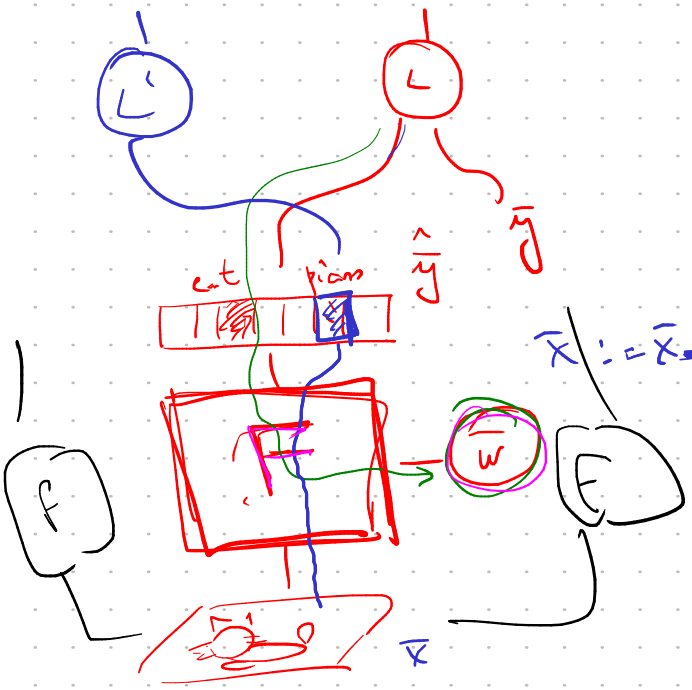
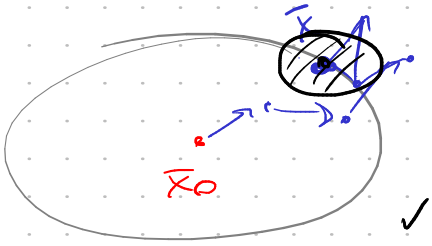


$$L(\bar{w}) = \sum_n l(\bar{w}, \bar{x}_n, y_n)$$

$$L'(\bar{x})$$

$$L'(\bar{x}, \bar{y}, \bar{w}) \xrightarrow{\bar{x}} \min \quad \text{s.t. } \|\bar{x} - \bar{x}_0\| \le \varepsilon$$

$$\bar{x} := \bar{x} - \eta \nabla_{\bar{x}} L'(\bar{x})$$



white-box attack

✓ grey-box attack

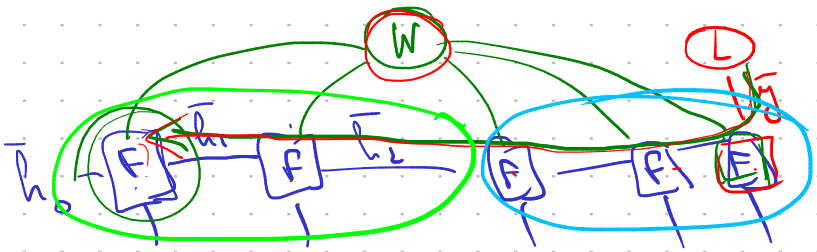
✓ black-box attack

$$\bar{w} := \bar{w} - \eta \cdot \nabla_{\bar{w}} L_t$$

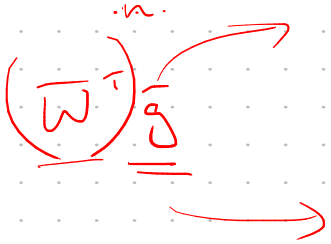
$$\bar{w} := \bar{w} - \eta \cdot \nabla_{\bar{w}} L\left(\bar{w}, \bar{x} - \lambda \cdot \nabla_{\bar{x}} L(\bar{w}, \bar{x})\right)$$

bag-of-freebies — Self-adversarial training — augmentation

YOLOv4



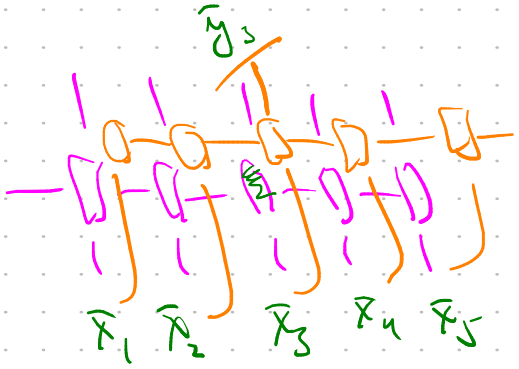
$$F(\bar{h}_k, \bar{x}_{k+1}) = (\bar{y}_{k+1}, \bar{h}_{k+1})$$



Self advers. training is cool

$\bar{x}_1$ $\bar{x}_2$ $\bar{x}_5$

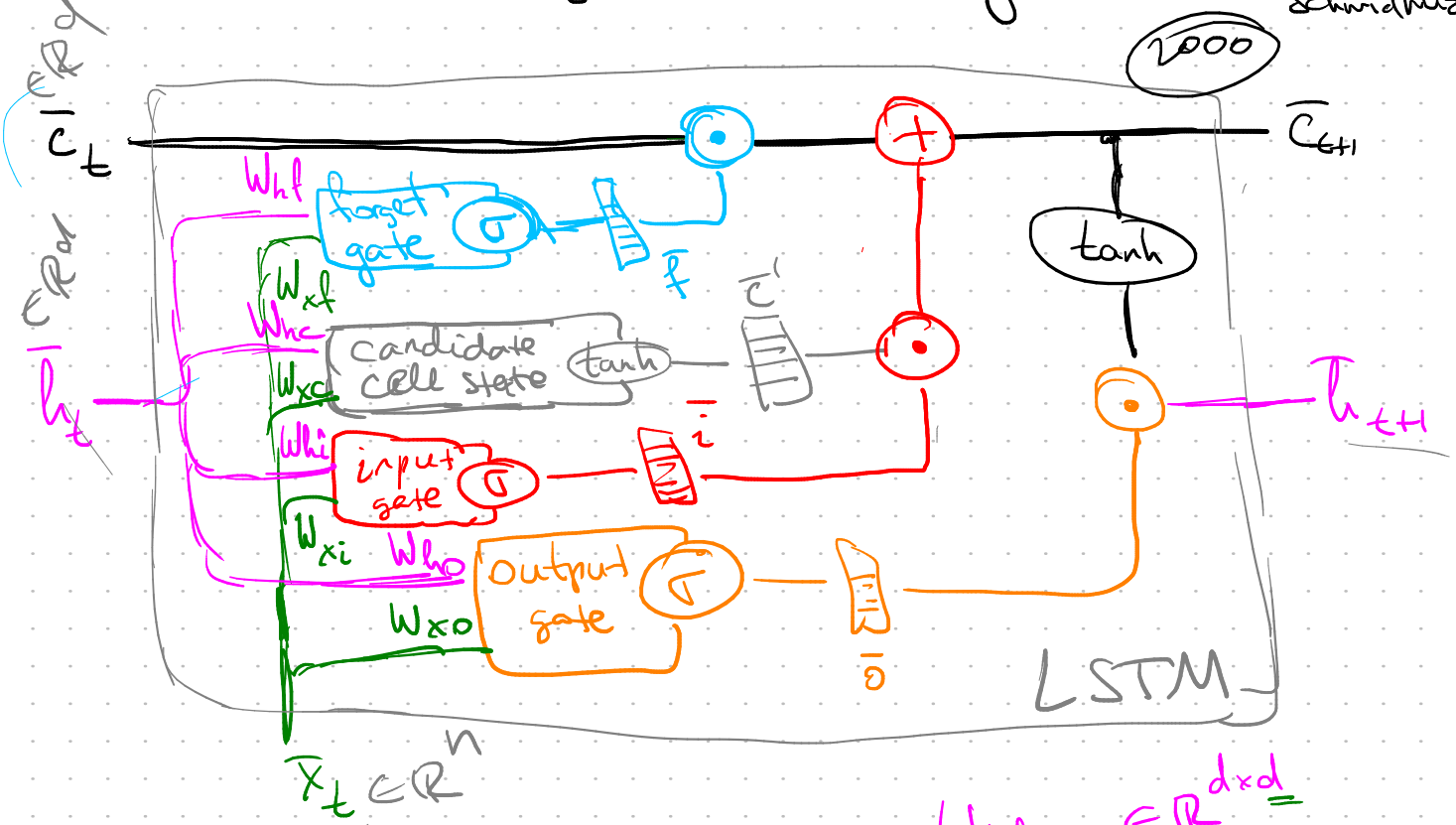
- exploding gradients
- vanishing gradients



LSTM – long short-term memory

1995 Hochreiter Schmidhuber

2000



$W_{hf}, \ldots \in \mathbb{R}^{d \times d}$

$W_{xf}, \ldots \in \mathbb{R}^{d \times n}$