NAME: _____

STUDENT ID: _____

Directions:

- ❏ This is an open-book, open-note and open-internet exam.
- ❏ You must answer the questions in your own words.
- ❏ Your answers should be brief, concise and readable.
- ❏ You have 120 minutes to complete the exam. Good luck!

1. (3 pts) Name **THREE** famous applications of CNN

2. (3 pts) True or False:

   a. ____ A convolution layer has far fewer trainable parameters than a fully-connected layer of the same output size.

   b. ____ A max-pool layer has the same number of trainable parameters as a convolution layer.

   c. ____ RELU activation function is often used in the convolution layers.

3. (8 pts) Assume the following 2D image (**I**), a 2x2 filter (**f**), and a 3x3 filter (**g**).

I =

1	2	1	5
0	5	0	3
3	7	4	1
1	8	1	0

f =

-1	1
1	-1

g =

-1	1	-1
1	-1	1
-1	1	-1

1) What is the output of convolving I with f using stride = 1 and padding = 0?

2) What is the output of convolving I with f using stride = 2 and padding = 0?

3) What is the output of convolving I with g using stride = 1 and padding = 1?

4) What is the output of max-pooling I with f using stride = 2 and padding = 0?

4. (6 pts) Consider the following CNN architecture. Fill in the table below.

CNN layers	Shape	# of activations	# of learnable params
Input	(100,100,3)		N/A
CONV (5x5,s=1,n=5,p=2)			
MAXPOOL (2x2, s=2)			
CONV (5x5,s=1,n=10,p=0)			
MAXPOOL (2x2, s=2)			
Flatten		N/A	N/A
DENSE (100)			
Softmax (10)	(10,1)	10	

5. (4 pts) Give **TWO** reasons why a convolution layer works better than a fully-connected layer in image classification tasks?

6. (4 pts) ResNet introduces the concept of “shortcuts”. What are they? How does it improve a CNN architecture?

7. (4 pts) Suppose you are to train a CNN to classify whether an image has MUIC logo or not. However, you don't have enough images with MUIC logo to train the network from scratch. What technique would you use to train the classifier? Also, explain the process.

8. (6 pts) From the face recognition application that we discussed in class,

    a. What is the triplet loss? (Write down the loss function and describe variables)

    b. Explain the training process

    c. Explain how to use the trained model in a real-time system

9. (3 pts) Give **THREE** kinds of data that are suitable for sequence models?

10. (3 pts) Name an application that uses each of the following sequence models.

    a. One-to-many

    b. Many-to-one

    c. Sequence-to-sequence

11. (6 pts) Simple RNN is vulnerable to exploding/vanishing gradient problems.
    a. Explain why is this the case?

    b. What can we do to get around the exploding gradient problem?

    c. What can we do to get around the vanishing gradient problem?

12. (6 pts) Consider the following notations in a seq2seq model. Let $[h_1,h_2,...h_n]$ be the encoding vectors of an input sequence $[w_1,w_2,...,w_n]$. Let s_t denote the current decoder hidden state.

    a. Write down the attention score of w_k at the current timestep t. (There are multiple solutions. You only need to suggest one)

    b. Suppose $\alpha_k{}^t$ is the normalized attention score of w_k at current timestep t. Write down the attention output vector at the current timestep t.

    c. List **TWO** benefits of attention mechanism.

13. (6 pts) A transformer model consists of two key components: the encoder and the decoder.
    a. What does the encoder do?

    b. What does the decoder do?

    c. BERT models are members of the transformer family. They only consist of _____(Choose one: encoders / decoders)

    d. GPT models are also members of the transformer models. They only consist of _____ (Choose one: encoders / decoders)

14. (5 pts) Explain how you would use a pre-trained BERT model to train sentiment analysis classifier