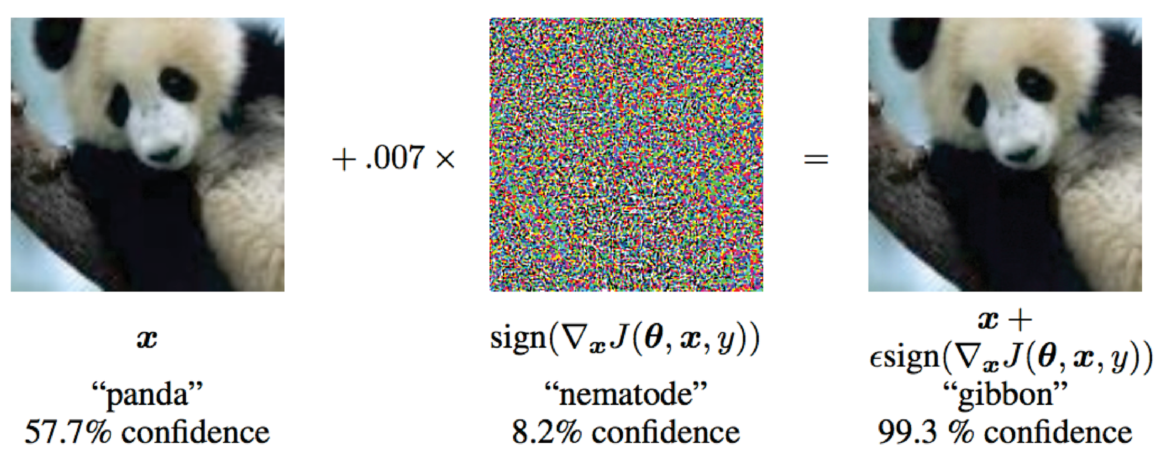
Generating Natural Adversarial Examples

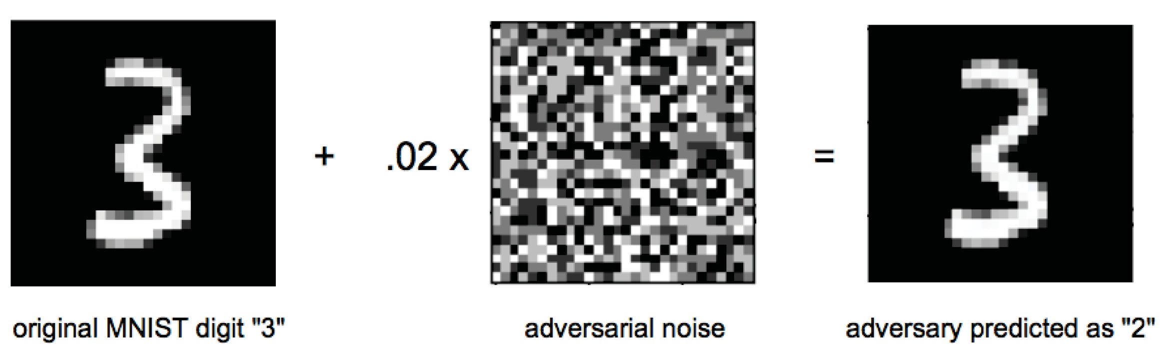
Zhengli Zhao, Dheeru Dua, Sameer Singh
University of California, Irvine

Motivation

- Adversarial examples [1]
  - $x^* = \mathrm{argmin}_{\tilde{x}} \|x - \tilde{x}\|_2$ s.t. $f(x) \neq f(\tilde{x})$.



An adversarial example in [1] applied to GoogLeNet

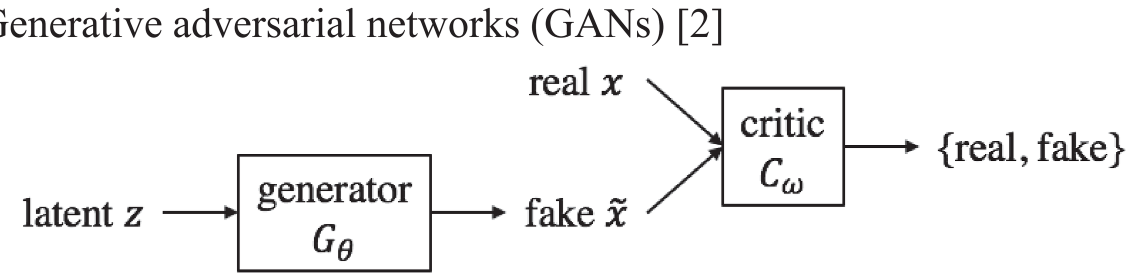


An adversarial example applied to MNIST classifier of Random Forests

- Disadvantages of these adversaries
  - Unnatural.
  - Added noise is imperceptible and uninterpretable.
  - There is mismatch between input space and *semantic space*.
  - Related approaches cannot be applied to heavily structured domains.

Background

- Generative adversarial networks (GANs) [2]



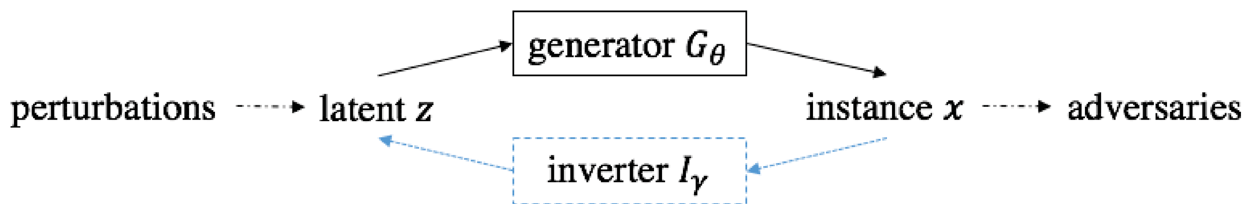
Architecture of GANs

- Minimax game between competing generator and critic.
- Generator maps prior distribution $p_z(z)$ to distribution of real instances $p_{\mathrm{real}}(x)$, generating fake $\tilde{x}$ which are close to real x.
  $\min_{G_\theta} \mathbb{E}_{z \sim p_z(z)}[\log(1 - C_\omega(G_\theta(z)))]$
- Critic discriminates between real instances x and generated fake $\tilde{x}$.
  $\max_{C_\omega} \mathbb{E}_{x \sim p_{\mathrm{real}}(x)}[\log(C_\omega(x))] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - C_\omega(G_\theta(z)))]$

- Wasserstein GAN [3]
  - Use Wasserstein-1 (also called Earth-Mover) distance instead.
  - Generator: $\max_{G_\theta} \mathbb{E}_{z \sim p_z(z)}[C_\omega(G_\theta(z))]$
  - Critic: $\max_{C_\omega} \mathbb{E}_{x \sim p_{\mathrm{real}}(x)}[C_\omega(x)] - \mathbb{E}_{z \sim p_z(z)}[C_\omega(G_\theta(z))]$

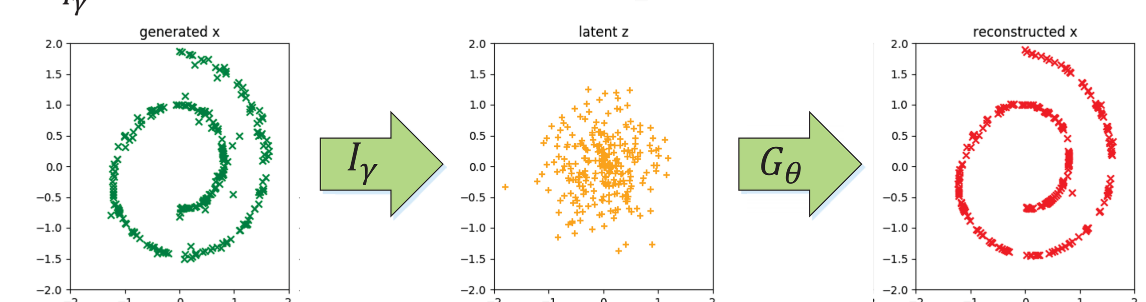
Proposed Approach

- Natural Adversarial Examples
  - $x^* = G_\theta(z^*)$ as $z^* = \mathrm{argmin}_{\tilde{z}} \|I_\gamma(x) - \tilde{z}\|_2$ s.t. $f(x) \neq f(G_\theta(\tilde{z}))$.



Architecture of our approach

- Inverter I_γ maps input x to corresponding z in semantic space.
  $\min_{I_\gamma} \lambda_1 \mathbb{E}_{x \sim p_{\mathrm{real}}(x)} \|x - G_\theta(I_\gamma(x))\|_2 + \lambda_2 \mathbb{E}_{z \sim p_z(z)} \|z - I_\gamma(G_\theta(z))\|_2$



Mapping from Gaussian to Swiss roll distribution

- Search for adversaries in dense and continuous representation z of the input data instead of in the data space directly.

Given input instance x, black-box classifier f
$y_{\mathrm{pred}} = f(x), \hat{z} = I_\gamma(x)$
for radius r in range$(0, R, \delta r)$:
  sample random noise ϵ within $(r, r + \delta r]$
  $\tilde{z} = \hat{z} + \epsilon, \tilde{x} = G_\theta(\tilde{z}), \tilde{y} = f(\tilde{x})$
  if there exists y^* in $\tilde{y}$ that $y^* \neq y_{\mathrm{pred}}$:
    return corresponding adversary x^*

Pseudo code for searching adversaries

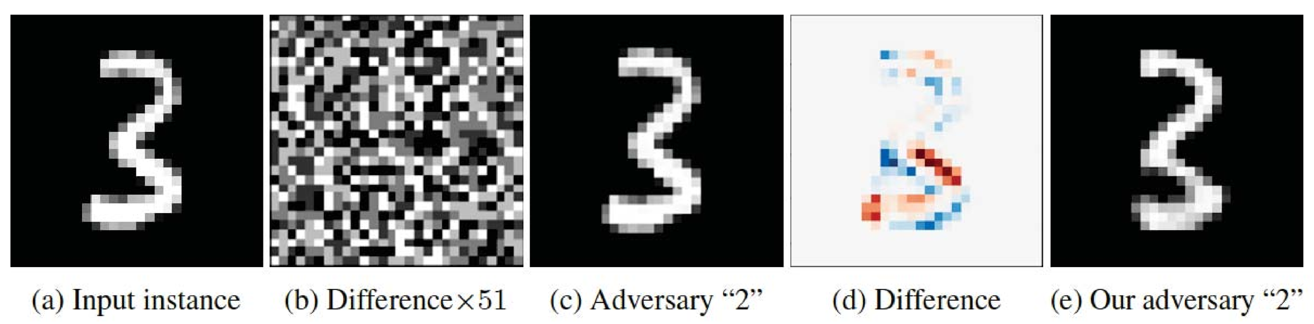
- For discrete text t, we incorporate a discrete structure auto-encoder proposed in [4], in order to encode sentence to continuous code space with $E_\phi(t)$, and decode continuous code to sentence with $D_\psi(x)$.
- Advantages of our adversaries
  - Natural.
  - Difference from the input is informative and interpretable.
  - Semantically close to the input.
  - Our approach can be applied to text data, generating grammatical adversarial sentences.

References

[1] Goodfellow et al, "Explaining and Harnessing Adversarial Examples", ICLR 2015.
[2] Goodfellow et al, "Generative Adversarial Nets", NIPS 2014.
[3] Arjovsky et al, "Wasserstein Generative Adversarial Networks", ICML 2017.
[4] Zhao et al, "Adversarially Regularized Autoencoders for Generating Discrete Structures", arXiv 2017.

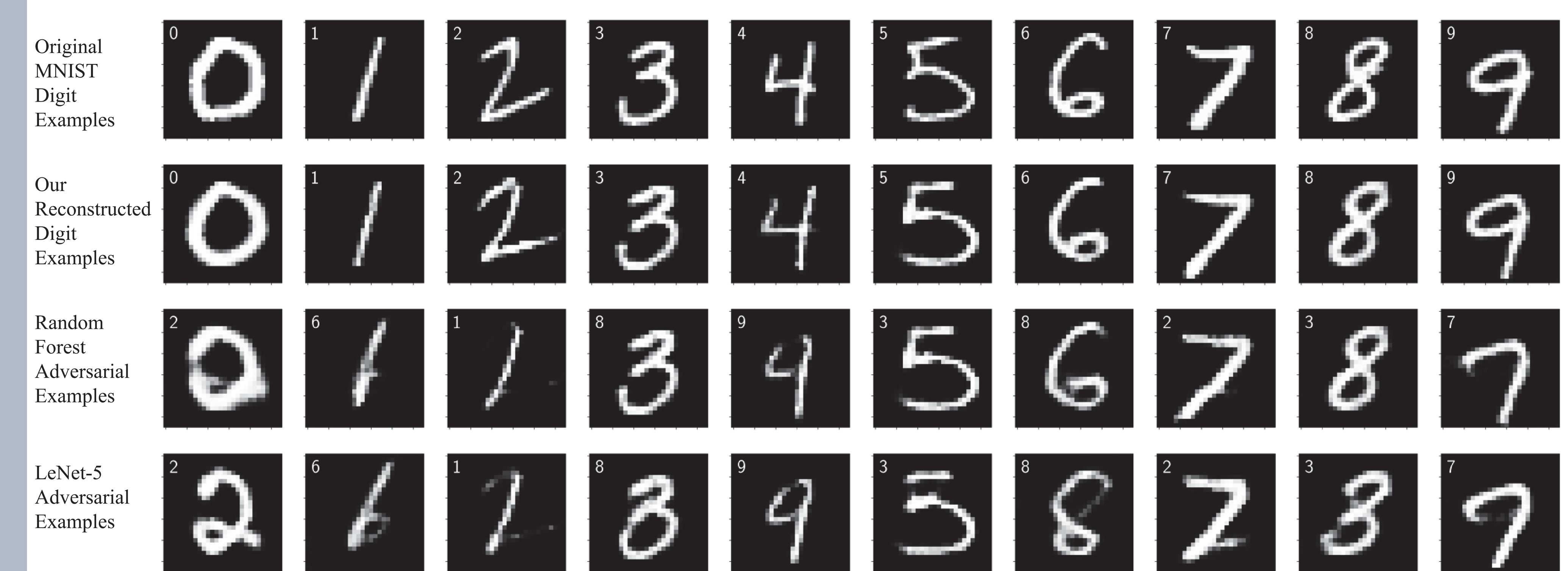
Experimental Results

- Experiments on images
  - Interpretable natural adversaries



Adversary via adding small scale noise *vs.* our natural adversary with interpretable difference.

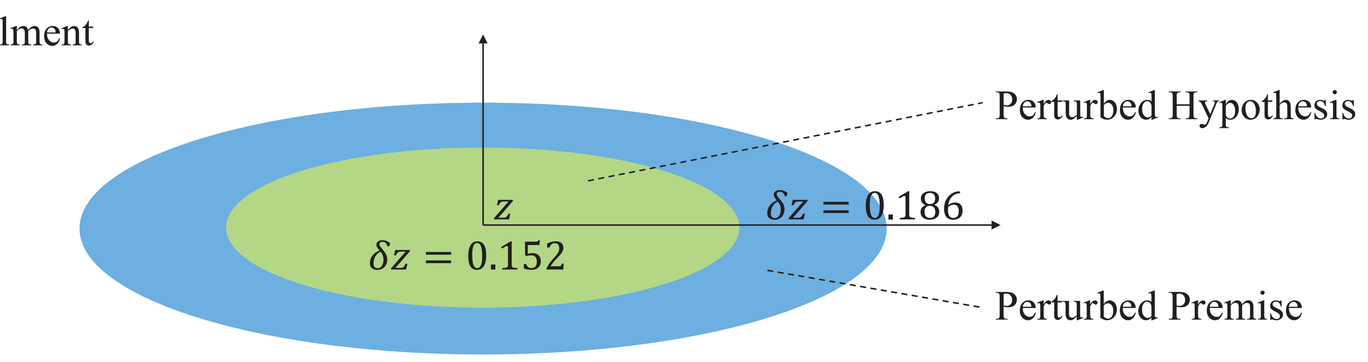
- Evaluation of black-box classifiers



Evaluation of black-box MNIST classifiers (Random Forests *vs.* LeNet-5) via analyzing generated adversaries.

	avg δx	P(larger δx)	avg δz	P(larger δz)	test accuracy
Random Forests	5.1097	0.27	1.3742	0.32	0.9045
LeNet-5	5.9665	0.73	1.7499	0.68	0.9871

- Experiments on text entailment



Premise	Hypothesis	Perturbed Hypothesis	Target Flip
the two boys are swimming with boogie boards .	the two boys are in their bath tub .	the two boys are in their room .	Contradiction => Neutral
an older women tending to a garden .	the lady is weeding her garden .	the lady is facing her garden .	Neutral => Entailment
boys with their backs against an incoming wave .	a group of people stand together .	a group of people are playing together .	Entailment => Neutral
workers standing on a lift .	workers walk off a lift .	there are some men climbing .	Contradiction => Entailment

For any questions, please email: **zhengliz@uci.edu**