Indian Institute of Technology Kharagpur
Class Test 02 2021-22

Date of Examination: 24 Feb. 2022 **Duration:** 45 minutes

Subject No.: CS60010 **Subject:** Deep Learning

Department/Center/School: Computer Science **Credits:** 3 **Full marks:** 20

Instructions

i. This question paper contains 2 pages and 3 questions. All questions are compulsory. Marks are indicated in parentheses. This question paper has been cross checked.

ii. Please write your name, roll number, subject name and code, date and time of examination on the answer script before attempting any solution.

iii. **Organize your work**, in a reasonably neat and coherent way. Work scattered all across the answer script without a clear ordering will receive very little marks.

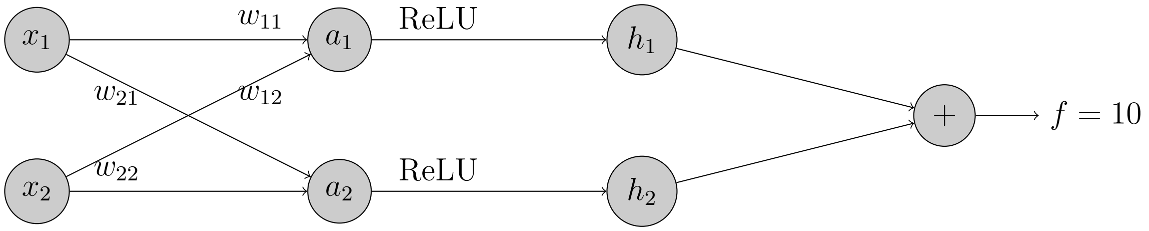
iv. **Mysterious or unsupported answers will not receive full marks**. A correct answer, unsupported by calculations, explanation, will receive no marks; an incorrect answer supported by substantially correct calculations and explanations may receive partial marks.

v. In the online mode of the quiz, you need to upload yuor answer scripts as **pdf file**. You can scan your worked out example or you can use latex to produce the pdf.

1. (a) (3 points) Show that for a matrix $\mathbf{A}$ and vector $\mathbf{x}$, $\frac{\partial}{\partial \mathbf{x}}(\mathbf{A}^{-1}) = -\mathbf{A}^{-1}\big(\frac{\partial}{\partial \mathbf{x}}\mathbf{A}\big)\mathbf{A}^{-1}$. Use the fact that for any two matrices $\mathbf{A}$ and $\mathbf{B}$, $\frac{\partial \mathbf{AB}}{\partial \mathbf{x}} = \frac{\partial \mathbf{A}}{\partial \mathbf{x}}\mathbf{B} + \mathbf{A}\frac{\partial \mathbf{B}}{\partial \mathbf{x}}$.

   (b) (3 points) Consider an image of shape $200 \times 200 \times 3$. Suppose depthwise separable convolution is applied on this image where the spatial size of the filter is 4×4. The stride, padding and the number of output channels are 1, 1 and 10 respectively. Note that for depthwise separable convolution you have to take both depthwise convolution and pointwise convolution into consideration. Assume that padding is applied in 'depthwise convolution' step only. What is the total number of computation here (considering multiplication operations only)? What are the output sizes after each of the depthwise convolution and pointwise convolution operation?

2. Consider the following simple neural network. It takes a two dimensional input $\mathbf{x} = [x_1, x_2]^T = [5, 2]^T$ followed by a hidden layer whose weights are given by, $w_{11} = -1, w_{12} = 3, w_{21} = 1, w_{22} = 2$. The resulting preactivations are given by $\mathbf{a} = [a_1, a_2]^T$ which are passed through ReLU activation function giving rise to the activations $\mathbf{h} = [h_1, h_2]^T$. These two activation values are summed up to get the final output f.



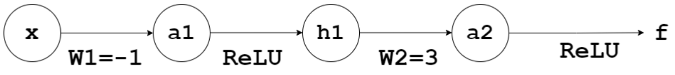
   (a) (2 points) Find the following values that will be observed in the forward pass: a_1, a_2, h_1 and h_2. Note that the value of f is provided.

(b) (4 points) Use chain rule of differentiation to find out the values of the following: $\frac{\partial f}{\partial h_1}, \frac{\partial f}{\partial h_2}, \frac{\partial f}{\partial a_1}$ and $\frac{\partial f}{\partial a_2}$.

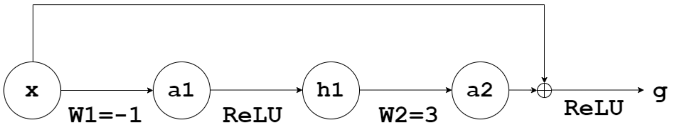
(c) (2 points) Find out the values of the following: $\frac{\partial f}{\partial w_{11}}, \frac{\partial f}{\partial w_{12}}, \frac{\partial f}{\partial w_{21}}$ and $\frac{\partial f}{\partial w_{22}}$.

(d) (2 points) Find out the values of the following: $\frac{\partial f}{\partial x_1}$ and $\frac{\partial f}{\partial x_2}$. [Hint: Note that if $u = f(x, y)$ where $x = \phi(z), y = \psi(z)$ then, $\frac{\partial u}{\partial z} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial z} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial z}$].

3. In the diagram below, we have two simple neural networks, Network-A and Network-B. Both the networks take a 1-dimensional input `x` and give the corresponding outputs `f` and `g`. The only difference between both networks is the presence of a 'skip-connection' in Network-B. The value of the weights `W1` and `W2` are -1 and 3 respectively, as mentioned below the arrows, which are used to obtain the pre-activations. The pre-activations are then passed through ReLU activation function to obtain the activations.



Network-A



Network-B

(a) (1 point) If for the same input `x`, the Network-B gives an output value of `g` $= 10$, find the corresponding output `f` of Network-A for the same `x`.

(b) (3 points) Find the values of $\frac{\partial \mathtt{f}}{\partial \mathtt{x}}$ and $\frac{\partial \mathtt{g}}{\partial \mathtt{x}}$ for Network-A and Network-B for the computed values of `x` in part (a) of the question. Use chain rule and show the steps. This would provide us the gradient value reaching the input from the output.

(c) (1 point) What difference in the gradient values did you observe between both the networks in part (b)? Why are 'skip-connections' useful? (Bonus)