NAME: VILYOM	Narola Wickys
STUDENT ID:	6081050

## **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

  - · Self-Driving Car · Face Recognition
- Video Analysis2. (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).

0	I = 0	0	0	0	0
0	1	2	1	5	0 (
Ø	0	5	0 -	3	<b>O</b> TIEL TI
0	3	7	4	1	0 (
8	1	8	1	0	0
Q	8	0	0	0	0
	f =				
	-1	1			

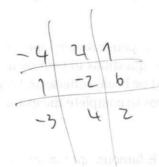
4	4	0
-9 .1 mil.	-2	Ь
-3	Ц	(-z

g =		
-1	1	-1
1	-1	1
-1	1	-1

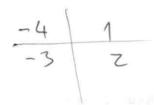
1) What is the output of convolving I with f using stride = 1 and padding = 0?

12-1

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)	3,0,000	N/A
CONV (5x5,s=1,n=5,p=2)	(100, 100, 31	30,000	228
MAXPOOL (2x2, s=2)	(50,50,3)	7,500	8
CONV (5x5,s=1,n=10,p=0)	(46,46,3)	6,348	2 2 8
MAXPOOL (2x2, s=2)	(23, 23, 3)	1,587	emana O sagar
Flatten	(15871)	N/A paged as	N/A
DENSE (100)	(1,387,1)	1,387	
Softmax (10)	(10,1)	10	13880

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

· Far less tunable purameters

· Geart Edge detaction which help reducing parameter without loss of details.

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

. Taking Activation from a layer and feed it to another deeper layer

. It greatly help reduing/eliminate vanishing and exploding gradient

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.

· Data Augmentation

- · Mirroring -> Dedecting text in logo better
- · Rotation > Rotate image by no
- · Shearing Apply linear transformation to deter
- · Color Shifting or Remar Gam, Charge Q, Ca, B

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)

L(A,B,C) = max (IfcA)-fcBs 12-IfcA)-fcN112+ alpha, 0) }

A,B,C

T = Sum (L(ACil, Bcil, Ccil, i) images

b. Explain the training process

- · Radon Pick Pics
- o Train on hard triplets
- · Signoid
- o Compale

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

Produce the woolet with pre train dertor and we take out put from the woole!

- 9. (3 pts) Give THREE kinds of data that are suitable for sequence models?
  - · Nove Sequence
  - · Text Sequence
  - · Video Frame
- 10. (3 pts) Name an application that uses each of the following sequence models.
  - a. One-to-many
    - · Music Creneration
  - b. Many-to-one
    - · Trext Analysis
  - c. Sequence-to-sequence
    - · Image Classi fication
- 11. (6 pts) Simple RNN is vulnerable to exploding/vanishing gradient problems.
  - a. Explain why is this the case?
    - · Cause it has long saquence size make unvitibling fraction to varish
  - b. What can we do to get around the exploding gradient problem?
    - · Gradient Clipping
    - · Truncate Buckprop ( not update all evelynt)
  - c. What can we do to get around the vanishing gradient problem?
    - · Weight Initialization
    - · Use CaRU
    - · Supervised learning

- 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<sub>k</sub> at the current timestep t. (There are multiple solutions. You only need to suggest one)

1 Stog(hn+Nn)

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

c. List TWO benefits of attention mechanism.

. Can pad rapidly analyse text

· Can pick up most important word

13. (6 pts) A transformer model consists of two key components: the encoder and the decoder.

a. What does the encoder do?

· Turn sequence to matrix

b. What does the decoder do?

· Generate output from matrix from ourcoder

- c. BERT models are members of the transformer family. They only consist of <u>lnolus</u> (Choose one: encoders / decoders )
- d. GPT models are also members of the transformer models. They only consist of decoders (Choose one: encoders (decoders)

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

Take the pretain model and Dut it to a Pt to train of the language