University Number: _____

Seat Number: _____

THE UNIVERSITY OF HONG KONG

FACULTY OF ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE

DASC7606 Deep Learning
(Subclasses A, B & C)

Date: Wednesday, May 8, 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: _____

(30 pts) Question 1: Convolutional Neural Network

Consider a convolutional neural network with the layers defined in the Layer Definition column of TABLE below where:

- CONVn-N denotes a convolutional layer with N n×n filters
- POOLn denotes a n×n max-pooling layer with stride of n and 0 padding
- FLATTEN flattens its inputs (identical to torch.nn.flatten / tf.layers.flatten)
- FC-N denotes a fully-connected layer with N neurons

TABLE

Layer #	Layer Definition	Dimension* (H×W×C) of Output	Number of Parameters	Number of Computation Operations	Local Receptive Field
0	Input	32×32×3			
1	Conv3-8				
2	Pool2				
3	Conv3-16				
4	Pool2				
5	FLATTEN				1
6	FC-4				

* Note: The dimensions of the outputs are in the format H × W× C where H, W, C are the height, width and channel dimensions, respectively.

(a) (2 pts) For CONV3-N, what padding size and stride size will keep the dimensions of the input and the output of the convolution layer the same? (Fill in table below.)

Answer:

Padding size	Stride size

(b) (15 pts) Fill in TABLE (see above) assuming the padding size and stride size in (a). (Hint: Work backwards to fill in Local Receptive Field column.)

(c) (5 pts) How would you convert the FLATTEN and FC-4 layers into a single convolution layer CONVn-N? (Fill in the table below.)

Answer:

n	N	Filter dimension	Number of Parameters	No. of computation operations

(d) (4 pts) After the conversion in (c), suppose we now want to process an image of size 256×256×3 as input using the resultant convolutional neural network (CNN), i.e. the dimension of the Input Layer is now 256×256×3. What is output dimension after Layer 4? What is the output dimension of the final layer of the CNN? (Fill in the table below.)

Answer:

Output dimension of Layer 4	Output dimension of final layer

(e) (4 pts) Suppose you revert back to an input dimension at Layer 0 of 32×32×3 and use a sliding window approach to process the image of size 256×256×3. How many sliding windows of size 32×32 would you use and with what stride size to get the same output dimension as in (d)? (Fill in the table below.)

Answer:

Number of sliding windows	Stride size

(30 pts) Question 2: Transformer and Attention

Consider the following set of words W = {"cat", "milk", "it", "sweet", "hungry"}.

Assume each word is represented by a word vector. Among the features in the word vector are two features: STATE and TASTE.

Below are 5 word vectors. For each word vector, the first number shown is the value for STATE and the second is the value for TASTE.

V1 = (..., 0, 4, ...)
V2 = (..., 1, 3, ...)
V3 = (..., 2, 2, ...)
V4 = (..., 2, 2, ...)
V5 = (..., 4, 0, ...)

The higher the value for STATE, the more related the word is to the word "hungry". The higher the value for TASTE, the more related the word is to the word "sweet"

Suppose we have a transformer with two attention heads: one for STATE and one for TASTE.

(a) (3 pts) Which word vector would you assign to the word "sweet"? What about the other words? (Fill in the table below with V1 to V5.)

Answer:

Word:	cat	milk	it	sweet	hungry
Word vector:					

(b) (3 pts) Let us consider the Q, K and V vectors for the attention head for STATE. Suppose we are processing the sentence: "***The cat drank the milk because it was sweet***" which contains the 4 words "cat", "milk", "it" and "sweet" from W.

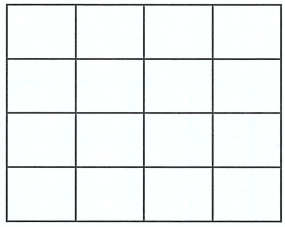
We assume that Q=K=V and they are each 4 × 1 matrix comprised of the STATE value for each of these 4 words in their word vectors. What does the Q=K=V matrix look like? (Fill in the blanks below.)

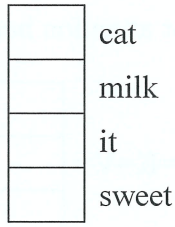
Answer: For attention head of STATE,

Q=K=V= [ ] cat
[ ] milk
[ ] it
[ ] sweet

(c) (5 pts) Compute the attention matrix $Q * K^T$ and $(Q * K^T) * V$.
(Note: this is a simplified form of the Attention equation used in Transformer without Softmax and normalization.)

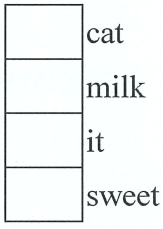
Answer: For attention head STATE,

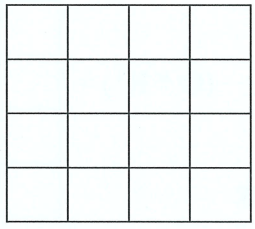
$(Q * K^T) =$ 

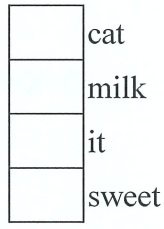
$(Q * K^T) * V =$ 

(d) (5 pts) Repeat (b) and (c) for the attention head of TASTE.

Answer: For attention head TASTE,

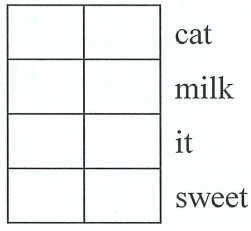
Q=K=V= 

$(Q * K^T) =$ 

$(Q*K^T)*V =$ 

(e) (2 pts) Combining the results of (c) and (d) will form the new embedding (2-dim) of each word, "cat", "milk", "it" and "sweet". Provide the new word embeddings.

Answer: New embedding for words…



(f) (2 pts) Show that which word "milk" or "cat" is more related to "it" by using cosine similarity. (Fill in the table below.)

Answer:

Dot-product of "milk" and "it"	
Dot-product of "cat" and "it"	
Which is more relate to "it"?	

(g) (10 pts) Repeat the above for the sentence "***The cat drank the milk because it was hungry***".

Answer:

For attention head STATE,

$Q=K=V=$ | | cat
milk
it
hungry

$(Q * K^T) =$ | | | | |

$(Q*K^T)*V =$ | | cat
milk
it
hungry

For attention head TASTE,

$Q=K=V=$ | | cat
milk
it
hungry

$(Q * K^T) =$ | | | | |

$(Q*K^T)*V =$ | | cat
milk
it
hungry

New embedding:

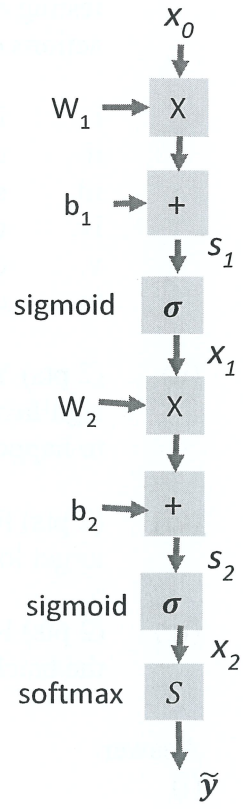
| | | cat
milk
it
hungry

Dot-product of "milk" and "it"	
Dot-product of "cat" and "it"	
Which is more relate to "it"?	

(40 pts) Question 3:

(a) **(10 pts) Backpropagation**

The figure to the right shows the computation graph for the full neural network to classify 10 objects as specified by the equations below:



$\mathbf{s}_1 = \mathbf{W}_1\mathbf{x}_0 + \mathbf{b}_1$
$\mathbf{x}_1 = \text{sigmoid}(\mathbf{s}_1) = \sigma(\mathbf{s}_1)$
$\mathbf{s}_2 = \mathbf{W}_2\mathbf{x}_1 + \mathbf{b}_2$
$\mathbf{x}_2 = \text{sigmoid}(\mathbf{s}_2) = \sigma(\mathbf{s}_2)$
$\tilde{\mathbf{y}} = [\text{activation}](\mathbf{x}_2)$
$J = [\text{cost}]\ (y, \tilde{\mathbf{y}})$

(i) (2 pts) What are the [activation] and [cost] functions used in the network?

(ii) (3 pts) Assuming the dimensions of $\mathbf{x}_0$, $\mathbf{x}_1$, $\mathbf{x}_2$ are $n_0 \times 1$, $n_1 \times 1$, $n_2 \times 1$ respectively, what are the dimensions of $\mathbf{s}_1$, $\mathbf{W}_1$, $\mathbf{b}_1$, $\mathbf{s}_2$, $\mathbf{W}_2$, $\mathbf{b}_2$ and $\mathbf{y}$?

(iii) (5 pts) Learning is to adjust $\mathbf{W}_1$, $\mathbf{b}_1$, $\mathbf{W}_2$, and $\mathbf{b}_2$ to minimize J. Show how the backpropagation values, $\boldsymbol{\delta}_2 = \frac{\partial J}{\partial \mathbf{s}_2}$ and $\boldsymbol{\delta}_1 = \frac{\partial J}{\partial \mathbf{s}_1}$, help the computation of the gradients $\frac{\partial J}{\partial \mathbf{W}_1}$, $\frac{\partial J}{\partial \mathbf{b}_1}$, $\frac{\partial J}{\partial \mathbf{W}_2}$ and $\frac{\partial J}{\partial \mathbf{b}_2}$ through chain rule.

Answer:

(i)

[activation]	
[cost]	

(ii)

$\mathbf{s}_1$	$\mathbf{W}_1$	$\mathbf{b}_1$	$\mathbf{s}_2$	$\mathbf{W}_2$	$\mathbf{b}_2$	$\mathbf{y}$

(iii)

$\frac{\partial J}{\partial \mathbf{W}_1}$	
$\frac{\partial J}{\partial \mathbf{b}_1}$	
$\frac{\partial J}{\partial \mathbf{W}_2}$	
$\frac{\partial J}{\partial \mathbf{b}_2}$	

(b) **(10 pts) Training**

(i) (4 pts) After you trained the network with a low training error, you have a high testing error. Explain the reason for this phenomenon. Which of the following actions can improve your result.

i. increase the size of the network
ii. use data augmentation
iii. stop training earlier
iv. use ReLU as the activation function
v. dropout is applied differently during training and testing
vi. change the cost function, explain how

(ii) (2 pts) You are training a neural network and notice that the testing error is significantly lower than the training error. Name any possible reasons for this to happen.

(iii) (2 pts) For small batch sizes, the number of iterations required to reach the target loss decreases as the batch size increases. Why is that?

(iv) (2 pts) For large batch sizes, the number of iterations does not change much as the batch size is increased. Why is that?

Answer:
(i)

(ii)

(iii)

(iv)

(c) **(20 pts) GAN**

(i) (4 pts) Draw the architecture diagram of GAN.
(ii) (4 pts) List some difficulties in training GAN.
(iii) (4 pts) How will the training be conducted when the discriminator gives 0.2 on a generated instance? (specify the loss value)
(iv) (4 pts) What type of neural network architecture would be used to color the BW pictures? Why?
(v) (4 pts) Can you describe the similarity of GAN with Turing test?

Answer:
(i)

(ii)

(iii)

(iv)

(v)

END OF PAPER