CSCE 676-HW2.

JinHyunPark
6770018229.

1.

$$E_{in}(\omega)=\frac{1}{N}\sum_{n=1}^{N}\left(\tanh(\omega^T x_n)-y_n\right)^2$$

$$\nabla E_{in}(\omega)=\frac{1}{N}\times 2\cdot\sum_{n=1}^{N}\left(\tanh(\omega^T x_n)-y_n\right)\cdot\nabla\tanh(\omega^T x_n)\quad\cdots ①$$

Let $\tanh(x)=g(x)$. We will find the derivative of g and use it to solve eq. ①

$$\frac{d}{dx}\cdot g(x)=\frac{d}{dx}\cdot\frac{e^x-e^{-x}}{e^x+e^{-x}}$$

$$=\frac{(e^x+e^{-x})\cdot(e^x+e^{-x})-(e^x-e^{-x})(e^x-e^{-x})}{(e^x+e^{-x})^2}$$

$$=1-\left(\frac{e^x-e^{-x}}{e^x+e^{-x}}\right)^2$$

$$\therefore\ \frac{d}{dx}g(x)=\nabla g(x)=1-\left(\frac{e^x-e^{-x}}{e^x+e^{-x}}\right)^2=1-(\tanh x)^2\quad\cdots ②$$

Plugging the result ② to ①, we get

$$\nabla E_{in}(\omega)=\frac{2}{N}\sum_{n=1}^{N}\left(\tanh(\omega^T x_n)-y_n\right)\left(1-\tanh^2(\omega^T x_n)\right)\cdot\frac{d}{d\omega}(\omega^T x_n)$$

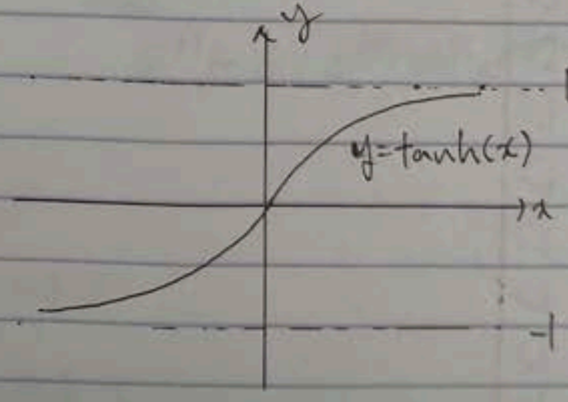
$$\therefore\ \nabla E_{in}(\omega)=\frac{2}{N}\cdot\sum_{n=1}^{N}\left(\tanh(\omega^T x_n)-y_n\right)\left(1-\tanh^2(\omega^T x_n)\right)\cdot x_n$$

= (gradient of in-sample error)

If $\omega\to\infty$, $\tanh^2(\omega^T x_n)\to 1$

Then, $\left(1-\tanh^2(\omega^T x_n)\right)\to 0$

This implies that $\nabla E_{in}(\omega)$ becomes 0 and this results in vanishing gradient issue. The weights won't be updated properly.



Q2.

$x^{0} \xrightarrow{W^{1}} S^{1} \xrightarrow{\theta} x^{1} \xrightarrow{W^{2}} \cdots \rightarrow x^{(L)} = h(x)$

Weight matrices are :

$s^{1} = W^{1^T} \cdot x^{0}$

$s^{(L)} \xrightarrow{\theta} x^{(L)}$

$$W^{(1)} = \begin{bmatrix} 0.1 & 0.2 \\ 0.3 & 0.4 \end{bmatrix} \quad W^{(2)} = \begin{bmatrix} 0.2 \\ 1 \\ -3 \end{bmatrix} \quad W^{(3)} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

and $x = 2$, $y = 1$.

$$x^{(0)} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad s^{(1)} = \underbrace{\begin{bmatrix} 0.1 & 0.3 \\ 0.2 & 0.4 \end{bmatrix}}_{W^{(1)^T}} \underbrace{\begin{bmatrix} 1 \\ 2 \end{bmatrix}}_{x^{(0)}} = \begin{bmatrix} 0.7 \\ 1 \end{bmatrix}$$

$$x^{(1)} = \begin{bmatrix} 1 \\ 0.6 \\ 0.76 \end{bmatrix} \quad s^{(2)} = \underbrace{\begin{bmatrix} 0.2 & 1 & -3 \end{bmatrix}}_{W^{(2)^T}} \underbrace{\begin{bmatrix} 1 \\ 0.6 \\ 0.76 \end{bmatrix}}_{x^{(1)}} = [-1.48]$$

$$x^{(2)} = \begin{bmatrix} 1 \\ -0.9 \end{bmatrix} \quad s^{(3)} = \underbrace{\begin{bmatrix} 1 & 2 \end{bmatrix}}_{W^{(3)^T}} \underbrace{\begin{bmatrix} 1 \\ -0.9 \end{bmatrix}}_{x^{(2)}} = [-0.8], \quad x^{(3)} = [-0.8]$$

output transformation is identity.

Output transformation is identity. This means that also, $\theta'(s^{(3)}) = 1$

We use the following equation to calculate δ.

things to remember:

$$\begin{bmatrix} \delta^{(l)} = \theta'(s^{(l)}) \otimes \left[W^{(l+1)} \delta^{(l+1)} \right]_{1}^{d^{(l)}} & l = L-1 \text{ to } 1 \\ \delta^{(L)} = 2(x^{(L)} - y)\, \theta'(s^{(L)}) & \end{bmatrix}$$

$\delta^{(3)} = 2(x^{(3)} - 1) \cdot 1 = 2(-1.8) = [-3.6]$

$\delta^{(2)} = \theta'(s^{(2)}) \otimes [W^{(3)} \cdot \delta^{(3)}] = (1 - \tanh^2(-1.48)) \otimes 2 \cdot (-3.6) = [-1.368]$

$$\delta^{(1)} = \theta'(s^{(1)}) \otimes [W^{(2)} \cdot \delta^{(2)}] = \begin{bmatrix} (1 - \tanh^2(0.7)) \cdot 1 \\ (1 - \tanh^2(1)) \cdot (-3) \end{bmatrix} \cdot (-1.368) = \begin{bmatrix} -0.876 \\ 1.774 \end{bmatrix}$$

We ignore the 0th row of W

☆ $\frac{\partial e}{\partial W^{(l)}} = x^{(l-1)} \cdot (\delta^{(l)})^T$: thing to remember.

$$\frac{\partial e}{\partial W^{(1)}} = x^{(0)} \cdot (\delta^{(1)})^T = \underset{2\times1}{\begin{bmatrix} 1 \\ 2 \end{bmatrix}} \cdot \underset{1\times2}{\begin{bmatrix} -0.876 & 1.734 \end{bmatrix}} = \begin{bmatrix} -0.876 & 1.734 \\ -1.752 & 3.468 \end{bmatrix}$$

$$\frac{\partial e}{\partial W^{(2)}} = x^{(1)} \cdot (\delta^{(2)})^T = \begin{bmatrix} 1 \\ 0.6 \\ 0.76 \end{bmatrix} \cdot \begin{bmatrix} -1.368 \end{bmatrix} = \begin{bmatrix} -1.368 \\ -0.821 \\ -1.040 \end{bmatrix}$$

$$\frac{\partial e}{\partial W^{(3)}} = x^{(2)} \cdot (\delta^{(3)})^T = \begin{bmatrix} 1 \\ -0.9 \end{bmatrix} \cdot \begin{bmatrix} -3.6 \end{bmatrix} = \begin{bmatrix} -3.6 \\ 3.24 \end{bmatrix}$$

3.

① standard residual block

input dim. = output dim

$= \underbrace{128}_{\text{Batch size}} \times \underbrace{16 \times 16}_{\text{feature map size}} \times \underbrace{32}_{\text{channels}}$

case1: Num. of param. with bias = 18496 since,

1st layer $\left(\left(\underbrace{32}_{\text{channel depth}} \times \underbrace{3 \times 3}_{\text{filter size (kernel size)}} \right) + 1 \right) \cdot \underbrace{32}_{\text{32 filters}} = 9248$

2nd layer $\left((32 \times 3 \times 3) + 1 \right) \cdot 32 = 9248$

case2: Num of param. without bias = 18432 since,

1st layer $(32 \times 3 \times 3) \times 32 = 9216$

2nd layer $(32 \times 3 \times 3) \times 32 = 9216$

② Bottleneck block

input dim = output dim

$= \underbrace{128}_{\text{Batchsize}} \times 16 \times 16 \times \underbrace{128}_{\text{Chanels}}$

case1: Num of param. with bias = 17600 since

1st layer: $\left((128 \times 1 \times 1) + 1 \right) \times 32 = 4128$

2nd layer: $\left((32 \times 3 \times 3) + 1 \right) \times 32 = 9248$

3rd layer $\left((32 \times 1 \times 1) + 1 \right) \times 128 = 4224$

case2. Num of param without bias = 17408 since,

1st layer : $(128 \times 1 \times 1) \times 32 = 4096$

2nd layer : $(32 \times 3 \times 3) \times 32 = 9216$

3rd layer : $(32 \times 1 \times 1) \times 128 = 4096$

We can summarize the num of param in the following table

	standard residual block	bottle-neck block
bias O	18496	17600
bias X	18432	17408

IMPORTANT NOTE!

If we consider batch-norm parameters then
(1) we should add 32 * 2 + 32 * 2
for the standard residual block.

(2) we should add 32 * 2 + 32 * 2 + 128 * 2
for the bottleneck block

Advantage of bottleneck over standard residual :

- As we use less parameters, its computational cost is lower than the standard residual block
- It can be used to obtain a representation with reduced dimensionality
- This is similar to Autoencoder where the latent space(= vector or layer) contains(= encodes) the important features of the image!

Disadvantag of using bottleneck

- We might loose some important features (information) of the image because it uses identity convolution.

(because we force dimensionality to be reduced which may lead to loose some important features)

4.

(a) shape of mean and variance is $1 \times C$

(b) As we normalise all the activation in a batch (batchsize = N) and this normalisation cover all pixels (or element) in $H \times W$, the shape of mean and variance is $1 \times 1 \times 1 \times C$

5(a)

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \end{bmatrix} \in R^{2\times4}$$

$$W^{ij} = [w_1^{ij}, w_2^{ij}, w_3^{ij}], \; i = 1.2 \,/\, j = 1.2$$

$$Y = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} \in R^{2\times2}$$

$$y_{11} = w_1^{11} x_{11} + w_2^{11} \cdot x_{12} + w_3^{11} x_{13} + w_1^{21} x_{21} + w_2^{21} \cdot x_{22} + w_3^{21} x_{23}$$

$$y_{12} = w_1^{11} x_{12} + w_2^{11} \cdot x_{13} + w_3^{11} \cdot x_{14} + w_1^{21} x_{22} + w_2^{21} \cdot x_{23} + w_3^{21} \cdot x_{24}$$

$$y_{21} = w_1^{12} x_{11} + w_2^{12} x_{12} + w_3^{12} x_{13} + w_1^{22} x_{21} + w_2^{22} x_{22} + w_3^{22} x_{23}$$

$$y_{22} = w_1^{12} x_{12} + w_2^{12} x_{13} + w_3^{12} x_{14} + w_1^{22} x_{22} + w_2^{22} x_{23} + w_2^{22} x_{24}$$

$$\underset{4\times1}{\tilde{Y}} = \underset{4\times8}{A}\, \underset{8\times1}{\tilde{X}}, \quad A \in R^{4\times8}$$

$$\underset{4\times1}{\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \end{bmatrix}} = \underset{4\times8}{\begin{bmatrix} w_1^{11} & w_2^{11} & w_3^{11} & 0 & w_1^{21} & w_2^{21} & w_3^{21} & 0 \\ 0 & w_1^{11} & w_2^{11} & w_3^{11} & 0 & w_1^{21} & w_2^{21} & w_3^{21} \\ w_1^{12} & w_2^{12} & w_3^{12} & 0 & w_1^{22} & w_2^{22} & w_3^{22} & 0 \\ 0 & w_1^{12} & w_2^{12} & w_3^{12} & 0 & w_1^{22} & w_2^{22} & w_3^{22} \end{bmatrix}} \underset{8\times1}{\begin{bmatrix} x_{11} \\ x_{12} \\ x_{13} \\ x_{14} \\ x_{21} \\ x_{22} \\ x_{23} \\ x_{24} \end{bmatrix}} \cdots$$

5(b) (Please see the next page for proof!)

$$\begin{bmatrix} \partial L/\partial x_{11} \\ \partial L/\partial x_{12} \\ \partial L/\partial x_{13} \\ \partial L/\partial x_{14} \\ \partial L/\partial x_{21} \\ \partial L/\partial x_{22} \\ \partial L/\partial x_{23} \\ \partial L/\partial x_{24} \end{bmatrix} = \begin{bmatrix} w_1^{11} & 0 & w_1^{12} & 0 \\ w_2^{11} & w_1^{11} & w_2^{12} & w_1^{12} \\ w_3^{11} & w_2^{11} & w_3^{12} & w_3^{12} \\ 0 & w_3^{11} & 0 & w_3^{12} \\ w_1^{21} & 0 & w_1^{22} & 0 \\ w_2^{21} & w_1^{21} & w_2^{22} & w_1^{22} \\ w_3^{21} & w_2^{21} & w_3^{22} & w_3^{22} \\ 0 & w_3^{21} & 0 & w_3^{22} \end{bmatrix} \begin{bmatrix} \partial L/\partial y_{11} \\ \partial L/\partial y_{12} \\ \partial L/\partial y_{21} \\ \partial L/\partial y_{22} \end{bmatrix}$$

$$\underset{\partial L/\partial \tilde{X}}{8\times 1} \qquad \underset{B}{8\times 4} \qquad \underset{\partial L/\partial \tilde{Y}}{4\times 1}$$

NOTE:

$$\tilde{Y} = A\tilde{X}$$

5(b)

$$\boxed{\partial L/\partial \tilde{Y}} = \left[\partial L/\partial y_{11} \quad \partial L/\partial y_{12} \quad \partial L/\partial y_{21} \quad \partial L/\partial y_{22}\right]^T$$

$$\boxed{\partial L/\partial \tilde{X} = B \cdot \partial L/\partial \tilde{Y} \Rightarrow \underset{8\times1}{\partial L/\partial \tilde{x}} = \underset{8\times4}{\partial \tilde{Y}/\partial \tilde{X}} \cdot \underset{4\times1}{\partial L/\partial \tilde{Y}}}$$

$$\partial L/\partial x_{11} = \partial y_{11}/\partial x_{11} \cdot \partial L/\partial y_{11} + \cancelto{0}{\partial y_{12}/\partial x_{11} \cdot \partial L/\partial y_{12}} + \partial y_{21}/\partial x_{11} \cdot \partial L/\partial y_{21} + \cancelto{0}{\partial y_{22}/\partial x_{11} \cdot \partial L/\partial y_{22}} \quad \cdots\cdots (1)$$

$$\partial L/\partial x_{12} = \partial y_{11}/\partial x_{12} \cdot \partial L/\partial y_{11} + \partial y_{12}/\partial x_{12} \cdot \partial L/\partial y_{12} + \partial y_{21}/\partial x_{12} \cdot \partial L/\partial y_{21} + \partial y_{22}/\partial x_{12} \cdot \partial L/\partial y_{22} \quad \cdots\cdots (2)$$

$$\partial L/\partial x_{13} = \partial y_{11}/\partial x_{13} \cdot \partial L/\partial y_{11} + \partial y_{12}/\partial x_{13} \cdot \partial L/\partial y_{12} + \partial y_{21}/\partial x_{13} \cdot \partial L/\partial y_{21} + \partial y_{22}/\partial x_{13} \cdot \partial L/\partial y_{22} \quad \cdots\cdots (3)$$

$$\partial L/\partial x_{14} = \cancelto{0}{\partial y_{11}/\partial x_{14} \cdot \partial L/\partial y_{11}} + \partial y_{12}/\partial x_{14} \cdot \partial L/\partial y_{12} + \cancelto{0}{\partial y_{21}/\partial x_{14} \cdot \partial L/\partial y_{21}} + \partial y_{22}/\partial x_{14} \cdot \partial L/\partial y_{22} \quad \cdots\cdots (4)$$

By simplifying eq (1), we get $w_1^{11} \partial L/\partial y_{11} + w_1^{12} \partial L/\partial y_{21}$

" (2), " $w_2^{11} \partial L/\partial y_{11} + w_1^{11} \partial L/\partial y_{12} + w_2^{12} \partial L/\partial y_{21} + w_1^{12} \partial L/\partial y_{22}$

(3), " $w_3^{11} \partial L/\partial y_{11} + w_2^{11} \partial L/\partial y_{12} + w_3^{12} \partial L/\partial y_{21} + w_2^{12} \partial L/\partial y_{22}$

(4), " $w_3^{11} \partial L/\partial y_{12} + w_3^{12} \partial L/\partial y_{22}$

And for $\partial L/\partial x_{2k}$ where $k = \{1, 2, 3, 4\}$, perform the same step. See the next page.

$$\frac{\partial L}{\partial x_{21}} = \frac{\partial y_{11}}{\partial x_{21}} \cdot \frac{\partial L}{\partial y_{11}} + \underbrace{\cancel{\frac{\partial y_{12}}{\partial x_{21}} \cdot \frac{\partial L}{\partial y_{12}}}}_{\to 0} + \frac{\partial y_{21}}{\partial x_{21}} \cdot \frac{\partial L}{\partial y_{21}} + \underbrace{\cancel{\frac{\partial y_{22}}{\partial x_{21}} \cdot \frac{\partial L}{\partial y_{22}}}}_{\to 0} \quad \cdots (5)$$

$$\frac{\partial L}{\partial x_{22}} = \frac{\partial y_{11}}{\partial x_{22}} \cdot \frac{\partial L}{\partial y_{11}} + \frac{\partial y_{12}}{\partial x_{22}} \cdot \frac{\partial L}{\partial y_{12}} + \frac{\partial y_{21}}{\partial x_{22}} \cdot \frac{\partial L}{\partial y_{21}} + \frac{\partial y_{22}}{\partial x_{22}} \cdot \frac{\partial L}{\partial y_{22}} \quad \cdots (6)$$

$$\frac{\partial L}{\partial x_{23}} = \frac{\partial y_{11}}{\partial x_{23}} \cdot \frac{\partial L}{\partial y_{11}} + \frac{\partial y_{12}}{\partial x_{23}} \cdot \frac{\partial L}{\partial y_{12}} + \frac{\partial y_{21}}{\partial x_{23}} \cdot \frac{\partial L}{\partial y_{21}} + \frac{\partial y_{22}}{\partial x_{23}} \cdot \frac{\partial L}{\partial y_{22}} \quad \cdots (7)$$

$$\frac{\partial L}{\partial x_{24}} = \underbrace{\cancel{\frac{\partial y_{11}}{\partial x_{24}} \cdot \frac{\partial L}{\partial y_{11}}}}_{\to 0} + \frac{\partial y_{12}}{\partial x_{24}} \cdot \frac{\partial L}{\partial y_{12}} + \underbrace{\cancel{\frac{\partial y_{21}}{\partial x_{24}} \cdot \frac{\partial L}{\partial y_{21}}}}_{\to 0} + \frac{\partial y_{22}}{\partial x_{24}} \cdot \frac{\partial L}{\partial y_{22}} \quad \cdots (8)$$

By simplifying eq. (5), we get $w_1^{21} \cdot \frac{\partial L}{\partial y_{11}} + w_1^{22} \cdot \frac{\partial L}{\partial y_{21}}$

" (6) " $w_2^{21} \cdot \frac{\partial L}{\partial y_{11}} + w_1^{21} \cdot \frac{\partial L}{\partial y_{12}} + w_2^{22} \cdot \frac{\partial L}{\partial y_{21}} + w_1^{22} \cdot \frac{\partial L}{\partial y_{22}}$

" (7) " $w_3^{21} \cdot \frac{\partial L}{\partial y_{11}} + w_2^{21} \cdot \frac{\partial L}{\partial y_{12}} + w_3^{22} \cdot \frac{\partial L}{\partial y_{21}} + w_2^{22} \cdot \frac{\partial L}{\partial y_{22}}$

(8) " $w_3^{21} \cdot \frac{\partial L}{\partial y_{12}} + w_3^{22} \cdot \frac{\partial L}{\partial y_{22}}$

$B \in R^{8 \times 4}$

$$B = \begin{bmatrix} w_1^{11} & 0 & w_1^{12} & 0 \\ w_2^{11} & w_1^{11} & w_2^{12} & w_1^{12} \\ w_3^{11} & w_2^{11} & w_3^{12} & w_2^{12} \\ w_1^{21} & 0 & w_1^{22} & 0 \\ w_2^{21} & w_1^{21} & w_2^{22} & w_1^{22} \\ w_3^{21} & w_2^{21} & w_3^{22} & w_2^{22} \\ 0 & w_3^{21} & 0 & w_3^{22} \end{bmatrix}$$

" $0 \quad w_3^{11} \quad 0 \quad w_3^{12}$ "

Relationship between A and B

$$A = B^T$$

5(c)

(spatial) input size of conv layer (includes padding)

$$\mathrm{Conv}_{out} = \frac{\mathrm{conv}_{in} - K}{S} + 1 \quad \cdots (A)$$

K: kernel size

S: stride

Conv_{out}: output size of a conv layer (spatial)

$$\underbrace{\partial L / \partial \tilde{X}}_{R^{2\times 4}} = B \cdot \underbrace{\partial L / \partial \tilde{Y}}_{R^{2\times 2}}$$

* We don't have kernels for convolution if there is no padding.
Consider (A), then

$4 = \frac{2-K}{S} + 1$, $3 = \frac{2-K}{S}$, $\underbrace{3S = 2-K}_{(B)}$. We cannot find such S and K that satisfies equation (B).

$\boxed{\text{Consider (A) again * with padding on } \partial L / \partial Y}$

$(2+2\cdot 2)$ $4 = \frac{6-3}{1} + 1$. $K=3$, $S=1$ then this condition satisfies 5(c). Therefore,

2: original conv_{in}; 2·2: padding left & right

$$\underbrace{\partial L / \partial Y}_{\text{with padding}} = \begin{bmatrix} 0 & 0 & \partial L/\partial y_{11} & \partial L/\partial y_{12} & 0 & 0 \\ 0 & 0 & \partial L/\partial y_{21} & \partial L/\partial y_{22} & 0 & 0 \end{bmatrix}$$

$$\partial L / \partial X = \begin{bmatrix} \partial L/\partial x_{11} & \partial L/\partial x_{12} & \partial L/\partial x_{13} & \partial L/\partial x_{14} \\ \partial L/\partial x_{21} & \partial L/\partial x_{22} & \partial L/\partial x_{23} & \partial L/\partial x_{24} \end{bmatrix}$$

$$\boxed{\text{kernel} \Rightarrow W^{ij} = \left[w_3^{ji}, w_2^{ji}, w_1^{ji} \right]} \quad i=1,2. \; j=1,2.$$

also consider that $A = B^T$! And equation (3) on Q5 in HW2 is $W^{ij} = \left[w_1^{ij}, w_2^{ij}, w_3^{ij} \right]$ $i=1,2$, $j=1,2$