

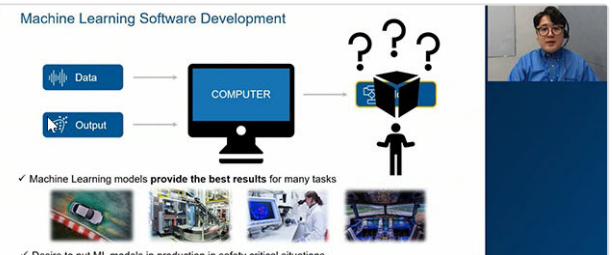
XAI_05. MATLAB 에서의 설명가능한 인공지능

🕒 생성일	@2022년 9월 3일 오후 9:17
🏷️ 유형	머신러닝/딥러닝
👤 작성자	

인공지능 해부하기 : 설명 가능한 인공지능(eXplainable AI) Video

설명 가능한 인공지능의 개요 및 필요성 등의 전반적인 소개와 더불어 실제 설명 가능한 인공지능의 구현 방식 및 활용 방안에 대하여 설명합니다.

🔗 <https://kr.mathworks.com/videos/explainable-ai-interpret-visualize-and-explain-your-deep-learning-model-1621527278198.html>



발표자료

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/e5dffc68-0335-42f8-9b77-0add5d75d30c/2021_%EB%A7%A4%ED%8A%B8%EB%9E%A9_%EC%97%91%EC%8A%A4%ED%8F%AC_-_EC%84%A4%EB%AA%85%EA%B0%80%EB%8A%A5%ED%95%9C_%EC%9D%B8%EA%B3%B5%EC%A7%80%EB%8A%A5.pdf

[새로알게 된 xai 를 분류하는 기준들]

Complexity

- Intrinsic : 간단한 모델은 원래 해석력을 갖고 있을 것.
Transparency(투명성) 을 갖고 있다는 표현이 쓰이기도 한다.
- Post-hoc : 복잡한 모델은 사후 해석

Scope (설명하는 범위에 따라서 해석 범위를 지정)

- Global : 전역적 해석, 모든 예측 결과에 대해서 설명력을 부여.
- Local : 지역적 해석, 하나 혹은 일부의 예측 결과에 대해서 설명력을 부여

Dependency

- Model-specific
- Model-agnostic : 모델 외부에서 설명력을 부여, 모델만의 전형적인 특징 사용 하지 않음.

Visual Interpretation of Features Across Layers

각각의 데이터에 대해서 레이어들이 어떻게 보고 있는지 활성화 함으로써 모델 학습에 대한 단서를 얻을 수 있다.

Grad-CAM


Occlusion sensitivity : 영상에 Occlusion 을 조금씩 주면서 결과값이 어떻게 변화는지를 관찰하는 방식.

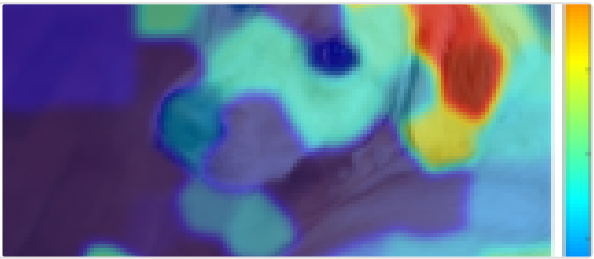
매트랩 온라인에서 LIME 활용해보기

참고자료

Explainable AI: interpreting the classification using LIME

This demo shows how to interpret the classification by CNN using LIME (Local Interpretable Model-agnostic Explanations) [1]. LIMEによる特徴量の可視化 [English] This demo shows how to interpret the classification by CNN using LIME (Local Interpretable Model-agnostic Explanations) [1]. This demo was created based on [1], but

 <https://kr.mathworks.com/matlabcentral/fileexchange/77828-explainable-ai-interpreting-the-classification-using-lime>



구현코드 및 데이터 폴더

: 매트랩 드라이브에 추가 → 매트랩 온라인에서 실행

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/df7f3ac0-9931-4f26-9e5c-1fcc63b4b7af/XAI_LIME.zip

생각보다 pre-trained 된 CNN 모델의 성능이 좋음. LIME 은 사실상 후처리 기법이라고 생각하는데, 구현 메서드가 간단하게 되어있음. 다만, 시각화 하는 부분에서 코드 조정이 필요함.

매트랩 온라인에서 Grad-CAM 활용해보기

<https://github.com/ogemarques/xai-matlab>

Experiment objective