ADVERSARIAL MACHINE LEARNING

ATTACK MODELS: L-BFGS

- Targeted attack proposed by Christian Szegedy et al.

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

17

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

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

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

ATTACK MODELS: FGSM

- Proposed by Goodfellow et al.
- Uses L_∞ distance metric
- General Attack Model

$$x_{adv} = x + \epsilon \cdot sign(\nabla_x \mathcal{L}(x, y, \theta))$$

- Targeted version

$$x_{adv} = x - \epsilon \cdot sign(\nabla_x \mathcal{L}(x, \ell, \theta))$$

- Fast but does not guarantee optimal or closest adversary

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples.

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.