2021 Semester Two (November 2021) Examination Period

Faculty of Information Technology

EXAM CODES:	FIT5215
TITLE OF PAPER:	Deep Learning
EXAM DURATION:	2 hours 10 mins

Rules

During your eExam, you must not have in your possession any item/material that has not been authorised for your exam. This includes books, notes, paper, electronic device/s, smart watch/device, or writing on any part of your body. Authorised items are listed above. Items/materials on your device, desk, chair, in your clothing or otherwise on your person will be deemed to be in your possession. Mobile phones must be switched off and placed face-down on your desk during your exam attempt.

You must not retain, copy, memorise or note down any exam content for personal use or to share with any other person by any means during or following your exam. You are not allowed to copy/paste text to or from external sources unless this has been authorised by your Chief Examiner.

You must comply with any instructions given to you by Monash exam staff.

As a student, and under Monash University's Student Academic Integrity procedure, you must undertake all your assessments with honesty and integrity. You must not allow anyone else to do work for you and you must not do any work for others. You must not contact, or attempt to contact, another person in an attempt to gain unfair advantage during your assessment. Assessors may take reasonable steps to check that your work displays the expected standards of academic integrity.

Failure to comply with the above instructions, or attempting to cheat or cheating in an assessment may constitute a breach of instructions under regulation 23 of the Monash University (Academic Board) Regulations or may constitute an act of academic misconduct under Part 7 of the Monash University (Council) Regulations.

<u>Authorised Materials</u>

CALCULATORS	☑ YES	☐ NO	
DICTIONARIES	☐ YES	☑ NO	
NOTES	☑ YES	☐ NO	Notes - Double sided A4 x 1 - Physical Only
WORKING SHEETS	☐ YES	☑ NO	

PERMITTED ITEM ☐ YES ☑ NO

if yes, items permitted are:

Instructions

Once your exam finishes, you will be given time to scan
a QR code and upload your answers using
your smartphone and laptop.

Here's how to do it.

This examination is designed for the FIT5215 Deep Learning unit, Semester 2, 2021.

- This is an closed book exam with specifically permitted items.
- Please answer **ALL** questions.

It contains THREE (3) parts with a total of **100 marks**:

- **Part A** contains **12** multiple-choice questions, together they are worth a total of **25 marks**. For questions with more than one answer, a correct choice will receive a partial mark and incorrect choices will reduce the mark. To receive full marks, only correct choices must be selected.
- **Part B** contains **8** short workout questions, worth **40 marks**. These questions typically require short knowledge answers and calculations based on the knowledge you have learned from the unit. Having the calculator handy for these questions is recommended.
- **Part C** contains **10** mixed and written-answer questions, worth **35 marks**. These questions typically assess the knowledge and understanding of lecture contents.
- Once the exam duration is finished, your exam will automatically submit. Please ensure you finalise your answers before the end of the allocated exam time.

Instructions

Information

You can review your exam instructions by clicking the 'Show Instructions' button above.

Good luck with your exam!

Part A - Multiple Choices (12 questions, 25 marks)

Information

Part A consists of multiple-choice or single-choice questions, worthing 25 marks. For multiple-choice questions, incorrect choices will be penalized.

Question 1

Consider the following implementation of a feedforward NN. What is the total number of parameters?

2
Marks

```
dnn_model = Sequential()
dnn_model.add(Dense(units= 20, input_shape= (10,), activation= 'relu'))
dnn_model.add(Dense(units= 40, activation= 'relu'))
```

Select one:

Select one:

- a. 1060
- b. 1070
- c. 1000
- d. 1020
- e. 1050

Question 2

Let $f(w) = 3w^2 - 4w + 10$. Assume that we use gradient descent with the learning rate $\eta = 0.05$ to solve $\min_w f(w)$. At the iteration t, we are at $w_t = 2$. What is the value of w_{t+1} at the next iteration?

Select one:

2
Marks

Select one:

- a. 1.6
- b. 1.9
- c. 1.7
- d. 2.0
- e. 1.8

Question 3

Consider the optimization problem:

2 Marks

$$\min_{\theta} L(D;\theta) := \frac{1}{N}\sum_{i=1}^{N} l(x_i, y_i; \theta)$$

with the model parameter θ and D={(x_1, y_1),...,(x_N, y_N)} is a training set. Let us sample a batch of indices i_1,...,i_b uniformly from {1,...,N}. Which statement is correct about the update rule of stochastic gradient descent?

Select one:

Select one:

- a. $\theta_{t+1} = \theta_t - \frac{\eta}{b}\sum_{k=1}^{b} \nabla_\theta l\left(x_{i_k}, y_{i_k}; \theta_t\right)$
- b. $\theta_{t+1} = \theta_t + \frac{\eta}{N}\sum_{i=1}^{N} \nabla_\theta l(x_i, y_i; \theta_t)$
- c. $\theta_{t+1} = \theta_t - \frac{1}{b}\sum_{k=1}^{b} \nabla_\theta l\left(x_{i_k}, y_{i_k}; \theta_t\right)$
- d. $\theta_{t+1} = \theta_t + \frac{\eta}{b}\sum_{k=1}^{b} \nabla_\theta l\left(x_{i_k}, y_{i_k}; \theta_t\right)$

Question 4

Given a 3D input tensor with shape [64, 64, 3] over which we apply a conv2D with 15 filters each of which has shape [5,5,3], strides [3,3], and padding same. What is the shape of the output tensor?

2 Marks

Select one:

Select one:

- a. [21, 21, 15]
- b. [20,20,15]
- c. [22, 22, 15]
- d. [23,23,15]

Question 5

Consider an image classification task with five classes {cat=1, dog=2, lion=3, flower=4, cow=5}. Consider an image x. Assume that a Convolutional Neural Network gives prediction probabilities f(x)= [0.4, 0.2, 0.1, 0.2, 0.1] and categorial ground-truth label of x is flower. What is the cross-entropy loss suffered by this prediction?

Select one:

2 Marks

Select one:

- a. -log 0.1
- b. -log 0.2
- c. log 0.1
- d. log 0.2

Question 6

Assume that we have 4 classes in {cat = 1, dog = 2, lion = 3, monkey = 4}. Given a data example x with ground-truth label dog, assume that a feed-forward NN gives discriminative scores to this x as h_1=-3, h_2=10, h_3=5, h_4=-1. What is the probability to predict x as lion or p(y=lion| x)?

Select one:

2 Marks

Select one:

- a. 1
- b. $$\frac{e^{5}}{e^{-3}+e^{10}+e^{5}+e^{-1}}$$
- c. $$\frac{e^{-3}}{e^{-3}+e^{10}+e^{5}+e^{-1}}$$
- d. $$\frac{e^{10}}{e^{-3}+e^{10}+e^{5}+e^{-1}}$$

Question 7

Assume that the tensor before the last tensor of a CNN has shape [64, 32, 32, 10] and we apply 20 filters each of which has the shape [5,5,10] and strides= [3,3] with padding = ‘same’ to obtain the last tensor. What is the shape of the output tensor?

2
Marks

Select one:

Select one:

- a. [64, 10, 10, 20]
- b. [64, 12, 12, 20]
- c. [64, 11, 11, 20]
- d. [64, 10, 10, 10]

Question 8

Given a DL model f(x;θ) parameterized by θ where f(x;θ) represents the prediction probabilities of x associated with a ground-truth label y∈{1,...,M} , we find an adversarial example by

2
Marks

$$x_{adv} = argmax_{x' \in B_{\epsilon}(x)} l(f(x';\theta), y)$$

Which statements are correct?

Select one or many:

Select one or more:

- a. We maximally increase the chance to predict x with label y.
- b. We maximally increase the chance to predict x with any else label y'≠y.
- c. It is an untargeted attack.
- d. It is a targeted attack.
- e. We maximally decrease the chance to predict x with label y.

Question 9

Given a skip-gram model with vocabulary size 500 and embedding size 150, we consider a pair of target and context words with indices 2 and 8 respectively. Let U and V be two weight matrices connecting the input to hidden layers and hidden to output layers. What statements are correct?

2
Marks

Select one or many:

Select one or more:

☐
a.
Shape of U is [500,500] and shape of V is [150,150]

☐
b.
Shape of U is [500,150] and shape of V is [150,500]

☐
c.
Input to the network is one-hot vector 1_8.

☐
d.
The hidden value h is the row 2 of the matrix U.

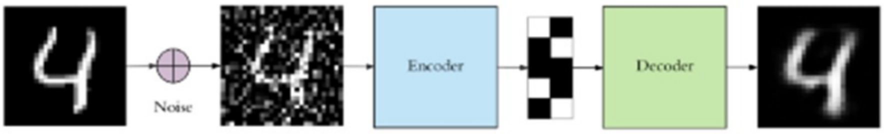
☐
e.
Input to the network is one-hot vector 1_2.

☐
f.
The hidden value h is the row 8 of the matrix U

Question 10

How to train a denoising auto-encoder with encoderf_θ and decoder g_Φ?

3
Marks



Select one or many:

Select one or more:

☐
a.
$$\min_{\theta,\Phi} \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}}\left[\mathbb{E}_{\boldsymbol{x}\prime\sim N(\boldsymbol{x},\eta I)}[d(\boldsymbol{x}, g_\Phi(f_\theta(\boldsymbol{x}')))]\right]$$

☐
b.
$$\min_{\theta,\Phi} \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}}\left[\mathbb{E}_{\epsilon\sim N(0,\eta I)}[d(\boldsymbol{x}, f_\theta(g_\Phi\ (\boldsymbol{x}+\boldsymbol{\epsilon}))]\right]$$

☐
c.
$$\min_{\theta,\Phi} \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}}\left[\mathbb{E}_{\epsilon\sim N(0,\eta I)}[d(\boldsymbol{x}, g_\Phi(f_\theta(\boldsymbol{x}+\boldsymbol{\epsilon}))]\right]$$

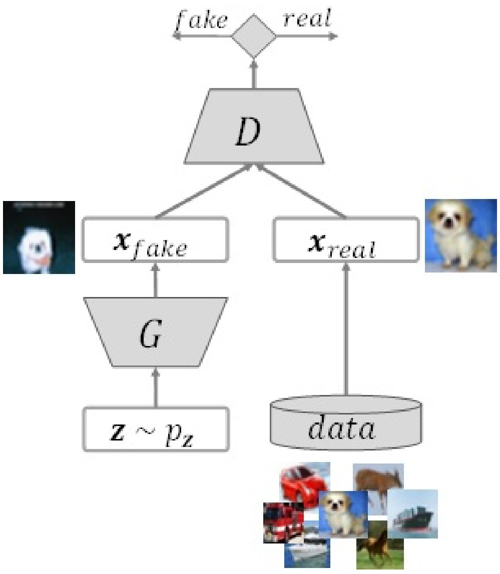
☐
d.
$$\min_{\theta,\Phi} \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}}\left[\mathbb{E}_{\boldsymbol{x}\prime\sim N(\boldsymbol{x},\eta I)}[d(\boldsymbol{x}, f_\theta(g_\Phi(\boldsymbol{x}')))]\right]$$

☐
e.
$$\min_{\theta,\Phi} \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}}\left[\mathbb{E}_{\epsilon\sim N(0,\eta I)}[d(\boldsymbol{x}, f_\theta(g_\Phi\ (\boldsymbol{x}+\boldsymbol{\epsilon}))]\right]$$

Question 11

How to train GANs?

2
Marks



Select one or many:

Select one or more:

☐ a. $\min_G \max_D J(G, D) = \mathbb{E}_{x \sim p_d(x)}[\log(1 - D(x))] + \mathbb{E}_{z \sim p(z)}[\log D(G(z))]$

☐ b. $\min_G \max_D J(G, D) = \mathbb{E}_{x \sim p_d(x)}[\log D(x)] + \mathbb{E}_{z \sim p(z)}[\log(1 - D(G(z))]$

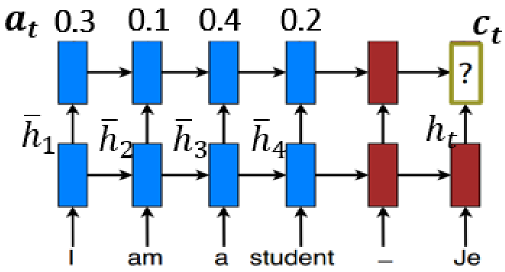
☐ c. $\min_{\theta,\Phi} \mathbb{E}_{x \sim \mathbb{P}}[d(\tilde{x}, g_\Phi(f_\theta(x))]$

☐ d. $\max_G \min_D J(G, D) = \mathbb{E}_{x \sim p_d(x)}[\log D(x)] + \mathbb{E}_{z \sim p(z)}[\log(1 - D(G(z))]$

Question 12

2
Marks

Consider the below seq2deq model. We apply global attention to compute the context vector c_t. What are correct?



Select many:

Select one or more:

☐
a.
The first word is more important than other words to the generation of the current output word.

☐
b.
$$c_t = 0.1\bar{h}_2 + 0.4\bar{h}_3 + 0.2\bar{h}_4$$

☐
c.
$$c_t = 0.3\bar{h}_1 + 0.1\bar{h}_2 + 0.4\bar{h}_3 + 0.2\bar{h}_4$$

☐
d.
$$c_t = 0.2\bar{h}_1 + 0.4\bar{h}_2 + 0.1\bar{h}_3 + 0.3\bar{h}_4$$

☐
e.
The third word is more important than other words to the generation of the current output word.

Part B - Short Workout & Knowledge Questions (8 questions, 40 marks)

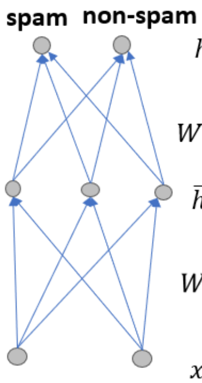
Information

Part B contains **8** short workout questions, worthing **40 marks**. These questions typically require short knowledge answers and calculations based on the knowledge you have learned from the unit. Having the calculator handy for these questions is recommended.

Question 13

Consider a feed-forward neural network as shown in the figure for spam email detection with two labels (spam=1 and non-spam=2). Assume that we feed a feature vector x=[-1,1]^T with true label y=2 to the network.

10
Marks



$$h^2, p = softmax(h^2)$$

$$W^2 = \begin{bmatrix} 1 & -1 & 1 \\ 0 & 1 & 1 \end{bmatrix}, b^2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$\bar{h}^1, h^1 = ReLU(\bar{h}^1)$$

$$W^1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \end{bmatrix}, b^1 = \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix}$$

$$x = [-1,1]^T = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, y = 2$$

13a)
What are the formulas and the values of $\bar{h}^1, h^1$?

3
Marks

Note: h_bar1 and h1 can be used to represent $\bar{h}^1, h^1$. You can use the notion [a,b,c]T or [[a], [b], [c]] to represent the corresponding column vector.

13b)
What are the formulas and the values for the logits h^2 and the prediction probabilities p?

3
Marks

Note: you can use h2 to represent h^2.

13c)
What are the predicted label $\hat{y}$ and the cross-entropy loss for this prediction? Is it a correct or incorrect prediction?

4
Marks

Note: you can use y_hat to represent $\hat{y}$.

Question 14

Consider a 4D tensor of a CNN with shape [32, 64, 64, 10]. We apply a Conv2D with 10 filters each of which has the shape [5,5,10] and strides= [3,3] with padding = 'valid' to obtain another tensor. On top of this tensor, we apply MaxPool2D with strides= (3,3), pool/kernel size = (2,2), and padding= 'same' to gain a 4D tensor. Finally, we flatten the last tensor to obtain a dense layer. What is the number of neurons on the dense layer? Show the steps of your answer.

5 Marks

Question 15

Assume that we have the following text corpus, vocabulary, and embedding matrix. We input to an RNN a mini-batch of two sentences as shown in the figure.

5 Marks

Movie review dataset

1. I really like this movie (**pos**:1).
2. This is a bad movie to watch (**neg**:0)
3. I really love this movie (**pos**: 1)
4. I do not recommend you to watch this movie (**neg**:0)
5. This movie is fantastic (**pos**:1)

Build up vocabulary

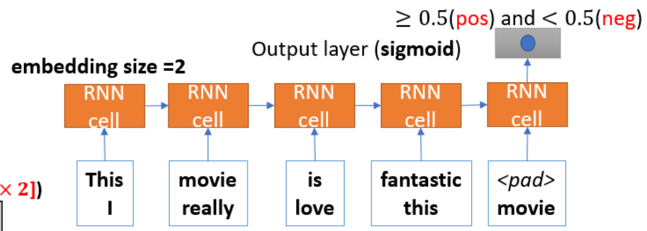
1. Like (index: 1)
2. Love (index: 2)
3. Bad (index: 3)
4. Fantastic (index: 4)
5. Not (index: 5)
6. Recommend (index: 6)

Not in vocabulary (out of vocabulary bucket: 2)

1. I, movie, this, watch (index: 7)
2. This, is, to, really, pad (index: 8)

Embedding matrix (U [8×2])

U_1	-1	2
U_2	1	-1
U_3	-1	-2
U_4	-1	3.5
U_5	1	-2.5
U_6	-1.5	-0.5
U_7	-4	2
U_8	1	-3



15a)

What are the numeric input items at each timestep really fed to the RNN?

2 Marks

15b)

What is the numeric value of the 3D tensor with the shape[batch_size, seq_len, embed_size] which is the output of the embedding layer?

3 Marks

<u>Note</u>: you can use NumPy format to represent a 3D array.

Question 16

Read the following code and provide the shapes of the tensors x, h1, h2, h3, h4, h5, h6.

5 Marks

```
embed_size = 64
vocab_size = 200
x = tf.keras.Input(shape=[5],dtype= 'int64')
h1 = tf.keras.layers.Embedding(vocab_size, embed_size)(x)
h2 = tf.keras.layers.GRU(16, return_sequences= True)(h1)
h3 = tf.keras.layers.GRU(8, return_sequences= True)(h2)
h4 = tf.keras.layers.GRU(16, return_sequences= True)(h3)
h5 = tf.keras.layers.Flatten()(h4)
h6 = tf.keras.layers.Dense(10, activation='softmax')(h5)
```

<u>Note</u>: the shapes should contain one dimension with the value None for batch size.

Question 17

Read the following code and provide the shape of the tensors x, h1, h2, h3, h4, h5, p.

5 Marks

```
x= Input((32,32,3))
h1 = Conv2D(filters=10, kernel_size= (3,3), strides=(1,1), padding= 'SAME')(x)
h2 = MaxPool2D(pool_size= (2,2), strides=(2,2), padding= 'VALID')(h1)
h3 = Conv2D(filters=20, kernel_size= (3,3), strides=(1,1), padding= 'VALID')(h2)
h4 = MaxPool2D(pool_size= (2,2), strides=(2,2), padding= 'SAME')(h3)
h5 = Flatten()(h4)
p = Dense(units=10, activation= 'softmax')(h5)
```

Note: the shapes must contain one dimension with the value None for batch size.

Question 18

Assume that we have [6,6,1] input tensor as shown below and applying max pooling or average pooling with kernel size = [2,2], strides = [2,2], padding= valid.

5 Marks

-3	1	4	2	-3	1
-2	1	-2	-3	3	2
-2	2	-5	2	-1	0
-3	3	6	-4	1	-2
2	-3	0	1	2	-1
-1	1	-1	1	-1	1

18a)

What is the output tensor if we apply max pooling?

Note: you can answer the output tensor by listing its rows (row1=[], row2=[], and so on)

2.5 Marks

18b)

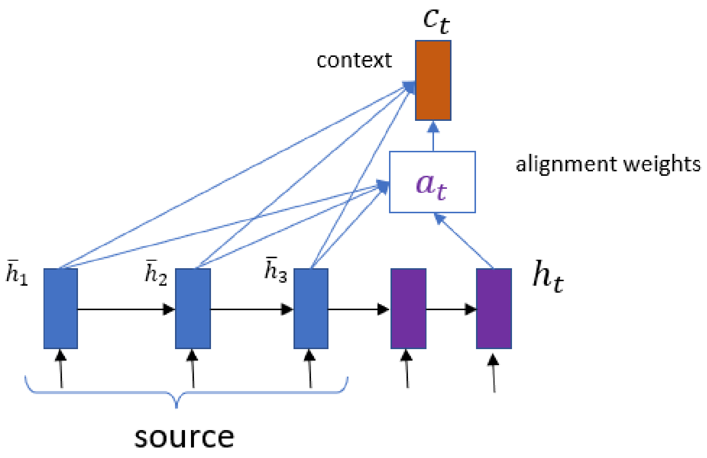
What is the output tensor if we apply average pooling?

2.5 Marks

Information

The statement below applies to Q.19-Q.20

Attention mechanism allows the decoding network to look back to the input sequence and is one of the most important techniques that has helped to achieve recent breakthroughs in NLP and seq2seq models. Consider the following encoding sequence of length 3 from the source and we are interested in using a global attention mechanism to decode the first target output. Answer the following questions:



Question 19

Assume that the encoder and decoder hidden states have the same dimension and we use the dot product to compute the alignment scores between them. Write down the analytical expressions to compute the alignment scores $score(h_t, \bar{h}_s)$ and alignment weights $a_t(s)$ for s=1,2,3.

2 Marks

Please answer the question on your blank piece of paper.

- After your exam finishes, you'll have **extra time** to access your phone to scan a **QR code** and **upload** your answer.
- Clearly label each page with **Student ID** and **this question number** (and **sub part** if applicable) (for example, 'Question 7a')
- **Do not** write your Name on it

No. of answer sheets: 1

Question 20

For simplicity, assuming the encoder hidden states are scalars with the following values $\bar{h}_1 = 1, \bar{h}_2 = -1, \bar{h}_3 = 2$, calculate the alignment weight vector a_t and then the context vector c_t with the decoder hidden state h_t=1. Note that we assume using the sign score function $score(h_t, \bar{h}_s) = sign(h_t \bar{h}_s)$ where sign(x) returns 1 if x≥0 and -1 if otherwise.

3 Marks

Please answer the question on your blank piece of paper.

- After your exam finishes, you'll have **extra time** to access your phone to scan a **QR code** and **upload** your answer.
- Clearly label each page with **Student ID** and **this question number** (and **sub part** if applicable) (for example, 'Question 7a')
- **Do not** write your Name on it

No. of answer sheets: 1

Part C - Mixed & Written-Answer Questions (10 questions, 35 marks)

Information

Part C contains 10 mixed and written-answer questions, worthing **35 marks** . These questions typically assess the knowledge and understanding of lecture contents.

Question 21

The following questions access your understanding of Word2Vec models. In your answer, to demonstrate your ideas, you can use the example sentence: "deep learning is really powerful". However, you can freely make your own example sentence.

5 Marks

21a)
What is the purpose of Word2Vec models?

1 Marks

21b)
Describe the pretext task of Skip-gram in training a Word2Vec model. What are the drawbacks of Skip-gram?

2 Marks

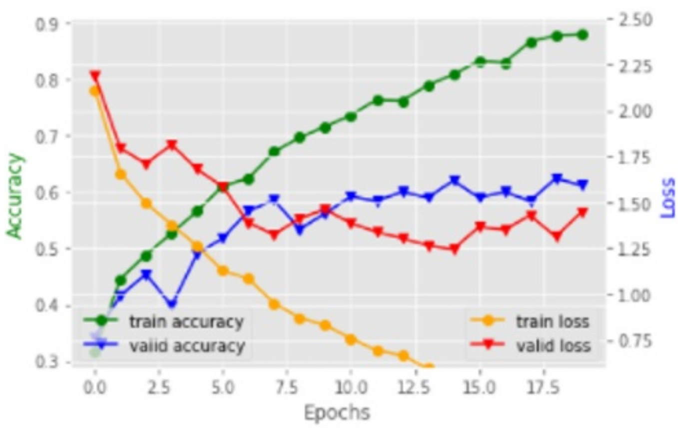
21c)
Describe Skip-gram with negative sampling.

2 Marks

Question 22

Considering training a deep learning model with the following plot. Describe the tendencies of the training loss, valid loss, and training accuracy, valid accuracy. Does the overfitting phenomenon happen? Explain your answer. When (i.e., at which epoch) should we do early stopping this training process?

5 Marks



Information

<u>**The statement below applies to Q.23-Q.24-Q.25**</u>

John is a research scientist who is doing a research project with a small-size image dataset. To enrich this dataset, John decides to use Generative Adversarial Network (GAN) to automatically produce novel high-quality fake images from noises $z \sim P_z = N(0,I)$. Assume that for his GAN, John uses a discriminator D and a generator G.

Question 23

Describe the roles of D and G in the min-max game of GAN.

3
Marks

Question 24

What are the optimization problems to train D and G?

3
Marks

Please answer the question on your blank piece of paper.

- After your exam finishes, you'll have **extra time** to access your phone to scan a **QR code** and **upload** your answer.
- Clearly label each page with **Student ID** and **this question number** (and **sub part** if applicable) (for example, 'Question 7a')
- **Do not** write your Name on it

No. of answer sheets: 1

Question 25

With an appropriate setting, John can train to reach the optimal D^* and G^* for which John observes that $D^*(x)=0.5$ for real and fake images x. Explain why it happens.

4
Marks

Please answer the question on your blank piece of paper.

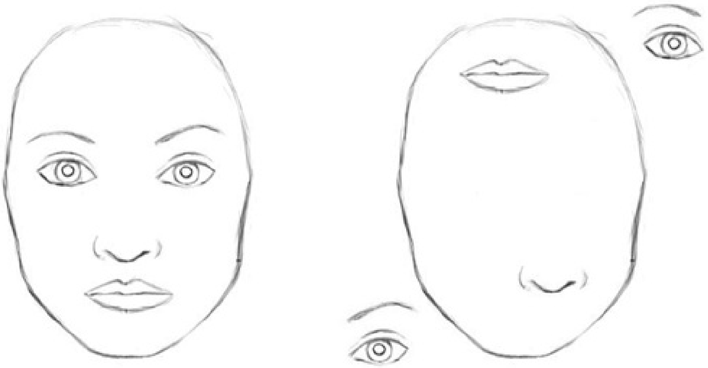
- After your exam finishes, you'll have **extra time** to access your phone to scan a **QR code** and **upload** your answer.
- Clearly label each page with **Student ID** and **this question number** (and **sub part** if applicable) (for example, 'Question 7a')
- **Do not** write your Name on it

No. of answer sheets: 1

Question 26

One potential problem with using CNNs is that they cannot realize the relative spatial relationship among objects in an image. Let us consider two images as below. Because CNNs have the power to learn the objects of eyes, nose, and mouth, but not spatial relationships among them, they could incorrectly classify the right image as a human face. Discuss why this is the case for CNNs.

5 Marks



Question 27

What is underfitting? In the context of deep learning, give an example of a scenario when underfitting can happen?

4 Marks

Information

Q.28-Q.29-Q.30 are relevant to the overfitting phenomenon.

Question 28

What is overfitting in machine learning?

2 Marks

Question 29

In the context of deep learning, give an example of a scenario when overfitting can happen?

2 Marks

Question 30

List TWO solutions to combat overfitting problems in deep learning and briefly explain why they can help to combat overfitting.

2 Marks