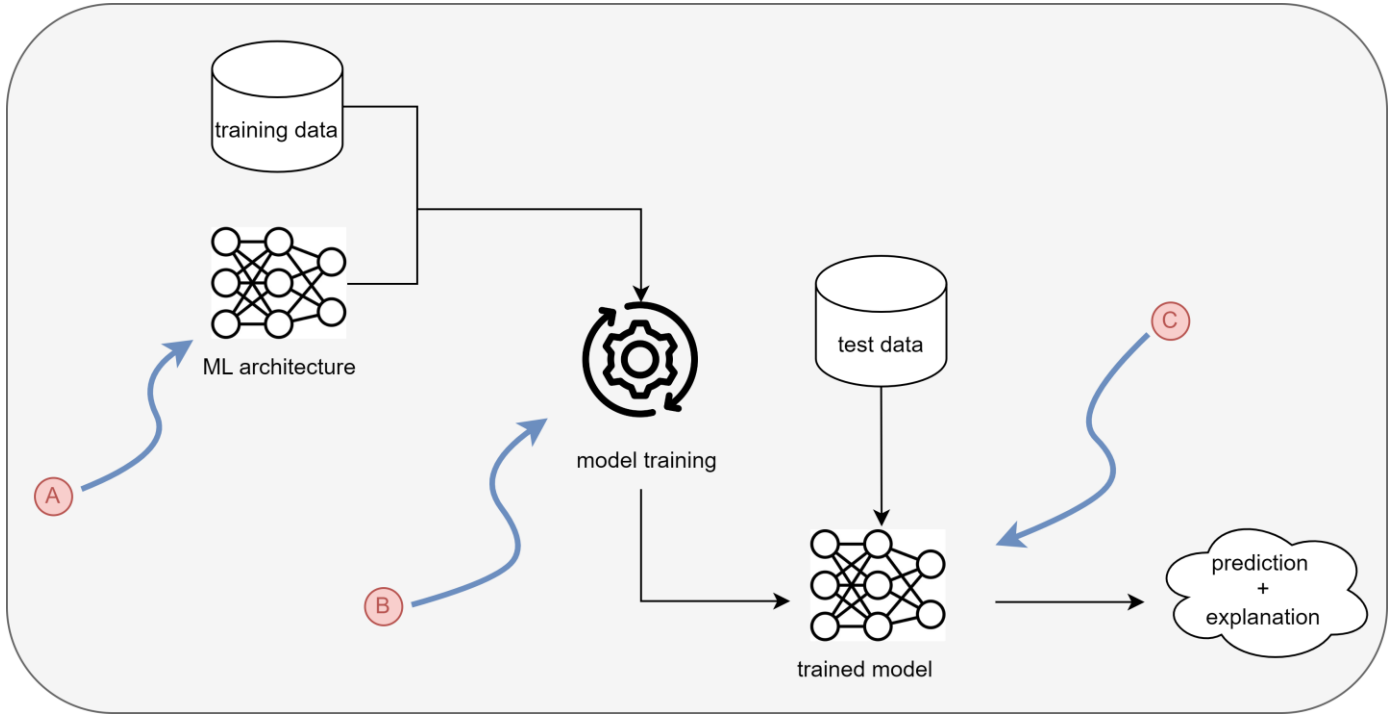
Evaluating Knowledge-Based XAI : A Focus on Plausibility and Similarity with Intrinsic Explanations

Rim El Cheikh, Issam Falih, Engelbert Mephu Nguifo

This work is funded by ANR-20-SFRI-0003

Knowledge-based XAI

- Integrates knowledge into the NN/XAI pipeline
- Anticipates human understandable explanatory elements
- Explanations belong to a vocabulary suitable for the user

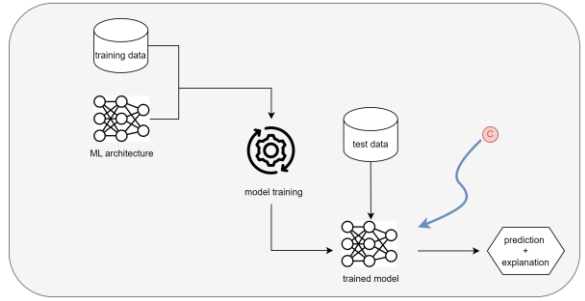


A At the design level

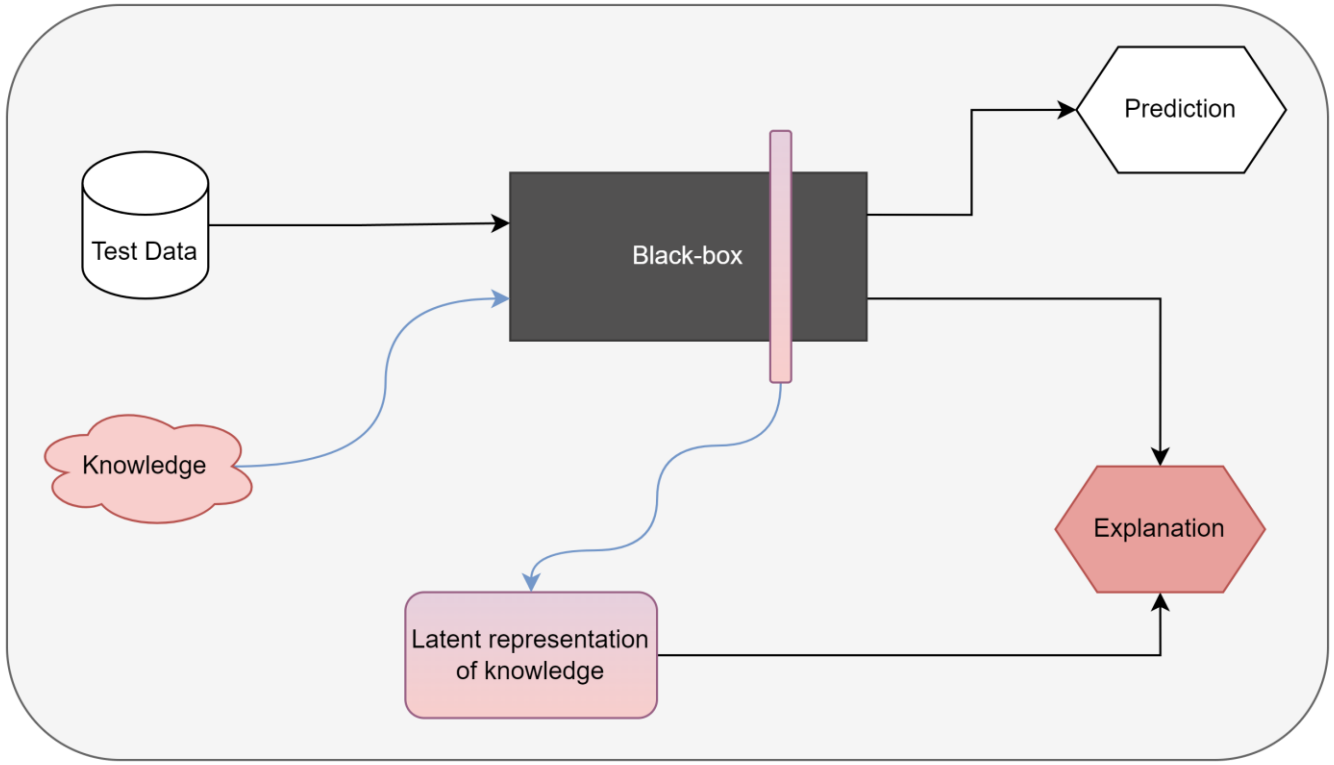
B At the training level

C At the post-hoc level

XAI with knowledge at the post-hoc level



- The most common
- Usually concept-based
- Examples : **TCAV**[1], CCE[2], CACE[3]



[1] Been Kim et al., Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), ICML 2018.
[2] Abubakar Abid et al., Meaningfully debugging model mistakes using conceptual counterfactual explanations, ICML 2022.
[3] Chih-Kuan Yeh et al., On completeness-aware concept-based explanations in deep neural networks, NIPS 2020.

Evaluating XAI methods

- Lack of standardized evaluation frameworks

- Existing approaches are not always adapted for knowledge-based methods
  - Explanations are scores that reflect class prediction sensitivity towards concepts

1. Plausibility

- How similar is the black-box decision process to human rationale ?

- We need a ground truth in terms of concepts scores

Evaluating knowledge-based XAI

1. Plausibility

- We use Osherson's class/attribute matrix[4] as ground truth
  - Classification for images of animals : Animals with Attributes dataset[5]
  - Provides association strength between a class and a concept according to **human users**

	black	white	blue	...	solitary	nestspot	domestic
antelope	-1.00	-1.00	-1.00	...	2.35	9.70	8.38
grizzly+bear	39.25	1.39	0.00	...	58.64	20.14	11.39
killer+whale	83.40	64.79	0.00	...	15.77	13.41	15.42
...	...	...	...	...	...	...	...
raccoon	63.57	43.10	0.00	...	35.95	28.26	5.00
cow	55.31	55.46	0.00	...	5.04	18.89	72.99
dolphin	10.22	21.53	27.73	...	3.96	14.05	37.98

[4] Daniel N Osherson et al., Default probability, Cognitive Science 1991.
[5] Yongqin Xian et al., Zero-Shot Learning - A Comprehensive Evaluation of the Good, the Bad and the Ugly, T-PAMI 2018.

2. Similarity with intrinsic explanations of an interpretable model

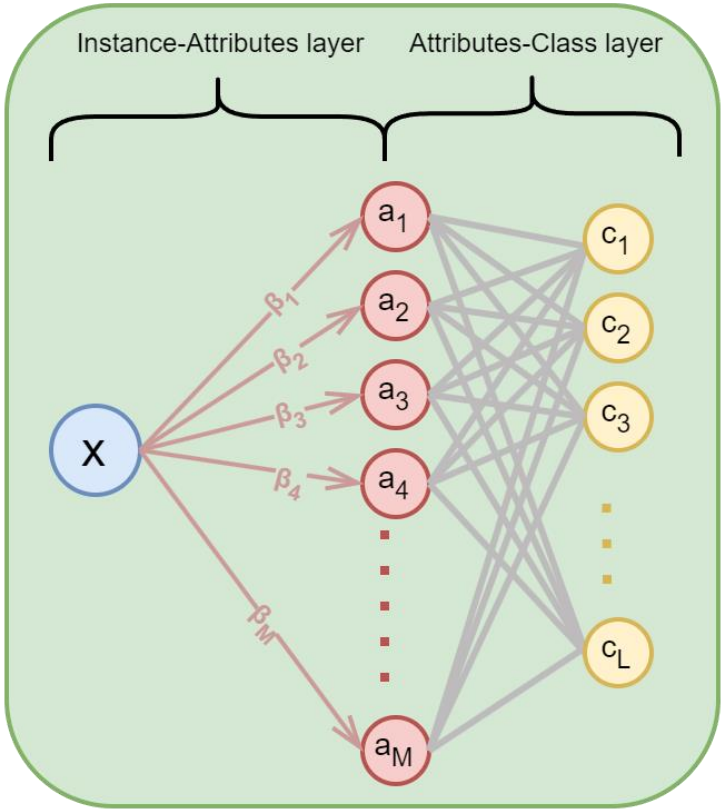
- Similarity between
  - XAI explanations of an interpretable/transparent model and,
  - Intrinsic explanations obtained by the same model

- Challenge :
  - “Classic transparent models” provide features scores as explanations

- We need an interpretable classifier that intrinsically provides concept importance scores for its predictions

2. Similarity with intrinsic explanations of an interpretable model

- We use Direct Attribute Prediction[6]
  - A classifier that computes class probabilities by passing through a layer of attributes/concepts
    - high-level semantically meaningful properties
  - The instance-attributes layer produces intrinsic explanations to the classifier's final output



[6] Christoph H Lampert et al., Attribute-based classification for zero-shot visual object categorization, T-PAMI 2013.

Thank you !
Happy to further discuss this work around the poster