ADVERSARIAL MACHINE LEARNING

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

16

ATTACK MODELS: L-BFGS

- Targeted attack proposed by Christian Szegedy et al.

$$\underset{||\delta||_2}{\arg\min} \text{ s.t. } \mathcal{F}(x+\delta) = \ell \ \& \ x+\delta \in [0,1]^m$$

- This is a hard problem. So instead they solves following problem using box constrained L-BFGS Method

$$\text{minimize } c||\delta||_2 + \mathcal{L}(x+\delta, \ell) \text{ subject to } x+\delta \in [0,1]^m$$

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks.

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