

2023 Poster Session 중간 발표

eXplainable AI

김채원 유광열 이서연



01

eXplainable AI?

02

연구 주제 및 배경

03

사전 연구

04

연구 진행 상황

05

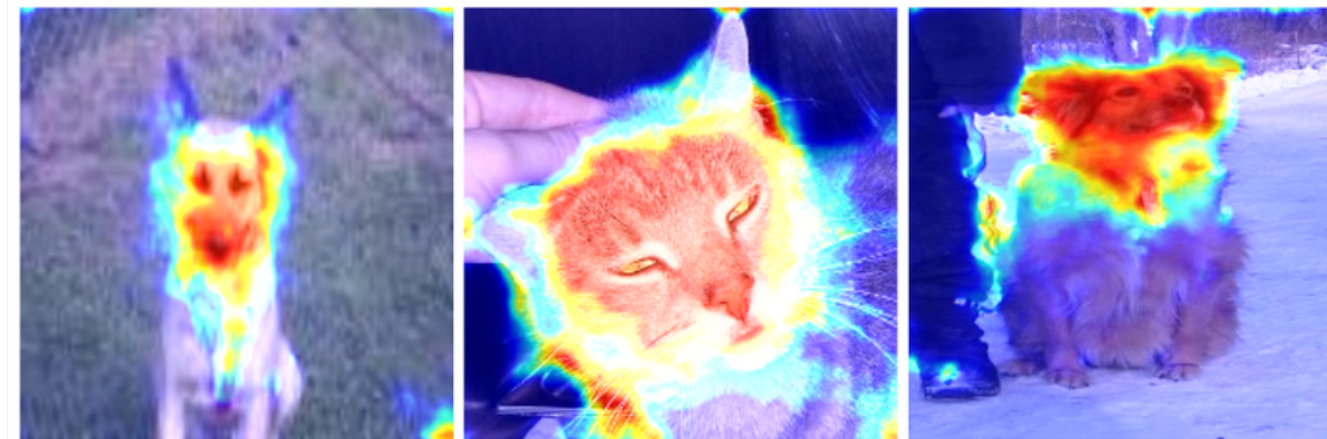
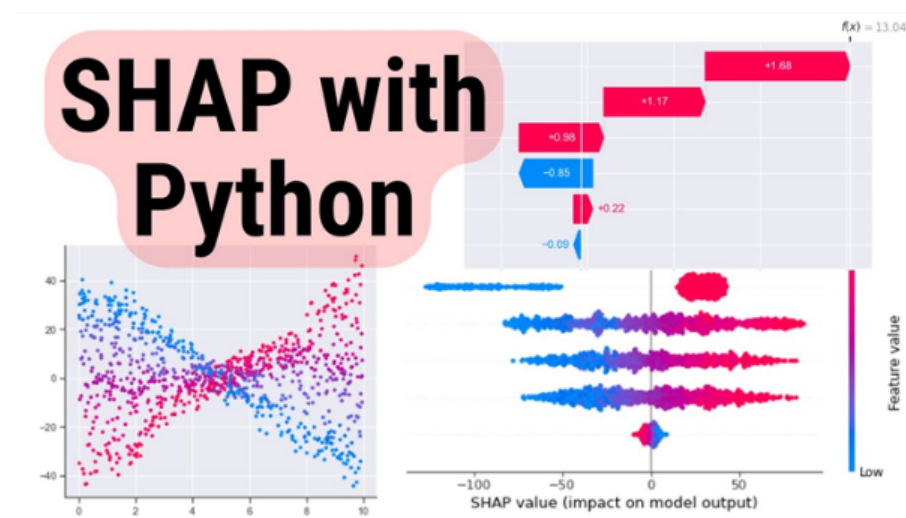
Plan B

eXplainable AI?



AI 모델을 설명하고 해석할 수 있게 하는 기술과 프레임워크를 의미

블랙박스 형태의 복잡한 AI 모델을 투명하고 해석 가능한 방식으로 만들어,
AI를 사용하는 사람들에게 모델의 작동 방식을 이해할 수 있는 기회를 제공



X:AI 적용 Task

현재 여러 Task에서 활발하게 연구중
(Tabular, Image, Text, Time Series 등)




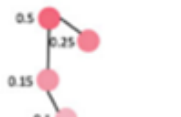


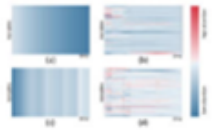





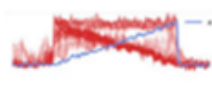

TABULAR	IMAGE	TEXT	TIME SERIES	GRAPHS
Feature Importance (FI) A vector containing a value for each feature. Each value indicates the importance of the feature for the classification. 	Saliency Maps (SM) A map that highlights the contribution of each pixel at the prediction. 	Sentence Highlighting (SH) A map that highlights the contribution of each word to the prediction. <p>the movie is not that bad</p>	Series Highlighting A score for every point in the series, which highlights the contribution to the prediction. 	Node Highlighting A score for every node in the graph which highlights the contribution of that node to the prediction. 
Rule-Based (RB) A set of premises that the record must satisfy in order to meet the rule's consequence. <p>$r = \text{Education} \leq \text{College} \rightarrow \leq 50k$</p>	Concept Attribution (CA) Compute attribution to a target "concept" given by the user. For example, how sensitive is the output (a prediction of zebra) to a concept (the presence of stripes)? 	Attention Based (AB) This type of explanation gives a matrix of scores which reveals how the words in the sentence are related to each other. 	Attention Based This type of explanation gives a matrix of scores that reveal how the points in the series are related to each other. 	Edge Highlighting A score for every edge in the graph which highlights the contribution of edges to the prediction. 
Prototypes (PR) The user is provided with a series of examples that characterize a class of the black box <p> $p = \text{Age} \in [35, 60],$ $\text{Education} \in [\text{College}, \text{Master}] \rightarrow \geq 50k$ </p> <p> $p =$  \rightarrow "cat" </p> <p> $p = \text{"... not bad ..."} \rightarrow$ "positive"  </p>				
Counterfactuals (CF) The user is provided with a series of examples similar to the input query but with different class prediction <p> $q = \text{Education} \leq \text{College} \rightarrow \leq 50k$ $c = \text{Education} \geq \text{Master} \rightarrow \geq 50k$ </p> <p> $q =$  \rightarrow "3" </p> <p> $c =$  \rightarrow "8" </p> <p> $q = \text{The movie is not bad} \rightarrow$ "positive" $c = \text{The movie is that bad} \rightarrow$ "negative"  </p>				
				Graph Prototypes Identifying which part of the graph has influenced the prediction 

Fig. 1 Explanation-based taxonomy with examples divided for different data types

X:AI 기술 분류

Complexity

Intrinsic

모델 자체가 해석 가능한 구조로 고안

Post-hoc

모델의 예측 결과를 사후에 해석

Scope

Global

예측 결과에 대해 항상 설명력을 가짐

Local

하나 또는 일부 예측 결과만 설명 가능

Dependency

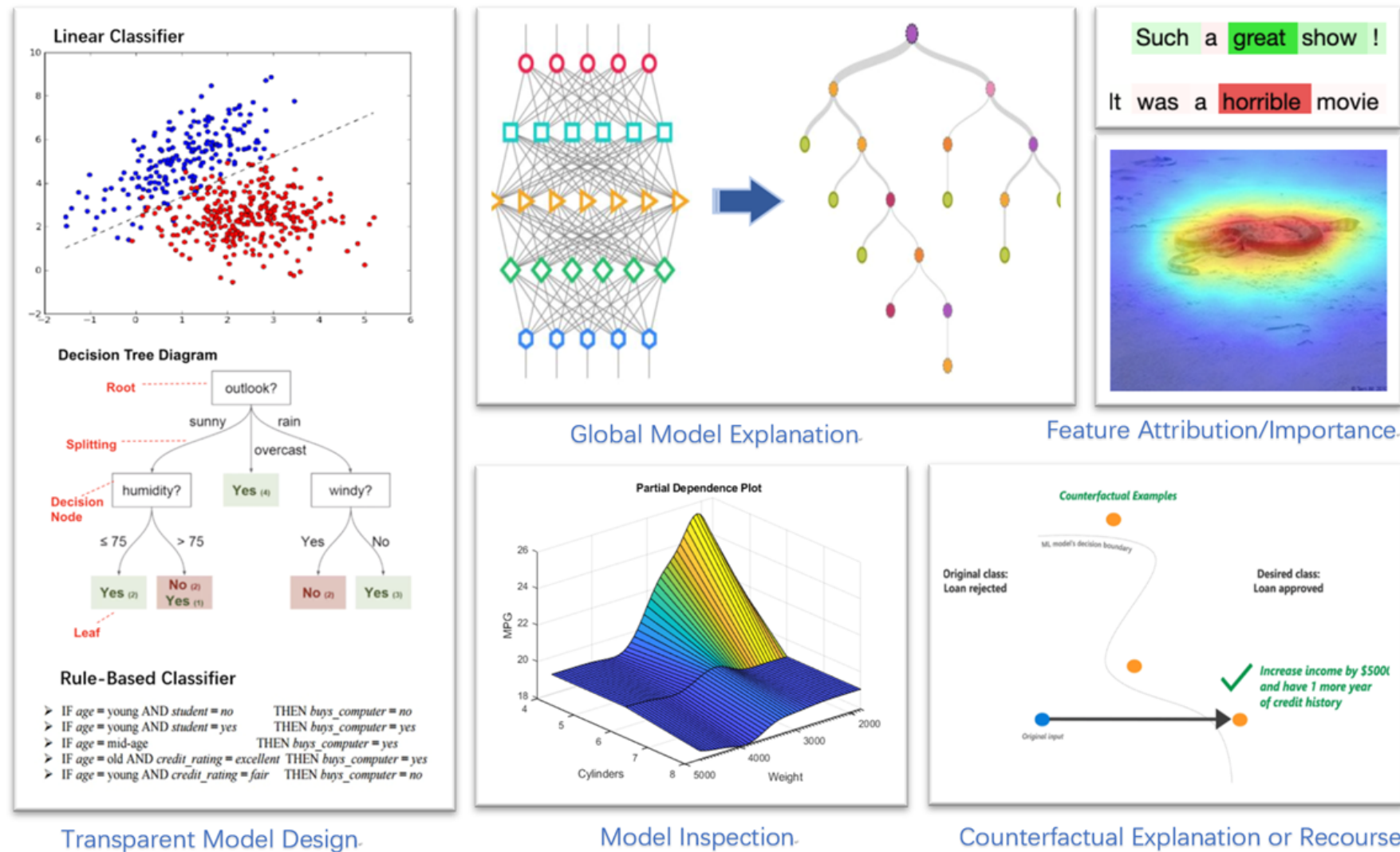
Specific

특정 종류의 모델만 적용할 수 있는 설명 기법

Agnostic

모델 밖에서 근거를 찾아, 모든 모델에 적용 가능

가장 많이 사용되는 X:AI 기술



Transparent Model Design

: Decision Tree와 같은 해석 가능한 모델 설계

Global model explanation

: 모델의 전반적으로 어떻게 작동하는지 이해하는 기술

Feature Attribution/Importance

: 모델이 예측을 진행할 때 어느 특성을 중요시 여기는지 해석하는 방법

Model Inspection

: 모델 검사방법을 통한 해석

Counterfactual Explanation or Recourse

: 예측 모델을 대변하는 설명 가능한 모델 생성

연구 주제

Explainable AI (XAI) for Regression Problem

XAI 기술을 활용하여 회귀 문제에 대해 모델의 설명가능성을 확인



이미지 화질 평가, 예술품 가치 평가 등의 회귀 문제에 대하여,
특정 이미지 영역이 모델의 의사결정에 어떠한 영향을 미쳤는지 파악하고자 함

주제 선정 배경

현재 XAI 분야의 주요 Task는 Classification



Original image
Egyptian cat

Grad-CAM for class:
Egyptian cat



“

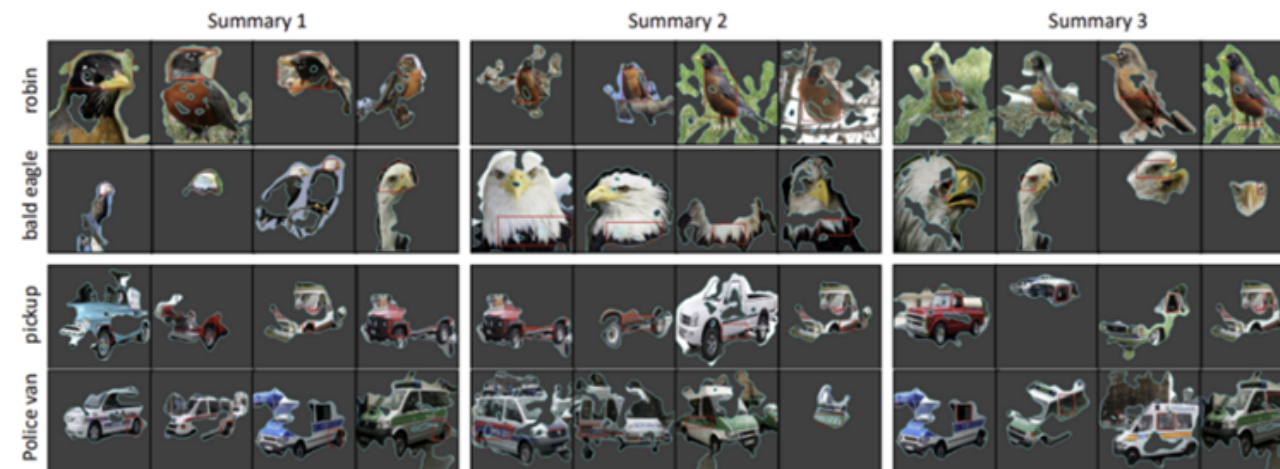
Regression에 적용해보자 !

”

사전 연구 : Image Classification

Understanding Deep Architectures by Visual Summaries (2019)

- 동일한 class에 속하는 여러 이미지에 대해서 네트워크가 인식한 이미지 class의 특징을 파악
- DNN이 특정 class를 결정할 때, 고르게 활용한 선명한 이미지 region을 중점으로 하는 clusters(=summaries)를 생성
 - 이미지 중요도 mask를 제공 & semantic flow similarity 측정을 통해 clustering 진행



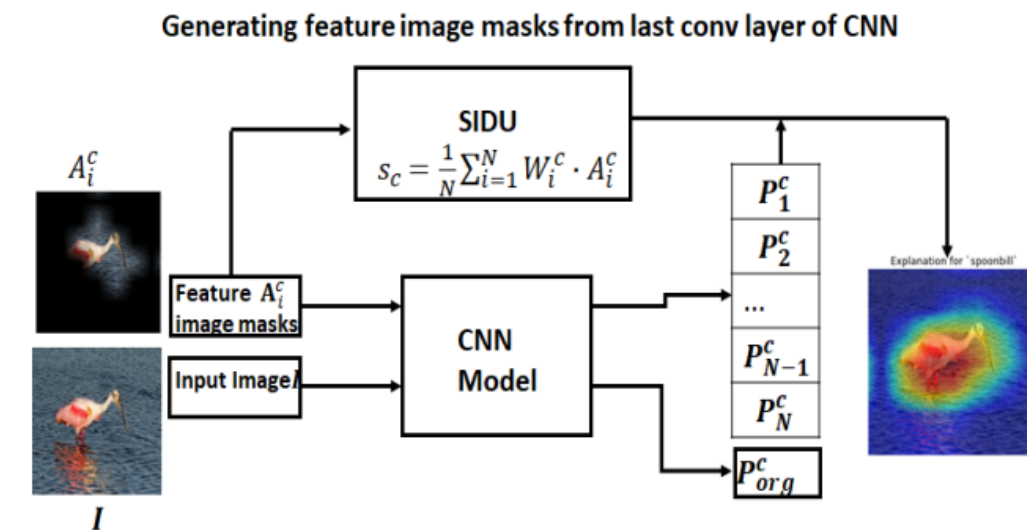
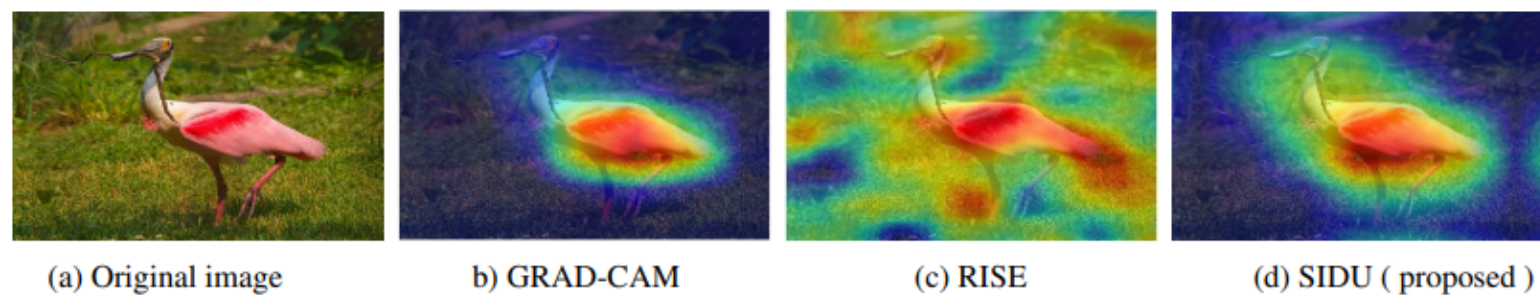
Segmentation Mask를 활용한 방법론

“ 모델이 class를 분류할 때 image의 어떤 region을 중점적으로 활용했는지 파악 가능 ”

사전 연구 : Image Classification

SIDU: SIMILARITY DIFFERENCE AND UNIQUENESS METHOD FOR EXPLAINABLE AI (2020)

- Mask 생성 → 유사성 차이 및 독특성 계산 → 예측에 대한 설명의 순서로 이미지의 중요 region을 시각화
- CNN 모델의 마지막 Conv layer에서 mask를 생성, N개의 각 Activation map을 각각의 binary mask 형태로 변환
 - 최종 시각적 설명 맵은 image mask의 가중합으로 계산



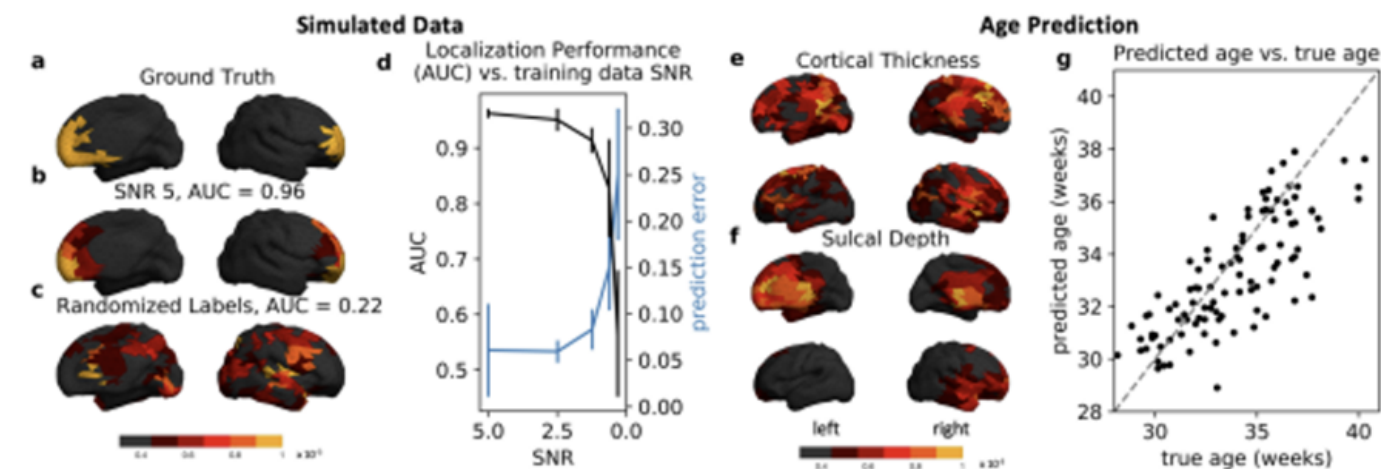
객체 영역 전체를 효과적으로 지역화할 수 있는 saliency map의 형태로 설명 방법 제시

“ 다른 영역에 비해서 픽셀값의 변화가 급격한 부분을 모아서 매핑한 후, 이를 강조하여 시각화 ”

사전 연구 : Regression

Regression activation mapping on the cortical surface using graph convolutional networks (2023)

- GCNs에 의해 식별된 중요 영역을 지역화하기 위한 회귀 활성화 매핑을 진행
 - 대뇌 표면 mesh의 정점 별 중요도 맵을 생성
- 대뇌 이미지와 그래프 개념을 결합하여, 이미지 상의 노드(node) 활성화를 측정



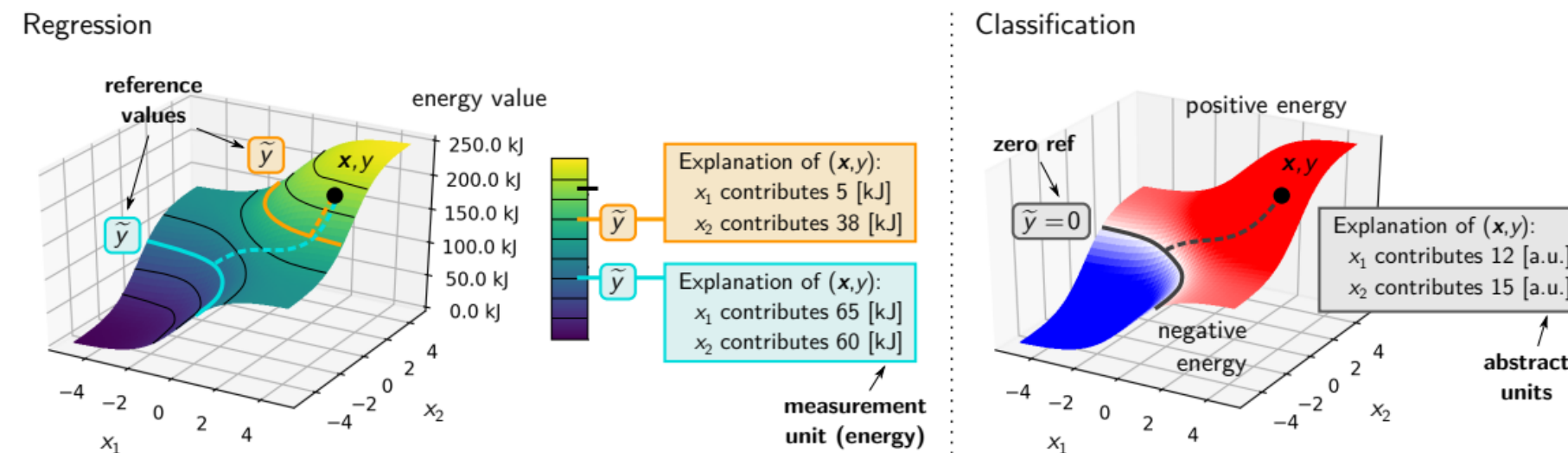
대뇌 이미지 상에 grid 형태로 node를 배치하여, 활성화된 node를 측정함으로써 중요도를 식별

“ 활성화된 node 아래 이미지 영역을 중요한 부분으로 간주 ”

사전 연구 : Regression

Toward Explainable AI for Regression Models (2023)

- Shapley value, Integrated gradient, LRP(Layer-wise Relevance Propagation)
- 위 방법론들을 통해 각 픽셀이나 특정 feature map이 모델의 예측에 기여하는 정도를 파악 가능
 - 일반적으로 회귀 모델의 실수 값 예측에는 분류보다 더 많은 정보가 포함되어 있음

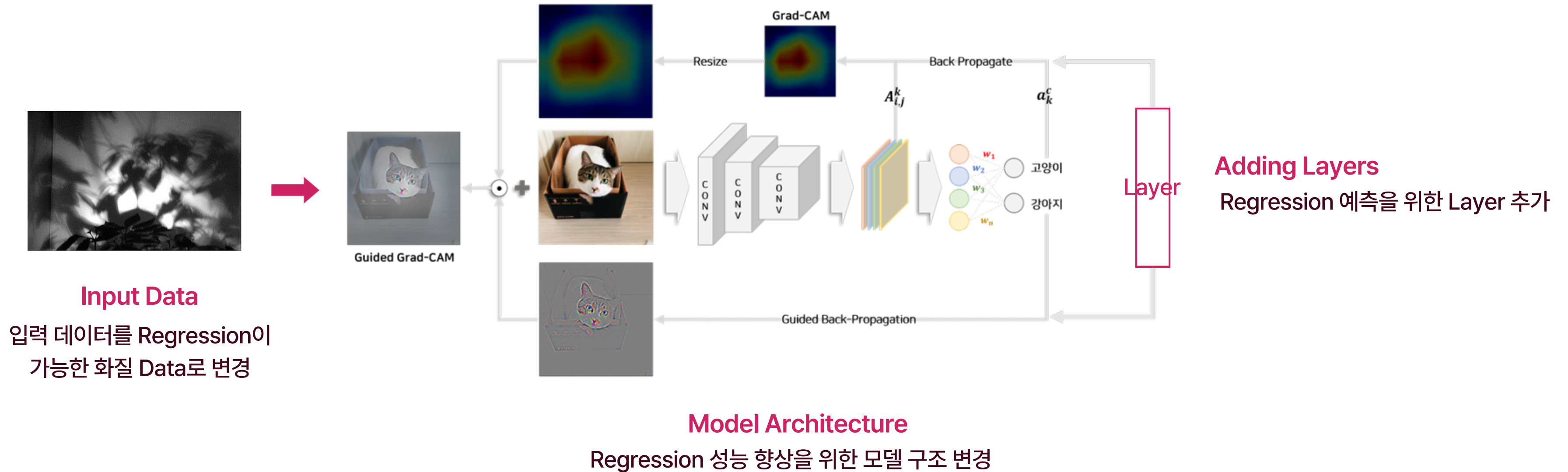


일반적으로 회귀 모델의 출력은 실수이므로 값의 단위 및 reference가 매우 중요

“ Classification 문제와 다르게, 출력값에 대한 사전 정보가 필요함을 강조 ”

연구 진행 상황

이미지 화질 데이터를 사용하여, 기존 GradCAM을 수정해 Regression XAI 진행

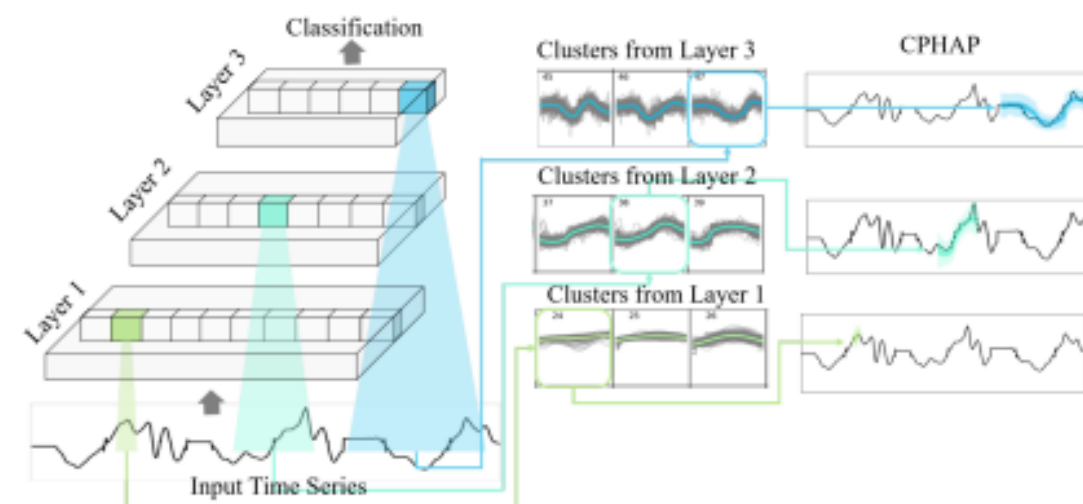


연구 방향

기존 모델에 Regression 예측을 위한
Layer를 추가



Time Series Data에 적용



Plan B

기존 방법론 결합을
통한 성능 개선

[Example]

- Encoder & Decoder로 재구성된 이미지, 모델이 이미지의 어떤 부분을 중요하게 여기는지를 나타내는 Saliency map과 결합
→ 모델이 어떤 부분을 주목하고, 어떻게 이해하는지 시각적으로 표현
- 데이터셋 전체에서 반복적으로 나타내는 패턴이나 특성 시각화 등

의사 결정 반영

예측 결과의 해석을 기반으로 하여
의사 결정에 반영할 수 있도록 함

Q & A

2023 Poster Session 중간 발표