Math of NN

1. Consider a single layer NN (matrix multiplication) without nonlinearity:

$$z = wx + b$$

Or in matrix form

$$\begin{vmatrix} z_1 \\ z_2 \end{vmatrix} = \begin{vmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{vmatrix} \begin{vmatrix} x_1 \\ x_2 \end{vmatrix} + \begin{vmatrix} b_1 \\ b_2 \end{vmatrix}$$

With loss function

$$L = l\left(\begin{vmatrix} z_1 \\ z_2 \end{vmatrix}\right)$$

We are looking for $\frac{\partial L}{\partial w}$ in order to minimize the loss function using gradient decent

$$\frac{\partial L}{\partial w} = \begin{vmatrix} \dfrac{\partial L}{\partial w_{11}} & \dfrac{\partial L}{\partial w_{12}} \\ \dfrac{\partial L}{\partial w_{21}} & \dfrac{\partial L}{\partial w_{22}} \end{vmatrix} = \begin{vmatrix} \dfrac{\partial L}{\partial z_1}\dfrac{\partial z_1}{\partial w_{11}} & \dfrac{\partial L}{\partial z_1}\dfrac{\partial z_1}{\partial w_{12}} \\ \dfrac{\partial L}{\partial z_2}\dfrac{\partial z_2}{\partial w_{21}} & \dfrac{\partial L}{\partial z_2}\dfrac{\partial z_2}{\partial w_{22}} \end{vmatrix} = \begin{vmatrix} \dfrac{\partial L}{\partial z_1}x_1 & \dfrac{\partial L}{\partial z_1}x_2 \\ \dfrac{\partial L}{\partial z_2}x_1 & \dfrac{\partial L}{\partial z_2}x_2 \end{vmatrix} = \begin{vmatrix} \dfrac{\partial L}{\partial z_1} \\ \dfrac{\partial L}{\partial z_2} \end{vmatrix} \begin{vmatrix} x_1 & x_2 \end{vmatrix}$$

Similarly, we have

$$\frac{\partial L}{\partial x} = \begin{vmatrix} \dfrac{\partial L}{\partial x_1} \\ \dfrac{\partial L}{\partial x_2} \end{vmatrix} = \begin{vmatrix} \dfrac{\partial L}{\partial z_1}\dfrac{\partial z_1}{\partial x_1} + \dfrac{\partial L}{\partial z_2}\dfrac{\partial z_2}{\partial x_1} \\ \dfrac{\partial L}{\partial z_1}\dfrac{\partial z_1}{\partial x_2} + \dfrac{\partial L}{\partial z_2}\dfrac{\partial z_2}{\partial x_2} \end{vmatrix} = \begin{vmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \end{vmatrix} \begin{vmatrix} \dfrac{\partial L}{\partial z_1} \\ \dfrac{\partial L}{\partial z_2} \end{vmatrix}$$

and

$$\frac{\partial L}{\partial b} = \begin{vmatrix} \dfrac{\partial L}{\partial b_1} \\ \dfrac{\partial L}{\partial b_2} \end{vmatrix} = \begin{vmatrix} \dfrac{\partial L}{\partial z_1}\dfrac{\partial z_1}{\partial b_1} \\ \dfrac{\partial L}{\partial z_2}\dfrac{\partial z_2}{\partial b_2} \end{vmatrix} = \begin{vmatrix} \dfrac{\partial L}{\partial z_1} \\ \dfrac{\partial L}{\partial z_2} \end{vmatrix}$$

Denote $\begin{vmatrix} \dfrac{\partial L}{\partial z_1} \\ \dfrac{\partial L}{\partial z_2} \end{vmatrix}$ as δ (the backpropagated/remain error), then we have

$$\frac{\partial L}{\partial w} = \delta x^T$$

$$\frac{\partial L}{\partial x} = w^T \delta$$

$$\frac{\partial L}{\partial b} = \delta$$

δ can be calculated from nest layer/operation of the current layer. For example, if this is what happened in the loss function

$$\begin{vmatrix} c_1 \\ c_2 \end{vmatrix} = \begin{vmatrix} z_1 \\ z_2 \end{vmatrix} - \begin{vmatrix} y_1 \\ y_2 \end{vmatrix}$$

$$L = \frac{1}{2}\sum c^2$$

Then we can have

$$\boldsymbol{\delta} = \begin{vmatrix} \dfrac{\partial L}{\partial c_1} \dfrac{\partial c_1}{\partial z_1} \\ \dfrac{\partial L}{\partial c_2} \dfrac{\partial c_2}{\partial z_2} \end{vmatrix} = \begin{vmatrix} 1 \cdot c_1 \\ 1 \cdot c_2 \end{vmatrix} = \begin{vmatrix} z_1 \\ z_2 \end{vmatrix} - \begin{vmatrix} y_1 \\ y_2 \end{vmatrix}$$

In other word, the $\boldsymbol{\delta}$ of a layer is the partial derivative of its output (which is used as the input of the next layer) regarding the output of the next layer/operation

If we have nonlinearity

$$\begin{vmatrix} a_1 \\ a_2 \end{vmatrix} = f\left(\begin{vmatrix} z_1 \\ z_2 \end{vmatrix}\right)$$

In which f is the sigmoid function

$$f(x) = \frac{1}{1 + e^{-x}}$$

$$f'(x) = f(x)f(-x) = f(x)(1 - f(x))$$

Then the loss value is calculated by

$$L = l\left(\begin{vmatrix} a_1 \\ a_2 \end{vmatrix}\right)$$

In this case,

$$\boldsymbol{\delta} = \begin{vmatrix} \dfrac{\partial L}{\partial a_1} \dfrac{\partial a_1}{\partial z_1} \\ \dfrac{\partial L}{\partial a_2} \dfrac{\partial a_2}{\partial z_2} \end{vmatrix} = \begin{vmatrix} \dfrac{\partial L}{\partial a_1} \\ \dfrac{\partial L}{\partial a_2} \end{vmatrix} \times \begin{vmatrix} \dfrac{\partial a_1}{\partial z_1} \\ \dfrac{\partial a_2}{\partial z_2} \end{vmatrix} = \begin{vmatrix} \dfrac{\partial L}{\partial a_1} \\ \dfrac{\partial L}{\partial a_2} \end{vmatrix} \times \begin{vmatrix} a_1(1 - a_1) \\ a_2(1 - a_2) \end{vmatrix}$$

Where $\times$ denotes element-wise multiplication. Means that the gradients of element wise operations are multiplied to the backpropagated error

2. Consider a two-layer NN

$$\begin{vmatrix} z_1^{(l1)} \\ z_2^{(l1)} \end{vmatrix} = \begin{vmatrix} w_{11}^{(l1)} & w_{12}^{(l1)} \\ w_{21}^{(l1)} & w_{22}^{(l1)} \end{vmatrix} \begin{vmatrix} x_1 \\ x_2 \end{vmatrix} + \begin{vmatrix} b_1^{(l1)} \\ b_2^{(l1)} \end{vmatrix}$$

$$\begin{vmatrix} z_1^{(l2)} \\ z_2^{(l2)} \end{vmatrix} = \begin{vmatrix} w_{11}^{(l2)} & w_{12}^{(l2)} \\ w_{21}^{(l2)} & w_{22}^{(l2)} \end{vmatrix} \begin{vmatrix} z_1^{(l1)} \\ z_2^{(l1)} \end{vmatrix} + \begin{vmatrix} b_1^{(l2)} \\ b_2^{(l2)} \end{vmatrix}$$

With loss function

$$L = l\left(\begin{vmatrix} z_1^{(l2)} \\ z_2^{(l2)} \end{vmatrix}\right)$$

To get the gradient of layer 1 we need the backpropagated error

$$\boldsymbol{\delta}^{(l1)} = \frac{\partial L}{\partial \boldsymbol{z}^{(l1)}} = \begin{vmatrix} \dfrac{\partial L}{\partial z_1^{(l1)}} \\ \dfrac{\partial L}{\partial z_1^{(l1)}} \end{vmatrix}$$

Actually, this is the $\frac{\partial L}{\partial x}$ shown in the 1-layer case as $\boldsymbol{z}^{(l1)}$ is at the position of $\boldsymbol{x}$ of the 2nd layer. Therefore,

$$\boldsymbol{\delta}^{(l1)} = \boldsymbol{w}^{(l2)T} \boldsymbol{\delta}^{(l2)}$$

This conclusion can be extended to multi-layer NNs

$$\boldsymbol{\delta}^{(l_n)} = \boldsymbol{w}^{(l_{n+1})T} \boldsymbol{\delta}^{(l_{n+1})}$$

3. Consider a nested NN in which w and b are reused

$$\begin{vmatrix} z_1^{(l1)} \\ z_2^{(l1)} \end{vmatrix} = \begin{vmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{vmatrix} \begin{vmatrix} x_1 \\ x_2 \end{vmatrix} + \begin{vmatrix} b_1 \\ b_2 \end{vmatrix}$$

$$\begin{vmatrix} z_1^{(l2)} \\ z_2^{(l2)} \end{vmatrix} = \begin{vmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{vmatrix} \begin{vmatrix} z_1^{(l1)} \\ z_2^{(l1)} \end{vmatrix} + \begin{vmatrix} b_1 \\ b_2 \end{vmatrix}$$

With loss function

$$L = l\left(\begin{vmatrix} z_1^{(l2)} \\ z_2^{(l2)} \end{vmatrix} \right)$$

We are interested in $\frac{\partial L}{\partial w}$ and $\frac{\partial L}{\partial b}$

$$\frac{\partial L}{\partial \boldsymbol{w}} = \begin{vmatrix} \dfrac{\partial L}{\partial w_{11}} & \dfrac{\partial L}{\partial w_{12}} \\ \dfrac{\partial L}{\partial w_{21}} & \dfrac{\partial L}{\partial w_{22}} \end{vmatrix}$$

Unlike the case of 1-layer NN, here $\frac{\partial z_2^{(l2)}}{\partial w_{11}}$, $\frac{\partial z_2^{(l2)}}{\partial w_{12}}$, $\frac{\partial z_1^{(l2)}}{\partial w_{21}}$ and $\frac{\partial z_1^{(l2)}}{\partial w_{22}}$ are not zeros due $\boldsymbol{z}^{l1}$. Therefore

$$\begin{vmatrix} \dfrac{\partial L}{\partial w_{11}} & \dfrac{\partial L}{\partial w_{12}} \\ \dfrac{\partial L}{\partial w_{21}} & \dfrac{\partial L}{\partial w_{22}} \end{vmatrix} = \begin{vmatrix} \dfrac{\partial L}{\partial z_1^{(l2)}} \dfrac{\partial z_1^{(l2)}}{\partial w_{11}} + \dfrac{\partial L}{\partial z_2^{(l2)}} \dfrac{\partial z_2^{(l2)}}{\partial w_{11}} & \dfrac{\partial L}{\partial z_1^{(l2)}} \dfrac{\partial z_1^{(l2)}}{\partial w_{12}} + \dfrac{\partial L}{\partial z_2^{(l2)}} \dfrac{\partial z_2^{(l2)}}{\partial w_{12}} \\ \dfrac{\partial L}{\partial z_2^{(l2)}} \dfrac{\partial z_2^{(l2)}}{\partial w_{21}} + \dfrac{\partial L}{\partial z_1^{(l2)}} \dfrac{\partial z_1^{(l2)}}{\partial w_{21}} & \dfrac{\partial L}{\partial z_2^{(l2)}} \dfrac{\partial z_2^{(l2)}}{\partial w_{22}} + \dfrac{\partial L}{\partial z_1^{(l2)}} \dfrac{\partial z_1^{(l2)}}{\partial w_{22}} \end{vmatrix}$$

Which equals to

$$\begin{vmatrix} \dfrac{\partial L}{\partial z_1^{(l2)}} \left(z_1^{(l1)} + w_{11} \dfrac{\partial z_1^{(l1)}}{\partial w_{11}} \right) + \dfrac{\partial L}{\partial z_2^{(l2)}} \dfrac{\partial z_2^{(l2)}}{\partial z_1^{(l1)}} \dfrac{\partial z_1^{(l1)}}{\partial w_{11}} & \dfrac{\partial L}{\partial z_1^{(l2)}} \left(z_2^{(l1)} + w_{11} \dfrac{\partial z_1^{(l1)}}{\partial w_{12}} \right) + \dfrac{\partial L}{\partial z_2^{(l2)}} \dfrac{\partial z_2^{(l2)}}{\partial z_1^{(l1)}} \dfrac{\partial z_{21}^{(l1)}}{\partial w_{12}} \\ \dfrac{\partial L}{\partial z_2^{(l2)}} \left(z_1^{(l1)} + w_{22} \dfrac{\partial z_2^{(l1)}}{\partial w_{21}} \right) + \dfrac{\partial L}{\partial z_1^{(l2)}} \dfrac{\partial z_1^{(l2)}}{\partial z_2^{(l1)}} \dfrac{\partial z_2^{(l1)}}{\partial w_{21}} & \dfrac{\partial L}{\partial z_2^{(l2)}} \left(z_2^{(l1)} + w_{22} \dfrac{\partial z_2^{(l1)}}{\partial w_{22}} \right) + \dfrac{\partial L}{\partial z_1^{(l2)}} \dfrac{\partial z_1^{(l2)}}{\partial z_2^{(l1)}} \dfrac{\partial z_2^{(l1)}}{\partial w_{22}} \end{vmatrix}$$

The above matrix may seem very complicated, but here is the trick:

Rule1. w_{11} and w_{12} have "direct" connect to $z_1^{(l2)}$ due to the matrix multiplication, the same as w_{21} and w_{22} to $z_2^{(l2)}$. Therefore we can get:

$$\frac{\partial z_1^{(l2)}}{\partial w_{11}} = \left(z_1^{(l1)} + w_{11} \frac{\partial z_1^{(l1)}}{\partial w_{11}} \right)$$

$$\frac{\partial z_1^{(l2)}}{\partial w_{12}} = \left(z_2^{(l1)} + w_{11} \frac{\partial z_1^{(l1)}}{\partial w_{12}} \right)$$

$$\frac{\partial z_2^{(l2)}}{\partial w_{21}} = \left(z_1^{(l1)} + w_{22} \frac{\partial z_2^{(l1)}}{\partial w_{21}} \right)$$

$$\frac{\partial z_2^{(l2)}}{\partial w_{22}} = \left(z_2^{(l1)} + w_{22} \frac{\partial z_2^{(l1)}}{\partial w_{22}} \right)$$

Rule 2. w_{11} and w_{12} are not "direct" connect to $z_2^{(l2)}$ in the 2nd layer, but due to $z_1^{(l1)}$, the term $\frac{\partial z_2^{(l2)}}{\partial w_{11}}$ and $\frac{\partial z_2^{(l2)}}{\partial w_{12}}$ are not zero, similarly as of w_{21} and w_{22} to $z_1^{(l2)}$ due to $z_2^{(l1)}$. Therefore we have

$$\frac{\partial z_2^{(l2)}}{\partial w_{11}} = \frac{\partial z_2^{(l2)}}{\partial z_1^{(l1)}} \frac{\partial z_1^{(l1)}}{\partial w_{11}}$$

$$\frac{\partial z_2^{(l2)}}{\partial w_{12}} = \frac{\partial z_2^{(l2)}}{\partial z_1^{(l1)}} \frac{\partial z_1^{(l1)}}{\partial w_{12}}$$

$$\frac{\partial z_1^{(l2)}}{\partial w_{21}} = \frac{\partial z_1^{(l2)}}{\partial z_2^{(l1)}} \frac{\partial z_2^{(l1)}}{\partial w_{21}}$$

$$\frac{\partial z_1^{(l2)}}{\partial w_{22}} = \frac{\partial z_1^{(l2)}}{\partial z_2^{(l1)}} \frac{\partial z_2^{(l1)}}{\partial w_{22}}$$

In other word, take w_{11} as an example, the connections are $w_{11} \rightarrow z_1^{(l2)}$ (Rule 1) and $w_{11} \rightarrow z_1^{(l1)} \rightarrow \left(z_1^{(l2)}, z_2^{(l2)} \right)$ (Rule 2)

Therefore, we have

$$\begin{vmatrix} \frac{\partial L}{\partial w_{11}} & \frac{\partial L}{\partial w_{12}} \\ \frac{\partial L}{\partial w_{21}} & \frac{\partial L}{\partial w_{22}} \end{vmatrix} = \begin{vmatrix} \frac{\partial L}{\partial z_1^{(l2)}} \left(z_1^{(l1)} + w_{11} x_1 \right) + \frac{\partial L}{\partial z_2^{(l2)}} w_{21} x_1 & \frac{\partial L}{\partial z_1^{(l2)}} \left(z_2^{(l1)} + w_{11} x_2 \right) + \frac{\partial L}{\partial z_2^{(l2)}} w_{21} x_2 \\ \frac{\partial L}{\partial z_2^{(l2)}} \left(z_1^{(l1)} + w_{22} x_1 \right) + \frac{\partial L}{\partial z_1^{(l2)}} w_{12} x_1 & \frac{\partial L}{\partial z_2^{(l2)}} \left(z_2^{(l1)} + w_{22} x_2 \right) + \frac{\partial L}{\partial z_1^{(l2)}} w_{12} x_2 \end{vmatrix}$$

Which equals to

$$\begin{vmatrix} \frac{\partial L}{\partial z_1^{(l2)}} z_1^{(l1)} + x_1 \left(\frac{\partial L}{\partial z_1^{(l2)}} w_{11} + \frac{\partial L}{\partial z_2^{(l2)}} w_{21} \right) & \frac{\partial L}{\partial z_1^{(l2)}} z_2^{(l1)} + x_2 \left(\frac{\partial L}{\partial z_1^{(l2)}} w_{11} + \frac{\partial L}{\partial z_2^{(l2)}} w_{21} \right) \\ \frac{\partial L}{\partial z_2^{(l2)}} z_1^{(l1)} + x_1 \left(\frac{\partial L}{\partial z_1^{(l2)}} w_{12} + \frac{\partial L}{\partial z_2^{(l2)}} w_{22} \right) & \frac{\partial L}{\partial z_2^{(l2)}} z_2^{(l1)} + x_2 \left(\frac{\partial L}{\partial z_1^{(l2)}} w_{12} + \frac{\partial L}{\partial z_2^{(l2)}} w_{22} \right) \end{vmatrix}$$

And

$$\begin{vmatrix} \dfrac{\partial L}{\partial z_1^{(l2)}} \\ \dfrac{\partial L}{\partial z_2^{(l2)}} \end{vmatrix} \begin{vmatrix} z_1^{(l1)} & z_2^{(l1)} \end{vmatrix} + \left(\begin{vmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \end{vmatrix} \begin{vmatrix} \dfrac{\partial L}{\partial z_1^{(l2)}} \\ \dfrac{\partial L}{\partial z_2^{(l2)}} \end{vmatrix} \right) \begin{vmatrix} x_1 & x_2 \end{vmatrix}$$

Therefore in the nested NN

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{\delta}^{(l2)} \boldsymbol{z}^{(l2)} + \left(\boldsymbol{w}^T \boldsymbol{\delta}^{(l2)} \right) \boldsymbol{x} = \boldsymbol{\delta}^{(l2)} \boldsymbol{z}^{(l2)} + \boldsymbol{\delta}^{(l1)} \boldsymbol{x}$$

Similarly

$$\frac{\partial L}{\partial \boldsymbol{b}} = \boldsymbol{\delta}^{(l2)} + \boldsymbol{w}^T \boldsymbol{\delta}^{(l2)} = \boldsymbol{\delta}^{(l2)} + \boldsymbol{\delta}^{(l1)}$$

This is a good news as the gradient of reused weights are the sum of gradients from the corresponding layers

4. Batch

Consider a 1 layer NN with a batch of 2

$$\begin{vmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{vmatrix} = \begin{vmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{vmatrix} \begin{vmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{vmatrix} + \begin{vmatrix} b_1 & b_1 \\ b_2 & b_2 \end{vmatrix}$$

With loss function

$$L = l\left(\begin{vmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{vmatrix} \right)$$

The gradient terms are

$$\frac{\partial L}{\partial \boldsymbol{w}} = \begin{vmatrix} \dfrac{\partial L}{\partial w_{11}} & \dfrac{\partial L}{\partial w_{12}} \\ \dfrac{\partial L}{\partial w_{21}} & \dfrac{\partial L}{\partial w_{22}} \end{vmatrix} = \begin{vmatrix} \dfrac{\partial L}{\partial z_{11}}\dfrac{\partial z_{11}}{\partial w_{11}} + \dfrac{\partial L}{\partial z_{12}}\dfrac{\partial z_{12}}{\partial w_{11}} & \dfrac{\partial L}{\partial z_{11}}\dfrac{\partial z_{11}}{\partial w_{12}} + \dfrac{\partial L}{\partial z_{12}}\dfrac{\partial z_{12}}{\partial w_{12}} \\ \dfrac{\partial L}{\partial z_{21}}\dfrac{\partial z_{21}}{\partial w_{21}} + \dfrac{\partial L}{\partial z_{22}}\dfrac{\partial z_{22}}{\partial w_{21}} & \dfrac{\partial L}{\partial z_{21}}\dfrac{\partial z_{21}}{\partial w_{22}} + \dfrac{\partial L}{\partial z_{22}}\dfrac{\partial z_{22}}{\partial w_{22}} \end{vmatrix}$$

$$= \begin{vmatrix} \dfrac{\partial L}{\partial z_{11}}x_{11} + \dfrac{\partial L}{\partial z_{12}}x_{12} & \dfrac{\partial L}{\partial z_{11}}x_{21} + \dfrac{\partial L}{\partial z_{12}}x_{22} \\ \dfrac{\partial L}{\partial z_{21}}x_{11} + \dfrac{\partial L}{\partial z_{22}}x_{12} & \dfrac{\partial L}{\partial z_{21}}x_{21} + \dfrac{\partial L}{\partial z_{22}}x_{22} \end{vmatrix} = \begin{vmatrix} \dfrac{\partial L}{\partial z_{11}} & \dfrac{\partial L}{\partial z_{12}} \\ \dfrac{\partial L}{\partial z_{21}} & \dfrac{\partial L}{\partial z_{22}} \end{vmatrix} \begin{vmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \end{vmatrix} = \boldsymbol{\delta} \boldsymbol{x}^T$$

$$\frac{\partial L}{\partial \boldsymbol{x}} = \begin{vmatrix} \dfrac{\partial L}{\partial x_{11}} & \dfrac{\partial L}{\partial x_{12}} \\ \dfrac{\partial L}{\partial x_{21}} & \dfrac{\partial L}{\partial x_{22}} \end{vmatrix} = \begin{vmatrix} \dfrac{\partial L}{\partial z_{11}}\dfrac{\partial z_{11}}{\partial x_{11}} + \dfrac{\partial L}{\partial z_{21}}\dfrac{\partial z_{21}}{\partial x_{11}} & \dfrac{\partial L}{\partial z_{12}}\dfrac{\partial z_{12}}{\partial x_{12}} + \dfrac{\partial L}{\partial z_{22}}\dfrac{\partial z_{22}}{\partial x_{12}} \\ \dfrac{\partial L}{\partial z_{11}}\dfrac{\partial z_{11}}{\partial x_{21}} + \dfrac{\partial L}{\partial z_{21}}\dfrac{\partial z_{21}}{\partial x_{21}} & \dfrac{\partial L}{\partial z_{12}}\dfrac{\partial z_{12}}{\partial x_{22}} + \dfrac{\partial L}{\partial z_{22}}\dfrac{\partial z_{22}}{\partial x_{22}} \end{vmatrix}$$

$$= \begin{vmatrix} \dfrac{\partial L}{\partial z_{11}}w_{11} + \dfrac{\partial L}{\partial z_{21}}w_{21} & \dfrac{\partial L}{\partial z_{12}}w_{11} + \dfrac{\partial L}{\partial z_{22}}w_{21} \\ \dfrac{\partial L}{\partial z_{11}}w_{12} + \dfrac{\partial L}{\partial z_{21}}w_{22} & \dfrac{\partial L}{\partial z_{12}}w_{12} + \dfrac{\partial L}{\partial z_{22}}w_{22} \end{vmatrix} = \begin{vmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \end{vmatrix} \begin{vmatrix} \dfrac{\partial L}{\partial z_{11}} & \dfrac{\partial L}{\partial z_{12}} \\ \dfrac{\partial L}{\partial z_{21}} & \dfrac{\partial L}{\partial z_{22}} \end{vmatrix} = \boldsymbol{w}^T \boldsymbol{\delta}$$

Since $\boldsymbol{b}$ is repeated column-wise, the gradient $\frac{\partial L}{\partial \boldsymbol{b}}$ will be the column-wise addition result of $\boldsymbol{\delta}$

$$\frac{\partial L}{\partial \boldsymbol{b}} = \begin{vmatrix} \dfrac{\partial L}{\partial z_{11}} + \dfrac{\partial L}{\partial z_{12}} \\ \dfrac{\partial L}{\partial z_{21}} + \dfrac{\partial L}{\partial z_{22}} \end{vmatrix}$$

Which means the batch version has no other difference to the no-batch version, except for element-wise operations (such as addition, bias, …)

5. Computational graph