Neural Networks 145 (2022) 199–208

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Detecting out-of-distribution samples via variational auto-encoder with reliable uncertainty estimation

Xuming Ran [a,e,*,1], Mingkun Xu [b,1], Lingrui Mei [c], Qi Xu [d,f], Quanying Liu [a,*]

[a] Shenzhen Key Laboratory of Smart Healthcare Engineering, Department of Biomedical Engineering, Southern University of Science and Technology, Shenzhen 518055, China
[b] Center for Brain Inspired Computing Research, Department of Precision Instrument, Tsinghua University, Beijing 100084, China
[c] China Automotive Engineering Research Institute, Chongqing 401122, China
[d] School of Artifical Intelligence, Electronic and Electrical Engineering, School of Artifical Intelligence Dalian University of Technology, Dalian 116024, China
[e] College of Mathematics and Statistics, Chongqing Jiaotong University, Chongqing 400074, China
[f] College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Article history:
Received 17 July 2020
Received in revised form 28 August 2021
Accepted 22 October 2021
Available online 28 October 2021

Keywords:
Variational auto-encoder
Out-of-distribution detection
Uncertainty estimation
Noise contrastive prior

ABSTRACT

Variational autoencoders (VAEs) are influential generative models with rich representation capabilities from the deep neural network architecture and Bayesian method. However, VAE models have a weakness that assign a higher likelihood to out-of-distribution (OOD) inputs than in-distribution (ID) inputs. To address this problem, a reliable uncertainty estimation is considered to be critical for in-depth understanding of OOD inputs. In this study, we propose an improved noise contrastive prior (INCP) to be able to integrate into the encoder of VAEs, called INCPVAE. INCP is scalable, trainable and compatible with VAEs, and it also adopts the merits from the INCP for uncertainty estimation. Experiments on various datasets demonstrate that compared to the standard VAEs, our model is superior in uncertainty estimation for the OOD data and is robust in anomaly detection tasks. The INCPVAE model obtains reliable uncertainty estimation for OOD inputs and solves the OOD problem in VAE models.

1. Introduction

The out-of-distribution (OOD) data has a significantly different distribution from the training in-distribution (ID) data. To make reliable and safe decisions, the deep learning models in real-world applications require to identify whether the testing data is the OOD data. Likelihood models are considered to naturally own the ideal capability of detecting OOD inputs, due to the intuitive assumption that these models assign lower likelihoods to the OOD inputs than the in-distribution (ID) inputs (Bishop, 1994). However, previous works have reported that some deep generative models, such as variational auto-encoders (VAEs) (Kingma & Welling, 2014; Rezende, Mohamed, & Wierstra, 2014), Pixel CNN (Van den Oord, Kalchbrenner, Espeholt, Vinyals, Graves, et al., 2016) and Glow (Kingma & Dhariwal, 2018), all based on likelihood models, are not able to correctly detect OOD inputs (Choi, Jang, & Alemi, 2018; Hendrycks, Mazeika, & Dietterich, 2019; Lee, Lee, Lee, & Shin, 2018b; Maaløe, Fraccaro, Liévin, & Winther, 2019; Nalisnick, Matsukawa, Teh, Gorur, & Lakshminarayanan, 2019a; Nalisnick, Matsukawa, Teh, & Lakshminarayanan, 2019b). Counter-intuitively, the OOD inputs are assigned higher likelihoods than the ID inputs, which is not in line with the assumption. Hence, when we employ the likelihood model as a detector on OOD detection tasks or general generation tasks, it is necessary to ensure that the adopted model possesses a good understanding and performance for OOD inputs.

The phenomenon that VAE models assign higher likelihoods to OOD inputs than ID inputs is called the OOD problem, and it was first reported by Nalisnick et al. (2019a) in 2018. Since then, it has been an increasingly popular topic in the field of generative models. Some studies have made great efforts to explain the reasons for this empirical phenomenon (Bütepage, Poklukar, & Kragic, 2019; Nalisnick et al., 2019b; Serrà et al., 2020). For instance, Bütepage et al. demonstrate that it is caused by model assumptions and evaluation schemes, where the oversimplified likelihood function (e.g., iid Bernoulli or iid Gaussian) assumed in the VAE model affects the judgment of the data distribution of the ID inputs (Bütepage et al., 2019). However, the true likelihood function is often unknown and more complicated, which has certain deviations from the assumed one. In some datasets, local

* Corresponding authors.
E-mail addresses: ranxuming@gmail.com (X. Ran), liuqy@sustech.edu.cn (Q. Liu).
[1] Equal contribution.

https://doi.org/10.1016/j.neunet.2021.10.020
0893-6080/

evaluations with the approximated posterior can lead to overconfidence. Nalisnick et al. conjecture that the high-likelihood region conflicts with the typical set of the model (Nalisnick et al., 2019b). Serrà et al. posit that the complexity of the input data will have a strong impact on likelihood-based models (Serrà et al., 2020).

Many approaches have been studied to solve the OOD detection problem in generative models. Some studies have suggested that likelihood models with reliable uncertainty estimates may help improve OOD detection (Choi et al., 2018; Nalisnick et al., 2019a). In addition, noise contrastive priors (NCPs) are a specific prior in the data space for neural networks, encouraging network weights to not only explain the ID inputs, but also capture the high uncertainty of OOD samples (Hafner, Tran, Irpan, Lillicrap, & Davidson, 2018). Thus, NCPs might help the uncertainty estimates of the OOD data. Inspired by these two viewpoints, we propose a novel method, named Improved Noise Contrastive Priors Variational Auto-encoder (INCPVAE), to allow VAE models to obtain reliable uncertainty estimates thereby solving the OOD detection problem. Although the original NCPs are often applied to classifier models, they cannot be directly applied to the VAE framework. Therefore, we have to improve the loss function of NCP (called the improved NCP, INCP) to make it suitable for the VAE framework. The INCP is integrated into the encoder of VAE, so that OOD samples can be generated by adding Gaussian noise to the origin ID inputs. Since using the simple likelihood function of VAE often leads to poor performance on OOD detection tasks, we exploit the INCP-KL divergence of INCPVAE, rather than the likelihood, for detecting OOD inputs. Our experiments show that compared to the traditional VAEs, our INCPVAE can reduce the overconfidence when facing OOD data and obtain better performances of OOD detection. The main contributions of this paper are as follows:

- We propose an improved noise contrastive prior to fit the VAE framework (Section 3.3). To the best of our knowledge, this is the first work to use the noise contrastive prior to obtain reliable uncertainty estimates in unsupervised generative models.
- We present a tailored metric (the ELBO Ratio) in the INCP-VAE framework to estimate the uncertainty (Section 3.4), which can achieve reliable uncertainty estimation and enhanced robustness (Section 4.2).
- We propose a novel OOD detection method by using the INCP-KL ratio of INCPVAE (Section 3.5). Through a number of experiments on the challenging OOD cases, we demonstrate that INCPVAE can learn the true characterization of OOD inputs, and achieves state-of-the-art (SOTA) performance in OOD detection (Section 4.3).

2. Related work

OOD detection: There are many neural network tools that can be used to perform pattern recognition, image classification, and OOD detection tasks, such as spike neural networks (Liu, Pan, Ruan, Xing, Xu, & Tang, 2020; Maciag, Kryszkiewicz, Bembenik, Lobo, & Ser, 2021; Xu, Qi, Yu, Shen, Tang, & Pan, 2018) and convolutional neural networks (Lee, Lee, Lee, & Shin, 2018a; Xu, Zhang, Gu, & Pan, 2019). The OOD detection permits a system to reject a novel input rather than assigning it an incorrect label; therefore the ability to detect OOD data is essential for machine learning models. From the algorithm perspective, there are two categories of mainstream approaches for OOD detection, (i) the supervised/discriminative approaches and (ii) the unsupervised/generative approaches (Daxberger & Hernández-Lobato, 2019). Most existing methods belong to the supervised model. For example, the classifiers are trained by both the OOD data and ID data to learn a decision boundary between ID and OOD inputs, which can be used for OOD detection. Liang et al. present an OOD detector with neural networks (called ODIN) which uses softmax function to maximize the difference between likelihoods of ID data and OOD data, while the model parameters are tailored to each OOD source (Liang, Li, & Srikant, 2018). Lakshminarayanan et al. propose an ensemble method for OOD detection, which independently trains multiple models with random initializations of network parameters and randomly shuffled training inputs (Lakshminarayanan, Pritzel, & Blundell, 2017). Some previous studies show that these supervised methods can to some extent prevent the poorly-calibrated neural networks from incorrectly high-confidence on OOD inputs (DeVries & Taylor, 2018; Lakshminarayanan et al., 2017; Liang et al., 2018). This capability can be used in various applications, including anomaly detection (Hendrycks & Gimpel, 2017; Pidhorskyi, Almohsen, Adjeroh, & Doretto, 2018; Vyas et al., 2018) and adversarial defense (Song, Shu, Kushman, & Ermon, 2018). However, these methods can only be applied to task-dependent scenarios. This is a severe limitation, for the anomalous data in real-world applications rarely knows in advance.

In contrast, the unsupervised approaches aim to solve the OOD detection problem by training deep generative models in a more general manner, among which density estimation is widely applied (Kingma & Dhariwal, 2018; Oord, Kalchbrenner, & Kavukcuoglu, 2016). For example, Choi et al. use generative model with Watanabe–Akaike information criterion (WAIC) for detecting OOD (Choi et al., 2018). Although this work performs well in practice, it does not explicitly solve the problem of typicality (Choi et al., 2018; Nalisnick et al., 2019b). Denouden et al. propose a method that incorporates both reconstruction loss and the Mahalanobis distance (Lee et al., 2018a) in the latent space as an OOD detection score (Denouden et al., 2018). Ren et al. propose a likelihood ratio method for deep generative models to detect the OOD data (Ren et al., 2019). Zhang et al. studied the intrinsic robustness of typical image distributions by using conditional generative models (Zhang, Chen, Gu, & Evans, 2020). They proved a fundamental bound on the intrinsic robustness, that is, the underlying data distribution can be captured by a conditional generative adversarial network. However, as mentioned, the likelihood estimation in deep generative models are not reliable for OOD detection. Many studies have attempted to explain the reasons and seek the solutions (Bütepage et al., 2019; Nalisnick et al., 2019b; Serrà et al., 2020). So far, an efficient and robust solution for OOD detection is still missing and urgently needed.

Uncertainty estimation: Uncertainty estimation is highly associated with OOD detection. The goal of uncertainty estimation is to generate a calibrated confidence measure for the predicted distribution which can be used in the OOD detection. The uncertainty estimation in MC Dropout (Gal & Ghahramani, 2016), Deep-Ensemble (Lakshminarayanan et al., 2017) and ODIN (Liang et al., 2018) involves presenting a calibrated predictive distribution by classifiers. Alternatively, variational information bottleneck (VIB) conducts OOD detection via divergence estimation in latent space (Alemi, Fischer, & Dillon, 2018). However, these existing methods are model-dependent and rely heavily on task-specific information to obtain a comprehensive estimate of uncertainty. Therefore, a more general and task-independent method is of high needs.

Recent studies have suggested that likelihood models with reliable uncertainty estimation can help to mitigate the high OOD likelihood problem for generative models in a task-independent manner (Choi et al., 2018; Nalisnick et al., 2019a). For example, Meronen et al. studied the influence of neural network activation functions and the Matérn family of kernels on the uncertainty estimation (Meronen, Irwanto, & Solin, 2020). Moreover, as an influential and generally-used class of likelihood-based generative models in unsupervised learning, VAEs may be a good OOD

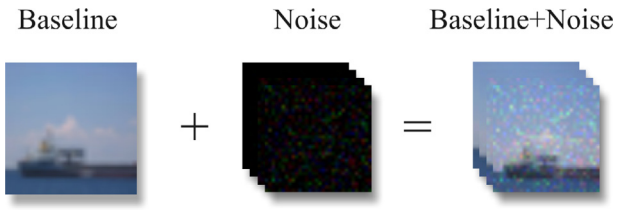


Fig. 1. Generating OOD samples by adding Gaussian Noise to the baseline data. The baseline data is sample from the original image dataset (e.g., FashionMNIST, MNIST, CIFAR10, SVHN). We add the Gaussian Noise at three levels to generate the OOD sample with different complexity. The Baseline + Noise is the generated OOD sample.

detector. It assumes that the model assigns higher likelihoods to the samples from the ID data than the OOD data. NCPs can inject variability or insensitivity into the model especially into regions that do not exhibit that otherwise after training. In this sense, NCPs can be considered as a part of model specification to get better estimation of uncertainty and therefore help model inference. In this study, we provide a novel hybrid framework that bridges NCPs with VAEs and generates OOD data by adding Gaussian noise, to help both the reliability of uncertainty estimation and model independence in OOD detection.

3. Method

3.1. Improved noise contrastive priors

NCPs has been proposed to obtain reliable uncertainty estimates by employing an input prior to the ID inputs $\boldsymbol{x}$ and OOD inputs $\tilde{\boldsymbol{x}}$ and an output prior which is a wide distribution given these inputs (Hafner et al., 2018). However, NCPs are not suitable for VAE framework. In this work, we modify the loss function to make the original NCPs fit the VAE framework, to obtain uncertainty through the VAE model. We add Gaussian noise to ID images to generate OOD data.

Generating OOD Inputs: OOD samples can be generated by sampling from the distribution boundary of the ID data with high uncertainty (Lee et al., 2018b). Inspired by noise contrastive estimation (Gutmann & Hyvärinen, 2010; Mnih & Kavukcuoglu, 2013), Hafner et al. (2018) proposed a NCP-based algorithm, where a complement distribution is approximated by random noise. To obtain OOD inputs $\tilde{\boldsymbol{x}}$, we add Gaussian noise $\boldsymbol{\epsilon}$ into the continuous ID inputs $\boldsymbol{x}$, formulated as $\tilde{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\epsilon}$ (See Fig. 1). The marginal distribution of OOD inputs $p_o(\tilde{\boldsymbol{x}})$ is derived in Eq. (1) as follows:

$$p_o(\tilde{\boldsymbol{x}}) = \int_{\boldsymbol{x}} p_i(\boldsymbol{x}) \mathcal{N}\left(\tilde{\boldsymbol{x}} - \boldsymbol{x} \mid \mu, \sigma^2 \boldsymbol{I}\right) d\boldsymbol{x}, \tag{1}$$

where $p_i(\boldsymbol{x})$ denotes the distribution density of ID inputs; μ and σ^2 are the mean and variance of Gaussian noise, respectively. In order to make the noise contrastive prior homogeneous in all directions of the data manifold, we set $\mu = 0$. The variance σ^2 is a hyperparameter to tune the sampling distance from the boundary of the training ID distribution. The higher the variance σ^2, the higher the complexity of OOD inputs.

Data Priors: The data priors consist of an input prior $p(\boldsymbol{x})$ and an output prior $p(\boldsymbol{z}|\boldsymbol{x})$. To obtain a reliable uncertainty estimation by the VAE model, appropriate input priors (including a prior on OOD inputs) should be set. A good output prior should be a high-entropy distribution, which serves as the high uncertainty of the VAE's target output for a given OOD input. The data priors in our model are listed as follows:

$$\begin{aligned} &\text{OOD input prior: } \tilde{p}(\tilde{\boldsymbol{x}}) = p_o(\tilde{\boldsymbol{x}}) \\ &\text{OOD output prior: } \tilde{p}(\tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{x}}) = \mathcal{N}\left(\tilde{\boldsymbol{z}} \mid \mu_{\tilde{\boldsymbol{x}}}, \sigma^2_{\tilde{\boldsymbol{x}}} \boldsymbol{I}\right), \end{aligned} \tag{2}$$

where $p_o(\tilde{\boldsymbol{x}})$ is the prior distribution of OOD inputs; $\mu_{\tilde{\boldsymbol{x}}}$ and $\sigma^2_{\tilde{\boldsymbol{x}}}$ are the hyperparameters of OOD output priors to tune the mean and the uncertainty in the target outputs.

Loss Function: KL divergence is not symmetric, and it has a forward version and a reverse version (Zhang, Bird, Habib, Xu, & Barber, 2019). In the original NCPs (Hafner et al., 2018), both the difference metrics between the distribution $p(\boldsymbol{z} \mid \boldsymbol{x})$ and $q_\theta(\boldsymbol{z} \mid \boldsymbol{x})$ and between $\tilde{p}(\tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{x}})$ and $q_\theta(\tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{x}})$ adopt the forward KL divergence. However, the VAE uses the reverse KL divergence as its basic metric in loss function, which leads to the inconsistency in optimization strategy and direction. Therefore, this poses an intractable challenge for constructing a unified optimization framework by incorporating NCP organically, where the forward KL divergence is not compatible for VAE. To better tackle the challenge and incorporate the NCP into the VAE framework, we proposed the improved NCP (INCP) method by integrating the reverse KL divergence into the NCP. To train INCPs, we modify the loss function as follows:

$$\begin{aligned} \mathcal{L}(\theta) = &\boldsymbol{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x})}\left[\boldsymbol{D}_{KL}\left[q_\theta(\boldsymbol{z} \mid \boldsymbol{x}) \,||\, p(\boldsymbol{z} \mid \boldsymbol{x})\right]\right] \\ &+ \gamma \boldsymbol{E}_{q_\theta(\tilde{\boldsymbol{z}}|\tilde{\boldsymbol{x}})}\left[\boldsymbol{D}_{KL}\left[q_\theta(\tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{x}}) \,||\, \tilde{p}(\tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{x}})\right]\right], \end{aligned} \tag{3}$$

where $\tilde{p}(\tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{x}})$ denotes OOD data priors, θ is the parameter of neural network. A hyper-parameter γ denotes the trade-off between the ID and OOD output priors. INCPs can be trained by minimizing this loss. Notice that in Eq. (3), by minimizing the reverse KL divergence in the first term, the neural network is trained to suit for the true ID data outputs prior. And an analogous term on the OOD data outputs prior is added in the second term. This loss function simultaneously optimizes the ID and OOD outputs prior for two distinct targets (*i.e.*, the true ID data outputs prior & the assumed OOD data outputs prior). In contrast, the origin NCP loss (Hafner et al., 2018) hardly integrates the ID and OOD conditional distribution into one target in the VAE framework.

3.2. Variational autoencoder

VAEs (Kingma & Welling, 2014; Rezende et al., 2014) are a class of latent variable models optimized by the maximum marginal likelihood of an observation variable. The marginal likelihood $p(\boldsymbol{x})$ can be written as follows:

$$\begin{aligned} \log p(\boldsymbol{x}) = &\boldsymbol{E}_{\boldsymbol{z}\sim q_\theta(\boldsymbol{z}|\boldsymbol{x})}[\log p_\phi(\boldsymbol{x} \mid \boldsymbol{z})] - \boldsymbol{D}_{KL}[q_\theta(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p(\boldsymbol{z})] \\ &+ \boldsymbol{D}_{KL}[q_\theta(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p(\boldsymbol{z} \mid \boldsymbol{x})], \end{aligned} \tag{4}$$

where $p(\boldsymbol{z})$ and $p(\boldsymbol{z} \mid \boldsymbol{x})$ are the ID input/output priors (e.g., Vamp Prior Tomczak & Welling, 2018, Resampled Prior Bauer & Mnih, 2019). In this study, $p(\boldsymbol{z})$ is instantiated by a standard normal distribution, and $p(\boldsymbol{z} \mid \boldsymbol{x})$ is the true posterior distribution corresponding to $p(\boldsymbol{z})$. The encoder $q_\theta(\boldsymbol{z} \mid \boldsymbol{x})$ and the decoder $p_\phi(\boldsymbol{x} \mid \boldsymbol{z})$ are modeled by two neural networks parameterized with θ, ϕ, respectively. Specifically, $q_\theta(\boldsymbol{z} \mid \boldsymbol{x})$ represents the variational posterior (the encoder) which is implemented by a Gaussian distribution, and $p_\phi(\boldsymbol{x} \mid \boldsymbol{z})$ is the generative model (the decoder) which is implemented by a Bernoulli distribution.

However, the true posterior $p(\boldsymbol{z} \mid \boldsymbol{x})$ cannot be computed analytically. Assuming that the variational posterior $q_\theta(\boldsymbol{z} \mid \boldsymbol{x})$ has a arbitrarily high-capacity for modeling, $q_\theta(\boldsymbol{z} \mid \boldsymbol{x})$ can learn to approximate the intractable $p(\boldsymbol{z} \mid \boldsymbol{x})$ and the reverse KL divergence between $q_\theta(\boldsymbol{z} \mid \boldsymbol{x})$ and $p(\boldsymbol{z} \mid \boldsymbol{x})$ goes to zero. Thus, we train the VAE with ID samples, or OOD samples, to maximize the following objective variational evidence lower bound, which are called ELBO_I for ID samples, and ELBO_O for OOD samples.

$$\begin{aligned} \text{ELBO}_I(\phi, \theta) &= \boldsymbol{E}_{\boldsymbol{z}\sim q_\theta(\boldsymbol{z}|\boldsymbol{x})}[\log p_\phi(\boldsymbol{x} \mid \boldsymbol{z})] - \boldsymbol{D}_{KL}[q_\theta(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p(\boldsymbol{z})] \\ \text{ELBO}_O(\phi, \theta) &= \boldsymbol{E}_{\tilde{\boldsymbol{z}}\sim q_\theta(\tilde{\boldsymbol{z}}|\tilde{\boldsymbol{x}})}[\log p_\phi(\tilde{\boldsymbol{x}} \mid \tilde{\boldsymbol{z}})] - \boldsymbol{D}_{KL}[q_\theta(\tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{x}}) \parallel \tilde{p}(\tilde{\boldsymbol{z}})] \end{aligned} \tag{5}$$

where $q_\theta(\boldsymbol{z} \mid \boldsymbol{x})$ and $q_\theta(\tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{x}})$ are the variational posteriors which approximate the true posteriors (*i.e.,* $p(\boldsymbol{z} \mid \boldsymbol{x})$ and $\tilde{p}(\tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{x}})$), given the ID input $\tilde{\boldsymbol{x}}$ and the OOD input $\boldsymbol{x}$, respectively. For a given dataset, the marginal likelihood $p(\boldsymbol{x})$ is a constant. Substituting Eq. (5) to Eq. (4), we obtain

$$\log p(\boldsymbol{x}) = \text{ELBO}_I(\phi, \theta) + \boldsymbol{D}_{KL}[q_\theta(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p(\boldsymbol{z} \mid \boldsymbol{x})] = \text{const.} \tag{6}$$

From Eq. (6), it is obvious that maximizing ELBO_I is equivalent to minimizing the KL-divergence between $q_\theta(\boldsymbol{z} \mid \boldsymbol{x})$ and $p(\boldsymbol{z} \mid \boldsymbol{x})$. Likewise, maximizing ELBO_O is equivalent to minimizing the reverse KL divergence between $q_\theta(\tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{x}})$ and $\tilde{p}_\theta(\tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{x}})$.

3.3. INCP variational autoencoder

INCPVAE consists of an encoder and a decoder, and the INCPs are imposed on the encoder network of VAE. The INCPVAE is trained on both ID and OOD inputs by minimizing ELBO_I and ELBO_O as shown in Eq. (5). We define the total ELBO of INCPVAE, $\text{ELBO}_{INCP}(\phi, \theta)$, as follows,

$$\text{ELBO}_{INCP}(\phi, \theta) = \text{ELBO}_I(\phi, \theta) + \gamma \text{ELBO}_O(\phi, \theta), \tag{7}$$

where the hyper-parameter γ is a setting as a trade-off between ELBO_I and ELBO_O.

We assume the variational posterior $q_\theta(\boldsymbol{z} \mid \boldsymbol{x})$ for ID inputs has high-capacity for modeling, then true posterior $p(\boldsymbol{z} \mid \boldsymbol{x})$ can be approximated by $q_\theta(\boldsymbol{z} \mid \boldsymbol{x})$. Since the OOD outputs prior $\tilde{p}(\tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{x}})$ is defined in Eq. (2), the true OOD data posterior $\tilde{p}(\tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{x}})$ is:

$$\tilde{p}(\tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{x}}) = \mathcal{N}\left(\tilde{\boldsymbol{z}} \mid \mu_{\tilde{\boldsymbol{x}}}, \sigma_{\tilde{\boldsymbol{x}}}^2 \boldsymbol{I}\right), \tag{8}$$

where $\mu_{\tilde{\boldsymbol{x}}} = \mu_{\boldsymbol{x}}$ and $(\mu_{\boldsymbol{x}} \sim q_\theta(\boldsymbol{z} \mid \boldsymbol{x}))$; $\sigma_{\tilde{\boldsymbol{x}}}^2$ is a hyper-parameter to tune the uncertainty in the outputs. The higher $\sigma_{\tilde{\boldsymbol{x}}}^2$, the higher the output uncertainty. The reverse KL divergence between $q_\theta(\tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{x}})$ and $\tilde{p}(\tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{x}})$ (called INCP-KL) becomes tractable and can be analytically computed. From Eq. (7) , maximizing the ELBO of INCPVAE can be replaced by minimizing the following loss function:

$$\mathcal{L}_{INCPVAE}(\phi, \theta) = -\text{ELBO}_I(\phi, \theta) + \gamma \underbrace{\boldsymbol{D}_{KL}[q_\theta(\tilde{\boldsymbol{z}}|\tilde{\boldsymbol{x}}) \parallel \tilde{p}(\tilde{\boldsymbol{z}}|\tilde{\boldsymbol{x}})]}_{\textbf{INCP}-\textbf{KL} \quad \textbf{Loss}} \tag{9}$$

Notably, the first term in Eq. (9) minimizes the negative ELBO_I, which is equivalent to maximizing ELBO_I. The second term in Eq. (9) minimizes INCP-KL for OOD data, which is equivalent to maximizing ELBO_O, according to Eq. (6). In this study, we set the hyperparameter $\gamma = 1$.

3.4. Metrics for uncertainty estimation: ELBO ratio

We proposed the objective variational evidence lower bound ratio (ELBO Ratio) for an uncertainty estimation metric of VAE. According to Eq. (5), we compute the ELBO of each ID sample and find the maximum one (called $\textbf{ELBO}_I(\boldsymbol{x}_{max})$). The ELBO Ratio for input data $\boldsymbol{x}_0$, $\mathcal{U}(\boldsymbol{x}_0)$, is defined as

$$\mathcal{U}(\boldsymbol{x}_0) = \frac{\textbf{ELBO}(\boldsymbol{x}_0)}{\textbf{ELBO}_I(\boldsymbol{x}_{max})}, \tag{10}$$

The ELBO ratio $\mathcal{U}(\boldsymbol{x}_0)$ measures the degree of uncertainty on data $\boldsymbol{x}_0$. The greater $\mathcal{U}(\boldsymbol{x}_0)$, the higher uncertainty $\boldsymbol{x}_0$.

3.5. OOD detection based on INCP-KL ratio

INCP-KL ratio: The likelihood of VAE has been used for OOD detection. However, it is reported that the OOD inputs have a higher likelihood than ID inputs that occur in some datasets (e.g., FashionMNIST vs MNIST, CIFAR10 vs SVHN). To solve this problem, Likelihood Ratios for OOD detection has been proposed (Ren et al., 2019). In Eq. (9), the second term of the INCPVAE loss is the reverse KL divergence between the OOD variational posterior ($q_\theta(\tilde{\boldsymbol{z}}|\tilde{\boldsymbol{x}})$) and the true OOD posterior ($\tilde{p}(\tilde{\boldsymbol{z}}|\tilde{\boldsymbol{x}})$), which is called INCP-KL. We find that INCP-KL of the OOD test samples (e.g., Baseline + Noise, in Fig. 1) is smaller than the ID samples from other distribution. Inspired by it, we use the INCP-KL Ratio for OOD detection. We calculate the INCP-KL divergence for all OOD training samples $\tilde{x}$ in the OOD dataset, and then find the OOD sample with maximum INCP-KL (called OOD_{max}). The INCP-KL Ratio for input data $\boldsymbol{x}_0$, $KLR(\boldsymbol{x}_0)$, is defined as

$$\begin{aligned} OOD_{max} &= \underset{\tilde{x}}{argmax}\, \boldsymbol{D}_{KL}[q_\theta(\tilde{\boldsymbol{z}}|\tilde{\boldsymbol{x}}) \parallel \tilde{p}(\tilde{\boldsymbol{z}}|\tilde{\boldsymbol{x}})] \\ KLR(\boldsymbol{x}_0) &= \frac{\boldsymbol{D}_{KL}[q_\theta(\boldsymbol{z}_0|\boldsymbol{x}_0) \parallel \tilde{p}(\tilde{\boldsymbol{z}}|\tilde{\boldsymbol{x}})]}{\boldsymbol{D}_{KL}(OOD_{max})} \end{aligned} \tag{11}$$

where $D_{KL}(OOD_{max})$ is INCP-KL divergence of the OOD sample, OOD_{max}.

OOD detection criterion: The OOD detection based on INCP-KL Ratio is as follows:

$$\textbf{\textit{Label}}(\boldsymbol{x}_0) = \begin{cases} 0, & \text{if} \quad KLR(\boldsymbol{x}_0) > \alpha \\ 1, & \text{if} \quad KLR(\boldsymbol{x}_0) \leq \alpha \end{cases} \tag{12}$$

where α is the decision threshold. In our study, we set $\alpha = 1$. $\textbf{\textit{Label}}(\boldsymbol{x}_0) = 1$ represents that the test sample $\boldsymbol{x}_0$ is detected as OOD data; $\textbf{\textit{Label}}(\boldsymbol{x}_0) = 0$ represents that $\boldsymbol{x}_0$ is detected as ID data.

4. Experiments and results

4.1. Experimental settings

To evaluate our method and compare with other existing methods, we conduct experiments on multiple datasets. There are two tasks involved in the experiments: the uncertainty estimation task and the OOD detection task.

To obtain the ground-truth OOD data, we synthesize OOD data by adding Gaussian noise to the baseline data (ID data), as Section 3.1 described. The baseline data are from FashionMNIST, MNIS, CIFAR10 and SVHN. Three levels of Gaussian noise ($\mu = 0$, $\sigma = \sigma_0, \sigma_1, \sigma_2$) are generated to represent three levels of uncertainty in OOD data. The detailed settings of the baseline ID data and the synthesized OOD data are in Appendix A.

To replicate the OOD phenomenon in VAE models, we conduct the likelihood tests of ID and OOD data, following the experimental settings in Nalisnick et al. (2019a) (Details are shown in Appendix B). Specifically, we train the traditional VAE on the training set (ID samples) and compute the likelihoods of 1000 random samples from the test set (including both ID samples and their corresponding OOD samples). We exhibit the histogram of the marginal likelihoods of the 1000 tests on VAE (Fig. B.6).

In the uncertainty estimation task 1, we set the standard deviation as the noise level to control the deviation of OOD data from the original data distribution. We run experiments on FashionMNIST, MNIST, CIFAR10, SVHN datasets, respectively. VAE and INCPVAE are trained with the data from the training sets, and then run the inference process with the test samples (OOD data with four levels of noise). The test samples are unseen by models during training process. We calculate the ELBO ratio of the traditional VAE and INCPVAE to estimate the uncertainty on these four datasets. The ELBO ratio is introduced in Section 3.5. Both the ELBO ratio of traditional VAE and INCPVAE are calculated for 1000 random samples from the testing sets. We then compare the ELBO ratio from INCPVAE and VAE (Fig. 2).

In the uncertainty estimation task 2, we train the model on FashionMNIST dataset, and compare the uncertainty estimation of INCPVAE and standard VAE on FashionMNIST (as ID data) and MNIST (as OOD data). This procedure tests whether the capability of uncertainty estimation in one specific data set can be transferred to another dataset. The results are shown in Fig. 3.

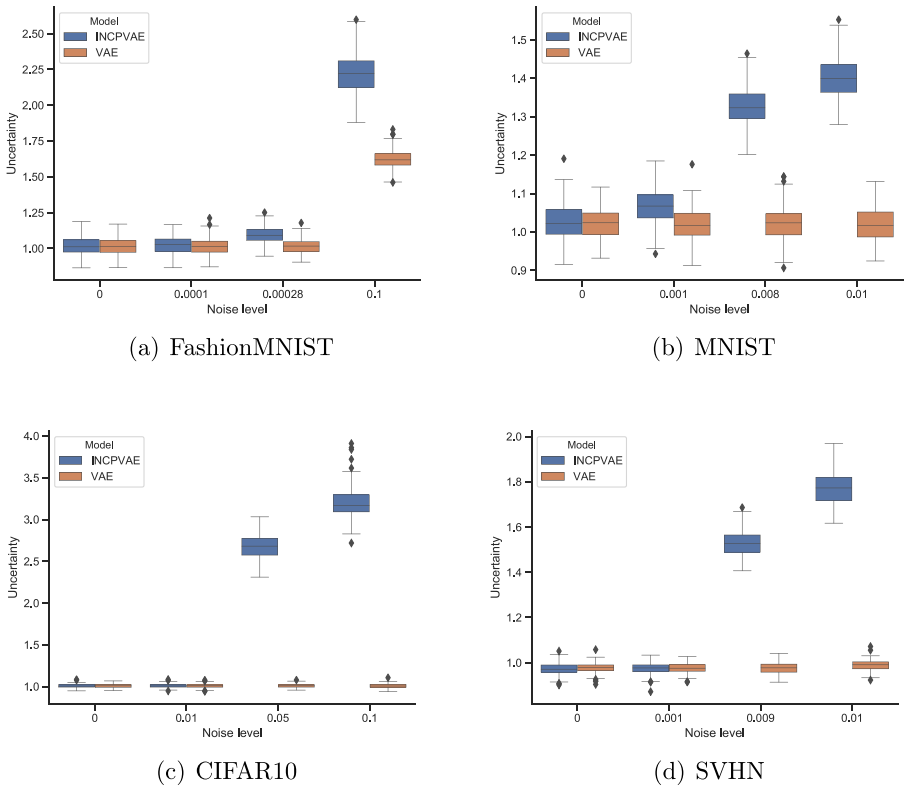


Fig. 2. Results of the uncertainty estimation task 1. The estimated uncertainty (ELBO ratio, $\mathcal{U}(x)$) from the INCPVAE and traditional VAE model on (a) FashionMNIST, (b) MNIST, (c) CIFAR10, (d) SVHN dataset are presented. Four levels of noise are tested.

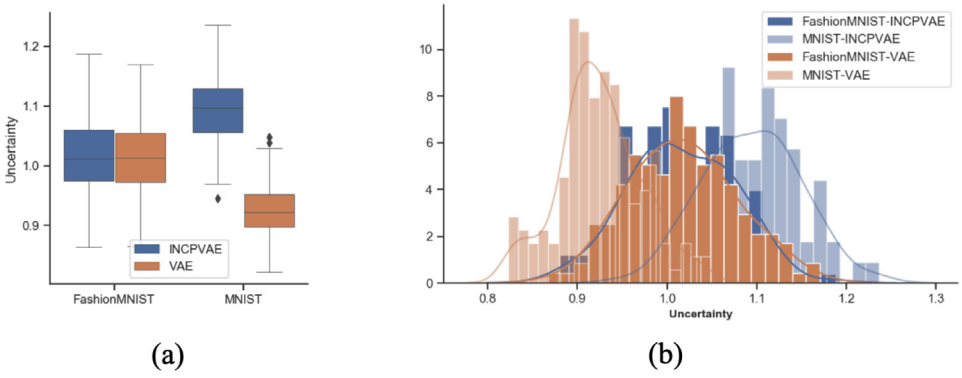


Fig. 3. Results of the uncertainty estimation task 2. The INCPVAE and VAE models are trained on FashionMNIST data, and tested on both FashionMNIST and MNIST. (a) The boxplot of the estimated uncertainty (ELBO ratio, $\mathcal{U}(x)$) from the INCPVAE and traditional VAE model on FashionMNIST and MNIST data. (b) The histogram of the estimated uncertainty from the INCPVAE and traditional VAE model on FashionMNIST and MNIST.

In the OOD detection task, we apply the INCP-KL Ratios (defined in Section 4.3) as a criterion for OOD detection using INCPVAE model. The tasks are conducted on four pairs of datasets (the training set and the test set). Specifically, we train INCPVAE and VAE with the samples only from the training set and then compute INCP-KL Ratios of 1000 random samples from the OOD test set. More details about the settings for OOD detection task are in Appendix B. We quantify the performance of OOD detection task with INCP-KL Ratios (See results in Fig. 4). Moreover, we compare INCPVAE method with 7 existing OOD detection methods, including two likelihood ratio methods, ONID, Mahalanobis distance method, Ensemble method, and WAIC method. We compare INCP-KL of INCPVAE and likelihood of VAE, as well as other baseline methods. The area under the ROC curve (AUROC) and the area under the precision–recall curve (AUPRC) are used as metrics for performance evaluation (See results in Tables 1 and 2).

More details related to the network architecture and implementation are shown in Appendix C. The code will be available at GitHub.

4.2. Results of uncertainty estimation

From Fig. 2, we obtained reliable patterns from these four datasets. When the testing data is drawn without additional perturbations (the noise level is 0), INCPVAE and VAE model present similar uncertainty, suggesting that our model is consistent with standard VAE when it is applied to the ID data. As the noise level increases from 0.01 to 0.1, the INCPVAE-estimated uncertainty of the OOD samples gradually increases in all four datasets, whereas the VAE-estimated uncertainty only shows a slight increase in FashionMNIST dataset and maintains unchanged in the other 3 datasets (MNIST, CIFAR10 and SVHN). These results

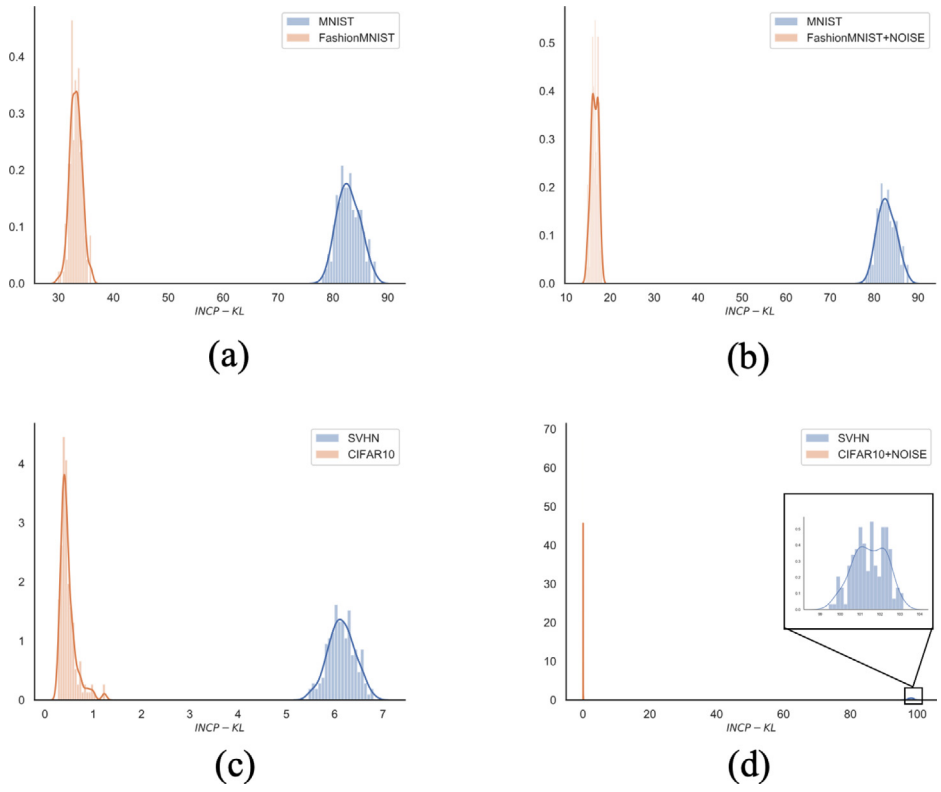


Fig. 4. The INCP-KL ratio of the INCPVAE in the OOD Detection task. The INCPVAE model is (a) trained on FashionMNIST (ID training set) and FashionMNIST + Noise (OOD training set), and tested on FashionMNIST (ID test set) and MNIST (OOD test set); (b) trained on FashionMNIST + Noise (ID) and FashionMNIST (OOD), and tested on FashionMNIST + Noise (ID) and MNIST (OOD); (c) trained on CIFAR10 (ID) and CIFAR10 + Noise (OOD), and tested on CIFAR10 (ID) and SVHN (OOD); (d) trained on CIFAR10 + Noise (ID) and CIFAR10 (OOD), and tested on CIFAR10 + Noise (ID) and SVHN (OOD). The orange lines are the INCP-KL ratios for ID test data, and the blue lines are for OOD test data. Our results show that the INCP-KL ratios of INCPVAE can largely separate ID and OOD inputs.

demonstrate that our INCPVAE model has a strong capability of capturing substantial peculiarity of ID and OOD data with outstanding robustness. Furthermore, we illustrate the $\text{ELBO}_I(x)$ from VAE and INCPVAE in Fig. A.5, where the standard VAE and our INCPVAE were trained and no noise were imposed during testing. Interestingly, we found that INCPVAE and VAE present almost coincident likelihood distributions in these four datasets, implying that INCPVAE model can reserve the generative ability of VAE model.

Moreover, Fig. 3 showed the estimated uncertainty of the test samples from FashionMNIST and MNIST dataset. The INCPVAE obtains higher uncertainty for the OOD data (MNIST) than the ID data (FashionMNIST) during the test step, suggesting that uncertainty estimation of INCPVAE trained in FashionMNIST can be successfully transferred to MNIST dataset. In contrast, VAE showed an opposite trend, which is contradictory to the reality.

4.3. Results of OOD detection

We firstly conduct the OOD detection experiments on FashionMNIST and CIFAR10 datasets using a standard VAE model. Fig. B.6 depicts that the VAE model assigns the OOD data higher likelihoods than training ID data, replicating the nerve-wracking and tricky OOD problem in the likelihood models.

Fig. 4 shows the INCP-KL ratio in 4 OOD detection tests using INCPVAE model. Specifically, in OOD test 1 Fig. 4(a), INCPVAE is trained on FashionMNIST (as ID training set) and FashionMNIST plus noise (as OOD training set), and then test on FashionMNIST (ID test set) and MNIST (OOD test set). In OOD test 2 (Fig. 4(b)), INCPVAE is trained on FashionMNIST plus noise (as ID training set) and FashionMNIST (as OOD training set), and then test on FashionMNIST (ID test set) and MNIST (OOD test set). In OOD test 3 (Fig. 4(c)), INCPVAE is trained on CIFAR10 (as ID training set) and CIFAR10 plus noise (as OOD training set), and then test on CIFAR10 (ID test set) and SVHN (OOD test set). In OOD test 4 (Fig. 4(d)), INCPVAE is trained on CIFAR10 plus noise (as ID training set) and CIFAR10 (as OOD training set), and then test

Table 1
AUROC and AUPRC for detecting OOD inputs using our INCP-KL Ratio method, likelihood method and other baseline methods on FashionMNIST vs. MNIST datasets.

Model	AUROC	AUPRC
INCP-KL Ratio(Baseline + Noise)	**1.000**	**1.000**
INCP-KL Ratio(Baseline)	**1.000**	**1.000**
Likelihood (Traditional VAE)	0.035	0.313
Likelihood Ratio(μ) (Ren et al., 2019)	0.973	0.951
Likelihood Ratio(μ, λ) (Ren et al., 2019)	0.994	0.993
ODIN (Liang et al., 2018)	0.752	0.763
Mahalanobis distance (Lee et al., 2018a)	0.942	0.928
Ensemble, 20 classifiers (Lakshminarayanan et al., 2017)	0.857	0.849
WAIC, 5 models (Choi et al., 2018)	0.221	0.401

Table 2
AUROC and AUPRC for detecting OOD inputs using INCP-KL Ratio method, likelihood method, and other baselines on CIFAR10 vs. SVHN datasets.

Model	AUROC	AUPRC
INCP-KL Ratio(Baseline + Noise)	**1.000**	**1.000**
INCP-KL Ratio(Baseline)	**1.000**	**1.000**
Likelihood (Traditional VAE)	0.057	0.314
Likelihood Ratio(μ) (Ren et al., 2019)	0.931	0.888
Likelihood Ratio(μ, λ) (Ren et al., 2019)	0.930	0.881

on CIFAR10 plus noise (ID test set) and SVHN (OOD test set). It is consistent that the OOD data have higher INCP-KL ratios than the ID data. Together, these results indicate that INCP-KL ratios for the ID test set and the OOD test set have no overlaps, thus a simple threshold on INCP-KL ratios can detect the OOD data.

To comprehensively compare the OOD detection performance of different methods, we perform the OOD detection task using INCPVAE and a variety of baseline models. Tables 1 and 2 list the AUROC and AUPRC metrics on the OOD detect tasks (FashionMNIST vs. MNIST, and CIFAR10 vs. SVHN, respectively). Evidently, our model achieves the highest AUROC and AUPRC scores on both tests, compared with other baseline methods.

5. Discussion and conclusion

In this study, we have proposed a novel VAE model, called INCPVAE, for reliable uncertainty estimation and OOD detection. Specifically, we firstly improve the noise contrastive prior, called INCP, to be suitable for VAE models, and then present a hybrid method combining INCP with the encoder of VAE framework. Using INCPVAE model, OOD samples can be generated by adding Gaussian noise into the ID samples; therefore, INCPVAE model can be jointly trained with ID data and OOD data. We define a new metric (ELBO Ratio) for uncertainty estimation and a new OOD detection criterion which is based on INCP-KL Ratio.

We reproduced the results that traditional VAE easily assigns higher likelihoods for OOD samples than ID samples (Fig. B.6). These results suggest that the likelihood in traditional VAEs is not a good metric to detect the OOD data, which is consistent with previous studies (Choi et al., 2018; Hendrycks et al., 2019; Lee et al., 2018b; Nalisnick et al., 2019a, 2019b) and the model with reliable uncertainty estimation can improve the performance of OOD detection. Firstly, we proposed a new metric, ELBO Ratio. The result of the uncertainty estimation task 1 across four datasets (Fig. 2) demonstrates that ELBO Ratio increases as the noise increases. The uncertainty estimation task 2 shows that INCPVAE trained with FashionMNIST data can accurately estimate the uncertainty in MNIST data, whereas the VAE model failed to transfer the uncertainty information (Fig. 3). Together, these results indicate that ELBO Ratio can reliably index the uncertainty in the input data.

Secondly, we proposed a metric called INCP-KL ratio to detect OOD data. A simple threshold on INCP-KL ratios (e.g. $\alpha = 1$ in Eq. (12)) can be used to detect OOD data in INCPVAE model. The results of OOD detection task demonstrate that our model achieves SOTA performance to differentiate OOD and ID data, compared with baseline methods (Tables 1 and 2). INCPVAE model, as a model-independent method to OOD detection, paves a way for future VAE applications on OOD detection. Also, INCPVAE can be easily extended to anomaly detection and adversarial example detection.

Despite the advantages of our work, there are still some limitations and future work worth mentioning. We only focused on the uncertainty estimation and OOD detection using VAE model in this study. It is interesting to extend INCP to other generative models, such as GAN. Moreover, we generate OOD data by adding Gaussian noise to ID data, which cannot capture the characteristics of the OOD data in the real applications. Other methods to generate appropriate OOD inputs are worthy of investigation; for example, using GAN to generate OOD data (Lee et al., 2018b). The realistic OOD data can help to train INCPVAE models, as it can be potentially used to generate priors of INCPVAE. Alternatively, adversarial examples (Goodfellow, Shlens, & Szegedy, 2015) may also be used to train INCPVAE, in order to enhance robustness of VAE.

In summary, we integrate INCP into VAE framework to solve the problem that the OOD detection techniques for deep generative models are hardly transferred to VAEs (Xiao, Yan, & Amit, 2020).

Table A.3
Baselines are FashionMNIST, MNIST, CIFAR10,SVHN. Noise is generated by Gaussian Noise(μ, σ^2), where $\mu = 0$, $\sigma = \sigma_0, \sigma_1, \sigma_2$.

Dataset	VAE	INCPVAE
ID training set	Baseline	Baseline
OOD training set	–	Baseline + Noise(σ_1)
ID testing set	Baseline	Baseline
OOD testing set0	Baseline + Noise(σ_0)	Baseline + Noise(σ_0)
OOD testing set1	Baseline + Noise(σ_1)	Baseline + Noise(σ_1)
OOD testing set2	Baseline + Noise(σ_2)	Baseline + Noise(σ_2)

Table A.4
The levels of noises added to four baseline datasets. Noise is generated by Gaussian Noise(μ, σ^2), where $\mu = 0$, $\sigma = \sigma_0, \sigma_1, \sigma_2$.

Noise level	FashionMNIST	MNIST	CIFAR10	SVHN
σ_0	0.0001	0.001	0.01	0.001
σ_1	0.00028	0.008	0.05	0.009
σ_2	0.1000	0.010	0.10	0.010

Table A.5
True OOD posterior of INCPVAE $\tilde{p}(\tilde{z} \mid \tilde{x})$ is employed by Gaussian distribution $\mathcal{N}(\mu_{\tilde{x}}, \sigma^2_{\tilde{x}})$.

Dataset	Uncertainty level ($\sigma_{\tilde{x}}$)
FashionMNIST	$e^{0.65}$
MNIST	$e^{0.65}$
CIFAR10	$e^{1.00}$
SVHN	$e^{1.00}$

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank the anonymous reviewers for their comments and Dr. Steffen Bollmann for his suggestion. This work was funded in part by the National Natural Science Foundation of China (62001205), Shenzhen Science and Technology Innovation Committee (20200925155957004, SGDX2020110309280100, KCXFZ2020122117340001), Guangdong Natural Science Foundation Joint Fund (2019A1515111038), Shenzhen Key Laboratory of Smart Healthcare Engineering (ZDSYS20200811144003009), CAAI-Huawei Mindspore Open Fund (CAAIXSJLJJ-2020-024A), Fundamental Research Funds for Central Universities (DUT21RC(3)091), Beijing Science and Technology Programs (Z191100007519009).

Appendix A. Settings for uncertainty estimation

In this section, we introduce detailed settings for uncertainty estimation. To evaluate uncertainty estimation from the traditional Variational Auto-encoder (VAE) and from the Improved Noise Contrastive Priors VAE (INCPVAE), we train VAE on in-distribution (ID) training set and INCPVAE on the ID and out-of-distribution (OOD) training set. Then we test both of VAE and INCPVAE on ID testing set and OOD testing set0/set1/set2, respectively. See full lists in Table A.3. The OOD training set and testing set0/set1/set2 are generated by adding three levels of Gaussian noise to the baseline (See Table A.4).

For each image dataset, the true OOD posterior of INCPVAE (or OOD data output prior) is assumed by Gaussian distribution with a specific variance (See Table A.5), which represents that these four datasets have various uncertainties.

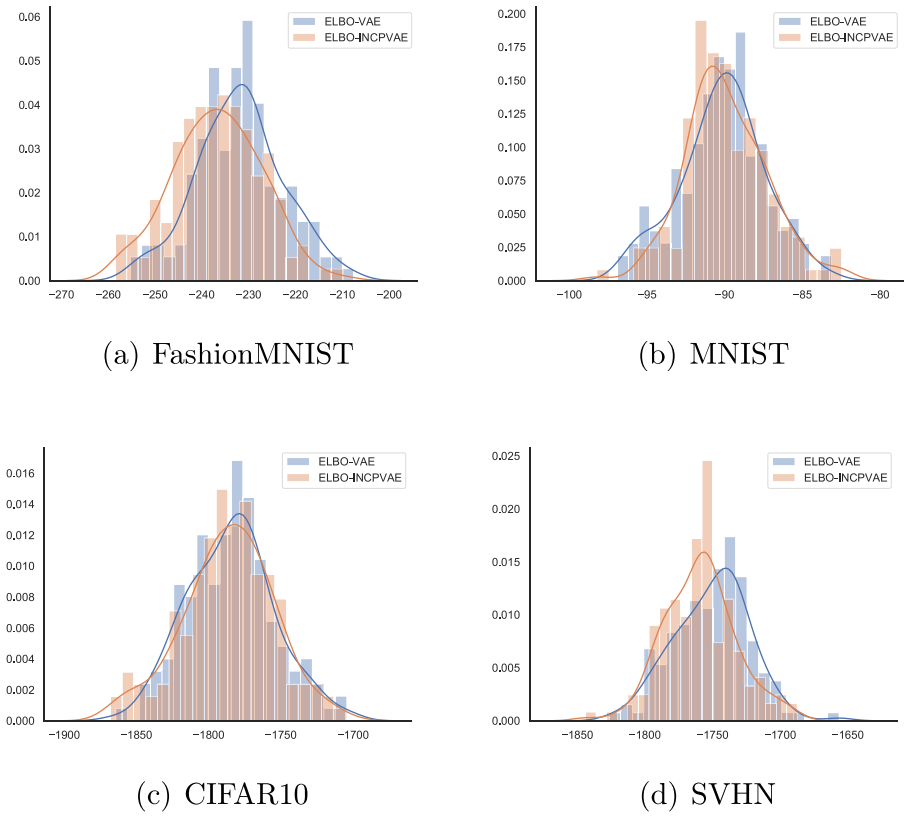


Fig. A.5. The histogram of the ELBO of the ID data, $\text{ELBO}_I(x)$, for VAE and INCPVAE. (a) FashionMNIST, (b) MNIST, (c) CIFAR10, (d) SVHN dataset. These results demonstrate that INCPVAE has similar $\text{ELBO}_I(x)$ with VAE.

Table A.6
Datasets: VAE for OOD detection.

Exp	ID training set	ID test set	OOD test set
Exp1	FashionMNIST	FashionMNIST	MNIST
Exp2	CIFAR10	CIFAR10	SVHN

Appendix B. Settings for OOD detection

In this section, we introduce detailed settings of OOD detection experiments. Firstly, following the most challenging experiment reported by Nalisnick et al., we train VAE on ID training set and test on ID and OOD testing set (See Table A.6). Secondly, to evaluate the OOD detection of INCPVAE, we train INCPVAE on the ID and OOD training set, and test INCPVAE on OOD testing set and OOD testing set1 (See Table A.7). The ID and OOD training set, as well as the OOD testing set, are generated by adding Gaussian noise with three levels to baseline (See Table B.8).

For different datasets, the true OOD posterior of INCPVAE (or OOD data output prior) is Gaussian distribution with different variance (See Table B.9), which represents that different datasets have different uncertainties.

Table B.8
Datasets for INCP-KL Ratios of INCPVAE. Noise is generated by Gaussian Noise(μ, σ^2), where set $\mu = 0$, $\sigma = \sigma_3, \sigma_4$.

Dataset	Noise level (σ_3)	Noise level (σ_4)
FashionMNIST	0.00028	0.00050
CIFAR10	0.05000	0.09000

Table B.9
True OOD posterior of INCPVAE $\tilde{p}(\tilde{z} \mid \tilde{x})$ is employed by Gaussian distribution $\mathcal{N}(\mu_{\tilde{x}}, \sigma^2_{\tilde{x}})$.

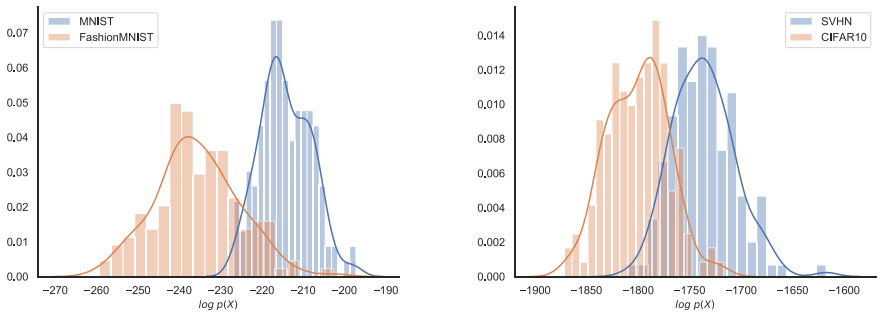
Dataset	Uncertainty level $(\sigma_{\tilde{x}})$
FashionMNIST	$e^{0.65}$
CIFAR10	$e^{1.00}$

Appendix C. Settings for implementation detail

In the experiments, VAE and INCPVAE are trained on FashionMNIST and CIFAR10. All models are trained with images normalized to [0, 1] on 1 × NVIDIA TITAN RTX GPU. In all experiments, VAE and INCPVAE consist of an encoder with the architecture given in Table B.10 and a decoder shown in Table C.11.

Table A.7
Datasets for INCP-KL Ratios of INCPVAE. Fashion is short for FashionMNIST.

Exp	ID training set	OOD training set	OOD test set1	OOD test set2
Exp1	Fashion	Fashion + Noise(σ_3)	Fashion + Noise(σ_3)	MNIST
Exp2	Fashion + Noise(σ_4)	Fashion	Fashion	MNIST
Exp3	CIFAR10	CIFAR10 + Noise(σ_3)	CIFAR10 + Noise(σ_3)	SVHN
Exp4	CIFAR10 + Noise(σ_4)	CIFAR10	CIFAR10	SVHN



(a) VAE (b) VAE

Fig. B.6. The histogram of the marginal likelihood of the VAE. (a) VAE trained on FashionMNIST (ID), and tested on FashionMNIST and MNIST (OOD); (b) VAE trained on CIFAR10 (ID), and tested on CIFRAR10 (ID) and SVHN (OOD). The orange lines are for ID data, and the blue lines are for OOD data.

Table B.10
Encoder architecture. This architecture was used for VAE and INCPVAE trained on FashionMNIST with linear layer units 3136 and CIFAR10 with 4096.

Operation	Kernel	Stride	Features	Padding
Input	–	–	–	–
Convolution	5 × 5	2 × 2	256	0
Convolution	5 × 5	2 × 2	32	0
Convolution	5 × 5	1 × 1	32	0
Dense	–	–	3136/4096	–

Table C.11
Decoder architecture. This architecture was used for VAE and INCPVAE trained on FashionMNIST with linear layer units 3136 and CIFAR10 with 4096.

Operation	Kernel	Stride	Features	Padding
Input $\boldsymbol{z}$	–	–	–	–
Dense	–	–	3136/4096	–
Dense	–	–	1568/2048	–
Transposed Convolution	5 × 5	1 × 1	32	0
Transposed Convolution	5 × 5	2 × 2	256	0
Transposed Convolution	5 × 5	2 × 2	3	0

Both VAE and INCPVAE use Leaky Relu activation function. We train the VAE for 200 epochs with a constant learning rate $1e^{-4}$, meanwhile using Adam optimizer and batch size 64 in each experiment.

References

Alemi, Alexander A, Fischer, Ian, & Dillon, Joshua V (2018). Uncertainty in the variational information bottleneck. arXiv preprint arXiv:1807.00906.

Bauer, Matthias, & Mnih, Andriy (2019). Resampled priors for variational autoencoders. In *International conference on artificial intelligence and statistics (aistats)*.

Bishop, Christopher M (1994). Novelty detection and neural network validation. *IEEE Proceedings-Vision, Image and Signal Processing*.

Bütepage, Judith, Poklukar, Petra, & Kragic, Danica (2019). Modeling assumptions and evaluation schemes: On the assessment of deep latent variable models. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr workshops)*.

Choi, Hyunsun, Jang, Eric, & Alemi, Alexander A (2018). Waic, but why? generative ensembles for robust anomaly detection. arXiv preprint arXiv:1810.01392.

Daxberger, Erik, & Hernández-Lobato, José Miguel (2019). Bayesian variational autoencoders for unsupervised out-of-distribution detection. arXiv preprint arXiv:1912.05651.

Denouden, Taylor, Salay, Rick, Czarnecki, Krzysztof, Abdelzad, Vahdat, Phan, Buu, & Vernekar, Sachin (2018). Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance. arXiv preprint arXiv:1812.02765.

DeVries, Terrance, & Taylor, Graham W (2018). Learning confidence for out-of-distribution detection in neural networks. arXiv preprint arXiv:1802.04865.

Gal, Yarin, & Ghahramani, Zoubin (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International conference on machine learning (icml)*.

Goodfellow, Ian J, Shlens, Jonathon, & Szegedy, Christian (2015). Explaining and harnessing adversarial examples. In *International conference on learning representations (iclr)*.

Gutmann, Michael, & Hyvärinen, Aapo (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *International conference on artificial intelligence and statistics (aistats)*.

Hafner, Danijar, Tran, Dustin, Irpan, Alex, Lillicrap, Timothy, & Davidson, James (2018). Reliable uncertainty estimates in deep neural networks using noise contrastive priors. arXiv preprint arXiv:1807.09289.

Hendrycks, Dan, & Gimpel, Kevin (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International conference on learning representations (iclr)*.

Hendrycks, Dan, Mazeika, Mantas, & Dietterich, Thomas G (2019). Deep anomaly detection with outlier exposure. In *International conference on learning representations (iclr)*.

Kingma, Durk P, & Dhariwal, Prafulla (2018). Glow: Generative flow with invertible 1x1 convolutions. In *Advances in neural information processing systems (neurips)*.

Kingma, Diederik P, & Welling, Max (2014). Auto-encoding variational bayes. In *International conference on learning representations (iclr)*.

Lakshminarayanan, Balaji, Pritzel, Alexander, & Blundell, Charles (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems (neurips)*.

Lee, Kimin, Lee, Kibok, Lee, Honglak, & Shin, Jinwoo (2018a). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in neural information processing systems (neurips)*.

Lee, Kimin, Lee, Honglak, Lee, Kibok, & Shin, Jinwoo (2018b). Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International conference on learning representations (iclr)*.

Liang, Shiyu, Li, Yixuan, & Srikant, R (2018). Enhancing the reliability of out-of-distribution image detection in neural networks. In *International conference on learning representations (iclr)*.

Liu, Qianhui, Pan, Gang, Ruan, Haibo, Xing, Dong, Xu, Qi, & Tang, Huajin (2020). Unsupervised aer object recognition based on multiscale spatio-temporal features and spiking neurons. *IEEE Transactions on Neural Networks and Learning Systems*.

Maaløe, Lars, Fraccaro, Marco, Liévin, Valentin, & Winther, Ole (2019). Biva: A very deep hierarchy of latent variables for generative modeling. *Advances in Neural Information Processing Systems (NeurIPS)*.

Maciag, Piotr S., Kryszkiewicz, Marzena, Bembenik, Robert, Lobo, Jesús López, & Ser, Javier Del (2021). Unsupervised anomaly detection in stream data with online evolving spiking neural networks. *Neural networks : the official journal of the International Neural Network Society*.

Meronen, Lassi, Irwanto, Christabella, & Solin, Arno (2020). Stationary activations for uncertainty calibration in deep learning. *Advances in Neural Information Processing Systems (NeurIPS)*.

Mnih, Andriy, & Kavukcuoglu, Koray (2013). Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in neural information processing systems (neurips)*.

Nalisnick, Eric, Matsukawa, Akihiro, Teh, Yee Whye, Gorur, Dilan, & Lakshminarayanan, Balaji (2019a). Do deep generative models know what they don't know? In *International conference on learning representations (iclr)*.

Nalisnick, Eric, Matsukawa, Akihiro, Teh, Yee Whye, & Lakshminarayanan, Balaji (2019b). Detecting out-of-distribution inputs to deep generative models using a test for typicality. arXiv preprint arXiv:1906.02994.

Van den Oord, Aaron, Kalchbrenner, Nal, Espeholt, Lasse, Vinyals, Oriol, Graves, Alex, et al. (2016). Conditional image generation with pixelcnn decoders. In *Advances in neural information processing systems (neurips)*.

Oord, Aaron van den, Kalchbrenner, Nal, & Kavukcuoglu, Koray (2016). Pixel recurrent neural networks. In *International conference on machine learning (icml)*.

Pidhorskyi, Stanislav, Almohsen, Ranya, Adjeroh, Donald A, & Doretto, Gianfranco (2018). Generative probabilistic novelty detection with adversarial autoencoders. In *Advances in neural information processing systems (neurips)*.

Ren, Jie, Liu, Peter J, Fertig, Emily, Snoek, Jasper, Poplin, Ryan, Depristo, Mark, et al. (2019). Likelihood ratios for out-of-distribution detection. In *Advances in neural information processing systems (neurips)*.

Rezende, Danilo Jimenez, Mohamed, Shakir, & Wierstra, Daan (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning (icml)*.

Serrà, Joan, Álvarez, David, Gómez, Vicenç, Slizovskaia, Olga, Núñez, José F, & Luque, Jordi (2020). Input complexity and out-of-distribution detection with likelihood-based generative models. In *International conference on learning representations (iclr)*.

Song, Yang, Shu, Rui, Kushman, Nate, & Ermon, Stefano (2018). Constructing unrestricted adversarial examples with generative models. In *Advances in neural information processing systems (neurips)*.

Tomczak, Jakub, & Welling, Max (2018). Vae with a VampPrior. In *International conference on artificial intelligence and statistics (aistats)*.

Vyas, Apoorv, Jammalamadaka, Nataraj, Zhu, Xia, Das, Dipankar, Kaul, Bharat, & Willke, Theodore L (2018). Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *Proceedings of the european conference on computer vision (eccv)* (pp. 550–564).

Xiao, Zhisheng, Yan, Qing, & Amit, Yali (2020). Likelihood regret: An out-of-distribution detection score for variational auto-encoder. arXiv preprint arXiv:2003.02977.

Xu, Qi, Qi, Yu, Yu, Hang, Shen, Jiangrong, Tang, Huajin, & Pan, Gang (2018). Csnn: an augmented spiking based framework with perceptron-inception. In *IJCAI*.

Xu, Qi, Zhang, Ming, Gu, Zonghua, & Pan, Gang (2019). Overfitting remedy by sparsifying regularization on fully-connected layers of cnns. *Neurocomputing*.

Zhang, Mingtian, Bird, Thomas, Habib, Raza, Xu, Tianlin, & Barber, David (2019). Variational f-divergence minimization. arXiv preprint arXiv:1907.11891.

Zhang, Xiao, Chen, Jinghui, Gu, Quanquan, & Evans, David (2020). Understanding the intrinsic robustness of image distributions using conditional generative models. In *International conference on artificial intelligence and statistics* (pp. 3883–3893). PMLR.