



# A Survey on Explainable Artificial Intelligence (XAI): towards Medical XAI



이상용 / 2020-02-14



Computational Data Science LAB



# CONTENTS

1. Introduction
  2. Types of interpretability
- 
- 

# 01 | Introduction

- ✓ Deep Learning (DL)의 성공으로 인해 machine learning (ML)은 다양한 분야에서 크게 성장함
- ✓ 하지만, 특정 분야에서는 학습된 모델이 잠시라도 제대로 된 기능을 하지 못하면 치명적인 결과를 도래
- ✓ 만약 모델이 잘못되면 왜 잘못되는지 설명할 수 있어야 하고, 모델이 잘 작동하면 어떤 이유로 잘 작동하는지 알아야 함
- ✓ 이 논문은 일반적인 interpretability에 관련된 연구를 조사한 후, 의료분야에 동일한 범주를 적용함

## 02 | Types of interpretability

### Perceptive Interpretability

- Perceptive Interpretability는 인간이 인식할 수 있는 해석능력, '명확한' 것으로 간주되는 것을 포함하는 범주
- Saliency
  - ✓ 모델의 특정한 decision에 대해서 input components의 기여도를 할당하는 것
  - ✓ 주로 확률 또는 heat-map과 같은 super-pixels로 표현
- Signal method
  - ✓ 뉴런의 자극 또는 뉴런의 집합을 관찰
  - ✓ 뉴런의 활성화 된 값은 해석 가능한 형태로 변환 될 수 있음
- Verbal interpretability
  - ✓ 인간이 자연스럽게 파악할 수 있는 Verbal chunk의 형태
  - ✓ *If A then B*의 구조를 가지지만 항상 명확하지 않음 (천식을 앓는다 → 폐렴으로 인한 사망 위험이 적다)

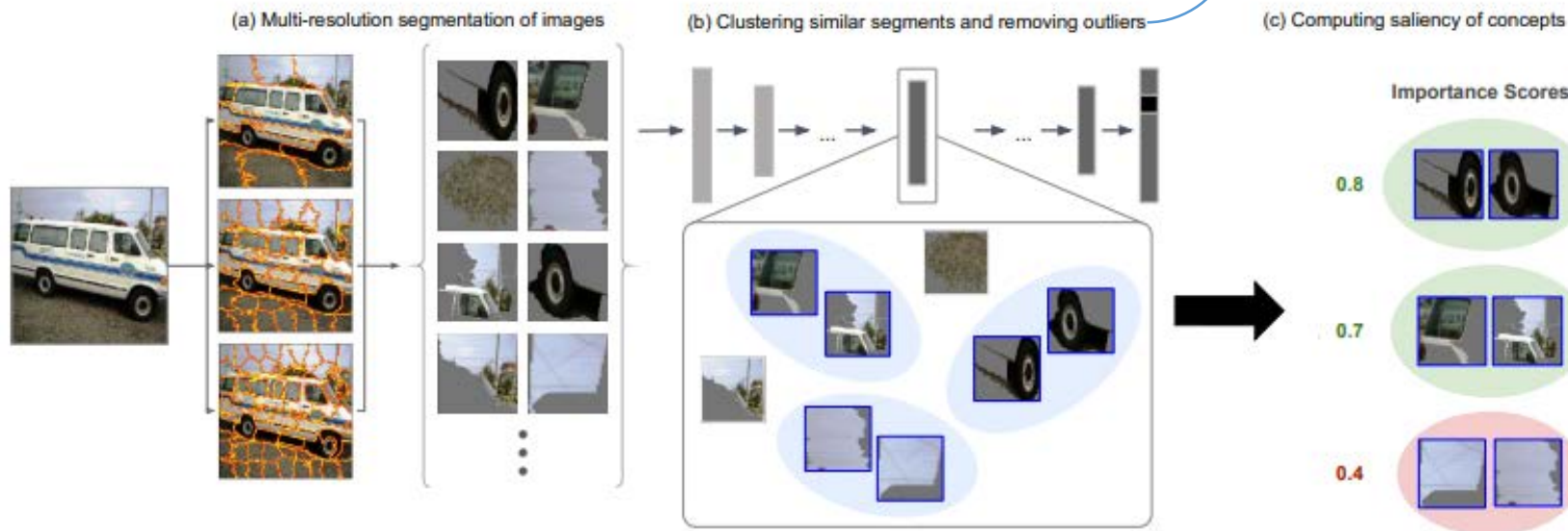
## 02 | Types of interpretability

### Perceptive Interpretability: Saliency

- Automatic Concept-based Explanations (ACE)

- ✓ ACE is a global explanation method that explains an entire class

Testing with Concept Activation Vectors (TCAV)

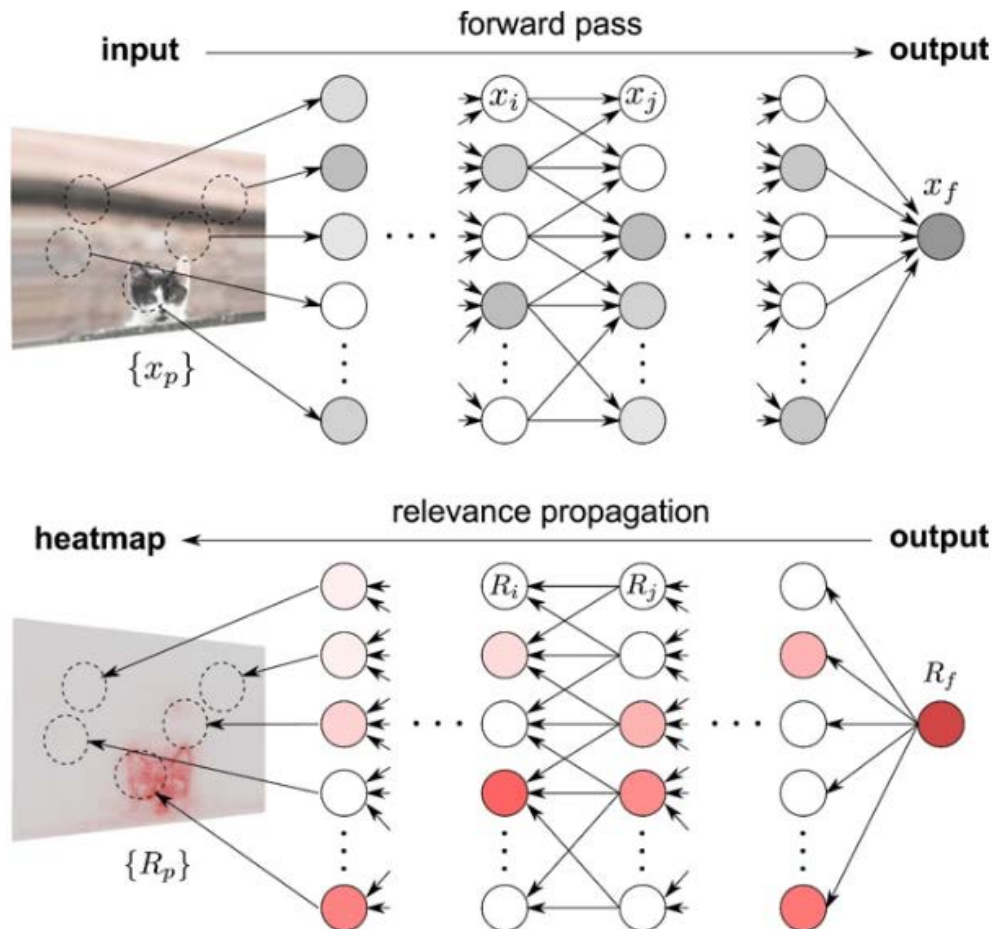


- ✓ 첫 번째 스텝 (a), 주어진 class images에 대해 segmentation
- ✓ 두 번째 스텝 (b), 동일한 concept으로 나누기 위해 유사한 segments를 그룹화
- ✓ 세 번째 스텝 (c), return important concepts (importance score 계산)

## 02 | Types of interpretability

### Perceptive Interpretability: Saliency

- layer-wise relevance propagation (LRP)



- ✓ Decomposition을 통한 explanation 방법
- ✓  $F(x)$ 를 얻기 위해 각 feature들이 기여하는 바를 계산

$$f(x) = b + \sum_d \alpha_d \phi_d(x_d)$$

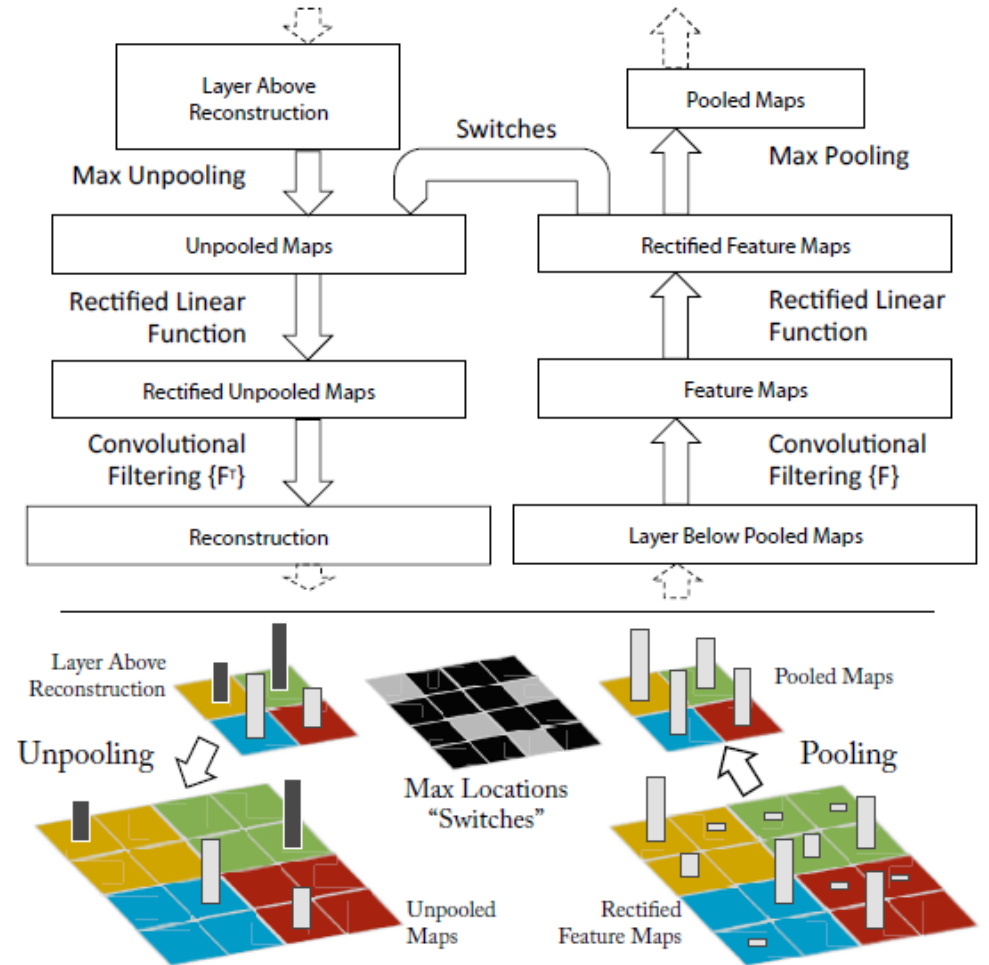
$$R_d^{(1)} = \begin{cases} f(x) \frac{|\alpha_d \phi_d(x_d)|}{\sum_d |\alpha_d \phi_d(x_d)|} & \text{if } \sum_d |\alpha_d \phi_d(x_d)| \neq 0 \\ \frac{b}{V} & \text{if } \sum_d |\alpha_d \phi_d(x_d)| = 0 \end{cases}$$

## 02 | Types of interpretability

### Perceptive Interpretability: Signal method

- Visualizing and Understanding Convolutional Networks (DeconvNet)

- ✓ 이미지로부터 특징이 어떻게 추출되고 학습되어 가는지 시각화를 통한 접근 방법
- ✓ Deconvolution이라는 개념을 사용하며
- ✓ Conv-ReLU-pooling(max) 과정의 반대로 수행
- ✓ Switch variables를 정의하여 max-pooling전에 이전 feature map에서 가장 큰 값들의 위치를 저장



# 02 | Types of interpretability

## Perceptive Interpretability

Methods	HSI	ANN	Mechanism		
LIME (Local Interpretable Model-agnostic Explanations) [30]	✓	✓	Optimization		
ACE (Automatic Concept-based Explanations) [32] uses TCAV [65]	✓	✓			
CAM uses global average pooling [33]	✗	✓			
Grad-CAM [34] generalizes CAM, utilizing gradient	✓	✓			
Guided Grad-CAM and Feature Occlusion [115]	✗	✓			
Respond CAM [35].	✗	✓			
Multi-layer CAM [98]	✗	✓			
**Listed elsewhere [52]	NA	NA			
LRP (Layer-wise Relevance Propagation) [12] [41]	✗	NA			
+ Image classifications. PASCAL VOC 2009 etc [36]	✗	✓			
+ Audio classification. AudioMNIST [37]	✗	✓			
+ LRP on DeepLight. fMRI data from Human Connectome Project. [38]	✗	✓			
+ LRP on CNN and on BoW(bag of words)/SVM [39]	✗	✓			
+ LRP on compressed domain action recognition algorithm [40]	✗	✗			
DeepLIFT [46]	✗	✓			
Prediction Difference Analysis [47]	✗	✓			
Slot Activation Vectors [31]	✗	✓			
Others. Also listed elsewhere: [105]	NA	NA			
+ Direct output labels. Training NN via multiple instance learning [49]	✗	✓			
+ Image corruption and testing Region of Interest statistically [50]	✗	✓			
+ Attention map with autofocus convolutional layer [51]	✗	✓			
DeconvNet [53]	✗	✓			
Inverting representation with natural image prior [54]	✗	✓			
Inversion using CNN [55]	✗	✓			
Activation maximization/optimization [27]	✗	✓			
Activation maximization on DBN (Deep Belief Network) [56]	✗	✓			
Activation maximization, multifaceted feature visualization [57]	✗	✓			
Visualization via regularized optimization [58]	✗	✓			
Semantic dictionary [28]	✗	✓			
Decision trees	NA	NA			
Propositional logic, rule-based [60]	✗	✗			
Sparse decision list [61]	✗	✗			
Decision sets, rule sets [62]	✓	✗			
Encoder-generator framework [63]	✗	✓			
Filter Attribute Probability Density Function [64]	✗	✗			



# Q&A

---

감사합니다.