COMP 5214 & ELEC 5680 **Name:** _____
Spring 2022
Midterm
7 April 2022
Time Limit: 80 minutes Student ID: _____

This exam contains 10 pages (including this cover page) and 3 questions.
Total of points is 100.

Grade Table (for teacher use only)

Question	Points	Score
1	39	
2	35	
3	26	
Total:	100	

- Prepare 3 white papers. Or you can print the PDF and write on the printed midterm. You can also use a tablet to answer the questions if you want.
- Prepare a black pen and a smartphone to capture images
- Sign the honor code. No communication among students
- Turn on video cameras (ensure that's you)
- Open book exam. Free to browse materials (including the hyperlinks to external websites) on the course website
- No Google. No external websites during the exam, but you can download external materials in advance
- Exam Time: 12:00 pm - 1:20 pm.
- A PDF of the midterm will be shared by 12:00 pm in the Zoom chatroom and via email announcement
- Write your name, student ID, and answers on your white papers
- Take clear images of your answers

Honor Code

Honesty and integrity are central to the academic work of HKUST. Students of the University must observe and uphold the highest standards of academic integrity and honesty in all the work they do throughout their program of study.

As members of the University community, students have the responsibility to help maintain the academic reputation of HKUST in its academic endeavors.

Sanctions will be imposed on students, if they are found to have violated the regulations governing academic integrity and honesty.

Please write "I have read and understood the honor code" on your white paper after your name and student ID.

1. (39 points) Short questions. Please choose the right choices for each question. There may be more than one correct choice. There is at least one correct answer.

   1. Which item(s) is NOT considered as the benefit of using recurrent neural networks (RNNs)?
      (a) Parameter sharing
      (b) Preventing gradient vanishing
      (c) Capable of modeling long-term dependencies
      (d) Capable of preserving order
   2. Select the metric(s) that is commonly used for image segmentation?
      (a) Mean Pixel Accuracy
      (b) Root Mean Square Error
      (c) Intersection Over Union
      (d) BLEU Score
   3. What kind of feature map(s) will be generated from the following convolution kernel?

      $$\begin{bmatrix} 2 & 4 & 2 \\ 4 & 8 & 4 \\ 2 & 4 & 2 \end{bmatrix}$$

      (a) Blurry feature map
      (b) Sharpen feature map
      (c) Edge feature map
      (d) A segmentation feature map
   4. Which method can add non-linearity to a linear neural network?
      (a) Activation function (e.g., ReLU)
      (b) Average pooling
      (c) Regularization loss with non-linear terms
      (d) SGD
   5. Which is TRUE about reflection removal?
      (a) After removing reflection from an image, the average intensity of an image will decrease
      (b) Reflection removal becomes easier with polarization images, compared to normal RGB images
      (c) Deep neural networks are the only feasible model for reflection removal
      (d) Reflection removal is a image decomposition problem that decomposes an image into two parts
   6. Consider a neural network trained with a total loss (cross entropy loss + regularization loss). However, the total loss decreases steadily but very slowly from the beginning. What could be the cause?

(a) Regularization loss is too large
(b) Learning rate is too low
(c) Learning rate is too large
(d) Missing activation function

7. Which description is correct for the transformer?

(a) There are usually three elements for attention: key, query, value.
(b) Self-attention needs to compute the similarity for each pair of keys and values.
(c) Adding position embedding helps solve the problem of permutation invariant.
(d) There is no non-linearity in transformers.

8. Which description is correct about the following convolution layer and pooling layer? The size of input is $32 \times 32 \times 1$. The first layer is a CONV layer with 8 filters, each of size $3 \times 3 \times 1$, 1 is the depth of the activation volume of previous layer. The total padding (left plus right padding) is 2, and the stride is 1. The next is 2×2 max-pooling layer with stride 2 (no padding). The input is fed into the CONV layer then the max-pooling layer.

(a) The activation volume for CONV layer is $32 \times 32 \times 8$.
(b) The number of parameters of CONV layer is $8 \times 1 \times 5 \times 5$.
(c) The size of the activation volume of the pooling layer is $16 \times 16 \times 8$.
(d) The parameters of the pooling layer is 2×2.

9. For kernel width k and stride s, input width w_{in} and total padding w_{pad}, output width w_{out} is?

(a) $(w_{in} + w_{pad} - k + 1)/s$
(b) $(w_{in} + w_{pad} - k - 1)/s$
(c) $(w_{in} + w_{pad} - k)/s + 1$
(d) $(w_{in} + w_{pad} - k)/s - 1$

10. Which can be the cause of the gradient becoming too small (vanishing gradient)?

(a) Choice of activation function is not suitable
(b) The input picture is too small
(c) The dataset is too small
(d) The loss function is not a simple function

11. Which of the following propositions are true about a CONV layer?

(a) The number of weights depends on the number of channels of the input volume
(b) The number of biases is equal to the number of filters
(c) The total number of parameters depends on the stride

(d) The total number of parameters depends on the padding

12. Which of the following are valid activation functions (elementwise non-linearity) you could use in a neural network? (That is, which functions could be effective when training a neural net in practice?)

(a) $f(x) = min(0, x)$
(b) $f(x) = 2x$
(c) $f(x) = max(0.1x, 0.9x)$
(d) $f(x) = \begin{cases} 1, & \text{if } x > 0 \\ \text{-1}, & \text{otherwise} \end{cases}$

13. Pretaining a big foundation model is beoming more and more popular today. Which is TRUE about pretrained models?

(a) Pretrained models are required to achieve state-of-the-art performance in computer vision tasks
(b) It is necessary to have labelled data to pretrain a model
(c) Transformers can be the network architecture for pretraining a model
(d) Pretrained models should be small, and can not have more than 1 billion parameters

2. (35 points) Short questions. Only the final answer is needed for each question.

   1. Let's consider the LeNet-5 model as follows.

Layer		Feature Map	Size	Kernel Size	Stride	Activation
Input	Image	1	32x32	-	-	-
1	Convolution	6	28x28	5x5	1	tanh
2	Average Pooling	6	14x14	2x2	2	tanh
3	Convolution	16	10x10	5x5	1	tanh
4	Average Pooling	16	5x5	2x2	2	tanh
5	Convolution	120	1x1	5x5	1	tanh
6	FC	-	84	-	-	tanh
Output	FC	-	10	-	-	softmax

      Please show the number of trainable parameters in Layer 1, 5, and 6. Bias is considered in each convolutional filter and fully-connected layer (FC).

   2. Consider we have n pairs of keys and values as well as m queries. Each of key, value, or query is a vector of dimension d. What is a time complexity to perform the attention operation? Please use the notation $O(\cdot)$ in terms of n, m, d for your result (assume matrix multiplication is implemented in a simple way).

3. Although a neural radiance field (NeRF) have strong novel view synthesis performance, it is slow to render an image at the test time: rendering a single 1280x720 image on one GPU takes about 30 seconds at test time. Why?

4. In style transfer, there are a style image and a content image with a style loss and a content loss. (1) What does the resulted image look like when we only use the style loss? (2) What does the resulted image look when we only use the content loss?

5. In internal learning (such as Deep Image Prior and Deep Video Prior), we do not need a training dataset for training. Therefore, given a test image or video, an internal learning model can run very fast (often less than 1 second) for this test image or video. Is this True or False? Please also explain why.

6. In a transformer, there is self-attention in its encoder and cross-attention in its decoder. Explain the difference between self-attention and cross-attention. Can we replace self-attention in the encoder with cross-attention?

7. If a graph is fully connected, then a graph neural network is the same as an MLP with fully connected layers. TRUE or FALSE. Please explain.

3. (26 points) **Encoder-Decoder**

Consider this simple encoder-decoder network:

$$x_1 = max(0, x * f_1 + a_1), \quad (1)$$
$$x_2 = max(0, x_1 * f_2 + a_2), \quad (2)$$
$$x_3 = max(0, x_2 \tilde{*} g_1 + b_1), \quad (3)$$
$$s = x_3 \tilde{*} g_2 + b_2, \quad (4)$$

where input x has size 785×1, convolutional filters f_1 and f_2 each has size 3×1 and is applied with stride 2 and zero-padding size of 1 (the total padding size is 2), and $\tilde{*}$ denotes the transposed convolution operation which is used to upsize the vector of responses so that x_3 has the same length as x_1 and s has the same length as x. The filters g_1 and g_2 both also have size 3×1. Input x and output s have size 785×1, x_1 and x_3 have size $m_1 \times 1$ and x_2 has size $m_2 \times 1$.

(a) [4 point] What are the values of m_1 and m_2?

(b) [8 points] We can write the convolution in equation (1) as the matrix multiplication:

$$x_1 = max(0, M_{f_1} x + a_1)$$

Write down the elements of M_{f_1} and also specify its dimension. Assume the filter weights of f_1 are $(f_{1,1}, f_{1,2}, f_{1,3})$. Also write the similar expression for M_{f_2} to write the convolution in Equation (2) as a matrix multiplication. It is sufficient to show the elements in the first three rows of M_{f_1} and M_{f_2}

(c) [6 points] The transposed convolution in Equation (3) corresponds to pre-multiplying x_2 by $M_{f_2}^T$ but with the non-zero entries of $M_{f_2}^T$ defined by f_2 replaced by the corresponding elements of g_1. What is the number of non-zero entries in $M_{f_2}^T$ (write your answer in terms of m_2).

(d) [8 points] Say the convolutions in the first two layers are changed to have stride 3 and zero-padding of 2 (total padding is 4). In this case what are the values of m_1 and m_2? Please write down M_{f_1} in this case. It is sufficient to show the elements in the first three rows of M_{f_1}.