AI Project

Robin MENEUST

August 2023

Table des matières

INTRODUCTION

- Let $a_i^{[l]}$ be the output i (neuron index) of the layer l. $l \in [\![0, n-1]\!]$ and $i \in [0, s(l)-1]$ where $s(l)$ is the size of the layer l.
- Let $z_i^{[l]}$ be the weighted output i of the layer l. $a_i^{[l]} = ReLU(z_i^{[l]}) = max(0, z_i^{[l]})$.

$$\begin{aligned} & Z^{[l]} = W^{[l]} \cdot A^{[l-1]} \\ \iff & \begin{pmatrix} z_0^{[l]} \\ z_1^{[l]} \\ ... \\ z_{s(l)}^{[l]} \end{pmatrix} = \begin{pmatrix} w_{0,0} & w_{0,1} & ... & w_{0,s(l-1)-1}^{[l]} \\ w_{1,0} & ... & ... & ... \\ ... & ... & ... & ... \\ w_{s(l)-1,0} & w_{0,1} & ... & w_{s(l)-1,s(l-1)-1}^{[l]} \end{pmatrix} \cdot \begin{pmatrix} a_0^{[l-1]} \\ a_1^{[l-1]} \\ ... \\ a_n^{[l-1]} \end{pmatrix} \end{aligned} \tag{1}$$

- Let $E_{tot} = \frac{1}{n} \sum_{i=0}^{n-1} E_i$ be the global error (loss), where $E_i = \left(a_i^{[n-1]} - t_i\right)^2$ is the error for the output i (t_i is the targetted value for the output i)
- Let $w_{i,j}^{[l]}$ be the weight between the neuron j of the layer $l-1$ and the neuron i of the layer l.

Back propagation

1 Useful derivatives

$$\frac{dReLU(x)}{dx} = \begin{cases} 0 \text{ if dx } \in \mathbb{R}_+^* \\ 1 \text{ if x } \in \mathbb{R}_-^* \end{cases} \tag{2}$$

$$\begin{aligned} \frac{\partial E_i}{\partial a_i^{[n-1]}} &= \frac{\partial (a_i^{[n-1]} - t_i)^2}{\partial a_i^{[n-1]}} \\ &= 2(a_i^{[n-1]} - t_i) \end{aligned} \tag{3}$$

$$\begin{aligned} \frac{\partial a_i^{[n-1]}}{\partial z_i^{[n-1]}} &= \frac{\partial ReLU(z_i^{[n-1]})}{\partial z_i^{[n-1]}} \\ &= \begin{cases} 0 \text{ if } z_i^{[n-1]} \in \mathbb{R}_-^* \\ 1 \text{ if } z_i^{[n-1]} \in \mathbb{R}_+^* \end{cases} \end{aligned} \tag{4}$$

$$\begin{aligned} \frac{\partial z_k^{[n-1]}}{\partial a_m^{[n-2]}} &= \frac{\partial \left(\sum\limits_{i=0}^{s(n-1)} w_{k,i}^{[n-1]} a_i^{[n-2]} \right)}{\partial a_m^{[n-2]}} \\ &= \sum_{i=0}^{s(n-1)} \left(\frac{\partial w_{k,i}^{[n-1]} a_i^{[n-2]}}{\partial a_m^{[n-2]}} \right) \\ &= w_{k,m}^{[n-1]} \end{aligned} \tag{5}$$

$$\begin{aligned} \frac{\partial z_k^{[n-1]}}{\partial w_{k,m}^{[n-1]}} &= \frac{\partial \left(\sum\limits_{i=0}^{s(n-1)} w_{k,i}^{[n-1]} a_i^{[n-2]} \right)}{\partial w_{k,m}^{[n-1]}} \\ &= \sum_{i=0}^{s(n-1)} \left(\frac{\partial w_{k,i}^{[n-1]} a_i^{[n-2]}}{\partial w_{k,m}^{[n-1]}} \right) \\ &= a_m^{[n-2]} \end{aligned} \tag{6}$$

2 Chain rule and more derivatives

$$\begin{aligned}
\frac{\partial a_k^{[n-1]}}{\partial a_m^{[n-2]}} &= \frac{\partial a_k^{[n-1]}}{\partial z_k^{[n-1]}} \cdot \frac{\partial z_k^{[n-1]}}{\partial a_m^{[n-2]}} \\
&= \begin{cases} 0 \text{ if } z_k^{[n-1]} \in \mathbb{R}_-^* \\ 1 \text{ if } z_k^{[n-1]} \in \mathbb{R}_+^* \end{cases} \cdot w_{k,m}^{[n-1]} \\
&= \begin{cases} 0 \text{ if } z_k^{[n-1]} \in \mathbb{R}_-^* \\ w_{k,m}^{[n-1]} \text{ if } z_k^{[n-1]} \in \mathbb{R}_+^* \end{cases}
\end{aligned} \tag{7}$$

$$\begin{aligned}
\frac{\partial a_k^{[n-1]}}{\partial w_{k,m}^{[n-1]}} &= \frac{\partial a_k^{[n-1]}}{\partial z_k^{[n-1]}} \cdot \frac{\partial z_k^{[n-1]}}{\partial w_{k,m}^{[n-1]}} \\
&= \begin{cases} 0 \text{ if } z_k^{[n-1]} \in \mathbb{R}_-^* \\ 1 \text{ if } z_k^{[n-1]} \in \mathbb{R}_+^* \end{cases} \cdot a_m^{[n-2]} \\
&= \begin{cases} 0 \text{ if } z_k^{[n-1]} \in \mathbb{R}_-^* \\ a_m^{[n-2]} \text{ if } z_k^{[n-1]} \in \mathbb{R}_+^* \end{cases}
\end{aligned} \tag{8}$$

$$\begin{aligned}
\frac{\partial E_i}{\partial z_i^{[n-1]}} &= \frac{\partial E_i}{\partial a_i^{[n-1]}} \cdot \frac{\partial a_i^{[n-1]}}{\partial z_i^{[n-1]}} \\
&= 2(a_i^{[n-1]} - t_i) \cdot \begin{cases} 0 \text{ if } z_i^{[n-1]} \in \mathbb{R}_-^* \\ 1 \text{ if } z_i^{[n-1]} \in \mathbb{R}_+^* \end{cases} \\
&= \begin{cases} 0 \text{ if } z_i^{[n-1]} \in \mathbb{R}_-^* \\ 2(a_i^{[n-1]} - t_i) \text{ if } z_i^{[n-1]} \in \mathbb{R}_+^* \end{cases}
\end{aligned} \tag{9}$$

$$\begin{aligned}
\frac{\partial E_i}{\partial a_m^{[n-2]}} &= \frac{\partial E_i}{\partial z_i^{[n-1]}} \cdot \frac{\partial z_i^{[n-1]}}{\partial a_m^{[n-2]}} \\
&= \begin{cases} 0 \text{ if } z_i^{[n-1]} \in \mathbb{R}_-^* \\ 2(a_i^{[n-1]} - t_i) \text{ if } z_i^{[n-1]} \in \mathbb{R}_+^* \end{cases} \cdot w_{i,m}^{[n-1]} \\
&= \begin{cases} 0 \text{ if } z_i^{[n-1]} \in \mathbb{R}_-^* \\ 2(a_i^{[n-1]} - t_i) \cdot w_{i,m}^{[n-1]} \text{ if } z_i^{[n-1]} \in \mathbb{R}_+^* \end{cases}
\end{aligned} \tag{10}$$

3 Steps

Step 1

First we evaluate, $\forall i \in [\![0, s(n-1)]\!]$, $\frac{\partial E_i}{\partial z_i^{[n-1]}}$ and we store these values in a temporary array to use them in the two following steps.

Step 2

We can now evaluate $\frac{\partial a_i^{[n-1]}}{\partial a_m^{[n-2]}}$ and use this value to calculate $\frac{\partial E_i}{\partial a_m^{[n-2]}} \forall i, m$. We store these values in a second temporary array since we will need them for the step 5.

Step 3

Similarly, we evaluate $\frac{\partial a_i^{[n-1]}}{\partial w_{i,j}^{[n-1]}}$ and use this value to calculate $D_{i,j} = \frac{\partial E_i}{\partial w_{i,j}^{[n-1]}} \forall i, j$. We don't store these values, but we use it to change the values $w_{i,j}$, that will become $w_{i,j} - (D_{i,j} \cdot learning\_rate)$. Modifying $learning\_rate$ will change the speed of the step in the gradient descent, it can be equal to 0.01 for instance.

Step 4

Delete the first temporary array (step 1), since we won't use these values anymore.

Step 5

We can now evaluate $\frac{\partial a_i^{[n-2]}}{\partial a_m^{[n-3]}}$ by using the array of the step 2, and use this value with the chain rule to calculate $\frac{\partial E_i}{\partial a_m^{[n-3]}} \forall i, m$. We store these values in a new array, and once they were all stored we can delete the array of the step 2.

Step 6

Similarly, we evaluate $\frac{\partial a_i^{[n-2]}}{\partial w_{i,j}^{[n-2]}}$ and use this value to calculate $D_{i,j} = \frac{\partial E_i}{\partial w_{i,j}^{[n-2]}} \forall i, j$. Just like before, we don't store these values, but we use it to change the values $w_{i,j}$, that will become $w_{i,j} - (D_{i,j} \cdot learning\_rate)$.

Next steps

Repeat the steps 5 and 6 until we reach the input layer. At the final step we should be calculating $\frac{\partial E_i}{\partial w_{i,j}^{[1]}}$. It's $[1]$ and not $[0]$ because the first layer is the input layer so there is not layer before, and thus no weights. The last weights that we change are those between stored in the layer $[1]$, that define the relation between the 2 first layers.