Adversarial Attacks on Variational Autoencoders

George Gondim-Ribeiro, *Pedro Tabacof* & Eduardo Valle

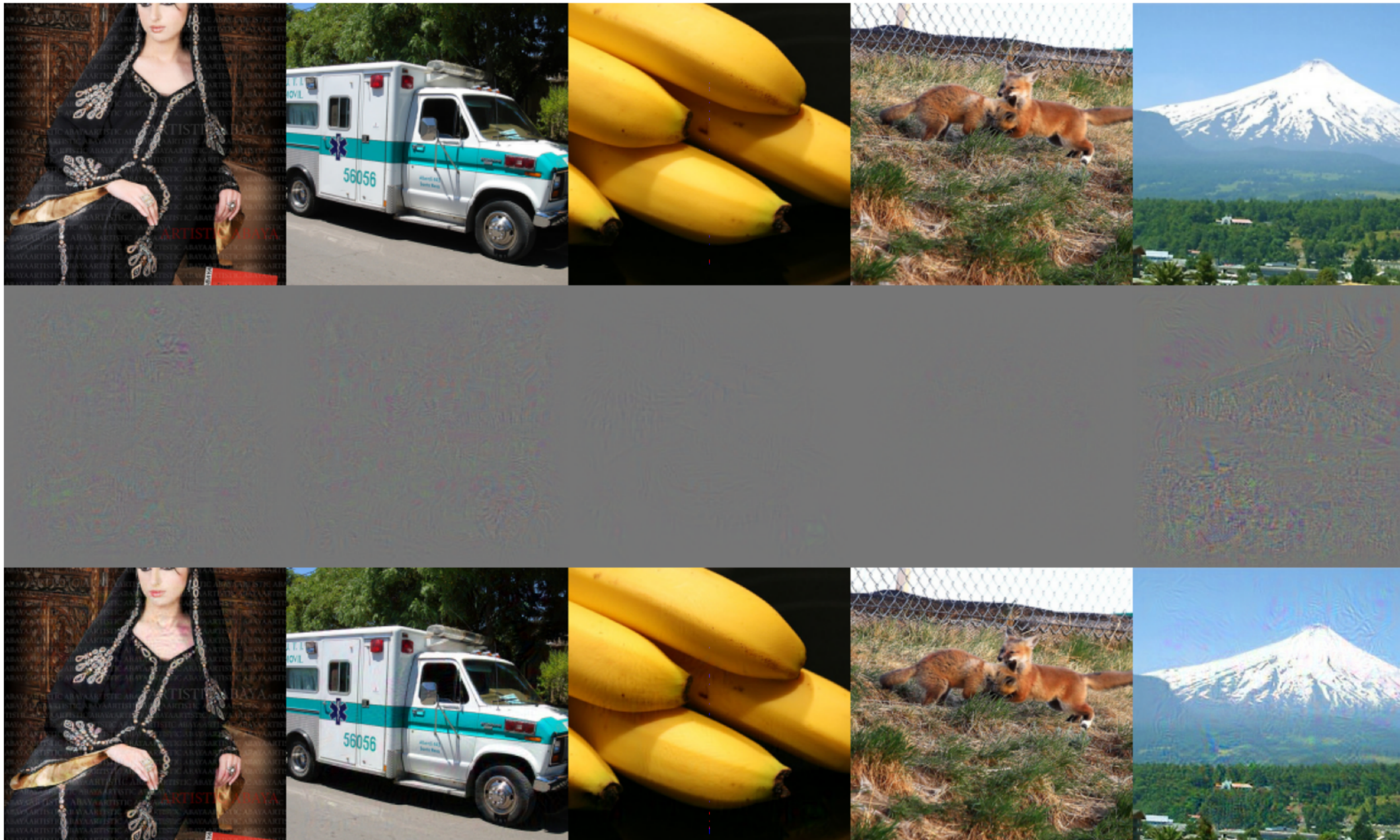
UNICAMP

RECOD reasoning for *complex* data

Who am I?

- PhD student at the University of Campinas in Brazil
- Data scientist at Nubank, credit card fintech
- Co-authored a few other papers on adversarial attacks, mostly during the Masters:
  - Exploring the space of adversarial images, 2016 IJCNN, with Eduardo Valle (55 citations)
  - Adversarial images for variational autoencoders, 2016 NIPS Adversarial Learning workshop, with Julia Tavares and Eduardo Valle (11 citations)
- Also interested in Bayesian deep learning and uncertainty in machine learning (current research topic)

Adversarial Images

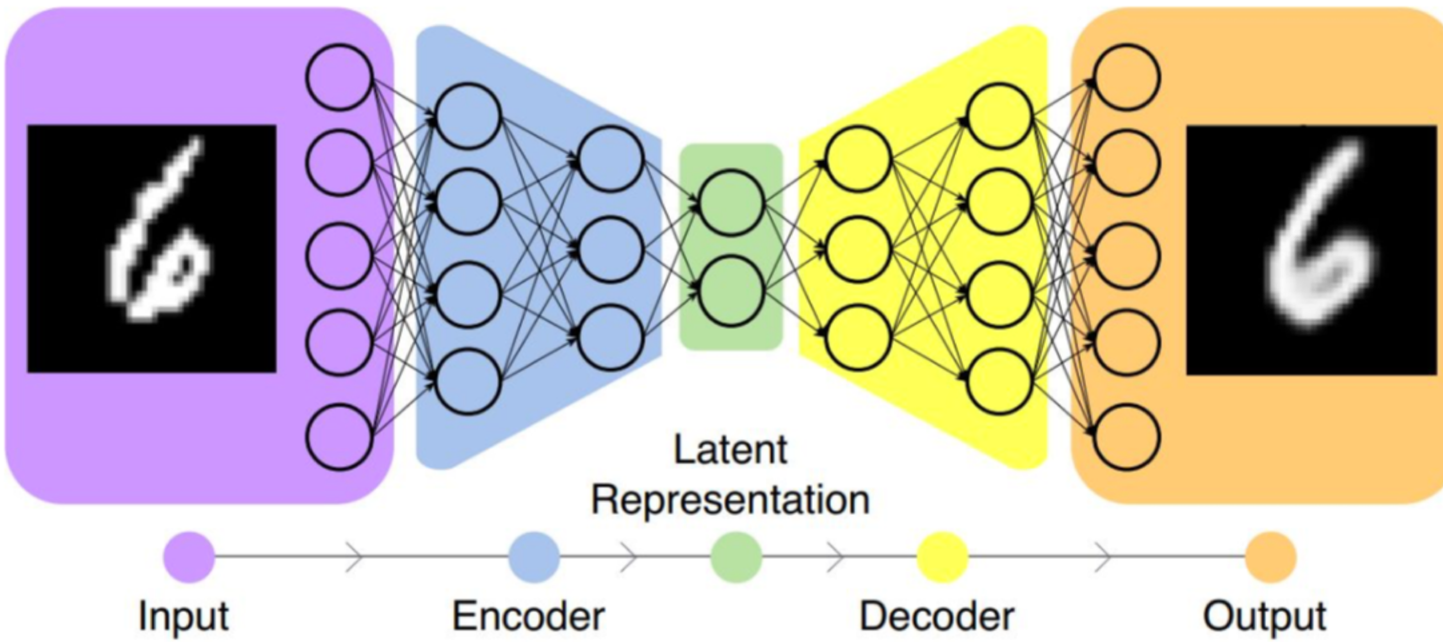


ADVERSARIAL IMAGES

$$\begin{aligned} \underset{\boldsymbol{d}}{\text{minimize}} \quad & \|\boldsymbol{d}\| \\ \text{subject to} \quad & L \le \boldsymbol{x} + \boldsymbol{d} \le U \\ & \boldsymbol{p} = f(\boldsymbol{x} + \boldsymbol{d}) \\ & \max(p_1 - p_c, \ldots, p_n - p_c) > 0 \end{aligned}$$

x is the original image, d is the distortion, x+d is the adversarial input, f is the classifier, p_i *are the scores for each class (where c is the correct class), and L and U are the bounds for the input space.*

VARIATIONAL AUTOENCODERS



Variational Autoencoders

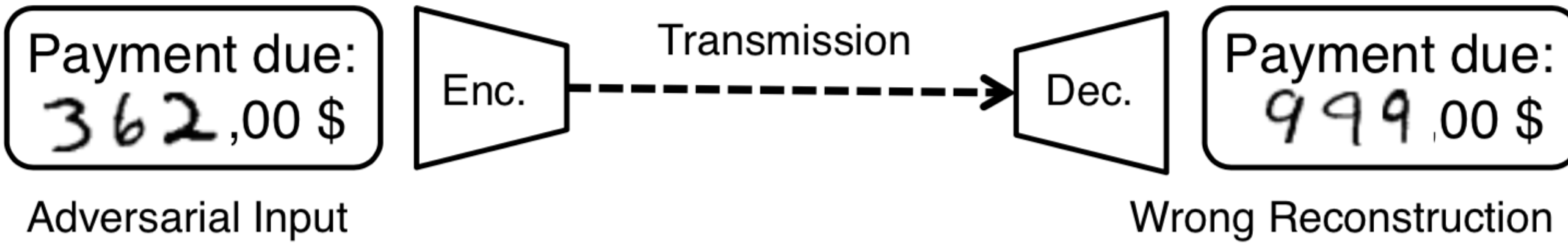
Maximize the Evidence Lower Bound (ELBO)

$$\log p_\theta(\boldsymbol{x}) \geq \mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\psi(\boldsymbol{x}|\boldsymbol{z})] - \mathrm{D}_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z}))$$

Reconstruction

Regularization

Motivation

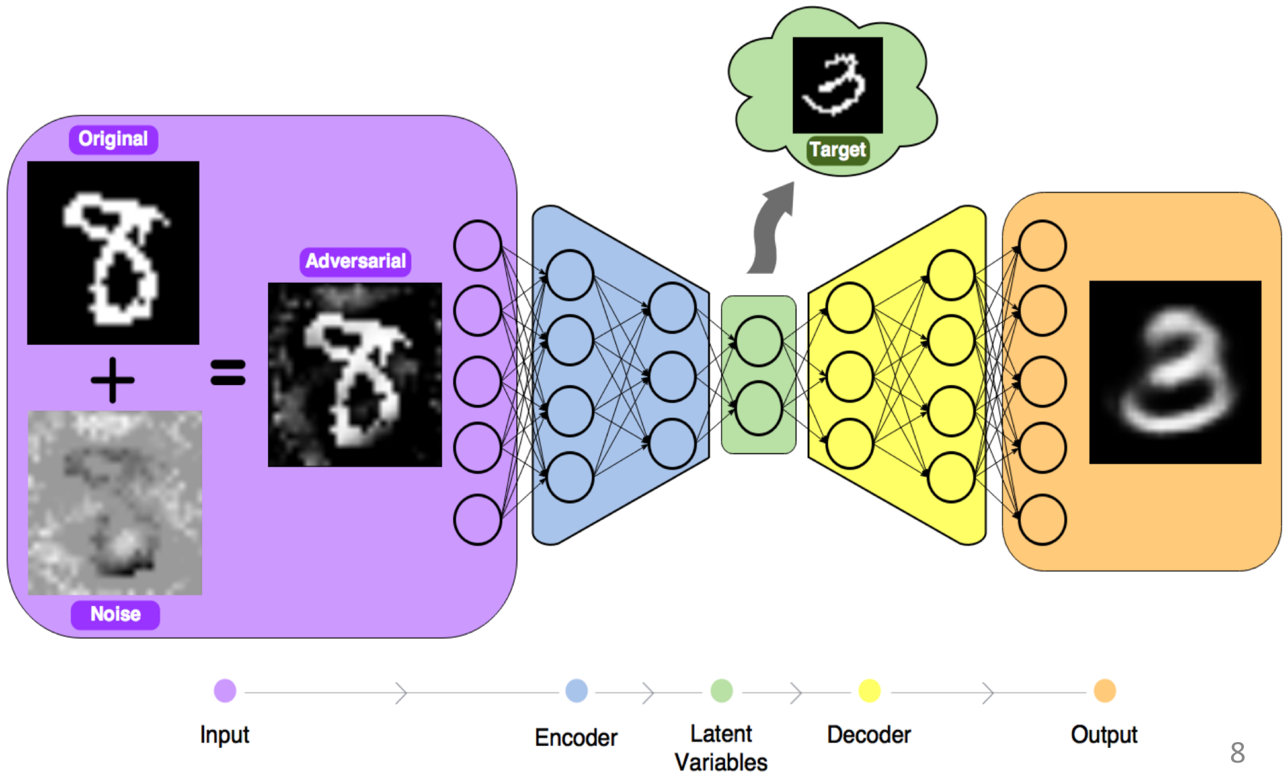


Scenario inspired from *Jernej Kos, Ian Fischer, and Dawn Song. Adversarial examples for generative models. 2017.*

Main Idea

We attack variational autoencoders with adversarial images. We aim not only to disturb the reconstruction, but also to fool the autoencoder into reconstructing a completely different target image.

We attack the latent representation, attempting to match it to the target image's, while keeping the input distortion as small as possible.



The Attack

We attack the *latent layer* — which is the information bottleneck of the autoencoder— with the optimization at the right.

The Δ function we used was the KL divergence.

$$\begin{aligned} \min_{\boldsymbol{d}} \quad & \Delta(\boldsymbol{z_a}, \boldsymbol{z_t}) + C\|\boldsymbol{d}\| \\ \text{s.t.} \quad & L \leq \boldsymbol{x} + \boldsymbol{d} \leq U \\ & \boldsymbol{z_a} = \text{encoder}(\boldsymbol{x} + \boldsymbol{d}) \end{aligned}$$

The Attack

We also attack the *output reconstruction* — with the optimization at the right.

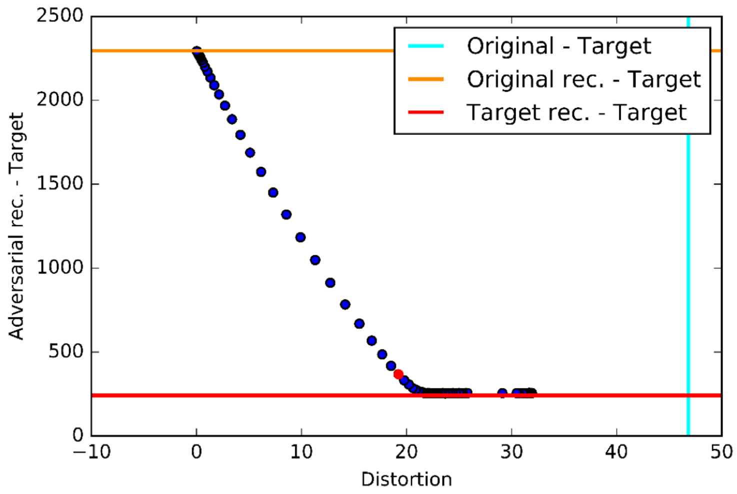
The Δ function is the ℓ_2–norm.

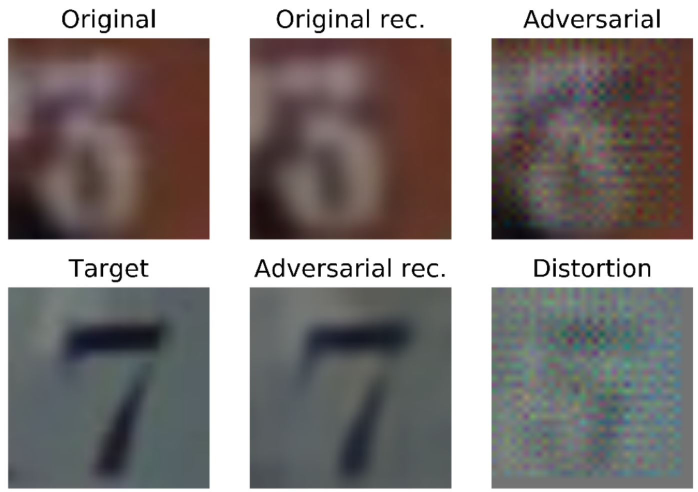
$$
\begin{aligned}
\min_{\boldsymbol{d}} \quad & \Delta(\boldsymbol{r_a}, \boldsymbol{I_t}) + C\|\boldsymbol{d}\| \\
\text{s.t.} \quad & L \leq \boldsymbol{x} + \boldsymbol{d} \leq U, \\
& \boldsymbol{z_a} = \text{encoder}(\boldsymbol{x} + \boldsymbol{d}), \\
& \boldsymbol{r_a} = \text{decoder}(\boldsymbol{z_a})
\end{aligned}
$$

The Attack

Decreasing the regularizer C allows for bigger distortions...

...bringing the adversarial reconstruction closer to the target

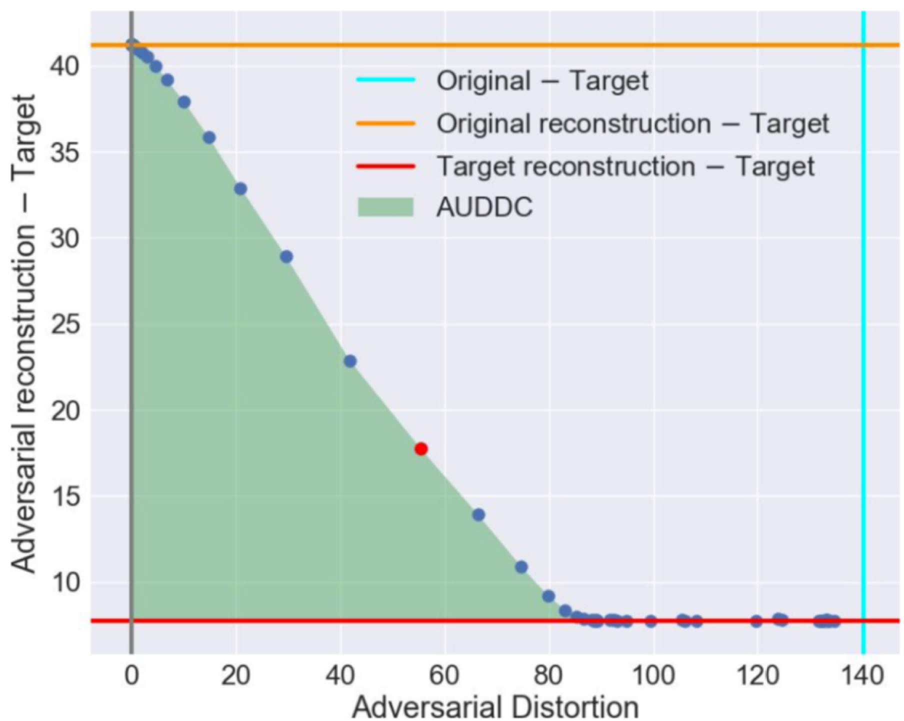


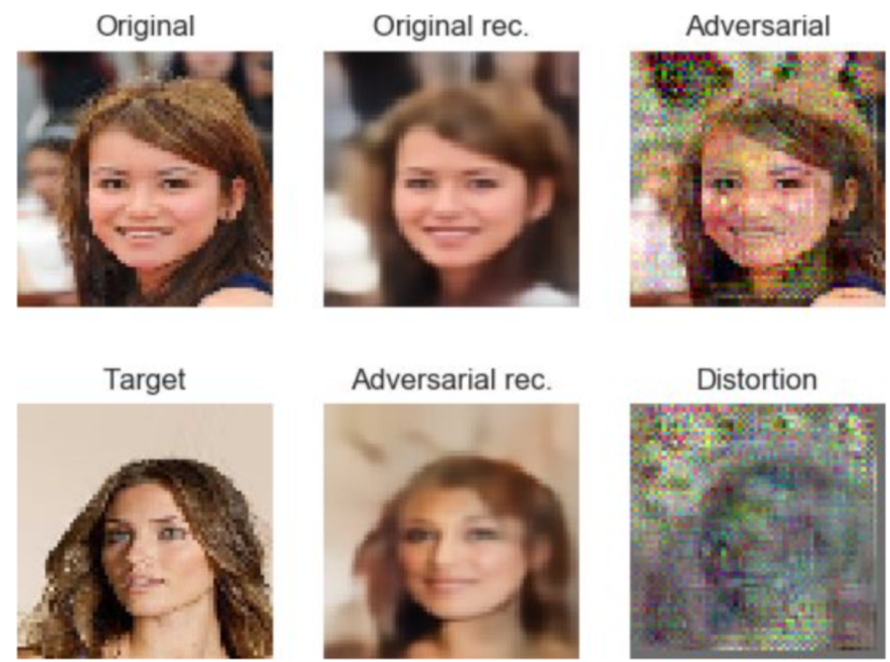


The Metric

AUDDC: Area Under the Distortion-Distortion Curve

From 0 (easiest attack possible) to 100 (hardest attack possible)





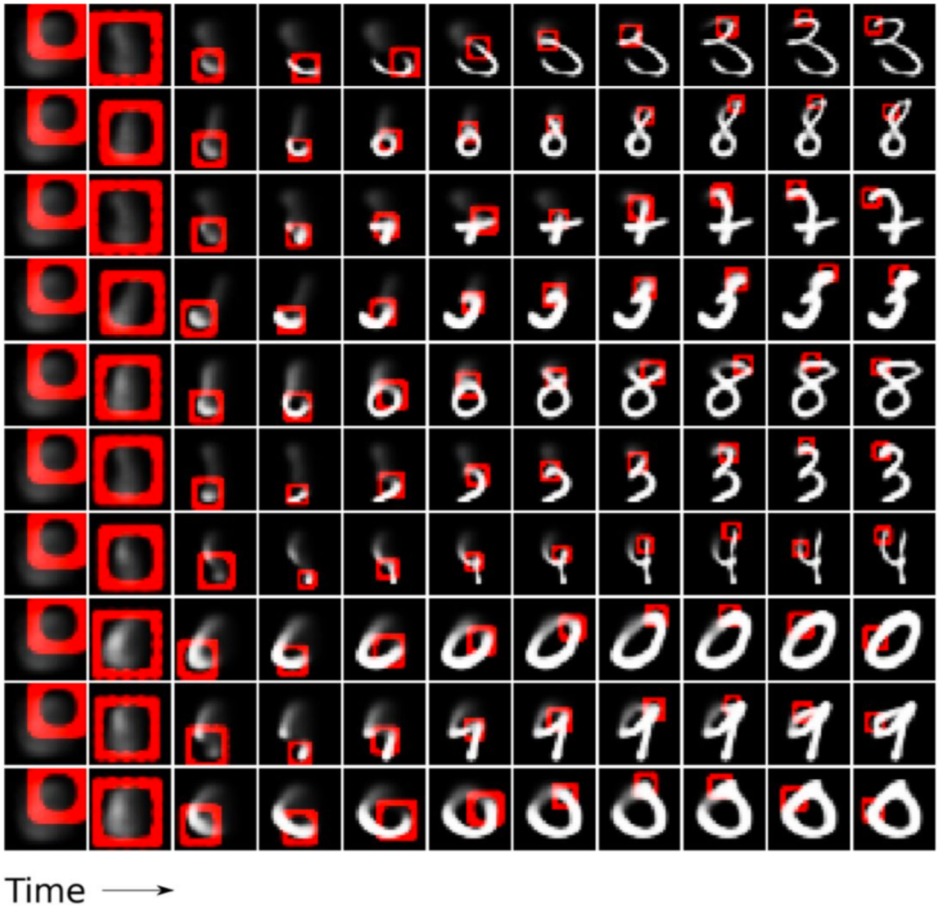
METHODOLOGY

- Three datasets: MNIST, SVHN, and CelebA



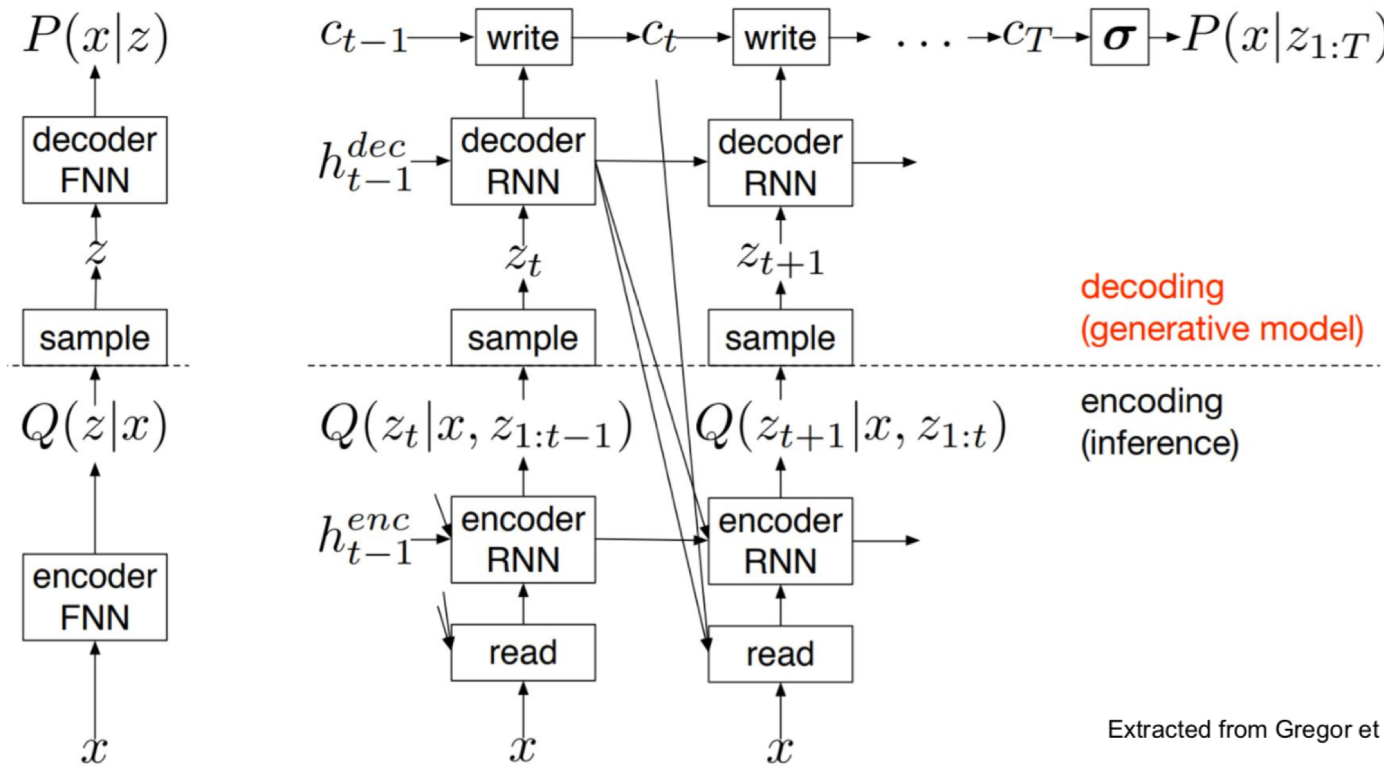
- Models: fully-connected VAEs, convolutional VAEs (CVAE), and DRAW
- A point in the Distortion-Distortion Curve is the average of 128 attacks

DRAW



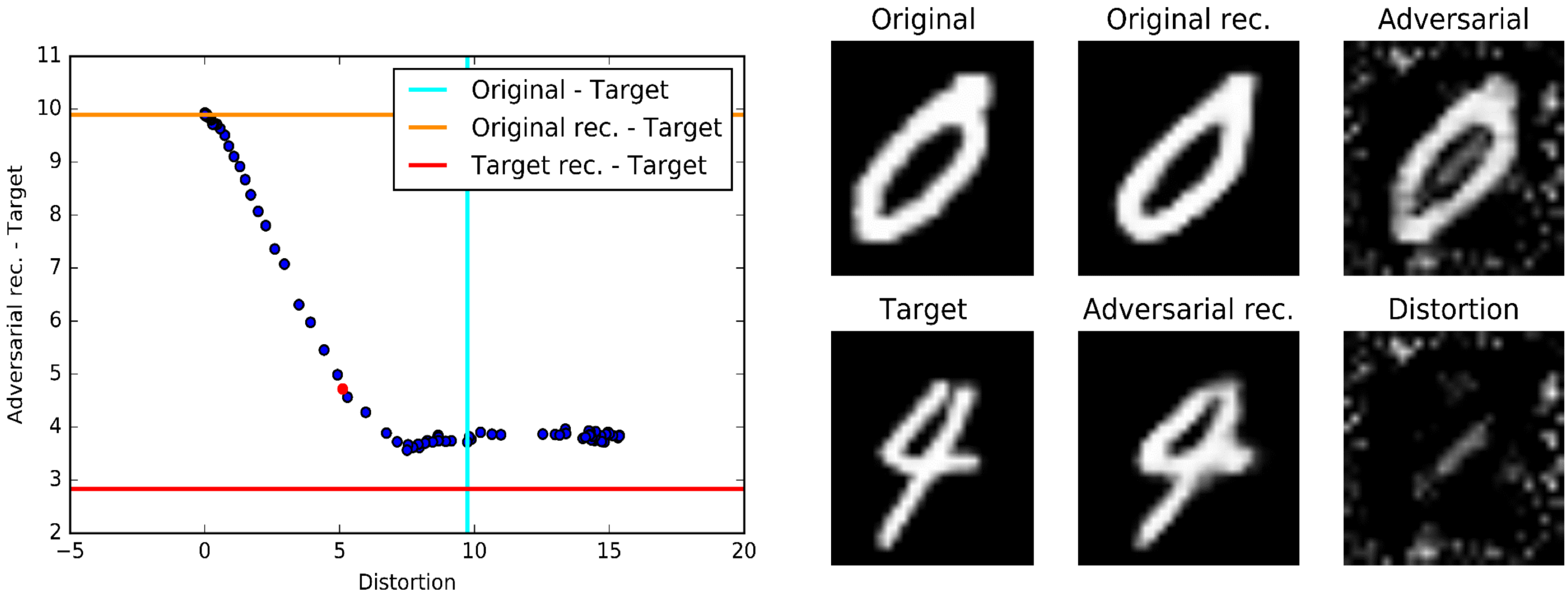
Extracted from Gregor et al., 2015

DRAW

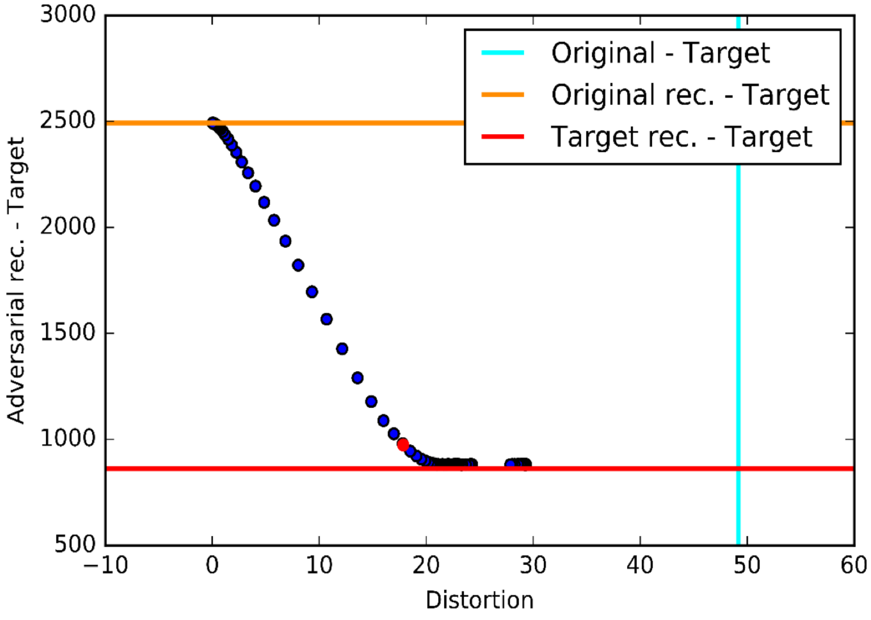


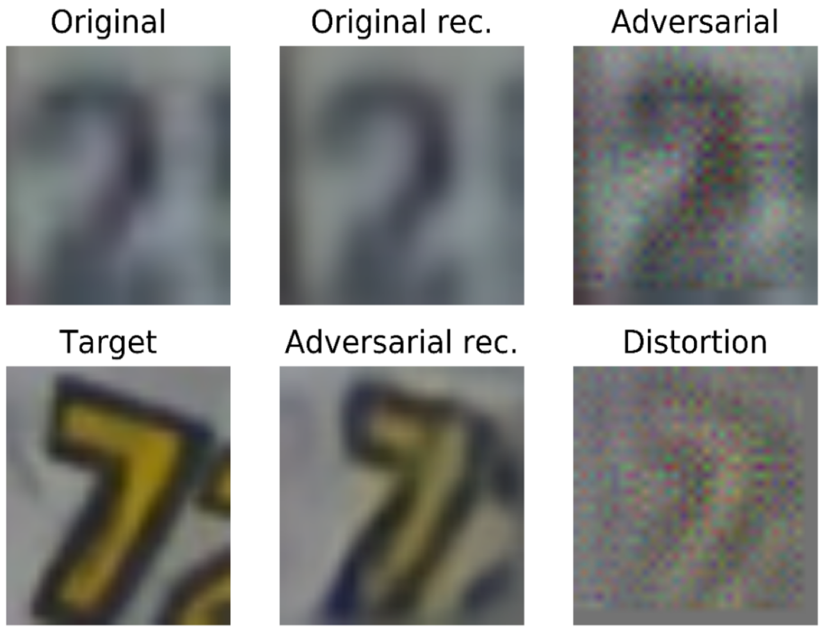
Extracted from Gregor et al., 2015

MAIN FINDINGS



Main Findings



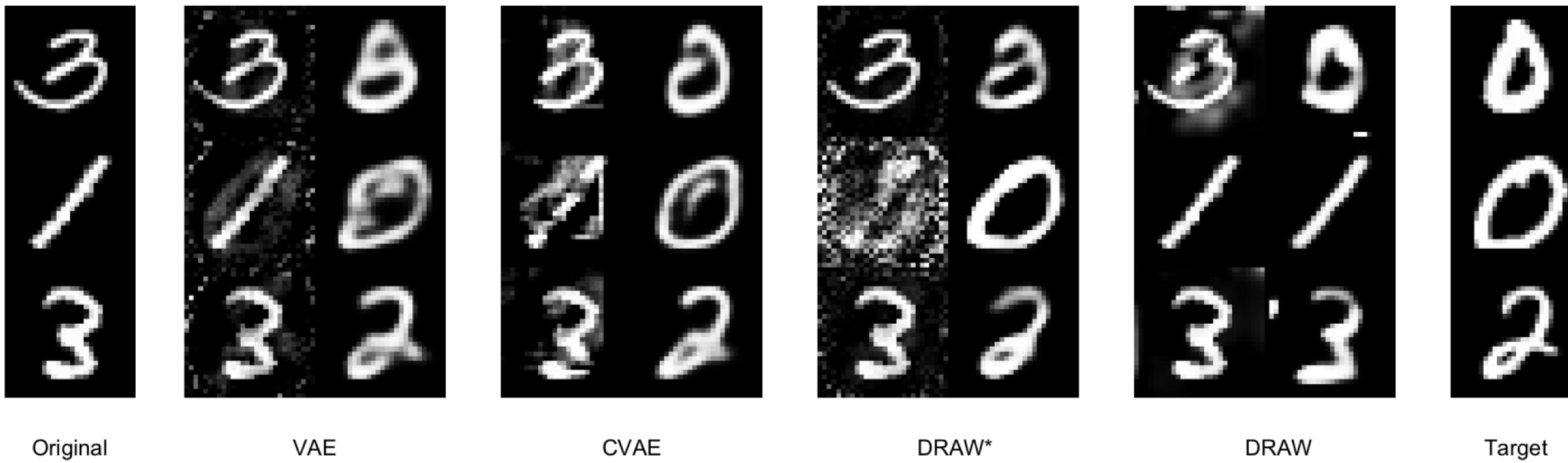


Main Findings

	VAE	CVAE	DRAW*	DRAW	DRAW*	DRAW	
Steps	—	—	1	1	16	16	
Attacks on latent representation							
MNIST	27 ± 2	35 ± 3	27 ± 1	35 ± 3	71 ± 5	**91 ± 3**	47 ± 3
SVHN	19 ± 1	18 ± 1	09 ± 1	27 ± 2	74 ± 6	**96 ± 2**	41 ± 4
CelebA	31 ± 1	28 ± 1	21 ± 2	36 ± 1	81 ± 4	**97 ± 1**	49 ± 4
	25 ± 1	27 ± 2	19 ± 2	33 ± 1	75 ± 3	**95 ± 1**	46 ± 2
Attacks on output							
MNIST	35 ± 2	56 ± 3	38 ± 2	48 ± 4	29 ± 3	**69 ± 4**	46 ± 2
SVHN	19 ± 1	19 ± 2	13 ± 1	27 ± 2	21 ± 2	**34 ± 2**	22 ± 1
CelebA	27 ± 1	24 ± 1	31 ± 3	35 ± 1	29 ± 2	**40 ± 1**	31 ± 1
	27 ± 1	33 ± 3	27 ± 2	37 ± 2	26 ± 1	**47 ± 3**	33 ± 1
All attacks							
MNIST	31 ± 2	45 ± 3	32 ± 2	42 ± 3	50 ± 5	**80 ± 3**	47 ± 2
SVHN	19 ± 1	19 ± 1	11 ± 1	27 ± 1	47 ± 7	**65 ± 7**	31 ± 2
CelebA	29 ± 1	26 ± 1	26 ± 2	36 ± 1	55 ± 6	**68 ± 7**	40 ± 2
	26 ± 1	30 ± 2	23 ± 1	35 ± 1	51 ± 4	**71 ± 3**	39 ± 1

* Attention mechanism disabled.

MAIN FINDINGS



* Attention mechanism disabled.

Main Findings

Original | Fully-connected VAE | DRAW | Target

Conclusions

- ✓ We can attack autoencoders with adversarial images, by targeting their internal representations;
- ✓ The attack forces the autoencoder to reconstruct a different image;
- ✓ Autoencoders are, however, robust: success cases are hard to find and must be regularized “by hand”;
- ✓ The attack has a linear “give-and-take”: success in approaching the target output is proportional to the distortion of the input;
- ✓ The proposed metric (AUDDC) correlates well with qualitative results and provides a measure of robustness;
- ✓ DRAW is the most resistant architecture: attention and recurrence hinders the attack.

Adversarial Attacks on Variational Autoencoders

tabacof@dca.fee.unicamp.br

UNICAMP

RECOD
reasoning for *complex* data