HLCV Assignment 2

2576831 Maria Sargsyan,
2571618 Hui-Syuan Yeh
2576656 Shivam Sharma
Course-Semester

May 10, 2019

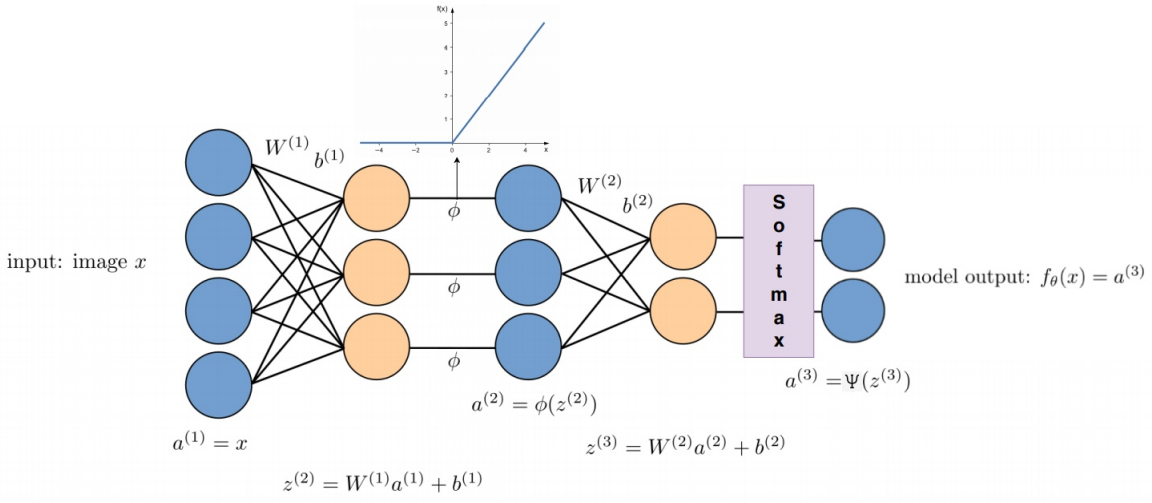


Figure 1: Visualisation of the two layer fully connected network, used in Q1-Q3

The equations from the ex2

$$\varphi(x) = max(0, x) \tag{1}$$
$$\psi(u)_i = \frac{\exp u_i}{\sum_{j=1}^{n} \exp u_j} \tag{2}$$
$$a^{(1)} = x \tag{3}$$
$$z^{(2)} = W^{(1)} a^{(1)} + b^{(1)} \tag{4}$$
$$a^{(2)} = \varphi(z^{(2)}) \tag{5}$$
$$z^{(3)} = W^{(2)} a^{(2)} + b^{(2)} \tag{6}$$
$$f_\theta(x) := a^{(3)} = \psi(z^(3)) \tag{7}$$
$$J(\theta, x_i, y_i) = -\log P(Y = y_i, X = x_i) \tag{8}$$
$$= -\log f_\theta(x_i)_{y_i} \tag{9}$$
$$= -\log \psi(z^{(3)})_{y_i} \tag{10}$$
$$J(\theta, x_i, y_i) = -\log \frac{\exp z^{(3)}_{y_i}}{\sum_{j=1}^{K} \exp z^{(3)}_{y_j}} \tag{11}$$
$$J(\theta, (x_i, y_i)_{i=1}^N) = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp z^{(3)}_{y_i}}{\sum_{j=1}^{K} \exp z^{(3)}_{y_j}} \tag{12}$$
$$\tilde{J(\theta)} = J(\theta, (x_i, y_i)_{i=1}^N) + \lambda(||W^{(1)}||_2^2 + ||W^{(2)}||_2^2) \tag{13}$$

Problem 2a. Verify that the loss function defined in Eq.12 has the gradient w.r.t z $z^{(3)}$ as below.

$$\frac{\partial J}{\partial z^{(3)}}((x_i, y_i)_{i=1}^N) = \frac{1}{N}(\varphi(z^{(3)}) - \triangle) \tag{14}$$

$$\triangle_{ij} = \begin{cases} 1 & if \quad y_i = j \\ 0 & else \end{cases}$$

Solution Let's first make the calculations for the case when $y_i \neq j$

$$\begin{aligned} \frac{\partial J}{\partial z_j^{(3)}} &= -\frac{1}{N} \frac{1}{a^{(3)}_{y_i}} \cdot -\frac{\exp z^{(3)}_{y_i} \cdot \exp z^{(3)}_j}{(\sum_{k=1}^{K} \exp z^{(3)}_{y_k})^2} \\ &= \frac{1}{N} \frac{1}{a^{(3)}_{y_i}} \cdot \frac{\exp z^{(3)}_{y_i}}{\sum_{k=1}^{K} \exp z^{(3)}_{y_k}} \cdot \frac{\exp z^{(3)}_j}{\sum_{k=1}^{K} \exp z^{(3)}_{y_k}} \\ &= \frac{1}{N} \frac{1}{a^{(3)}_{y_i}} \cdot a^{(3)}_{y_i} \cdot a^{(3)}_j = \frac{1}{N} a^{(3)}_j = \frac{1}{N} \psi^{(3)}_j \end{aligned}$$

If $y_i = j$

$$\frac{\partial J}{\partial z_j^{(3)}} = -\frac{1}{N}\frac{1}{a_{y_i}^{(3)}} \cdot \Big(a_{y_i}^{(3)} - \Big(\frac{\exp z_{y_i}^{(3)}}{\sum_{k=1}^{K} \exp z_{y_k}^{(3)}}\Big)^2\Big)$$
$$= -\frac{1}{N}\frac{1}{a_{y_i}^{(3)}} \cdot \big(a_{y_i}^{(3)} - (a_{y_i}^{(3)})^2\big) = \frac{1}{N} \cdot \big(a_{y_i}^{(3)} - 1\big)$$

Problem 2b. To compute the effect of the weight matrix W_2 on the loss in Eq.12 incurred by the network, we compute the partial derivatives of the loss function with respect to W_2. This is done by applying the chain rule. Verify that the partial derivative of the loss w.r.t W_2 is :

$$\frac{\partial J}{\partial W^{(2)}}((x_i, y_i)_{i=1}^N) = \frac{\partial J}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial W^{(2)}} \tag{15}$$
$$= \frac{1}{N}(\psi(z^{(3)}) - \triangle)a^{(2)\prime} \tag{16}$$

Similarly verify that:

$$\frac{\partial \tilde{J}}{\partial W^{(2)}} = \frac{1}{N}(\psi(z^{(3)}) - \triangle)a^{(2)\prime} + 2\lambda W^{(2)} \tag{17}$$

Solution Let us calculate the derivatives for a single entry in the matrix W_2, namely w_{ij} . Also, note that $\frac{\partial J}{\partial z^{(3)}}$ and $\frac{\partial z^{(3)}}{\partial W_{ij}^{(2)}}$ have the same dimensions. Now

$$\frac{\partial z_{ek}^{(3)}}{\partial W_{ij}^{(2)}} = \begin{cases} a_{ej}^{(2)} & if \quad k = j \\ 0 & else \end{cases}$$

Similarly,

$$\frac{\partial\ \lambda||W^{(2)}||_2^2}{\partial W_{ij}^{(2)}} = 2\lambda W_{ij}^{(2)} \tag{18}$$

Now,

$$\frac{\partial J}{\partial W_{ij}^{(2)}}((x_k, y_k)_{k=1}^N) = \frac{\partial J}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial W_{ij}^{(2)}}$$
$$= \sum_{e=1}^{N}\sum_{c=1}^{H_2} \frac{\partial J}{\partial z_{ec}^{(3)}} \cdot \frac{\partial z_{ec}^{(3)}}{\partial W_{ij}^{(2)}} = \sum_{e=1}^{N} \frac{\partial J}{\partial z_{ei}^{(3)}} \cdot \frac{\partial z_{ei}^{(3)}}{\partial W_{ij}^{(2)}} = \sum_{e=1}^{N} \frac{\partial J}{\partial z_{ei}^{(3)}} \cdot a_{ej}^{(2)}$$

Thus,

$$\frac{\partial J}{\partial W_{ij}^{(2)}}((x_k, y_k)_{k=1}^N) = a^{(2)\intercal} \cdot \frac{\partial J}{\partial z^{(3)}} \tag{19}$$

Finally,

$$\frac{\partial \tilde{J}}{\partial W^{(2)}} = a^{(2)\intercal} \cdot \frac{1}{N}(\psi(z^{(3)}) - \triangle) + 2\lambda W^{(2)} \tag{20}$$

Problem 2c. Derive the expressions for the derivatives of the regularized loss in w.r.t other variables

Solution Let's calculate $\frac{\partial \tilde{J}}{\partial b^{(2)}}$ which is the same as $\frac{\partial J}{\partial b^{(2)}}$

$$\frac{\partial z_{ec}^{(3)}}{\partial b_j^{(2)}} = \begin{cases} 1 & if \quad c = j \\ 0 & else \end{cases}$$

$$\frac{\partial J}{\partial b_j^{(2)}}((x_k, y_k)_{k=1}^N) = \frac{\partial J}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial b_j^{(2)}}$$
$$= \sum_{e=1}^{N} \sum_{c=1}^{H_2} \frac{\partial J}{\partial z_{ec}^{(3)}} \cdot \frac{\partial z_{ec}^{(3)}}{\partial b_j^{(2)}} = \sum_{e=1}^{N} \frac{\partial J}{\partial z_{ej}^{(3)}} \cdot \frac{\partial z_{ej}^{(3)}}{\partial b_j^{(2)}} = \sum_{e=1}^{N} \frac{\partial J}{\partial z_{ej}^{(3)}}$$

Thus,

$$\frac{\partial J}{\partial b^{(2)}}((x_k, y_k)_{k=1}^N) = \mathbf{1}^\intercal \frac{\partial J}{\partial z^{(3)}}, \tag{21}$$

where $\mathbf{1} \in \mathbb{R}^N$ is a vector full of ones Now, to calculate the rest of the gradients we need to calculate first $\frac{\partial J}{\partial z^{(2)}}$ since

$$\frac{\partial J}{\partial b^{(1)}} = \frac{\partial J}{\partial z^{(2)}} \cdot \frac{\partial z^{(2)}}{\partial b^{(1)}}$$

$$\frac{\partial J}{\partial W^{(1)}} = \frac{\partial J}{\partial z^{(2)}} \cdot \frac{\partial z^{(2)}}{\partial W^{(1)}} \tag{22}$$

So,

$$\frac{\partial J}{\partial z^{(2)}} = \frac{\partial J}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial a^{(2)}} \cdot \frac{\partial a^{(2)}}{\partial z^{(2)}} \tag{23}$$

$$\frac{\partial J}{\partial a_{ij}^{(2)}} = \frac{\partial J}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial a_{ij}^{(2)}} = \sum_{e=1}^{N} \sum_{c=1}^{H_2} \frac{\partial J}{\partial z_{ec}^{(3)}} \cdot \frac{\partial z_{ec}^{(3)}}{\partial a_{ij}^{(2)}}$$
$$= \sum_{c=1}^{H_2} \frac{\partial J}{\partial z_{ic}^{(3)}} \cdot \frac{\partial z_{ic}^{(3)}}{\partial a_{ij}^{(2)}} = \sum_{c=1}^{H_2} \frac{\partial J}{\partial z_{ic}^{(3)}} \cdot W_{jc}^{(2)}$$

Thus

$$\frac{\partial J}{\partial a^{(2)}} = \frac{\partial J}{\partial z^{(3)}} \cdot W^{(2)\intercal} \tag{24}$$

$$\frac{\partial a_{ec}^{(2)}}{\partial z_{ij}^{(2)}} = \begin{cases} \phi'(z_{ij}^{(2)}) & if \quad e = i \ and \ c = j \\ 0 & else \end{cases}$$

$$\phi'(x) = \begin{cases} 1 & if \ x \geq 0 \\ 0 & else \end{cases}$$

Now,

$$\frac{\partial J}{\partial z^{(2)}{}_{ec}} = \frac{\partial J}{\partial a^{(2)}} \cdot \frac{\partial a^{(2)}}{\partial z_{ec}^{(2)}} = \frac{\partial J}{\partial a_{ec}^{(2)}} \cdot \frac{\partial a_{ec}^{(2)}}{\partial z_{ec}^{(2)}} \tag{25}$$

Which is just element-wise multiplication

Following similar arguments to 21

$$\frac{\partial J}{\partial b^{(1)}}((x_k, y_k)_{k=1}^N) = \mathbf{1}^\intercal \frac{\partial J}{\partial z^{(3)}} \cdot W^{(2)} \cdot \phi'(z^{(2)}) \tag{26}$$

$$\frac{\partial J}{\partial W^{(1)}}((x_k, y_k)_{k=1}^N) = a^{(1)\intercal}\Big(\frac{\partial J}{\partial z^{(3)}} \cdot W^{(2)} \cdot \phi'(z^{(2)})\Big) \tag{27}$$

Finally

$$\frac{\partial \tilde{J}}{\partial W^{(2)}} = \frac{\partial J}{\partial W^{(2)}} + 2\lambda W^{(2)} = a^{(2)\intercal} \cdot \frac{1}{N}(\psi(z^{(3)}) - \triangle) + 2\lambda W^{(2)} \tag{28}$$

$$\frac{\partial \tilde{J}}{\partial W^{(1)}} = \frac{\partial J}{\partial W^{(1)}} + 2\lambda W^{(1)} = a^{(1)\intercal} \cdot \Big(\frac{\partial J}{\partial z^{(3)}} \cdot W^{(2)} * \phi'(z^{(2)})\Big) + 2\lambda W^{(1)} \tag{29}$$

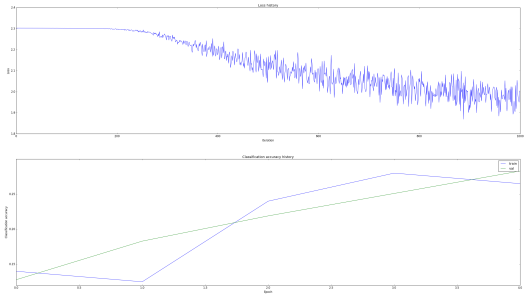
$$\frac{\partial \tilde{J}}{\partial b^{(2)}} = \frac{\partial J}{\partial b^{(2)}} = \mathbf{1}^\intercal \frac{\partial J}{\partial z^{(3)}} \tag{30}$$

$$\frac{\partial \tilde{J}}{\partial b^{(1)}} = \frac{\partial J}{\partial b^{(1)}} = \mathbf{1}^\intercal \frac{\partial J}{\partial z^{(3)}} \cdot W^{(2)} * \phi'(z^{(2)}) \tag{31}$$

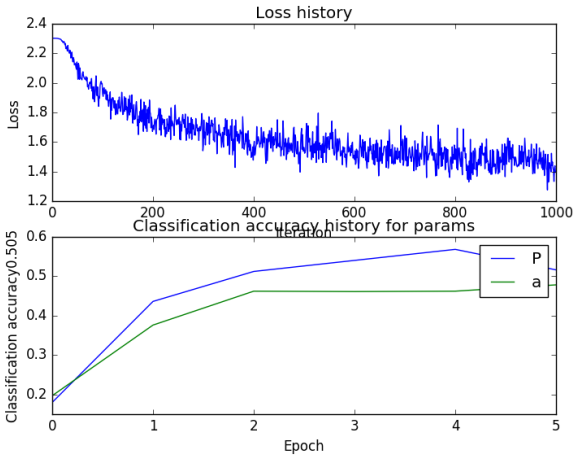
$$\frac{\partial J}{\partial z^{(3)}}((x_i, y_i)_{i=1}^N) = \frac{1}{N}(\phi(z^{(3)}) - \triangle) \tag{32}$$

Problem 3b. There are several pointers provided in the comments in the ex2 FCnet.py to help you understand why the network might be underperforming (Line 224-250). Once you have tuned your hyper parameters, and get validation accuracy greater than 48report the performance

As, it was suggested in the code. From the graph above, it is clearly seen the loss during the first iterations is decreasing very slowly. Actually, only by changing the learning rate from 0.0001 to 0.001, the accuracy increased up to 47 percents.



We could increase the ac curacies up to 0.505 percents by increasing the number of hidden layers to 80 and setting batch size to 250, learning decay to 0.99 and regularization to 0.12 and test accuracy is 0.483

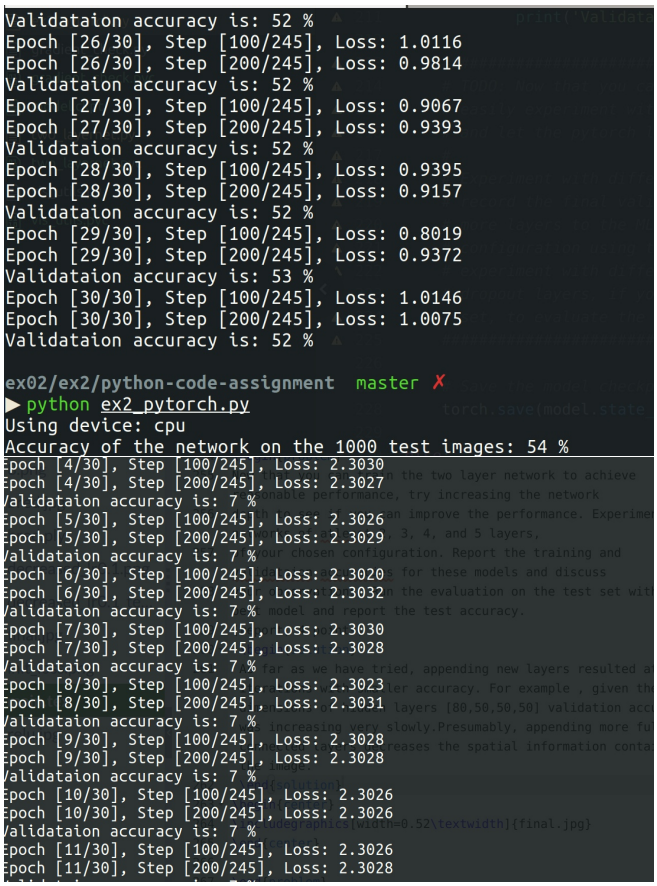



```
('Validation data shape: ', (1000, 3072))
('Validation labels shape: ', (1000,))
('Test data shape: ', (1000, 3072))
('Test labels shape: ', (1000,))
iteration 0 / 1000: loss 2.302849
iteration 100 / 1000: loss 1.926791
iteration 200 / 1000: loss 1.696625
iteration 300 / 1000: loss 1.576330
iteration 400 / 1000: loss 1.576058
iteration 500 / 1000: loss 1.562522
iteration 600 / 1000: loss 1.482224
iteration 700 / 1000: loss 1.642342
iteration 800 / 1000: loss 1.482232
iteration 900 / 1000: loss 1.558295
('Validation accuracy: ', 0.505)
('Test accuracy: ', 0.483)

ex02/ex2/python-code-assignment  master ✗
```

Problem N. ow that you can train the two layer network to achieve reasonable performance, try increasing the network depth to see if you can improve the performance. Experiment with networks of atleast 2, 3, 4, and 5 layers, of your chosen configuration. Report the training and validataion accuracies for these models and discuss your observations. Run the evaluation on the test set with your best model and report the test accuracy. (report, 4 points)

Solution As far as we have tried, models with several fully connected layers were not learning. For example , given the dimensions of hidden layers [80,50,50,50] validation accuracy was increasing very slowly.Presumably, appending more fully connected layers may decrease the spatial information contained in the image which is better captured in the first layer.



The first image refers to the final results. The second refers to the training the net with more than 2 layers