JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

GEN: Pushing the Limits of Softmax-Based Out-of-Distribution Detection

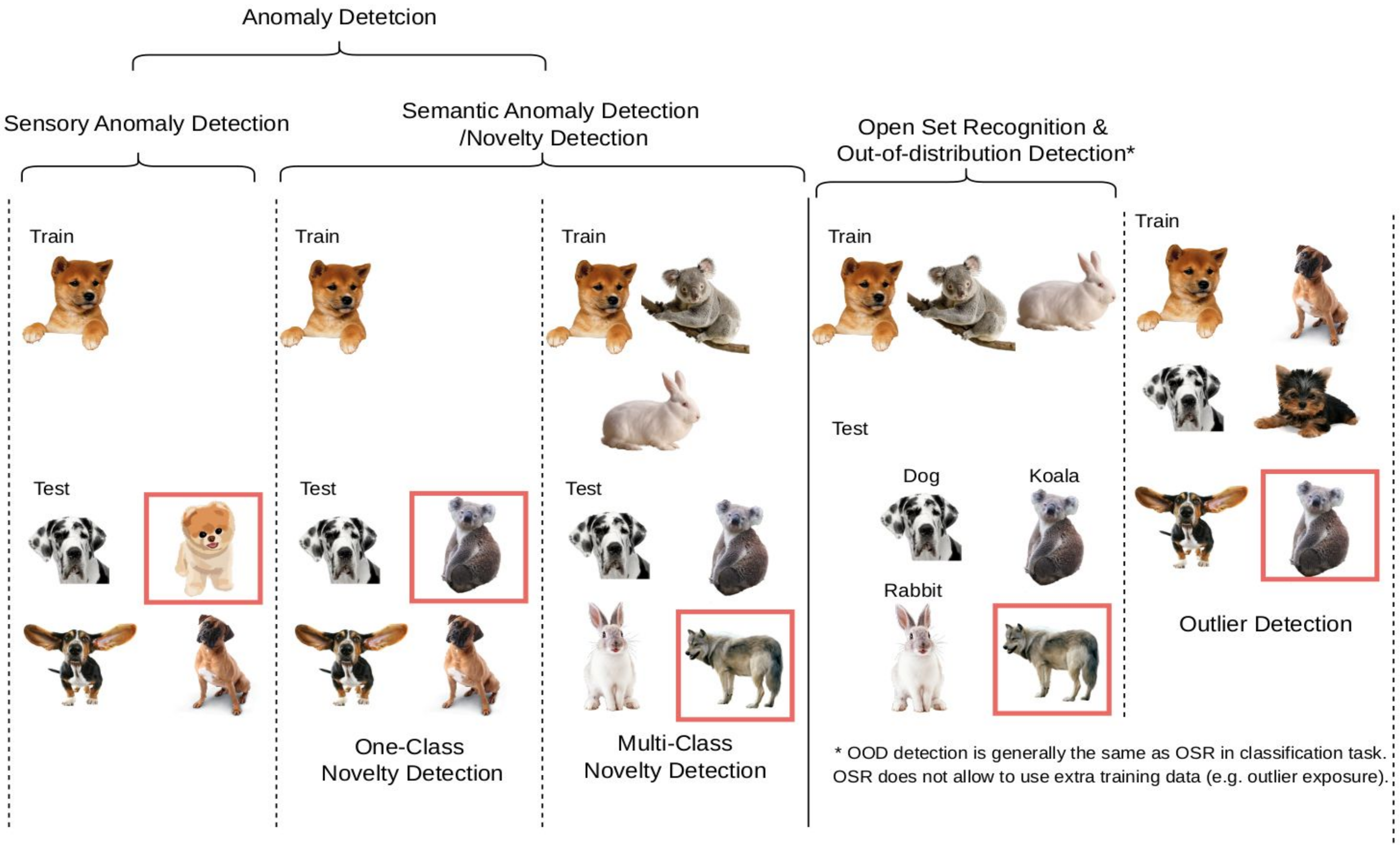
Xixi Liu, Yaroslava Lochman, Christopher Zach

Chalmers University of Technology

Why OOD detection?

- Diverse inputs
- To ensure the reliability of the deep learning models
  - high-stake tasks, such as medical image analysis and autonomous driving.
- OOD detection:
  - differentiates between in-distribution (ID) and out-of-distribution (OoD) inputs at test time.
- A model should know what they do not know.

Generalized OOD detection



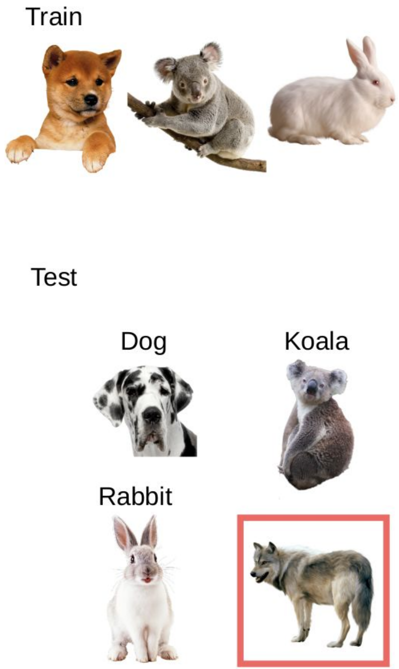
Reproduced from "Generalized OOD Detection: A Survey", Jingkang Yang et al., 2021.

Distributional shift

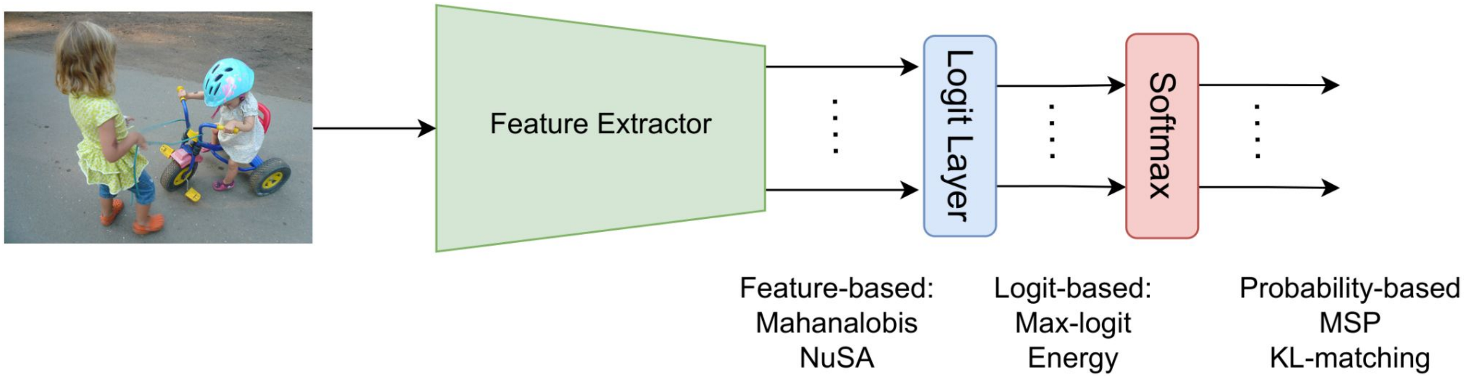
- Semantic shift
  - $P_{\text{train}}(Y) \neq P_{\text{test}}(Y)$
  - The occurrence of new classes
  - Novelty detection and OOD detection
- Covariate shift
  - $P_{\text{train}}(X) \neq P_{\text{test}}(X)$
  - Style change or adversarial examples
  - Sensory anomaly detection

Classification-based method

- Require a softmax-based (pre-trained) classifier
  - **Post-hoc method**
    - aims to design a suitable score function for distinguishing between ID and OOD data accurately given a pre-trained classifier.
  - Enhancing method
    - modifies features from intermediate layers to enhance OOD performance for given score functions.
  - Training loss modification
    - incorporates OOD samples (e.g. outlier exposure/synthesis) in the training procedure to perform OOD detection.

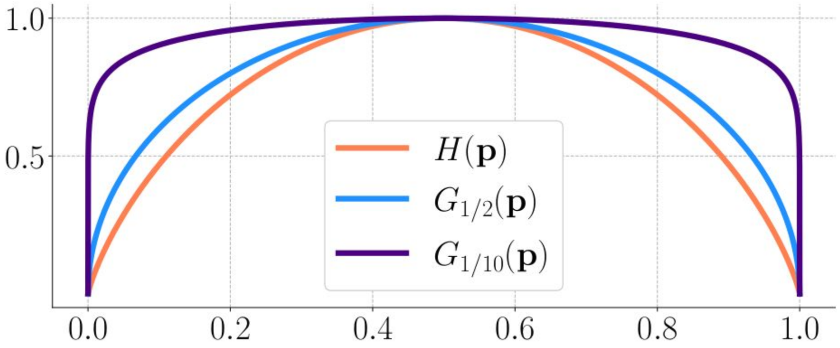


Post-hoc methods



- Overhead cost of retraining is avoided
- Use both feature and predictive distribution
  - GradNorm and predictive normalized maximum likelihood (pNML)
- Use both feature and logit:
  - Virtual-logit Matching (ViM)

Generalized entropy



Assumption: In-distribution test samples close to the training data are expected to result in a confident prediction.

$$G_\gamma(\boldsymbol{p}) = \sum_j p_j^\gamma (1 - p_j)^\gamma, \gamma \in (0, 1).$$

Truncation

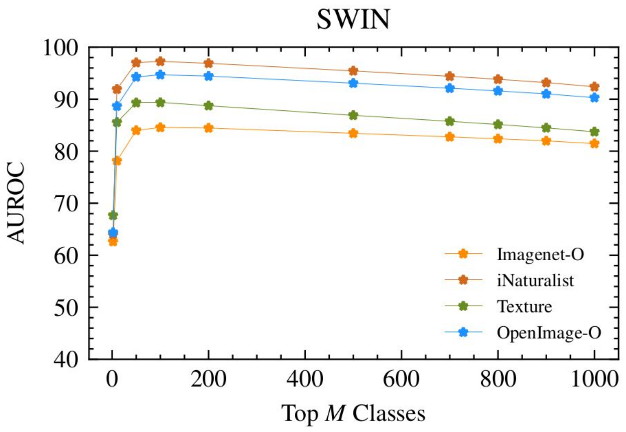
Considering sorted predictive probabilities, $p_{j_1} \geq p_{j_2} \geq \cdots \geq p_{i_C}$,

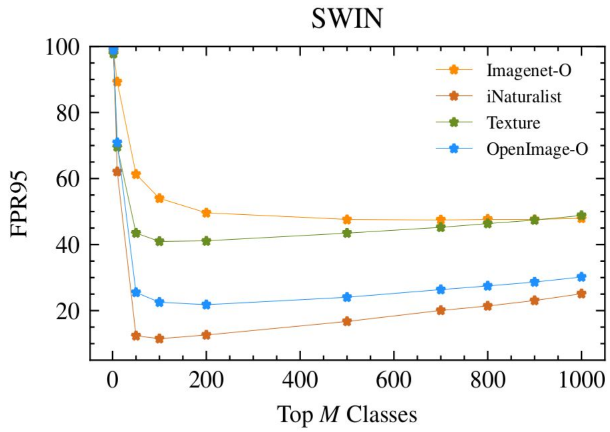
our score is designed to capture small entropy variations in the top-M classes.

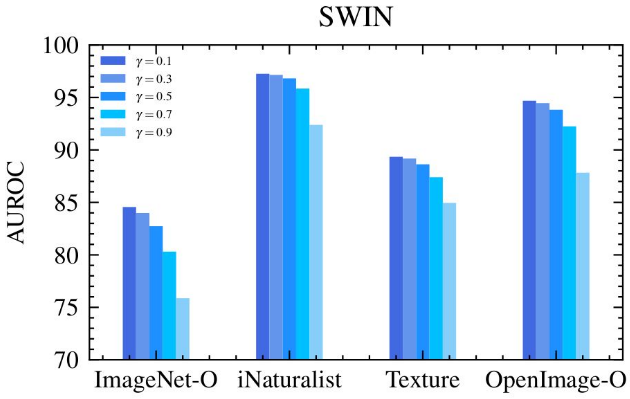
The final score reads as

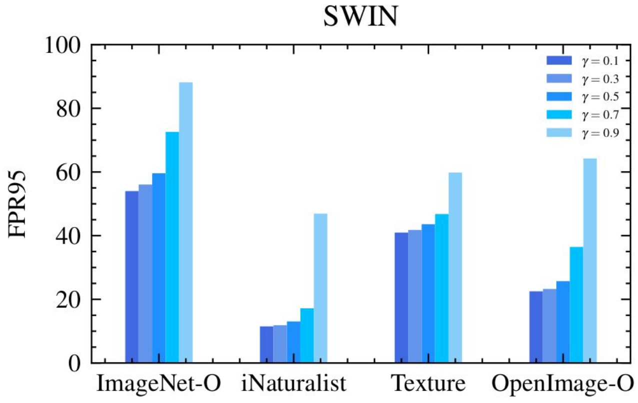
$$-G_\gamma(\boldsymbol{p}) = -\sum_{m=1}^{M} p_{i_m}^\gamma (1 - p_{i_m})^\gamma$$

Selected Results









Selected Results

	OOD Method	OpenImage-O		Textures		iNaturalist		ImageNet-O		Average	
		AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
Post-hoc	MSP [4]	86.62	55.87	82.58	63.20	90.45	44.01	66.56	82.97	81.55	61.51
	MaxLogit [1]	86.26	52.33	82.57	59.18	89.82	43.41	68.77	76.47	81.85	57.85
	EnergyBased [5]	83.91	55.87	80.52	62.79	86.89	51.55	69.01	**73.99**	80.08	61.05
	GradNorm [7]	54.82	78.12	60.31	76.58	56.83	75.14	51.02	85.47	55.75	78.83
	ODIN [6]	86.80	50.74	83.10	58.12	89.62	43.79	68.42	77.09	81.98	57.44
	ReAct*	84.21	55.69	80.96	62.70	87.03	51.29	69.34	74.10	80.39	60.94
	Shannon	81.98	52.06	83.97	59.18	91.48	41.56	68.99	70.71	83.09	57.63
	GEN	89.83	49.04	86.19	**55.65**	93.27	35.59	73.69	77.83	85.74	54.53
	GEN + ReAct*	**90.07**	**49.00**	**86.62**	55.66	**93.38**	**35.54**	**74.11**	77.87	**86.04**	**54.52**
Require ID	KL Matching [1]	89.03	50.57	86.10	55.86	92.45	36.05	72.69	77.97	85.07	55.11
	Mahalanobis [2]	89.56	50.86	91.99	37.62	92.37	42.05	81.89	71.57	88.95	50.52
	ReAct [8]	79.84	54.40	81.92	54.44	82.80	46.29	69.03	72.87	78.40	57.00
	pNML [9]	90.61	**41.76**	89.91	37.20	93.49	**31.42**	73.94	71.12	86.99	45.38
	Residual [3]	87.14	56.00	91.90	36.84	89.41	48.04	81.22	71.20	87.42	53.02
	ViM [3]	91.85	43.16	**93.43**	**30.04**	93.47	37.41	**83.07**	**66.72**	90.45	44.33
	GEN + ReAct [8]	90.59	46.94	88.76	50.91	93.89	32.70	75.76	76.76	87.25	51.83
	GEN + Residual	**92.23**	42.05	93.01	31.69	**94.36**	33.85	82.58	69.24	**90.55**	**44.21**

Conclusion

✓ GEN uses output probabilities only.

✓ It does not use any training data statistics.

✓ It does not require re-training and/or outlier exposure.

Yet it performs very well across four datasets and six architectures, meaning that it can potentially be used in more constrained model deployment scenarios!