CS286 homework2

Problem 1 (20 points)
Instructor: Jie Zheng (SIST)

Oct,2020

Requirements:

- The following are theory problems, so that you don't need coding. (Of course you can coding to verify your answers).
- Please write down the key derivation and calculation steps, because only the answer will not be accepted.
- This part we only accept solutions in **pdf** format, but do not accept figure with handwriting. If you cannot handle LaTeXwell, you can write it with **Word** first and then convert it into a **pdf** file.
- Please name your file in the format of **name_student_id_hw2**.

1. *Convolutional Neural Network*(10 points).

   (1) Let input $\boldsymbol{x}$ be a matrix with shape of $64 \times 64 \times 3$,
   layer $\boldsymbol{L_1}$ with 10 4×4 filters with stride=2, no padding, dilation=1;
   layer $\boldsymbol{L_2}$ with 20 3×3 filters with stride=4, padding=2, dilation=1;
   layer $\boldsymbol{L_3}$ with 10 3×3 filters with stride=2, no padding, dilation=2.
   Denote $\boldsymbol{x_1} = \boldsymbol{L_1(x)}$, $\boldsymbol{x_2} = \boldsymbol{L_2(x_1)}$ and $\boldsymbol{x_3} = \boldsymbol{L_3(x_2)}$.
   Compute the shapes of $\boldsymbol{x_1, x_2, x_3}$ and the numbers of parameters of layer $\boldsymbol{L_1, L_2}$ and $\boldsymbol{L_3}$. You must consider about bias(5 points).

   Solution: According to the definition, we can get: $w_{\text{out}} = \frac{w_{\text{in}}+2* \text{ padding } -K}{\text{stride}} + 1$
   where $K = kernel + (kernel - 1) \cdot (dilation - 1)$
   Shape of x_1:
   $\therefore w_1 = 31 and n_1 = 10$
   $\therefore L_1 = (4 \times 4 \times 3 + 1) \times 10 = 490$
   Shape of x_2:
   $\therefore w_2 = 9 and n_2 = 20$
   $\therefore L_2 = (3 \times 3 \times 10 + 1) \times 20 = 1820$
   Shape of x_3:
   $\therefore w_3 = 3 and n_3 = 10$
   $\therefore L_3 = (3 \times 3 \times 20 + 1) \times 10 = 1810$

   (2) Given a tensor $\boldsymbol{x} =$

1	2	-1	0	3
2	-3	6	4	1
1	5	-2	0	3
1	3	4	5	7
-2	-1	0	3	4

   as input,

   layer $\boldsymbol{L_1}$ with kernel

-1	2
3	1

   , dilation=2,no padding, stride=1;

layer $\boldsymbol{L_2}$ with kernel

1	2
-1	3

, dilation=1,no padding, stride=1;

layer $\boldsymbol{L_3}$ is a max pooling layer with kernel size 2×2.

Denote $\boldsymbol{x_1} = \boldsymbol{L_1(x)}$, $\boldsymbol{x_2} = \boldsymbol{L_2(x_1)}$ and $\boldsymbol{x_3} = \boldsymbol{L_3(x_2)}$.

Compute $\boldsymbol{x_1, x_2}$ and $\boldsymbol{x_3}$(5 points).

Solution: Considering the influence of dilation, L_1's kernel should be

-1	0	2
0	0	0
3	0	1

$\therefore \boldsymbol{x_1} =$

-2	13	4
17	25	15
-11	-5	12

$\therefore \boldsymbol{x_2} =$

82	41
63	96

$\therefore \boldsymbol{x_3} = 96$

2. *Long Short Term Memory networks*(5 points). Suppose
$\boldsymbol{W_f} = \begin{bmatrix} 1 & 2 & 4 & 5 \\ -1 & 2 & 0 & 1 \end{bmatrix}$, $\boldsymbol{W_i} = \begin{bmatrix} -1 & 2 & 0 & 1 \\ 1 & -2 & 3 & -1 \end{bmatrix}$, $\boldsymbol{W_c} = \begin{bmatrix} 1 & -2 & 3 & -1 \\ 2 & 4 & 1 & -3 \end{bmatrix}$, $\boldsymbol{W_o} = \begin{bmatrix} 2 & 4 & 1 & -3 \\ 1 & 2 & 4 & 5 \end{bmatrix}$;
$\boldsymbol{b_f} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$, $\boldsymbol{b_i} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$, $\boldsymbol{b_c} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\boldsymbol{b_o} = \begin{bmatrix} 0.5 \\ -0.1 \end{bmatrix}$;
$\boldsymbol{x_t} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $\boldsymbol{h_{t-1}} = \begin{bmatrix} 0.4 \\ -0.2 \end{bmatrix}$, $\boldsymbol{c_{t-1}} = \begin{bmatrix} -0.5 \\ 0.3 \end{bmatrix}$.
Compute $\boldsymbol{c_t}$ and $\boldsymbol{h_t}$ based on standard LSTM structure, preserving 4 decimal places in your result.
(**Hint:** You can refer to this link.)

Solution:

$$f_t = \sigma(W_f \times [h_{t-1}^T, x_t^T]^T + b_f) = \sigma(\begin{bmatrix} 1 & 2 & 4 & 5 \\ -1 & 2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0.4 \\ -0.2 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ -2 \end{bmatrix}) = \sigma(\begin{bmatrix} 5 \\ -2.8 \end{bmatrix} = \begin{bmatrix} 0.9933 \\ 0.0573 \end{bmatrix})$$

$$i_t = \sigma(W_i \times [h_{t-1}^T, x_t^T]^T + b_i) = \sigma(\begin{bmatrix} -1 & 2 & 0 & 1 \\ 1 & -2 & 3 & -1 \end{bmatrix} \begin{bmatrix} 0.4 \\ -0.2 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} -1 \\ 1 \end{bmatrix}) = \sigma(\begin{bmatrix} -1.8 \\ 4.8 \end{bmatrix}) = \begin{bmatrix} 0.1419 \\ 0.9918 \end{bmatrix}$$

$$\tilde{C}_t = \tanh(\begin{bmatrix} 1 & -2 & 3 & -1 \\ 2 & 4 & 1 & -3 \end{bmatrix} \begin{bmatrix} 0.4 \\ -0.2 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix}) = \tanh \begin{bmatrix} 3.8 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.9990 \\ 0.7616 \end{bmatrix}$$

$$C_t = \begin{bmatrix} 0.9933 \\ 0.0573 \end{bmatrix} \times \begin{bmatrix} -0.5 \\ 0.3 \end{bmatrix} + \begin{bmatrix} 0.1419 \\ 0.9918 \end{bmatrix} \times \begin{bmatrix} 0.9990 \\ 0.7616 \end{bmatrix} = \begin{bmatrix} -0.3549 \\ 0.7725 \end{bmatrix}$$

$$O_t = \sigma(\begin{bmatrix} 2 & 4 & 1 & -3 \\ 1 & 2 & 4 & 5 \end{bmatrix} \begin{bmatrix} 0.4 \\ -0.2 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0.5 \\ -0.1 \end{bmatrix}) = \sigma(\begin{bmatrix} 1.5 \\ 3.9 \end{bmatrix}) = \begin{bmatrix} 0.8176 \\ 0.9802 \end{bmatrix}$$

$$\therefore h_t = O_t \times \tanh(C_t) = \begin{bmatrix} -0.3407 \\ 0.6484 \end{bmatrix}$$

3. *Variational Autoencoder*(5 points). Denote $\boldsymbol{P_1} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, $\boldsymbol{P_2} \sim \mathcal{N}(0, 1)$, prove that

$$\boldsymbol{KL}(\boldsymbol{P_1}||\boldsymbol{P_2}) = \frac{1}{2}\left(\boldsymbol{\mu}^2 + \boldsymbol{\sigma}^2 - \log\boldsymbol{\sigma}^2 - 1\right),$$

where $\mathcal{N}()$ is the Gaussian distribution, and $\boldsymbol{KL}$ is the function of Kullback-Leibler divergence.

Solution:
According to the definition:

$$
\begin{aligned}
\boldsymbol{KL}(\boldsymbol{P_1}||\boldsymbol{P_2}) &= \int_x \frac{1}{\sqrt{2\pi}\sigma} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}} \log \frac{\dfrac{1}{\sqrt{2\pi}\sigma} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}}{\dfrac{1}{\sqrt{2\pi}} e^{-\dfrac{x^2}{2}}} dx \\
&= \int_x \frac{1}{\sqrt{2\pi}\sigma} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}} \left[\log \frac{1}{\sigma} - \frac{(x-\mu)^2}{2\sigma^2} + \frac{x^2}{2} \right] dx \\
&= \log \frac{1}{\sigma} \int_x \frac{1}{\sqrt{2\pi}\sigma} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}} dx - \frac{1}{2\sigma^2} \int_x (x-\mu)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}} dx + \frac{1}{2} \int_x x^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}} dx \\
&= \log \frac{1}{\sigma} - \frac{1}{2\sigma^2} \times D(X-\mu) + \frac{1}{2} EX^2 \\
&= \log \frac{1}{\sigma} - \frac{1}{2\sigma^2} \times \sigma^2 + \frac{1}{2}(DX + E^2 X) \\
&= \log \frac{1}{\sigma} - \frac{1}{2} + \frac{\sigma^2 + \mu^2}{2}
\end{aligned}
$$

Proved.