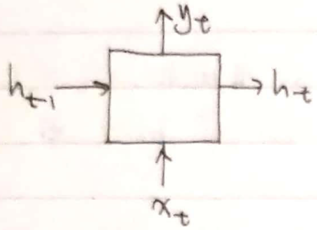
CS7643: Deep Learning
Fall 2020
HW4 Solutions

Ruize Yang

October 22, 2020

Problem 1

1.



x_t is input, y_t is output.

make $y_t = h_t$, so the input hidden state would be indicator of # 1's at last step, and output hidden state would be the indicator of # 1's at current step, same as y_t.

Relationship of x_t, h_{t-1}, $h_t = y_t$:

h_{t-1}	x_t	h_t / y_t
1	1	0
1	0	1
0	1	1
0	0	0.

So $h_t = h_{t-1}$ XOR x_t. If use OR 、AND:

$h_t = (h_{t-1}$ OR $x_t) - (h_{t-1}$ AND $x_t)$.

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

h_{t-1}	x_t	$\min(h_{t-1} + x_t, 1)$ (OR)	$\max(h_{t-1} + x_t - 1, 0)$ (AND).	h_t = OR − AND $= y_t$.
1	1	1	1	0
1	0	1	0	1
0	1	1	0	1
0	0	0	0	0

Problem 2

2. C_t is the parity of string, but x_t is computed with h_{t-1},
So make $h_t = C_t$.

$h_t = O_t \cdot \tanh(C_t)$, $\tanh(C_t) = C_t$, so make $O_t = 1$.

$O_t = \sigma(W_o [h_{t-1}\ x_t]^T + b_o)$, let $W_o = [0, 0]$, $b_o = 1$,

$\Rightarrow O_t = \sigma(b_o) = 1$.

$C_t = x_t \text{ XOR } C_{t-1} = (x_t \wedge \bar{C}_{t-1}) \vee (\bar{x}_t \wedge C_{t-1})$, $C_{t-1} = h_{t-1}$

Make $f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t = (x_t \wedge \bar{C}_{t-1}) \vee (\bar{x}_t \wedge C_{t-1})$:

there is a C_{t-1} at both side, so $f_t = \bar{x}_t$.

$f_t = \sigma(W_f [h_{t-1}\ x_t]^T + b_f)$, let $W_f = [0, -1]$, $b_f = 1$.

$f_t = \sigma(1 - x_t) = \bar{x}_t$

Make one of i_t and $\tilde{C}_t$ x, the other $\bar{C}_{t-1}$:

$i_t = \sigma(W_i [h_{t-1}\ x_t]^T + b_i)$, let $W_i = [0, 1]$, $b_i = 0$.

$i_t = \sigma(x_t) = x_t$

$\tilde{C}_t = \tanh(W_c [h_{t-1}\ x_t]^T + b_c)$, let $W_c = [-1, 0]$, $b_c = 1$.

$\tilde{C}_t = \tanh(1 - h_{t-1}) = \bar{h}_{t-1} = \bar{C}_{t-1}$

$\Rightarrow W_f = [0, -1]$, $b_f = 1$, $W_i = [0, 1]$, $b_i = 0$, $W_c = [-1, 0]$, $b_c = 1$.
$W_o = [0, 0]$, $b_o = 1$.

Problem 3

3. At time t, $B_t = (\langle y^1, s^1 \rangle \cdots \langle y^b, s^b \rangle)$, $best_{\le t} = \langle y^1, s^1 \rangle$.

For any $\langle y^i, s^i \rangle \in B_t$, $s^i \le s^1$. $s^i = \sum_{j=1}^{t} \log p(y^i_j \mid x, y^i_{<j})$.

At next step, $s^i_{t+1} = \sum_{j=1}^{t+1} \log p(y^i_j \mid x, y^i_{<j})$

$$= s^i + \log p(y^i_{t+1} \mid x, y^i_{<t+1}).$$

As p is a probability, $p \in [0, 1]$, $\log p \le 0$.

So $s^i_{t+1} \le s^i \le s^1$, any ~~futher~~ further extension of y^i in B_t will not give a score higher than s^1, $\langle y^1, s^1 \rangle$ is the highest-probability one.

Problem 4

4. Assume W has n linearly independent eigenvectors.

The eigendecomposition of W is $W = Q\Lambda Q^{-1}$, $W^T = (Q\Lambda Q^{-1})^T$.

$(Q^{-1})^T = (Q^T)^{-1}$, $\Lambda^T = \Lambda$, $\Rightarrow W^T = (Q^T)^{-1} \Lambda Q^T$.

$$h_t = W^T h_{t-1} = (W^T)^t h_0 = (Q^T)^{-1} \Lambda Q^T \cdot \dots \cdot (Q^T)^{-1} \Lambda Q^T h_0$$
$$= (Q^T)^{-1} \Lambda^t Q^T h_0.$$

Λ is a diagonal matrix of eigenvalues $\lambda_1 \dots \lambda_n$ of W.

If $\rho(W) < 1$, then when $t \to \infty$, $\rho(W)^t \to 0$.

then ~~every~~ entry on the diagonal of $\Lambda^t \to 0$,

then $\frac{\partial h_t}{\partial h_0} = (Q^T)^{-1} \Lambda^t Q^T \to 0$, the gradient vanish.

If $\rho(W) > 1$, then when $t \to \infty$, $\rho(W)^t \to \infty$.

at least one entry on the diagonal of $\Lambda^t \to \infty$.. let $\rho(W) = \lambda_i$

then $\frac{\partial h_t}{\partial h_0} = (Q^T)^{-1} \Lambda^t Q^T$, every entry is λ_i^t multiply with some number, so every entry of $\frac{\partial h_t}{\partial h_0} \to \infty$. the gradient explode.

Problem 5a

5. (a). $h_i^{t+1} = q\left(h_i^t, \sum_{v_j \in N(v_i)} f_{ij}(h_j^t)\right)$.

Problem 5b

(b) $Agg(H'_{1-t}) = [0.6 \;\; 0.2 \;\; 0.2] \begin{bmatrix} f([-1,1]) \\ f([0,-1]) \\ f([1,0]) \end{bmatrix}$

$$= [0.6,\, 0.2,\, 0.2] \begin{bmatrix} -2 & 2 \\ 0 & -2 \\ 2 & 0 \end{bmatrix} = [-0.8 \;\; 0.8].$$

$$h_1^{t+1} = W(h_1^t)^T + \max\{Agg(H'_{1-0}),\, 0\}$$

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

$$= [0,\, 0.8].$$

Problem 5c

(c) 24.

Problem 5d

(d) Use attention as the aggregation, which is a weighted sum of all elements in H_i^t.

$$h_i^{t+1} = q(h_i^t, Agg(\{f_{ij}(h_j^t)\}))$$

$$= \sum_j \text{softmax}(Q^t h_i^t, K^t h_j^t) \cdot V^t h_j^t$$

$f_{ij}(h_j^t) = V^t h_j^t$, then summed over their weights $\text{softmax}(Q^t h_i^t; K^t h_j^t)$.

Problem 5e

(e) Transformer is a special case of GNN, ~~then~~ for a graph with n nodes representing a ~~sentence~~ sentence of n words, the edges between words are $\sim n^2$. In very long term dependency, n is large, computation on n^2 dependencies is difficult.

Problem 6

The author develop a multimodal explanation system, for visual question answering task and activity recognition task. The system answers the question in answering model, then gives textual explanation and points out the corresponding area in the image in multimodal explanation model, and is the first model that does both. The model is trained on image with human annotation of descriptions and explanations. Their model outperforms several baselines, and using explanations and both textual and visual evidence increase model performance.

One question is that in training the model on different parts of dataset, the features of those data could influence the model performance, as the metrics is similarity between ground truth and output sentences. For example, a model trained on explanation learns more about the structure of explanation than a model trained on description does, so it is not guaranteed that the better performance is from learning to explain.

Problem 7

The authors can also use accuracy on the tasks to compare the models, instead of using similarity of ground truth and output, so that only the explanation power of models are compared. The the parameters of answering model are frozen on VQA task, they can also be trained jointly.