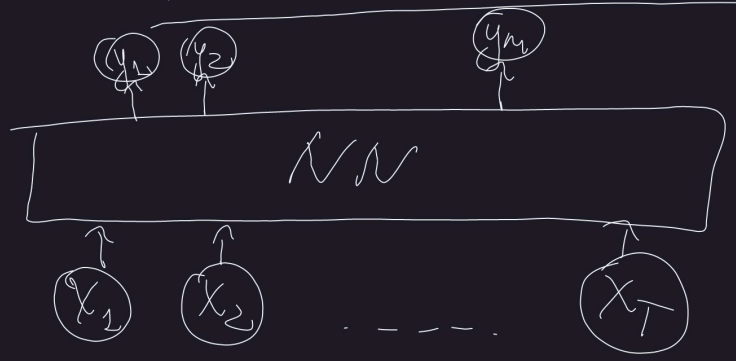
Recurrent Neural Networks and Attention



$$h_t = g(W_x x_t + W_h h_{t-1} + b)$$

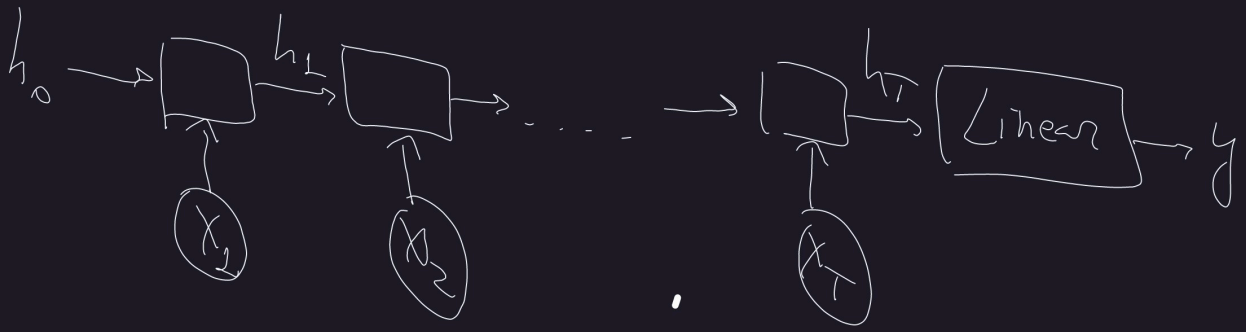
$x_t \in R^{\text{input}}$

$h_t \in R^{\text{hidden}}$

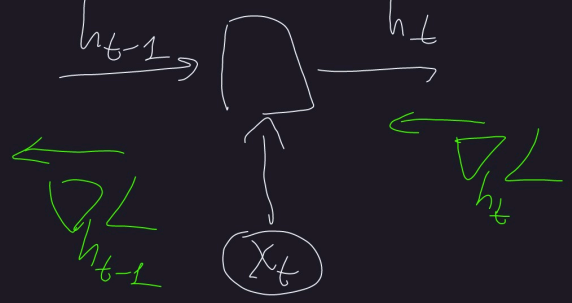
$W_x \in R^{\text{hidden} \times \text{input}}$

$W_h \in R^{\text{hidden} \times \text{hidden}}$

$b \in R^{\text{hidden}}$



RNN



$$\nabla_{h_{t-1}} L = W_h^T \left(\frac{\partial g}{\partial z_t}\right)^T \nabla_{h_t} L = W_h^T \frac{\partial g}{\partial z_t} \cdot W_h^T \frac{\partial g}{\partial z_{t-1}} \nabla_{h_{t-1}} L =$$

$$= \left(\prod_{i=t}^{1} W_h^T \cdot \frac{\partial g}{\partial z_i} \right) \cdot \nabla_{h_T} L$$

$$dL = \nabla_{h_t} L^T dh_t = \nabla_{h_t} L^T \frac{\partial g}{\partial z_t} \cdot W_h \cdot dh_{t-1}$$

$$h_t = g(\underbrace{W_x x_t + W_h \cdot h_{t-1} + b}_{z_t})$$

$$\| \nabla_{h_{t-1}} L \| \leq \prod_{i=t}^{1} \| W_h \| \cdot \left\| \frac{\partial g}{\partial z_i} \right\| \cdot \| \nabla_{h_T} L \|$$

exploding gradients: gradient clipping

$$clip_\varepsilon(g) = \begin{cases} \frac{g}{\|g\|} \cdot \varepsilon, & \text{if } \|g\| > \varepsilon \\ g, & \text{otherwise} \end{cases}$$

vanishing gradients: Relu activ. +

orth. W_h : - init.

- regul. $+\lambda \|W_h^T W_h - I\|_F^2$
- reparam. $\exp_m(V - V^T)$

$$W^T W = I$$

LSTM

$$\frac{\partial c_t}{\partial c_{t-1}} = I$$

Long Short Term Memory Network

c_t – long term memory

h_t – short-term memory

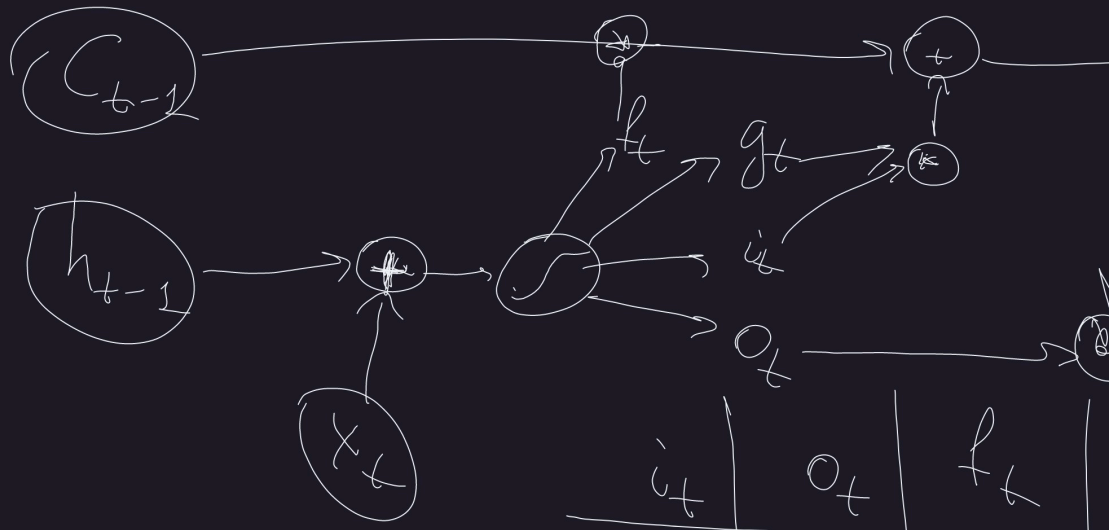
$$c_t = f_t \odot c_{t-1} + g_t \odot i_t$$

$$h_t = o_t \odot \tanh(c_t)$$

$$g_t = g\left(W_x^g x_t + W_h^g \cdot h_{t-1} + b^g\right) \qquad f_t = \sigma(\dots)$$

$$i_t = \sigma\left(W_x^i x_t + W_h^i \cdot h_{t-1} + b^i\right) \in (0,1) \qquad \text{Init: } b^f >> 1$$

$$o_t = \sigma\left(W_x^o \cdot x_t + W_h^o \cdot h_{t-1} + b^o\right)$$



i_t	o_t	f_t	Behavior of LSTM cell
			RNN
+	+	−	Storing inform. to memory
+	−	+	Loading inform.
−	+	+	preserving inf.
−	−	−	erasing inf.

GRU

gated recurrent unit

$$z_t = \sigma(W_x^z \cdot x_t + W_h^z \cdot h_{t-1} + b^z)$$

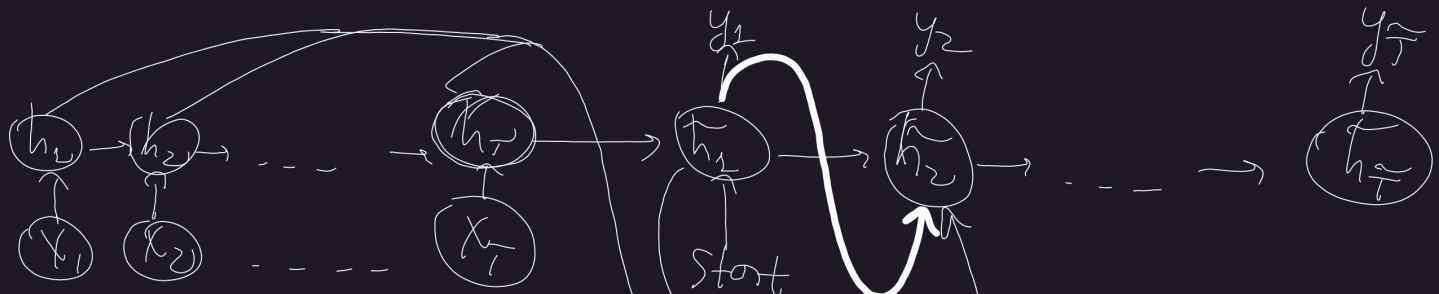
$$u_t = \sigma(\text{----})$$

$$g_t = g(W_x^g \cdot x_t + W_h^g \cdot (h_{t-1} \circ z_t) + b^g)$$

$$h_t = (1 - u_t) \circ g_t + u_t \circ h_{t-1}$$

Init: $b^u >> 1$

abaeebcd ⟶ abaeebcdabaeebcd



$$score(\tilde{h}_{t-1}, h_i) \in \mathbb{R}$$

$$Attention(\tilde{h}_{t-1}, h_1, \dots, h_T)$$

$$\alpha_i = \frac{\exp(score(\tilde{h}_{t-1}, h_i))}{\sum_{j=1}^{T} \exp(score(\tilde{h}_{t-1}, h_j))}, \quad i = 1, \dots, T$$

$$Att = \sum_{i=1}^{T} \alpha_i h_i$$

score(x, y): ① $x^T y$ ② $\frac{x^T y}{\sqrt{dim(x)}}$ ③ $w^T \tanh(W_x \cdot x + W_y \cdot y) + b$

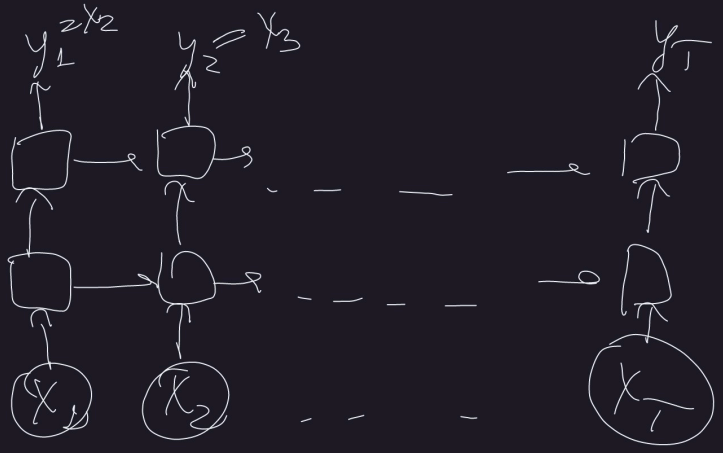
$$x^T y = \sum_{i=1}^{d} x_i y_i$$

$x_i, y_i \sim \mathcal{N}(0, 1)$

$$Var(x^T y) = Var \sum_i x_i y_i = \sum_{i=1}^{d} Var(x_i) \cdot Var(y_i) = 1 \cdot 1 = d$$

$$\mathbb{E}\, x^T y = \mathbb{E} \sum_i x_i y_i = \sum_i \mathbb{E} x_i \, \mathbb{E} y_i = 0 \cdot 0$$

$$Var\left(\frac{x^T y}{\sqrt{dim(x)}}\right) = 1$$


$y_1 = x_2$
$y_2 = x_3$
y_{T-1}
x_1
x_2
x_{T-1}