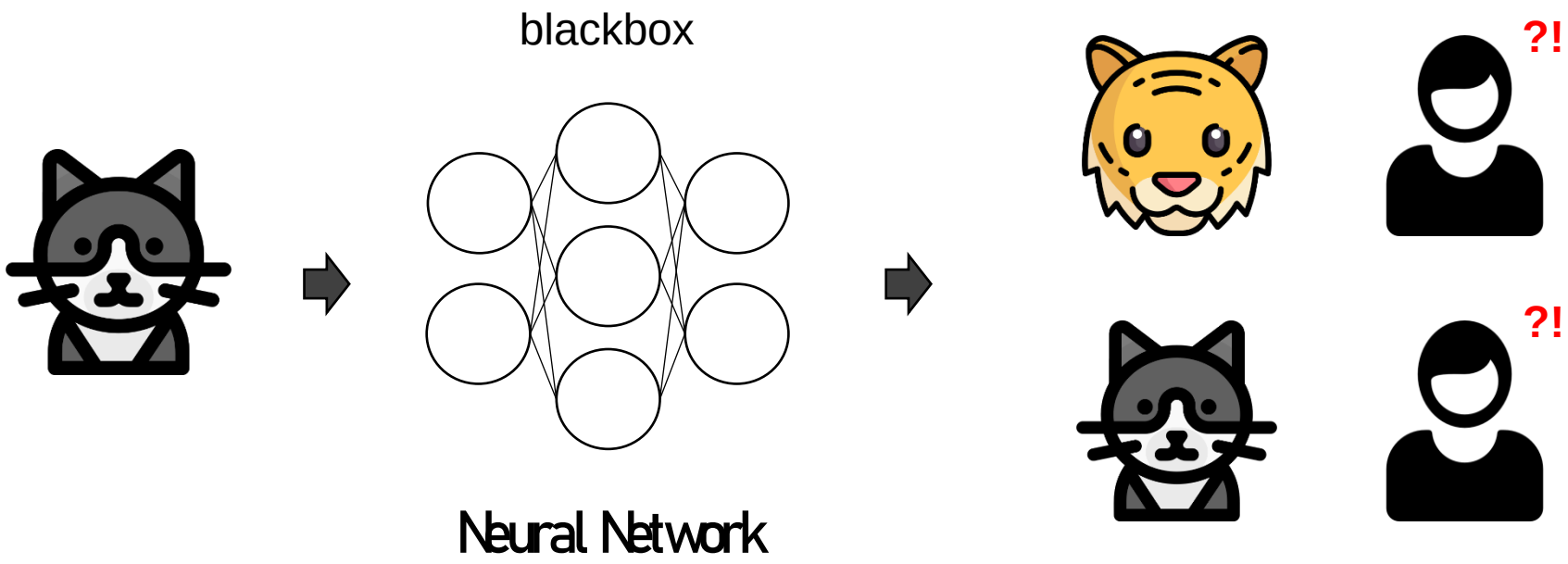
eXplainable Artificial Intelligence (XAI)

https://youtu.be/9mm1rvlfcPk

Contents

CryptoCraft LAB

설명 가능한 인공지능

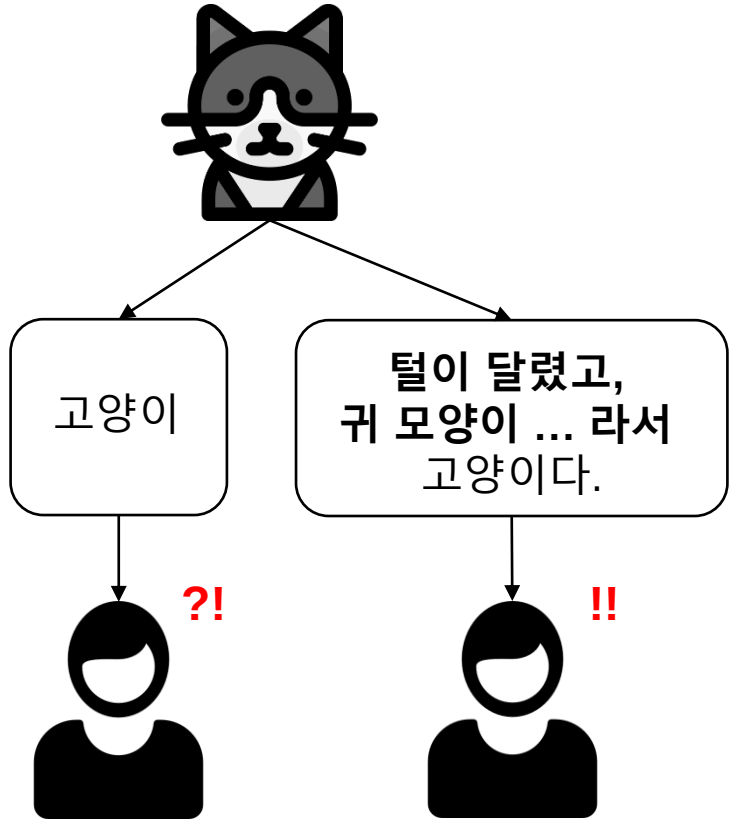


설명 가능한 인공지능

- 다양한 패턴을 추출·분석 → 드러나지 않았던 법칙, 전략 등을 도출
  - 왜 고양이가 고양이로 분류되었는지?
  - 어떤 특징때문에 잘못 분류되었는지?
  - 왜 고양이가 호랑이로 분류되었는지?

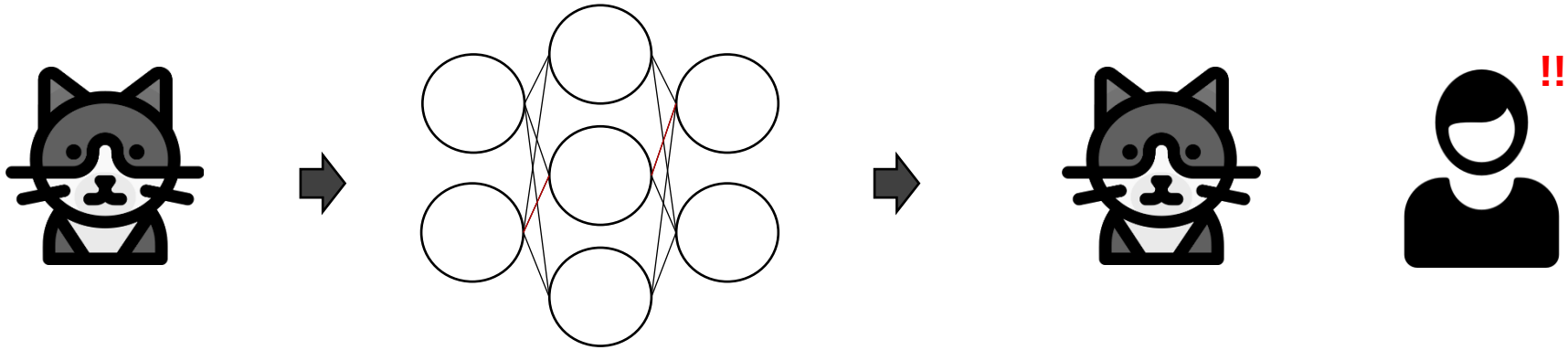
  → 이유 설명 가능
- 이유를 알 수 있기 때문에, 오류 수정도 가능

  →Human-computer interaction 통한 개선
- 결과에 대한 신뢰성 증진



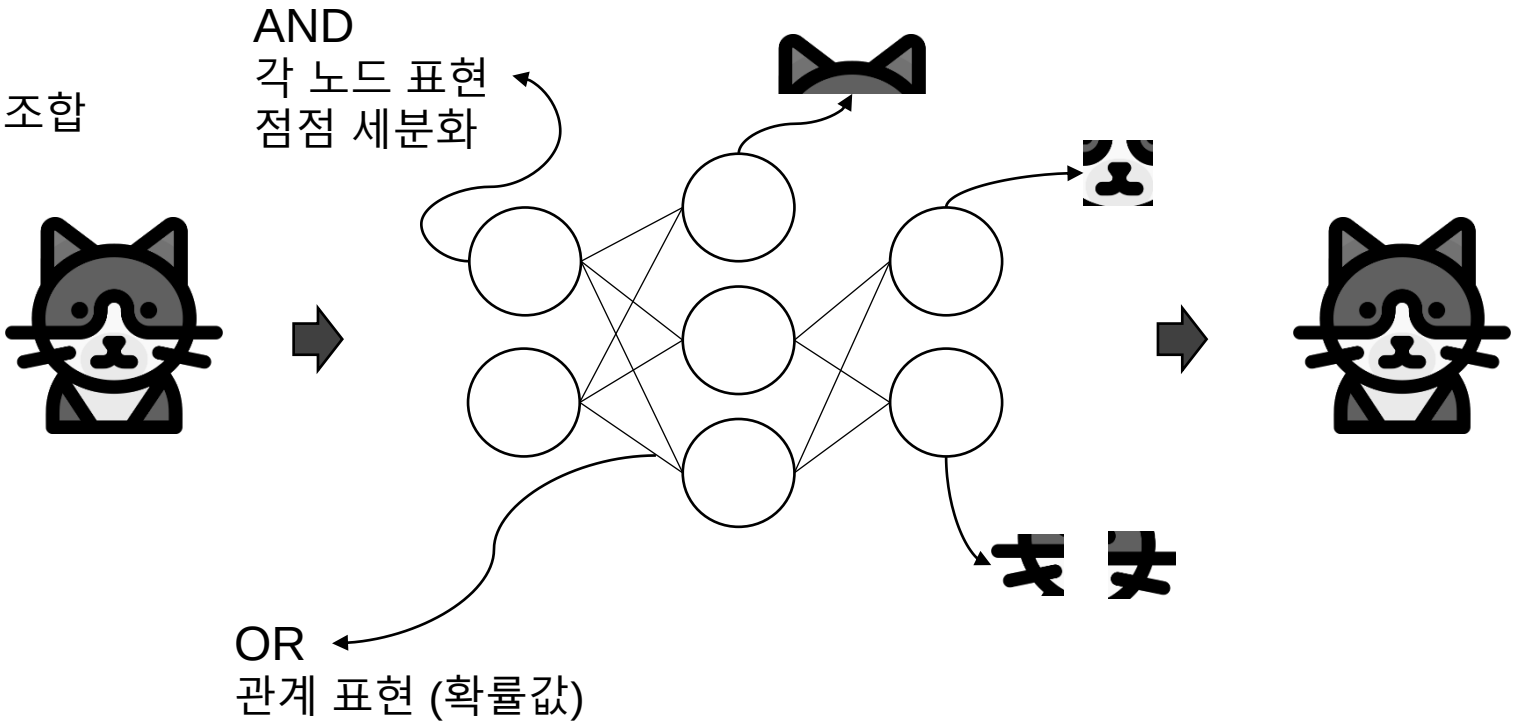
심층 설명 학습

- 설명 가능한 특징들을 학습하도록 함
  - 각 은닉층이 고양이의 귀, 꼬리, 발 등(의미 있는 속성)을 나타내도록 함
  - 학습 후, 귀, 꼬리, 발 중 어떤 것을 근거로 판단했는지 알 수 있음



해석 가능한 모델

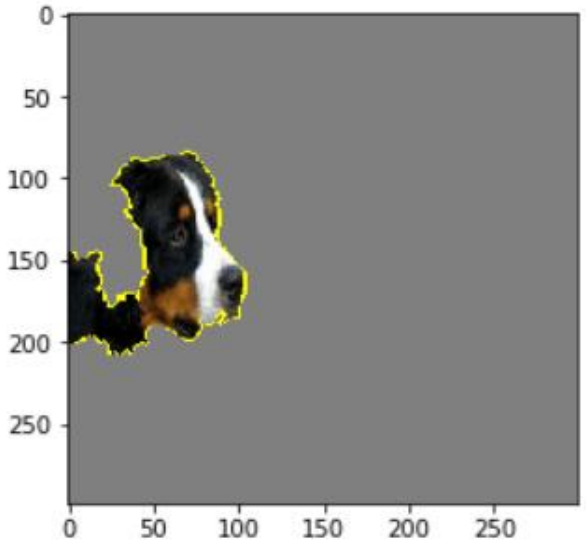
- 해석 가능한 인과 관계 모델
- 작은 단위로 나누어 학습
- 확률적 AND-OR 그래프 기반
  - 각 터미널 노드들이 나타내는 특징을 조합
  - 결과에 이르는 과정과 확률 제공

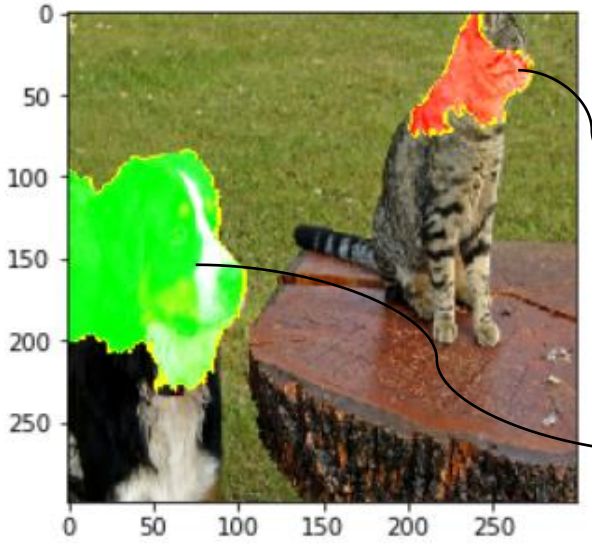


모델 귀납

- Local Interpretable Model-agnostic Explanations (LIME)
  - 이미지를 판단한 결과를 주어진 이미지에서 제시
  - idea : 입력값이 조금 바뀔 때 예측값이 많이 변한다면 중요한 변수
  - 해석 가능한 요소 (super pixel)로 쪼갠 후 가림

    → 여러 번 예측을 통해 강아지를 표현하는데에 가장 중요한 superpixel 을 추출
  - local → 각 데이터에 대해 설명





강아지라고 판단하지 않은 이유

강아지라고 판단한 이유

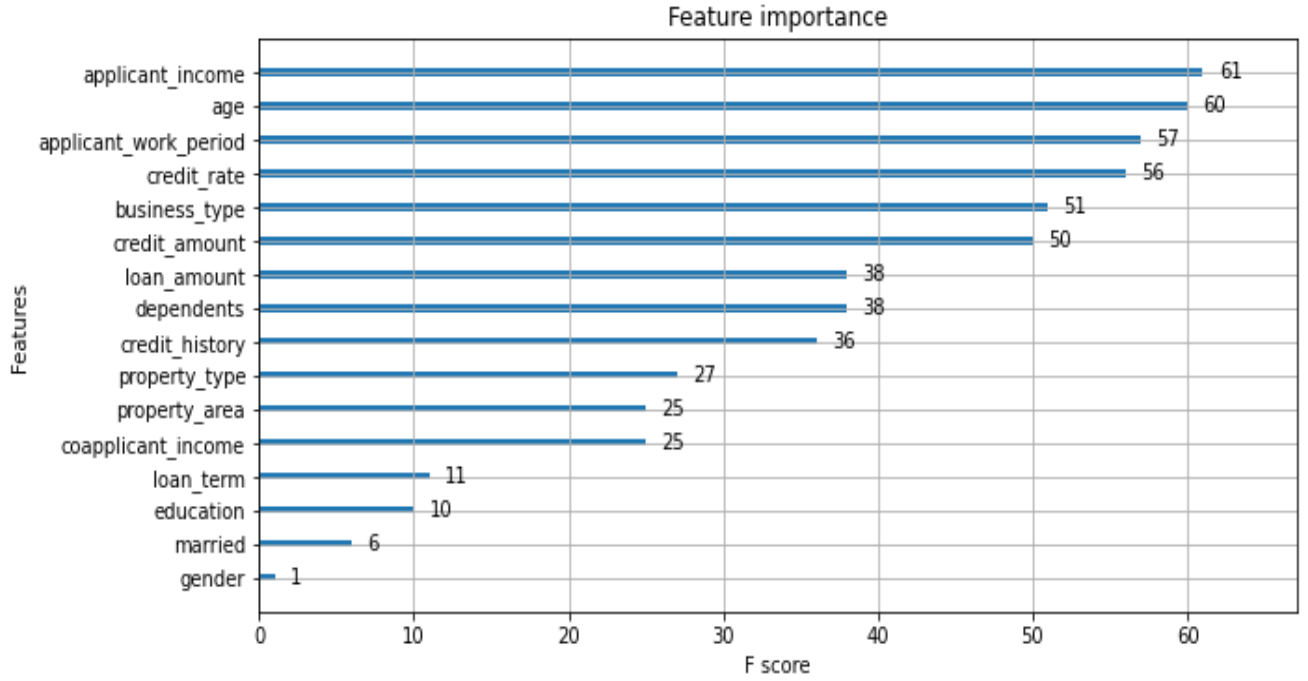
모델 귀납

- SHapley Additive exPlanation(SHAP)
  - 특징별 기여도를 파악
  - 다양한 형태로 표현 가능
  - 전체데이터 또는 일부 표현 가능
  - XGBClassifier() 통해 중요 변수 확인

```
model = XGBClassifier(booster='gbtree', objective='binary:logistic',)
model.fit(x_train, y_train)
```

```
xgboost.plot_importance(model)
```

```
gender                1.0
age                  30.0
married               1.0
dependents            0.0
education             1.0
self_employed         0.0
business_type        74.0
applicant_income   1809.0
applicant_work_period  35.0
coapplicant_income 1868.0
credit_history        1.0
credit_amount      3212.0
property_area         1.0
property_type         1.0
credit_rate           3.0
loan_amount        2520.0
loan_term            30.0
Name: 498, dtype: float64
```

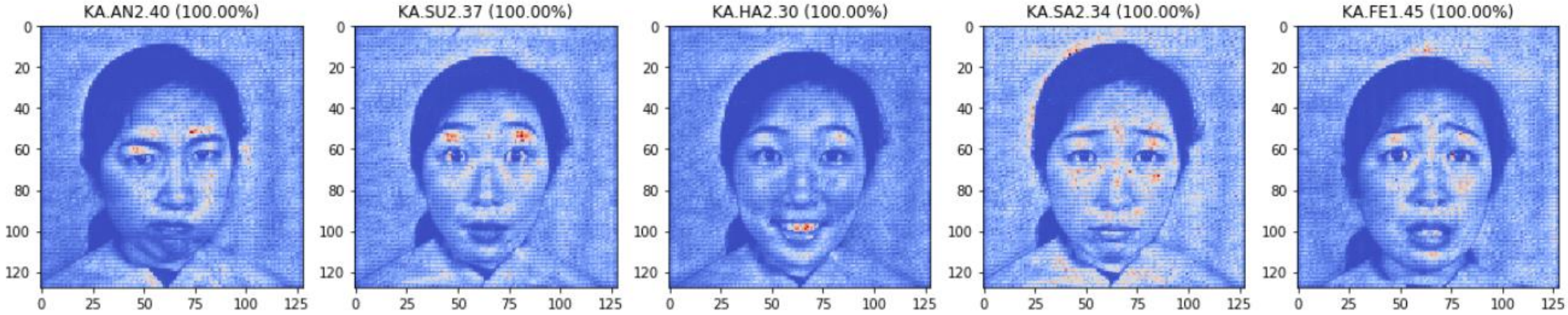


모델 귀납

- Layer-wise Relevance Progagation(LRP)
  - 각 계층의 기여도를 역전파하여 히트맵형태로 표현

    → 결과를 역추적하여 입력 이미지에 표현

*빨간 부분이 가장 크게 기여




```
Y = cnn.forward(X)
D = cnn.relprop(Y*T)
```

Q & A