Homework 4 Neural Networks

$u_0 = x_0 w_0 \quad = 0.3$

$y_0 = u_0 + b_0 \quad = 0.4$

$x_1 = ReLU(y_0) = 0.4$

$u_1 = x_1 w_1 \quad = -0.8$

$y_1 = u_1 + b_1 \quad = -0.38$

$v_1 = y_1 + x + 0 \quad = 0.62$

$z = ReLU(v_1) = \quad 0$

Instructions

Answer the questions and upload your answers to courseville. Answers can be in Thai or English. Answers can be either typed or handwritten and scanned. the assignment is divided into several small tasks. Each task is weighted equally (marked with **T**). For this assignment, each task is awarded equally. There are also optional tasks (marked with **OT**) counts for half of the required task.

The Basics

In this section, we will review some of the basic materials taught in class. These are simple tasks and integral to the understanding of deep neural networks, but many students seem to misunderstand.

T1. Compute the forward and backward pass of the following computation. Note that this is a simplified residual connection.

$$x_1 = ReLU(x_0 * w_0 + b_0)$$
$$y_1 = x_1 * w_1 + b_1$$
$$z = ReLU(y_1 + x_0)$$

$$\frac{\partial z}{\partial w_0} = \frac{\partial ReLU(v_1)}{\partial v_1} \frac{\partial v_1}{\partial u_1} \frac{\partial y_1}{\partial u_1} \frac{\partial u_1}{\partial x_1} \frac{\partial x_1}{\partial y_0} \frac{\partial y_0}{\partial u_0} \frac{\partial u_0}{\partial w_0}$$
$$= (1)(1)(1)(w_1)(1)(1)(x_0) = -0.2$$

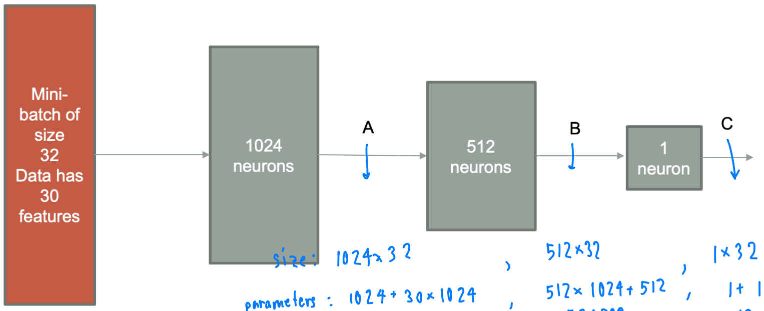
$$\frac{\partial z}{\partial w_1} = \frac{\partial ReLU(v_1)}{\partial v_1} \frac{\partial v_1}{\partial y_1} \frac{\partial y_1}{\partial u_1} \frac{\partial u_1}{\partial w_1}$$
$$= (1)(1)(1)(x_1) = 0.4$$

$$\frac{\partial z}{\partial b_0} = (1)(1)(1)(w_1)(1)(1)$$
$$= -0.2$$

$$\frac{\partial z}{\partial b_1} = (1)(1)(1)$$
$$= 1$$

Let $x_0 = 1.0$, $w_0 = 0.3$, $w_1 = -0.2$, $b_0 = 0.1$, $b_1 = -0.3$. Find the gradient of z with respect to w_0, w_1, b_0, and b_1.

T2. Given the following network architecture specifications, determine the size of the output A, B, and C.



size: 1024×32 , 512×32 , 1×32

parameters : $1024 + 30 \times 1024 = 31744$, $512 \times 1024 + 512 = 524800$, $1 + 1 \times 512 = 513$

T3. What is the total number of learnable parameters in this network? (Don't forget the bias term)

Deep Learning from (almost) scratch

In this section we will code simple a neural network model from scratch (numpy). However, before we go into coding let's start with some loose ends, namely the gradient of the softmax layer.

i=j ;

$$o_j = \frac{exp(h_j)}{\Sigma_k exp(h_k)}$$

$$\frac{\partial o_j}{\partial h_i} = \frac{\frac{\partial exp(h_i)}{\Sigma_k exp(h_k)}}{\partial h_i}$$

$$= o_i(1-o_i)$$

Recall in class we define the softmax layer as:

i ≠ j ;

$$\frac{\partial o_j}{\partial h_i} = \frac{\frac{\partial exp(h_j)}{\Sigma_k exp(h_k)}}{\partial h_i}$$

$$= -o_i o_j$$

$$P(y=j) = \frac{exp(h_j)}{\Sigma_k exp(h_k)} \tag{1}$$

where h_j is the output of the previous layer for class index j The cross entropy loss is defined as:

$$L = -\Sigma_j y_j log P(y=j) \tag{2}$$

where y_j is 1 if y is class j, and 0 otherwise.

$$\frac{\partial L}{\partial o_j} = \frac{-\Sigma_j y_j \log(o_j)}{\partial o_j}$$

$$= -\Sigma_j y_j \frac{\partial \log(o_j)}{\partial o_j}$$

$$= -\Sigma_j \frac{y_j}{o_j}$$

T4. Prove that the derivative of the loss with respect to h_i is $P(y=i) - y_i$. In other words, find $\frac{\partial L}{\partial h_i}$ for $i \in \{0, ..., N-1\}$ where N is the number of classes. Hint: first find $\frac{\partial P(y=j)}{\partial h_i}$ for the case where $j = i$, and the case where $j \neq i$. Then, use the results with chain rule to find the derivative of the loss.

Next, we will code a simple neural network using numpy. Use the starter code `hw4.zip` on github. There are 8 tasks you need to complete in the starter code.

i=j ;

$$\frac{\partial L}{\partial h_i}$$

$$= \frac{-\Sigma_i y_j \log(a_i)}{\partial o_i} \frac{\partial o_i}{\partial h_i}$$

$$= -\Sigma_i \frac{y_i}{o_i} o_i(1-o_i)$$

$$= y_i o_i - y_i$$

Hints: In order to do this part of the assignment, you will need to find gradients of vectors over matrices. We have done gradients of scalars (Traces) over matrices before, which is a matrix (two-dimensional). However, gradients of vectors over matrices will be a tensor (three-dimensional), and the properties we learned will not work. I highly recommend you find the gradients in parts. In other words, compute the gradient for each element in the the matrix/vector separately. Then, combine the result back into matrices. For more information, you can read this simple guide `http://cs231n.stanford.edu/vecDerivs.pdf`

Happy coding.

i ≠ j ;

$$\frac{\partial h}{\partial h_i} = \frac{-\Sigma_j y_j \log(o_j)}{\partial o_j} \frac{\partial o_j}{\partial h_i}$$

$$= -\Sigma_j y_j \frac{1}{o_j} \frac{\partial o_j}{\partial h_i}$$

$$= -y_i(1-o_i) - \Sigma_{i \neq j} y_j \frac{1}{o_j}(-o_j o_i)$$

$$= -y_i + y_i p_i + \Sigma_{i \neq j} y_j o_i$$

$$= -y_i + p_i \Sigma_j y_j$$

$$= -y_i + p_i$$