# Adversarial

$(x, y_{true})$ - object $f_\theta(x)$ - neural network

$\mathcal{L}(\theta, x, y)$ - loss value

$f_\theta(x) = y_{true}$ $\quad \tilde{x} \approx x: f_\theta(\tilde{x}) \neq y_{true}$

$\|\tilde{x} - x\|_\infty < \frac{\varepsilon}{255}$ $\quad \varepsilon$ - attack budget

$\|$

$\max_i |\tilde{x}_i - x_i|$

- White Box - gradient are available
- Black Box - gradient not available

- Targeted attack: $f_\theta(x) = y_{target}$
- Untargeted attack: $f_\theta(x) \neq y_{true}$

FGSM (Fast Gradient Sign Method)

- untargeted

$$\tilde{x} = x + \varepsilon \cdot \text{sign}\left(\nabla_x \mathcal{L}(\theta, x, y_{true})\right)$$

- targeted

$$\tilde{x} = x - \varepsilon \,\text{sign}\left(\nabla_x \mathcal{L}(\theta, x, y_{target})\right)$$

# I-FGSM (iterative FGSM)

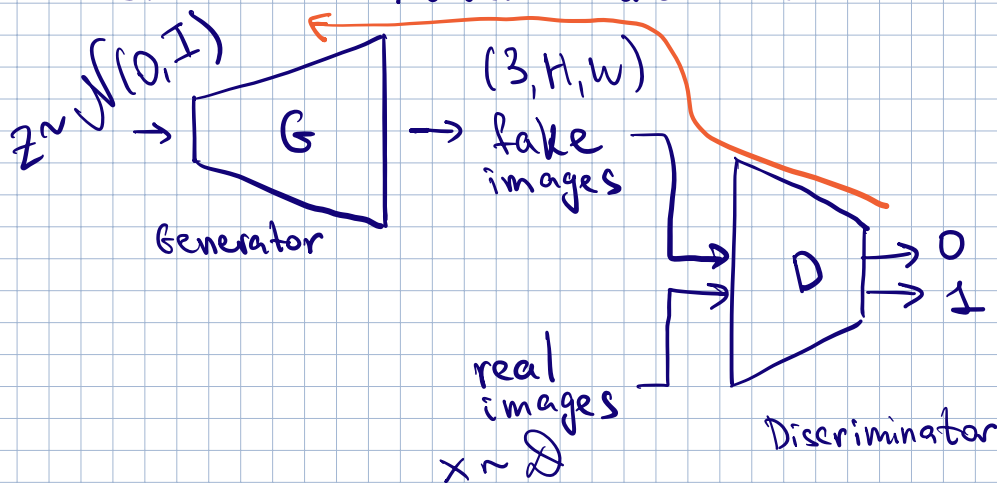
$\tilde{x}_0 = x$

for $i = 1, \dots, N$

$$\tilde{x}_i = \tilde{x}_{i-1} + \frac{\varepsilon}{N} \operatorname{sign}\left(D_x h(\theta, \tilde{x}_{i-1}, y_{true})\right)$$

$\tilde{x} = \tilde{x}_N$

# Generative Adversarial Networks (GANs)




$D(x) \in [0, 1]$

$$\min_G \max_D \left\{ \mathbb{E}_{x \sim p_{data}} [\log D(x)] + \mathbb{E}_{z \sim \mathcal{N}(0,I)} [\log (1 - D(G(z)))] \right\}$$

StyleGAN

cGAN (conditional GAN)

$G_\theta(z, c)$

$D^1_\varphi(x)$ - head for GAN loss

$D^2_\varphi(x)$ - head for classification loss

# Img 2 Img models

Pix 2 Pix $\quad x_S \sim p_S(x),\ x_T \sim p_T(x)$

$G(x_S) \simeq x_T$

MSE loss: $\mathcal{L}_{MSE} = \mathbb{E}_{x_S, x_T} \| G(x_S) - x_T \|_2^2 \to \min\limits_G$

GAN loss: $\mathcal{L}_{GAN}(G, D)$

MSE + GAN loss: $\alpha \mathcal{L}_{MSE} + (1-\alpha)\mathcal{L}_{GAN}$

$$\mathcal{L}_{GAN} = \mathbb{E}_{x \sim p_{data}} [\log D(x)] + \mathbb{E}_{x \sim p_{gen}} [\log (1 - D(x))] \to \max\limits_D$$

$D^*(x) - ?$

$$\mathcal{L}_{GAN} = \int_{x \in \Sigma} [p_{data}(x) \log D(x) + p_{gen}(x) \log (1 - D(x))] \, dx$$

$$g(c) = p_{data}(x) \cdot \log c + p_{gen}(x) \log (1 - c)$$

$$g'(c) = \frac{p_{data}(x)}{c} - \frac{p_{gen}(x)}{1 - c} = 0$$

$$(1 - c) p_{data}(x) - c\, p_{gen}(x) = 0$$

$$D^*(x) = c = \frac{p_{data}(x)}{p_{data}(x) + p_{gen}(x)}$$

× $\quad p_{gen}(x) \gg 1 \qquad D^*(x) \simeq 0$

$p_{data}(x) \simeq 0$

$$\mathcal{L}_{GAN} = \int\Big[ p_{data}(x) \log \frac{p_{data}(x)}{p_{data}(x) + p_{gen}(x)} + p_{gen}(x) \log \frac{p_{gen}(x)}{p_{data}(x) + p_{gen}(x)}$$

$$+ p_{data}(x) \log 2 - p_{data}(x) \log 2 + p_{gen}(x) \log 2 - p_{gen}(x) \log 2 \Big] dx$$

$$\int\Big[ p_{data}(x) \log \frac{2\, p_{data}(x)}{p_{data}(x) + p_{gen}(x)} + p_{gen}(x) \log \frac{2 \cdot p_{gen}(x)}{p_{data}(x) + p_{gen}(x)} \Big] dx$$

$$- \log 2 \underbrace{\int (p_{data}(x) + p_{gen}(x))\, dx}_{=2} =$$

$$= KL\left(p_{data} \,\middle\|\, \frac{p_{data} + p_{gen}}{2}\right) + KL\left(p_{gen} \,\middle\|\, \frac{p_{data} + p_{gen}}{2}\right)$$

$$- 2\log 2 = 2\, JS(p_{data}, p_{gen}) - 2\log 2$$

Mode collapsing