CS5242 Neural Networks and Deep Learning
Quiz 1 - Answers

September 27, 2021

1 Question 1

What are the advantages of training neural networks with mini-batch gradient vs. full-batch gradient?

Answer:
Using mini-batch gradient descent allows for faster updates to the weights and regularizes training which improves generalization.

2 Question 2

Consider a three-layer Multi-Layer Perceptron (MLP) network for a 10-class classification task with an input layer, a hidden layer and an output layer. The input data is of 784-dimension and there are 100 neurons in the hidden layer. What is the total number of learnable parameters in this MLP network (including the biases in the linear layers)?

Answer:
From input layer to hidden layer, the number of neurons are: $784 * 100$ (weight) + 100 (bias) = 78500. From input layer to output layer, the number of neurons are $100 * 10$ (weight) + 10 (bias) = 1010. Thus, the total number of neurons is $78500 + 1010 = 79510$.

3 Questions 3

Consider the following softmax operation: $\text{Softmax}\left(\begin{bmatrix}1 & 1\\ 2 & 0\\ 1 & 0\end{bmatrix} \times \begin{bmatrix}2\\ 1\end{bmatrix}\right) = \begin{bmatrix}0.24\\ 0.67\\ x\end{bmatrix}$.
What is the value of x?

Answer:
1 - 0.24 - 0.67 = 0.09

4 Question 4

Consider the following 3-layer MLP network with the input layer (i), the hidden layer (h), the output layer (o) and the biases for the hidden layer and the output layer (b).The initial weight and bias values are as follows:

$\omega_{i_1h_1} = 0.1, \omega_{i_1h_2} = 0.1, \omega_{i_1h_3} = 0.2, \omega_{i_2h_1} = 0.2, \omega_{i_2h_2} = 0.2, \omega_{i_2h_3} = 0.3$
$\omega_{h_1o_1} = 0.4, \omega_{h_1o_2} = 0.4, \omega_{h_2o_1} = 0.5, \omega_{h_2o_2} = 0.5, \omega_{h_3o_1} = 0.6, \omega_{h_3o_2} = 0.7$
$b_1 = 0.03, b_2 = 0.6$

The values of the inputs are $i_1 = 0.5, i_2 = 0.1$ What are the output values o1 and o2 (round the value to the nearest 2 decimal points, i.e. x.xx)?

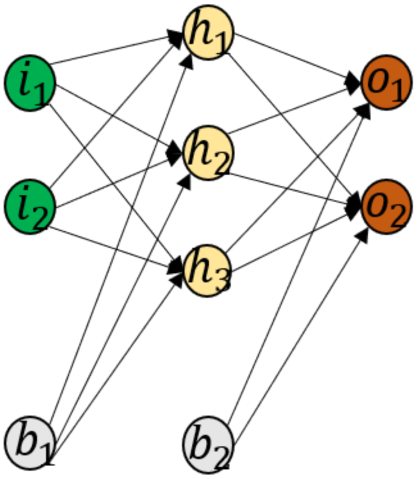

Figure 1: MLP Network

Answer:

$out_{h_1} = \omega_{i_1h_1} \cdot i_1 + \omega_{i_2h_1} \cdot i_2 + b_1 = 0.1 \cdot 0.5 + 0.2 \cdot 0.1 + 0.03 = 0.1$
$out_{h_2} = \omega_{i_1h_2} \cdot i_1 + \omega_{i_2h_2} \cdot i_2 + b_1 = 0.1 \cdot 0.5 + 0.2 \cdot 0.1 + 0.03 = 0.1$
$out_{h_3} = \omega_{i_1h_3} \cdot i_1 + \omega_{i_2h_3} \cdot i_2 + b_1 = 0.2 \cdot 0.5 + 0.3 \cdot 0.1 + 0.03 = 0.16$
$out_{o_1} = \omega_{h_1o_1} \cdot out_{h_1} + \omega_{h_2o_1} \cdot out_{h_2} + \omega_{h_3o_1} \cdot out_{h_3} + b_2 = 0.4 \cdot 0.1 + 0.5 \cdot 0.1 + 0.6 \cdot 0.16 + 0.6 = 0.786$
$out_{o_2} = \omega_{h_1o_2} \cdot out_{h_1} + \omega_{h_2o_2} \cdot out_{h_2} + \omega_{h_3o_1} \cdot out_{h_3} + b_2 = 0.4 \cdot 0.1 + 0.5 \cdot 0.1 + 0.7 \cdot 0.16 + 0.6 = 0.802$

Then the answers are 0.79 and 0.80.

5 Question 5

Let us consider a 3-class classification task and a forward pass defined as: $\hat{\mathbf{y}} = \sigma(\mathbf{WX})$, where σ is the softmax function, $\mathbf{X}$ is the input data matrix of size $\mathbb{R}^{d\times n}$: $\mathbf{X} = \begin{bmatrix} 2 & 1 & 1 \\ -1 & 0 & -1 \\ 1 & 0 & 0 \end{bmatrix}$, where d is the number of features and n is the number of data points. $\mathbf{W}$ represents the $\mathbb{R}^{d\times d}$ matrix of parameters of a linear layer $\mathbf{W} = \begin{bmatrix} 1 & 2 & 0 \\ 1 & 1 & 2 \\ -1 & 2 & 1 \end{bmatrix}$. The labels of the data is $\mathbf{y} = [1\ 2\ 0]$. What is the value of the mean cross-entropy loss? Note: The logarithm used is the natural logarithm, i.e., the logarithm to the base e: $ln(x) = log_e(x)$.

Answer:

$$WX = \begin{bmatrix} 1 & 2 & 0 \\ 1 & 1 & 2 \\ -1 & 2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 \\ -1 & 0 & -1 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & -1 \\ 3 & 1 & 0 \\ -3 & -1 & -3 \end{bmatrix}$$

Given $y = \begin{bmatrix} 1 & 2 & 0 \end{bmatrix}$, the mean cross-entropy is

$$-\frac{1}{3} * [\ln \frac{e^3}{e^0 + e^3 + e^{-3}} + \ln \frac{e^{-1}}{e^1 + e^1 + e^{-1}} + \ln \frac{e^{-1}}{e^{-1} + e^0 + e^{-3}}] \approx 1.386$$

Answers of 1.38 ± 0.01 should be correct.

6 Question 6

Suppose we have a loss function $L(\mathbf{w}) = \frac{1}{2}\|\mathbf{wX} - \mathbf{y}\|^2 + \frac{1}{2}\|\mathbf{w}\|^2$, where $\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 0 & 0.5 \end{bmatrix}$ and $\mathbf{y} = [0\ 1]$. What is the value of the gradient $\frac{\partial L}{\partial \mathbf{w}}$ for $\mathbf{w} = [1\ 0]$?

Answer:
$\frac{\partial L}{\partial \mathbf{w}} = (\mathbf{wX} - \mathbf{y})\mathbf{X}^T + \mathbf{w} = [2.00\ 0.00]$

7 Question 7

Let us define the Leaky ReLU activation function as:

$$f(x) = \begin{cases} 0.01x & \text{if } x < 0 \\ x & \text{otherwise} \end{cases}$$

Suppose we have an MLP where one of the linear layers is defined as

$$\mathbf{W} = \begin{bmatrix} 2 & 1 \\ 3 & 2 \end{bmatrix}$$

, and its output is activated by a Leaky ReLU function. Denote the input to that layer $x = [-2, -1]^\intercal$ and the output of the Leaky ReLU by $\mathbf{y}$.

During the backpropagation process, we obtained $\frac{\partial \mathcal{L}}{\partial \mathbf{y}} = [-1.5, 2.5]^\intercal$.

After backpropagation, $\mathbf{W}$ is updated as $\mathbf{W}' = \mathbf{W} - \frac{\partial \mathcal{L}}{\partial \mathbf{W}}$.

What is the value of $\mathbf{W}'_{11}$? Answer (round the value to the nearest 2 decimal points, i.e. x.xx): ___1___

What is the value of $\mathbf{W}'_{22}$? Answer (round the value to the nearest 2 decimal points, i.e. x.xx): ___2___

Answer: 1.97; 2.03

$$\begin{aligned}
\mathbf{W}'_{11} &= \mathbf{W}_{11} - \frac{\partial \mathcal{L}}{\partial \mathbf{W}_{11}} \\
&= \mathbf{W}_{11} - \frac{\partial \mathcal{L}}{\partial \mathbf{y}_1} \cdot \frac{\partial \text{LeakyReLU}(\mathbf{W}_{11}\mathbf{x})}{\partial \mathbf{W}_{11}} \\
&= 2 - (-1.5) \cdot 0.01 \cdot \mathbf{x}_1 \\
&= 1.97
\end{aligned}$$

$$\begin{aligned}
\mathbf{W}'_{22} &= \mathbf{W}_{22} - \frac{\partial \mathcal{L}}{\partial \mathbf{W}_{22}} \\
&= \mathbf{W}_{22} - \frac{\partial \mathcal{L}}{\partial \mathbf{y}_2} \cdot \frac{\partial \text{LeakyReLU}(\mathbf{W}_{22}\mathbf{x})}{\partial \mathbf{W}_{22}} \\
&= 2 - 2.5 \cdot 0.01 \cdot \mathbf{x}_2 \\
&= 2.025
\end{aligned}$$

8 Question 8

Let LogSumExp (LSE) function be a **smooth maximum** function defined as: $LSE(a_1, ..., a_n) = log\big(\sum_{i=1}^{n} exp(a_i)\big)$, i.e. $LSE(a_1, ..., a_n) \approx max(a_1, ..., a_n)$. Let us consider the output $y = LSE(w_1x_1, w_2x_2, w_3x_3, w_4x_4)$, where $\mathbf{x} = [x_1, x_2, x_3, x_4]$ and $\mathbf{w} = [w_1, w_2, w_3, w_4]$ are the input and weight parameters respectively. In other words, the input $\mathbf{x}$ and the parameter $\mathbf{w}$ are multiplied element-wise and then passed through the LSE function to get the output y.

Given the input value $\mathbf{x} = [x_1, x_2, x_3, x_4] = [1, 2, 10, 3]$ and the weight parameters $\mathbf{w} = [w_1, w_2, w_3, w_4] = [1, 1, 1, 1]$, what is the closet **integer** value to the gradient $\frac{\partial y}{\partial w_3}$?

Answer:

$$\frac{\partial y}{\partial w_3} = \frac{1}{\sum_{i=1}^{4} exp(w_i x_i)} \cdot exp(w_3 x_3) \cdot x_3$$

Due to the property of LogSumExp (LSE) function, we have:

$$log\sum_{i=1}^{4} exp(w_i x_i) \approx \max(w_i x_i), i \in 1,2,3,4$$

Given the input value, we have

$$\sum_{i=1}^{4} exp(w_i x_i) \approx exp(w_3 x_3)$$

Then,

$$\frac{\partial y}{\partial w_3} \approx 1 \cdot x_3 = 10.$$

9 Question 9

Consider a single linear layer binary-classification vanilla neural network with a softmax output.

If the prediction accuracy is 100% for the following test dataset:

$$\begin{bmatrix} 0 & 0 & 1 & -1 \\ 2 & -2 & -2 & 1 \\ 2 & 0 & 2 & 1 \end{bmatrix}$$

with label $[0,1,0,1]$, then what is the prediction accuracy for the following test dataset

$$\begin{bmatrix} 1 & 1 & 2 & -1 \\ 1 & 0 & -2 & -1 \\ 1 & 2 & -2 & -1 \end{bmatrix}$$

with label $[0,0,1,1]$?

Answer (in %, i.e. an integer in [0,100]) : ________

Answer: 75

Single-layer vanilla neural network preserves the linearity of the input, and ensures that if $\mathbf{X}_i$ has label 1, then $-\mathbf{X}_i$ has the label 0 (since there's no bias term in single-layer vanila NN).

Denote the first dataset $\mathbf{X}^1$ and the second dataset $\mathbf{X}^2$. Through observation, we can find that

$$\begin{aligned}\mathbf{X}_1^2 &= \mathbf{X}_1^1 - \mathbf{X}_4^1\\ \mathbf{X}_2^2 &= \mathbf{X}_3^1 - \mathbf{X}_2^1\\ \mathbf{X}_3^2 &= -2\mathbf{X}_4^1\\ \mathbf{X}_4^2 &= \mathbf{X}_4^1 - \mathbf{X}_1^1\end{aligned}$$

So we can conclude that $\mathbf{X}_1^2$, $\mathbf{X}_2^2$ and $\mathbf{X}_4^2$ have correct labels, while X_3^2 have the incorrect label. Thus the result is 75%.

10 Question 10

Given two d-dimensional vector $\mathbf{a} = [a_1, a_2, ..., a_d]$ and $\mathbf{b} = [b_1, b_2, ..., b_d]$, where each element in the vector is sampled from a standard Gaussian distribution, i.e. $a_i \sim \mathbf{N}(0, 1), b_j \sim \mathbf{N}(0, 1), i, j \in [1, d]$. After an inner-product, we get the activation $\hat{o} = \mathbf{a}^T\mathbf{b} = \sum_{i=1}^{d} a_i b_i$. To guarantee the final output o follows a standard Gaussian distribution, we need to scale the activation $\hat{o}$ with the constant $\frac{1}{k}$, i.e., the final output is defined as $o = \frac{\hat{o}}{k}$. When $d = 100$, what is the value of k that guarantees the output o to be a standard Gaussian distribution?

Answer:
Suppose two independent random variables A, B are sampled from standard Gaussian Distribution, then $E(A) = E(B) = 0, D(A) = D(B) = 1$ (E for expectation, D for variance). Then,

$$E(AB) = E(A)E(B) = 0$$

$$\begin{aligned}D(AB) &= E(A^2B^2) - [E(AB)]^2\\ &= E(A^2)E(B^2) - [E(A)E(B)]^2\\ &= E(A^2 - 0)E(B^2 - 0)\\ &= D(A)D(B)\\ &= 1\end{aligned}$$

Therefore,

$$\begin{aligned}E(\hat{o}) &= 0\\ D(\hat{o}) &= D(a_1b_1 + a_2b_2 + ... + a_db_d)\\ &= D(a_1b_1) + D(a_2b_2) + ... + D(a_db_d)\\ &= 1 + 1 + ... + 1\\ &= d\end{aligned}$$

Then,

$$E(o) = \frac{E(\hat{o})}{k} = 0$$

$$
\begin{aligned}
D(o) &= D(\frac{\hat{o}}{k}) \\
&= \frac{D(\hat{o})}{k^2} \quad (k\ is\ a\ constant) \\
&= \frac{d}{k^2}
\end{aligned}
$$

To make $D(o) = \frac{d}{k^2} = 1$, we need to set $k = \sqrt{d}$. Then, when $d = 100$, $k = \sqrt{1}00 = 10$.