THE UNIVERSITY OF HONG KONG

SCHOOL OF COMPUTING AND DATA SCIENCE
DEPARTMENT OF COMPUTER SCIENCE

DASC7606 Deep Learning
(Subclass C)

Date: December 10, 2024 **Time: 6.30pm – 8.30pm**

There are 4 questions in total. The 1st and 2nd questions are worth 20/100 points each, while the third and 4th questions are worth 30/100 points each. Answer ALL questions. They are all COMPULSORY.

You should write your answers on the exam paper in the corresponding spaces left blank after the corresponding questions. Write only the final answers to the questions, without providing the intermediary calculations.

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.

University Number: ____________________

Brand and Type of Calculator: ____________________

Question 1 (General questions, 20/100.)

a) (6/100) Why is it preferable that the output of the activation functions in a neural network be zero-centered?
   1) If their output is zero-centered, then gradient descent will be constrained to take only ``good'' directions, thereby converging much faster;
   2) If their output is zero-centered, then the gradients typically take a wider range of values, which typically leads to better convergence;
   3) Because this solves the vanishing gradient problem;
   4) If the output of the activation functions are all positive, and all activation functions are linear functions, then the gradients are either all non-positive or all non-negative.
   5) Because this guarantees there is no overfitting.

   Provide all correct statements in numerical order (e.g. 1,3,5):

   ____

b) (7/100) Consider the following statements concerning the backpropagation algorithm:
   1) The backpropagation algorithm allows to efficiently compute the gradients in a neural network.
   2) The backpropagation algorithm consists of two passes: in the first pass, it computes the gradients. In the second pass, it typically runs one single step of the gradient descent algorithm. This typically leads to better convergence of the latter algorithm.
   3) The backward pass of the backpropagation algorithm can be efficiently parallelized: in this case the running time does not depend on the number of layers of the underlying neural network but only on the number of GPUs.

   Provide all correct statements in numerical order (e.g. 1,3):

   ____

c) (7/100) Consider the following statements concerning ResNet
   1) The main goal of ResNet is to approximate any identity function arbitrarily well. Because most real-world problems can be approximated by the identity function, ResNet provides very good results
   2) Deeper networks should provide better training and test error, however, they are more difficult to train.
   3) ResNet alleviates the problem of learning the identity function in deeper networks, while also allowing to learn more complex functions.

   Provide all correct statements in numerical order (e.g. 1,3):

   ____

Question 2 (Backpropagation, 20/100). Consider the computational graph shown in Figure 1, where the x_i's are the input data, the w_i's are the parameters of a neural network, while σ denotes the sigmoid function. F_1 denotes the value in the corresponding position of the computational graph, during the forward pass of the backpropagation algorithm. B_0, B_1 and B_2 denote the corresponding values during the backward pass of such an algorithm. We let $w_0=1$, $w_1=-1$, $w_2=-1$ and $x_0=-1$, $x_1=2$, $x_2=1$. You should run the backpropagation algorithm and provide below the values of F_1, B_0, B_1 and B_2. Always truncate the results to the first 2 decimal digits, e.g. 0.213, 0.3191 are truncated to 0.21 and 0.31, respectively.

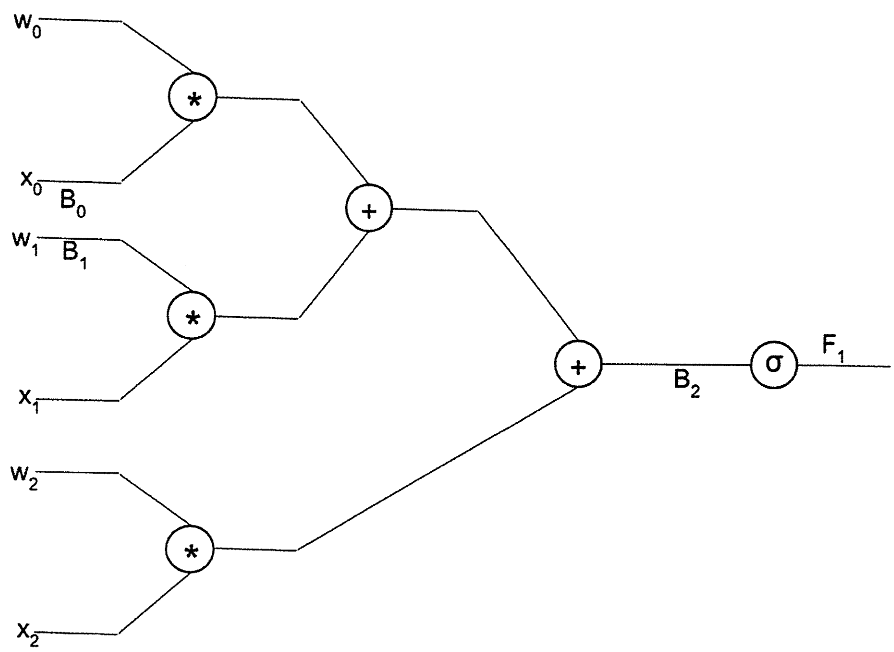

Figure 1.

a) (4/100) Provide the value of F_1 ____
b) (4/100) Provide the value of B_0 ____
c) (4/100) Provide the value of B_1 ____
d) (4/100) Provide the value of B_2 ____
e) (4/100) Suppose there is an additional node right after the σ gate which multiplies by 2 its input value (i.e. there is a node labeled '*2' right after the σ gate). In such a case:
   i) Provide the value of B_0 ____
   ii) Provide the value of B_2 ____

Question 3 (Reinforcement Learning, 30/100). Consider the following simplified version of a popular card game. The game starts with a card dealer dealing a card to the player, whose value is an integer in [1,3] chosen uniformly at random. The player can decide whether to ``hit'' (H), i.e. to ask for another card or to ``stand'' (S) i.e. not to take any more cards. This is iterated until either the player decides to ``stand'' or the total value of the cards received by the player is larger than 3. After that, the dealer deals one single card to herself, whose value is an integer in [1,3] chosen uniformly at random. Let x be the total value of the cards received by the player, while let y be the value of the card received by the dealer. The player wins if:

1. the $x \leq 3$ and $x > y$ or
2. $x = y$ and ``head'' is the result of flipping an unbiased coin (i.e. with probability 1/2).

The player loses in any other case, (i.e. $x > 3$, $x < y$, or $x = y$ and ``tail'' is the result of the coin flip). We wish to learn an optimal policy for the player, that is, a strategy that maximizes the probability that the player wins. To this end, we model the game with a Markov Decision Process (MDP) which is represented in Figure 2.

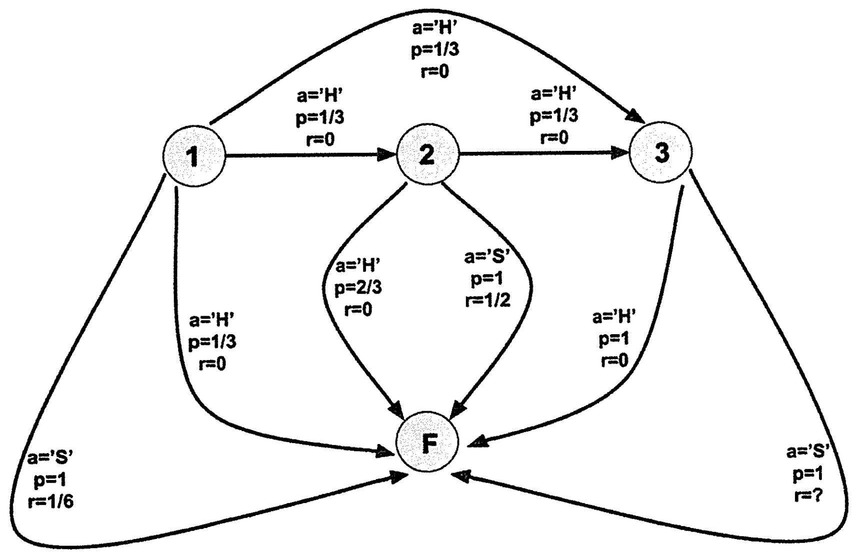

Figure 2.

In the MDP depicted in Figure 2, there are 4 states, with 1,2,3 representing the current total value of the cards received by the player, while 'F' is the final state; a denotes the action of the player ('H' for 'hit' and 'S' for 'stand'), p denotes the transition probability between the corresponding states when the corresponding action is taken, while r denotes the expected reward when the corresponding action is taken. For example, when in state 1, if the player decides to hit ('H') then with probability 1/3, the MDP moves to state 2 and gets expected reward 0. If the player decides to stand ('S') when in state 1, then with probability 1 the MDP moves to F and gets expected reward r=1/6.

The latter one corresponds to the probability that the player wins when x=1 (which is the probability that y=1 and 'head' is the result of the coin flip). Observe
that the expected reward when in state 3 and a='S' is not provided, you should provide such a missing value in question a).

a) (10/100) Provide the expected reward *r* when in state 3 and a='S' so that this corresponds to the probability that the player wins when x=3: ____
b) (10/100) What is an optimal policy? Specify the action to take for each state from the smallest to the largest one. E.g. HSH denotes Hit when in state 1, Stand when in state 2, and Hit when in state 3 _______
c) (10/100) You should run the value iteration algorithm to compute an optimal policy. Let $q_i(s,a)$ be the *q* value for state *s* and action *a* at the end of the *i*th iteration of the value iteration algorithm. Run the value iteration algorithm *undiscounted,* that is γ=1, starting with $q_0(s,a)=0$ for every *s* and *a*.
   i) Using the same notation specified above, write down an optimal policy that can be obtained from the $q_1(s,a)$ values (e.g. SHH): ______
   ii) Using the same notation specified above, write down an optimal policy that can be obtained from the $q_2(s,a)$ values (e.g. SHH): ______
   iii) What is the value of $q_2(1,\text{'H'})$? _______

Question 4 (Convolutional Neural Networks, 30/100). We assume that our input data consists of $n \geq 1$ binary matrices of size 4x4, where each such a matrix is equal to one of the following matrices:

$$
M_1 = \begin{matrix} 0&0&0&1 \\ 0&0&1&0 \\ 0&1&0&0 \\ 1&0&0&0 \end{matrix} \qquad
M_2 = \begin{matrix} 1&1&1&1 \\ 0&0&0&0 \\ 0&0&0&0 \\ 0&0&0&0 \end{matrix} \qquad
M_3 = \begin{matrix} 0&0&0&0 \\ 1&1&1&1 \\ 0&0&0&0 \\ 0&0&0&0 \end{matrix} \qquad
M_4 = \begin{matrix} 0&0&0&0 \\ 0&0&0&0 \\ 1&1&1&1 \\ 0&0&0&0 \end{matrix} \qquad
M_5 = \begin{matrix} 0&0&0&0 \\ 0&0&0&0 \\ 0&0&0&0 \\ 1&1&1&1 \end{matrix}
$$

Where M_1 depicts a diagonal, while each other matrix depicts a horizontal line. We wish to develop a classifier which classifies diagonals (i.e. any copy of M_1) with the class 1, while horizontal lines are classified with the class 0.

Develop a CNN classifier as follows:

1) There is zero padding on the input matrix with P=1
2) The first layer is a CNN layer with one 2x2 filter and stride S=1
3) Then there are k 2x2 max pooling layers, where k is such that the size of the matrix in the k+1-*th* layer is 1x1
4) There is a flattening layer turning the 1x1 matrix into a scalar with value x
5) The activation function on the last layer is a threshold function f(x,t) which outputs 1 if $x \geq t$, and 0 otherwise (where x is the scalar obtained after flattening). All other activation functions are the identity function.

Questions:

a) (5/100) What is the size of the feature map in the first layer? _______
b) (5/100) What is the value of k? _____
c) (5/100) What is the value of t in the threshold activation function f? _______
d) (15/100) Provide all possible solutions for the filter in the first layer assuming that such a filter is a 0-1 matrix (with entries either 0 or 1). Write down all possible solutions below as a list of 2x2 matrices allowing some space between the matrices (if more than one).

END OF PAPER

ROUGH WORKSHEET

NOT TO BE TAKEN AWAY

Anything on this page will NOT be graded.

ROUGH WORKSHEET

NOT TO BE TAKEN AWAY

Anything on this page will NOT be graded.