READ: Reconstruction Error Aggregated Out-of-Distribution Detection

Wenyu Jiang Yuxin Ge Hao Cheng Mingcai Chen Shuai Feng
Chongjun Wang

State Key Laboratory for Novel Software Technology, Nanjing University

Thirty-Seventh AAAI Conference on Artificial Intelligence
Washington, D.C., 2023

Outline

1. Background
   - Out-of-Distribution Detection
   - Distance-based Methods
   - Idea
2. READ
   - Training
   - Transformed Reconstruction Error
   - Adjustment Coefficient based on Image Complexity
   - Inference
   - Overall Concept
3. Experiments
   - Main Results
   - Ablation Results
4. Summary

- **Task:** The trained network deployed in the wild would be exposed to the unknown out-of-distribution (OOD) data, which is different from the known in-distribution (ID) training samples.
- **Aim:** The model should predict correctly on the ID data, and refuse to make inference when the test input is from OOD.
- **Challenge:** The network makes **overconfident** prediction on the OOD data [1].

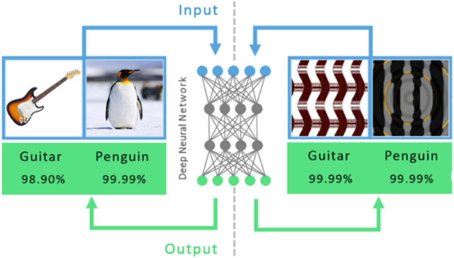

Figure 1: The model makes **overconfident** prediction on unrecognizable OOD!

Background

Distance-based Methods

- Distance-based methods assume that the ID test data is closer to the known training samples with same category than the OOD data. Considering the limitations of classifier retraining in practical scenario, there are two different strategies:
  - **Mahalanobis distance [2] for pre-training**

$$\hat{\mu}_i = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{X}_{\mathrm{in}}^{\mathrm{train}}, y=i}[f_{\mathrm{fe}}(\boldsymbol{x})] \tag{1}$$

$$\hat{\boldsymbol{\Sigma}} = \mathbb{E}_{(\boldsymbol{x}, y) \sim \mathcal{D}_{\mathrm{in}}^{\mathrm{train}}}[(f_{\mathrm{fe}}(\boldsymbol{x}) - \hat{\mu}_y)(f_{\mathrm{fe}}(\boldsymbol{x}) - \hat{\mu}_y)^{\mathsf{T}}] \tag{2}$$

$$Score_{\mathrm{cla}} = -\min_i (f_{\mathrm{fe}}(\boldsymbol{x}) - \hat{\mu}_i)^{\mathsf{T}} \hat{\boldsymbol{\Sigma}}^{-1} (f_{\mathrm{fe}}(\boldsymbol{x}) - \hat{\mu}_i)) \tag{3}$$

  - **Euclidean distance [3] for retraining**

$$f_{\mathrm{ch}_i}(\boldsymbol{z}) = \frac{h_i(\boldsymbol{z})}{g(\boldsymbol{z})} = \frac{-\|\boldsymbol{z} - \omega_i\|_2^2}{\sigma(BN(\omega_{\mathrm{g}} \boldsymbol{z} + b_{\mathrm{g}}))} \tag{4}$$

$$Score_{\mathrm{cla}} = -\min_i (\|\boldsymbol{z} - \omega_i\|_2^2) \tag{5}$$

Background

Idea

- We further extend the above discrepancy of distance to the closest class in latent space with **reconstruction error** from autoencoder.
  - The extracted representations by autoencoder are enforced to contain **important regularities** of the ID data.
  - OOD inputs are **poorly reconstructed** from the resulting representations due to the irregular patterns.

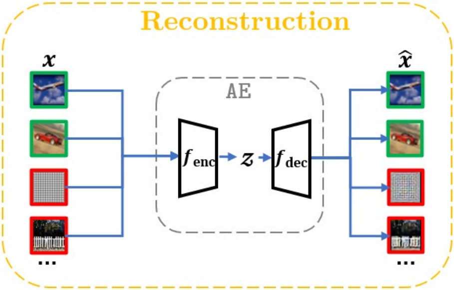

Figure 2: Reconstruction error from autoencoder.

- **Loss function:** The formulation is defined as follows:
  - Classifier (CLF)

$$\mathcal{L}_{\mathrm{CLF}} = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_{\mathrm{in}}^{\mathrm{train}}}[-\log F_y(\boldsymbol{x})] \tag{6}$$

  - Autoencoder (AE)

$$\mathcal{L}_{\mathrm{AE}} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{X}_{\mathrm{in}}^{\mathrm{train}}}[\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2] \tag{7}$$

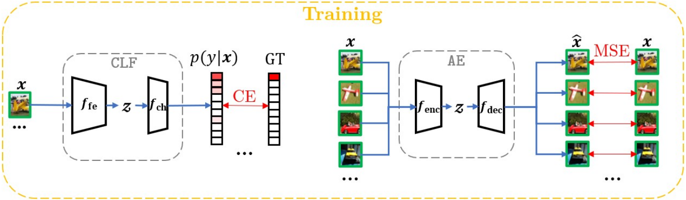

Figure 3: Illustration of training process.

- The CLF and AE are **independent** components.

READ

Transformed Reconstruction Error

- The reconstruction error is measured in the classifier **latent space** instead of raw pixel space.
  - Reach unification with distance measurement.
  - Bridge the semantic gap & show competitive distinguishability.

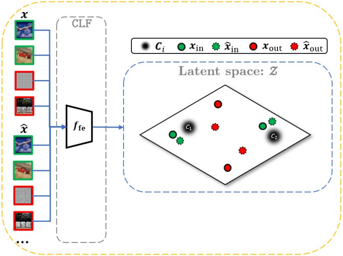

Figure 4: Transformed reconstruction error.

- Pre-training (READ-MD)

$$Score_{\text{rec}} = -((f_{\text{fe}}(\boldsymbol{x}) - f_{\text{fe}}(\hat{\boldsymbol{x}}))^{\top}\hat{\boldsymbol{\Sigma}}^{-1}(f_{\text{fe}}(\boldsymbol{x}) - f_{\text{fe}}(\hat{\boldsymbol{x}})) \quad (8)$$

- Retraining (READ-ED)

$$Score_{\text{rec}} = -(\|\boldsymbol{z} - \hat{\boldsymbol{z}}\|_2^2) \quad (9)$$

- **Overconfidence!** The transformed reconstruction error is small for specific OOD data.

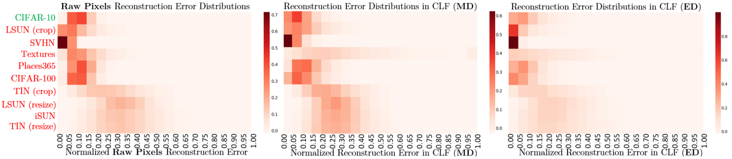

Figure 5: Overconfident reconstruction error.

- **Observation:** The reconstruction error and image complexity is correlated. Simpler representations are required for easy image description, thus bring smaller reconstruction error.

- **Adjustment.** Adjust the overconfident reconstruction error with image complexity.
  - **Characterization of OODs:** A proxy to quantify the "easiness" of OOD by off-the-shelf lossless image compression algorithm [4, 5].
  - **Re-scale reconstruction error:** The transformed reconstruction error for OOD input with small image complexity is re-scaled by coefficient λ.

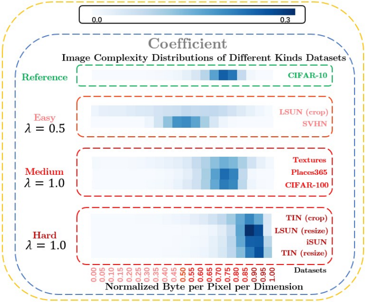

Figure 6: Adjust reconstruction error.

- **Score function:** The definition of score function is defined as follows:

$$Score = -Score_{\mathrm{cla}} - \lambda * Score_{\mathrm{rec}} \tag{10}$$

- **Input perturbation:** This strategy brings larger gain on *Score* for ID samples [6].

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} - \epsilon * \mathrm{sign}(-\nabla_{\boldsymbol{x}}(Score_{\mathrm{cla}}(\boldsymbol{x}) + Score_{\mathrm{rec}}(\boldsymbol{x}, \hat{\boldsymbol{x}}))) \tag{11}$$

  - Considering that test time OOD data is unavailable, the choice of hyperparameters depends on metric FPR@TPR95 of ID and synthesized OOD data.

- **Illustration of the proposed method.**

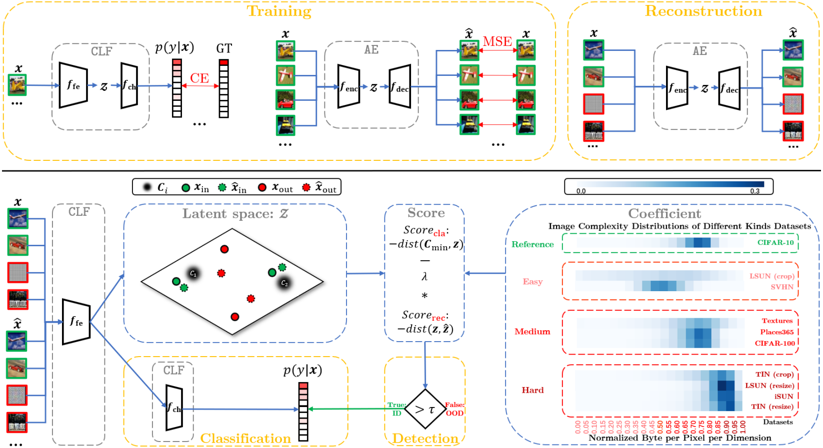

Figure 7: READ

- We do not rely on real auxiliary OOD training data.
- READ achieves **SOTA** performance under both pre-training and retraining scenarios.
  - READ-MD

ID	OOD	FPR@95TPR ↓	AUROC ↑
		MSP / ODIN / Maha / Energy / READ-MD (ours)	
CIFAR-10	SVHN	48.3 / 33.2 / 15.3 / 35.4 / **12.0**	91.9 / 92.0 / 97.0 / 91.1 / **97.5**
	LSUN (c)	42.4 / 29.7 / 31.6 / **19.1** / 28.3	93.6 / 92.8 / 94.1 / **96.0** / 94.9
	Textures	59.5 / 49.5 / 18.0 / 52.5 / **10.3**	88.4 / 84.7 / 96.3 / 85.4 / **98.0**
	Places365	60.5 / 57.7 / 74.2 / **40.9** / 75.5	88.1 / 84.3 / 80.3 / **89.7** / 80.7
	CIFAR-100	62.9 / 60.7 / 71.8 / **50.5** / 76.5	**87.8** / 82.7 / 79.7 / 87.1 / 79.2
	TIN (c)	54.3 / 37.3 / 37.7 / 38.3 / **19.9**	90.5 / 91.6 / 92.9 / 91.5 / **96.5**
	LSUN (r)	52.0 / 26.5 / 34.1 / 27.9 / **9.4**	91.5 / 94.6 / 94.2 / 94.1 / **98.3**
	TIN (r)	60.8 / 39.1 / 34.1 / 46.5 / **12.3**	88.2 / 91.3 / 93.5 / 89.0 / **97.7**
	iSUN	56.4 / 32.4 / 33.5 / 33.9 / **12.5**	89.9 / 93.4 / 93.9 / 92.6 / **97.6**
	average	55.2 / 40.7 / 38.9 / 38.3 / **28.5**	90.0 / 89.7 / 91.3 / 90.7 / **93.4**
CIFAR-100	SVHN	85.0 / 82.1 / **58.0** / 92.2 / 67.9	70.3 / 69.1 / **85.3** / 73.6 / 81.8
	LSUN (c)	79.0 / 66.8 / 63.5 / 75.4 / **61.7**	77.6 / 81.2 / 82.0 / 83.1 / **83.1**
	Textures	83.1 / 78.8 / 36.9 / 78.0 / **35.6**	73.4 / 72.9 / 90.9 / 76.0 / **92.1**
	Places365	82.9 / 88.4 / 90.6 / **81.3** / 91.7	73.4 / 70.5 / 64.5 / **75.4** / 63.3
	CIFAR-10	**81.8** / 89.2 / 93.9 / 82.4 / 95.0	75.1 / 70.1 / 61.9 / **77.2** / 69.3
	TIN (c)	78.5 / 74.4 / 41.5 / 63.1 / **29.8**	76.5 / 80.0 / 91.0 / 81.2 / **93.6**
	LSUN (r)	82.5 / 73.9 / 22.7 / 62.0 / **10.9**	74.5 / 80.3 / 95.7 / 79.1 / **97.6**
	TIN (r)	82.3 / 71.6 / 25.3 / 63.5 / **14.7**	73.7 / 80.2 / 94.8 / 77.5 / **97.0**
	iSUN	83.1 / 70.6 / 26.2 / 62.3 / **15.5**	75.0 / 81.4 / 94.3 / 78.9 / **96.3**
	average	82.0 / 77.3 / 51.0 / 73.4 / **47.0**	74.4 / 76.2 / 84.5 / 78.0 / **84.9**

Table 1: Comparison with post-hoc methods. ↑ (↓) indicates larger (smaller) values are better. **Bold** numbers are superior.

- READ achieves **SOTA** performance under both pre-training and retraining scenarios.
  - READ-ED

ID	OOD	FPR@95TPR ↓	AUROC ↑
		G-ODIN-I / G-ODIN-C / G-ODIN-E / READ-ED (ours)	
CIFAR-10	SVHN	11.1 / 9.7 / **8.3** / 10.3	98.0 / 98.1 / **98.2** / 97.9
	LSUN (c)	6.1 / 11.0 / 3.1 / **2.8**	98.9 / 97.9 / 99.3 / **99.4**
	Textures	26.6 / 22.0 / 19.3 / **14.9**	94.9 / 96.0 / 96.7 / **97.4**
	Places365	42.0 / 34.1 / 25.8 / **25.7**	91.4 / 92.6 / 94.6 / **94.6**
	CIFAR-100	53.7 / 45.2 / 45.1 / **44.7**	88.3 / 89.9 / 90.7 / **90.8**
	TIN (c)	8.1 / 20.9 / 8.1 / **4.2**	98.5 / 96.2 / 98.5 / **99.1**
	LSUN (r)	3.0 / 13.4 / 2.7 / **1.3**	99.3 / 97.4 / 99.3 / **99.7**
	TIN (r)	6.2 / 24.0 / 8.6 / **4.5**	98.8 / 95.6 / 98.3 / **99.1**
	iSUN	2.8 / 16.1 / 2.7 / **1.5**	99.3 / 97.0 / 99.3 / **99.6**
	average	17.7 / 21.8 / 13.7 / **12.2**	96.4 / 95.6 / 97.2 / **97.5**
CIFAR-100	SVHN	65.6 / 78.2 / **36.6** / 63.9	85.2 / 83.6 / **94.0** / 89.5
	LSUN (c)	35.3 / 46.2 / **25.4** / 31.1	93.3 / 90.4 / **95.4** / 94.6
	Textures	80.0 / 40.7 / 21.7 / **17.9**	77.2 / 91.7 / 95.5 / **96.3**
	Places365	79.5 / **76.6** / 81.4 / 83.3	76.8 / **77.5** / 76.4 / 75.7
	CIFAR-10	**83.6** / 84.1 / 87.1 / 90.5	71.2 / **75.0** / 70.5 / 69.3
	TIN (c)	63.1 / 51.0 / 25.9 / **14.5**	87.1 / 90.1 / 95.3 / **97.5**
	LSUN (r)	75.6 / 56.7 / 22.9 / **6.5**	85.2 / 88.6 / 95.7 / **98.7**
	TIN (r)	73.5 / 51.0 / 20.6 / **7.9**	84.6 / 89.8 / 96.0 / **98.5**
	iSUN	78.6 / 57.0 / 24.7 / **10.5**	83.8 / 88.7 / 95.2 / **97.9**
	average	69.5 / 60.1 / 38.5 / **36.2**	82.7 / 86.1 / 90.4 / **90.9**

Table 2: Comparison with retraining methods. ↑ (↓) indicates larger (smaller) values are better. **Bold** numbers are superior.

- The extension from transformed reconstruction error improves the discrimination between ID & OOD.
- Adjustment coefficient and perturbation strategy play a vital role in READ.

Method	FPR@95TPR ↓	AUROC ↑
	$-Score_{\text{cla}}$ / $-Score_{\text{rec}}$ / $-(Score_{\text{cla}} + Score_{\text{rec}})$	
READ-MD	46.3 / 55.3 / **37.6**	90.2 / 75.4 / **90.8**
READ-ED	13.7 / 78.7 / **12.4**	97.2 / 59.1 / **97.5**

Table 3: OOD detection results for combination study. ↑ (↓) indicates larger (smaller) values are better. The results are averaged on nine OOD test datasets. **Bold** numbers are superior results.

Method	Adj	Pert	FPR@95TPR ↓	AUROC ↑
READ-MD	-	-	37.6	90.8
	-	✓	29.9	92.7
	✓	-	33.3	92.3
	✓	✓	**28.5**	**93.4**

Table 4: OOD detection results for ablation study. ↑ (↓) indicates larger (smaller) values are better. **Bold** numbers are superior results. Adj and Pert mean adjustment and perturbation respectively.

Summary

- **Contribution**
  - We propose a novel reconstruction error aggregated detector (READ) and its two variants, READ-MD and READ-ED, which combine the distance to the closest class and reconstruction error in the latent space of classifier.
  - Against the overconfidence of transformed reconstruction error, we explain and alleviate this problem by a fine-grained characterization of OOD data and an image complexity based adjustment coefficient.
  - We conduct comprehensive analysis with experiments under both scenarios to demonstrate the effectiveness of the proposed methods.
- Learn More!
  - Paper: `https://arxiv.org/abs/2206.07459`
  - Code: `https://github.com/lygjwy/READ`
  - Contact: `https://lygjwy.github.io`

References I

Anh Nguyen, Jason Yosinski, and Jeff Clune.
Deep neural networks are easily fooled: High confidence predictions for unrecognizable images.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin.
A simple unified framework for detecting out-of-distribution samples and adversarial attacks.
In *NeurIPS*, 2018.

Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira.
Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10951–10960, 2020.

References II

Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F Núñez, and Jordi Luque.
Input complexity and out-of-distribution detection with likelihood-based generative models.
arXiv preprint arXiv:1909.11480, 2019.

Ziqian Lin, Sreya Dutta Roy, and Yixuan Li.
Mood: Multi-level out-of-distribution detection.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15313–15323, 2021.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant.
Enhancing the reliability of out-of-distribution image detection in neural networks.
arXiv preprint arXiv:1706.02690, 2017.