

CSC 144V: Autonomous Vehicles

Lecture 5

Object Detection and Segmentation Exercises Solution

Department of Computer Science,
Hofstra University

Exercise: IoU and NMS

Four detections of two pedestrians, boxes (x_1, y_1, x_2, y_2) in pixels. $\text{IoU} = |A \cap B| / |A \cup B|$.

(a) Compute the IoU of each box with the top-scoring box A. (b) Run greedy NMS: keep the best box, suppress boxes with $\text{IoU} > \text{threshold}$ with a kept box. Threshold 0.5. (c) Which boxes remain with threshold 0.4?

Box	(x_1, y_1, x_2, y_2)	Score	IoU with A	NMS, thr. 0.5
A	(10, 10, 60, 60)	0.9		
B	(20, 20, 70, 70)	0.8		
C	(12, 12, 62, 62)	0.7		
D	(100, 100, 150, 160)	0.6		

Exercise Solution: IoU and NMS

- IoU(A, B): intersection $40 \times 40 = 1,600$, union $2,500 + 2,500 - 1,600 = 3,400 \rightarrow 0.47$. IoU(A, C): $48 \times 48 = 2,304$, union $2,696 \rightarrow 0.85$. IoU(A, D) = 0.
- Threshold 0.5: keep A; suppress C (0.85); B survives ($0.47 \leq 0.5$); D survives $\rightarrow$ A, B, D.
- Threshold 0.4: B is also suppressed ($0.47 > 0.4$) $\rightarrow$ A, D. A low threshold risks removing a second object that overlaps the first.

Box	(x_1, y_1, x_2, y_2)	Score	IoU with A	NMS, thr. 0.5
A	(10, 10, 60, 60)	0.9	1	keep
B	(20, 20, 70, 70)	0.8	0.47	keep
C	(12, 12, 62, 62)	0.7	0.85	suppress
D	(100, 100, 150, 160)	0.6	0	keep

Exercise: Precision, recall and AP

There are **5** ground-truth objects. The detector returns 6 detections sorted by confidence: TP, TP, FP, TP, FP, TP.

(a) Precision and recall after each rank.

(b) Average precision using the all-point interpolated envelope: precision at recall r is the maximum precision at any recall $\geq r$.

Rank	TP / FP	Precision	Recall
1	TP		
2	TP		
3	FP		
4	TP		
5	FP		
6	TP		

Exercise Solution: Precision, recall and AP

- Precision = TP/rank; recall = TP/5.
- Envelope: recall 0.2 $\rightarrow$ 1, 0.4 $\rightarrow$ 1, 0.6 $\rightarrow$ 0.75, 0.8 $\rightarrow$ 0.667, 1.0 $\rightarrow$ 0 (never reached).
- AP = $0.2 \times (1 + 1 + 0.75 + 0.667 + 0) \approx \mathbf{0.683}$.

Rank	TP / FP	Precision	Recall
1	TP	1.00	0.2
2	TP	1.00	0.4
3	FP	0.67	0.4
4	TP	0.75	0.6
5	FP	0.60	0.6
6	TP	0.67	0.8

Exercise: YOLO label of a grid cell

A **300 × 300** image is divided into a **3 × 3** grid; classes: Pedestrian, Car, Motorcycle. A car has its mid-point at pixel (170, 240); its box is **150 px wide** and **60 px high**.

- (a) Which grid cell (row, col from 0, top-left) is responsible?
- (b) Fill in the label $y = (pc, bx, by, bh, bw, c_1, c_2, c_3)$ of that cell.
- (c) What is the shape of the whole output tensor?

pc	bx	by	bh	bw	c ₁	c ₂	c ₃

Exercise Solution: YOLO label of a grid cell

- (a) Cell size 100 px: $\text{col} = \lfloor 170/100 \rfloor = 1$, $\text{row} = \lfloor 240/100 \rfloor = 2 \rightarrow$ **cell (2, 1)**.
- (b) $\text{bx} = (170 - 100)/100 = 0.7$, $\text{by} = (240 - 200)/100 = 0.4$, $\text{bh} = 60/100 = 0.6$, $\text{bw} = 150/100 = 1.5$ (can exceed 1). Car is class 2.
- (c) 8 numbers per cell: **$3 \times 3 \times 8$** .

pc	bx	by	bh	bw	c ₁	c ₂	c ₃
1	0.7	0.4	0.6	1.5	0	1	0

Exercise: Realistic YOLO dimensions

A realistic YOLO takes a $608 \times 608 \times 3$ image and outputs a $19 \times 19 \times 5 \times 85$ tensor.

- (a) What do 5 and 85 stand for?
- (b) How many numbers are predicted, and how many candidate boxes enter NMS?
- (c) What is the stride (pixels per grid cell)? In which cell (row, col from 0) does an object with mid-point at pixel (330, 200) fall, and what are (bx, by) of its label?

Exercise Solution: Realistic YOLO dimensions

- (a) **5 anchor boxes** per cell; $85 = 5 + 80$: (pc, bx, by, bh, bw) plus **80 class scores**.
- (b) $19 \times 19 \times 5 \times 85 = \mathbf{153,425}$ numbers; $19 \times 19 \times 5 = \mathbf{1,805}$ candidate boxes before NMS.
- (c) Stride $608/19 = \mathbf{32 \text{ px}}$. col = $\lfloor 330/32 \rfloor = 10$, row = $\lfloor 200/32 \rfloor = 6 \rightarrow \mathbf{\text{cell (6, 10)}}$; bx = $(330 - 320)/32 = \mathbf{0.31}$, by = $(200 - 192)/32 = \mathbf{0.25}$.

Exercise: Upsampling

Upsample the 2×2 feature map to 4×4 with a stride of 2.

(a) **Bed of nails**: insert zeros; each input goes to the top-left corner of its 2×2 block.

(b) **Nearest neighbor**: duplicate each element.

(c) Name two other upsampling methods used in FCNs.

Input (2×2)

1	2
3	4

Bed of nails (4×4)

Nearest neighbor (4×4)

Exercise Solution: Upsampling

- (a) Bed of nails keeps the values at fixed positions and fills the rest with 0.
- (b) Nearest neighbor replicates each value over its 2×2 block.
- (c) Max unpooling (restores each value to the position of the max in the matching pooling layer), bilinear / bicubic interpolation, transposed convolution (learnable).

Input (2×2)

1	2
3	4

Bed of nails (4×4)

1	0	2	0
0	0	0	0
3	0	4	0
0	0	0	0

Nearest neighbor (4×4)

1	1	2	2
1	1	2	2
3	3	4	4
3	3	4	4

Exercise: Transposed convolution

Transposed convolution (stride 1, no padding): each input element multiplies the whole kernel, the products are placed at its position, and overlapping outputs add up.

(a) Compute the 3×3 output.

(b) Output size formula for input N, kernel K, stride S?

(c) With L stacked 3×3 CONV layers (stride 1) the receptive field is $2L + 1$. For $L = 10$? How many layers for a receptive field ≥ 224 ? How does downsampling help?

Input

1	2
3	4

Kernel

1	2
0	1

Output (3×3)

Exercise Solution: Transposed convolution

- (a) e.g. $O[1][1] = 1 \cdot 1 + 2 \cdot 0 + 3 \cdot 2 + 4 \cdot 1 = 11$ (four overlapping contributions).
- (b) $(N - 1) \cdot S + K = (2 - 1) \cdot 1 + 2 = 3$.
- (c) $L = 10 \rightarrow 21 \times 21$. $2L + 1 \geq 224 \rightarrow L \geq 111.5 \rightarrow 112$ layers: too slow. Pooling / strided convolution makes the receptive field grow much faster and shrinks the cost; upsampling then restores the output size.

Input

1	2
3	4

Kernel

1	2
0	1

Output (3 × 3)

1	4	4
3	11	10
0	3	4