Object Detection

Sanjiban Choudhury
What is an object? Why should robots detect them?
What about more complex scenes like a real kitchen?
Activity!
Last Lecture: Image Classification

(assume given a set of possible labels)
{dog, cat, truck, plane, ...}

---

This image by Nikita is licensed under CC-BY 2.0

Slides from Stanford CS231N: Object Detection and Image Segmentation
Think-Pair-Share!

Think (30 sec): How do we extend our image classifiers to classify objects in an image? What are some of the challenges?

Pair: Find a partner

Share (45 sec): Partners exchange ideas
Increasing complexity of computer vision tasks

Slides from Stanford CS231N: Object Detection and Image Segmentation
Increasing complexity of computer vision tasks

Classification

CAT

No spatial extent

Slides from Stanford CS231N: Object Detection and Image Segmentation
Increasing complexity of computer vision tasks

Classification

Semantic Segmentation

CAT

No spatial extent

GRASS, CAT, TREE, SKY

No objects, just pixels

Slides from Stanford CS231N: Object Detection and Image Segmentation
Increasing complexity of computer vision tasks

Classification

Semantic Segmentation

Object Detection

CAT

GRASS, CAT, TREE, SKY

DOG, DOG, CAT

No spatial extent

No objects, just pixels

Multiple Object

Slides from Stanford CS231N: Object Detection and Image Segmentation
Increasing complexity of computer vision tasks

**Classification**
- CAT
- No spatial extent

**Semantic Segmentation**
- GRASS, CAT, TREE, SKY
- No objects, just pixels

**Object Detection**
- DOG, DOG, CAT
- Multiple Object

**Instance Segmentation**
- DOG, DOG, CAT

Slides from Stanford CS231N: Object Detection and Image Segmentation
Increasing complexity of computer vision tasks

Classification

No spatial extent

Semantic Segmentation

GRASS, CAT, TREE, SKY

No objects, just pixels

Object Detection

Multiple Object

Instance Segmentation

DOG, DOG, CAT

Slides from Stanford CS231N: Object Detection and Image Segmentation
Semantic Segmentation: The Problem

**GRASS, CAT, TREE, SKY, ...**

Paired training data: for each training image, each pixel is labeled with a semantic category.
Semantic Segmentation: The Problem

Paired training data: for each training image, each pixel is labeled with a semantic category.

At test time, classify each pixel of a new image.
Semantic Segmentation Idea: Sliding Window

Can you classify this pixel?
Semantic Segmentation Idea: Sliding Window

Can you classify this pixel?

Pretty hard without context!
Semantic Segmentation Idea: Sliding Window

Full image

Extract patch
Semantic Segmentation Idea: Sliding Window

Classify each patch!

Pinheiro and Collobert, “Recurrent Convolutional Neural Networks for Scene Labeling”, ICML 2014
Semantic Segmentation Idea: Sliding Window

Problem: Very inefficient! Not reusing shared features between overlapping patches

Pinheiro and Collobert, “Recurrent Convolutional Neural Networks for Scene Labeling”, ICML 2014
Semantic Segmentation Idea: Convolution

Full image
Semantic Segmentation Idea: Convolution

Full image

An intuitive idea: encode the entire image with conv net, and do semantic segmentation on top.
Semantic Segmentation Idea: Convolution

An intuitive idea: encode the entire image with conv net, and do semantic segmentation on top.

Problem: classification architectures often reduce feature spatial sizes to go deeper, but semantic segmentation requires the output size to be the same as input size.
Semantic Segmentation Idea: Fully Convolutional

Design network as a bunch of convolutional layers, with **downsampling** and **upsampling** inside the network!

Input: 3 x H x W

High-res: $D_1 \times H/2 \times W/2$

Low-res: $D_3 \times H/4 \times W/4$

Med-res: $D_2 \times H/4 \times W/4$

Med-res: $D_2 \times H/4 \times W/4$

High-res: $D_1 \times H/2 \times W/2$

Predictions: $H \times W$

---


Semantic Segmentation: Summary

Slides from Stanford CS231N: Object Detection and Image Segmentation
Semantic Segmentation

Label each pixel in the image with a category label

Don’t differentiate instances, only care about pixels
Increasing complexity of computer vision tasks

Classification

Semantic Segmentation

Object Detection

Instance Segmentation

- No spatial extent
- No objects, just pixels
- Multiple Object

Slides from Stanford CS231N: Object Detection and Image Segmentation
Increasing complexity of computer vision tasks

Classification

Semantic Segmentation

Object Detection

Instance Segmentation

CAT

GRASS, CAT, TREE, SKY

DOG, DOG, CAT

DOG, DOG, CAT

No spatial extent

No objects, just pixels

Multiple Object

Slides from Stanford CS231N: Object Detection and Image Segmentation
Object Detection: Single Object
(Classification + Localization)
Object Detection: Single Object
(Classification + Localization)

- Class Scores
  - Cat: 0.9
  - Dog: 0.05
  - Car: 0.01
  - ...  

- Fully Connected: 4096 to 1000

- Vector: 4096

- Box Coordinates: $(x, y, w, h)$

This image is in CC0 public domain.

Slides from Stanford CS231N: Object Detection and Image Segmentation
Object Detection: Single Object
(Classification + Localization)

Correct label: Cat

Class Scores
Cat: 0.9
Dog: 0.05
Car: 0.01
...

Fully Connected: 4096 to 1000

Softmax Loss

Vector: 4096

Box Coordinates (x, y, w, h)

L2 Loss

Fully Connected: 4096 to 4

Correct box: (x’, y’, w’, h’)

Treat localization as a regression problem!

Slides from Stanford CS231N: Object Detection and Image Segmentation
What about multiple objects? Would this idea work?
Object Detection: Multiple Objects

CAT: (x, y, w, h)

DOG: (x, y, w, h)
DOG: (x, y, w, h)
CAT: (x, y, w, h)

DUCK: (x, y, w, h)
DUCK: (x, y, w, h)

...
Object Detection: Multiple Objects

Each image needs a different number of outputs!

CAT: (x, y, w, h)  
4 numbers

DOG: (x, y, w, h)  
DOG: (x, y, w, h)  
12 numbers

DUCK: (x, y, w, h)  
DUCK: (x, y, w, h)  
Many numbers!

....

Slides from Stanford CS231N: Object Detection and Image Segmentation
What if we tried to detect a SINGLE object in a PATCH?
Object Detection: Multiple Objects

Apply a CNN to many different crops of the image, CNN classifies each crop as object or background

Dog? NO
Cat? NO
Background? YES

Slides from Stanford CS231N: Object Detection and Image Segmentation
Object Detection: Multiple Objects

Apply a CNN to many different crops of the image, CNN classifies each crop as object or background.

Dog? YES
Cat? NO
Background? NO
Object Detection: Multiple Objects

Apply a CNN to many different crops of the image, CNN classifies each crop as object or background

Dog? YES
Cat? NO
Background? NO
Object Detection: Multiple Objects

Apply a CNN to many different crops of the image, CNN classifies each crop as object or background

Dog? NO
Cat? YES
Background? NO

Q: What's the problem with this approach?
Object Detection: Multiple Objects

Apply a CNN to many different crops of the image, CNN classifies each crop as object or background.

Problem: Need to apply CNN to huge number of locations, scales, and aspect ratios, very computationally expensive!

Dog? NO
Cat? YES
Background? NO
What if we had a SMART path proposer?
Region Proposals: Selective Search

- Find “blobby” image regions that are likely to contain objects
- Relatively fast to run; e.g. Selective Search gives 2000 region proposals in a few seconds on CPU

Alexe et al., "Measuring the objectness of image windows", TPAMI 2012
Uijlings et al., "Selective Search for Object Recognition", IJCV 2013
Cheng et al., "BING: Binarized normed gradients for objectness estimation at 300fps", CVPR 2014
Zitnick and Dollar, "Edge boxes: Locating object proposals from edges", ECCV 2014
R-CNN

Figure copyright Ross Girshick, 2015; source. Reproduced with permission.
R-CNN

Regions of Interest (RoI) from a proposal method (~2k)

R-CNN

Warped image regions (224x224 pixels)

Regions of Interest (RoI) from a proposal method (~2k)

Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Slides from Stanford CS231N: Object Detection and Image Segmentation
R-CNN

Forward each region through ConvNet (ImageNet-pretrained)

Warped image regions (224x224 pixels)

Regions of Interest (RoI) from a proposal method (~2k)

Input image

R-CNN

- **Input image**
- **Regions of Interest (RoI)** from a proposal method (~2k)
- **Warped image regions** (224x224 pixels)
- **Forward each region through ConvNet** (ImageNet-pretrained)
- **Classify regions with SVMs**

Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Slides from Stanford CS231N: Object Detection and Image Segmentation
R-CNN

Predict “corrections” to the RoI: 4 numbers: (dx, dy, dw, dh)

Classify regions with SVMs

Forward each region through ConvNet (ImageNet-pretrained)

Warped image regions (224x224 pixels)

Regions of Interest (RoI) from a proposal method (~2k)

Input image

Figure copyright Ross Girshick, 2015; source. Reproduced with permission.
Isn’t calling a CNN for each patch super duper slow?
Predict “corrections” to the RoI: 4 numbers: (dx, dy, dw, dh)

**Problem:** Very slow! Need to do ~2k independent forward passes for each image!

- **Bbox reg**
- **SVMs**

Classify regions with SVMs

Forward each region through ConvNet

Warped image regions (224x224 pixels)

Regions of Interest (RoI) from a proposal method (~2k)

Input image


Slides from Stanford CS231N: Object Detection and Image Segmentation
“Slow” R-CNN

Predict “corrections” to the RoI: 4 numbers: (dx, dy, dw, dh)

Classify regions with SVMs

Forward each region through ConvNet

Warped image regions (224x224 pixels)

Regions of Interest (RoI) from a proposal method (~2k)

Input image

Problem: Very slow! Need to do ~2k independent forward passes for each image!

Idea: Pass the image through convnet before cropping! Crop the conv feature instead!


Slides from Stanford CS231N: Object Detection and Image Segmentation
Instead of running $N$ ConvNets, run just ONE!
Fast R-CNN

“Slow” R-CNN

Input image

Girshick, “Fast R-CNN”, ICCV 2015. Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Slides from Stanford CS231N: Object Detection and Image Segmentation
Fast R-CNN

“Backbone” network: AlexNet, VGG, ResNet, etc

“conv5” features
Run whole image through ConvNet

ConvNet
Input image

“Slow” R-CNN

Conv Net
Conv Net
Conv Net

SVMs
SVMs
SVMs

Input image

Girshick, “Fast R-CNN”, ICCV 2015. Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Slides from Stanford CS231N: Object Detection and Image Segmentation
Fast R-CNN

Regions of Interest (RoIs) from a proposal method

“Backbone” network: AlexNet, VGG, ResNet, etc

“conv5” features

Run whole image through ConvNet

Input image

“Slow” R-CNN

ConvNet

Conv Net

SVMs

SVMs

Input image

Girshick, “Fast R-CNN”, ICCV 2015. Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Slides from Stanford CS231N: Object Detection and Image Segmentation
Fast R-CNN

Regions of Interest (RoIs) from a proposal method

“Backbone” network: AlexNet, VGG, ResNet, etc

Crop + Resize features

“conv5” features

Run whole image through ConvNet

Input image

“Slow” R-CNN

Girshick, “Fast R-CNN”, ICCV 2015. Figure copyright Ross Girshick, 2015; source, Reproduced with permission.

Slides from Stanford CS231N: Object Detection and Image Segmentation
Fast R-CNN

**Object category**

Regions of Interest (RoIs) from a proposal method

“Backbone” network: AlexNet, VGG, ResNet, etc

**CNN**

Per-Region Network

Crop + Resize features

“conv5” features

Run whole image through ConvNet

Linear + softmax

Linear

Box offset

ConvNet

Input image

“Slow” R-CNN

SVMs

SVMs

Conv Net

Conv Net

Conv Net

Input image

Slides from Stanford CS231N: Object Detection and Image Segmentation
Fast R-CNN

Object category

Regions of Interest (Rois) from a proposal method

“Backbone” network: AlexNet, VGG, ResNet, etc

Linear + softmax

Linear

Box offset

Per-Region Network

Crop + Resize features

“conv5” features

Run whole image through ConvNet

Input image

“Slow” R-CNN

Girshick, “Fast R-CNN”, ICCV 2015. Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Slides from Stanford CS231N: Object Detection and Image Segmentation
R-CNN vs Fast R-CNN

Training time (Hours)

- R-CNN: 84 hours
- SPP-Net: 25.5 hours
- Fast R-CNN: 8.75 hours

Test time (seconds)

- Including Region proposals:
  - R-CNN: 49 seconds
  - SPP-Net: 4.3 seconds
  - Fast R-CNN: 2.3 seconds

- Excluding Region proposals:
  - R-CNN: 47 seconds
  - SPP-Net: 2.3 seconds
  - Fast R-CNN: 0.32 seconds

Problem: Runtime dominated by region proposals!

He et al., “Spatial pyramid pooling in deep convolutional networks for visual recognition”, ECCV 2014
Can we get rid of the hacky region proposal algorithm?
Learn region proposal in an end to end manner!
Faster R-CNN:
Make CNN do proposals!

Insert Region Proposal Network (RPN) to predict proposals from features

Otherwise same as Fast R-CNN:
Crop features for each proposal, classify each one

Figure copyright 2015, Ross Girshick; reproduced with permission

Slides from Stanford CS231N: Object Detection and Image Segmentation
Faster R-CNN:
Make CNN do proposals!

Jointly train with 4 losses:
1. RPN classify object / not object
2. RPN regress box coordinates
3. Final classification score (object classes)
4. Final box coordinates

Figure copyright 2015, Ross Girshick; reproduced with permission

Slides from Stanford CS231N: Object Detection and Image Segmentation
Faster R-CNN:
Make CNN do proposals!

R-CNN Test-Time Speed

- R-CNN: 49
- SPP-Net: 4.3
- Fast R-CNN: 2.3
- Faster R-CNN: 0.2

Slides from Stanford CS231N: Object Detection and Image Segmentation
Instance Segmentation

Classification

Semantic Segmentation

Object Detection

Instance Segmentation

CAT

GRASS, CAT, TREE, SKY

DOG, DOG, CAT

DOG, DOG, CAT

No spatial extent

No objects, just pixels

Multiple Object

Slides from Stanford CS231N: Object Detection and Image Segmentation
Object Detection: Faster R-CNN

Slides from Stanford CS231N: Object Detection and Image Segmentation
Instance Segmentation: Mask R-CNN

Object Detection

Instance Segmentation

Classification loss

Bounding-box regression loss

proposals

Region Proposal Network

feature map

CNN

image

Classification loss

Bounding-box regression loss

Mask Prediction

Add a small mask network that operates on each RoI and predicts a 28x28 binary mask

He et al, "Mask R-CNN", ICCV 2017

Slides from Stanford CS231N: Object Detection and Image Segmentation
Mask R-CNN

He et al., “Mask R-CNN”, arXiv 2017

Slides from Stanford CS231N: Object Detection and Image Segmentation
Mask R-CNN: Example Mask Training Targets
Mask R-CNN: Example Mask Training Targets
Mask R-CNN: Example Mask Training Targets
Mask R-CNN: Example Mask Training Targets
Mask R-CNN: Very Good Results!

He et al, "Mask R-CNN", ICCV 2017
Mask R-CNN
Also does pose

He et al, "Mask R-CNN", arXiv 2017

Slides from Stanford CS231N: Object Detection and Image Segmentation
Mask R-CNN
Also does pose

He et al, "Mask R-CNN", ICCV 2017
Is 2D instance segmentation enough for robots?

No!
Object Detection + Captioning = Dense Captioning

Figure copyright IEEE, 2016. Reproduced for educational purposes.
Dense Video Captioning

Ranjay Krishna et al., “Dense-Captioning Events in Videos”, ICCV 2017
Figure copyright IEEE, 2017. Reproduced with permission.
Scene Graph Prediction

Xu, Zhu, Choy, and Fei-Fei, “Scene Graph Generation by Iterative Message Passing”, CVPR 2017
Figure copyright IEEE, 2018. Reproduced for educational purposes.
3D Object Detection: Monocular Camera

Candidate sampling in 3D space

projection

Faster R-CNN

Scoring & NMS

Proposals

2D candidate boxes
- Same idea as Faster RCNN, but proposals are in 3D
- 3D bounding box proposal, regress 3D box parameters + class score


Slides from Stanford CS231N: Object Detection and Image Segmentation
tl;dr

Increasing complexity of computer vision tasks

Classification
- CAT
- No spatial extent

Semantic Segmentation
- Grass, Cat, Tree, Sky
- No objects, just pixels

Object Detection
- Dog, Dog, Cat
- Multiple Object

Instance Segmentation
- Dog, Dog, Cat
- Multiple Object

Region Proposals: Selective Search
- Find "blobby" image regions that are likely to contain objects
- Relatively fast to run; e.g., Selective Search gives 2000 region proposals in a few seconds on CPU

Mask R-CNN
- Classification Scores: C
  - Box coordinates (per class): 4 * C
- CNN + RPN
- Conv
- Conv
- Roll Align
- 256 x 14 x 14
- 256 x 14 x 14
- Predict a mask for each of C classes
- C x 26 x 28