50.039 - Theory and Practice of Deep Learning
Week 8 Homework

Krishna Penukonda - 1001781

1 Task 1 - LSTM

1.1 Is Previous Cell State c_{t-1} a function of the Hidden State h_{t-1}?

c_t is a function of h_{t-1}. Previous cell state c_{t-1} can thus be written as:

$$c_{t-1} = f_{t-1} \circ c_{t-2} + i_{t-1} \circ \tanh(W^c x_{t-1} + U^c h_{t-2})$$

c_{t-1} is therefore *not* a function of h_{t-1}.

1.2 Derivative of the Hidden State

Current hidden state:

$$h_t = o_t \circ \tanh(c_t) \tag{1}$$

Current cell state:

$$c_t = f_t \circ c_{t-1} + i_t \circ u_t \tag{2}$$

Expanding eq. (1) using the value of c_t from eq. (2),

$$h_t = o_t \circ \tanh(f_t \circ c_{t-1} + i_t \circ u_t) \tag{3}$$

Taking the derivative of eq. (3) w.r.t h_{t-1},

$$\frac{\partial h_t}{\partial h_{t-1}} = \frac{\partial o_t}{\partial h_{t-1}} \circ \tanh(c_t) + o_t \circ \left(\left[c_{t-1} \circ \frac{\partial f_t}{\partial h_{t-1}} + f_t \circ \frac{\partial c_{t-1}}{\partial h_{t-1}} + i_t \circ \frac{\partial u_t}{\partial h_{t-1}} + u_t \circ \frac{\partial i_t}{\partial h_{t-1}} \right] (1 - \tanh^2(c_t)) \right) \tag{4}$$

We know that c_{t-1} is not a function of h_{t-1}. Therefore,

$$\frac{\partial c_{t-1}}{\partial h_{t-1}} = 0 \tag{5}$$

Substituting (5) into (4), we get the final result:

$$\frac{\partial h_t}{\partial h_{t-1}} = \frac{\partial o_t}{\partial h_{t-1}} \circ \tanh(c_t) + o_t \circ \left(\left[c_{t-1} \circ \frac{\partial f_t}{\partial h_{t-1}} + i_t \circ \frac{\partial u_t}{\partial h_{t-1}} + u_t \circ \frac{\partial i_t}{\partial h_{t-1}} \right] (1 - \tanh^2(c_t)) \right)$$

1.3 Sigmoid derivative

$$\begin{aligned}
\sigma(z) &= \frac{1}{1+e^{-z}} \\
\implies \frac{\mathrm{d}\sigma(z)}{\mathrm{d}z} &= \frac{1}{e^z} \cdot \frac{1}{(1+e^{-z})^2} \\
&= \frac{1}{1+e^{-z}} \cdot \frac{e^{-z}}{1+e^{-z}} \\
&= \sigma(z) \left(\frac{1+e^{-z}}{1+e^{-z}} - \frac{1}{1+e^{-z}} \right) \\
&= \sigma(z)(1-\sigma(z))
\end{aligned} \tag{6}$$

1.4 Derivative of the Forget Gate

$$\begin{aligned} f_t &= \sigma(W^f x_t + U^f h_{t-1}) \\ \implies \frac{\partial f_t}{\partial h_{t-1}} &= \frac{\partial (W^f x_t + U^f h_{t-1})}{\partial h_{t-1}} \cdot \sigma'(W_f x_t + U_f h_{t-1}) \\ &= U^f \cdot \sigma'(W_f x_t + U_f h_{t-1}) \end{aligned}$$

Using the derivative of $\sigma(z)$ we calculated in eq. (6),

$$\frac{\partial f_t}{\partial h_{t-1}} = U_f \cdot \sigma(W_f x_t + U_f h_{t-1}) \cdot (1 - \sigma(W_f x_t + U_f h_{t-1})) \tag{7}$$

1.5 Gate Activation

-

2 Task 2 - Convolution

2.1 Feature Map Spatial Size

$$D_{out} = \left\lfloor \frac{D_{in} + 2P - k}{S} + 1 \right\rfloor \tag{8}$$

Where:

D_{out} is the output dimension

D_{in} is the input dimension

P is the padding size

k is the kernel size

S is the stride length

2.1.1

$$H_{in} = 78, \qquad W_{in} = 84, \qquad P = 2, \qquad k = (5,5), \qquad S = 3$$

$$\begin{aligned} H_{out} &= \left\lfloor \frac{78 + 2 \cdot 2 - 5}{3} + 1 \right\rfloor \\ &= \left\lfloor \frac{77}{3} + 1 \right\rfloor \\ &= 26 \end{aligned}$$

$$\begin{aligned} W_{out} &= \left\lfloor \frac{84 + 2 \cdot 2 - 5}{3} + 1 \right\rfloor \\ &= \left\lfloor \frac{83}{3} + 1 \right\rfloor \\ &= 28 \end{aligned}$$

2.1.2

$$H_{in} = 64, \qquad W_{in} = 64, \qquad P = 0, \qquad k = (3,5), \qquad S = 2$$

$$\begin{aligned} H_{out} &= \left\lfloor \frac{64 + 2 \cdot 0 - 3}{2} + 1 \right\rfloor \\ &= \left\lfloor \frac{63}{2} + 1 \right\rfloor \\ &= 32 \end{aligned}$$

$$\begin{aligned} W_{out} &= \left\lfloor \frac{64 + 2 \cdot 0 - 5}{2} + 1 \right\rfloor \\ &= \left\lfloor \frac{61}{2} + 1 \right\rfloor \\ &= 31 \end{aligned}$$

2.1.3

$$D_{out} = 16, \quad P = 1, \quad k = 9, \quad S = 3$$

$$\begin{aligned} &\left\lfloor \frac{D_{in} + 2 \cdot 1 - 9}{3} + 1 \right\rfloor = 16 \\ \implies &\left\lfloor \frac{D_{in} - 4}{3} \right\rfloor = 16 \\ \implies &D_{in} = (16 \cdot 3) + 4 = 52 \end{aligned}$$

2.2 Trainable Parameters

Trainable parameters = $Channels_{in} * Kernel_x * Kernel_y * Channels_{out}$
Multiplications = $Channels_{in} * Kernel_x * Kernel_y * Channels_{out} * H_{out} * W_{out}$
Sum operations = $Channels_{in} * Channels_{out} * H_{out} * W_{out}$

2.2.1

Trainable parameters = $32 * 7 * 7 * 64 = 100352$
Multiplications = $32 * 7 * 7 * 64 * 5 * 5 = 2508800$
Sum operations = $32 * 64 * 5 * 5 = 51200$

2.2.2

Trainable parameters = $512 * 1 * 1 * 128 = 65536$