XAI

조서윤

XAI

설명 가능한 인공지능(Explainable Artificial Intelligence)의 약자로, 인공지능의 행위와 도출한 결과를 사람이 이해할 수 있는 형태로 설명하는 방법

인공지능 시스템의 도출 결과를 이해하지 못하는, 블랙박스 한계를 해결하고자 등장

과정

1. 기존 머신러닝 모델에 설명 가능한 기능 추가
2. 머신러닝 모델에 HCI(Human Computer Interaction) 기능 추가
3. XAI를 통한 현재 상황의 개선

XAI 목적

- 하나의 입력 데이터가 AI에 전달되어 최종적인 예측을 내릴 때 데이터 중에서 어떠한 부분이 더 중요시되고 있는가?
- 다양한 데이터를 입력 받는 AI에서 예측을 분별하는 요소, 판단 기준이 되는 요소는 무엇인가?
- 블랙박스와 같은 구조의 AI에 대한 입출력 관계를 보다 알기 쉬운 구조로 만들어 생각하는 경우, 어떠한 판단 과정에 도달하는가?

국소 설명 & 전역 설명

국소 설명

'각 예측 결과의 판단 사유를 이해하는 것'을 목적으로 하는 설명 방법

주어진 하나하나의 사례(입력 데이터)에 대한 예측 경과를 설명

전역 설명

'AI 모델의 전체적인 동작'을 이해하는 것을 목적으로 하는 설명 방법

하나의 사례(입력 데이터)에 대한 예측 과정을 설명하는 것이 아닌 다양한 사례에서의

예측을 통합해 보았을 때 'AI 내부의 지배적인 경향'을 설명

신고 내용의 타당성 검증

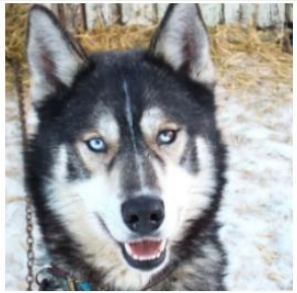
국소 설명은 입력을 활용한 특징량(변수) 중 특히 중요한 요소나 불안 요소를 지정해 제시

어떤 신고의 심사 결과로 AI가 내린 판단했을 때 사유를 설명할 수 있다.

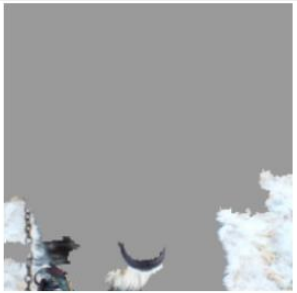
신고 내용이 타당한지를 심사하는 업무에 국소 설명을 사용해 업무 내용과 정합성이 있는

판정 이유를 설명할 수 있다.

의도와 다른 학습 재검토



(a) Husky classified as wolf



(b) Explanation

단순히 예측 결과를 설득시키기 위해 사용하는 것이 아니라 AI 학습의 타당성에 대해서도 시사점을 얻을 수 있음을 기대할 수 있다.

설명 방법의 차이

특징량을 활용한 설명

어떤 입력 데이터에 대해 특징량이 예측 결과에 어느정도 영향도가 있는지를 산출

판단 규칙을 활용한 설명

예측에 이르게 된 근거를 이해하기 위해 판단 규칙의 형태로 설명

데이터를 사용한 설명

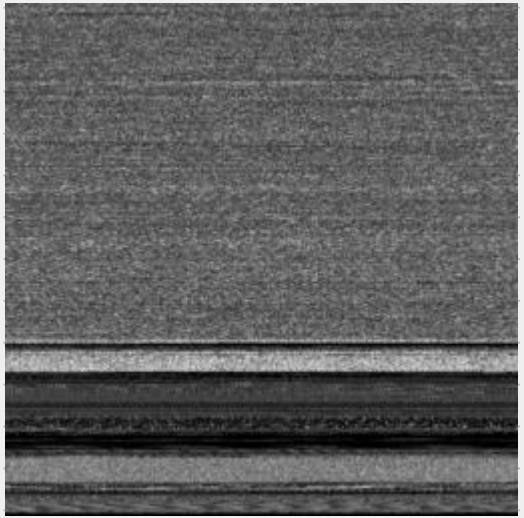
데이터가 입력되어 예측을 수행하는 경우, 그 예측을 판단할 때 많이 참고한 학습 데이터를 제시하여 판단 이유 설명

LIME

LIME은 어떠한 설명 대상 데이터 1건에 대한 AI 모델의 결과 예측에 기여한 데이터의 특징을 산출한다.

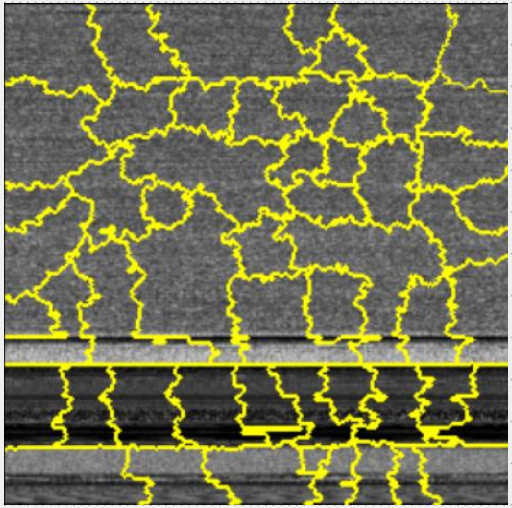
입력 데이터에 '섭동'을 추가해 섭동에 대한 예측 결과의 '변동'으로부터 특징량이 목적 변수에 미치는 영향의 강도를 측정할 수 있다.

AI 모델의 종류에 제약이 없으며 모델이 변경되더라도 다른 XAI로 변경할 필요가 없다.

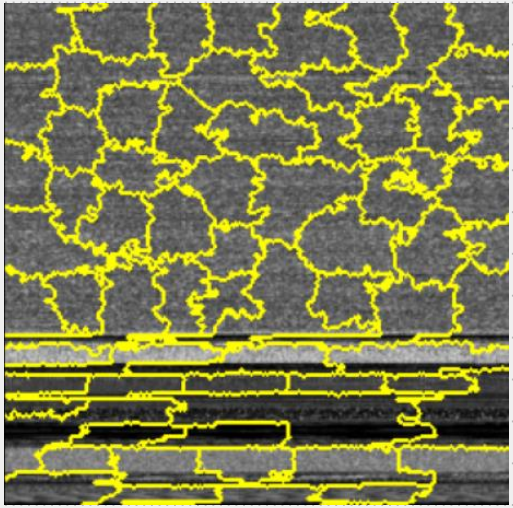


```
(0.39672533, 1, 'Label B')
(0.2885969, 0, 'Label A')
```

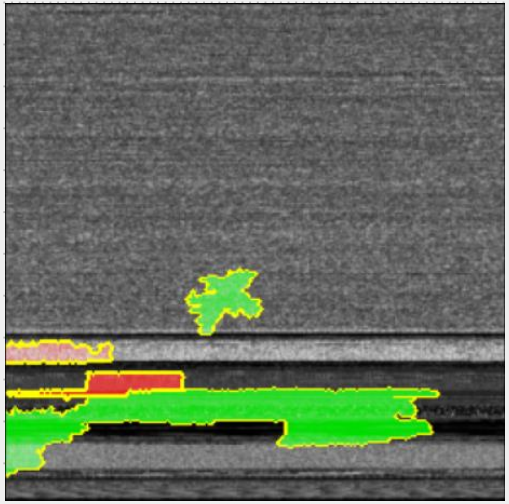
quickshift



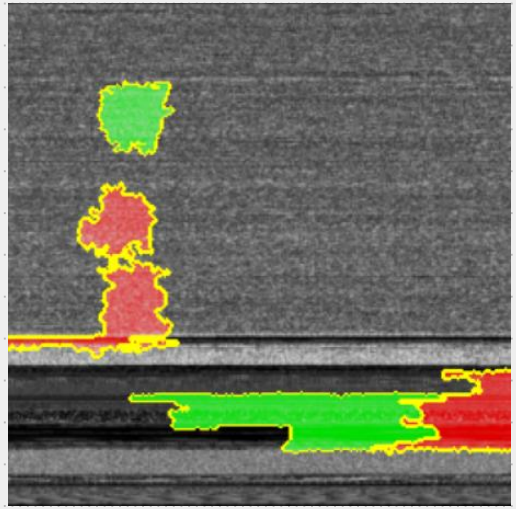
slic



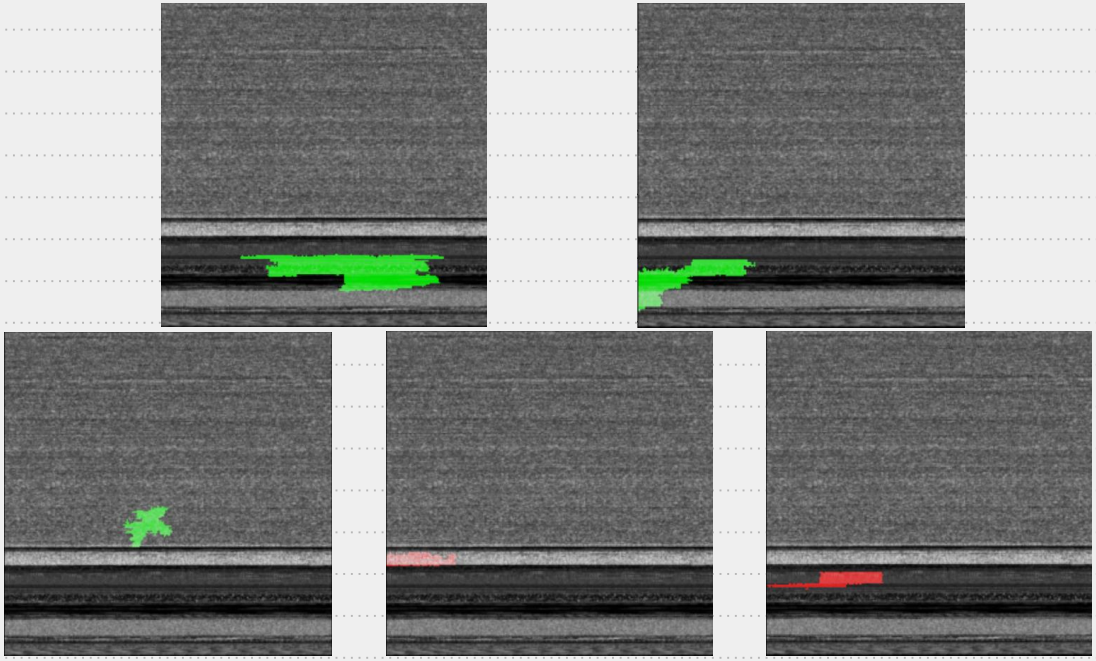
Label A



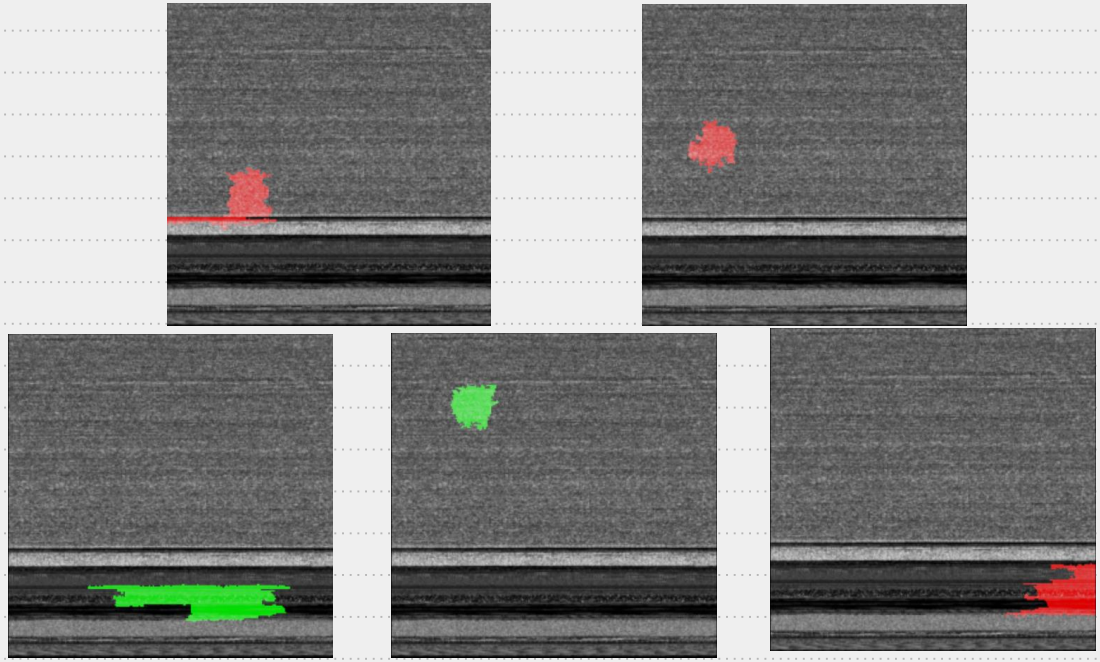
Label B



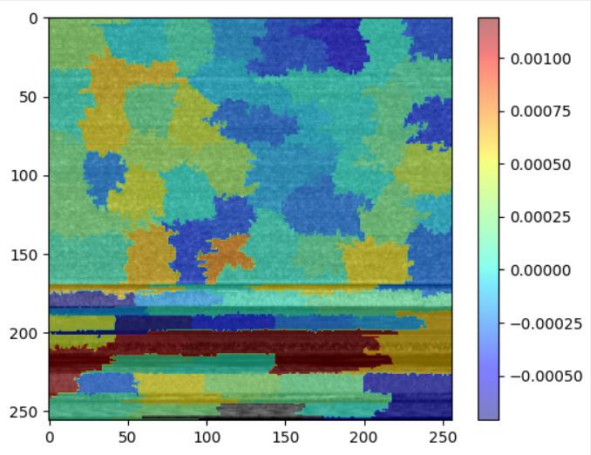
Label A



Label B



Label A



Label B