This document explains the structure and the forward propogation process of the neural nets used in Programming Assignment 3. It also provides the gradient computation of the crossentropy loss w.r.t. the input image.

1 Structure and Forward Propogation

The neural net is a fully-connected multi-layer perceptron with three hidden layers. The hidden layers contains 2048, 512 and 512 hidden nodes respectively. We use ReLU as the activation function at each hidden node. The last intermediate layer's output is passed through a softmax function, and the loss is measured as the cross-entropy between the resulted probability vector and the true label.

We use the following notations.

1. x: the input image vector with dimension 1x784.
2. y: the true class label of x.
3. z^i: the value of the i-th intermediate layer *before* activation, with dimension 1x2048, 1x512, 1x512 and 1x10 for $i = 1, 2, 3, 4$.
4. h^i: the value of the i-th intermediate layer *after* activation, with dimension 1x2048, 1x512 and 1x512 for $i = 1, 2, 3$.
5. p: the predicted class probability vector after the softmax function, with dimension 1x10.
6. W^i: the weights between the $(i-1)$-th and the i-th intermediate layer. For simplicity, we use h^0 as an alias to x. Each W^i has dimension l_{i-1}xl_i, where l_i is the number of nodes in the i-th layer. For example, W^1 has dimension 784x2048.
7. b^i: the bias between the $(i-1)$th and the i-th intermediate layer. The dimension is 1xl_i.

The forward propagation rules are as follows.

$$z^i = h^{i-1}W^i + b^i \quad \text{for} \quad i = 1, 2, 3, 4 \tag{1}$$

$$h^i = ReLU(z^i) \quad \text{for} \quad i = 1, 2, 3 \tag{2}$$

$$p = Softmax(z^4) \tag{3}$$

2 Gradient Calculation

Let L denote the cross entropy loss of an image-label pair (x, y). We are interested in the gradient of L w.r.t. x, and move x in the direction of (the sign of) the gradient to increase L. If L becomes large, the new image x_{adv} will likely be misclassified.

We use chain rule for gradient computation. Again, let h^0 be the alias of x. We have

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial h^0} = \frac{\partial L}{\partial z^4}\frac{\partial z^4}{\partial h^3}\prod_{i=1}^{3}\frac{\partial h^i}{\partial h^{i-1}} = \frac{\partial L}{\partial z^4}\frac{\partial z^4}{\partial h^3}\prod_{i=1}^{3}\left(\frac{\partial h^i}{\partial z^i}\frac{\partial z^i}{\partial h^{i-1}}\right). \tag{4}$$

The intermediate terms can be computed as follows.

$$\frac{\partial L}{\partial z^4} = p - y \tag{5}$$

$$\frac{\partial z^i}{\partial h^{i-1}} = (W^i)^T \tag{6}$$

$$\frac{\partial h^i}{\partial z^i} = diag(1(h^i > 0)). \tag{7}$$