ATTACK MODELS: FGSM^k OR PGD

- This is an iterative version of FGSM:
- Targeted version:
- Madry et al. proposes PGD as universal first order attack method
- This means defence against this attack would guarantee defence against all gradient based method

19

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks

$$x_{adv}^{0} = x; \quad x_{adv}^{t+1} = Clip_{x,\epsilon}\{x_{adv}^{t} + \alpha \cdot sign(\nabla_x \mathcal{L}(x, \theta, y))$$

$$x_{adv}^{0} = x; \quad x_{adv}^{t+1} = Clip_{x,\epsilon}\{x_{adv}^{t} - \alpha \cdot sign(\nabla_x \mathcal{L}(x, \theta, \ell))$$

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.

ATTACK MODELS: FGSM^K OR PGD

- This is an iterative version of FGSM:

$$x^0_{adv} = x; \quad x^{t+1}_{adv} = Clip_{x,\epsilon}\{x^t_{adv} + \alpha \cdot sign(\nabla_x \mathcal{L}(x, \theta, y))$$

- Targeted version:

$$x^0_{adv} = x; \quad x^{t+1}_{adv} = Clip_{x,\epsilon}\{x^t_{adv} - \alpha \cdot sign(\nabla_x \mathcal{L}(x, \theta, \ell))$$

- Madry et al. proposes PGD as universal first order attack method
- This means defence against this attack would guarantee defence against all gradient based method

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks