ADVERSARIAL MACHINE LEARNING

ATTACK MODELS: FGSM

- Proposed by Goodfellow et al.
- Uses L_∞ distance metric
- General Attack Model
- Targeted version
- Fast but does not guarantee optimal or closest adversary

18

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

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

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

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

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.