



Concept-based Explanations for Out-Of-Distribution Detectors

Jihye Choi¹, Jayaram Raghuram¹, Ryan Feng², Jiefeng Chen¹,
Somesh Jha¹, Atul Prakash²

¹ University of Wisconsin-Madison ² University of Michigan

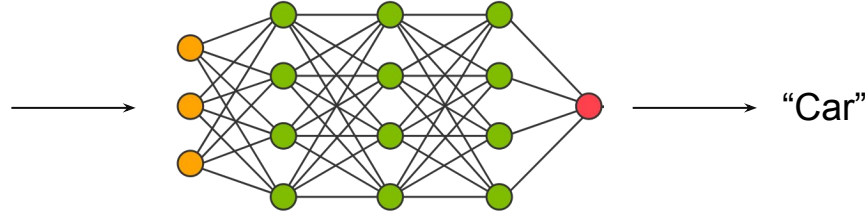


Standard Machine Learning (ML) Models

Training



ML model in self-driving car



"Car"

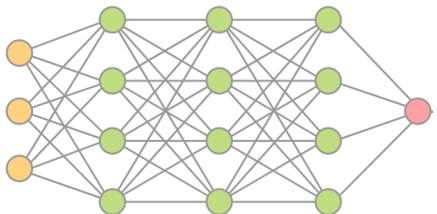


Standard Machine Learning (ML) Models

Training



ML model in self-driving car

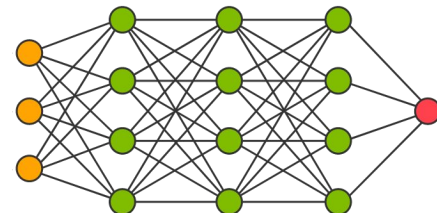


“Car”

Testing



Out-of-distribution
(OOD)



“Car”

Overconfident prediction
for unseen object “Buffalo”

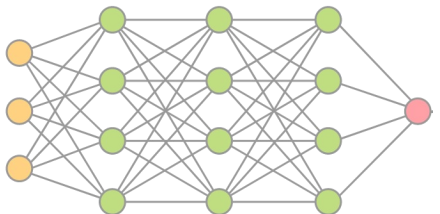


Out-Of-Distribution Detection

Training



ML model in self-driving car

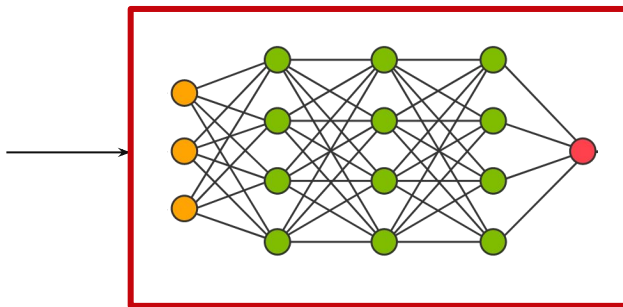


“Car”

Testing



Out-of-distribution
(OOD)

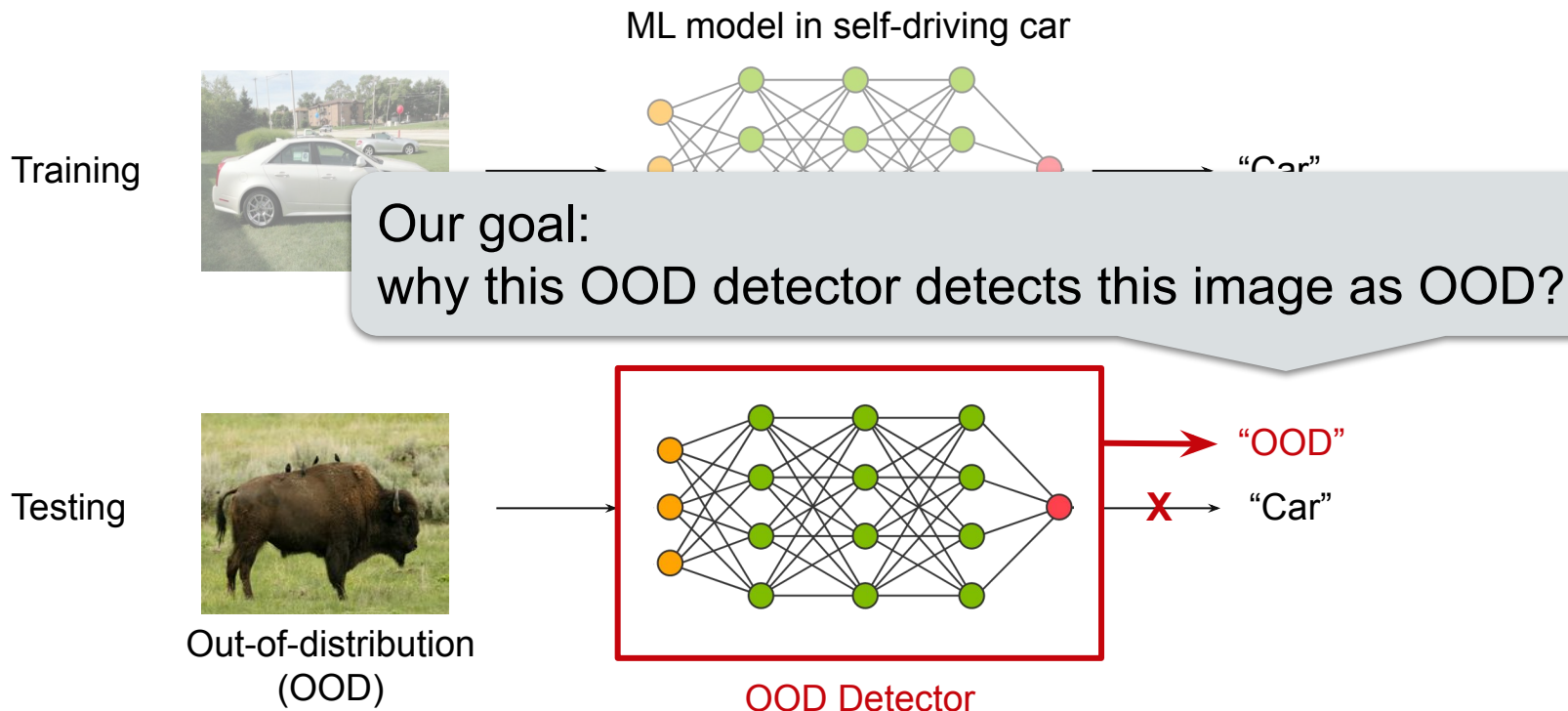


OOD Detector

“OOD”
X
“Car”



Understanding OOD Detection

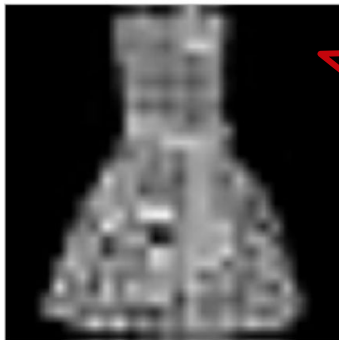




ML Explanations for Classification

[Type 1] Feature Attributions

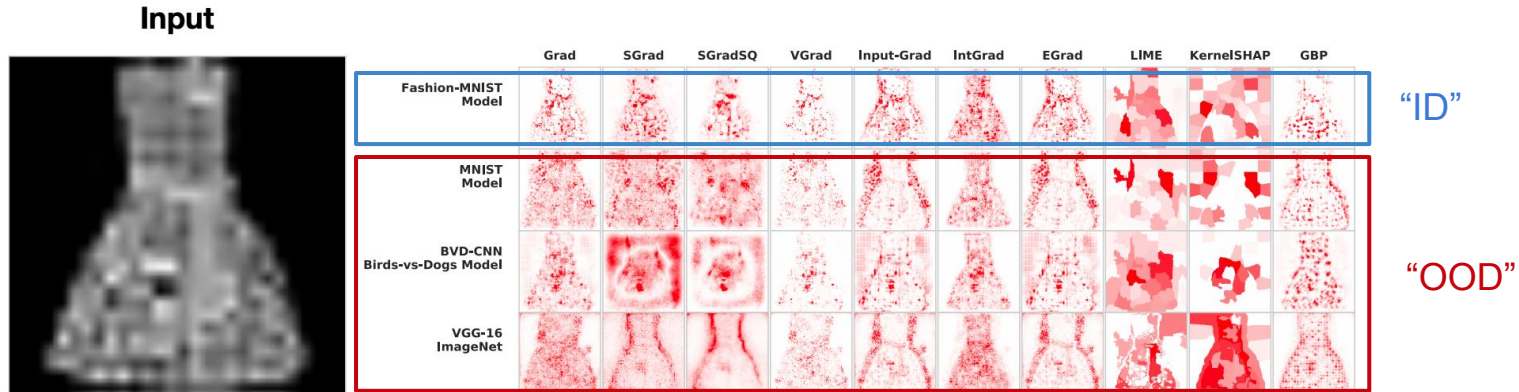
Input



$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_{i,j}} \quad \leftarrow \quad \begin{array}{l} \text{a logit} \\ (i, j)^{\text{th}} \text{ pixel} \end{array}$$

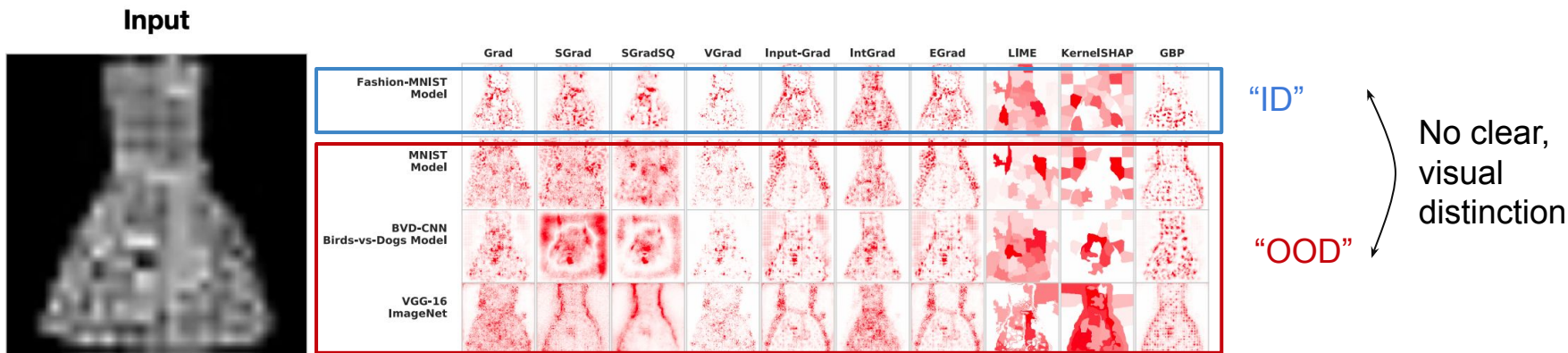
ML Explanations for Classification

[Type 1] Feature Attributions



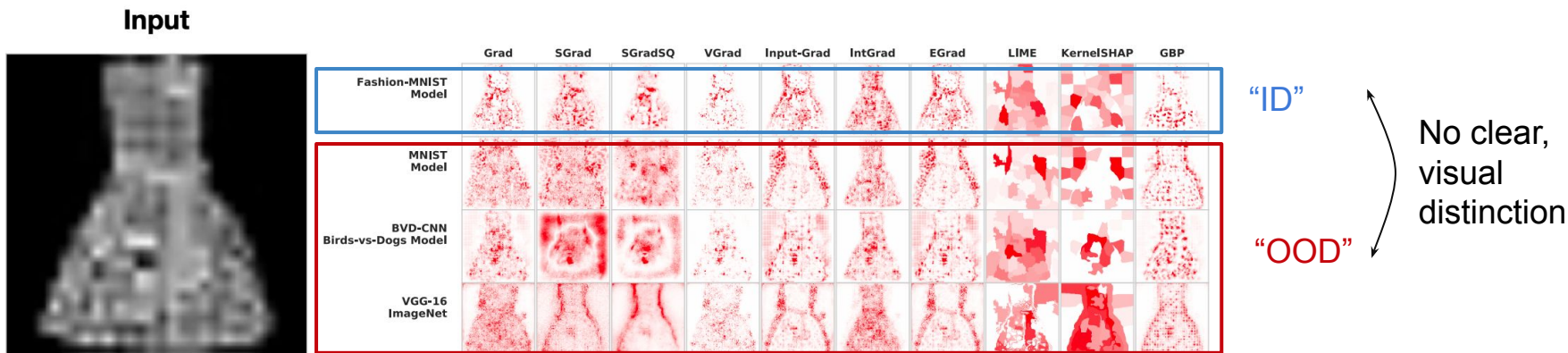
ML Explanations for Classification

[Type 1] Feature Attributions



ML Explanations for Classification

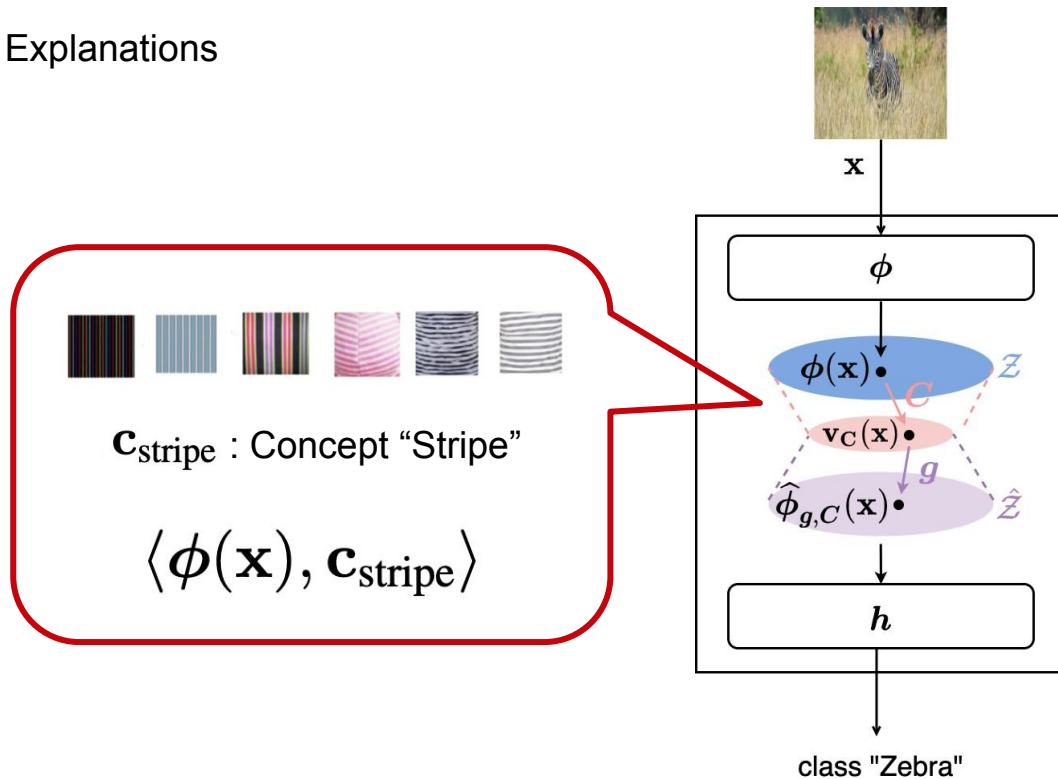
[Type 1] Feature Attributions



Pixel-level activations might not be the most intuitive form of explanations for humans

ML Explanations for Classification

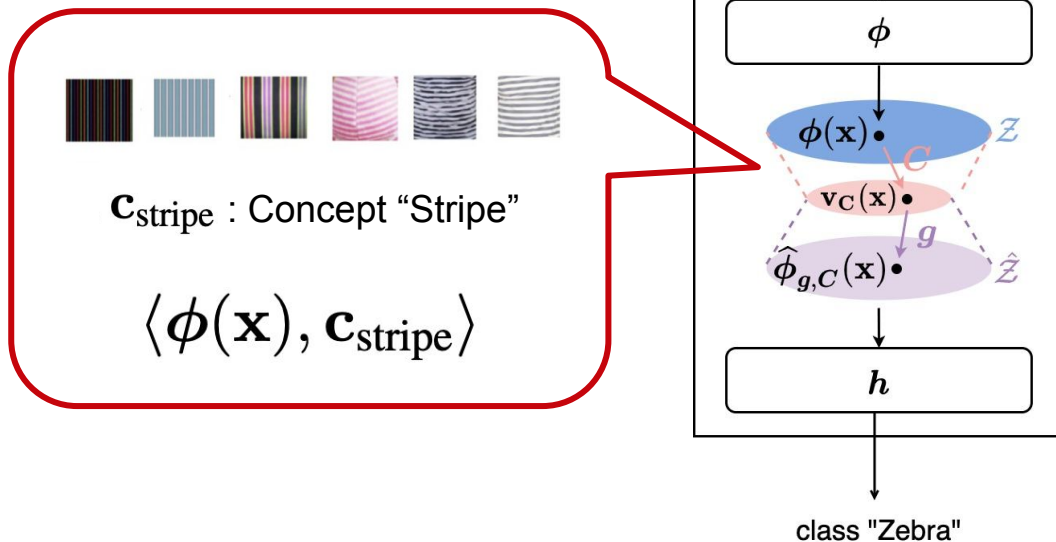
[Type 2] Concept-based Explanations



ML Explanations for Classification

[Type 2] Concept-based Explanations

The use of concept-based explanations for OOD detectors remains unexplored

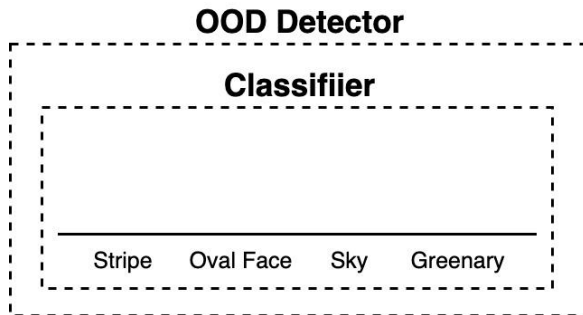




Concept-based Explanation for OOD Detection

Our work: the first method to understand the decisions of an OOD detector in terms of *high-level concepts*

Given DNN classifier, OOD detector, and a set of concepts that sufficiently explain their behaviors.

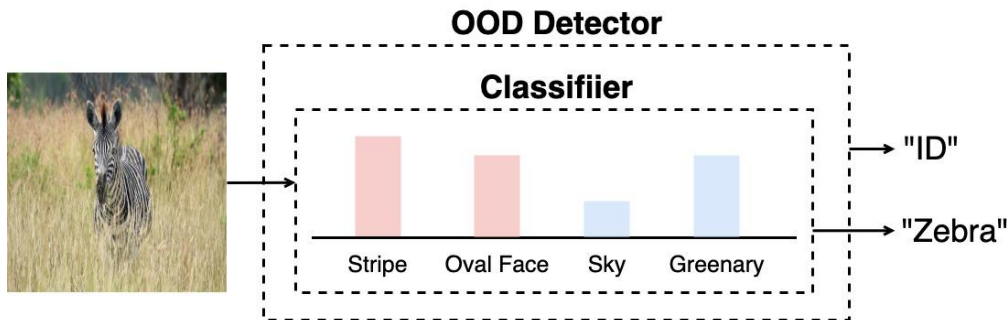




Concept-based Explanation for OOD Detection

Our work: the first method to understand the decisions of an OOD detector in terms of *high-level concepts*

Observe normal concept activations patterns given ID inputs.

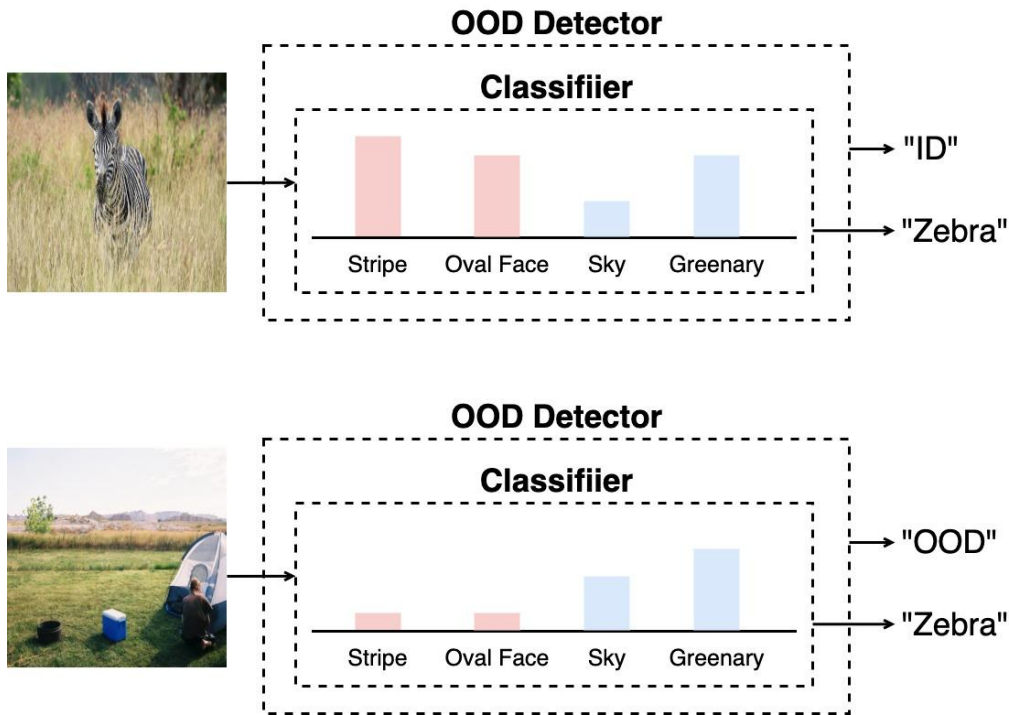




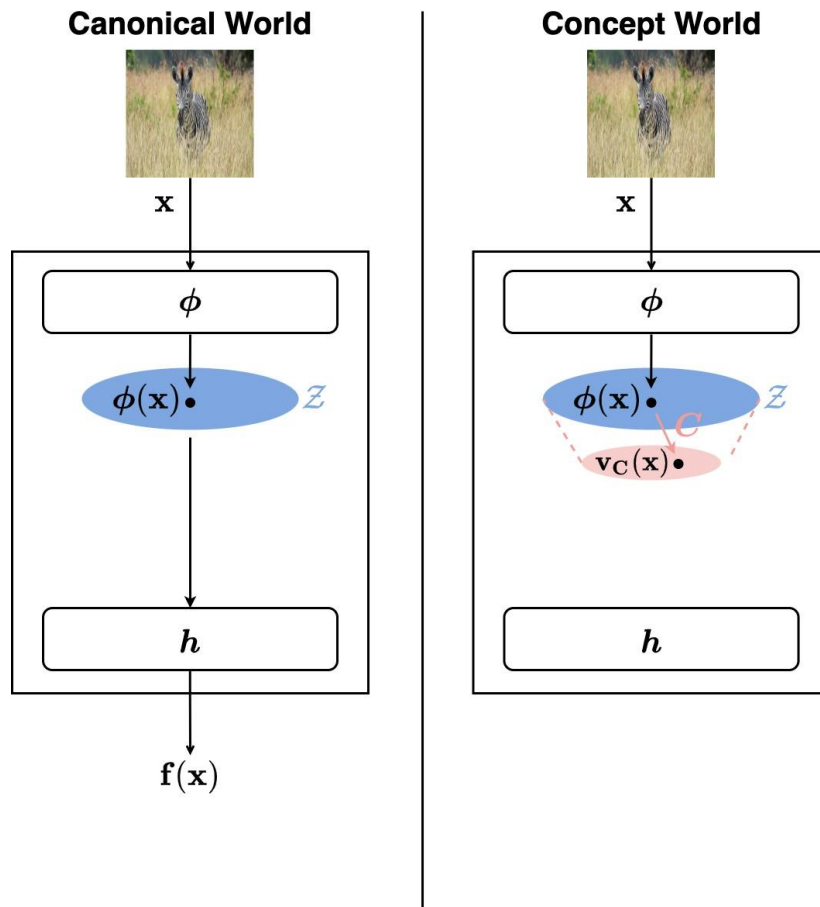
Concept-based Explanation for OOD Detection

Our work: the first method to understand the decisions of an OOD detector in terms of *high-level concepts*

Given OOD inputs, we observe different concept activation patterns compared to that of ID inputs.



Our Method



Our Method

$$\operatorname{argmax}_{\mathbf{C}, \mathbf{g}} \underbrace{\mathbb{E}_{(\mathbf{x}, y) \sim P_{\text{in}}} [\log h_y(\mathbf{g}(\mathbf{v}_{\mathbf{C}}(\mathbf{x})))]}_{\text{Accuracy in Concept World}} + \underbrace{\lambda_{\text{expl}} R_{\text{expl}}(\mathbf{C})}_{\text{Interpretability of concepts}}$$

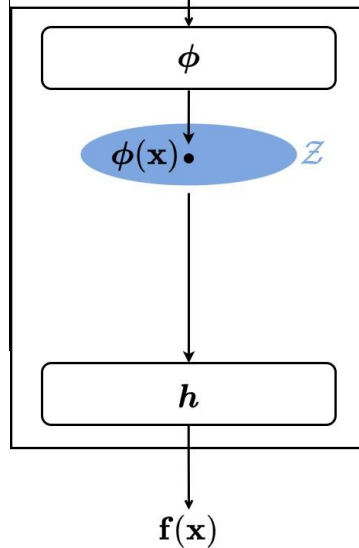
*Accuracy
in Concept World*

*Interpretability
of concepts*

Canonical World



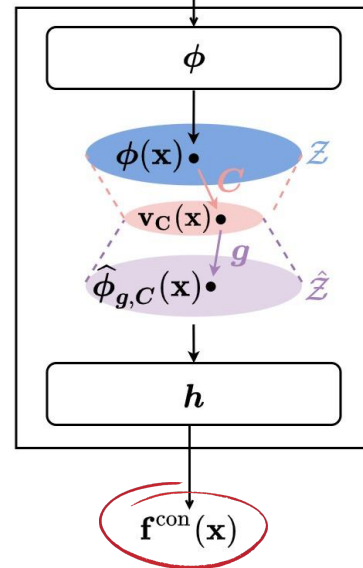
\mathbf{x}



Concept World



\mathbf{x}



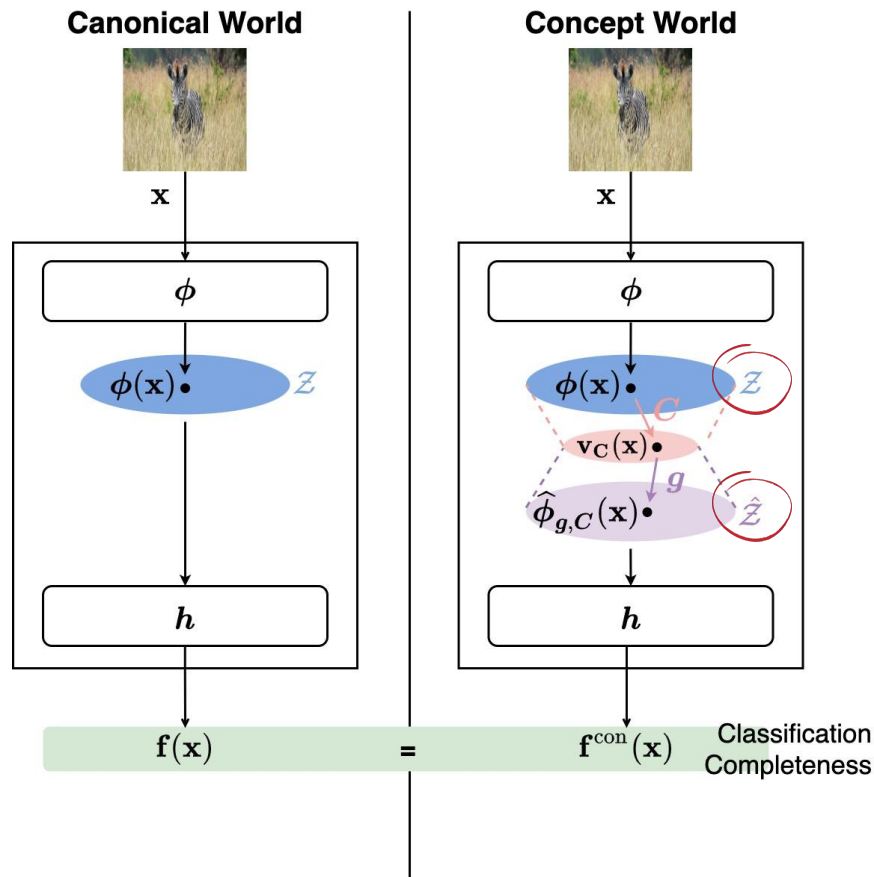
Our Method

$$\operatorname{argmax}_{\mathbf{C}, \mathbf{g}} \mathbb{E}_{(\mathbf{x}, y) \sim P_{\text{in}}} [\log h_y(\mathbf{g}(\mathbf{v}_{\mathbf{C}}(\mathbf{x})))] + \lambda_{\text{expl}} R_{\text{expl}}(\mathbf{C})$$

$$- \underbrace{\lambda_{\text{norm}} \mathbb{E}_{\mathbf{x} \sim P_{\text{in}}} \|\phi(\mathbf{x}) - \hat{\phi}_{\mathbf{g}, \mathbf{C}}(\mathbf{x})\|^2}_{\text{Accurate reconstruction of } Z}$$

Accurate reconstruction of Z

\Rightarrow high Classification Completeness

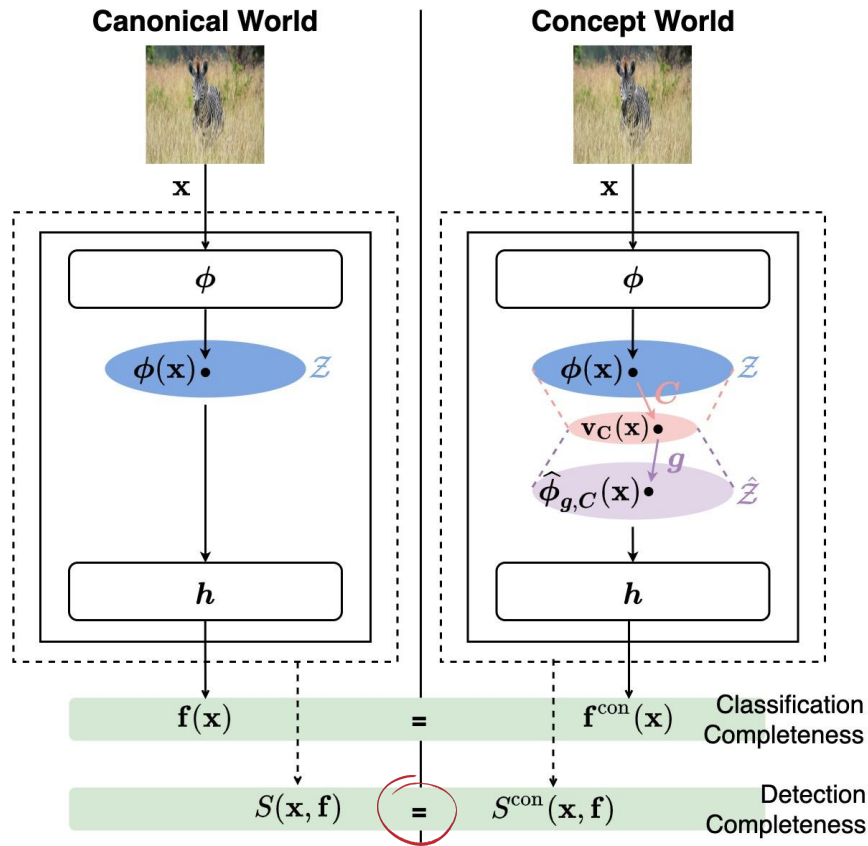


Our Method

$$\begin{aligned}
 & \operatorname{argmax}_{\mathbf{C}, \mathbf{g}} \mathbb{E}_{(\mathbf{x}, y) \sim P_{\text{in}}} [\log h_y(\mathbf{g}(\mathbf{v}_{\mathbf{C}}(\mathbf{x})))] + \lambda_{\text{expl}} R_{\text{expl}}(\mathbf{C}) \\
 & - \lambda_{\text{norm}} \mathbb{E}_{\mathbf{x} \sim P_{\text{in}}} \|\phi(\mathbf{x}) - \hat{\phi}_{\mathbf{g}, \mathbf{C}}(\mathbf{x})\|^2 \\
 & - \lambda_{\text{mse}} \left(\mathbb{E}_{\mathbf{x} \sim P_{\text{in}}} (S(\mathbf{x}, \mathbf{h} \circ \hat{\phi}_{\mathbf{g}, \mathbf{C}}) - S(\mathbf{x}, \mathbf{f}))^2 \right. \\
 & \quad \left. + \mathbb{E}_{\mathbf{x} \sim P_{\text{out}}} (S(\mathbf{x}, \mathbf{h} \circ \hat{\phi}_{\mathbf{g}, \mathbf{C}}) - S(\mathbf{x}, \mathbf{f}))^2 \right)
 \end{aligned}$$

*Similar OOD detector behavior
in both worlds*

⇒ high Detection Completeness



Our Method

$$\operatorname{argmax}_{\mathbf{C}, \mathbf{g}} \mathbb{E}_{(\mathbf{x}, y) \sim P_{\text{in}}} [\log h_y(\mathbf{g}(\mathbf{v}_{\mathbf{C}}(\mathbf{x})))] + \lambda_{\text{expl}} R_{\text{expl}}(\mathbf{C})$$

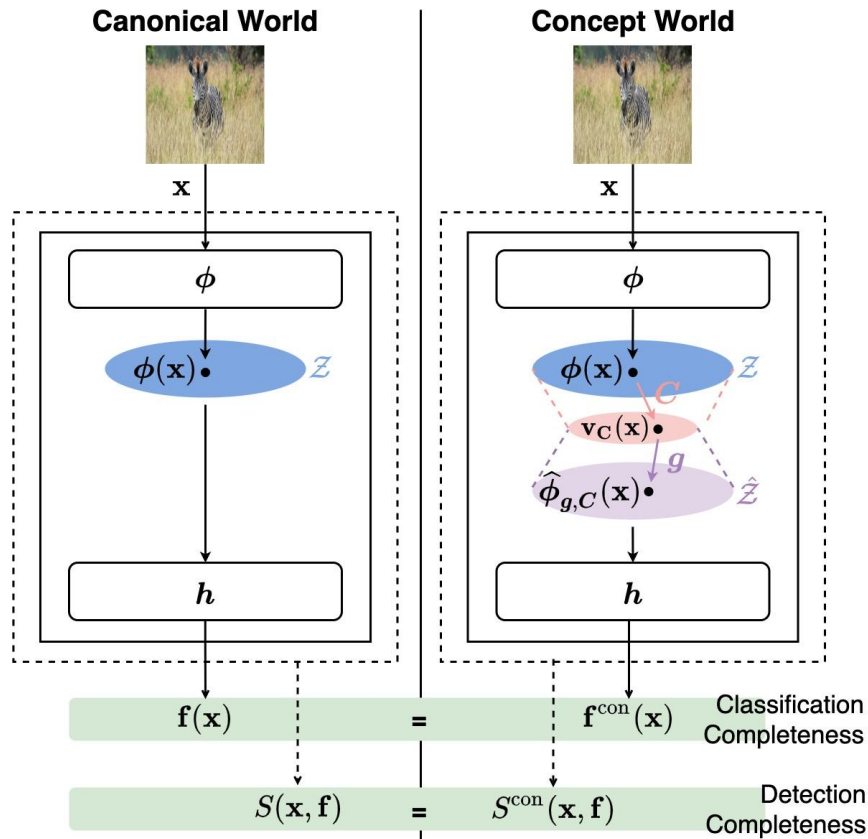
$$- \lambda_{\text{norm}} \mathbb{E}_{\mathbf{x} \sim P_{\text{in}}} \|\phi(\mathbf{x}) - \hat{\phi}_{\mathbf{g}, \mathbf{C}}(\mathbf{x})\|^2$$

$$- \lambda_{\text{mse}} \left(\mathbb{E}_{\mathbf{x} \sim P_{\text{in}}} (S(\mathbf{x}, \mathbf{h} \circ \hat{\phi}_{\mathbf{g}, \mathbf{C}}) - S(\mathbf{x}, \mathbf{f}))^2 \right. \\ \left. + \mathbb{E}_{\mathbf{x} \sim P_{\text{out}}} (S(\mathbf{x}, \mathbf{h} \circ \hat{\phi}_{\mathbf{g}, \mathbf{C}}) - S(\mathbf{x}, \mathbf{f}))^2 \right)$$

$$+ \underbrace{\lambda_{\text{sep}} J_{\text{sep}}(\mathbf{C})}_{\text{Separability between ID vs OOD inputs in concept space}}$$

*Separability between ID vs OOD inputs
in concept space*

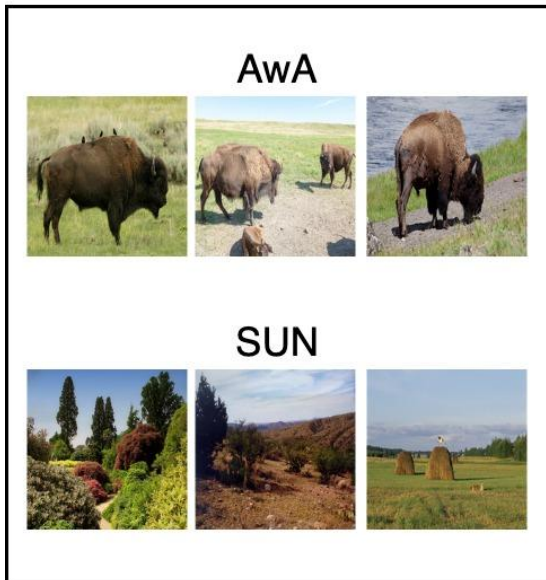
⇒ high Concept Separability



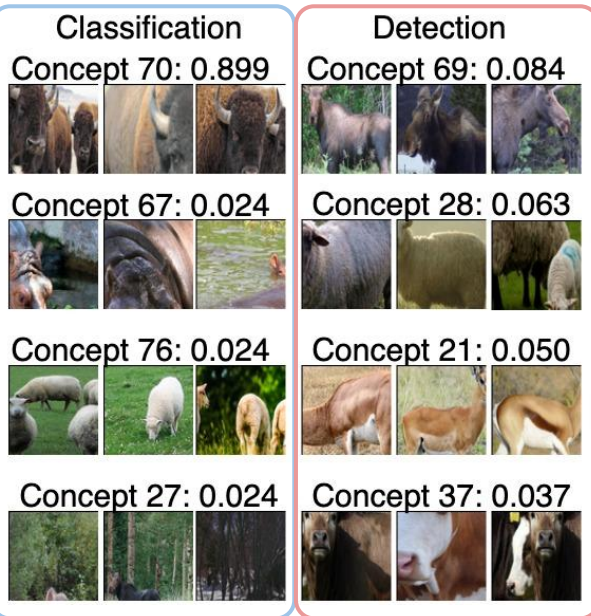
Results

Concept-based explanations given Inception-V3 classifier and Energy detector:

Classified to "Buffalo"



Ours



How much each concept contributes to the prediction of "Buffalo"?

How much each concept contributes to the detection results for inputs predicted to "Buffalo"?



Our Contributions

Our work is the **first method** for providing **concept-based attributions** of the **decision of an OOD detector** based on high-level concepts. Specifically,

1. We propose metrics to quantify the effectiveness of concept-based explanation for OOD detection:
 - a. **Detection Completeness**: are the concept scores sufficient statistics for class predictions and OOD detection?
 - b. **Concept Separability**: are ID and OOD inputs clearly distinctive in terms of concepts?



Our Contributions

Our work is the **first method** for providing **concept-based attributions** of the **decision of an OOD detector** based on high-level concepts. Specifically,

1. We propose metrics to quantify the effectiveness of concept-based explanation for OOD detection:
 - a. **Detection Completeness**: are the concept scores sufficient statistics for class predictions and OOD detection?
 - b. **Concept Separability**: are ID and OOD inputs clearly distinctive in terms of concepts?
2. We introduce **general concept learning framework** that discovers a set of concepts that have good detection completeness and concept separability.



Our Contributions


Our work is the **first method** for providing **concept-based attributions** of the **decision of an OOD detector** based on high-level concepts. Specifically,

1. We propose metrics to quantify the effectiveness of concept-based explanation for OOD detection:
 - a. **Detection Completeness**: are the concept scores sufficient statistics for class predictions and OOD detection?
 - b. **Concept Separability**: are ID and OOD inputs clearly distinctive in terms of concepts?
2. We introduce **general concept learning framework** that discovers a set of concepts that have good detection completeness and concept separability.
3. By using the concepts learned by our framework, show how to identify prominent concepts that contribute to an OOD detector's decisions, and provide insights for popular OOD detectors.



Thank you

For the complete description of our work, please check out our paper!

 > cs > arXiv:2203.02586

Search...
Help | Advanced

Computer Science > Machine Learning

[Submitted on 4 Mar 2022]

Concept-based Explanations for Out-Of-Distribution Detectors

Jihye Choi, Jayaram Raghuram, Ryan Feng, Jiefeng Chen, Somesh Jha, Atul Prakash

Out-of-distribution (OOD) detection plays a crucial role in ensuring the safe deployment of deep neural network (DNN) classifiers. While a myriad of methods have focused on improving the performance of OOD detectors, a critical gap remains in interpreting their decisions. We help bridge this gap by providing explanations for OOD detectors based on learned high-level concepts. We first propose two new metrics for assessing the effectiveness of a particular set of concepts for explaining OOD detectors: 1) detection completeness, which quantifies the sufficiency of concepts for explaining an OOD-detector's decisions, and 2) concept separability, which captures the distributional separation between in-distribution and OOD data in the concept space. Based on these metrics, we propose a framework for learning a set of concepts that satisfy the desired properties of detection completeness and concept separability and demonstrate the framework's effectiveness in providing concept-based explanations for diverse OOD techniques. We also show how to identify prominent concepts that contribute to the detection results via a modified Shapley value-based importance score.

Comments: 19 pages, 9 figures

Subjects: **Machine Learning** (cs.LG); Computer Vision and Pattern Recognition (cs.CV)

Cite as: arXiv:2203.02586 [cs.LG]
(or arXiv:2203.02586v1 [cs.LG] for this version)
<https://doi.org/10.48550/arXiv.2203.02586>