Hierarchical VAEs Know What They Don't Know

Jakob D. Havtorn [1 2] **Jes Frellsen** [1] **Søren Hauberg** [1] **Lars Maaløe** [1 2]

Abstract

Deep generative models have been demonstrated as state-of-the-art density estimators. Yet, recent work has found that they often assign a higher likelihood to data from outside the training distribution. This seemingly paradoxical behavior has caused concerns over the quality of the attained density estimates. In the context of hierarchical variational autoencoders, we provide evidence to explain this behavior by out-of-distribution data having in-distribution low-level features. We argue that this is both expected and desirable behavior. With this insight in hand, we develop a fast, scalable and fully unsupervised likelihood-ratio score for OOD detection that requires data to be in-distribution across all feature-levels. We benchmark the method on a vast set of data and model combinations and achieve state-of-the-art results on out-of-distribution detection.

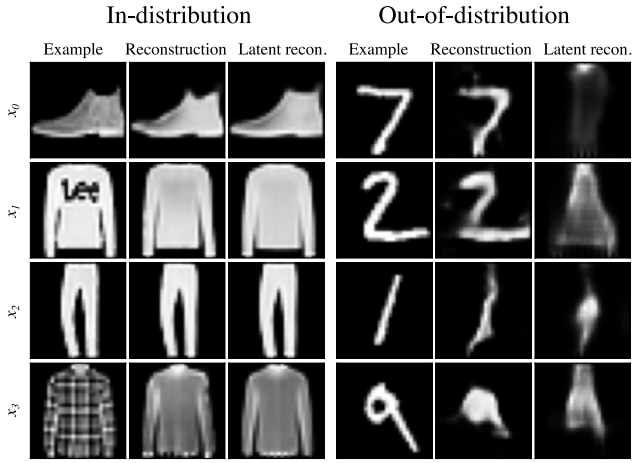


Figure 1. Reconstructions using a hierarchical VAE trained on FashionMNIST. Reconstruction quality of OOD data is comparable to in-distribution data, resulting in high likelihoods and poor OOD discrimination. By sampling the k bottom-most latent variables from the conditional prior distribution $p(\mathbf{z}_{\geq l}|\mathbf{z}_{>l})$ (latent reconstructions) instead of the approximate posterior $q(\mathbf{z}_{>l}|\mathbf{z}_{<l})$, the model reconstructs from the training distribution resulting in lower $p(\mathbf{x}|\mathbf{z})$ for OOD data.

1. Introduction

The reliability and safety of machine learning systems applied in the real-world is contingent on the ability to detect when an input is different from the training distribution. Supervised classifiers built as deep neural networks are well-known to misclassify such *out-of-distribution* (OOD) inputs to known classes with high confidence (Goodfellow et al., 2015; Nguyen et al., 2015). Several approaches have been suggested to equip deep classifiers with OOD detection capabilities (Hendrycks & Gimpel, 2017; Lakshminarayanan et al., 2017; Hendrycks et al., 2019; DeVries & Taylor, 2018). But, such methods are inherently supervised and require in-distribution labels or examples of OOD data limiting their applicability and generality.

Unsupervised generative models that estimate an explicit likelihood should understand what it means to be in- and out-of-distribution without requiring labels or examples of OOD data. By directly modeling the training distribution, such models are expected to assign low likelihoods to OOD data as it originates from regions of little or no support under the learned density (Bishop, 1994). Recent advances in deep generative models (Kingma & Welling, 2014; Rezende et al., 2014; Oord et al., 2016b; Salimans et al., 2017; Kingma & Dhariwal, 2018) have enabled learning high quality generative models on complex data such as natural images, sequences including audio (Oord et al., 2016a) and graphs (Kipf & Welling, 2016). However, recent observations have brought into question the quality of the learned density estimates by showing that they often assign higher likelihoods to OOD data than to in-distribution data (Nalisnick et al., 2019a; Choi et al., 2019). Many complex data distributions can be explained to a large degree by low-level features, e.g. edges in images. However, such features do not explain high-level semantics of the data and may inhibit OOD detection (Ren et al., 2019; Nalisnick et al., 2019a)

In this paper, we examine the failure cases of deep generative models on OOD detection tasks within the context of

[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kongens Lyngby, Denmark [2]Corti AI, Copenhagen, Denmark. Correspondence to: Jakob D. Havtorn <jdh@corti.ai>, Lars Maaløe <lm@corti.ai>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

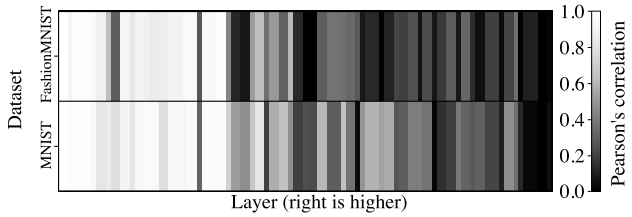


Figure 2. Absolute correlations between data representations in all layers of the inference network of a hierarchical VAE trained on FashionMNIST and of another trained on MNIST. We compute the correlation between the representations of the two different models given the same data, FashionMNIST (top) and MNIST (bottom).

Figure 3. Reconstructions of in-distribution data (CelebA) of the BIVA model using higher latent variables (Maaløe et al., 2019). The higher the latent variable, the more the reconstructions fall into the mode of the learned distribution. It is more common to wear regular glasses than sunglasses but most common not to wear glasses at all. A man with long hair collapses into the mode of the more common long-haired woman.

hierarchical VAEs, and make the following contributions:

(i) We provide evidence that the root cause of OOD failures is that learned low-level features generalize well across datasets and dominate the estimated likelihoods.

(ii) We then propose a fast, scalable, and fully unsupervised likelihood-ratio score for OOD detection that is explicitly developed to ensure that data should be in-distribution across all feature levels, which prevents the low-level features from dominating.

(iii) With the likelihood-ratio score, we demonstrate state-of-the-art performance across a wide range of known OOD failure cases.

2. Why does OOD detection fail?

The inability to detect out-of-distribution data with deep generative models is surprising. Before the advent of deep generative models, this was not considered a major issue for probabilistic models (Bishop, 1994). Is the failure due to model pathologies or something different?

Deep learning models are generally believed to form hierarchies of representations that range from low-level features to more conceptual ones related to semantics (Bengio et al., 2013). This has also been observed within deep generative models (Maaløe et al., 2019; Child, 2021). For image data there is a trend that the low-level features are quite similar across models (edge detectors, etc.). This raises the question to what extend such features are relevant when detecting OOD data, also suggested by (Nalisnick et al., 2019a). To investigate, we train two hierarchical VAEs (subsection 3.2) on FashionMNIST and MNIST, respectively, and compute the between-models correlation of the extracted features of in-distribution data and OOD data. The result appears in Figure 2. We observe that features extracted in the early layers (low-level features) correlate strongly between the two models, and that this correlation drops as we get into later layers. This suggests that low-level features do not carry much information for OOD detection.

To shed further light on the impact of semantic versus low-level features, we look at model reconstructions of images with a hierarchical VAE (Figure 3). To study the feature hierarchy, we replace the inference distribution with the corresponding conditional prior in the first layers of the model to see what information is lost. We observe that as more layers rely on the prior, more details are lost. Sunglasses, which are uncommon, are first replaced by more common glasses, and then finally disappear. This suggests that as we fall back to the conditional priors of each layer, we are pushed closer to local modes of the modeled distribution.

Finally, we look at reconstructions of out-of-distribution data. Figure 1 illustrates that MNIST data is surprisingly well reconstructed by a hierarchical VAE trained on FashionMNIST. Similar results have been found elsewhere (Xiao et al., 2020). We repeat the previous experiment and replace inference distributions by their corresponding conditional prior, and now observe that reconstructions from higher latent layers become increasingly similar to the data on which the model was trained. The reliance on conditional priors seems to prevent accurate reconstruction of out-of-distribution data. Some details are lost on in-distribution data too, but the distinction between that and out-of-distribution data becomes more clear.

These observations lead to our main hypothesis. We hypothesize that the lowest latent variables in a hierarchical VAE learn generic features that can be used to describe a wide range of data. This enables the model to achieve high rates of compression, and hence high likelihoods, even on out-of-distribution data as long as the learned low-level features are appropriate. We further suggest that OOD data are in-distribution with respect to these low-level features, but not with respect to semantic ones.

3. Background and related work

3.1. Variational autoencoders

The variational autoencoder (VAE) (Kingma & Welling, 2014; Rezende et al., 2014) is a framework for constructing

deep generative models defined by an observed variable $\mathbf{x}$ and a stochastic latent variable $\mathbf{z}$. Typically, a neural network with parameters θ is chosen to parameterize the generative distribution $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})$, where the prior $p(\mathbf{z})$ is commonly a standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The true posterior $p(\mathbf{z}|\mathbf{x})$ is generally not analytically tractable and is approximated by a variational distribution $q_\phi(\mathbf{z}|\mathbf{x})$ parameterized via another neural network with parameters ϕ. The approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ is most often a diagonal covariance Gaussian. The model parameters θ and variational parameters ϕ are jointly optimized by maximizing the *evidence lower bound* (ELBO),

$$\log p_\theta(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] \equiv \mathcal{L}(\mathbf{x}; \theta, \phi) \ . \quad (1)$$

For brevity, we will denote $\mathcal{L}(\mathbf{x}; \theta, \phi)$ as $\mathcal{L}(\mathbf{x})$ or $\mathcal{L}$. The reparameterization trick is used to backpropagate gradients through the stochastic latent variables with low variance.

The VAE is defined with a single latent variable which limits the ability to learn a high likelihood representation of complex input distributions, e.g. natural images. There exists a few complementary approaches to make the VAE more flexible: (i) model a more expressive variational distribution $q_\phi(\mathbf{z}|\mathbf{x})$ or prior distribution $p_\theta(\mathbf{z})$ (Rezende & Mohamed, 2015; Kingma et al., 2016), (ii) model a more expressive posterior distribution $p_\theta(\mathbf{x}|\mathbf{z})$ e.g. with a autoregressive decoder (van den Oord et al., 2016) and (iii) learn a deeper hierarchy of latent variables (Burda et al., 2016; Sønderby et al., 2016). Here, we focus on the latter.

3.2. Hierarchical variational autoencoders

Hierarchical VAEs are a family of probabilistic latent variable models which extends the basic VAE by introducing a hierarchy of L latent variables $\mathbf{z} = \mathbf{z}_1, \ldots, \mathbf{z}_L$. The most common generative model is defined from the top down as $p_\theta(\mathbf{x}|\mathbf{z}) = p(\mathbf{x}|\mathbf{z}_1)p_\theta(\mathbf{z}_1|\mathbf{z}_2) \cdots p_\theta(\mathbf{z}_{L-1}|\mathbf{z}_L)$. The inference model can then be defined in two ways respectively referred to as *bottom-up* (Burda et al., 2016)

$$q_\phi(\mathbf{z}|\mathbf{x}) = q_\phi(\mathbf{z}_1|\mathbf{x}) \prod_{i=2}^{L} q_\phi(\mathbf{z}_i|\mathbf{z}_{i-1}) \quad (2)$$

and *top-down* (Sønderby et al., 2016)

$$q_\phi(\mathbf{z}|\mathbf{x}) = q_\phi(\mathbf{z}_L|\mathbf{x}) \prod_{i=L-1}^{1} q_\phi(\mathbf{z}_i|\mathbf{z}_{i+1}) \ . \quad (3)$$

Regardless of the choice of inference model, a hierarchical VAE is still trained using the ELBO (1).

Until recently, hierarchical VAEs gave inferior likelihoods compared to state-of-the-art autoregressive (Ho et al., 2019) and flow-based models (Salimans et al., 2017). This was changed by Maaløe et al. (2019), Vahdat & Kautz (2020), and Child (2021), which introduced complementary methods to extend the number of latent variables to a very deep hierarchy resulting in state-of-the-art likelihood performance.

In this paper we employ a simple hierarchical VAE with bottom-up inference paths and the more powerful BIVA variant with a bidirectional (top-down and bottom-up) inference model (Maaløe et al., 2019). We employ skip connections between latent variables but omit them for brevity.

3.3. Out-of-distribution detection

So far, no reliable direct likelihood-based method has been found for fully unsupervised deep generative model OOD detection. A major line of work considers developing new scores that are more reliable than the likelihood. This includes the *typicality* test presented by Nalisnick et al. (2019b) which is an OOD detection test based on the typicality of a batch of potentially OOD examples. This approach however requires a batch of examples from the same class (OOD or not) which limits its practical applicability. In Ren et al. (2019), the *likelihood ratio* between a primary model and a background model was shown to be an effective score for OOD detection. However, to train the background model, the in-distribution data is perturbed via a data augmentation technique that is designed with knowledge about the confounding factors between the in-distribution data and the OOD data. Furthermore, it is tuned towards high performance on a known OOD dataset. Serrà et al. (2020) take a similar approach and attribute the failure to detect OOD data to the high influence of the input complexity on the likelihood and choose a generic lossless compression algorithm as the background model. Although this method gives good results, no single best choice of compression algorithm exists for all types of OOD data, and any particular choice encodes prior knowledge about the data into the detection method. Both these methods can be seen as correcting for low-level features of the OOD data being assigned high model likelihood by using a second model focused exclusively on these features.

Similar to these methods, the majority of the approaches to OOD detection make assumptions about the nature of the OOD data. The assumptions encompass using labels on the in-distribution data (Hendrycks & Gimpel, 2017; Liang et al., 2018; Alemi et al., 2018; Lee et al., 2018; Lakshminarayanan et al., 2017), examples of OOD data (Hendrycks et al., 2019), augmenting in-distribution data to mimic it (Ren et al., 2019), or assuming a certain data type (Serrà et al., 2020). Any of these assumptions encode implicit biases into the model about the attributes of OOD data which, in turn, might impair performance on truly unknown data examples (unknown unknowns).

While some of these methods achieve very good results on OOD detection with autoregressive models (Oord et al., 2016b; Salimans et al., 2017) and invertible flow-based models (Kingma & Dhariwal, 2018), it was recently shown that they can be much less effective for VAEs (Xiao et al.,

2020) highlighting the need for a more reliable OOD score for VAEs. Although VAEs have the same failure cases as autoregressive and flow-based models, the caveat is that the difference in the likelihood is generally not as big and reconstructions of OOD can be surprisingly good (Xiao et al., 2020). Xiao et al. (2020) alleviate this by refitting the inference network, as previously proposed by Cremer et al. (2018); Mattei & Frellsen (2018), to a potentially OOD example and measuring the so-called *likelihood regret*. However, refitting the inference network can be computationally expensive, especially for the large hierarchical VAEs that are used to model complex data (Maaløe et al., 2019; Vahdat & Kautz, 2020; Child, 2021). Furthermore, this scales poorly to large amounts of potentially OOD examples as the optimization is done per example.

A few methods have approached OOD detection in a completely unsupervised fashion (Maaløe et al., 2019; Choi et al., 2019; Xiao et al., 2020). The work of Maaløe et al. (2019) is the most related to ours. They introduce BIVA, a deep hierarchy of stochastic latent variables with a top-down and bottom-up inference model and achieve state-of-the-art likelihood scores. They also provide early results indicative that a looser likelihood bound may have value in OOD detection. In this paper, we provide an explanation of those results, and significantly improve upon them.

4. OOD detection with hierarchical VAEs

4.1. A bound for semantic OOD detection

If the lowest latent variable in the VAE hierarchy codes for a large part of the low-level features required to reconstruct the input with high accuracy, as exemplified in Figure 1-3, then $p_\theta(\mathbf{x}|\mathbf{z}_1)$ will be high for both in- and out-of-distribution data. Hence, any OOD detection capabilities based on the ELBO $\mathcal{L} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z}_1)] - D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$ from (1) relies on the KL-term for OOD detection. For a bottom-up hierarchical VAE, the KL-term $D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$ can be expressed by a hierarchical sum

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\sum_{i=1}^{L-1} \log \frac{p_\theta(\mathbf{z}_i|\mathbf{z}_{i+1})}{q_\phi(\mathbf{z}_i|\mathbf{z}_{i-1})} + \log \frac{p_\theta(\mathbf{z}_L)}{q_\phi(\mathbf{z}_L|\mathbf{z}_{L-1})}\right] . \quad (4)$$

In general, the absolute log-ratios grow with $\dim(\mathbf{z}_i)$ as the individual log probability terms are computed by summing over the dimensionality of $\mathbf{z}_i$. This means that the value of the KL-term is dominated by terms where $\mathbf{z}_i$ is high-dimensional. We refer to Appendix C for a more detailed argument. Since hierarchical VAEs are generally constructed with a bottleneck type structure, the terms corresponding to latent variables towards the top of the hierarchy will have a vanishing influence on the value of the KL-term. However, as the semantic information most relevant for OOD detection has a tendency to be represented in the top-most latent variables, this makes OOD detection using the regular ELBO difficult, even for state-of-the-art models. This behavior has also been reported by Xiao et al. (2020).

To shift the ELBO from primarily being based on the approximate posterior of the lowest latent variables to instead focus on the conditional prior, Maaløe et al. (2019) introduced slightly different likelihood lower bound defined as

$$\mathcal{L}^{>k} = \mathbb{E}_{p_\theta(\mathbf{z}_{\leq k}|\mathbf{z}_{>k})q_\phi(\mathbf{z}_{>k}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z}_{>k})}{q_\phi(\mathbf{z}_{>k}|\mathbf{x})}\right] \quad (5)$$

where $k \in \{0, 1, \ldots, L\}$ (see Appendix for the derivation). We note that $\mathcal{L}^{>0}$ is the regular ELBO (1) and that empirically we always observe that $\mathcal{L} \geq \mathcal{L}^{>k} \,\forall\, k$ although this need not hold in general. The core idea behind this variation on the ELBO is to sample the k lowest latent variables from the conditional prior $\mathbf{z}_1, \ldots, \mathbf{z}_l \sim p_\theta(\mathbf{z}_{\leq k}|\mathbf{z}_{>k})$ and only the $L-k$ highest from the approximate posterior $\mathbf{z}_{k+1}, \ldots, \mathbf{z}_L \sim q_\phi(\mathbf{z}_{>k}|\mathbf{x})$. Importantly, this has the effect that the data likelihood $p(\mathbf{x}|\mathbf{z})$ is dependent on the approximate posterior through a latent variable $\mathbf{z}_{k+1}$ different from $\mathbf{z}_1$ for all $k \geq 1$. Thereby, the likelihood can be evaluated with a reconstruction from each of the latent variables $\mathbf{z}_k$ of the hierarchical VAE. Hence, we can now test how well the input $\mathbf{x}$ is reconstructed from each latent variable. The notation $\mathcal{L}^{>k}$ highlights that for latent variables $\mathbf{z}_{>k}$, the bound is the regular ELBO while for the latent variables $\mathbf{z}_{\leq k}$, the bound is evaluated using the (conditional) prior rather than the approximate posterior as the proposal distribution.

4.2. A likelihood-ratio score for all feature levels

While the $\mathcal{L}^{>k}$ bound provides a score for performing semantic OOD detection, it still relies on the data space likelihood function (see equation (7) below), which is known to be problematic for OOD detection (section 3.3). To alleviate this, we phrase OOD detection as a likelihood ratio test of being *semantically* in-distribution. A standard likelihood ratio test (Buse, 1982) suggests to consider the ratio between the associated likelihoods, which we can approximate on a log-scale by the corresponding lower bounds $\mathcal{L}$ and $\mathcal{L}^{>k}$,

$$LLR^{>k}(\mathbf{x}) = \mathcal{L}(\mathbf{x}) - \mathcal{L}^{>k}(\mathbf{x}) . \quad (6)$$

Since, empirically, $\mathcal{L} \geq \mathcal{L}^{>k}$, the ratio is always positive as is standard for likelihood ratio tests. A low value of $LLR^{>k}(\mathbf{x})$ means that the ELBO and $\mathcal{L}^{>k}$ are almost equally tight for the data. On the contrary, a high value indicates that $\mathcal{L}^{>k}$ is looser on the data than the ELBO; hence, the data may be OOD.

We can gather further insights about this score if we write the regular ELBO and the $\mathcal{L}^{>k}$ bounds in the exact form that includes the intractable KL-divergence between the approximate and true posteriors,

$$\begin{aligned} \mathcal{L} &= \log p_\theta(\mathbf{x}) - D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})\right) , \quad (7) \\ \mathcal{L}^{>k} &= \log p_\theta(\mathbf{x}) - D_{\mathrm{KL}}\left(p_\theta(\mathbf{z}_{\leq k}|\mathbf{z}_{>k})q_\phi(\mathbf{z}_{>k}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})\right) . \end{aligned}$$

Subtracting these cancel out the two data likelihood terms $\log p_\theta(\mathbf{x})$ and only the KL-divergences from the approximate to the true posterior remain,

$$LLR^{>k}(\mathbf{x}) = -D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})\right) + D_{\mathrm{KL}}\left(p_\theta(\mathbf{z}_{\leq k}|\mathbf{z}_{>k})q_\phi(\mathbf{z}_{>k}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})\right) . \quad (8)$$

Hence, it is clear that compared to the likelihood bound $\mathcal{L}^{>k}$, this likelihood-ratio measures divergence exclusively in the latent space whereas $\mathcal{L}^{>k}$ includes the $\log p_\theta(\mathbf{x})$ term similar to the ELBO. Therefore, the $LLR^{>k}$ score should be an improved method for semantic OOD detection compared to $\mathcal{L}^{>k}$. Now, it can be noted that if we replace the regular ELBO, $\mathcal{L}$, in (7) with the strictly tighter importance weighted bound (Burda et al., 2016),

$$\mathcal{L}_S = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\log \frac{1}{N}\sum_{s=1}^{S}\frac{p(\mathbf{x},\mathbf{z}^{(s)})}{q(\mathbf{z}^{(s)}|\mathbf{x})}\right] , \quad (9)$$

then, in the limit $S \to \infty$, we have $\mathcal{L}_S \to \log p_\theta(\mathbf{x})$ and the likelihood ratio reduces to

$$LLR_S^{>k}(\mathbf{x}) \to D_{\mathrm{KL}}(p(\mathbf{z}_{\leq k}|\mathbf{z}_{>k})q(\mathbf{z}_{>k}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})) \quad (10)$$

which, in practice, is well-approximated for a finite S. We expect this importance weighted likelihood ratio to monotonically improve upon the one in (8) as S increases and the KL-divergence in the regular ELBO that contains terms for which $\mathbf{z}_i$ is high-dimensional goes to zero.

Since the scores in (8) and (10) are estimated by sampling their estimators are stochastic objects with nonzero variance. We note that $\mathrm{Var}(\widehat{LLR}^{>k}) = \mathrm{Var}(\hat{\mathcal{L}}) + \mathrm{Var}(\hat{\mathcal{L}}^{>k}) - 2\,\mathrm{Cov}(\hat{\mathcal{L}}, \hat{\mathcal{L}}^{>k})$. Since $\log p_\theta(\mathbf{x})$ and part of the KL divergence are identical in the expressions of $\mathcal{L}$ and $\mathcal{L}^{>k}$ we expect $\mathrm{Cov}(\hat{\mathcal{L}}, \hat{\mathcal{L}}^{>k})$ to be positive which reduces the total variance. Empirical results indeed show that $\mathrm{Var}(\widehat{LLR}^{>k})$ is larger than $\mathrm{Var}(\hat{\mathcal{L}})$ but smaller than $\mathrm{Var}(\hat{\mathcal{L}}^{>k})$. Nevertheless, the variance of the estimators is guaranteed to go to zero as the number of samples is increased.

The OOD scores considered in this research all assume that what discriminates an out-of-distribution from an in-distribution data point are semantic, high-level features. Clearly, if this is not the case and the difference instead lies in low-level statistics, the scores would likely fail. We hypothesize that a complementary bound to (5), $\mathcal{L}^{<l}$ described in Appendix E, might be useful in these cases, but leave further examination to future work.

5. Experimental setup

Tasks: We follow existing literature (Nalisnick et al., 2019a; Hendrycks et al., 2019) and evaluate our method by setting up OOD detection tasks from FashionMNIST (Xiao et al., 2017) to MNIST (LeCun et al., 1998) and from CIFAR10 (Krizhevsky, 2009) to SVHN (Netzer et al., 2011). For each experiment we train our model on the train split of the former dataset and test its ability to recognize the test split of the latter dataset as OOD from the test split of the former dataset. We use the standard train/test splits for the datasets. More details on the datasets can be found in the Appendix.

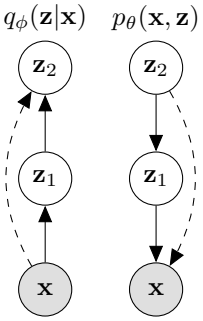


Figure 4. The inference and generative models, q_ϕ and p_θ, for an $L = 2$ layered bottom-up hierarchical VAE as the one used in our experiments. Dashed lines indicate deterministic skip connections which are employed in both networks. Skip connections are found to be useful for optimizing latent variable models (Dieng et al., 2019; Maaløe et al., 2019).

Models: For each OOD task, we train a simple bottom-up hierarchical VAE with L stochastic layers which we will refer to as "HVAE". To alleviate posterior collapse we include skip-connections that connect $\mathbf{z}_i$ to $\mathbf{z}_{i+2}$ for $i \in \{0, L-2\}$ and $\mathbf{z}_0 \equiv \mathbf{x}$ in both the inference and generative models (Dieng et al., 2019) and employ the *free bits* scheme with $\lambda = 2$ (Kingma et al., 2016). We use weight-normalization (Salimans & Kingma, 2016) on all weights and residual networks in the deterministic paths. A graphical representation of this model can be seen in Figure 4. We use a Bernoulli output distribution for FashionMNIST/MNIST and a discretized mixture of logistics output distribution (Salimans et al., 2017) for CIFAR10/SVHN. We use $L = 3$ for grey-scale images and $L = 4$ for natural images. For CIFAR/SVHN, we also train a BIVA model (Maaløe et al., 2019) with $L = 10$ and similar configuration as used by the original paper[1]. All models are trained by optimizing the ELBO in (1). We implement our models in PyTorch (Paszke et al., 2017)[2]. Full model details are in the Appendix.

Baselines: We group baselines into those that use prior knowledge about OOD data, ones that use labels associated with the in-distribution data and purely unsupervised approaches that do not make such assumptions. Our method falls into the latter category. For more information on each baseline, we refer to the original literature.

Evaluation: Following previous work (Hendrycks & Gimpel, 2017; Hendrycks et al., 2019; Alemi et al., 2018; Ren et al., 2019; Choi et al., 2019) we use the threshold-independent evaluation metrics of Area Under the Receiver

[1]Source code available at `github.com/larsmaaloee/BIVA` and `github.com/vlievin/biva-pytorch`

[2]Source code available at `github.com/jakobhavtorn/hvae-oodd`

Operator Characteristic (AUROC↑), Area Under the Precision Recall Curve (AUPRC↑) and False Positive Rate at 80% true positive rate (FPR80↓) where the arrow indicates the direction of improvement. Note that these metrics are only computable given examples of OOD data but faced with truly OOD data (unknown unknowns), there are many ways to select thresholds to use in practice e.g. as the one that yields a specific tolerable false positive rate on the in-distribution test data. To compute the metrics, we use an equal number of samples from the in-distribution and OOD datasets by including all examples in the smallest of the two sets and randomly sampling equally many from the larger. We compute the $LLR^{>k}$ score with one and S importance samples denoted by $LLR_S^{>k}$.

Selection of k: To determine whether an example is OOD in practice, the value of $LLR^{>k}$ is computed on the in-distribution test set for all k and the resulting empirical distribution is used as reference. If for any value of k, the $LLR^{>k}$ score of a new input differs significantly from the empirical distribution, it is regarded OOD. If it differs for multiple values of k, the value for which it differs the most is selected. In our experiments, we consider an entire dataset at a time and report the results of $LLR^{>k}$ with the value of k that yielded the highest AUROC↑ for that dataset in a threshold-free manner. In practice, slightly better performance may be achieved by choosing k per example. This would not exclude the use of batching in our method, since $LLR^{>k}$ is computed after the forward pass.

6. Results

The likelihoods for our trained models are in Table 1 alongside baseline results for in-distribution and OOD data. The main results of the paper on the OOD tasks can be seen along with comparisons to the baseline methods in Table 2. We note that for all our results, the value of the score ($\mathcal{L}^{>k}$ and $LLR^{>k}$) for the training and test splits of the in-distribution data was observed to have the same empirical distribution to within sampling error hence yielding an AUROC score of ≈ 0.5 as expected. Results on additional commonly used datasets are found in Appendix G.

6.1. Likelihood-based OOD detection

We first report the results of the different variations of the $\mathcal{L}^{>k}$ bound for OOD detection. We reconfirm the results of Nalisnick et al. (2019a) by observing that our hierarchical latent variable models also assign higher $\mathcal{L}^{>0}$ to the OOD dataset in the FashionMNIST/MNIST and CIFAR10/SVHN cases resulting in an AUROC↑ inferior to random (Table 2).

[2]Serrà et al. (2020) performs the best when high likelihoods are assigned to OOD data such that the overlap with in-distribution data is low. Performance is worse when the overlap is high, cf. Serrà et al. (2020, Table 1), as seen with complex images.

Method	Dataset	Avg. bits/dim			
		$\log p(x)$	$\mathcal{L}^{>1}$	$\mathcal{L}^{>2}$	$\mathcal{L}^{>3}$
Trained on FashionMNIST					
Glow	FashionMNIST	2.96	-	-	
	MNIST	1.83	-	-	
HVAE (Ours)	FashionMNIST	0.420	0.476	0.579	-
	MNIST	0.317	0.601	0.881	-
Trained on CIFAR10					
Glow	CIFAR10	3.46	-	-	
	SVHN	2.39	-	-	
HVAE (Ours)	CIFAR10	3.74	17.8	54.3	75.7
	SVHN	2.62	10.2	64.0	93.9
BIVA (Ours)	CIFAR10	3.46	8.74	19.7	37.3
	SVHN	2.35	6.62	25.1	59.0

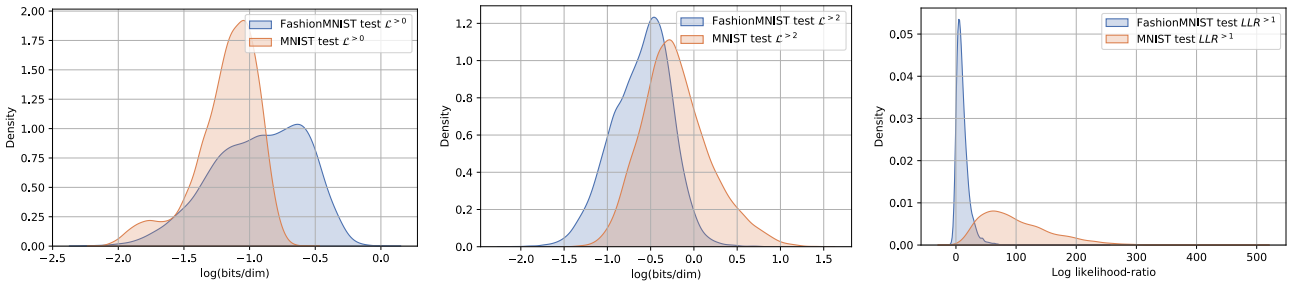
Table 1. Average bits per dimension of different datasets for models trained on FashionMNIST and CIFAR10. For the hierarchical models we include the $\mathcal{L}^{>k}$ bounds. The likelihoods of training and test splits of the in-distribution data are all cases close. Since we train on dynamically binarized FashionMNIST, our bits/dim are smaller than for Glow. As k is increased for the $L^{>k}$ bound, the bound gets looser but the model eventually assigns higher likelihood to the in distribution data than to the OOD data. Glow refers to Kingma & Dhariwal (2018); Nalisnick et al. (2019a). BIVA refers to our implementation of Maaløe et al. (2019).

Switching the in-distribution data for the OOD data in both cases result in correctly detecting the OOD data; an asymmetry also reported by Nalisnick et al. (2019a). Figure 5a shows the density of $\mathcal{L}^{>0}$ in bits per dimension (Theis et al., 2016) by the model trained on FashionMNIST when evaluated on the FashionMNIST and MNIST test sets. We observe a high degree of overlap, with less separation of the OOD data compared to similar results of autoregressive and flow-based models, like Xiao et al. (2020).

We then evaluate the looser $\mathcal{L}^{>k}$ (5) for $k \in \{1, L\}$. Figure 5b shows the result for $\mathcal{L}^{>2}$, which yielded the highest AUCROC↑, only slightly better than random. Like Maaløe et al. (2019), we see that increasing the value of k generally leads to improved OOD detection. However, we also observe that the two empirical distributions never cease to overlap. Importantly, depending on the OOD dataset, the amount of remaining overlap can be high which limits the discriminatory power of the likelihood-based $\mathcal{L}^{>k}$ bound. This is in-line with the pathological behavior of the raw likelihood of latent variable models when used for OOD detection (Xiao et al., 2020). Since a high degree of overlap also seems present in Maaløe et al. (2019), and we see the same problem for our BIVA model trained on CIFAR10, we do not expect this to be due to the less expressive HVAE.

6.2. Likelihood-ratio-based OOD detection

We now move to the likelihood ratio-based score. We find that $LLR^{>k}$ separates the OOD MNIST data from in-distribution FashionMNIST to a higher degree than the



(a) (b) (c)

Figure 5. Empirical densities of FashionMNIST (in-distribution) and MNIST (OOD) using the raw likelihood (a), the $\mathcal{L}^{>2}$ bound (b) and the $LLR^{>1}$ score (c). All densities are computed using the HVAE model.). For the regular likelihood MNIST is very clearly more likely on average than the FashionMNIST test data while with the $\mathcal{L}^{>2}$ bound separation is better but significant overlap remains. The $LLR^{>1}$ provides a high degree of separation. Likelihoods are reported in units of the natural log of the number of bits per dimension.

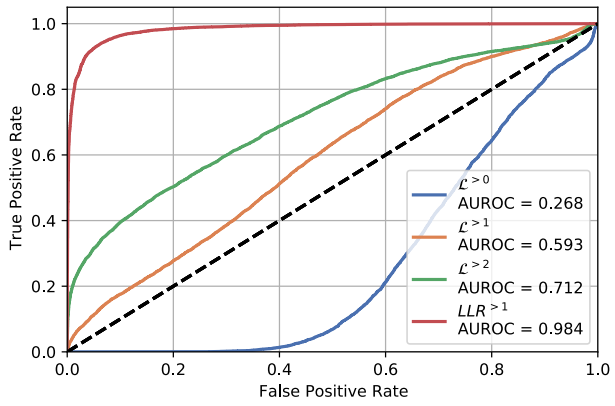


Figure 6. ROC curves with AUROC score for detecting MNIST as OOD with the HVAE model trained on FashionMNIST. A ROC curve is plotted for each of the $\mathcal{L}^{>k}$ bounds including the ELBO along with one for the best-performing log likelihood-ratio $LLR^{>1}$.

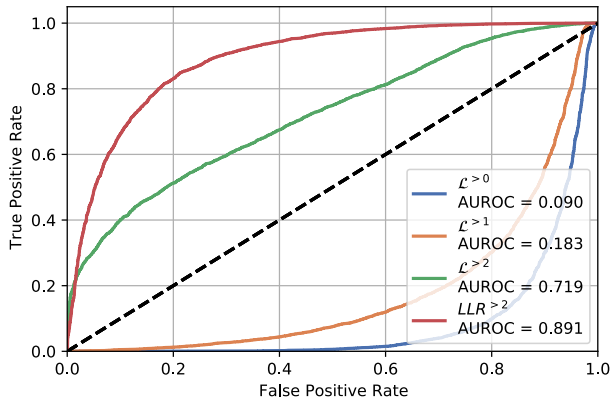


Figure 7. ROC curves with AUROC score for detecting SVHN as OOD with the BIVA model trained on CIFAR10. A ROC curve is plotted for each of the $\mathcal{L}^{>k}$ bounds including the ELBO along with one for the best-performing log likelihood-ratio $LLR^{>2}$.

likelihood estimates as can be seen by the empirical densities of the score in Figure 5c. We note that the likelihood ratio between the ELBO and the $\mathcal{L}^{>k}$ bound provides the highest degree of separation of MNIST and FashionMNIST as measured by the AUROC↑ for $k = 1$ smaller than L. This is not surprising since the value of k that provides the maximal separation to the reference in-distribution dataset need not be the one for which $\mathcal{LR}^{>k}$ is overall maximal for the OOD dataset. We also visualize the ROC curves resulting from using the $LLR^{>k}$ score for OOD detection on both FashionMNIST/MNIST and CIFAR10/SVHN and compare it to the ROC curves resulting from the different $\mathcal{L}^{>k}$ bounds in Figures 6 and 7, respectively. On both datasets we see significantly better discriminatory performance when using the $LLR^{>k}$ score.

Table 2 shows that BIVA improves upon the HVAE model for OOD detection on CIFAR while Table 1 shows that the BIVA model also improves upon the HVAE in terms of likelihood. We hypothesize that models larger than our implementation of BIVA, with better likelihood scores may perform even better (Maaløe et al., 2019; Vahdat & Kautz, 2020; Child, 2021).

6.3. Comparison to baselines

Performance: Table 2 summarize our results compared to baselines based on the commonly used AUROC↑, AUPRC↑ and FPR80↓ metrics. Our method outperforms other generative model-based methods such as WAIC (Choi et al., 2019) with Glow model and performs similarly to the likelihood regret method of (Xiao et al., 2020). Furthermore, our method performs similarly to the background constrative likelihood ratio method of Ren et al. (2019) on FashionMNIST/MNIST but contrary to the failure of that method on CIFAR10/SVHN reported by (Xiao et al., 2020), our method performs very well on this task too. Our approach outperforms all supervised approaches that use in-distribution la-

Method	AUROC↑	AUPRC↑	FPR80↓
FashionMNIST (in) / MNIST (out)			
Use prior knowledge of OOD			
Backgr. contrast. LR (PixelCNN) [1]	0.994	0.993	0.001
Backgr. contrast. LR (VAE) [7]	0.924	-	-
Binary classifier [1]	0.455	0.505	0.886
$p(\hat{y}\|\mathbf{x})$ with OOD as noise class [1]	0.877	0.871	0.195
$p(\hat{y}\|\mathbf{x})$ with calibration on OOD [1]	0.904	0.895	0.139
Input complexity (S, Glow) [9]	0.998	-	-
Input complexity (S, PixelCNN++) [9]	0.967	-	-
Use in-distribution data labels y			
$p(\hat{y}\|\mathbf{x})$ [1], [2]	0.734	0.702	0.506
Entropy of $p(y\|\mathbf{x})$ [1]	0.746	0.726	0.448
ODIN [1, 3]	0.752	0.763	0.432
VIB [4, 7]	0.941	-	-
Mahalanobis distance, CNN [1]	0.942	0.928	0.088
Mahalanobis distance, DenseNet [5]	0.986	-	-
Ensemble, 20 classifiers [1, 6]	0.857	0.849	0.240
No OOD-specific assumptions			
- *Ensembles*			
WAIC, 5 models, VAE [7]	0.766	-	-
WAIC, 5 models, PixelCNN [1]	0.221	0.401	0.911
- *Not ensembles*			
Likelihood regret [8]	**0.988**	-	-
$\mathcal{L}^{>0}$ + HVAE (ours)	0.268	0.363	0.882
$\mathcal{L}^{>1}$ + HVAE (ours)	0.593	0.591	0.658
$\mathcal{L}^{>2}$ + HVAE (ours)	0.712	0.750	0.548
$LLR^{>1}$ + HVAE (ours)	0.964	0.961	0.036
$LLR^{\geq 1}_{250}$ + HVAE (ours)	0.984	**0.984**	**0.013**
CIFAR10 (in) / SVHN (out)			
Use prior knowledge of OOD			
Backgr. contrast. LR (PixelCNN) [1]	0.930	0.881	0.066
Backgr. contrast. LR (VAE) [8]	0.265	-	-
Outlier exposure [9]	0.984	-	-
Input complexity (S, Glow) [10]	0.950	-	-
Input complexity (S, PixelCNN++) [10]	0.929	-	-
Input complexity (S, HVAE) (Ours) [10][3]	0.833	0.855	0.344
Use in-distribution data labels y			
Mahalanobis distance [5]	0.991	-	-
No OOD-specific assumptions			
- *Ensembles*			
WAIC, 5 models, Glow [7]	1.000	-	-
WAIC, 5 models, PixelCNN [1]	0.628	0.616	0.657
- *Not ensembles*			
Likelihood regret [8]	0.875	-	-
$LLR^{>2}$ + HVAE (ours)	0.811	0.837	0.394
$LLR^{>2}$ + BIVA (ours)	**0.891**	**0.875**	**0.172**

Table 2. AUROC↑, AUPRC↑ and FPR80↓ for OOD detection for a FashionMNIST model using scores on the FashionMNIST test set as reference. We bold the best results within the "No OOD-specific assumptions" group since we only compare directly to those. HVAE (ours) refers to our hierarchical bottom-up VAE. BIVA (ours) refers to our implementation of the hierarchical BIVA model (Maaløe et al., 2019). [1] is (Ren et al., 2019), [2] is (Hendrycks & Gimpel, 2017), [3] is (Liang et al., 2018), [4] is (Alemi et al., 2018), [5] is (Lee et al., 2018), [6] is (Lakshminarayanan et al., 2017), [7] is (Choi et al., 2019), [8] is (Xiao et al., 2020), [9] is (Hendrycks et al., 2019), [10] is (Serrà et al., 2020).

bels or synthetic examples of OOD data derived from the in-distribution data including ODIN (Liang et al., 2018) and the predictive distribution of a classifier $p(\hat{y}|\mathbf{x})$ trained and evaluated in various ways (see Ren et al. (2019)).

Runtime: For a full evaluation of a single example across all feature levels of a model with L stochastic layers, our method requires $L-1$ forward passes through the inference and generative networks as well as computing the likelihood ratio, of which the forward passes are dominant. For a typical forward pass that is linear in the input dimensionality, D, and the number of stochastic layers, L, this amounts to computation of $\mathcal{O}(DL)$. Compared to some related work that either requires an $M > 1$ sized batch of inputs of which either all or none are OOD (Nalisnick et al., 2019b) or cannot be applied to batches due to the required per-example optimization (Xiao et al., 2020), our method additionally is applicable to batches of any size that may consist of both OOD and in-distribution examples which provides drastic speed-ups via vectorization and parallelization. Furthermore, the method of Xiao et al. (2017) requires refitting the inference network of a VAE which can be computationally demanding. Compared to the likelihood ratio proposed in Ren et al. (2019), our method requires training only a single model on a single dataset.

7. Discussion

Deep generative models are state-of-the-art density estimators, but the OOD failures reported in recent years have raised concerns about the limitations of such density estimates. Recent work on improving OOD detection has largely sidestepped this concern by relying on additional assumptions that strictly should not be needed for models with explicit likelihoods. While the engineering challenge of building reliable OOD detection schemes is important, it is of more fundamental importance to understand *why* the naive likelihood test fails. We have provided evidence that low-level features of the neural nets dominate the likelihood, which gives a *cause* to the *why*. The fact that a simple score for measuring the importance of semantic features yield state-of-the-art results on OOD detection without access to additional information gives validity to our hypothesis.

The findings from, amongst others, Nalisnick et al. (2019a); Serrà et al. (2020) have a clear relation to information theory and compression. Semantically complex in-distribution data yields models with diverse low-level feature sets that enable generalization across datasets. Simpler datasets can only yield models with less diverse low-level feature sets compared to complex training data. Hence, there can be an asymmetry where the likelihoods of simple OOD data can be high for a model trained on complex data, but not the other way around. Loosely put, the minimal number of bits required to losslessly compress data sampled from some distribution is the entropy of the generating process (Shannon, 1948; MacKay, 2003). Townsend et al. (2019) recently showed that VAEs can be used for lossless compression at rates superior to more generic algorithms.

We also note that since the hierarchical VAE is a probabilistic graphical latent variable model, it lends itself very

naturally to manipulation at the feature level (Kingma et al., 2014; Maaløe et al., 2016; 2017). This property sets it apart from other generative models that do not explicitly define such a hierarchy of features. This in turn enables reliable OOD detection with our methodology while making no explicit assumptions about the nature of OOD data and only using a single model. This has not been achieved with autoregressive or flow-based models.

8. Conclusion

In this paper we study unsupervised out-of-distribution detection using hierarchical variational autoencoders. We provide evidence that highly generalizeable low-level features contribute greatly to estimated likelihoods resulting in poor OOD detection performance. We proceed to develop a likelihood-ratio based score for OOD detection and define it to explicitly ensure that data must be in-distribution across all feature levels to be regarded in-distribution. This ratio is mathematically shown to perform OOD detection in the latent space of the model, removing the reliance on the troublesome input-space likelihood. We point out that contrary to much recent literature on OOD detection, our approach is fully unsupervised and does not make assumptions about the nature of OOD data. Finally, we demonstrate state-of-the-art performance on a wide range of OOD failure cases.

Acknowledgements

This research was partially funded by the Innovation Fund Denmark via the Industrial PhD Programme (grant no. 0153-00167B). JF and SH were funded in part by the Novo Nordisk Foundation (grant no. NNF20OC0062606) via the Center for Basic Machine Learning Research in Life Science (MLLS, https://www.mlls.dk). JF was further funded by the Novo Nordisk Foundation (grant no. NNF20OC0065611) and the Independent Research Fund Denmark (grant no. 9131-00082B). SH was further funded by VILLUM FONDEN (15334) and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 757360).

References

Alemi, A. A., Fischer, I., and Dillon, J. V. Uncertainty in the Variational Information Bottleneck. July 2018. URL http://arxiv.org/abs/1807.00906. arxiv: 1807.00906.

Bengio, Y., Courville, A. C., and Vincent, P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8): 1798–1828, 2013. doi: 10.1109/TPAMI.2013.50. URL https://doi.org/10.1109/TPAMI.2013.50.

Bishop, C. M. Novelty Detection and Neural-Network Validation. *IEE Proceedings - Vision, Image and Signal Processing*, 141(4):217–222, 1994. ISSN 1350245x, 13597108. doi: 10.1049/ip-vis:19941330.

Burda, Y., Grosse, R., and Salakhutdinov, R. R. Importance Weighted Autoencoders. In *Proceedings of the 4th International Conference on Learning Representations (ICLR)*, pp. 8, San Juan, Puerto Rico, 2016. URL https://arxiv.org/abs/1509.00519.

Buse, A. The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. *The American Statistician*, 36(3a):153–157, 1982.

Child, R. Very Deep VAEs Generalize Autoregressive Models and Can Outperform Them on Images. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/pdf/2011.10650.pdf.

Choi, H., Jang, E., and Alemi, A. A. WAIC, but Why? Generative Ensembles for Robust Anomaly Detection. May 2019. URL http://arxiv.org/abs/1810.01392. arxiv: 1810.01392.

Cremer, C., Li, X., and Duvenaud, D. Inference Suboptimality in Variational Autoencoders. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of machine learning research*, pp. 1078–1086, Stockholmsmässan, Stockholm, Sweden, July 2018. PMLR. URL http://proceedings.mlr.press/v80/cremer18a.html.

DeVries, T. and Taylor, G. W. Learning Confidence for Out-of-Distribution Detection in Neural Networks. February 2018. URL http://arxiv.org/abs/1802.04865. arxiv: 1802.04865.

Dieng, A. B., Kim, Y., Rush, A. M., and Blei, D. M. Avoiding latent variable collapse with generative skip models. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89, pp. 2397–2405, Naha, Okinawa, Japan, 2019. PMLR. URL http://proceedings.mlr.press/v89/dieng19a.html.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In Bengio, Y. and LeCun, Y. (eds.), *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015. URL http://arxiv.org/abs/1412.6572.

Hendrycks, D. and Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *Proceedings of the 5th International Conference on Learning Representations (ICRL)*, Toulon, France, 2017. URL http://arxiv.org/abs/1610.02136.

Hendrycks, D., Mazeika, M., and Dietterich, T. G. Deep anomaly detection with outlier exposure. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA, 2019. URL https://openreview.net/forum?id=HyxCxhRcY7.

Ho, J., Chen, X., Srinivas, A., Duan, Y., and Abbeel, P. Flow++: Improving Flow-Based Generative Models with Variational Dequantization and Architecture Design. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 9, Long Beach, CA, USA, 2019. URL http://proceedings.mlr.press/v97/ho19a/ho19a.pdf.

Kingma, D. P. and Dhariwal, P. Glow: Generative Flow with Invertible 1×1 Convolutions. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*, pp. 10, Montréal, Canada, 2018.

Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, Banff, AB, Canada, 2014. URL http://arxiv.org/abs/1312.6114. arXiv: 1312.6114.

Kingma, D. P., Rezende, D. J., Mohamed, S., and Welling, M. Semi-Supervised Learning with Deep Generative Models. In *Proceedings of the 28th International Conference on Neural Information Processing Systems (NeurIPS)*, Montréal, Quebec, Canada, June 2014. URL http://arxiv.org/abs/1406.5298. arXiv: 1406.5298.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improved variational inference with inverse autoregressive flow. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS)*, NIPS'16, pp. 4743–4751, Barcelona, Spain, 2016. ISBN 978-1-5108-3881-9. URL http://arxiv.org/abs/1606.04934.

Kipf, T. N. and Welling, M. Variational Graph Auto-Encoders. November 2016. URL http://arxiv.org/abs/1611.07308. arxiv: 1611.07308.

Krizhevsky, A. *Learning Multiple Layers of Features from Tiny Images*. PhD thesis, University of Toronto, 2009. arXiv: 1011.1669v3 ISBN: 9788578110796 ISSN: 1098-6596.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *In Proceddings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, 2017. URL http://arxiv.org/abs/1612.01474.

LeCun, Y. A., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2323, 1998.

Lee, K., Lee, K., Lee, H., and Shin, J. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*, pp. 11, Montréal, Quebec, Canada, 2018. URL https://papers.nips.cc/paper/2018/file/abdeb6f575ac5c6676b747bca8d09cc2-Paper.pdf.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018. URL https://openreview.net/forum?id=H1VGkIxRZ.

Maaløe, L., Sønderby, C. K., Sønderby, S. K., and Winther, O. Auxiliary deep generative models. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48 of *Proceedings of machine learning research*, pp. 1445–1453, New York, New York, USA, June 2016. PMLR. URL http://proceedings.mlr.press/v48/maaloe16.html.

Maaløe, L., Fraccaro, M., and Winther, O. Semi-Supervised Generation with Cluster-aware Generative Models. April 2017. URL http://arxiv.org/abs/1704.00637. arxiv: 1704.00637.

Maaløe, L., Fraccaro, M., Liévin, V., and Winther, O. BIVA: A Very Deep Hierarchy of Latent Variables for Generative Modeling. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*, pp. 6548–6558, Vancouver, Canada, February 2019. URL http://arxiv.org/abs/1902.02102.

MacKay, D. J. C. *Information theory, inference, and learning algorithms*. Cambridge University Press, 1 edition, 2003. ISBN 978-0-521-64298-9.

Mattei, P.-A. and Frellsen, J. Refit your encoder when new data comes by. In *3rd NeurIPS Workshop on Bayesian Deep Learning*, 2018.

Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., and Lakshminarayanan, B. Do Deep Generative Models Know What They Don't Know? In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA, 2019a. URL http://arxiv.org/abs/1810.09136. arXiv: 1810.09136.

Nalisnick, E., Matsukawa, A., Teh, Y. W., and Lakshminarayanan, B. Detecting Out-of-Distribution Inputs to Deep Generative Models Using Typicality. pp. 15, 2019b. URL https://arxiv.org/abs/1906.02994. arxiv: 1906.02994.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading Digits in Natural Images with Unsupervised Feature Learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. URL http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

Nguyen, A., Yosinski, J., and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CCVPR)*, volume 07-12-June, pp. 427–436, 2015. ISBN 978-1-4673-6964-0. doi: 10.1109/CVPR.2015.7298640.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. WaveNet: A Generative Model for Raw Audio. In *In Proceedings of the 9th ISCA Speech Synthesis Workshop*, Sunnyval, CA, USA, September 2016a. URL http://arxiv.org/abs/1609.03499.

Oord, A. v. d., Kalchbrenner, N., and Kavukcuoglu, K. Pixel Recurrent Neural Networks. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, New York, NY, USA, August 2016b. Journal of Machine Learning. URL http://arxiv.org/abs/1601.06759.

Paszke, A., Chanan, G., Lin, Z., Gross, S., Yang, E., Antiga, L., and Devito, Z. Automatic differentiation in PyTorch. In *In Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*, 2017. URL https://pytorch.org/.

Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., Depristo, M., Dillon, J., and Lakshminarayanan, B. Likelihood Ratios for Out-of-Distribution Detection. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 12, Vancouver, Canada, 2019. URL https://papers.nips.cc/paper/2019/file/1e79596878b2320cac26dd792a6c51c9-Paper.pdf.

Rezende, D. J. and Mohamed, S. Variational Inference with Normalizing Flows. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, Lille, France, 2015. URL http://arxiv.org/abs/1505.05770.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *Proceedings of Machine Learning Research*, volume 32, pp. 1278–1286, Beijing, China, January 2014. PMLR. URL http://proceedings.mlr.press/v32/rezende14.pdf.

Salimans, T. and Kingma, D. P. Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks. In *Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS)*, Barcelona, Spain, February 2016. URL http://arxiv.org/abs/1602.07868.

Salimans, T., Karpathy, A., Chen, X., and Kingma, D. P. PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, Toulon, France, April 2017. URL http://arxiv.org/abs/1701.05517.

Serrà, J., Álvarez, D., Gómez, V., Slizovskaia, O., Núñez, J. F., and Luque, J. Input complexity and out-of-distribution detection with likelihood-based generative models. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, 2020. URL https://openreview.net/forum?id=SyxIWpVYvr.

Shannon, C. E. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(July 1948):379–423, 1948. ISSN 07246811. doi: 10.1145/584091.584093. arXiv: chao-dyn/9411012 ISBN: 0252725484.

Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., and Winther, O. Ladder Variational Autoencoders. In *Proceedings of the 29th Conference on Neural Information Processing Systems (NeurIPS)*, Barcelona, Spain, December 2016. URL http://arxiv.org/abs/1602.02282.

Theis, L., Oord, A. v. d., and Bethge, M. A note on the evaluation of generative models. In *Proceedings of the 4th International Conference on Learning Representations (ICLR)*, San Juan, Puerto Rico, May 2016. URL http://arxiv.org/abs/1511.01844.

Townsend, J., Bird, T., and Barber, D. Practical Lossless Compression With Latent Variables Using Bits Back Coding. In *7th International Conference on Learning Representations (ICLR)*, pp. 13, New Orleans, LA, USA, 2019.

Vahdat, A. and Kautz, J. NVAE: A Deep Hierarchical Variational Autoencoder. In *34th Conference on Neural Information Processing Systems (NeurIPS)*, Virtual, July 2020. URL http://arxiv.org/abs/2007.03898.

van den Oord, A., Kalchbrenner, N., Espeholt, L., Kavukcuoglu, K., Vinyals, O., and Graves, A. Conditional image generation with PixelCNN decoders. In *Proceedings of the 29th International Conference on Neural Information Processing Systems*, pp. 4790–4798, Barcelona, Spain, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/b1301141feffabac455e1f90a7de2054-Abstract.html.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. 2017. URL https://arxiv.org/abs/1708.07747. arXiv:1708.07747 [cs.LG].

Xiao, Z., Yan, Q., and Amit, Y. Likelihood Regret: An Out-of-Distribution Detection Score for Variational Auto-Encoder. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, Virtual, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/eddea82ad2755b24c4e168c5fc2ebd40-Abstract.html.