深度學習 報告

1. Introduction

用兩層的神經網路預測二維座標的 01 標籤

2. Experiment setups

A. Sigmoid functions

Activation function 使用 sigmoid，也就是 1/1+exp(-x)，他的導數是 sigmoid(x)*(1-sigmoid(x))

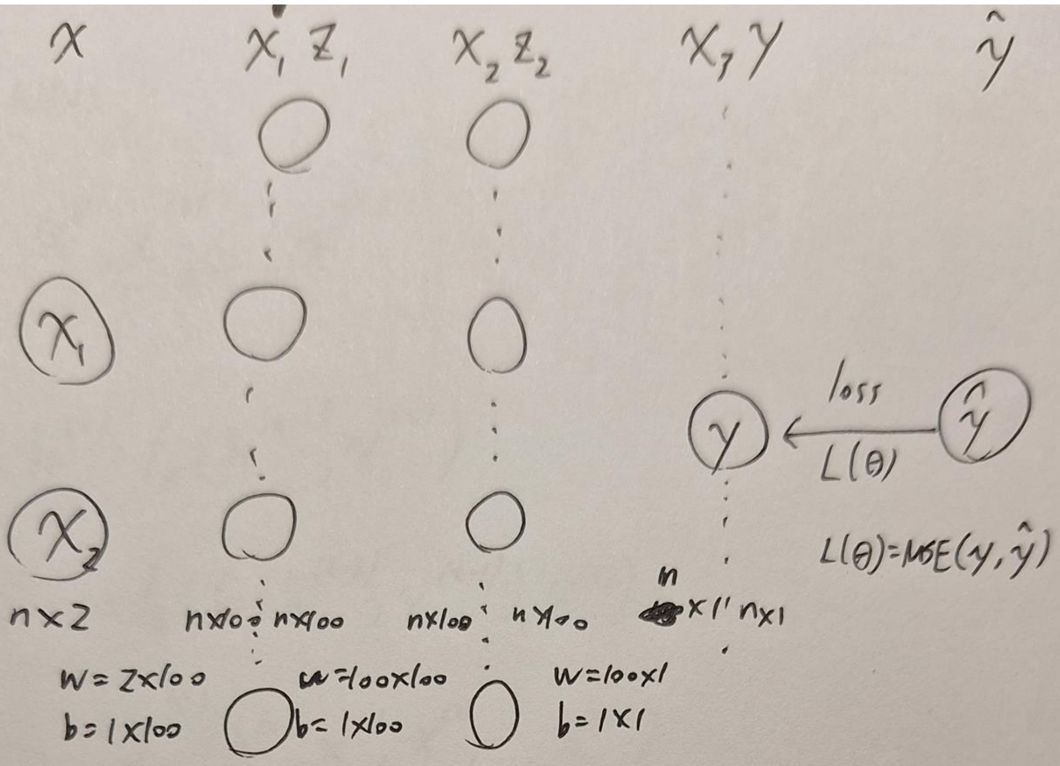
B. Neural network

網路結構使用兩層 hidden layer，每層 hidden layer 傳遞 100 維的向量，每層 layer 之間做線性變換加上偏差後通過 activation function，參數是線性變換的權重和偏差

C. Backpropagation

初始梯度是 loss 函數對輸出的導數，由後往前，每層 layer 都會貢獻一部份的梯度，使得前面 layer 的參數對 loss 造成得影響更難計算，因此 loss 對越前面的 layer 參數的導數越複雜

梯度計算方式如下:



$X_1 = XW_1 + b_1$ $\quad X_2 = Z_1 W_2 + b_2$ $\quad X_3 = Z_2 W_3 + b_3$

$Z_1 = \sigma(X_1)$ $\quad Z_2 = \sigma(X_2)$ $\quad y = \sigma(X_3)$

$Y = WX$ $\quad \nabla_X Z = W^T \nabla_Y Z$ $\quad \frac{\partial Y}{\partial X} = W^T$

$Y = XW$ $\quad \nabla_X Z = (\nabla_Y Z) W^T$ $\quad \frac{\partial Y^T}{\partial X^T} = W^T$

$Y = \sigma(X)$ $\quad (\nabla_X Z)[i,j] = (\nabla_Y Z)[i,j] \cdot \sigma'(X)[i,j]$ $\quad \sigma'(X) = \sigma(X)(1-\sigma(X)) \approx X(1-X)$

$L = MSE(y, \hat{y}) = \frac{1}{n} \sum (y[:,i] - \hat{y}[:,i])^2$ $\quad \frac{\partial L}{\partial y} = \frac{2}{n}(y - \hat{y})$

$$\overset{100\times 1}{\frac{\partial L(\theta)}{\partial W_3}} = \frac{\partial x_3}{\partial W_3}\frac{\partial y}{\partial x_3}\frac{\partial L(\theta)}{\partial y} = \overset{100\times n}{Z_2^{T}} \cdot \overset{n\times 1}{\sigma'(x_3) \odot \frac{2}{n}(y-\hat{y})}$$

$$\overset{1\times 1}{\frac{\partial L(\theta)}{\partial b_3}} = \frac{\partial x_3}{\partial b_3}\frac{\partial y}{\partial x_3}\frac{\partial L(\theta)}{\partial y} = \underset{\text{on axis}=0}{\text{sum}}\left(1 \cdot \overset{n\times 1}{\sigma'(x_3) \odot \frac{2}{n}(y-\hat{y})}\right)$$

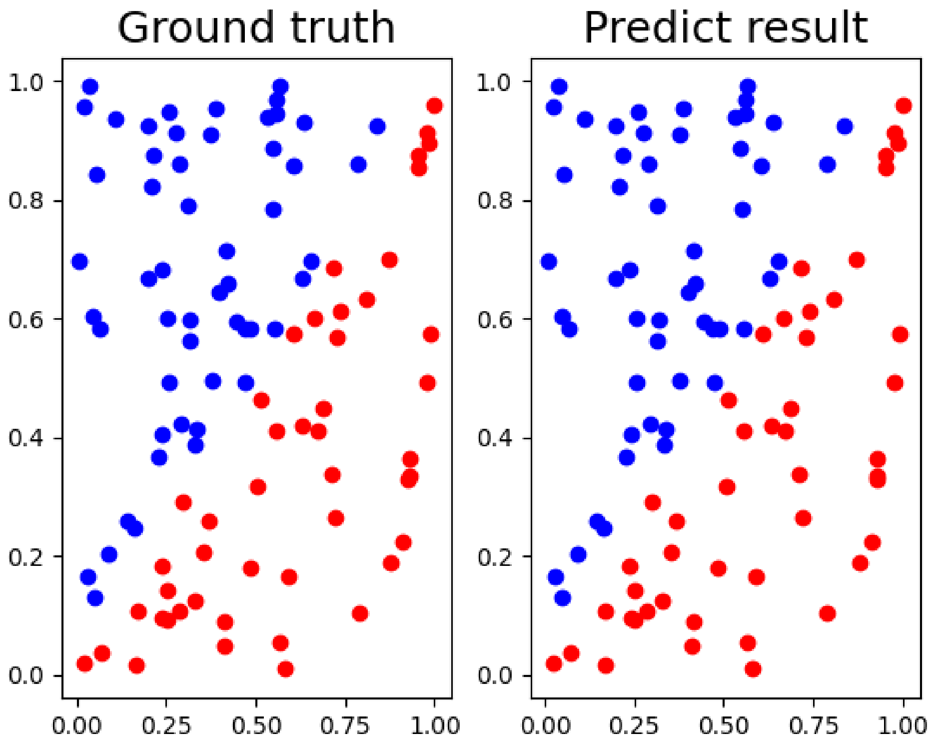
$$\overset{100\times 100}{\frac{\partial L(\theta)}{\partial W_2}} = \frac{\partial x_2}{\partial W_2}\frac{\partial Z_2}{\partial x_2}\frac{\partial L(\theta)}{\partial Z_2} = \overset{100\times n}{Z_1^{T}} \cdot \sigma'(x_2) \odot \overset{n\times 100}{\frac{\partial L(\theta)}{\partial Z_2}}$$

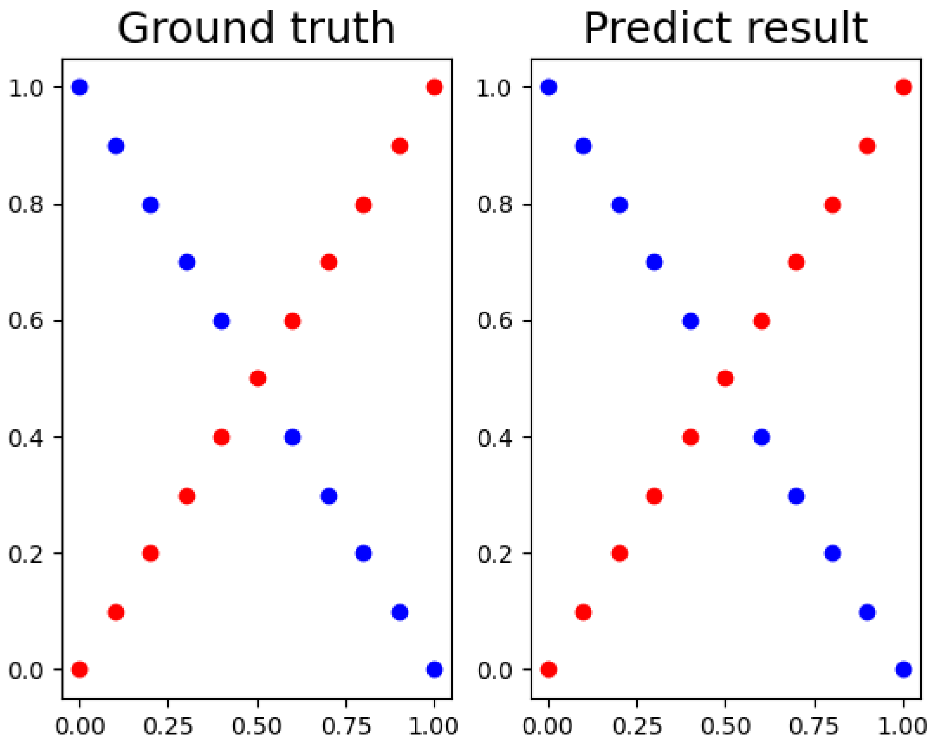
$$\overset{1\times 100}{\frac{\partial L(\theta)}{\partial b_2}} = \frac{\partial x_2}{\partial b_2}\frac{\partial Z_2}{\partial x_2}\frac{\partial L(\theta)}{\partial Z_2} = \underset{\text{on axis}=0}{\text{sum}}\left(1 \cdot \sigma'(x_2) \odot \overset{n\times 100}{\frac{\partial L(\theta)}{\partial Z_2}}\right)$$

$$\overset{n\times 100}{\frac{\partial L(\theta)}{\partial Z_2}} = \frac{\partial x_3}{\partial Z_2}\frac{\partial y}{\partial x_3}\frac{\partial L(\theta)}{\partial y} = \overset{n\times 1}{\left(\sigma'(x_3) \odot \frac{2}{n}(y-\hat{y})\right)} \overset{1\times 100}{W_3^{T}}$$

3. Results of your testing

A. Screenshot and comparison figure



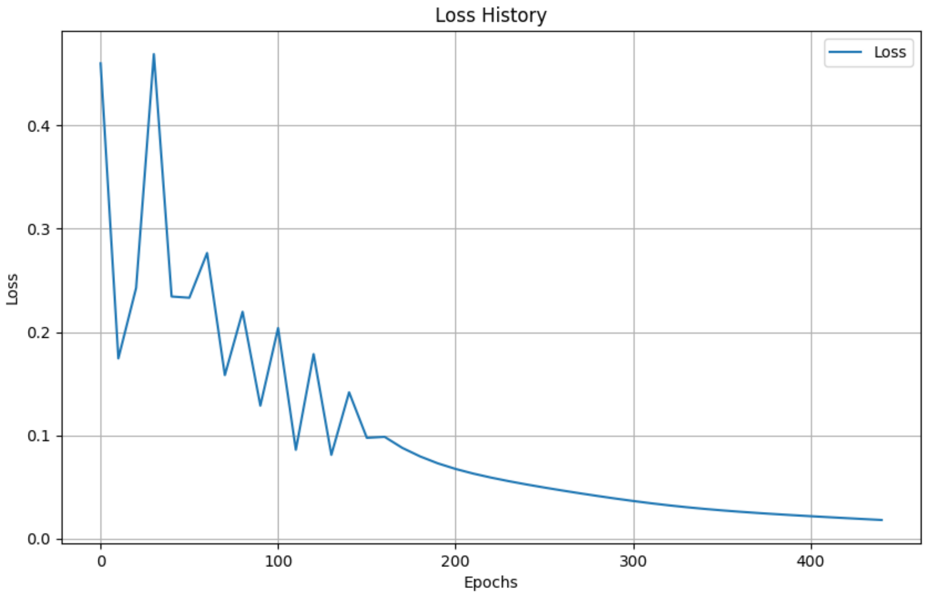


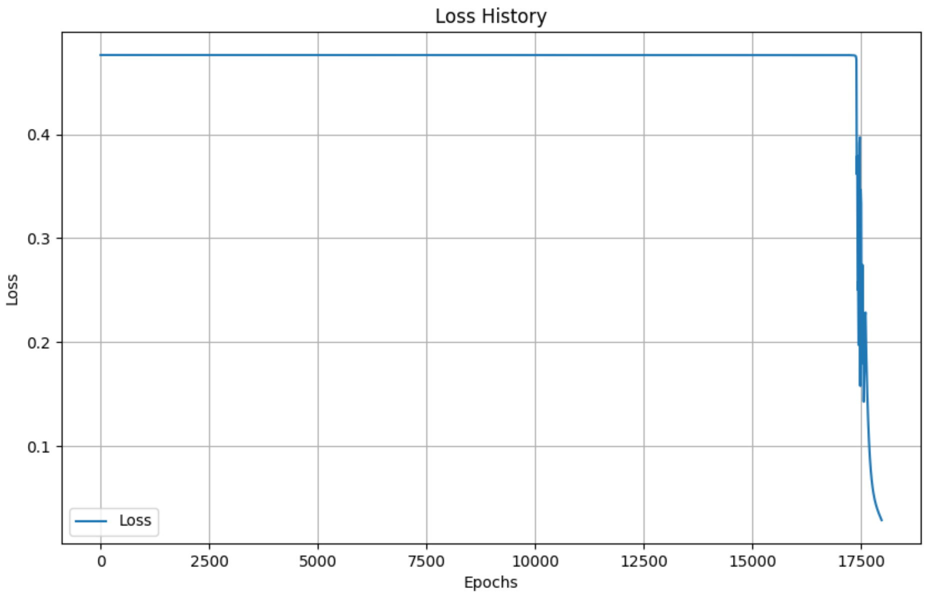
B. Show the accuracy of your predictions

```
Iter95 |         Ground truth: 0 |        prediction: 0.11986657256034744|
Iter96 |         Ground truth: 0 |        prediction: 6.183882565296563e-05|
Iter97 |         Ground truth: 1 |        prediction: 0.9998937812652718|
Iter98 |         Ground truth: 0 |        prediction: 0.06839439389740344|
Iter99 |         Ground truth: 0 |        prediction: 6.075906924579294e-06|
loss=0.017539261634822754 accuracy=100.0%
```

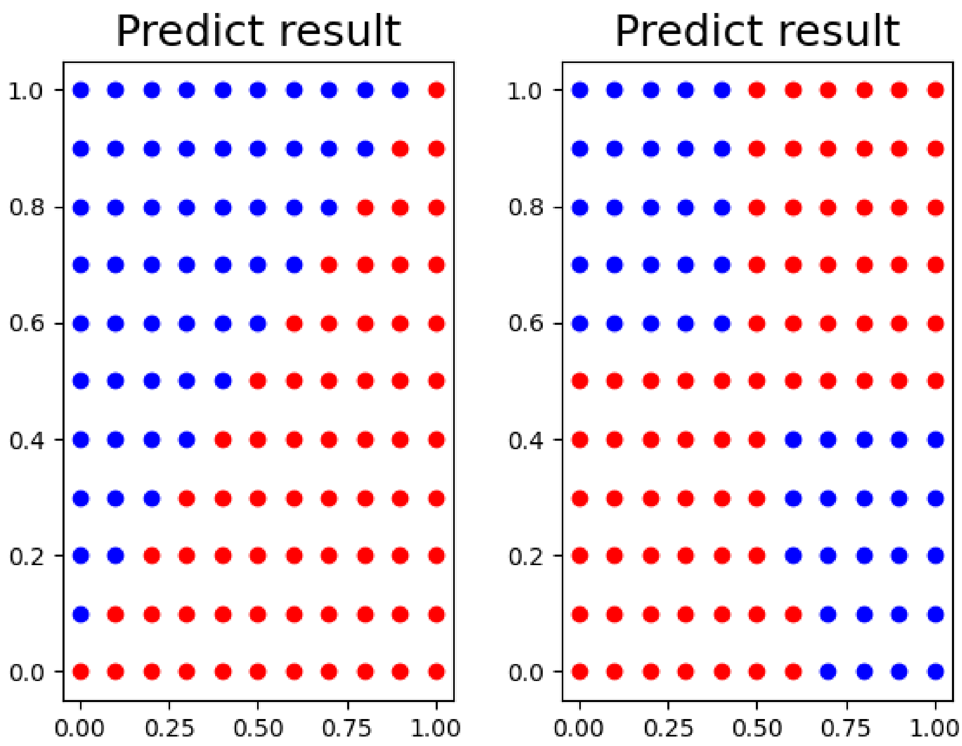
```
Iter16 |         Ground truth: 1 |        prediction: 0.9299231253024557|
Iter17 |         Ground truth: 0 |        prediction: 0.003692685329021936 7|
Iter18 |         Ground truth: 1 |        prediction: 0.9517528814786125|
Iter19 |         Ground truth: 0 |        prediction: 0.002598450095680138|
Iter20 |         Ground truth: 1 |        prediction: 0.959479395123099|
loss=0.028999113468982513 accuracy=100.0%
```

C. Learning curve (loss, epoch curve)





D. Anything you want to present



4. Discussion

A. Try different learning rates

Learning rate 越大，訓練越快，且經測試沒有 learning rate 太大的問題

B. Try different numbers of hidden units

hidden units 越少訓練越慢，遇到初始參數很差的狀況時，hidden units 越多訓練也越慢，100 做為 hidden units 的數量是個相當不錯的數字

C. Try without activation functions

沒有 activation function 限制數值的範圍，所有的變數無限制的膨脹，最後直接爆炸了，包括 loss 全部變成 nan 無法計算

D. Anything you want to share

講義上提供的 derivative_sigmoid 函數是 x*(1-x)，正確的公式是 sigmoid(x)*(1-sigmoid(x))，請出題方修正或加註如何正確使用，以免坑害未來學弟們

5. Extra

A. Implement different optimizers

未測試

B. Implement different activation functions

未測試

C. Implement convolutional layers

未測試