Diagnosing Vulnerability of Variational Auto-Encoders to Adversarial Attacks

Anna Kuzina
Vrije Universiteit Amsterdam
a.kuzina@vu.nl

Max Welling
Universiteit van Amsterdam
m.welling@uva.nl

Jakub M. Tomczak
Vrije Universiteit Amsterdam
j.m.tomczak@vu.nl

Abstract

In this work, we explore adversarial attacks on the Variational Autoencoders (VAE). We show how to modify data point to obtain a prescribed latent code (supervised attack) or just get a drastically different code (unsupervised attack). We examine the influence of model modifications (β-VAE, NVAE) on the robustness of VAEs and suggest metrics to quantify it. [1]

1 Introduction

Variational Autoencoders (VAEs) (Kingma & Welling, 2013; Rezende et al., 2014) are deep generative models used in various domains. Recent works show that deep Variational Autoencoders models can generate high-quality images (Vahdat & Kautz, 2020; Child, 2020). These works employ hierarchical structure (Ranganath et al., 2016), coupled with skip-connections (Maaløe et al., 2019; Sønderby et al., 2016). An additional advantage of VAE is that it has a meaningful latent space induced by the Encoder. This motivates us to explore the *robustness* of the resulting latent representations. *Adversarial attack* is one way to assess the robustness of a deep neural network. Robustness to the attacks is a crucial property for VAEs, especially in such applications as anomaly detection (An & Cho, 2015; Maaløe et al., 2019) or compression (Ballé et al., 2018).

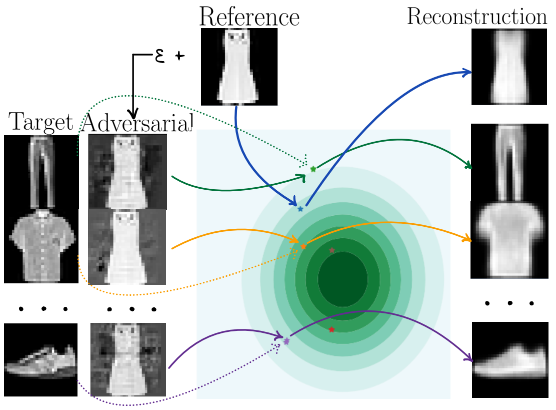


Figure 1: Example of a supervised attack on VAE with 2D latent space. Given a single reference point we learn additive perturbation ϵ, s.t. perturbed input has the same latent representation as a target image. We observe that a single reference point can be mapped to extremely different regions of the latent space.

Gondim-Ribeiro et al. (2018) propose to minimize the KL-divergence between an adversarial and a target input to learn an adversarial attack on the standard VAE model. We show that we can use a similar strategy to attack hierarchical VAEs. Willetts et al. (2019) suggest a modified VAE objective coupled with hierarchical structure as a way to increase VAE robustness to adversarial attacks. Camuto et al. (2020) show that β-VAE tend to be more robust to adversarial attacks in terms of proposed r-metric. In our work, we test both β-VAE and hierarchical NVAE and show that we can attack them successfully.

In Figure 1 we show an example of the *supervised* attack on VAE with the 2D latent space. We show that we can add noise ϵ to a reference image so that the resulting adversarial input (second column) is encoded to a new point in the latent space. This new point is defined by the target image (first column). In *unsupervised* attacks, on the other hand, we assume that the target image is not given. We show that it is still possible to construct an effective attack in this setting. The research goals of this work are the following:

- Defining robustness measures to understand how VAEs behave for adversarial attacks both in the latent space and the pixel space.
- Assessing robustness of hierarchical and β-VAEs to adversarial attacks.

[1] The code is published at https://github.com/AKuzina/attack_vae

2 METHODOLOGY

2.1 VARIATIONAL AUTOENCODERS

Let us consider a vector of observable random variables, $\mathbf{x} \in \mathcal{X}^D$ (e.g., $\mathcal{X} = \mathbb{R}$) sampled from the empirical distribution $p_e(\mathbf{x})$, and vectors of latent variables $\mathbf{z}_l \in \mathbb{R}^{M_l}$, $l = 1, 2, \ldots, L$. First, we focus on a model with $L = 1$ and the joint distribution $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})$. The marginal likelihood is then equal to $p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{z})\mathrm{d}\mathbf{z}$. VAEs exploit variational inference (Jordan et al., 1999) with a family of variational posteriors $\{q_\phi(\mathbf{z}|\mathbf{x})\}$, also referred to as encoders, that results in a tractable objective function, i.e., Evidence Lower BOund (ELBO):

$$\mathcal{L}(\phi, \theta) = \mathbb{E}_{p_e(\mathbf{x})}\left(\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \ln p_\theta(\mathbf{x}|\mathbf{z}) - \mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})\right]\right). \tag{1}$$

β-VAE (Higgins et al., 2016) uses a slightly modifier objective, weighting the second term by $\beta > 0$. In case of $L > 1$, we consider a hierarchical latent structure with the generative model of the following form: $p_\theta(\mathbf{x}, \mathbf{z}_1, \ldots, \mathbf{z}_L) = p_\theta(\mathbf{x}|\mathbf{z}_1)\prod_{l=1}^{L} p(\mathbf{z}_l|\mathbf{z}_{l+1})$, where $\mathbf{z}_{L+1} \equiv \emptyset$. There are various possible formulations of the family of variational posteriors, however, here we follow the proposition of Sønderby et al. (2016) where the inference model with skip-connections was proposed, namely:

$$q_\phi(\mathbf{z}_1, \ldots, \mathbf{z}_L|\mathbf{x}) = q_\phi(\mathbf{z}_L|\mathbf{x}) \prod_{i=1}^{L-1} q_{\theta,\phi}(\mathbf{z}_i|\mathbf{z}_{>i}, \mathbf{x}). \tag{2}$$

This formulation was used in NVAE (Vahdat & Kautz, 2020). It allows to share data-dependent information between the inference model and the generative model, because of the top-down structure.

2.2 ADVERSARIAL ATTACKS

An *adversarial attack* is a slightly deformed data point $\mathbf{x}$ that results in an undesired or unpredictable performance of a model. In the case of a VAE, we construct an adversarial input $\mathbf{x}^a$ as a deformation of a reference point $\mathbf{x}^r$ to satisfy the following conditions: (*i*) $\mathbf{x}^a$ should be close[2] to $\mathbf{x}^r$; (*ii*) the variational distribution $q_\phi(\mathbf{z}|\mathbf{x})$ for the $\mathbf{x}^r$ and the $\mathbf{x}^a$ should be different, and the same should hold for the conditional likelihood $p_\theta(\mathbf{x}|\mathbf{z})$. Next, we define a framework that can be used to define and compare adversarial attacks and, most importantly, evaluate the VAE robustness to them.

Attacks construction We define an adversarial point as a result of the additive perturbation of the reference point, $\mathbf{x}^a = \mathbf{x}^r + \epsilon^*$, where the perturbation ϵ^* is a solution of an optimization problem. We distinguish between the *supervised attack*, when we have access to a target point, $\mathbf{x}^t$ and the *unsupervised attack*, when $\mathbf{x}^t$ is not available. In the former case, we propose to solve the following optimization problem:[3]

$$\epsilon^* = \arg\min_{\|\epsilon\| \le 1} \mathrm{SKL}[q_\phi(\mathbf{z}|\mathbf{x}^r + \epsilon), q_\phi(\mathbf{z}|\mathbf{x}^t)]. \tag{3}$$

In the unsupervised case, we formulate the following optimization problem:

$$\epsilon^* = \arg\max_{\|\epsilon\| \le 1} \Delta[q_\phi(\mathbf{z}|\mathbf{x}^r + \epsilon), q_\phi(\mathbf{z}|\mathbf{x}^r)] = \arg\max_{\|\epsilon\| \le 1} \|\mathbf{J}^q_{\mathbf{x}^r}\epsilon\|_2^2 \tag{4}$$

where $\mathbf{J}^q_{\mathbf{x}^r}$ is the Jacobian of $q_\phi(\mathbf{z}|\mathbf{x}^r)$ at point $\mathbf{x}^r$. See Appx. A for the details for this objective.

Robustness measures It is important to define proper quantitative measures that will reflect our expectations from adversarial attacks discussed in Section 2.2. First, we focus on measuring differences in the latent spaces. For this purpose, we propose to use the following measure:

$$\Omega = \sum_{\mathbf{x}^r} \sum_{\mathbf{x}^a|\mathbf{x}^r} \mathrm{SKL}[q_\phi(\mathbf{z}|\mathbf{x}^a), q_\phi(\mathbf{z}|\mathbf{x}^r)], \tag{5}$$

where the value of ϵ is a solution from either (3) or (4).

Further, we would like to measure similarities between $\mathbf{x}^r$ and $\mathbf{x}^a$, and their reconstructions. Here, we propose to use the Multi-Scale Structural Similarity Index Measure (MSSSIM) (Wang et al., 2003) which is a perception-based measure calculated at different scales, $\mathrm{MSSSIM} \in [0, 1]$:

[2] Either visually or in terms of the pixel values.

[3] $\mathrm{SKL}[p_1, p_2] = \frac{1}{2}\mathrm{KL}[p_1||p_2] + \frac{1}{2}\mathrm{KL}[p_2||p_1]$ is the symmetric version of the Kullback-Leibler divergence

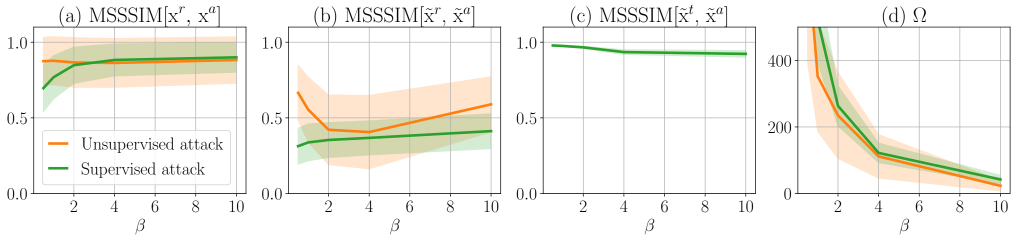


Figure 2: Robustness results for β-VAEs trained on Fashion MNIST dataset.

- $\mathrm{MSSSIM}[\mathbf{x}^r, \mathbf{x}^a]$: the similarity between a reference and the corresponding adversarial input;
- $\mathrm{MSSSIM}[\widetilde{\mathbf{x}}^r, \widetilde{\mathbf{x}}^a]$: the similarity between reconstructions of $\mathbf{x}^r$ and the corresponding $\mathbf{x}^a$;

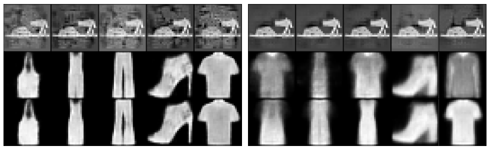
A successful adversarial attack would have large $\mathrm{MSSSIM}[\mathbf{x}^r, \mathbf{x}^a]$ (close to 1) and small $\mathrm{MSSSIM}[\widetilde{\mathbf{x}}^r, \widetilde{\mathbf{x}}^a]$. Moreover, for *supervised attacks* we will measure similarity between reconstructions of the target and the adversarial image, $\mathrm{MSSSIM}[\widetilde{\mathbf{x}}^t, \widetilde{\mathbf{x}}^a]$. Large value of the latter would indicate a successful supervised attack.

3 Experiments

In this section we consider attacking a 1-level VAE trained on the Fashion MNIST dataset (Xiao et al., 2017) and NVAE trained on the CelebA dataset (Liu et al., 2015). All the metrics are averaged over the reference, target, and adversarial inputs. In Appx. B.2 we provide details on the selection of the reference and target points for both datasets.

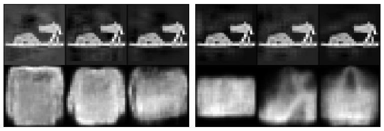
3.1 VAE and β-VAE

We start with the experiments on the VAE with one level of latent variables. We train both VAE and β-VAE (Higgins et al., 2016). The latter weights the KL-term in eq. (1) with $\beta > 0$. It is said that the larger values of β encourage disentangling of latent representations (Chen et al., 2018) and improve the model robustness as observed by Camuto et al. (2020). In Appx. B.1 we provide details on the architecture, optimization, and results on the test dataset for VAE trained with different values of β. We observe that optimal value in terms of NLL is $\beta = 1$. Larger values of β are supposed to improve robustness in exchange for the reconstruction quality.



(a) $\beta = 1$ (b) $\beta = 10$

Figure 3: Supervised attack: adversarial inputs (*row 1*), their reconstructions (*row 2*) and reconstructions of the corresponding target points (*row 3*).

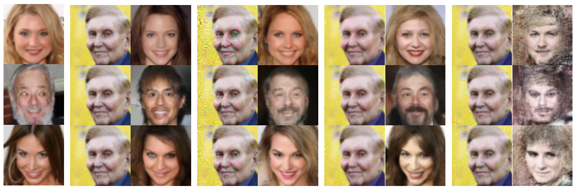


(a) $\beta = 1$ (b) $\beta = 10$

Figure 4: Unsupervised attack: adversarial inputs (*row 1*) and their reconstructions (*row 2*).

Supervised attack We train supervised attacks using eq. (3). Figure 2 depicts result for different values of β. We observe that robustness of the encoder increases if measured by distance between adversarial and reference point in the latent space (Ω). On the other hand, we still observe that in terms of reconstructions the adversarial inputs are closer to the target points than to the reference (plot (b) ad (c)). Moreover, we do not observe higher distortion levels, that is, adversarial inputs itself are still close to the reference (plot (a)). In Figure 3 we provide examples of adversarial inputs for a single reference point and 5 different targets.

Unsupervised attack We present results for the unsupervised attacks trained with eq. (4) in Figure 2 and examples of learned adversarial inputs for a single reference point in Figure 4. We observe behavior similar to supervised tasks, where even for large values of β we can construct a successful adversarial attack.



(a) $\mathbf{x}^t$ (b) $k_A = 1$ (c) $k_A = 2$ (d) $k_A = 4$ (e) $k_A = 8$

Figure 5: Supervised attacks on NVAE. We plot target images in (a). In (b) - (e) we plot adversarial inputs (*column 1*) and their reconstructions (*column 2*). k_A stands for number of top level latent variables considered while learning $\mathbf{x}^a$. We observe that adversarial reconstructions are able to mimic such high level features of the target as face orientation, hairstyle and smile.

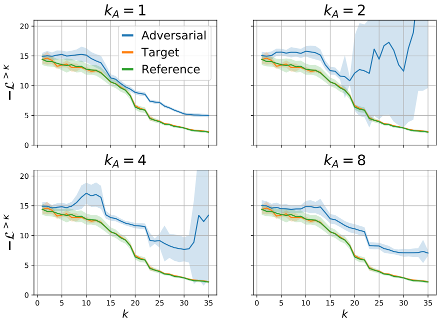


Figure 6: Negative ELBO $\mathcal{L}_k$ from Maaløe et al. (2019). Higher values indicate anomaly.

Table 1: Supervised adversarial attacks on NVAE in terms of proposed metrics.

k_A	MSSSIM$[\mathbf{x}^r, \mathbf{x}^a]$	MSSSIM$[\widetilde{\mathbf{x}}^r, \widetilde{\mathbf{x}}^a]$	MSSSIM$[\widetilde{\mathbf{x}}^t, \widetilde{\mathbf{x}}^a]$	Ω
1	0.99	0.25	0.51	270
2	0.97	0.25	0.65	281
4	0.98	0.30	0.55	328
8	0.99	0.46	0.42	803

3.2 Hierarchical VAE: NVAE

In this section, we explore the robustness of deep hierarchical VAE. It was also studied in Willetts et al. (2019), where authors notice that pure hierarchical VAE with up to 5 levels of latent variables are not gaining any robustness. We construct a supervised adversarial attack for NVAE (Vahdat & Kautz, 2020), a recently proposed VAE with state-of-the-art performance in terms of image generation. We use a model trained on CelebA dataset using the official NVAE implementation[4].

We notice that to effectively attack a hierarchical VAE model, one has to consider only higher-order levels of latent variables, e.g. $\{z_{L-k_A}, z_{L-k_A+1}, \ldots, z_L\}$. That being said, we formulate adversarial attacks using eq. (3), with $q_\phi(\mathbf{z}|\mathbf{x}) = \prod_{i=L-k_A}^{L} q_\phi(\mathbf{z}_i|\mathbf{x}, \mathbf{z}_{>i})$, where k_A is number of latent variables considered during attack construction. We assume that lower-order latent variables are responsible for the specific details of an image and are less useful to learn an adversarial input. A similar approach was suggested in Maaløe et al. (2019) for anomaly detection. They use modified ELBO $\mathcal{L}^{>k}$, where they use prior instead of variational approximation for the first k latent variables.

In Table 1 we present numerical results for attacks with $k_A = \{1, 2, 4, 8\}$. We also plot examples of the learned adversarial inputs with their reconstructions and corresponding target images in Figure 5. In Figure 6 we plot $-\mathcal{L}^{>k}$ from Maaløe et al. (2019) for all possible values of k. We see that curves for adversarial inputs are always above those for the real images from the dataset (either target or reference ones). Moreover, according to our metrics, we are able to obtain an adversarial input that has reconstructions close to the target images rather than to the reference ones. This makes us question the robustness of the hierarchical models to adversarial attacks.

4 Conclusion

In this work, we have explored the robustness of VAEs to adversarial attacks. We have suggested metrics that are easily interpretable. We have used these metrics as well as the proposed definition of supervised and unsupervised attacks to show that VAE and β-VAE are prone to adversarial attacks. Moreover, we were able to attack deep hierarchical VAE to produce high quality adversarial inputs and reconstructions.

[4] The code and model weight were taken from https://github.com/NVlabs/NVAE

ACKNOWLEDGEMENTS

This research was (partially) funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl.

REFERENCES

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

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

Alexander Camuto, Matthew Willetts, Stephen Roberts, Chris Holmes, and Tom Rainforth. Towards a theoretical understanding of the robustness of variational autoencoders. *arXiv preprint arXiv:2007.07365*, 2020.

Ricky TQ Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942*, 2018.

Rewon Child. Very deep VAEs generalize autoregressive models and can outperform them on images. September 2020.

George Gondim-Ribeiro, Pedro Tabacof, and Eduardo Valle. Adversarial attacks on variational autoencoders. June 2018.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

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

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

Lars Maaløe, Marco Fraccaro, Valentin Liévin, and Ole Winther. BIVA: A very deep hierarchy of latent variables for generative modeling. In *Advances in Neural Information Processing Systems*, February 2019.

Rajesh Ranganath, Dustin Tran, and David Blei. Hierarchical variational models. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 324–333, New York, New York, USA, 2016. PMLR.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pp. 1278–1286. PMLR, 2014.

Casper Kaae Sønderby, Tapani Raiko, Lars Maalø e, Sø Ren Kaae Sø nderby, and Ole Winther. Ladder variational autoencoders. In *Advances in Neural Information Processing Systems*, volume 29, pp. 3738–3746. Curran Associates, Inc., 2016.

Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. July 2020.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pp. 1398–1402. Ieee, 2003.

Matthew Willetts, Alexander Camuto, Tom Rainforth, Stephen Roberts, and Chris Holmes. Improving VAEs' robustness to adversarial attack. June 2019.

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

Appendix

A Δbjective for unsupervised attacks

We start with defining the adversarial objective as a difference between the mean evaluated on the adversarial input ($\mathbf{x}^a = \mathbf{x}^r + \epsilon$) and the mean evaluated on the reference point, namely:

$$\widetilde{\Delta}[q_\phi(\mathbf{z}|\mathbf{x}_r + \epsilon), q_\phi(\mathbf{z}|\mathbf{x}_r)] \overset{df}{=} \|\mu(\mathbf{x}^r + \epsilon) - \mu(\mathbf{x}^r)\|_2^2. \tag{6}$$

Since we consider a small perturbation ϵ, we assume that it is reasonable to use linear approximation of the change in μ. That is, we can approximate $\mu(\mathbf{x}^r + \epsilon)$ using on its value in $\mathbf{x}^r$ and Jacobian evaluated at $\mathbf{x}^r$.

$$\mu(\mathbf{x}^r + \epsilon) \approx \mu(\mathbf{x}^r) + \mathbf{J}^q_{\mathbf{x}^r}\epsilon^\top \tag{7}$$

If we plug that into the equation 6, we get an objective:

$$\widetilde{\Delta}[q_\phi(\mathbf{z}|\mathbf{x}_r + \epsilon), q_\phi(\mathbf{z}|\mathbf{x}_r)] \overset{df}{=} \|\mu(\mathbf{x}^r + \epsilon) - \mu(\mathbf{x}^r)\|_2^2 \tag{8}$$

$$\approx \|\mu(\mathbf{x}^r) + \mathbf{J}^q_{\mathbf{x}^r}\epsilon^\top - \mu(\mathbf{x}^r)\|_2^2 \tag{9}$$

$$= \|\mathbf{J}^q_{\mathbf{x}^r}\epsilon^\top\|_2^2 \tag{10}$$

Note, that matrix $\mathbf{J}^q_{\mathbf{x}^r}$ does not depend on ϵ. Therefore, we only need to compute it once for a given reference point.

B Details of the experiments

B.1 β-VAE

Architecture We use fully convolutional architecture with latent dimension 128. In Table 2 we provide detailed scheme of the architecture. We use `Conv(3x3, 1->32)` to denote convolution with kernel size `3x3`, `1` input channel and `32` output channels. We denote stride of the convolution with `s` and padding with `p`. The same notation applied for the transposed convolutions (`ConvTranspose`).

Table 2: Convolutional architecture for VAE.

Encoder	Decoder
Conv(3x3, 1->32, s=1, p=1)	ConvTranspose(3x3,128->256,s=1,p=0)
ReLU()	ReLU()
Conv(5x5, 32->64, s=2, p=0)	ConvTranspose(3x3,256->128,s=2,p=0)
ReLU()	ReLU()
Conv(5x5, 64->128, s=2, p=0)	ConvTranspose(4x4,128->64,s=2,p=0)
ReLU()	ReLU()
Conv(3x3,128->256,s=2,p=1)	ConvTranspose(4x4,64->1,s=2,p=1)
ReLU()	$\mu_x \leftarrow$ Sigmoid()
$\mu_z \leftarrow$ Conv(3x3,256->128,s=2,p=1)	
$\log \sigma_z^2 \leftarrow$ Conv(3x3,256->128,s=2,p=1)	

Optimization We use Adam to perform the optimization. We start from the learning rate $5e-4$ and reduce it by the factor of 0.9 if the validation loss does not decrease for 10 epochs. We train a model for 500 epochs with the batch size 256.

Results In Table 3 we report negative log-likelihood (NLL) of the VAE with different values of parameter β. We observe that the optimal value in terms of NLL is $\beta = 1$. Larger values of β are supposed to improve robustness in exchange for the reconstruction quality.

Table 3: Test performance of the β-VAE with different values of β. Negative loglikelihood is estimated with importance sampling as suggested in (Burda et al., 2015)

β	$-\log p(\mathbf{x})$	$\mathrm{KL}\left[q_\phi(\mathbf{z}\vert\mathbf{x})\Vert p(\mathbf{z})\right]$
.5	234.9	22.5
1	**233.9**	15.1
2	235.5	10.2
4	239.0	6.8
10	250.6	**3.9**

B.2 Adversarial Attack

In all the experiments we randomly select reference and target points from the test dataset. For the Fashion MNIST, we also ensure that the resulting samples are properly stratified — include an even number of points from each of the classes.

For supervised attacks, we learn one adversarial input for each possible pair of the target and reference point. For unsupervised attack we train 6 adversarial inputs for each reference point since we have noticed, that different initialization results in different adversarial inputs. In Table 4 we provide the summary of the total number of points considered.

Table 4: Number of reference, target and adversarial points considered for adversarial attacks.

	Dataset	# of reference points	# of target points	# of adversarial points
Unsupervised	Fashion MNIST	50	—	300
Supervised	Fashion MNIST	50	10	500
Supervised	CelebA	15	3	45