Name:

Student ID:

Q1: A convolutional layer in a CNN has 32 filters, each with a kernel size of 3 × 3, applied to an input with 16 channels. Assuming each filter has a bias term, calculate the total number of trainable parameters in this layer.

Q2: A fully connected (linear) layer in a neural network has 128 input neurons and 64 output neurons. Each neuron in the output layer has a bias term.

Q3: What is the purpose of batch normalization

Q4: What does top-1 and top-5 error rates mean? Please compare top-17.0% and top-5 37.5% and top-1 37.5% and top-5 17.0%? Which one is the correct order?

Q5: Please compare ReLU and Tanh activation. Elaborate as much as possible

Q6: True or False: In a Convolutional Neural Network (CNN), the deeper layers (further from input) are responsible for extracting fundamental visual patterns such as edges, textures, and simple geometric shapes. These layers work at a granular level, ensuring that basic visual structures are well understood before passing the information to higher layers. On the other hand, the shallower layers (closer to input) take this information and transform it into more complex, abstract representations, allowing the network to recognize high-level concepts like object categories and scene understanding. This hierarchical structure ensures that the network can gradually build up to meaningful representations from the simplest building blocks.

Q7: Why do we use dropout?

Q8: In inceptionNet: why such large percentage of dropout is used

Q9: 1x1 CNN is widely used in deep learning architectures like GoogleNet (Inception) and ResNet to reduce computational cost and enhance efficiency. Please explain with example how this is the case

Q10:A Vision Transformer (ViT) takes an input image of size 224 × 224 with 3 channels (RGB) and splits it into fixed-size patches before processing them as tokens in a transformer model.

(a) If the patch size is 16 × 16, how many patches (tokens) will the ViT generate?

(b) What is the dimensionality of each patch (token) before flattening?

(c) After flattening each patch into a token, what is the final shape of the input sequence fed into the Transformer? *(Assume no additional positional embedding for now.*

Mark Scheme

Question	Answer - Total = 25 points
Q1: A convolutional layer in a CNN has 32 filters, each with a kernel size of 3 × 3, applied to an input with 16 channels. Assuming each filter has a bias term, calculate the total number of trainable parameters in this layer.	**Total - 3 points** Option1: Correct answer 3 points without showing your working Option2: Correct formula 1 point Correct substitution 1 point Correct answer 1 point Option3: Almost correct answer, the only thing wrong is your bias term 2 points Total Parameters = (KW × KH ×IC + BT) × #filters = (3 × 3 × 16 + 1) × 32 = 4640 KW = Kernel Width KH = Kernel Height IC = Input Channels BT = bias term #Fitler = number of filters
Q2: A fully connected (linear) layer in a neural network has 128 input neurons and 64 output neurons. Each neuron in the output layer has a bias term.	**Total - 3 points** Option1: Correct answer 3 points without showing your working Option2: Correct formula 1 point Correct substitution 1 point Correct answer 1 point Option3: Almost correct answer, the only thing wrong is your bias term 2 points Total Parameters = (IN x ON) + BT = 128 x 64 + 64 = 8256 IN = Input Neurons ON = Output Neurons

	BT = Bias terms
Q3: What is the purpose of batch normalization	**Total - 1 point** ● Prevent exploding gradient ● Facilitate faster convergence ● Regularization *Open to any other sound answers
Q4: What does top-1 and top-5 error rates mean? Please compare top-1 17.0% and top-5 37.5% and top-1 37.5% and top-5 17.0%? Which one is the correct order?	**Total - 3 points** Correct meaning - 2 point Correct order - 1 point Top-1 Error Rate: The percentage of times the model's most confident prediction is incorrect. Top-5 Error Rate: The percentage of times the correct label is not within the model's top 5 predictions. The correct relationship is Top-5 error rate ≤ Top-1 error rate
Q5: Please compare ReLU and Tanh activation. Elaborate as much as possible	**Total - 2 points** *Open to any other sound comparison and elaboration
Q6: True or False: In a Convolutional Neural Network (CNN), the deeper layers (further from input) are responsible for extracting fundamental visual patterns such as edges, textures, and simple geometric shapes. These layers work at a granular level, ensuring that basic visual structures are well understood before passing the information to higher layers. On the other hand, the shallower layers (closer to input) take this information and transform it into more complex, abstract representations, allowing the network to recognize high-level concepts like object categories and scene understanding. This hierarchical structure	**Total - 1 point** Correct answer - 2 points False

ensures that the network can gradually build up to meaningful representations from the simplest building blocks.	
Q7: Why do we use dropout?	**Total - 1 point** Correct answer - 1 point ● Prevent over relying on certain features ● Make other neurons learn other features *Open to any other sound reasons
Q8: In inceptionNet: why such large percentage of dropout is used	**Total - 1 point** ● InceptionNet is super large, we want to apply dropout so that all parts of the network is used *Open to any other discussion
Q9: 1x1 CNN is widely used in deep learning architectures like GoogleNet (Inception) and ResNet to reduce computational cost and enhance efficiency. Please explain with example how this is the case	**Total - 3 points** Correct explanation 1 point Example 2 points 1x1 reduces cost Example: Case 1 3x3 CNN H = 3, W = 3, IC = 128, OC = 128 Total = 3x3x64x128= 73,000 Case2 1x1 followed by 3x3 H = 1, W = 1, IC = 64, OC = 16 Total(1x1) = 1x1x64x16 = 1024 H = 3, W = 3, IC = 16, OC = 128 Total(3x3) = 3x3x16x64 = 18432 Therefore, #params(case2) < #params(case1) *Open to any other sound explanations
Q10:A Vision Transformer (ViT) takes an input image of size 224 × 224 with 3 channels	**(a) Total - 3 points** Option1: Correct answer 3 points without

(RGB) and splits it into fixed-size patches before processing them as tokens in a transformer model. (a) If the patch size is 16 × 16, how many patches (tokens) will the ViT generate? (b) What is the dimensionality of each patch (token) before flattening? (c) After flattening each patch into a token, what is the final shape of the input sequence fed into the Transformer? *(Assume no additional positional embedding for now.*)	showing your working Option2: Correct width 1 point Correct height 1 point Correct answer 1 point Total = Image size/ patch size = (224 x 224) / (16x 16) H = 224/16 W = 224/16 Total = 196 **(b) Total - 3 points** Option1: Correct answer 3 points without showing your working Option2: Correct patch size 1 point Correct channel 1 point Correct answer 1 point Dim = patch size * channel = 16 x 16 x 3 = 768 **(c) Total - 1 point** Correct answer - 1 point Final shape: (196,768)