Vinh Thinh Ho (2562630)
Noshaba Cheema (2562653)

Exercise 5.1

a)
With $ReLU(z) = max(0, z)$:

$$\frac{\partial ReLU(z)}{\partial z} = \begin{cases} \frac{\partial 0}{\partial z} = 0 \ when \ z \leq 0 \\ \frac{\partial z}{\partial z} = 1 \ when \ z \geq 0 \end{cases}$$

With sigmoid function $g(z) = \frac{1}{1+e^{-z}}$:

$$\begin{aligned} \frac{\partial g(z)}{z} &= \frac{-1}{(1+e^{-z})^2}.\frac{\partial(1+e^{-z})}{\partial z} \\ &= \frac{e^{-z}}{(1+e^{-z})^2} = \frac{1}{1+e^{-z}}.\frac{e^{-z}}{1+e^{-z}} \\ &= g(z).(1-g(z)) = g(z).g(-z) \end{aligned}$$

b) We can see that the value of $g'(z)$ can be derived from $g(z)$. Hence, for a given layer that use sigmoid function as activation function, we can just store the value $g(z)$, then the gradient $g'(z)$ of that layer could be evaluated directly from $g(z)$ by simple multiplication and subtraction.

Exercise 5.2

See **nn.py**
We use sigmoid activation function in the hidden layer only, and linear activation in output layer. In addition the dimension for hidden layer is 256. We trained the model with 30 epochs and learning rate 0.05
Below is the graph of MSE loss function after each epoch and accuracy after each epoch (training process uses MNIST train data and accuracy is evaluated basing on MNIST test data)

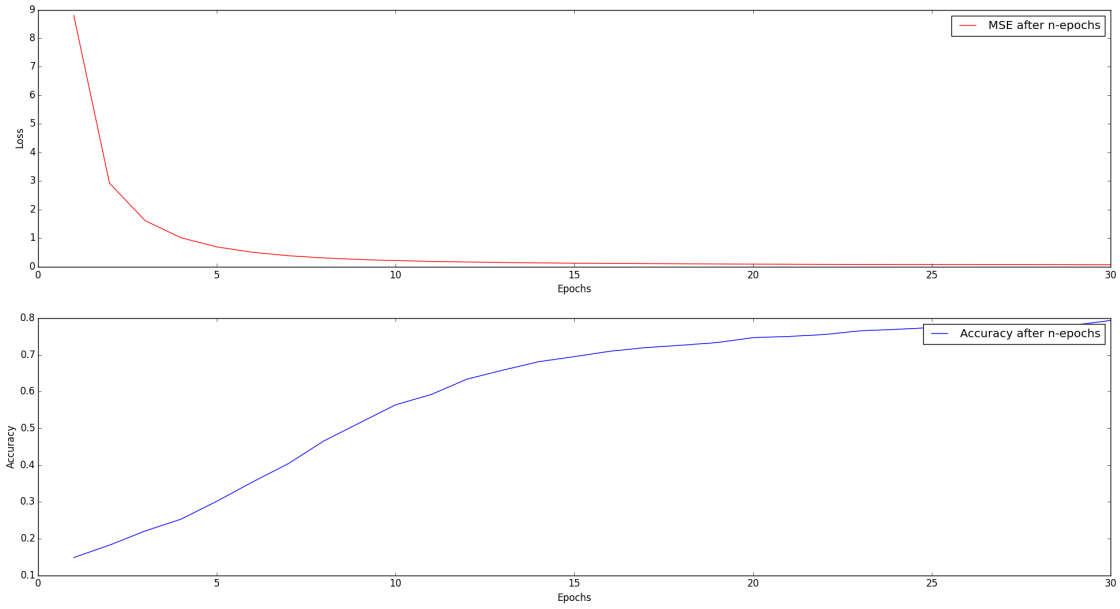

Figure 1: Classification Result