**Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies**

<https://www.sciencedirect.com/science/article/pii/S0004370221000102>

1. **Introduction**

|  |
| --- |
| XAI는 다음과 같이 구분된다.   * 모델의 **작동 방법을 직접적으로 전달**하기 위한 **interpretability**를 발전시키기 위한 방법 * 모델이 **어떻게/왜 해당 prediction을 도출했는지(transparency)**에 대해 설명을 하기 위한 방법 |

1. **Related work: post-hoc explanation, trust in systems, and debugging tasks**

|  |
| --- |
| **< Twin-Systems Explanation Framework >**  딥러닝 모델이 **MNIST test image에서 misclassification을 발생**시켰을 때를 나타낸 그림으로, 6을 0으로 잘못 분류하였다.   * 이것은 네트워크의 feature-weights와 이것을 twinned k-NN에 적용해서 **training set에 있는 nearest neighbor**들을 추출한 이미지를 통해 설명된다. * 설명되는 이미지는 **6과 비슷하게 생긴 0 이미지를 이용해서 0이라는 prediction을 생성했다고 설명**한다.   **< Post-hoc explanation of XAI >**  몇 개의 최근의 XAI review들에서는 explanation, interpretability, transparency에 대한 정의가 불분명하다고 한다.   * 최근에는 이 구분이 interpretability를 **transparency (또는 simulatability)**와 **post-hoc interpretability**로 구분하는 것으로 이어진다.   Post-hoc explanation은 다음과 같이 구분되고, 따라서 post-hoc explainability는 **“explanation-by-justification”**이라고 할 수 있다.   * Textual explanations of system outputs * Visualizations of learned representations or models (saliency map) * Explanations-by-example |

1. **Twin systems: feature weighting to find post-hoc explanations**

|  |
| --- |
| 현재의 **user study는 twin-system을 평가**하며, 이것은 **CNN을 k-NN과 pairing**한 것이다.   * **Black-box classifier**의 **MNIST dataset**에 대한 prediction에 대해서 **factual한 post-hoc explanations by-example**을 찾는다.   여기서는 정밀한 계산 방법을 이용하여 **이전의 weighting 알고리즘인 COLE을 CNN에 적용**하려고 한다.   * Saliency map이나 휴리스틱이 불필요하다. * **기존 기술이 휴리스틱에 의존하는 Saliency map을 사용**하는데, 이 문제를 해결한다.   **< Contributions Oriented Local Explanations (COLE) >**  이전 연구에서 COLE이 여러 개의 **fully-connected layer로 구성된 CNN system에 성공적으로 가중치를 적용**할 수 있다는 것을 증명하였다.  COLE feature-weighting method는 **feature의 모델의 예측에 대한 기여는 explanatory case들을 찾기 위한 가장 좋은 source**라는 아이디어에서 기반한다.   * 이 방법은 **CNN function을 보다 해석 가능한 k-NN function으로 추상화**시켜서, **설명을 도출하기 위한 유사한 case**들을 보여준다.   Instance x에 대한 **contribution vector c**는 다음과 같다.  여기서는 **c가 linear regression model f의 실제 predictive logic을 더 잘 표현**할 수 있으므로 **x와 w보다는 c를 이용한 유사도 metric**을 이용하여 explanatory case들을 찾으려고 한다.   * Output logit은 이다.   **< COLE Hadamard Product (COLE-HP) >**  CNN은 일반적으로 입력 이미지의 tensor로부터 latent feature의 집합을 추출하는 **feature extractor network 와 classification network** 로 구성되어 있다. 따라서 전체 네트워크는 다음과 같다.  f의 구조가 layer X에 있는 원래의 latent representation과 output SoftMax layer만을 포함하는 간단한 linear architecture이고 **hidden layer가 없으므로**, COLE을 구현하고 feature contribution을 계산하는 것은 다음과 같다.  위 수식은 모든 training data로부터 얻어지고, **COLE Hadamard product를 구현하기 위한 k-NN classifier를 fitting**하는 데 사용된다. 이때 **k-NN classifier에는 Euclidean Distance**를 사용한다. |

1. **Twin systems: feature weighting to find post-hoc explanations (Cont.)**

|  |
| --- |
| **< Running COLE-HP on MNIST >**  MNIST dataset에 대해서 **Algorithm 1을 작동시켜서 post-hoc explanation을 생성**한다.  다음 그림은 **right classification 및 wrong classification**에 대한 system output을 설명하기 위한 **sample explanation**을 나타낸다.   * CNN은 숫자 6과 7에 대한 prediction은 맞았지만, 숫자 8과 9에 대한 prediction은 일부 부정확하다. * 여기서는 k-NN을 이용하여 3개의 explanatory example을 표시한다. |

1. **User tests on post-hoc explanations: an experimental paradigm**

|  |
| --- |
| **< The current paradigm : task context and variables tested >**  현재의 설명 패러다임은 debugging-a-classifier task로, 참여자들은 **컴퓨터 프로그램(CNN-CBR MNIST twin-system)이 생성한 prediction을 평가**하라는 요청을 받는다.   * 이들 classification의 correctness와 reasonableness를 평가한다. * 구체적으로는 “the program labelled the number this way because of what it learned from the human-labelled numbers it was shown”이라는 말을 들었다. |

1. **User tests on post-hoc explanations: an experimental paradigm (Cont.)**

|  |
| --- |
| **< The current paradigm : task context and variables tested >**  참여자는 시스템에 의해 생성된 서로 다른 25~30개의 classification을 평가하였다. 또한 **3가지의 key manipulation**이 explore되었는데, 각각 다음과 같다.   * The **presence/absence** of the explanations * The **presentation** of right/wrong classifications (Classification-Type) * **Relative proportions** of right/wrong classifications (Error-Rate)     모든 실험에서 **classification에 대한 explanation의 presence/absence를 조작**했으며, 다음 그림은 참여자에게 보이는 matched material들 중 일부를 보여준다. |

**3. User tests on post-hoc explanations: an experimental paradigm (Cont.)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **< Current paradigm : item-level and system-level measures >**  몇 가지의 서로 다른 measure들이 이와 같은 experimental paradigm에 사용되었다.   |  |  | | --- | --- | | **Item-level** measures | 시스템에 의해 생성된 **각각의 prediction을 measure**한다.   * 각각의 제시된 classification에 대한 correctness 및 reasonableness에 대한 평가 | | **System-level** measures | **시스템에 대한 전반적인 적합성**을 평가한다.   * 전체 시스템에 대한 사람들의 신뢰와 만족에 대한 평가 | |

1. **Mental models in XAI: taxonomies, frameworks, and theory**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **< A taxonomy of XAI tasks >**  Doshi-Velez and Kim은 XAI user study의 influential taxonomy를 제안했는데, 다음과 같은 **three-levelled taxonomy**를 제안했다.   |  |  | | --- | --- | | Application grounded evaluation | 인간 사용자가 완전한 AI 모델을 이용하여 task를 수행한다. | | Human-grounded evaluation | 인간 사용자가 **real application task를 선택적 관점**에서 수행한다. | | Functionally-grounded evaluation | 평가가 실제 사람의 평가에 대한 **computational-proxy**에 기반한다. | |

1. **Mental models in XAI: taxonomies, frameworks, and theory (Cont.)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **< Tricorn user model of XAI >**  Tricorn User Model은 **사용자의 mental model을 다음과 같은 서로 구분되는 3가지의 sub-model로 나눈 것**이다.   |  |  | | --- | --- | | User Model of the **Domain** () | 사용자는 **domain에 대한 어느 정도의 이해**가 있고, expertise의 수준은 높을 수도 있고 낮을 수도 있다.   * 이 이해는 AI 시스템 및 설명 전략과 상호작용하면서 발전될 수 있다. | | User Model of the **AI System** () | 사용자는 domain model에 대한 것과는 다른 **AI model에 대한 이해**가 있다. | | User Model of the **Explanation** () | 사용자는 **설명 전략이 어떻게 시스템을 설명하는지**에 대한 이해를 가지고 있다.   * **Post-hoc explanatory case**가 사용될 때 사람들은 **이웃한 example 간의 evidential dependency**를 이해할 수 있다. |   이 서로 다른 mental model들은 **고려되는 아이템이 서로 같을 때 task-goal context에 의존하여 서로 다르게 develop**될 수 있다. User model의 이러한 관점은 XAI 시스템 평가에 대한 접근에 대해 다음과 같이 암시한다.   * 모든 XAI user-study에서 연구자들은 AI 모델을 단지 평가한다고 하지만, 그들은 **이 3가지의 sub-model을 평가**하고 있다. * 사람들은 주어진 user-study에서 각각의 sub-model에 대한 **expertise level이 서로 다를 수 있다.** * 이들 sub-model 중 **일부 또는 전체가 user study 과정에서 바뀌고 발전**될 수 있다. * 우리는 이 3가지 sub-model 간의 다양한 변화와 상호작용이 **task-goal context에 따라 다양**하다는 것을 기억해야 한다. * 이러한 tricorn analysis는 XAI user study 설계자에게 **설명 방법에 대해 어떤 sub-model을 타겟팅할 것인지**를 생각하게 한다. |

1. **Experiment 1: post-hoc explanations of a CNN classifier**

|  |
| --- |
| **Explanation이 있을 때와 없을 때의 Mean Correctness Rating과 Classification의 종류 (right, errors):**    **Misclassification에 대한 ANOVA 분석:** |

1. **Experiment 1: post-hoc explanations of a CNN classifier (Cont.)**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Classification의 종류에 따른 Correctness 및 Reasonableness:**    **<분석>** (Correctness Ratings) (Reasonableness Ratings)   |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Item** | **p <=** | **F(a, b)** | **Wilks’** | **Partial** | | Main effects of Explanation and Classification-Type | 0.001 | F(2, 200) = 1235.9 |  | 0.93 | | Reliable interaction between Explanation and Classification-Type | 0.005 | F(2, 200) = 5.4 |  | 0.05 | | Main effect of Classification-Type | 0.001 | F(2, 200) = 335.3 |  | 0.77 | | Reliable interaction between Explanation and Classification-Type | 0.05 | F(2, 200) = 3.84 |  | 0.037 | |

1. **Experiment 2: explanations and high-end error-rates**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Right Classification 및 Wrong Classification의 Percent-Error Level 및 Explanation의 유무에 따른 평균 correctness ratings**    **<분석>** (Correctness Ratings) (Reasonableness Ratings) (Trust) (Satisfaction)   |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Item** | **p <=** | **F(a, b)** | **Wilks’** | **Partial** | | Correctness ratings revealed significant effects for Explanation | 0.05 | F(2, 158) = 3.129 | 0.962 | 0.04 | | … for Error-Rate | 0.001 | F(4, 316) = 14.976 | 0.707 | 0.16 | | Error-rate impacts correctness ratings for wrong classifications | 0.001 | F(2, 159) = 31.61 |  | 0.284 | | … for right classifications | >0.20 | F(2, 159) = 1.45 |  | 0.018 | | Significant interaction between Explanation and Error-rate | 0.05 | F(4, 316) = 2.53 | 0.94 | 0.03 | | Explanation x Error-rate interaction occurs only for right classfications | 0.05 | F(2, 159) = 4.27 |  | 0.05 | | Trust survey revealed a significant effect of Error-Rate | 0.001 | F(16, 304) = 5.234 | 0.616 | 0.22 | | Satisfaction ratings of the overall system revealed a significant effect for Error-rate | 0.05 | F(16, 304) = 1.755 | 0.838 | 0.85 |  * **3%-error** 조건 하에서의 실험자들의 시스템에 대한 confidence level이 가장 높았음 **(M=4.28, SD=0.67)**, 분류가 가장 **예측 가능함** **(M=4.21, SD=0.82)**, 시스템을 **enjoy했음** **(M=3.11, SD=0.86)** (60%-error 조건에서는 **M=2.55, SD=1.15**) |

1. **Experiment 3: the effects of explanation and low-end error-rates**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **<분석>** (Correctness and Reasonableness Ratings) (Overall correctness and overall reasonableness) (Trust and Satisfaction)   |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Item** | **p <=** | **F(a, b)** | **Wilks’** | **Partial** | | Correctness ratings revealed significant effects for Explanation | 0.05 | F(2, 175) = 4.12 | 0.955 | 0.045 | | … for Error-Rate | 0.001 | F(6, 350) = 3.89 | 0.879 | 0.063 | | Error-rate variable’s effects does not occur in right classifications | >0.30 | F(3, 176) = 1.23 |  | 0.02 | | Overall correctness and reasonableness ratings revealed significant effects for error-rate | 0.001 | F(6, 348) = 9.43 | 0.74 | 0.14 | | Trust ratings of overall system also revealed significant effect for Error-rate | 0.009 | F(24, 490) = 1.85 | 0.78 | 0.08 |  * **4%-error** 조건 하에서의 실험자들의 시스템에 대한 confidence level이 가장 높았음 **(M=4.43, SD=0.54)**, 이 시스템이 **가장 믿을 만함** **(M=3.24, SD=1.02)**, 시스템이 **정확히 분류할 것으로 믿음** **(M=3.93, SD=0.8)** |

1. **Discussion and conclusion (생략)**