CNN Classification and Object Detection Questions

1. **How does the architecture of a CNN designed for image classification differ from one used for object detection?**
    - A CNN used for image classification typically consists of several convolutional layers followed by pooling layers and finally fully connected layers. The output of the fully connected layers is used to classify the entire image into one of the pre-defined categories. The model predicts a single label corresponding to the image. Whereas, a CNN in object detection not only classifies objects in an image but also predicts their locations, which is known as **localisation**. This requires more architectural components such as region proposal networks (RPN). These components help in predicting multiple bounding boxes around objects, along with the classes of those objects. **Basically, classification predicts a single label and object detection predicts multiple bounding boxes and labels for objects in the image.**
    - Example: In image classification, the CNN might take an input image of a cat and predict a single label "cat" without determining where the cat is located in the image. In object detection, the CNN would take the same image and predict both the label "cat" and the location of the cat within the image by drawing a bounding box around it.

2. What is the role of a Region Proposal Network (RPN) in object detection models like Faster R-CNN, and how does it help in identifying objects in an image?

Ans: Region proposals are candidate regions in an image that are likely to contain objects. They serve as inputs for further processing in object detection models. Traditionally, region proposals were found using techniques like **selective search**, which grouped similar pixels in an image to propose regions that might contain objects. However, selective search was slow and not learnable since it used a CPU, limiting efficiency. In models like Faster R-CNN, the network uses an **RPN** to replace selective search. The RPN operates by sliding a small network over the convolutional feature map of the image. At each sliding window position, the RPN generates multiple **anchor boxes** (predefined bounding boxes of different scales and aspect ratios) and predicts whether each box contains an object or not. The RPN also refines the anchor box locations based on the image content. The RPN narrows down the search space by generating region proposals, allowing the network to focus on areas likely to contain objects. This speeds up the object detection process while improving accuracy, as the network does not need to evaluate the entire image exhaustively.

3. Explain how transfer learning can be applied to a CNN for both image classification and object detection tasks?

- **Image Classification:** Use a pre-trained CNN (e.g., on ImageNet), freeze early layers, and fine-tune the later layers on your specific classification dataset.
- **Object Detection:** Use a pre-trained model's backbone (like ResNet) for feature extraction, then attach an object detection head (like RPN or SSD) for bounding box and class predictions.

4. What is the significance of anchor boxes in object detection models, and how do they assist CNNs in predicting object locations?

Ans: Anchor Boxes are predefined boxes of different sizes and aspect ratios used to predict objects in various scales and shapes. The model adjusts these boxes to fit the objects in the image, helping predict object locations more accurately without needing to generate bounding boxes from scratch.

5. Compare the loss functions used in CNN-based image classification (e.g., cross-entropy loss) and object detection (e.g., localization loss and classification loss). How are they combined in object detection tasks?

- **Image Classification:** Uses **cross-entropy loss** to measure classification accuracy.
- **Object Detection:** Uses both **classification loss** (e.g., cross-entropy) and **localization loss** (e.g., smooth L1 loss) to measure bounding box accuracy.
- **Combination:** In object detection, these two losses are combined (often added) to train the model to predict both the object class and its location.

6. How does the role of fully connected layers in CNNs for image classification differ from their role (or absence) in object detection networks like YOLO and SSD?

- **Image Classification:** Fully connected layers are used at the end to output class probabilities.
- **Object Detection (YOLO/SSD):** Fully connected layers are often removed. Instead, these models use convolutional layers to directly predict bounding boxes and class probabilities across the entire image grid.

7. What are the key architectural characteristics of the VGG network, and how does its deep, sequential structure contribute to improved performance in image classification tasks?

Ans: VGG uses 2 small 3x3 convolution filters, instead of 5x5 as used in AlexNet, deep layers (16 or 19), and a simple sequential structure. The deep architecture allows the model to learn more complex features, leading to better performance in image classification.

8. Explain how Non-Maximum Suppression (NMS) is used in object detection models to eliminate redundant bounding boxes and improve detection accuracy.
Ans: NMS removes the redundant and overlapping bounding boxes by selecting the box with the highest confidence score and discarding others that have a high overlap. If all the boxes have a high IOU, the threshold will be higher and we cannot give false positives, in that case all the redundant boxes are removed. The purpose is to keep only the best bounding box for each detected object, improving detection accuracy.

9. In a CNN-based object detection model like YOLO, how is the concept of grid cells used to predict multiple bounding boxes in an image, and how does it affect the model's efficiency and accuracy?
Ans: The image is divided into a grid, and each cell is responsible for predicting bounding boxes and object classes within that cell. This method allows YOLO to predict multiple objects at different locations simultaneously, improving efficiency and speed, though it may reduce accuracy for small objects.