University Number: _____

Seat Number: _____

THE UNIVERSITY OF HONG KONG

SCHOOL OF COMPUTING AND DATA SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

DASC7606 Deep Learning

Date: Monday, May 19, 2025 **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 a 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.

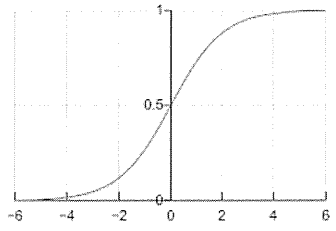
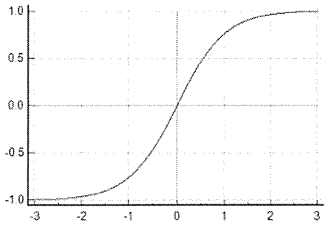
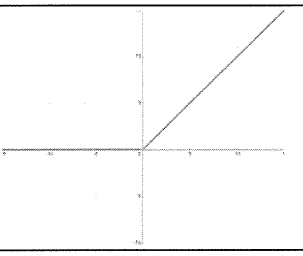
Brand and Type of Calculator: _____

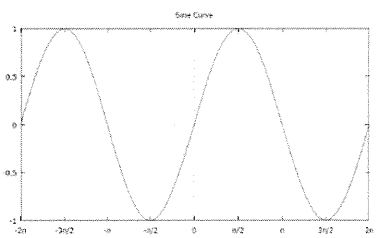
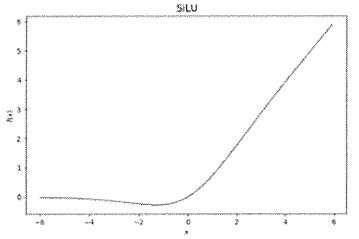
Question	Max	Marks
1	12	
2	10	
3	10	
4	10	
5	10	
6	24	
7	12	
8	12	
Total	**100**	

University Number: _____

(12 marks) Question 1: Activation Functions

Consider the following activation functions:

Sigmoid	Tanh	ReLU
		
$g(s) = \frac{1}{1+e^{-s}}$	$g(s) = \frac{2}{1+e^{-2s}} - 1$	$g(s) = \begin{cases} 0 \text{ for } s < 0 \\ s \text{ for } s \geq 0 \end{cases}$

Sine	SiLU (Sigmoid Linear Unit)
	
$g(s) = \sin(s)$	$g(s) = \frac{s}{1+e^{-s}}$

(a) (2 marks) Which function(s) are zero-centred?

(b) (2 marks) Which function(s) are computationally efficient?

(c) (2 marks) Which function(s) help to normalise or scale the data?

(d) (2 marks) Which function(s) are monotonic, i.e. as the input increases, the output does not decrease?

(e) (4 marks) The swish function is a family of function defined as:

$$\text{swish}_\beta(x) = x\,\text{sigmoid}(\beta x) = \frac{x}{1+e^{-\beta x}}$$

where β can be a constant or trainable. Researchers found that using swish as an activation function in artificial neural networks improves performance, compared to using ReLU and sigmoid. Why do you think this is so? [Hint: Consider the case where $\beta = 0$ and $\beta = 1$ and large values for β.]

(10 marks) Question 2: Distillation and Model Size

Consider a large language model whose size in terms of parameters is approximated by the following equation:

$$2VE + L(4E^2 + 3EH + 2E)$$

where
- V = 32000 = number of tokens in the vocabulary
- L = 40 = number of transformer layers
- E = 5120 = size of the embedding vector
- H = 13824 = size of the hidden layer in MLP

(a) (4 marks) Suppose you want to halve the size of the model for distillation by changing only one of the variables. Which of the variables could do the job? (The answer could be more than one.)

(b) (3 marks) What is the hardest variable to change and why?

(c) (3 marks) In general, which part of the language model contributes the most parameters to the model's size?

(10 marks) Question 3: Quantization

Suppose we want to quantize the weights of a pre-trained neural networks into a 4-bit format, assuming that the weights follow a zero-centred normal distribution and are normalised in the range of [-1, 1]. We can use the table below to map actual weight values to their nearest value in the table and then to its 4-bit representation:

value	-1.00	-0.70	-0.53	-0.39	-0.28	-0.18	-0.09	0.00
4-bit	0000	0001	0010	0011	0100	0101	0110	0111

value	0.08	0.16	0.25	0.34	0.44	0.56	0.72	1.00
4-bit	1000	1001	1010	1011	1100	1101	1110	1111

(a) (2 marks) What is the 4-bit representation for the value 0.90? What about -0.60?

(b) (2 marks) Suppose only the following GPUs are available to you:

GPU:	H100	L40S	V100	RTX4090	T4
VRAM:	80GB	48GB	32GB	24GB	16GB

Given 4-bit quantization, what is the smallest GPU in terms of VRAM that you could use to store a 70B parameter model)?

(c) (4 marks) Now suppose the distribution of the weights (instead of following a zero-centred normal distribution) follow a uniform distribution (i.e. all values in [-1, 1] are equally likely). What would your table to map with values to their 4-bit representation look like? Answer by filling in the values in the table below.

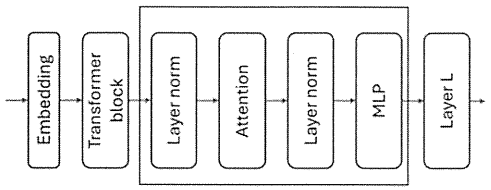
value								
4-bit	0000	0001	0010	0011	0100	0101	0110	0111

value								
4-bit	1000	1001	1010	1011	1100	1101	1110	1111

(d) (2 marks) Give a reason why the mapping in (a) would be used instead of the one you gave in (c).

(10 marks) Question 4: Low-Rank Adaptation (LoRA)

Suppose you want to use LoRA to fine-tune a LLM with the architecture shown in the diagram below and specifications shown in the table below.



Number of parameters	70B
Size of embedding vector = E	5120
Size of hidden layer in MLP = H	13824

(a) (3 marks) Focusing only on the attention component of the model, suppose
- there are four E × E weight matrices whose weights are to be learnt; and
- the appropriate change ΔW to each matrix W of these four matrices for the fine-tuning can be decomposed into the multiplication of two matrices A and B where A has the dimension of E × r and B has the dimension of r × E.

If the goal is to reduce the number of parameters by a factor of at least 1000, what is the maximum value you could pick for r? Show your chain-of-draft calculations for r.

(b) (3 marks) Focusing on the multi-layer perceptron (MLP) component of the model, suppose
- there are two E × H weight matrices whose weights are to be learnt; and
- the appropriate change ΔW to each matrix W of these two matrices for the fine-tuning can be decomposed into the multiplication of two matrices A and B where A has the dimension of E × r and B has the dimension of r × H.

If the goal is to reduce the number of parameters by a factor of at least 1000, what is the maximum value you could pick for r? Show your chain-of-draft calculations for r.

For each of the following multiple-choice questions, please select one or more answers as appropriate by circling the letter before the selected answer(s).

(c) (2 marks) Which of the following factors will affect the choice of r?

A. The size of the GPU used for training
B. The size of the GPU used for inference
C. The size of the data for fine-tuning
D. The performance of the fine-tuned LLM in terms of the quality of the output

(d) (2 marks) Which of the following are advantages of using LoRA?

A. Efficient training
B. Prevents overfitting
C. Adaptation to multiple tasks
D. More accurate weights
E. Improves the quality of the output of the LLM

(10 marks) Question 5: Transformer and Word Embeddings

The table on the left (below) maps tokens of a vocabulary to their corresponding embeddings. For example, the embedding of the token "books" is the 4-element vector (-4, 0, 0, 0). Meanwhile, the table on the right (below) maps positions (1 to 7) to their corresponding embeddings. For example, position 3 is mapped to the 4-element vector (0.0, 0.5, 0.0, 0.0).

Token	Embedding of Token			
Ann	1	0	0	1
Bill	1	0	0	-1
like	0	0	1	1
hate	0	0	1	-1
cat	-1	0	0	0
dog	-2	0	0	0
book	-4	0	0	0
white	0	-1	0	1
brown	0	-1	0	2
black	0	-1	0	3
big	0	0	-1	1
small	0	0	-1	-1
#s	0	0	0	1

Position	Embedding of Position			
1	0.0	0.0	0.0	0.5
2	0.0	0.0	0.5	0.0
3	0.0	0.5	0.0	0.0
4	0.5	0.0	0.0	0.0
5	0.0	0.0	0.5	0.5
6	0.0	0.5	0.0	0.5
7	0.5	0.0	0.0	0.5

(a) (3 marks) The addition of the token embedding and the position embedding is used to obtain the following embedding for the input to a transformer:

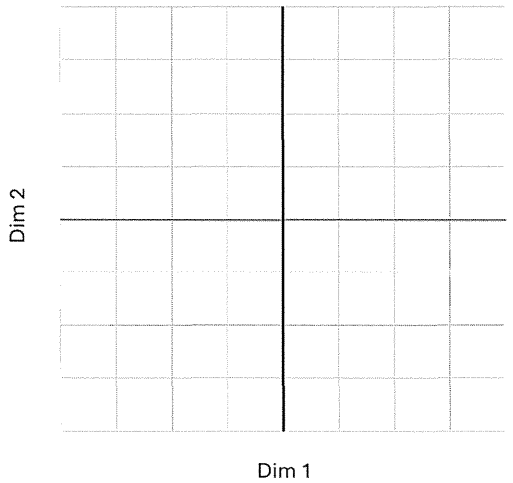
1.0	0.0	0.0	1.5
0.0	0.0	1.5	1.0
0.0	0.5	0.0	1.0
0.5	0.0	-1.0	-1.0
0.0	-1.0	0.5	1.5
-1.0	0.5	0.0	0.5
0.5	0.0	0.0	1.5

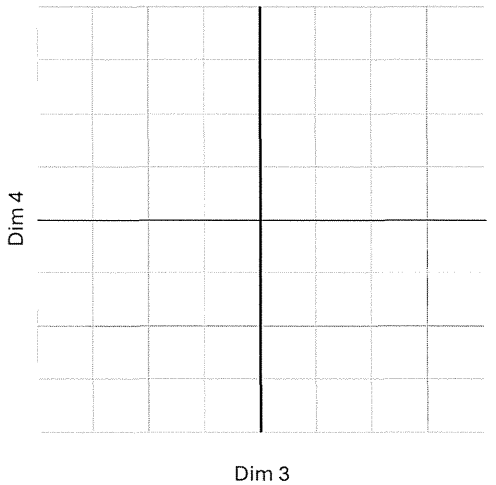
What are the tokens of the input?

(b) (3 marks) Consider the embedding of the input below.

Plot and label the input tokens "big", "black" and "dog" in the chart on the left (below) based on their embedding value for the first two dimensions (i.e. Dim 1 and Dim 2) and in the chart on the right (below) based on their embedding value for the last two dimensions (i.e. Dim 3 and Dim 4).

Input	Embedding of Input			
Bill	1.0	0.0	0.0	-0.5
hate	0.0	0.0	1.5	-1.0
#s	0.0	0.5	0.0	1.0
big	0.5	0.0	-1.0	1.0
black	0.0	-1.0	0.5	3.5
dog	-2.0	0.5	0.0	0.5
#s	0.5	0.0	0.0	1.5





(c) (4 marks) How would you expect the embedding of the token "dog" to change after the self-attention mechanism of the transformer is applied to these input embeddings?

(24 marks) Question 6: CNN and ImageNet

Please give short and concise answers to the following prompts/questions in the space provided following the prompt/question.

(a) (3 marks) Explain why applying 3×3 filters to three consecutive convolutional layers has the same effective local receptive field (LRF) as a 7×7 filter applied to a single convolution layer.

(b) (3 marks) Further to (a) above, explain why 3 × 3 filters applied to three convolutional layers uses less parameters than a 7 × 7 filter applied to a single convolution layer.

(c) (3 marks) What padding size and stride would you use for a 3×3 filter applied to a 28×28×192 input to obtain a 28×28×128 output?

(d) (3 marks) What is the cost in terms of the number of multiplication operations if a 3×3 filter is applied to a 28×28×192 input to obtain a 28×28×128 output?

(e) (3 marks) Alternatively to (d) above, what is the cost in terms of the number of multiplication operations if a 1×1 filter is applied to the 28×28×192 input to first obtain a 28×28×64 intermediate output, followed by applying a 3×3 filter to the intermediate output to obtain the final 28×28×128 output? What is the name of this technique?

(f) (3 marks) How much cost is saved (if any) if the technique in (e) is adopted instead of the technique in (d)?

(g) (3 marks) What technique was used in ResNet to enable the training of a very deep neural network without significant performance loss?

(h) (3 marks) What technique was used in GoogleNet to enable the training of a deep neural network without significant performance loss?

(12 marks) Question 7: GAN

For each of the following multiple-choice questions, please select one or more answers as appropriate by circling the letter before the selected answer(s).

(a) (4 marks) Which loss function is commonly used in GANs?

A. Cross-entropy loss
B. Mean squared error loss
C. Binary logistic loss
D. Softmax loss

(b) (4 marks) Which of the following are related to the generation of sample variations?

A. Input in the generator component of a GAN
B. Adjusting the weights and biases of the generator
C. Mode collapse in GANs
D. Conditional GAN

(c) (4 marks) How does the training process of a GAN usually work with a generated sample?

A. Only the generator is trained
B. Either the generator or discriminator is trained
C. Both the generator and discriminator are trained for score 0.5
D. The generator is trained for score 1 and the discriminator for score 0

(12 marks) Question 8: Masking

Please give short and concise answers to the following prompts in the space provided following the prompt.

(a) (4 marks) Please explain how masking is used in the transformer.

(b) (4 marks) Please explain how masking is used in BERT.

(c) (4 marks) Please explain how masking applies to occlusion sensitivity in vision.

END OF PAPER