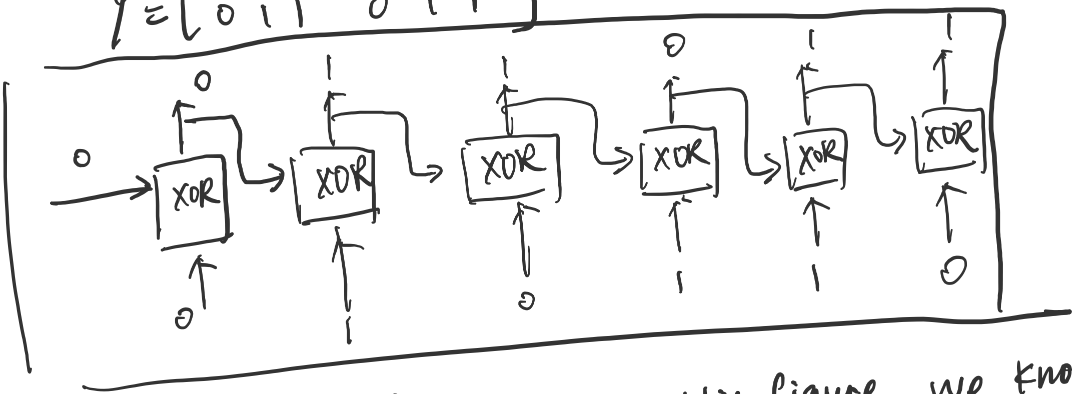
Q1

$x = [0\ 1\ 0\ 1\ 1\ 0]$

$y = [0\ 1\ 1\ 0\ 1\ 1]$



y_{t-1}	x_t	y_t
0	0	0
0	1	1
1	0	1
1	1	0

From this figure, we know it is a XOR operator

$$\begin{cases} y_t = y_{t-1}\overline{x_t} + \overline{y_{t-1}}\,x_t \\ y_0 = 0 \end{cases}$$

2.

$$\begin{cases} y_t = h_t \\ h_t = h_{t-1}(1-x_t) + (1-h_{t-1})x_t \\ \quad = h_{t-1} - h_{t-1}x_t + x_t - x_t h_{t-1} \\ \quad = h_{t-1} + x_t - 2h_{t-1}x_t \\ h_0 = 0. \end{cases}$$

Q2 LSTM

(4): $C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$

we assume $C_t = h_t$ $\tilde{C}_t = \overline{h_{t-1}}$, $f_t = \overline{x_t}$, $i_t = x_t$.

so (4): $h_t = \overline{x_t} \cdot h_{t-1} + x_t \cdot \overline{h_{t-1}}$

so (2): $i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) = x_t$

$\therefore W_i = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ $b_i = 0$

(1): $f_t = \sigma(W_f[h_{t-1}\ x_t] + b_f) = \overline{x_t} = 1 - x_t$

$\therefore W_f = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$ $b_f = 1$

(3): $\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) = 1 - h_{t-1}$

$\therefore W_c = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$ $b_c = 1$

(5): $O_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$

we make $O_t \equiv 1$.

so $W_o = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. $b_o = 1$

(6): $h_t = O_t * \tanh(C_t) = \tan(h_t) \Rightarrow h_t = \tan(h_t)$

that make sense.

so the assumption we make is right.

in all.

$W_f = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$ $W_i = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ $W_c = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$ $W_o = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$b_f = 1$ $b_i = 0$ $b_c = 1$ $b_o = 1$

Problem 3.

$\log p(y_t \mid x, y_{<t}) \leq 0.$

$\therefore \quad \max\{y \in B_{t+1}\} \leq \max\{y \in B_t\}$

hence if $\max\{y \in B_t\} \leq best_{<t}$, $\max\{y \in B_{t+1}\} \leq best_{<t}$

As a result, future steps will be no better and current highest scoring beam is the overall highest probability completed beam.

Q4.

$h_T = (W^T)^t h_0$ (1)

$W \in C^{n\times n}$, so $W = V \Lambda V^T$, V is eigenvectors,
Λ is eigenvalues

$h_T = \left[(V\Lambda V^T)_1 (V\Lambda V^T)_2 \cdots\cdots (V\Lambda V^T)_t \right] h_0$

since $V^T V = E$

$h_T = [V \Lambda^t V^T] h_0$

$\frac{d h_T}{d h_0} = V \Lambda^t V^T$

if $t \gg 0$, if $\rho(W) < 1$, then $\Lambda^t \to 0$, resulting in vanishing gradient

if $\rho(W) > 1$ $\Lambda^t \to \infty$. resulting in exploding gradient.

Q5 (a) $Agg(H'_{it}) = \sum_{j=1}^{\#N(V_i)} f_{ji}(h_j^t)$

$$h_i^{t+1} = g(h_i^t, \sum_{j=1}^{\#N(V_i)} f_{ji}(h_j^t))$$

(b).

$$Agg(H'_{1t}) = [0.6 \quad 0.2 \quad 0.2] \begin{bmatrix} f(h_2^t) \\ f(h_3^t) \\ f(h_4^t) \end{bmatrix} = [0.6 \ 0.2 \ 0.2] \begin{bmatrix} [-2 \ 2] \\ [0, -2] \\ [2, 0] \end{bmatrix}$$

$$= 0.6 [-2 \ 2] + 0.2 [0 \ -2] + 0.2 [2 \ 0]$$

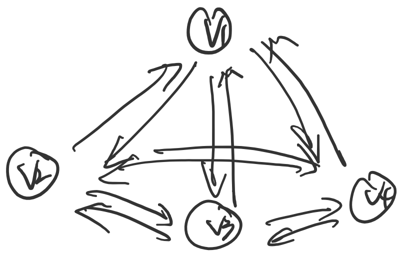
$$= [-0.8 \quad 0.8]$$

$$h_1^{t+1} = g([1, -1], [-0.4 \ 0.4]) = W(h_1^t)^T + \max([-0.8 \ 0.8], 0)$$

$$= [1 \ 1] \begin{bmatrix} 1 \\ -1 \end{bmatrix} + [0, 0.8] = [0 \quad 0.8]$$

(C)

I think every token has connection to each other. so it would be like



so it has 12 edges

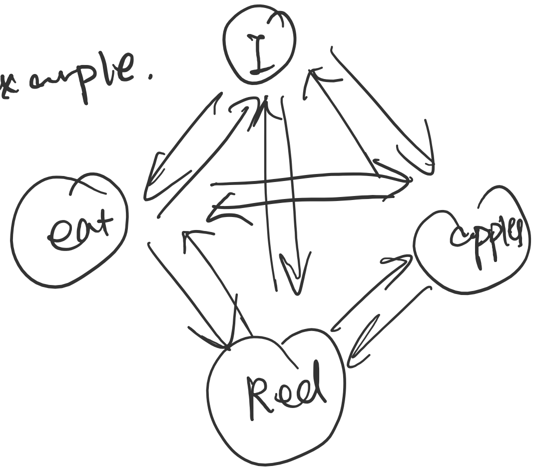
(d) $Agg(H_{it}^{l}) = \text{softmax}_j (Q^l h_i^l \cdot K^l h_j^l)$

$$h_i^{l+1} = g(h_i^{t}, Agg(H_{it}^{l})) = \sum_{j \in S} (V^l h_j^l \cdot Agg(H_{it}^{l}))$$

From above, we see equation 15 has the same format as (17). So the Transformer model's single-head attention mechanism is equivalent to a special case of a GNN.

From (C) For GNN to represent a sentence, it is fully connected. AND the fully connected graph doesn't show any information about sequence order.

For example.



From this figure, we don't know if the original sequence is "I eat Red apple", or "Red apple I eat".