

Rethinking the Evaluation of Out-of-Distribution Detection: A Sorites Paradox

Xingming Long, Jie Zhang, Shiguang Shan, Xilin Chen

Key Laboratory of AI Safety of CAS, Institute of Computing Technology,
Chinese Academy of Sciences (CAS)

University of Chinese Academy of Sciences

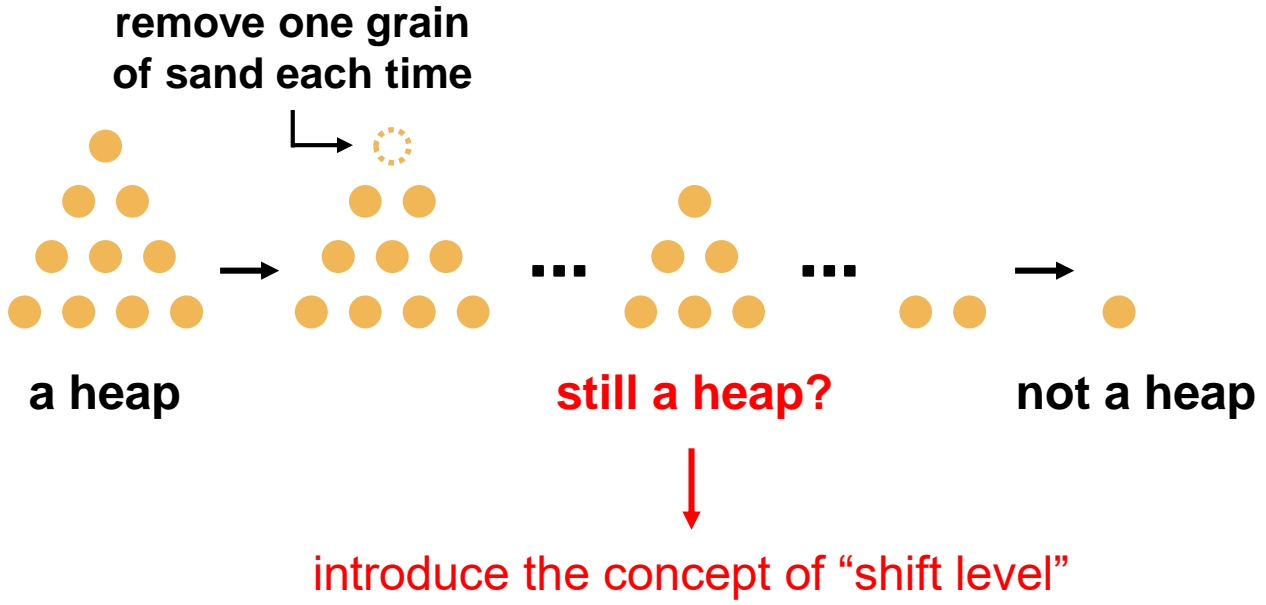








中国科学院计算技术研究所
Institute of Computing Technology, Chinese Academy of Sciences

Introduction

■ Motivation

- Marginal OOD samples are ambiguous (Sorites Paradox)
- Current label-based division introduces confusion

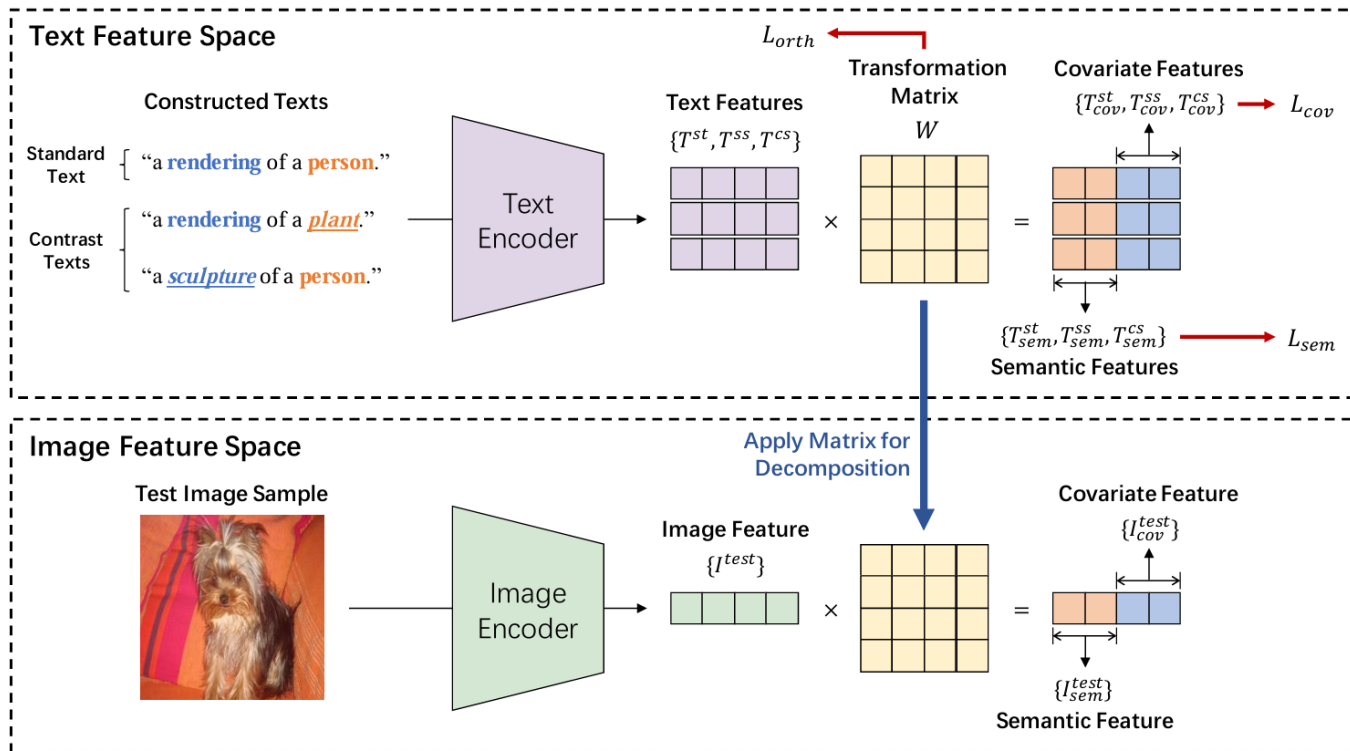


	Similar Labels	Overlapping Labels	Insufficient Labels
ImageNet-1K	 African bush elephant	 corn	 T-shirt
ImageNet-21K	 African elephant	 food	 athlete

Method

■ Feature Decomposition

□ Text feature space => Image feature space



Triplet mining

$$L_{sem} = dist(T_{sem}^{st}, T_{sem}^{cs}) - dist(T_{sem}^{st}, T_{sem}^{ss}) + \alpha$$

$$L_{cov} = dist(T_{cov}^{st}, T_{cov}^{ss}) - dist(T_{cov}^{st}, T_{cov}^{cs}) + \alpha$$

Orthogonal regularization

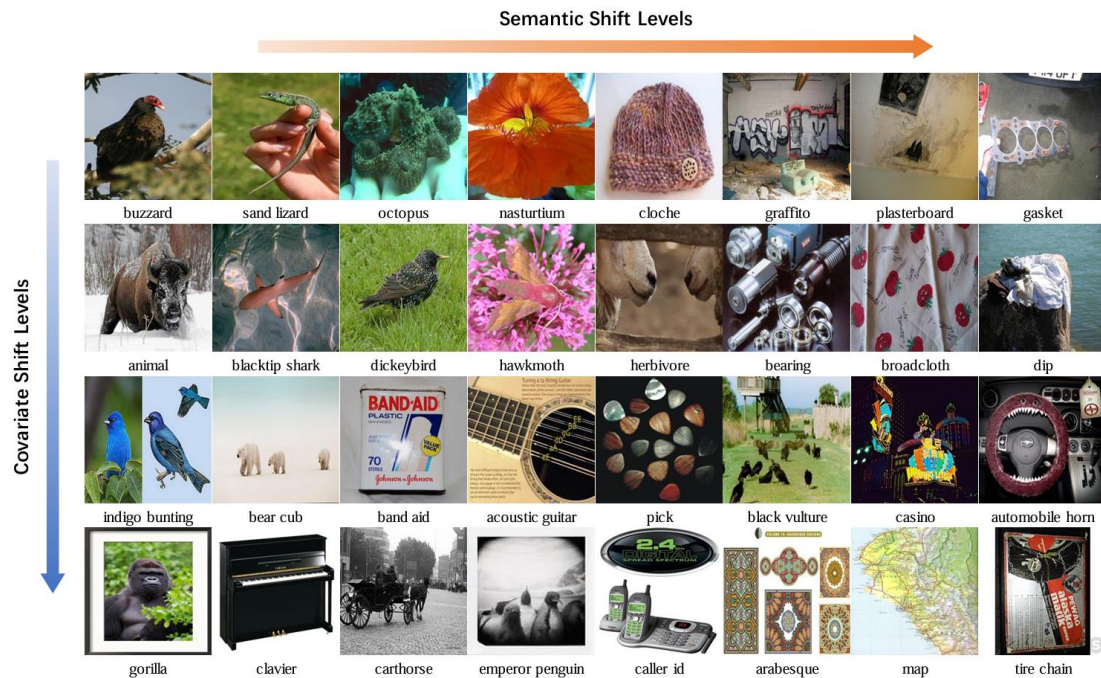
$$L_{orth} = \|W^T W - I\|_2^2$$

Language Aligned Image feature Decomposition (LAID)

Method

■ Benchmark Construction

- Divide ImageNet-21K according to shift levels
- Generate Syn-IS with enhanced covariate diversity



ImageNet-21K



Syn-IS

Method

■ Metrics

- Study the variations in model performance

$$\textit{correlation} = \frac{\sum_{i=1}^n (x_i - \bar{x})(i - \frac{n+1}{2})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (i - \frac{n+1}{2})^2}}$$

$$\textit{sensitivity} = \left| \frac{\sum_{i=1}^n (x_i - \bar{x})(i - \frac{n+1}{2})}{\sum_{i=1}^n (i - \frac{n+1}{2})^2} \right|$$

(“i” denotes the shift levels)

Experiments on ImageNet-21K

- Evaluate on each subset
 - OOD detection methods perform better when there is a large semantic shift and a small covariate shift

Semantic Shift Levels

	1	2	3	4	5	6	7	8
1	46.0	61.3	72.0	81.1	86.3	86.3	86.2	88.1
2	45.6	60.3	70.4	78.7	84.3	85.7	86.3	85.7
3	44.8	57.6	66.3	73.6	80.3	83.3	84.9	84.8
4	44.1	55.3	63.1	69.9	76.5	80.9	83.2	83.7
5	41.9	52.0	60.0	66.0	73.1	78.2	81.4	81.9
6	45.6	45.1	54.1	62.0	69.4	75.1	78.9	80.2
7	N/A	42.3	49.5	58.6	66.2	72.8	76.6	78.7
8	N/A	N/A	44.8	54.9	64.7	70.4	73.3	78.1

MSP

Semantic Shift Levels

	1	2	3	4	5	6	7	8
1	65.9	64.7	70.0	78.7	85.4	81.7	76.8	80.6
2	65.1	63.5	67.0	74.7	80.8	79.0	77.3	74.7
3	66.1	63.7	64.6	68.1	73.0	73.7	73.6	73.3
4	66.7	65.2	64.6	65.7	68.1	69.9	70.7	71.8
5	65.1	66.0	65.3	64.4	65.8	67.7	69.2	72.0
6	73.6	64.9	66.2	65.6	66.4	67.9	69.3	74.0
7	N/A	71.4	66.4	68.3	69.4	68.7	70.2	74.1
8	N/A	N/A	65.5	71.4	74.9	71.3	69.5	76.0

GradNorm

Semantic Shift Levels

	1	2	3	4	5	6	7	8
1	48.3	60.5	71.3	81.8	88.6	87.8	87.3	90.4
2	51.5	62.8	72.1	81.1	87.3	88.0	88.4	88.3
3	52.5	62.0	70.1	77.5	84.0	86.6	87.8	87.7
4	52.2	60.7	68.3	75.4	81.7	85.5	87.3	87.4
5	50.1	59.8	67.2	72.7	79.5	84.2	86.7	87.1
6	56.0	56.5	63.9	70.4	77.3	82.6	85.3	87.1
7	N/A	61.7	61.0	67.1	75.0	81.2	84.4	86.9
8	N/A	N/A	57.6	70.1	76.1	80.1	82.2	88.5

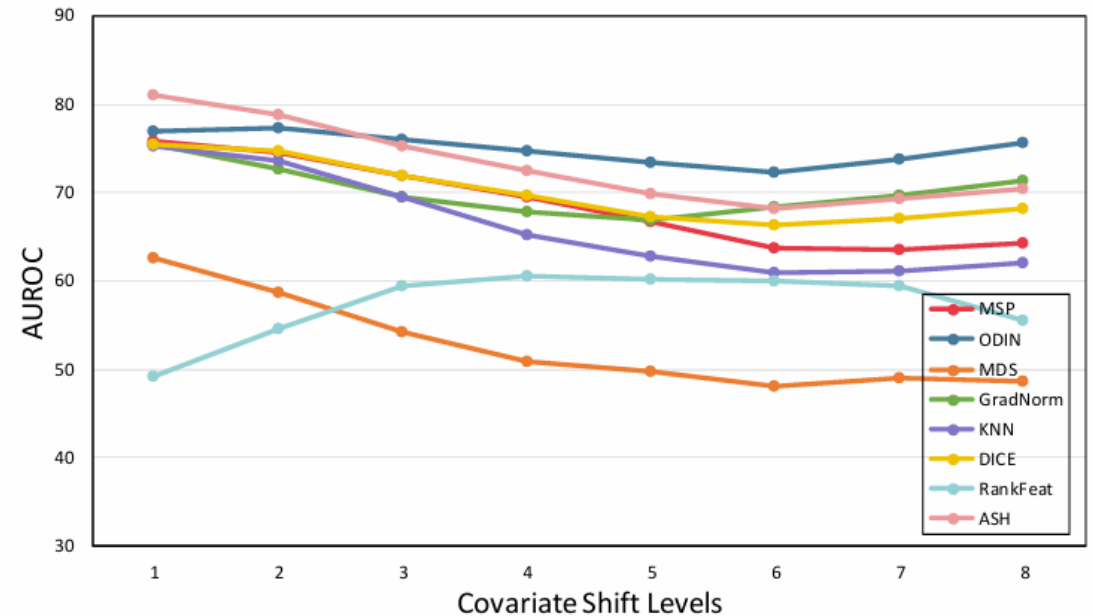
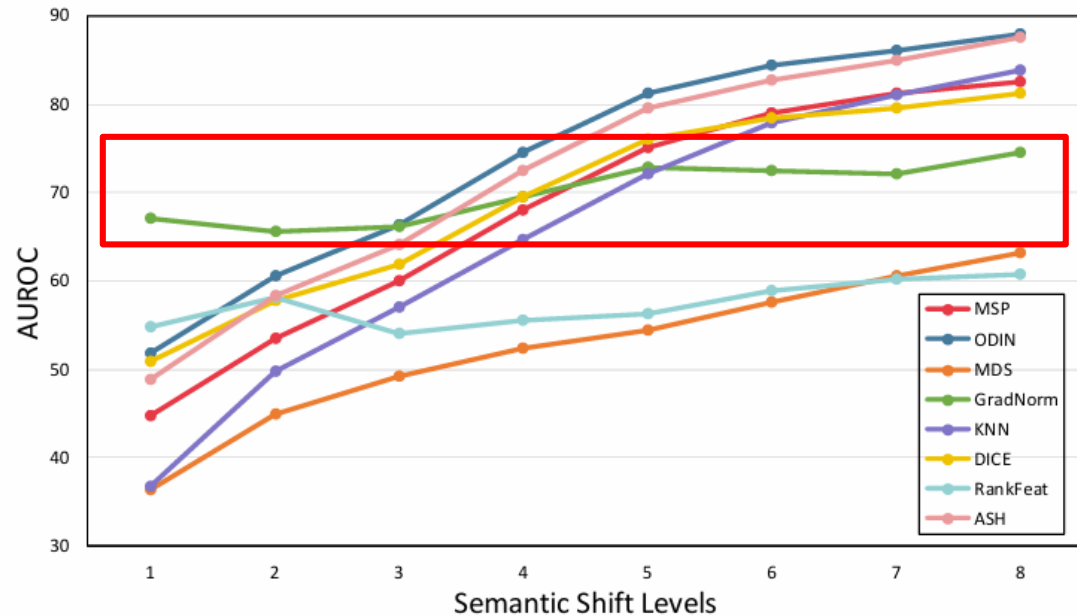
ODIN

(“N/A” indicates the number of data in this subset is too small for a fair evaluation)

Experiments on ImageNet-21K

■ Curve of performance

- Performance of most methods significantly improves as the semantic shift increases
- Some methods rely less on semantic shifts



Experiments on ImageNet-21K

■ Correlation & Sensitivity

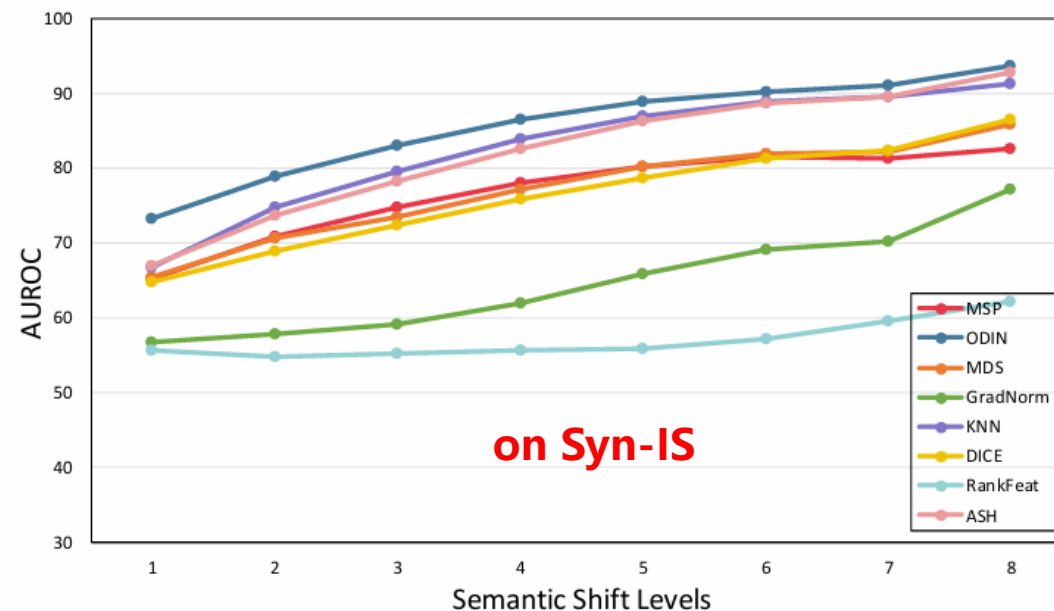
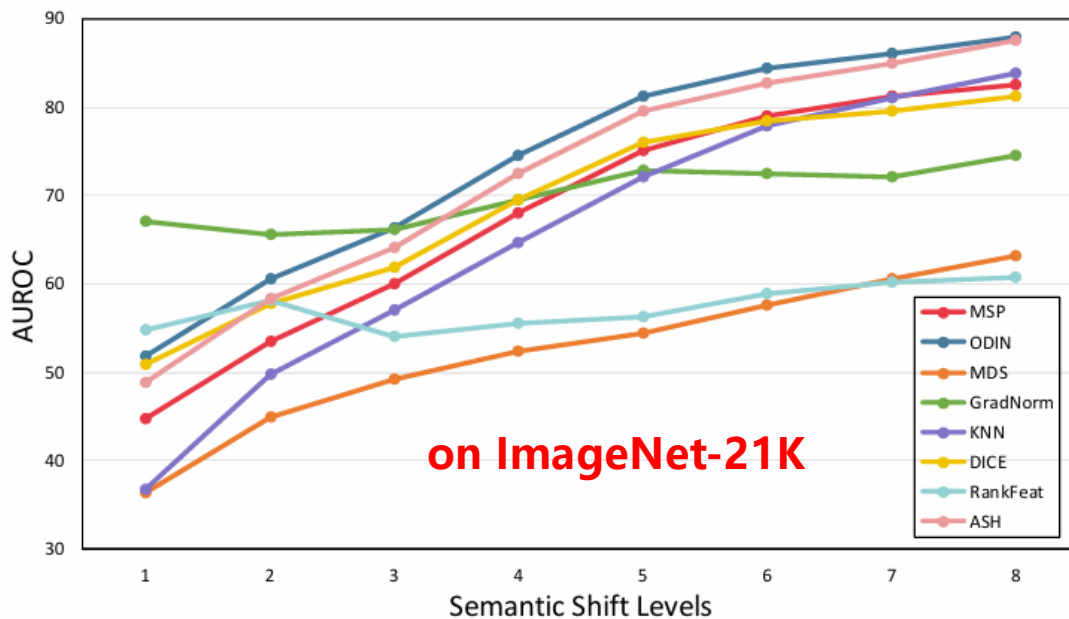
- Most methods exhibit positive correlation and higher sensitivity to semantic shifts, but show the opposite for covariate shifts

	Semantic		Covariate	
	correlation	sensitivity	correlation	sensitivity
MSP [1]	0.97	5.59	-0.96	1.95
ODIN [2]	0.97	5.26	-0.63	0.46
MDS [3]	0.98	3.50	-0.91	1.98
GradNorm [4]	0.91	1.27	-0.49	0.56
KNN [5]	0.98	6.64	-0.93	2.18
DICE [6]	0.97	4.52	-0.88	1.29
RankFeat [7]	0.78	0.79	0.51	0.83
ASH [8]	0.98	5.56	-0.89	1.73

Experiments on Syn-IS

■ Curve of performance

- Performances of most methods improve on Syn-IS
- Methods like GradNorm perform unsatisfactorily



Experiments on Syn-IS

■ Correlation & Sensitivity

- Many methods show a positive correlation with the covariate shift levels on Syn-IS

	Semantic		Covariate	
	correlation	sensitivity	correlation	sensitivity
MSP [1]	0.93	2.33	0.36	0.23
ODIN [2]	0.96	2.71	0.92	0.72
MDS [3]	0.98	2.71	0.99	2.91
GradNorm [4]	0.98	2.85	-0.85	0.72
KNN [5]	0.95	3.30	0.92	1.71
DICE [6]	0.99	2.96	-0.41	0.13
RankFeat [7]	0.87	0.91	-0.94	1.03
ASH [8]	0.98	3.52	0.92	0.75

Conclusion

■ Take home messages

- Most OOD detection methods are **sensitive to semantic shifts**, which aligns with common sense
- **Excessive covariate shifts** can also impact detection methods, a factor worth noting
- Methods like GradNorm have **potential limitations** and require further investigation

Conclusion

- Codebase

- <https://github.com/qqwsad5/IS-OOD>



- Contact us

- xingming.long@vipl.ict.ac.cn