Are generative classifiers more robust to adversarial attacks?

Spoiler: Yes (when you have a good generative model)

Yingzhen Li[1], John Bradshaw[2,3], Yash Sharma[4]

[1]Microsoft Research Cambridge, [2]University of Cambridge, [3]MPI Tubingen, [4]Eberhard Karls University of Tubingen

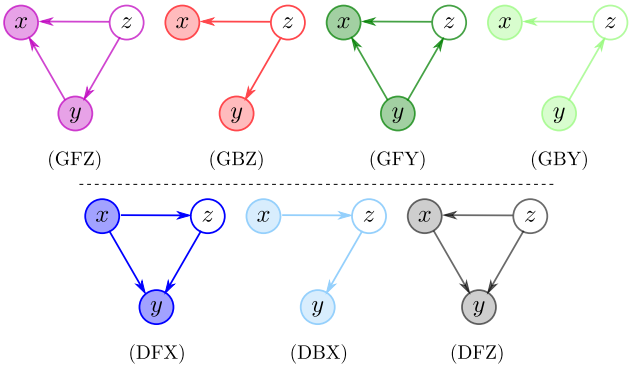
Our contributions

white-box/black-box attacks
FGSM, PGD, MIM, CW, SPSA
transferred adversarial examples

Bayesian NNs
Discriminative LVM classifiers
Generative LVM classifiers

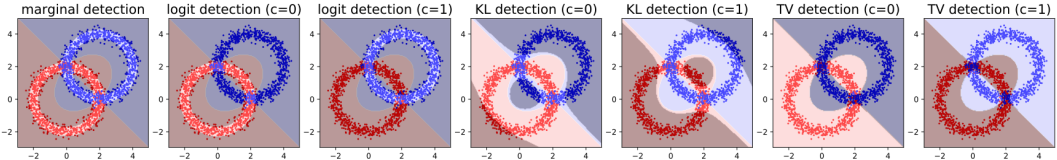
Marginal detection
Logit detection
KL detection

Deep latent variable model based classifiers



- Tested both **G**enerative (deep Bayes) and **D**iscriminative classifiers
- Model has either **F**ully connected or **B**ottleneck structure
- Approximate Bayes' rule with amortised approximate inference (IWAE)

Detecting adversarial examples



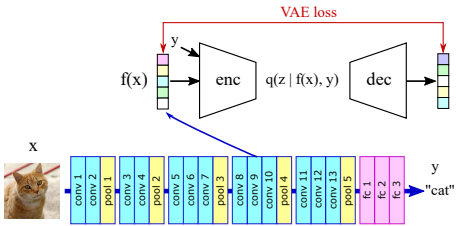
- Marginal detection: reject $\boldsymbol{x}$ if $-\log p(\boldsymbol{x}) > \delta$.
- Logit detection: reject $(\boldsymbol{x}, y_{\text{pred}})$ if $-\log p(\boldsymbol{x}, y_{\text{pred}}) > \delta$.
- Divergence detection: reject under-/over-confident predictions using a divergence metric
- δ is selected to achieve FPR on clean training data $= 5\%$.

Comes for free!

(no need to train extra detection networks/generative models/de-noising auto-encoders)

Observations

- Generative classifiers are more robust on MNIST
  - graphical model architecture does matter!
- Generative classifiers have worse clean accuracy on CIFAR-10
  - build a fusion model to help
- The tested attacks cannot fool both the generative classifier and the detection methods together
  - they share the same generative model



The proposed fusion model

Promising directions

better Bayesian NNs
better generative classifiers
fusion model

build causally informed generative models
explicitly model manipulations
unsupervised training on adversarial examples

welcome to discuss at 6:30pm tonight, poster #3