**Black Box Attacks on Explainable Artificial Intelligence (XAI) methods in Cyber Security**

[**https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9206780**](https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9206780)

**0. Role of XAI in Security Domain**

|  |
| --- |
| **Explainable Security (XSec)**은 XAI를 security domain으로 확장시킨 것으로, 다음의 요소를 갖는다.   * Who? What? Where? When? Why? And How? |

**1. Taxonomy of XAI in Security Domain**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 여기서는 **Explainability space를 다음의 3가지로** 나눈다.   |  |  | | --- | --- | | X-PLAIN | **예측(prediction) 또는 데이터 그 자체(data)**에 대한 설명 | | XSP-PLAIN | 예측 또는 데이터의 **security와 privacy property**들을 포함하는 설명 | | XT-PLAIN | 예측 또는 데이터의 **threat model**을 포함하는 설명 |   각각은 다음에 대한 설명을 포함한다.  **X-PLAIN:**   * 설명에 대한 **static vs. interactive한 변경 사항** * **Local vs. global**한 설명 * 모델을 cover하는, **In-model vs. post-hoc model** 설명   + 직접적으로 보일 수도 있지만, 모델이 훈련된 후 auxiliary method를 사용할 수도 있음 * **Surrogate model**은 더 복잡한 모델에 접근하는, 보통 directly interpretable한 모델이다.   **XSP-PLAIN:**   * 데이터와 모델에 대한 **Confidentiality property**   + 예를 들어, 데이터의 어떤 feature가 시스템 소유자에 의해 보호되는가? * 데이터와 모델의 **integrity property**   + 데이터가 언제 어떻게 수집되고 모델이 훈련되었는가?   + Fairness 속성은 model integrity의 속성 중 일부일 수 있다. * 데이터와 모델에 대한 **Privacy property**   + 데이터 또는 예측의 어떤 부분이 누구에게 노출되었는가?   + 공공적으로 노출된 training data와 model에서는 noise가 추가되어 data right 또는 model privacy가 compromise되지 않았는가?   + Global explainability method는 model weight에 대한 detail을 설명하지 않고도 모델에 대해 설명할 수 있는 방법을 찾아야 한다. |
| **(NEXT PAGE)** |

**1. Taxonomy of XAI in Security Domain**

|  |
| --- |
| **(PREVIOUS PAGE)** |
| **XT-PLAIN:**  여기서는 **training과 deployment 과정에서 고려되는 threat model의 속성**에 대해서 설명한다. (데이터 오염 보호, 사용된 threshold 등)   |  |  | | --- | --- | | Correctness | Explainer의 성능을 측정하는 지표로, **입력의 component들을 정확하게 identify**하는 것이다. | | Consistency | **관련된 component를 capture**할 수 있는 explainer의 성능에 대한 지표이다.   * 특히 classifier가 **원본 입력과 transform된 입력을 서로 같은 class로 분류**하면, transform된 입력에 대한 생성된 설명이 original 입력에 대한 생성된 설명과 비슷한지 판단한다. | | Transferability | Explainability는 **model에 영향을 주는 boundary를 더 자세히 설명**하는 작업을 쉽게 할 수 있게 하므로, **Explainability는 transferability의 지지도**라고 할 수 있다.  마찬가지로, **model 내부적인 inner relation**에 대한 이해는 **사용자 또는 공격자에게 이 지식을 재사용하는 것을 더 쉽게** 한다. | | Confidence | Confidence는 **reliability가 예상되는 모델에서 항상 평가**되어야 한다.   * **Stability**는 특정한 모델에서 interpretation을 생성할 때 필요하다. * Trustworthy interpretation은 stable하지 않은 모델에 의해 생성되어서는 안 된다.   따라서, **explainable model은 자신의 동작 체제에 대한 정보를 포함**하고 있어야 한다. | | Fairness | 사회적인 관점에서 explainability는 **머신러닝 모델이 fairness에 도달하고 그것을 보장할 수 있는 능력**으로 간주된다.   * 알고리즘과 모델의 지원은 인간 생활에서 커지고 있으므로, **explainability는 불공정하거나 비윤리적인 알고리즘 사용을 피할 수 있게** 한다. | | Privacy | ML model의 explainability에 의해 가능해진 한 가지는 그것의 **privacy를 평가하는 능력**이다.   * ML model은 학습한 패턴에 대한 복잡한 설명을 할 수 있다. * 모델에 의해 무엇이 capture되었는지에 대한 설명을 할 수 없고 **internal representation으로 저장되어 있다는 것은 privacy를 위반**하는 것이 될 수 있다. * 학습된 모델의 **제3자에 의한 Inner relation에 대한 설명** 능력은 **데이터의 근원에 대한 differential privacy를 compromise**할 수 있다. | |

**2. Proposed Method**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 기반 시스템의 confidence를 compromise하는 공격(attack)은 **CI-Attack과 I-Attack으로 구분**할 수 있다.   |  |  | | --- | --- | | CI-Attack | interpreter뿐만 아니라 관련된 classifier도 공격한다. | | I-Attack | 기반 시스템의 interpreter만 공격한다. |   예를 들어, ReLU 활성화 함수를 사용하는 **neural network 가 를 K개의 카테고리 중 하나로 분류**할 때, explanation map은 가 되고 같은 차원의 벡터의 샘플과 연관된다.   * 이때 이 샘플들의 component는 **신경망의 예측을 위한 각 feature의 relevance score를 인코딩**한다. * **가 주어진 target explanation map이고 가 input sample**일 때, 각 공격 방법에서는 다음과 같이 공격한다.  |  |  | | --- | --- | | I-Attack | **설명이 와 매우 비슷하지만 output은 거의 constant**인 조작된 샘플 를 구성한다. | | CI-Attack | 다음 2가지 조건을 만족시킨다. |   Threat model에 따라 attacker에게는 다음과 같은 constraint가 부가된다.   * **공격이 black-box setting에서 이루어진다.** 이때 공격자는 모델의 기반 구조, 하이퍼파라미터, 학습 데이터의 분포에 대한 지식이 없다. * 원래 instance x를 **로 변환시키는 perturbation 이 sparse**할 수 있다. * **Adversarial sample 가 interpretable**해야 할 수 있다.   + 어떤 **는 모델의 training data distribution에 가까이 있을 때 interpretable**하다고 할 수 있다.  |  |  | | --- | --- | | Step 1 | **Black-box model attack**을 사용한다.   * 공격자는 n개의 test data를 수집하고, 각 샘플에 대한 예측과 설명을 요구한다. * 여기서는 **Manifold Approximation Algorithm (MAA)를 n개의 data point**에 적용한다.   + 이를 통해 **best piece 또는 subspace** 을 찾고, 다음의 mean square error가 최소화되는 **projection map** 을 찾는다.   Step 1의 output은 **n개의 sample을 데이터 분포로 나누는** 데 도움을 준다. | |
| **NEXT PAGE** |

**2. Proposed Method**

|  |
| --- |
| **PREVIOUS PAGE** |
| Step 2에서는 **decision boundary를 를 따라 로 이동**시키기 위해 **입력 분포 에 대한 최소한의 distortion**을 유도해야 한다.   * 여기서 **는 각각 natural, target sample distribution**이다. * 데이터 분포에 따른 perturbation은 **adversarial sample 가 interpretable해야 한다는 data manifold constraint**를 만족시키는 데 도움을 준다.   이를 달성하기 위하여 다음 방정식을 푼다.  공격의 성공 정도를 다음 수식을 이용하여 평가한다. |

**3. Evaluation and Discussion**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **System 1:** **Mimicus** – 해로운 PDF 문서를 탐지할 수 있는 multilayer perceptron  **System 2:** **Drebin** – Android Malware를 identify하기 위한 multilayer perceptron  **System 3:** AAE(Adversarial Auto Encoder)를 이용하는 Intrusion Detection System으로 **AAE-BotNet AD**라고 함   |  |  | | --- | --- | |  |  | | 실험에 사용된 모델의 성능 | 제안된 CI-Attack과 I-Attack 공격 결과  (Attack Success Rate) | |
| **NEXT PAGE** |

**3. Evaluation and Discussion**

|  |
| --- |
| **PREVIOUS PAGE** |
| |  | | --- | |  | | **해로운(Malicious)** PDF 파일 268a에 대한 Explanation Map | |  | | Mimicus System에 대한 IC-Attack 결과 **classification이 Benign으로 바뀜** | |  | | Mimicus System에 대한 I-Attack 결과 **원래 파일의 분류가 그대로 Malicious임** | |
| **NEXT PAGE** |

**3. Evaluation and Discussion**

|  |
| --- |
| **PREVIOUS PAGE** |
| |  | | --- | |  | | **해로운(Malicious)** Android 파일 eb1b에 대한 Explanation Map | |  | | Drebin System에 대한 IC-Attack 결과 **classification이 Benign으로 바뀜** | |  | | Drebin System에 대한 I-Attack 결과 **원래 파일의 분류가 그대로 Malicious임** | |