Albert-Ludwigs-Universität Freiburg

Foundations of Deep Learning

Winter Semester 2018/2019

Exercise 02

Group members: Eduardo Alvarado, Manav Madan, Jatin Dhawan

November 3, 2018

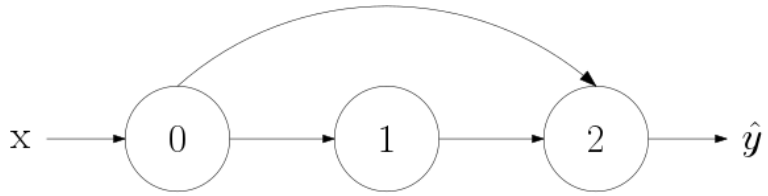
Part I

Pen and Paper Backpropagation

Answer the questions without python code.

(a) Backpropagation algorithm.

Perform a forward and backward pass to calculate the gradients for the weights w_0, w_1, w_2, w_s in the following MLP. Each node represents one unit with a weight $w_i, i \in \{0, 1, 2\}$ connecting it to the previous node. The connection from unit 0 to unit 2 is called a *skip connection*, which means unit 2 receives input from two sources and thus has an additional weight w_s. The weighted inputs are added before the nonlinearity is applied.



We assume that we want to solve a regression task.We use an L1-loss $L(\hat{y}, y) = |y - \hat{y}|$

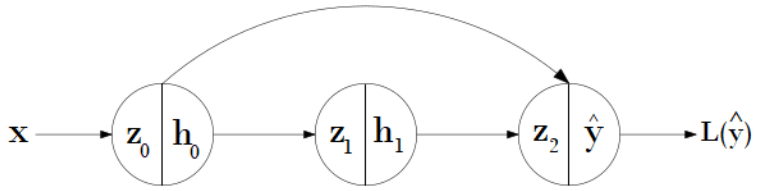
The nonlinearities for the first two units are rectified linear functions/units (ReLU): $g_0(z) = g_1(z) = \begin{cases} 0, z < 0 \\ z, else \end{cases}$.

We do not use a nonlinearity for the second unit: $g_2(z_2) = z_2$.

Note: We use the notation of the Deep Learning book here, i.e. $z = Wx + b$. If you attended the Machine Learning course, you might be used to the different notation used in the Bishop Book, where z denotes the value after applying the activation function. Here, z is the value before applying the activation function.

Solution:

The MLP can be represented in the following way, for better understanding:



We will assume that there is a general bias b for each layer, and w_s belongs to the skip connection.

The Forward and Backward Passes are described by:

Forward Pass:	**Backward Pass:**
$\hat{y} = g_2(z_2) = g_{linear}(z_2)$	$\frac{\partial L}{\partial \hat{y}} = \frac{\hat{y} - y}{\lvert \hat{y} - y \rvert}$
$z_2 = w_2 h_1 + w_s h_0 + b_2$	$\frac{\partial L}{\partial z_2} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_2} = \frac{\partial L}{\partial \hat{y}} \cdot g_2^{'}(z_2) = \frac{\partial L}{\partial \hat{y}} \cdot 1$
	$\rightarrow \frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial z_2} \frac{\partial z_2}{\partial w_2} = \frac{\partial L}{\partial z_2} \cdot h_1$
	$\rightarrow \frac{\partial L}{\partial w_s} = \frac{\partial L}{\partial z_2} \frac{\partial z_2}{\partial w_s} = \frac{\partial L}{\partial z_2} \cdot h_0$
	$\rightarrow \frac{\partial L}{\partial b_2} = \frac{\partial L}{\partial z_2} \frac{\partial z_2}{\partial b_2} = \frac{\partial L}{\partial z_2} \cdot 1$
$h_1 = g_1(z_1) = g_{ReLu}(z_1)$	$\frac{\partial L}{\partial h_1} = \frac{\partial L}{\partial z_2} \frac{\partial z_2}{\partial h_1} = \frac{\partial L}{\partial z_2} \cdot w_2$
$z_1 = w_1 h_0 + b_1$	$\frac{\partial L}{\partial z_1} = \frac{\partial L}{\partial h_1} \frac{\partial h_1}{\partial z_1} = \frac{\partial L}{\partial h_1} \cdot g_1^{'}(z_1) =$
	$= \begin{cases} \frac{\partial L}{\partial h_1}, & \text{if } z_1 \geq 0. \\ 0, & \text{otherwise.} \end{cases}$
	$\rightarrow \frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial z_1} \frac{\partial z_1}{\partial w_1} = \frac{\partial L}{\partial z_1} \cdot h_0$
	$\rightarrow \frac{\partial L}{\partial b_1} = \frac{\partial L}{\partial z_1} \frac{\partial z_1}{\partial b_1} = \frac{\partial L}{\partial z_1} \cdot 1$
$h_0 = g_0(z_0) = g_{ReLu}(z_0)$	$\frac{\partial L}{\partial h_0} = \frac{\partial L}{\partial z_1} \frac{\partial z_1}{\partial h_0} + \frac{\partial L}{\partial z_2} \frac{\partial z_2}{\partial h_0} =$
	$= \frac{\partial L}{\partial z_1} \cdot w_1 + \frac{\partial L}{\partial z_2} \cdot w_s$
$z_0 = w_0 x + b_0$	$\frac{\partial L}{\partial z_0} = \frac{\partial L}{\partial h_0} \frac{\partial h_0}{\partial z_0} = \frac{\partial L}{\partial h_0} \cdot g_0^{'}(z_0) =$
	$= \begin{cases} \frac{\partial L}{\partial h_0}, & \text{if } z_0 \geq 0. \\ 0, & \text{otherwise.} \end{cases}$
	$\rightarrow \frac{\partial L}{\partial w_0} = \frac{\partial L}{\partial z_0} \frac{\partial z_0}{\partial w_0} = \frac{\partial L}{\partial z_0} \cdot x$
	$\rightarrow \frac{\partial L}{\partial b_0} = \frac{\partial L}{\partial z_0} \frac{\partial z_0}{\partial b_0} = \frac{\partial L}{\partial z_0} \cdot 1$

What difference does the skip connection make when propagating back the error?

With this skip connection, the third layer receives input from two different sources, such that the gradient flows not only from the third to the second layer in the backward pass, but also from the third to the first layer. This adds an additional term in the equations.

The main reason why this is done, is to create a bypass to reduce the **vanishing gradient problem**. When we have a network with many layers, it can happen that the magnitude of the gradients shrink towards the first layers, reaching a point where they are not trained.

Thank to this bypass, we propagate the gradient without any interruption from the first layers, avoiding non-trained layers due to the vanishing problem.

(b) Gradient descent.

Calculate the gradients for the given datapoint and the given initial weights (calculating the gradients requires to calculate a forward pass first). Also calculate the weights and the loss after one gradient descent step.

$$(x_1, y_1) = (1, -3)$$
$$w_0 = w_1 = w_2 = w_s = 0.5$$
$$LearningRate = 1$$

Solution:

$z_0 = w_0 \cdot x = 0.5 \cdot 1 = 0.5 \rightarrow z_0 = 0.5$
$h_0 = g_0(z_0) = g_{relu}(z_0) = g_{relu}(0.5) = 0.5 \rightarrow h_0 = 0.5$
$z_1 = w_1 \cdot h_0 = 0.5 \cdot 0.5 = 0.25 \rightarrow z_1 = 0.25$
$h_1 = g_1(z_1) = g_{relu}(z_1) = g_{relu}(0.25) = 0.25 \rightarrow h_1 = 0.25$
$z_2 = w_2 \cdot h_1 + w_s \cdot h_0 = 0.5 \cdot 0.25 + 0.5 \cdot 0.5 = 0.125 + 0.25 = 0.375 \rightarrow z_2 = 0.375$
$\hat{y}_1 = g_2(z_2) = g_{linear}(z_2) = g_{linear}(0.375) = 0.375 \rightarrow \hat{y}_1 = 0.375$

$$\frac{\partial L}{\partial \hat{y}} = \frac{\hat{y} - y}{|\hat{y} - y|} = \frac{0.375 - (-3)}{|0.375 - (-3)|} = 1$$
$$\frac{\partial L}{\partial z_2} = \frac{\partial L}{\partial \hat{y}} \cdot 1 = 1$$
$$\rightarrow \frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial z_2} \cdot h_1 = 1 \cdot 0.25 = 0.25$$
$$\rightarrow \frac{\partial L}{\partial w_s} = \frac{\partial L}{\partial z_2} \cdot h_0 = 1 \cdot 0.5 = 0.5$$
$$\frac{\partial L}{\partial h_1} = \frac{\partial L}{\partial z_2} \cdot w_2 = 1 \cdot 0.5 = 0.5$$
$$\frac{\partial L}{\partial z_1} = \frac{\partial L}{\partial h_1} \cdot g_1^{'}(z_1) = \frac{\partial L}{\partial h_1} \cdot g_1^{'}(0.25) = \frac{\partial L}{\partial h_1} \cdot 1 = 0.5$$
$$\rightarrow \frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial z_1} \cdot h_0 = 0.5 \cdot 0.5 = 0.25$$

$$\frac{\partial L}{\partial h_0} = \frac{\partial L}{\partial z_1} \cdot w_1 + \frac{\partial L}{\partial z_2} \cdot w_s = 0.5 \cdot 0.5 + 1 \cdot 0.5 = 0.25 + 0.5 = 0.75$$
$$\frac{\partial L}{\partial z_0} = \frac{\partial L}{\partial h_0} \cdot g_0^{'}(z_0) = \frac{\partial L}{\partial h_0} \cdot g_0^{'}(0.5) = \frac{\partial L}{\partial h_0} \cdot 1 = 0.75$$
$$\rightarrow \frac{\partial L}{\partial w_0} = \frac{\partial L}{\partial z_0} \cdot x = 0.75 \cdot 1 = 0.75$$

After one gradient descent step:

$$w_0 = w_0 - \eta \cdot \frac{\partial L}{\partial w_0} = 0.5 - 1 \cdot 0.75 = -0.25$$
$$w_1 = w_1 - \eta \cdot \frac{\partial L}{\partial w_1} = 0.5 - 1 \cdot 0.25 = 0.25$$
$$w_s = w_s - \eta \cdot \frac{\partial L}{\partial w_s} = 0.5 - 1 \cdot 0.5 = 0$$
$$w_2 = w_2 - \eta \cdot \frac{\partial L}{\partial w_2} = 0.5 - 1 \cdot 0.25 = 0.25$$

(c) For course improvements, we would like your feedback about this question. At least tell us how much time you did invest, if you had major problems and if you think it's useful.

The first part of the assignment (Pen and Paper Backpropagation) was a good way to remember previous concepts from Machine Learning and have a clear idea about what MLP is. We think is always important to try and understand the concepts, before starting to code them. For that reason we believe that it was an useful way to begin the assignment.

We spent in this exercise around 60-90 minutes.

Part II

MLP Implementation

For the MLP Implementation, the code can be found in the Jupyter Notebook.