CS6501: Deep Learning for Visual Recognition

Object Detection I:
RCNN, Fast-RCNN, Faster-RCNN
Today’s Class

• Object Detection
• The RCNN Object Detector (2014)
• The Fast RCNN Object Detector (2015)
• The Faster RCNN Object Detector (2016)
Object Detection
Object Detection as Classification

CNN

deer?
cat?
background?
Object Detection as Classification

- deer?
- cat?
- background?
Object Detection as Classification

cat?
deer?
background?
Object Detection as Classification with Sliding Window

CNN

deer?
cat?
background?
Object Detection as Classification with Box Proposals
Rich feature hierarchies for accurate object detection and semantic segmentation.
Girshick et al. CVPR 2014.
First stage: generate category-independent region proposals.

- 2000 Region proposals for every image

Selective Search: combine the strength of both an exhaustive search and segmentation. Uijlings et al. IJCV 2013. ref
**RCNN**

First stage: generate category-independent region proposals.
- 2000 Region proposals for every image

Second stage: extracts a fixed-length feature vector from each region.
- A 4096-dimensional feature vector from each region proposal

Arbitrary rectangles? A fixed size input? 227 x 227

5 conv layers + 2 fully connected layers
RCNN

First stage: generate category-independent region proposals.
- 2000 Region proposals for every image

Second stage: extracts a fixed-length feature vector from each region.
- a 4096-dimensional feature vector from each region proposal

Third stage: a set of class-specific linear SVMs.
- object category and location
• Simple and scalable.
• improves mAP.

• A multistage pipeline.
• Training is expensive in space and time (features are extracted from each region proposal in each image and written into disk).
• Object detection is slow.

?
Fast-RCNN

Idea: No need to recompute features for every box independently

https://arxiv.org/abs/1504.08083
Fast R-CNN. Girshick. ICCV 2015.
Fast-RCNN

- Process the whole image with several convolutional (conv) and max pooling layers to produce a conv feature map.

- A region of interest (RoI) pooling layer extracts a fixed-length feature vector from the region feature map.

- For each RoI:
  - FC+ softmax
  - K + 1 categories
  - FC+ regressor
  - Four real-valued numbers for each of the K object classes.
**RCNN**
- Simple and scalable.
- Improves mAP.
- A multistage pipeline.
- Training is expensive in space and time (features are extracted from each region proposal in each image and written into disk).
- Object detection is slow.

**Fast-RCNN**
- Higher mAP.
- Single stage, end-to-end training.
- No disk storage is required for feature caching.
- Proposals are the computational bottleneck in detection systems.

**Faster-RCNN**

?
Faster-RCNN

Idea: Integrate the Bounding Box Proposals as part of the CNN predictions

https://arxiv.org/abs/1506.01497
Ren et al. NIPS 2015.
Faster-RCNN

Region Proposal Networks:

- Sliding window, nxn
- Conv layer
- 1x1 conv layer
- cls layer
- reg layer
- nxn conv layer
- 2k scores: object or not object
- 4k coordinates: bounding box proposal
- k anchors boxes
- Shared conv layers
- RPN
- Classification loss
- Bounding-box regression loss
- Feature map
- Pre-train ImageNet
- Fast-RCNN
RCNN

• Simple and scalable.
• Improves mAP.
• A multistage pipeline.
• Training is expensive in space and time (features are extracted from each region proposal in each image and written into disk).
• Object detection is slow.

Fast-RCNN

• Higher mAP.
• Single stage, end-to-end training.
• No disk storage is required for feature caching.
• Proposals are the computational bottleneck in detection systems.

Faster-RCNN

• Compute proposals with a deep convolutional neural network -- Region Proposal Network (RPN)
• Merge RPN and Fast R-CNN into a single network, enabling nearly cost-free region proposals.
YOLO- You Only Look Once

Idea: No bounding box proposal. A single regression problem, straight from image pixels to bounding box coordinates and class probabilities.

- extremely fast
- reason globally
- learn generalizable representations

https://arxiv.org/abs/1506.02640
Redmon et al. CVPR 2016.
YOLO- You Only Look Once

Divide the image into 7x7 cells.
Each cell trains a detector.
The detector needs to predict the object’s class distributions.
The detector has 2 bounding-box predictors to predict bounding-boxes and confidence scores.
Idea: Similar to YOLO, but denser grid map, multiscale grid maps. + Data augmentation + Hard negative mining + Other design choices in the network.

Liu et al. ECCV 2016.
Questions?