Homework 4
Neural Networks Instructions

Pupipat Singkhorn

March 12, 2024

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.

$$\begin{aligned} x_1 &= ReLU(x_0 * w_0 + b_0) \\ y_1 &= x_1 * w_1 + b_1 \\ z &= ReLU(y_1 + x_0) \end{aligned}$$

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.

Answer:

Forward Pass

$$
\begin{aligned}
x_1 &= ReLU(x_0 * w_0 + b_0) \\
x_1(x_0 = 1.0, w_0 = 0.3, b_0 = 0.1) &= ReLU(1.0 * 0.3 + 0.1) \\
&= ReLU(0.4) \\
&= max(0, 0.4) \\
\therefore x_1(x_0 = 1.0, w_0 = 0.3, b_0 = 0.1) &= 0.4
\end{aligned}
$$

$$
\begin{aligned}
y_1 &= x_1 * w_1 + b_1 \\
y_1(x_1 = 0.4, w_1 = -0.2, b_1 = -0.3) &= 0.4 * -0.2 + -0.3 \\
\therefore y_1(x_1 = 0.4, w_1 = -0.2, b_1 = -0.3) &= -0.38
\end{aligned}
$$

$$
\begin{aligned}
z &= ReLU(y_1 + x_0) \\
z(y_1 = -0.38, x_0 = 1.0) &= ReLU(-0.38 + 1.0) \\
&= ReLU(0.62) \\
&= max(0, 0.62) \\
\therefore z(y_1 = -0.38, x_0 = 1.0) &= 0.62
\end{aligned}
$$

Backward Pass
$function : x_1(x_0, w_0, b_0),\ y_1(x_1, w_1, b_1),\ z(y_1, x_0)$

$$\begin{aligned}
\frac{\partial z}{\partial w_0} &= \frac{\partial z}{\partial y_1} * \frac{\partial y_1}{\partial x_1} * \frac{\partial x_1}{\partial w_0} \\
&= \frac{\partial}{\partial y_1} ReLU(y_1 + x_0) \\
&\quad * \frac{\partial}{\partial x_1}(x_1 * w_1 + b_1) \\
&\quad * \frac{\partial}{\partial w_0} ReLU(x_0 * w_0 + b_0) \\
\frac{\partial z}{\partial w_0}\Big|_{x_0=1.0,\ w_1=-0.2,\ b_0=0.1,\ b_1=-0.3} &= \frac{\partial}{\partial (y_1 + x_0)} ReLU(y_1 + x_0)\Big|_{x_0=1.0,\ y_1=-0.38} * w_1 * x_0
\end{aligned}$$

$$\bullet\ \frac{\partial}{\partial x} ReLU(x) = \begin{cases} 1 & ;\ x > 0 \\ 0 & ;\ x \leqslant 0 \end{cases}$$

$$\begin{aligned}
\frac{\partial z}{\partial w_0}\Big|_{x_0=1.0,\ w_1=-0.2,\ b_0=0.1,\ b_1=-0.3} &= 1 * -0.2 * 1 \\
&= -0.2
\end{aligned}$$

$\therefore$ Gradient of z with respect to $w_0 = -0.2$

$$\begin{aligned}
\frac{\partial z}{\partial w_1}\Big|_{x_0=1.0,\ w_0=0.3,\ b_0=0.1,\ b_1=-0.3} &= \frac{\partial z}{\partial y_1} * \frac{\partial y_1}{\partial w_1} \\
&= \frac{\partial}{\partial (y_1 + x_0)} ReLU(y_1 + x_0)\Big|_{x_0=1.0,\ y_1=-0.38} * x_1 \\
&= 1 * 0.4 \\
&= 0.4
\end{aligned}$$

$\therefore$ Gradient of z with respect to $w_1 = 0.4$

$$
\begin{aligned}
\frac{\partial z}{\partial b_0}\Big|_{x_0=1.0,\ w_0=0.3,\ w_1=-0.2,\ b_1=-0.3} &= \frac{\partial z}{\partial y_1} * \frac{\partial y_1}{\partial x_1} * \frac{\partial x_1}{\partial b_0} \\
&= \frac{\partial}{\partial (y_1 + x_0)} ReLU(y_1 + x_0)\Big|_{x_0,\ y_1} \\
&\quad * w_1 \\
&\quad * \frac{\partial}{\partial (x_0 * w_0 + b_0)} ReLU(x_0 * w_0 + b_0)\Big|_{x_0,\ w_0,\ b_0} \\
&= 1 * -0.2 * 1 \\
&= -0.2
\end{aligned}
$$

$\therefore$ Gradient of z with respect to $b_0 = -0.2$

$$
\begin{aligned}
\frac{\partial z}{\partial b_1}\Big|_{x_0=1.0,\ w_0=0.3,\ w_1=-0.2,\ b_0=0.1,} &= \frac{\partial z}{\partial y_1} * \frac{\partial y_1}{\partial b_1} \\
&= \frac{\partial}{\partial (y_1 + x_0)} ReLU(y_1 + x_0)\Big|_{x_0,\ w_0,\ w_1,\ b_0} * 1 \\
&= 1 * 1 \\
&= 1
\end{aligned}
$$

$\therefore$ Gradient of z with respect to $b_1 = 1$

T2.

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

$\therefore$ Size of the output A = 32 * 1024 = 32,768
$\therefore$ Size of the output B = 32 * 512 = 16,384
$\therefore$ Size of the output C = 32 * 1 = 32

T3.

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

$$\begin{aligned} Total &= First\ layer + Second\ layer + Third\ layer \\ &= (30+1) * 1024 + (1024+1) * 512 + (512+1) * 1 \\ &= 557,057 \end{aligned}$$

$\therefore$ Total learnable parameters = 557,057

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.

Recall in class we define the softmax layer as:

$$P(y=j) = \frac{\exp(h_j)}{\sum_k \exp(h_k)}$$

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

$$L = -\sum_j y_j \log P(y=j)$$

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

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 $i \in \{0, \ldots, N-1\}$ 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 the chain rule to find the derivative of the loss. To find the derivative

of the loss with respect to h_i, we need to compute $\frac{\partial P(y=j)}{\partial h_i}$ for the case where $j = i$ and the case where $j \neq i$, and then use the chain rule to find $\frac{\partial L}{\partial h_i}$.

Answer:

Compute $\frac{\partial P(y=j)}{\partial h_i}$ for $j = i$:

$$\begin{aligned}
\frac{\partial P(y=i)}{\partial h_i} &= \frac{\partial}{\partial h_i}\left(\frac{\exp(h_i)}{\sum_k \exp(h_k)}\right) \\
&= \frac{\exp(h_i)\sum_k \exp(h_k) - \exp(h_i)\exp(h_i)}{(\sum_k \exp(h_k))^2} \\
&= \frac{\exp(h_i)}{\sum_k \exp(h_k)} - \frac{(\exp(h_i))^2}{(\sum_k \exp(h_k))^2} \\
&= P(y=i) - (P(y=i))^2 \\
\therefore \ \frac{\partial P(y=i)}{\partial h_i} &= P(y=i)\cdot(1 - P(y=i))
\end{aligned}$$

Compute $\frac{\partial P(y=j)}{\partial h_i}$ for $j \neq i$:

$$\begin{aligned}
\frac{\partial P(y=j)}{\partial h_i} &= \frac{\partial}{\partial h_i}\left(\frac{\exp(h_j)}{\sum_k \exp(h_k)}\right) \\
&= -\frac{\exp(h_j)\exp(h_i)}{(\sum_k \exp(h_k))^2} \\
\therefore \ \frac{\partial P(y=j)}{\partial h_i} &= -P(y=j)\cdot P(y=i)
\end{aligned}$$

Now, we can compute $\frac{\partial L}{\partial h_i}$

$$
\begin{aligned}
\frac{\partial L}{\partial h_i} &= \frac{\partial}{\partial h_i}\left(-\sum_j y_j \log P(y=j)\right) \\
&= \sum_j -\frac{\partial}{\partial h_i} y_j \log P(y=j) \\
&= -\frac{\partial}{\partial h_i} y_i \log P(y=i) - \sum_{j\neq i} \frac{\partial}{\partial h_i} y_j \log P(y=j) \\
&= -y_i \frac{\partial}{\partial h_i} \log P(y=i) - \log P(y=i) \frac{\partial}{\partial h_i} y_i \\
&\quad - \sum_{j\neq i}\left(y_j \frac{\partial}{\partial h_i} \log P(y=j) + \log P(y=j) \frac{\partial}{\partial h_i} y_j\right) \\
&= -y_i \frac{\partial}{\partial h_i} \log P(y=i) - \sum_{j\neq i} y_j \frac{\partial}{\partial h_i} \log P(y=j) \\
&= -\frac{y_i}{P(y=i)} \frac{\partial}{\partial h_i} P(y=i) - \sum_{j\neq i} \frac{y_j}{P(y=j)} \frac{\partial}{\partial h_i} P(y=j) \\
&= -\frac{y_i}{P(y=i)} P(y=i)(1-P(y=i)) - \sum_{j\neq i} \frac{y_j}{P(y=j)} \cdot (-P(y=j)P(y=i)) \\
&= -y_i + y_i P(y=i) + \sum_{j\neq i} y_i P(y=i) \\
&= -y_i + y_i P(y=i) + P(y=i) \sum_{j\neq i} y_i \\
&= -y_i + y_i P(y=i) + P(y=i)\left(\sum_j y_j - y_i\right) \\
&= -y_i + y_i P(y=i) + P(y=i)\,(1-y_i) \\
&= -y_i + y_i P(y=i) + P(y=i) - y_i P(y=i) \\
\therefore\ \frac{\partial L}{\partial h_i} &= P(y=i) - y_i
\end{aligned}
$$

Therefore, the derivative of the loss with respect to h_i is $P(y=i) - y_i$.