ASSIGNMENT - 12

1. Describe the Quick R-CNN architecture.

Ans: Quick R-CNN (Regions with CNN features) builds upon R-CNN by sharing convolutional features across all proposals. Here's the breakdown:

* Input: An image and a set of region proposals (bounding boxes likely to contain objects).
* Feature Extraction: A pre-trained convolutional neural network (CNN) like VGG-16 extracts features from the entire image. Only the convolutional layers are used, discarding fully connected layers.
* Region of Interest (ROI) Pooling: Each proposal is warped into a fixed size, preserving spatial information.
* Classification and Bounding Box Regression: Fully connected layers predict class probabilities and bounding box offsets for each proposal.

2. Describe two Fast R-CNN loss functions.

Ans: Fast R-CNN addresses R-CNN's slow training by introducing two loss functions:

* Classification Loss: This is typically softmax loss, measuring the difference between predicted class probabilities and ground truth labels for each proposal.
* Bounding Box Regression Loss: This is smooth L1 loss, which is less sensitive to outliers compared to L2 loss. It penalizes the model for inaccurate bounding box offsets predicted for each proposal.

3. Describe the DISABILITIES OF FAST R-CNN

Ans: Region Proposal Dependence: Fast R-CNN relies on an external algorithm (e.g., Selective Search) to generate region proposals, which can be computationally expensive and introduce errors.

Multiple Stages: The separate stages of proposal generation and classification/regression can be inefficient.

4. Describe how the area proposal network works.

Ans: The Address R-CNN introduced the APN, which is a small CNN that shares convolutional features with the main network. It efficiently predicts multiple bounding boxes (anchors) and their confidence scores at different scales and aspect ratios within the feature map. This eliminates the need for a separate proposal generation step.

5. Describe how the RoI pooling layer works.

Ans: The RoI pooling layer takes a feature map and RoIs (regions of interest) as input. It performs spatial binning to ensure each RoI has the same fixed size output, regardless of its size in the feature map. This allows for processing RoIs of different sizes through subsequent fully connected layers. Common binning methods include max pooling and average pooling.

6. What are fully convolutional networks and how do they work? (FCNs)

Ans: FCNs are CNN architectures that don't use fully connected layers for image segmentation or dense prediction tasks. Instead, they employ convolutional layers throughout the network to produce a pixel-wise output map. This allows for predictions at every location in the input image. Here's a common FCN structure:

* Encoder: A pre-trained CNN (e.g., VGG or ResNet) extracts features at different scales.
* Decoder: Upsampling layers or transposed convolutions increase the resolution of the feature maps, allowing for precise localization.
* Output Layer: A convolutional layer with 1 filter per class predicts the probability of each class belonging to each pixel.

7. What are anchor boxes and how do you use them?

Ans: Anchor boxes are a set of predefined bounding boxes at various scales and aspect ratios that are placed on a regular grid across the feature maps of an image. In SSD, for instance, multiple anchor boxes are associated with each location in the feature map. The model predicts offsets (offsets) to adjust these anchors to better match the actual objects in the image.

8. Describe the Single-shot Detector’s architecture (SSD)

Ans: SSD is a one-stage object detection model that directly predicts bounding boxes and class probabilities from a single pass through the network. Here's a breakdown:

* Base Network: A pre-trained CNN (like VGG-16) is used as a feature extractor, with its fully connected layers removed.
* Multi-Scale Feature Maps: Additional convolutional layers are added on top of the base network to create feature maps at various scales. This allows for detection of objects of different sizes.
* Default Boxes: A set of default bounding boxes (anchors) with different aspect ratios and scales are predefined for each feature map location.
* Prediction Layers: Each feature map location has multiple convolution filters to predict:
  + Class probabilities for each default box
  + Offsets (offsets) to adjust the default boxes to better match objects

9. HOW DOES THE SSD NETWORK PREDICT?

Ans: SSD performs these steps:

* The input image is passed through the base network and additional convolutional layers to generate feature maps at multiple scales.
* For each feature map location and its associated default boxes:
  + The model predicts class probabilities for each default box using class filters.
  + It predicts offsets (offsets) to refine the default boxes using offset filters.
* Non-maximum suppression (NMS) is applied to remove redundant bounding boxes and select the most confident detections for each class.

10. Explain Multi Scale Detections?

Ans: Multi-scale detection is a technique used in object detection models to improve their ability to find objects of different sizes in an image. Here are some common approaches:

* Feature Pyramids: This approach creates feature maps at different scales by:
  + Downsampling the input image at different scales to generate a pyramid of images.
  + Passing each image in the pyramid through a convolutional neural network (CNN) to extract features at corresponding scales.
  + Combining these feature maps in a way that preserves information from all scales. This can be done using techniques like feature fusion, where information from different scales is concatenated or summed.
* Anchor Boxes with Different Scales and Aspect Ratios: This approach uses a set of predefined bounding boxes (anchor boxes) at various scales and aspect ratios. These anchors are placed on a regular grid across the feature maps of an image. The model predicts offsets (offsets) to adjust these anchors to better match the actual objects in the image. This is commonly used in models like SSD (Single-Shot Detector) and YOLO (You Only Look Once).
* Resizing the Input Image: Some models might resize the input image to multiple scales and pass them through the network independently. This can be computationally expensive but effective for detecting very small or large objects.

11. What are dilated (or atrous) convolutions?

Ans: Dilated convolutions, also called atrous convolutions, are a type of convolutional operation used in CNNs to control the dilation rate of the filter (kernel) during convolution. Here's how they work:

* Standard Convolution: In a standard convolution, the filter slides across the input feature map with a stride of 1, meaning it moves one unit at a time. Each output element depends on a corresponding neighborhood of input elements defined by the filter size.
* Dilation Rate: Dilated convolutions introduce a dilation rate (also called a "hole rate") greater than 1. This inserts spaces (holes) between filter elements during the convolution operation. This allows the filter to capture a larger context from the input feature map without increasing its size.