Alleviating Adversarial Attacks on Variational Autoencoders with MCMC

VRIJE UNIVERSITEIT AMSTERDAM
UNIVERSITY OF AMSTERDAM
Hybrid Intelligence

Anna Kuzina, Max Welling, Jakub M. Tomczak

Paper: Code:

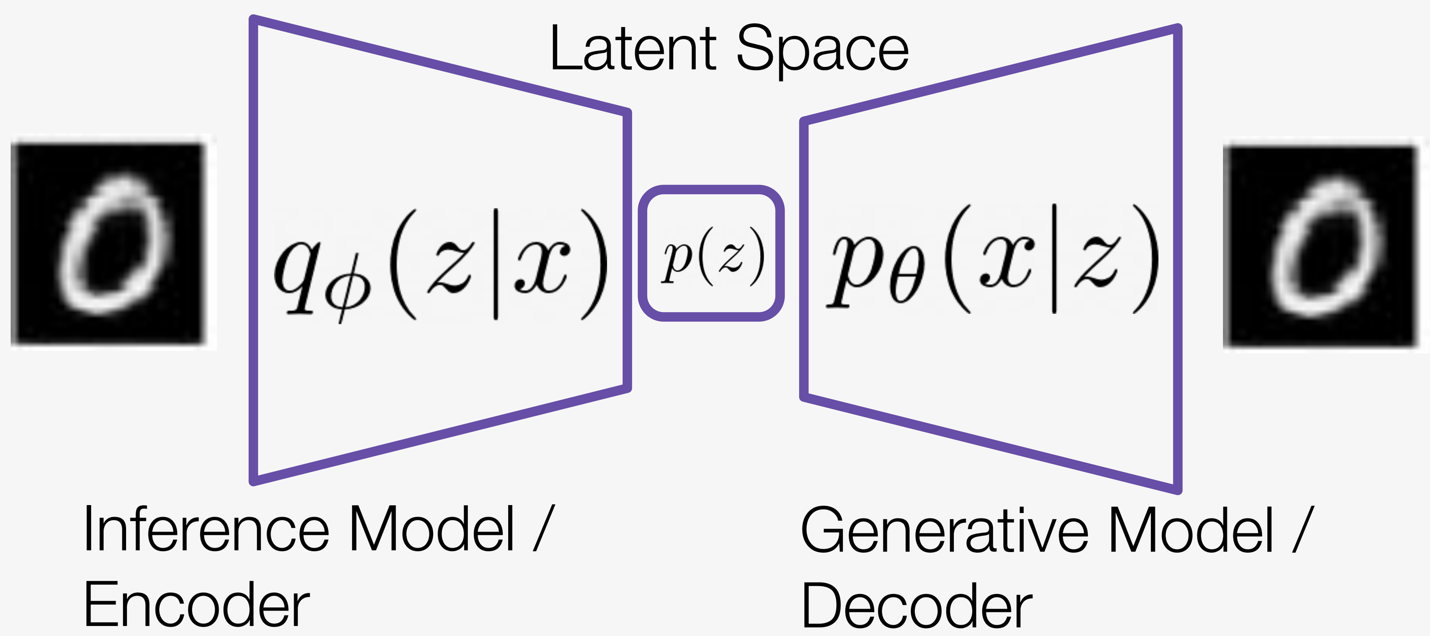
Summary

We propose the way to alleviate the effect of adversarial attacks on the VAE's encoder.

Method does not require changing the training procedure: we use the decoder to protect the encoder.

We provide theoretical justification for it to work.

Variational Auto-Encoder

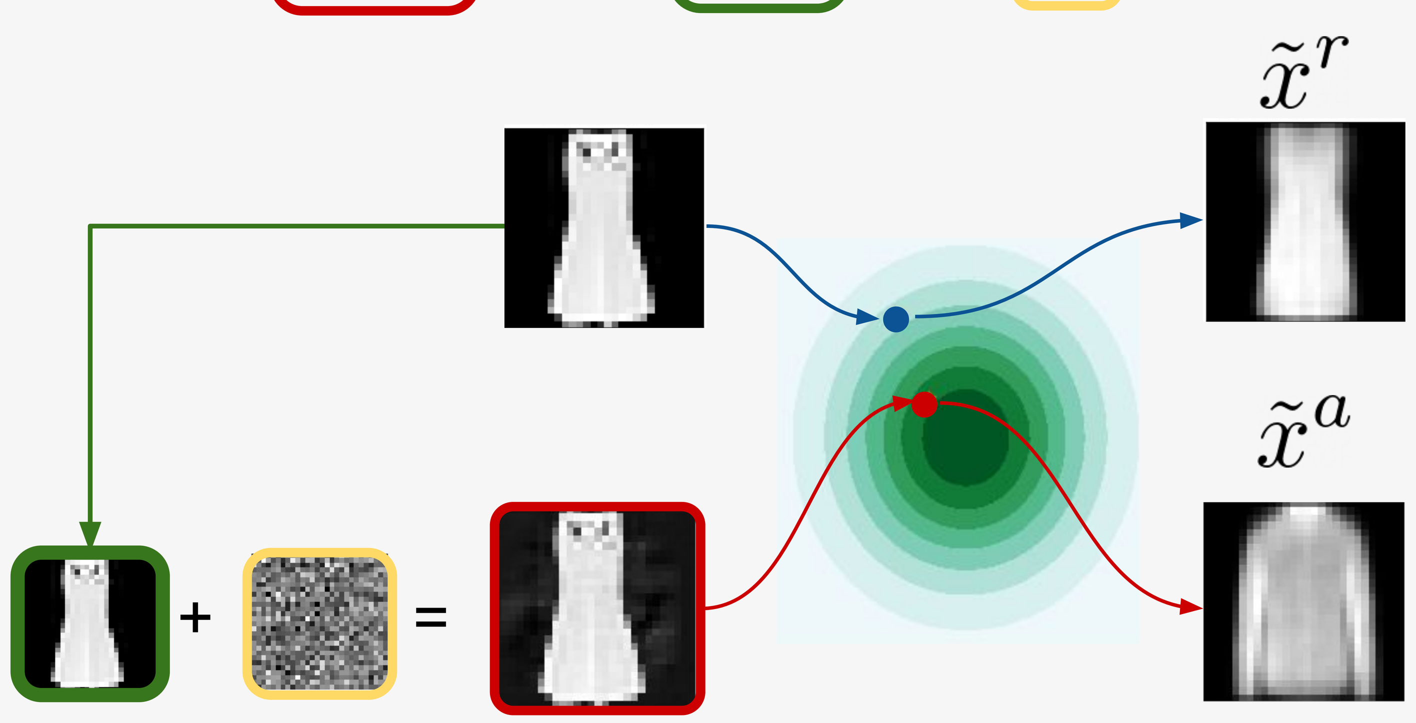


Hierarchical VAE

L latent variables $\mathbf{z} = (z_1, \dots, z_L)$

Adversarial Input

$$x^a = x^r + \varepsilon$$



Can we reduce the effect of an attack?

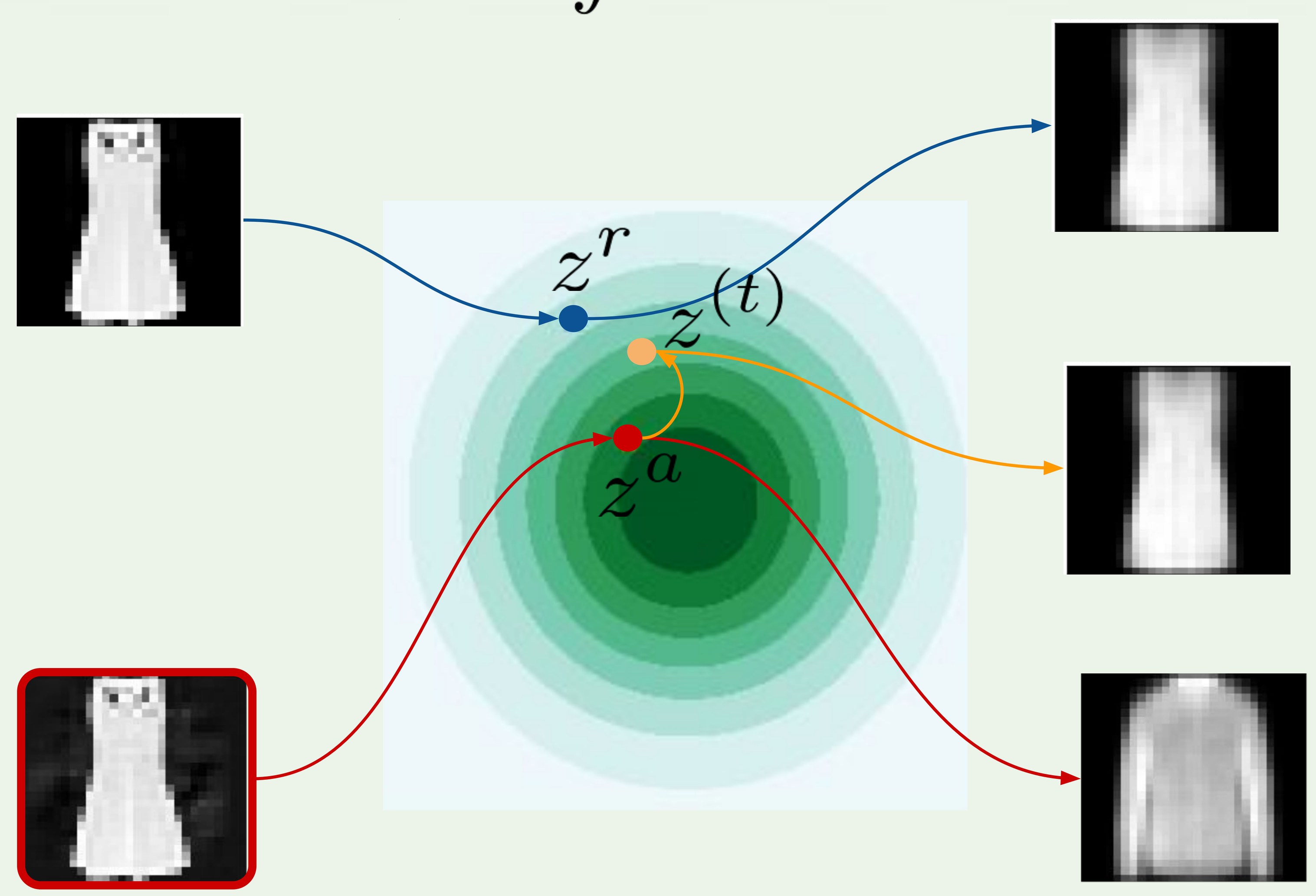
$z^r \sim q_\phi(z|x^r)$ is what we want

$z^a \sim q_\phi(z|x^a)$ is what we get instead

Let's sample from the true posterior: $p_\theta(z|x^a) \propto p(z)p_\theta(x^a|z)$

We use MCMC to get a sample:

$$z^{(t)} \sim q^{(t)}(z|x^a) = \int q_\phi(z_0|x^a)Q^{(t)}(z|z_0)dz_0$$



Why does it work?

Gets smaller with each MCMC step

$$\mathrm{TV}[q^{(t)}(z|x^a)\|q_\phi(z|x^r)] \leq \sqrt{\tfrac{1}{2}\mathrm{KL}\left[q^{(t)}(z|x^a)\|p_\theta(z|x^a)\right]}$$

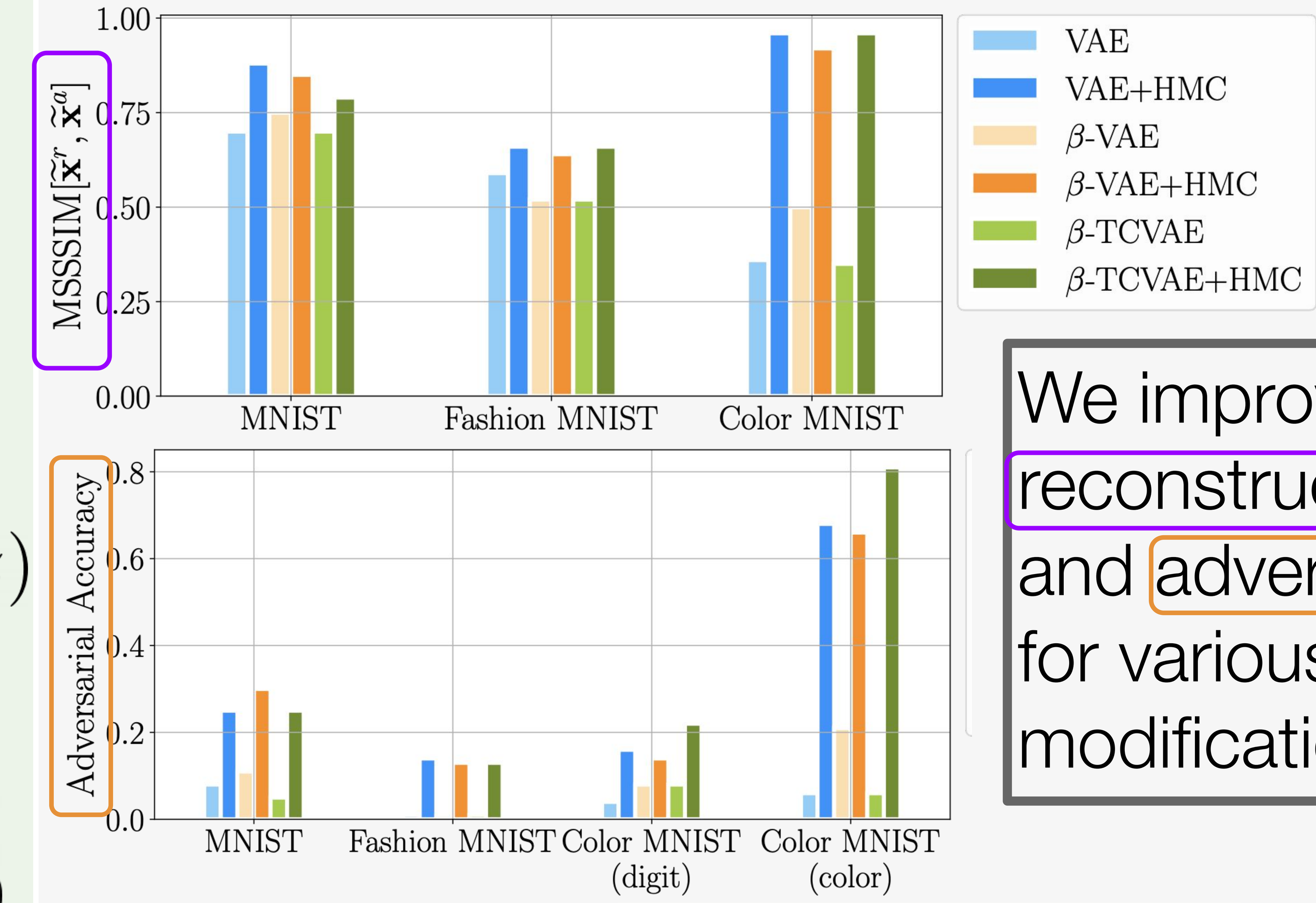
VAE amortization gap

$$+\sqrt{\tfrac{1}{2}\mathrm{KL}\left[q_\phi(z|x^r)\|p_\theta(z|x^r)\right]}$$

Attack radius

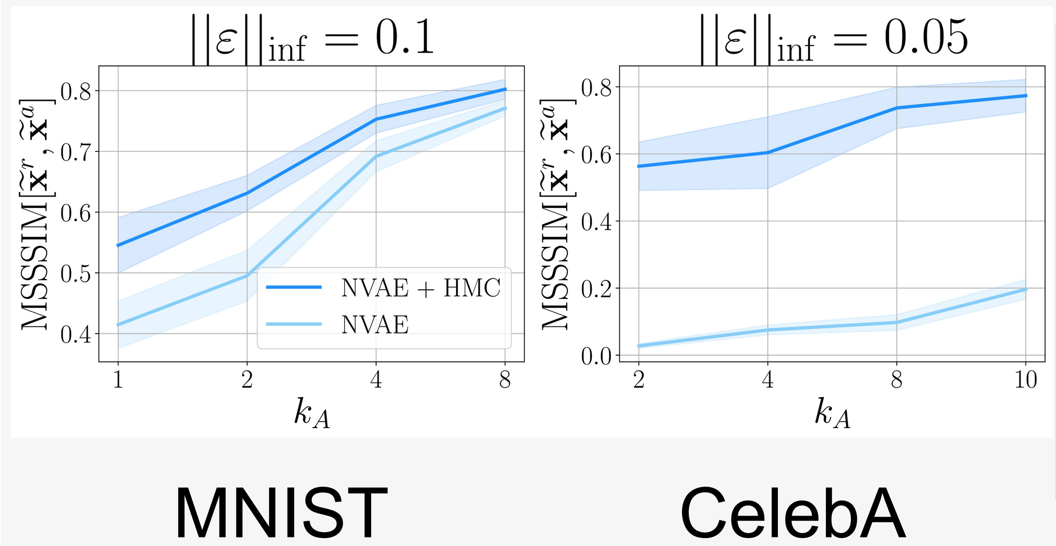
$$+\,o(\sqrt{\|\varepsilon\|})$$

Results: VAE



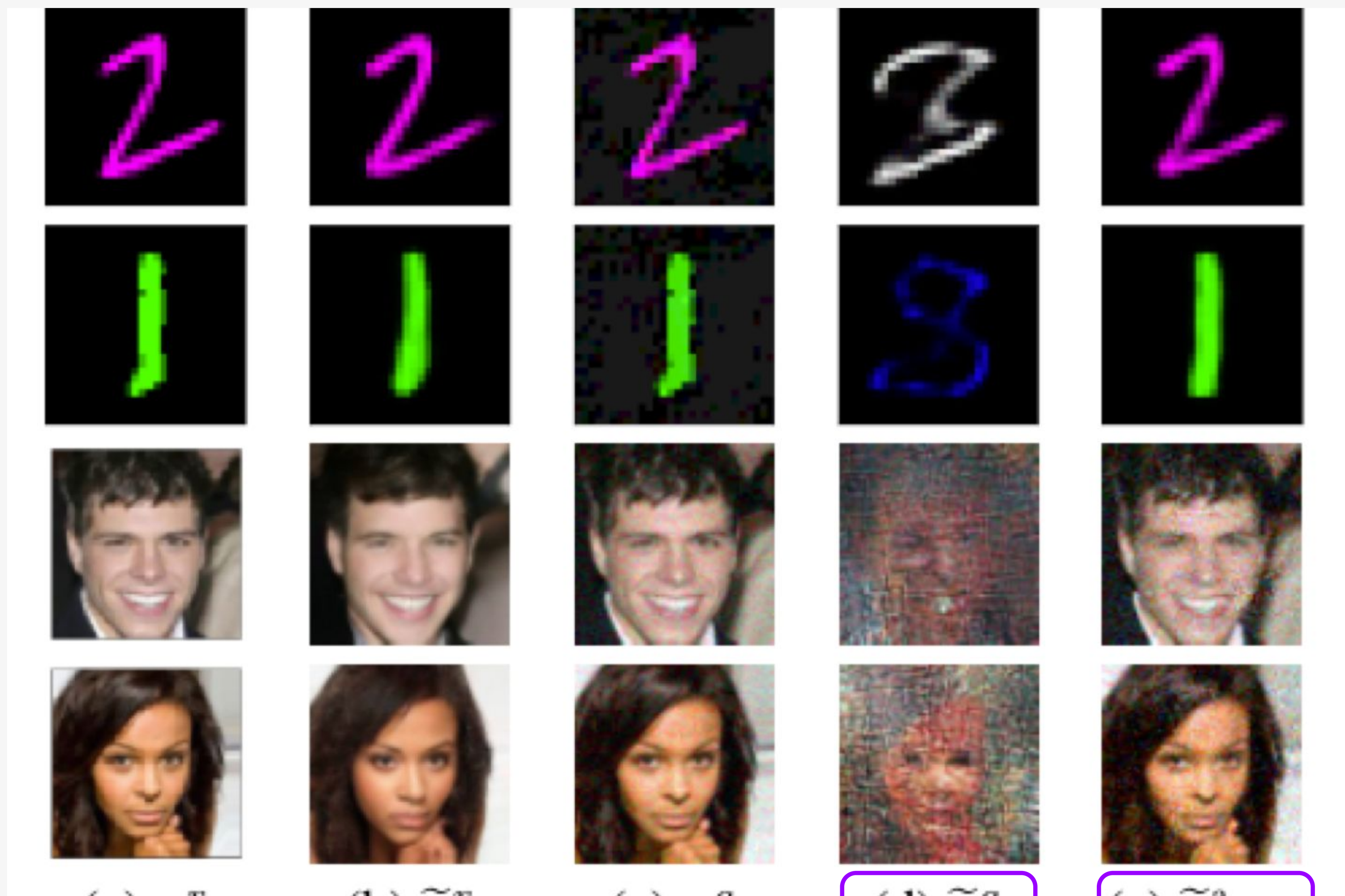
We improve both reconstruction quality and adversarial accuracy for various VAE modifications

Results: NVAE



Consider only last k_A latent variables for attack construction. Our method improves reconstruction similarity

We observe that reconstructions are more similar to the reference when we use the proposed method



(a) $\mathbf{x}^r$ (b) $\tilde{\mathbf{x}}^r$ (c) $\mathbf{x}^a$ (d) $\tilde{\mathbf{x}}^a$ (e) $\tilde{\mathbf{x}}^a_{\text{HMC}}$