Homework 3

Phillip Wang pw1287

October 2021

1.1

a.

Along the y-axis(row-wise), $n = \lfloor \frac{10-3}{2} \rfloor + 1 = 4$, along the x-axis(column-wise), $m = \lfloor \frac{11-3}{2} \rfloor + 1 = 5$. Therefore, the output size is 4×5

b.

With padding P, the input size is effectively changed to $C \times (H+2P) \times (W+2P)$. Since the kernel is dilated by D, the kernel size is effectively $(K + (K-1)(D-1)) \times (K + (K-1)(D-1)) = ((K-1)D+1) \times ((K-1)D+1)$. Following the equations in section 1.1, for a single filter and mono channel, we have a height of $\frac{(H+2P)-((K-1)D+1)}{S} + 1$ and a weight of $\frac{(W+2P)-((K-1)D+1)}{S} + 1$. So the final output size is

$$F \times [\frac{(H+2P)-((K-1)D+1)}{S} + 1] \times [\frac{(W+2P)-((K-1)D+1)}{S} + 1] \times C$$

c.

i)

$f_W(x) \in \mathbb{R}^{1\times 1\times 2}$, where $C = 1, H = 1, W = 2$. Let us use $x[i]$ to express the i-th element of $x[n], i \in 1, 2, 3, 4, 5$ and $x[i][j]$ to express the j-th channel of i-th element in $x[n]$. Then, we have $f_W(x)[i,j,k] = \sum_{i=0}^{2} x[k+i,1] \cdot W[1,1,i+1] + x[k+i,2] \cdot W[1,2,i+1]$

ii)

Using the numerator layout to take element-wise derivative of $f_W(x)$, we have

$$\frac{\partial f_W(x)}{\partial \mathbf{W}} \in \mathbb{R}^{3\times 2} \text{ where } f_W(x) \in \mathbb{R}^2, \mathbf{W} \in \mathbb{R}^{2\times 3}$$

For simplicity, let's denote $f_W(x)$ as y, and we use w'_1 to denote the weight corresponding to the second channel at the first location.

$$\frac{\partial f_W(x)}{\partial \mathbf{W}} = \begin{bmatrix} \frac{\partial y_1}{\partial w_1} + \frac{\partial y_2}{\partial w_1} & \frac{\partial y_1}{\partial w_1'} + \frac{\partial y_2}{\partial w_1'} \\ \frac{\partial y_1}{\partial w_2} + \frac{\partial y_2}{\partial w_2} & \frac{\partial y_1}{\partial w_2'} + \frac{\partial y_2}{\partial w_2'} \\ \frac{\partial y_1}{\partial w_3} + \frac{\partial y_2}{\partial w_3'} & \frac{\partial y_1}{\partial w_3'} + \frac{\partial y_2}{\partial w_3'} \end{bmatrix} \in \mathbb{R}^{3\times 2}$$

iii)

For the first convolution, $x_1 \in \mathbb{R}^{2\times 3}$ is used. Therefore, $\frac{\partial y_1}{\partial x_1} \in \mathbb{R}^{3\times 2}$. Following the same logic for the second convolution, $\frac{\partial y_2}{\partial x_2} \in \mathbb{R}^{3\times 2}$, with one element overlapped in both gradients, which gives us $\frac{\partial f_W(x)}{\partial x} \in \mathbb{R}^{5\times 2}$. For simplicity, let's denote the second channel of x_5 as x_5', we have

$$\frac{\partial f_W(x)}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_1'} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_1}{\partial x_2'} \\ \frac{\partial y_1}{\partial x_3} + \frac{\partial y_2}{\partial x_3} & \frac{\partial y_1}{\partial x_3} + \frac{\partial y_2}{\partial x_3'} \\ \frac{\partial y_2}{\partial x_4} & \frac{\partial y_2}{\partial x_4'} \\ \frac{\partial y_2}{\partial x_5} & \frac{\partial y_2}{\partial x_5'} \end{bmatrix} \in \mathbb{R}^{5\times 2}$$

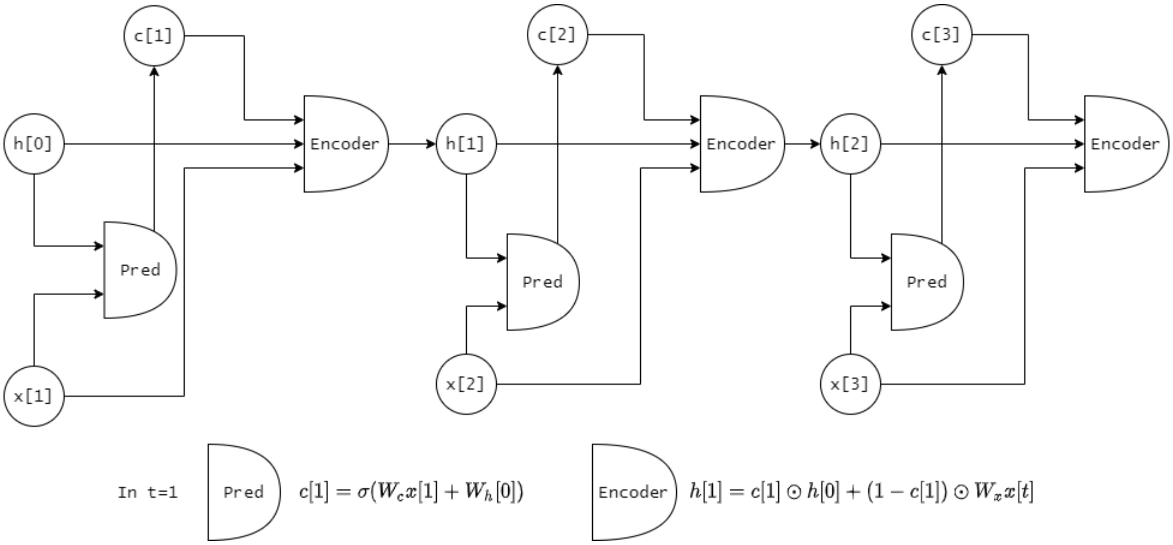
iiii)

Since $f_W(x) \in \mathbb{R}^2$, if we treat it as a column vector then we have $\frac{\partial l}{\partial f_W(x)}$ as a row vector of length 2.

$$\begin{aligned} \frac{\partial l}{\partial W} &= \frac{\partial l}{\partial f_W(x)} \frac{\partial f_W(x)}{\partial W} \\ &= \begin{bmatrix} \frac{\partial l}{\partial y_1}\frac{\partial y_1}{\partial w_1} + \frac{\partial l}{\partial y_2}\frac{\partial y_2}{\partial w_1} & \frac{\partial l}{\partial y_1}\frac{\partial y_1}{\partial w_1'} + \frac{\partial l}{\partial y_2}\frac{\partial y_2}{\partial w_1'} \\ \frac{\partial l}{\partial y_1}\frac{\partial y_1}{\partial w_2} + \frac{\partial l}{\partial y_2}\frac{\partial y_2}{\partial w_2} & \frac{\partial l}{\partial y_1}\frac{\partial y_1}{\partial w_2'} + \frac{\partial l}{\partial y_2}\frac{\partial y_2}{\partial w_2'} \\ \frac{\partial l}{\partial y_1}\frac{\partial y_1}{\partial w_3} + \frac{\partial l}{\partial y_2}\frac{\partial y_2}{\partial w_3'} & \frac{\partial l}{\partial y_1}\frac{\partial y_1}{\partial w_3'} + \frac{\partial l}{\partial y_2}\frac{\partial y_2}{\partial w_3'} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial l}{\partial y_1}x[1,1] + \frac{\partial l}{\partial y_2}x[3,1] & \frac{\partial l}{\partial y_1}x[1,2] + \frac{\partial l}{\partial y_2}x[3,2] \\ \frac{\partial l}{\partial y_1}x[2,1] + \frac{\partial l}{\partial y_2}x[4,1] & \frac{\partial l}{\partial y_1}x[2,2] + \frac{\partial l}{\partial y_2}x[4,2] \\ \frac{\partial l}{\partial y_1}x[3,1] + \frac{\partial l}{\partial y_2}x[5,1] & \frac{\partial l}{\partial y_1}x[3,2] + \frac{\partial l}{\partial y_2}x[5,2] \end{bmatrix} \in \mathbb{R}^{3\times 2} \end{aligned}$$

1.2

a


c[1]
c[2]
c[3]
h[0]
Encoder
h[1]
Encoder
h[2]
Encoder
Pred
Pred
Pred
x[1]
x[2]
x[3]
In t=1
Pred
$c[1] = \sigma(W_c x[1] + W_h[0])$
Encoder
$h[1] = c[1] \odot h[0] + (1 - c[1]) \odot W_x x[t]$

b

$c[t] \in \mathbb{R}^m$ since $W_c \in \mathbb{R}^{m\times n}, x[t] \in \mathbb{R}^n$.

c

For consistency, let us assume all one-dimensional vectors are column vectors. Since $h[t] \in \mathbb{R}^m$, by numerator layout, we have $\frac{\partial l}{\partial h[t]}$ as a row vector $\in \mathbb{R}^{1\times m}$. Then we have $\frac{\partial l}{\partial W_x} = \frac{\partial l}{\partial h[t]} \frac{\partial h[t]}{\partial W_x x[t]} \frac{\partial W_x x[t]}{\partial W_x}$. Since $\frac{\partial x[t]}{\partial W_x}$ is a tensor $\in \mathbb{R}^{m\times n\times m}$ by numerator layout and that $h[t] = (1-c[t]) \odot W_x x[t] \implies \frac{\partial h[t]}{\partial w_x x[t]} = (1-c[t])^T \in \mathbb{R}^{1\times m}$, we have $\frac{\partial l}{\partial W_x} = \frac{\partial l}{\partial h[t]} \sum_i \frac{\partial h[t]}{\partial w_x x[t]} \frac{\partial (w_x x[t])_i}{\partial w_x}$

Since we have

$$\frac{\partial (w_x x[t])_i}{\partial w_x} = \begin{bmatrix} | & & | & & | \\ 0 & \dots & x[t] & \dots & 0 \\ | & & | & & | \end{bmatrix} \in \mathbb{R}^{n\times m} \text{ where the i-th column is } x[t] \text{ and 0 elsewhere}$$

Hence, we have $\sum_i \frac{\partial h[t]}{\partial w_x x[t]} \frac{\partial (w_x x[t])_i}{\partial w_x} = \sum_i \frac{\partial (w_x x[t])_i}{\partial w_x} (1 - c[t][i]) = \sum_i (1 - c[t][i]) \cdot x[t] \in \mathbb{R}^{n\times 1}$, denoted by $sum(1 - c[t]) \cdot x[t]$.
So, we have $\frac{\partial l}{\partial W_x}[i][j] = sum(1 - c[t]) \times x[t][i] \times \frac{\partial l}{\partial h[t]}[j] \in \mathbb{R}^{n\times m}$ where $i = 1 \dots n, j = 1 \dots m$.

In backward propagation, the gradients for W_x is computed using $h[t], x[t], c[t]$, which is in line with the forward propagation since $h[t]$ is computed using the rest of them. The difference is that there is a bias that involves the previous hidden state in the forward propagation, which is discarded in the backward propagation.

d

Yes. With the sigmoid non-linearity, it is likely that the RNN will suffer from vanishing gradients since there is not much to learn when the gradients approaches 0. The exploding gradients typically occurs when there is no non-linearity.