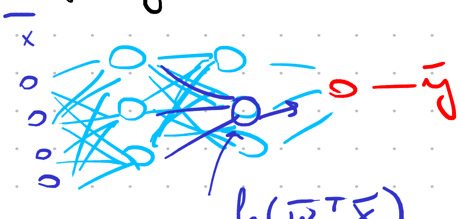
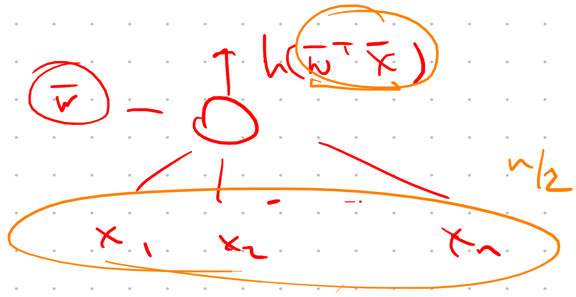
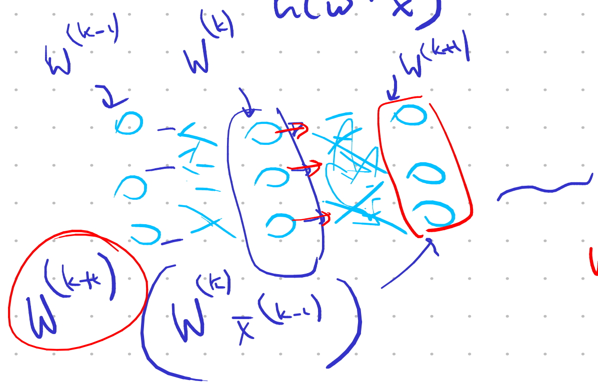
① Dropout | fully connected / feedforward



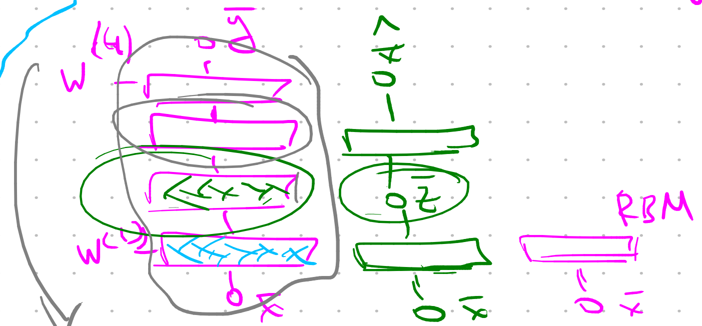




$$W^{(k+1)} W^{(k)} \bar{x} = (W^{(k+1)} \cdot A)(A^{-1} W^{(k)})\bar{x}$$

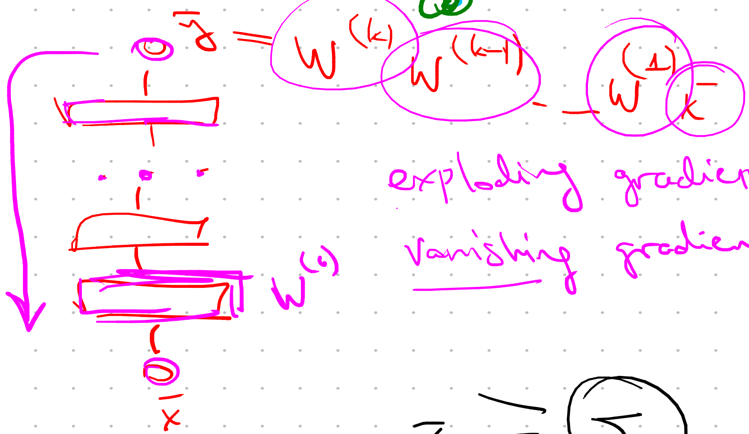
② Weight initialization

unsupervised pretraining



fine-tuning

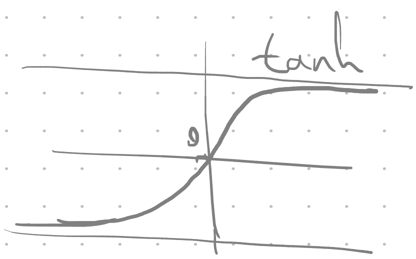
autoencoders



$= W^{(k)} W^{(k-1)} \ldots W^{(1)} \bar{x}$

exploding gradients

vanishing gradients



$$\bar{x} \to \Sigma \to \underline{\bar{w}^T\bar{x} = y} \to h \to h(y)$$

$$Var[y] = \sum_i Var[y_i] \qquad y = \sum_{i=1}^{n} \underbrace{w_i x_i}_{y_i} = \sum_i y_i$$

$$Var[y_i] = Var[w_i x_i] = E[w_i^2 x_i^2] - E[w_i x_i]^2 =$$

$$= \cancel{E[x_i]^2}\, Var[w_i] + \cancel{E[w_i]^2\, Var[x_i]} + Var[w_i] \cdot Var[x_i]$$

$E[x_i] = 0$ $\qquad$ $E[w_i] = 0$

Symm. act. func.:

$$Var[y_i] = Var[w_i]\,Var[x_i]$$

$$Var[y] = n \cdot Var[w_i] \cdot Var[x_i]$$

≈ 1

$\times 1/3$

$$Var[w_i] = O(1/n)$$

$$w_i \sim Unif\left(\left[-\frac{\sqrt{3}}{\sqrt{n}}, \frac{\sqrt{3}}{\sqrt{n}}\right]\right)$$

Xavier init

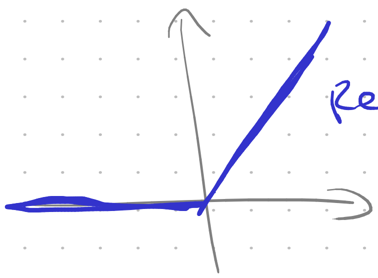
Xavier Glorot (2010)

Y. Bengio

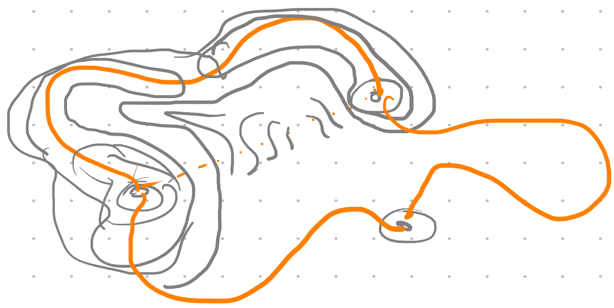
$$Var(Unif([a,b])) = \frac{(b-a)^2}{12}$$

$$Var\left(Unif\left(\left[-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right]\right)\right) = \frac{(2/\sqrt{n})^2}{12} = \frac{1}{3n}$$

Kaiming He



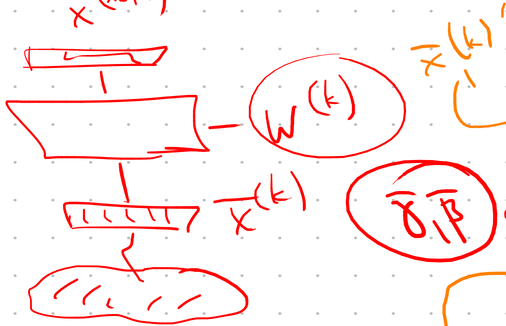
ReLU $\rightsquigarrow$ $Var[w_i] = 2/n$ He init

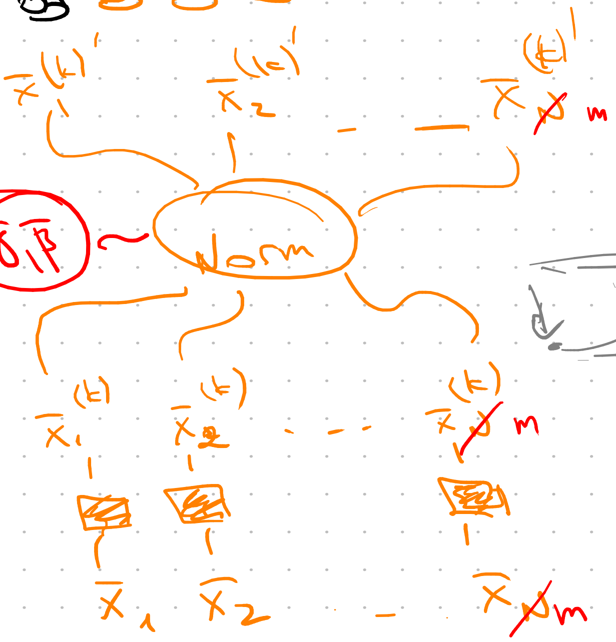


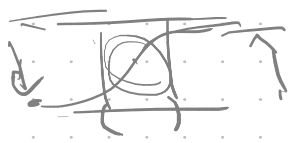
3 Batch normalization

Covariate shift

Internal covariate shift



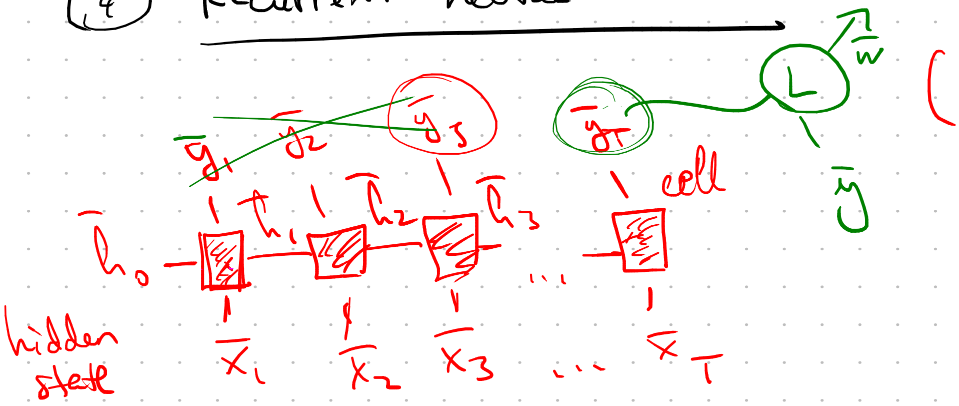




$$\bar{x} := \frac{x - E[x]}{\sqrt{Var[x]}}$$

$$BN(\bar{x}) = \frac{\bar{x} - E_m[\bar{x}]}{\sqrt{Var_m[\bar{x}]}} \odot \underbrace{\bar{\gamma}}_{\text{scale}} + \underbrace{\bar{\beta}}_{\text{shift}}$$

(4) Recurrent neural networks

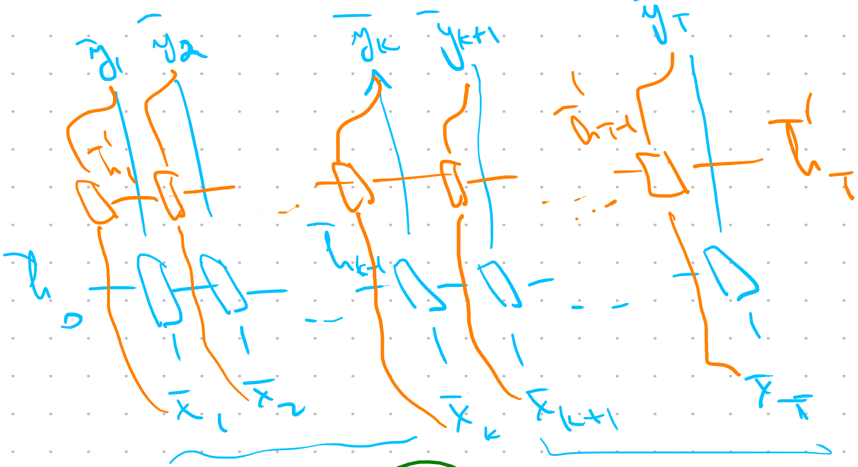


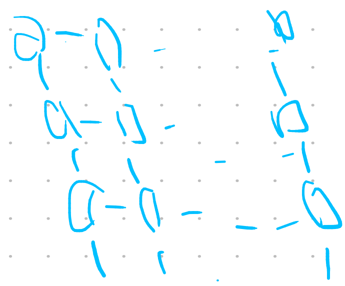
$$(\bar{h}_t, \bar{y}_t) = F(\bar{x}_t, \bar{h}_{t-1}; \bar{w})$$

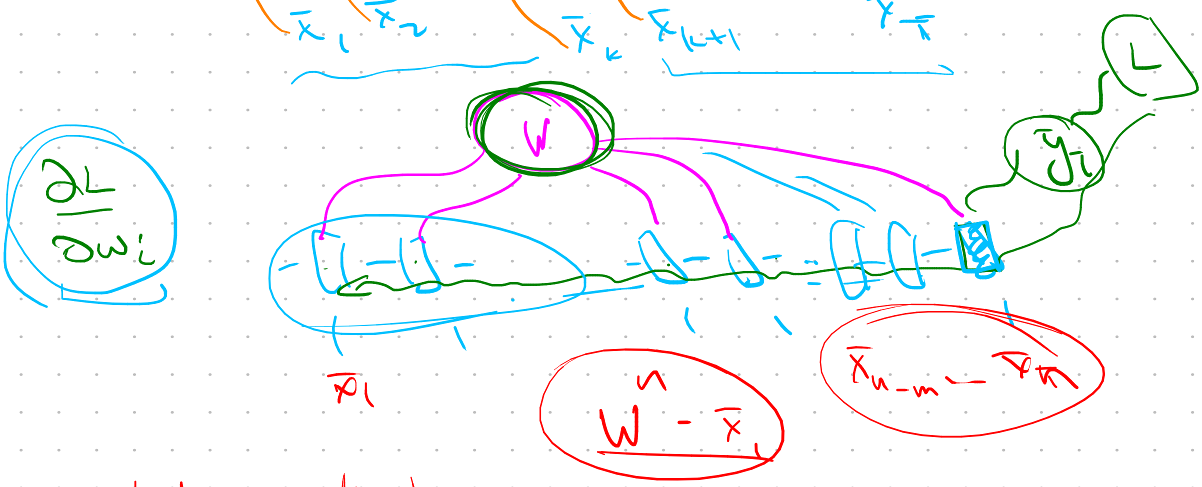
autoregressive

left context

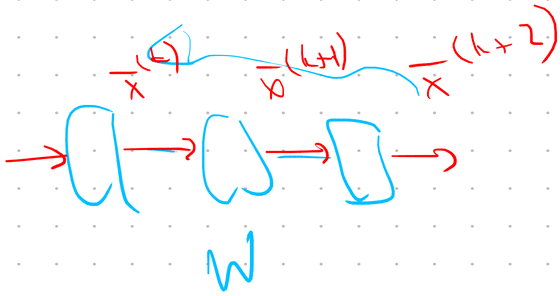
Bidirectional RNN





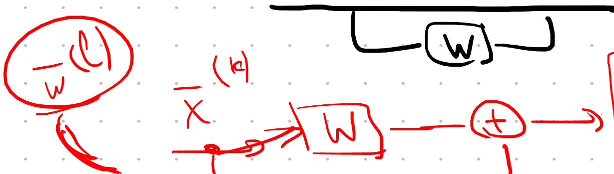


- exploding gradients
- vanishing gradients



$$\bar{x}^{(k+1)} = W\bar{x}^{(k)}$$

$\boxed{W}$



$$\boxed{\bar{x}^{(k+1)} = F(\bar{x}^{(k)}) + \bar{x}^{(k)}}$$

L

$F \sim \bar{x}^{(k)} \mapsto \bar{x}^{(k+1)} - \bar{x}^{(k)}$

- residual connection

$$\frac{\partial L}{\partial \bar w^{(\ell)}} = \frac{\partial L}{\partial \bar x^{(k+1)}} \cdot \frac{\partial \bar x^{(k+1)}}{\partial \bar w^{(\ell)}} =$$

$$= \frac{\partial L}{\partial \bar x^{(k+1)}} \cdot \boxed{\frac{\partial \bar x^{(k+1)}}{\partial \bar x^{(k)}}} \cdot \frac{\partial \bar x^{(k)}}{\partial \bar w^{(\ell)}} \quad \rightarrow 0$$

Constant error carousel

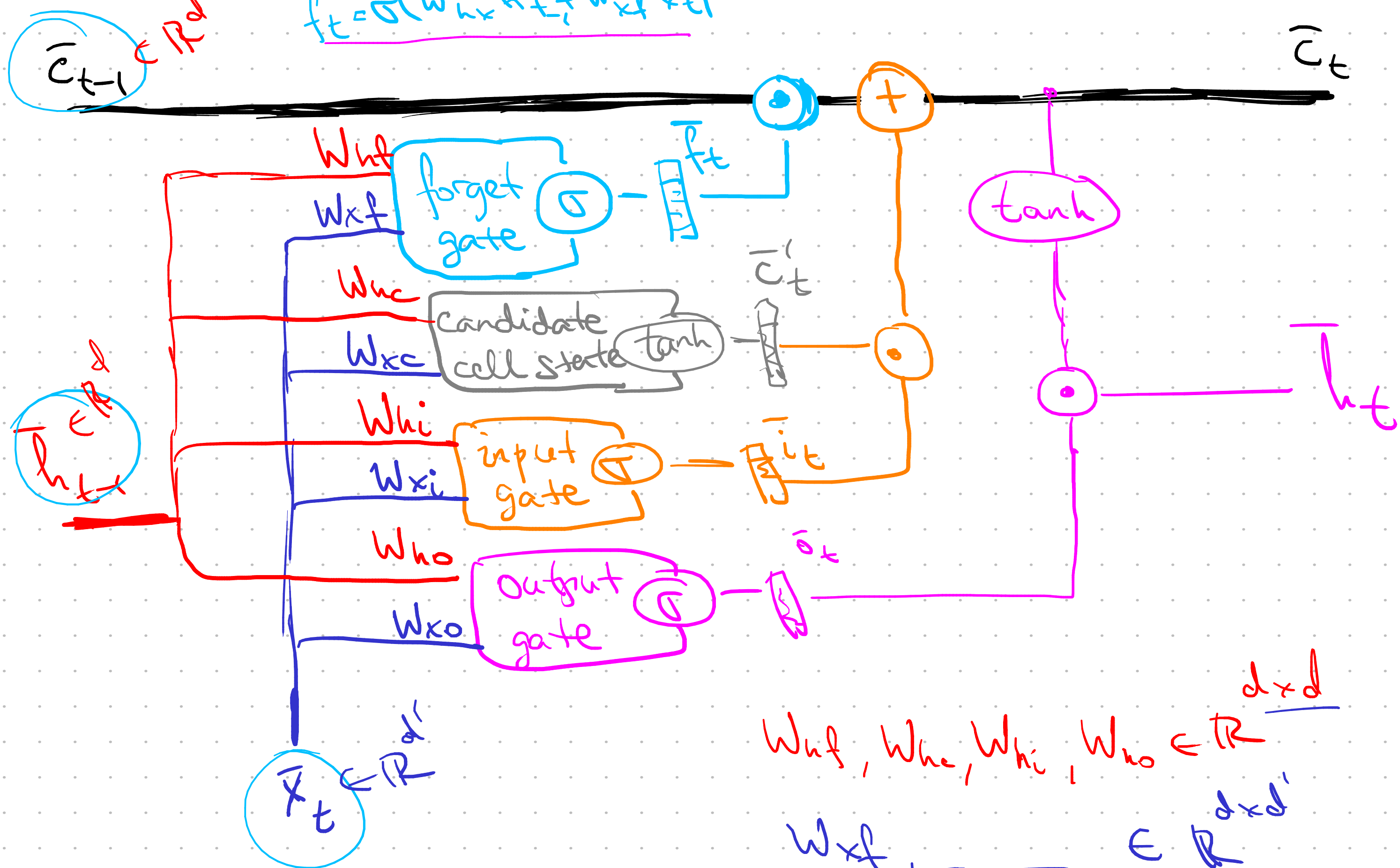
Hochreiter Schmidhuber 1995/2000

$$— '' — \cdot \left(\frac{\partial \bar f}{\partial \bar x} + I \right) \cdot — '' —$$

$\frac{\partial \bar f}{\partial \bar x} \rightarrow 0$

LSTM (long short-term memory)

$\bar f_t = \sigma(W_{hf} \bar h_{t-1} + W_{xf} \bar x_t)$



$W_{hf}, W_{hc}, W_{hi}, W_{ho} \in \mathbb{R}^{d\times d}$

$W_{xf}, — — \in \mathbb{R}^{d\times d'}$