ADVERSARIAL MACHINE LEARNING

ATTACK MODELS: CARLINI WAGNER ATTACKS

- Nicholas Carlini and David Wagner proposer three attack models with L_0, L_2 and L_∞ distance
- L2 attack is most optimal and broke all the existing defences
- The defined the problem as following:

22

Carlini, N., & Wagner, D. (2017, May). Towards evaluating the robustness of neural networks.

$$\text{minimize } \mathcal{D}(x, x+\delta)$$

$$\text{such that } \mathcal{C}(x+\delta) = \ell$$

$$x + \delta \in [0,1]^n$$

$\mathcal{D}$ can be L_0, L_2, L_∞ distance

ATTACK MODELS: CARLINI WAGNER L2 ATTACK

- This is a hard problem. So they defined f such that:

$$\mathcal{C}(x+\delta) = \ell \text{ if and if only } f(x+\delta) \leq 0$$

- So the optimization problem becomes:

$$\text{minimize } \mathcal{D}(x, x+\delta) + c \cdot f(x+\delta) \quad \text{s.t. } x+\delta \in [0,1]^n$$

- Variable replacement

$$\text{They replaced } \delta_i \text{ as } \delta_i = \tfrac{1}{2}(\tanh(w_i)+1) - x_i$$

- So the optimization problem becomes

$$\text{minimize } \|\frac{1}{2}(\tanh(w)+1) - x\|_2^2 + c \cdot f(\frac{1}{2}(\tanh(w)+1)$$

$$\text{with } f \text{ defined as}$$

$$f(x') = \max(\max\{\mathcal{H}(x')_i : i \neq l\} - \mathcal{H}(x')_l, -\kappa).$$

- They solve optimization using multiple starting point gradient descent

Carlini, N., & Wagner, D. (2017, May). Towards evaluating the robustness of neural networks.

ATTACK MODELS: CARLINI WAGNER ATTACKS

- Nicholas Carlini and David Wagner proposer three attack models with L_0, L_2 and L_∞ distance
- L2 attack is most optimal and broke all the existing defences
- The defined the problem as following:

$$\begin{aligned} &\text{minimize } \mathcal{D}(x, x+\delta) \\ &\text{such that } \mathcal{C}(x+\delta) = \ell \\ &x + \delta \in [0,1]^n \end{aligned}$$

$$\mathcal{D} \text{ can be } L_0, L_2, L_\infty \text{ distance}$$

Carlini, N., & Wagner, D. (2017, May). Towards evaluating the robustness of neural networks.