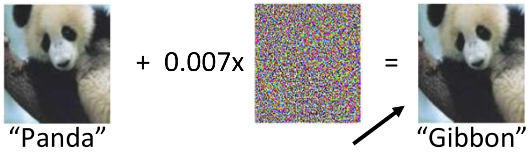
Adversarial Training with a Surrogate

Keane Lucas, Alec Jasen, and Lujo Bauer
Kjlucas, ajasen, lbauer@andrew.cmu.edu

2020 NeurIPS NewInML Workshop

Introduction

Machine-learning (ML) algorithms are fragile:

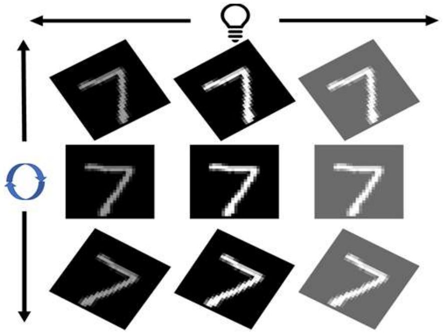


Adversarial Example
(image from Goodfellow 2014)

Typical attack approach: first-order gradient-based optimization (FGSM or PGD)

Perturbation set *P(x)* - set of images formed by small changes to *x* in which all members have the same classification, according to humans

How do we find adversarial examples on perturbation sets without using first-order gradients?

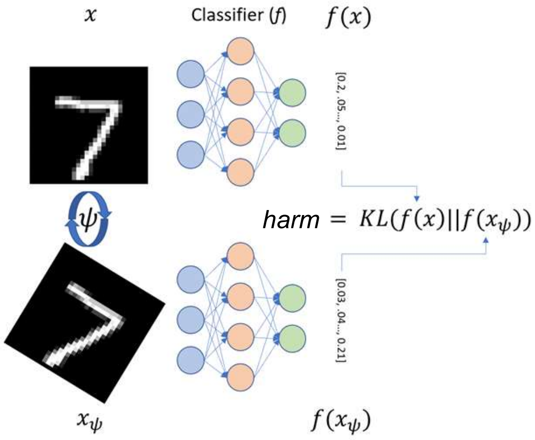


Perturbation set containing images of a `7' with changes in brightness and rotation

Explicit pixel-wise first-order gradients for these perturbations are not available and would need to be approximated or derived by alternate means

Estimating Harm—Method

If we train a surrogate neural network to estimate **harm** *h* of applying a perturbation *ψ* to an input *x* …

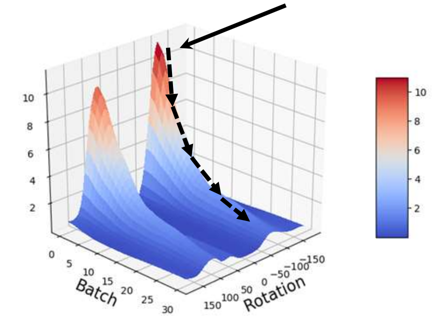


$$harm = KL(f(x)||f(x_\psi))$$

… then we can use the surrogate first-order gradients to directly approach effective adversarial examples

$$s : (X, \Psi) \rightarrow \mathbb{R}$$

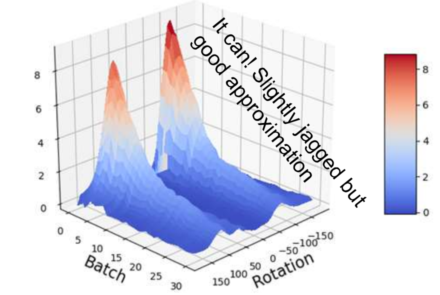
How does harm behave? - Harm distribution should flatten over the perturbation set as a result of adversarial training



Mean harm of perturbations (rotations) on the MNIST digit '1' of a classifier during adversarial training

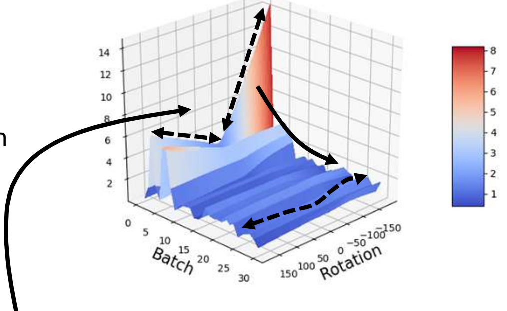
Estimating Harm - Viability

Can a surrogate neural network accurately predict model harm?



Harm surface estimated by surrogate trained using many queries from classifier

However, to limit extra computation, we should only use harm calculations derived from regular adversarial training

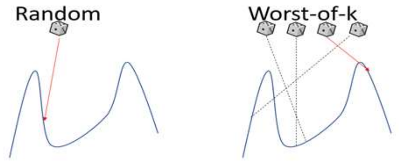


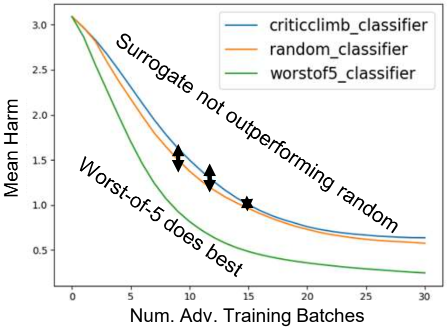
Harm surface estimated by surrogate neural network with only harm information gained during adversarial training

Initially, the surrogate estimates a rough approximation of the harm/perturbation relationship as linear, but develops more complex throughout training

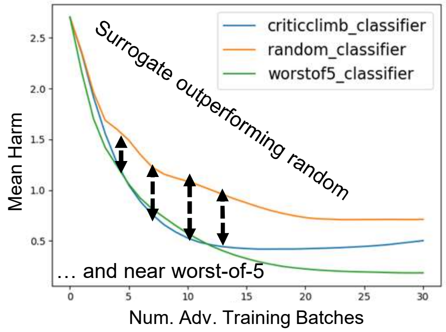
Adversarial Training Results

Does this trained surrogate help adversarial training?





MNIST (6, 7 removed) – Surrogate does not provide advantage above random



CIFAR10 – Surrogate does provide advantage compared to random and is competitive with worst-of-5

Electrical & Computer ENGINEERING

CyLab Carnegie Mellon University Security and Privacy Institute