Improving the out-of-distribution performance of score-based generative models via self-supervision

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this work, we first examine the efficacy of score-based generative models (SGMs) for out-of-distribution (OOD) detection. We show previously proposed OOD detection metrics based on SGMs fail to address OODs that share similar textures but different object shapes. Based on the observation, we construct RotNCSN, a novel OOD detection method based-on the score matching and data augmentation. RotNCSN first applies random rotation to the perturbed data and forces its output to be rotation-invariant. Therefore, RotNCSN becomes more shape-aware. Experiment results show that RotNCSN consistently improves over the baseline metric based on the SGMs. Furthermore, RotNCSN also achieves competitive OOD detection performance in the FashionMNIST domain.

1 Introduction

Score-based generative models (SGMs) Song and Ermon [2019], Ho et al. [2020], Song et al. [2021] have emerged as a promising method for deep generative modelling on various domains Dhariwal and Nichol [2021], Xu et al. [2022] due to its competitive performance and stable training. Furthermore, they have been successfully applied in various image-based subtasks, including stroke-based editing Meng et al. [2022], super-resolution Hoogeboom et al. [2022], and segmentation Baranchuk et al. [2022]. However, most of the applications are based on image generation and relatively little work has been devoted to applying SGMs to hypotheis testing, including out-of-distribution (OOD) detection. OOD detection aims to design a reliable metric that discriminates the given distribution from the others. Since generative models naturally model in-distribution images, they are widely applied for OOD detection Havtorn et al. [2021], Xiao et al. [2020]. However, there are few investigations on the potential of SGM in OOD detection.

In this paper, we first question whether previous OOD detection metrics based on SGMs determine data based on its object shape. For example, in the work of Yang et al. [2021], OOD detection methods based on the classifier trained in the CIFAR-10 cat and dog images assign higher confidence to the CIFAR-10 data that come from different classes than the dog image from the ImageNet data with negligible covariate shift. We take this analog to an unsupervised setting and design OOD data that share a similar texture to the in-distribution data. OOD detection metrics based on the norm of the score function Mahmood et al. [2021] and reconstruction loss are vulnerable to such OOD data that share a similar texture.

To overcome such issues, we propose RotNCSN, an OOD detection method that integrates score matching with data augmentation. RotNCSN is trained to be rotation-invariant and therefore be more shape-aware. We test RotNCSN in various OOD detection benchmarks. Compared to the previous score matching-based methods, RotNCSN achieves considerable improvement. Moreover, RotNCSN shows competitive performance against state-of-the-art baselines on unsupervised OOD detection.

Under review at the NeurIPS 2022 Workshop on Score-Based Methods. Do not distribute.

2 Background

In this section, we introduce the formulation of OOD detection. Furthermore, we study examples of score-based methods applied to OOD detection.

2.1 OOD detection

In OOD detection, we want to distinguish the given distribution $\mathcal{D}$ from the others. Therefore, given $\mathcal{D}_{\text{train}} \subset \mathcal{D}$, we design a metric f that can discriminate samples from $\mathcal{D}_{\text{test}} \subset \mathcal{D}$ from the other outlier distributions. We use binary hypothesis testing measures to evaluate the model; area under the ROC curve (AUROC) or detection accuracy.

In unsupervised OOD detection where only the data information is applied, generative models have been widely used to extract reliable metrics. In the beginning, the likelihood or discriminator output of the generative models, including VAE Kingma and Welling [2014], GAN Goodfellow et al. [2014], and GLOW Kingma and Dhariwal [2018], is used. However, Nalisnick et al. [2019] found that such metrics can be vulnerable to distinguishing simple OODs, such as SVHN from CIFAR-10 Krizhevsky and Hinton [2009]. Various methods are proposed to explain such phenomena; complexity of the data Choi and Chung [2020], Serrà et al. [2020], overfitting to low-level features Havtorn et al. [2021], Kirichenko et al. [2020], Schirrmeister et al. [2020], Zhang et al. [2021], and overfitting into backgrounds ren et al. [2019]. While SGMs do considerably better in such dataset Mahmood et al. [2021], we show they are still vulnerable to OODs generated by geometrical transformations.

2.2 SGMs and application to OOD detection

In this section, we introduce MSMA Mahmood et al. [2021], an OOD detection method that utilizes the score function of SGM. MSMA use NCSN Song and Ermon [2019] as a base model for training the score function. NCSN utilizes a score function $s_\theta(x, l)$ that takes perturbed data and the degree of perturbation as input. NCSN trains to match the gradient of log-likelihood of the perturbed data in multiple scales $(\sigma_i)_{i=1}^{L}$. We follow the denoising score matching version of NCSN that is optimized to minimize the loss function below.

$$\frac{1}{2L}\sum_{i=1}^{L}\sigma_i^2\mathbb{E}_{x\in D_{\text{train}}}\mathbb{E}_{\tilde{x}\sim\mathcal{N}(x,\sigma_i I)}\left[\left\|s_\theta(\tilde{x},\sigma_i)+\frac{\tilde{x}-x}{\sigma_i^2}\right\|_2^2\right] \tag{1}$$

MSMA utilizes the normality vector $f_{\text{msma}}(x)$ for the observed data x, which is based on the multi-scale score function. f_{msma} is a L-dimensional vector where each indice is defined as below.

$$f_{\text{msma}}(x)_i = \|\sigma_i s_\theta(x,\sigma_i)\| \tag{2}$$

MSMA trains an unsupervised one-class classification model based on the normality vector and uses its likelihood to detect OOD data. Alongside the score function, we also explore the choice of normality vector f_{rec} based on the reconstruction error of the perturbation.

$$f_{\text{rec}}(x)_i = \mathbb{E}_{\tilde{x}\sim\mathcal{N}(x,\sigma_i I)}\left\|s_\theta(\tilde{x},\sigma_i)+\frac{\tilde{x}-x}{\sigma_i^2}\right\| \tag{3}$$

While computing the expectation may require multiple iterations, we found that the number of iterations does not affect the normality vector much. MSMA learns an unsupervised model (e.g GMM) over the normality vector of training data to extract scalar metrics. While MSMA shows state-of-the-art performance in some OOD detection tasks Mahmood et al. [2021], we show that they show underwhelming performance in detecting various OODs.

Method	CIFAR-100	rot90	rot180	rot270	Patch shuffle
f_{msma}	0.615	0.554	0.542	0.562	0.737
f_{rec}	0.607	0.556	0.546	0.570	0.737

Table 1: AUROC of previously proposed SGM-based OOD detection methods trained in the CIFAR-10 dataset tested in the proposed OOD dataset. All the methods show underwhelming performance on the OOD data.

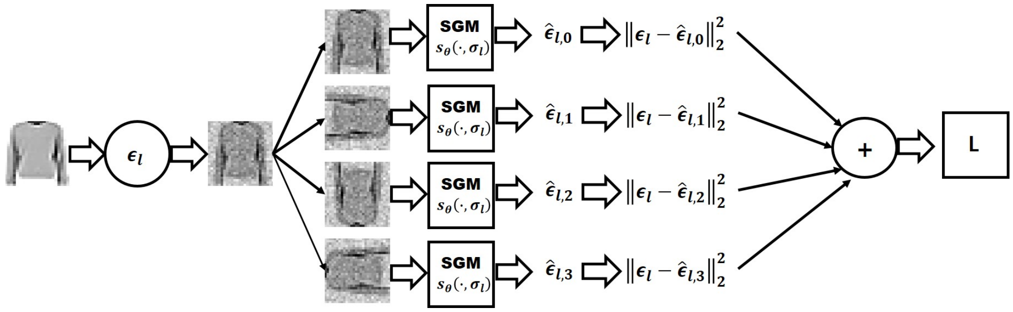

Figure 1: Visual schematic of RotNCSN. RotNCSN forces the output to be rotation-invariant concerning the perturbed data.

3 Methodology

3.1 Proposed Method

3.2 Motivation

We first ask our motivating question: do SGMs recognize in-distribution data via object shape instead of texture? For example, if a classifier model changes its prediction when the texture changes, the model is likely to predict the data by the texture, not by the object shape Geirhos et al. [2019]. Since we are dealing with the unsupervised setting, we test SGMs against the model that shares similar backgrounds but differs in shape. Specifically, we test the MSMA Mahmood et al. [2021] method on the NCSN trained in the CIFAR-10 Krizhevsky and Hinton [2009] dataset against the following OODs.

• **CIFAR-100** is known as near-OOD data for CIFAR-10 since they share similar textures. Since CIFAR-100 and CIFAR-10 are both subsets of the 80-million image dataset, an OOD detector should be aware of class-wise discriminability. This is challenging for an unsupervised OOD detection setting.

• **Rotation** is also a plausible OOD to check the dependence of metric to object orientation. In the CIFAR-10 dataset, most data share a similar shape orientation. For example, there is no deer standing upside-down in the CIFAR-10 dataset. Therefore, we regard rotated data as OOD data Gossweiler et al. [2009]. We test the CIFAR-10 dataset that rotated 90, 180, and 270 degrees counter-clockwise and refer to them as rot90, rot180, and rot270.

• **Patch shuffling** extracts the patch from the image and shuffles the order of the patch to construct the OOD data. We divide the image into 16 8×8-sized patches and shuffle them in random order. This operation relatively destroys the object's shape compared to the texture Noroozi and Favaro [2016].

We now report the performance of GMMs trained on f_{msma} Mahmood et al. [2021] and f_{rec} in the following OODs in Table 1. Both metrics show underwhelming performance in detecting OOD images that share similar textures. We further show that the trend is consistent when evaluated in the alternative dataset, SVHN. We refer the results to Appendix.

We now introduce our proposed scheme RotNCSN. In the training phase, RotNCSN applies random rotation $r \in \mathcal{R} = \{\text{Rot}(X, 0), \text{Rot}(X, 90), \text{Rot}(X, 180), \text{Rot}(X, 270)\}$. Then, RotNCSN minimizes the loss function below.

OOD	RotNCSN (**ours**)	f_{msma}	f_{rec}	LR	IC	Likelihood Ratio
EMNIST	**0.982**	0.961	0.937			
MNIST	**0.994**	0.828	0.842	0.967	0.946	0.924
NotMNIST	0.978	0.932	0.892	**1.0**	0.923	0.996
KMNIST	**0.988**	0.901	0.893	0.983	0.708	0.983

OOD	RotNCSN	f_{msma}	f_{rec}
CIFAR-100	**0.678**	0.615	0.607

Table 2: AUROC of RotNCSN and other OOD detection metrics on FashionMNIST (**up**) and CIFAR-10 (**down**) in-distribution datasets. Results on LR, IC, and Likelihood Ratio are from Xiao et al. [2020]

$$\frac{1}{2L}\sum_{i=1}^{L}\sigma_i^2\mathbb{E}_{x\in D_{\text{train}}}\mathbb{E}_{\tilde{x}\sim\mathcal{N}(x,\sigma_i I)}\mathbb{E}_{r\in\mathcal{R}}\left[\left\|s_\theta\left(r\left(\tilde{x}\right),\sigma_i\right)+\frac{\tilde{x}-x}{\sigma_i^2}\right\|_2^2\right] \tag{4}$$

The score function of RotNCSN outputs the original noise matrix instead of the rotated noise matrix. Therefore, RotNCSN naturally learns to discriminate in-distribution data from the rotated data. Moreover, we expect RotNCSN to be more shape-aware since rotation-based self-supervised learning methods show efficacy in various downstream tasks Gidaris et al. [2018], Hendrycks et al. [2019].

When new sample x is given, RotNCSN outputs the normality vector $f_{\text{rot}}(x)$ as follows.

$$f_{\text{rot}}(x)_i = \mathbb{E}_{\tilde{x}\sim\mathcal{N}(x,\sigma_i I)}\sum_{r\in\mathcal{R}}\left\|s_\theta\left(r\left(\tilde{x}\right),\sigma_i\right)+\frac{\tilde{x}-x}{\sigma_i^2}\right\| \tag{5}$$

RotNCSN then trains a unsupervised model (e.g. GMM) over the extracted normality vector from the training data. This is consistent to the evaluation of Mahmood et al. [2021].

4 Discussion

We first evaluate the performance of RotNCSN in the following OOD detection tasks. We refer training of the RotNCSN in the Appendix.

FashionMNIST: we evaluate RotNCSN trained in FashionMNIST Xiao et al. [2017] dataset against the various OOD dataset; EMNIST, MNIST LeCun et al. [2010], NotMNIST, and KMNIST Clanuwat et al. [2018].

CIFAR-10: we evaluate RotNCSN trained in the CIFAR-10 dataset against the challenging CIFAR-100 dataset. Since the two datasets show similar textures, we expect the conventional normality vector to struggle in this task.

We sample $\tilde{x}$ once w.r.t each x while we compute $f_{\text{rot}}(x)$. For the baseline, we compare with conventional normality vector f_{msma} and f_{recon}. We directly use the trained score model of Mahmood et al. [2021] for evaluating the basline method. Since all score-based methods output the normality vector, we train GMM over the normality vector of the train data to output the explicit likelihood. Furthermore, as a competitive unsupervised OOD detection baseline, we compare the results of Xiao et al. [2020], Serrà et al. [2020], and ren et al. [2019] and refer to them as LR, IC, and Likelihood Ratio. For the result of LR and Likelihood Ratio, we refer to the result of Xiao et al. [2020] that use VAE.

We show the OOD detection result in Table 2. In the FashionMNIST domain, our proposed RotNCSN consistently improves over the conventional NCSN-based OOD detection metrics (MSMA, Recon). Furthermore, RotNCSN is competitive with various OOD detection methods. Finally, in the CIFAR-10 domain, RotNCSN improves over MSMA in the challenging CIFAR-100 detection task. We leave further analysis in the Appendix.

References

D. Baranchuk, A. Voynov, I. Rubachev, V. Khrulkov, and A. Babenko. Label-efficient semantic segmentation with diffusion models. In *ICLR*, 2022.

S. Choi and S.-Y. Chung. Novelty detection via blurring. In *ICLR*, 2020.

T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, and D. Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

G. Cohen, S. Afshar, J. Tapson, and A. van Schaik. Emnist: an extension of mnist to handwritten letters. *arXiv preprint arXiv:1702.05373*, 2017.

P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021.

R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased toward texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019.

S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018.

I. J. Goodfellow, J. Pouget-Abadie, B. X. Mehdi Mizra, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. In *NeurIPS*, 2014.

R. Gossweiler, M. Kamvar, and S. Baluja. What's up captcha: a captcha based on image orientation. In *WWW*, 2009.

J. D. Havtorn, J. Frellsen, S. Hauberg, and L. Maaløe. Hierarchical vaes know what they don't know. In *ICML*, 2021.

D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song. Using self-supervised learning can improve model robustness and uncertainty. In *NeurIPS*, 2019.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

E. Hoogeboom, A. A. Gritsenko, J. Bastings, B. Poole, R. van den Berg, and T. Salimans. Autoregressive diffusion models. In *ICLR*, 2022.

D. P. Kingma and P. Dhariwal. Glow: generative flow with invertible 1x1 convolutions. In *NeurIPS*, 2018.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

P. Kirichenko, P. Izmailov, and A. G. Wilson. Why normalizing flows fail to detect out-of-distribution data. In *NeurIPS*, 2020.

A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Technical Report*, 2009.

Y. LeCun, C. Cortes, and C. Burges. Mnist handwritten digit database. 2010.

A. Mahmood, J. Oliva, and M. Styner. Multiscale score matching for out-of-distribution detection. In *ICLR*, 2021.

C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon. Sdedit: guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022.

E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan. Do deep generative models know what they don't know? In *ICLR*, 2019.

M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016.

J. ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. A. DePristo, J. V. Dillon, and B. Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *NeurIPS*, 2019.

R. T. Schirrmeister, Y. Zhou, T. Ball, and D. Zhang. Understanding anomaly detection with deep invertible networks through hierarchies of distributions and features. In *NeurIPS*, 2020.

J. Serrà, D. Álvarez, V. Gómez, O. Slizovskaia, J. F. Núñez, and J. Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. In *ICLR*, 2020.

Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dastaset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Z. Xiao, Q. Yan, and Y. Amit. Likelihood regret: an out-of-distribution detection for variational autoencoder. In *NeurIPS*, 2020.

M. Xu, L. Yu, Y. Song, C. Shi, S. Ermon, and J. Tang. Geodiff: a geometric diffusion model for molecular conformation generation. In *ICLR*, 2022.

J. Yang, H. Wang, L. Feng, X. Yan, H. Zheng, W. Zhang, and Z. Liu. Semantically coherent out-of-distribution detection. In *ICCV*, 2021.

M. Zhang, A. Zhang, and S. McDonagh. On the out-of-distribution generalization of probabilistic image modelling. In *NeurIPS*, 2021.

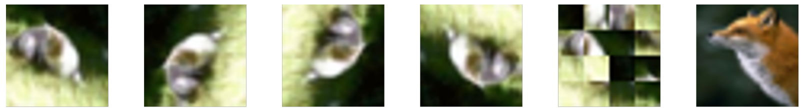

Figure 2: Examples of the tested OODs. (a): Original CIFAR-10 image. (b),(c),(d): rotated CIFAR-10 images. (e): patch-shuffled CIFAR-10 image. (f): CIFAR-100 images.

Method	rot90	rot180	rot270	Patch shuffle
f_{msma}	0.635	0.520	0.630	0.897
f_{rec}	0.640	0.527	0.634	0.865

Table 3: AUROC of previously proposed SGM-based OOD detection methods trained in the SVHN dataset tested in the proposed OOD dataset.

A Appendix

A.1 Further results on the motivation

We further provide the visualization and additional results that support our hypothesis. First, we provide a visualization of the OOD data used in Figure 2. We further test our motivation in the alternative SVHN dataset. We refer the result to Table 3. Similar to CIFAR-10, SGM trained in the SVHN dataset also struggles to detect rotated OODs although they are not in the training dataset. Nevertheless, SGM trained in the SVHN dataset is more robust to patch shuffling compared to the CIFAR-10 dataset.

A.2 Experiment settings

In training the RotNCSN, we do not change any training details (e.g noise scale) except the training epoch. We train RotNCSN for 400000 steps in the FashionMNIST dataset and 600000 steps in the CIFAR-10 dataset.

A.3 Further analysis

In this section, we further analyze the performance of RotNCSN compared to the NCSN. We visualize top-9 samples that RotNCSN and NCSN output the lowest likelihood on the FashionMNIST domain in Figure 3. While NCSN assigns higher uncertainty to relatively complex data, RotNCSN assigns higher uncertainty to data with a possible anomaly. For example, in the third row of the left of Figure 3, we observe cracks in the object.

We further analyze the effect of each noise level for OOD detection. While our method originates on the multi-scale vector, we extract the score of each noise level and use them independently for OOD detection. We simply use the distance from the mean of the training dataset's score as the detection metric. Instead of AUROC, we plot TNR at 95% TPR for the distinct visualization.

We report the result in Figure 4. Since we did not train over GMM nor use multi-scale score matching, the performance is less than the reported one in Table 2. The OOD detection performance increases while the noise level increases generally. However, the performance diminishes after 0.359. One hypothesis to explain this behavior is that reconstructing the original image at a high noise level may be an ill-posed problem and therefore become unsuitable for OOD detection.

Figure 3: Top-9 samples from the Fashion-MNIST dataset where RotNCSN (**left**) and NCSN (**right**) assign highest uncertainty.

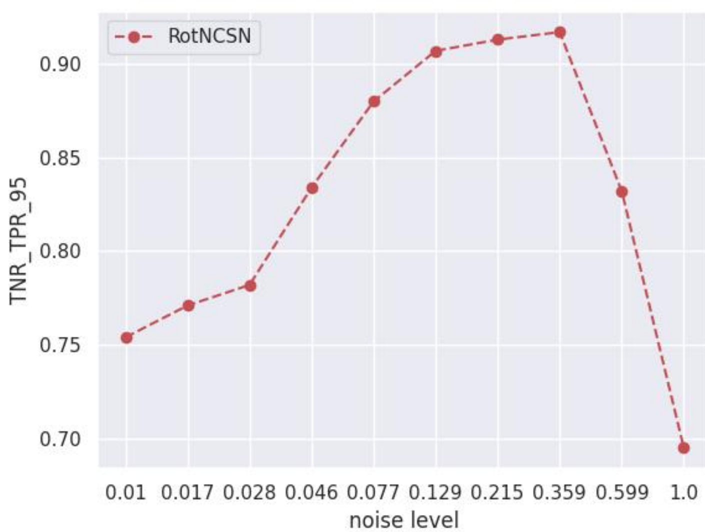

Figure 4: True negative rate (TNR) at 95% True positive rate (TPR) performance of RotNCSN w.r.t noise-level.