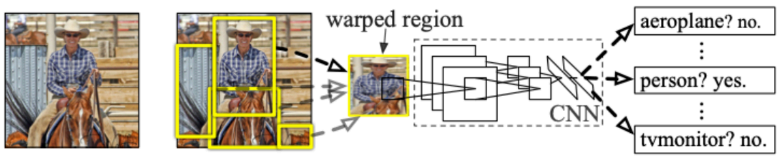
Rich feature hierarchies for accurate object detection and semantic segmentation

Boulder

YouTube Playlist

R-CNN: *Regions with CNN features*



1. Input image | **2.** Extract region proposals (~2k) | **3.** Compute CNN features | **4.** Classify regions

Region proposals: selective search

Feature extraction: 4096-dimensional feature vector from AlexNet

Warped training samples

Classify regions: class-specific linear SVM

Greedy Non-maximum Suppression

```
For each class independently reject a region if it has an
intersection-over-union (IoU) overlap with a higher scoring
selected region larger than a learned threshold.
```

Domain-specific fine-tuning

21-way classification layer

```
Treat all region proposals with ≥ 0.5 IoU overlap with a ground-truth
box as positives for that box's class and the rest as negatives.
```

Object category classifiers

IoU $< 0.3 \rightarrow$ negative examples

ground-truth bounding boxes $\rightarrow$ positive examples

VOC 2010 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
DPM v5 [17]†	49.2	53.8	13.1	15.3	35.5	53.4	49.7	27.0	17.2	28.8	14.7	17.8	46.4	51.2	47.7	10.8	34.2	20.7	43.8	38.3	33.4
UVA [32]	56.2	42.4	15.3	12.6	21.8	49.3	36.8	46.1	12.9	32.1	30.0	36.5	43.5	52.9	32.9	15.3	41.1	31.8	47.0	44.8	35.1
Regionlets [35]	65.0	48.9	25.9	24.6	24.5	56.1	54.5	51.2	17.0	28.9	30.2	35.8	40.2	55.7	43.5	14.3	43.9	32.6	54.0	45.9	39.7
SegDPM [15]†	61.4	53.4	25.6	25.2	35.5	51.7	50.6	50.8	19.3	33.8	26.8	40.4	48.3	54.4	47.1	14.8	38.7	35.0	52.8	43.1	40.4
R-CNN	67.1	64.1	46.7	32.0	30.5	56.4	57.2	65.9	27.0	47.3	40.9	66.6	57.8	65.9	53.6	26.7	56.5	38.1	52.8	50.2	50.2
R-CNN BB	**71.8**	**65.8**	**53.0**	**36.8**	**35.9**	**59.7**	**60.0**	**69.9**	**27.9**	**50.6**	**41.4**	**70.0**	**62.0**	**69.0**	**58.1**	**29.5**	**59.4**	**39.3**	**61.2**	**52.4**	**53.7**

Visualizing learned features

```
The idea is to single out a particular unit (feature) in the network and use
it as if it were an object detector in its own right.
```

Bounding-box regression

$G = (G_x, G_y, G_w, G_h) \rightarrow$ ground-truth bounding box

$P = (\underbrace{P_x, P_y,}_{\text{center}} \quad \underbrace{P_w, P_h}_{\text{width \& height}}) \rightarrow$ proposal bounding box

$$\mathbf{w}_\star = \underset{\hat{\mathbf{w}}_\star}{\operatorname{argmin}} \sum_i^N (t_\star^i - \hat{\mathbf{w}}_\star^{\mathrm{T}} \phi_5(P^i))^2 + \lambda \|\hat{\mathbf{w}}_\star\|^2$$

$$\boxed{\begin{aligned} t_x &= (G_x - P_x)/P_w \\ t_y &= (G_y - P_y)/P_h \\ t_w &= \log(G_w/P_w) \\ t_h &= \log(G_h/P_h). \end{aligned}}$$

$$\boxed{\begin{aligned} \hat{G}_x &= P_w d_x(P) + P_x \\ \hat{G}_y &= P_h d_y(P) + P_y \\ \hat{G}_w &= P_w \exp(d_w(P)) \\ \hat{G}_h &= P_h \exp(d_h(P)). \end{aligned}}$$

$$d_\star(P) = \mathbf{w}_\star^{\mathrm{T}} \underbrace{\phi_5(P)}_{\text{pool}_5 \text{ features}}$$

Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.