HW to Chapter 17 “Overlapping Objects and Semantic Segmentation”

***Non-programming Assignment:***

1. What are anchor boxes and how do they work?
2. What is bounding box prediction and how does it work?
3. Describe R-CNN
4. What are advantages and disadvantages of R-CNN?
5. What is semantic segmentation?
6. How does deep learning work for semantic segmentation?
7. What is transposed convolution?
8. Describe U-Net.

Answer

**1. What are anchor boxes and how do they work?**

Anchor boxes are basically like pre-drawn rectangles. Imagine you're trying to find objects in a picture. Instead of starting from scratch, you've got these pre-made boxes with different sizes and shapes already laid out.

* **They're like templates:** They act as a guide for the computer to find objects.
* **Placed strategically:** They're spread across the image in a grid, kind of like a chessboard, to cover all areas.
* **Variety is key:** They come in different sizes and shapes to catch all sorts of objects.

They help detect objects of different scales and shapes by being placed at various locations across an image in a grid-like pattern. During training and inference, the model predicts adjustments to anchor boxes to match actual object locations, rather than directly predicting object coordinates from scratch. This eliminates the need for a sliding window approach, making detection more efficient.

Basically, they help the computer find things faster. Instead of checking every single pixel, the computer just tweaks these boxes to fit the objects it sees. This is how models like SSD, YOLO, and Faster R-CNN work – they use anchor boxes to get a head start.

**2. What is bounding box prediction and how does it work?**

Bounding box prediction determines where objects are in an image. A bounding box is a rectangle enclosing an object, defined by:

* **(x-min, y-min)** – Top-left corner
* **(x-max, y-max)** – Bottom-right corner

**How it works:**

During training, the model learns to adjust anchor boxes to match the actual object locations by minimizing differences (typically using IoU – Intersection over Union). At inference, it predicts the necessary refinements to fine-tune the anchor boxes, resulting in precise bounding boxes around detected objects.

**Key challenges:**

* Handling objects of different scales
* Managing aspect ratio variations
* Ensuring accurate localization across multiple object classes

**3. Describe R-CNN**

R-CNN (Region-based Convolutional Neural Network) is an object detection model that first proposes candidate object regions and then classifies them using a CNN.

**How it works:**

1. **Region Proposal** – The image is segmented into ~2000 possible object regions using Selective Search.
2. **Feature Extraction** – Each proposed region is resized and passed through a CNN to generate feature vectors.
3. **Classification & Refinement** – The features are fed into an SVM to classify objects, and a regression model adjusts the bounding boxes.

This method was groundbreaking but had inefficiencies, leading to faster models like Fast R-CNN and Faster R-CNN.

**4. What are the advantages and disadvantages of R-CNN?**

***Advantages:***

* Provides accurate object detection by leveraging convolutional features.
* Separates complex detection task into manageable subtasks
* Handles objects of different sizes, orientations, and scales well.
* Can be adapted for tasks like instance segmentation and object tracking.

***Disadvantages:***

* Computationally expensive due to processing each of the 2000 region proposals separately.
* Slow inference time (~47 seconds per image), making it unsuitable for real-time applications.
* Redundant region proposals lead to inefficiencies.
* Not an end-to-end trainable model since region proposal and classification are separate steps. Which is essentially a complex, multi-stage training pipeline.
* Significant storage requirements for region features

**5. What is semantic segmentation?**

Semantic segmentation is a technique where every pixel in an image is assigned a category label. Unlike object detection, which uses bounding boxes, semantic segmentation provides a dense classification map.

**Key characteristics:**

* **Per-pixel classification** – Each pixel is labeled with a class (e.g., road, car, pedestrian).
* **Spatial understanding** – Helps in tasks requiring detailed object boundaries.
* **Used in applications like** – Autonomous driving, medical imaging, and scene understanding.

The main difference between **semantic segmentation** and **instance segmentation** is that semantic segmentation does not differentiate between individual objects of the same class.

**6. How does deep learning work for semantic segmentation?**

Deep learning models for semantic segmentation use CNNs with some modifications:

1. **Encoder (Feature Extraction)** – A CNN extracts features while reducing spatial resolution.
2. **Decoder (Upsampling)** – Convolution layers restore spatial resolution for pixel-wise classification.
3. **Fully Convolutional Networks (FCNs)** – Replace fully connected layers with transposed convolutions to retain spatial information.
4. **Skip Connections (e.g., in U-Net)** – Help recover fine details lost during downsampling.

These architectures allow models to achieve high accuracy in tasks requiring detailed object segmentation.

**7. What is transposed convolution?**

Transposed convolution (also called deconvolution or up-convolution) increases the spatial resolution of feature maps, making it essential for segmentation models.

**How it works:**

* **Expands feature maps** rather than reducing them (like regular convolution).
* **Uses learnable filters** to reconstruct spatial details.
* **Common in segmentation models** to upsample low-resolution feature maps back to the original image size.

**Challenges:**

* Can be computationally expensive.
* May introduce artifacts or blurriness if not designed properly.

Despite these challenges, transposed convolution plays a crucial role in image generation and segmentation tasks.

**8. Describe U-Net**

U-Net is a CNN architecture designed for **biomedical image segmentation**, but it’s widely used for other segmentation tasks as well.

**Architecture:**

1. **Contracting Path (Encoder)** – Extracts features using convolution and max-pooling.
2. **Expansive Path (Decoder)** – Uses transposed convolution to restore spatial resolution.
3. **Skip Connections** – Link encoder and decoder layers to retain fine details.
4. **Final Layer** – Uses a 1×1 convolution to assign class labels to each pixel.

**Key Features:**

* Retains fine details using **skip connections**.
* Works well even with **limited training data**.
* Popular in **medical imaging** and other tasks requiring high-precision segmentation.

Because of its effectiveness in preserving spatial information, U-Net remains one of the top choices for segmentation problems.