Reagan Kan | rkan3 | 903 404 746

CS 4803 | PS 3

1 Parity, RNN

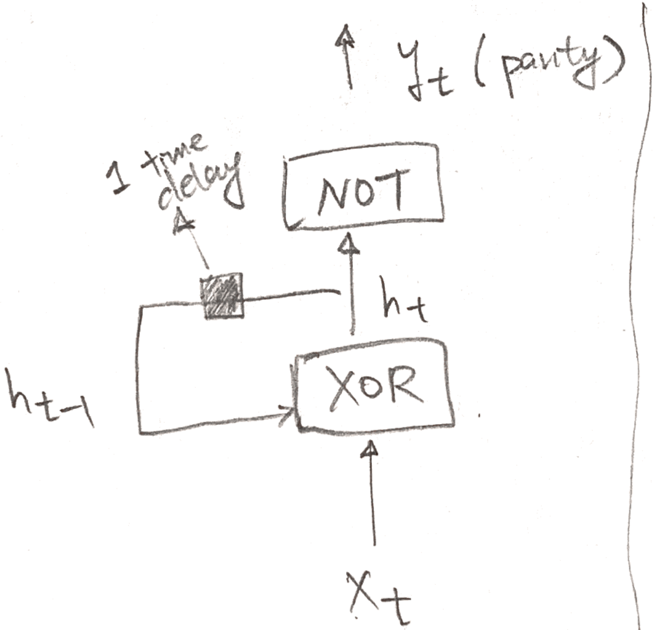
consider given example, $X = 010110$

parity offset by 1	X	parity
h_0	0	1
1	1	0
0	0	0
0	1	1
1	1	0
0	0	0

this is $XOR(C_{t-1}, X_t)$

if we set $h_0 = 0$, we need to use NOT before outputting.

So RNN layer has this structure



y_t	1	0	0	1	0	0
X_t	0	1	0	1	1	0
h_{t-1}	0	0	1	1	0	1
h_t	0	1	1	0	1	1

2| Parity. LSTM

want to find W_f b_f, W_i, b_i, W_c, b_c, W_o, b_o

s.t. C_t = parity.

if $h_0 = 0$, then C_t = parity = $XOR(h_{t-1}, X_t)$

$$= (h_{t-1} \wedge \bar{X}_t) \vee (\overline{h_{t-1}} \wedge X_t) \quad \text{(from hint)}$$

$$= h_{t-1} \circ \bar{X}_t + h_{t-1} \cdot X_t$$

$$= C_{t-1} \circ f_t + i_t \bullet \tilde{C}_t \quad \text{(eq 4)}$$

Let's set

$$\underset{(a)}{f_t = \overline{X_t}}, \quad \underset{(b)}{i_t = \overline{h_{t-1}}}, \quad \underset{(c)}{\tilde{C}_t = X_t}, \quad \underset{(d)}{C_{t-1} = h_{t-1}}$$

(a) $f_t = \bar{X}_t = \sigma(W_f[h_{t-1}, X_t]) = \begin{cases} 1 & \text{if } W_f[h_{t-1}, 0] > 0 \\ 0 & \text{if } W_f[h_{t-1}, 1] \le 0. \end{cases}$

$\boxed{\text{Letting } W_f = \begin{bmatrix} 0 \\ -2 \end{bmatrix} \text{ and } b_f = 1}$

will satisfy the piecewise function

$f_t = \bar{X}_t = 1$ if $0(h_{t-1}) - 2(0) + 1 = (1 > 0) \equiv T$

$= \bar{X}_t = 0$ if $0(h_{t-1}) - 2(1) + 1 = (-1 \le 0) \equiv T$

(b) $i_t = \overline{h_{t-1}} = \begin{cases} 1 & \text{if } W_i [\underset{}{\overset{0}{h_{t-1}}}, x_t] > 0 \\ 0 & \text{if } W_i [\underset{1}{h_{t-1}}, x_t] \le 0 \end{cases}$

to satisfy above equation,

set $\boxed{W_i = \begin{bmatrix} -2 \\ 0 \end{bmatrix} \quad b_i = 1}$

$i_t = 1$ if $(-2(0) + 0 + 1 > 0) \equiv (1 > 0) \equiv T$

$= 0$ if $(-2(1) + 0 + 1 \le 0) \equiv (-1 \le 0) \equiv T$

(c) $\tilde{c}_t = x_t = \begin{cases} 1 & \text{if } w_c [h_{t-1}, \overset{1}{x_t}] > 0 \\ 0 & \text{if } w_c [h_{t-1}, \underset{0}{x_t}] = 0 \end{cases}$

to satisfy this, set $\boxed{W_c = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad b_c = 0}$

$\tilde{c}_t = 1$ if $(0(h_{t-1}) + (1)(1) > 0) \equiv (1 > 0) \equiv T$

$= 0$ if $(0(h_{t-1}) + (0)(1) = 0) \equiv (0 = 0) \equiv T$

(d) from eq 6, $h_t = o_t * \tanh(c_t)$

if $o_t = 1$ then $h_t = \tanh(c_t) = \begin{cases} 1 & \text{if } c_t = 1 \\ 0 & \text{if } c_t = 0 \end{cases}$

$= c_t$

(d continued).

if $o_t = 1$, then $1 = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$

$$= \begin{cases} 1 & \text{if } W_o [h_{t-1}, x_t] + b_o > 0 \end{cases}$$

Setting $\boxed{W_o = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad b_o = 1}$ ensures

$o_t = 1$, hence ensures $\underline{h_t = c_t}$.

So, in order for (a) (b) (c) (d) to be true,

we have:

$$W_f = \begin{bmatrix} 0 \\ -2 \end{bmatrix}, \quad b_f = 1$$

$$W_i = \begin{bmatrix} -2 \\ 0 \end{bmatrix}, \quad b_i = 1$$

$$W_c = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad b_c = 0$$

$$W_o = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad b_o = 1$$

3] Beam Search.

Want to Show: $\forall i$, if $best_{\leq i}$ exists with score S

and S is better than score for highest scoring item in B_i

then $\neg \exists\ y''$, with $|y''| > |best_{\leq i}|$

s.t. y'' has better score than S.

$\forall i$, suppose $y = best_{\leq i}$, with score S.

and y_{B_i} be the highest scoring item in B_i

with score S_{B_i} s.t. $S_{B_i} \leq S$.

Let's try to grow y_{B_i} in-to y'', with j additional Beam Search steps.

Case 1 $\quad comp(y_{B_i}) \equiv T$, we cannot do additional steps

so y'' does not exist.

Case 2 $\quad comp(y_{B_i}) \equiv F$, We can grow w/ j more steps.

So, $y'' = y_{B_i} \circ y_{i+1} \circ y_{i+2} \circ \cdots \circ y_{i+j}$

Consider the score for y''. Call it S''

$$S'' = S_{B_i} \circ S_{i+1} \circ \cdots \circ S_{i+j}$$

$$= S_{B_i} \circ p(y_{i+1} \mid X, y_{B_i}) \circ p(y_{i+2} \mid X, y_{B_i} \circ y_{i+1})$$

$$\circ \cdots \circ p(y_{i+j} \mid X, y_{B_i} \circ y_{i+1} \circ \cdots \circ y_{i+j-1}).$$

Note: the maximum possible values for $S_{i+1} \cdots S_{i+j}$ is 1, since they are probabilities.

thus. the max value for $S'' = S_{B_i} \circ 1^j = S_{B_i}$

$$\leq S_{B_i}$$

So, y'' does <u>not</u> have score higher than S_{B_i}.

4| Exploding Gradients.

$$h_t = W^T h_{t-1} = W^T W^T h_{t-2} = \cdots = (W^T)^t h_0$$

W can be decomposed, i.e.

$W = Q \Lambda Q^{-1}$, when Λ is a diagonal matrix with elements $\lambda_1 \cdots \lambda_n$. $\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}$.

let $p = t$

$$h_p = (W^T)^p h_0 \quad \text{with} \quad p >> 0$$

$$= \left((Q \Lambda Q^{-1})^T \right)^p h_0 = \left((Q^{-1})^T \Lambda^T Q^T \right)^p h_0$$

$$= \left((Q^T)^{-1} \Lambda^T Q^T \right)^p h_0$$

$$= \left((Q^T)^{-1} \Lambda^T \underbrace{Q^T (Q^T)^{-1}}_{= I} \Lambda^T \cdots \underbrace{(Q^T)^{-1}}_{= I} \Lambda Q^T \right)$$

$$= (Q^T)^{-1} \Lambda^P Q^T = h_p$$

Consider $\Lambda^P = \begin{bmatrix} \lambda_1^P & & 0 \\ & \ddots & \\ 0 & & \lambda_n^P \end{bmatrix}$

if $\rho(W) > 1$, then $(\rho(W))^P$ explodes with $P >> 0$

if $\rho(W) < 1$, then $(\rho(W))^P$ vanishes with $P >> 0$

Λ^P contains $(\rho(W))^P$, so Λ^P explodes/vanish accordingly.

as does $h_P = (Q^T)^{-1} \Lambda^P Q^T$