ADVERSARIAL MACHINE LEARNING

ATTACK MODELS: JSMA

- Saliency Map based greedy approach
- Modify the pixel who will impact the classifier output most
- Saliency Map is defined as:

20

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016, March). The limitations of deep learning in adversarial settings.

$$\mathcal{S}(x,\ell)[i] = \begin{cases} 0 \text{ if } \frac{\partial \mathcal{P}_l(\mathbf{x})}{\partial \mathbf{x}_i} < 0 \text{ or } \sum_{j \neq t} \frac{\partial \mathcal{P}_j(\mathbf{x})}{\partial \mathbf{x}_i} > 0 \\ \left(\frac{\partial \mathcal{P}_l(\mathbf{x})}{\partial \mathbf{x}_i}\right) \left|\sum_{j \neq t} \frac{\partial \mathcal{P}_j(\mathbf{x})}{\partial \mathbf{x}_i}\right| \text{ otherwise} \end{cases}$$

ATTACK MODEL : JSMA ALGORITHM

Algorithm 1 Crafting adversarial samples

$\mathbf{x}$ is the benign sample, ℓ is the target network output, $\mathcal{F}$ is the function learned by the network during training, Υ is the maximum distortion, and θ is the change made to features.

Input: $\mathbf{x}$, ℓ, $\mathcal{F}$, Υ, θ

```
x* ← x
Γ = {1 ... |x|}
while F(x*) ≠ l and ||δ_x|| < Υ do
    Compute forward derivative ∇P(x*)
    S = saliency_map (∇P(x*), Γ, l)
    Modify x*_{i_max} by θ s.t. i_max = arg max_i S(x, l)[i]
    δ_x ← x* − x
end while
return x*
```

$\mathbf{x}^* \leftarrow \mathbf{x}$

$\Gamma = \{1 \dots |\mathbf{x}|\}$

while $\mathcal{F}(\mathbf{x}^*) \neq l$ and $||\delta_{\mathbf{x}}|| < \Upsilon$ **do**

  Compute forward derivative $\nabla\mathcal{P}(\mathbf{x}^*)$

  $S = \texttt{saliency\_map}\,(\nabla\mathcal{P}(\mathbf{x}^*), \Gamma, l)$

  Modify $\mathbf{x}^*_{i_{max}}$ by θ s.t. $i_{max} = \arg\max_i S(\mathbf{x}, l)[i]$

  $\delta_{\mathbf{x}} \leftarrow \mathbf{x}^* - \mathbf{x}$

end while

return $\mathbf{x}^*$

ATTACK MODELS: JSMA

- Saliency Map based greedy approach
- Modify the pixel who will impact the classifier output most
- Saliency Map is defined as:

$$\mathcal{S}(x,\ell)[i] = \begin{cases} 0 \text{ if } \frac{\partial \mathcal{P}_l(\mathbf{x})}{\partial \mathbf{x}_i} < 0 \text{ or } \sum_{j \neq t} \frac{\partial \mathcal{P}_j(\mathbf{x})}{\partial \mathbf{x}_i} > 0 \\ \left(\frac{\partial \mathcal{P}_l(\mathbf{x})}{\partial \mathbf{x}_i}\right) \left|\sum_{j \neq t} \frac{\partial \mathcal{P}_j(\mathbf{x})}{\partial \mathbf{x}_i}\right| \text{ otherwise} \end{cases}$$

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016, March). The limitations of deep learning in adversarial settings.