Homework 2: Theory

Yuxiang Chai

October 13, 2022

1.1 Convolutional Neural Networks

(a) The output dimension is 5×2

(b) The output width will be $\lfloor \frac{W-KD+D+2P-1}{S} \rfloor + 1$.

The output height will be $\lfloor \frac{H-KD+D+2P-1}{S} \rfloor + 1$.

So the output shape will be $F \times (\lfloor \frac{H-KD+D+2P-1}{S} \rfloor + 1) \times (\lfloor \frac{W-KD+D+2P-1}{S} \rfloor + 1)$

(c) i. The output dimension will be $f_W(x) \in \mathbb{R}^{1\times 1\times 3}$,
and

$$f_W(x)[1,1,p] = \sum_{c=1}^{5} \sum_{m=1}^{3} x[c, 2p+m-2] W[1, c, 2+m-2]$$

, where $p = 1,2,3$

ii. $\frac{\partial f_W(x)}{\partial W} \in \mathbb{R}^{3\times 3\times 5}$, where $\frac{\partial f_W(x)}{\partial W}[p,m,c] = x[c, 2p+m-2]$

iii. $\frac{\partial f_W(x)}{\partial x} \in \mathbb{R}^{3\times 7\times 5}$, where

$$\frac{\partial f_W(x)}{\partial x}[p,j,i] = \begin{cases} W[1,i,j] & 2p-1 \le j \le 2p+1 \\ 0 & else \end{cases}$$

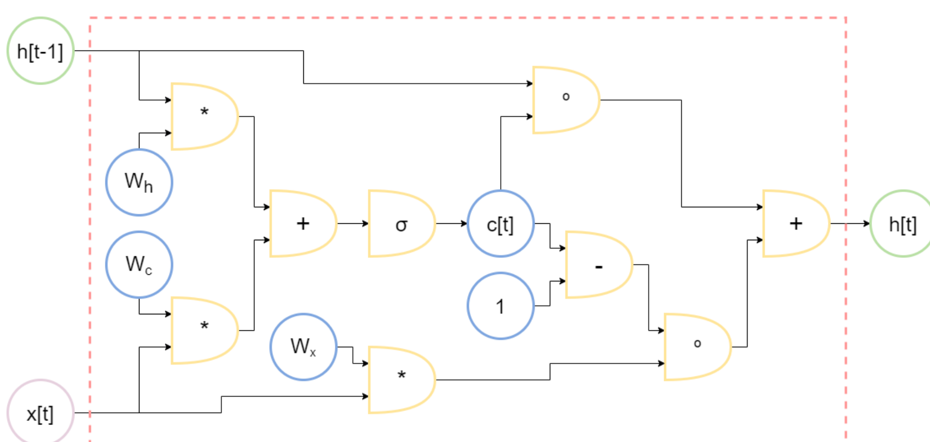
iv. $\frac{\partial \ell}{\partial W} = \frac{\partial \ell}{\partial f_W(x)} \frac{\partial f_W(x)}{\partial W}$, so $\frac{\partial \ell}{\partial W} \in \mathbb{R}^{3\times 3\times 5}$, where $\frac{\partial \ell}{\partial W}[p,m,c] = \frac{\partial \ell}{\partial f_W(x)}[p] \cdot x[c, 2p+m-2]$

The similarity is that they both use x value. And the difference is that this equation doesn't use summation.

1.2 Recurrent Neural Networks

1.2.1 Part 1

(a)



(b) The dimension of $c[t]$ is $\mathbb{R}^m$

(c) $\frac{\partial \ell}{\partial W_x} = \sum_{t=1}^{k} \frac{\partial \ell}{\partial h[t]} \frac{\partial h[t]}{\partial W_x}$, let $h[t] = f(W_x, h[t-1])$, then

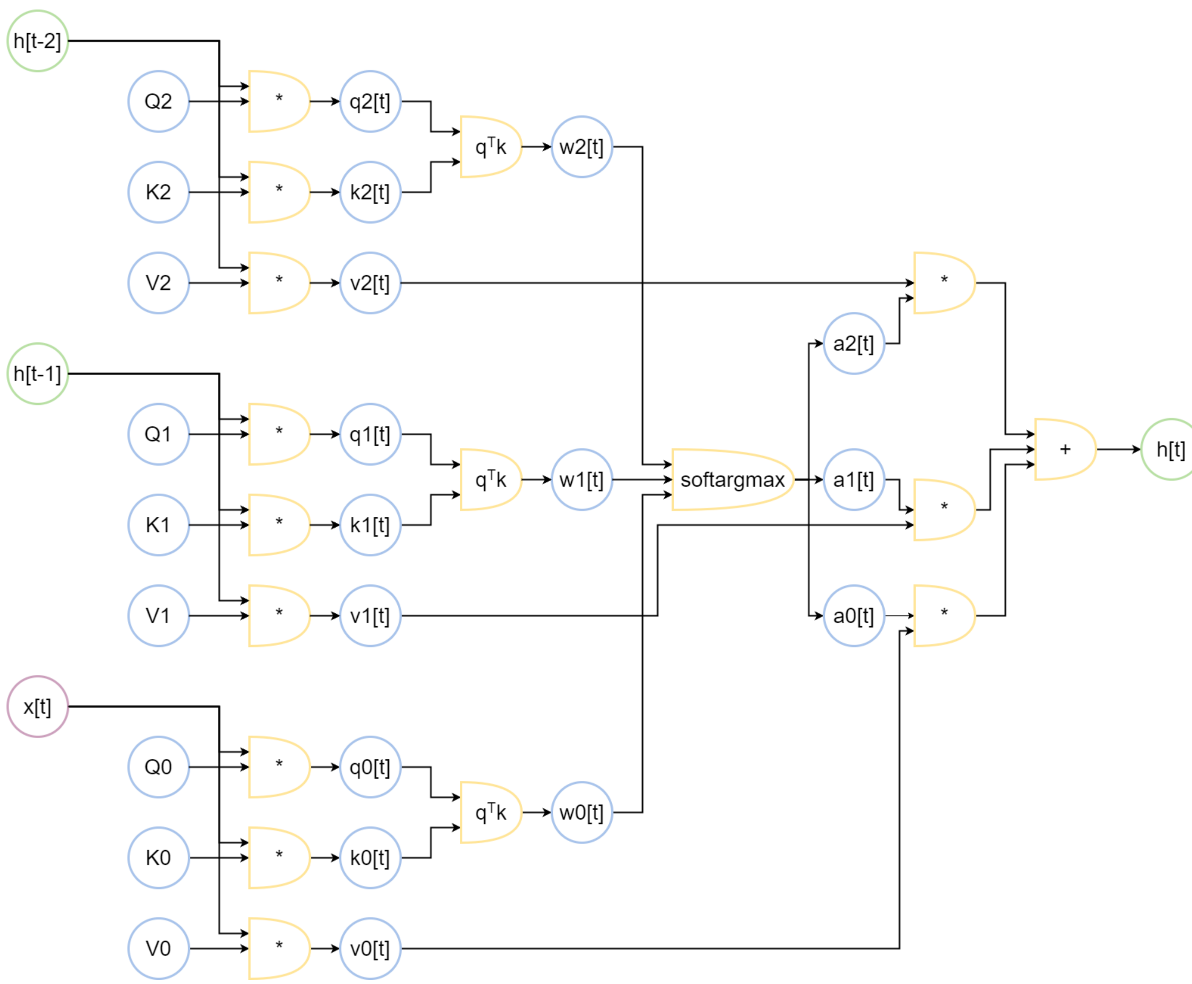
$$\begin{aligned}
\frac{\partial h[t]}{\partial W_x} &= \frac{\partial f(W_x, h[t-1])}{\partial W_x} + \frac{\partial f(W_x, h[t-1])}{\partial h[t-1]} \frac{\partial h[t-1]}{\partial W_x} \\
&= \frac{\partial f(W_x, h[t-1])}{\partial W_x} + \sum_{i=1}^{t-1} \left(\prod_{j=i+1}^{t} \frac{\partial f(W_x, h[j-1])}{\partial h[j-1]} \right) \frac{\partial f(W_x, h[i-1])}{\partial W_x} \\
&= (1 - c[t]) x[t]^\top + \sum_{i=1}^{t-1} \left(\prod_{j=i+1}^{t} \frac{f(W_x, h[j-1])}{\partial h[j-1]} \right) (1 - c[i]) x[i]^\top
\end{aligned}$$

The dimension of $\frac{\partial \ell}{\partial W_x}$ is $\mathbb{R}^{n \times m}$. The similarity is that they both need to use the recurrent way to compute the values.

(d) Yes. Because with sigmoid and multiplication, the derivative will approach 0 after many multiplications, which is the vanishing gradient.

1.2.2 Part 2

(a)



(b) The dimension of $a[t]$ is $\mathbb{R}^3$

(c)

$$
\begin{aligned}
q_0[t], q_1[t], ..., q_k[t] &= Q_0 x[t], Q_1 h[t-1], ..., Q_k h[t-1] \\
k_0[t], k_1[t], ..., k_k[t] &= K_0 x[t], K_1 h[t-1], ..., K_k h[t-1] \\
v_0[t], v_1[t], ..., v_k[t] &= V_0 x[t], V_1 h[t-1], ..., V_k h[t-1] \\
w_i[t] &= q_i[t]^\top k_i[t] \\
a[t] &= \texttt{softargmax}([w_0[t], w_1[t], ..., w_k[t]]) \\
h[t] &= \sum_{i=0}^{k} a_i[t] v_i[t]
\end{aligned}
$$

(d) Assume that t starts from 1.

$$
\begin{aligned}
q_0[t], q_1[t], ..., q_{t-1}[t] &= Q_0 x[t], Q h[t-1], Q h[t-2], ..., Q h[1] \\
k_0[t], k_1[t], ..., k_{t-1}[t] &= K_0 x[t], K h[t-1], K h[t-2], ..., K h[1] \\
v_0[t], v_1[t], ..., v_{t-1}[t] &= V_0 x[t], V h[t-1], V h[t-2], ..., V h[t-1] \\
w_i[t] &= q_i[t]^\top k_i[t] \\
a[t] &= \texttt{softargmax}([w_0[t], w_1[t], ..., w_{t-1}[t]]) \\
h[t] &= \sum_{i=0}^{k} a_i[t] v_i[t]
\end{aligned}
$$

(e)

$$
\begin{aligned}
\frac{\partial h[t]}{\partial h[t-1]} &= \frac{\partial a_0[t] v_0[t]}{\partial h[t-1]} + \frac{\partial a_1[t] v_1[t]}{\partial h[t-1]} + \frac{\partial a_2[t] v_2[t]}{\partial h[t-1]} \\
&= v_0[t] \frac{\partial a_0[t]}{\partial h[t-1]} + a_1[t] \frac{\partial v_1[t]}{\partial h[t-1]} + v_1[t] \frac{\partial a_1[t]}{\partial h[t-1]} + v_2[t] \frac{\partial a_2[t]}{\partial h[t-1]} \\
&= v_0[t](-a_0[t] a_1[t]) \frac{\partial w_1}{\partial h[t-1]} + a_1[t] V_1 \\
&\quad + v_1[t] a_1[t](1 - a_1[t]) \frac{\partial w_1}{\partial h[t-1]} + v_2[t](-a_2[t] a_1[t]) \frac{\partial w_1}{\partial h[t-1]} \\
&= a_1[t] \left(Q_1^\top h[t-1]^\top (K_1 + K_1^\top)(v_1[t] - v_0[t] a_0[t] - v_1[t] a_1[t] - v_2[t] a_2[t]) + V_1 \right)
\end{aligned}
$$

(f)

$$
\frac{\partial \ell}{\partial h[T]} = \sum_{t=T+1}^{T+k} \frac{\partial \ell}{\partial h[t]} \frac{\partial h[t]}{\partial h[T]}
$$

1.3 Debugging Loss Curves

1. The spikes are the results of gradient explosion.
2. Because if gradient explosion exists, then the recurrent value will be huge and the initial loss value is only random output.
3. We can add a clip or normalization to constrain the gradient.
4. For the accuracy, because the weights are randomly initialized and there are 4 classes, so the accuracy should be 0.25. And apply this probability to the cross entropy loss, we can get $\ell = 4 \times (-0.25 \times \log 0.25) \approx 1.39$.