Explainable Artificial Intelligence (XAI) - Concepts, taxonomies, opportunities and challenges toward responsible AI

HYU AI LAB 세미나

2021.06.21

조환희

목차

1. XAI?
2. XAI in Deep Learning
3. Challenges of XAI

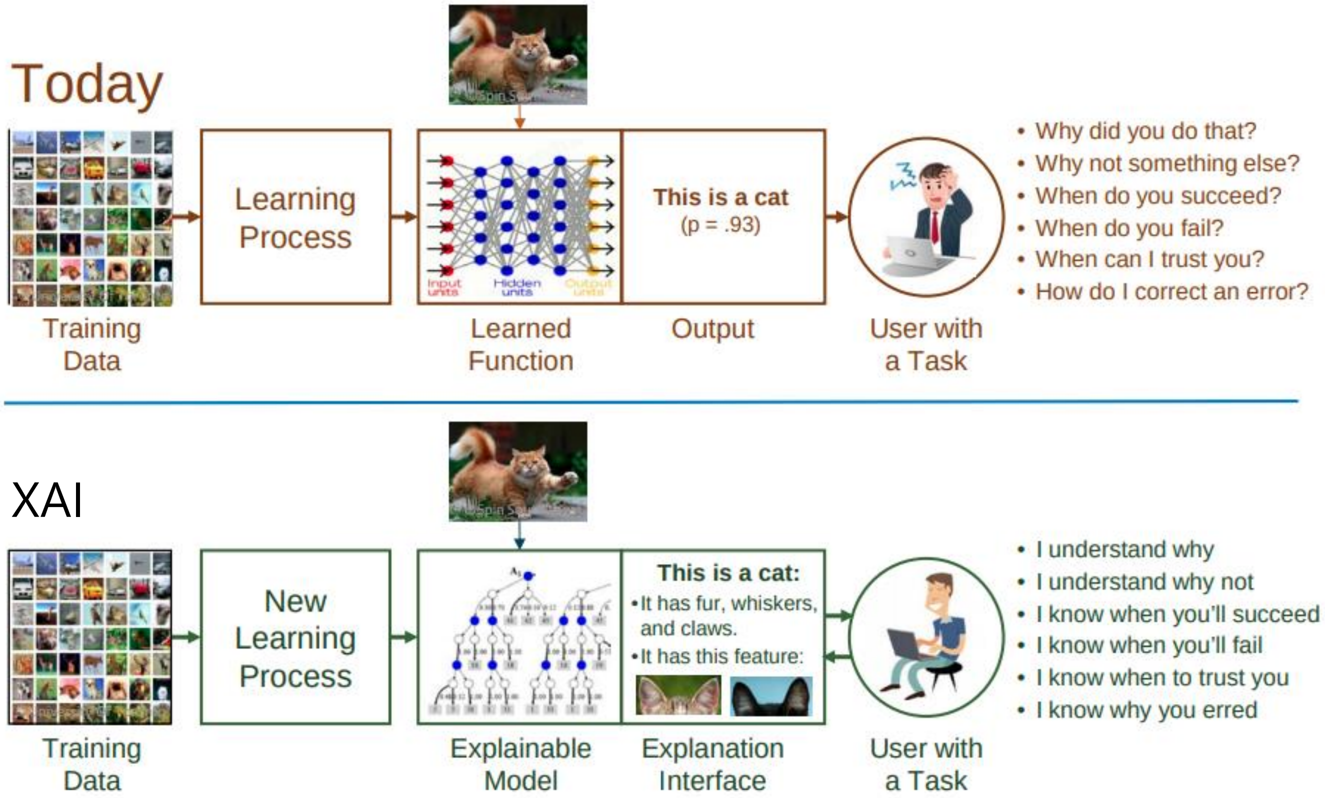
1. XAI?

XAI

=eXplainable Artificial Intelligence

: artificial intelligence (AI) in which the results of the solution can be understood by humans

1. XAI?


Today
Learning Process
Input units
Hidden units
Output units
This is a cat
(p = .93)
Training Data
Learned Function
Output
User with a Task
• Why did you do that?
• Why not something else?
• When do you succeed?
• When do you fail?
• When can I trust you?
• How do I correct an error?
XAI
New Learning Process
This is a cat:
•It has fur, whiskers, and claws.
•It has this feature:
Training Data
Explainable Model
Explanation Interface
User with a Task
• I understand why
• I understand why not
• I know when you'll succeed
• I know when you'll fail
• I know when to trust you
• I know why you erred

1. XAI?

Deep Learning

→ black-box

Increasingly being employed to make important predictions in critical contexts

Decisions are made not justifiable, legitimate, that do not allow obtaining detailed explanations

→ XAI가 필요하다!!

2. XAI in Deep Learning

How to implement XAI in Deep Learning?

: Transparent models and Post-hoc explainability

2. XAI in Deep Learning

1)Transparent models : convey interpretability by themselves

Linear/Logistic Regression

Decision Trees

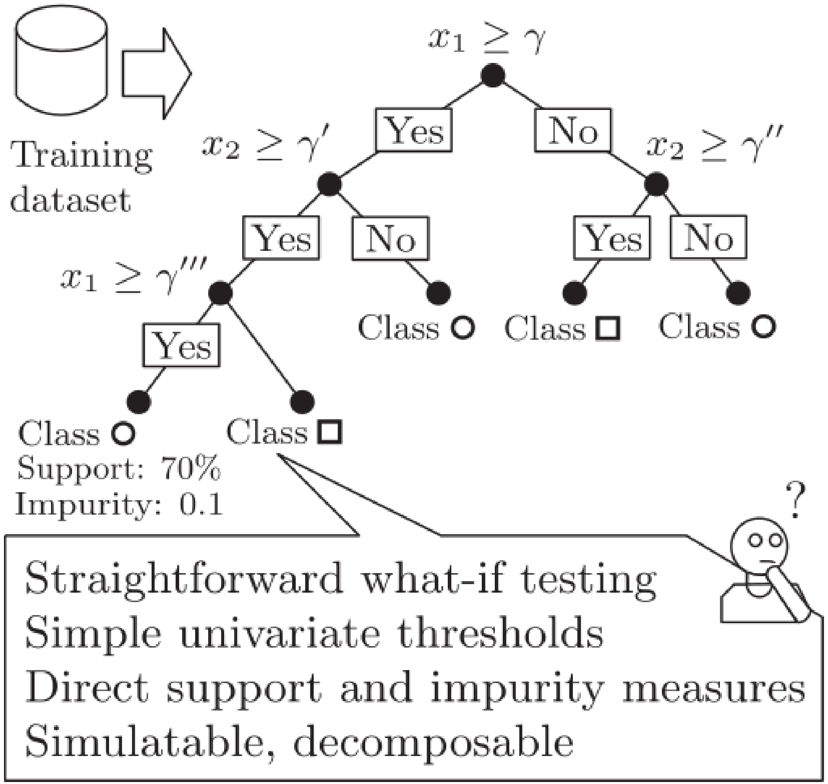
K-Nearest Neighbors

2. XAI in Deep Learning

ex) Decision Trees

features : x_1, x_2

thresholds : r, r′, r′′, …



→results can be understood by humans

2. XAI in Deep Learning

2) Post-hoc explainability : models that are not readily interpretable by desi gn by resorting to diverse means to enhance their interpretability

text explanations

visual explanations

local explanations

explanations by example

explanations by simplifications

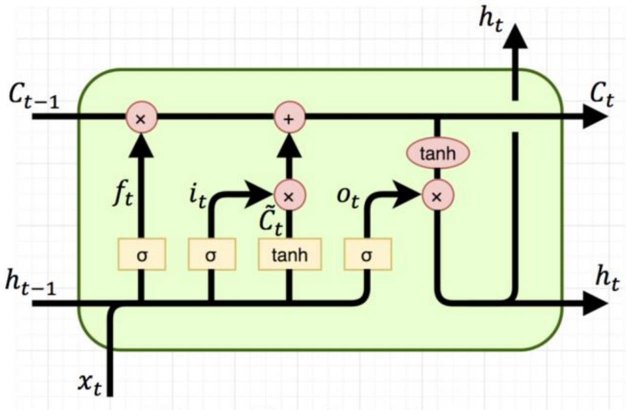
feature relevance explanations

2. XAI in Deep Learning

Visual Explanations : interpretable, long-range **LSTM cells**

LSTMs can use its **memory cells** to remember long-range informati on.

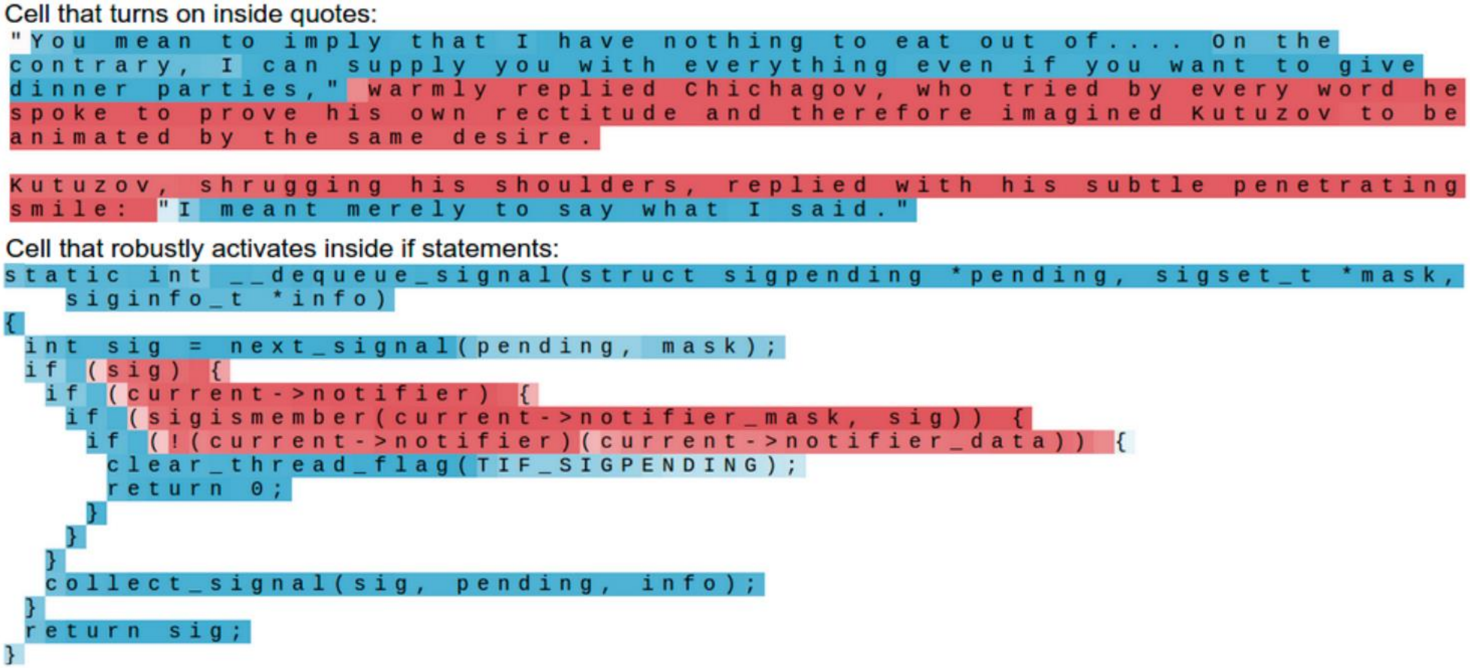
Demonstrate that LSTM learn powerful, and often interpretable long-range interactions on real-world data.



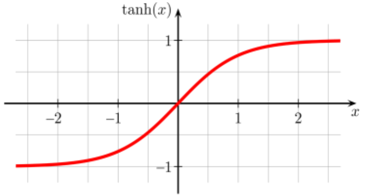
(a) Long Short-Term Memory

using tanh(c) for visualization

2. XAI in Deep Learning



tanh(c): -1 red
tanh(c): +1 blue



2. XAI in Deep Learning

feature relevance explanations : **Layer-Wise Relevance Backpropagation**

decompose the network classification decision into **contributions** of its **input elements**. → **Relevance Score**

Relevance score : score that how much pixel p contributes to explaining the classification decision, $R_p(\boldsymbol{x})$

2. XAI in Deep Learning

Consider each **neuron** as an object that can **be decomposed and expanded**.

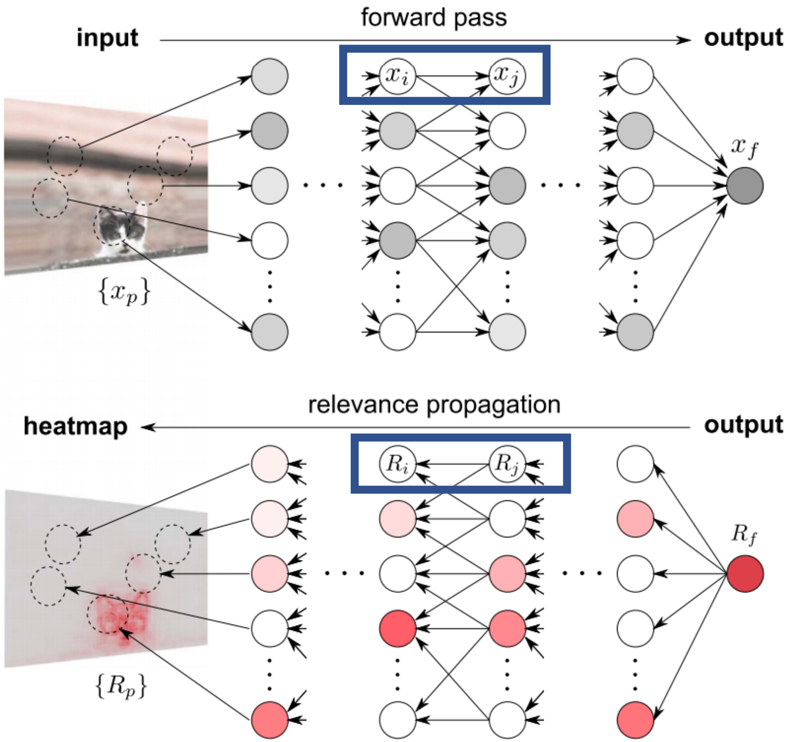
Taylor Decomposition

$$f(\boldsymbol{x}) = f(\widetilde{\boldsymbol{x}}) + \left(\frac{\partial f}{\partial \boldsymbol{x}}\Big|_{\boldsymbol{x}=\widetilde{\boldsymbol{x}}}\right)^{\top} \cdot (\boldsymbol{x} - \widetilde{\boldsymbol{x}}) + \varepsilon = 0 + \sum_{p} \underbrace{\frac{\partial f}{\partial x_p}\Big|_{\boldsymbol{x}=\widetilde{\boldsymbol{x}}} \cdot (x_p - \widetilde{x}_p)}_{R_p(\boldsymbol{x})} + \varepsilon$$

well-chosen *root point* : $f(\widetilde{\boldsymbol{x}}) = 0$.

the relevances $R_p(\boldsymbol{x})$ assigned to pixels in the image.

2. XAI in Deep Learning



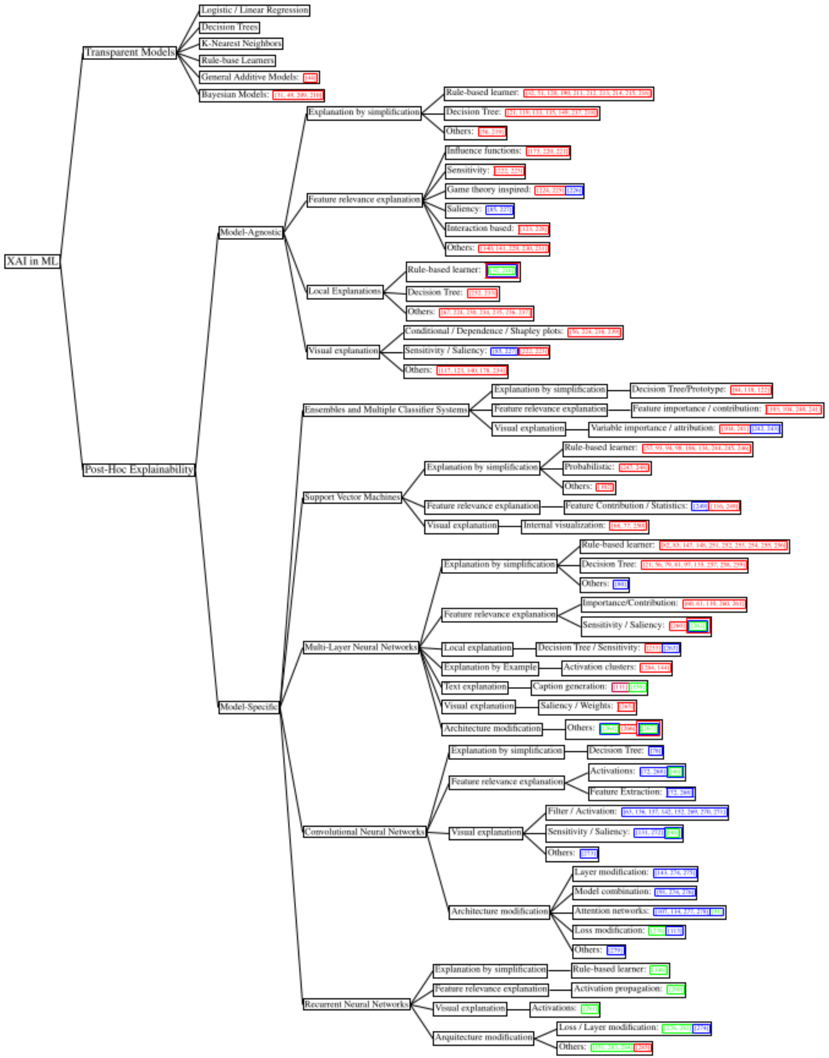
Decompose R_j on the set of lower layer neurons x_i which x_j is conne cted

$$R_j = \left(\frac{\partial R_j}{\partial\{x_i\}}\Big|_{\{\tilde{x}_i\}^{(j)}}\right)^{\mathsf{T}} \cdot (\{x_i\} - \{\tilde{x}_i\}^{(j)}) + \varepsilon_j = \sum_i \underbrace{\frac{\partial R_j}{\partial x_i}\Big|_{\{\tilde{x}_i\}^{(j)}} \cdot (x_i - \tilde{x}_i^{(j)})}_{R_{ij}} + \varepsilon_j$$

$$R_i = \sum_j R_{ij}.$$

2. XAI in Deep Learning

다양한 taxonomies



3. Challenges of XAI

1. Lack of agreement on the vocabulary and definitions.

2. Trade-off between interpretability and performance

3. Providing explanations that are accessible for society, p olicy makers and the law.

Thank you

References

A. Karpathy, J. Johnson, L. Fei-Fei, Visualizing and understanding recurrent networks, 2015, https://arxiv.org/pdf/1506.02078.pdf

S. Bach , A. Binder , G. Montavon , F. Klauschen , K.-R. Müller , W. Samek , On pixel–wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PloS one 10 (7) (2015) e0130140, https://www.sciencedirect.com/science/article/pii/S0031320316303582