Out-of-Distribution Detection with Likelihoods Assigned by Deep Generative Models Using Multimodal Prior Distributions

Ryo Kamoi, Kei Kobayashi
Keio University, Japan
ryo_kamoi_st@keio.jp, kei@math.keio.ac.jp

Abstract

Modern machine learning systems can exhibit undesirable and unpredictable behavior in response to out-of-distribution inputs. Consequently, applying out-of-distribution detection to address this problem is an active subfield of safe AI. Probability density estimation is one popular approach for out-of-distribution detection of low-dimensional data. However, for high dimensional data, recent work has reported that deep generative models can assign higher likelihoods to out-of-distribution data than to training data. We propose a new method to detect out-of-distribution inputs using deep generative models with multimodal prior distributions. Our experimental results show that our models trained on Fashion-MNIST successfully assign lower likelihoods to MNIST, and successfully function as out-of-distribution detectors.

1 Introduction

The field of machine learning has experienced rapid progress in various areas including computer vision and natural language processing. However, modern machine learning systems can return predictions with high confidence even for out-of-distribution inputs (Goodfellow, Shlens, and Szegedy 2015; Gal 2016). This is a serious problem in terms of the safety of machine learning. As a real world example, in 2016, an autonomous car collided with a tractor trailer on a highway with no warning (Natural Highway Traffic Safety Administration 2016). They reported that the situation was outside the expected performance capabilities of the system.

To avoid this problem, out-of-distribution detection is an important field of study within safe AI. For low-dimensional data, many studies have been performed over the past few decades. The review paper by Pimentel et al. (2014) categorized detection methods into five groups: probabilistic, distance-based, reconstruction-based, domain-based, and information-theoretic methods. However, it is known that those methods cannot be applied to high dimensional cases straightforwardly, so new detection methods for high dimensional data have been proposed recently (Theis, Van Den Oord, and Bethge 2016; Hendrycks and Gimpel 2017; Liang, Li, and Srikant 2018). In this work, we focus on probabilistic approaches, which estimate the distribution of training data via a probabilistic model and are based on the intuition that out-of-distribution inputs locate in low-density areas (Bishop 1994). For high dimensional data, recent work (Choi, Jang, and Alemi 2019; Nalisnick et al. 2019a) has reported that deep generative models cannot detect out-of-distribution inputs via assigned likelihoods. Methods for alleviating this problem have been proposed from various perspectives (Hendrycks, Mazeika, and Dietterich 2019; Choi, Jang, and Alemi 2019; Nalisnick et al. 2019b).

We propose the use of deep generative models with multimodal prior distributions to alleviate the out-of-distribution problem. Although a typical choice of the prior is the standard normal distribution, various studies have proposed the use of alternatives (Dilokthanakul et al. 2016; Chen et al. 2017; Tomczak and Welling 2017). Previous work on the choice of a prior distribution for deep generative models have criterion based on the representative ability, natural fit to data sets, and the likelihood or reconstruction quality of in-distribution inputs. To the best of our knowledge, this is the first work focusing on the relationships between the prior distribution and likelihood assigned to out-of-distribution data. Here, we consider data sets which can be naturally partitioned into clusters, so its underlying distribution can be approximated as a multimodal distribution with components located far away from each other. This assumption is reasonable for many data sets found in the wild such as Fashion-MNIST containing different types of images such as T-shirts, shoes, and bags. If a unimodal prior distribution is used to train generative models on such data sets, the models are forced to learn the mappings between unimodal and multimodal distributions. We consider this inconsistency is an important factor causing the assignment of high likelihoods to out-of-distribution areas.

We evaluate our method on Fashion-MNIST, and show that models with multimodal prior distributions assign lower likelihoods to out-of-distribution inputs. In our experiments, we use Gaussian mixture distributions that are not trainable, and manually assign each data to a component of the prior distribution based on the labels in the data set. While it is difficult to apply this method to more complex data sets, our observation motivates further work on the relationships between prior distributions and the out-of-distribution likelihoods.

2 Related Work

Our work is directly motivated by the recent observation that deep generative models can assign higher likelihoods to out-of-distribution inputs (Nalisnick et al. 2019a; Choi, Jang, and Alemi 2019).

2.1 Out-of-Distribution Detection by Deep Generative Models

Nalisnick et al. (2019a) reported that deep generative models such as Variational Autoencoders (VAEs), flow-based models, and PixelCNN can assign higher likelihoods to out-of-distribution inputs. Solutions have been proposed from various perspectives. Hendrycks et al. (2019) proposed "outlier exposure", a technique using outlier data during training to lower the likelihoods assigned to out-of-distribution inputs. Another line of study is to use alternative metrics. Choi et al. (2019) proposed the use of the Watanabe-Akaike Information Criterion (WAIC). Nalisnick et al. (2019b) proposed the use of a hypothesis test to check whether an input resides in the model's typical set. Grathwohl et al. (2020) proposed to use the l_2 norm of the gradient of the log-likelihood as a score. To our knowledge, no prior work has focused on the relationship between prior distributions and out-of-distribution likelihoods.

2.2 Prior distribution

The standard Gaussian distributions are typically used as prior distributions for deep generative models such as VAEs and flow-based models. However, various studies propose different options. One line of study suggests more expressive prior distributions: multimodal distributions (Johnson et al. 2016; Dilokthanakul et al. 2016; Tomczak and Welling 2017; Nalisnick and Smyth 2017), stochastic processes (Nalisnick and Smyth 2017; Goyal et al. 2017; Casale et al. 2018), and an autoregressive models (Chen et al. 2017; van den Oord, Vinyals, and Kavukcuoglu 2017).

3 Proposed Method

Motivation If a data distribution is unimodal, intermediate images of two in-distribution data should have high likelihoods. However, this assumption is not reasonable for many data sets including Fashion-MNIST that contains dissimilar images such as T-shirts and bags whose intermediate images may not be in-distribution data. Therefore, we assume that data distributions can be approximated by multimodal distributions. Under the assumption that the data distribution is multimodal with components located far away from each other, high likelihood areas of the prior and data distribution have differing topologies if the prior distribution is unimodal. Therefore, some high likelihood areas in the prior distribution will be mapped to out-of-distribution areas in the data distribution if we assume that deep generative models learn topology preserving mappings between prior and data distributions. While the probability density of latent variables in the prior distribution is not the only factor influencing the likelihoods assigned by the models, we consider this inconsistency an important factor in the out-of-distribution phenomenon.

Model We replace the prior distributions of deep generative models with mixture distributions $\sum_{i=1}^{K} p_i/K$ that are not trainable, and we assume that all components are uniformly weighted for simplicity. We manually assign each input to a component of the prior distribution based on the labels in the data set. During training, the likelihood for each input is evaluated with a different unimodal prior distribution p_i (different i for each input), which is the component of the multimodal prior distribution assigned to each input. The test likelihood is evaluated on a mixture prior distribution $\sum_{i=1}^{K} p_i/K$ without using the label information used during training.

Evaluation We evaluate the generative models as out-of-distribution detectors by interpreting the log-likelihoods assigned to inputs as classifier scores. Here, we consider out-of-distribution data as the negative class. We evaluate our models with four different metrics: the false positive rate (FPR) at 95% true positive rate (TPR), Detection Error at 95% TPR, the Area Under the Receiver Operating Characteristic curve (AUROC), and the Area Under the Precision-Recall curve (AUPR). The Detection Error is defined as $P_e = 0.5(1-\mathrm{TPR})+0.5\mathrm{FPR}$. Our evaluation assumes that in-distribution and out-of-distribution inputs have an equal probability of appearing in the test set.

4 Experiments

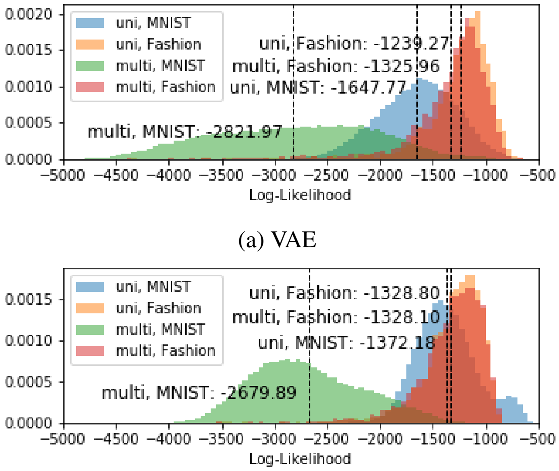
We assess deep generative models with multimodal prior distributions as out-of-distribution detectors trained on Fashion-MNIST (Xiao, Rasul, and Vollgraf 2017) and evaluated on MNIST (LeCun et al. 1998) as the out-of-distribution inputs.

4.1 Model Structure and Training Details

VAE Our implementation is based on the architecture described in (Rosca, Lakshminarayanan, and Mohamed 2018; Nalisnick et al. 2019a). The encoder is comprised of five convolutional layers with 5×5 kernels. The output channels are $[8, 16, 32, 64, 64]$, the strides are $[2, 1, 2, 1, 2]$, and the paddings are $[1, 1, 1, 1, 1]$. After the convolutional layers, two fully connected layers project into 50 dimensional means and log-variances. The latent variables are projected into 3,136 dimensions with a fully connected layer, and reshaped into $7 \times 7 \times 64$. The decoder is comprised of five convolutional layers. The first four layers use 5×5 kernels, and the last layer uses a 4×4 kernel. The output channels are $[64, 32, 64, 256]$, the strides are $[2, 2, 1, 1]$, and the paddings are $[2, 1, 1, 1]$. We assume i.i.d. categorical distributions on pixels. We train for 1,000 epochs using the Adam optimizer (Kingma and Lei Ba 2014) with parameters $\beta_1 = 0.5, \beta_2 = 0.9$, and a constant learning rate of $1\mathrm{e}{-3}$. We use 5,000 samples to approximate the test likelihood.

Glow Our implementation is based on the code hosted in OpenAI's open source repository[1]. We use 1 block of 32

[1] https://github.com/openai/glow



(a) VAE

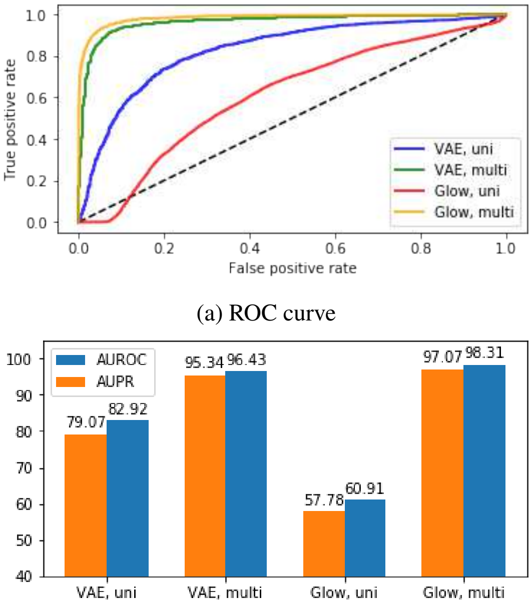
(b) Glow

Figure 1: Histograms of log-likelihoods assigned by VAEs and Glow trained on Fashion-MNIST (label 1 and 7). “uni” denotes a standard Gaussian prior distribution, and “multi” denotes a bimodal Gaussian Mixture prior distribution. For Fashion-MNIST, we report the likelihoods evaluated on test data. Models using multimodal prior distributions alleviate the out-of-distribution problem.

affine coupling layers, squeezing the spatial dimension after the 16-th layer. To mitigate spatial dependencies on the latent variables, we do not use the multi-scale architecture, which splits the latent variables after squeezing (Dinh, Sohl-Dickstein, and Bengio 2017). Additionally, we apply 1×1 convolution over width, height, and channel after the encoder, and the inverse operation before the decoder. We train for 1,000 epochs using the Adam optimizer in accordance with the OpenAI’s code. We use a learning rate of $1\mathrm{e}{-3}$, which is linearly annealed from zero over the first 10 epochs.

4.2 Evaluation

Here, we evaluate the deep generative models trained only on label 1 (Trouser) and 7 (Sneaker) of Fashion-MNIST. We compare two types of prior distributions: a standard Gaussian and a bimodal Gaussian Mixture distribution. The means of the bimodal prior are $[\pm 75, 0, \ldots, 0]$ for VAE, and $[\pm 50, 0, \ldots, 0]$ for Glow. The variances are $\mathrm{diag}([1, \ldots, 1])$ for all components. In the training phase, images with different labels are allocated to different components. Figure 1 shows that the models using multimodal prior distributions successfully assign lower likelihood to MNIST, the out-of-distribution data, while the models using unimodal prior distributions assign high likelihood to MNIST.



(a) ROC curve

(b) AUROC and AUPR

Figure 2: ROC curves, AUROC, and AUPR of the out-of-distribution detectors using the log-likelihood assigned by the models with unimodal and multimodal prior distributions. Higher values are better for AUROC and AUPR.

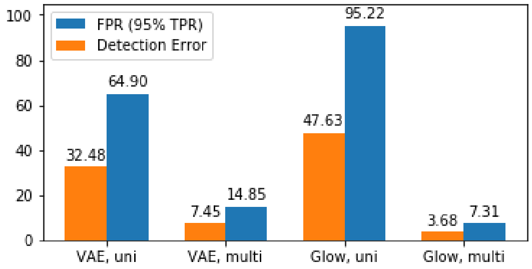


Figure 3: False positive rate (FPR) at 95% true positive rate (TPR) and Detection Error $P_e = 0.5(1 - \mathrm{TPR}) + 0.5\mathrm{FPR}$ at 95% TPR of the out-of-distribution detector using the log-likelihood assigned by the models with unimodal and multimodal prior distributions. Lower values are better for both metrics.

We evaluate the models as out-of-distribution detectors. Figure 2 shows the ROC curve, AUROC, and AUPR of the detectors. The models using multimodal prior distributions increase both AUROC and AUPR for VAE and Glow. Figure 3 shows FPR at 95% TPR and Detection Error of the detectors. Our models reduce both metrics significantly on VAE and Glow. Improvements in all metrics evaluated demonstrate that models using multimodal prior distributions improve the performance as out-of-distribution detectors.

5 Conclusion and Discussion

We propose a new method for out-of-distribution detection using deep generative models with multimodal prior distributions. Recent work (Nalisnick et al. 2019a; Choi, Jang, and Alemi 2019) has shown that deep generative models can assign higher likelihoods to out-of-distribution inputs than to training data, and the reported results suggest that they cannot be used as out-of-distribution detectors. We show that our model lowers the out-of-distribution likelihoods, and functions as an out-of-distribution detector on Fashion-MNIST vs. MNIST. To the best of our knowledge, this is the first work on the relationship between the choice of a prior distribution and the likelihoods assigned to out-of-distribution inputs.

However, it is difficult to apply our method to complex data as it would require a large number of components, better data allocation strategy, and more sophisticated prior distributions. Our observations motivate further work on latent variable space and prior distribution design for deep generative models.

Acknowledgements

This paper has benefited from advice and English language editing from Masayuki Takeda. This work was supported by JSPS KAKENHI (JP19K03642, JP19K00912) and RIKEN AIP Japan.

References

Bishop, C. M. 1994. Novelty Detection and Neural Network Validation. *IEE Proceedings: Vision, Image and Signal Processing* 141(4):217–222.

Casale, F. P.; Dalca, A. V.; Saglietti, L.; Listgarten, J.; and Fusi, N. 2018. Gaussian process prior variational autoencoders. In *Conference on Neural Information Processing System (NeurIPS)*.

Chen, X.; Kingma, D. P.; Salimans, T.; Duan, Y.; Dhariwal, P.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2017. Variational Lossy Autoencoder. In *International Conference on Learning and Representation (ICLR)*.

Choi, H.; Jang, E.; and Alemi, A. A. 2019. WAIC, but Why? Generative Ensembles for Robust Anomaly Detection. *arXiv preprint arXiv:1810.01392*.

Dilokthanakul, N.; Mediano, P. A. M.; Garnelo, M.; Lee, M. C. H.; Salimbeni, H.; Arulkumaran, K.; and Shanahan, M. 2016. Deep Unsupervised Clustering with Gaussian Mixture Variational Autoencoders. *arXiv preprint arXiv:1611.02648*.

Dinh, L.; Sohl-Dickstein, J.; and Bengio, S. 2017. Density estimation using Real NVP. In *International Conference on Learning and Representation (ICLR)*.

Gal, Y. 2016. *Uncertainty in Deep Learning*. Ph.D. Dissertation, University of Cambridge.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations (ICLR)*.

Goyal, P.; Hu, Z.; Liang, X.; Wang, C.; and Xing, E. P. 2017. Nonparametric Variational Auto-encoders for Hierarchical Representation Learning. In *IEEE International Conference on Computer Vision (ICCV)*.

Grathwohl, W.; Wang, K.-C.; Jacobsen, J.-H.; Duvenaud, D.; Swersky, K.; and Norouzi, M. 2020. Your Classifier is Secretly an Energy Based Model and You Should Treat it Like One. In *International Conference on Learning and Representation (ICLR)*.

Hendrycks, D., and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning and Representation (ICLR)*.

Hendrycks, D.; Mazeika, M.; and Dietterich, T. 2019. Deep Anomaly Detection with Outlier Exposure. In *International Conference on Learning and Representation (ICLR)*.

Johnson, M. J.; Duvenaud, D.; Wiltschko, A. B.; Datta, S. R.; and Adams, R. P. 2016. Composing graphical models with neural networks for structured representations and fast inference. In *Conference on Neural Information Processing Systems (NIPS)*.

Kingma, D. P., and Lei Ba, J. 2014. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE* 86(11):2278 – 2324.

Liang, S.; Li, Y.; and Srikant, R. 2018. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *International Conference on Learning and Representation (ICLR)*.

Nalisnick, E., and Smyth, P. 2017. Stick-Breaking Variational Autoencoders. In *International Conference on Learning Representations (ICLR)*.

Nalisnick, E.; Matsukawa, A.; Teh, Y. W.; Gorur, D.; and Lakshminarayanan, B. 2019a. Do Deep Generative Models Know What They Don't Know? In *International Conference on Learning and Representation (ICLR)*.

Nalisnick, E.; Matsukawa, A.; Teh, Y. W.; and Lakshminarayanan, B. 2019b. Detecting Out-of-Distribution Inputs to Deep Generative Models Using a Test for Typicality. *arXiv preprint arXiv:1906.02994*.

Natural Highway Traffic Safety Administration. 2016. PE 16-007.

Pimentel, M. A.; Clifton, D. A.; Clifton, L.; and Tarassenko, L. 2014. A review of novelty detection. *Signal Processing* 99:215–249.

Rosca, M.; Lakshminarayanan, B.; and Mohamed, S. 2018. Distribution Matching in Variational Inference. *arXiv preprint arXiv:1802.06847*.

Theis, L.; Van Den Oord, A.; and Bethge, M. 2016. A Note on the Evaluation of Generative Models. In *International Conference on Learning and Representation (ICLR)*.

Tomczak, J. M., and Welling, M. 2017. VAE with a VampPrior. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

van den Oord, A.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural Discrete Representation Learning. In *Conference on Neural Information Processing Systems (NIPS)*.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv:1708.07747*.