A1) Image Classification CNNs are simpler and designed to output a single class label for an image.

Object Detection CNNs are more complex, designed to detect multiple objects with both class predictions and bounding box regressions.

A2) The Region Proposal Network (RPN) in faster R-CNN plays a key role in identifying potential regions in an image that may contain objects. It does this by applying anchor boxes of various sizes and aspect ratios across the feature maps, and then predicting both the likelihood that each box contains an object (objectness score) and the precise bounding box coordinates. This approach greatly accelerates the object detection process by efficiently learning to focus on relevant regions, allowing faster R-CNN to detect objects more effectively and with better precision.

A3) Transfer learning is a powerful technique that can be applied to CNNs for both image classification and object detection tasks. The idea is to take a pre-trained CNN model (usually trained on a large dataset like ImageNet) and adapt it to a new, often smaller, task-specific dataset. In image classification, I can use a ResNet model pre-trained on ImageNet, replace the last classification layer with one suitable for my specific categories (e.g., 5 classes instead of 1,000), and fine-tune it on my smaller dataset. In object detection, I could take Faster R-CNN with a pre-trained ResNet backbone. I'd keep the backbone's weights, replace the region proposal network (RPN) and classification head with ones suited to my dataset, and then fine-tune these detection-specific layers on my object detection dataset.

A4) Anchor boxes are a crucial component in object detection models as they provide predefined reference bounding boxes of various sizes and aspect ratios, allowing CNNs to detect objects of different shapes and scales within an image. At each location on the feature map, the model generates multiple anchor boxes and predicts adjustments to these boxes to fit the objects more precisely. By associating class probabilities with each anchor box, the CNN can predict the presence and location of multiple objects in a single image, even if they overlap.

A5) In CNN-based image classification, the primary loss function that is used is cross-entropy loss, which measures the difference between the predicted class probabilities and the true labels. For object detection tasks, the loss is more complex, combining two main components: classification loss and localization loss. Classification loss, often still cross-entropy or focal loss, is used to predict the correct object class, while localization loss, typically smooth L1 or IoU-based loss, measures how accurately the predicted bounding box aligns with the actual object. In object detection models, these two losses are combined in a weighted sum, where both the ability to classify objects correctly and the precision of their localization are optimized together, ensuring that the network can both identify and accurately locate multiple objects within an image.

A6) In CNNs designed for image classification, fully connected layers play a key role in transforming the high-level feature maps into class probabilities. These layers take the output of the final convolutional layers and flatten them into a vector, which is then passed through dense layers to produce the final class predictions. In contrast, object detection networks like YOLO and SSD generally minimize or avoid using fully connected layers, opting instead for convolutional approaches to predict bounding boxes and class scores directly from the feature maps. By doing this, these models maintain spatial information better, allowing them to detect objects at various scales and locations across the image in a more efficient, end-to-end manner without the need for flattening or dense layers. This helps these networks process images more quickly and with fewer parameters compared to classification-based CNNs.

A7) The VGG network is characterized by its deep, sequential architecture, which primarily consists of a series of convolutional layers followed by small 3x3 filters, often stacked in groups, along with 2x2 max pooling layers for downsampling. This design enables the network to capture intricate features at various levels of abstraction while maintaining spatial hierarchies. The depth of the VGG architecture, which can include up to 19 layers, allows it to learn increasingly complex representations of the input data. By using small filter sizes and deep layers, VGG effectively increases the receptive field without a significant increase in the number of parameters, resulting in improved performance on image classification tasks.

A8) Non-Maximum Suppression (NMS) is a technique used in object detection models to refine predictions by eliminating redundant bounding boxes. After generating multiple boxes for the same object, NMS selects the box with the highest confidence score and compares it to the others based on their Intersection over Union (IoU) values. If the IoU of any lower-scoring box exceeds a predefined threshold with the selected box, that lower-scoring box is discarded. This iterative process continues until all boxes are evaluated, resulting in a final set of bounding boxes that reduces overlap and improves detection accuracy.

A9) In a CNN-based object detection model like YOLO, the image is divided into a grid of cells, and each cell is responsible for predicting bounding boxes and class probabilities for objects whose centers fall within that cell. Specifically, each grid cell outputs a fixed number of bounding boxes, along with confidence scores and class predictions. This approach allows the model to detect multiple objects within the same grid cell, enhancing its efficiency by enabling real-time processing. The use of grid cells also improves accuracy by providing a structured way to localize objects, as each cell can focus on detecting objects in a specific region of the image, leading to better spatial awareness and reducing the chances of missing smaller or overlapping objects.