ATTACK MODELS: CARLINI WAGNER L2 ATTACK

23

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

- This is a hard problem. So they defined f such that:
- So the optimization problem becomes:
- Variable replacement
- So the optimization problem becomes
- They solve optimization using multiple starting point gradient descent

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

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

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

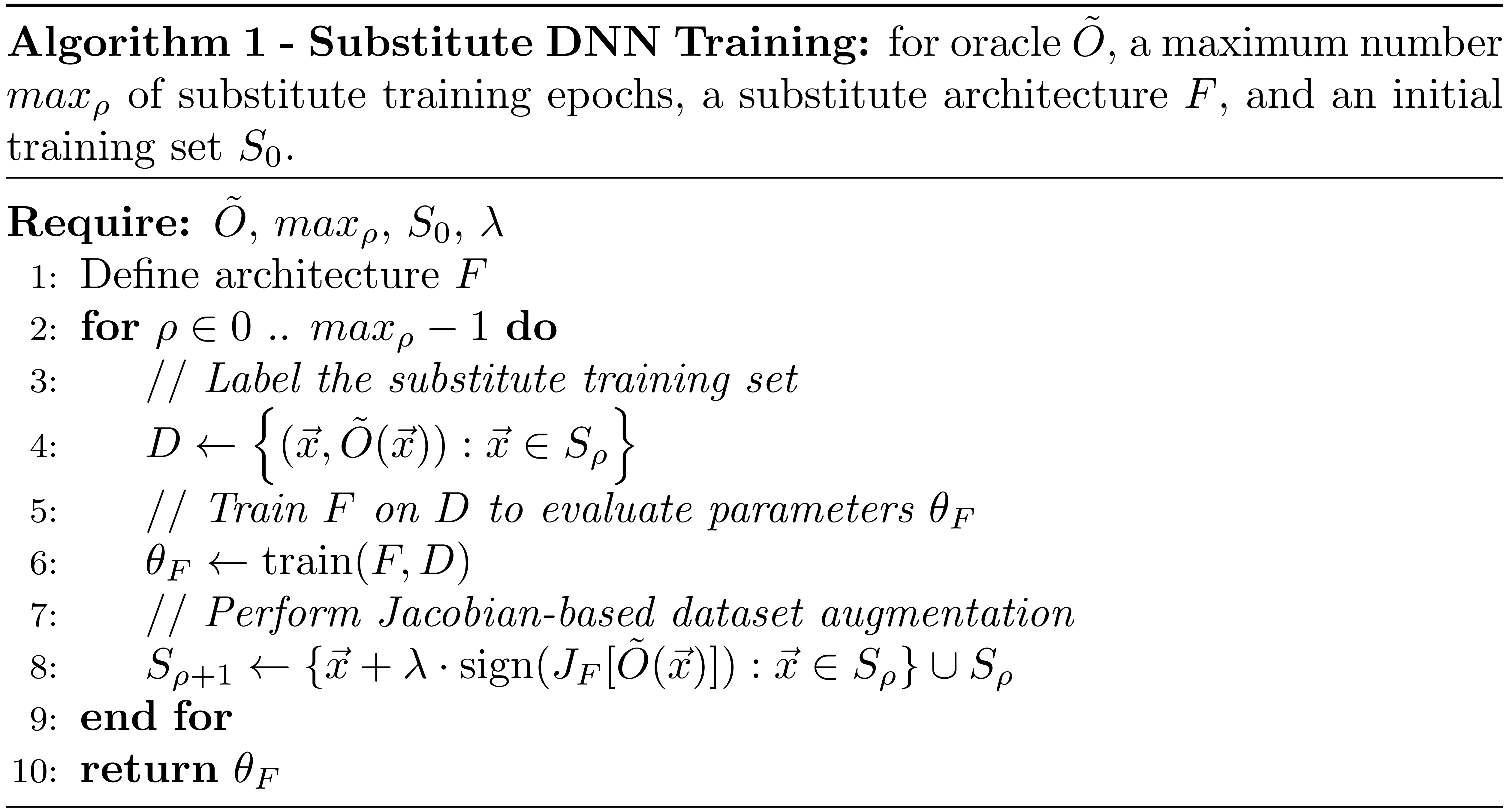
$$\text{minimize} \;\; \|\frac{1}{2}(\tanh(w)+1)-x\|_2^2 + c\cdot f(\frac{1}{2}(\tanh(w)+1)$$

with f defined as

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

PRACTICAL BLACK BOX ATTACK

- Papernot et al. proposed a practical black box attack on CNN.
- They first train a substitute DNN on target classifier (oracle õ)
- They used Adversarial Examples crafted on substitute DNN to attack oracle.



Algorithm 1 - Substitute DNN Training: for oracle $\tilde{O}$, a maximum number max_ρ of substitute training epochs, a substitute architecture F, and an initial training set S_0.

```
Require: $\tilde{O}$, $max_\rho$, $S_0$, $\lambda$
Define architecture $F$
for $\rho \in 0 .. max_\rho - 1$ do
    // Label the substitute training set
    $D \leftarrow \left\{ (\vec{x}, \tilde{O}(\vec{x})) : \vec{x} \in S_\rho \right\}$
    // Train $F$ on $D$ to evaluate parameters $\theta_F$
    $\theta_F \leftarrow \mathrm{train}(F, D)$
    // Perform Jacobian-based dataset augmentation
    $S_{\rho+1} \leftarrow \{ \vec{x} + \lambda \cdot \mathrm{sign}(J_F[\tilde{O}(\vec{x})]) : \vec{x} \in S_\rho \} \cup S_\rho$
end for
return $\theta_F$
```

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017, April). Practical black-box attacks against machine learning.

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

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

with f 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.