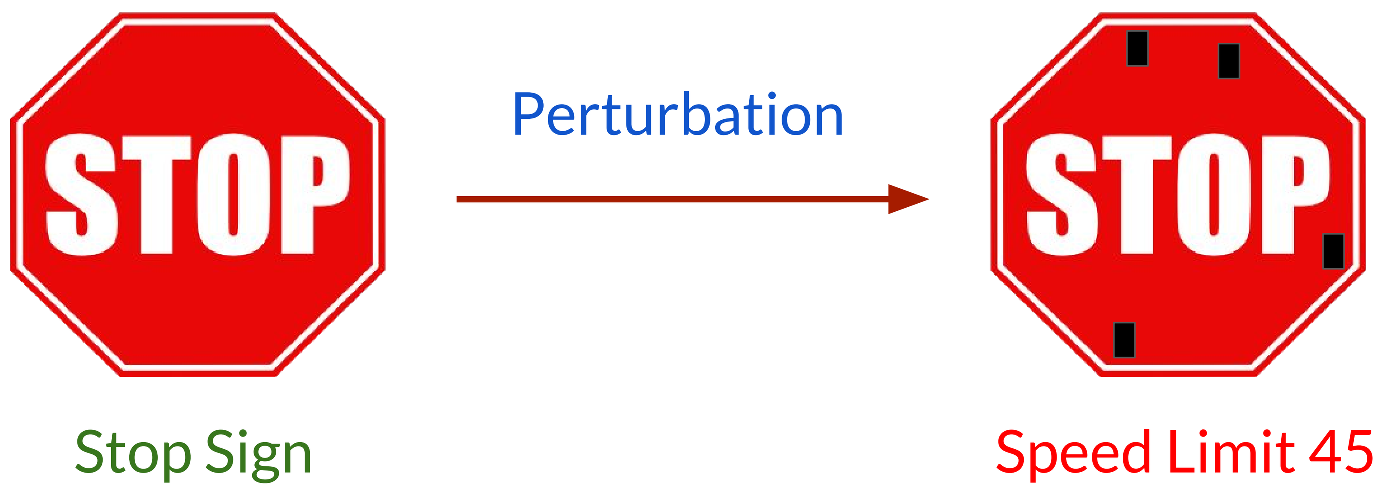
Ensembling as a Defense Against Adversarial Examples

Evan Liu (evanliu) Brendon Go (bgo)

Motivation

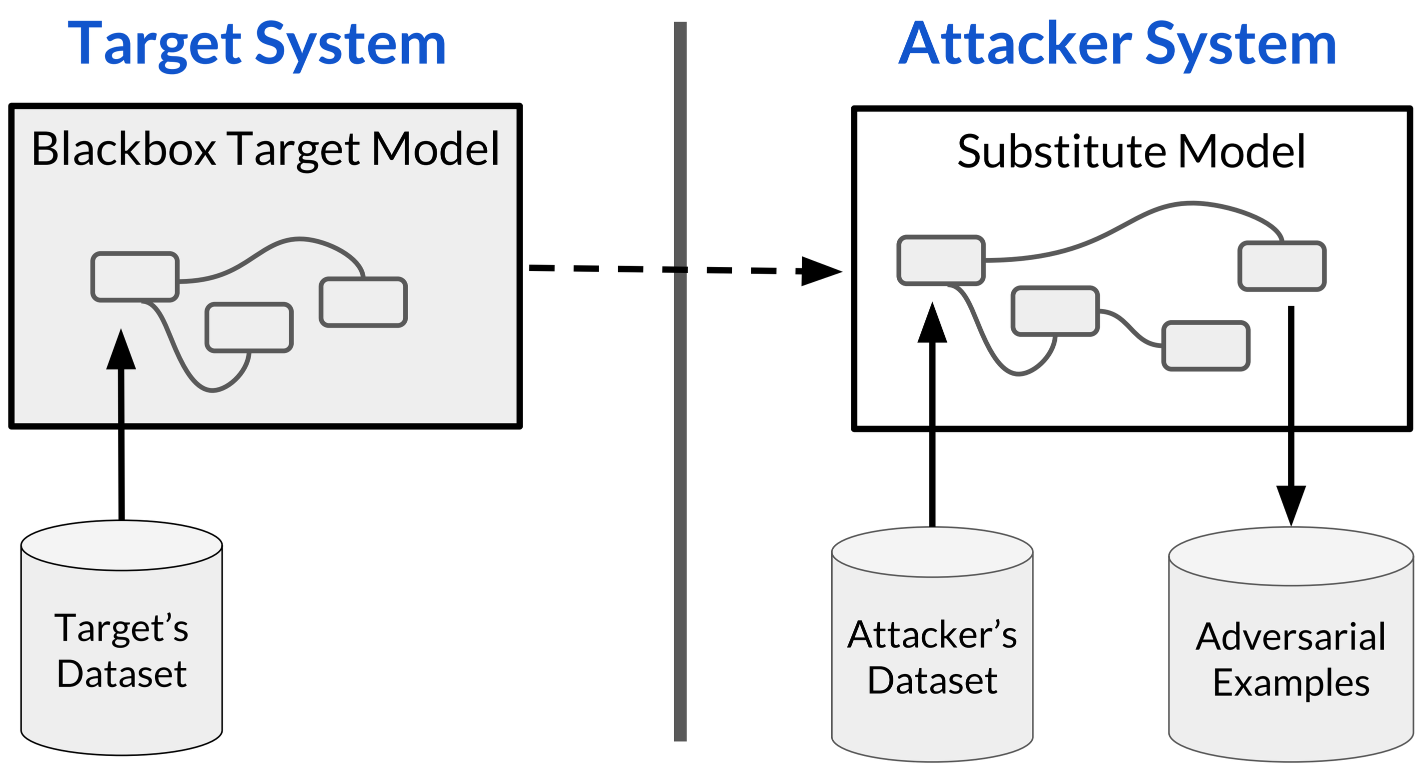
Adversarial Example: a maliciously crafted input that is easily classified correctly by humans, but is misclassified by a machine learning system



Attack Setting

Attacker Knowledge:

- does not know target's model internals or have the target database
- does know target's architecture and has own dataset



Transferability:

- Adversarial Examples generated on one system tend to generalize well to other systems

Ensembling Defense

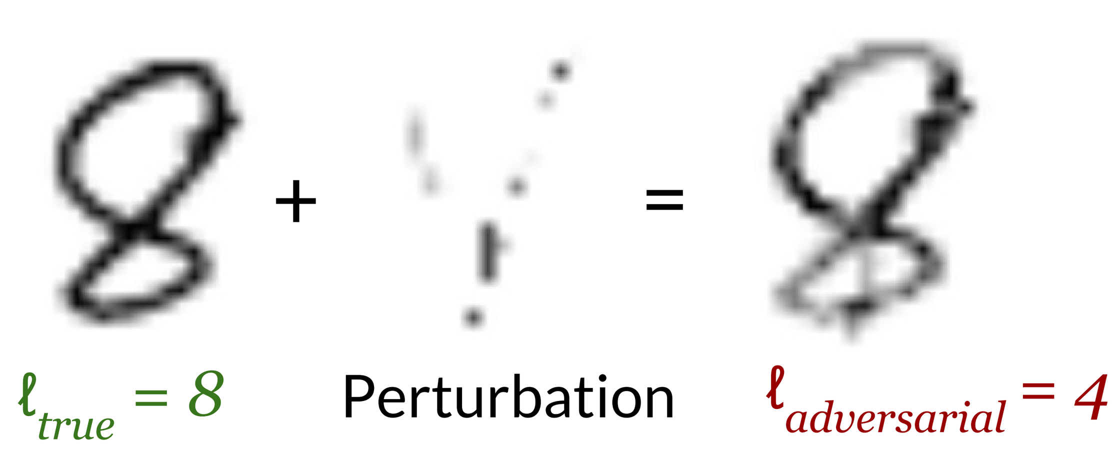
Adversarial examples built for some model transfer best to other models of same type

- Examples for kNN have relatively low transferability to CNN and vice versa
- Ensemble to take advantage of low transferability

Attack Generation

Change classification of example x from true label ℓ_{true} to adversarial label $\ell_{adversarial}$ without modifying many pixels

- Find some adversarial input x^* within an ε-ball of x by taking gradient steps toward $\ell_{adversarial}$
- f selects the most important pixels

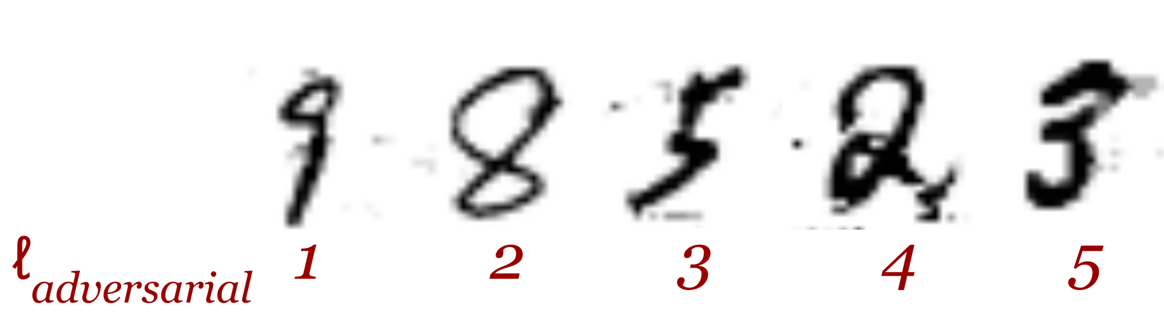


$x^* = x$

while $model(x^*)$ *is not* $\ell_{adversarial}$:

$\quad x^* := clip(x^* + \alpha f(\nabla_x s_{\ell_{adversarial}}(x^*)))$

$\quad$ *if* $x^* \notin B_\varepsilon$: *fail*



Data

MNIST Dataset: Labeled dataset of grayscaled images of handwritten images as an array of 784 pixel intensities

Split Train Data: Divide the 55000 training data as follows: 27000 to train blackbox, 27000 to train attacker, 1000 to generate adversarial examples for. Similarly split the validation data

Test Data: We test accuracy with MNIST test dataset (10000 examples) and we test adversarial success on ~1000 adversarial examples generated with each substitute model

Model

k-Nearest Neighbors (kNN):

- k = 5 in target, k = 3 in substitute, both with l_2-distance
- In substitute, gradient is approximated with soft-min:

$$s_k(x) = \frac{\sum_{z \in class_k(X)} e^{-||z-x||^2}}{\sum_{z \in X} e^{-||z-x||^2}}$$

Convolutional Neural Network (CNN):

- Substitute model is simple two convolutional layers followed by two fully connected layers
- Target model is two max pool layers followed by two fully connected layers, trained with dropout

Ensemble Model (Ens):

- Learns a parameter α and score examples with

$$s_{ensemble}(x) = \alpha s_{kNN}(x) + (1 - \alpha) s_{CNN}(x)$$

Results

Adversarial Success % / Partial Success %

Substitute \ Target	kNN	CNN	Ens
kNN	18.7%/ 30.1%	12.0%/ 24.3%	11.4%/ 19.9%
CNN	4.2%/ 12.4%	11.2%/ 18.4%	9.6%/ 15.6%
Ens	4.7%/ 12.9%	14.0%/ 22.8%	8.8%/ 15.2%

Accuracy

	Substitute	Target
kNN	96.3%	96.1%
CNN	96.3%	97.7%
Ens	96.1%	97.7%

Adversarial examples x^* created for substitute

- success if model(x^*) = $\ell_{adversarial}$
- partial success if model(x^*) ≠ ℓ_{true}

Results show that Ensemble model is more robust to adversarial examples crafted for kNN and for CNN

Furthermore, Ensemble model is more robust to adversarial examples crafted for other Ensemble models

Ensembling achieves this without sacrificing test accuracy

Future Work

Future work include examining ensembling as a defense for adversarial examples crafted using other attack generation methods, effectiveness with other datasets (CIFAR10), investigating other ensemble models, and how ensembling complements other defences like distillation.

References

Y. LeCun and C. Cortes. The mnist database of handwritten digits, 1998.

Papernot, N., McDaniel, P., and Goodfellow, I. Transferability in Machine Learning: from Phenomena to Blackbox Attacks using Adversarial Samples. ArXiv e-prints, May 2016b. URL http://arxiv.org/abs/1605.07277.

Paperno, N., McDaniel, P., Goodfellow, I., Jha, S., and al. Practical black-box attacks against deep learning systems using adversarial examples. arXiv preprint arXiv:1602.02697, 2016.

I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in Proceedings of the International Conference on Learning Representations, 2015. [7] J. Stallkamp, M. Schlipsing, J. Salmen, and