

THE UNIVERSITY OF HONG KONG
SCHOOL OF COMPUTING AND DATA SCIENCE
DEPARTMENT OF COMPUTER SCIENCE

DASC7606 Deep Learning
(Subclasses A, B & D)

Date: Tuesday, December 10, 2024

Time: 6:30 p.m. – 8:30 p.m.

Answer ALL questions. They are all COMPULSORY.

The mark value of each question (or part of a question) is indicated before the question (or part of the question).

Please write your answers on this examination paper in the space provided.

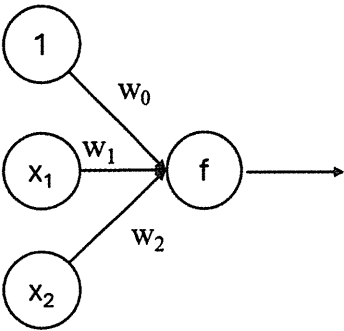
Only approved calculators as announced by the Examinations Secretary can be used in this examination. It is candidates' responsibility to ensure that their calculator operates satisfactorily, and candidates must record the name and type of the calculator used on the front page of the examination script.

Brand and Type of Calculator: _____

(20 pts) Question 1: Basic Concepts

At the right (see diagram) is a neural network where

- the two inputs $x_1, x_2 \in \{0,1\}$ and
- the activation function f is the step function, i.e., $f(z) = 1$ if $z > 0$, and 0 otherwise, with z being the weighted summation (with weights w_0, w_1 and w_2) of the 3 inputs (including 1)



- (a) (4 pts) Assign values to w_0, w_1 , and w_2 to compute the “OR” logic function. Fill in tables below with your answer.

Answer:

x_1	x_2	OR
0	0	
0	1	
1	0	
1	1	

w_0	w_1	w_2

- (b) (4 pts) Can you adjust only one of the values w_0, w_1 , or w_2 to change the logic function to “AND”? Fill in table for “AND” and give reasons to explain why this can/cannot be done.

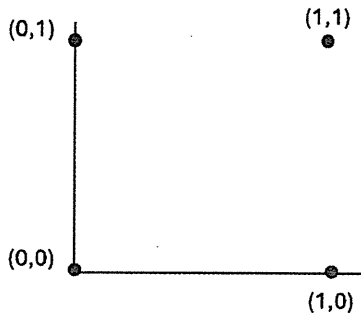
Answer:

x_1	x_2	AND
0	0	
0	1	
1	0	
1	1	

Explanation:

- (c) (4 pts) Consider inputs $x_1, x_2 \in \{0,1\}$ as coordinates of data points on a 2D plane. Think of the neural network shown in the diagram near the beginning of Question 1 as a classifier dividing the 4 possible data points (see diagram below) by a dividing (classification) line. Draw a dividing line for the “OR” logic function in the diagram below. Based on the properties of any dividing (classification) line for “OR”, list the relationships among w_0, w_1 , and w_2 .

Answer:



- (d) (4 pts) Can you adjust only one of the values w_0, w_1 , or w_2 of *some* neural network for the “OR” logic function to compute instead the L1 function which is set out in the table at the right? Give reasons to explain why this can/cannot be done.

x_1	x_2	L1
0	0	0
0	1	0
1	0	1
1	1	1

Answer:

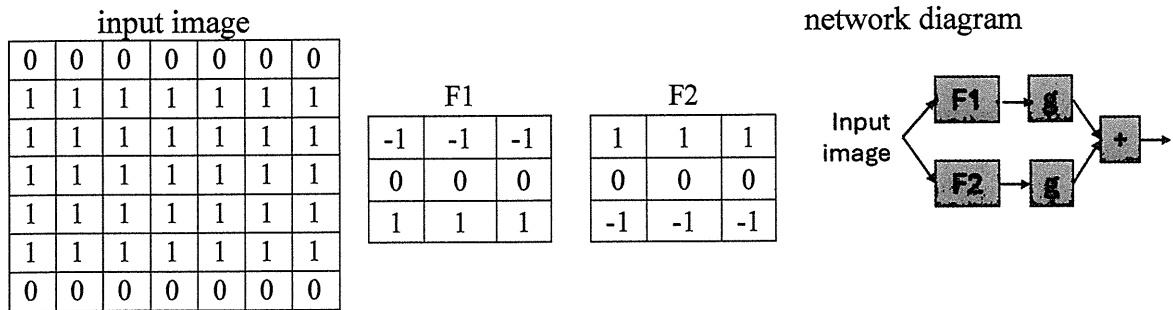
- (e) (4 pts) Can you adjust only one of the values w_0, w_1 , or w_2 of *some* neural network for the “OR” logic function to compute instead the L2 function which is set out in the table at the right? Give reasons to explain why this can/cannot be done.

x_1	x_2	L2
0	0	0
0	1	0
1	0	1
1	1	0

Answer:

(15 pts) Question 2: Convolutional Neural Networks and Filters

Consider the 7x7 input image, two 3x3 filters F1 and F2, and the network diagram below:



- (a) (5 pts) Assuming no padding and a stride of 1, what is the result after applying filter F1 to the input image? What is the result after applying filter F2 to the input image?

Answer:

- (b) (5 pts) What are the purposes of filter F1 and F2?

Answer:

F1	
F2	

- (c) (5 pts) What would be a good choice for the activation function g (see network diagram) to apply to the results of F1 and F2 before adding them together? Give reasons for your answer.

Answer:

(15 pts) Question 3: Vanishing Gradient and Skip Connections

Consider the diagram on the right which shows the computation graph for a single hidden layer of a neural network with

$$s_i = W_i x_i + b_i$$

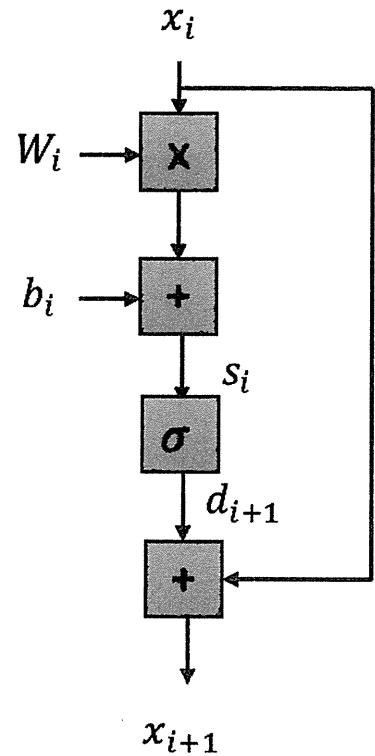
$$d_{i+1} = \text{sigmoid}(s_i) = \sigma(s_i)$$

$$x_{i+1} = d_{i+1} + x_i$$

Assume the dimensions of x_i, s_i, d_{i+1} and x_{i+1} are the same and the value of $\delta_{i+1} = \frac{\partial J}{\partial x_{i+1}}$ is known from backpropagation, where J is the loss function.

- (a) (10 pts) What is $\delta_i = \frac{\partial J}{\partial x_i}$?

Answer:

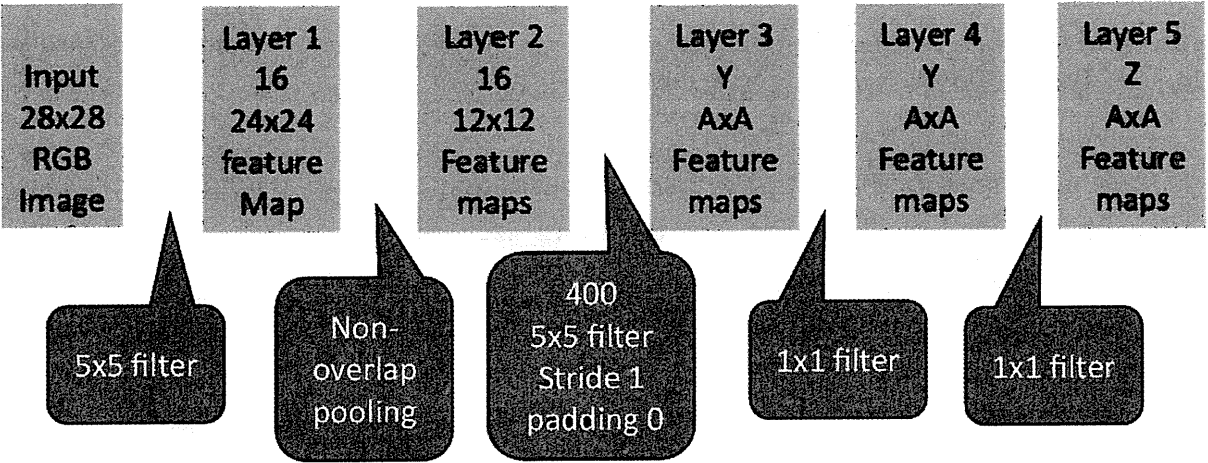


- (b) (5 pts) Based on (a), explain why the skip connection in the above computation graph can alleviate the vanishing gradient problem.

Answer:

(27 pts) Question 4: CNN and Sliding Windows

Consider a CNN for implementing the sliding window approach to classifying inputs of 28x28 RGB images into 4 categories (pedestrian, car, motorcycle, background) with the following architecture:



- (a) (3 pts) Describe the 5x5 filter between the Input and Layer 1 by filling in the table below.

Answer:

Number of filters	
Padding	
Stride	

- (b) (3 pts) How many parameters are there between the Input and Layer 1?

Answer:

- (c) (3 pts) What is the computation complexity (which is expressed in terms of total number of multiplication operations) for computing all the Layer 1 feature maps from the Input?

Answer:

- (d) (2 pts) Describe the non-overlap pooling in terms of its filter size and stride.

Answer:

- (e) (3 pts) Is there any reason why 400 filters are used? What is the value of A and Y at Layer 3?

Answer:

- (f) (4 pts) How many parameters are there between Layer 2 and Layer 3? What is the number of computations used to produce Layer 3 (also known as the number of connections between Layer 2 and Layer 3)?

Answer:

- (g) (3 pts) What is the purpose of the 1x1 filter between Layer 3 and 4?

Answer:

- (h) (3 pts) What is the size of the output layer?

Answer:

- (i) (3 pts) How many sliding windows have been considered and what is the size of each window?

Answer:

(23 pts) Question 5: Transformer

- (a) (6 pts) What are three main advantages that make the transformer architecture superior to the RNN and/or LSTM? (Hint: Two about interaction between tokens and one about embedding size.)

Answer:

- (b) (4 pts) How can the transformer solve the problems with RNN/LSTM you pointed out in (a)?

Answer:

- (c) (6 pts) Give the interpretation of K, Q and V in the attention mechanism related to the following scenarios:

- (i) A group of people wants to pick up items from supermarket. Each person has an item name and looks for the item by scanning the label name at the supermarket shelf.

- (ii) Assign the nearest Uber car to the passenger.

Answer:

(i)

(ii)

- (d) (3 pts) Describe how the encoder-decoder attention mechanism is implemented and the reason of this implementation.

Answer:

- (e) (4 pts) What are the shortcomings of transformer architecture?

Answer:

END OF PAPER