# AI6126: Homework 1

Deadline: 31 August 2020 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.

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

**(6 marks)**

**Answer**:

```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Conv2d-1            [-1, 6, 28, 28]             156
              ReLU-2            [-1, 6, 28, 28]               0
         MaxPool2d-3            [-1, 6, 14, 14]               0
            Conv2d-4           [-1, 16, 10, 10]           2,416
              ReLU-5           [-1, 16, 10, 10]               0
         MaxPool2d-6             [-1, 16, 5, 5]               0
            Conv2d-7            [-1, 120, 1, 1]          48,120
              ReLU-8            [-1, 120, 1, 1]               0
            Linear-9                  [-1, 84]          10,164
             ReLU-10                  [-1, 84]               0
           Linear-11                  [-1, 10]             850
       LogSoftmax-12                  [-1, 10]               0
================================================================
Total params: 61,706
Trainable params: 61,706
Non-trainable params: 0
```

Conv2d-1: (5x5x1+1)*6 = 156 (1 mark)
Conv2d-4: (5x5x6+1)*16 = 2416 (1 mark)
Conv2d-7: (5x5x16+1)*120 = 48120 (1 mark)
Linear-9: 120*84+84 = 10164 (1 mark)
Linear-11: 84*10+10 = 850 (1 mark)
Total parameters = 61,706 (1 mark)

**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)**

**Answer**:

We first convert the kernel into a sparse Toeplitz circulant matrix

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

(2 marks for the correct matrix)

The input $\boldsymbol{x}$ is flattened

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

(1 mark for flattening the input)

Then,

$$\boldsymbol{W}v(\boldsymbol{x}) = (-40 \quad -40 \quad -40 \quad -40)^{\mathrm{T}}$$

(1 mark for getting result)

which we can reshape to a 2x2 matrix to obtain the final convolution result.

(1 mark for reshaping)

**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.

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 $\hat{y}$.

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 $\hat{Y} = (0.2, 0.5, 0.1)^{\mathsf{T}}$. There is no penalty on the weights.

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?

**Answer**:

i) When performing binary classification, the outputs are constrained from 0 to 1. When using MSE, we place an upper bound of 1 on the loss, when intuitively it should be infinity (you are being incorrect as possible). BCE does this, and would be a more natural choice.

(1 mark)

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

(1 mark)

iii) J = −1/3 (log 0.2 + log 0.5 + log 0.9).

(2 marks, the final result is optional)

iv) The weights of model A will generally be smaller in magnitude than those of model B.

(1 mark)

**Question 4:** Why might we prefer to minimize the sum of absolute residuals (L1 loss) instead of the residual sum of squares for some data sets (L2 loss)? (*Hint*: What is one of the flaws of least-squares regression?)

**Answer**: The sum of absolute residuals is less sensitive to outliers than the residual sum of squares.

(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.

**Answer**: Normalizing over a batch doesn't make sense when you just have a small number of examples per batch. Batch normalization at train time will get messed up because your mean/variance estimates will be super noisy (using just a few samples we are trying to estimate the true population/distribution).

(2 marks)