COMP 5214 & ELEC 5680 **Name:** ____________________
Spring 2021
Midterm
8 April 2021
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. Select the method(s) that DOES NOT prevent a model from overfitting to the training set?

      (a) Early stopping
      (b) Dropout
      (c) Data augmentation
      (d) Pooling

      Answer: D

   2. How should we tune the hyper-parameters of a machine learning model?

      (a) Tune hyper-parameters on the training set
      (b) Tune hyper-parameters on the validation set
      (c) Tune hyper-parameters on the test set
      (d) Tune hyper-parameters on a single example

      Answer: B

   3. Which of the following is/are permutation invariant or equivariant? A function f is permutation invariant if $f(\pi x) = f(x)$ for any permutation π. A function f is permutation equivariant if $f(\pi x) = \pi f(x)$ for any permutation π.

      (a) PointNet
      (b) Global pooling
      (c) Attention
      (d) Graph convolution layer

      Answer: A B C D

   4. Which of the following is/are true about RAW sensor data and RGB images?

      (a) An image signal processing pipeline can take RAW sensor data as input and generate an RGB image.
      (b) RAW sensor data has 3 channels of RGB colors at each pixel
      (c) RAW sensor data often has more than 8 bits per pixel, such as 12-14 bits
      (d) RAW sensor data has a linear relationship with light intensity while an RGB image does not

      Answer: A C D

   5. What can we do to train a deep neural network when we have limited GPU memory (without considering the influence on accuracy)?

      (a) Increase the learning rate
      (b) Reduce the batch size
      (c) Add batch normalization
      (d) Train on smaller images with lower resolutions

Answer: B D

6. Which of the following is true about deep 3D learning?
   (a) PointNet has universal approximation ability
   (b) Octree can be used to increase the resolution in deep 3D models
   (c) Spatial transform networks can learn scaling, rotation, and translation
   (d) It is impossible to generate a point cloud with a neural network

   Answer: A B C

7. Which of the following statements on backpropagation is/are true?
   (a) Pooling layers are not differentiable
   (b) The gradients will become smaller in magnitude flowing through tanh layers
   (c) The gradients will be unchanged flowing through ReLU layers
   (d) When performing SGD, the loss will always drop after each update

   Answer: B

8. Which of the following operation can increase the receptive field of a CNN?
   (a) Add residual connections
   (b) Use smaller strides
   (c) Increase dilation rates
   (d) Add more pooling layers

   Answer: C D

9. What can be the input for the problem of novel view synthesis?
   (a) An image
   (b) A video
   (c) Stereo images
   (d) Multi-view images

   Answer: A B C D

10. Which of the following loss function(s) is/are per-pixel loss, where the loss can computed independently on each pixel.
    (a) MSE
    (b) PSNR
    (c) Content loss in style transfer
    (d) Style loss in style transfer

    Answer: A B

11. Which of the following is FALSE about the one-hot node feature for graph neural networks?
    (a) It is usually used on graphs with inductive settings
    (b) It is computationally expensive when the number of node is large

(c) It is easy to compute

(d) It does not generalize well to new nodes

Answer: A

12. Which one of the following is not an element attention operate on?

(a) Queries

(b) Masks

(c) Keys

(d) Values

Answer: B

13. Which one of the following is an internal learning method? An internal method does not require a training set, but trains a model on the test example.

(a) NeRF (Lecture 12)

(b) SynSin (Lecture 12)

(c) Deep Image Prior (Lecture 10)

(d) Perceptual loss for real-time style transfer and super-resolution, ECCV 2016 (Lecture 8)

Answer: A C

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

   1. In a simple MLP model with 8 neurons in the input layer, 5 neurons in the hidden layer and 1 neuron in the output layer. Bias is used in each layer. How many parameters are in this MLP.

      The total number of parameters is

      $$(8 \times 5 + 5) + (5 \times 1 + 1) = 51$$

   2. Explain how attention can solve one flaw of encoder-decoder architectures for machine translation.

      An encoder-decoder architecture without attention needs to store all the necessary information about the input sentence in its code vector (final hidden state). An attentionbased architecture can refer to the annotation vectors for particular words, reducing the memory pressure on its hidden state.

   3. Show a problem in which Pix2pix (conditional GANs) can not solve but CycleGAN can solve.

      As long as the problem is an unpair image-to-image translation. For example, style transfer based on a set of style images.

   4. TRUE or FALSE. The a deep neural network model with cross-entropy loss for image classification has a training loss of zero if every training image is classified correctly? Explain why.

      False.

   5. You are given a content image X_C and a style image, X_S. You would like to apply neural style transfer to obtain an output image Y , with the content of X_C and the style of X_S, as discussed in section. You are told that you need a pretrained VGG-16 network to do this. Explain mathematically how do we compute the content loss and the style loss between X_C and X_S.

      Here is the content loss:

      $$l_{\text{content}}(\hat{y}, y) = \sum_{i \in I_\phi} \frac{1}{N_i(\phi)} \|\phi_i(\hat{y}) - \phi_i(y)\|_2^2,$$

      For style loss, we can use the Frobenius norm of differences of the Gram matrices of $\hat{y}$ and y_s:

      $$l_{\text{style}}(\hat{y}, y_s) = \sum_{i \in I_\phi} \frac{1}{N_i(\phi)} \|G_i^\phi(\hat{y}) - G_i^\phi(y_s)\|_F^2.$$

   6. Explain why the activation function needs to be an non-linear function in a deep neural network.

      If the activation function is also nonlinear, then the whole model is also linear.

7. TRUE or FALSE. Two graphs which contain the same number of graph vertices connected in the same way are said to be isomorphic. Consider two graphs G_1 and G_2, where G_1 and G_2 are isomorphic and the initial feature at each node in both graphs is the SAME. Let f be a graph convolutional neural network with random weights:

$$h_v^{(l)} = \sigma\left(W^{(l)} \sum_{u \in N(v)} \frac{h_u^{(l-1)}}{|N(v)|}\right).$$

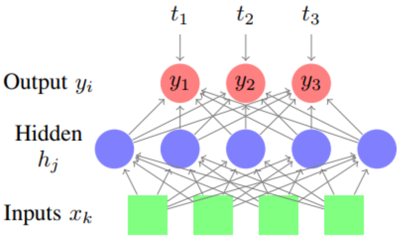
Then the corresponding nodes in $f(G_1)$ and $f(G_2)$ have the same output embedding.

True. Graph convolution layers are permutation equivariant.

3. (26 points) **Backpropagation and Cross Entropy Error**

In a classification task with two classes, it is standard to use a neural network architecture with a single logistic output unit and the cross-entropy loss function (as opposed to, for example, the sum-of-squared loss function). With this combination, the output prediction is always between zero and one, and is interpreted as a probability. Training corresponds to maximizing the conditional log-likelihood of the data, and as we will see, the gradient calculation simplifies nicely with this combination.

We can generalize this slightly to the case where we have multiple, independent, two-class classification tasks. In order to demonstrate the calculations involved in backpropagation, we consider a network with a single hidden layer of logistic units, multiple logistic output units, and where the total error for a given example is simply the cross-entropy error summed over the output units.



The cross entropy error for a single example with *nout* independent targets is given by the sum

$$E = -\sum_{i=1}^{nout}(t_i \log(y_i) + (1-t_i)\log(1-y_i)),$$

where $\mathbf{t}$ is the target vector, $\mathbf{y}$ is the output vector. In this architecture, the outputs are computed by applying the logistic function to the weighted sums of the hidden layer activations, $\mathbf{s}$,

$$\begin{aligned} y_i &= \frac{1}{1+e^{-s_i}} \\ s_i &= \sum_{j=1} h_j w_{ji}. \end{aligned}$$

We can compute the derivative of the error with respect to each weight connecting the hidden units to the output units using the chain rule.

$$\frac{\partial E}{\partial w_{ji}} = \frac{\partial E}{\partial y_i}\frac{\partial y_i}{\partial s_i}\frac{\partial s_i}{\partial w_{ji}}$$

(a) [4 points] What is $\frac{\partial E}{\partial y_i}$?

$$\begin{aligned}\frac{\partial E}{\partial y_i} &= \frac{-t_i}{y_i} + \frac{1-t_i}{1-y_i} \\ &= \frac{y_i - t_i}{y_i(1-y_i)}\end{aligned}$$

(b) [4 points] What is $\frac{\partial y_i}{\partial s_i}$?

$$\frac{\partial y_i}{\partial s_i} = y_i(1-y_i)$$

(c) [4 points] What is $\frac{\partial s_i}{\partial w_{ji}}$?

$$\frac{\partial s_i}{\partial w_{ji}} = h_j$$

(d) [4 points] What is $\frac{\partial E}{\partial w_{ji}}$?

$$\frac{\partial E}{\partial w_{ji}} = (y_i - t_i)h_j$$

The above gives us the gradients of the error with respect to the weights in the last layer of the network, but computing the gradients with respect to the weights in lower layers of the network (i.e. connecting the inputs to the hidden layer units) requires another application of the chain rule. This is the backpropagation algorithm.

Here it is useful to calculate the quantity $\frac{\partial E}{\partial s_j^1}$ where j indexes the hidden units, s_j^1 is the weighted **input sum** at hidden unit j, and $h_j = \frac{1}{1+e^{-s_j^1}}$ is the activated at unit j.

(e) [5 points] What is $\frac{\partial E}{\partial s_j^1}$?

$$\begin{aligned}\frac{\partial E}{\partial s_j^1} &= \sum_{i=1}^{nout} \frac{\partial E}{\partial s_i}\frac{\partial s_i}{\partial h_j}\frac{\partial h_j}{\partial s_j^i} \\ &= \sum_{i=1}^{nout}(y_i - t_i)(w_{ji})(h_j(1-h_j))\end{aligned}$$

(f) [5 points] Then, what is $\frac{\partial E}{\partial w_{kj}^1}$, the gradient with respect to a weight w_{kj}^1 connecting input unit k to hidden unit j?

$$\begin{aligned}\frac{\partial E}{\partial w_{kj}^1} &= \frac{\partial E}{\partial s_j^1}\frac{\partial s_j^1}{\partial w_{kj}^1} \\ &= \sum_{i=1}^{nout}(y_i - t_i)(w_{ji})(h_j(1-h_j))(x_k)\end{aligned}$$

By recursively computing the gradient of the error with respect to the activity of each neuron, we can compute the gradients for all weights in a network.