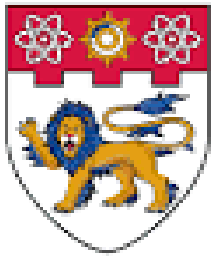

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Master in Science (Artificial Intelligence)

(MSAI)

AI-6126

Homework 1

Name of Student:
Teo Lim Fong

Matriculation No: G2101964G

Contents

Question 1

Output for Conv2d-1 $= \frac{N - F + 2P}{S} + 1$

-F = S(Output – 1 – 2P - N)

-F = 1(28 – 1 – 2(0) - 32)

-F = -5

F = 5

Output for Conv2d-4:

-F = 1(10 – 1 – 2(0) - 14)

F = 5

Output for Conv2d-7:

-F = 1(1 – 1 – 2(0) – 5)

F = 5

Layer (type)	Output Shape	Parameters
Conv2d-1	[-1, 6, 28 ,28]	(5x5 x1+1) x 6 = 156
ReLU-2	[-1, 6, 28 ,28]	0
MaxPool2d -3	[-1, 6, 14 ,14]	0
Conv2d-4	[-1, 16, 10, 10]	(5x5x6+1) x16 = 2416
ReLU-5	[-1, 16, 10, 10]	0
MaxPool2d -6	[-1, 16, 5, 5]	0
Conv2d-7	[-1, 120, 1, 1]	(5x5x16+1) x120 = 48 120
ReLU-8	[-1, 120, 1, 1]	0
Linear-9	[-1, 84]	(84 x 120) + 84 = 10164
ReLU-10	[-1, 84]	0
Linear- 11	[-1, 10]	(10 x 84) + 10 = 850
LogSoftmax-12	[-1, 10]	(10 x 10) + 10 = 110
Total Parameters	-	61 816

Question 2

Question 2: Let us consider the convolution of single-channel tensors $\mathbf{x} \in \mathbb{R}^{4\times4}$ and $\mathbf{w} \in \mathbb{R}^{3\times3}$

$$\boldsymbol{w} \star \boldsymbol{x} = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 10 & 10 & 0 & 0 \\ 10 & 10 & 0 & 0 \\ 10 & 10 & 0 & 0 \\ 10 & 10 & 0 & 0 \end{pmatrix}$$

Perform convolution as matrix multiplication by converting the kernel into sparse Toeplitz circulant matrix. Show your steps.

(5 marks)

Converting matrix $\boldsymbol{w}$ into sparse Toeplitz circulant matrix

$$\mathbf{W} = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix} =$$

$$\begin{pmatrix} -1 & 0 & 1 & 0 & -2 & 0 & 2 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & -2 & 0 & 2 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & -2 & 0 & 2 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & -2 & 0 & 2 & 0 & -1 & 0 & 1 \end{pmatrix}$$

$\mathbf{v}(\boldsymbol{x})$ = $(10 \quad 10 \quad 0 \quad 0 \quad 10 \quad 10 \quad 0 \quad 0 \quad 10 \quad 10 \quad 0 \quad 0 \quad 10 \quad 10 \quad 0 \quad 0)^T$

Wv(*x*) = (((-1 x 10) + 3(0) + (-2 x 10) + 3(0) + (-1 x 10) + 3(0) + 4 (0)) ((-1 x 10) + 3(0) + (-2 x 10) + 3(0) + (-1 x 10) + 3(0) + 4(0)) (4(0) + (-1 x10) + 3(0) + (-2 x10) + 3(0) + (-1x10) + 3(0)) (4(0) + (-1 x10) + 3(0) + (-2 x10) + 3(0) + (-1 x10) + 3(0)))

Wv(*x*) = (((-10) + (-20) + (-10) + 0) ((-10) + (-20) + (-10) + 0) (0 + (-10) + (-20) + -10) (0 + (-10) + (-20) + (-10)))

Wv(*x*) = $(-40 \quad -40 \quad -40 \quad -40)^T$

Question 3

Question 3: Many people in Singapore like to eat durian. Many customers believe that a perfectly oval and rounded durian is not always the best. An odd-shaped fruit that comes in slightly curved and crescent shape may taste better. You decide to train an image classifier to predict whether a durian is with rounded shape (label=0) or odd shape (label=1).

i) Mean square error (MSE) cost function is non-convex especially in binary classification. Thus, it is not guaranteed to minimize the loss function and is harder to coverage.

ii) BCE = $-\frac{\mathbf{1}}{N}\sum_{i}^{N} \boldsymbol{y}\, log\, \hat{y} + (1 - y) \log (1 - \hat{y})$

iii) $\boldsymbol{log_2(x)} = \frac{\boldsymbol{log_{10}(x)}}{\boldsymbol{log_{10}(2)}}$

$J = -\frac{1}{3}\sum_{i=2}^{3} y\, log\, \hat{y} + (1 - y) \log (1 - \hat{y})$

$= -\frac{1}{3}\,[\,(\sum_{i=1}^{3} 1\, log_2\, 0.2 + (1 - 1)\, log_2\, (1 - 0.2)) + (\sum_{i=2}^{3} 0\, log_2\, 0.5 + (1 - 0)\, log_2\, (1 - 0.5)) + (\sum_{i=3}^{3} 0\, log_2\, 0.1 + (1 - 0)\, log_2\, (1 - 0.1))$

$= -\frac{1}{3}\,[\,(\sum_{i=1}^{3} 1 . \frac{log_{10}(0.2)}{log_{10}(2)} + (1 - 1). \frac{log_{10}(1 - 0.2)}{log_{10}(2)}) + (\sum_{i=2}^{3} 0 . \frac{log_{10}(0.5)}{log_{10}(2)} + (1 - 0). \frac{log_{10}(1-0.5)}{log_{10}(2)})$

$+ (\sum_{i=3}^{3} 0 . \frac{log_{10}(0.1)}{log_{10}(2)} + (1 - 0) . \frac{log_{10}(1-0.1)}{log_{10}(2)})]$

$= -\frac{\mathbf{1}}{\mathbf{3}}\,[\,(\sum_{i=1}^{3} \cdot \frac{log_{10}(0.2)}{log_{10}(2)} + 0 + (\sum_{i=2}^{3} \mathbf{0} + \frac{log_{10}(1-0.5)}{log_{10}(2)}) + (\sum_{i=3}^{3} \mathbf{0} + 1\, \frac{log_{10}(1-0.1)}{log_{10}(2)})$

$= -\frac{\mathbf{1}}{\mathbf{3}}\,[-\mathbf{2.3219} + (-1) + (-0.152)]$

$= 1.15797$

iv) For model A (with L2 regularization):

The cost function is: $-\frac{\mathbf{1}}{N}\sum_{i}^{N} \boldsymbol{y}\, log\, \hat{y} + (1 - y) \log (1 - \hat{y}) + \frac{\lambda}{2m}\sum ||\boldsymbol{w}||^{\mathbf{2}}$

λ = regularization parameter

For model B:

The cost function is: $-\frac{\mathbf{1}}{N}\sum_{i}^{N} \boldsymbol{y}\, log\, \hat{y} + (1 - y) \log (1 - \hat{y})$

Normally when training with big datasets, the network will experience overfitting issue. Overfitting means that the weights are huge and cannot minimize the losses toward zero. For example, without regularization, the model will stop training when it stuck at some local minimum point. Adding additional weights will help the model to not only prevent the model from stop training but also give the model some weights to 'move out' of the local minimum

point to the global minimum point. Therefore, regularization is added to force the weight to decay toward zero and prevent overfitting.

Question 4

Before we start, we need to analysis the difference between L1 loss and L2 loss. L1 loss is given as $||w||_1 = \sum_1^n w$ and L2 loss is given as $||w||_2^2 = \sum_1^n w^2$. From the formula, we can see that L2 loss squared the error. If the error is larger than 1, then for L2 norm the error will be much larger compared to L1 norm losses. If the error is too large, it will be harder to minimize toward zero. This also mean that L2 norm loss will be sensitive to outliers. Therefore, L1 norm loss is more robust than L2 norm loss in term of minimizing the errors during training.

Question 5

1) The training duration will be very long if the datasets is very large.

2) A very small mini-batch size will not provide accurate statistics for batch normalization because a small proportion of the datasets cannot represent the whole datasets.

3) If you use a very small mini-batch size, the chances of getting positive and negative datasets will be unequal. Negative datasets normally refer to background images or images that are not related to the datasets to learn a larger distribution to reduce false positives.