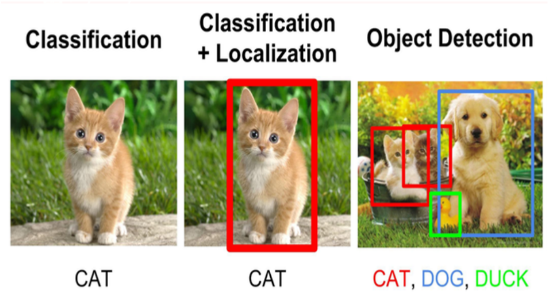
Two-Stage Object Detection

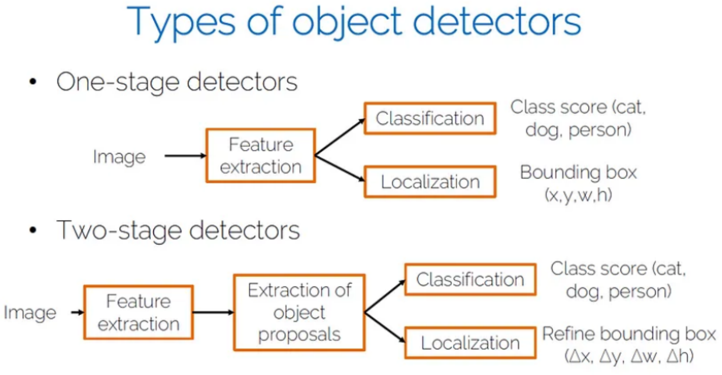
ANAND M K

Introduction to Object Detection

- **Image Classification:** Takes an image and predicts the object in the image.
- **Object Localization:** Locates the presence of an object in the image and represents it with a bounding box.
- **Object Detection:** Combines image classification and object localization. It takes an image as input and produces one or more bounding boxes with the class label attached to each bounding box.



Single-Stage vs. Two-Stage Object Detectors



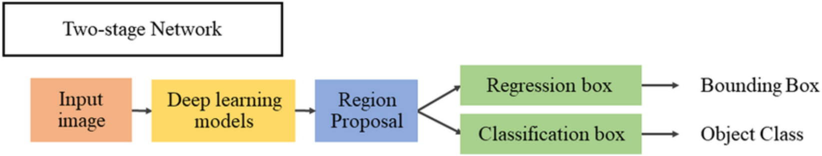
Types of Object Detector

Single-Stage vs. Two-Stage Object Detectors

- **Single-Stage Object Detector:**
  - Directly goes from the image to classification and bounding box coordinates.
  - Features are extracted using a CNN, which are then used for classification and regression.
  - **Advantages:**
    - Very fast, suitable for real-time object detection.
  - **Disadvantages:**
    - Performance can be poorer than two-stage detectors.
  - **Examples:** YOLO family, SSD, RetinaNet.
- **Two-Stage Object Detector:**
  - Divides the process into two steps:
    1. Extracts features using a CNN.
    2. Extracts regions of interest (object proposals) for classification and localization.
  - **Advantages:**
    - Extremely accurate with high mean Average Precision (mAP).
    - More suitable for applications where accuracy is prioritized over speed (e.g., medical imaging).
  - **Examples:** R-CNN family.

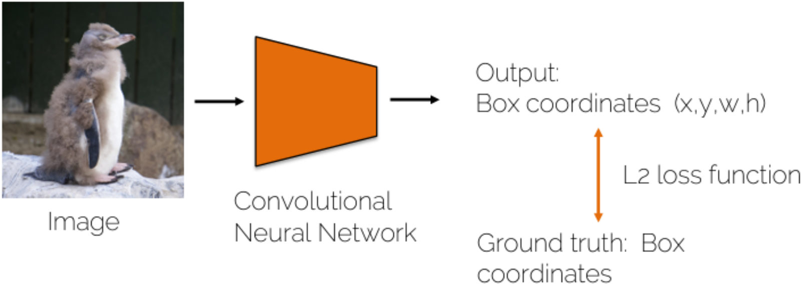
Steps of Two-Stage Object Detection

- **Step 1: Region Proposal**
  - The model generate candidate regions, known as region proposals.
  - These regions are likely to contain objects.
- **Step 2: Classification and Bounding Box Refinement**
  - Each proposed region is classified to determine the object category.
  - The bounding box is adjusted to accurately surround the detected object.



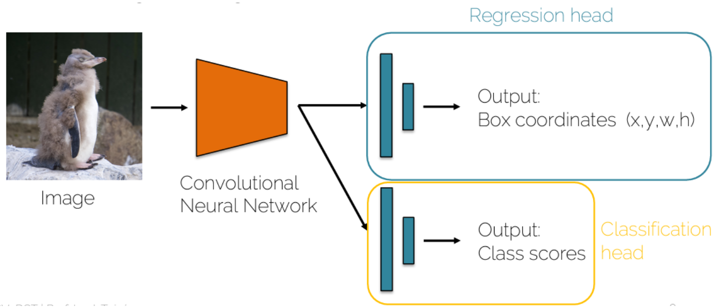
Localization: Bounding Box Regression

- Bounding Box Regression is the process of refining the predicted object location by learning four key coordinates:
  - x, y (center of the box)
  - w, h (width and height of the box)
- A CNN extracts features from the image and predicts the coordinates of the object.
- The goal is to minimize the difference between the predicted box and the ground truth box using a loss function, typically L2 loss.

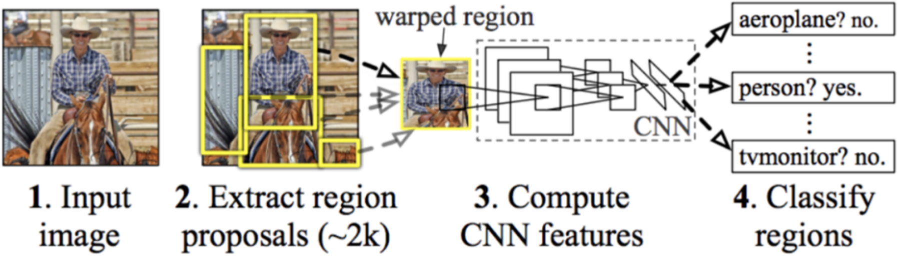


Localization and Classification

- **Classification:**
  - The task is to classify the object in the bounding box.
  - The CNN extracts features and uses fully connected layers to predict class scores.
  - The loss function used for class score prediction is **Softmax loss**.



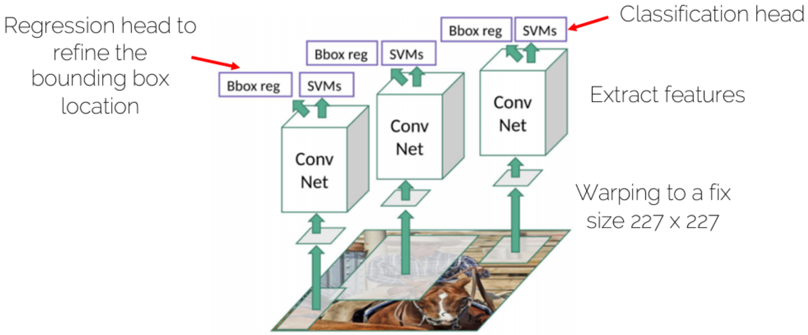
R-CNN (Regions with Convolutional Neural Networks



- **Step 1: Region Proposal Generation**
  - Use selective search to generate around 2000 candidate regions.

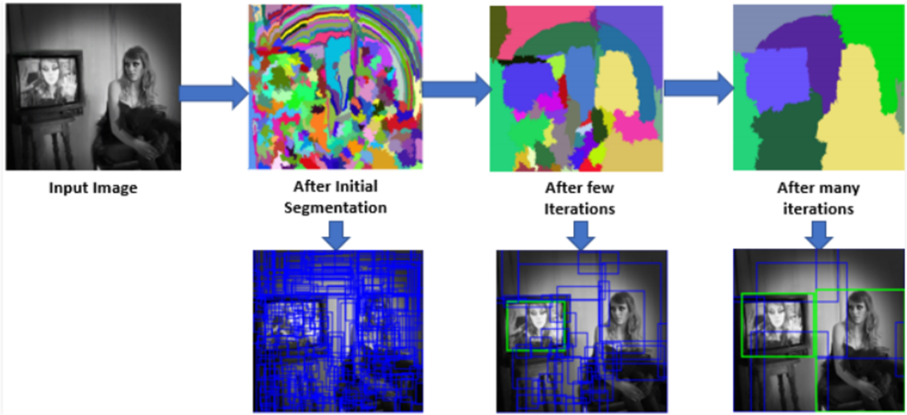
R-CNN: Steps 2 and 3 (Feature Extraction and Classification)

- **Step 2: Feature Extraction**
  - Resize each region to a fixed size and extract features using a CNN.
- **Step 3: Classification and Bounding Box Regression**
  - Classify each region proposal using an SVM (Support Vector Machine).
  - Refine bounding box coordinates using linear regression.



Selective Search for Region Proposals

- Selective Search is a region proposal algorithm used in object detection.
- It generates candidate object locations by grouping similar regions in an image.



Selective Search for Region Proposals

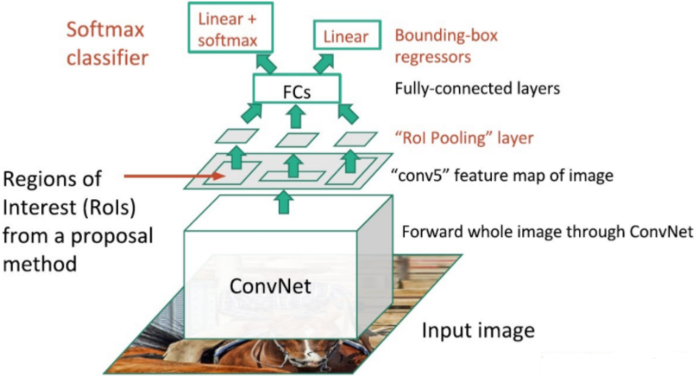
- **Step 1: Over-Segmentation**
  - Break the image into many small regions based on pixel color and intensity.
- **Step 2: Initial Region Proposal**
  - Treat each small region as a starting point for object candidates.
- **Step 3: Merge Similar Regions**
  - Combine neighboring regions that are similar in color, texture, size, and shape.
- **Step 4: Generate Region Proposals**
  - As regions merge, create larger candidate regions for potential objects.
- **Step 5: Repeat Merging**
  - Continue merging until the entire image is covered by larger regions, generating multiple proposals at different scales.

Limitations of R-CNN

- **Slow Processing Speed:**
  - Each region proposal requires a separate forward pass through the object detector, making it resource-intensive and slow.
- **Fixed Object Proposal Algorithm:**
  - The use of a fixed selective search algorithm does not allow for improvement through training.
- **Separate Training:**
  - Feature extraction and SVM classifier are trained separately, limiting the model's ability to exploit learning potential.

Introduction to Fast R-CNN

- Fast R-CNN is an advanced object detection framework that enhances the original R-CNN approach.
- By employing a single forward pass through a convolutional neural network (CNN), it efficiently extracts features from the entire image, minimizing computational overhead.

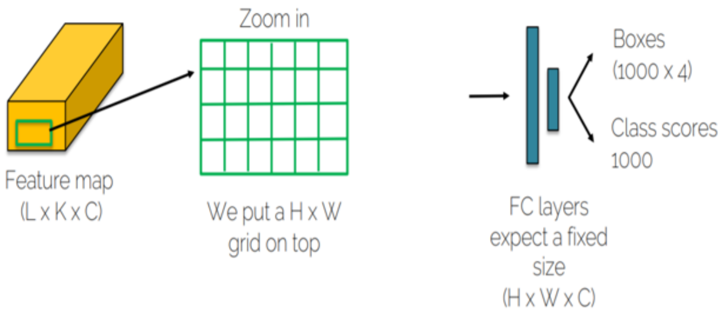


Overview of Fast R-CNN Working

- **Single Forward Pass:**
  - Fast R-CNN processes the entire image through a convolutional neural network (CNN) in a single forward pass, generating a feature map.
- **Region Proposal Generation:**
  - An external algorithm (e.g., Selective Search) generates a set of candidate region proposals from the image.
- **Region of Interest (RoI) Pooling:**
  - The feature map is used to extract features for each region proposal.
  - RoI pooling converts these features into a fixed size to enable processing by fully connected layers.
- **Fully Connected Layers:**
  - The pooled features are fed into fully connected layers for classification and bounding box regression.
- **Output:**
  - The softmax layer predicts class probabilities for each region.
  - The bounding box regression layer refines the bounding box coordinates for more accurate localization.

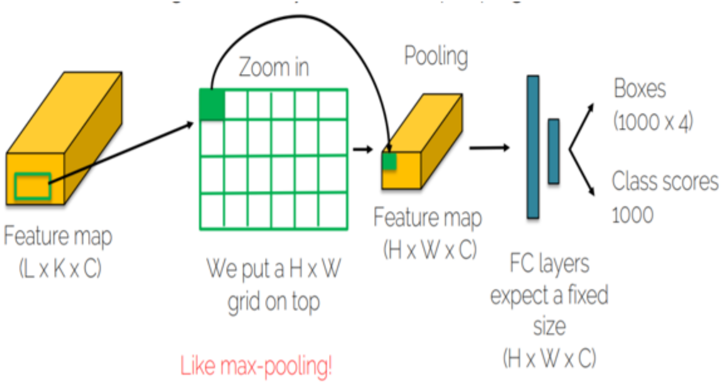
ROI Pooling: Key Concepts

Feature Map: Dimensions $L \times K \times C$, where $L \times K$ is the spatial size and C is the number of channels.
Object Proposals: Define regions (green box) of interest in the feature map, which vary in size.
Problem: FC layers need fixed-size input $H \times W \times C$.



ROI Pooling

Pooling: A grid $H \times W$ is placed over the region, and max-pooling is applied in each cell to produce a fixed-size $H \times W \times C$ feature map.
FC Layer Input: The output is resized to fit the FC layers.

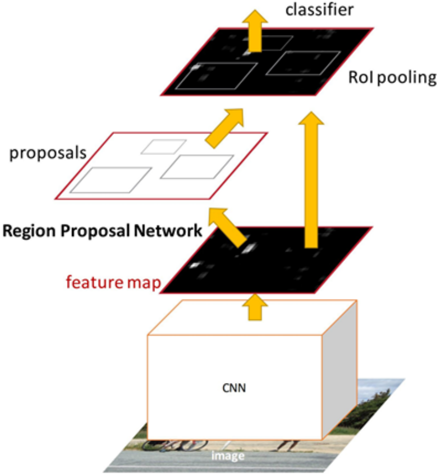


Challenges of Fast R-CNN

- **Selective Search Dependency:**
  - Fast R-CNN uses Selective Search as a method for generating Regions of Interest (RoIs).
  - This approach is inherently slow and time-consuming.
- **Performance Issues:**
  - Detection time is approximately 2 seconds per image, an improvement over R-CNN.
  - However, in large real-life datasets, this speed may not be sufficient, making Fast R-CNN less effective.

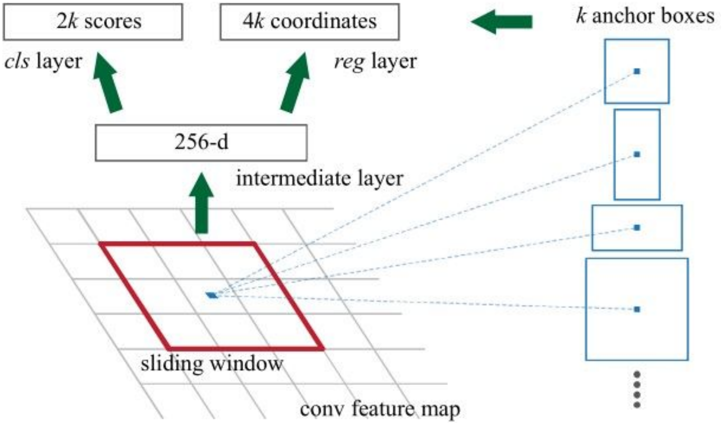
Faster R-CNN

- Faster R-CNN is an extension of Fast R-CNN with a Region Proposal Network (RPN) that enhances speed and efficiency.
- RPN replaces traditional region proposal methods (Selective Search) with a fully convolutional network.
- RPN improves object detection by proposing regions (bounding boxes) directly from feature maps.



Region Proposal Network (RPN) Overview

- RPN generates region proposals (bounding boxes) directly from feature maps.
- RPN is fully convolutional and slides over feature maps to generate proposals for objects in the image.



RPN - Working Mechanism

- **Sliding Window:**
  - RPN uses a sliding window that moves across the convolutional feature map to detect potential objects.
- **256-d Intermediate Layer:**
  - At each sliding window position, a 256-dimensional feature vector is extracted to capture visual information.
- **k Anchor Boxes:**
  - For each sliding window, k anchor boxes (default: 9) are generated with various scales and aspect ratios.
- **2k Scores (cls layer):**
  - The classification layer predicts whether each of the k anchor boxes contains an object (positive) or not (negative).
  - The result is $2k$ scores—two for each anchor box: one score for object presence and one for absence.
- **4k Coordinates (reg layer):**
  - The regression layer refines the bounding box coordinates for each anchor by predicting 4 values: (x, y, width, height).
  - The result is $4k$ coordinates for anchor box refinement.

Advantages of Faster R-CNN

- **Speed:**
  - Replaces traditional region proposal methods with a Region Proposal Network (RPN), significantly improving detection speed.
- **Unified Framework:**
  - Integrates region proposal and object detection into a single network, streamlining the process and reducing computational overhead.
- **Scalability:**
  - Capable of handling a large number of object categories without separate steps for region proposal and detection.
- **Improved Accuracy:**
  - Jointly trained RPN and Fast R-CNN network improve localization and classification accuracy.

Conclusion

- Two-stage object detection algorithms like Faster R-CNN separate the process into region proposal and object classification, offering higher accuracy by refining object localization.
- The introduction of Region Proposal Network (RPN) significantly improved both speed and accuracy in the two-stage detection pipeline.
- While two-stage methods are generally more accurate, one-stage detectors (like YOLO, SSD) offer faster inference by directly predicting bounding boxes and class scores in a single step.
- Each approach has its own trade-offs between speed and accuracy, with Faster R-CNN excelling in applications requiring precise detection.

References

- Jun Xiao, Jinlong Chen, Yi Ning, and Yun Jiang. 2024. Object Detection in Recent Years: An Overview. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering (CAICE '24)*.
- **Faster R-CNN**: Shaoqing Ren, Kaiming He, Ross B. Girshick, Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.
- University of Michigan, Fall 2019 Object Detection Lecture Slides, `https://web.eecs.umich.edu/~justincj/teaching/eecs498/Fall2019/`
- R. Girshick, "Fast R-CNN," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.
- **R-CNN**: R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," *CVPR*, 2014.