



电子科技大学

University of Electronic Science and Technology of China

Can Interpretability Help?



Data Mining Lab, Big Data Research Center, USETC

Wei Han, wei.hb.han@gmail.com

1. Interpretability in Medical
2. Interpretability in Industrial
3. From Human to Model
4. From Model to Human
5. Discussion

0 Preliminary



• Interpretability of Neural Network

Dimension 1 — Passive vs. Active Approaches

Passive	Post-hoc explain trained neural networks
Active	Actively change the network architecture or training process for better interpretability

Dimension 2 — Type of Explanations (in the order of increasing explanatory power)

To explain a prediction/class by

Examples	Provide example(s) which may be considered similar or as prototype(s)
Attribution	Assign credit (or blame) to the input features (e.g. feature importance, saliency masks)
Hidden semantics	Make sense of certain hidden neurons/layers
Rules	Extract logic rules (e.g. decision trees, rule sets and other rule formats)

Dimension 3 — Local vs. Global Interpretability (in terms of the input space)

Local	Explain network's <i>predictions on individual samples</i> (e.g. a saliency mask for a input image)
Semi-local	In between, for example, explain a group of similar inputs together
Global	Explain the network <i>as a whole</i> (e.g. a set of rules/a decision tree)

0 Preliminary



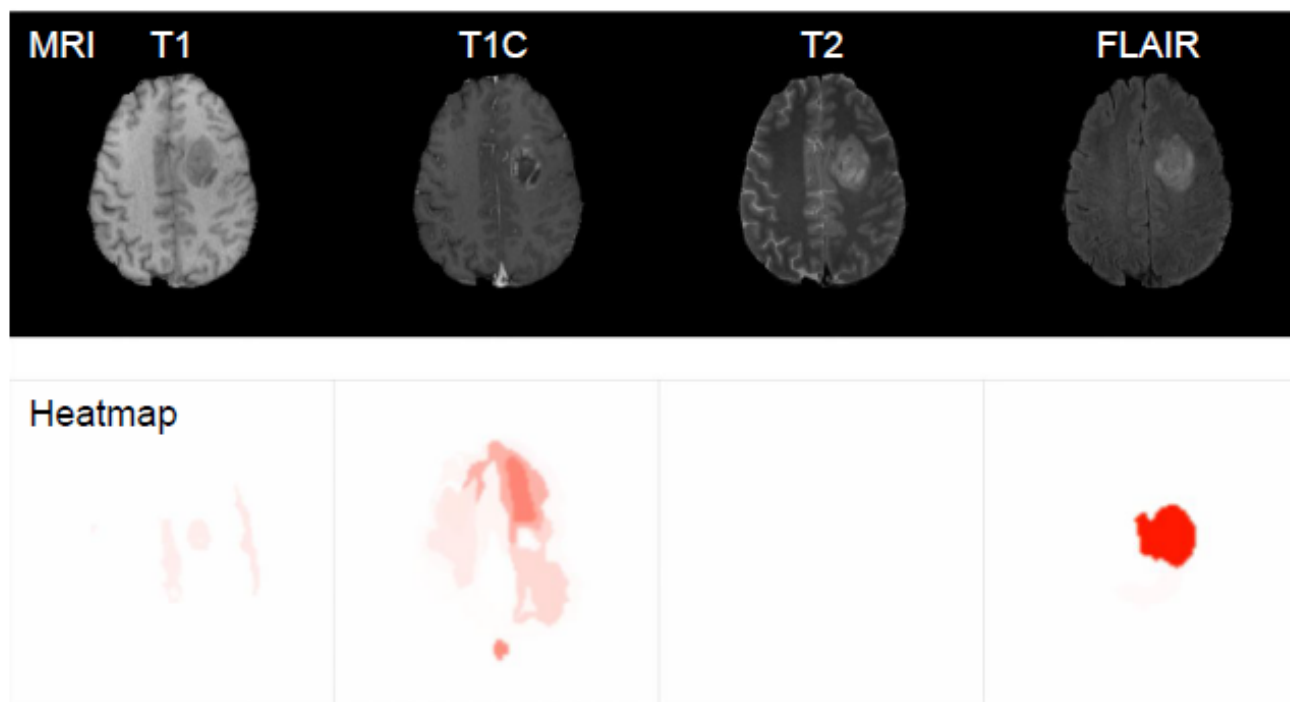
- Attribute-based Interpretability
 - Activation-based
 - e.g. CAM



1 Interpretability in Medical



- Glioma Grading Task [AAAI 2022]
 - Multi-modal medical imaging
 - Input-level multi-modal fusion



1 Interpretability in Medical



- Glioma Grading Task [AAAI 2022]
 - Modality priority
 - New METRICS

We extracted **clinical interpretation patterns of multi-modal explanations** using qualitative data analysis on clinicians' comments. When interpreting multi-modal images, physicians tend to **prioritize modalities** for a given task.

“Many of us just look at FLAIR and T1C. 90% of my time (interpreting the MRI) is on the T1C, and then I will spend 2% on each of the other modalities.”

Jin, W., Li, X., & Hamarneh, G. (2022). Evaluating Explainable AI on a Multi-Modal Medical Imaging Task: Can Existing Algorithms Fulfill Clinical Requirements?. AAAI

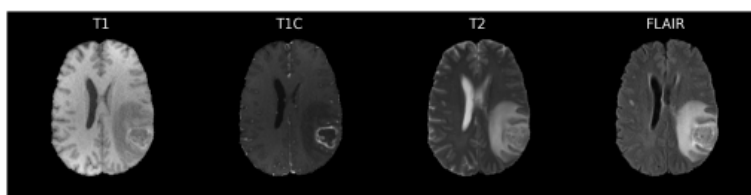
1 Interpretability in Medical



• Glioma Grading Task [AAAI 2022]

- MI: Coefficient of modality
- MSFI: Weighted sum

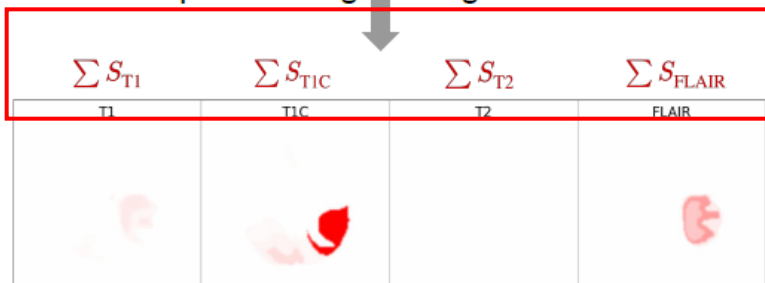
Modality Importance (MI) Correlation



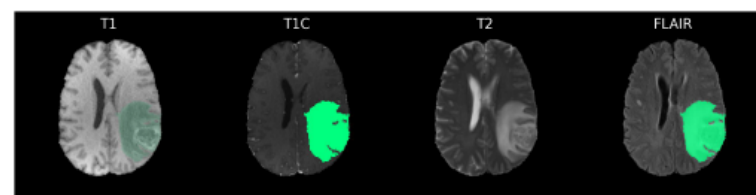
$$\varphi^{\text{mod}} = \begin{bmatrix} 0.1 & 0.5 & 0 & 0.4 \end{bmatrix}$$

compare ranking using Kendall's tau

Estimated MI
Heatmap

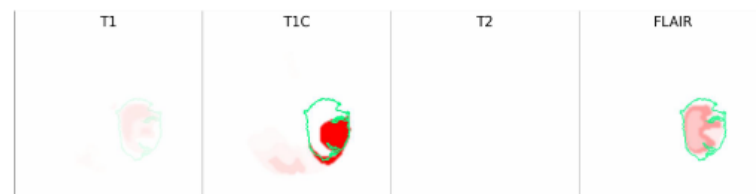


Modality-Specific Feature Importance (MSFI)



compare similarity using MSFI

$$\text{MSFI} = 0.1 \times \text{mask}_1 + 0.5 \times \text{mask}_2 + 0 \times \text{mask}_3 + 0.4 \times \text{mask}_4$$



1 Interpretability in Medical



- Glioma Grading Task [AAAI 2022]
 - What's pros

	MSFI (BraTS)	Stat. Sig.	MSFI (Synthetic)	MI Correlation	diffAUC	FP	IoU	Doctors' Rating	Speed (second)
Guided BackProp	0.48±0.33	NS	0.49±0.21	0.80±0.27	0.21±0.24	0.34±0.29	0.02±0.01	0.6±0.1	1.7±1.1
Guided GradCAM	0.50±0.36	★★	0.42±0.29	0.81±0.26	0.26 ± 0.25	0.37±0.31	0.02±0.02	0.1±0.0	2.2±1.4
InputXGradient	0.51±0.32	★	0.23±0.14	0.87±0.16	0.17 ± 0.12	0.40±0.30	0.08±0.05	0.1±0.0	1.7±1.1
DeepLift	0.54±0.34	★	0.22±0.23	0.53±0.45	0.19 ± 0.14	0.43±0.32	0.08±0.05	0.6±0.2	3.8±2.0
Integrated Gradients	0.48±0.31	★	0.22±0.19	0.73±0.39	0.17 ± 0.12	0.36±0.28	0.08±0.05	0.5±0.0	62±29
Occlusion	0.28±0.26	★★★	0.22±0.25	0.60±0.33	0.13 ± 0.15	0.18±0.19	0.03±0.02	0.6±0.2	989±835
Gradient Shap	0.48±0.31	★	0.22±0.19	0.53±0.40	0.17 ± 0.12	0.36±0.28	0.08±0.05	0.5±0.0	6.8±3.0
Feature Ablation	0.48±0.30	★★★	0.19±0.23	0.27±0.44	0.30 ± 0.15	0.35±0.28	0.05±0.06	0.4±0.4	74±23
Gradient	0.34±0.23	NS	0.19±0.13	0.47±0.16	0.05 ± 0.09	0.20±0.16	0.02±0.01	0.6±0.6	1.8±1.1
Shapley Value Sampling	0.38±0.24	★★★	0.10±0.10	0.47±0.65	0.35 ± 0.04	0.25±0.21	0.04±0.05	0.2±0.1	2018±654
Kernel Shap	0.28±0.25	★★	0.08±0.08	NaN	0.26 ± 0.16	0.18±0.20	0.06±0.08	0.1±0.0	194±100
Feature Permutation	0.23±0.26	NS	0.08±0.07	NaN	0.05 ± 0.05	0.13±0.18	0.05±0.07	0.1±0.0	14±2.2
Lime	0.24±0.21	★★	0.05±0.07	0.53±0.58	0.37 ± 0.08	0.14±0.16	0.05±0.06	0.1±0.0	341±181
Deconvolution	0.26±0.23	NS	0.04±0.02	0.73±0.39	0.11 ± 0.21	0.17±0.17	0.02±0.01	0.4±0.4	1.8±1.0
Smooth Grad	0.27±0.17	★	0.03±0.02	0.67±0.00	0.29 ± 0.25	0.16±0.12	0.02±0.01	0.7±0.1	12±6
GradCAM	0.04±0.03	★★★	0.02±0.02	NaN	0.16 ± 0.19	0.02±0.01	0.02±0.01	0.0±0.0	0.6±0.3

1 Interpretability in Medical



- Glioma Grading Task [AAAI 2022]
 - What's cons
 - Fairness and effectiveness of indicators
 - Depth and width of the research

Shapley value-based MI ground-truth We define the modality Shapley value φ_m to be the ground truth MI value for a modality m . It is calculated as:

$$\varphi_m(v) = \sum_{c \subseteq \mathcal{M} \setminus \{m\}} \frac{|c|!(M - |c| - 1)!}{M!} (v(c \cup \{m\}) - v(c)),$$

1 Interpretability in Medical



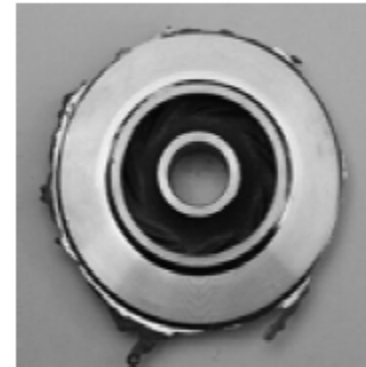
- Discussion

- Width of data sets and models
- Deep participation of doctors
- Interaction between doctor and system
- Large scale practical application

2 Interpretability in Industrial



- Defect Detection [AAAI 2022]
 - Reasonable prediction
 - Interactive human-in-the-loop approach



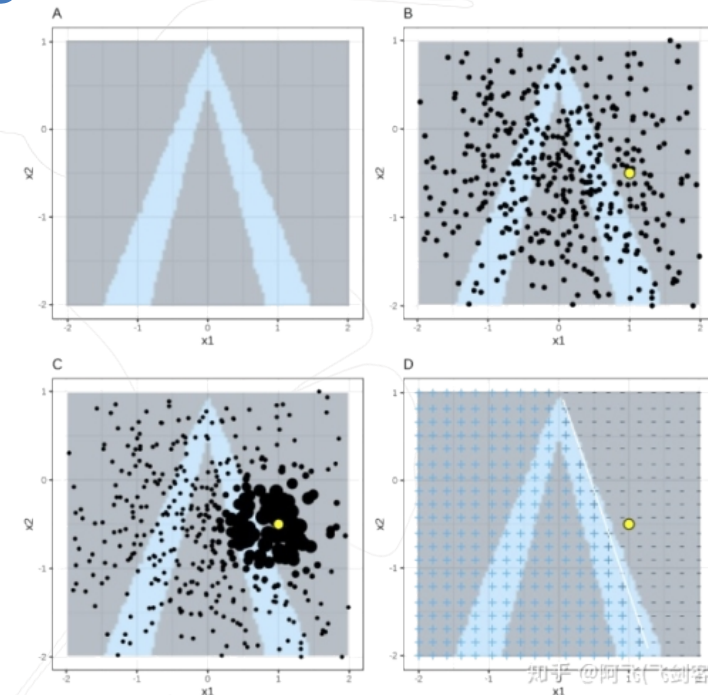
Müller, D., März, M., Scheele, S., & Schmid, U. (2022). An Interactive Explanatory AI System for Industrial Quality Control. AAAI

2 Interpretability in Industrial



- LIME (Local Interpretable Model-Agnostic Explanations)

- Perturbation-based



- ILP (Inductive Logic Programming)

- Rule-based

吸收(absorption):

辨识(identification):

内构(intra-construction):

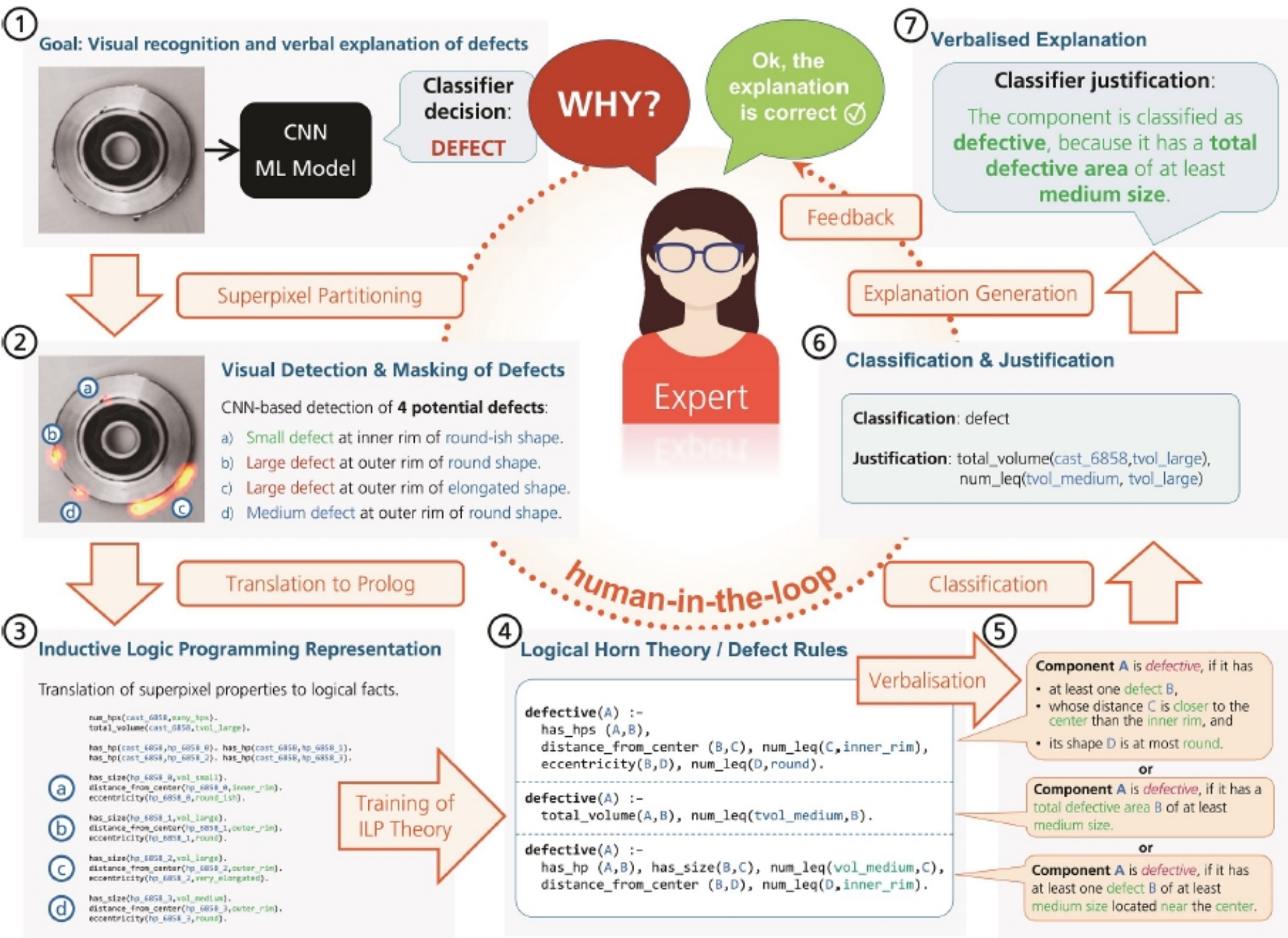
互构(inter-construction):

$$\frac{p \leftarrow A \wedge B \quad q \leftarrow A}{p \leftarrow q \wedge B \quad q \leftarrow A}$$

$$\frac{p \leftarrow A \wedge B \quad p \leftarrow A \wedge q}{q \leftarrow B \quad p \leftarrow A \wedge q}$$

$$\frac{p \leftarrow A \wedge B \quad p \leftarrow A \wedge C}{q \leftarrow B \quad p \leftarrow A \wedge q \quad q \leftarrow C}$$

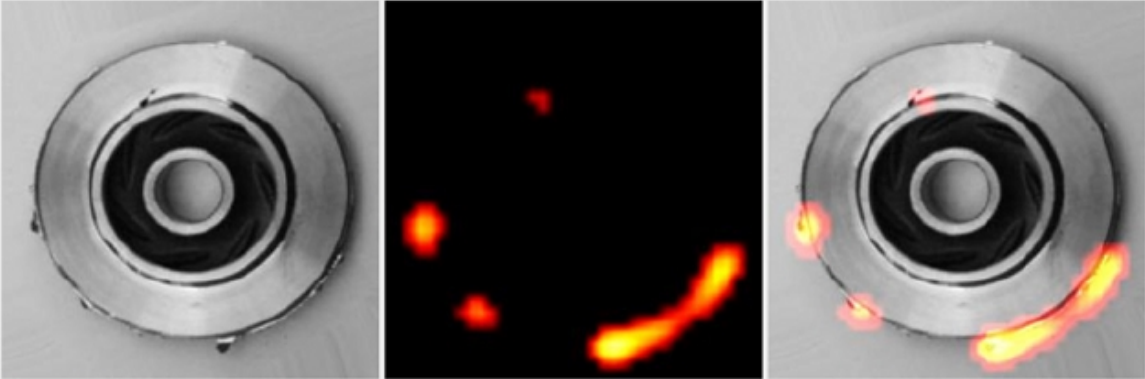
$$\frac{p \leftarrow A \wedge B \quad q \leftarrow A \wedge C}{p \leftarrow r \wedge B \quad r \leftarrow A \quad q \leftarrow r \wedge C}$$



2 Interpretability in Industrial



localhost:8080



4 potential defects (many_hpx)
Total defective area: 7352.7568627450055 (tvol_large)

1. Volume: 153.874509803922 (vol_small), Eccentricity: 0.7651099649246286 (round_ish), Distance from center: 75.03472113231672 (inner_rim)
2. Volume: 1134.725490196076 (vol_large), Eccentricity: 0.5533964261233965 (round), Distance from center: 121.75694714676452 (outer_rim)
3. Volume: 5446.6392156861875 (vol_large), Eccentricity: 0.9833548917696631 (very_elongated), Distance from center: 129.83038516676203 (outer_rim)
4. Volume: 617.5176470588204 (vol_medium), Eccentricity: 0.692298609348484 (round), Distance from center: 122.7094752336706 (outer_rim)

Classification: defective ☒ Correct Save & Next >

Explanation:

Correct: ☒ The total defect-volume is large

Current rules for defective components:
A component is considered defective, iff:

1. There is a defect, whose location is at most on the inner rim and whose eccentricity is at most round
2. The total defect-volume is at least medium
3. There is a defect, whose size is at least medium and whose location is at most on the inner rim

2 Interpretability in Industrial



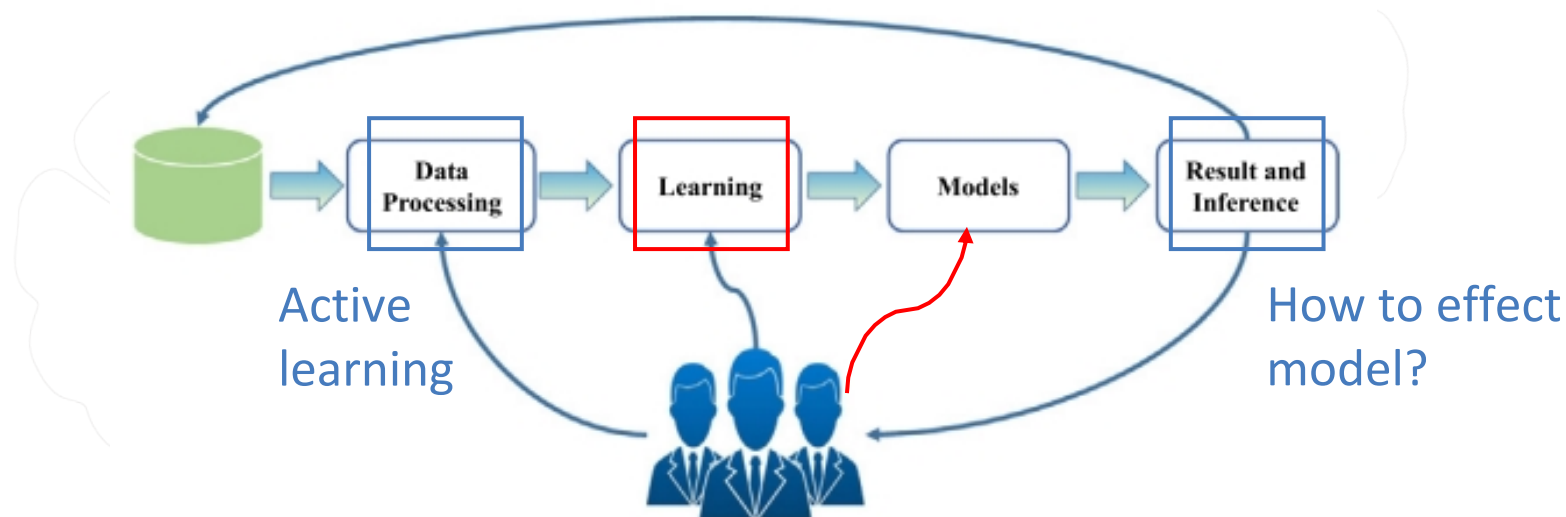
- Discussion

- System components are independent
- Human-in-the-loop in name, but just labeling
- Can interpretability help?
 - Trust, faithfulness, robustness
 - From interpretability to model

3 From Human to Model



- Human-in-the-loop
 - Towards performance & trust
 - Involve additional info into model training



3 From Human to Model



- Natural Language Inference [AAAI 2022]
 - Reasoning with textual corpus

Premise:

Wet brown **dog swims** towards camera.

Hypothesis:

A **dog** is **sleeping** in his bed.

Explanation for contradiction class:

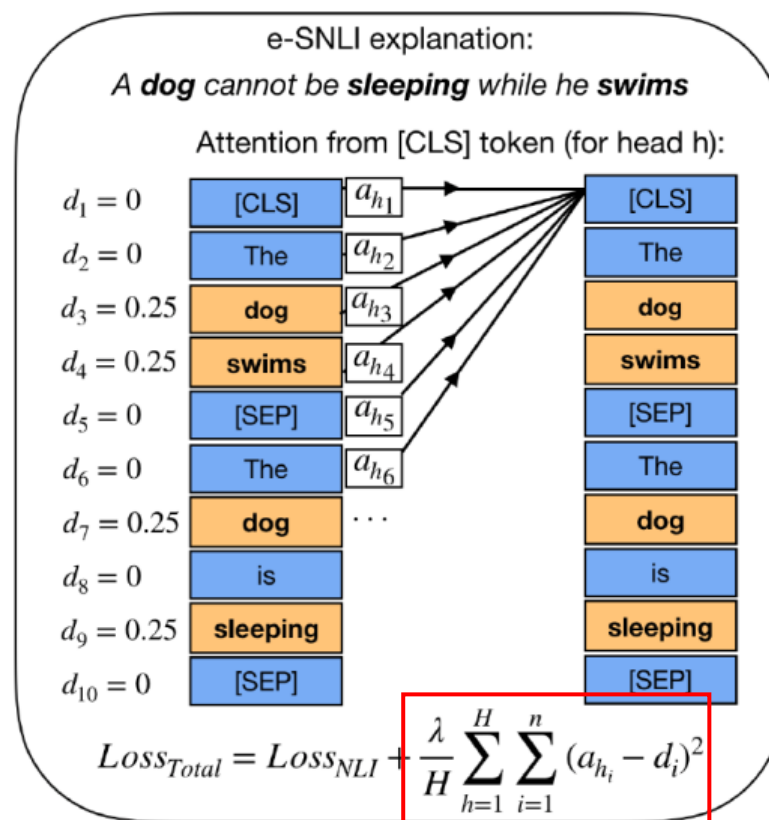
A **dog** cannot be **sleeping** while he **swims**.

Stacey, J., Belinkov, Y., & Rei, M. (2022). Supervising model attention with human explanations for robust natural language inference. AAAI

3 From Human to Model



- Natural Language Inference [AAAI 2022]
- Additional loss from manual labeling

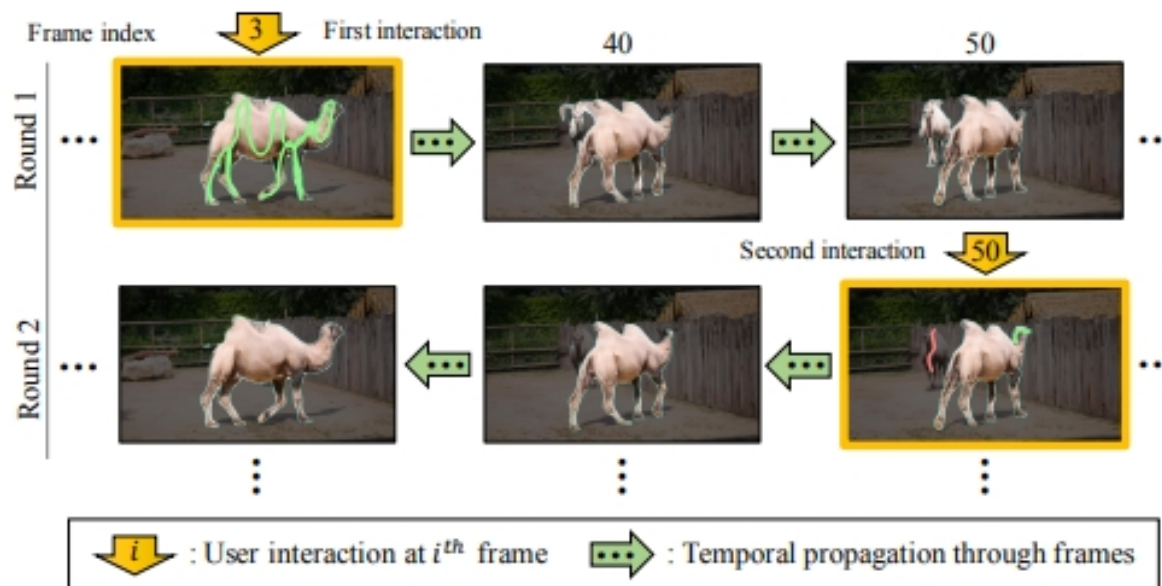


How much labor cost is required?

3 From Human to Model



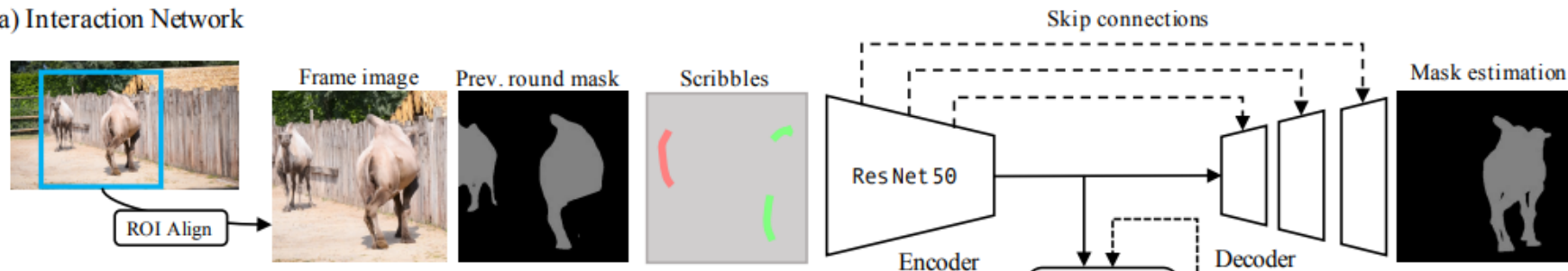
- Video Object Segmentation [CVPR 2019]
 - Video with sequential frames
 - Human scribbles & temporal propagation



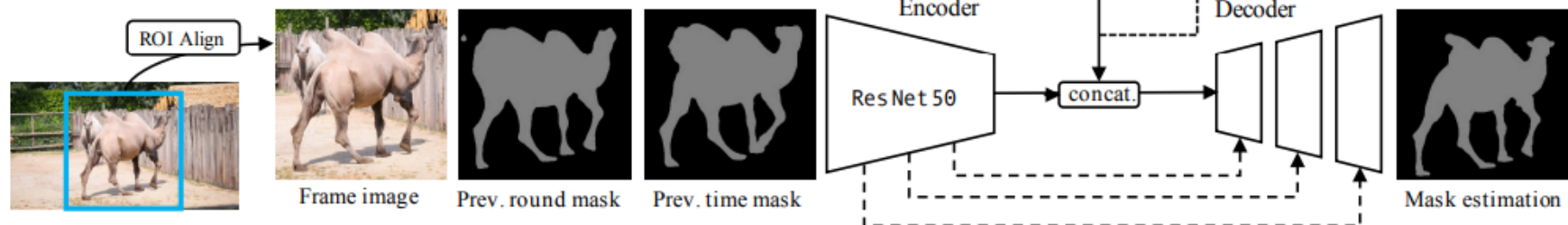
Oh, S. W., Lee, J. Y., Xu, N., & Kim, S. J. (2019). Fast user-guided video object segmentation by interaction-and-propagation networks. CVPR

3 From Human to Model

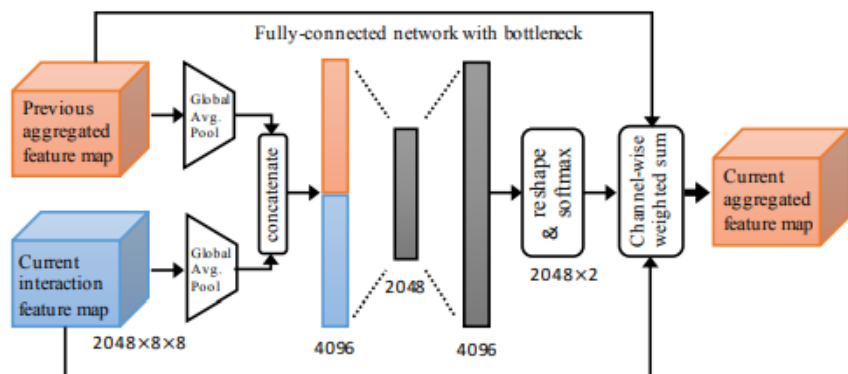
(a) Interaction Network



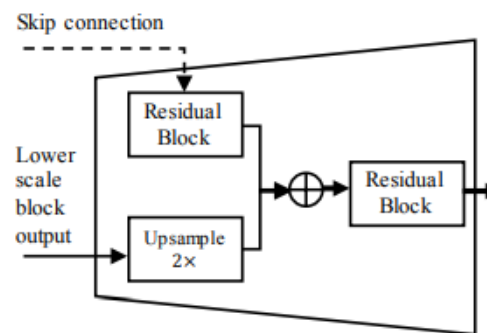
(b) Propagation Network



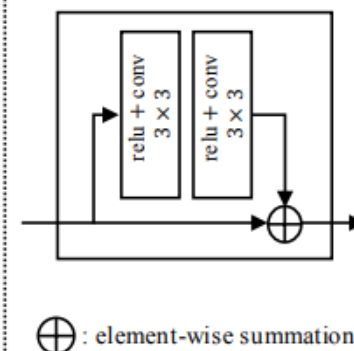
(c) Feature Aggregation Module



(d) Decoder block



(e) Residual Block



\oplus : element-wise summation

3 From Human to Model

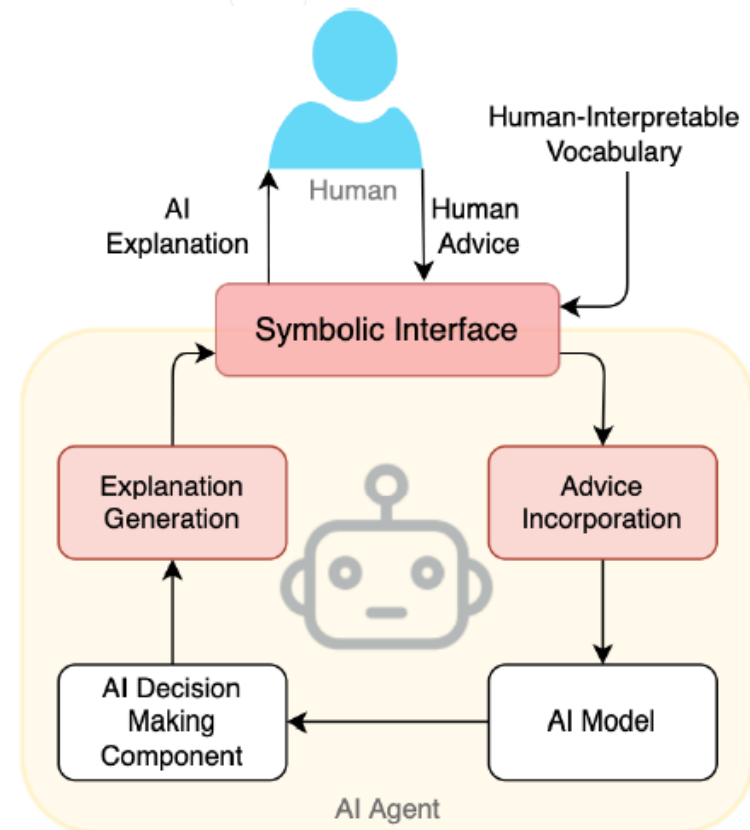


- Discussion
 - Human-in-the-loop for performance
 - Interpret human for model
 - Applicable? Scalable? Interpretable?
 - Independent elements

4 From Model to Human



- Symbols for Bridging Human-AI Chasm [AAAI 2022]
 - Research proposal
 - Potential challenges

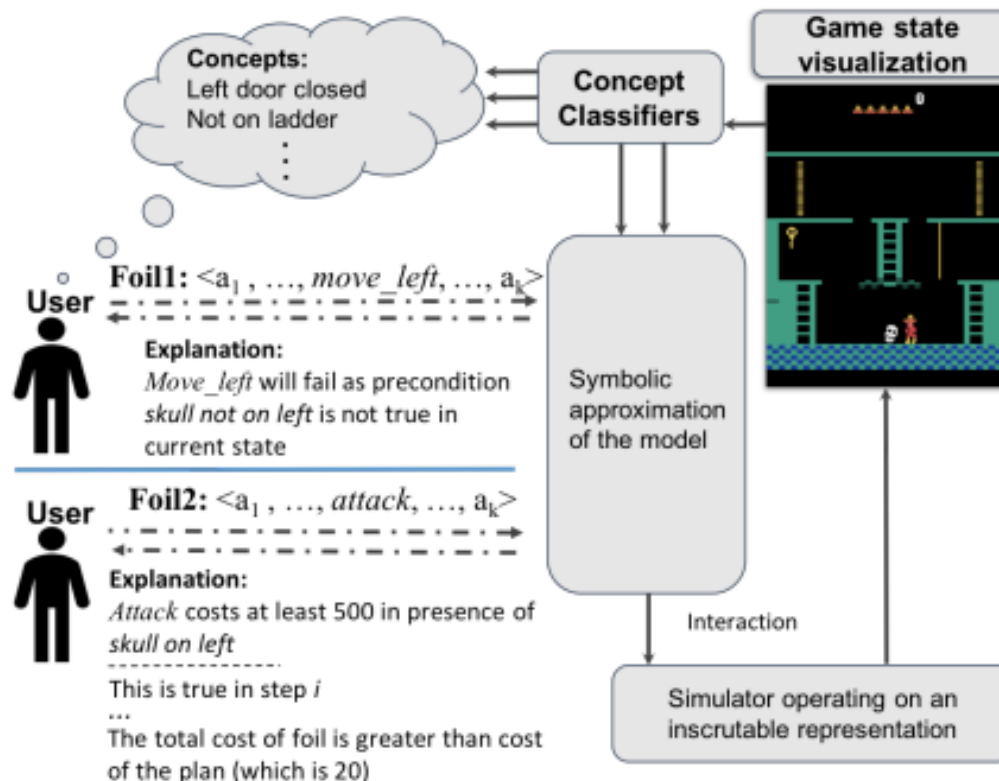


Kambhampati, S., Sreedharan, S., Verma, M., Zha, Y., & Guan, L. (2022). Symbols as a lingua franca for bridging human-ai chasm for explainable and advisable ai systems. AAAI

4 From Model to Human



- Sequential Decision-Making [ICLR 2022]
 - Collect predefined concept & Classify
 - Explain failing precondition & cost



4 From Model to Human



- Potential challenges [AAAI 2022]
 - Approximating BOTH explanations
 - Assembling the symbolic interface
 - getting the symbolic vocabulary
 - grounding it in the model representations
 - Expand the symbolic vocabulary

5 Discussion



- What function network
 - Represent *func* of each neuron into embedding space
 - Activate only when processing related *func*
 - *Func Module* for friend interpretability and robustness
- Why function network
 - Embodied neuron func & Precise control
 - General 'language' between machines
 - Adaptively expand concept

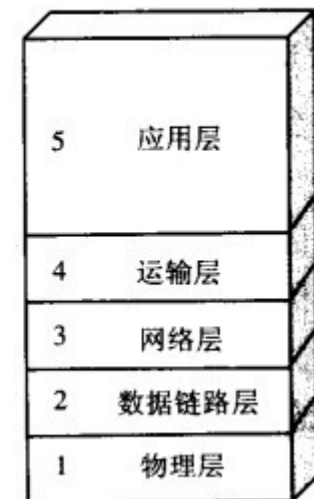
5 Discussion



• Promising Framework

- Model interpretability (机器码/物理层)
- Mapping/Matching (驱动/数据链路层)
- Knowledge graph (路由/网络层)
- NLP/CV/Audio trans. (运输层)
- Human-friendly interpret. (应用层)

五层协议的体系结构



5 Discussion



- Benefits

- Approximating BOTH ex

- Assembling the symbolic

- getting the symbolic

- grounding it in the n

- Expand the symbolic vo

1. **Constant memory.** To avoid unbounded systems, the consumed memory should be constant w.r.t. the number of tasks or length of the data stream.
2. **No task boundaries.** Learning from the input data without requiring clear task divisions makes continual learning applicable to any never-ending data stream.
3. **Online learning without demanding offline training** of large batches or separate tasks introduces fast acquisition of new information.
4. **Forward transfer or zero-shot learning** indicates the importance of previously acquired knowledge to aid the learning of new tasks by increased data efficiency.
5. **Backward transfer** aims at retaining previous knowledge and preferably improving it when learning future related tasks.
6. **Problem agnostic continual learning** is not limited to a specific setting (e.g. only classification).
7. **Adaptive systems** learn from available unlabeled data as well, opening doors for adaptation to specific user data.
8. **No test time oracle** providing the task label should be required for prediction.
9. **Task revisiting** of previously seen tasks should enable enhancement of the corresponding task knowledge.
10. **Graceful forgetting.** Given an unbounded system and infinite stream of data, selective forgetting of trivial information is an important mechanism to achieve balance between stability and plasticity.

5 Discussion



- What's the problem
 - Design of architecture
 - Design of losses
 - Performance
 - Interpretability?
 - Evaluation

5 Discussion



- Time Schedule

- End of Nov.
 - Interpretability & continual learning pipeline
- End of Dec.
 - Experiments or Re-design
- Beginning of Jan.
 - Writing & Polishing



电子科技大学

University of Electronic Science and Technology of China



Thanks



Data Mining Lab, Big Data Research Center, USETC

Wei Han, wei.hb.han@gmail.com