SyntheticData4ML Workshop in NeurIPS 2023

Diffusion-based Semantic-Discrepant Outlier Generation for Out-of-Distribution Detection

Suhee Yoon*, Sanghyu Yoon*, Hankook Lee, Sangjun Han,
Ye Seul Sim, Kyungeun Lee, Hyeseung Cho, Woohyung Lim

LG AI Research

INTRODUCTION

- **Out-of-distribution (OOD) detection** aims to detect whether a given sample is drawn from the in-distribution (ID) or not. Among a number of OOD detection methods, one promising approach is learning a detector using auxiliary OOD dataset, as pioneered by *Outlier Exposure (OE)*. This makes learning easier since such OOD dataset can provide additional information about discrepancy between ID and OOD.
- The crucial properties for effective synthetic OOD dataset are outliers should be OOD with respect to semantics while preserving nuisances (e.g., background) which have no intrinsic relevance to the semantic.
- **Contribution**
  - We introduce a novel and effective detection framework that consists of **Semantic-Discrepant (SD) Outlier** generation via a diffusion model, and OOD detection with SD outliers.

METHODOLOGY

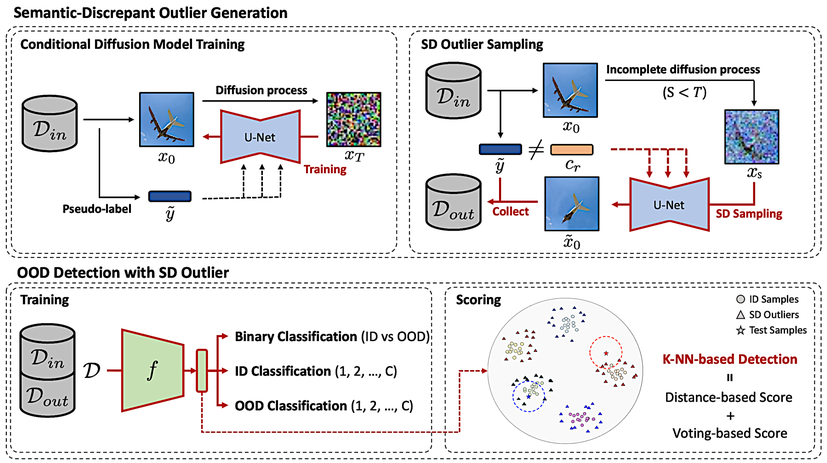


Figure 1. Illustration of our proposed methods framework. This first generates Semantic-Discrepant (SD) Outliers using a conditional diffusion model (top), and then trains a detector using both ID samples and the generated outliers (bottom). For scoring, we use distance-based and voting-based detection scores based on K nearest neighbors on the embedding space.

- **Semantic-Discrepant Outlier Generation**
  - a. **Semantic-Aware Diffusion Model Training** is based on a conditional diffusion model with a pseudo-label $\tilde{y}$ obtained from SCAN clustering, $\mathcal{D}_{in} \coloneqq \{\mathbf{x}^{(i)}, \tilde{y}^{(i)}\}$. We utilize Classifier-free Diffusion Guidance by optimizing with a guidance on the condition $\boldsymbol{c}_d = (\tilde{y}, t)$.

$$\mathcal{L}(\theta) = \mathbb{E}_{(\mathbf{x}_0,\tilde{y})\sim\tilde{\mathcal{D}}_{\text{in}},t\sim\text{Uniform}([0,T]),\epsilon\sim\mathcal{N}(\mathbf{0},I)}[\|\epsilon_\theta(\mathbf{x}_t,\mathbf{c}_d)-\epsilon\|_2^2].$$

  - b. **Semantic-Discrepant Outlier Sampling.** Our key idea is *semantic-discrepant guidance*, sampling with $\boldsymbol{c}_r$ which is inconsistent with $\tilde{y}$. We obtain semantically shifted $\tilde{\mathbf{x}}$ by sampling from $\mathbf{x}_S$ that incompletely diffused to a limited timesteps S, $\mathcal{D}_{in} \coloneqq \{\tilde{\mathbf{x}}^{(i)}, \tilde{y}^{(i)}\}$.

$$c_r \sim \text{Uniform}(\{1,\dots,C|\, c_r \neq \tilde{y}\}). \qquad \epsilon_t = (1+w)\epsilon_\theta(\mathbf{x}_t,\mathbf{c}_r) - w\epsilon_\theta(\mathbf{x}_t)$$

- **OOD Detection with Semantic-Discrepant Outlier**
  - a. **Training with SD Outlier.** Our loss force the model distinguish ID samples from SD outliers while exposing the original semantic to both ID and outlier samples in a different degree.

$$\mathcal{L}(\mathbf{x},\tilde{y}) = \mathcal{L}_{\text{CE}}(g_{\text{bin}}(f(\mathbf{x})),I(\mathbf{x})) + (1-I(\mathbf{x}))\mathcal{L}_{\text{CE}}(g_{\text{in}}(f(\mathbf{x})),\tilde{y}) + I(\mathbf{x})\lambda\mathcal{L}_{\text{CE}}(g_{\text{out}}(f(\mathbf{x})),\tilde{y}) \qquad I(\mathbf{x}) = \begin{cases} 0, & \mathbf{x}\in\mathcal{D}_{\text{in}} \\ 1, & \mathbf{x}\in\mathcal{D}_{\text{out}} \end{cases}$$

  - b. **Scoring with SD Outlier.** We derive OOD score function in two different ways based on k nearest neighbors.

$$\text{OOD Score}(x) = \underbrace{\sum_{i=1}^{k}\|f(\mathbf{x})-f(\mathbf{x}_d^i)\|^2}_{\text{distance-based score}} + \underbrace{\alpha\sum_{i=1}^{k}\mathbb{1}(I(\mathbf{x}_v^i)=1)}_{\text{voting-based score}}$$

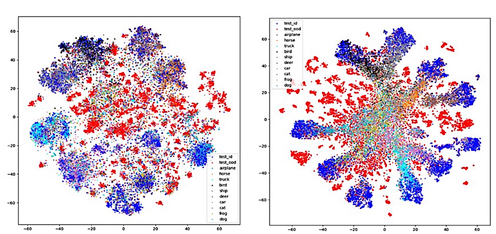
EXPERIMENTS

- **Main Results**
- Our method outperforms previous methods in all benchmarks dataset. Especially, one notable result is we almost reaching to ground-truth level performance in CIFAR-100 dataset (98.2%).
- Our samples successfully maintain nuisances,but cause crucial semantic corruption. Furthermore, ours exhibit a highly realistic appearance with FID score less than 8 while original sampling method shows 2.97 and Fake-it 45.

Table 1. OOD Detection AUROC (%) on various benchmark datasets.

	Methods	Networks	(In) CIFAR-10		
			CIFAR-100	SVHN	LSUN
Likelihood	Likelihood	PixelCNN++	52.6	8.3	-
	Likelihood ratio [39]	PixelCNN++	-	91.2	
	Input Complexity [40]	Glow	73.6	95.0	
Self-supervised	Rot [41]	ResNet-18	79.0	97.6	89.2
	GOAD [42]	ResNet-18	77.2	96.3	89.3
	CSI [43]	ResNet-18	89.2	99.8	97.5
	SSD [44]	ResNet-18	89.6	-	-
Pre-trained	DN2 [45]	ResNet-18	83.3	88.9	91
	DN2 [45]	ResNet-152	86.5	96.2	88.7
	MSCL [46]	ResNet-152	90.0	98.6	90.6
	Multi-class AD [38]	ResNet-18	90.8	98.6	98.6
	Multi-class AD [38]	ResNet-152	93.3	99.8	95.4
	Multi-class AD [38]	ViT-B/16	96.7	99.9	99.3
	Fake-it [29]	ViT-B/16	95.7	99.9	99.4
	Ours	ViT-B/16	**98.0**	**99.9**	**99.9**

Figure 2. Comparing generated outlier examples on CIFAR-10 (32x32 resolution) with diffusion-based methods.



(a) (b)

Figure 3. t -SNE visualization of the embedding space. Blue points is ID (CIFAR-10), red for OOD (CIFAR-100) and SD-outliers are different colors for each pseudo-label ỹ. (a) Eearly epoch of training, (b) last epoch of training

- **Ablation Study on CIFAR-10 (ID) vs CIFAR-100 (OOD)**
- We confirm the robust to sampling hyperparameters, diffusion timesteps S,and guidance weight w,
- Merging with voting score consistently improves in various sampling timesteps S,.
- The SOTA baselines show severe performance degradation on our SD outlier test dataset.

Table 2: Ablation study of sampling hyperparameter,

	Sampling timesteps S				
	50	80	100	150	200
w=2.0	97.6	97.8	**98.0**	97.8	97.4
w=3.0	97.5	97.8	97.9	**98.0**	97.6
w=4.0	97.7	97.7	**98.0**	97.8	97.6

Table 3. Ablation study of OOD scoring function

	Sampling timesteps S				
	50 (FID=4.45)	80 (FID=5.95)	100 (FID=6.88)	150 (FID=7.85)	200 (FID=7.59)
Distance score	97.2	97.5	**97.7**	97.6	97.3
Distance + Voting Score	97.6	97.8	**98.0**	97.8	97.4

Table 4. Performance of SD outlier as test dataset

Setting	Multi-class AD	Fake-it	Ours
CIFAR-10 vs CIFAR-100	96.7	95.7	**98.0**
CIFAR-10 vs SD outliers	74.3	81.4	**98.2**

CONCLUSION

- In this paper, we introduce Semantic-Discrepant (SD) outlier generation and application to OOD detection framework. Our key concept is semantic-discrepant guidance, generating realistic outliers that semantically shifted while retaining nuisances found in ID.
- Experimental results demonstrate the effectiveness of our approach on several OOD detection benchmarks. It has been proven that our SD outliers can be served as effective auxiliary OOD to learn detector without any additional dataset acquisition efforts.

Paper