AI Summer School 2025

Medical Imaging Informatics

University of Pittsburgh

Introduction to Object Detection

Instructor: Nick Littlefield, MS

Pitt SHRS School of Health and Rehabilitation Sciences

Computational Pathology & AI Center of Excellence

IEEE COMPUTER SOCIETY

Learning Objectives

After completing this lecture, you should be able to:

- Explain the purpose of object detection in computer vision and its application in medical imaging.
- Describe how sliding windows are used to scan images for object detection.
- Identify the limitations of traditional sliding window methods, including computational inefficiency.
- Understand how convolutional layers simulate sliding windows to improve efficiency in modern deep learning models.
- Define and interpret bounding boxes, including their coordinate format and practical use cases.
- Understand how to evaluate object detectors

Outline

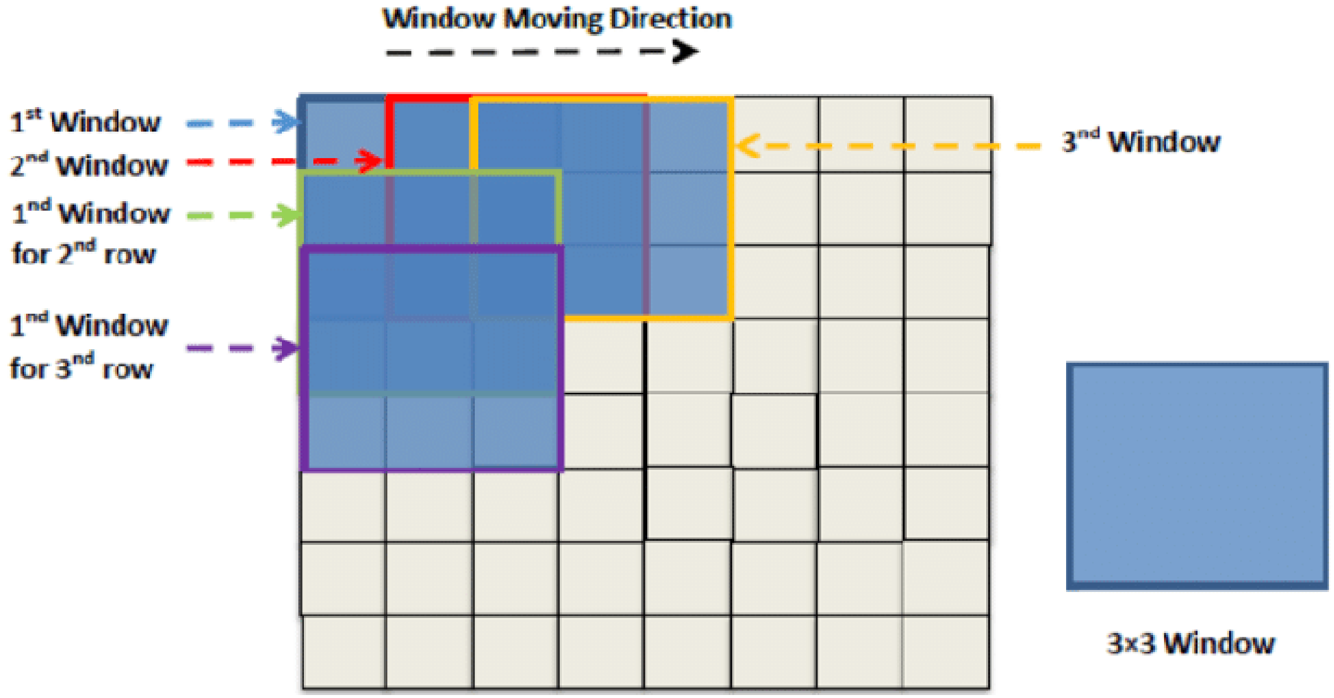
- Object detection
- Sliding windows
- Limitation of sliding windows
- Sliding windows using convolution
- Bounding box generation
- Evaluating Object Detectors

Object Detection

- **Object detection** identifies a specific object in an image.
- Combines deep learning and spatial reasoning to solve critical real-world problems.
  - Classification: What is in the image?
  - Detection: What *and* where is it?
  - Bounding Boxes localize objects in an image
  - Used in medical imaging to find tumors, organs, fractures, etc.
- Before the rise of deep learning, object detection was approached through exhaustive search using sliding windows in conjunction with hand-crafted features and traditional classifiers.

Sliding Windows

- A fixed-size rectangular window is moved across the image at regular intervals (a stride), scanning every region for the presence of an object.



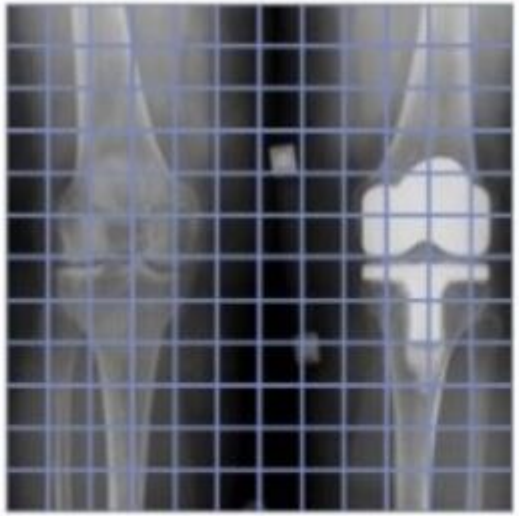
Sliding Windows

Algorithm:

1. **Define the Window Size:** Choose the dimensions of the window (e.g., 32x32 pixels for an image).
2. **Set the Stride:** Determine the step size for moving the window (e.g., moving 1 pixel at a time or larger steps like 5 pixels).
3. **Slide the Window:** Move the window across the image starting from the top-left corner, shifting by the stride amount each time.
4. **Extract Segments:** At each position, extract the segment of the image that falls within the window.
5. **Process Each Segment:** Apply the desired processing to each segment (e.g., pass it through a classifier to detect objects).
6. **Post-Processing**: Due to overlapping windows from small strides, many nearby windows may generate positive detections for the same object.

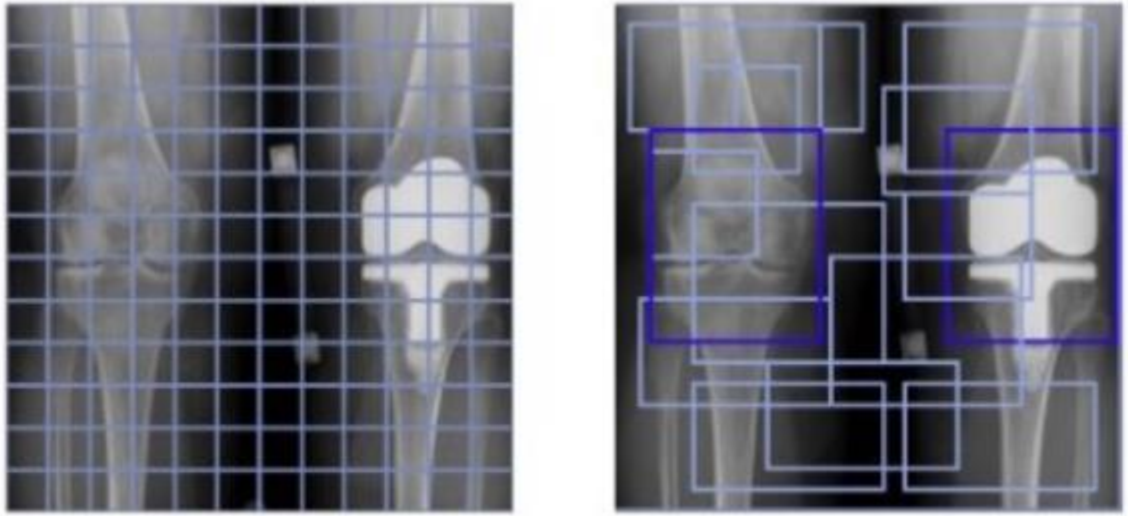
Sliding Windows: Example

- Each square is one window position; for example, a 32×32 pixel window sliding with a stride of 16 pixels.
- At each grid location, the algorithm extracts that patch and passes it through a feature extractor and classifier.



Sliding Windows: Example

- Some windows may partially contain objects of interest (e.g., prosthetic knee implant or native knee joint), while others may capture only background.
- Multiple adjacent windows may classify the same object as positive due to overlap.



Sliding Windows: Example

- Eliminate redundant detections and keep the best bounding box around the object.

Sliding Windows: Advantages/Disadvantages

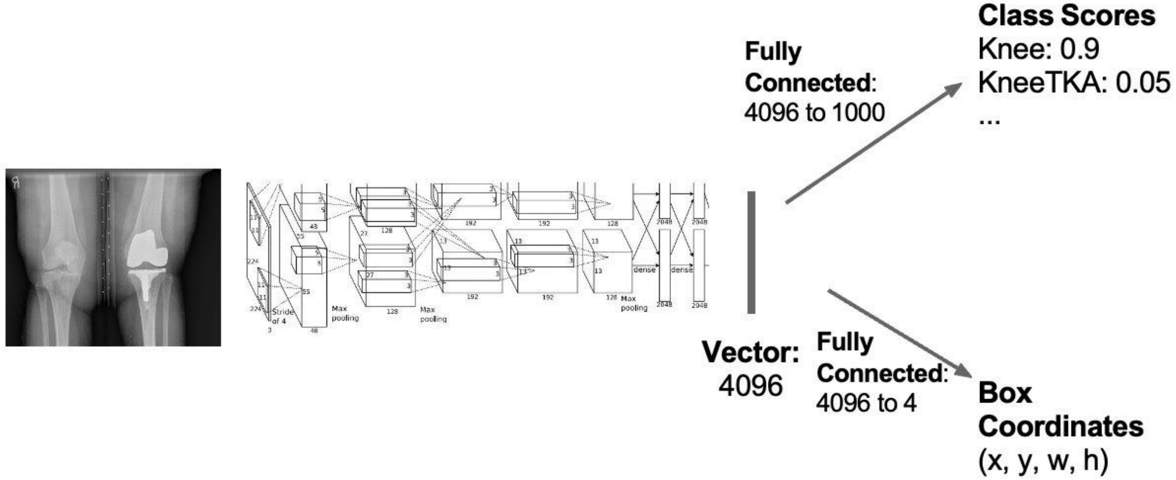
- **Advantages:**
  - **Simplicity:** Easy to implement and understand.
  - **Flexibility:** Can be adapted to different data types and processing tasks.
  - **Local Analysis:** Allows for the detailed examination of local regions within the data.
- **Disadvantages:**
  - **Computational Cost:** Especially with small strides, the number of segments can be very large, leading to high computational cost.
  - **Redundancy:** Overlapping windows mean the same data is processed multiple times, which can be inefficient.
  - **Scalability:** Not suitable for large-scale problems without optimization, due to the high number of operations required.

Sliding Windows: Optimizations

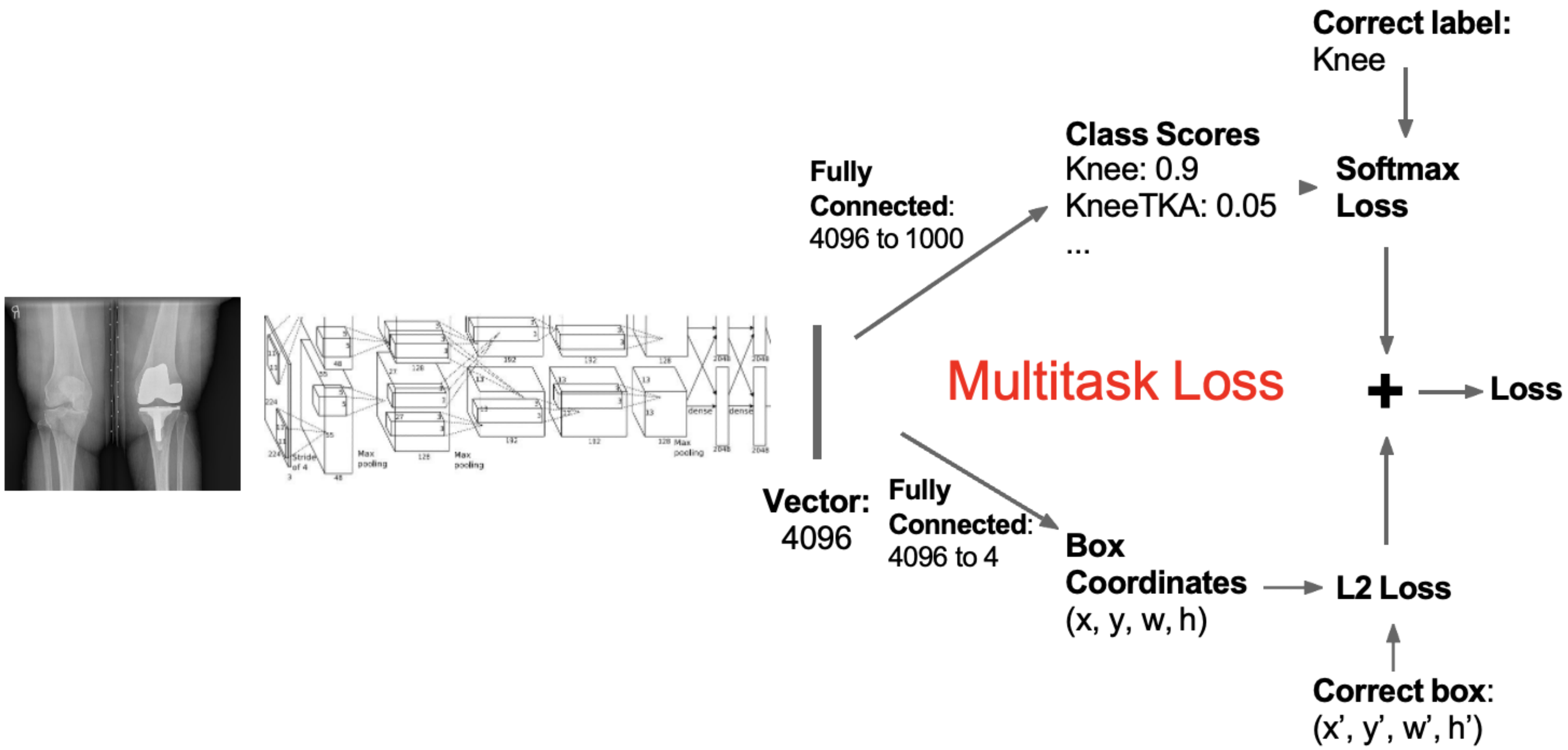
- **Optimizations:**
  - **Adjusting Stride:** Increasing the stride reduces the number of segments but may miss smaller objects or details.
  - **Multi-scale Sliding Windows:** Using windows of different sizes to capture objects at different scales.
  - **Feature Maps:** Using precomputed feature maps from neural networks to reduce the amount of data processed directly.

Object Detection using CNNs

- Uses a set of learned filters that slide across the entire image simultaneously.
- Shared weights reduce the number of parameters and computational cost.
- Can process the entire image in one pass, capturing spatial hierarchies of features.
- Can treat object detection as classification and localization to classify both the object and determine a bounding box for the object



Object Detection using CNNs



Advantages of CNNs

- **Advantages of Convolutional Implementation:**
  - **Efficiency:** Shared weights and fewer parameters lead to faster processing.
  - **Performance:** Ability to learn hierarchical features improves detection and classification accuracy.
  - **Scalability:** Suitable for large-scale problems and high-resolution images.
  - **End-to-End Learning:** Filters are learned directly from data, optimizing feature extraction for the specific task.

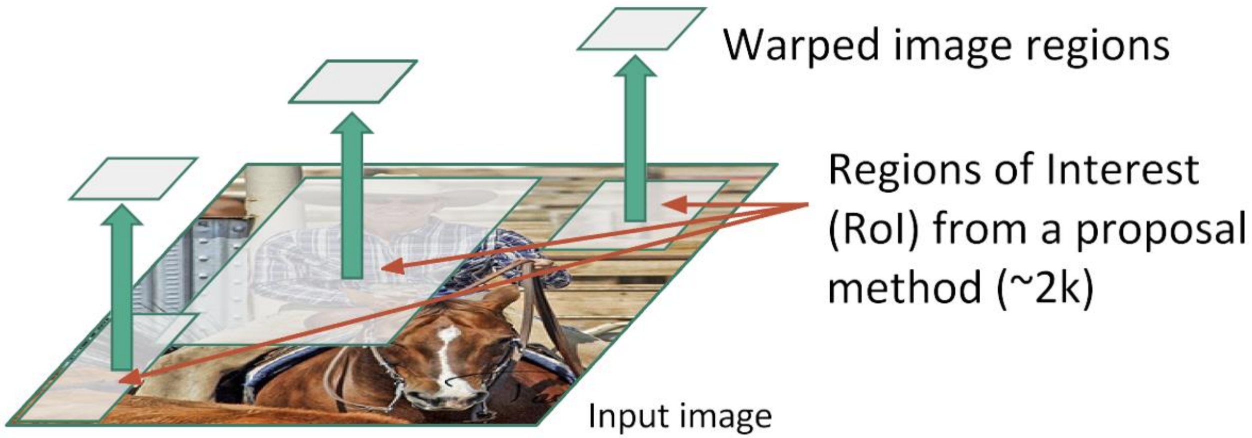
Selective Search using Region-based CNN (R-CNN)

- **Overview:** Combines hierarchical grouping of similar regions with exhaustive search to propose object regions.
  - **Process:**
    - Start with initial segmentation of the image into superpixels.
    - Merge superpixels based on color, texture, size, and shape.
    - Propose regions (bounding boxes) around merged segments.
- **Advantages:**
  - Does not require training data.
  - Provides a moderate number of object proposals.
- **Disadvantages:**
  - Computationally expensive and slow.
  - May produce redundant proposals.

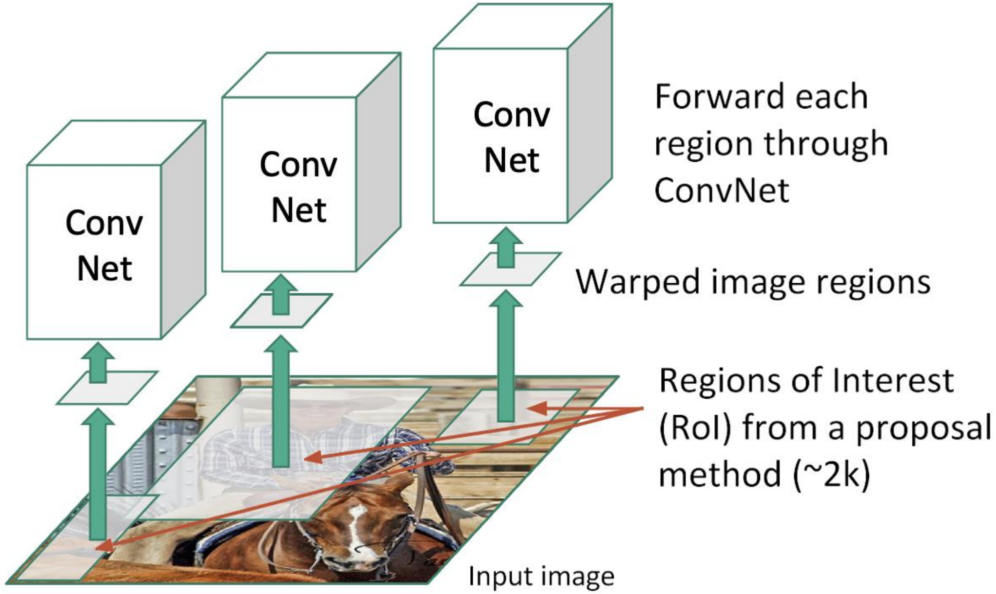
Selective Search using Region-based CNN (R-CNN)

- **Overview:** Combines hierarchical grouping of similar regions with exhaustive search to propose object regions.
  - **Process:**
    - Start with initial segmentation of the image into superpixels.
    - Merge superpixels based on color, texture, size, and shape.
    - Propose regions (bounding boxes) around merged segments.
- **Advantages:**
  - Does not require training data.
  - Provides a moderate number of object proposals.
- **Disadvantages:**
  - Computationally expensive and slow.
  - May produce redundant proposals.

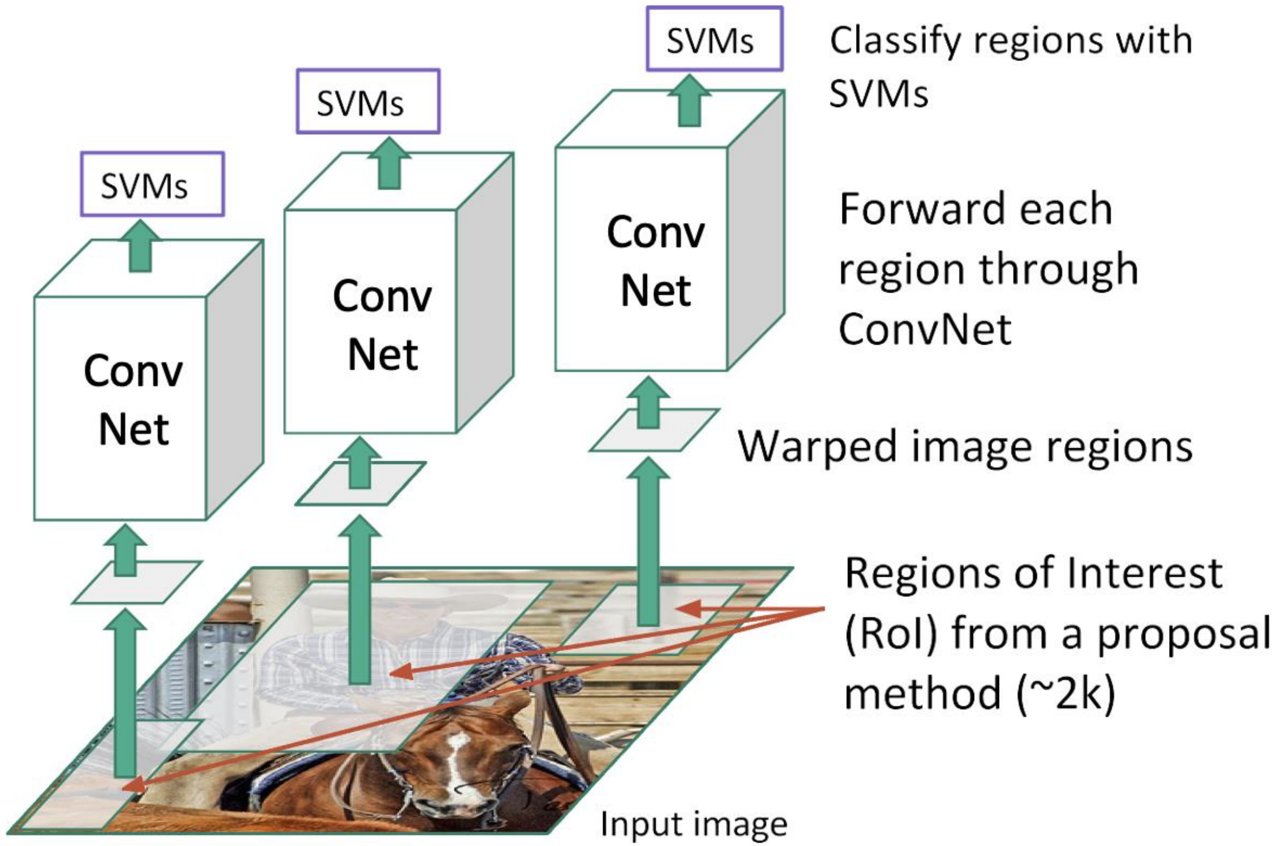
Selective Search using Region-based CNN (R-CNN)



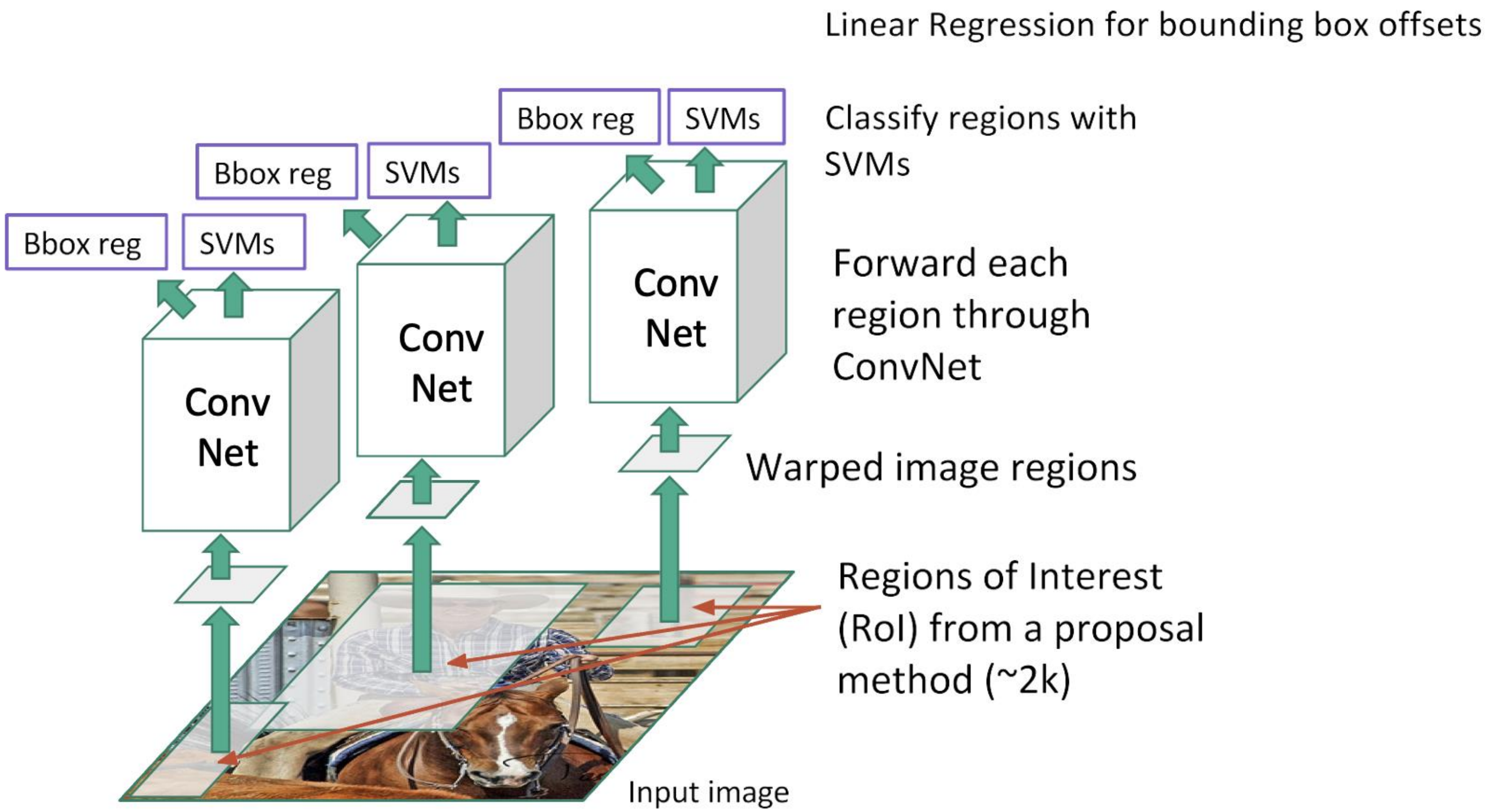
Selective Search using Region-based CNN (R-CNN)



Selective Search using Region-based CNN (R-CNN)



Selective Search using Region-based CNN (R-CNN)



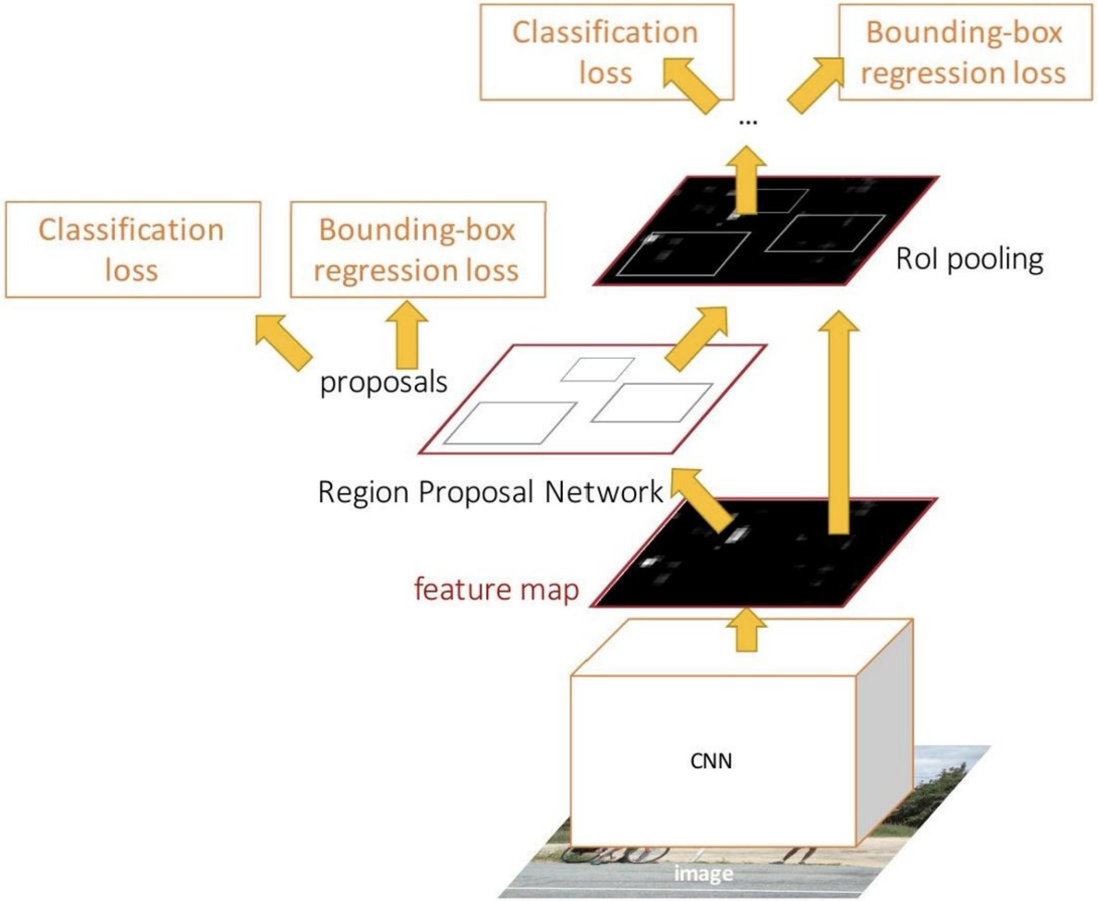
Problems with R-CNN

- Ad-hoc training objectives
  - Train post-hoc linear SVMs (hinge loss)
  - Train post-hoc bounding-box regressions (L2 loss)
- Training is slow (84h), takes a lot of disk space
  - Need to store all region crops
- Inference (detection) is slow
- Solution: Fast R-CNN

Faster R-CNN

- Introduces Region Proposal Networks (RPNs) which generate object proposals directly from feature maps
- **Process:**
  - A small network slides over the convolutional feature map.
  - For each sliding window, it predicts multiple bounding boxes and objectness scores.
- **Advantages:**
  - Integrated into the CNN, allowing for end-to-end training.
  - Faster and more efficient than traditional methods like Selective Search.
- **Disadvantages:**
  - Requires training on large datasets.

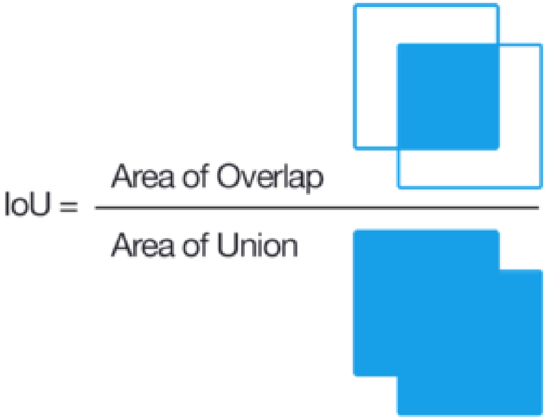
Faster R-CNN



Classification loss
Bounding-box regression loss
...
Classification loss
Bounding-box regression loss
RoI pooling
proposals
Region Proposal Network
feature map
CNN
image

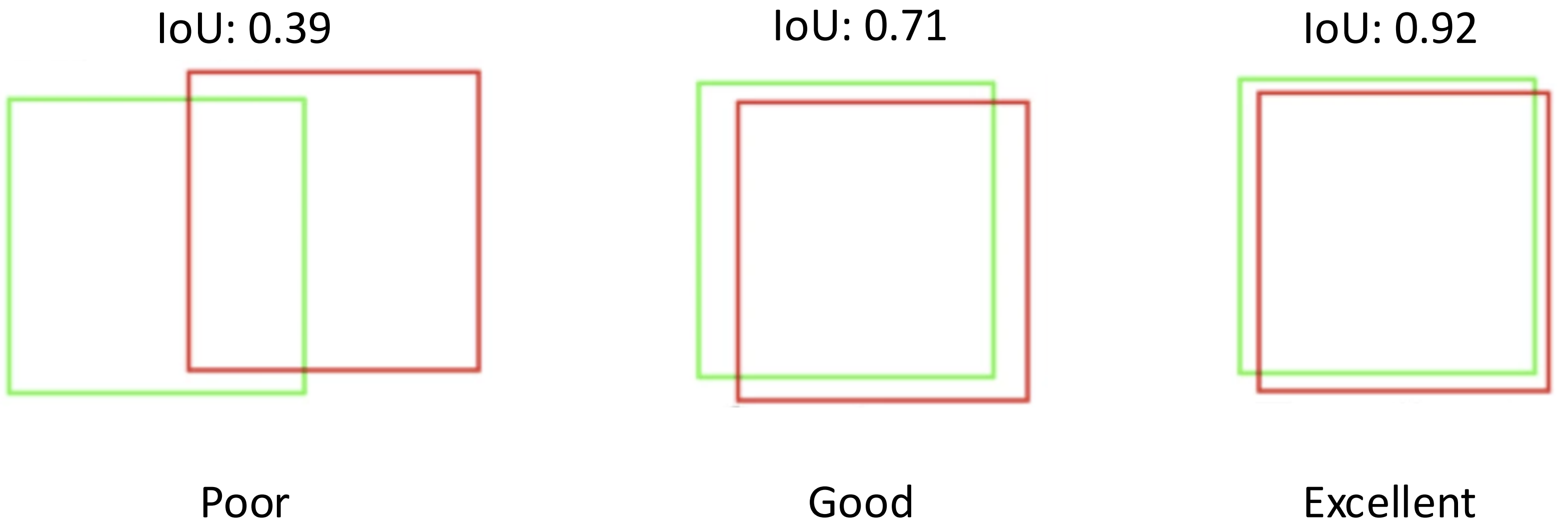
Evaluating Object Detectors: Intersection over Union

- When evaluating both **manual annotations** and **object detection** algorithms we need a metric to measure overall **quality** of the annotations or **performance** of the model
- **Intersection over Union (IoU)** measures the amount overlap between
  - Groups of annotators annotating medical images
  - The predicted bounding box and the ground truth bounding box



- The better the overlap between the groups of annotators or between the predicted bounding box and ground truth bounding box the better the inter-rater agreement between annotators and better the predictions

Evaluating Object Detectors: Intersection over Union



Evaluating Object Detectors: Mean Average Precision

$$mAP = \frac{1}{|classes|} \sum_{c \in classes} \frac{\#TP(c)}{\#TP(c) + \#FP(c)}$$

- True Positive - TP(c): a predicted bounding box (pred_bb) was made for class c, there is a ground truth bounding box (gt_bb) of class c, and IoU(pred_bb, gt_bb) >= 0.5.
- False Positive - FP(c): a pred_bb was made for class c, and there is no gt_bb of class c. Or there is a gt_bb of class c, but IoU(pred_bb, gt_bb) < 0.5.

Summary: Pre-CNN vs CNN

Feature	Pre-CNN	CNN-Based
Feature Type	Hand-crafted	Learned
Processing Methods	Per-window	Full-image convolution
Speed	Very slow	Real-time
Bounding Box Generation	Manual Regression	Integrated with CNN
Scalability	Poor	Excellent

Thank you!

Questions!