Vacant Holes for Unsupervised Detection of the Outliers in Compact Latent Representation

Abstract

Detection of the outliers is pivotal for any machine learning model deployed and operated in real-world. It is essential for the Deep Neural Networks that were shown to be overconfident with such inputs. Moreover, even deep generative models that allow estimation of the probability density of the input fail in achieving this task. In this work, we concentrate on the specific type of these models: Variational Autoencoders (VAEs). First, we unveil a significant theoretical flaw in the assumption of the classical VAE model. Second, we enforce an accommodating topological property to the image of the deep neural mapping to the latent space: compactness to alleviate the flaw and obtain the means to provably bound the image within the determined limits by squeezing both inliers and outliers together. We enforce compactness using two approaches: (i) Alexandroff extension and (ii) fixed Lipschitz continuity constant on the mapping of the encoder of the VAEs. Finally and most importantly, we discover that the anomalous inputs predominantly tend to land on the vacant latent holes within the compact space, enabling their successful identification. For that reason, we introduce a specifically devised score for hole detection and evaluate the solution against several baseline benchmarks achieving promising results.

1 INTRODUCTION

Deep Generative Models (DGMs) allow for estimating the probability density of the input. This capability may appear tempting to utilize in the tasks of the detection of the outliers by casting all of the inputs that lie Out-of-Distribution (OoD) with the low density as anomalous. Nevertheless, empirical evidence shows that DGMs may sometimes be overconfident in their density estimation over OoDs [Nalisnick et al., 2018]. Overconfidence is observed in all types of DGMs, including autoregressive models [Oord et al., 2016], normalizing flows [Dinh et al., 2017], and VAEs [Kingma and Welling, 2013, Rezende et al., 2014]. This fact may appear especially intriguing, considering the difference in the techniques used for density estimation among these three distinct modeling approaches. However, from the theoretical perspective, there is nothing peculiar in such performance. It can be easily demonstrated that it is possible to learn an invertible reparametrization of the actual density of the data in a way that assigns an arbitrary density to each point in the new representation even in the models with perfect densities and in a low-dimensional setting [Lan and Dinh, 2020]. It means that the outlier detection is infeasible while relying only on the arbitrary learned probability density.

There are several alternative approaches aiming at tackling this issue that can be coarsely classified into one of the following categories: (i) methods that augment the input data by outliers [Hendrycks et al., 2018, Ren et al., 2019], (ii) ensemble-based methods [Daxberger and Hernández-Lobato, 2019, Glazunov and Zarras, 2022, Choi et al., 2019], (iii) methods that introduce new scores [Nalisnick et al., 2019, Serrà et al., 2019], (iv) methods based on the model modification [Hernández-Lobato et al., 2016, Schirrmeister et al., 2020], (v) and methods that involve retraining of the models [Xiao et al., 2020].

In this work, we refrain from augmenting data with outliers during training since it is not always feasible; we do not retrain the model to check every input as it is time-consuming, and due to the same reason, we do not apply ensemble-based methods. Instead, we utilize a model modification by introducing a new score. Specifically, we address the outlier detection from the perspective of general topology. Namely, we consider the property of compactness of the mapped image in the latent space. This property satisfies the necessary condition for the modeling assumption of a classical VAE from the viewpoint of the Universal Approximation Theorem (UAT) [Cybenko, 1989, Hornik, 1991, Pinkus, 1999].

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

First, we implement compactification using the Alexandroff extension of a flat subspace to a hypersphere. Second, we utilize a related topological property: bounded continuity. It equips us with two additional valuable tools. In particular, it lets enforce the Lipschitz-continuity constraints on the mappings used in the model. These constraints, in turn, permit both to establish the compactness of the mapped image and simultaneously control its boundaries in the case of the flat latent space. In addition, it helps to identify if the continuity holes in the latent prior play a significant role in the outlier detection during the ablation study.

Constraining the mapped image of the encoder may at first sound counterintuitive since the common choice of a prior over the latent is used to be the standard normal distribution with the infinite support that explicitly implies that outliers should be placed in some different location, distinctly separated from the inliers. It includes the low-dimensional cases where such inputs are placed in the tails far from the mode and the high-dimensional cases where the outliers are located outside the typical set. However, as we already indicated, there is no guarantee for such behavior even in perfect density models since any density function can be manipulated by an arbitrary choice of representation. Since there is no control over the mapped compact in the latent space, the choice of the bounds of the learned factors of variations of the VAE is basically arbitrary. In some situations, it can be the case that the outliers are indeed placed far from the inliers, which gives an excellent separation based only on the density values; however, in other situations, the outliers and inliers may overlap, which in some cases results in the overconfidence of the model. Hence the purposeful control over the compactness of the mapped image enforces the model to bind the learned factors of variations for *any* input within the predefined limits. If these limits are chosen in such a way that enforces the model to squeeze all of its inputs in the properly bounded space, then the model would have no other choice than to map the outliers somewhere *within* the same space that is used for the inliers in the latent representation. Experimental evidence shows that when the model is confronted with such tight condensing, it tends to place the outliers into the vacant latent continuity holes allowing their successful detection.

In summary, we make the following main contributions:

- We reveal the persistent theoretical flaw in the modeling assumption of VAEs.
- We mitigate this shortcoming by enforcing controlled compactness of the latent space.
- By bounding the image of the encoder, we discover that the outliers tend to gravitate toward the vacant latent holes and devise an appropriate score for their detection.
- We empirically evaluate the suggested approach based on several datasets.

2 BACKGROUND

2.1 NOTATION

We use nonbold x's to denote elements of general topological spaces, including the ones equipped with the appropriate metric. In the case of the normed vector spaces and random vectors within such spaces, we adhere to traditional usage in the literature, namely $\mathbf{x}$. When it comes to the particular elements comprising the random vector, we utilize x. The spaces are denoted as a pair $(\mathcal{X}, \mathcal{T})$ for topological spaces with the corresponding topology $\mathcal{T}$. In the specific case of metric spaces, we indicate the appropriate metric d that induces topology: $(\mathcal{X}, d_{\mathcal{X}})$.

2.2 VAES

VAE represents a DGM that allows to get an approximate value of the density of the input $\mathbf{x}$. It is based on the optimization of the evidence lower bound (ELBO), that provides joint optimization w.r.t variational parameters $\boldsymbol{\phi}$ of the encoder responsible for variational approximation of the posterior $q_{\boldsymbol{\phi}}$ over the latent variable $\mathbf{z}$, and the generative parameters $\boldsymbol{\theta}$ of the decoder responsible for the parameterization of the likelihood $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$:

$$\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\Big[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) - \log \frac{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{z})}\Big] \quad (1)$$

This equation involves a data likelihood term (used for generative purposes) and a regularization term (the KL divergence between the variational family $q_{\phi}(\mathbf{z}|\mathbf{x})$ and the prior distribution over the latent variables).

The final estimation of the marginal likelihood is done using importance sampling. Backpropagation through the random variable $\mathbf{z}$ is performed utilizing the standard reparameterization trick [Kingma and Welling, 2013].

2.3 COMPACTNESS

A topological space is compact or, equivalently, possesses a compactness property if every of its open cover has a finite subcover. In the case of the Eucledian spaces, the following specific result exists.

Theorem 2.1 (Heine-Borel)**.** *Let $K \subset \mathbb{R}^n$ then K is compact if and only if K is closed and bounded.*

Compactification is the process of turning a topological space into a compact one.

Definition 2.2. Let $(\mathcal{X}, \mathcal{T})$ be a topological space and let $(\mathcal{X}^*, \mathcal{T}^*)$ be a compact topological space s.t. $\mathcal{X}$ is homeomorphic to a dense subspace of $\mathcal{X}^*$. Then $(\mathcal{X}^*, \mathcal{T}^*)$ is called a compactification of $(\mathcal{X}, \mathcal{T})$. Thus, a compact space

$(\mathcal{X}^*, \mathcal{T}^*)$ is a compactification of a space $(\mathcal{X}, \mathcal{T})$ if and only if there exists a mapping f of $\mathcal{X}$ into $\mathcal{X}^*$ s.t. f is homeomorphism of $\mathcal{X}$ onto the subspace $f(\mathcal{X})$ of $\mathcal{X}^*$ and $f(\mathcal{X})$ is dense in $\mathcal{X}^*$.

An illustrative example of a frequently used compactification is an extension of $\mathbb{R}$ to $\mathbb{R} \cup \{-\infty, +\infty\}$.

Besides, there is a specific type of compactification by adjoining only one point: the Alexandroff extension.

Definition 2.3. Let $(\mathcal{X}, \mathcal{T})$ be a topological space and let ∞ be an object not belonging to $\mathcal{X}$. Let $\mathcal{X}^* = \mathcal{X} \cup \infty$ and let a topology $\mathcal{T}^*$ on $\mathcal{X}^*$ defined as follows: $\mathcal{T}^* = \mathcal{T} \cup \{V \subset \mathcal{T}^* : \infty \in V \text{and } \mathcal{X} \setminus V \text{ is closed and compact in } \mathcal{X}\}$. Then $(\mathcal{X}^*, \mathcal{T}^*)$ is the Alexandroff extension of $(\mathcal{X}, \mathcal{T})$.

An intuitive example of the Alexandroff extension is the inverse stereographic projection from the Euclidean plane to the sphere with the addition of a point at infinity.

2.4 LIPSCHITZ CONTINUITY

Definition 2.4. A map $f : \mathcal{X} \to \mathcal{Y}$, where $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ are metric spaces with the corresponding metrics $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$, is called Lipschitz continuous if for any $x_1, x_2 \in \mathcal{X}$, there exists a constant $M \in \mathbb{R}^+$ such that:

$$d_{\mathcal{Y}}(f(x_1), f(x_2)) \leq M d_{\mathcal{X}}(x_1, x_2) \tag{2}$$

M is called a Lipschitz constant. In this work, we refer to the Lipschitz constant as the smallest possible M. A mapping with such a constant is called an M-Lipschitz map. If not explicitly indicated otherwise, we let $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y} = \mathbb{R}^m$.

Recall that widely used activation functions such as sigmoid, tanh, and ReLU [Jarrett et al., 2009] are already generally scaled to be 1-Lipschitz. Hence, due to the composition property of the Lipschitz mappings, the first intuitive attempt to enforce the desirable Lipschitz property on the mapping would be to constrain the operator norm of the weights of each layer of the Deep Neural Network (DNN) [Yoshida and Miyato, 2017, Cisse et al., 2017]. However, it was proven that such an approach could not approximate even a simple absolute value function [Huster et al., 2018]. To tackle the issue, Anil et al. [2018] observed the critical component that influences the expressive power of any DNN, namely, the gradient-preserving property of its transformations. Therefore, they introduced the appropriate linear transformations and the 1-Lipschitz activation function, GroupSort, both of which are gradient preserving. They provably allow setting a Lipschitz constant on a DNN mapping. Moreover, DNNs utilizing them represent universal approximators of any Lipschitz mapping.

2.4.1 Latent Holes

Falorsi et al. [2018] introduced a score for detecting continuity holes in the latent space based on the ratio of the distances between two nearly located points in the input space and the distances of their corresponding latent codes:

$$\mathcal{F}_{Lip} = d_{\mathcal{Y}}(f(x_1), f(x_2)) / d_{\mathcal{X}}(x_1, x_2) \tag{3}$$

Xu et al. [2020] discovered that there exist vacant regions of low density in the aggregated posterior where prior assigns a relatively high density. They suggested detecting these regions by estimating the negative log-likelihood of the manipulated reference latent codes under the aggregated posterior:

$$\mathcal{F}_{Agg} = -\log p(\mathbf{z} \pm \epsilon) \tag{4}$$

where ϵ represents a magnitude of manipulation. It was demonstrated by Li et al. [2021] that both of these scores are connected despite the different motivation meaning that if the hole is detected by the score $\mathcal{F}_{Agg}$ then it will be also detected by $\mathcal{F}_{Lip}$.

2.5 UNIVERSAL APPROXIMATION THEOREM

The theoretical underpinning of DNNs is rooted in the results obtained in the approximation theory that is commonly referred to as the universal approximation theorem [Cybenko, 1989, Hornik, 1991, Pinkus, 1999].

Theorem 2.5. *Let $C(\mathcal{X}, \mathcal{Y})$ denote the set of all continuous mappings from $\mathcal{X}$ to $\mathcal{Y}$. Let $\sigma \in C(\mathbb{R}, \mathbb{R})$ represent an element-wise activation function. Then let $\mathcal{N}^{\sigma}_{n,m}$ represent the class of feedforward neural networks with activation function σ, with n neurons in the input layer, m neurons in the output layer, and one hidden layer with an arbitrary number of neurons. Let $K \subseteq \mathbb{R}^n$ be **compact**. Then σ is nonpolynomial if and only if $\mathcal{N}^{\sigma}_{n,m}$ is dense in $C(K, \mathcal{Y})$.*

The activation functions currently used in DNNs are nonpolynomial, so they fulfill the main requirement of the theorem. However, we deliberately emphasize that the results of the Universal Approximation Theorem (UAT) apply only in the cases when the input of the neural network is a compact set that is often overlooked.

3 RELATED WORK

New Scores-Based Methods. Nalisnick et al. [2019] conjecture that considering the high dimensionality of inputs, the over-confidence of DGMs may be because in-distribution images lie in the typical set as opposed to the tested OoDs that concentrate in the high-density region. They introduce the test for typicality that treats all input sequences as inliers if their entropy is close to the model's entropy.

Since the likelihood of generative models is biased by the complexity of the inputs, Serrà et al. [2019] propose to offset this bias by a factor that measures the input complexity and use the length of lossless compression of the image as the complexity factor, which is used to determine OoD. However, they do not evaluate their method on VAEs.

Ensemble-Based Methods. Choi et al. [2019] use an ensemble of independently trained DGMs that allow to get the density value and score them against the WAIC.

Bayesian DGMs. Although the BDGMs represent a single model, the Bayesian inference over model parameters allows building ensembles on the fly. The theoretical justification for the Bayesian VAE has been first laid out by Kingma and Welling [2013]. Several works are dedicated to OoD detection using Bayesian inference [Daxberger and Hernández-Lobato, 2019, Glazunov and Zarras, 2022]. They introduced new scores, such as the disagreement score and entropy. Both are based on the discrepancy between the models' estimations within the ensemble that achieved state-of-the-art results.

Lipschitz Continuity Methods. Several works utilize the Lipschitz continuity to improve the robustness of discriminative models against adversarial examples [Hein and Andriushchenko, 2017, Tsuzuku et al., 2018, Yang et al., 2020]. Barrett et al. [2022] apply the gradient-preserving transformations from Anil et al. [2018] in a similar to our approach manner. However, their main focus is to use Lipschitz mappings for certifiable robustness against adversarial examples.

4 METHODOLOGY

4.1 COMPACTNESS OF THE LEARNED LATENT REPRESENTATION

A usual assumption for VAE models is that the prior follows the standard normal distribution: $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$. It is a meaningful choice from the perspective of the generative process since it provides a clear and simple way of sampling. Moreover, it is a natural candidate for the ELBO objective's regularization term in learning a Gaussian posterior per each input of VAE. However, it additionally implies an infinite support of the latent prior. We show that such an assumption contradicts the UAT (Theorem 2.5).

Lemma 4.1. *Let $f : \mathcal{X} \to \mathcal{Y}$ be a continuous mapping from a topological space $\mathcal{X}$ to a topological space $\mathcal{Y}$. If $\mathcal{X}$ is compact, then its image $f[\mathcal{X}]$ is also compact (for proof, see Appendix A).*

Hence, by combining both Theorem 2.5 and Lemma 4.1 it follows that the image of any DNN trained on a compact set is also compact. This conclusion contradicts the infinite support assumption of the standard normal prior in the case of VAEs. Any DNN used as an encoder will map all inputs to the compact subset of the latent space.

In the case of in-distribution inputs, this conclusion may be considered subtle since all such inputs should be assigned the appropriate density under the model learned during the DNN training. However, it plays a significant role as soon as the model starts dealing with the OoD inputs. These are the different inputs that the model has not seen before and has not been able to generalize during training. Therefore, as it was demonstrated in Lan and Dinh [2020] the model is not constrained in putting those inputs anywhere within the whole available support or, more precisely, within the learned image of the encoder mapping. The properties of the compactness of the latent space become of great importance. One of the essential questions concerns the locations where the model tends to map the OoD inputs within the image compact space. As it was demonstrated by Nalisnick et al., the DGMs and VAEs, in particular, tend to be overconfident with OoD inputs. There were several attempts to explain this type of behaviour Nalisnick et al. [2019], Kirichenko et al. [2020], but none addressed the issue of compactness.

In this paper, we deliberately enforce the latent space's compactness. The reason for that is twofold. First, it should alleviate the contradiction above in the modeling assumption of the VAE by providing a principled way to set the compactness of the image of the learned mapping. Furthermore, the input support for the decoder also gets a compact space during training which is again in line with UAT. Second, it allows us to conduct experiments with the outliers' detection in the controlled environment with the desirable compactness properties so that all the holes will be located within the predefined boundaries.

In principle, this approach can be implemented utilizing the following two methods: (i) by Alexandroff extension and (ii) by setting a predefined Lipschitz constant of the encoder. The first method implies a change of the intrinsic curvature of the latent space by switching from a Euclidean to a non-Euclidean manifold. On the other hand, the second method allows keeping a flat latent space by only enforcing specific bounds on a mapped compact.

4.2 COMPACTIFICATION OF THE LATENT SPACE

4.2.1 Compactification of the Latent Space to the Hypersphere

The Alexandroff extension (Definition 2.3) of $\mathbb{R}^n$ can be done by adjoining a single point at infinity, turning the flat Euclidean space into a hypersphere $\mathcal{S}^n$ embedded into $\mathbb{R}^{n+1}$.

Lemma 4.2. *Let $\mathcal{S}^n := \{\mathbf{x} \in \mathbb{R}^{n+1} : ||\mathbf{x}|| = 1\}$ be a*

hypersphere with radius $r = 1$ centered at $\mathbf{0}$ and embedded in $\mathbb{R}^{n+1}$ then $\mathcal{S}^n$ is compact (for proof, see Appendix B).

The appropriate type of distribution that can be utilized on the hyperspherical surface is the von Mises-Fisher distribution which is parameterized by mean μ and concentration κ. If the concentration parameter κ is greater than zero, then the distribution has properties similar to normal; however, when $\kappa = 0$, then it is a uniform distribution. It allows choosing the uniform prior and calculating the corresponding KL-divergence term for the regularization within the latent space. We utilize the same algorithm introduced by Davidson et al. [2018] for the sampling and reparametrization trick. They named it a Hyperspherical VAE (HVAE).

4.2.2 Enforcing Compactness by Setting a Lipschitz Constant on the Encoder Mapping

Although the Alexandroff extension of the Euclidean space to the hypersphere is theoretically appealing, it has an issue with the surface area collapse, which makes it infeasible to use in high-dimensional settings (see Appendix C). To alleviate these issues, we implement our own method of ensuring the compactness of the latent codes. This method is beneficial since it keeps the flat Euclidean space for the latent representation and provides the necessary means to control the boundaries of the resulting compact.

Theorem 4.3. *Image of an M-Lipschitz mapping $f : \mathcal{X} \to \mathcal{Z}$ from a compact $K \subseteq \mathcal{X}$ with $x, y \in K$: $\|f(x) - f(y)\| \leq M \|x - y\|$ is bounded by both a corresponding Lipschitz constant M* ***and*** *by a radius R of a closed ball in the input support.*

Proof. By the Heine-Borel (Theorem 2.1), a compact $K \subset \mathbb{R}^n$ is closed and bounded, meaning that the set is contained in some closed ball with a finite radius R. Hence, for any $x, y \in K$:

$$\|x - y\| \leq R \tag{5}$$

Therefore, by combining the two inequalities above, we get:

$$\|f(x) - f(y)\| \leq MR \tag{6}$$

so the mapping f is bounded by the constant MR. □

Note that it is necessary to consider three components simultaneously to set a bound on the DNN output: bounds of the input compact, a norm being used, and, finally, an M-Lipschitz constant. In this work, we normalize the input support to the following compact vector space: $[0, 1]^n$. It conveniently allows constraining $R \leq 1$ by applying an L_∞-norm.

Moreover, to preserve both the generative functionality and the comparable log-likelihood values with the non-compact latent prior, it is important to consider the properties of the standard normal prior distribution. In the case of the low-dimensional setting, it is natural to bind the resulting compact with some standard deviation multiplier depending on the condensing tightness one wants to obtain. However, in the high-dimensional setting, the typical set should be considered. For that reason, the actual values for the Lipschitz constant of the encoder should be based on the dimensionality of the latent space. Namely, an upper bound on the mapped image should depend on the location of the typical set of the prior and its width. Recall that the center of the typical set of a centered normal distribution is located at the distance of $\sigma\sqrt{m}$ from the mode. In our experiments, we set the width equal to two standard deviations, and we choose the closest whole number:

$$M := \lfloor \sigma\sqrt{m} + 2\sigma \rceil \tag{7}$$

where m is the dimensionality of the latent representation and $\sigma = 1$ for the standard deviation of the prior.

In our work, we ensure the Lipschitz constant of the mapping utilizing the GroupSort activation function together with a projection of the weights of each layer on L_∞-ball during the forward-pass of the DNN. The constant is set layerwise in the following way: for a DNN with K number of layers in order to guarantee the M-lipschitz constant of the entire network mapping, we enforce the $M^{\frac{1}{K}}$ constant per each of its layer. It relies on the fact that the finite composition of Lipschitz functions is also Lipschitz with the product of the corresponding constants used in composition:

$$M = \prod_{n=1}^{K} M^{\frac{1}{K}} \tag{8}$$

The main building blocks are both 1-Lipschitz non-linearity and 1-Lipschitz linear mapping per each layer. The appropriate scaling of the results makes them equal to $M^{\frac{1}{K}}$. For the complete algorithm, see Appendix E.

4.3 SCORES

As we indicated before, currently available scores for the holes' detection are based either on the availability of the suitable metric in the input space [Falorsi et al., 2018] or on the computationally expensive estimation of the aggregated posterior based on all the training samples Xu et al. [2020]. The motivation for that was clearly because these scores are based on the inlier inputs; hence the search for the holes starts from their corresponding latent codes. However, in the case of outlier detection, we can directly check if the mapped input lands within the hole. For this purpose, we sample the approximated posterior $q_\phi(\mathbf{z}|\mathbf{x})$ with several latent codes $\mathbf{z}$ under a particular input $\mathbf{x}$ and compute the

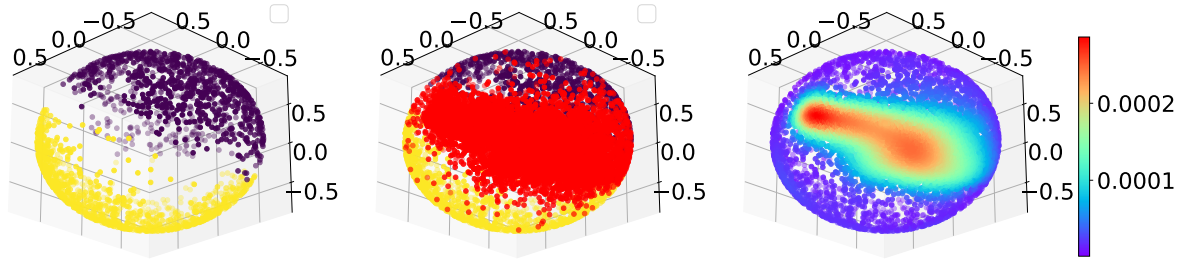

Figure 1: Compact S^2 latent space of VAE trained only on the two digits of the MNIST dataset 0's and 1's. The outliers densely land on the hole. **From left to right**: yellow depicts means of the estimated posteriors for 1's and purple for 0's; red represents the mapped means for a held-out outlier: a class of digits 9's; kernel density estimation of all means with the densest region in the hole packed with the outliers.

sample standard deviation of the log-likelihoods $\log p(\mathbf{x}|\mathbf{z})$.

$$\Sigma_{\mathbf{z}}[\mathbf{x}] = \sqrt{\frac{1}{N-1}\sum_{\mathbf{z}}\left(\log p(\mathbf{x}|\mathbf{z}) - \overline{\log p(\mathbf{x}|\mathbf{z})}\right)^2} \quad (9)$$

The approximated posterior under the input provides a locality within the latent space. Based on this locality—the samples from the posterior give us the notion of *nearness* around this specific locality. Finally, the standard deviation of the log-likelihoods based on the samples indicates *how far* from each other the sampled codes are mapped back into the input space.

There is also an alternative but still connected way of scoring the presence of a hole. Recall that density calculation of the given input under probability models with latent variables can be done through marginal likelihood. It is defined as the expected model likelihood marginalized over the latent's prior:

$$p(\mathbf{x}) = \mathbb{E}_{p(\mathbf{z})}[p(\mathbf{x}|\mathbf{z})] \quad (10)$$

First, let $\mathbf{z} \in \mathbb{L}$ and $|\mathbb{L}| < \infty$ then the marginal likelihood can be considered as a finite mixture of different $p(\mathbf{x}|\mathbf{z})$ with different constant weights $w = p(\mathbf{z})$ s.t. $\sum_{i=1}^{|\mathbb{L}|} w_i = 1$:

$$p(\mathbf{x}) = \sum_{i=1}^{|\mathbb{L}|} w_i p(\mathbf{x}|\mathbf{z}_i) \quad (11)$$

Suppose that $p(\mathbf{x}|\mathbf{z})$ is fully factorized, then the variance of the mixture of individual random components x's comprising $\mathbf{x}$ is given by:

$$\mathrm{Var}_{p(\mathbf{x})}(\mathrm{x}) = \underbrace{\sum_{i=1}^{|\mathbb{L}|} w_i \mathrm{Var}_{p(\mathbf{x}|\mathbf{z}_i)}[\mathrm{x}]}_{\text{Weighed individual variances}} + \underbrace{\sum_{i=1}^{|\mathbb{L}|} w_i \left(\mathbb{E}_{p(\mathbf{x}|\mathbf{z}_i)}[\mathrm{x}]\right)^2 - \left(\sum_{i=1}^{|\mathbb{L}|} w_i \mathbb{E}_{p(\mathbf{x}|\mathbf{z}_i)}[\mathrm{x}]\right)^2}_{\text{Jensen's gap}} \quad (12)$$

The first term is a weighted sum of variances of individual model likelihoods under all latent codes. Note that the difference of second and third terms is always non-negative due to Jensen's inequality. This difference represents a *Jensens's gap* and can be interpreted as the variance of the means of the likelihoods weighted by the appropriate prior probabilities of the latent. Hence, by computing the variance of the marginal likelihood under importance sampling due to this Jensen's gap, it is possible to estimate the variance of the means of the likelihoods, which can be utilized for hole detection with outlier inputs. For that reason, we apply the sample standard deviation of the estimated marginal likelihoods under importance sampling and test the performance of this score in our thorough experiments. To distinguish between the two scores, we label the first as the hole indicator and the second as the standard deviation of marginal log-likelihoods (Stds of LLs for short).

Threshold. For identifying the best threshold for the scores, we utilize threshold-independent metrics (these metrics are calculated for all possible thresholds) such as the *Area Under the Receiver Operating Characteristic Curve* (AUROC), the *Area Under Precision-Recall curve* (AUPR), and the *False-Positive Rate at 80% of True-Positive Rate* (FPR80) [Davis and Goadrich, 2006].

5 EVALUATION

5.1 TOY EXPERIMENTS WITH COMPACT $\mathcal{S}^2$

We begin with the simple held-out experiments based on the MNIST dataset [LeCun and Cortes, 2010]. For that reason, we utilize HVAE. [1] It is trained with the hyperspherical uniform prior on $\mathcal{S}^2$ only on two digits as inliers, namely zeros and ones. The rest of the handwritten digits are considered outliers. These experiments assist in acquiring a fundamental intuition in the way how the encoder of the model maps the outliers in the compact latent space. As

[1] For HVAE, we used the implementation available at `https://github.com/nicola-decao/s-vae-pytorch`

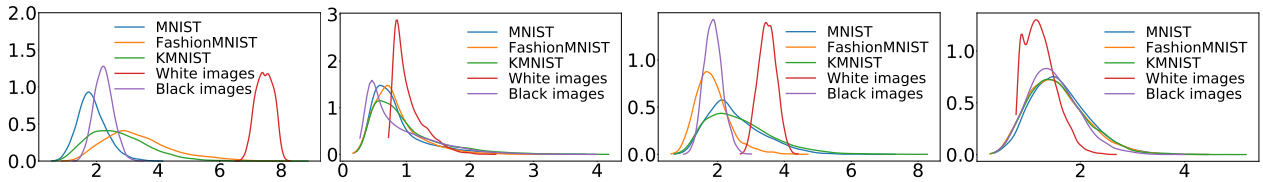

Figure 2: Estimated Gaussian kernel densities of L_∞-norms of the approximated posterior means in the latent space for datapoints from MNIST as inliers, and datapoints from FashionMNIST, KMNIST, white and black images as outliers. **From left to right**: classical VAE trained on MNIST; VAE with a fixed Lipschitz constant $M = 1$ for encoder trained on MNIST; classical VAE trained on Fashion-MNIST; VAE with a fixed Lipschitz constant $M = 1$ for encoder trained on Fashion-MNIST.

Table 1: Hole indicator scores for toy experiments with $\mathcal{S}^2$. The held-out outliers are all digits except 0's and 1's.

	MNIST held-out	MNIST vs. Fashion-MNIST
ROC AUC↑	89.38	94.67
AUPRC↑	99.40	95.38
FPR80↓	14.99	5.67

it can be observed in Figure 1, the two inlier classes are separated from each other on the sphere surface. There is also a hole between the clusters formed by these classes. Next, we try to map to the latent space held-out classes. As a result, we visually demonstrate that the encoder is forced to place the unseen during training classes somewhere within the constrained space and choose to land the outliers into latent holes. It happens when the model is confronted with the bounded factors of variation. In addition, we run experiments with our hole detection score $\Sigma_{\mathbf{z}}[\mathbf{x}]$ first with all held-out classes as outliers and second with Fashion-MNIST [Xiao et al., 2017] as outliers. The obtained results strongly support our hypothesis about holes as centers of attraction for the outliers (see Table 1).

5.2 EXPLORING COMPACTNESS ENFORCED BY LIPSCHITZ CONTINUITY

We continue probing compactness properties based on the constrained Lipschitz mapping to the latent space. We run experiments with both the classical VAE models and the VAE models with the enforced Lipschitz constant $M = 1$ for the encoder. We trained four separate models (for the used DNN architectures, see Appendix D): on MNIST and Fashion-MNIST, with and without continuity constraints—the dimensionality of the latent space across all models: $m = 10$. We evaluate the means of the approximated posteriors for the outliers from KMNIST [Clanuwat et al., 2018] (and analogously from Fashion-MNIST for the models trained on MNIST and vice versa). In addition, we run the same tests with the specially forged datasets. One contains non-realistic images, but all of their pixels tend to the black color; another contains images that tend to the white color. The idea behind the two latter datasets is that they represent extreme values of the compact support of the input data. As shown in Figure 2, the possible range of the values achievable by the classical VAE is considerably broad based on the limited number of the outlier datasets. For the model trained on MNIST, it goes as far as seven standard deviations from the mean.

Meanwhile, the unconstrained model trained on Fashion-MNIST has a range with a maximum of around four standard deviations. It demonstrates the arbitrariness of the mapped compact and its limits. Note, however, that when we bound the continuity of the encoder, then both inliers and outliers are squeezed together in a compact within the appropriate limits, which experimentally confirms our theoretical result (see Theorem 4.3). It allows the enforcement of a controlled and bounded compactness on the flat prior.

5.3 DETECTING OUTLIERS

As we indicated before, due to the surface collapse of the sphere, it is infeasible to use HVAE with high-dimensional priors. Hence, we apply the fixed Lipschitz mapping together with the appropriate input normalization (all inputs are normalized to $[0, 1]^n$). We evaluate our approach against several baseline methods. For them, we choose the classical VAE, the ensemble-based VAE, namely, the one based on the Bayesian inference over the weights of the DNN, and several approaches based on the new scores, namely, typicality score and input complexity. For scoring the Bayesian VAE, we utilize three available scores: WAIC, a disagreement score, and entropy. Bayesian inference is implemented utilizing the Bayes by Backpropagation Blundell et al. [2015]. The corresponding hyper-parameters and the training protocol are based on the work by Glazunov and Zarras [2022]. All models trained on MNIST and Fashion-MNIST have the dimensionality of the latent space equal to 10, and models trained on the CIFAR10 dataset have the latent of 70 dimensions. For our suggested Lipschitz-based model, we compute the appropriate Lipschitz constant for the decoder based on the dimensionality of the latent space in order to preserve the comparable log-likelihood values of the classical VAE and also to be able to sample the prior in

Table 2: Scoring values for the Lipschitz constrained VAEs trained on MNIST and Fashion-MNIST data.

	MNIST vs. Fashion-MNIST			Fashion-MNIST vs. MNIST			CIFAR10 vs. SVHN		
	ROC AUC↑	**AUPRC↑**	**FPR80↓**	**ROC AUC↑**	**AUPRC↑**	**FPR80↓**	**ROC AUC↑**	**AUPRC↑**	**FPR80↓**
					Vanilla VAE				
Log likelihood	99.99	99.99	0.00	54.03	57.37	84.70	61.08	53.92	56.25
Input complexity	0.00	32.91	100.00	99.17	99.24	0.00	95.87	95.36	9.09
Typicality test	100.00	100.00	0.00	53.81	50.78	70.74	59.75	64.06	80.20
					Bayesian VAE				
WAIC	99.99	99.99	0.00	59.53	59.35	71.88	61.15	54.22	57.15
Disagreement score	98.95	99.01	0.23	96.44	97.22	1.11	81.16	84.82	38.47
Entropy	99.42	99.47	0.02	97.97	98.43	0.19	84.76	88.21	29.31
					Lipschitz VAE				
Stds of LLs	99.78	99.79	0.06	99.21	99.16	0.84	86.40	84.88	21.59
Hole indicator (ours)	**99.87**	**99.87**	**0.00**	**99.69**	**99.65**	**0.28**	**91.76**	**89.58**	**12.30**

The most robust scores are in bold. The highest values are in gray.

* 0's in FPR80 are possible since it is a value for false-positive rate at 80% of true-positive rate

a standard way. For MNIST and Fashion-MNIST, it is equal to 5, and for CIFAR10, it is equal to 10. The results can be observed in Table 2. Our hole indicator demonstrates the best results among the scores that consistently perform well across all datasets. Moreover, the standard deviation of the likelihoods is the second most robust score, which agrees with our theoretical derivation (see Equation 12). Despite the high values for the typicality test on MNIST vs. Fashion-MNIST datasets and input complexity on CIFAR10 vs. SVHN datasets, they are not consistent across all of the datasets, which makes them unreliable in practical applications.

5.4 ABLATION STUDY

To check if the continuity holes are responsible for the obtained results, we conduct experiments with the gradual reduction of the holes in the latent space. This can be done by smoothing out the decoder mapping. This approach is advantageous since it affects all holes in the latent space. Hence, if our assumption is correct, then the results of the outlier detection based on the holes should degrade according to the strength of the smoothing. We enforce smoothing by setting the corresponding Lipschitz constants on the decoder mapping in the same way as it was done for the encoder in previous experiments. We train six separate models, all of which have the Lipschitz encoder with $M = 1$. Decoder is enforced with the values of the Lipschitz constants M from the following set: $\{1, 2, 3, 4, 5, 10\}$. As can be seen in Figure 3, there is an apparent performance degradation of the hole indicator for the outliers with decreasing of the corresponding Lipschitz constant enforced on the decoder, which is in line with our hypothesis that outliers land on the latent holes.

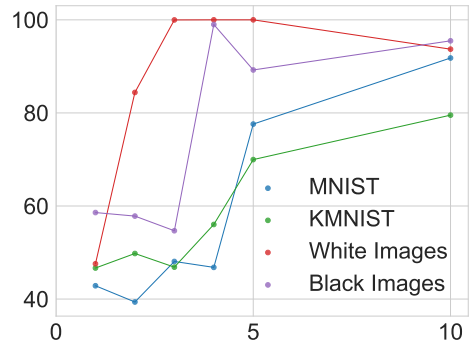

Figure 3: ROC AUC values from the ablation study. VAE with different Lipschitz constants enforced on the decoder, namely, $M = 1, M = 2, M = 3, M = 4, M = 5$ and $M = 10$, all plotted along x-axis. VAE is trained on Fashion-MNIST with the Encoder Lipschitz constant $M = 1$ for all tests and evaluated on several outlier datasets.

6 CONCLUSION

In this paper, we identified an implicit theoretical inconsistency from the perspective of general topology between the VAE modeling and the UAT. We addressed this discrepancy utilizing the compactness of the mapped image to the latent space. In order to enforce the compactness, we devised a provable method for controlling the bounds of the resulting compact. The experimental evidence revealed that constraining the limits of the factors of variation is beneficial for outlier detection. In particular, we discovered that outlier inputs tend to be mapped to the latent continuity holes. By devising a special score for the hole indicator, we conducted several experiments aimed at their detection. Utilizing this score, we achieved promising results in unsupervised outlier detection based on the latent representation. Specifically, the suggested method and score demonstrated the most robust performance across all the used benchmarks and datasets.

References

Cem Anil, James Lucas, and Roger Grosse. Sorting out lipschitz function approximation, 2018. URL https://arxiv.org/abs/1811.05381.

Ben Barrett, Alexander Camuto, Matthew Willetts, and Tom Rainforth. Certifiably robust variational autoencoders. In *International Conference on Artificial Intelligence and Statistics*, pages 3663–3683. PMLR, 2022.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015.

Hyunsun Choi, Eric Jang, and Alexander A. Alemi. Waic, but why? generative ensembles for robust anomaly detection, 2019.

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples, 2017. URL https://arxiv.org/abs/1704.08847.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Syst.*, 2(4):303–314, 1989.

Tim R Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M Tomczak. Hyperspherical variational auto-encoders. *arXiv preprint arXiv:1804.00891*, 2018.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

Erik Daxberger and José Miguel Hernández-Lobato. Bayesian variational autoencoders for unsupervised out-of-distribution detection, 2019.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HkpbnH9lx.

Luca Falorsi, Pim de Haan, Tim R Davidson, Nicola De Cao, Maurice Weiler, Patrick Forré, and Taco S Cohen. Explorations in homeomorphic variational auto-encoding. In *TADGM Workshop @ ICML*, 2018.

Misha Glazunov and Apostolis Zarras. Do bayesian variational autoencoders know what they don't know? In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022. URL https://openreview.net/forum?id=SSr4JOIs5l5.

Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2263–2273. Curran Associates, Inc., 2017.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure, 2018.

Daniel Hernández-Lobato, Thang D Bui, Yinzhen Li, José Miguel Hernández-Lobato, and Richard E Turner. Importance weighted autoencoders with random neural network parameters. In *Workshop on Bayesian Deep Learning, NIPS*, volume 2016, 2016.

K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Netw.*, 4(2):251–257, 1991.

Todd Huster, Cho-Yu Jason Chiang, and Ritu Chadha. Limitations of the lipschitz constant as a defense against adversarial examples, 2018. URL https://arxiv.org/abs/1807.09705.

Kevin Jarrett, Koray Kavukcuoglu, Marc'Aurelio Ranzato, and Yann LeCun. What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th International Conference on Computer Vision*, pages 2146–2153, 2009. doi: 10.1109/ICCV.2009.5459469.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Why normalizing flows fail to detect out-of-distribution data, 2020.

C. L. Lan and Laurent Dinh. Perfect density models cannot guarantee anomaly detection. *ArXiv*, abs/2012.03808, 2020.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. http://yann.lecun.com/exdb/mnist/, 2010.

Ruizhe Li, Xutan Peng, and Chenghua Lin. On the latent holes of vaes for text generation, 2021. URL https://arxiv.org/abs/2110.03318.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using typicality, 2019.

Aäron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with pixelcnn decoders. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4797–4805, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

A. Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numer.*, 8:143–195, 1999.

Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection, 2019.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

Robin Schirrmeister, Yuxuan Zhou, Tonio Ball, and Dan Zhang. Understanding anomaly detection with deep invertible networks through hierarchies of distributions and features. *ArXiv*, abs/2006.10848, 06 2020.

Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. *arXiv preprint arXiv:1909.11480*, 2019.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 6542–6551, Red Hook, NY, USA, 2018. Curran Associates Inc.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Zhisheng Xiao, Qing Yan, and Yali Amit. Likelihood regret: An out-of-distribution detection score for variational auto-encoder, 2020.

Peng Xu, Jackie Chi Kit Cheung, and Yanshuai Cao. On variational learning of controllable representations for text without supervision. In *International Conference on Machine Learning*, pages 10534–10543. PMLR, 2020.

Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Ruslan Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness, 2020. URL `https://arxiv.org/abs/2003.02460`.

Yuichi Yoshida and Takeru Miyato. Spectral norm regularization for improving the generalizability of deep learning, 2017. URL `https://arxiv.org/abs/1705.10941`.