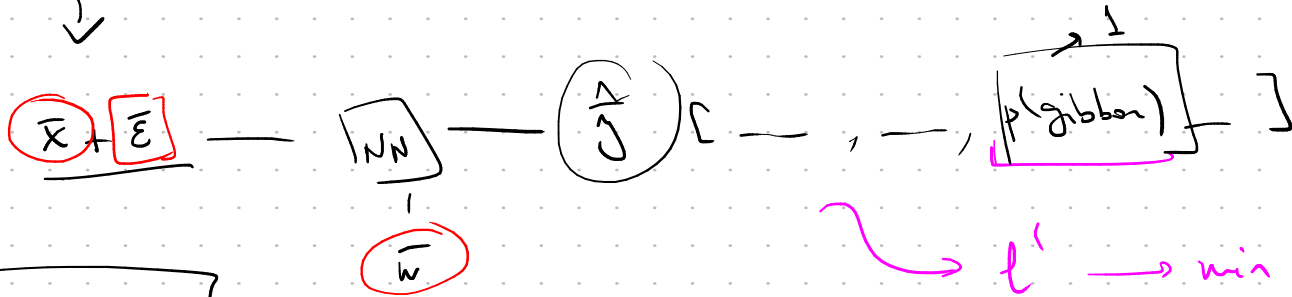
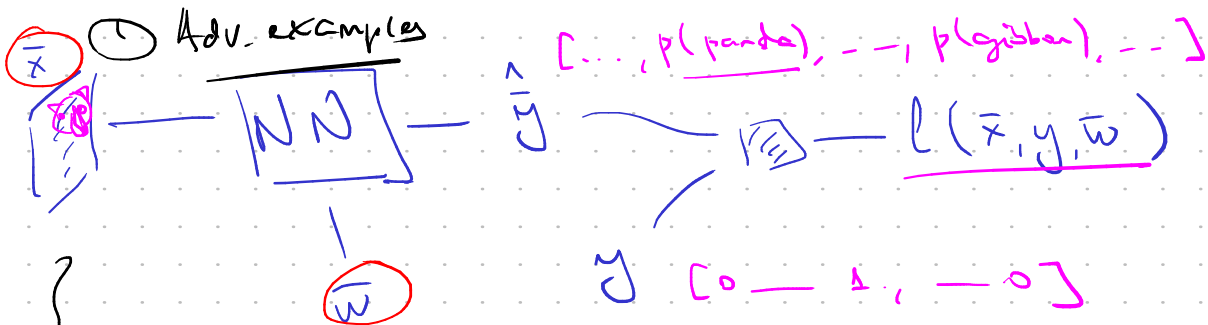


Adv. examples



$$F(\bar{x}, \bar{w})$$

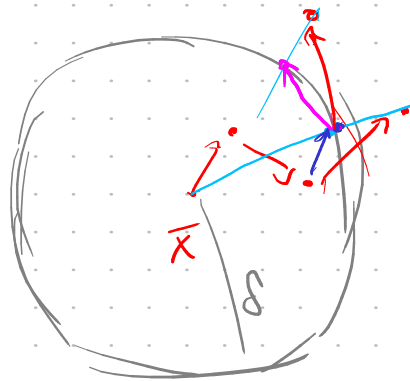
$$\nabla_{\bar{w}} F$$

$$\nabla_{\bar{x}} F$$

$$\nabla_{\bar{\epsilon}} l'(\bar{x} + \bar{\epsilon}, y, \bar{w})$$

$$l' \rightarrow \min$$

$$\|\bar{\epsilon}\| \leq \delta$$

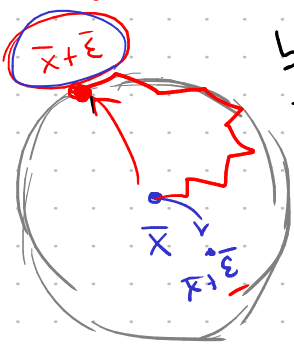


white-box attack

grey-box

black-box

self-adversarial augmentation



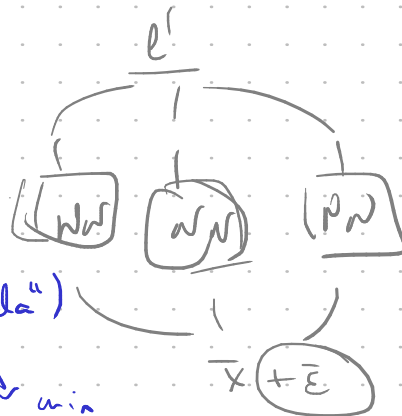
bag-of-freelbies

$$l(\bar{x}, \bar{w}) = p(\text{"panda"})$$

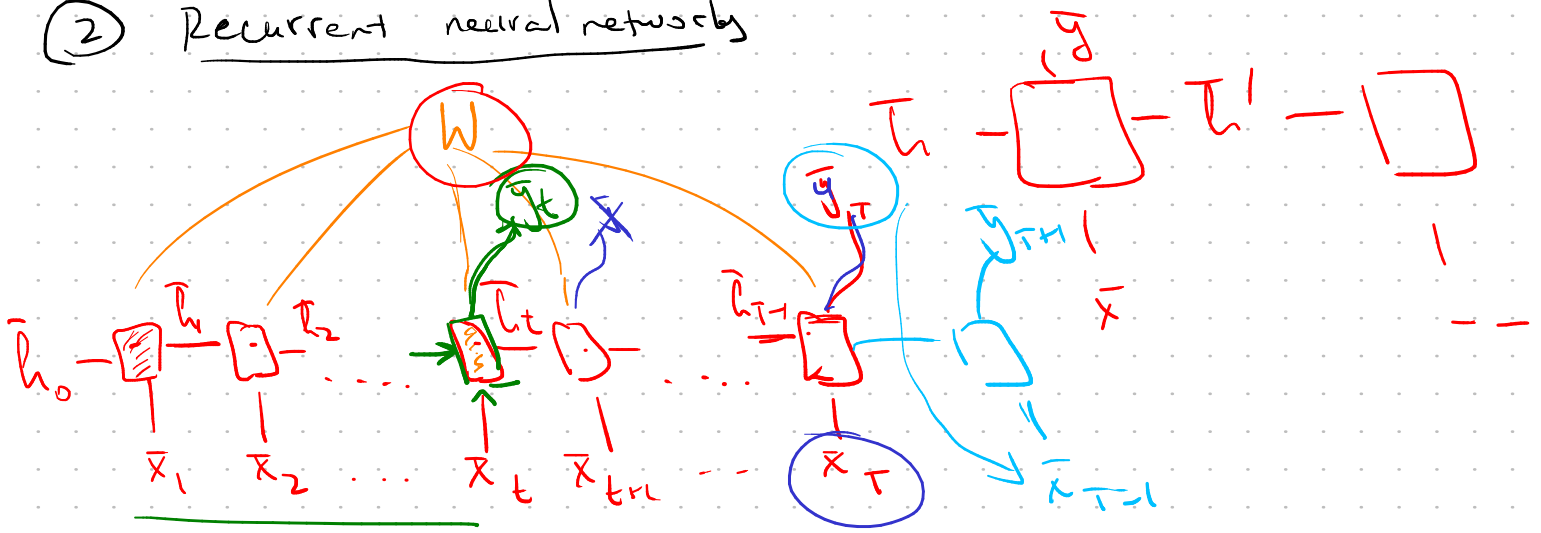
$\downarrow \min$

$$\bar{x} + \bar{\epsilon} : \arg\min_{\bar{x}} l(\bar{x}, \bar{w})$$

$$(\bar{x}, y = \text{"panda"})$$



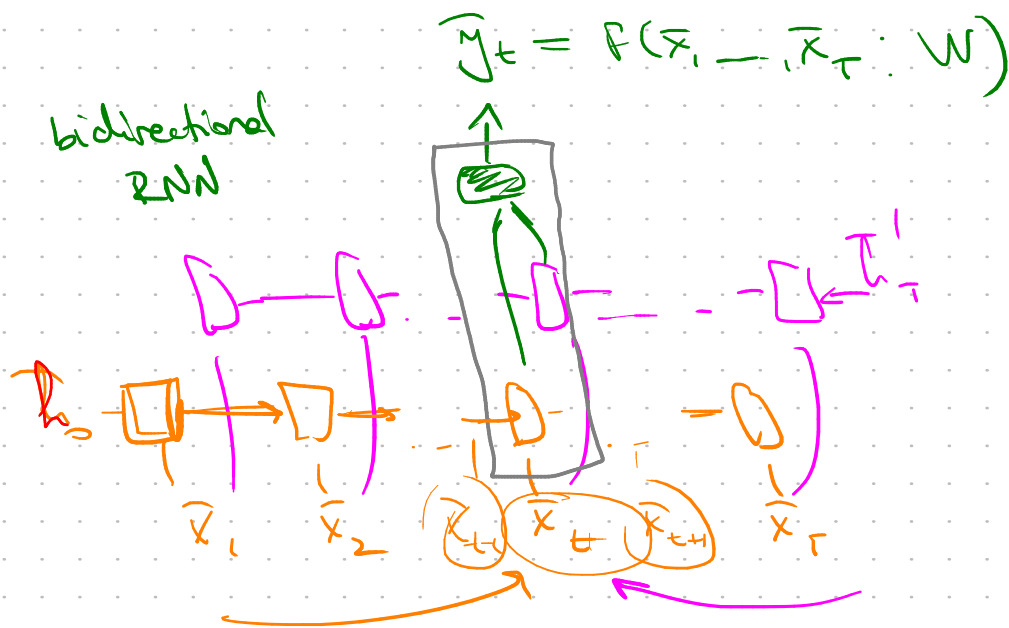
② Recurrent neural networks



autoregressive

$$\bar{y}_t = F(\bar{x}_1, \dots, \bar{x}_t; W)$$

left context



bidirectional
RNN

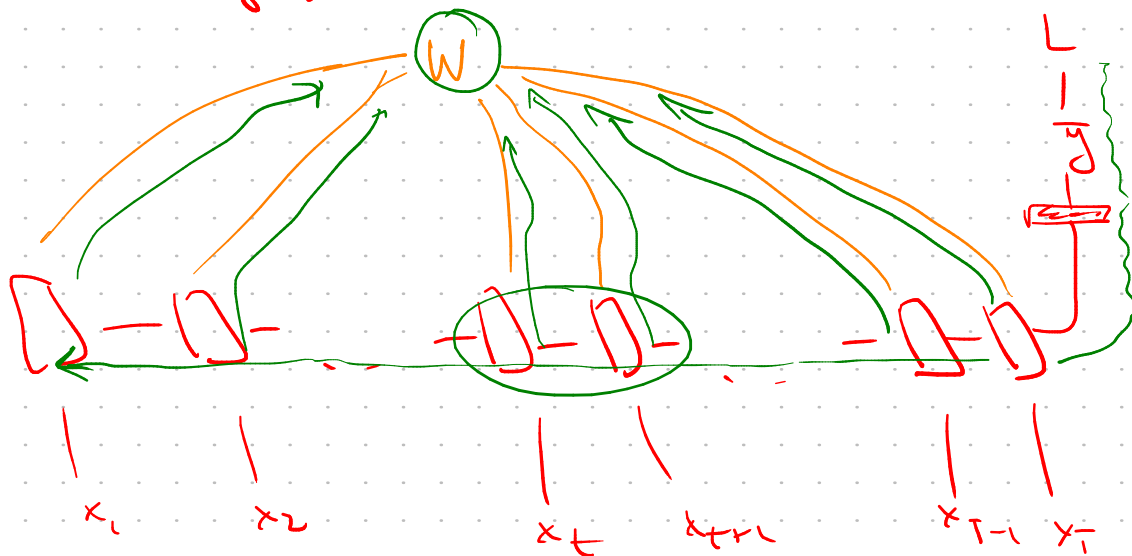
$$\bar{y}_t = F(\bar{x}_1, \dots, \bar{x}_T; W)$$

$$h_0 \rightarrow W h_0 \rightarrow W^2 h_0 \rightarrow \dots \rightarrow W^T h_0$$

- exploding gradients

$$\|g\| \rightarrow \infty, \|L\| \rightarrow \infty$$

- vanishing gradients

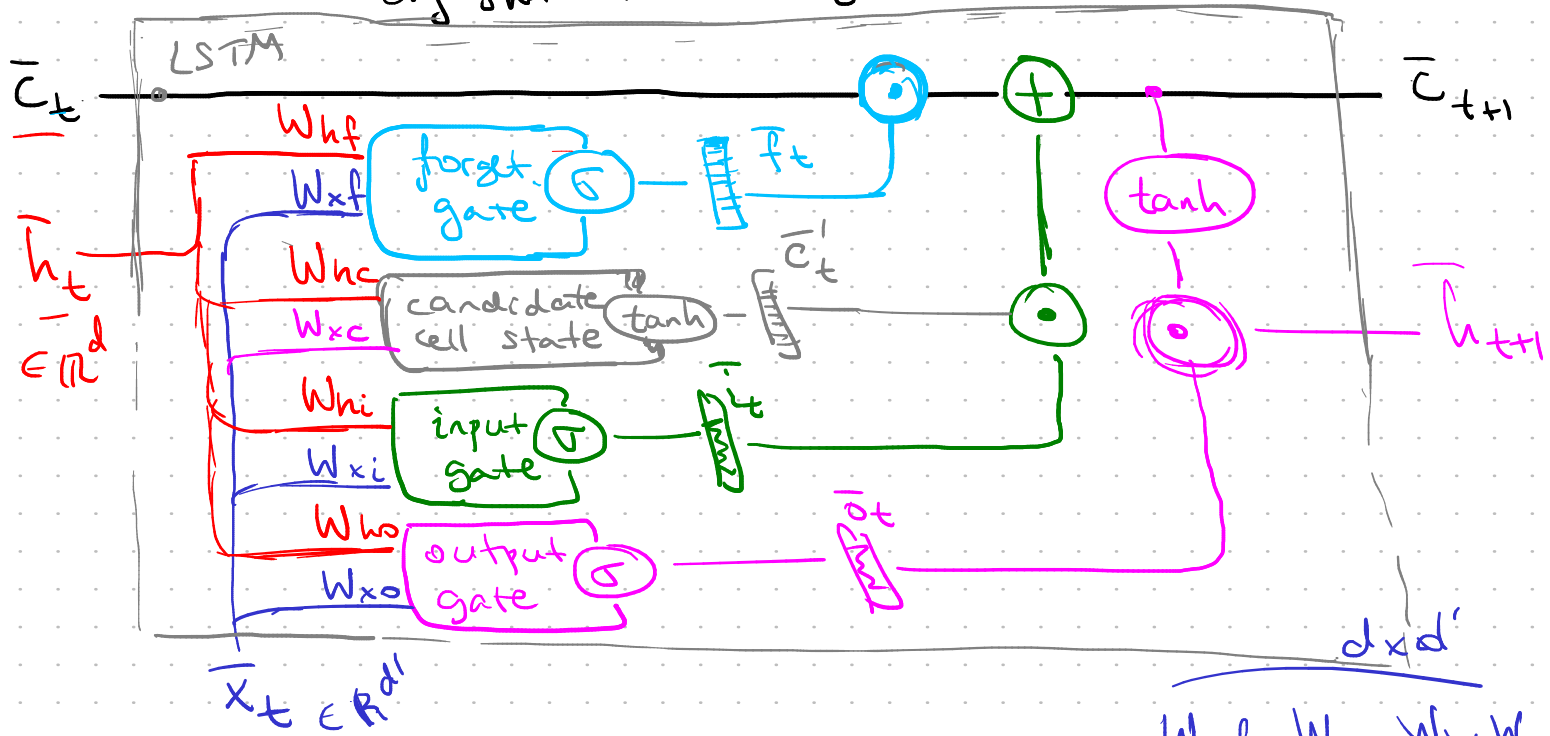


$$\frac{\partial L}{\partial w_i} = \sum_{t=1}^T \left[\frac{\partial L}{\partial y_t} \frac{\partial y_t}{\partial w_i} \right]$$

③ LSTM & Constant error carousel

long short-term memory

1995, 2000 Gers
Kochreiter, Schmidhuber



$d \times d'$
 $W_{xf}, W_{xc}, W_{xi}, W_{xo}$
 $W_{hf}, W_{hc}, W_{hi}, W_{ho}$
 $d \times d$

