ADVERSARIAL MACHINE LEARNING

EXPLANATION FOR ADVERSARIAL EXAMPLES

15

- In *Explaining and Harnessing Adversarial Examples* Goodfellow et al. argues that adversarial examples exist because of the piece wise linearity in deep neural models
- Adversarial Examples occur at the difference between model and real class boundary.
- So purpose of the attack model is to perturb the original input so that it falls into desired region.

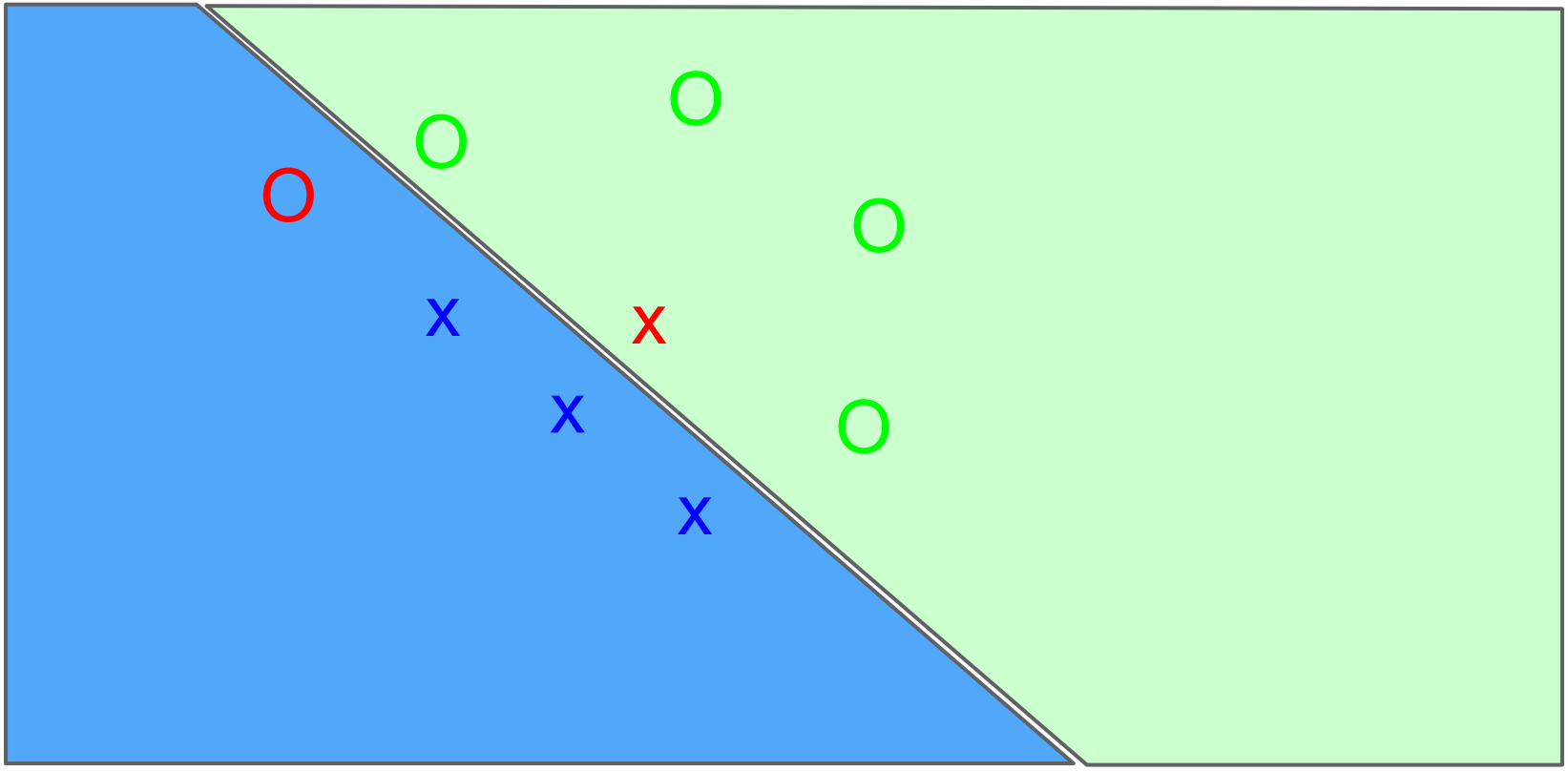
O
O
O
O
O
x
x
x
x

Image Source: iangoodfellow.com

ATTACK MODELS

- L-BFGS (Broyden–Fletcher–Goldfarb–Shanno)
- The Fast Gradient Sign Method (FGSM)
- Projected Gradient Descent (PGD or $FGSM^k$)
- Jacobean Based Saliency Map Approach (JSMA)
- Carlini Wagner Attack
- Black Box attack

EXPLANATION FOR ADVERSARIAL EXAMPLES

- In *Explaining and Harnessing Adversarial Examples* Goodfellow et al. argues that adversarial examples exist because of the piece wise linearity in deep neural models
- Adversarial Examples occur at the difference between model and real class boundary.
- So purpose of the attack model is to perturb the original input so that it falls into desired region.

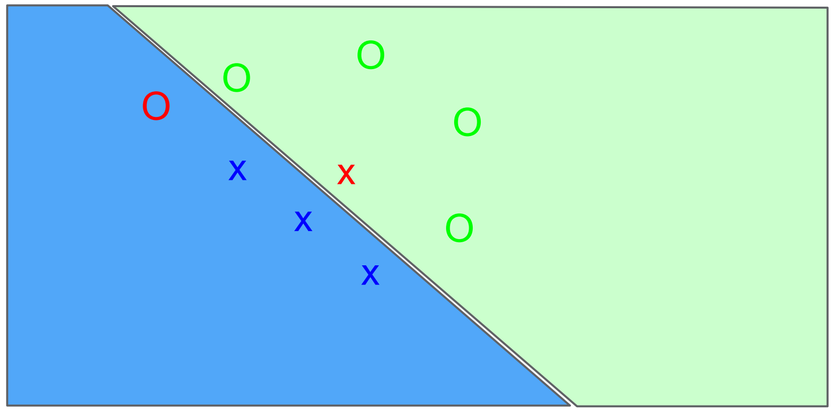

Image Source: iangoodfellow.com