A Numpy Neural Network

Saffet Gökçen Şen

May 31, 2020

A neural network with 3 hidden layers is to be constructed. The gradient of the loss function with respect to the weights and biases of the hidden units are derived. In a separate Jupyter notebook, the forward propagation and backward propagation are implemented using numpy to fit the network to a dataset. The output data and the input data are created randomly. The activation functions of the hidden units are all relu except the last hidden layer.

The dimension of the input layer is d_{in}. It means that there are d_{in} features. The dimensions of the first, second and third layers are d_1, d_2 and d_{out} respectively. Hence, there are d_{out} outputs. Let the batch size be n. Then, the input to the first hidden layer is a matrix $\mathbf{x}$ of dimension $n \times d_{in}$. The weight matrix of the first hidden layer is $\mathbf{w}_1$ of dimension $d_{in} \times d_1$. The bias matrix of the first hidden layer is $\mathbf{b}_1$ of dimension $1 \times d_1$. The output of the first hidden layer is $\mathbf{h}_1$ given by

$$\mathbf{h}_1 = \mathbf{x}.\mathbf{w}_1 + \mathbf{1}_{\mathrm{n}\times 1}.\mathbf{b}_1 \tag{1}$$

$$\mathbf{h}_{1_{\mathrm{relu}}} = \mathrm{Relu}\left(\mathbf{h}_1\right) \tag{2}$$

$\mathbf{1}_{\mathrm{n}\times 1}$ is a matrix of ones with dimension $\mathrm{n} \times 1$. Let $\mathbf{1}_{\mathrm{n}\times 1}.\mathbf{b}_1$ be denoted by $\mathbf{B}_1$. $\mathbf{B}_1$, $\mathbf{h}_1$ and $\mathbf{h}_{1_{\mathrm{relu}}}$ are of dimension $n \times d_1$. The weight matrix of the second hidden layer is $\mathbf{w}_2$ of dimension $d_1 \times d_2$. The bias matrix of the second hidden layer is $\mathbf{b}_2$ of dimension $1 \times d_2$. The output of the second hidden layer is $\mathbf{h}_2$ given by

$$\mathbf{h}_2 = \mathbf{h}_{1_{\mathrm{relu}}}.\mathbf{w}_2 + \mathbf{1}_{\mathrm{n}\times 1}.\mathbf{b}_2 \tag{3}$$

$$\mathbf{h}_{2_{\mathrm{relu}}} = \mathrm{Relu}\left(\mathbf{h}_2\right) \tag{4}$$

Let $\mathbf{1}_{\mathrm{n}\times 1}.\mathbf{b}_2$ be denoted by $\mathbf{B}_2$. $\mathbf{B}_2$, $\mathbf{h}_2$ and $\mathbf{h}_{2_{\mathrm{relu}}}$ are of dimension $n \times d_2$. The weight matrix of the third hidden layer is $\mathbf{w}_3$ of dimension $d_2 \times d_{\mathrm{out}}$. The bias matrix of the third hidden layer is $\mathbf{b}_3$ of dimension $1 \times d_{\mathrm{out}}$. The output of the third hidden layer is y_{pred} given by

$$\mathbf{y}_{\text{pred}} = \mathbf{h}_{2_{\text{relu}}}.\mathbf{w}_3 + \mathbf{1}_{\text{n}\times 1}.\mathbf{b}_3 \tag{5}$$

Let $\mathbf{1}_{\text{n}\times 1}.\mathbf{b}_3$ be denoted by $\mathbf{B}_3$. $\mathbf{B}_3$ and $\mathbf{y}_{\text{pred}}$ are of dimension $n \times d_{\text{out}}$. The loss function is the mean squared error function. Hence,

$$\text{L} = \sum_{i=1}^{n}\sum_{j=1}^{d_{\text{out}}} \left(y_{\text{pred}_{\text{ij}}} - y_{\text{ij}}\right)^2 \tag{6}$$

The derivative of the loss function L with respect to the matrix $\mathbf{w}_3$ is a matrix of dimension $1 \times (\text{d}_2.\text{d}_{\text{out}})$. The element in the column k.l of this matrix is given by

$$\left(\frac{\partial \text{L}}{\partial \mathbf{w}_3}\right)_{1(\text{k.l})} = \frac{\partial L}{\partial w_{3_{\text{kl}}}} = \frac{\partial}{\partial w_{3_{\text{kl}}}}\left[\sum_{i=1}^{n}\sum_{j=1}^{d_{\text{out}}} \left(y_{\text{pred}_{\text{ij}}} - y_{\text{ij}}\right)^2\right] = \tag{7}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{d_{\text{out}}} \frac{\partial}{\partial w_{3_{\text{kl}}}}\left[\left(y_{\text{pred}_{\text{ij}}} - y_{\text{ij}}\right)^2\right] = \sum_{i=1}^{n}\sum_{j=1}^{d_{\text{out}}} 2\left(y_{\text{pred}_{\text{ij}}} - y_{\text{ij}}\right)\frac{\partial y_{\text{pred}_{\text{ij}}}}{\partial w_{3_{\text{kl}}}} \tag{8}$$

$y_{\text{pred}_{\text{ij}}}$ is equal to the following:

$$y_{\text{pred}_{\text{ij}}} = \sum_{m=1}^{d_2} h_{2_{\text{relu}_{\text{im}}}} w_{3_{\text{mj}}} + B_{3_{\text{ij}}} \Rightarrow \tag{9}$$

$$\frac{\partial y_{\text{pred}_{\text{ij}}}}{\partial w_{3_{\text{kl}}}} = h_{2_{\text{relu}_{\text{im}}}} I\left(\text{m=k \& j=l}\right) \tag{10}$$

Let the derivation for $\left(\frac{\partial \text{L}}{\partial \mathbf{w}_3}\right)_{1(\text{k.l})}$ be completed:

$$\left(\frac{\partial \text{L}}{\partial \mathbf{w}_3}\right)_{1(\text{k.l})} = \sum_{i=1}^{n}\sum_{j=1}^{d_{\text{out}}} 2\left(y_{\text{pred}_{\text{ij}}} - y_{\text{ij}}\right) h_{2_{\text{relu}_{\text{im}}}} I\left(\text{m=k \& j=l}\right) \Rightarrow \tag{11}$$

$$\left(\frac{\partial \mathrm{L}}{\partial \mathbf{w}_3}\right)_{1(\mathrm{k.l})} = \sum_{i=1}^{n} 2\left(y_{\mathrm{pred_{il}}} - y_{\mathrm{il}}\right) h_{2_{\mathrm{relu_{ik}}}} \Rightarrow \tag{12}$$

$$\left(\frac{\partial \mathrm{L}}{\partial \mathbf{w}_3}\right)_{1(\mathrm{k.l})} = \left[\mathbf{h}_{2_{\mathrm{relu}}}^{T}.2\left(\mathbf{y}_{\mathrm{pred}} - \mathbf{y}\right)\right]_{\mathrm{kl}} \tag{13}$$

If $\frac{\partial \mathrm{L}}{\partial \mathbf{w}_3}$ is put into the form of a matrix of dimension $\mathrm{d}_2 \times \mathrm{d}_{\mathrm{out}}$, then it's equal to $\mathbf{h}_{2_{\mathrm{relu}}}^{T}.2\left(\mathbf{y}_{\mathrm{pred}} - \mathbf{y}\right)$.

Now, let the derivative of the loss with respect to the $\mathbf{b}_3$ matrix be derived. It is a matrix of dimension $1 \times \mathrm{d}_{\mathrm{out}}$. The element in its k^{th} column is given as follows:

$$\left(\frac{\partial \mathrm{L}}{\partial \mathbf{b}_3}\right)_{1\mathrm{k}} = \frac{\partial L}{\partial b_{3_{\mathrm{k}}}} = \frac{\partial}{\partial b_{3_{\mathrm{k}}}}\left[\sum_{i=1}^{n}\sum_{j=1}^{d_{\mathrm{out}}}\left(y_{\mathrm{pred_{ij}}} - y_{\mathrm{ij}}\right)^2\right] = \tag{14}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{d_{\mathrm{out}}}\frac{\partial}{\partial b_{3_{\mathrm{k}}}}\left[\left(y_{\mathrm{pred_{ij}}} - y_{\mathrm{ij}}\right)^2\right] = \sum_{i=1}^{n}\sum_{j=1}^{d_{\mathrm{out}}} 2\left(y_{\mathrm{pred_{ij}}} - y_{\mathrm{ij}}\right)\frac{\partial y_{\mathrm{pred_{ij}}}}{\partial b_{3_{\mathrm{k}}}} \tag{15}$$

$$y_{\mathrm{pred_{ij}}} = \sum_{m=1}^{d_2} h_{2_{\mathrm{relu_{im}}}} w_{3_{\mathrm{mj}}} + B_{3_{\mathrm{ij}}} \Rightarrow \frac{\partial y_{\mathrm{pred_{ij}}}}{\partial b_{3_{\mathrm{k}}}} = \frac{\partial B_{3_{\mathrm{ij}}}}{\partial b_{3_{\mathrm{k}}}} = \frac{\partial b_{3_{\mathrm{j}}}}{\partial b_{3_{\mathrm{k}}}} \Rightarrow \tag{16}$$

$$\frac{\partial y_{\mathrm{pred_{ij}}}}{\partial b_{3_{\mathrm{k}}}} = I\left(\mathrm{j{=}k}\right) \tag{17}$$

Hence:

$$\left(\frac{\partial \mathrm{L}}{\partial \mathbf{b}_3}\right)_{1\mathrm{k}} = \sum_{i=1}^{n}\sum_{j=1}^{d_{\mathrm{out}}} 2\left(y_{\mathrm{pred_{ij}}} - y_{\mathrm{ij}}\right)\frac{\partial y_{\mathrm{pred_{ij}}}}{\partial b_{3_{\mathrm{k}}}} \Rightarrow \tag{18}$$

$$\left(\frac{\partial \mathrm{L}}{\partial \mathbf{b}_3}\right)_{1\mathrm{k}} = \sum_{i=1}^{n}\sum_{j=1}^{d_{\mathrm{out}}} 2\left(y_{\mathrm{pred_{ij}}} - y_{\mathrm{ij}}\right) I\left(\mathrm{j{=}k}\right) \Rightarrow \tag{19}$$

$$\left(\frac{\partial \mathrm{L}}{\partial \mathbf{b}_3}\right)_{1\mathrm{k}} = \sum_{i=1}^{n} 2\left(y_{\mathrm{pred_{ik}}} - y_{\mathrm{ik}}\right) \tag{20}$$

The k^{th} column of $\frac{\partial \mathrm{L}}{\partial \mathbf{b}_3}$ is equal to the sum of the elements of the k^{th} column of $2\left(\mathbf{y}_{\mathrm{pred}} - \mathbf{y}\right)$.

The derivative of L with respect to the matrix $\mathbf{w}_2$ is to be derived. It is a matrix of dimension 1 × ($\mathrm{d_1.d_2}$). The element in the column k.l of this matrix is given by

$$\left(\frac{\partial \mathrm{L}}{\partial \mathbf{w}_2}\right)_{1(\mathrm{k.l})} = \frac{\partial L}{\partial w_{2_{\mathrm{kl}}}} = \frac{\partial}{\partial w_{2_{\mathrm{kl}}}}\left[\sum_{i=1}^{n}\sum_{j=1}^{d_{\mathrm{out}}}\left(y_{\mathrm{pred}_{\mathrm{ij}}} - y_{\mathrm{ij}}\right)^2\right] = \tag{21}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{d_{\mathrm{out}}}\frac{\partial}{\partial w_{2_{\mathrm{kl}}}}\left[\left(y_{\mathrm{pred}_{\mathrm{ij}}} - y_{\mathrm{ij}}\right)^2\right] = \sum_{i=1}^{n}\sum_{j=1}^{d_{\mathrm{out}}} 2\left(y_{\mathrm{pred}_{\mathrm{ij}}} - y_{\mathrm{ij}}\right)\frac{\partial y_{\mathrm{pred}_{\mathrm{ij}}}}{\partial w_{2_{\mathrm{kl}}}} \tag{22}$$

$$y_{\mathrm{pred}_{\mathrm{ij}}} = \sum_{m=1}^{d_2} h_{2_{\mathrm{relu}_{\mathrm{im}}}} w_{3_{\mathrm{mj}}} + B_{3_{\mathrm{ij}}} = \sum_{m=1}^{d_2} h_{2_{\mathrm{relu}_{\mathrm{im}}}} w_{3_{\mathrm{mj}}} + B_{3_{\mathrm{ij}}} \Rightarrow \tag{23}$$

$$\frac{\partial y_{\mathrm{pred}_{\mathrm{ij}}}}{\partial w_{2_{\mathrm{kl}}}} = \sum_{m=1}^{d_2}\frac{\partial h_{2_{\mathrm{relu}_{\mathrm{im}}}}}{\partial w_{2_{\mathrm{kl}}}} w_{3_{\mathrm{mj}}} = \sum_{m=1}^{d_2}\frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial w_{2_{\mathrm{kl}}}} w_{3_{\mathrm{mj}}} = \tag{24}$$

$$\sum_{m=1}^{d_2}\frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial h_{2_{\mathrm{im}}}}\frac{\partial h_{2_{\mathrm{im}}}}{\partial w_{2_{\mathrm{kl}}}} w_{3_{\mathrm{mj}}} = \tag{25}$$

$$\sum_{m=1}^{d_2}\frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial h_{2_{\mathrm{im}}}}\frac{\partial}{\partial w_{2_{\mathrm{kl}}}}\left(\sum_{p=1}^{d_1} h_{1_{\mathrm{relu}_{\mathrm{ip}}}} w_{2_{\mathrm{pm}}} + B_{2_{\mathrm{im}}}\right) w_{3_{\mathrm{mj}}} = \tag{26}$$

$$\sum_{m=1}^{d_2}\frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial h_{2_{\mathrm{im}}}} h_{1_{\mathrm{relu}_{\mathrm{ip}}}} I\left(\mathrm{p{=}k\ \&\ m{=}l}\right) w_{3_{\mathrm{mj}}} \Rightarrow \tag{27}$$

$$\frac{\partial y_{\mathrm{pred}_{\mathrm{ij}}}}{\partial w_{2_{\mathrm{kl}}}} = \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{il}}}\right)}{\partial h_{2_{\mathrm{il}}}} h_{1_{\mathrm{relu}_{\mathrm{ik}}}} w_{3_{\mathrm{lj}}} \tag{28}$$

Going on with the derivation:

$$\left(\frac{\partial \mathrm{L}}{\partial \mathbf{w}_2}\right)_{1(\mathrm{k.l})} = \sum_{i=1}^{n}\sum_{j=1}^{d_{\mathrm{out}}} 2\left(y_{\mathrm{pred}_{\mathrm{ij}}} - y_{\mathrm{ij}}\right)\frac{\partial y_{\mathrm{pred}_{\mathrm{ij}}}}{\partial w_{2_{\mathrm{kl}}}} = \tag{29}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{d_{\mathrm{out}}} 2\left(y_{\mathrm{pred}_{\mathrm{ij}}} - y_{\mathrm{ij}}\right)\frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{il}}}\right)}{\partial h_{2_{\mathrm{il}}}} h_{1_{\mathrm{relu}_{\mathrm{ik}}}} w_{3_{\mathrm{lj}}} = \tag{30}$$

$$\sum_{i=1}^{n} \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{il}}}\right)}{\partial h_{2_{\mathrm{il}}}} h_{1_{\mathrm{relu}_{\mathrm{ik}}}} \left(\sum_{j=1}^{d_{\mathrm{out}}} 2\left(y_{\mathrm{pred}_{\mathrm{ij}}} - y_{\mathrm{ij}}\right) w_{3_{\mathrm{lj}}}\right) \Rightarrow \tag{31}$$

$$\left(\frac{\partial \mathrm{L}}{\partial \mathbf{w}_2}\right)_{1(\mathrm{k.l})} = \sum_{i=1}^{n} h_{1_{\mathrm{relu}_{\mathrm{ik}}}} \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{il}}}\right)}{\partial h_{2_{\mathrm{il}}}} \left[2\left(\mathbf{y}_{\mathrm{pred}} - \mathbf{y}\right).\mathbf{w}_3^T\right]_{\mathrm{il}} \tag{32}$$

Let the following definition be made for the matrix $\mathbf{A}$:

$$\mathrm{A}_{\mathrm{il}} = \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{il}}}\right)}{\partial h_{2_{\mathrm{il}}}} = \begin{cases} 1 & \text{if } h_{2_{\mathrm{il}}} \geq 0 \\ 0 & \text{if } h_{2_{\mathrm{il}}} < 0 \end{cases} \tag{33}$$

If $\frac{\partial \mathrm{L}}{\partial \mathbf{w}_2}$ is put into a matrix of dimension $d_1 \times d_2$, then it is equal to

$$\mathbf{h}_{1_{\mathrm{relu}}}^T . \left(\mathbf{A} \circ \left(2\left(\mathbf{y}_{\mathrm{pred}} - \mathbf{y}\right).\mathbf{w}_3^T\right)\right) \tag{34}$$

where $\circ$ denotes the Hadamard or Schur product of two matrices.

The derivative of L with respect to the bias matrix $\mathbf{b}_2$ is going to be obtained. It is a matrix of $1 \times \mathrm{d}_2$. The element in its k^{th} column is given as follows:

$$\left(\frac{\partial \mathrm{L}}{\partial \mathbf{b}_2}\right)_{1\mathrm{k}} = \frac{\partial L}{\partial b_{2_{\mathrm{k}}}} = \frac{\partial}{\partial b_{2_{\mathrm{k}}}} \left[\sum_{i=1}^{n} \sum_{j=1}^{d_{\mathrm{out}}} \left(y_{\mathrm{pred}_{\mathrm{ij}}} - y_{\mathrm{ij}}\right)^2\right] = \tag{35}$$

$$\sum_{i=1}^{n} \sum_{j=1}^{d_{\mathrm{out}}} \frac{\partial}{\partial b_{2_{\mathrm{k}}}} \left[\left(y_{\mathrm{pred}_{\mathrm{ij}}} - y_{\mathrm{ij}}\right)^2\right] = \sum_{i=1}^{n} \sum_{j=1}^{d_{\mathrm{out}}} 2\left(y_{\mathrm{pred}_{\mathrm{ij}}} - y_{\mathrm{ij}}\right) \frac{\partial y_{\mathrm{pred}_{\mathrm{ij}}}}{\partial b_{2_{\mathrm{k}}}} \tag{36}$$

$$y_{\mathrm{pred}_{\mathrm{ij}}} = \sum_{m=1}^{d_2} h_{2_{\mathrm{relu}_{\mathrm{im}}}} w_{3_{\mathrm{mj}}} + B_{3_{\mathrm{ij}}} \Rightarrow \tag{37}$$

$$\frac{\partial y_{\mathrm{pred}_{\mathrm{ij}}}}{\partial b_{2_{\mathrm{k}}}} = \sum_{m=1}^{d_2} \frac{\partial h_{2_{\mathrm{relu}_{\mathrm{im}}}}}{\partial b_{2_{\mathrm{k}}}} w_{3_{\mathrm{mj}}} = \sum_{m=1}^{d_2} \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial b_{2_{\mathrm{k}}}} w_{3_{\mathrm{mj}}} = \tag{38}$$

$$\sum_{m=1}^{d_2} \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial h_{2_{\mathrm{im}}}} \frac{\partial h_{2_{\mathrm{im}}}}{\partial b_{2_{\mathrm{k}}}} w_{3_{\mathrm{mj}}} = \tag{39}$$

$$\sum_{m=1}^{d_2} \frac{\partial \text{Relu}\left(h_{2_{\text{im}}}\right)}{\partial h_{2_{\text{im}}}} \frac{\partial}{\partial b_{2_{\text{k}}}} \left(\sum_{p=1}^{d_1} h_{1_{\text{relu}_{\text{ip}}}} w_{2_{\text{pm}}} + B_{2_{\text{im}}} \right) w_{3_{\text{mj}}} = \tag{40}$$

$$\sum_{m=1}^{d_2} \frac{\partial \text{Relu}\left(h_{2_{\text{im}}}\right)}{\partial h_{2_{\text{im}}}} \frac{\partial b_{2_{\text{m}}}}{\partial b_{2_{\text{k}}}} w_{3_{\text{mj}}} = \sum_{m=1}^{d_2} \frac{\partial \text{Relu}\left(h_{2_{\text{im}}}\right)}{\partial h_{2_{\text{im}}}} I\left(\text{m=k}\right) w_{3_{\text{mj}}} \Rightarrow \tag{41}$$

$$\frac{\partial y_{\text{pred}_{\text{ij}}}}{\partial b_{2_{\text{k}}}} = \frac{\partial \text{Relu}\left(h_{2_{\text{ik}}}\right)}{\partial h_{2_{\text{ik}}}} w_{3_{\text{kj}}} \tag{42}$$

Going on with the derivation:

$$\left(\frac{\partial \text{L}}{\partial \mathbf{b}_2}\right)_{\text{1k}} = \sum_{i=1}^{n} \sum_{j=1}^{d_{\text{out}}} 2\left(y_{\text{pred}_{\text{ij}}} - y_{\text{ij}}\right) \frac{\partial y_{\text{pred}_{\text{ij}}}}{\partial b_{2_{\text{k}}}} = \tag{43}$$

$$\sum_{i=1}^{n} \sum_{j=1}^{d_{\text{out}}} 2\left(y_{\text{pred}_{\text{ij}}} - y_{\text{ij}}\right) \frac{\partial \text{Relu}\left(h_{2_{\text{ik}}}\right)}{\partial h_{2_{\text{ik}}}} w_{3_{\text{kj}}} = \tag{44}$$

$$\sum_{i=1}^{n} \frac{\partial \text{Relu}\left(h_{2_{\text{ik}}}\right)}{\partial h_{2_{\text{ik}}}} \left(\sum_{j=1}^{d_{\text{out}}} 2\left(y_{\text{pred}_{\text{ij}}} - y_{\text{ij}}\right) w_{3_{\text{kj}}} \right) \Rightarrow \tag{45}$$

$$\left(\frac{\partial \text{L}}{\partial \mathbf{b}_2}\right)_{\text{1k}} = \sum_{i=1}^{n} \frac{\partial \text{Relu}\left(h_{2_{\text{ik}}}\right)}{\partial h_{2_{\text{ik}}}} \left[2\left(\mathbf{y}_{\text{pred}} - \mathbf{y}\right).\mathbf{w}_3^T\right]_{\text{ik}} \tag{46}$$

The k$^{\text{th}}$ column of $\frac{\partial \text{L}}{\partial \mathbf{b}_2}$ is equal to the sum of the elements in the k$^{\text{th}}$ column of $\mathbf{A} \circ \left[2\left(\mathbf{y}_{\text{pred}} - \mathbf{y}\right).\mathbf{w}_3^T\right]$.

The derivative of the loss L is to be derived with respect to the weight matrix $\mathbf{w}_1$. This derivative will be of dimension $1 \times (\text{d}_{\text{in}}.\text{d}_1)$.

$$\left(\frac{\partial \text{L}}{\partial \mathbf{w}_1}\right)_{\text{1(k.l)}} = \frac{\partial L}{\partial w_{1_{\text{kl}}}} = \frac{\partial}{\partial w_{1_{\text{kl}}}} \left[\sum_{i=1}^{n} \sum_{j=1}^{d_{\text{out}}} \left(y_{\text{pred}_{\text{ij}}} - y_{\text{ij}}\right)^2 \right] = \tag{47}$$

$$\sum_{i=1}^{n} \sum_{j=1}^{d_{\text{out}}} \frac{\partial}{\partial w_{1_{\text{kl}}}} \left[\left(y_{\text{pred}_{\text{ij}}} - y_{\text{ij}}\right)^2 \right] = \sum_{i=1}^{n} \sum_{j=1}^{d_{\text{out}}} 2\left(y_{\text{pred}_{\text{ij}}} - y_{\text{ij}}\right) \frac{\partial y_{\text{pred}_{\text{ij}}}}{\partial w_{1_{\text{kl}}}} \tag{48}$$

$$y_{\mathrm{pred}_{\mathrm{ij}}} = \sum_{m=1}^{d_2} h_{2_{\mathrm{relu}_{\mathrm{im}}}} w_{3_{\mathrm{mj}}} + B_{3_{\mathrm{ij}}} = \sum_{m=1}^{d_2} \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right) w_{3_{\mathrm{mj}}} + B_{3_{\mathrm{ij}}} \Rightarrow \tag{49}$$

$$\frac{\partial y_{\mathrm{pred}_{\mathrm{ij}}}}{\partial w_{1_{\mathrm{kl}}}} = \sum_{m=1}^{d_2} \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial h_{2_{\mathrm{im}}}} \frac{\partial h_{2_{\mathrm{im}}}}{\partial w_{1_{\mathrm{kl}}}} w_{3_{\mathrm{mj}}} = \tag{50}$$

$$\sum_{m=1}^{d_2} w_{3_{\mathrm{mj}}} \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial h_{2_{\mathrm{im}}}} \frac{\partial}{\partial w_{1_{\mathrm{kl}}}} \left(\sum_{p=1}^{d_1} h_{1_{\mathrm{relu}_{\mathrm{ip}}}} w_{2_{\mathrm{pm}}} + B_{2_{\mathrm{im}}} \right) = \tag{51}$$

$$\sum_{m=1}^{d_2} w_{3_{\mathrm{mj}}} \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial h_{2_{\mathrm{im}}}} \sum_{p=1}^{d_1} \frac{\partial \mathrm{Relu}\left(h_{1_{\mathrm{ip}}}\right)}{\partial h_{1_{\mathrm{ip}}}} \frac{\partial h_{1_{\mathrm{ip}}}}{\partial w_{1_{\mathrm{kl}}}} w_{2_{\mathrm{pm}}} = \tag{52}$$

$$\sum_{m=1}^{d_2} w_{3_{\mathrm{mj}}} \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial h_{2_{\mathrm{im}}}} \sum_{p=1}^{d_1} \frac{\partial \mathrm{Relu}\left(h_{1_{\mathrm{ip}}}\right)}{\partial h_{1_{\mathrm{ip}}}} \frac{\partial}{\partial w_{1_{\mathrm{kl}}}} \left(\sum_{q=1}^{d_{\mathrm{in}}} x_{\mathrm{iq}} w_{1_{\mathrm{qp}}} + B_{1_{\mathrm{ip}}} \right) w_{2_{\mathrm{pm}}} = \tag{53}$$

$$\sum_{m=1}^{d_2} w_{3_{\mathrm{mj}}} \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial h_{2_{\mathrm{im}}}} \sum_{p=1}^{d_1} \frac{\partial \mathrm{Relu}\left(h_{1_{\mathrm{ip}}}\right)}{\partial h_{1_{\mathrm{ip}}}} x_{\mathrm{iq}} I\left(\mathrm{q{=}k\ \&\ p{=}l}\right) w_{2_{\mathrm{pm}}} \Rightarrow \tag{54}$$

$$\frac{\partial y_{\mathrm{pred}_{\mathrm{ij}}}}{\partial w_{1_{\mathrm{kl}}}} = \sum_{m=1}^{d_2} w_{3_{\mathrm{mj}}} \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial h_{2_{\mathrm{im}}}} \frac{\partial \mathrm{Relu}\left(h_{1_{\mathrm{il}}}\right)}{\partial h_{1_{\mathrm{il}}}} x_{\mathrm{ik}} w_{2_{\mathrm{lm}}} \tag{55}$$

Going on with the derivation of $\frac{\partial L}{\partial \mathbf{w}_1}$:

$$\left(\frac{\partial \mathrm{L}}{\partial \mathbf{w}_1}\right)_{1(\mathrm{k.l})} = \frac{\partial L}{\partial w_{1_{\mathrm{kl}}}} = \sum_{i=1}^{n} \sum_{j=1}^{d_{\mathrm{out}}} 2\left(y_{\mathrm{pred}_{\mathrm{ij}}} - y_{\mathrm{ij}}\right) \frac{\partial y_{\mathrm{pred}_{\mathrm{ij}}}}{\partial w_{1_{\mathrm{kl}}}} = \tag{56}$$

$$\sum_{i=1}^{n} \sum_{j=1}^{d_{\mathrm{out}}} 2\left(y_{\mathrm{pred}_{\mathrm{ij}}} - y_{\mathrm{ij}}\right) \left(\sum_{m=1}^{d_2} w_{3_{\mathrm{mj}}} \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial h_{2_{\mathrm{im}}}} \frac{\partial \mathrm{Relu}\left(h_{1_{\mathrm{il}}}\right)}{\partial h_{1_{\mathrm{il}}}} x_{\mathrm{ik}} w_{2_{\mathrm{lm}}} \right) = \tag{57}$$

$$\sum_{i=1}^{n} \sum_{j=1}^{d_{\mathrm{out}}} \sum_{m=1}^{d_2} 2\left(y_{\mathrm{pred}_{\mathrm{ij}}} - y_{\mathrm{ij}}\right) w_{3_{\mathrm{mj}}} \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial h_{2_{\mathrm{im}}}} \frac{\partial \mathrm{Relu}\left(h_{1_{\mathrm{il}}}\right)}{\partial h_{1_{\mathrm{il}}}} x_{\mathrm{ik}} w_{2_{\mathrm{lm}}} = \tag{58}$$

$$\sum_{i=1}^{n} \sum_{m=1}^{d_2} \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial h_{2_{\mathrm{im}}}} \frac{\partial \mathrm{Relu}\left(h_{1_{\mathrm{il}}}\right)}{\partial h_{1_{\mathrm{il}}}} x_{\mathrm{ik}} w_{2_{\mathrm{lm}}} \sum_{j=1}^{d_{\mathrm{out}}} 2\left(y_{\mathrm{pred}_{\mathrm{ij}}} - y_{\mathrm{ij}}\right) w_{3_{\mathrm{mj}}} = \tag{59}$$

$$\sum_{i=1}^{n} \frac{\partial \mathrm{Relu}\left(h_{1_{\mathrm{il}}}\right)}{\partial h_{1_{\mathrm{il}}}} x_{\mathrm{ik}} \sum_{m=1}^{d_2} w_{2_{\mathrm{lm}}} \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial h_{2_{\mathrm{im}}}} \left(2\left(\mathbf{y}_{\mathrm{pred}} - \mathbf{y}\right).\mathbf{w}_3^T\right)_{\mathrm{im}} = \tag{60}$$

$$\sum_{i=1}^{n} x_{\mathrm{ik}} \frac{\partial \mathrm{Relu}\left(h_{1_{\mathrm{il}}}\right)}{\partial h_{1_{\mathrm{il}}}} \left(\left(\mathbf{A} \circ \left(2\left(\mathbf{y}_{\mathrm{pred}} - \mathbf{y}\right).\mathbf{w}_3^T\right)\right).\mathbf{w}_2^T\right)_{\mathrm{il}} = \tag{61}$$

$$\left(\frac{\partial \mathrm{L}}{\partial \mathbf{w}_1}\right)_{1(\mathrm{k.l})} = \left(\mathbf{x}^T.\left(\mathbf{C} \circ \left(\left(\mathbf{A} \circ \left(2\left(\mathbf{y}_{\mathrm{pred}} - \mathbf{y}\right).\mathbf{w}_3^T\right)\right).\mathbf{w}_2^T\right)\right)\right)_{\mathrm{kl}} \tag{62}$$

where the matrix $\mathbf{C}$ is defined as:

$$\mathrm{C}_{\mathrm{il}} = \frac{\partial \mathrm{Relu}\left(h_{1_{\mathrm{il}}}\right)}{\partial h_{1_{\mathrm{il}}}} = \begin{cases} 1 & \text{if } h_{1_{\mathrm{il}}} \geq 0 \\ 0 & \text{if } h_{1_{\mathrm{il}}} < 0 \end{cases} \tag{63}$$

If $\frac{\partial \mathrm{L}}{\partial \mathbf{w}_1}$ is put into the form of a matrix with dimension $\mathrm{d}_{\mathrm{in}} \times \mathrm{d}_1$, then this matrix is given by $\mathbf{x}^T.\left(\mathbf{C} \circ \left(\left(\mathbf{A} \circ \left(2\left(\mathbf{y}_{\mathrm{pred}} - \mathbf{y}\right).\mathbf{w}_3^T\right)\right).\mathbf{w}_2^T\right)\right)$.

It is now the turn of the derivative of L with respect to the bias matrix $\mathbf{b}_1$. This derivative is a matrix with dimension $1 \times \mathrm{d}_1$:

$$\left(\frac{\partial \mathrm{L}}{\partial \mathbf{b}_1}\right)_{1\mathrm{k}} = \frac{\partial L}{\partial b_{1_{\mathrm{k}}}} = \frac{\partial}{\partial b_{1_{\mathrm{k}}}} \left[\sum_{i=1}^{n} \sum_{j=1}^{d_{\mathrm{out}}} \left(y_{\mathrm{pred}_{\mathrm{ij}}} - y_{\mathrm{ij}}\right)^2\right] = \tag{64}$$

$$\sum_{i=1}^{n} \sum_{j=1}^{d_{\mathrm{out}}} \frac{\partial}{\partial b_{1_{\mathrm{k}}}} \left[\left(y_{\mathrm{pred}_{\mathrm{ij}}} - y_{\mathrm{ij}}\right)^2\right] = \sum_{i=1}^{n} \sum_{j=1}^{d_{\mathrm{out}}} 2\left(y_{\mathrm{pred}_{\mathrm{ij}}} - y_{\mathrm{ij}}\right) \frac{\partial y_{\mathrm{pred}_{\mathrm{ij}}}}{\partial b_{1_{\mathrm{k}}}} \tag{65}$$

$$y_{\mathrm{pred}_{\mathrm{ij}}} = \sum_{m=1}^{d_2} h_{2_{\mathrm{relu}_{\mathrm{im}}}} w_{3_{\mathrm{mj}}} + B_{3_{\mathrm{ij}}} = \sum_{m=1}^{d_2} \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right) w_{3_{\mathrm{mj}}} + B_{3_{\mathrm{ij}}} \Rightarrow \tag{66}$$

$$\frac{\partial y_{\mathrm{pred}_{\mathrm{ij}}}}{\partial b_{1_{\mathrm{kl}}}} = \sum_{m=1}^{d_2} \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial h_{2_{\mathrm{im}}}} \frac{\partial h_{2_{\mathrm{im}}}}{\partial b_{1_{\mathrm{k}}}} w_{3_{\mathrm{mj}}} = \tag{67}$$

$$\sum_{m=1}^{d_2} w_{3_{\mathrm{mj}}} \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial h_{2_{\mathrm{im}}}} \frac{\partial}{\partial b_{1_{\mathrm{k}}}} \left(\sum_{p=1}^{d_1} h_{1_{\mathrm{relu}_{\mathrm{ip}}}} w_{2_{\mathrm{pm}}} + B_{2_{\mathrm{im}}}\right) = \tag{68}$$

$$\sum_{m=1}^{d_2} w_{3_{\mathrm{mj}}} \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial h_{2_{\mathrm{im}}}} \sum_{p=1}^{d_1} \frac{\partial \mathrm{Relu}\left(h_{1_{\mathrm{ip}}}\right)}{\partial h_{1_{\mathrm{ip}}}} \frac{\partial h_{1_{\mathrm{ip}}}}{\partial b_{1_{\mathrm{k}}}} w_{2_{\mathrm{pm}}} = \tag{69}$$

$$\sum_{m=1}^{d_2} w_{3_{\mathrm{mj}}} \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial h_{2_{\mathrm{im}}}} \sum_{p=1}^{d_1} \frac{\partial \mathrm{Relu}\left(h_{1_{\mathrm{ip}}}\right)}{\partial h_{1_{\mathrm{ip}}}} \frac{\partial}{\partial b_{1_{\mathrm{k}}}} \left(\sum_{q=1}^{d_{\mathrm{in}}} x_{\mathrm{iq}} w_{1_{\mathrm{qp}}} + B_{1_{\mathrm{ip}}}\right) w_{2_{\mathrm{pm}}} = \tag{70}$$

$$\sum_{m=1}^{d_2} w_{3_{\mathrm{mj}}} \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial h_{2_{\mathrm{im}}}} \sum_{p=1}^{d_1} \frac{\partial \mathrm{Relu}\left(h_{1_{\mathrm{ip}}}\right)}{\partial h_{1_{\mathrm{ip}}}} \frac{\partial b_{1_{\mathrm{p}}}}{\partial b_{1_{\mathrm{k}}}} w_{2_{\mathrm{pm}}} = \tag{71}$$

$$\sum_{m=1}^{d_2} w_{3_{\mathrm{mj}}} \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial h_{2_{\mathrm{im}}}} \sum_{p=1}^{d_1} \frac{\partial \mathrm{Relu}\left(h_{1_{\mathrm{ip}}}\right)}{\partial h_{1_{\mathrm{ip}}}} I\left(\mathrm{p{=}k}\right) w_{2_{\mathrm{pm}}} \Rightarrow \tag{72}$$

$$\frac{\partial y_{\mathrm{pred}_{\mathrm{ij}}}}{\partial b_{1_{\mathrm{k}}}} = \sum_{m=1}^{d_2} w_{3_{\mathrm{mj}}} \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial h_{2_{\mathrm{im}}}} \frac{\partial \mathrm{Relu}\left(h_{1_{\mathrm{ik}}}\right)}{\partial h_{1_{\mathrm{ik}}}} w_{2_{\mathrm{km}}} \tag{73}$$

Going on with the derivation of $\frac{\partial L}{\partial \mathbf{b}_1}$:

$$\left(\frac{\partial \mathrm{L}}{\partial \mathbf{b}_1}\right)_{1\mathrm{k}} = \frac{\partial L}{\partial b_{1_{\mathrm{k}}}} = \sum_{i=1}^{n} \sum_{j=1}^{d_{\mathrm{out}}} 2\left(y_{\mathrm{pred}_{\mathrm{ij}}} - y_{\mathrm{ij}}\right) \frac{\partial y_{\mathrm{pred}_{\mathrm{ij}}}}{\partial b_{1_{\mathrm{k}}}} = \tag{74}$$

$$\sum_{i=1}^{n} \sum_{j=1}^{d_{\mathrm{out}}} 2\left(y_{\mathrm{pred}_{\mathrm{ij}}} - y_{\mathrm{ij}}\right) \left(\sum_{m=1}^{d_2} w_{3_{\mathrm{mj}}} \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial h_{2_{\mathrm{im}}}} \frac{\partial \mathrm{Relu}\left(h_{1_{\mathrm{ik}}}\right)}{\partial h_{1_{\mathrm{ik}}}} w_{2_{\mathrm{km}}}\right) = \tag{75}$$

$$\sum_{i=1}^{n} \sum_{j=1}^{d_{\mathrm{out}}} \sum_{m=1}^{d_2} 2\left(y_{\mathrm{pred}_{\mathrm{ij}}} - y_{\mathrm{ij}}\right) w_{3_{\mathrm{mj}}} \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial h_{2_{\mathrm{im}}}} \frac{\partial \mathrm{Relu}\left(h_{1_{\mathrm{ik}}}\right)}{\partial h_{1_{\mathrm{ik}}}} w_{2_{\mathrm{km}}} = \tag{76}$$

$$\sum_{i=1}^{n} \sum_{m=1}^{d_2} \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial h_{2_{\mathrm{im}}}} \frac{\partial \mathrm{Relu}\left(h_{1_{\mathrm{ik}}}\right)}{\partial h_{1_{\mathrm{ik}}}} w_{2_{\mathrm{km}}} \sum_{j=1}^{d_{\mathrm{out}}} 2\left(y_{\mathrm{pred}_{\mathrm{ij}}} - y_{\mathrm{ij}}\right) w_{3_{\mathrm{mj}}} = \tag{77}$$

$$\sum_{i=1}^{n} \frac{\partial \mathrm{Relu}\left(h_{1_{\mathrm{ik}}}\right)}{\partial h_{1_{\mathrm{ik}}}} \sum_{m=1}^{d_2} w_{2_{\mathrm{km}}} \frac{\partial \mathrm{Relu}\left(h_{2_{\mathrm{im}}}\right)}{\partial h_{2_{\mathrm{im}}}} \left(2\left(\mathbf{y}_{\mathrm{pred}} - \mathbf{y}\right).\mathbf{w}_3^T\right)_{\mathrm{im}} \Rightarrow \tag{78}$$

$$\left(\frac{\partial \mathrm{L}}{\partial \mathbf{b}_1}\right)_{1\mathrm{k}} = \sum_{i=1}^{n} \frac{\partial \mathrm{Relu}\left(h_{1_{\mathrm{ik}}}\right)}{\partial h_{1_{\mathrm{ik}}}} \left(\left(\mathbf{A} \circ \left(2\left(\mathbf{y}_{\mathrm{pred}} - \mathbf{y}\right).\mathbf{w}_3^T\right)\right).\mathbf{w}_2^T\right)_{\mathrm{ik}} \tag{79}$$

The k$^{\mathrm{th}}$ column of $\frac{\partial \mathrm{L}}{\partial \mathbf{b}_1}$ is equal to the sum of the elements in the k$^{\mathrm{th}}$ column of

$\mathbf{C} \circ \left(\left(\mathbf{A} \circ \left(2\left(\mathbf{y}_{\text{pred}} - \mathbf{y}\right).\mathbf{w}_3^T\right)\right).\mathbf{w}_2^T\right)$.

The gradient results can be summarized as follows:

$$\frac{\partial \mathrm{L}}{\partial \mathbf{w}_3} \rightarrow \mathbf{h}_{2_{\text{relu}}}^T.2\left(\mathbf{y}_{\text{pred}} - \mathbf{y}\right) \tag{80}$$

$$\left(\frac{\partial \mathrm{L}}{\partial \mathbf{b}_3}\right)_{\mathrm{k}} = \sum_{\mathrm{i}=1}^{n} \left(2\left(\mathbf{y}_{\text{pred}} - \mathbf{y}\right)\right)_{\mathrm{ik}} \tag{81}$$

$$\frac{\partial \mathrm{L}}{\partial \mathbf{w}_2} \rightarrow \mathbf{h}_{1_{\text{relu}}}^T.\left(\mathbf{A} \circ \left(2\left(\mathbf{y}_{\text{pred}} - \mathbf{y}\right).\mathbf{w}_3^T\right)\right) \tag{82}$$

$$\left(\frac{\partial \mathrm{L}}{\partial \mathbf{b}_2}\right)_{\mathrm{k}} = \sum_{\mathrm{i}=1}^{n} \left(\mathbf{A} \circ \left(2\left(\mathbf{y}_{\text{pred}} - \mathbf{y}\right).\mathbf{w}_3^T\right)\right)_{\mathrm{ik}} \tag{83}$$

$$\frac{\partial \mathrm{L}}{\partial \mathbf{w}_1} \rightarrow \mathbf{x}^T.\left(\mathbf{C} \circ \left(\left(\mathbf{A} \circ \left(2\left(\mathbf{y}_{\text{pred}} - \mathbf{y}\right).\mathbf{w}_3^T\right)\right).\mathbf{w}_2^T\right)\right) \tag{84}$$

$$\left(\frac{\partial \mathrm{L}}{\partial \mathbf{b}_1}\right)_{\mathrm{k}} = \sum_{\mathrm{i}=1}^{n} (\mathbf{C} \circ \left(\left(\mathbf{A} \circ \left(2\left(\mathbf{y}_{\text{pred}} - \mathbf{y}\right).\mathbf{w}_3^T\right)\right).\mathbf{w}_2^T\right))_{\mathrm{ik}} \tag{85}$$

The numerical implementation of the above gradients, the updates of the weight matrices and the bias matrices are carried out in a separate Jupyter notebook.