Probabilistic Auto-Encoder

Vanessa Böhm
Berkeley Center for Cosmological Physics
Department of Physics
University of California
Berkeley, CA, USA
Lawrence Berkeley National Laboratory
`vboehm@berkeley.edu`

Uroš Seljak
Berkeley Center for Cosmological Physics
Department of Physics
University of California
Berkeley, California, USA
Lawrence Berkeley National Laboratory
`useljak@berkeley.edu`

Abstract

We introduce the probabilistic auto-encoder (PAE), a generative model with a lower dimensional latent space that is based on an auto-encoder which is interpreted probabilistically after training using a normalizing flow. The PAE is fast and easy to train, achieves small reconstruction errors, high sample quality and good performance in downstream tasks. Compared to a VAE and its common variants, the PAE trains faster, reaches a lower reconstruction error and produces state of the art sample quality without requiring special tuning parameters or training procedures. We further demonstrate that the PAE is a powerful model for performing the downstream tasks of outlier detection and probabilistic image reconstruction: 1) We identify a PAE-based outlier detection metric which achieves state of the art results and outperforms other likelihood based estimators. 2) We perform high dimensional data inpainting and denoising with uncertainty quantification by means of posterior analysis in the PAE latent space. Most generative models are specifically tuned to excel in one or two applications. With the PAE we introduce an easy-to-train, simple, but at the same time powerful model that performs well and reliably in many tasks without requiring special fine-tuning or training procedures. We make all PAE codes publicly available.[1]

1 Introduction

Deep generative models are powerful machine learning models designed to perform tasks such as realistic artificial data generation and data likelihood evaluation. Many generative models are laid out for one specific task. For example, generative adversarial networks (Goodfellow et al., 2014) produce high quality samples, but do not allow for a direct evaluation of the data likelihood. Their samples often fail to cover the entire data space. A probabilistic model that covers the data space are variational auto-encoders (VAEs) (Kingma & Welling, 2013; Rezende et al., 2014), which are trained to maximize a lower bound to the data likelihood (ELBO). VAEs allow for expressive architectures but require fine-tuning additional parameters (Higgins et al., 2017), architectural tweaks or special training procedures to reach good sample quality (Alemi et al., 2018b; Rezende & Mohamed, 2015; Larsen et al., 2016). Both GANs and VAEs use lower dimensional latent spaces. Normalizing flows (NFs) (Rippel & Adams, 2013; Dinh et al., 2014, 2016; Kingma & Dhariwal, 2018a; Grathwohl et al., 2018; Kingma & Dhariwal, 2018b; Durkan et al., 2019; Papamakarios et al., 2019) model the probability density $p(\boldsymbol{x})$ of high-dimensional data directly. They introduce a bijective mapping from $\boldsymbol{x}$ to an underlying latent representation $\boldsymbol{z}$. The latter is enforced to follow a tractable prior distribution $p(\boldsymbol{z})$. NFs do not reduce the dimensionality and achieve high data likelihoods on many standard data sets (Gritsenko et al., 2019). The GLOW model (Kingma & Dhariwal, 2018b) achieves

---

[1]`https://github.com/VMBoehm/PAE`

visually impressive results in artificial image generation, but requires the user to make a trade-off between image qualities and diversities. This trade-off is chosen by setting a temperature parameter. In addition, the data likelihood of NFs has been shown to be a poor out-of-distribution (OoD) metric (Nalisnick et al., 2018). This finding similarly applies to the ELBO of VAEs (Nalisnick et al., 2018).

The failure of VAEs and NFs in OoD detection shows that maximizing the data likelihood does not necessarily correspond to optimizing a model for a specific downstream task. For example, a phenomenon which is well known in linear methods such as probabilistic PCA (Tipping & Bishop, 1999) are singular latent space variables. They have vanishing effect on the data distribution, yet dominate the likelihood: the contribution of the smallest eigenvalues, λ_i, to the log likelihood of a Gaussian, linear model is $\Delta \ln p(\boldsymbol{x}) = -0.5[1 + \ln(2\pi\lambda_i)]$, which diverges for $\lambda_i \to 0$. This suggests that maximizing the likelihood under a model can be dominated by these eigenvalues, while they may have no relevance for tasks such as generative sampling or outlier detection. In linear methods this problem is addressed with shrinkage regularization (Ledoit & Wolf, 2004, 2015) or dimensionality reduction (low rank PCA) (Tipping & Bishop, 1999).

Here we propose a new composite generative model, the Probabilistic auto-encoder (PAE), which is not trained to maximize the data likelihood, but achieves state of the art results in many downstream tasks. Specifically, we show that the PAE produces high quality samples, that a PAE-based density estimator outperforms other generative model-based outlier detectors, and that the PAE can be used for probabilistic data denoising and inpainting. The PAE maps the data to a lower dimensional latent space with an auto-encoder (AE) and performs density estimation in the latent space of the AE with a normalizing flow. In the limit of no dimensionality reduction, the PAE becomes a normalizing flow. The PAE can thus be interpreted as a regularized normalizing flow, in which the effect of singular dimensions is suppressed by the AE compression, in analogy to a low rank probabilistic PCA regularizing the density estimation in the linear case. An advantage of the PAE over many other generative models is that it requires little hyperparameter tuning, no special training procedures or architectural tweaks to achieve good results. It is thus particularly useful for non-ML experts, in particular scientists, who are looking for an easy-to-use and reliable generative model that also performs well in downstream tasks.

2 PAE training, sampling and density estimation

To construct the PAE we start by training a standard auto-encoder on the N-dimensional data $\boldsymbol{x}$. The latent space $\boldsymbol{z}$ is of dimensionality $K < N$. Encoder $\boldsymbol{f}$ and decoder $\boldsymbol{g}$ are deep neural networks with trainable parameters ϕ and θ, respectively.

$$\boldsymbol{f}_\phi : \mathbb{R}^N \to \mathbb{R}^K, \boldsymbol{x} \mapsto \boldsymbol{f}_\phi(\boldsymbol{x}), \ \boldsymbol{g}_\theta : \mathbb{R}^K \to \mathbb{R}^N, \boldsymbol{z} \mapsto \boldsymbol{g}_\theta(\boldsymbol{z}). \tag{1}$$

The training objective of the AE is the reconstruction error

$$\mathcal{L}_{\text{AE}} = \mathbb{E}_{p(x)}[\boldsymbol{x} - \boldsymbol{g}_\theta(\boldsymbol{f}_\phi(\boldsymbol{x}))]^2. \tag{2}$$

The auto-encoder by itself is not a probabilistic model. To construct the PAE, we interpret it probabilistically *after* training. The latent space prior, $p(\boldsymbol{z})$, of the PAE is found by performing a density estimation on the AE-encoded training data. Formally, this can be be derived from

$$p(\boldsymbol{z}) = \int \mathrm{d}\boldsymbol{x} \, p(\boldsymbol{x}) \, p_\phi(\boldsymbol{z}|\boldsymbol{x}) = \mathbb{E}_{p(x)}\left[p_\phi(\boldsymbol{z}|\boldsymbol{x})\right] = \mathbb{E}_{p(x)}\left[\delta_D(\boldsymbol{z} - \boldsymbol{f}_\phi(\boldsymbol{x}))\right], \tag{3}$$

where in the last step we have identified the AE posterior, $p_\phi(\boldsymbol{z}|\boldsymbol{x})$, as a Dirac delta distribution. In our experiments we learn $p(\boldsymbol{z})$ using a normalizing flow (NF), but other density estimators could be used instead. Note that the NF training is performed *after we have trained the auto-encoder*. We do not retrain the generator or decoder of the AE in this step. The AE latent space is of relatively low dimensionality, $K \ll N$, which allows for computationally tractable density estimation: the NF models require little tuning in terms of architecture and are fast to train. The performance of different density estimators can be compared at this step.

The NF is a bijective mapping from the latent space of the AE-encoder, $\boldsymbol{z}$, to the latent space of the NF, $\boldsymbol{u}$, which is parametrized by deep networks with parameters, γ,

$$\boldsymbol{b}_\gamma : \mathbb{R}^K \to \mathbb{R}^K, \boldsymbol{z} \mapsto \boldsymbol{u} = \boldsymbol{b}_\gamma(\boldsymbol{z}). \tag{4}$$
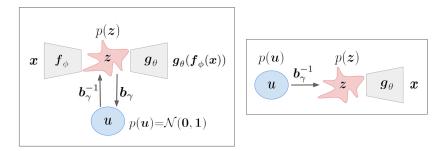
Figure 1: Schematic diagram of the PAE (left panel) and an illustration of the sampling procedure from the PAE (right panel). The auto-encoder networks are depicted as gray trapezia, the normalizing flow is represented by black arrows and the latent spaces of the auto-encoder and normalizing flow are shown in red and blue, respectively.

The training objective of the NF is the negative log likelihood of the encoded samples,

$$\mathcal{L}_{\mathrm{NF}} = -\ln p_\gamma(\boldsymbol{z}) = -\ln p(\boldsymbol{u}) - \ln \left| \det \frac{\partial \boldsymbol{b}_\gamma^{-1}(\boldsymbol{u})}{\partial \boldsymbol{u}} \right|. \tag{5}$$

where the NF prior is $p(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{1})$ and the determinant of the NF Jacobian is given by

$$|\det J_\gamma| = \left| \det \frac{\partial \boldsymbol{b}_\gamma(\boldsymbol{z})}{\partial \boldsymbol{z}} \right| = \left| \det \frac{\partial \boldsymbol{b}_\gamma^{-1}(\boldsymbol{u})}{\partial \boldsymbol{u}} \right|. \tag{6}$$

2.1 Sampling from the PAE

To sample from $p(\boldsymbol{x})$ we draw a sample, $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{1})$, from the NF latent distribution and pass it through both the NF and AE generators (Figure 1),

$$\boldsymbol{x} = \boldsymbol{g}_\theta(\boldsymbol{b}_\gamma^{-1}(\boldsymbol{u})). \tag{7}$$

It is often argued that AEs are problematic because they can map very different images into very nearby latent space points. This is cured by the NF, which maps well separated samples, $\boldsymbol{u}$, to nearby samples in the AE latent space $\boldsymbol{z}$. It makes sure that many samples land in regions of high $p(\boldsymbol{z})$ and that the entire AE latent space is covered. This approach achieves an excellent sample quality and diversity, as well as continuous interpolations between data points. We note that a lower dimensional latent space is usually not a hindrance to high sample quality, as long as its dimensionality is chosen high enough to achieve reconstructions with the desired quality.

The PAE is trained to first achieve optimal reconstruction quality. The NF then allows to draw samples of that same quality. This is different from a VAE which needs to balance the reconstruction error (the likelihood term) with the sample quality (the KL term). If the former dominates the loss during training, the encoded distribution and prior do not match well. In this case, samples from the prior can land outside of the encoded domain resulting in a low sample quality. If the KL term dominates, some latent variables will not encode sufficient information and the reconstruction quality is low (a problem known as posterior collapse). A number of solutions have been proposed to overcome these problems. They generally add complexity in terms of model design and training procedure to the VAE. The differences between PAE and modified VAEs are discussed in more detail in Section 3.

2.2 A PAE-based density estimator for outlier detection

The PAE is not trained to maximize the data likelihood $p(\boldsymbol{x})$. While we have shown in Section 2.1 that this is not required to obtain a probabilistic generative model, an estimate of $p(\boldsymbol{x})$ can be useful for certain downstream tasks such as out-of-distribution detection.

In latent space models a $p(\boldsymbol{x})$ estimate is obtained by marginalization over the latent space

$$p_{\theta,\gamma}(\boldsymbol{x}) = \int \mathrm{d}\boldsymbol{z} \, p_\gamma(\boldsymbol{z}) p_\theta(\boldsymbol{x}|\boldsymbol{z}). \tag{8}$$

3

The prior, $p_\gamma(\boldsymbol{z})$, is modelled by the NF (Eq. 5), but Eq. 8 further requires defining an implicit likelihood, $p(\boldsymbol{x}|\boldsymbol{z})$. For this we make the Gaussian ansatz

$$p_\theta(\boldsymbol{x}|\boldsymbol{z}) = [2\pi \prod_i \sigma_i^2]^{-\frac{N}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^N \frac{[x_i - g_{\theta i}(\boldsymbol{z})]^2}{\sigma_i^2}\right\}, \tag{9}$$

where σ_i is the mean reconstruction error of the AE in dimension i, which we measure on the validation data. We denote the noise covariance of the implicit likelihood, which is the diagonal matrix with values σ_i, as $\boldsymbol{\sigma}$.

The integral in Eq.8 is generally not tractable and needs to be solved approximately. In VAEs it is approximated with variational inference (VI). Here, we follow a different approach and approximate it with Laplace's method: We approximate the exponent in the integrand of Eq.8 by its expansion to second order around the maximum and perform the remaining Gaussian integral. For a function $A : \mathbb{R}^{\mathbb{K}} \to \mathbb{R}, \boldsymbol{z} \mapsto A(\boldsymbol{z}), A \in \mathcal{C}^2$ with unique maximum $A(\boldsymbol{z}')$ the Laplace approximation to the integral $\int \mathrm{d}\boldsymbol{z} \exp\left[-A(\boldsymbol{z})\right]$ is

$$\int \mathrm{d}\boldsymbol{z}\, e^{-A(\boldsymbol{z})} \approx e^{-A(\boldsymbol{z}')} \int \mathrm{d}\boldsymbol{z}\, e^{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{z}')^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{z}')} = e^{-A(\boldsymbol{z}')}\sqrt{\frac{(2\pi)^K}{\det \boldsymbol{\Sigma}^{-1}}}, \tag{10}$$

with the covariance given by the inverse Hessian at the maximum $\boldsymbol{\Sigma} = \boldsymbol{H}^{-1}$, $H_{ij} = \partial_i \partial_j A(\boldsymbol{z}')$. The PAE density estimate for a data point under the Laplace approximation is thus,

$$\ln p(\boldsymbol{x}) \approx \ln p_\theta(\boldsymbol{x}|\boldsymbol{z}') + \ln p_\gamma(\boldsymbol{z}') - \frac{N}{2}\ln 2\pi + \frac{1}{2}\ln \det \boldsymbol{\Sigma}_{z'}. \tag{11}$$

Evaluating Eq. 11 proceeds in two steps:

1. Find the maximum of the posterior (MAP), $\boldsymbol{z}'$, by minimizing
$$\mathcal{L}_{\mathrm{MAP}} = \arg\min_{\boldsymbol{z}}\{-\ln p_\theta(\boldsymbol{x}|\boldsymbol{z}) - \ln p_\gamma(\boldsymbol{z})\}. \tag{12}$$

   The encoded value, $\boldsymbol{f}_\phi(\boldsymbol{x})$, can be used as initial value. The minimization is generally computationally tractable and fast since it is performed in the lower dimensional latent space. We have found that using the amortized $\boldsymbol{f}_\phi(\boldsymbol{x})$ in lieu of the MAP already achieves excellent outlier detection accuracy.

2. Obtain the Hessian at the MAP value
$$\boldsymbol{\Sigma}_{z'}^{-1} = \boldsymbol{J}_\gamma^T \boldsymbol{J}_\gamma + \left[\frac{\partial \boldsymbol{g}_\theta}{\partial \boldsymbol{z}}(\boldsymbol{z}')\right]^T \boldsymbol{\sigma}^{-2}\left[\frac{\partial \boldsymbol{g}_\theta}{\partial \boldsymbol{z}}(\boldsymbol{z}')\right], \tag{13}$$

   where we have used the Gauss-Newton approximation to the Hessian, meaning that we have dropped all terms with higher derivatives.

Given the MAP position and the Hessian we can then estimate $p(\boldsymbol{x})$ under the Laplace approximation,

$$\ln p(\boldsymbol{x}) \approx -\frac{1}{2}[\boldsymbol{x} - \boldsymbol{g}_\theta(\boldsymbol{z}')]^T \boldsymbol{\sigma}^{-2}[\boldsymbol{x} - \boldsymbol{g}_\theta(\boldsymbol{z}')] - \frac{1}{2}\boldsymbol{b}_\gamma(\boldsymbol{z}')^T \boldsymbol{b}_\gamma(\boldsymbol{z}') + \ln|\det \boldsymbol{J}_\gamma|$$
$$- [\ln \det \boldsymbol{\sigma}^2 - \ln \det \boldsymbol{\Sigma}_{z'} + N \ln 2\pi]/2. \tag{14}$$

We found in our experiments that the Hessian tends to suffer from numerical noise, which decreases the accuracy of the density estimation. To obtain a more reliable OoD metric, we drop the Hessian term,

$$\ln \tilde{p}(\boldsymbol{x}) \approx -\frac{1}{2}[\boldsymbol{x} - \boldsymbol{g}_\theta(\boldsymbol{z}')]^T \boldsymbol{\sigma}^{-2}[\boldsymbol{x} - \boldsymbol{g}_\theta(\boldsymbol{z}')] - \frac{1}{2}\boldsymbol{b}_\gamma(\boldsymbol{z}')^T \boldsymbol{b}_\gamma(\boldsymbol{z}') + \ln|\det \boldsymbol{J}_\gamma|. \tag{15}$$

The metric in Eq. 15 has been derived from a normalized probability density estimator but can no longer be interpreted as one. Intuitively, it can be interpreted as the conjunction of two outlier metrics: a (weighted) reconstruction error (first term of Eq. 15) and a density estimator in the latent space of the auto-encoder (second and third term of Eq. 15).

We will show that in high latent space dimensions we can also drop the reconstruction error (first term) from Eq. 15: in this case the density estimator becomes a regularized NF density estimator, where regularization has been achieved by dimensionality reduction. We find that Eq. 15 is an excellent outlier detection metric for high-dimensional data $\boldsymbol{x}$, outperforming AE-based outlier detection (AE-based outlier detection uses the reconstruction error, which corresponds to the first term of Eq. 15 if all σ_i are equal.), the ELBO, IWAE (Burda et al., 2016), and the data likelihood of normalizing flows.

2.3 PAE-based denoising and inpainting with uncertainty quantification

Another important downstream task of generative models is image denoising and inpainting. This is often performed with point estimators and the important task of uncertainty quantification (UQ) is neglected. The probabilistic PAE framework can be used to address both with posterior inference (Böhm et al., 2019). For Gaussian noise, n, with noise covariance matrix σ_{noise} (typically a diagonal matrix) and a pixel-wise mask M, the log posterior of a corrupted data point, $\tilde{x} = Mx + n$, under the PAE is

$$\ln p_{\theta,\gamma}(u|\tilde{x}, M, \sigma_{\text{noise}}) = \ln p_{\theta,\gamma}(\tilde{x}|u, M, \sigma_{\text{noise}}) + \ln p(u) - \text{const.} \qquad (16)$$

The prior is $p(u) = \mathcal{N}(0, 1)$ and the implicit likelihood is given by

$$p_{\theta,\gamma}(\tilde{x}|u, M, \sigma_{\text{noise}}) = \mathcal{N}\left(M\left[g_\theta(b_\gamma^{-1}(u))\right], \sigma^2 + \sigma_{\text{noise}}^2\right). \qquad (17)$$

The covariance of this Gaussian likelihood is composed of the PAE reconstruction error, σ_i^2, and the noise level in the corrupted data, $\sigma_{\text{noise},i}^2$. For sufficiently high latent space dimensionalities the latter dominates, $\sigma_i \ll \sigma_{\text{noise},i}$, ensuring that the likelihood is well approximated by a Gaussian. Note that we have replaced x by its generative process, $g_\theta(b_\gamma^{-1}(u))$, thus bringing the inference problem to the low dimensional latent space of the PAE. Posterior analysis is computationally more tractable in this lower dimensional space.

To denoise and inpaint a corrupted image we perform latent space posterior analysis. A point estimate is given by the MAP, u', the maximum of Eq. 16, which forward modeled into data space, $x_{\text{recon}} = g_\theta(b_\gamma^{-1}(u'))$, yields the most likely underlying image. For a full posterior analysis any technique, such as MAP+Lapace, VI or MCMC sampling could be used. Given the multi-modal posterior of some of our examples we fit a full rank Gaussian mixture model to Eq. 16 following Seljak & Yu (2019). We can then sample from this model to obtain other solutions that are compatible with the data, thereby obtaining an uncertainty estimate of the reconstruction. In Sec. 4 we show that this procedure recovers high quality images from badly corrupted inputs.

3 Related Work

The PAE can be viewed as a generalization of probabilistic PCA (Tipping & Bishop, 1999) and reduces to it for linear, Gaussian models. The PAE replaces the low rank PCA components by the low dimensional AE latent space. Both approaches treat the reconstruction error that arises from the compression as Gaussian noise (Eq. 9). The low rank PCA representation removes components with very small eigenvalues that would otherwise dominate the likelihood. A Gaussian prior is imposed on the PCA weights to make the PCA probabilistic. The PAE imposes a normalizing flow prior on the latent space.

The PAE is also a form of regularized normalizing flow, since an NF can be viewed as a nonlinear generalization of a full rank probabilistic PCA. The PAE regularizes the NF by reducing its dimensionality. A similar reinterpretation of an NF as a VAE has been discussed in Gritsenko et al. (2019). The PAE uses an AE as a dimensionality reduction tool, which for powerful architectures and high latent dimensions has a small reconstruction error, yet regularizes the sensitivity of the NF to low variance components.

The PAE uses an NF prior, which differs from a VAE with an NF posterior (Rezende & Mohamed, 2015): for the latter the training relies on the standard ELBO optimization. The PAE uses an NF to learn the prior after the AE training is complete. VI and ELBO are never used. This separation allows for fast, easy and stable training of both the AE and the NF.

Other approaches that try to improve the VAE sample quality include β-VAEs (Higgins et al., 2017) and 2-Stage-VAEs (Dai & Wipf, 2019). β-VAEs need to balance reconstruction error and sample quality and require an additional costly hyperparameter tuning (Eq. 18). 2-stage-VAEs combine two consecutive VAE's, one for the purpose of data compression, where the KL term in the ELBO is suppressed (small β in Eq. 18), and a second one for latent space density estimation. This is similar to the PAE, and achieves similarly high quality samples. Our approach shows that VI and ELBO are not essential to obtain these results: in the first stage the KL term of the ELBO (Eq. 18) can be completely dropped and no sampling from q is needed (the VAE becomes an AE), and in the second stage the VAE can be replaced by a powerful NF, which is useful for downstream tasks such as OoD detection.

Table 1: FID scores and reconstruction errors at latent space dimensionality $K=32$ measured on 10000 samples/test data. The errors on the FIDs are smaller than the differences (order 1).

Model	AE	VAE	β-VAE	PAE
sample FID score $\downarrow$	102.9	102.4	56.3	**35.4**
reconstruction FID score $\downarrow$	**31.5**	32.3	51.3	**31.5**
mean reconstruction error $\downarrow$	**0.077**	0.084	0.114	**0.077**

Figure 2: Sample qualities at $K=32$: From left to right we show samples from an AE, samples from an equivalent VAE without parameter-tuning, samples from a β-VAE with tuned parameters ($\beta=100, C=15$), samples from the PAE and reconstructions of test data. Corresponding FID scores are listed in Table 1: the PAE reaches better FIDs than the β-VAE without requiring parameter-tuning.

Xiao et al. (2019) propose a similar conjunction of AE and normalizing flow (the generative latent flow) and achieve high sample quality, but do not explore the probabilistic interpretation for downstream tasks. Generative moment matching networks (Li et al., 2015) have also been proposed to be used in a 2-stage PAE-like set up, consisting of an auto-encoder and a mapping of a Gaussian to the encoded distribution. The second stage is non-invertible and trained with a moment-matching objective. GMMNs have been shown to produce high quality samples on MNIST and the Toronto face dataset. Wasserstein auto-encoders (Tolstikhin et al., 2018) employ yet another training objective to match the encoded distribution to a given prior. WAEs achieve high sample quality with comparable FID scores to the PAE, but do not provide a density estimate that is required for other tasks.

Other approaches that can be interpreted as regularized normalizing flows and that address cases where the data is confined to a lower dimensional manifold include $\mathcal{M}_{(e)}$-flows (Brehmer & Cranmer, 2020) and relaxed injective probability flows (Kumar et al., 2020). Instead of separating the tasks of compression and density estimation of the manifold, these models regularize the flow itself.

Out-of-distribution detection with generative models has recently attracted a lot of attention, triggered by the finding that state-of-the art generative models such as VAE, GLOW and MAF fail in this task on a number of standard datasets (Nalisnick et al., 2018). (Nalisnick et al., 2018) proposed a fix that uses likelihood ratios as an OoD metric instead of the likelihood itself. While achieving good OoD accuracy, their method requires fine-tuning a noise parameter, which in turn requires some kind of correctly labeled in- and out-of-distribution data sets. We demonstrate in our experiments (Section 4.2) that the PAE density estimate achieves high OoD accuracy without fine-tuning or requiring a labeled outlier training set.

There is a long list of works that have addressed the problem of image inputation and denoising with generative models, (Rezende et al., 2014; Mattei & Frellsen, 2018a,b; Dong et al., 2015; Jin et al., 2016; Putzky & Welling, 2017; Ulyanov et al., 2017; Bora et al., 2018). Most of them, however, do not address the crucial task of uncertainty estimation. A single point estimate is useless in many applications if its precision cannot be quantified. Our approach has two further advantages that many other methods do not share: 1) It only requires a single training of the generative model and can be applied to different corruption types without retraining. 2) It uses a robust Bayesian framework that comes with well established theoretical guarantees and well known behavior in reconstruction and uncertainty estimation, while it limits the use of the generative model to the task of dimensionality reduction, where its performance can be easily tested and verified.

4 Experiments

We train PAEs on the MNIST (Lecun et al., 1998), Fashion-MNIST (Xiao et al., 2017) and Celeb-A (Liu et al., 2015) training data sets. We preprocess these data sets with standard procedures: we dequantize and rescale pixel values to the interval [-0.5,0.5]. The celeb-A samples are cropped to the central 128x128 pixels and then downsampled to 64x64 pixels. The first building block of the PAE is an auto-encoder: The AE encoder and decoder networks are based on the infoGAN architecture (Chen et al., 2016). We train the AEs until convergence (at least 500 epochs) using an initial learning rate of 1e-3 which we reduce to 1e-4 after 400 epochs. The second building block of the PAE is a normalizing flow. This NF is trained on the AE latent space (Eq. 4). Depending on the latent space dimensionality we use different architectures for the NF. For $K<=32$, we find that a realNVP architecture (Dinh et al., 2016) with random permutations is sufficient to achieve samples of the same quality as the reconstructions. For higher K, we switch to more expressive neural spline flows (Durkan et al., 2019) and trainable permutations (Kingma & Dhariwal, 2018a). All NFs are trained for at least 100 epochs until convergence. For comparison we also train (β)-VAEs using the same architecture and training procedure as for the AE. The VAEs are trained with a Gaussian likelihood (allowing for a direct comparison with the PAE), a standard normal distribution as prior and using the mean field approximation for the posterior. In β-VAEs (Higgins et al., 2017) the training objective is modified by two tunable parameters, β and C,

$$\mathcal{L}_{\beta-\text{VAE}} = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \left[p_\theta(\boldsymbol{x}|\boldsymbol{z}) \right] - \beta |\text{D}_{\text{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) - C|. \tag{18}$$

We iterate over several parameter combinations until we achieve optimal sample quality.

4.1 Sample and reconstruction quality

We compare the sample and reconstruction quality of the PAE with an equivalent VAE. We quantify sample qualities with the Frechet-Inception Distance (FID) (Heusel et al., 2017) to the test data. We also quote FIDs for reconstructions, not only because they quantify reconstruction quality, but also because they constitute a lower bound to the sample FID that is achievable with a model. The reconstruction quality is further quantified with the mean reconstruction error on the test data (averaged over all pixels).

In Fig. 2 we show F-MNIST samples of different models at latent space dimensionality $K=32$. The samples of the VAE without parameter-tuning are very similar to samples from an AE (to sample from the AE, we draw latent space points from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{1})$ and pass them through the AE generator). The low VAE sample quality is a consequence of the likelihood term dominating the ELBO in this setting. This can be cured by fine-tuning the relative weighting of likelihood and KL terms (Eq. 18). We run 5 different parameter combinations (changing the parameters based on the result of the previous combination) and show results for the combination with the highest FID score. In contrast to the β-VAE, the PAE does not require any iterative parameter search to achieve high quality samples. In Table 1 we list the corresponding FID scores.

While the β-VAE balances a trade-off between reconstruction error and sample quality, the PAE is trained to reach an optimal reconstruction error. Matching the latent space distribution of the AE with an NF then allows to achieve samples with similar quality as the reconstructions. Because of this, the PAE reaches lower reconstruction errors (Table 1) and higher sample quality than the β-VAE. At the same time it does so at lower computational (and human) cost: Several β-VAE models have to be trained to find the optimal parameter combination, while the PAE only needs to be trained once. We show PAE samples for F-MNIST and PAE samples and interpolations for celeb-A in Figs. 3 and 4. The interpolations are obtained by mapping two random images from the test data set into the latent space, fitting a linear interpolation between the latent points and mapping equidistant points along this line back into the image space.

4.2 Downstream tasks: Out of Distribution (OoD) detection, denoising, inpainting

In Section 2.2 we derived an approximate, PAE-based density estimator which we apply as an outlier detector. OoD detection with generative models has recently attracted a lot of attention, since their probability estimates have been shown to be poor outlier-detectors: Nalisnick et al. (2018) found that different generative models, including VAEs and NFs, can assign consistently higher log probabilities to OoD data than training data. One particular combination of data sets for which this

Figure 3: F-MNIST samples from the PAE at K=64 with FID=31.7 (left) and K=128 with FID=28.0 (right).
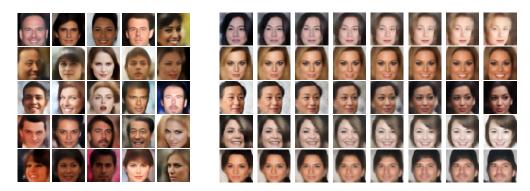

Figure 4: PAE performance on Celeb-A at K=64. Samples (left) reach FID=49.2 (reconstructions FID=44.0). Right: Interpolations between samples from the test set.

has been observed is Fashion-MNIST and MNIST, where a model trained on the former assigns higher probability to the latter. We use this pair to demonstrate the discriminating power of the PAE outlier detector. We compare our results to other recently suggested outlier detectors: Choi et al. (2018) propose the use of the Watanabe-Akaike information criterion (WAIC) and compare it to several other techniques. Here, we also quote their results using the variational information bottleneck OoD detector (VIB) (Alemi et al., 2018a), since it performed best in their experiments on the data pairs we consider here, and the importance weighted auto-encoder (IWAE) (Burda et al., 2016). Different to our PAE-based detector, the VIB OoD-detector requires a labeled training data set. Ren et al. (2019) propose to use the likelihood ratio between two models: one trained on in-distribution data, the other trained on perturbed in-distribution data. Their method requires hyper-parameter tuning (the amount of perturbation) for which some OoD data needs to be used.

The PAE-based outlier detector, Eq. 15 combines two terms, a likelihood term measuring the reconstruction error, and a latent space density term measuring the probability of the encoded data point in latent space. We find that the reconstruction error is more informative about outliers at low latent space dimensionalities (K<16), while it quickly becomes irrelevant for K>16. At higher K, the reconstruction error becomes very small (as illustrated in Fig. 5), while the latent space density becomes very informative about outliers. If we allowed for an additional hyper-parameter, we could find an optimal weighting between likelihood and latent space term for each K. However, the aim of the PAE is to omit hyper-parameter tuning, and so we quote results for the two extreme settings: 1) OoD detection with only the latent space density at K=64. 2) OoD detection with ony the AE reconstruction error at K=4. We compare our results for these settings to other methods in terms of the Area Under Receiver Operating Characteristic (AUROC) curve in Table 2. As OoD data we use MNIST, OMNIGLOT (Lake et al., 2015) as well as vertically (vflip) and horizontaly (hflip) flipped F-MNIST test data. The in distribution data is the F-MNIST test set. We find that the PAE outlier detector yields consistently high AUROC values without any parameter tuning, outperforming the other methods in MNIST, OMNIGLOT and hflip. The good performance of the PAE-detector is owed to the regularization imposed by dimensionality reduction, as it eliminates the low variance components that dominate the likelihood but are not informative about outliers. The likelihood ratio (Ren et al., 2019) similarly upweights the more informative part of the data by adding noise, and

Table 2: OoD detection accuracy quantified by the AUROC (↑) for models trained on F-MNIST.

Outlier	MNIST	OMNIGLOT	FMNIST-hflip	FMNIST-vflip
PAE density (this work)	**0.997**	**0.981**	**0.698**	0.891
AE reconstruction error (this work)	0.986	0.916	0.689	0.880
likelihood ratios (Ren et al. 2019)	0.996	-	-	-
VIB (Choi et al. 2018)	0.941	0.943	0.667	**0.902**
WAIC (Choi et al. 2018)	0.766	0.796	0.624	0.704
IWAE (Choi et al. 2018)	0.423	0.568	0.594	0.668

Figure 5: MNIST input images (outermost left plot) and their reconstructions with with AEs trained on F-MNIST. As K increases from left to right (4,8,16,32), the reconstructions transition from being visually close to F-MNIST (high reconstruction error) to visually close to the actual input (low reconstruction error). On the right we compare OoD detection based on reconstruction error (yellow) to latent space density (blue) as a function of K in terms of the ratio of True Positive (TP) over False Positive (FP) rate at a False Negative rate of 5%. At low K the reconstruction error is a better OoD detector, at higher K the latent space density outperforms it.

can be viewed as a form of regularization since it also eliminates information from low variance components. However, it requires fine-tuning the type and amount of perturbation on some additional OoD data set. We achieve competitive results without parameter tuning and without requiring any kind of OoD data for calibration.

Another downstream task that deep probabilistic models are useful for is posterior analysis of incomplete or corrupted data. We address this task with the PAE following Section 2.3. In the left panel of Figure 6 we show 3 examples of corrupted MNIST test data (left column), reconstructions at the peak of the posterior (MAP) (middle) and and true images (right column). The first example of the masked 3 illustrates why uncertainty quantification in terms of posterior analysis is important: The corrupted data is compatible not only with a 3, but also a 5. We see in the middle panel[2] that the corresponding posterior is complicated, possibly multi-modal. Drawing from this posterior results in the samples at the top of the right panel. The masked samples in the bottom show that they are all compatible with the input data. The samples are mostly 3s, with occasional occurrence of 5 (possibly

---

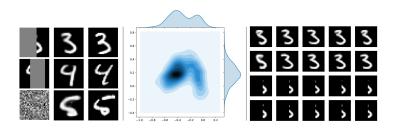[2]plot created with the mcmcplot package (Miles, 2018)

Figure 6: Left panel: corrupted input data with masks shown in gray (left column), MAP reconstructions (middle column) and underlying truth (right column). Middle panel: posterior density of two latent space variables (out of 8) for the first example of the masked 3. Right panel: Samples drawn from the posterior, the bottom rows show the same samples with the mask.

associated with the secondary peak in the latent space posterior). While the digit 5 can be consistent with the data, this is true for only a small fraction of the prior and the posterior reflects this.

5   Discussion and conclusion

We show that an auto-encoder combined with a normalizing flow is a powerful, easy and fast to train probabilistic model (PAE). The PAE can be used for sample generation, outlier detection and probabilistic image inpainting and denoising. When compared to baselines such as (β-)VAEs it outperforms these in terms of reconstruction error, sample quality, computational time, and downstream tasks such as outlier detection, without the need of additional parameter-tuning.

The PAE can be viewed either as a nonlinear generalization of a low rank probabilistic PCA or as a regularized form of an normalizing flow. The regularization through dimensionality reduction avoids the issues that plague normalizing flows, such as zero or near zero variance pixels, which can flaw density estimation based outlier detection. This allows the PAE to achieve state of the art results in out-of-distribution problems where normalizing flows are known to fail.

Acknowledgements

References

Alemi, A. A., Fischer, I., and Dillon, J. V. Uncertainty in the variational information bottleneck. *CoRR*, abs/1807.00906, 2018a. URL `http://arxiv.org/abs/1807.00906`.

Alemi, A. A., Poole, B., Fischer, I., Dillon, J. V., Saurous, R. A., and Murphy, K. Fixing a broken ELBO. In Dy, J. G. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 159–168. PMLR, 2018b. URL `http://proceedings.mlr.press/v80/alemi18a.html`.

Böhm, V., Lanusse, F., and Seljak, U. Uncertainty Quantification with Generative Models. *arXiv e-prints*, art. arXiv:1910.10046, October 2019.

Bora, A., Price, E., and Dimakis, A. G. Ambientgan: Generative models from lossy measurements. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL `https://openreview.net/forum?id=Hy7fDog0b`.

Brehmer, J. and Cranmer, K. Flows for simultaneous manifold learning and density estimation. *arXiv e-prints*, art. arXiv:2003.13913, March 2020.

Burda, Y., Grosse, R. B., and Salakhutdinov, R. Importance weighted autoencoders. In Bengio, Y. and LeCun, Y. (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL `http://arxiv.org/abs/1509.00519`.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *CoRR*, abs/1606.03657, 2016. URL `http://arxiv.org/abs/1606.03657`.

Choi, H., Jang, E., and Alemi, A. Waic, but why?: Generative ensembles for robust anomaly detection. *CoRR*, abs/1810.01392, 2018. URL `http://arxiv.org/abs/1810.01392`.

Dai, B. and Wipf, D. P. Diagnosing and enhancing VAE models. *CoRR*, abs/1903.05789, 2019. URL `http://arxiv.org/abs/1903.05789`.

Dinh, L., Krueger, D., and Bengio, Y. NICE: non-linear independent components estimation. *CoRR*, abs/1410.8516, 2014. URL `http://arxiv.org/abs/1410.8516`.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real NVP. *CoRR*, abs/1605.08803, 2016. URL `http://arxiv.org/abs/1605.08803`.

Dong, C., Loy, C. C., He, K., and Tang, X. Image super-resolution using deep convolutional networks. *CoRR*, abs/1501.00092, 2015. URL `http://arxiv.org/abs/1501.00092`.

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. Neural spline flows. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 7509–7520, 2019. URL `http://papers.nips.cc/paper/8969-neural-spline-flows`.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing*

Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, pp. 2672–2680, 2014. URL `http://papers.nips.cc/paper/5423-generative-adversarial-nets`.

Grathwohl, W., Chen, R. T. Q., Bettencourt, J., Sutskever, I., and Duvenaud, D. K. FFJORD: free-form continuous dynamics for scalable reversible generative models. *CoRR*, abs/1810.01367, 2018. URL `http://arxiv.org/abs/1810.01367`.

Gritsenko, A. A., Snoek, J., and Salimans, T. On the relationship between normalising flows and variational- and denoising autoencoders, 2019. URL `https://openreview.net/forum?id=HklKEUUY_E`.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 6626–6637, 2017. URL `http://papers.nips.cc/paper/7240-gans-trained-by-a-two-time-scale-update-rule-converge-to-a-local-nash-equilibrium`.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=Sy2fzU9gl`.

Jin, K. H., McCann, M. T., Froustey, E., and Unser, M. Deep convolutional neural network for inverse problems in imaging. *CoRR*, abs/1611.03679, 2016. URL `http://arxiv.org/abs/1611.03679`.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pp. 10236–10245, 2018a. URL `http://papers.nips.cc/paper/8224-glow-generative-flow-with-invertible-1x1-convolutions`.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pp. 10236–10245, 2018b. URL `http://papers.nips.cc/paper/8224-glow-generative-flow-with-invertible-1x1-convolutions`.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. URL `http://arxiv.org/abs/1312.6114`.

Kumar, A., Poole, B., and Murphy, K. Regularized autoencoders via relaxed injective probability flow. *CoRR*, abs/2002.08927, 2020. URL `https://arxiv.org/abs/2002.08927`.

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. ISSN 0036-8075. doi: 10.1126/science.aab3050. URL `https://science.sciencemag.org/content/350/6266/1332`.

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., and Winther, O. Autoencoding beyond pixels using a learned similarity metric. In Balcan, M. and Weinberger, K. Q. (eds.), *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 1558–1566. JMLR.org, 2016. URL `http://proceedings.mlr.press/v48/larsen16.html`.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pp. 2278–2324, 1998.

Ledoit, O. and Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.*, 88(2):365–411, February 2004. ISSN 0047-259X. doi: 10.1016/S0047-259X(03)00096-4. URL `https://doi.org/10.1016/S0047-259X(03)00096-4`.

Ledoit, O. and Wolf, M. Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions. *J. Multivariate Analysis*, 139:360–384, 2015. doi: 10.1016/j.jmva.2015.04.006. URL https://doi.org/10.1016/j.jmva.2015.04.006.

Li, Y., Swersky, K., and Zemel, R. S. Generative moment matching networks. In Bach, F. R. and Blei, D. M. (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 1718–1727. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/li15.html.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Mattei, P.-A. and Frellsen, J. MIWAE: Deep Generative Modelling and Imputation of Incomplete Data. *arXiv e-prints*, art. arXiv:1812.02633, Dec 2018a.

Mattei, P.-A. and Frellsen, J. Leveraging the Exact Likelihood of Deep Latent Variable Models. *arXiv e-prints*, art. arXiv:1802.04826, Feb 2018b.

Miles, P. prmiles/mcmcplot: v0.0.0. *Zenodo*, Aug 2018. doi: 10.5281/zenodo.1342783.

Nalisnick, E. T., Matsukawa, A., Teh, Y. W., Görür, D., and Lakshminarayanan, B. Do deep generative models know what they don't know? *CoRR*, abs/1810.09136, 2018. URL http://arxiv.org/abs/1810.09136.

Papamakarios, G., Nalisnick, E. T., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. Normalizing flows for probabilistic modeling and inference. *CoRR*, abs/1912.02762, 2019. URL http://arxiv.org/abs/1912.02762.

Putzky, P. and Welling, M. Recurrent Inference Machines for Solving Inverse Problems. *arXiv e-prints*, art. arXiv:1706.04008, Jun 2017.

Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., DePristo, M. A., Dillon, J. V., and Lakshminarayanan, B. Likelihood ratios for out-of-distribution detection. *CoRR*, abs/1906.02845, 2019. URL http://arxiv.org/abs/1906.02845.

Rezende, D. J. and Mohamed, S. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pp. 1530–1538, 2015. URL http://jmlr.org/proceedings/papers/v37/rezende15.html.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pp. 1278–1286, 2014. URL http://jmlr.org/proceedings/papers/v32/rezende14.html.

Rippel, O. and Adams, R. P. High-dimensional probability estimation with deep density models. *CoRR*, abs/1302.5125, 2013. URL http://arxiv.org/abs/1302.5125.

Seljak, U. and Yu, B. Posterior inference unchained with EL_2O. *arXiv e-prints*, art. arXiv:1901.04454, Jan 2019.

Tipping, M. E. and Bishop, C. M. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999. doi: 10.1111/1467-9868.00196. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00196.

Tolstikhin, I. O., Bousquet, O., Gelly, S., and Schölkopf, B. Wasserstein auto-encoders. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=HkL7n1-0b.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. Deep Image Prior. *arXiv e-prints*, art. arXiv:1711.10925, Nov 2017.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017. URL http://arxiv.org/abs/1708.07747.

Xiao, Z., Yan, Q., Chen, Y., and Amit, Y. Generative latent flow: A framework for non-adversarial image generation. *CoRR*, abs/1905.10485, 2019. URL `http://arxiv.org/abs/1905.10485`.