Submission for Deep Learning Exercise 3

Team: dl2022-ryd
Students: Yumna Ali, Deepu K Reddy, Rean Fernandes

November 8, 2022

1 Task 1

1.1 a

We first specify the forward pass equations for the sake of clarity.

$$z_0 = w_0 \cdot x \tag{1a}$$

$$h_0 = g_0(z_0) = max(0, z_0) \tag{1b}$$

$$z_1 = w_1 \cdot h_0 \tag{1c}$$

$$h_1 = g_1(z_1) = max(0, z_1) \tag{1d}$$

$$z_2 = w_s \cdot h_0 + w_2 \cdot h_1 \tag{1e}$$

$$\hat{y} = z_2 \tag{1f}$$

$$\mathscr{L}(y, \hat{y}) = |(y - \hat{y})| \tag{1g}$$

To show the backpropagation we will write the derivative with respect to all the variables for the sake of clarity again, starting from the bottom i.e. the output.

Output Layer:

$$\frac{\partial \mathscr{L}(y,\hat{y})}{\partial \hat{y}} = \frac{\partial\,|y-\hat{y})|}{\partial z_2} = \frac{(y-\hat{y})}{|(y-\hat{y})|} \tag{2}$$

$$\frac{\partial \mathscr{L}(y,\hat{y})}{\partial w_s} = \frac{\partial \mathscr{L}(y,\hat{y})}{\partial z_2} \cdot \frac{\partial z_2}{\partial w_s} \tag{3a}$$

$$\frac{\partial \mathscr{L}(y,\hat{y})}{\partial w_2} = \frac{\partial \mathscr{L}(y,\hat{y})}{\partial z_2} \cdot \frac{\partial z_2}{\partial w_2} \tag{3b}$$

$$\frac{\partial \mathscr{L}(y,\hat{y})}{\partial h_0} = \frac{\partial \mathscr{L}(y,\hat{y})}{\partial z_2} \cdot \frac{\partial z_2}{\partial h_0} \tag{3c}$$

$$\frac{\partial \mathscr{L}(y,\hat{y})}{\partial h_1} = \frac{\partial \mathscr{L}(y,\hat{y})}{\partial z_2} \cdot \frac{\partial z_2}{\partial h_1} \tag{3d}$$

Layer 1:

$$\frac{\partial \mathscr{L}(y,\hat{y})}{\partial z_1} = \frac{\partial \mathscr{L}(y,\hat{y})}{\partial h_1} \cdot \frac{\partial h_1}{\partial g_1(z_1)} \cdot \frac{\partial g_1(z_1)}{\partial z_1} \tag{4}$$

$$\frac{\partial \mathscr{L}(y,\hat{y})}{\partial w_1} = \frac{\partial \mathscr{L}(y,\hat{y})}{\partial z_1} \cdot \frac{\partial z_1}{\partial w_1} \tag{5a}$$

$$\frac{\partial \mathscr{L}(y,\hat{y})}{\partial h_0} = \frac{\partial \mathscr{L}(y,\hat{y})}{\partial z_1} \cdot \frac{\partial z_1}{\partial h_0} + \frac{\partial \mathscr{L}(y,\hat{y})}{\partial z_2} \cdot \frac{\partial z_2}{\partial h_0} \tag{5b}$$

Layer 0:

$$\frac{\partial \mathscr{L}(y,\hat{y})}{\partial z_0} = \frac{\partial \mathscr{L}(y,\hat{y})}{\partial h_0} \cdot \frac{\partial h_0}{\partial g_0(z_0)} \cdot \frac{\partial g_0(z_0)}{\partial z_0} \tag{6}$$

$$\frac{\partial \mathscr{L}(y,\hat{y})}{\partial x} = \frac{\partial \mathscr{L}(y,\hat{y})}{\partial z_0} \cdot \frac{\partial z_0}{\partial x} \tag{7a}$$

$$\frac{\partial \mathscr{L}(y,\hat{y})}{\partial w_0} = \frac{\partial \mathscr{L}(y,\hat{y})}{\partial z_0} \cdot \frac{\partial z_0}{\partial w_0} \tag{7b}$$

We note that we do not need to calculate the gradient for the input, as it does not change except in the case of backpropagation for GANs, we have just worked this out to better show how we have applied the chain rule here. The addition of 3a to 5b is due to the backward gradient being added from the skip connection. Finally, the final gradients with respect to the weights after simplification is:

$$\frac{\partial \mathscr{L}(y,\hat{y})}{\partial w_2} = \frac{(y-\hat{y})}{|(y-\hat{y})|} \cdot h_1 \tag{8a}$$

$$\frac{\partial \mathscr{L}(y,\hat{y})}{\partial w_s} = \frac{(y-\hat{y})}{|(y-\hat{y})|} \cdot h_0 \tag{8b}$$

$$\frac{\partial \mathscr{L}(y,\hat{y})}{\partial w_1} = \frac{(y-\hat{y})}{|(y-\hat{y})|} \cdot w_2 \cdot h_0 \cdot ReLU'(z_1) \tag{8c}$$

$$\frac{\partial \mathscr{L}(y,\hat{y})}{\partial w_0} = \frac{(y-\hat{y})}{|(y-\hat{y})|} \cdot w_2 \cdot (1 + w_1 \cdot ReLU'(z_1)) \cdot ReLU'(z_0) \cdot x \tag{8d}$$

Where

$$ReLU'(x) = \begin{cases} 0 & : \ x \leq 0 \\ 1 & : \ x < 0 \end{cases} \tag{9}$$

1.2 b

The adding of the skip connection leads prevents the gradient from shrinking exceedingly, as a larger number of layers will lead to the issue of vanishing gradients. Even if the gradient shrinks due to the many multiplications that will be done during backpropagation, the gradient of the final output with respect to the node from which the skip connection passes forward will still be add to the total gradient.

1.3 c

Performing one forwards pass with the given initial values:

$$z_0 = w_0 \cdot x = 1 \cdot 0.5 = 0.5 \tag{10a}$$

$$h_0 = g_0(z_0) = max(0, z_0) = max(0, 0.5) = 0.5 \tag{10b}$$

$$z_1 = w_1 \cdot h_0 = 0.5 \cdot 0.5 = 0.25 \tag{10c}$$

$$h_1 = g_1(z_1) = max(0, z_1) = max(0, 0.25) = 0.25 \tag{10d}$$

$$z_2 = w_s \cdot h_0 + w_2 \cdot h_1 = 0.5 \cdot 0.5 + 0.5 \cdot 0.25 = 0.375 \tag{10e}$$

$$\hat{y} = z_2 = 0.375 \tag{10f}$$

$$\mathscr{L}(y,\hat{y}) = |(y-\hat{y})| = |(-3 - 0.375)| = 3.375 \tag{10g}$$

Now to calculate the gradients, performing the backward pass and substituting the equations obtained from 8,

$$\frac{\partial \mathscr{L}(y,\hat{y})}{\partial w_2} = \frac{(y-\hat{y})}{|(y-\hat{y})|} \cdot h_1 = \frac{-3.375}{|-3.375|} \cdot 0.25 = -0.25 \tag{11a}$$

$$\frac{\partial \mathscr{L}(y,\hat{y})}{\partial w_s} = \frac{(y-\hat{y})}{|(y-\hat{y})|} \cdot h_0 = \frac{-3.375}{|-3.375|} \cdot 0.5 = -0.5 \tag{11b}$$

$$\frac{\partial \mathscr{L}(y,\hat{y})}{\partial w_1} = \frac{(y-\hat{y})}{|(y-\hat{y})|} \cdot w_2 \cdot h_0 \cdot ReLU'(z_1) = \frac{-3.375}{|-3.375|} \cdot 0.5 \cdot 0.5 \cdot 1 = -0.25 \tag{11c}$$

$$\frac{\partial \mathscr{L}(y,\hat{y})}{\partial w_0} = \frac{(y-\hat{y})}{|(y-\hat{y})|} \cdot w_2 \cdot (1 + w_1 \cdot ReLU'(z_1)) \cdot ReLU'(z_0) \cdot x = \frac{-3.375}{|-3.375|} \cdot 0.5 \cdot (1+0.5\cdot 1) \cdot 1 \cdot 1 = -0.75 \tag{11d}$$

Finally performing gradient descent with the obtained gradients, the new weights are now :

$$w_0^{new} = w_0 - \alpha \cdot \frac{\partial \mathscr{L}(y,\hat{y})}{\partial w_0} = 0.5 - 1 \cdot -0.75 = 1.25 \tag{12a}$$

$$w_1^{new} = w_1 - \alpha \cdot \frac{\partial \mathscr{L}(y,\hat{y})}{\partial w_1} = 0.5 - 1 \cdot -.25 = 0.75 \tag{12b}$$

$$w_2^{new} = w_2 - \alpha \cdot \frac{\partial \mathscr{L}(y,\hat{y})}{\partial w_2} = 0.5 - 1 \cdot -0.25 = 0.75 \tag{12c}$$

$$w_s^{new} = w_s - \alpha \cdot \frac{\partial \mathscr{L}(y,\hat{y})}{\partial w_s} = 0.5 - 1 \cdot -0.5 = 1.00 \tag{12d}$$

2 Coding assignment 1

3 Coding assignment 2

4 Questions on Experiments

4.1 a

The loss should have gone down, but it remains the same after one step. However over the subsequent steps it reduces.

4.2 b

The loss decreases after multiple steps for our current problem, however it need not always decrease. this could be due to exceedingly small changes in the gradient, which might lead to stagnating of the descent, or a local minima.

4.3 c

Not every run gives correct results, this might be due to local minima.

4.4 d

The variable lr is the learning rate, which decides the rate at which we perform our gradient descent. We would set the learning rate higher when we are further away from the minimum to ensure convergence faster, and smaller when we are close to converging in order to not go beyond the minimum, which could lead to oscillations around the minimum, in which case we might never reach the minimum.