Solution for the seventh assignment — 27.11.2023

1 Backpropagation

1.1 First Question

Without using backpropagation we would need to calculate the chain gradients for each individual weight. When using backpropagation, we can reuse the gradient calculations from the previous layers.

1.2 Second Question

We can write the given network as

$$h^{(1)} = \text{ReLU}\left(a^{(1)}\right), \quad a^{(1)} = W^{(1)} \cdot x \tag{1}$$

$$h^{(2)} = \text{ReLU}\left(a^{(2)}\right), \quad a^{(2)} = W^{(2)} \cdot h^{(1)} \tag{2}$$

$$h^{(3)} = W^{(3)} \cdot h^{(2)} \tag{3}$$

with weights

$$W^{(1)} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad W^{(2)} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}, \quad W^{(3)} = \begin{pmatrix} -1 & 1 \end{pmatrix}. \tag{4}$$

With this, we can compute the *forward pass* as

- Layer 1:

$$a^{(1)} = W^{(1)} \cdot x = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \cdot 3 = \begin{pmatrix} 3 \\ -3 \end{pmatrix} \tag{5}$$

$$h^{(1)} = \text{ReLU}\left(a^{(1)}\right) = \text{ReLU}\left(\begin{pmatrix} 3 \\ -3 \end{pmatrix}\right) = \begin{pmatrix} 3 \\ 0 \end{pmatrix} \tag{6}$$

- Layer 2:

$$a^{(2)} = W^{(2)} \cdot h^{(1)} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} 3 \\ 0 \end{pmatrix} = \begin{pmatrix} -3 \\ 3 \end{pmatrix} \tag{7}$$

$$h^{(2)} = \text{ReLU}\left(a^{(2)}\right) = \text{ReLU}\left(\begin{pmatrix} -3 \\ 3 \end{pmatrix}\right) = \begin{pmatrix} 0 \\ 3 \end{pmatrix} \tag{8}$$

- Layer 3:

$$h^{(3)} = W^{(3)} \cdot h^{(2)} = \begin{pmatrix} -1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 3 \end{pmatrix} = 3 \tag{9}$$

The prediction of our network for $x = 3$ is $\hat{y} = h^{(3)} = 3$. The ground truth value is $y = 6$. Therefore, the loss can be computed as

$$J(w) = (y - \hat{y})^2 \tag{10}$$

$$= (6 - 3)^2 = 9 \tag{11}$$

Next, we can compute the *backward pass.* For this, we start off with the loss and propagate the gradients backwards:

- Loss: First we compute the gradient of the loss in respect to the input of the loss function $\hat{y} = h^{(3)}$:

$$\frac{\partial J(w)}{\partial h^{(3)}} = \frac{\partial}{\partial h^{(3)}} \left(y - h^{(3)}\right)^2 \tag{12}$$

$$= -2\left(y - h^{(3)}\right). \tag{13}$$

  Then we can plug in our values and get

$$\frac{\partial J(w)}{\partial h^{(3)}} = -2\,(6-3) = -6. \tag{14}$$

- Layer 3 weights: To obtain the gradients for the layer 3 weights, we can use the chain rule:

$$\frac{\partial J(w)}{\partial w^{(3)}_{1,1}} = \frac{\partial J(w)}{\partial h^{(3)}} \cdot \frac{\partial h^{(3)}}{\partial w^{(3)}_{1,1}} \tag{15}$$

  and

$$\frac{\partial J(w)}{\partial w^{(3)}_{1,2}} = \frac{\partial J(w)}{\partial h^{(3)}} \cdot \frac{\partial h^{(3)}}{\partial w^{(3)}_{1,2}}. \tag{16}$$

  Since we already know the first part, i.e., the derivative of the loss function with regards to $h^{(3)}$, we now only need to compute

$$\frac{\partial h^{(3)}}{\partial w^{(3)}_{1,1}} = \frac{\partial}{\partial w^{(3)}_{1,1}} \left(W^{(3)} \cdot h^{(2)}\right) = h^{(2)}_1 \tag{17}$$

  and

$$\frac{\partial h^{(3)}}{\partial w^{(3)}_{1,2}} = \frac{\partial}{\partial w^{(3)}_{1,2}} \left(W^{(3)} \cdot h^{(2)}\right) = h^{(2)}_2. \tag{18}$$

  Plugging in our values for $h^{(2)}$ from the forward pass, we get

$$\frac{\partial h^{(3)}}{\partial w^{(3)}_{1,1}} = 0 \quad \text{and} \quad \frac{\partial h^{(3)}}{\partial w^{(3)}_{1,2}} = 3. \tag{19}$$

  Via Equations (15) and (16) we then get

$$\frac{\partial J(w)}{\partial w^{(3)}_{1,1}} = \frac{\partial J(w)}{\partial h^{(3)}} \cdot \frac{\partial h^{(3)}}{\partial w^{(3)}_{1,1}} = -6 \cdot 0 = 0 \tag{20}$$

  and

$$\frac{\partial J(w)}{\partial w^{(3)}_{1,2}} = \frac{\partial J(w)}{\partial h^{(3)}} \cdot \frac{\partial h^{(3)}}{\partial w^{(3)}_{1,2}} = -6 \cdot 3 = -18. \tag{21}$$

- Layer 2 activation inputs ($a^{(2)}$): We again apply the chain rule to compute
$$\frac{\partial J(w)}{\partial a_1^{(2)}} = \frac{\partial J(w)}{\partial h^{(3)}} \cdot \frac{\partial h^{(3)}}{\partial h_1^{(2)}} \cdot \frac{\partial h_1^{(2)}}{\partial a_1^{(2)}} \tag{22}$$
and
$$\frac{\partial J(w)}{\partial a_2^{(2)}} = \frac{\partial J(w)}{\partial h^{(3)}} \cdot \frac{\partial h^{(3)}}{\partial h_2^{(2)}} \cdot \frac{\partial h_2^{(2)}}{\partial a_2^{(2)}}. \tag{23}$$
Again, the first term is known. Therefore, we need to determine the derivatives of third layer outputs with respect to its inputs
$$\frac{\partial h^{(3)}}{\partial h_1^{(2)}} = \frac{\partial}{\partial h_1^{(2)}} \left(W^{(3)} \cdot h^{(2)}\right) = w_{1,1}^{(3)} = -1, \tag{24}$$
and
$$\frac{\partial h^{(3)}}{\partial h_2^{(2)}} = \frac{\partial}{\partial h_2^{(2)}} \left(W^{(3)} \cdot h^{(2)}\right) = w_{1,2}^{(3)} = 1, \tag{25}$$
as well as the derivatives of the activation function on the second layer
$$\frac{\partial h_1^{(2)}}{\partial a_1^{(2)}} = \begin{Bmatrix} 1 & \text{iff } a_1^{(2)} > 0 \\ 0 & \text{else} \end{Bmatrix} = 0 \tag{26}$$
and
$$\frac{\partial h_2^{(2)}}{\partial a_2^{(2)}} = \begin{Bmatrix} 1 & \text{iff } a_2^{(2)} > 0 \\ 0 & \text{else} \end{Bmatrix} = 1 \tag{27}$$
Thus the gradient values according to equations (22) and (23) are
$$\frac{\partial J(w)}{\partial a_1^{(2)}} = \frac{\partial J(w)}{\partial h^{(3)}} \cdot \frac{\partial h^{(3)}}{\partial h_1^{(2)}} \cdot \frac{\partial h_1^{(2)}}{\partial a_1^{(2)}} = -6 \cdot (-1) \cdot 0 = 0 \tag{28}$$
and
$$\frac{\partial J(w)}{\partial a_2^{(2)}} = \frac{\partial J(w)}{\partial h^{(3)}} \cdot \frac{\partial h^{(3)}}{\partial h_2^{(2)}} \cdot \frac{\partial h_2^{(2)}}{\partial a_2^{(2)}} = -6 \cdot 1 \cdot 1 = -6. \tag{29}$$

- Layer 2 weights (we can again plug in the values we computed in the previous steps):
$$\frac{\partial J(w)}{\partial w_{1,1}^{(2)}} = \frac{\partial J(w)}{\partial a_1^{(2)}} \cdot \frac{\partial a_1^{(2)}}{\partial w_{1,1}^{(2)}} = \frac{\partial J(w)}{\partial a_1^{(2)}} \cdot h_1^{(1)} = 0 \cdot 3 = 0 \tag{30}$$
$$\frac{\partial J(w)}{\partial w_{1,2}^{(2)}} = \frac{\partial J(w)}{\partial a_1^{(2)}} \cdot \frac{\partial a_1^{(2)}}{\partial w_{1,2}^{(2)}} = \frac{\partial J(w)}{\partial a_1^{(2)}} \cdot h_2^{(1)} = 0 \cdot 0 = 0 \tag{31}$$
$$\frac{\partial J(w)}{\partial w_{2,1}^{(2)}} = \frac{\partial J(w)}{\partial a_2^{(2)}} \cdot \frac{\partial a_2^{(2)}}{\partial w_{2,1}^{(2)}} = \frac{\partial J(w)}{\partial a_2^{(2)}} \cdot h_1^{(1)} = -6 \cdot 3 = -18 \tag{32}$$
$$\frac{\partial J(w)}{\partial w_{2,2}^{(2)}} = \frac{\partial J(w)}{\partial a_2^{(2)}} \cdot \frac{\partial a_2^{(2)}}{\partial w_{2,2}^{(2)}} = \frac{\partial J(w)}{\partial a_2^{(2)}} \cdot h_2^{(1)} = -6 \cdot 0 = 0 \tag{33}$$

- Layer 1 activation inputs ($a^{(1)}$): We apply the chain rule and compute

$$\begin{aligned}
\frac{\partial J(w)}{\partial a_1^{(1)}} &= \frac{\partial J(w)}{\partial a_1^{(2)}} \cdot \frac{\partial a_1^{(2)}}{\partial a_1^{(1)}} + \frac{\partial J(w)}{\partial a_2^{(2)}} \cdot \frac{\partial a_2^{(2)}}{\partial a_1^{(1)}} && (34)\\
&= \frac{\partial J(w)}{\partial a_1^{(2)}} \cdot \frac{\partial a_1^{(2)}}{\partial h_1^{(1)}} \cdot \frac{\partial h_1^{(1)}}{\partial a_1^{(1)}} + \frac{\partial J(w)}{\partial a_2^{(2)}} \cdot \frac{\partial a_2^{(2)}}{\partial h_1^{(1)}} \cdot \frac{\partial h_1^{(1)}}{\partial a_1^{(1)}} && (35)\\
&= \frac{\partial J(w)}{\partial a_1^{(2)}} \cdot w_{1,1}^{(2)} \cdot \begin{Bmatrix} 1 & \text{iff } a_1^{(1)} > 0 \\ 0 & \text{else} \end{Bmatrix}\\
&\qquad + \frac{\partial J(w)}{\partial a_2^{(2)}} \cdot w_{2,1}^{(2)} \cdot \begin{Bmatrix} 1 & \text{iff } a_1^{(1)} > 0 \\ 0 & \text{else} \end{Bmatrix} && (36)\\
&= 0 \cdot (-1) \cdot 1 + (-6) \cdot 1 \cdot 1 && (37)\\
&= -6 && (38)
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial J(w)}{\partial a_2^{(1)}} &= \frac{\partial J(w)}{\partial a_1^{(2)}} \cdot \frac{\partial a_1^{(2)}}{\partial a_2^{(1)}} + \frac{\partial J(w)}{\partial a_2^{(2)}} \cdot \frac{\partial a_2^{(2)}}{\partial a_2^{(1)}} && (39)\\
&= \frac{\partial J(w)}{\partial a_1^{(2)}} \cdot \frac{\partial a_1^{(2)}}{\partial h_2^{(1)}} \cdot \frac{\partial h_2^{(1)}}{\partial a_2^{(1)}} + \frac{\partial J(w)}{\partial a_2^{(2)}} \cdot \frac{\partial a_2^{(2)}}{\partial h_2^{(1)}} \cdot \frac{\partial h_2^{(1)}}{\partial a_2^{(1)}} && (40)\\
&= \frac{\partial J(w)}{\partial a_1^{(2)}} \cdot w_{1,2}^{(2)} \cdot \begin{Bmatrix} 1 & \text{iff } a_2^{(1)} > 0 \\ 0 & \text{else} \end{Bmatrix}\\
&\qquad + \frac{\partial J(w)}{\partial a_2^{(2)}} \cdot w_{2,2}^{(2)} \cdot \begin{Bmatrix} 1 & \text{iff } a_2^{(1)} > 0 \\ 0 & \text{else} \end{Bmatrix} && (41)\\
&= 0 \cdot 1 \cdot 0 + (-6) \cdot (-1) \cdot 0 && (42)\\
&= 0 && (43)
\end{aligned}$$

Note that this time around we have two paths in our computation graph to get from $a_1^{(1)}$ to the loss (summed up in the last layer), leading to the two terms being added (the same holds for $a_2^{(1)}$).

- Layer 1 weights (similar to before):

$$\frac{\partial J(w)}{\partial w_{1,1}^{(1)}} = \frac{\partial J(w)}{\partial a_1^{(1)}} \cdot \frac{\partial a_1^{(1)}}{\partial w_{1,1}^{(1)}} = \frac{\partial J(w)}{\partial a_2^{(1)}} \cdot x = -6 \cdot 3 = -18 \tag{44}$$

$$\frac{\partial J(w)}{\partial w_{1,2}^{(1)}} = \frac{\partial J(w)}{\partial a_2^{(1)}} \cdot \frac{\partial a_2^{(1)}}{\partial w_{1,2}^{(1)}} = \frac{\partial J(w)}{\partial a_2^{(1)}} \cdot x = 0 \cdot 3 = 0 \tag{45}$$

With this, we now have all gradient values for each weight. We can therefore compute the updated weights as

$$w_{j,k}^{(i)} \leftarrow w_{j,k}^{(i)} - \eta \cdot \frac{\partial J(w)}{\partial w_{j,k}^{(i)}} \tag{46}$$

Skipping all weights with zero gradients (those values remain the same), we get for $\eta = 0.01$

$$w_{1,2}^{(3)} \leftarrow 1 - 0.1 \cdot (-18) = 1.18 \tag{47}$$

$$w_{2,1}^{(2)} \leftarrow 1 - 0.1 \cdot (-18) = 1.18 \tag{48}$$

$$w_{1,1}^{(1)} \leftarrow 1 - 0.1 \cdot (-18) = 1.18 \tag{49}$$

2 Convolutional Neural Networks (CNNs)

2.1 First Question

A feed forward neural network is a neural network where the units are fully connected. Every individual unit in a layer l is connected to all the previous units in the layer $l-1$.

A convolutional neural network features sliding kernels that offer local and not full connectivity. The kernel is applied over the input to find patterns and the pattern is valued the same at every location of the input that it can be matched.

2.2 Second Question

1. Output dimensions with stride 1: $2 \times 3 \times 3$

2. Output dimensions with stride 2: $2 \times 2 \times 2$

2.3 Third Question

Given:

$$I = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 1 & 5 & 3 \end{pmatrix}, \quad K = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{50}$$

we need to compute the output matrix O:

$$O = \begin{pmatrix} O_{1,1} & O_{1,2} \\ O_{2,1} & O_{2,2} \end{pmatrix} \tag{51}$$

For computing $O_{1,1}$ we multiply element-wise the following submatrix of I and the kernel:

$$\begin{pmatrix} 1 & 2 \\ 4 & 5 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix} \tag{52}$$

Then we add the elements of the matrix, thus $O_{1,1} = 1+0+0+5 = 6$. Similarly we compute the values of the other elements:

$$\begin{pmatrix} 2 & 3 \\ 5 & 6 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 6 \end{pmatrix} \rightarrow O_{1,2} = 2 + 6 = 8 \tag{53}$$

$$\begin{pmatrix}4 & 5\\ 1 & 5\end{pmatrix} \cdot \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} = \begin{pmatrix}4 & 0\\ 0 & 5\end{pmatrix} \to O_{2,1} = 4 + 5 = 9 \tag{54}$$

$$\begin{pmatrix}5 & 6\\ 5 & 3\end{pmatrix} \cdot \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} = \begin{pmatrix}5 & 0\\ 0 & 3\end{pmatrix} \to O_{2,2} = 5 + 3 = 8 \tag{55}$$

The final matrix is:

$$O = \begin{pmatrix}O_{1,1} & O_{1,2}\\ O_{2,1} & O_{2,2}\end{pmatrix} = \begin{pmatrix}6 & 8\\ 9 & 8\end{pmatrix} \tag{56}$$

3 Neural Networks and Regularization

We have given the loss

$$\tilde{J} = J(\mathbf{w}) + \alpha \|\mathbf{w}\|_p^p \tag{57}$$

and p-norm of the weights as

$$\|\mathbf{w}\|_p^p = \sum_i w_i^p; \tag{58}$$

Our task is to compute the gradient $\nabla_w \tilde{J}$. For this, we can start with

$$\nabla_{\mathbf{w}} \tilde{J} = \nabla_{\mathbf{w}} J + \nabla_{\mathbf{w}} \left(\alpha \|\mathbf{w}\|_p^p\right) \tag{59}$$

and then compute the element-wise gradients independently as

$$\begin{aligned}
\nabla_{w_i} \tilde{J} &= \nabla_{w_i} J + \alpha \nabla_{w_i} \|\mathbf{w}\|_p^p &(60)\\
&= \nabla_{w_i} J + \alpha \nabla_{w_i} \sum_j w_j^p &(61)\\
&= \nabla_{w_i} J + \alpha \cdot \nabla_{w_i} w_i^p &(62)\\
&= \nabla_{w_i} J + \alpha \cdot p \cdot w_i^{p-1} &(63)
\end{aligned}$$