Generating Natural Adversarial Examples
Zhengli Zhao, Dheeru Dua, Sameer Singh
University of California, Irvine

Motivation

- Adversarial examples [1]
  - $x' = \arg\min_{x \in [0,1]^D} \mathbb{E}_y [f(x+y) - f(x)]$.
  - $x' = \arg\min_{x \in [0,1]^D} \mathbb{E}_y [f(x+y) - f(x)]$.
  - $x' = \arg\min_{x \in [0,1]^D} \mathbb{E}_y [f(x+y) - f(x)]$.
  - $x' = \arg\min_{x \in [0,1]^D} \mathbb{E}_y [f(x+y) - f(x)]$.
  - $x' = \arg\min_{x \in [0,1]^D} \mathbb{E}_y [f(x+y) - f(x)]$.

An adversarial example applied to MNIST classifier of Random Forests.

- Disadvantages of these adversarial examples:
  - 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 $p_z(x)$ to distribution of real instances $p_{real}(x)$, generating fake $\hat{x}$ which are close to real $x$.
  - Critic discriminates between real instances $x$ and generated fake $\hat{x}$.
  - Wasserstein GAN [3]
    - Use Wasserstein-1 (also called Earth-Mover) distance instead.
    - Generator: $\max_{G(z)} \mathbb{E}_{x \sim p_{real}} [\log (1 - C_G(G(z)))]$.
    - Critic: $\max_{C(z)} \mathbb{E}_{x \sim p_{real}} [\log C_G(G(z))] + \mathbb{E}_{z \sim p_z} [\log (1 - C_G(G(z)))]$.
  - $\text{Perturbed Hypothesis}$
  - $\text{Target Flip}$

Proposed Approach

- Natural Adversarial Examples
  - For any question, please email: zhengliz@uci.edu

- Experiments on images
  - Interpretable natural adversaries

- Natural Adversarial Examples
  - $x' = G_G(x')$ as $z' = \arg\min_{z} \mathbb{E}_{x \sim p_{real}} [f(G(x')) - f(G_G(z))].$

- Experiments on images
  - Interpretive natural adversaries

Evaluation of black-box classifiers

- $\text{Original MNIST Digit Examples}$
  - $\text{Reconstructed Digit Examples}$
  - $\text{Random Forest Adversarial Examples}$
  - $\text{LeNet-5 Adversarial Examples}$

References