University Number: _____

Seat Number: _____

THE UNIVERSITY OF HONG KONG

FACULTY OF ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE

DASC7606 Deep Learning
(Subclass D)

Date: May, 8th, 2024 **Time: 6.30 PM – 8.30 PM**

There are 3 questions in total. The first question gives 20/100 points, the second one and third one 40/100 each. Answer ALL questions. They are all COMPULSORY.

Write your answers on this examination paper in the corresponding space after the questions.

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: _____

University Number: _____

Question 1. (20/100) (General Questions)

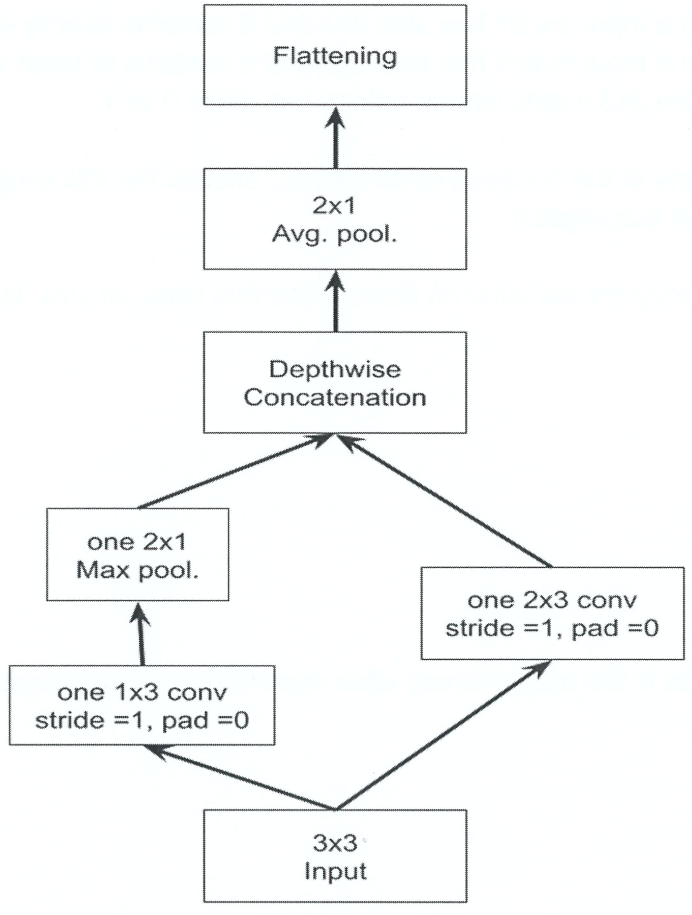
1.1 (7/100) Why do we ideally use the test set only once to evaluate the accuracy of a ML model?_____

_____

1.2 (7/100) Mention at least one variant of gradient descent (among those we have seen in our course) that aims at alleviating the issues with ill-conditioned loss functions. _____

1.3 (6/100) What are the three main so-called gates in an LSTM? (you need to only specify their names) _____

Question 2. (40/100) CNN. Consider the following CNN architecture inspired by the Google inception module presented in our course.



The input is a binary matrix whose entries are either 0 or 1. After flattening, for simplicity we use a threshold function f(**x**,**w**,t) as an activation function, which outputs 1 if the dot product **xw** >=t, 0 otherwise (**w**,t >= 0). We assume w=(1,3/2).
All other activation functions are the identity function.
You should determine the filters so that the CNN will classify 1 the following input matrix:

1 0 0

0 0 0

1 1 1

and classify 0 any other matrix. To this end, we need some assumptions on the input matrix. Choose one of the following assumptions, with the goal of choosing the least restrictive one, that is, you should try to maximize the number of matrices your CNN receives in input while always providing correct classifications.

2.1 (10/100) Which of the following assumptions do you choose? _____

a) The input matrix has size 3x3 and it contains at most 4 ones.
b) The input matrix has size 3x3 and it contains exactly 4 ones.
c) The input matrix has size 3x3 and it contains at least 4 ones.
d) Any 3x3 matrix whose entries are either 0 or 1.

After choosing one of the previous assumptions, answer the following questions while making that assumption.

2.2 (10/100) Specify the convolution filters in the first layer, of size 1x3 and 2x3 respectively

2.3 (10/100) What is the output tensor after depthwise concatenation?

2.4 (10/100) What is the value of t? _____

Question 3 (Language Models, 40/100). Consider an RNN where x_t , y_t , s_t ,denote the input, the output and the state of the RNN at step t, respectively. Recall that given matrices W,U,V we have that:

$$s_t = g(Wx_t+Us_{t-1}) \text{ and } y_t = g(Vs_t),$$

at every step t>0, where we are ignoring the biases, for simplicity. We consider a language model (LM) where the input x_t is a character in the English alphabet and y_t is a probability vector over a subset of those characters. Recall that at step t, a character is sampled from y_t and then fed in input to the RNN at step t+1. For this question, we consider only the following characters with the following ordering (in alphabetical order except <start> and <stop>):

<start>, 'a', 'e', 'i', 'n', 'r', 's', 't', 'u', <stop>

with <start>,<stop> being special characters denoting the beginning and the end of the sequence, respectively.

Activation function: We use pointwise ReLu as activation function, i.e. $g(x_1,...,xd) = (\max(0,x_1),...,\max(0,x_d))$.

x_t vector: We use the one-hot encoding representation for the 10 characters, using the ordering specified above. E.g. x_t='1000000000' for <start>, x_t='0100000000' for 'a', x_t='0010000000' for 'e', x_t='0000000001' for <stop>. Observe that <stop> is never received in input, however, we use this encoding for simplicity.

y_t vector: probability distribution over 'a', 'e', 'i', 'n', 'r', 's', 't', 'u', <stop> stored in that order, that is y_t^1 is the probability that 'a' is generated next, y_t^2 the probability that 'e' is generated next, etc.

Moreover, observe that s_t is a vector with 10 dimensions, with s_0=**0**.
There is a special symbol 'f' that should be understood to understand what our RNN does. W has size 10 x 10 with *W*['f','r']=1, *W*[i,i]=1 for every *i* in {<start>, 'a', 'e', 'i', 'n', 'r', 's', 't', 'u'}, W[i,j]=0 otherwise. V is specified in Table1. Observe that U is not specified yet, this will be one of your tasks.

	f	<start>	a	e	i	n	r	s	t	u
a	-1			1						
e					1/2		1/2			
i		1								
n										1
r					1/2					
s	1	-1	-1		-1	-1	-1	-1	-1	-1
t			1							
u							1/2			
<stop>						1			1	

Table 1. Matrix V of size 9 x 10 (cells in grey contain row/column names). V entries are 0, unless otherwise specified.'f' plays an important role that should be understood to understand what our RNN does.

3.1 (10/100) What is the size of U? _____

3.2 (10/100) Specify U so that the RNN produces only English sentences that are grammarly correct and with some meaning in English (e.g. ``she swims'').

_____

3.3 (10/100) What are the sentences computed by the RNN and with which probability? _____

3.4 (10/100) Change V, W and U so that the RNN produces the same sentences as specified above but uniformly at random, that is, each sentence must have the same probability to be generated.

END OF PAPER