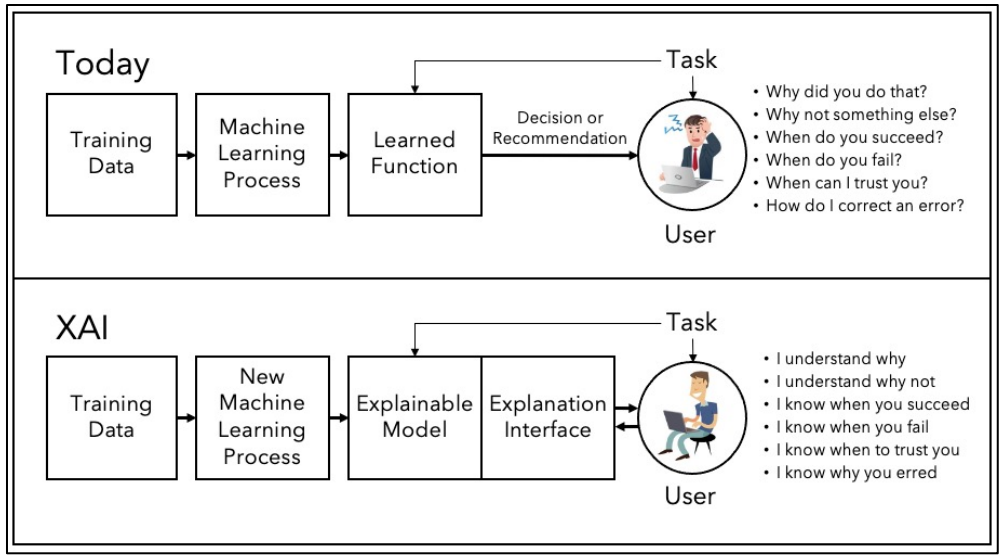
그 외 XAI

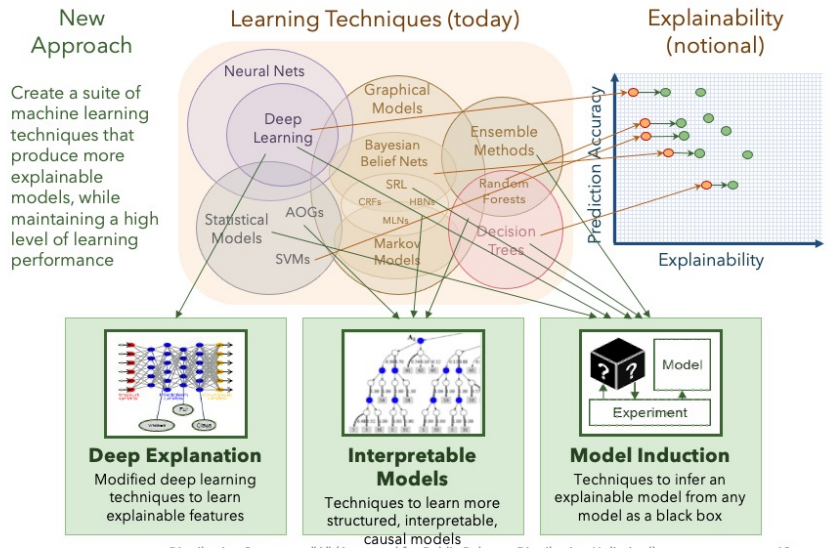
이재준

Contents

XAI의 이해

- XAI
  - DARPA (Defense Advanced Research Projects Agency)
    - 신기술을 제약 없이 만드는 기관 (XAI 대중화의 시초)
    - XAI의 과정을 지침으로 정리
      - 기존 Machine Learning model에 설명 가능한 기능 추가 **(현재 대부분의 연구는 여기)**
      - Machine Learning model에 HCI(Human Computer Interaction) 기능 추가
      - XAI를 통한 현재 상황의 개선
  - XAI의 의미
    - Model이 어떻게 개선되어야 하는지에 대한 직접적인 단서 제공
    - Model 구축 후 어떻게 데이터를 받아들이고 있는지에 대한 해설 필요





XAI의 이해 (2)

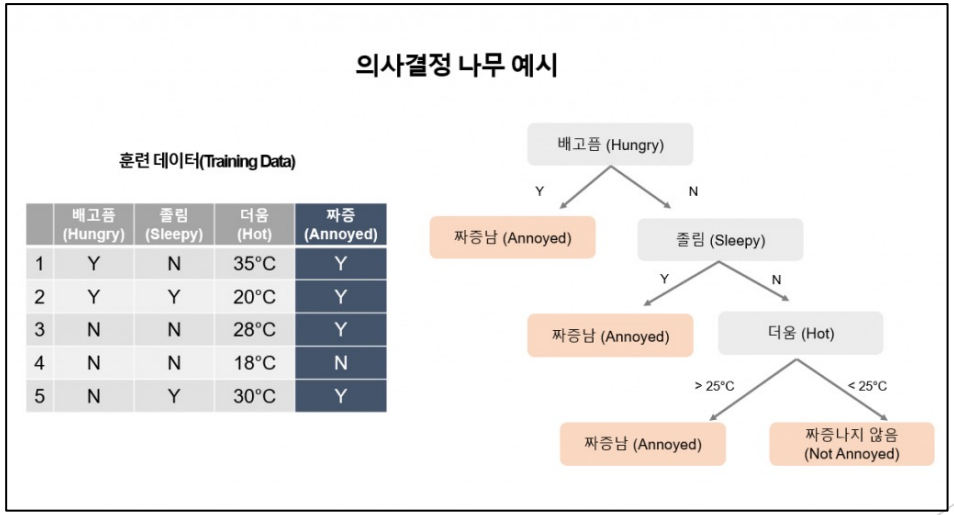
- XAI를 잘하기 위한 조건:
  1. 기존 machine learning 이론을 충분히 이해하기
  2. 설명 모델을 어떻게 접목할지 생각하기
- XAI의 trade-off 관계
  - Deep Explanation
  - Interpretable Models
  - Model Induction (Agnostic)

XAI의 이해 (3)

- *시각화와 XAI의 차이
  - Machine Learning Model의 과정 시각화 ≠ XAI
  - XAI의 핵심: “해석 가능성”
    ➔ 왜 model을 신뢰 하는지, 모델이 왜 특정 결정을 했는지에 대한 근거가 있는지 등
  - XAI ➔ Surrogate Analysis, Partial Dependence Plots (PDP), Similarity Measure, Feature Importance 등으로 **설명**

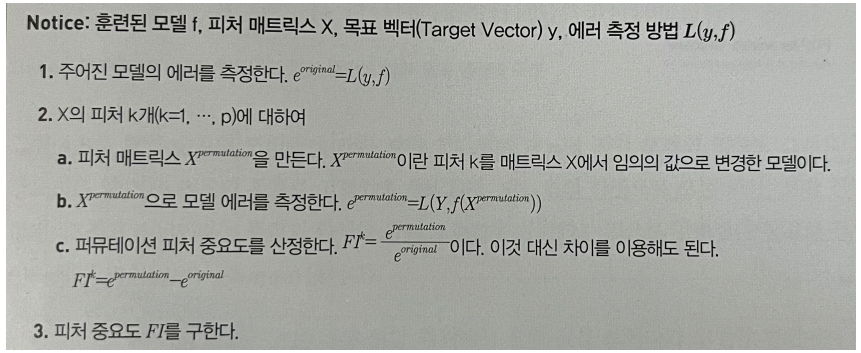
XAI 이해의 기본 ➔ Decision Tree (1)

- Decision Tree: 질문을 던지고 답을 하는 과정을 연쇄적으로 반복
  → Classification or Prediction 진행
- 정보 이득 수치를 계산해 최적의 목표를 달성하는 것이 목표 ➔ Entropy 변화량
  - $Entropy(A) = -\sum_{k=1}^{n} p_k \log_2 p_k$
  - A → 전체 영역, n → 범주 개수, p_k → 전체 영역 중 k 범주에 속한 data
  - Decision Tree 시각화:



Decision Tree (2)

- Decision Tree의 분해 (1): **Feature Importance**
  - Data의 feature가 model/algorithm에 어느정도 영향을 미치는지…
  - 특정 feature를 변형했을 때 model의 예측 결과가 크게 달라졌으면 해당 model은 feature의 의존도가 높음 → feature importance가 높음
  - 계산 방법:



- Decision Tree의 분해 (2): Partial Dependence Plots (PDP)
  - Feature의 수치를 linear하게 변형 → 해석 능력이 얼마나 증감했는지 관찰
  - 자세한 내용: https://eair.tistory.com/20 , https://www.youtube.com/watch?v=hV6FsDBMtx0

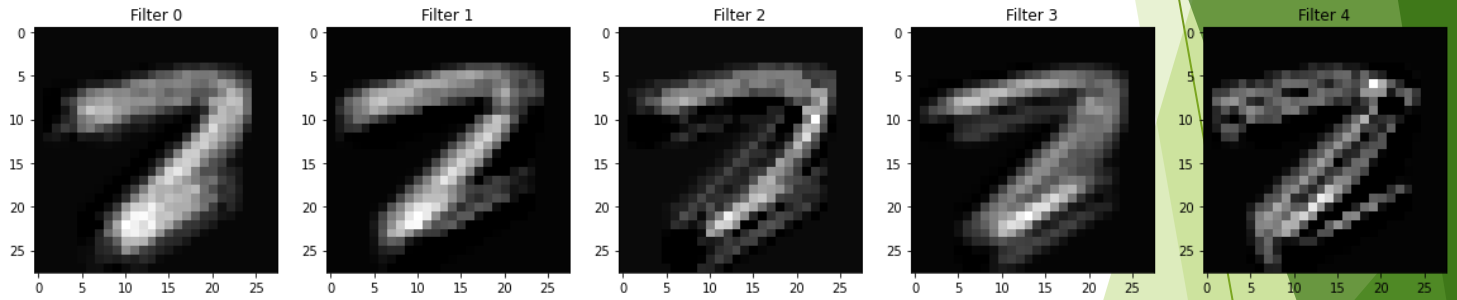
Decision Tree (3)

- Decision Tree의 발전: XGBoost
  - 약한 classifier를 여러개 쌓아서 복잡한 classifier를 형성
  - 장점:
    - 병력 처리로 인해 학습과 classification이 빠름
    - 좋은 유연성 → evaluation function 등의 optimized option 들 제공
    - Greedy Algorithm을 사용해 자동으로 forest 형성 → avoid overfitting
    - 여러가지 algorithm을 연계해서 사용 가능 → ensemble 학습
  - XGBoost를 정확하게 이해하면 이해할수록 XAI에 대한 원리 및 접근이 용이해짐
  - 자세한 내용:
    - 이론: https://www.youtube.com/watch?v=VHky3d_qZ_E
    - 실습: https://www.youtube.com/watch?v=4Jz4_lOgS4c

Filter Visualization (1)

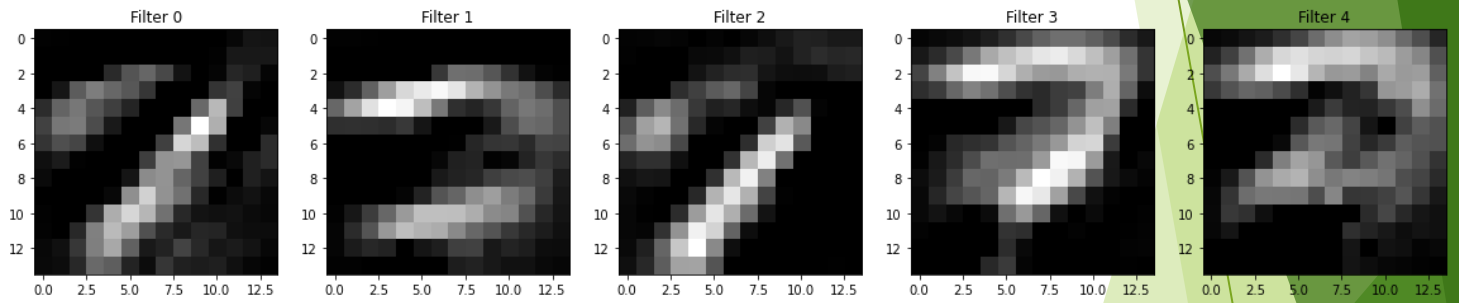
- 보통의 deep learning(CNN, LSTM 등): input layer, hidden layer, output layer
- Hidden Layer
  - Computation 시 backpropagation 진행
  - But.. black box 형태 ➔ 내부에서 어떤 식으로 학습이 진행되는지 알기 힘듦
- Hidden Layer에 대한 별도의 연구 진행 → Visualizing Image Filter
  - Filter / Kernel: raw data를 해석하기 위한 일정한 matrix
  - Hidden Layer에 대한 연구를 하기 위해서는 filter 분석이 필수
  - Filter를 통과할 때 어떤 식으로 값이 바뀌는지 확인 및 비교 → 해석 가능
  - 실습 (colab): http://bit.ly/37M96YV

Filter Visualization (2) - example

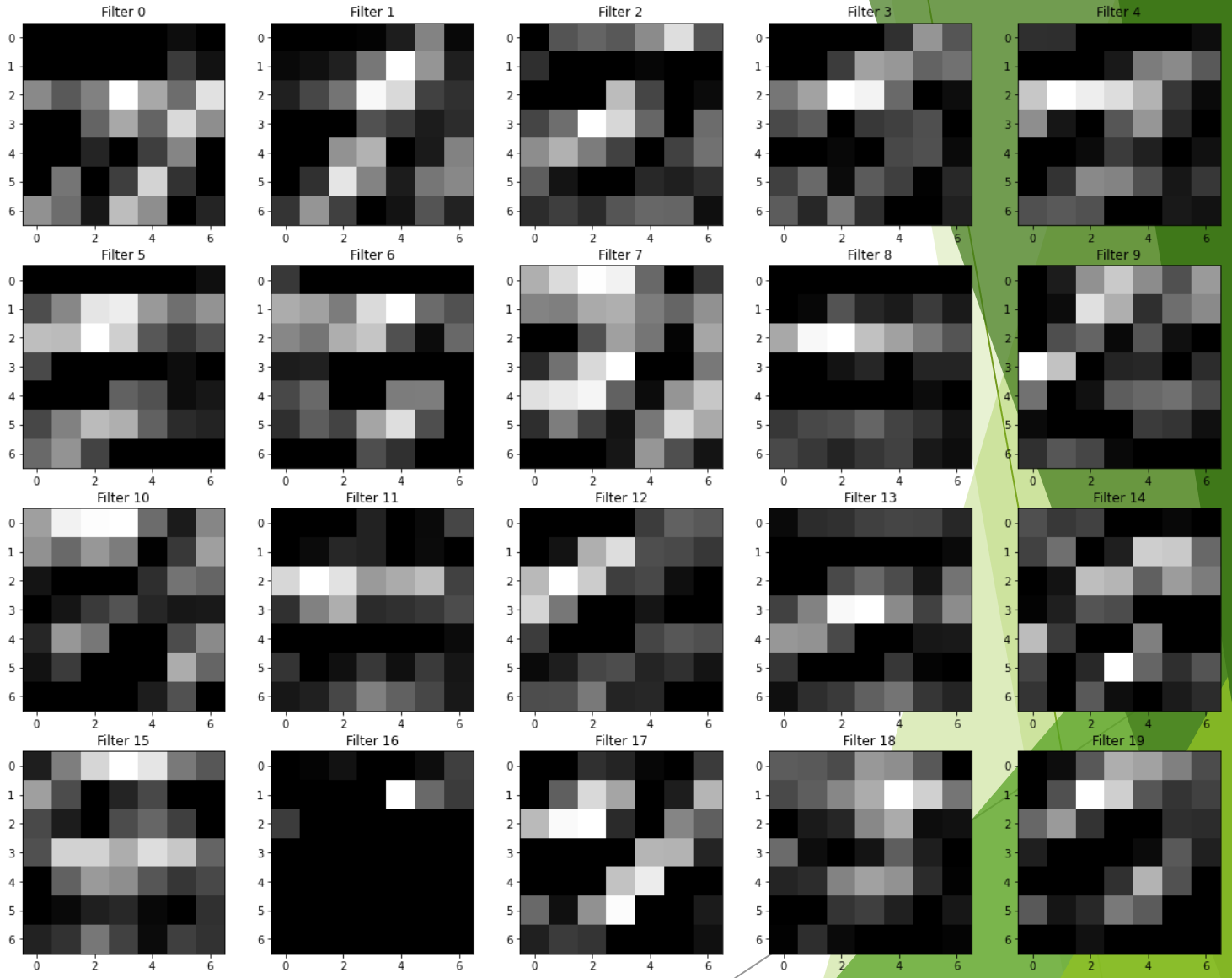


- 손글씨 데이터 (MNIST) 2에 대한 각 hidden layer의 filter 확인
  - **Hidden Layer 1의 filter를 시각화**
  - Hidden Layer 2의 filter를 시각화
  - Hidden Layer 3의 filter를 시각화

Filter Visualization (2) - example



- 손글씨 데이터 (MNIST) 2에 대한 각 hidden layer의 filter 확인
  - Hidden Layer 1의 filter를 시각화
  - **Hidden Layer 2의 filter를 시각화**
  - Hidden Layer 3의 filter를 시각화

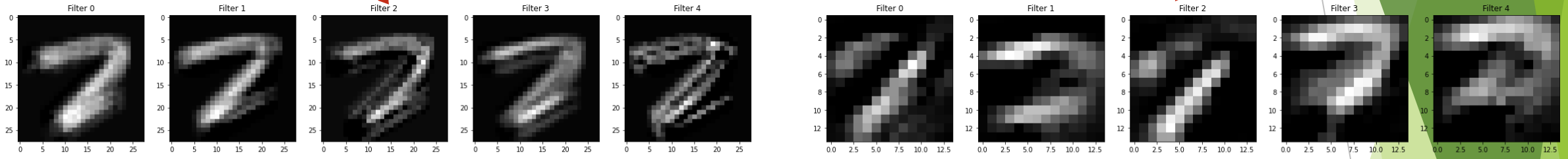


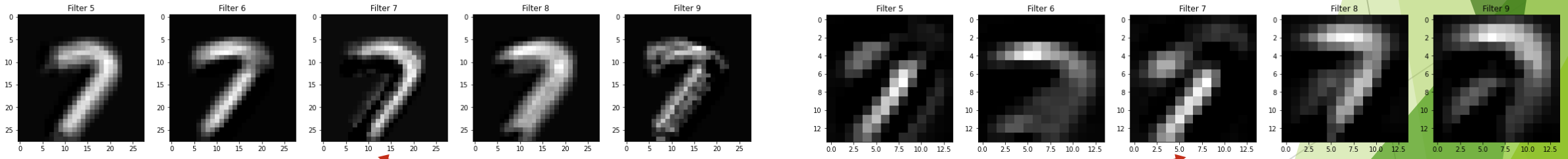
Filter Visualization (2) - example

- 손글씨 데이터 (MNIST) 2에 대한 각 hidden layer의 filter 확인
  - Hidden Layer 1의 filter를 시각화
  - Hidden Layer 2의 filter를 시각화
  - **Hidden Layer 3의 filter를 시각화 ➔ 매우 세분화됨.**

Filter Visualization (3) - Comparison

MNIST - 2

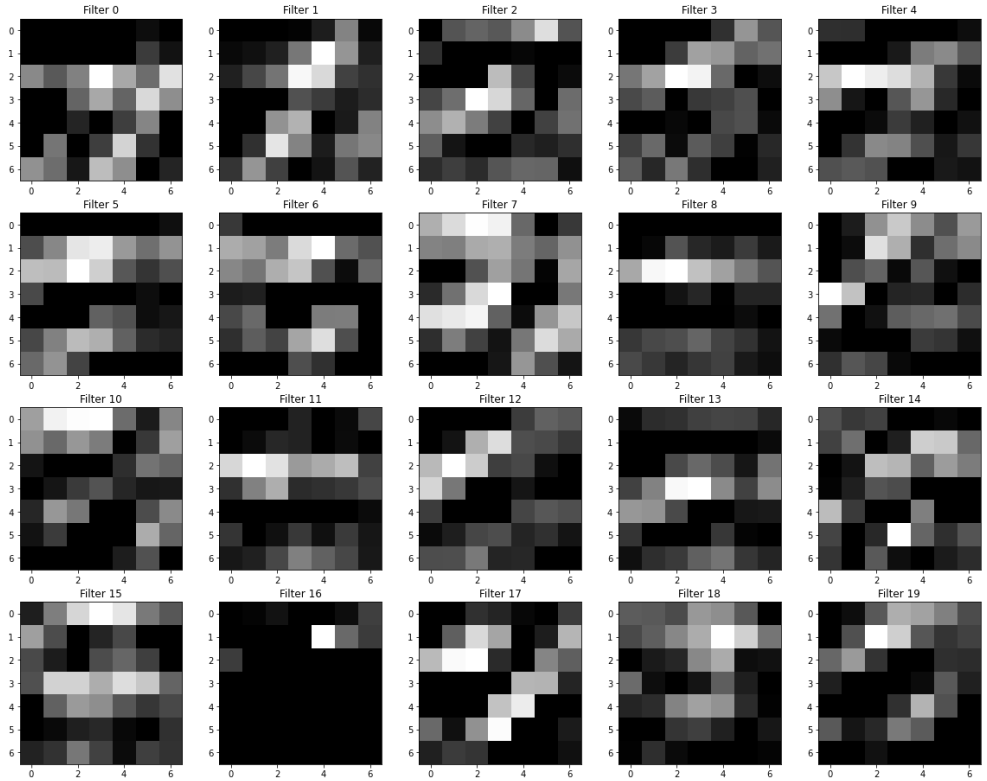




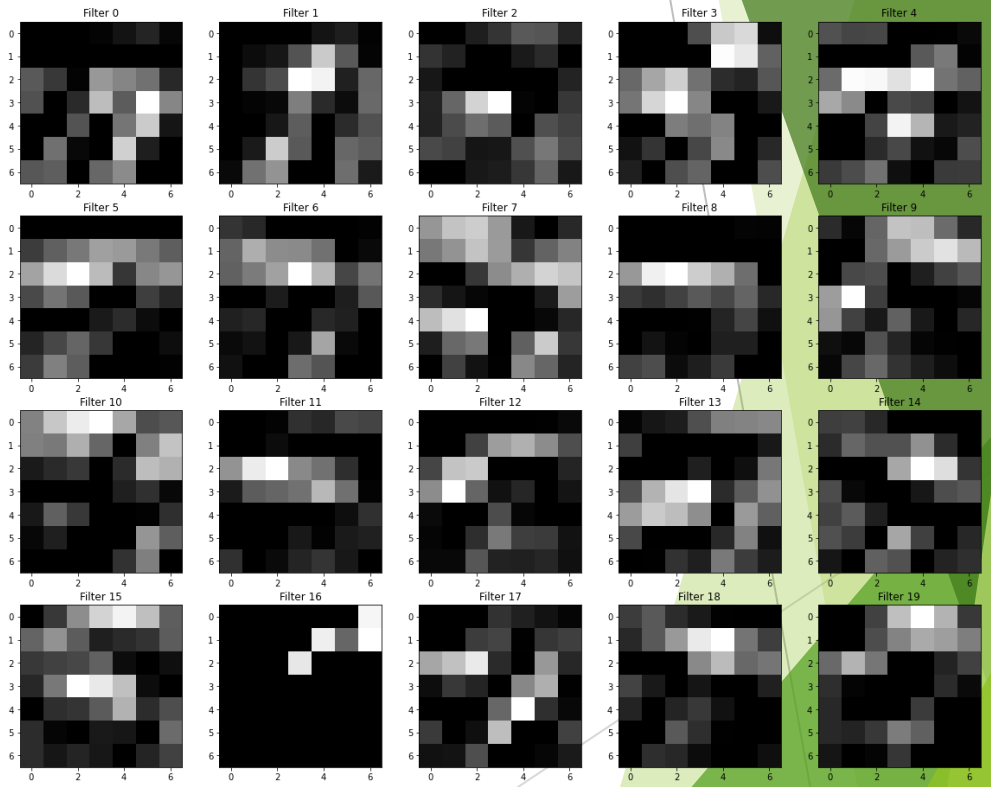
MNIST - 7

Filter Visualization (3) - Comparison

****비교하기 매우 힘듦 → 오류 발생 가능성 높음**



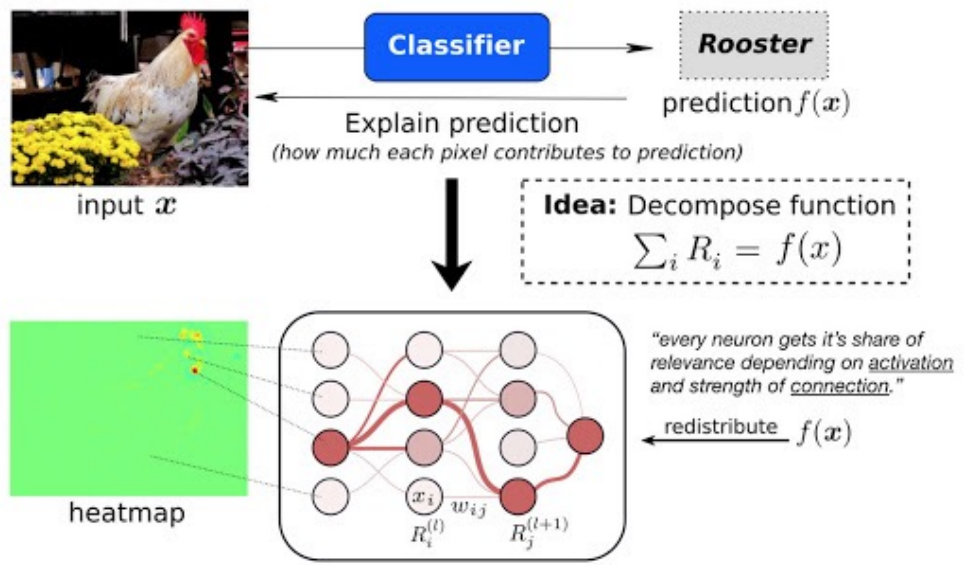
MNIST - 2

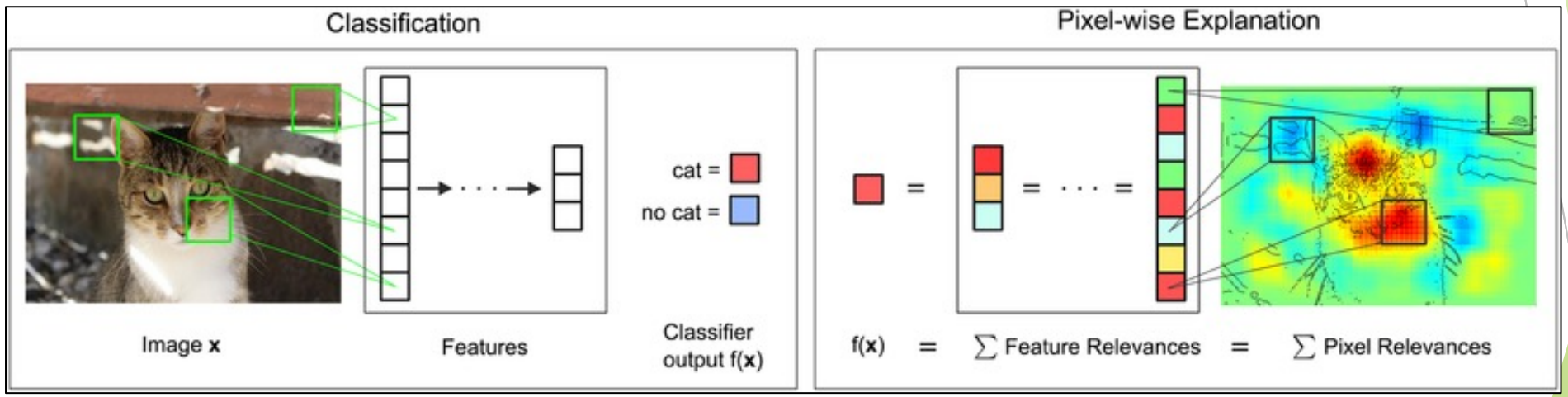


MNIST - 7

LRP (Layer-wise Relevance Propagation)

- 핵심 keyword: 역추적 → input image에 heatmap으로 표현
- Background:
  - Decomposition:
    input의 feature가 결과에 얼만큼의 영향을 미쳤는지 분해
  - Relevance Propagation:
    기여도(relevance)를 top-down 형식으로 재분배
- 실습: http://bit.ly/37SDpwX





LRP (2) - Decomposition

LRP (3) – Relevance Propagation

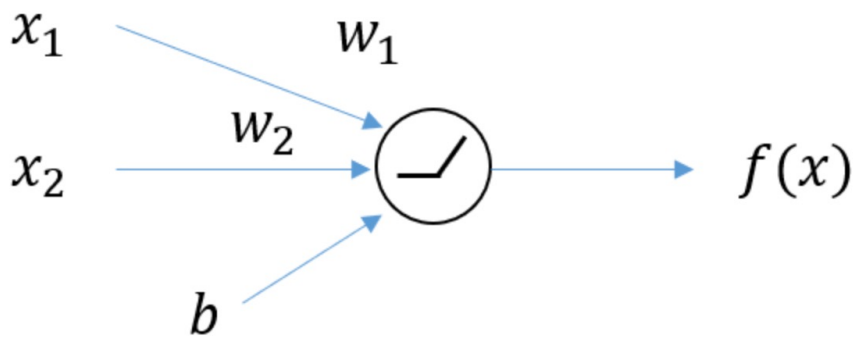
- Taylor Series 를 활용하여 기여도 (Relevance) 계산
- 수학적 기법이 많이 사용되므로… 다음 사이트 참조:
- https://velog.io/@tobigs_xai/3%EC%A3%BC%EC%B0%A8LRPLayer-wise-Relevance-Propagation

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n$$

$$= f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots$$

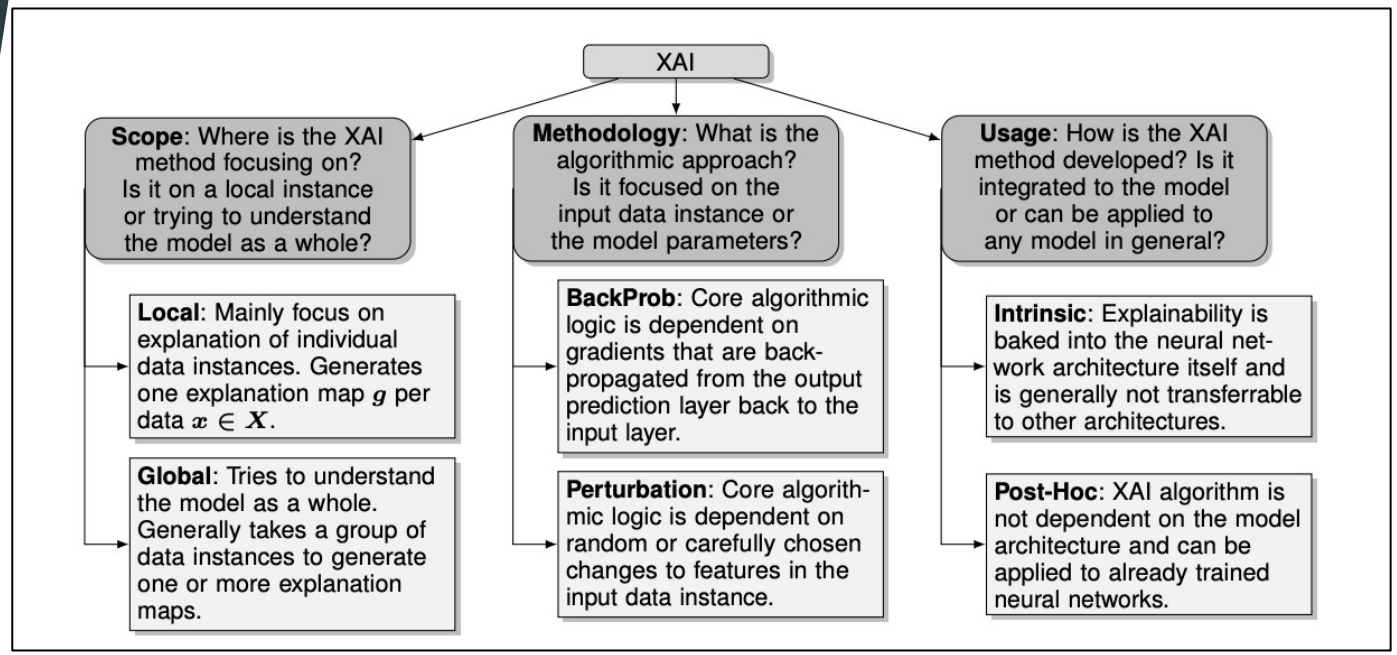
1차 Taylor Series: $f(x) = f(a) + \frac{d}{dx}f(x)\Big|_{x=a}(x-a) + \epsilon$

- ϵ : 2차 이상의 항들
- $\frac{d}{dx}f(x)\Big|_{x=a}(x-a)$: relevance score 결정
  x가 변할 때 $f(x)$ 가 얼마나 변하는지 알 수 있음
- $f(a) = 0, \epsilon = 0$일 때 relevance를 계산할 수 있음



Conclusion

- XAI에 대한 고찰은 강력한 Machine Learning 기법에 대한 이해부터 출발
- 해석을 위한 simple 하면서도 accurate 한 model 필요 (i.g. surrogate model)
- 관점 분류 (Taxonomy): Scope, Methodology, Usage
- 정확한 평가방법 필요: (i.g. 신뢰성, 편향성, 투명성 등)
- 출처 survey: https://arxiv.org/abs/2006.11371



앞으로 계획

AI engineer로써의 간단한 포부(?) 나누기!!