Adversarial robustness of $\beta-$VAE through the lens of local geometry

Asif Khan [1] **Amos Storkey** [1]

Abstract

Variational autoencoders (VAEs) are susceptible to adversarial attacks. An adversary can find a small perturbation in the input sample to change its latent encoding non-smoothly, thereby compromising the reconstruction. A known reason for such vulnerability is the latent space distortions arising from a mismatch between approximated latent posterior and a prior distribution. Consequently, a slight change in the inputs leads to a significant change in the latent space encodings. This paper demonstrates that the sensitivity at any given input exploits the directional bias of a stochastic pullback metric tensor induced by the encoder network. The pullback metric tensor captures how the infinitesimal volume changes from the input space to the latent space. Thus, it can be viewed as a lens to analyse the effect of small changes in the input leading to distortions in the latent space. We propose robustness evaluation scores using the eigenspectrum of a pullback metric. Moreover, we empirically show that the scores correlate with the robustness parameter β of the $\beta-$VAE.

1. Introduction

Variational autoencoders (VAEs) (Kingma and Welling, 2013) are deep generative models with an encoder-decoder network. The encoder parameterises the variational distribution over latent variables conditioned on data samples, and the decoder estimates the data distribution through the latent distribution. Thus, VAEs serve a dual purpose of estimating data density and providing a rich representation space with uncertainty quantification. Recently several works have shown application of VAEs to high fidelity image generation (Vahdat and Kautz, 2020), music generation (Roberts et al., 2017), video generation (Wu et al., 2021), and many more. However, like other machine learning models, VAEs are also vulnerable to adversarial attacks, as demonstrated in several recent works (Tabacof et al., 2016; Gondim-Ribeiro et al., 2018; Willetts et al., 2019; Kos et al., 2018; Kuzina et al., 2021). In a typical setup, an adversary can attack a VAE by learning the small perturbation resulting in a significant change in latent encoding. This mechanism takes the form of an optimisation problem which is generally solved using stochastic gradient methods (Chakraborty et al., 2018; Szegedy et al., 2013). (Sun et al., 2021) provide a comprehensive overview of different types of attacks on VAEs as well as other generative models.

The primary reason for the vulnerability of VAEs is the distortion in the latent space resulting due to the mismatch between approximated posterior and latent space prior, also known as the *KL gap*. Hence, the latent space is non-smooth, and the representations of similar inputs tend to be significantly distant in the space under the Euclidean metric. Methods such as (Mathieu et al., 2019; Willetts et al., 2019) have emphasised the importance of reducing the KL gap for learning disentangled latent space that also improves the robustness of VAEs. $\beta-$ VAE (Higgins et al., 2016) formulation introduces a parameter β to directly control the gap. Other methods (Chen et al., 2018b; Kim and Mnih, 2018; Esmaeili et al., 2019) utilise the total correlation term to disentangle the latent coordinates of VAEs that also smooths the latent space proving helpful in improving the robustness. The limitation with these approaches is that they cause over smoothing of the reconstructed samples and require a careful training mechanism to balance the regularisation term. To address the problem (Willetts et al., 2019) utilise the TC term in hierarchical VAEs that provides robustness along with sharp reconstruction. However, most robustness methods use regularisation terms, which do not provide meaningful insights for quantifying robustness. Therefore, a notion of small change in the input to the large change in latent space is not well established. Moreover, the comparison of these schemes relies on the visual inspection of distorted images at a different magnitude of adversarial loss.

Much recently (Camuto et al., 2021) proposed a theoretical framework that takes into account the uncertainty of encoder for studying the robustness of VAEs. However, the attacks in the input space do not consider the effect of geometry. Another recent work (Kuzina et al., 2021) proposed asym-

[1]School of Informatics, University of Edinburgh, United Kingdom. Correspondence to: Asif Khan <asif.khan@ed.ac.uk>.

The 2022 ICML Workshop on New Frontiers in Adversarial Machine Learning. Baltimore, Maryland, USA. Copyright 2022 by the author(s).

metric KL term to capture the difference between the latent representation of input and its perturbation. They obtain the ϵ using the Jacobian of the mean latent code evaluated at the input perturbation. Nevertheless, their optimisation objective does not provide any geometrical insights. Likewise, they do not consider the standard deviation term when computing Jacobian. Thus, do not account for the uncertainty in the representation space. Our paper shows that the geometry induced by the stochastic encoder mapping provides the intuition behind the sensitivity of VAEs that can be a valuable tool for understanding robustness.

The central theme of our paper is to view the adversarial attack problem through the lens of manifold geometry. Unlike existing approaches treating the input space as euclidean, we propose to utilise the stochastic pullback-metric tensor induced by the encoder map to measure the distance in the image space. We demonstrate that the distortion in the latent space results in a directional bias appearing in the form of an anisotropic metric tensor. We show that an adversary can design a one-step attack by moving along the dominant eigendirection of the local metric tensor. To quantify robustness, we propose scores based on the eigenspectrum of the local metric tensor. We hypothesise that defence methods would influence the local geometry of the metric tensor. We use $\beta-$VAE and demonstrate that the proposed scores correlate with the β parameter of $\beta-$VAE used to control robustness. To our knowledge, such a geometric view of the robustness of VAEs has not been previously investigated.

2. Background

2.1. $\beta-$Variational Autoencoder

$\beta-$VAE (Higgins et al., 2016) is a probabilistic encoder-decoder framework that simultaneously parameterises the latent distribution and emission distribution using deep neural networks. Consider a sample $\mathbf{x} \in \mathbf{X} = \mathbb{R}^N$ drawn from unknown data distribution $p(\mathbf{x})$, VAE learns an approximate posterior distribution $\mathsf{q}_\theta(\mathbf{z}|\mathbf{x})$ over latent variables $\mathbf{z} \in \mathbf{Z} = \mathbb{R}^{d_z}$ using a stochastic encoder network, and an emission distribution $p_\phi(\mathbf{x}|\mathbf{z})$ using a stochastic decoder network. The parameters θ of encoder network and ϕ of a decoder network are obtained by maximising the evidence lower bound (ELBO),

$$\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta KL[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})] \quad (1)$$

where KL stands for Kullback-Leibler divergence (Kullback and Leibler, 1951), and a parameter β controls the smoothness of latent distribution, setting $\beta = 1$ is equivalent to a standard VAE (Kingma and Welling, 2013).

2.2. Adversarial Attacks on VAE

The adversarial attacks on VAEs assume access to a pre-trained encoder-decoder network. The adversary aims to exploit the capacity of VAE by finding small perturbations in the input sample that lead to a large change in its latent encoding or its reconstruction. In a supervised scenario, the adversary starts with the target image and finds a minimal change that can match the reconstruction to the target. In an unsupervised, the aim is to simply maximise the distance in latent codes, which in turn will compromise the reconstruction. Several recent developments have proposed mechanisms for designing an adversary as well as evaluating the robustness of existing deep generative models (Tabacof et al., 2016; Willetts et al., 2019). In our paper, we propose a geometrical viewpoint of unsupervised attack by analysing the metric tensor induced by the stochastic encoder network. We first introduce the unsupervised variational attack problem and later in Section 3 present our approach.

Consider, an encoder network $f_\theta : \mathbf{X} \rightarrow \mathbf{Z}$ the unsupervised attack optimises the objective (Gondim-Ribeiro et al., 2018),

$$\max_{\boldsymbol{\eta}} \quad d(f_\theta(\mathbf{x}), f_\theta(\mathbf{x}+\boldsymbol{\eta})) \qquad \text{subject to} \quad ||\boldsymbol{\eta}||_2 = \eta_0 \quad (2)$$

where η_0 is a small constant that decides the severity of the attack, and $d(.,.)$ is a distance function that measures the proximity in the latent space. A common approach of finding a corruption $\boldsymbol{\eta}$ is to use stochastic gradient methods (Sun et al., 2021; Willetts et al., 2019).

2.3. Latent Space Geometry

In this section, we introduce definitions from Riemannian geometry relevant for our work and then discuss related work on latent space geometry.

Definition 2.1. A $n-$dimensional manifold $\mathcal{M}$ is a topological space where for every $\mathbf{x} \in \mathcal{M}$ there exist a neighborhood region $\mathcal{V}_\mathbf{x}$ homeomorphic to $\mathcal{R}^n$ (Lee, 2006).

Definition 2.2. A Riemannian metric for a smooth manifold $\mathcal{M}$ is a bilinear, symmetric, positive definite map $\mathbf{G}_x : \mathcal{T}_x\mathcal{M} \times \mathcal{T}_\mathbf{x}\mathcal{M} \rightarrow \mathbb{R}$ for all $\mathbf{x} \in \mathcal{M}$, where $T_\mathbf{x}\mathcal{M}$ is a tangent plane at point x on the manifold (Lee, 2006).

Definition 2.3. A smooth manifold $\mathcal{M}$ with a Riemannian metric $\mathbf{G}$ defined on every point of a manifold is called a Riemannian Manifold (Lee, 2006).

Definition 2.4. Given a mapping $f : \mathcal{M} \rightarrow \mathcal{N}$ from smooth manifold $\mathcal{M}$ to $\mathcal{N}$, for any $\mathbf{x} \in \mathcal{M}$ the pull-back metric $\mathbf{G}_\mathbf{x}$ induced by the mapping f is given as $\mathbf{G}_\mathbf{x} = \mathbf{J}_{f(\mathbf{x})}^T \mathbf{G}_{f(\mathbf{x})} \mathbf{J}_{f(\mathbf{x})}$, where $\mathbf{J}_{f(\mathbf{x})} = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}$.

Several recent works treat the decoder mapping of VAEs as a smooth immersion and use the pullback as an induced metric in the latent space. The computation of such metrics has been useful in various applications such as drawing on manifold samples, latent space interpolation, clustering, motion planning and many more (Arvanitidis et al., 2017;

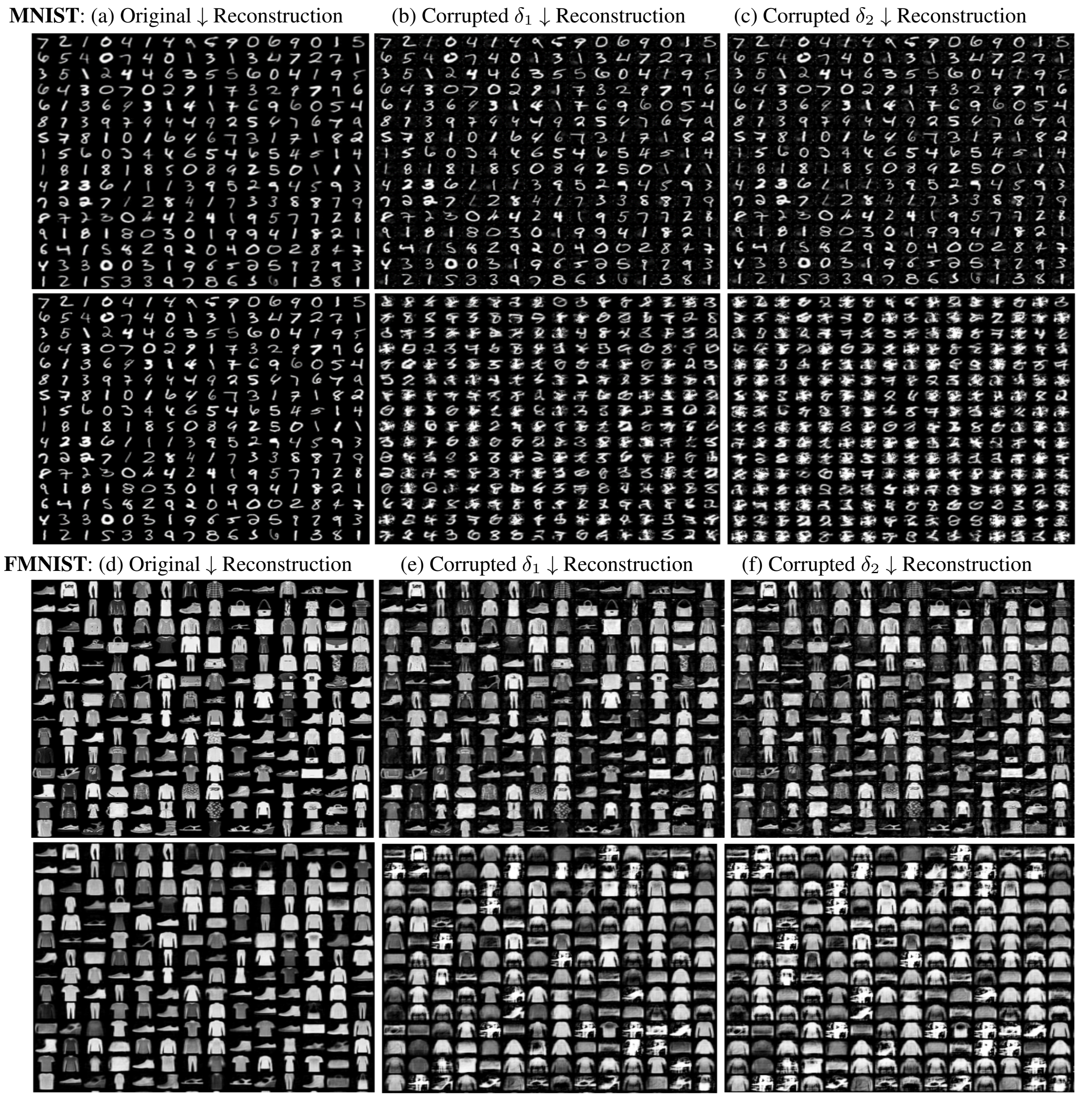


Figure 1. Illustration of adversarial attack along the dominant eigenvector of a stochastic pullback metric tensor. The first two rows are results on MNIST data and the bottom two on FashionMNIST dataset. We evaluate the reconstruction for original images and its two corrputed version with different step sizes $\delta_1 = 0.5233$ and $\delta_2 = 0.7443$. We observe moving along eigendirection doesn't effect the input image but significantly changes its reconstruction.

Yang et al., 2018; Hauberg, 2018; Chen et al., 2018a; Shao et al., 2018; Arvanitidis et al., 2020; Beik-Mohammadi et al., 2022). The computation of a pull-back in (Chen et al., 2018a; Yang et al., 2018) does not take into account the contribution of the uncertainty in the decoder mapping and are therefore limited in their ability to capture topological properties of manifold. (Arvanitidis et al., 2017; Hauberg, 2018) proposed to consider uncertainty of the decoder by treating the Gaussian decoder as a random projection of a deterministic manifold. This allows them to treat the reconstruction space as a random manifold and treat the pullback metric tensors as stochastic which proves useful in handling topological holes and low-density regions.

In our paper, we establish the connection between the directional bias of the stochastic pullback metric tensor of an encoder and the adversarial robustness of $\beta-$VAE. Previously (Zhao et al., 2019; Sun et al., 2020; Martin and Elster, 2020) studied the spectrum of Fisher information (pullback from probability simplex to the input space) of a classifier to investigate the robustness to adversarial perturbations. However, there is no such study for generative models to our knowledge. Also, the metric tensor considered in our paper considers the effect of uncertainty in the latent space, which is vital for understanding the latent distortions.

3. Method

We consider the encoder f_θ mapping samples from the data manifold to the latent manifold as a smooth immersion. We then utilise the pullback metric tensor induced by f_θ to connect the infinitesimal volume element around an input sample on data manifold with the infinitesimal volume around its representation on the latent space. Figure 2 shows an example of change in volume element under f. Unlike the existing methods relying on euclidean distance in the input space, we can use the pullback metric tensor to measure the infinitesimal distance given by a local inner product. We first express the adversarial optimisation problem in terms of the pullback metric tensor. Next, we show that due to the anisotropic nature of the metric tensor, the adversary can exploit the directional bias to design attacks. Finally, we propose an evaluation criterion based on the eigenspectrum of the metric tensor to evaluate the robustness of VAEs.

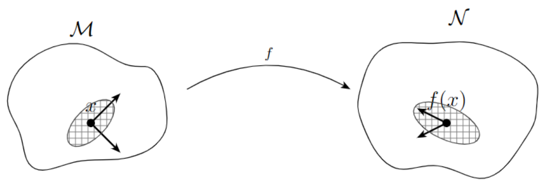


Figure 2. Illustration of the change in local geometry under a mapping f from the data manifold $\mathcal{M}$ to the latent manifold $\mathcal{N}$, the pullback metric tensor provides a connection between infinitesimal volume on the latent manifold and a input manifold.

Infinitesimal distance Since perturbation $\boldsymbol{\eta}$ is small in norm, we can approximate the distance $d(.,.)$ in Equation 5 as an infinitesimal distance along the latent manifold using the Taylor expansion of squared distance,

$$d(f_\theta(\mathbf{x}), f_\theta(\mathbf{x}+\boldsymbol{\eta})) = ||(f_\theta(\mathbf{x}) - f_\theta(\mathbf{x}+\boldsymbol{\eta})||_2^2 = \boldsymbol{\eta}^T \mathbf{G}_\mathbf{x} \boldsymbol{\eta},$$
$$\mathbf{G}_\mathbf{x} = \mathbf{J}_{f_\theta(\mathbf{x})}^T \mathbf{J}_{f_\theta(\mathbf{x})}, \quad \mathbf{J}_{f_\theta(\mathbf{x})} = \frac{\partial f_\theta}{\partial \mathbf{x}} \tag{3}$$

where $\mathbf{J}_{f_\theta(\mathbf{x})} \in \mathbb{R}^{d_z \times N}$ is a Jacobian matrix, d_z is the dimensionality of $\mathbf{Z}$ and N is the dimensionality of $\mathbf{X}$. The matrix $\mathbf{G}_\mathbf{x}$ is a symmetric, positive definite matrix known as a pullback metric tensor under the mapping f_θ. We can use it to measure local inner product for every $\mathbf{x}$ in the input space $\mathbf{x} \in \mathbf{X}$.

Stochastic pull-back metric tensor The encoder is a stochastic map given by the combination of $\boldsymbol{\mu}_\theta(\mathbf{x})$ and $\boldsymbol{\sigma}_\theta(\mathbf{x})$ as $f_\theta(\mathbf{x}) = \boldsymbol{\mu}_\theta(\mathbf{x}) + \boldsymbol{\epsilon} \odot \boldsymbol{\sigma}_\theta(\mathbf{x})$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_d)$. As a result, the Jacobian of f_θ is also a combination of the two maps $\mathbf{J}_f = \mathbf{J}_{\boldsymbol{\mu}(\mathbf{x})} + \boldsymbol{\epsilon} \odot \mathbf{J}_{\boldsymbol{\sigma}(\mathbf{x})}$ and the final pullback matrix $\mathbf{G}_\mathbf{x}$ is given by,

$$\mathbf{G}_\mathbf{x} = (\mathbf{J}_{\boldsymbol{\mu}(\mathbf{x})} + \boldsymbol{\epsilon} \odot \mathbf{J}_{\boldsymbol{\sigma}(\mathbf{x})})^T (\mathbf{J}_{\boldsymbol{\mu}(\mathbf{x})} + \boldsymbol{\epsilon} \odot \mathbf{J}_{\boldsymbol{\sigma}(\mathbf{x})})$$
$$= \mathbf{J}_{\boldsymbol{\mu}(\mathbf{x})}^T \mathbf{J}_{\boldsymbol{\mu}(\mathbf{x})} + \mathbf{J}_{\boldsymbol{\mu}(\mathbf{x})}^T \boldsymbol{\epsilon} \mathbf{J}_{\boldsymbol{\mu}(\mathbf{x})} + \mathbf{J}_{\boldsymbol{\sigma}(\mathbf{x})}^T \boldsymbol{\epsilon} \mathbf{J}_{\boldsymbol{\mu}(\mathbf{x})} + \mathbf{J}_{\boldsymbol{\sigma}(\mathbf{x})}^T \boldsymbol{\epsilon}^2 \mathbf{J}_{\boldsymbol{\sigma}(\mathbf{x})}$$

We can view the latent space as a random projection of a deterministic manifold. Under the assumption the sample paths from f_θ are smooth we can treat the metric tensor as a stochastic matrix. The metric of the random latent manifold can be estimated in expectation as $\hat{\mathbf{G}}_\mathbf{x} = \mathbb{E}_{\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})}[\mathbf{G}_\mathbf{x}]$. Since, $\boldsymbol{\epsilon}$ is a zero mean and unit covariance the $\mathbb{E}[\boldsymbol{\epsilon}] = 0$ and $\mathbb{E}[\boldsymbol{\epsilon}^2] = 1$ the final expected metric tensor is,

$$\hat{\mathbf{G}}_\mathbf{x} = \mathbf{J}_{\boldsymbol{\mu}(\mathbf{x})}^T \mathbf{J}_{\boldsymbol{\mu}(\mathbf{x})} + \mathbf{J}_{\boldsymbol{\sigma}(\mathbf{x})}^T \mathbf{J}_{\boldsymbol{\sigma}(\mathbf{x})} \tag{4}$$

Adversarial attack under expected local geometry We now use the expected metric tensor to reformulate the adversarial attack optimisation in Equation 5 as,

$$\max_{\boldsymbol{\eta}} \quad \boldsymbol{\eta}^T \hat{\mathbf{G}}_\mathbf{x} \boldsymbol{\eta} \tag{5}$$
$$\text{subject to} \quad ||\boldsymbol{\eta}||_2 = \eta_0 \tag{6}$$

To solve the problem we combine the constraints by introducing Lagrange multiplier $\boldsymbol{\lambda}$,

$$\max_{\boldsymbol{\eta}} \quad \boldsymbol{\eta}^T \hat{\mathbf{G}}_\mathbf{x} \boldsymbol{\eta} + \boldsymbol{\lambda}(\eta_0 - ||\boldsymbol{\eta}||_2^2) \tag{7}$$

The closed form solution of the above optimisation takes the form $\hat{\mathbf{G}}_\mathbf{x} \boldsymbol{\eta} = \boldsymbol{\lambda} \boldsymbol{\eta}$, where a pair $(\boldsymbol{\lambda}, \boldsymbol{\eta})$ represents the eigenvalue and eigenvector of the stochastic pullback metric tensor, the eigenvector with the largest eigenvalue corresponds to the direction of maximal change. Thus, for a given input $\mathbf{x}$ an adversary can exploit this directional bias to find perturbations that maximally change the encoding of the input sample as $\mathbf{x}_c = \mathbf{x} + \delta \boldsymbol{\lambda} \boldsymbol{\eta}$, where δ is a magnitude of the step along the eigendirection. This way adversary can devise a one-shot attack by choosing step size δ based on the distance between reconstruction of original input and its corrupted version as $||\hat{\mathbf{x}} - \hat{\mathbf{x}}_c||$.

We next, present the two evaluation scores and study their effect for different values of β.

Robustness evaluation A robustness method should suppress the maximum eigenvalue of the pullback metric tensor. Moreover, it should eliminate the directional bias resulting from the anisotropic distribution of eigenvalues of a pullback metric tensor. To quantify these two effects, we report the following two scores,

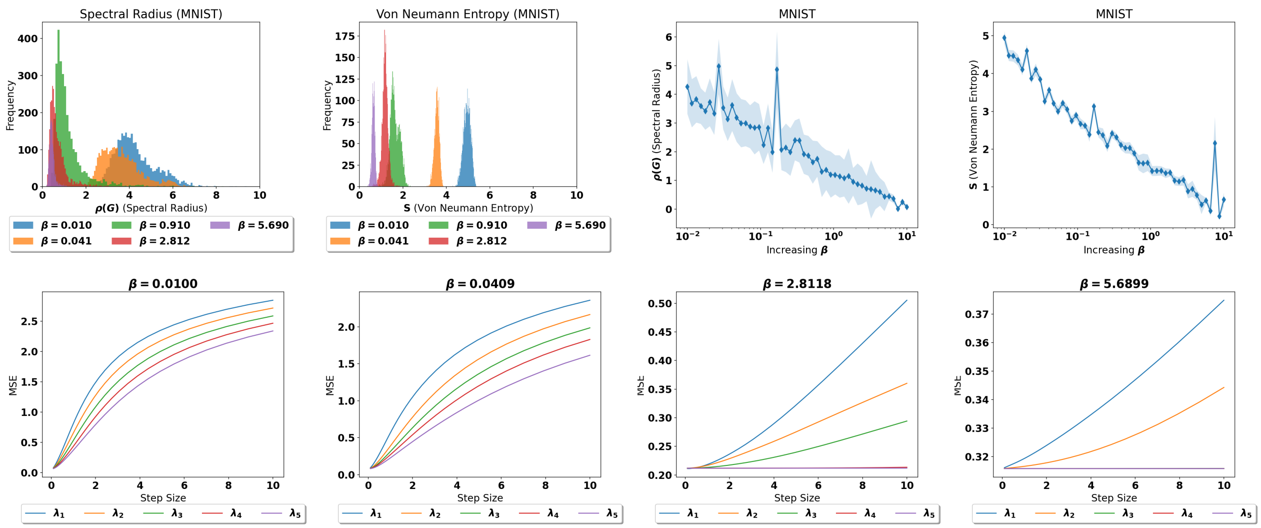
Spectral Radius for a matrix $\mathbf{G}$ is defined as,

$$\rho(\mathbf{G}) = \max\{|\boldsymbol{\lambda}|, \boldsymbol{\lambda} \text{ is an eigenvalue of } \mathbf{G}\} \tag{8}$$

A robust model will have a low value of spectral radius.

Von Neumann Entropy (Bengtsson et al., 2008) $\mathbf{S}$ of a metric tensor $\mathbf{G}$ is given by the Shannon entropy of its eigenvalues $\mathbf{S} = -\sum_k \boldsymbol{\lambda}_k \log \boldsymbol{\lambda}_k$. The high value would imply the eigendirections are anisotrpic resulting in a directional bias. Thus, a robust model will have a low value of $\mathbf{S}$.

(a) Robustness evaluation of $\beta-$VAE on MNIST.



(b) Robustness evaluation of $\beta-$VAE on FashionMNIST.

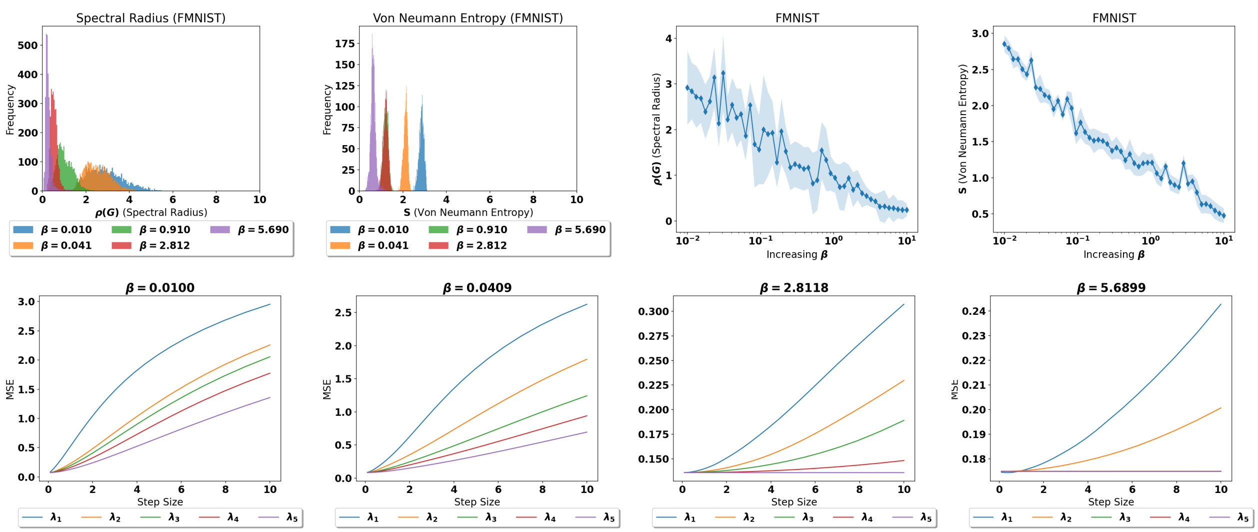


Figure 3. Figure (a), on left, we report the histogram of spectral radius and Von Neumann entropy (on test samples) for different values of β in β-VAE. On right, we report the average of two scores across test samples for increasing value of β. We observe, increasing the value of β suppresses the maximum eigenvalue of metric tensor, and the distribution of eigenspectrum gets more isotropic. In second row, we corrupt the test images along top five eigendirection (denoted by $\lambda_1, \lambda_2, \lambda_3, \lambda_4,$ and λ_5) with an increasing step size for different values of β. The plots describe the average MSE across test samples. We observe for higher value of β the average step size increases. Increasing the value of β reduces the KL gap which in turn minimises distortion in the latent space. Figure (b), demonstrates similar observations on FashionMNIST dataset.

4. Experiments and Results

In our experiments, we first demonstrate the vulnerability of a VAE along the dominant eigendirection of various input samples. Next, we report the proposed scores and study their relationship with the robustness parameter β in $\beta-$VAE. For this purpose, we sample 50 values of β with a logarithmic spacing between $[0.01, 10]$. We train the encoder-decoder model for each parameter and evaluate the scores on test data. We run experiments on MNIST (Deng, 2012) and FashionMNIST (Xiao et al., 2017) datasets.

Implementation details We use the same encoder-decoder architecture across all the experiments. The encoder network is a four-layer MLP with 256, 256, 512 and 32 hidden units. The latent space distribution is a multivariate Gaussian with mean and standard deviation parameterised by two 32×32 linear mappings. We use the standard zero mean and unit covariance prior in latent space. The decoder network is the inverse of an encoder with 32, 512, 256 and 256 hidden units. We use `tanh` as an activation function and batch-normalisation (Ioffe and Szegedy, 2015) before all activations.

4.1. Adversarial Attack

Figure 1 demonstrates the two instances of corruption along the dominant eigenvector of $\beta = 1$ VAE on MNIST and FashionMNIST datasets. For each dataset, the three columns in the first row are a set of original images and their corrupted version with a step size of $\delta = 0.5223$ and $\delta = 0.7443$. In a second row, we report their respective reconstructions. We observe that with $\delta = 0.5223$, the reconstruction significantly differs from the original images, and for $\delta = 0.7443$ gets much more severe, exposing the capacity of VAE. This result proves an attacker can exploit the directional bias of metric tensor to design one-shot attack.

4.2. Adversarial Robustness

Here, we investigate the connection between our proposed scores and the robustness of $\beta-$VAE. For this purpose, we took a sample size of 4000 test images and computed two scores $\rho(\hat{\mathbf{G}})$ and $\mathbf{S}$ for every sample point. Figure 3 first and the second column of row on (MNIST) and three (FashionMNIST) reports the histogram of the scores for four different values of β. We observe that the higher value of β suppresses the spectral radius. Similarly, the von Neumann entropy is decreased, demonstrating that the local directions get isotropic. Importantly, this indicates the adversary cannot exploit the directional bias for high values of β with $\boldsymbol{\eta}$ small in the norm. In the third and fourth columns of rows one and three, we report the mean and standard deviation of the scores computed for fifty increasing values of β. The results demonstrated that by reducing the KL gap, parameter β prevents distortion in the latent space eliminating the directional bias exploited by an adversary.

Next, we examine the connection between the step size δ and the strength of attacks under different values of β. We report the mean squared error (MSE) between an original image and its reconstruction under varying corruption rates along five dominant eigendirections. We generated 40 logarithmic spacing step size in range $[0.01, 10]$. Figure 3 second row (MNIST) and fourth row (FashionMNIST) demonstrate the MSE vs step size averaged across all test samples for four different values of β. We observe that for the small β, all five directions tend to get high MSE, and as β increases, it requires a larger step size to get a significant change in MSE. This effect shows that the probabilistic encoder-decoder model is more robust than its deterministic counterpart. The stochastic metric tensor comprises two terms resulting from the latent space's variational distribution. For small values of β, the encoder does not do well in quantifying the uncertainty and fails to match the prior; as an outcome, the latent space is more distorted, resulting in empty and low-density regions given by dominating eigendirections of the metric tensor. The distortions are reduced for a high value of β; accordingly, the second term can account for the uncertainty in latent space preventing the eigenvalues from getting large.

5. Conclusion and Future Scope

We have presented a geometrical perspective of adversarial attacks and introduced scores for measuring the robustness of $\beta-$VAE. We have shown that the sensitivity of the encoder at a given input can be interpreted using a stochastic pullback metric tensor. The spectral radius, and Von Neumann's entropy of a metric tensor correlate to robustness parameter β of $\beta-$VAE suggesting the scores are valuable criterion for evaluating robustness. A caveat with $\beta-$VAE is that increasing the β parameters tradesoff the capacity of representations with the quality of reconstruction resulting in mode collapse for large β. This further implies a tradeoff between spectral radius and the robustness. Therefore, to develop a good robustness scheme without compromising the capacity of representation we want to restrict the spectral radius within a specified range. Recently few mechanisms have been proposed to reduce the distortion in latent space and improve the robustness of VAEs (Willetts et al., 2019; Kuzina et al., 2021). We hypothesise that the benefits of such robustness measures can be better established geometrically by investigating their pullback metric tensors. We wish to investigate it in the future work.

In this paper, we consider that the local metric tensor in latent space as $\mathbf{G_z} = \mathbf{I}$. However, the stochastic nature of the decoder map results in a curved geometry in the latent space (Arvanitidis et al., 2017). We can take this geometry into account by computing $\mathbf{G_z}$ using the stochastic decoder network and using it to express the metric in the input space as, $\mathbf{G_x} = \mathbf{J}^T_{f_\theta(\mathbf{x})}\mathbf{G_z}\mathbf{J}_{f_\theta(\mathbf{x})}$, where $\mathbf{G_z} = \mathbf{J}^T_{g_\omega(\mathbf{z})}\mathbf{J}_{g_\omega(\mathbf{z})}$ can be computed with respect to stochastic decoder network g_ω. Connecting the two pullback tensor will be useful for understanding how the changes in input space are propagated to the reconstruction via latent space and might be a more valuable score for evaluating the robustness to adversarial perturbations. We are currently investigating these directions as a future extension.

A limitation of our current work is that we only consider the unsupervised attack when a target sample is not known. Furthermore, there can be different forms of attack by replacing l_2-norm with more general $p-$norms. We wish to study these in the future work.

References

Arvanitidis, G., Hansen, L. K., and Hauberg, S. (2017). Latent space oddity: on the curvature of deep generative models. *arXiv preprint arXiv:1710.11379*.

Arvanitidis, G., Hauberg, S., and Schölkopf, B. (2020). Geometrically enriched latent spaces. *arXiv preprint arXiv:2008.00565*.

Beik-Mohammadi, H., Hauberg, S., Arvanitidis, G., Neumann, G., and Rozo, L. (2022). Reactive motion generation on learned riemannian manifolds. *arXiv preprint arXiv:2203.07761*.

Bengtsson, I., Zyczkowski, K., and Milburn, G. J. (2008). Geometry of quantum states: an introduction to quantum entanglement by ingemar bengtsson and karol zyczkowski. *Quantum Information & Computation*, 8(8):860–860.

Camuto, A., Willetts, M., Roberts, S., Holmes, C., and Rainforth, T. (2021). Towards a theoretical understanding of the robustness of variational autoencoders. In *International Conference on Artificial Intelligence and Statistics*, pages 3565–3573. PMLR.

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., and Mukhopadhyay, D. (2018). Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*.

Chen, N., Klushyn, A., Kurle, R., Jiang, X., Bayer, J., and Smagt, P. (2018a). Metrics for deep generative models. In *International Conference on Artificial Intelligence and Statistics*, pages 1540–1550. PMLR.

Chen, R. T., Li, X., Grosse, R. B., and Duvenaud, D. K. (2018b). Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31.

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142.

Esmaeili, B., Wu, H., Jain, S., Bozkurt, A., Siddharth, N., Paige, B., Brooks, D. H., Dy, J., and Meent, J.-W. (2019). Structured disentangled representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2525–2534. PMLR.

Gondim-Ribeiro, G., Tabacof, P., and Valle, E. (2018). Adversarial attacks on variational autoencoders. *arXiv preprint arXiv:1806.04646*.

Hauberg, S. (2018). Only bayes should learn a manifold (on the estimation of differential geometric structure from data). *arXiv preprint arXiv:1806.04994*.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2016). beta-vae: Learning basic visual concepts with a constrained variational framework.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR.

Kim, H. and Mnih, A. (2018). Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kos, J., Fischer, I., and Song, D. (2018). Adversarial examples for generative models. In *2018 ieee security and privacy workshops (spw)*, pages 36–42. IEEE.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Kuzina, A., Welling, M., and Tomczak, J. M. (2021). Diagnosing vulnerability of variational auto-encoders to adversarial attacks. *arXiv preprint arXiv:2103.06701*.

Lee, J. M. (2006). *Riemannian manifolds: an introduction to curvature*, volume 176. Springer Science & Business Media.

Martin, J. and Elster, C. (2020). Inspecting adversarial examples using the fisher information. *Neurocomputing*, 382:80–86.

Mathieu, E., Rainforth, T., Siddharth, N., and Teh, Y. W. (2019). Disentangling disentanglement in variational autoencoders. In *International Conference on Machine Learning*, pages 4402–4412. PMLR.

Roberts, A., Engel, J., and Eck, D. (2017). Hierarchical variational autoencoders for music. In *NIPS Workshop on Machine Learning for Creativity and Design*, volume 3.

Shao, H., Kumar, A., and Thomas Fletcher, P. (2018). The riemannian geometry of deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 315–323.

Sun, H., Zhu, T., Zhang, Z., Xiong, D. J., Zhou, W., et al. (2021). Adversarial attacks against deep generative models on data: A survey. *arXiv preprint arXiv:2112.00247*.

Sun, K., Koniusz, P., and Wang, Z. (2020). Fisher-bures adversary graph convolutional networks. In *Uncertainty in Artificial Intelligence*, pages 465–475. PMLR.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Tabacof, P., Tavares, J., and Valle, E. (2016). Adversarial images for variational autoencoders. *arXiv preprint arXiv:1612.00155*.

Vahdat, A. and Kautz, J. (2020). Nvae: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 33:19667–19679.

Willetts, M., Camuto, A., Rainforth, T., Roberts, S., and Holmes, C. (2019). Improving vaes' robustness to adversarial attack. *arXiv preprint arXiv:1906.00230*.

Wu, B., Nair, S., Martin-Martin, R., Fei-Fei, L., and Finn, C. (2021). Greedy hierarchical variational autoencoders for large-scale video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2318–2328.

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Yang, T., Arvanitidis, G., Fu, D., Li, X., and Hauberg, S. (2018). Geodesic clustering in deep generative models. *arXiv preprint arXiv:1809.04747*.

Zhao, C., Fletcher, P. T., Yu, M., Peng, Y., Zhang, G., and Shen, C. (2019). The adversarial attack and detection under the fisher information metric. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5869–5876.