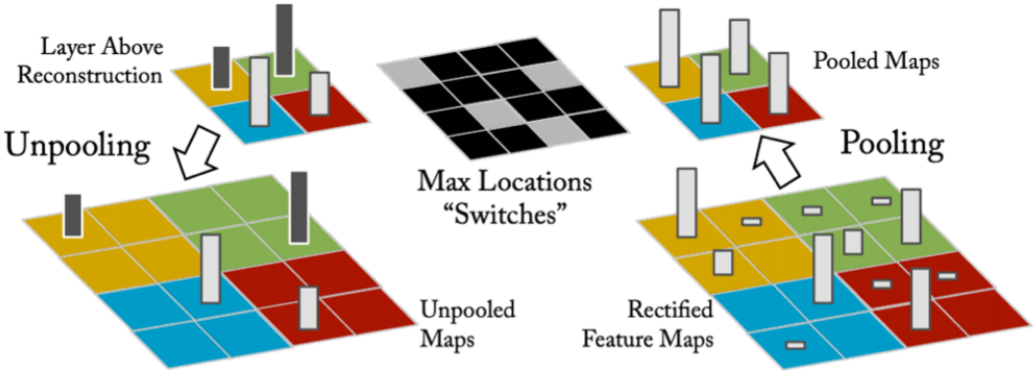
(Integration Over the whole training distribution이 intractable 하기 때문에 integral 대신에 Expectation 사용)

▼ **Perturbation based Explainable Artificial Intelligence (XAI)**

(1) DeConvolution Nets for Convolution Visualization

CNN의 정반대 방향으로 진행되는 Network을 사용하여 Activation Map을 생성함

(Activation Map : input pixel space로 매핑하는 역할로, 각 layer가 무엇을 어떻게 학습하였는지 알수 있음)



[그림 4] DeConvolution Nets architecture

즉, 입력 이미지로 부터 대략적인 Approximate version of convolution feature를 생성하여, feature상의 변화를 de-convolution함으로 변화가 어떻게 감지 / 학습 되는지 파악 가능

(2) RISE (Randomized Input Sampling for Explanation)

input 이미지에 random mask를 곱하여 perturbation을 가한 후 prediction 수행하는데 이때 **classification에 대해 confidence score를 return, mask와 confidence score을 가중합**하여 **saliency map을 구함**

3. Metric of Explainable Artificial Intelligence (XAI)

- **SCS (System Causality Scale)**

  'User facing Human-AI, Machine-interfase' 에 대한 설명을 위해 고안됨 (척도로된 평가표)

▼ **BAM (Benchmarking Attribution Methods)**

모델별로, 같은 모델 내에서 입력별로, functionality equivalent input에 따라 explainer의 성능 추정

(1) **Model Constrast scores** : object로 학습된 모델과 scene label로 학습된 모델의 attribution 차이

(2) **Input dependence rate** : scene only image로 train된 경우 input dependence는 추가된 물체가 loss importance라고 attributed된 region을 리턴하는 비율을 측정

(3) **Input independence rate :** scene-only로만 train된 model이 주어졌을때 input independence는 기능적으로 크게 중요하지 않은 patch가 붙어 있는 이미지가 크게 explanation에 영향을 미치지 않는 비율

- **Faithfulness**

  importance score of feature와 performance effect of each feature간의 상관 관계

- **Monotonicity**

  feature importance에 따라 model에 feature를 점진적으로 더해가며, 각 data feature의 성능에 기여한 정도를 지표화

- **Human-grounded evaluation Benchmark**

  사람의 평가 방식을 benchmark삼아 설명 모델을 평가하는 방식으로, 다양한 의견을 수집하여 최대한 human bias를 제거하고자 함

4. Limitation of Explainable Artificial Intelligence (XAI)

- **사람이 XAI의 Explanation maps를 추론하기 힘듦**
- **XAI에 대한 정량적인 평가가 어려움**
- **input layer에 perturbation을 가하면 prediction accuracy 보다 feature importance map이 확연하게 변화**되는 한계

Reference

[1] **Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey**

[2] [DSBA] XAI Review - 1. Overview