Chapter24 Attack and Defense

1. Motivation
    a) We seek to deploy ml not only in the labs, but also in real world
    b) We want the models robust to the inputs that are built to fool the model
    c) Especially useful for spam classification, malware detection, etc.
2. Attack
    a) Loss Function

       Training : $L_{train}(\theta) = C(y^0, y^{true})$ x fixed

       Non-targeted Attack : $L(x') = -C(y', y^{true})$ θ fixed

       Targeted Attack : $L(x') = -C(y', y^{true}) + C(y', y^{false})$

       Constraint : $d(x^0, x') \leq \varepsilon$
    b) Constraint
        i) L2-norm

           $$d(x^0, x') = ||x^0 - x'||_2$$
           $$= (\Delta x_1)^2 + (\Delta x_2)^2 + (\Delta x_3)^2 + \cdots$$
        ii) L-infinity-norm

           $$d(x^0, x') = ||x^0 - x'||_\infty$$
           $$= \max\{\Delta x_1, \Delta x_2, \Delta x_3, \dots\}$$
    c) How to attack

       $$x^* = arg \underbrace{min}_{d(x^0,x')\leq\varepsilon} L(x')$$

       Start from original image x^0

       for t = 1 to T:

           $x^t \leftarrow x^{t-1} - \eta \nabla L(x^{t-1})$

           if $d(x^0, x^t) > \varepsilon$:

               $x^t \leftarrow fix(x^t)$

       def $fix(x^t)$:

           for all x fulfill $d(x^0, x^t) \leq \varepsilon$

           return the one closest to x^t
    d) Attack Approaches
        i) Different optimization methods & Different constraints
        ii) Fast Gradient Sign Method

           $$x^t \leftarrow x^0 - \varepsilon \Delta x$$

           $$\Delta x = \begin{bmatrix} sign(\partial L/\partial x_1) \\ sign(\partial L/\partial x_2) \\ sign(\partial L/\partial x_3) \\ \cdots \end{bmatrix}, only\ have\ 1\ or -1$$
    e) White Box v.s. Black Box
        i) In the previous attack, we fix network parameters θ to find optimal x'
        ii) To attack, we need to know network θ, this is called white box attack
        iii) Black Box Attack is possible

f) Black Box Attack

  If you have the training data of the target network

    Train a proxy network yourself

    Using the proxy network to generate attacked objects

  Otherwise, obtaining input-output pairs from target network

g) Universal Adversarial Attack

h) Beyond Images

  i) Attack Audio https://adversarial-attacks.net

  ii) Attack Text

3. Defense

  a) Adversarial Attack cannot be defended by weight regularization, dropout and model ensemble

  b) Two types of defense

    Passive defense : finding the attached image without modifying the model

    Proactive defense : training a model that is robust to adversarial attack

  c) Passive Defense

    Smoothing

    Feature Squeeze

    Randomization at Inference Phase

  d) Proactive Defense

    Find adversarial input $\tilde{x}^n$ given x^n by an attack algorithm

    We have new training data, ensemble to data augmentation

    This method would stop algorithm A, but is still vulnerable for algorithm B