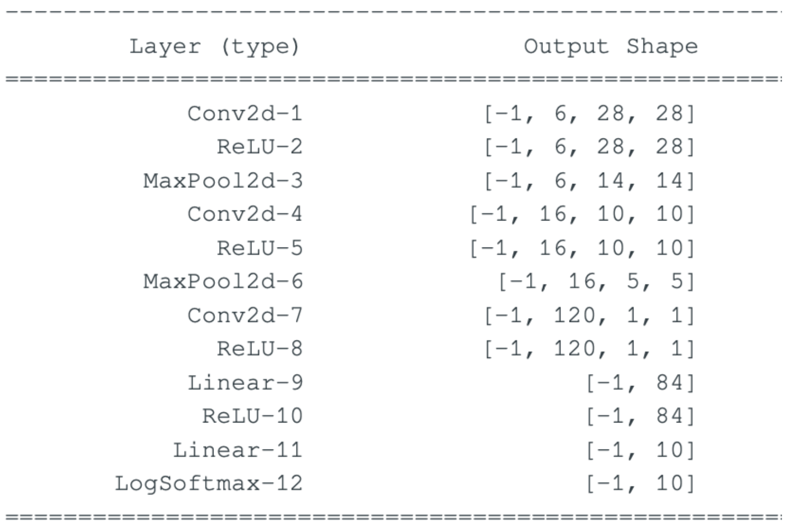
AI6126: Homework 1

Deadline: 12 February 2023 11:59PM

Question 1: A network with the type of each layer and the corresponding output shape is given as follows



```
----------------------------------------------------
        Layer (type)               Output Shape
====================================================
            Conv2d-1            [-1, 6, 28, 28]
              ReLU-2            [-1, 6, 28, 28]
         MaxPool2d-3            [-1, 6, 14, 14]
            Conv2d-4           [-1, 16, 10, 10]
              ReLU-5           [-1, 16, 10, 10]
         MaxPool2d-6             [-1, 16, 5, 5]
            Conv2d-7            [-1, 120, 1, 1]
              ReLU-8            [-1, 120, 1, 1]
            Linear-9                   [-1, 84]
             ReLU-10                   [-1, 84]
           Linear-11                   [-1, 10]
       LogSoftmax-12                   [-1, 10]
====================================================
```

The input has a shape of 1x32x32. The output shape of each layer is provided as [<ignore>, output channels, height, width]. For instance, at layer 'Conv2d-1', the output shape is [6, 28, 28], i.e., six feature maps of spatial size 28x28. Each conv filter and neuron of linear layer has a bias term and stride = 1. No padding is assumed.

Calculate the number of parameters for each layer and finally the total number of parameters of this network.

(6 marks)

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

$$\boldsymbol{w} \star \boldsymbol{x} = \begin{pmatrix} -1 & 0 & 1 \\ -3 & 0 & 3 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 4 & 7 & 0 & 0 \\ 5 & 8 & 0 & 0 \\ 4 & 8 & 0 & 0 \\ 5 & 7 & 0 & 0 \end{pmatrix}$$

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

(5 marks)

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) You've collected your own labeled dataset, chosen a neural network architecture, and are thinking about using the mean squared error (MSE) loss to optimize model parameters. Give one reason why MSE might not be a good choice for your loss function.

(1 mark)

ii) You decide to use the binary cross-entropy (BCE) loss to optimize your network. Write down the formula for this loss (for a single example) in terms of the label y and prediction $\hat{y}$.

(1 mark)

iii) Compute the total cost, J, of the network averaged across the following dataset of three examples using the binary cross entropy loss. $Y = (1, 0, 0)^{\mathsf{T}}$, and $\hat{Y} = (0.3, 0.4, 0.1)^{\mathsf{T}}$. There is no penalty on the weights.

(2 mark)

iv) You decide to train one model with L2 regularization (model A) and one without (model B). How would you expect model A's weights to compare to model B's weights?

(1 mark)

Question 4: Give one reason of using mini-batch gradient descent over batch gradient descent.

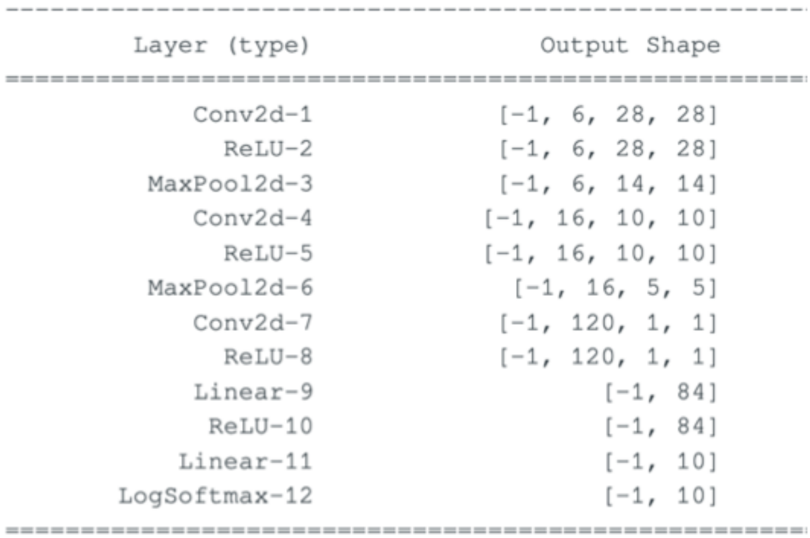
(2 marks)

Question 5: You want to apply batch normalization in your network. Explain why you shouldn't choose a very small mini-batch size during your training.

(2 marks)

(END)

Question 1: A network with the type of each layer and the corresponding output shape is given as follows



```
-------------------------------------------------------
        Layer (type)               Output Shape
=======================================================
            Conv2d-1            [-1, 6, 28, 28]
              ReLU-2            [-1, 6, 28, 28]
         MaxPool2d-3            [-1, 6, 14, 14]
            Conv2d-4           [-1, 16, 10, 10]
              ReLU-5           [-1, 16, 10, 10]
         MaxPool2d-6             [-1, 16, 5, 5]
            Conv2d-7            [-1, 120, 1, 1]
              ReLU-8            [-1, 120, 1, 1]
            Linear-9                   [-1, 84]
             ReLU-10                   [-1, 84]
           Linear-11                   [-1, 10]
       LogSoftmax-12                   [-1, 10]
=======================================================
```

The input has a shape of 1x32x32. The output shape of each layer is provided as [<ignore>, output channels, height, width]. For instance, at layer 'Conv2d-1', the output shape is [6, 28, 28], i.e., six feature maps of spatial size 28x28. Each conv filter and neuron of linear layer has a bias term and stride = 1. No padding is assumed.

Calculate the number of parameters for each layer and finally the total number of parameters of this network.

(6 marks)

$W_2 = \frac{W_1 - F + 2P}{S} + 1$ $\quad H_2 = \frac{H_1 - F + 2P}{S} + 1$ $\quad \therefore F = 5$

$28 = \frac{32 - F + 2(0)}{1} + 1$ $\quad 28 = \frac{32 - F + 2(0)}{1} + 1$

$D_1 = 1, D_2 = 6$ $\quad \therefore K = 6$

Layer 1 no. of params $= (F \cdot F \cdot D_1 + 1) \cdot K$

$= (5 \times 5 \times 1 + 1) \times 6$

$= 156$

Layer 2 no. of params $= 0$ [ReLU has no params]

Layer 3 no. of params $= 0$ [Pooling has no params]

$W_4 = \frac{W_3 - F + 2P}{S} + 1$ $\quad \therefore F = 5$

$10 = \frac{14 - F + 2(0)}{1} + 1$ $\quad D_3 = 6, D_4 = 16$ $\quad \therefore K = 16$

Layer 4 no. of params $= (F \cdot F \cdot D_3 + 1) \cdot K$

$= (5 \times 5 \times 6 + 1) \times 16$

$= 2416$

Layer 5 no. of params $= 0$

Layer 6 no. of params $= 0$

$W_7 = \frac{W_6 - F + 2P}{S} + 1$ $\quad \therefore F = 5$

$1 = \frac{5 - F + 2(0)}{1} + 1$ $\quad D_6 = 16, D_7 = 120$ $\quad \therefore K = 120$

Layer 7 no. of params $= (F.F.D_6 + 1) \cdot K$

$= (5 \times 5 \times 16 + 1) \times 120$

$= 48120$

Layer 8 no. of params $= 0$

Layer 9 no. of params $= (W_8 H_8 D_8 + 1) D_9$

$= (1 \times 1 \times 120 + 1) \times 84$

$= 10164$

Layer 10 no. of params $= 0$

Layer 11 no. of params $= (D_{10} + 1) D_{11}$

$= (84 + 1) \times 10$

$= 850$

Layer 12 no. of params $= 0$

Total no. of params $= 156 + 2416 + 48120 + 10164 + 850$

$= 61706$

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

$$\boldsymbol{w} \star \boldsymbol{x} = \begin{pmatrix} -1 & 0 & 1 \\ -3 & 0 & 3 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 4 & 7 & 0 & 0 \\ 5 & 8 & 0 & 0 \\ 4 & 8 & 0 & 0 \\ 5 & 7 & 0 & 0 \end{pmatrix}$$

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

(5 marks)

$$W = \begin{pmatrix} -1 & 0 & 1 & 0 & -3 & 0 & 3 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & -3 & 0 & 3 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & -3 & 0 & 3 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & -3 & 0 & 3 & 0 & -1 & 0 & 1 \end{pmatrix}$$

$$v(x) = (4 \quad 7 \quad 0 \quad 0 \quad 5 \quad 8 \quad 0 \quad 0 \quad 4 \quad 8 \quad 0 \quad 0 \quad 5 \quad 7 \quad 0 \quad 0)^T$$

$u(1,1) = -1(4)+0(7)+1(0)+0(0)-3(5)+0(8)+3(0)+0(0)-1(4)+0(8)+1(0)+0(0)+0(5)+0(7)+0(0)+0(0)$

$= -23$

$u(1,2) = 0(4)-1(7)+0(0)+1(0)+0(5)-3(8)+0(0)+3(0)+0(4)-1(8)+0(0)+1(0)+0(5)+0(7)+0(0)+0(0)$

$= -39$

$u(2,1) = 0(4)+0(7)+0(0)+0(0)-1(5)+0(8)+1(0)-0(0)-3(4)+0(8)+3(0)-0(0)-1(5)+0(7)+1(0)+0(0)$

$= -22$

$u(2,2) = 0(4)+0(7)+0(0)+0(0)+0(5)-1(8)+0(0)+1(0)+0(4)-3(8)+0(0)+3(0)+0(5)-1(7)+0(0)+1(0)$

$= -39$

$\therefore W_v(x) = (-23 \quad -39 \quad -22 \quad -39)^T$

$$w \ast x = \underline{\begin{pmatrix} -23 & -39 \\ -22 & -39 \end{pmatrix}}$$

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) You've collected your own labeled dataset, chosen a neural network architecture, and are thinking about using the mean squared error (MSE) loss to optimize model parameters. Give one reason why MSE might not be a good choice for your loss function.

(1 mark)

ii) You decide to use the binary cross-entropy (BCE) loss to optimize your network. Write down the formula for this loss (for a single example) in terms of the label y and prediction $\hat{y}$.

(1 mark)

iii) Compute the total cost, J, of the network averaged across the following dataset of three examples using the binary cross entropy loss. $Y = (1, 0, 0)^{\mathsf{T}}$, and $\hat{Y} = (0.3, 0.4, 0.1)^{\mathsf{T}}$. There is no penalty on the weights.

(2 mark)

iv) You decide to train one model with L2 regularization (model A) and one without (model B). How would you expect model A's weights to compare to model B's weights?

(1 mark)

i) MSE is not suitable for a binary classification problem as it calculates the difference between the actual and predicted class without taking false positives or false negatives into consideration

ii) $L(\hat{y}, y) = -y \log(\hat{y}) - (1-y)\log(1-\hat{y})$

iii) $J = \frac{1}{3}\sum_{i=1}^{3}\left[-y_i \log(\hat{y}_i) - (1-y_i)\log(1-\hat{y}_i)\right.$

$= \frac{-1}{3}\left[(1)\log(0.3) + (1-1)\log(1-0.3) + (0)\log(0.4) + (1-0)\log(1-0.4) + (0)\log(0.1) + (1-0)\log(1-0.1)\right]$

$= 0.2635$ (4 d.p.)

iv) Adding L2 regularization prevents overfitting by reducing the weights, thus Model A's weights are expected to be smaller than those of Model B's

Question 4: Give one reason of using mini-batch gradient descent over batch gradient descent.

(2 marks)

Mini-batch gradient descent updates the parameters more often as compared to batch gradient descent, thus allowing faster convergence, especially on large datasets.

Question 5: You want to apply batch normalization in your network. Explain why you shouldn't choose a very small mini-batch size during your training.

(2 marks)

A very small mini-batch would lead to large variations in the mean and variance when applying batch normalization, resulting in poor generalization and slow convergence.