NEGATIVE SAMPLING IN VARIATIONAL AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose negative sampling as an approach to improve the notoriously bad out-of-distribution likelihood estimates of Variational Autoencoder models. Our model pushes latent images of negative samples away from the prior. When the source of negative samples is an auxiliary dataset, such a model can vastly improve on baselines when evaluated on OOD detection tasks. Perhaps more surprisingly, we present a fully unsupervised variant that can also significantly improve detection performance: using the output of the generator as negative samples results in a fully unsupervised model that can be interpreted as adversarially trained.

1 INTRODUCTION

Learning semantically meaningful and useful representations for downstream tasks in an unsupervised manner is a big promise of generative modeling. While a plethora of work demonstrates the effectiveness of deep generative models in this regard, recent work of Nalisnick et al. (2018) and Choi et al. (2018) show that these models often fail even at a task that is supposed to be close to their original goal of learning densities. Variational Autoencoders, PixelCNN and flow-based models cannot distinguish common objects like cats and dogs from house numbers. That is, when trained e.g., on CIFAR-10, the models consistently assign higher likelihoods for the elements of the SVHN test set than for the elements of the CIFAR-10 test set or even the elements of the CIFAR-10 train set. As generative models are becoming more and more ubiquitous due to the massive progress in this area in recent years, it is of fundamental importance to understand these phenomena.

In this work we study Variational Autoencoder (VAE) models, and besides the likelihood, we also investigate to what extent the latent representation of a data point can be used to identify out-of-distribution (OOD) samples (points that are not from the true data distribution learned by the model). In particular, we utilize the KL divergence between the prior and the posterior distribution of a data point as a score to distinguish inliers and outliers.

Our contributions are summarized as follows:

- We demonstrate empirically that the extent of this notorious phenomenon — of bad out-of-distribution likelihood estimates — present in VAEs largely depends on the observation model of the VAE. In particular, our experiments show that it diminishes when a Gaussian noise model is considered (with a reasonably sized fixed or learned variance) instead of a Bernoulli. Meanwhile, when examining only the KL divergence between the prior and the posterior distributions in the latent space (instead of the full likelihood), the weak separating capability between inliers and outliers more consistently prevails.
- We propose *negative sampling in Variational Autoencoders* as an approach to alleviate the above weaknesses of the model family. In this method, we introduce an additional prior distribution $\bar{p}(\boldsymbol{z})$ in the latent space, where the representations of negative samples are meant to be mapped by the inference model of the VAE machinery. Negative samples can be obtained from an auxiliary dataset, or — to remain completely in the unsupervised setting — from an adversarial training scheme using generated images as negative samples.
- We present empirical evidence that utilizing negative samples either from an auxiliary dataset or from an adversarial training scheme significantly and consistently improves the discriminative power of VAE models regarding out-of-distribution samples.

The general intuition behind our approach is that if the posterior distribution of each and every point is pulled towards the prior then it is rather natural to expect that the system will map out-of-distribution samples close to the prior, as well. This viewpoint suggests that providing negative signals throughout the learning process would be beneficial to enhance the OOD discriminative power of the system.

Hendrycks et al. (2018) demonstrates that utilizing auxiliary datasets as OOD examples (as a supervised signal) significantly improves the performance of existing anomaly detection models on image and text data. First, we study how this approach can be employed in the VAE setting. Beyond that, we also propose a method which remains completely in the unsupervised learning paradigm (without using an auxiliary dataset for supervised signal). The core idea of this unsupervised approach is to provide near-manifold negative samples throughout the training process for which the model is explicitly encouraged to give low likelihood estimates. The near-manifold negative samples are obtained from the generative model itself by utilizing the generated samples.

2 Background

The generative modeling task aims to model a ground truth data density $p^*(\boldsymbol{x})$ on a space $\mathcal{X}$ by learning to generate samples from the corresponding distribution. The learning is done in an unsupervised manner with sampled observables $\mathbf{X} = \{\mathbf{x}^{(i)}\}_{i=1}^N$ as training points assumed to be drawn independently from $p^*(\boldsymbol{x})$, where N is the sample size. In latent variable models the observables are modeled together with hidden variables $\boldsymbol{z}$ on which a prior distribution $p(\boldsymbol{z})$ is imposed.

The Variational Autoencoder (VAE) (Kingma & Welling, 2013) is a latent variable model that takes the maximum likelihood approach and maximizes a lower bound of the sample data log likelihood $\sum_{i=1}^N \log p_\theta(\mathbf{x}^{(i)})$, where θ are the model parameters. The utilized lower bound $\mathcal{L}(\theta, \phi, \mathbf{x}^{(i)})$ (called the ELBO) comes from a variational approximation $q_\phi(\boldsymbol{z}|\mathbf{x}^{(i)})$ of the intractable posterior $p_\theta(\boldsymbol{z}|\mathbf{x}^{(i)})$, where ϕ are the variational parameters:

$$\log p_\theta(\mathbf{x}^{(i)}) = \log \int p_\theta(\mathbf{x}^{(i)}|\boldsymbol{z})p(\boldsymbol{z}) \geq$$

$$\geq \mathbb{E}_{q_\phi(\boldsymbol{z}|\mathbf{x}^{(i)})} \log p_\theta(\mathbf{x}^{(i)}|\boldsymbol{z}) - D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\mathbf{x}^{(i)}) \parallel p(\boldsymbol{z})) \triangleq \mathcal{L}(\theta, \phi, \mathbf{x}^{(i)}).$$

In the VAE model the parametrized distributions p_θ and q_ϕ are modeled with neural networks and are trained jointly to maximize $\mathcal{L}$ with some variant of the SGD. The prior is often chosen to be the multivariate standard normal distribution, and a Bernoulli or Gaussian noise model is considered in the observable space to define the likelihood. Throughout our paper we follow the convention of minimizing the negative log likelihood, so all of our loss terms are meant to be minimized.

To give likelihood estimates for unseen data points at test time, one can use the trained inference model $q_\phi(\boldsymbol{z}|\mathbf{x}^{(i)})$ and generative model $p_\theta(\mathbf{x}^{(i)}|\boldsymbol{z})$ to estimate the ELBO, thus to give a lower bound of the likelihood. Throughout our paper, we are considering these ELBO estimates to measure the likelihood of data points.

3 Negative sampling in Variational Autoencoders

To incorporate negative samples in the VAE training process, we introduce an additional prior distribution $\bar{p}(\boldsymbol{z})$ for the *negative samples* on the latent variables $\boldsymbol{z}$ into which the representations of negative samples $\overline{\mathbf{X}} = \{\bar{\mathbf{x}}^{(i)}\}_{i=1}^M$ are meant to be mapped by the inference model. This is encouraged in the training process by adding a new loss term to the regular ELBO which pulls the KL divergence of the posterior distributions of negative samples to this negative prior. The joint loss function thus is as follows:

$$L = -\mathcal{L}(\theta, \phi, \mathbf{x}^{(i)}) + D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\bar{\mathbf{x}}^{(i)}) \parallel \bar{p}(\boldsymbol{z})).$$

To motivate this extra loss term, we now compare our model with a simple variational model that can work both as a generator and as a classifier between the $\mathbf{X}$ and $\overline{\mathbf{X}}$ distributions. This graphical model has an extra observable besides $\boldsymbol{x}$, the latent variable y, which is a Bernoulli random variable with $p = 1/2$, $y = 0$ meaning a choice from $\mathbf{X}$ and $y = 1$ meaning a choice from $\overline{\mathbf{X}}$, giving rise to the joint density function $p(\boldsymbol{x}, y)$. Let $p(\boldsymbol{z}|y)$ be a normal distribution with parameters depending on y. In our graphical model, y is screened from $\boldsymbol{x}$ by $\boldsymbol{z}$, that is, $p_\theta(\boldsymbol{x}|\boldsymbol{z}) = p_\theta(\boldsymbol{x}|\boldsymbol{z}; y)$. Similarly, our variational posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is chosen to be independent from y. Writing up the log-likelihood:

$$\mathbb{E}_{p(\boldsymbol{x},y)} \log p_\theta(\boldsymbol{x}, y) = p(y = 0)\mathbb{E}_{p(\boldsymbol{x}|y=0)} \log p_\theta(\boldsymbol{x}, 0) + p(y = 1)\mathbb{E}_{p(\boldsymbol{x}|y=1)} \log p_\theta(\boldsymbol{x}, 1) =$$
$$= \frac{1}{2}(\mathbb{E}_{p_\theta(\boldsymbol{x})} \log p_\theta(\boldsymbol{x}) + \mathbb{E}_{\bar{p}_\theta(\bar{\mathbf{x}})} \log \bar{p}_\theta(\bar{\mathbf{x}})),$$

where $\bar{p}$ is the density function of the negative samples. Sampling $\mathbf{x}^{(i)}$ and $\bar{\mathbf{x}}^{(i)}$ from the positive and negative samples respectively, and writing up the ELBO for both of the terms:

$$\log p_\theta(\mathbf{x}^{(i)}) + \log \bar{p}_\theta(\bar{\mathbf{x}}^{(i)}) \geq$$
$$\geq \mathbb{E}_{q_\phi(\boldsymbol{z}|\mathbf{x}^{(i)})} \log p_\theta(\mathbf{x}^{(i)}|\boldsymbol{z}) - D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\mathbf{x}^{(i)}) \parallel p(\boldsymbol{z}|y = 0))+$$
$$+\mathbb{E}_{q_\phi(\boldsymbol{z}|\bar{\mathbf{x}}^{(i)})} \log p_\theta(\bar{\mathbf{x}}^{(i)}|\boldsymbol{z}) - D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\bar{\mathbf{x}}^{(i)}) \parallel p(\boldsymbol{z}|y = 1)).$$

Note that while the encoder itself is unaware of the y label, the whole maximum likelihood model is aware of it, via the conditional prior $p(\boldsymbol{z}|y)$. Technically, the generator is also unaware of the y label, but in our experiments we choose priors with such a small overlap in support between the positive and negative priors that $\boldsymbol{z}$ "leaks" all information about y. The small overlap in support, in effect, enforces the encoder to operate as a classifier.

The above graphical model is symmetric with respect to the roles of $\mathbf{X}$ and $\overline{\mathbf{X}}$. Our loss formula deviates from it by omitting the reconstruction loss term for the negative samples, motivated by the fact that we do not intend to generate from the negative samples, sparing information bandwidth for the reconstruction of positive samples.

One has numerous options to choose the positive and negative priors. In our experiments we simply choose a standard normal for the positive prior, and a shifted standard normal for the negative prior. With a rotationally symmetric posterior distribution, the distance between the two priors would be the only unspecified hyperparameter of such a model. The assumption of diagonal covariance matrix posterior breaks rotational symmetry in principle, but our exploratory experiments have demonstrated that the magnitude of the shift is a more significant modeling choice than the direction/sparsity of the shift.

Negative samples can also be obtained in different ways. We conduct experiments with several variants:

- The data with isotropic Gaussian noise added.
- Samples from an auxiliary dataset.
- Generated samples from the trained model itself.

Except for the variant using auxiliary data, these methods are fully unsupervised. The third variant, where the negative samples are coming from the generated distribution can be interpreted as a form of generative adversarial training: one direction of the latent space is dedicated to discriminating a newly generated sample from previously generated samples.

The task of our models is to generalize from the negative samples as much as possible to all possible out-of-distribution samples, so that they can push down out-of-distribution likelihood estimates. Depending on the source of negative samples, this generalization can be easier or harder. Negative samples that are very far from the data manifold do not facilitate generalization. Noise added to data points is a simple and principled way to sample from the vicinity of the data manifold, but as we will see it does not provide good generalization. We argue that the reason for this is that discriminating between noisy and noiseless points is too easy for the encoder, so "semantically" the noisy versions are far from the data manifold. In contrast, samples generated from the trained model are a more robust way to achieve good out-of-distribution likelihood estimates, as we will experimentally demonstrate. We hypothesize that the reason for this is that near-manifold points obtained this way are semantically more meaningful in the above sense. See Lee et al. (2017) for an incarnation of this idea in the context of classification and generative adversarial networks.

4 Experimental results

Our main concern is on the discriminative power of VAE models regarding out-of-distribution samples. The general experimental setup in this section is as follows: we train a model on a train set of a dataset (e.g. train set of Fashion-MNIST) and then require the model to discriminate between the test set of the train dataset (e.g. test set of Fashion-MNIST) and the test set of an out-of-distribution dataset (e.g. test set of MNIST). During the training phase, the models do not encounter examples from the OOD dataset, only at test time are they expected to able to distinguish between inliers and out-of-distribution samples.

For quantitative assessment, we use the threshold independent AUC metric calculated with the bits-per-dimension score (denoted by AUC BPD throughout this section) and also with the KL divergence of the posterior distribution of a data point to the prior as a score (denoted by AUC KL). We also report average bits-per-dimension (BPD) scores on the test set of both the training and the out-of-distribution datasets (denoted by Test BPD and OOD BPD, respectively). All reported numbers in this section are averages of 5 runs with standard deviations denoted in parentheses.

We conduct experiments on two sets of datasets: color images of size 32x32 (CIFAR-10, SVHN, downscaled ImageNet) and grayscale images of size 28x28 (MNIST, Fashion-MNIST, Kuzushiji-MNIST, EMNIST-Letters). For both cases, the prior is chosen to be standard normal and the second prior is standard normal with a shifted mean. For color images, the latent dimension is set to 100, and the negative prior is centered at $25 \cdot \mathbf{1}$. For grayscale images, the latent dimension is set to 10 and the negative prior is centered at $8 \cdot \mathbf{1}$. (In both cases, the magnitude of the shift for the negative prior is set to be large enough for the typical regions of the prior and the negative prior not to overlap.)

For a detailed description of the utilized datasets, models, and training methodology, see Appendix A.

4.1 Experiment 1: the effect of the noise model

In this experiment, we examine baseline VAE models (i.e., models without negative sampling) and investigate the effect of the choice of distributions in the observable space. We conduct experiments with two dataset pairs and compare the behavior of the Bernoulli and the Gaussian noise models. Table 1 and Table 2 summarizes the results.

First, we examine the importance of the choice of the noise model. Results of experiments conducted with grayscale images from the first two columns in Table 1 suggest that the intriguing phenomenon in VAEs discussed by Nalisnick et al. (2018) and Choi et al. (2018) is highly dependent on modelling choices. In the case of Gaussian noise model the issue of assigning higher likelihood estimates to OOD samples simply does not occur, however, one can observe that discrimination between inliers and OOD samples based on the KL-divergence between approximate posterior and prior is hardly feasible, with below-$1/2$ AUC scores. Meanwhile, with a Bernoulli noise model (also used in Nalisnick et al. (2018)) both the likelihood-estimates and the KL-divergences fail to discriminate. The other results in the table (where models are trained on MNIST) confirm the assymetric behaviour of the phenomenon already described by Nalisnick et al. (2018).

Concerning experiments with color images, the last two columns of Table 2 again shows the importance of modelling choices, while when CIFAR-10 is the training set, the phenomenon persistently occurs with Bernoulli, Gaussian and Quantized Gaussian noise model as well.

4.2 Experiment 2: the effectiveness of negative sampling

To demonstrate the effectiveness of negative sampling we present two different sets of experiments: on one hand we incorporate negative samples from an auxiliary dataset (here we use the EMNIST-Letters dataset for grayscale images and Downscaled ImageNet for color images), and on the other hand we also explore the use of adversarially generated negative samples.

Table 1: Comparing the out-of-distribution discriminative power of baseline VAE models with different noise models using datasets Fashion-MNIST and MINST.

Train	Fashion-MNIST		MNIST	
OOD test set	MNIST		Fashion-MNIST	
Noise model	Bernoulli	Gaussian	Bernoulli	Gaussian
AUC BPD	0.46 (0.05)	0.98 (0.00)	1.00 (0.00)	0.97 (0.00)
AUC KL	0.61 (0.09)	0.26 (0.03)	0.73 (0.14)	0.71 (0.04)
Test BPD	0.30 (0.00)	0.94 (0.00)	0.13 (0.00)	0.94 (0.00)
OOD BPD	0.35 (0.08)	0.96 (0.00)	1.36 (0.03)	0.99 (0.00)

Table 2: Comparing the out-of-distribution discriminative power of baseline VAE models with different noise models using datasets CIFAR-10 and SVHN. (Q. Gaussian refers to Quantized Gaussian.)

Train	CIFAR-10			SVHN	
OOD test set	SVHN			CIFAR-10	
Noise model	Bernoulli	Gaussian	Q. Gaussian	Bernoulli	Gaussian
AUC BPD	0.59 (0.00)	0.25 (0.02)	0.19 (0.00)	0.51 (0.00)	0.92 (0.00)
AUC KL	0.29 (0.00)	0.25 (0.01)	0.28 (0.01)	0.87 (0.00)	0.74 (0.01)
Test BPD	0.59 (0.00)	0.93 (0.00)	0.93 (0.00)	0.60 (0.00)	0.93 (0.00)
OOD BPD	0.60 (0.00)	0.93 (0.00)	0.93 (0.00)	0.62 (0.01)	0.94 (0.00)

Table 3 shows that using the auxiliary dataset as source of negative samples proved to result in models that are capable of nearly perfectly distinguishing between inliers and OOD samples, as the AUC scores from the two middle columns in Table 3 indicate. This is also the case with color images, as experimental results in Table 4 show.

The last two columns in Table 3 show the effectiveness of the fully unsupervised approach: both with a Gaussian and a Bernoulli noise model, the trained models achieve notably higher AUC scores than the baseline.

Table 3: Table summarizing experimental results for models trained on Fashion-MNIST with MNIST as OOD dataset. First two rows show means of AUCs (higher is better) calculated with bits-per-dimension (BPD) score (first row) or only the KL-divergence in latent space (second row), averages of 5 runs with standard deviations in parentheses. Last two rows show the calculated reconstruction losses (lower is better) of inlier and OOD test samples averaged over a minibatch.

	Baseline VAE		VAE with negative sampling auxiliary: EMNIST-Letters		VAE with negative sampling auxiliary: generated	
	Bernoulli	Gaussian	Bernoulli	Gaussian	Bernoulli	Gaussian
AUC BPD	0.46 (0.05)	0.98 (0.00)	1.00 (0.00)	1.00 (0.00)	0.70 (0.13)	0.80 (0.04)
AUC KL	0.61 (0.09)	0.26 (0.03)	1.00 (0.00)	1.00 (0.00)	0.88 (0.07)	0.74 (0.05)
Test BPD	0.30 (0.00)	0.94 (0.00)	0.30 (0.00)	0.94 (0.00)	0.47 (0.09)	1.04 (0.01)
OOD BPD	0.35 (0.08)	0.96 (0.00)	1.45 (0.19)	1.38 (0.01)	10^{18} (10^{19})	42.40 (76.21)

4.3 Experiment 3: utilizing different sources for negative samples

In this experiment, we investigate how the choice of the auxiliary dataset influences the performance of the trained model. We train models with Fashion-MNIST as the inlier dataset and employ MNIST as outlier dataset. What we vary in this experiment is the source of the utilized negative samples, which are as follows: EMNIST-Letters, Kuzushiji-MNIST (KMNIST)[1], random noise (in which we

[1] EMNIST-Letters, Kuzushiji-MNIST and Fashion-MNIST are datasets that can be utilized as drop-in replacements for MNIST.

Table 4: Comparing VAE with negative sampling with Bernoulli, Gaussian, and Quantized Gaussian noise models trained on CIFAR-10, and using SVHN as OOD, with auxiliary dataset and adversarial training as source of negative samples.

	VAE with negative sampling auxiliary: Downscaled ImageNet			VAE with negative sampling auxiliary: generated		
	Bernoulli	Gaussian	Q. Gaussian	Bernoulli	Gaussian	Q. Gaussian
AUC BPD	$0.90\ (0.05)$	$0.93\ (0.01)$	$0.92\ (0.03)$	$0.53\ (0.09)$	$0.63\ (0.15)$	$0.50\ (0.08)$
AUC KL	$0.90\ (0.06)$	$0.93\ (0.01)$	$0.92\ (0.03)$	$0.53\ (0.10)$	$0.63\ (0.14)$	$0.51\ (0.08)$
Test BPD	$0.77\ (0.15)$	$1.02\ (0.02)$	$1.06\ (0.08)$	$1.99\ (0.16)$	$2.29\ (0.36)$	$2.29\ (0.27)$
OOD BPD	$2.09\ (0.34)$	$2.63\ (0.13)$	$2.51\ (0.53)$	$2.52\ (0.86)$	$4.43\ (2.30)$	$2.39\ (0.48)$

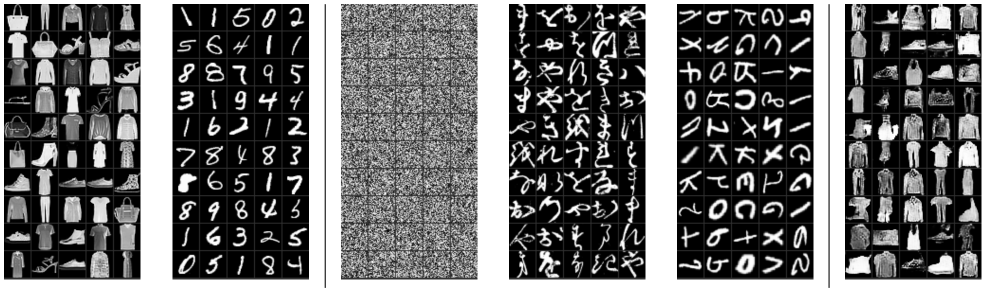
sample each pixel intensity from the uniform distribution on $[0, 1]$ — modeling a dataset with less structure). We also experiment here with an adversarial training scheme, where the negative samples are coming from a model itself by utilizing the generated samples of the generator. In this setup, the generator gets gradient signal to map the generated images into the prior. In this experiment, we use a Bernoulli noise model. The results are summarized in Table 5.

The results show that utilizing either KMNIST or MNIST-Letters results in perfect separation of the inliers (Fashion-MNIST) and outliers (MNIST). Employing adversarial negatives (last column) also significantly improves the performance over the baseline with remarkably better separation measured in AUC KL metric.

The weak results with random noise as negative samples show the significance of the choice of negative samples. We also experimented with utilizing the training set itself with an additive isotropic Gaussian noise as negative samples — a rather natural choice to provide near-manifold examples. With an additive noise of $\sigma = 0.25$, the results for the AUC BPD metric is $0.44\ (0.01)$ and 0.70 (0.09) for the AUC KL, showing weak discriminative power.

Table 5: Comparing baseline model and negative sampling with auxiliary datasets and adversarial training. Columns correspond to different sources for negative samples. Results for the baseline (i.e., VAE without negative sampling) are indicated again in the first column for comparison. Samples from the different data sets are also depicted in the last row to show their general visual characteristics.

Trained on Fashion-MNIST	OOD test set MNIST	Trained with negatives from other datasets Random	KMNIST	Letters	Adversarial Negatives
AUC BPD	$0.46\ (0.05)$	$0.47\ (0.05)$	$1.00\ (0.00)$	$1.00\ (0.00)$	$0.76\ (0.14)$
AUC KL	$0.61\ (0.09)$	$0.56\ (0.08)$	$1.00\ (0.00)$	$1.00\ (0.00)$	$0.89\ (0.08)$
Test set BPD	$0.30\ (0.00)$	$0.30\ (0.00)$	$0.30\ (0.00)$	$0.30\ (0.00)$	$0.47\ (0.09)$
OOD set BPD	$0.35\ (0.08)$	$0.32\ (0.04)$	$1.10\ (0.09)$	$1.44\ (0.20)$	$10^{18}\ (10^{19})$



5 Related Work

Our investigations are mostly inspired by and related to recent work on the evaluation of generative models on OOD data (Shafaei et al., 2018; Nalisnick et al., 2018; Choi et al., 2018; Hendrycks et al., 2018).

As several concurrent works report, despite intuitive expectations, generative models — including but not limited to VAEs — consistently fail at distinguishing OOD data from the training data, yielding likelihood estimates higher on unseen OOD samples, e.g. a VAE trained on the train set of CIFAR-10 dataset produces higher likelihoods on the unseen SVHN test dataset than on CIFAR-10 train or test set. Nalisnick et al. (2018) examine the phenomenon in detail, focusing on finding the cause of it by analyzing flow-based models that allow exact likelihood calculation. Our work aligns with their empirical results regarding VAEs: the asymmetric behaviour with Bernoulli noise model is also confirmed by our results.

Choi et al. (2018) also notice the above mentioned phenomenon, while they address the task of OOD sample detection with Generative Ensembles. They decrease the weight of the KL-divergence term in the ELBO in order to alleviate the wrong likelihood estimation of a single model, contrarily to what is promoted by the β-VAE loss function. One line of work uses the reconstruction error of a VAE to distinguish between inliers and outliers (An & Cho, 2015).

The same observation is presented by Hendrycks et al. (2018), but they concentrate on improving the OOD data detection with Outlier Exposure. Their work demonstrates that utilizing samples from auxiliary data set as OOD examples i.e. training models to discriminate between training and auxiliary samples, significantly improves on the performance of existing OOD detection models on image and text data.

Within the context of uncertainty estimation, Lee et al. (2017) demonstrate that adversarially generated samples improve the confidence of classifiers in their correct predictions. They train a classifier simultaneously with a GAN and require from it to have lower confidence on GAN samples. For each class distribution, they tune the classifier and GAN using samples from that OOD dataset. Their method of utilizing generated samples of GANs is closest to our approach of using generated data points as negative samples, but Lee et al. (2017) work within a classification setting.

Nalisnick et al. (2019) propose a solution that can alleviate the issue without modifying existing generative models, but the issue they aim to address (distributional shift) is very different from the standard concerns of OOD sample detection. Their model works by using the likelihood estimates coming from likelihood-based models as inputs to detect distributional shift, as opposed to using them as raw OOD sample detectors. The model operates under the assumption that at evaluation time, samples come in batches, and thus can be the inputs of statistical tests differentiating between likelihood estimates for inlier datasets and likelihood estimates for evaluation datasets. In the limiting case where the evaluation dataset has batch-size 1, the performance of this model can meaningfully be compared with our unsupervised models. We leave this for future work.

6 Conclusions

In this work we studied Variational Autoencoder (VAE) models, and investigated to what extent the latent representations of data points or the likelihood estimates given by the model can be used to identify out-of-distribution (OOD) samples (points that are not from the true data distribution learned by the model). We demonstrated empirically that the extent of the notorious phenomenon of wrong out-of-distribution likelihood estimates present in VAEs is highly dependent on the observation model. We introduced negative sampling as an approach to alleviate a weakness of the Variational Autoencoder model family of assigning incorrect likelihood estimations to out-of-distribution samples. We presented empirical evidence that utilizing negative samples either from an auxiliary dataset or from an adversarial training scheme significantly and consistently improves the discriminative power of VAE models regarding out-of-distribution samples.

References

Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, 2:1–18, 2015.

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

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

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017. doi: 10.1109/ijcnn.2017.7966217.

Dan Hendrycks, Mantas Mazeika, and Thomas G Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

Diderik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, 2, 2010.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

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

Eric Nalisnick et al. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. *Advances in Neural Information Processing Systems (NIPS)*, 2011.

Mihaela Rosca, Balaji Lakshminarayanan, and Shakir Mohamed. Distribution matching in variational inference. *arXiv preprint arXiv:1802.06847*, 2018.

Alireza Shafaei, Mark Schmidt, and James J Little. Does your model know the digit 6 is not a cat? a less biased evaluation of" outlier" detectors. *arXiv preprint arXiv:1809.04729*, 2018.

Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *CoRR*, abs/1601.06759, 2016. URL `http://arxiv.org/abs/1601.06759`.

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

A Experimental details

A.1 Datasets and preprocessing

We conduct experiments with two types of data set: color images of size 32x32 and grayscale images of size 28x28. The utilized datasets are listed below.

Datasets of color images of size 32x32:

- CIFAR-10 (Krizhevsky, 2009): 32x32x3 images, 50.000 train + 10.000 test, 10 classes
- SVHN (cropped) (Netzer et al., 2011): 32x32x3 images, 73.257 train + 26,032 test (+ 531.131 extra unlabeled), 10 classes
- Downsampled ImageNet (van den Oord et al., 2016): 32x32x3 images, 1.281.149 train + 49.999 validation, 1000 classes

Datasets of grayscale images of size 28x28:

- MNIST (LeCun et al., 2010): 28x28x1, 60.000 train + 10.000 test, 10 classes
- Fashion-MNIST (Xiao et al., 2017): 28x28x1, 60.000 train + 10.000 test, 10 classes
- Kuzushiji-MNIST (Clanuwat et al., 2018): 28x28x1, 60.000 train + 10.000 test, 10 classes
- EMNIST-Letters (Cohen et al., 2017): 28x28x1, 60.000 train + 10.000 test, 10 classes

We apply no preprocessing step other than normalizing the input images to $[0, 1]$.

A.2 Network architecture and training details

A.2.1 Details for grayscale images

Following Nalisnick et al. (2018), for grayscale images, we use the encoder architecture described in Rosca et al. (2018) in appendix K table 4. Also, as in Rosca et al. (2018), all of the models are trained with the RMSProp optimizer with learning rate set to 10^{-4}. We train the models for 100 epochs with mini-batch size of 50. We update the parameters of the encoder and decoder network iteratively/separately.

A.2.2 Details for color images

We use a DCGAN-style CNN architecture with Conv–BatchNorm–ReLU modules for both the encoder and the decoder. The size of the kernels are 4×4, and the number of filters are $32, 64, 128$ for the encoder; and $128, 64, 1$ for the decoder. All of the models are trained with the Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$) for 100 epochs with mini-batch size 50. The learning rate is set to 10^{-4}. We update the parameters of the encoder and decoder network iteratively/separately. For all of our experiments, we employ a standard normal latent prior.