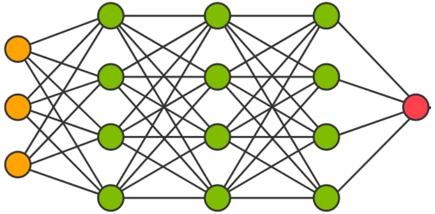
Concept-based Explanations for Out-Of-Distribution Detectors

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

Standard Machine Learning (ML) Models

ML model in self-driving car

Training

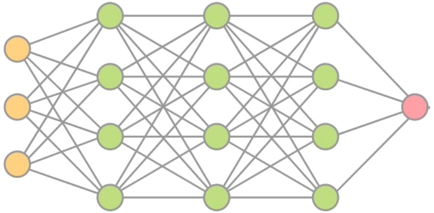


"Car"

Standard Machine Learning (ML) Models

ML model in self-driving car

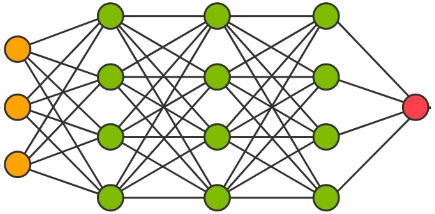
Training



→ "Car"

Testing

In-distribution (ID)

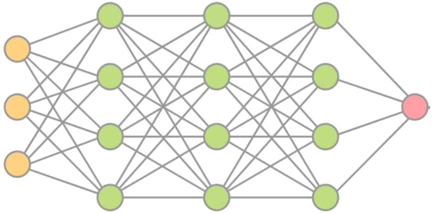


→ "Car"

Standard Machine Learning (ML) Models

ML model in self-driving car

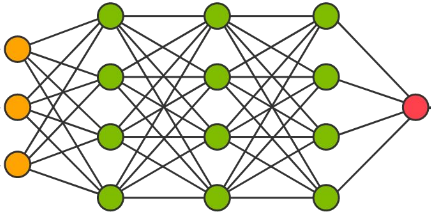
Training



“Car”

Testing

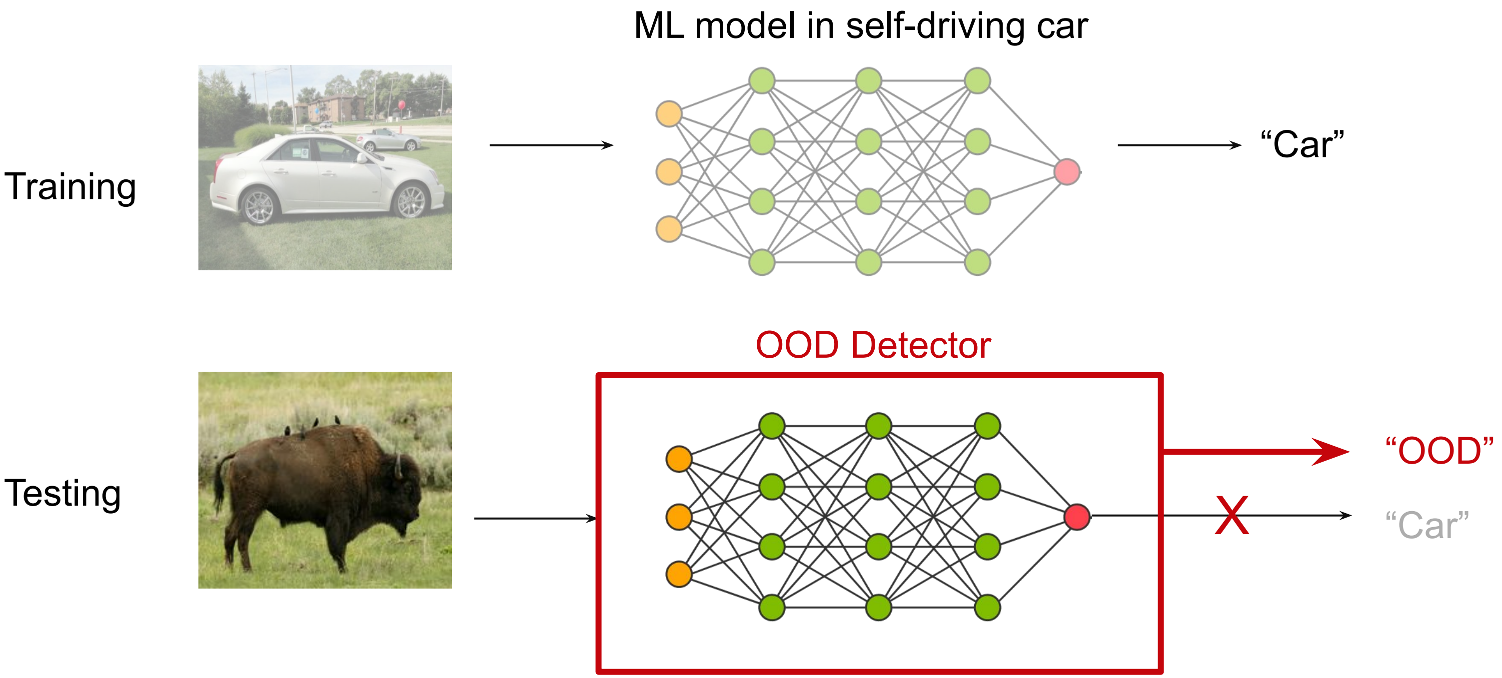
Out-of-distribution
(OOD)



“Car”

Overconfident prediction
for unseen object “Buffalo”

Out-Of-Distribution (OOD) Detection

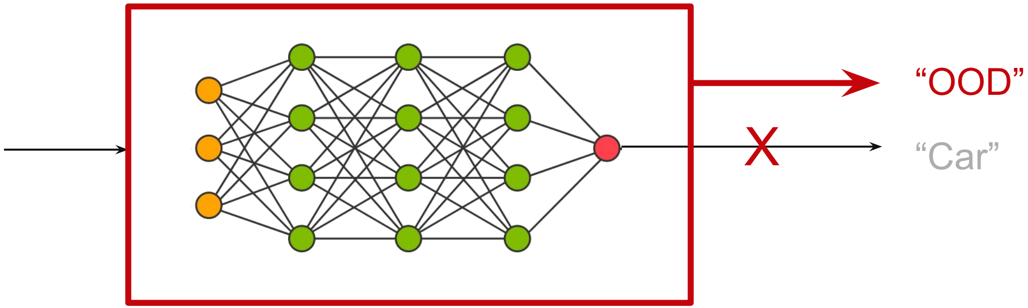


Understanding OOD Detection

Why a given OOD detector decides an input to be OOD?

Can we use existing ML explanation methods for classification to interpret OOD detection results?

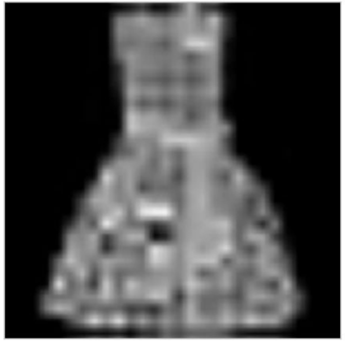
Testing



ML Explanations for Classification

[Type 1] Feature Attributions

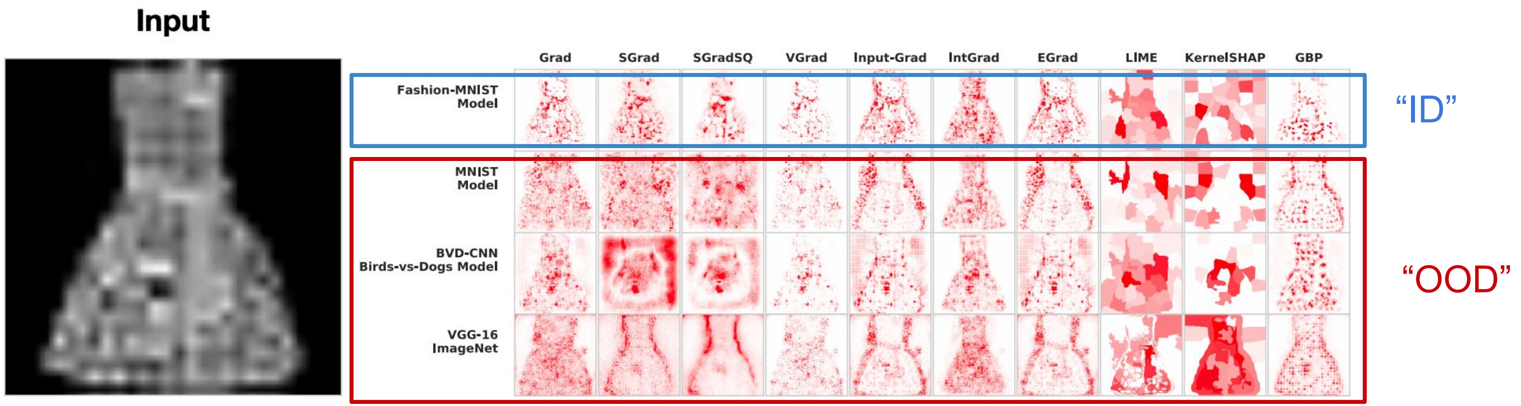
Input



[Adebayo et al., NeurIPS'20]

ML Explanations for Classification

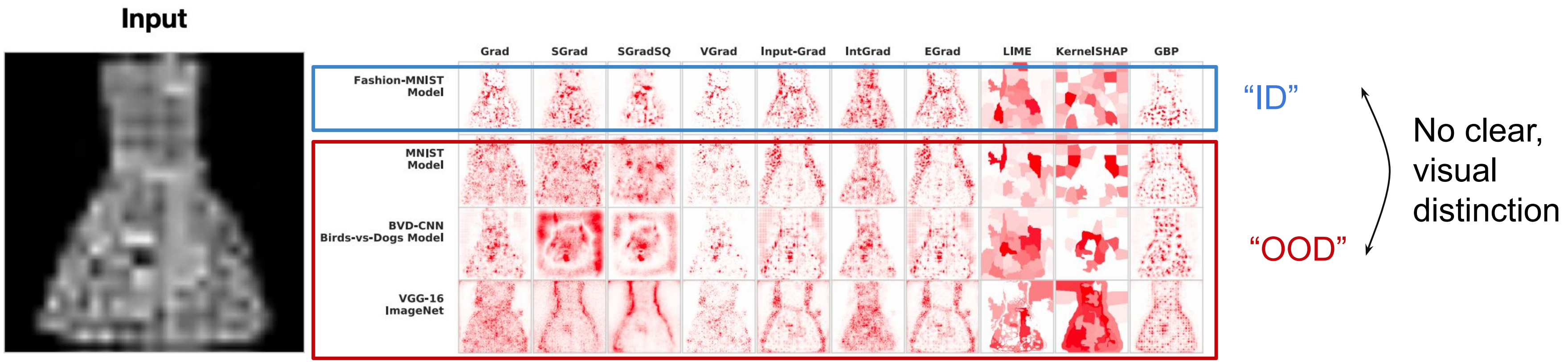
[Type 1] Feature Attributions



[Adebayo et al., NeurIPS'20]

ML Explanations for Classification

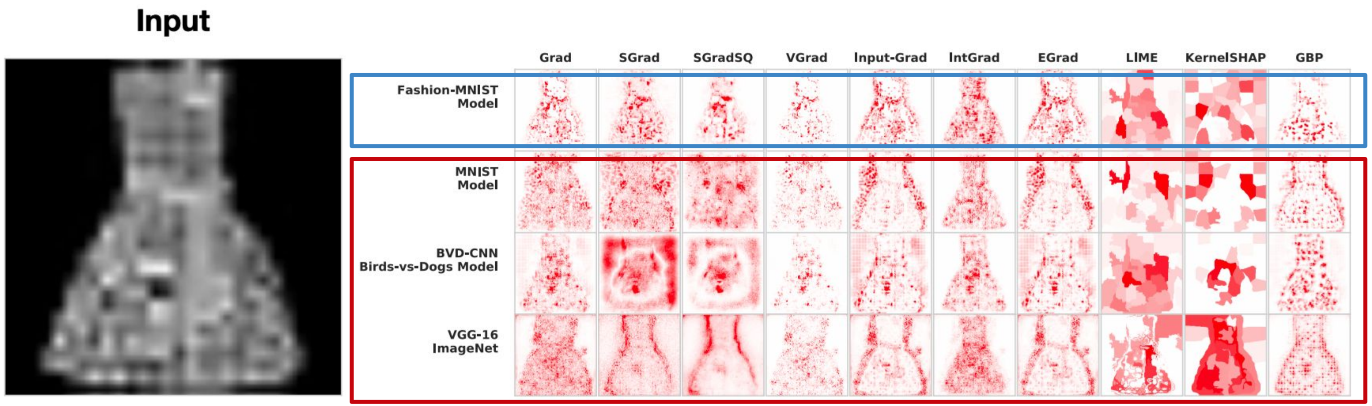
[Type 1] Feature Attributions



[Adebayo et al., NeurIPS'20]

ML Explanations for Classification

[Type 1] Feature Attributions



"ID"

"OOD"

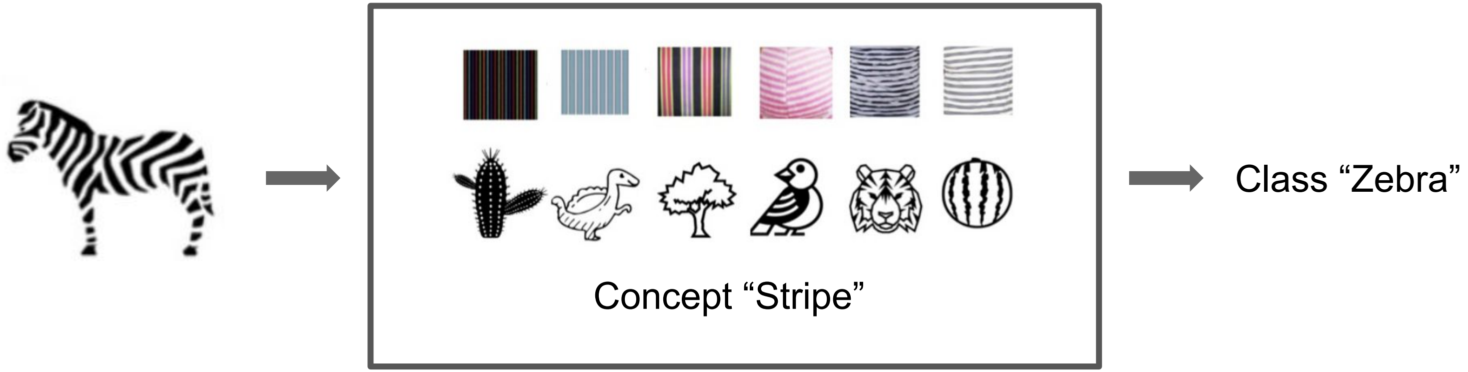
No clear, visual distinction

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

[Adebayo et al., NeurIPS'20]

ML Explanations for Classification

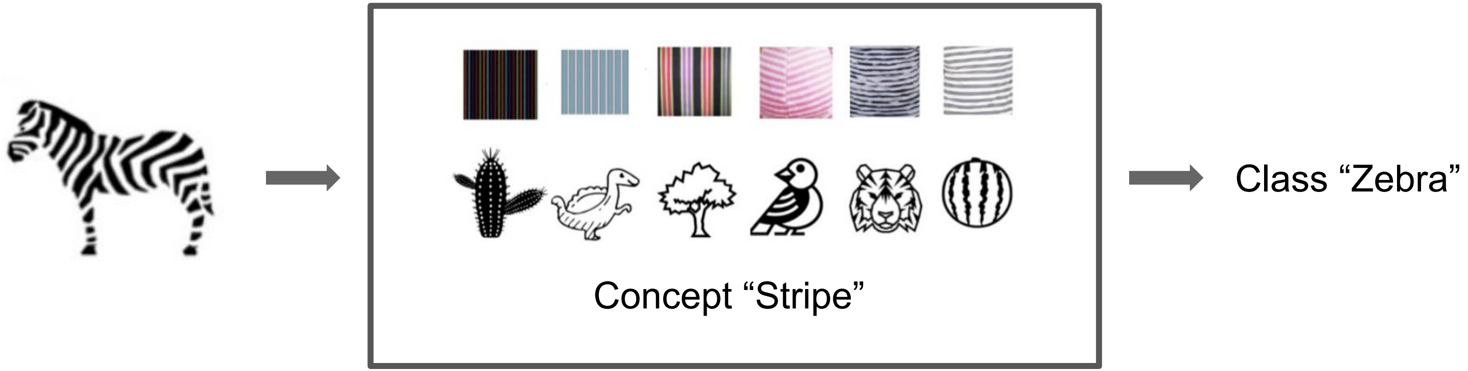
[Type 2] Concept-based Explanations



[Kim et al., ICML’18]

ML Explanations for Classification

[Type 2] Concept-based Explanations



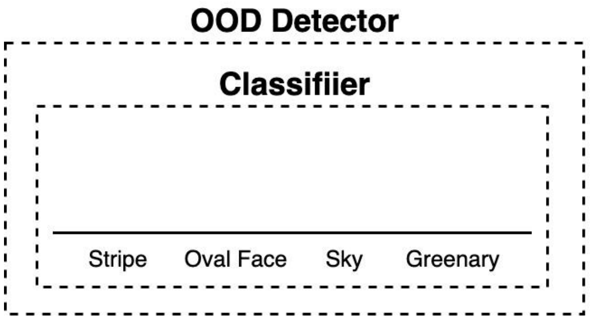
The use of concept-based explanations for OOD detectors remains unexplored

[Kim et al., ICML’18]

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*

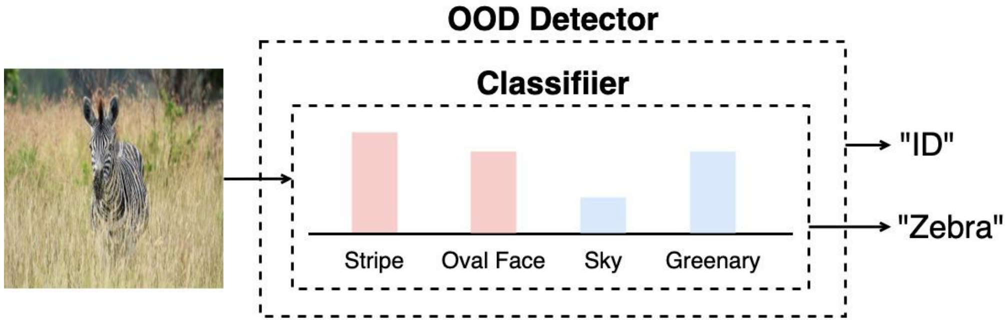
GIven DNN classifier and OOD detector, find a set of concepts that sufficiently explain their behaviors.



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*

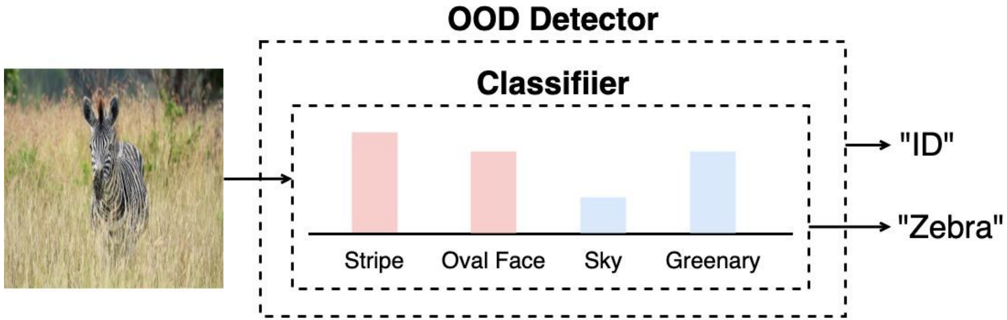
Observe normal concept activations patterns given ID inputs.

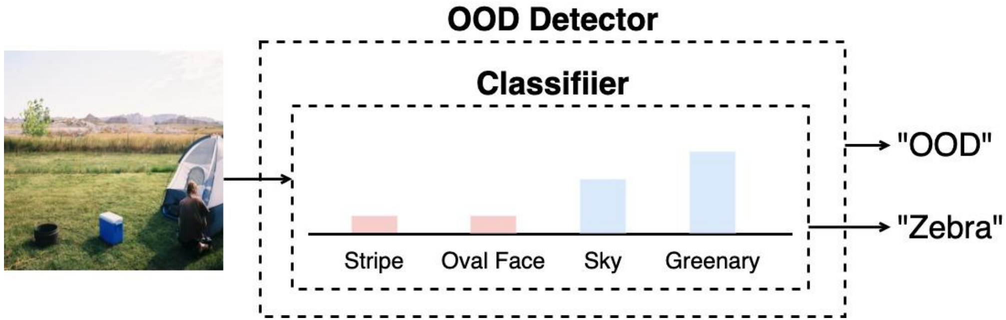


Concept-based Explanation for OOD Detection

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 Contributions

Given DNN classifier and OOD detector, we

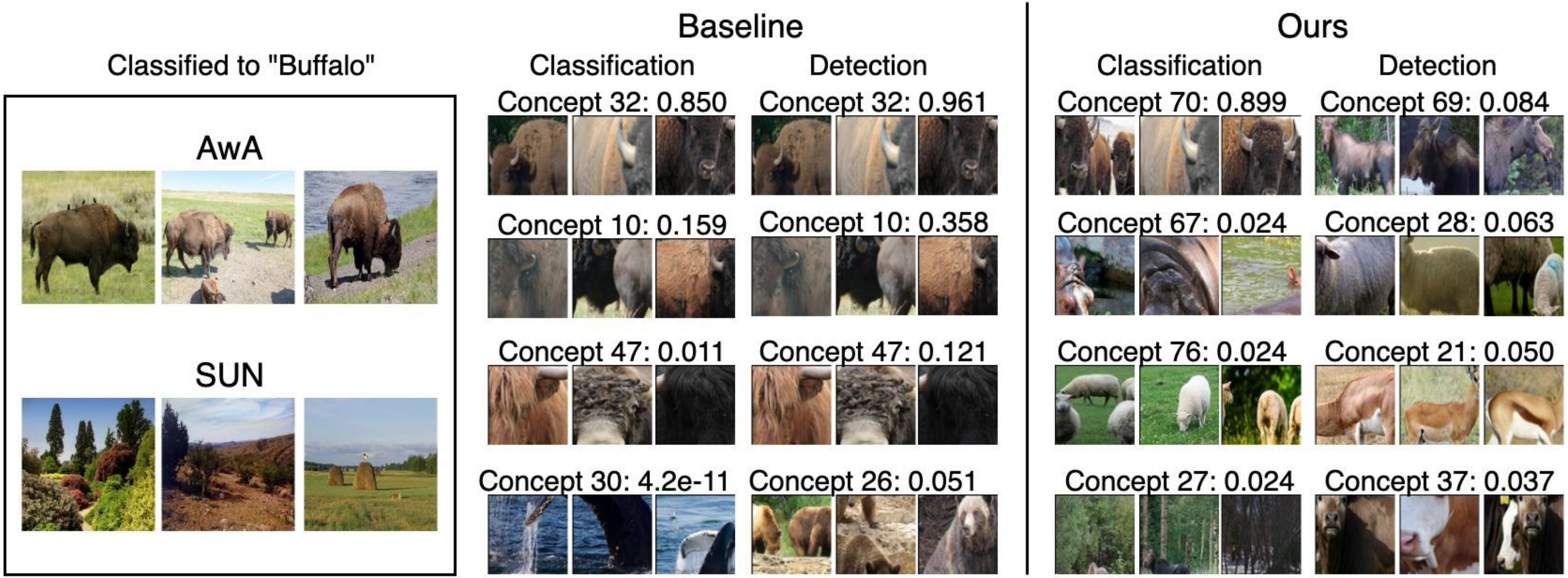
1. propose metrics to quantify the effectiveness of concept-based explanation for OOD detection:
    a. *Detection Completeness*: are the concept scores sufficient statistics for class predictions and OOD detection?
    b. *Concept Separability*: are ID and OOD inputs clearly distinctive in terms of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

2. 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.

Results

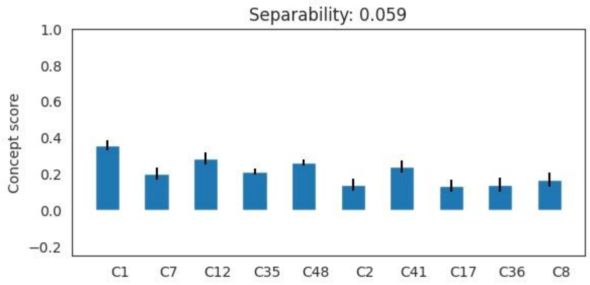
Concept-based explanations for Energy detector



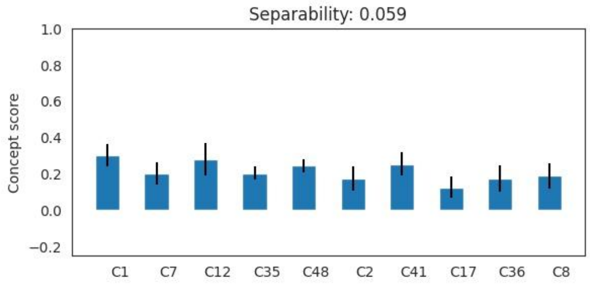
Energy detector: [Liu et al., NeurIPS'20]
Baseline: [Yeh et al., NeurIPS'20]

Results

Concept score patterns between inputs detected as ID vs OOD



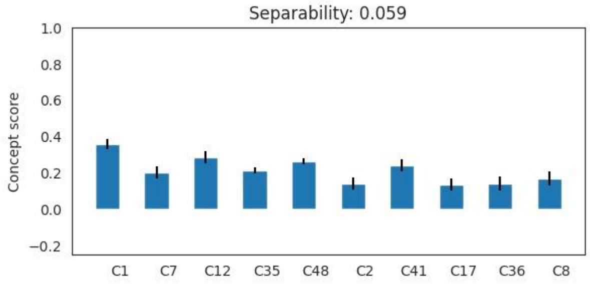
low separability, inputs detected as ID



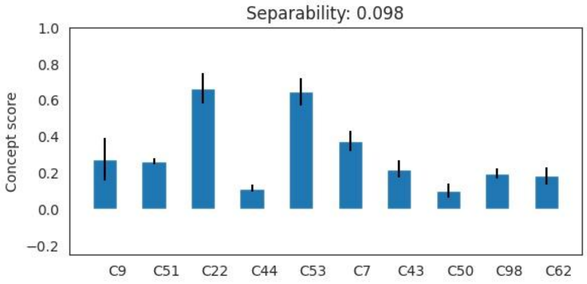
low separability, inputs detected as ID

Results

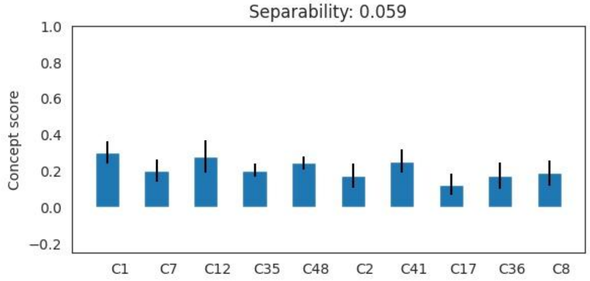
Concept score patterns between inputs detected as ID vs OOD



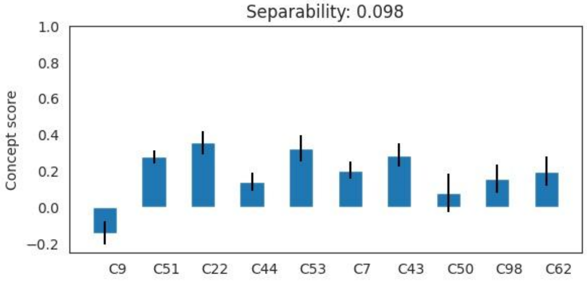
low separability, inputs detected as ID



high separability, inputs detected as ID



low separability, inputs detected as ID



high separability, inputs detected as ID

Thank you

For complete description of our method and full results, please check out our paper!

arXiv > 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