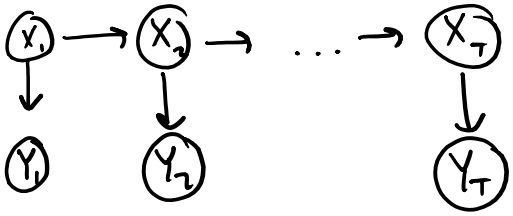
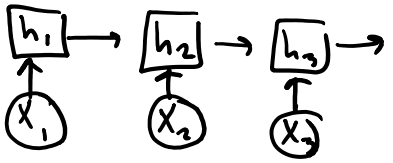
Recurrent Neural Nets

Wednesday, May 29, 2019 12:47 PM



DAG for the HMM

node: random variable



computation graph

random variables are inputs ○

hidden units / other tensors □

Ex. $X_t \in \{0,1\}^p$ and it is a 1-hot encoding of word in a dictionary (p elements)

$$a_t = b + W h_{t-1} + U X_t$$

$$h_t = \tanh(a_t)$$

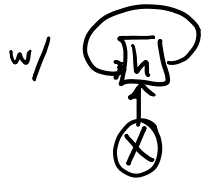
$a_t, h_t \in \mathbb{R}^d$

U $d \times p$

W $d \times d$

$b \in \mathbb{R}^d$

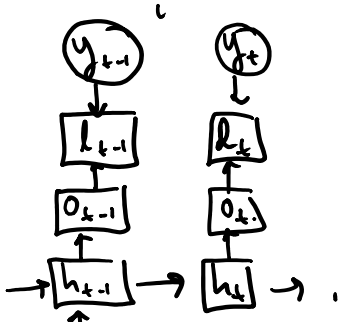
RNN can be written w/



y_t is some response at time t (word in another language)

$$o_t = c + V h_t$$

$$\hat{y}_t = \frac{e^{o_t}}{\sum_i e^{o_{t,i}}} \quad \text{(softmax)}$$



Neural machine translation v1

observe X_t English word

predict $\hat{y}_t$

observe y_t French word

Back Propagation Through Time (BPTT) $L = \sum_{t=1}^{T} l_t$

$$\frac{\partial L}{\partial o_{t,i}} = \frac{\partial}{\partial o_{t,i}} l(o_t, y_t) = \hat{y}_{t,i} - y_{t,i}$$

$$\frac{\partial L}{\partial h_T} = V^T \frac{\partial l_T}{\partial o_T} \qquad V^T$$

$$\frac{\partial L}{\partial h_t} = \left(\frac{\partial h_{t+1}}{\partial h_t}\right)^T \frac{\partial L}{\partial h_{t+1}} + \left(\frac{\partial o_t}{\partial h_t}\right)^T \frac{\partial L}{\partial o_t} \quad (\star)$$

$$\frac{\partial}{\partial h_t}\left(\tanh\left(b + W h_t + \dots\right)\right) = W^T \operatorname{diag}\left(1 - h_{t+1}^2\right)$$

$$(\star) = W^T \operatorname{diag}\left(1 - h_{t+1}^2\right) \frac{\partial L}{\partial h_{t+1}} + V^T \frac{\partial L}{\partial o_t}$$

BPTT: recursively calculate $\frac{\partial L}{\partial h_t}$

RNN Architectures

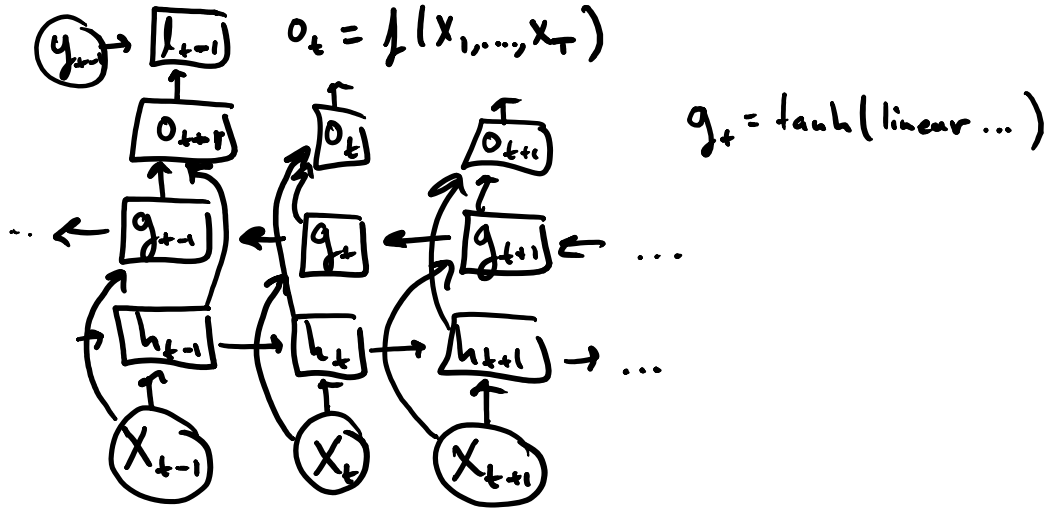
Wednesday, May 29, 2019 12:48 PM

Bidirectional RNN

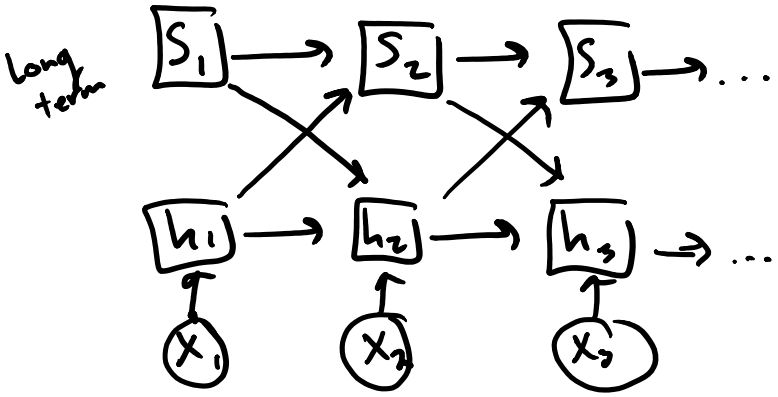
Observe $X_1, \ldots, X_T$ and predict $y_1, \ldots, y_T$

$o_t = f(X_1, \ldots, X_T)$

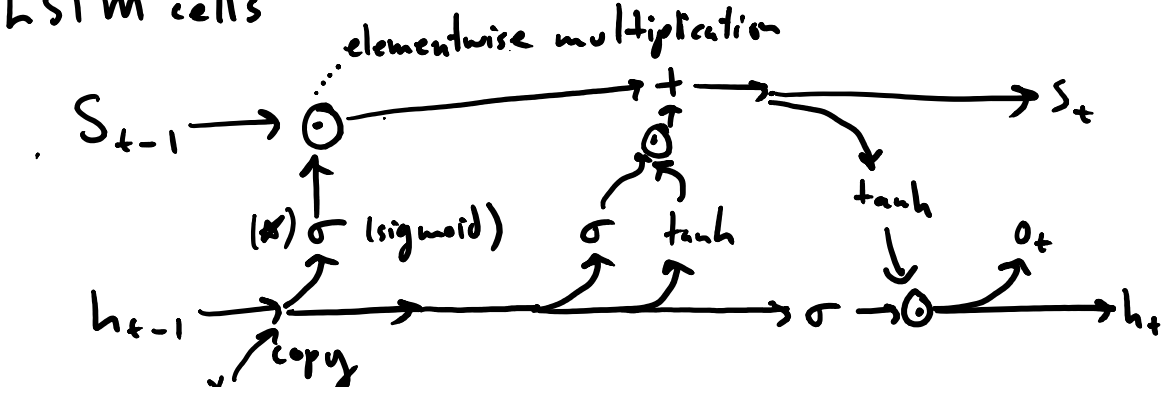
$g_t = \tanh(\text{linear} \ldots)$



Long - Short term memory (LSTM)



LSTM cells



(*) $a_t = b_a + W_a h_{t-1} + U_a x_t$

$\beta_t = \sigma(a_t)$

$\beta_{t,i} \cdot s_{t-1,i}$ "forget" gate