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

[**https://arxiv.org/pdf/2006.11371.pdf**](https://arxiv.org/pdf/2006.11371.pdf)

**0. NOTATIONS AND DEFINITIONS**

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
| Interpretability | 알고리즘의 **동작 원리를 이해할 수 있는 충분한 expressive data**를 제공하는 것에 대한 feature 또는 quality |
| Interpretation | Complex domain에 대한 **단순화된 표현**  (머신러닝 모델에 의한 출력을 사람이 이해할 수 있는 적절한 개념으로 바꾸는 것 등) |
| Explanation | 머신러닝 **모델 자체 또는 external algorithm**에 의해 생성된, 입력 데이터의 출력 데이터에 대한 **feature importance 또는 relevance**를 설명하기 위한 추가적인 메타 정보 |
| White-box  Black-box | 딥러닝 모델 f에 대하여 **모델 파라미터와 모델 아키텍쳐 정보가 알려진** 경우의 모델  딥러닝 모델 f에 대하여 **모델 파라미터와 모델 아키텍쳐 정보가 알려지지 않은** 경우의 모델 |
| Transparent | 딥러닝 모델이 **사람이 충분히 이해할 수 있게 표현**된 것   * Transparancy는 알고리즘의 일부이거나 external한 수단을 사용 |
| Trustability | 주어진 모델이 다양한 실제 환경에서 **최종 사용자가 의도한 작업을 수행할 때의 confidence** |
| Bias | 딥러닝 알고리즘을 통해 학습된 모델의, **data subset에 대한 불균형한 weight, 편견**   * 데이터를 수집하는 사람의 bias, 머신러닝 알고리즘의 결점 등이 원인 |
| Fairness | **공정하며 판단만을 내리는** 것에 대한 모델의 quality   * 입력 데이터 분포에 어떤 favor도 들어가지 않아야 한다. |

**1. SCOPE OF EXPLANATION**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **[LOCAL EXPLANATIONS]**  일반적으로 1개의 input instance가 설명을 위해 사용된다.   * **1개의 input instance x**에 대하여 **모델 f의 결정에 대한 설명 g를 생성**하는 것을 목표로 한다.  |  |  | | --- | --- | | Activation Maximization | * 복잡한 feature를 학습하는 **첫 번째 레이어**에서는 CNN 모델의 **layer-wise feature importance를 interpret**하는 것이 간단하다. * CNN에서 **더 깊이 들어가면** 특정 prediction에 대해 **특정 layer의 중요도는 표현하기 어렵다.**   + Subsequent layer의 파라미터는 previous layer의 파라미터의 영향을 받으므로 | | Saliency Map Visualization | 입력 이미지에 대한 **출력 분류의 gradient를 계산**하는 방법   * Gradient를 시각화하여 **각 픽셀의 중요도가 fair summary**된다. * Class score function: | | LRP  (Layer-wise Relevance Backpropag-  ation) | **LRP (Layer-wise Relevance Backpropagation):** Deep Neural Network의 **output prediction을 decompose**하여 **개별적인 feature에 대한 relevance score**를 계산한다.   * Input instance **x**, linear output **y**, activation output **z**가 있는 간단한 신경망에서는 다음과 같다. * 이때 relevance of activation output 에 대하여 는 | | LIME  (Local Interpretable Model-Agnostic Explanations) | **LIME (Local Interpretable Model-Agnostic Explanations):** 원본 이미지에서 output class에 대한 **인접한 superpixel의 중요도**를 계산한다.   * Class output에 대한 **가장 큰 representation을 제공하는 superpixel**이 존재하는지 아닌지를 **binary vector** 을 이용하여 구한다. | |

**[NEXT PAGE]**

**[PREVIOUS PAGE]**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **[LOCAL EXPLANATIONS]**   |  |  | | --- | --- | | LIME  (Local Interpretable Model-Agnostic Explanations) | **Explanation complexity** , 2개의 instance x와 z 사이의 **proximity measure** , 에 의해 정의된 대략적인 **f의 faithfulness를** 라 하면, input data sample x에 대한, **LIME equation에 의한 explanation** 는 다음과 같다.   * 여기서 LIME 최적화는 locality-aware loss **를 model- agnostic way를 통해 최소화**하는 것이다. * 여기서 input instance에 대하여 **superpixel of information**을 찾는 방법으로 permute한다. (fake data)   + 이때 **original input과 새로운 fake data 사이에 class score가 어떻게 다른지** 알기 때문에 새로운 fake data에 대하여 예측할 수 있다.     **<single instance에서의 LIME 알고리즘>** | | SHAP  (SHapley Additive exPlanations) | **SHapley Additive exPlanations (SHAP):** Shapley value를 이용한, **model explainability를 위한, 게임 이론에 의한** optimal solution   * Data feature는 **tabular data에서의 개별적인 분류** 또는 **이미지에서의 superpixel group**으로, LIME에서와 비슷하다. * **g**가 ML model의 explanation model f이고 **이 coalition vector**이며 **M이 maximum coalition size**이고 **이 feature j에 대한 attribution**일 때, bias and individual feature contribution의 합은 다음과 같다. | |

**[NEXT PAGE]**

**[PREVIOUS PAGE]**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **[GLOBAL EXPLANATIONS]**  Globally explainable 방법들은 **입력 배열을 요약하여 black box model의 전반적인 특성을 설명**한다.   |  |  | | --- | --- | | Global Surrogate Models | **비선형 AI 모델의 예측을 approximate**하기 위한 방법으로 쓰일 수 있다.   * **XAI의 How**에 해당하는 부분을 답변한다. 즉 **“How generalized is my AI model?”, “How do variations of my AI model perform?”** 등. | | Class Model Visualization | **[LOCAL EXPLANATIONS]** 의 **Activation maximization**에서 **Class Model Visualization**을 이용하면 global method로 사용할 수 있다.   * **CNN f**와 **class of interest c**에 대하여 **c를 대표하는 image visualization I’**를 생성할 때 다음 수식을 이용한다. | | LIME | **[LOCAL EXPLANATIONS]** 의 LIME 모델은 **submodular pick 알고리즘 (SP-LIME)을 통해 확장**될 수 있으며, 이때 global correlation을 이해할 수 있다. | | CAVs (Concept Activation Vectors) | **신경망의 내부 상태를 사용자 친화적인 concept domain에서 interpret**할 수 있는 global explainability method   * 머신러닝 모델 **f(.)을 기저 벡터 에서 span된 vector space** 이라고 하면 사용자의 이해는 vector space 로 모델링될 수 있다. 즉 이 모델의 **explanation function은** 이다. * **TCAV (Testing with CAVs):** f의 특정 레이어 j에 대해, 주어진 input에서 **concept C 방향으로의 class 예측에서의 sensitivity**를 측정하기 위해, gradient based method와 유사한 **directional derivatives**를 이용한다.   + **Layer j, class k**에 대하여 directional derivative는   + **TCAV score는 C로 향하는 input의 영향력 계산**에 활용된다. | | SpRAy (Spectral Relevance Analysis) | **LRP에 대한 설명에 기반한 local instance의 top** 부분에 build된다.   * 모델의 의사결정 과정을 global하게 알기 위하여, author들이 **LRP의 local explanation에서 특정한 클러스터링 알고리즘을 설명**한다. | |

**[NEXT PAGE]**

**[PREVIOUS PAGE]**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **[GLOBAL EXPLANATIONS]**   |  |  | | --- | --- | | Global Attribution Mapping | **feature들이 well defined semantic**을 가지고 있을 때 **attribution들을 weighted conjoined rankings로 간주**할 수 있다.   * 이때 각 **feature는 rank vector** 가 된다. | | Neural Additive Models (NAMs) | **여러 개의 신경망을 학습**시키고, 이때 **각 신경망은 single input feature**에 해당한다.   * **Genralized additive models (GAM)**을 대신 이용할 수 있다.   + GAM은 다음과 같은 수식으로 나타내어질 수 있다. | |

**2. DIFFERENCES IN THE METHODOLOGY**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **[PERTURBATION-BASED]**   |  |  | | --- | --- | | DeConvolution nets for Convolution Visualizations | Randomized Input Sampling for Explanation (RISE) | |  |  |   이 외에도 **Prediction Difference Analysis, Randomization and Feature Testing** 등이 있다. |

**[NEXT PAGE]**

**[PREVIOUS PAGE]**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **[BACKPROPAGATION- OR GRADIENT-BASED]**  **Saliency Maps, Gradient class activation mapping (CAM), Salient Relevance (SR) Maps, Attribution Maps** 등이 있다.   |  |  | | --- | --- | | Gradient class activation mapping (CAM) 에서의 segmentation method | Algorithm of Salient Relevance (SR) | |  |  | | Integrated Gradients for Attribution Maps | Expected Gradients (EG) for Attribution Maps | | x의 model f에 대한 attribution은 **baseline x’에서 x까지의 선분상의 모든 점에서의 gradient를 적분**하여 얻는다. | **integrated gradient에서 baseline image의 영향을 제거**하기 위한 수식 |   **Gradient-based Method에 대한 요구 사항:**   |  |  | | --- | --- | | **Sensitivity** | 모든 input과 baseline에 대해서 단 **1개의 feature가 다를 때 예측이 다르다**면 그 feature는 **non-zero attribution**을 갖는다. | | **Implementation invariance** | 2개의 신경망이 **기능적으로 동일하고 그 출력이 모든 입력에 대하여 동일**하다면 (구현은 다를 수 있음) **attribution들은 이들 신경망에 대해서 항상 identical**해야 한다. (즉 attribution들이 다르면 출력이 달라야 함) | | **Completeness** | attribution들은 입력 이미지 **x**와 baseline 이미지 **x’**에 대해, **모델 함수 f의 출력값의 차이**가 되어야 한다. | | **Linearity** | 선형적으로 구성된 신경망 모델 에 대하여 이것이 **2개의 신경망 모델 과 의 선형 결합** 이면 **의 attribution들은 과 의 attribution에 대한, weight이 각각 와 인 weighted sum**일 것이다. | |

**3. MODEL USAGE OR IMPLEMENTATION LEVEL**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **[MODEL INTRINSIC]**  오른쪽 그림의 모델의 explainability는 **자기 자신인 블랙박스 모델 f와 분리될 수 없다.**   * 이때 모델 f는 **naturally explainable**하다.  |  |  | | --- | --- | | Trees and Rule-based Models | Decision Tree와 같은 **shallow rule-based model은 자체적으로 interpretable**하다.  **LIME, SHAP**과 같은 알고리즘들이 linear 또는 tree-based 모델을 사용한다. | | Generalized additive models (GAMs) | **Generalized additive models with pairwise interactions ()**를 통해 GAM의 interpretability를 유지하면서 정확도를 높일 수 있다. | | Sparse LDA and Discriminant Analysis | **Bayesian non-parametric model, Graph-Sparse LDA, Sparse Penalized Discriminant Analysis (SPDA)** 등 |   **[POST-HOC]**  오른쪽 그림의 **post-hoc model explainability 알고리즘이 모델 f에 적용**되었고, 이때 모델 f는 **explainable externally**이다. |

**4. EVALUATION METHODOLOGIES, ISSUES, AND FUTURE DIRECTIONS**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 일반적으로 각 explanation은 **실제 세계에서 사용되기 위하여** 다음의 constraint를 따라야 한다.   |  |  | | --- | --- | | **1. Identity or Invariance** | **Identical한 data instance**는 반드시 **identical한 attribution 또는 explanation**을 생성해야 한다. | | **2. Stability** | **같은 class c**에 속하는 data instance는 **비교할 수 있는 설명 g**를 생성해야 한다. | | **3. Consistency** | Data instance의 **단 1개의 feature의 변동**에 대해서 **어떤 것이 변화되었는지 설명**해야 한다. | | **4. Separability** | **서로 다른 population의 data instance**에 대해서는 **서로 다른 형태의 explanation**을 생성해야 한다. | | **5. Similarity** | Class에 무관하게 **서로 가까운 data instance**는 **비슷한 explanation**을 생성해야 한다. |   이외에도 **Implementation Constraints, Bias Detection**이 있다. |

**[NEXT PAGE]**

**[PREVIOUS PAGE]**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Evaluation Scheme은 다음과 같다.   |  |  | | --- | --- | | **System Causability Scale (SCS)** | Explanability method가 데이터를 자동으로 분석하고 인간과 상호작용하는 AI에 적용되는 만큼, **human-AI interface**에 대한 평가는 중요하다. | | **Benchmarking Attribution Methods (BAM)** | **Feature attribution의 정확도와 상대적인 중요도**를 평가한다. | | **Faithfulness and Monotonicity** | 각 feature의 **importance score**와 해당 feature에 대한 **performance effect** (예측 정확도) 간의 **correlation**을 평가한다. | | **Human-grounded Evaluation Benchmark** | XAI 알고리즘이 생성하는 **local explanation**을 평가한다. |   관련된 소프트웨어 패키지에는 **Interpret, IML, DeepExplain, DrWhy** 등이 있다.  **<XAI 시각화의 한계점>**   * 의사결정을 위한 **XAI explanation map을 추론**하기 위해 **인간이 개입할 수 없음** * Explanation map의 completeness와 correctness에 대한 **quantitative measure를 이용할 수 없음** * Input instance에서의 **small perturbation이 출력값을 크게 변화**시킬 수 있음 * 신경망에 대한 **explanation**은 **간단한 변환으로 쉽게 조작**할 수 있음 * **Gradient-based method**는 선천적으로 **모델과 데이터 생성 프로세스에 의존적**이다. |