Feedforward Neural Networks in Depth, Part 2: Activation Functions

Dec 21, 2021

This is the second post of a three-part series in which we derive the mathematics behind feedforward neural networks. We worked our way through forward and backward propagations in the first post, but if you remember, we only mentioned activation functions in passing. In particular, we did not derive an analytic expression for $\partial a_{j,i}^{[l]} / \partial z_{j,i}^{[l]}$ or, by extension, $\partial J / \partial z_{j,i}^{[l]}$. So let us pick up the derivations where we left off.

ReLU

The rectified linear unit, or ReLU for short, is given by

$$\begin{aligned}
a_{j,i}^{[l]} &= g_j^{[l]}(z_{1,i}^{[l]}, \dots, z_{j,i}^{[l]}, \dots, z_{n^{[l]},i}^{[l]}) \\
&= \max(0, z_{j,i}^{[l]}) \\
&= \begin{cases} z_{j,i}^{[l]} & \text{if } z_{j,i}^{[l]} > 0, \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

In other words,

$$\mathbf{A}^{[l]} = \max(0, \mathbf{Z}^{[l]}). \tag{1}$$

Next, we compute the partial derivatives of the activations in the current layer:

$$\begin{aligned}
\frac{\partial a_{j,i}^{[l]}}{\partial z_{j,i}^{[l]}} &:= \begin{cases} 1 & \text{if } z_{j,i}^{[l]} > 0, \\ 0 & \text{otherwise,} \end{cases} \\
&= I(z_{j,i}^{[l]} > 0), \\
\frac{\partial a_{p,i}^{[l]}}{\partial z_{j,i}^{[l]}} &= 0, \quad \forall p \neq j.
\end{aligned}$$

input and output of 2 different units > 0

It follows that

$$
\begin{aligned}
\frac{\partial J}{\partial z_{j,i}^{[l]}} &= \sum_p \frac{\partial J}{\partial a_{p,i}^{[l]}} \frac{\partial a_{p,i}^{[l]}}{\partial z_{j,i}^{[l]}} \\
&= \frac{\partial J}{\partial a_{j,i}^{[l]}} \frac{\partial a_{j,i}^{[l]}}{\partial z_{j,i}^{[l]}} + \sum_{p \neq j} \frac{\partial J}{\partial a_{p,i}^{[l]}} \frac{\partial a_{p,i}^{[l]}}{\partial z_{j,i}^{[l]}} \\
&= \frac{\partial J}{\partial a_{j,i}^{[l]}} I(z_{j,i}^{[l]} > 0),
\end{aligned}
$$

which we can vectorize as

$$
\frac{\partial J}{\partial \mathbf{Z}^{[l]}} = \frac{\partial J}{\partial \mathbf{A}^{[l]}} \odot I(\mathbf{Z}^{[l]} > 0), \tag{2}
$$

where $\odot$ denotes element-wise multiplication.

Sigmoid

The sigmoid activation function is given by

$$
\begin{aligned}
a_{j,i}^{[l]} &= g_j^{[l]}(z_{1,i}^{[l]}, \dots, z_{j,i}^{[l]}, \dots, z_{n^{[l]},i}^{[l]}) \\
&= \sigma(z_{j,i}^{[l]}) \\
&= \frac{1}{1 + \exp(-z_{j,i}^{[l]})}.
\end{aligned}
$$

Vectorization yields

$$
\mathbf{A}^{[l]} = \frac{1}{1 + \exp(-\mathbf{Z}^{[l]})}. \tag{3}
$$

To practice backward propagation, first, we construct a computation graph:

$$
\begin{aligned}
u_0 &= z_{j,i}^{[l]}, \\
u_1 &= -u_0, \\
u_2 &= \exp(u_1), \\
u_3 &= 1 + u_2, \\
u_4 &= \frac{1}{u_3} = a_{j,i}^{[l]}.
\end{aligned}
$$

Then, we perform an outside first traversal of the chain rule:

$$
\begin{aligned}
\frac{\partial a_{j,i}^{[l]}}{\partial u_4} &= 1, \\
\frac{\partial a_{j,i}^{[l]}}{\partial u_3} &= \frac{\partial a_{j,i}^{[l]}}{\partial u_4}\frac{\partial u_4}{\partial u_3} = -\frac{1}{u_3^2} = -\frac{1}{(1+\exp(-z_{j,i}^{[l]}))^2}, \\
\frac{\partial a_{j,i}^{[l]}}{\partial u_2} &= \frac{\partial a_{j,i}^{[l]}}{\partial u_3}\frac{\partial u_3}{\partial u_2} = -\frac{1}{u_3^2} = -\frac{1}{(1+\exp(-z_{j,i}^{[l]}))^2}, \\
\frac{\partial a_{j,i}^{[l]}}{\partial u_1} &= \frac{\partial a_{j,i}^{[l]}}{\partial u_2}\frac{\partial u_2}{\partial u_1} = -\frac{1}{u_3^2}\exp(u_1) = -\frac{\exp(-z_{j,i}^{[l]})}{(1+\exp(-z_{j,i}^{[l]}))^2}, \\
\frac{\partial a_{j,i}^{[l]}}{\partial u_0} &= \frac{\partial a_{j,i}^{[l]}}{\partial u_1}\frac{\partial u_1}{\partial u_0} = \frac{1}{u_3^2}\exp(u_1) = \frac{\exp(-z_{j,i}^{[l]})}{(1+\exp(-z_{j,i}^{[l]}))^2}.
\end{aligned}
$$

Let us simplify:

$$
\begin{aligned}
\frac{\partial a_{j,i}^{[l]}}{\partial z_{j,i}^{[l]}} &= \frac{\exp(-z_{j,i}^{[l]})}{(1+\exp(-z_{j,i}^{[l]}))^2} \\
&= \frac{1+\exp(-z_{j,i}^{[l]})-1}{(1+\exp(-z_{j,i}^{[l]}))^2} \\
&= \frac{1}{1+\exp(-z_{j,i}^{[l]})} - \frac{1}{(1+\exp(-z_{j,i}^{[l]}))^2} \\
&= a_{j,i}^{[l]}(1-a_{j,i}^{[l]}).
\end{aligned}
$$

We also note that

$$
\frac{\partial a_{p,i}^{[l]}}{\partial z_{j,i}^{[l]}} = 0, \quad \forall p \neq j.
$$

Consequently,

$$
\begin{aligned}
\frac{\partial J}{\partial z_{j,i}^{[l]}} &= \sum_p \frac{\partial J}{\partial a_{p,i}^{[l]}}\frac{\partial a_{p,i}^{[l]}}{\partial z_{j,i}^{[l]}} \\
&= \frac{\partial J}{\partial a_{j,i}^{[l]}}\frac{\partial a_{j,i}^{[l]}}{\partial z_{j,i}^{[l]}} + \sum_{p\neq j} \frac{\partial J}{\partial a_{p,i}^{[l]}}\frac{\partial a_{p,i}^{[l]}}{\partial z_{j,i}^{[l]}} \\
&= \frac{\partial J}{\partial a_{j,i}^{[l]}} a_{j,i}^{[l]}(1-a_{j,i}^{[l]}).
\end{aligned}
$$

Lastly, no summations mean trivial vectorization:

$$\frac{\partial J}{\partial \mathbf{Z}^{[l]}} = \frac{\partial J}{\partial \mathbf{A}^{[l]}} \odot \mathbf{A}^{[l]} \odot (1 - \mathbf{A}^{[l]}). \tag{4}$$

Tanh

The hyperbolic tangent function, i.e., the tanh activation function, is given by

$$\begin{aligned} a_{j,i}^{[l]} &= g_j^{[l]}(z_{1,i}^{[l]}, \dots, z_{j,i}^{[l]}, \dots, z_{n^{[l]},i}^{[l]}) \\ &= \tanh(z_{j,i}^{[l]}) \\ &= \frac{\exp(z_{j,i}^{[l]}) - \exp(-z_{j,i}^{[l]})}{\exp(z_{j,i}^{[l]}) + \exp(-z_{j,i}^{[l]})}. \end{aligned}$$

By utilizing element-wise multiplication, we get

$$\mathbf{A}^{[l]} = \frac{1}{\exp(\mathbf{Z}^{[l]}) + \exp(-\mathbf{Z}^{[l]})} \odot (\exp(\mathbf{Z}^{[l]}) - \exp(-\mathbf{Z}^{[l]})). \tag{5}$$

Once again, let us introduce intermediate variables to practice backward propagation:

$$\begin{aligned} u_0 &= z_{j,i}^{[l]}, \\ u_1 &= -u_0, \\ u_2 &= \exp(u_0), \\ u_3 &= \exp(u_1), \\ u_4 &= u_2 - u_3, \\ u_5 &= u_2 + u_3, \\ u_6 &= \frac{1}{u_5}, \\ u_7 &= u_4 u_6 = a_{j,i}^{[l]}. \end{aligned}$$

Next, we compute the partial derivatives:

$$
\begin{aligned}
\frac{\partial a_{j,i}^{[l]}}{\partial u_7} &= 1, \\
\frac{\partial a_{j,i}^{[l]}}{\partial u_6} &= \frac{\partial a_{j,i}^{[l]}}{\partial u_7}\frac{\partial u_7}{\partial u_6} = u_4 = \exp(z_{j,i}^{[l]}) - \exp(-z_{j,i}^{[l]}), \\
\frac{\partial a_{j,i}^{[l]}}{\partial u_5} &= \frac{\partial a_{j,i}^{[l]}}{\partial u_6}\frac{\partial u_6}{\partial u_5} = -u_4\frac{1}{u_5^2} = -\frac{\exp(z_{j,i}^{[l]}) - \exp(-z_{j,i}^{[l]})}{(\exp(z_{j,i}^{[l]}) + \exp(-z_{j,i}^{[l]}))^2}, \\
\frac{\partial a_{j,i}^{[l]}}{\partial u_4} &= \frac{\partial a_{j,i}^{[l]}}{\partial u_7}\frac{\partial u_7}{\partial u_4} = u_6 = \frac{1}{\exp(z_{j,i}^{[l]}) + \exp(-z_{j,i}^{[l]})}, \\
\frac{\partial a_{j,i}^{[l]}}{\partial u_3} &= \frac{\partial a_{j,i}^{[l]}}{\partial u_4}\frac{\partial u_4}{\partial u_3} + \frac{\partial a_{j,i}^{[l]}}{\partial u_5}\frac{\partial u_5}{\partial u_3} \\
&= -u_6 - u_4\frac{1}{u_5^2} \\
&= -\frac{1}{\exp(z_{j,i}^{[l]}) + \exp(-z_{j,i}^{[l]})} - \frac{\exp(z_{j,i}^{[l]}) - \exp(-z_{j,i}^{[l]})}{(\exp(z_{j,i}^{[l]}) + \exp(-z_{j,i}^{[l]}))^2} \\
&= -\frac{2\exp(z_{j,i}^{[l]})}{(\exp(z_{j,i}^{[l]}) + \exp(-z_{j,i}^{[l]}))^2}, \\
\frac{\partial a_{j,i}^{[l]}}{\partial u_2} &= \frac{\partial a_{j,i}^{[l]}}{\partial u_4}\frac{\partial u_4}{\partial u_2} + \frac{\partial a_{j,i}^{[l]}}{\partial u_5}\frac{\partial u_5}{\partial u_2} \\
&= u_6 - u_4\frac{1}{u_5^2} \\
&= \frac{1}{\exp(z_{j,i}^{[l]}) + \exp(-z_{j,i}^{[l]})} - \frac{\exp(z_{j,i}^{[l]}) - \exp(-z_{j,i}^{[l]})}{(\exp(z_{j,i}^{[l]}) + \exp(-z_{j,i}^{[l]}))^2} \\
&= \frac{2\exp(-z_{j,i}^{[l]})}{(\exp(z_{j,i}^{[l]}) + \exp(-z_{j,i}^{[l]}))^2}, \\
\frac{\partial a_{j,i}^{[l]}}{\partial u_1} &= \frac{\partial a_{j,i}^{[l]}}{\partial u_3}\frac{\partial u_3}{\partial u_1} \\
&= \left(-u_6 - u_4\frac{1}{u_5^2}\right)\exp(u_1) \\
&= -\frac{2\exp(z_{j,i}^{[l]})\exp(-z_{j,i}^{[l]})}{(\exp(z_{j,i}^{[l]}) + \exp(-z_{j,i}^{[l]}))^2}, \\
\frac{\partial a_{j,i}^{[l]}}{\partial u_0} &= \frac{\partial a_{j,i}^{[l]}}{\partial u_1}\frac{\partial u_1}{\partial u_0} + \frac{\partial a_{j,i}^{[l]}}{\partial u_2}\frac{\partial u_2}{\partial u_0} \\
&= -\left(-u_6 - u_4\frac{1}{u_5^2}\right)\exp(u_1) + \left(u_6 - u_4\frac{1}{u_5^2}\right)\exp(u_0) \\
&= \frac{2\exp(z_{j,i}^{[l]})\exp(-z_{j,i}^{[l]})}{} + \frac{2\exp(z_{j,i}^{[l]})\exp(-z_{j,i}^{[l]})}{}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\quad}{(\exp(z^{[l]}_{j,i}) + \exp(-z^{[l]}_{j,i}))^2} \qquad \frac{\quad}{(\exp(z^{[l]}_{j,i}) + \exp(-z^{[l]}_{j,i}))^2} \\
&= \frac{4\exp(z^{[l]}_{j,i})\exp(-z^{[l]}_{j,i})}{(\exp(z^{[l]}_{j,i}) + \exp(-z^{[l]}_{j,i}))^2}.
\end{aligned}
$$

It follows that

$$
\begin{aligned}
\frac{\partial a^{[l]}_{j,i}}{\partial z^{[l]}_{j,i}} &= \frac{4\exp(z^{[l]}_{j,i})\exp(-z^{[l]}_{j,i})}{(\exp(z^{[l]}_{j,i}) + \exp(-z^{[l]}_{j,i}))^2} \\
&= \frac{\exp(z^{[l]}_{j,i})^2 + 2\exp(z^{[l]}_{j,i})\exp(-z^{[l]}_{j,i}) + \exp(-z^{[l]}_{j,i})^2}{(\exp(z^{[l]}_{j,i}) + \exp(-z^{[l]}_{j,i}))^2} \\
&\quad - \frac{\exp(z^{[l]}_{j,i})^2 - 2\exp(z^{[l]}_{j,i})\exp(-z^{[l]}_{j,i}) + \exp(-z^{[l]}_{j,i})^2}{(\exp(z^{[l]}_{j,i}) + \exp(-z^{[l]}_{j,i}))^2} \\
&= 1 - \frac{(\exp(z^{[l]}_{j,i}) - \exp(-z^{[l]}_{j,i}))^2}{(\exp(z^{[l]}_{j,i}) + \exp(-z^{[l]}_{j,i}))^2} \\
&= 1 - a^{[l]}_{j,i}a^{[l]}_{j,i}.
\end{aligned}
$$

Similiar to the sigmoid activation function, we also have

$$
\frac{\partial a^{[l]}_{p,i}}{\partial z^{[l]}_{j,i}} = 0, \quad \forall p \neq j.
$$

Thus,

$$
\begin{aligned}
\frac{\partial J}{\partial z^{[l]}_{j,i}} &= \sum_p \frac{\partial J}{\partial a^{[l]}_{p,i}} \frac{\partial a^{[l]}_{p,i}}{\partial z^{[l]}_{j,i}} \\
&= \frac{\partial J}{\partial a^{[l]}_{j,i}} \frac{\partial a^{[l]}_{j,i}}{\partial z^{[l]}_{j,i}} + \sum_{p\neq j} \frac{\partial J}{\partial a^{[l]}_{p,i}} \frac{\partial a^{[l]}_{p,i}}{\partial z^{[l]}_{j,i}} \\
&= \frac{\partial J}{\partial a^{[l]}_{j,i}}(1 - a^{[l]}_{j,i}a^{[l]}_{j,i}),
\end{aligned}
$$

which implies that

$$
\frac{\partial J}{\partial \mathbf{Z}^{[l]}} = \frac{\partial J}{\partial \mathbf{A}^{[l]}} \odot (1 - \mathbf{A}^{[l]} \odot \mathbf{A}^{[l]}). \tag{6}
$$

Softmax

The softmax activation function is given by

$$
\begin{aligned}
a_{j,i}^{[l]} &= g_j^{[l]}(z_{1,i}^{[l]}, \dots, z_{j,i}^{[l]}, \dots, z_{n^{[l]},i}^{[l]}) \\
&= \frac{\exp(z_{j,i}^{[l]})}{\sum_p \exp(z_{p,i}^{[l]})}.
\end{aligned}
$$

Vectorization results in

$$
\mathbf{A}^{[l]} = \frac{1}{\text{broadcast}(\underbrace{\sum_{\text{axis}=0} \exp(\mathbf{Z}^{[l]})}_{\text{row vector}})} \odot \exp(\mathbf{Z}^{[l]}). \tag{7}
$$

To begin with, we construct a computation graph for the jth activation of the current layer:

$$
\begin{aligned}
u_{-1} &= z_{j,i}^{[l]}, \\
u_{0,p} &= z_{p,i}^{[l]}, && \forall p \neq j, \\
u_1 &= \exp(u_{-1}), \\
u_{2,p} &= \exp(u_{0,p}), && \forall p \neq j, \\
u_3 &= u_1 + \sum_{p \neq j} u_{2,p}, \\
u_4 &= \frac{1}{u_3}, \\
u_5 &= u_1 u_4 = a_{j,i}^{[l]}.
\end{aligned}
$$

By applying the chain rule, we get

$$
\begin{aligned}
\frac{\partial a_{j,i}^{[l]}}{\partial u_5} &= 1, \\
\frac{\partial a_{j,i}^{[l]}}{\partial u_4} &= \frac{\partial a_{j,i}^{[l]}}{\partial u_5}\frac{\partial u_5}{\partial u_4} = u_1 = \exp(z_{j,i}^{[l]}), \\
\frac{\partial a_{j,i}^{[l]}}{\partial u_3} &= \frac{\partial a_{j,i}^{[l]}}{\partial u_4}\frac{\partial u_4}{\partial u_3} = -u_1\frac{1}{u_3^2} = -\frac{\exp(z_{j,i}^{[l]})}{(\sum_p \exp(z_{p,i}^{[l]}))^2}, \\
\frac{\partial a_{j,i}^{[l]}}{\partial u_1} &= \frac{\partial a_{j,i}^{[l]}}{\partial u_3}\frac{\partial u_3}{\partial u_1} + \frac{\partial a_{j,i}^{[l]}}{\partial u_5}\frac{\partial u_5}{\partial u_1} \\
&= -u_1\frac{1}{u_3^2} + u_4 \\
&= -\frac{\exp(z_{j,i}^{[l]})}{(\sum_p \exp(z_{p,i}^{[l]}))^2} + \frac{1}{\sum_p \exp(z_{p,i}^{[l]})}, \\
\frac{\partial a_{j,i}^{[l]}}{\partial u_{-1}} &= \frac{\partial a_{j,i}^{[l]}}{\partial u_1}\frac{\partial u_1}{\partial u_{-1}} \\
&= \left(-u_1\frac{1}{u_3^2} + u_4\right)\exp(u_{-1}) \\
&= -\frac{\exp(z_{j,i}^{[l]})^2}{(\sum_p \exp(z_{p,i}^{[l]}))^2} + \frac{\exp(z_{j,i}^{[l]})}{\sum_p \exp(z_{p,i}^{[l]})}.
\end{aligned}
$$

Next, we need to take into account that $z_{j,i}^{[l]}$ also affects other activations in the same layer:

$$
\begin{aligned}
u_{-1} &= z_{j,i}^{[l]}, \\
u_{0,p} &= z_{p,i}^{[l]}, && \forall p \neq j, \\
u_1 &= \exp(u_{-1}), \\
u_{2,p} &= \exp(u_{0,p}), && \forall p \neq j, \\
u_3 &= u_1 + \sum_{p \neq j} u_{2,p}, \\
u_4 &= \frac{1}{u_3}, \\
u_5 &= u_{2,p}u_4 = a_{p,i}^{[l]}, && \forall p \neq j.
\end{aligned}
$$

Backward propagation gives us the remaining partial derivatives:

$$
\begin{aligned}
\frac{\partial a_{p,i}^{[l]}}{\partial u_5} &= 1, \\
\frac{\partial a_{p,i}^{[l]}}{\partial u_4} &= \frac{\partial a_{p,i}^{[l]}}{\partial u_5}\frac{\partial u_5}{\partial u_4} = u_{2,p} = \exp(z_{p,i}^{[l]}), \\
\frac{\partial a_{p,i}^{[l]}}{\partial u_3} &= \frac{\partial a_{p,i}^{[l]}}{\partial u_4}\frac{\partial u_4}{\partial u_3} = -u_{2,p}\frac{1}{u_3^2} = -\frac{\exp(z_{p,i}^{[l]})}{(\sum_p \exp(z_{p,i}^{[l]}))^2}, \\
\frac{\partial a_{p,i}^{[l]}}{\partial u_1} &= \frac{\partial a_{p,i}^{[l]}}{\partial u_3}\frac{\partial u_3}{\partial u_1} = -u_{2,p}\frac{1}{u_3^2} = -\frac{\exp(z_{p,i}^{[l]})}{(\sum_p \exp(z_{p,i}^{[l]}))^2}, \\
\frac{\partial a_{p,i}^{[l]}}{\partial u_{-1}} &= \frac{\partial a_{p,i}^{[l]}}{\partial u_1}\frac{\partial u_1}{\partial u_{-1}} = -u_{2,p}\frac{1}{u_3^2}\exp(u_{-1}) = -\frac{\exp(z_{p,i}^{[l]})\exp(z_{j,i}^{[l]})}{(\sum_p \exp(z_{p,i}^{[l]}))^2}.
\end{aligned}
$$

We now know that

$$
\begin{aligned}
\frac{\partial a_{j,i}^{[l]}}{\partial z_{j,i}^{[l]}} &= -\frac{\exp(z_{j,i}^{[l]})^2}{(\sum_p \exp(z_{p,i}^{[l]}))^2} + \frac{\exp(z_{j,i}^{[l]})}{\sum_p \exp(z_{p,i}^{[l]})} \\
&= a_{j,i}^{[l]}(1 - a_{j,i}^{[l]}), \\
\frac{\partial a_{p,i}^{[l]}}{\partial z_{j,i}^{[l]}} &= -\frac{\exp(z_{p,i}^{[l]})\exp(z_{j,i}^{[l]})}{(\sum_p \exp(z_{p,i}^{[l]}))^2} \\
&= -a_{p,i}^{[l]}a_{j,i}^{[l]}, \quad \forall p \neq j.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\frac{\partial J}{\partial z_{j,i}^{[l]}} &= \sum_p \frac{\partial J}{\partial a_{p,i}^{[l]}}\frac{\partial a_{p,i}^{[l]}}{\partial z_{j,i}^{[l]}} \\
&= \frac{\partial J}{\partial a_{j,i}^{[l]}}\frac{\partial a_{j,i}^{[l]}}{\partial z_{j,i}^{[l]}} + \sum_{p \neq j}\frac{\partial J}{\partial a_{p,i}^{[l]}}\frac{\partial a_{p,i}^{[l]}}{\partial z_{j,i}^{[l]}} \\
&= \frac{\partial J}{\partial a_{j,i}^{[l]}}a_{j,i}^{[l]}(1 - a_{j,i}^{[l]}) - \sum_{p \neq j}\frac{\partial J}{\partial a_{p,i}^{[l]}}a_{p,i}^{[l]}a_{j,i}^{[l]} \\
&= a_{j,i}^{[l]}\left(\frac{\partial J}{\partial a_{j,i}^{[l]}}(1 - a_{j,i}^{[l]}) - \sum_{p \neq j}\frac{\partial J}{\partial a_{p,i}^{[l]}}a_{p,i}^{[l]}\right) \\
&= a_{j,i}^{[l]}\left(\frac{\partial J}{\partial a_{j,i}^{[l]}}(1 - a_{j,i}^{[l]}) - \sum_{p}\frac{\partial J}{\partial a_{p,i}^{[l]}}a_{p,i}^{[l]} + \frac{\partial J}{\partial a_{j,i}^{[l]}}a_{j,i}^{[l]}\right) \\
&= a_{j,i}^{[l]}\left(\frac{\partial J}{\partial a_{j,i}^{[l]}} - \sum_{p}\frac{\partial J}{\partial a_{p,i}^{[l]}}a_{p,i}^{[l]}\right),
\end{aligned}
$$

which we can vectorize as

$$\frac{\partial J}{\partial \mathbf{z}_{:,i}^{[l]}} = \mathbf{a}_{:,i}^{[l]} \odot \Big(\frac{\partial J}{\partial \mathbf{a}_{:,i}^{[l]}} - \underbrace{\mathbf{a}_{:,i}^{[l]\,\mathsf{T}} \frac{\partial J}{\partial \mathbf{a}_{:,i}^{[l]}}}_{\text{scalar}} \Big).$$

Let us not stop with the vectorization just yet:

$$\frac{\partial J}{\partial \mathbf{Z}^{[l]}} = \mathbf{A}^{[l]} \odot \Big(\frac{\partial J}{\partial \mathbf{A}^{[l]}} - \text{broadcast}\big(\underbrace{\sum_{\text{axis}=0} \frac{\partial J}{\partial \mathbf{A}^{[l]}} \odot \mathbf{A}^{[l]}}_{\text{row vector}} \big) \Big). \tag{8}$$

Subscribe

Jonas Lalin

Yet another blog about deep learning.