Do Diffusion Models Know What They Don't Know in Representation Space?

Yifan Ding, Arturas Aleksandrauskas, Amirhossein Ahmadian,
Jonas Unger, Fredrik Lindsten and Gabriel Eilertsen
Linköping University, Norrköping, Sweden
{firstname.lastname}@liu.se

Introduction

Likelihood-based Deep Generative Models (DGMs) are commonly used for Out-Of-Distribution (OOD) detection, but often struggle, frequently assigning higher likelihoods to OOD images than to In-Distribution (ID) data [2]. We introduce a practical OOD detection method using diffusion models applied to image representations. Our approach matches state-of-the-art performance on large-scale benchmarks and show that conditional training using logits from supervised encoder further enhances detection. Additionally, we provide empirical evidence that the entropy of ID representations, estimated by the diffusion model, correlates with OOD detection performance, guiding the selection of encoders.

Methods

OOD Detection without Labels

Access to annotated ID data points is not possible in many cases (the dashed line case in Fig. 1). We can train a density model $p_\theta(\mathbf{x})$ to approximate the true distribution of the training inputs $p(\mathbf{x})$, given only $x \sim \mathcal{X}_{\text{ID}}$. Any x that has a sufficiently low density under $p_\theta(\mathbf{x})$ is assigned to the $\mathcal{D}_{\text{OOD}}$ space if $p_\theta(\mathbf{x})$ is a reasonably good estimation of $p(\mathbf{x})$. However, $p_\theta(\mathbf{x})$ estimated on image space for OOD detection has been observed unreliable [2].

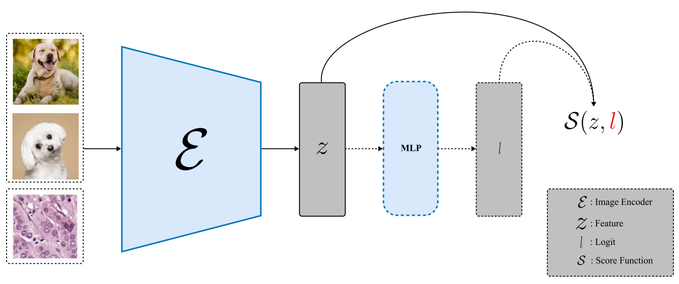

Figure 1: OOD detection distance function design.

Representation Diffusion Model

Given the image space dataset $\{\mathbf{x}_i\}_{i=0}^N$, representations $\{\mathbf{z}_i\}_{i=0}^N$ are calculated by a pre-trained encoder

$$\mathbf{z} = \mathcal{E}(\mathbf{x}).$$

Training score-based diffusion models can be formulated as reverse-time SDE learning, and the corresponding probability flow ODE of such SDE, can be derived as

$$d\mathbf{z} = \{\mathbf{f}(\mathbf{z}, t) - \frac{1}{2} g(t) g(t)^T \nabla_{\mathbf{z}} \log p_t(\mathbf{z})\} dt,$$

With the instantaneous change of variables formula [1], we can compute the precise representation likelihood $p(\mathbf{z}(0))$ using

$$\log p(\mathbf{z}(0)) = \log p(\mathbf{z}(1)) + \int_0^1 \nabla \cdot f_\theta(\mathbf{z}(t), t) dt.$$

Our method can be enhanced with logit from supervised encoder in conditional training, we name our method RDM and ConRDM.

Entropy and Encoder

Given likelihood estimation $p(\mathbf{z})$ from the diffusion model, entropy on ID representation can be approximated as

$$H(\mathbf{z}) = -\int_{\mathcal{Z}} \log p(\mathbf{z}) p(\mathbf{z}) d\mathbf{z},$$

entropy estimated on different representation are correlated with OOD detection performance.

Results

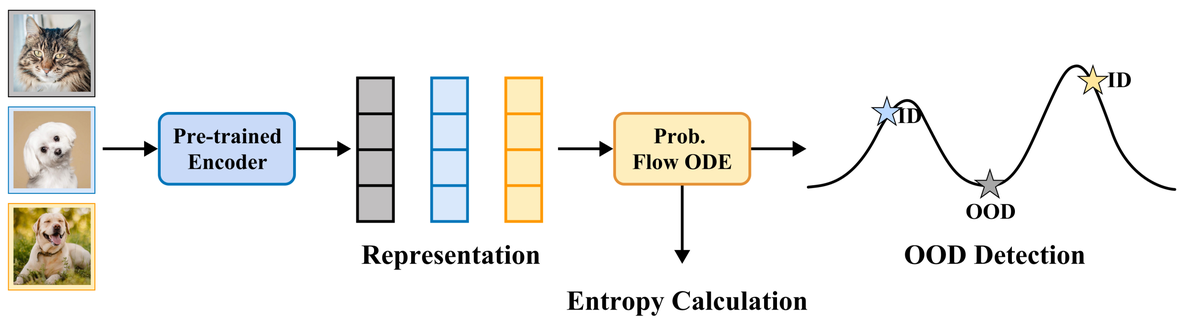

Figure 2: Our method can be used for both OOD detection and evaluate the representation space for OOD detection using information theory.

OOD Method	OpenImage-O		Textures		iNaturalist		ImageNet-O		Average	
	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
MAE										
KNN	**60.54**	**89.03**	89.04	41.51	**48.02**	**97.69**	68.64	81.20	**66.56**	77.36
Residual w/o offset	59.52	89.22	**90.33**	**38.90**	42.47	98.87	**69.60**	**79.85**	65.48	**76.71**
RDM	58.15	91.50	89.06	43.80	41.44	99.20	66.40	86.25	63.76	80.19
DINO										
KNN	85.26	65.25	94.15	25.39	88.30	67.62	81.55	74.70	87.31	58.23
Residual w/o offset	**87.57**	**54.77**	**97.84**	**11.10**	**92.71**	**42.76**	**81.98**	**68.40**	**90.02**	**44.25**
RDM	85.68	64.73	96.59	17.17	86.67	70.98	79.80	73.90	87.18	56.69
DINOv2										
KNN	**95.05**	**25.66**	91.65	35.33	99.06	3.47	**86.67**	**57.55**	93.10	**30.50**
Residual w/o offset	92.61	35.53	**93.60**	33.41	**99.32**	**1.74**	83.23	70.40	92.19	35.26
RDM	94.06	31.07	93.32	**32.5**	99.30	1.83	85.97	63.30	**93.16**	32.17

Table 1: OOD detection with self-supervised encoder: AUROC and FPR95 are reported as percentages. Results for MAE, DINO and DINOv2 with ImageNet-1K as ID data and four OOD datasets: OpenImage-O, Textures, iNaturalist, and ImageNet-O. Since logits are not available, we only compare with KNN and Residual. The best method is marked in bold.

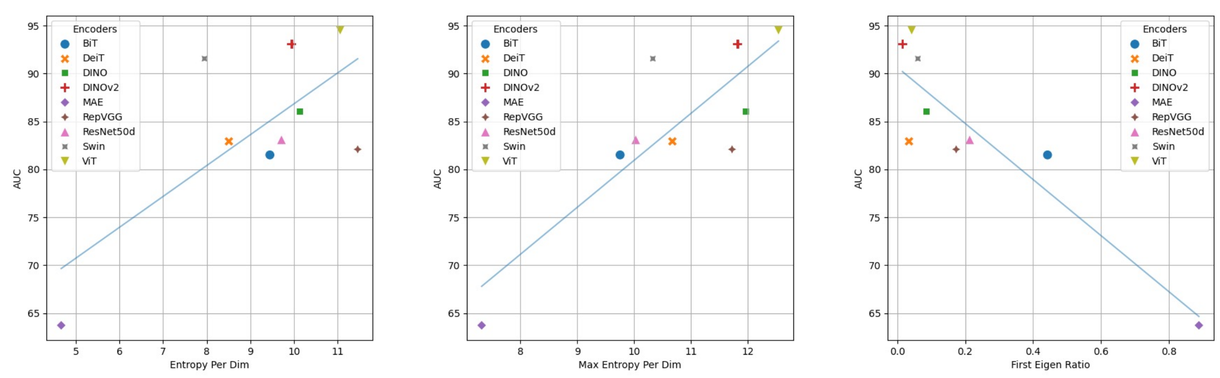

Figure 3: From left to right, entropy per dimension, maximum entropy per dimension, and first eigenvalue rate. These are calculated on the ID dataset to compare with OOD performance (AUC).

Contribution

- A simple, yet powerful likelihood-based OOD detection method using score-based diffusion models operating in the representation space.
- To our knowledge, it is the first likelihood-based OOD detection method with performance on par with state-of-the-art (SOTA) methods on large-scale benchmarks.
- An in-depth empirical analysis of different representations, providing important insights for selecting appropriate encoders for OOD detection.

Acknowledgment

This work was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

[1] Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. *Advances in neural information processing systems 31* (2018).

[2] Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., and Lakshminarayanan, B. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136* (2018).

LINKÖPING UNIVERSITY

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM