Toward Metrics for Differentiating Out-of-Distribution Sets

M. Abbasi[1]; C. Shui[1]; A. Rajabi[2]; C. Gagné[1,3]; R. Bobba[2]

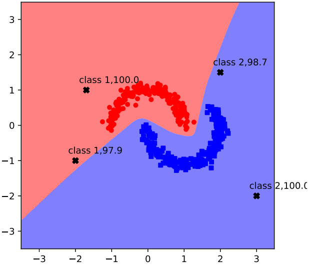
Presented at ECAI 2020

1. IID, Université Laval, Québec, Canada
2. Oregon State University, Corvallis, USA
3. Mila, Canada CIFAR AI Chair

Aug. 29 - Sept. 5, 2020

Out-of-Distribution (OOD): a risk for vanilla CNNs

- **Unreliable** models (e.g. vanilla CNNs) are uncalibrated:
  - High confidence for most samples, drawn from any data distributions.



A vanilla MLP classifies the entire input space into two classes.

- **Reliable CNNs** are calibrated:
  - High confidence on in-distribution samples but low confidence predictions for out-of-distribution ones.

How to detect OOD samples?

End-to-end models by OOD learning; a promising avenue to detect OOD samples:

1. **Explicitly** train a vanilla CNN to output calibrated prediction on OOD samples, then use a **threshold** on the calibrated predictions for detecting OOD samples [1]–[3].
2. **Explicitly** train an Augmented CNN (A-CNN) – a vanilla CNN with an extra class added to its output – with an extra class to assign OOD samples. *[threshold-free]*

An unaddressed central question in OOD learning

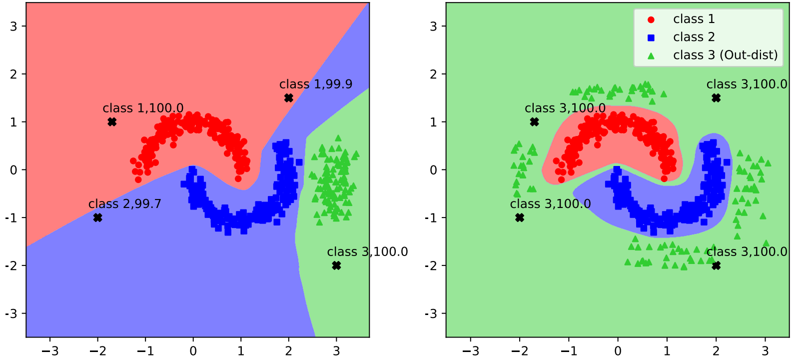
Research question

Among several OOD sets available, how can one identify the most appropriate set for training a calibrated CNN with high detection rate over unseen OOD samples?

Previous methods selected an OOD set manually, without a rigorous justification for their selection.

Our proposal: protective OOD set

- We characterize OOD sets with their **level of protection** of the in-distribution sub-manifolds.
  - How well an OOD set can cover all in-distribution sub-manifolds.



Our metrics for measuring protection level

I) **Softmax-based Entropy**

II) **Coverage Ratio**

III) **Coverage Distance**

Notation:

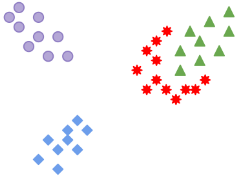
- $\mathcal{S}_O = \{\mathbf{x}_O^j\}_{j=1}^M$: OOD set of M samples
- $\mathcal{S}_I = \{\mathbf{x}_I^i\}_{i=1}^N$: in-distribution training set of N samples
- $h(\cdot)$: a pre-trained vanilla CNN trained on $\mathcal{S}_I$

I) Softmax-based Entropy (SE)

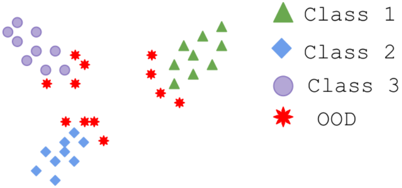
Goal: measure how uniformly the OOD samples $\mathcal{S}_O$ are distributed to the in-distribution sub-manifolds.

$$H(\mathcal{S}_O) = -\sum_{k=1}^{K} p(c = k|\mathcal{S}_O) \log p(c = k|\mathcal{S}_O).$$

$p(c = k|\mathcal{S}_O)$: the ratio of OOD samples classified as k-th class by <u>the vanilla h</u>.



(a) **Small SE**: OOD samples collapse to one manifold.

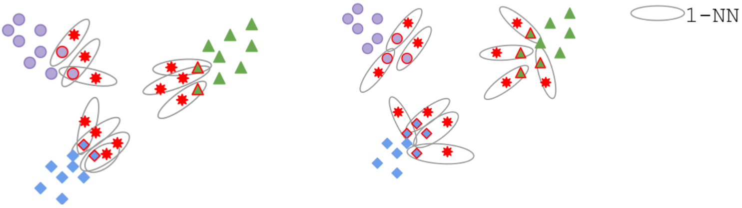


(b) **Large SE**: OOD samples uniformly distributed over all manifolds

II) Coverage Ratio (CR)

Goal: measuring coverage of the sub-manifolds by the OOD samples.

Adjacency matrices $\mathbf{z}_I^i$: in-dist.; $\mathbf{z}_O^j$: out-dist; both in feature space	**Coverage Ratio (CR)**
$\mathrm{W}_{i,j} = \begin{cases} \lVert \mathbf{z}_I^i - \mathbf{z}_O^j \rVert_2 & \text{if } \mathbf{z}_I^i \in \text{k-NN}(\mathbf{z}_O^j, S_I) \\ 0 & \text{otherwise} \end{cases}$ $A_{i,j} = \mathbb{I}(W_{i,j} > 0)$	$\mathrm{R}(\mathrm{S}_I, \mathcal{S}_O) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(\sum_{j=1}^{M}(A_{i,j}) > 0\right)$



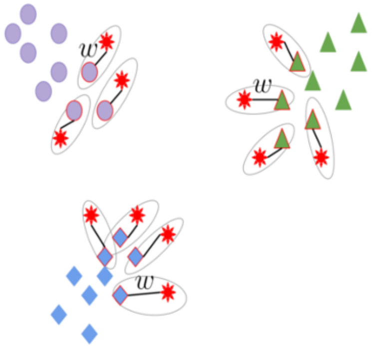
(a) High SE but low CR

(b) High SE and CR

III) Coverage Distance (CD)

Goal: measuring the average distance between an OOD set $\mathcal{S}_O$ and the in-distribution sub-manifolds

$$D(S_I, S_O) = \frac{\sum_{i,j} W_{ij}}{\sum_{i,j} A_{ij}} = \frac{1}{kM} \sum_{i,j} W_{ij}.$$



Protective OOD set

- A protective OOD set has a **high softmax-based Entropy (SE)** and **Coverage Ratio (CR)**
- Preferably, it also has a small coverage distance, placing near to the in-distribution sub-manifolds

Experimentation: benchmark datasets

- **Image classification**
  - In-distribution: SVHN and CIFAR-10
  - Natural OOD sets: LSUN, ISUN, CIFAR-100 and TinyImageNet
  - Synthetic OOD set: Gaussian noise
- **Sound classification**
  - In-distribution: Urban Sound
  - Natural OOD sets: TuT, Google Command, and ECS
  - Synthetic OOD set: White noise

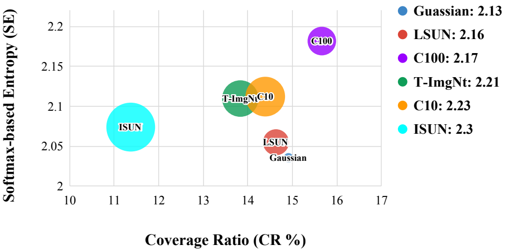
Assessment of our metrics by A-CNNs

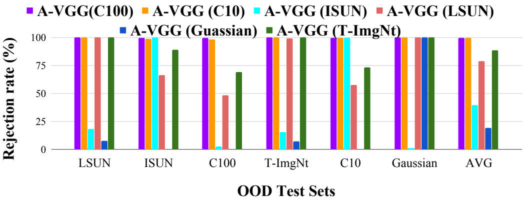
Metrics assessment approach:

1. Identify the most protective OOD set w.r.t an in-distribution task
2. Show that an A-CNN trained on **most protective** OOD set has a **high average detection rate** on unseen OOD sets
3. Show that an A-CNN trained on the **least protective** OOD sets has a **low average detection rate** on unseen OOD sets

Assessment of our metrics (A-CNNs): SVHN

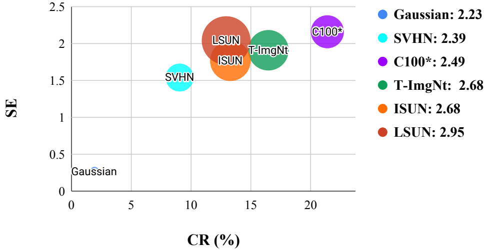
- **Most protective OOD set**: CIFAR-100 (highest SE and CR)
- **Least protective**: ISUN and Gaussian noise

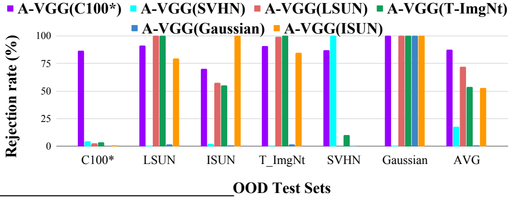




Assessment of our metrics (A-CNN): CIFAR-10

- **Most protective OOD set**: C100*[1] (highest SE and CR)
- **Least protective**: SVHN and Gaussian noise

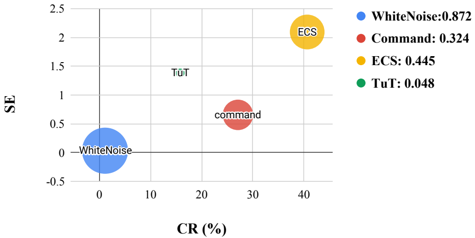


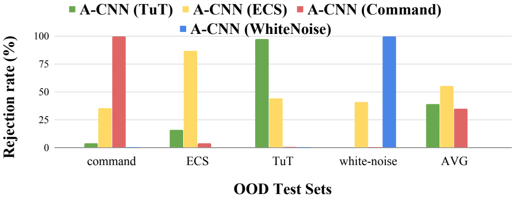


[1] C100* is the modified C100 by removing its classes that have an overlap with those of C10.

Assessment of our metrics (A-CNN): Urban Sound

- **Most protective OOD set**: ECS (highest SE and CR)
- **Least protective**: Command, TuT, and white noise (due to their very low SE)





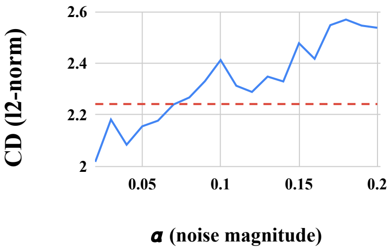
Assessment of metrics: explicitly-calibrated CNNs

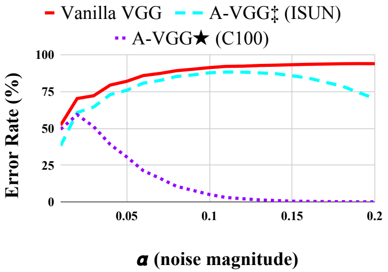
Likewise A-CNNs, **the most protective OOD set** induces to an explicitly-calibrated CNN with **higher average AUROC** and **lower average False Positive Rate** (FPR) @95% TPR.

In-distribution	Training OOD set	Test OOD sets Avg AUROC(↑) / *AvgFPR*(↓)
SVHN	ISUN‡	94.73 / 31.97
	LSUN	99.25 / 4.39
	C10	99.75 / 0.41
	T-ImgNt	99.75 / 1.10
	C100⋆	**99.86** / **0.07**
CIFAR-10	SVHN‡	86.38 / 75.04
	ISUN	86.20 / 77.03
	LSUN	93.31 / 38.59
	T-ImgNt	**93.89** / 34.44
	C100*⋆	93.03 / **26.13**
Urban-Sound	Command‡	59.15 / 63.06
	TuT‡	45.40 / 85.08
	ECS⋆	**71.41** / **60.67**

FGS adversaries rejection

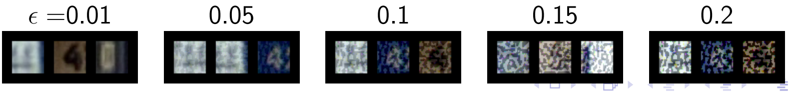
- SVHN adversarial examples
- A-CNN*: trained on the most protective OOD set.
- A-CNN‡: trained on the least protective OOD set.



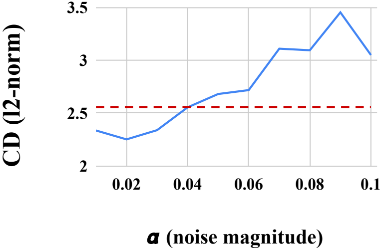


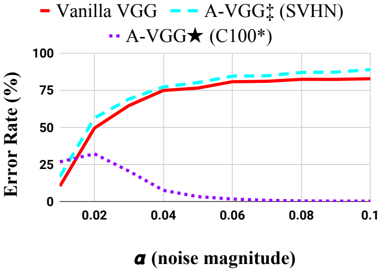
(a) (blue line) CD of SVHN adversaries; (red dotted line) CD of C100 as an OOD*

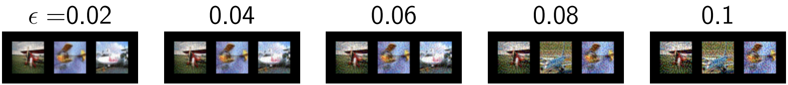
(b) Err. rate = 1-(Acc. + Rej.)



CIFAR-10 adversarial examples







Conclusion

- OOD sets are not equivalent for training a well-generalized A-CNN and explicitly-calibrated CNN with **high OOD detection rate**
- The **protection level is a valid property** to guide selection of an appropriate OOD set
- **Our Metrics** can successfully reveal the most protective OOD set.

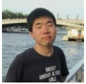
Thanks for your attention

Q&A

Link to our paper: https://arxiv.org/abs/1910.08650

mahdieh.abbasi.1@ulaval.ca

christian.gagne@gel.ulaval.ca



changjian.shui.1@ulaval.ca

rajabia@oregonstate.edu

rakesh.bobba@oregonstate.edu

Bibliography

[1] A. Meinke and M. Hein, "Towards neural networks that provably know when they don't know," in *International Conference on Learning Representations (ICLR)*, 2020.

[2] D. Hendrycks, M. Mazeika, and T. G. Dietterich, "Deep anomaly detection with outlier exposure," *International Conference on Learning Representations (ICLR)*, 2019.

[3] K. Lee, H. Lee, K. Lee, and J. Shin, "Training confidence-calibrated classifiers for detecting out-of-distribution samples," *International Conference on Learning Representations (ICLR)*, 2017.