NAME: RIZU JAIN
UIN: 430000753

CSCE 636: HW2

Q-1

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

$$\nabla E_{in}(w) = \frac{2}{N}\sum_{n=1}^{N}\Big[\left(\tanh\left(w^T x_n\right) - y_n\right)\cdot \nabla\left(\tanh\left(w^T x_n\right) - y_n\right)\Big] \quad \text{——(1)}$$

Let $g(x) = \tanh(x)$

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

By rule of quotient,

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

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

$$\Rightarrow \nabla g(x) = 1 - (\tanh x)^2 \quad \text{——(2)}$$

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

$$= \frac{2}{N} \sum_{n=1}^{N} \left(\tanh(w^T x_n) - y_n \right) \cdot \left(1 - \tanh(w^T x_n)^2 \right) \cdot \frac{d}{dw}(w^T x_n)$$

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

As $w \to \infty$,

$$\tanh(w^T x_n)^2 \to 1$$

$$\Rightarrow \left(1 - \tanh(w^T x_n)^2 \right) \to 0$$

$\Rightarrow$ Gradient tends to become 0.

This results in vanishing gradients and weights will not change significantly.

Q-2

$$W^1 = \begin{bmatrix} 0.1 & 0.2 \\ 0.3 & 0.4 \end{bmatrix}$$

$$W^2 = \begin{bmatrix} 0.2 \\ 1 \\ -3 \end{bmatrix}$$

$$W^3 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Given $x^0 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

$$s^1 = (W^1)^T x^0$$

$$= \begin{bmatrix} 0.1 & 0.2 \\ 0.3 & 0.4 \end{bmatrix}^T \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$= \begin{bmatrix} 0.1 & 0.3 \\ 0.2 & 0.4 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 0.7 \\ 1 \end{bmatrix}$$

$$x^1 = \begin{bmatrix} \text{bias} \\ \tanh(s^1) \end{bmatrix}$$

$$= \begin{bmatrix} 1 \\ 0.60 \\ 0.76 \end{bmatrix}$$

$$s^2 = (W^2)^T (x^1)$$

$$= \begin{bmatrix} 0.2 \\ 1 \\ -3 \end{bmatrix}^T \begin{bmatrix} 1 \\ 0.60 \\ 0.76 \end{bmatrix}$$

$$= \begin{bmatrix} 0.2 & 1 & -3 \end{bmatrix} \begin{bmatrix} 1 \\ 0.6 \\ 0.76 \end{bmatrix} = \begin{bmatrix} -1.48 \end{bmatrix}$$

$$x^2 = \begin{bmatrix} \text{bias} \\ \tanh(s^2) \end{bmatrix}$$

$$= \begin{bmatrix} 1 \\ -0.90 \end{bmatrix}$$

$$s^3 = (W^3)^T (x^2)$$

$$= \begin{bmatrix} 1 \\ 2 \end{bmatrix}^T \begin{bmatrix} 1 \\ -0.90 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ -0.90 \end{bmatrix} = \begin{bmatrix} -0.80 \end{bmatrix}$$

$$x^3 = \begin{bmatrix} \text{Identity}(s^2) \end{bmatrix} = \begin{bmatrix} -0.80 \end{bmatrix}$$

Using back-propagation,

$$\delta^3 = 2\left(x^3 - y\right)\theta'(s^3)$$

$$\left\{\begin{array}{l}\theta(s) = s \\ \theta'(s) = 1\end{array}\right\}$$

$$\delta^3 = 2\left(-0.8 - 1\right)(1)$$
$$= -3.6.$$

$$\delta^2 = \theta'(s^2)\, w^3 \delta^3$$

$$\left\{\begin{array}{l}\theta(s) = \tanh(s) \\ \theta'(s) = 1 - (\tanh(s))^2 \\ \qquad = 1 - x^2\end{array}\right\}$$

$$\delta^2 = \left(1 - 0.9^2\right)(2)(-3.6)$$
$$= -1.368.$$

$$\delta^1 = \theta'(s^1)\, w^2 \delta^2$$

$$= \left(1 - \begin{bmatrix}0.6 \\ 0.76\end{bmatrix}^2\right)\left(\begin{bmatrix}1 \\ -3\end{bmatrix}\right)(-1.368)$$

$$= \begin{bmatrix}-0.875 \\ 1.736\end{bmatrix}$$

$$\frac{\partial e}{\partial w^1} = x^0 \cdot (\delta^1)^T$$

$$= \begin{bmatrix} 1 \\ 2 \end{bmatrix} \begin{bmatrix} -0.875 \\ 1.736 \end{bmatrix}^T$$

$$= \begin{bmatrix} -0.875 & 1.736 \\ -1.75 & 3.467 \end{bmatrix}$$

$$\frac{\partial e}{\partial w^2} = x^1 \cdot (\delta^2)^T$$

$$= \begin{bmatrix} 1 \\ 0.6 \\ 0.76 \end{bmatrix} \begin{bmatrix} -1.368 \end{bmatrix}^T$$

$$= \begin{bmatrix} -1.368 \\ -0.821 \\ -1.04 \end{bmatrix}$$

$$\frac{\partial e}{\partial w^3} = x^2 \cdot (\delta^3)^T$$

$$= \begin{bmatrix} 1 \\ -0.90 \end{bmatrix} \begin{bmatrix} -3.6 \end{bmatrix}^T$$

$$= \begin{bmatrix} -3.6 \\ 3.24 \end{bmatrix}$$

Q-3

(a) In a standard residual block,

i/p dimensions = o/p dimensions

= 128 x 16 x 16 x 32.

batch size; spatial size; # of channels

With Bias :

i Layer 1, No. of parameters :

$= ((32 \times 3 \times 3) + 1) \times 32$

(+1: bias)

$= 9248.$

ii Layer 2, No. of parameters :

$= ((32 \times 3 \times 3) + 1) \times 32.$

(+1: bias)

$= 9248.$

Total No. of parameters = 18496.

Without bias # of parameters:

(i) Layer 1:

$(32 \times 3 \times 3) \times 32 = 9216$

(ii) Layer 2:

$(32 \times 3 \times 3) \times 32 = 9216$

Total parameters = 18432.

(b) Bottleneck Layer:

I/p, o/p dimensions: $128 \times 16 \times 16 \times 128$.

batch size ↙ | spatial size ↘ | # of channels ↓

With bias, Number of parameters:

(i) Layer 1:

$((128 \times 1 \times 1) + 1) \times 32$

(+1 ↓ bias)

$= 4128.$

ii) Layer 2:

$((32 \times 3 \times 3) + 1) \times 32$

↓ bias.

$= 9248$

(iii) Layer 3:

$((32 \times 1 \times 1) + 1) \times 128$

↓ bias

$= 4224.$

Total parameters = 17600

Without Bias, Number of parameters:

(i) Layer 1.

$= (128 \times 1 \times 1) \times 32 = 4096$

(ii) Layer 2:

$= (32 \times 3 \times 3) \times 32 = 9216$

(iii) Layer 3

$= (32 \times 1 \times 1) \times 128 = 4096$

Total Parameters = 17408

(c) Bottleneck Layer:

# of channels = 128
# of parameters = 17600 (with bias)

Standard residual:

# of channels = 32
# of parameters = 18496 (with bias)

Advantage of bottleneck over standard residual:
- computationally less costly since we use less parameters
- we can have more channels with approximately same # of parameters.

Disadvantages:
- bottleneck design uses identity convolution. Hence we lose on benifits of bigger kernel size.

Q-4

(a)

Shape of x is $N \times C$

$\Rightarrow$ shape of mean $= 1 \times C$

$\Rightarrow$ shape of variance $= 1 \times C$

(b)

2-D convolution batch-norm :

As mentioned in the paper, we jointly normalize all the activations in a mini-batch, over all the locations.

Since, shape of $x = N \times H \times W \times C$

$\Rightarrow$ shape of mean $= 1 \times 1 \times 1 \times C$

$\Rightarrow$ shape of variance $= 1 \times 1 \times 1 \times C$

Q-5

(a) $$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \end{bmatrix} \in R^{2\times 4}$$

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

$$w_1^{ij} = \begin{bmatrix} w_1^{ij}, & w_2^{ij}, & w_3^{ij} \end{bmatrix}$$
$$i = 1,2, \quad j = 1,2.$$

where w^{ij} scans i-th channel of inputs and contributes to j-th channel of outputs.

$$y^{11} = w_1^{11} x_{11} + w_2^{11} x_{12} + w_3^{11} x_{13} + w_1^{21} x_{21} + w_2^{21} x_{21} + w_3^{21} x_{21}$$

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

$$y^{21} = w_1^{12} x_{11} + w_2^{12} x_{12} + w_3^{12} x_{13} + w_1^{22} x_{12} + w_2^{22} x_{13} + w_3^{22} x_{14}$$

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

$$\therefore \hat{Y} = \begin{bmatrix} w_1^{11} & w_2^{11} & w_3^{11} & 0 & w_1^{21} & w_2^{21} & w_3^{21} & 0 \\ 0 & w_1^{11} & w_2^{11} & w_3^{11} & 0 & w_1^{21} & w_2^{21} & w_3^{21} \\ w_1^{12} & w_2^{12} & w_3^{12} & 0 & w_1^{22} & w_2^{22} & w_3^{22} & 0 \\ 0 & w_1^{12} & w_2^{12} & w_3^{12} & 0 & w_1^{22} & w_2^{22} & w_3^{22} \end{bmatrix} \begin{bmatrix} x_{11} \\ x_{12} \\ x_{13} \\ x_{14} \\ x_{21} \\ x_{22} \\ x_{23} \\ x_{24} \end{bmatrix}$$

$$\Downarrow$$

$$A$$

(b)

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial Y} \times \frac{\partial Y}{\partial x}$$

$$\frac{\partial L}{\partial x_{11}} = w_1^{11} \frac{\partial L}{\partial Y_{11}} + w_1^{12} \frac{\partial L}{\partial Y_{21}}$$

$$\frac{\partial L}{\partial x_{12}} = w_2^{11} \frac{\partial L}{\partial Y_{11}} + w_1^{11} \frac{\partial L}{\partial Y_{12}} + w_2^{12} \frac{\partial L}{\partial Y_{21}} + w_1^{12} \frac{\partial L}{\partial Y_{22}}$$

$$\frac{\partial L}{\partial x_{13}} = w_3^{11} \frac{\partial L}{\partial Y_{11}} + w_2^{11} \frac{\partial L}{\partial Y_{12}} + w_3^{12} \frac{\partial L}{\partial Y_{21}} + w_2^{12} \frac{\partial L}{\partial Y_{22}}$$

$$\frac{\partial L}{\partial x_{14}} = w_3^{11} \frac{\partial L}{\partial Y_2} + w_3^{12} \frac{\partial L}{\partial Y_{22}}$$

Similarly, we can write equations for

$$\frac{\partial L}{\partial x_{21}}, \quad \frac{\partial L}{\partial x_{22}}, \quad \frac{\partial L}{\partial x_{23}}, \quad \frac{\partial L}{\partial x_{24}}.$$

$$\therefore \frac{\partial L}{\partial x} = \begin{bmatrix} w_1^{11} & 0 & w_1^{12} & 0 \\ w_2^{11} & w_1^{11} & w_2^{12} & w_1^{12} \\ w_3^{11} & w_2^{11} & w_3^{12} & w_2^{12} \\ 0 & w_3^{11} & 0 & w_3^{12} \\ w_1^{21} & 0 & w_1^{22} & 0 \\ w_2^{21} & w_1^{21} & w_2^{22} & w_1^{22} \\ w_3^{21} & w_2^{21} & w_3^{22} & w_2^{22} \\ 0 & w_3^{21} & 0 & w_3^{22} \end{bmatrix} \cdot \frac{\partial L}{\partial y}$$

$$\Downarrow$$

$$B.$$

$$\Rightarrow \boxed{B = A^T}.$$

(c)

Lets pad $\frac{\partial L}{\partial Y}$ on both sides

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{\partial L}{\partial y_{11}} & \frac{\partial L}{\partial y_{12}} & 0 \\ 0 & \frac{\partial L}{\partial y_{21}} & \frac{\partial L}{\partial y_{22}} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Let kernel.

$$W = \left[w_3^{ij}, w_2^{ij}, w_1^{ij} \right] \quad \begin{matrix} i=1,2 \\ j=1,2. \end{matrix}$$

where $i \rightarrow$ i/p channel scanned

$j \rightarrow$ o/p channel populated

$\Rightarrow$ after running convolution,

for first row first element in o/p,

$$= 0+0+w_1^{11}\frac{\partial L}{\partial y_{11}} + 0+0+w_1^{21}\frac{\partial L}{\partial y_{21}}$$

$$= w_1^{11}\frac{\partial L}{\partial y_{11}} + w_1^{21}\frac{\partial L}{\partial y_{21}}.$$

which is current $\frac{\partial L}{\partial x_{11}}$. Similarly we will get all $\frac{\partial L}{\partial X}$.

$\therefore$ So kernels with conv is possible $\left[w_3^{ij}, w_2^{ij}, w_1^{ij} \right]$ $\begin{matrix} i=1,2 \\ j=1,2. \end{matrix}$

(c)

Lets pad $\frac{\partial L}{\partial Y}$ on both sides

$$\begin{bmatrix} 0 & 0 & \partial L/\partial Y_{11} & \partial L/\partial Y_{12} & 0 & 0 \\ 0 & 0 & \partial L/\partial Y_{21} & \partial L/\partial Y_{22} & 0 & 0 \end{bmatrix}$$

Let kernel.

$$W = \left[W_3^{ij}, W_2^{ij}, W_1^{ij} \right] \quad \begin{matrix} i = 1,2 \\ j = 1,2. \end{matrix}$$

where $i \rightarrow$ i/p channel scanned

$j \rightarrow$ o/p channel populated

$\Rightarrow$ after running convolution,

for first row first element in o/p,

$$= 0+0+W_1^{11}\frac{\partial L}{\partial Y_{11}} + 0+0+W_1^{21}\frac{\partial L}{\partial Y_{21}}$$

$$= W_1^{11}\frac{\partial L}{\partial Y_{11}} + W_1^{21}\frac{\partial L}{\partial Y_{21}}.$$

which is current $\frac{\partial L}{\partial X_{11}}$.

$\therefore$ So conv is possible & kernels are with $\left[W_3^{ij}, W_2^{ij}, W_1^{ij} \right]$ $\begin{matrix} i = 1,2 \\ j = 1,2. \end{matrix}$