National University of Computer and Emerging Sciences

Islamabad Campus

Computer Vision (AI4002)

Final Examination

Course Instructor(s):
Ms. Khadija Mahmood
Section(s): (if applicable)
Part:A

Total Time (Hrs): 1
Total Marks: 50
Total Questions: 1

Date: Dec 18, 2024

_____ _____ _____

Roll No **Course Section** **Student Signature**

Do not write below this line.

Attempt all the questions.

[CLO:3 & 5. Apply appropriate image processing methods for image filtering, image restoration, image reconstruction, segmentation, classification and representation]

Q1: MCQ's, Fill the bubble sheet provided at last page. **[50 marks]**

1. **What is the primary purpose of an activation function in a neural network?**

A) To introduce non-linearity to the model
B) To prevent overfitting
C) To calculate gradients during backpropagation
D) To initialize the weight

2. **In a Convolutional Neural Network (CNN), which layer is responsible for reducing the spatial dimensions of the input?**

A) Convolution Layer
B) Pooling Layer
C) Fully Connected Layer
D) ReLU Layer

3. **Which of the following optimizers is known for combining the benefits of both Adagrad and RMSprop?**

A) SGD (Stochastic Gradient Descent)
B) Adam
C) Adadelta
D) RMSprop

4. **What is the main advantage of using the batch normalization technique in deep neural networks?**

A) It speeds up the training process by normalizing the inputs to each layer
B) It reduces the need for regularization
C) It helps in handling missing values
D) It prevents vanishing gradients

5. An autoencoder consists of an encoder and a decoder. What is the primary purpose of the decoder in an autoencoder?

A) To encode the input into a compressed form.
B) To reconstruct the original input from the compressed encoding.
C) To perform classification on the encoded data.
D) To perform unsupervised learning on the data.

6. When applying transfer learning, what aspect of the pre-trained model is most crucial for generalization to a new task?

A) The pre-trained model should have been trained on the same dataset as the new task.
B) The pre-trained model should have learned general features that are applicable to many tasks.
C) The pre-trained model should have been trained with the same architecture as the new model.
D) The pre-trained model should not include any dropout layers.

7. What happens to the key points detected in low-contrast regions of an image in the SIFT algorithm?

A) They are always discarded.
B) They are detected but with lower confidence.
C) They are detected but assigned lower importance in the matching process.
D) They are detected and classified as high-confidence key points.

8. Which of the following transformations is SIFT invariant to?

A) Affine transformations, such as scaling, rotation, and translation.
B) Only affine transformations involving scaling.
C) Only affine transformations involving rotation.
D) SIFT is not invariant to affine transformations.

9. How is the descriptor for each key point in the SIFT algorithm created?

A) By calculating the pixel intensity of the key point.
B) By computing a histogram of image gradients around the key point and normalizing it.
C) By measuring the distance between the key point and the center of the image.
D) By using a pre-trained neural network to classify the key point.

10. In the context of CNNs, what does the term "weight sharing" mean, and why is it used?

A) It refers to using identical filters across different layers to reduce overfitting
B) It refers to using the same filter across different spatial locations of the input to reduce the number of parameters
C) It refers to applying the same weights for different classes during classification
D) It refers to applying shared weights for pooling layers to make the network more efficient.

11. In CNNs, how does the choice of kernel size (e.g., 3x3 vs 5x5) affect the receptive field of the network?

A) Larger kernels increase the receptive field and capture more contextual information with fewer layers
B) Larger kernels reduce the receptive field and cause the network to focus on more localized features
C) Smaller kernels always reduce the complexity of the network and lead to better generalization
D) Kernel size does not affect the receptive field; only the number of layers matters

12. In a typical CNN, which of the following scenarios is most likely to cause overfitting in a network?

A) Increasing the number of convolutional layers while keeping the number of parameters fixed
B) Using too much data augmentation without regularization
C) Using an excessively small learning rate
D) Using a very large fully connected layer after the convolutional layers

13. In a CNN, what is the effect of using a smaller stride in the convolution operation?

A) It leads to an increase in the output feature map size and requires more computations
B) It decreases the receptive field of the network
C) It prevents the network from learning high-level features
D) It reduces the depth of the output feature maps

14. What additional feature does Mask R-CNN provide compared to Faster R-CNN?

A) It performs image segmentation in addition to object detection.
B) It performs keypoint detection for human poses.
C) It only detects objects in grayscale images.
D) It uses a sliding window for localization.

15. In Faster R-CNN, what role does the Region Proposal Network (RPN) play in object detection?

A) It generates fixed-size bounding boxes for every object in the image.
B) It generates region proposals that likely contain objects, which are then used by the classifier.
C) It directly classifies each pixel in the image as belonging to an object.
D) It segments the image into smaller regions based on color and texture.

16. How do Feature Pyramid Networks (FPN) improve object detection performance, especially for small objects?

A) By applying multi-scale feature maps to handle objects at various scales more effectively.
B) By using a single convolutional layer to detect objects at different scales.
C) By focusing only on the most prominent objects in the image.
D) By introducing an additional attention mechanism to refine object localization.

17. In the context of object detection, what is the role of the Intersection Over Union (IoU) threshold in non-maximum suppression (NMS)?

A) The IoU threshold determines the minimum overlap required to consider two bounding boxes as separate detections.
B) The IoU threshold determines the maximum amount of overlap allowed between different bounding boxes before suppression occurs.
C) The IoU threshold specifies the degree of classification confidence needed to retain a bounding box.
D) The IoU threshold filters out low-confidence bounding boxes to increase detection accuracy.

18. In YOLO (You Only Look Once), the image is divided into a grid of cells, and each cell predicts bounding boxes for objects within its region. What is the impact of the grid system on the localization accuracy of YOLO?

A) The grid system allows YOLO to predict objects with high accuracy regardless of their size and location.
B) The grid system improves YOLO's performance by ensuring that each cell predicts only one bounding box per object.
C) The grid system restricts the number of objects a cell can detect, which limits the model's ability to localize small objects in the image.
D) The grid system improves localization accuracy by providing precise bounding boxes for all objects, regardless of size.

19. You are tasked with developing an object detection system for an autonomous mobile robot that will operate in a real-time, resource-constrained environment. The robot has limited computational power, memory, and battery life. You have the choice between several object detection architectures: GoogleNet, VGG, ResNet, MobileNet, and InceptionNet.

Which of the following models would be most suitable for this task, considering the need for a balance between detection accuracy and computational efficiency?

A) MobileNet - MobileNet is designed specifically for mobile and embedded vision applications. It uses depth-wise separable convolutions to reduce the number of parameters and computations, making it well-suited for real-time applications on resource-constrained devices.
B) VGG - VGG has a simple architecture but uses a large number of parameters, making it computationally expensive and not ideal for mobile or resource-limited environments.
C) InceptionNet - InceptionNet is powerful and accurate but involves complex multi-scale convolutions, which may lead to a higher computational load and is not ideal for real-time performance on mobile devices.
D) ResNet - ResNet, with its deep residual learning approach, is great for handling very deep networks, but its high computational cost makes it less suitable for mobile deployment.

20. **What is the primary advantage of using Inception modules in InceptionNet?**

A) To reduce the number of parameters by using depth-wise convolutions
B) To extract multi-scale features by combining convolutions of different kernel sizes in parallel
C) To increase model depth using residual connections
D) To use a smaller kernel size to improve computation efficiency

21. **Which of the following is a key feature of MobileNet that makes it suitable for real-time object detection on mobile devices?**

A) Use of large convolution kernels to capture detailed features
B) Depth-wise separable convolutions to reduce the number of parameters
C) Residual connections that improve training stability
D) Use of multi-scale feature extraction for better object localization

22. **What is the primary characteristic of VGG architecture that distinguishes it from other CNN architectures?**

A) Use of small 3x3 convolutional filters and deep stacking of layers
B) Use of large 7x7 convolution filters with fewer layers
C) Use of depth-wise separable convolutions to reduce computation
D) Use of global average pooling instead of fully connected layers

23. **In GoogleNet (InceptionNet), how is computational efficiency achieved while maintaining high accuracy?**

A) By using depth-wise separable convolutions
B) By using 1x1 convolutions to reduce dimensionality before applying larger convolutions
C) By combining convolutional and fully connected layers
D) By using residual connections in each inception module

24. Consider the following architecture for MobileNet: The number of input channels is 256, the kernel size is 3x3, and depth-wise separable convolutions are used. If the number of output channels is 512, what is the total number of parameters in this convolutional layer? If the value is not in option mark the nearest value to your answer.

A) 1,440
B) 4,608
C) 9,216
D) 16,384

25. In an InceptionNet model, if the input image size is 299x299, and the first convolution layer uses a 7x7 kernel with a stride of 2, what will be the output size after this layer? If the value is not in option mark the nearest value to your answer.

A)149x149
B) 148x148
C) 150x150
D) 151x151

26. **What is the primary advantage of using Inception modules in InceptionNet?**

A) To reduce the number of parameters by using depth-wise convolutions
B) To extract multi-scale features by combining convolutions of different kernel sizes in parallel
C) To increase model depth using residual connections
D) To use a smaller kernel size to improve computation efficiency

27. You are using an object detection system to detect small objects (e.g., screws, tools, or components) in a factory floor with cluttered, complex scenes. The detection system needs to maintain accuracy in highly varied conditions, including varying object scales and occlusions.

Which of the following architectures would be most effective for detecting small objects in cluttered scenes?

A) GoogleNet - GoogleNet uses an inception module that extracts features at multiple scales, making it effective at detecting small objects within complex scenes. Its multi-scale convolutional layers can handle different object sizes, making it a good choice for this scenario.
B) VGG - While VGG can extract strong feature representations, its large convolutional layers may not be as effective for detecting small objects at different scales.
C) ResNet - ResNet is highly effective for deep feature learning but might not perform as well on small objects in complex scenes unless specifically adapted.
D) MobileNet - MobileNet, while efficient, may struggle with small object detection due to its trade-off between accuracy and efficiency, which can affect its performance in complex environments.

28. **What is the primary purpose of the "skip connections" in the U-Net architecture?**

A) To reduce the number of parameters by eliminating unnecessary layers
B) To help the model focus on high-level features and ignore low-level details
C) To preserve spatial information and allow the decoder to reconstruct finer details
D) To enable deeper networks by adding more layers between the encoder and decoder

29. **Which of the following operations is typically used in the encoder part of the U-Net architecture?**

A) Transposed convolution
B) Max pooling
C) Up-sampling
D) Dropout

30. In a U-Net architecture, if the number of filters in each convolutional layer is doubled after each down-sampling step (i.e., 64 filters in the first layer, 128 in the second, etc.), What will be

the depth of filter will be used in the 4th layer? If the value is not in option mark the nearest value to your answer.

A) 256 filters
B) 128 filters
C) 512 filters
D) 64 filters

31. Why is the dot product in the attention mechanism scaled by the square root of the key dimension?

A) To ensure stability of gradients during training
B) To reduce the number of attention heads required
C) To prevent overly large dot product values for high-dimensional keys
D) To normalize the weights across all layers

32. The sequence length of a first invented Transformer is fixed at 100 and the embedding size is 768, how many trainable parameters are introduced by the positional encoding layer?

A) 76,800
B) 768
C) 100
D) 0

33. In a Transformer decoder processing a sequence of length 50, how many elements are masked in the attention matrix to prevent information leakage from future tokens?

A) 1,225
B) 1,250
C) 1,275
D) 1,300

34. What is the main purpose of the class token (CLS token) used in the Vision Transformer (ViT)?

A) It serves as an auxiliary token to assist with regularization during training
B) It collects the global information from all patches to make the final classification decision
C) It reduces the size of the input feature maps
D) It is used for data augmentation purposes

35. What is the effect of increasing the number of attention heads in a transformer model for computer vision?

A) It increases the amount of parallelization in the computation, improving performance
B) It reduces the model's capacity to capture long-range dependencies

C) It results in less accurate results by overfitting the model
D) It reduces the size of the input feature maps

36. In Vision Transformers, what is the primary disadvantage of using large patch sizes?

A) It increases the amount of computation required in each self-attention layer
B) It leads to a loss of spatial information and finer-grained details
C) It results in better feature extraction from the image
D) It increases the number of parameters needed to train the model

37. If the embedding dimension of each patch in a Vision Transformer (ViT) is 768 and the image is divided into 196 patches, what is the total number of input tokens for the transformer?

A) 768
B) 196
C) 76752
D) 1536

38. In a Vision Transformer, if each patch is represented by a 768-dimensional embedding, and there are 196 patches, what is the total number of parameters required for a fully connected classification layer with 1000 output classes? If the value is not in options, mark the nearest value to your answer.

A) 768,000
B) 196,000
C) 768,000,000
D) 768,000,000,000

39. If the patch size is 32x32 and the input image is 256x256x3, how many patches will be created after splitting the image? If the value is not in options, mark the nearest value to your answer.

A) 64
B) 1024
C) 256
D) 2048

40. In the Swin Transformer, what is the role of "patch merging"?

A) To reduce the dimensionality of each patch for more efficient processing
B) To combine multiple patches into a single patch to reduce computational cost
C) To combine features from different layers of the network into a final prediction
D) To combine local patches into larger patches, increasing the receptive field

41. In the Swin Transformer, the window-based self-attention is applied within non-overlapping local windows. How does this affect the flow of information between distant patches in the image?

A) Information flows only within the individual window, and no global context is captured in the original window.
B) Information flows across all patches at each layer, eliminating the need for window-based attention.
C) Information flows through the entire image, ensuring that all patches interact globally, without any local restrictions.
D) Information is captured only locally within a window at the first few layers but gradually merges with global information at higher layers.

42. How does the Swin Transformer balance the need for global context and computational efficiency while using window-based attention?

A) By using local windows only at the first layer and switching to global attention throughout the model.
B) By applying attention to all patches globally in early layers and locally in later layers.
C) By dividing the image into local windows, applying attention locally, and then progressively merging patches and shifting windows to capture global context at higher layers.
D) By increasing the window size to capture global information at every layer.

43. View coordinates are different from canonical coordinates because:

A) View coordinates are always fixed regardless of the camera's position
B) View coordinates depend on the position and orientation of the camera relative to the object
C) View coordinates are not used in 3D computer vision tasks
D) View coordinates are used to represent textures on 3D models

44. In a scenario where you are training the Pixel2Mesh model on 3D human faces, the model learns to deform the initial mesh from the encoder to better match the target face shape. What is the role of the deformation module in this process?

A) It ensures that the mesh has the correct color by altering vertex colors
B) It refines the mesh by adjusting vertex positions to reduce the reconstruction error
C) It reduces the complexity of the mesh by eliminating unnecessary vertices
D) It increases the resolution of the texture map used in the mesh generation

45. During training of the Pixel2Mesh model, a loss function is used to guide the model in improving the 3D mesh. This loss function typically compares the predicted mesh to a ground truth mesh. Which of the following types of loss is most likely to be used?

A) Cross-entropy loss
B) Chamfer distance loss
C) L2 loss on image pixels
D) Binary classification loss

46. Voxel grids, while useful in 3D modeling, can be computationally expensive. Which of the following is a key reason for this?

A) They require large amounts of computational power for ray tracing
B) They store information as discrete volumes, leading to high memory usage
C) They simplify object shapes, which makes them less computationally expensive
D) They do not support the processing of large datasets

47. When using voxel grids for 3D representation, which of the following problems is commonly encountered due to the grid-based structure?

A) Inability to perform 3D transformations
B) Limited ability to represent fine details in objects, especially for high-resolution models
C) Difficulty in integrating textures with 3D models
D) Difficulty in converting voxel grids to mesh representations

48. Mesh R-CNN uses a technique for 3D object detection that involves both 2D and 3D information. Which of the following is the primary benefit of this approach?

A) It increases the speed of generating 3D meshes by skipping the 2D feature extraction
B) It enhances the quality of the texture mapping applied to 3D models
C) It leverages 2D object detection to propose potential 3D object locations, improving detection accuracy
D) It simplifies the generation of depth maps by using only 2D features

49. In Mesh R-CNN, after detecting the region of interest (RoI) using the RPN, what happens next?

A) The 3D shape of the object is generated using voxel-based methods
B) The detected 3D object is segmented and refined using the Mesh Decoder module
C) The image is re-projected onto a 3D plane to generate depth information
D) A texture map is applied to the 3D mesh

50. You are using Pixel2Mesh to generate 3D faces from 2D images, but the reconstructed 3D faces appear too smooth and lack fine facial details. Which part of the model would you likely modify to improve the detail level?

A) The image encoder network, to provide more detailed image features
B) The deformation module, to allow for more iterations of mesh refinement
C) The graph convolution layers, to allow for more vertex interactions
D) The texture mapping module, to add more fine-grained texture details

Instruction for filling the sheet 1. This sheet should not be folded or crushed 2. Use only blue/black ball pen or 2HB pencil 3. Circle should darkened completely and properly 4. Erase marked circle completely for deselect	
NAME :	
EXAM :	DATE :