

Rich feature hierarchies for accurate object detection
and semantic segmentation

(R-CNNの論文)

Contents

- Introduction
- R-CNN design
- Training method
- Test result
- Subjects

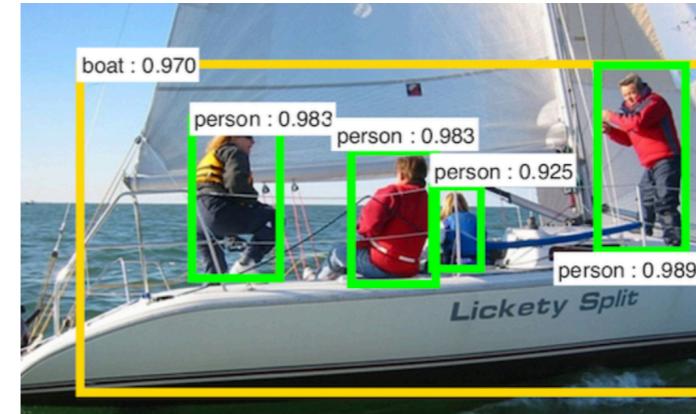
Introduction

1. この論文について

- 物体識別にCNNを用いたモデル(R-CNN)を生み出したパイオニア的論文
- PASCAL VOC, ILSVRCで既存のシステムを大きく上回る識別精度を記録した。

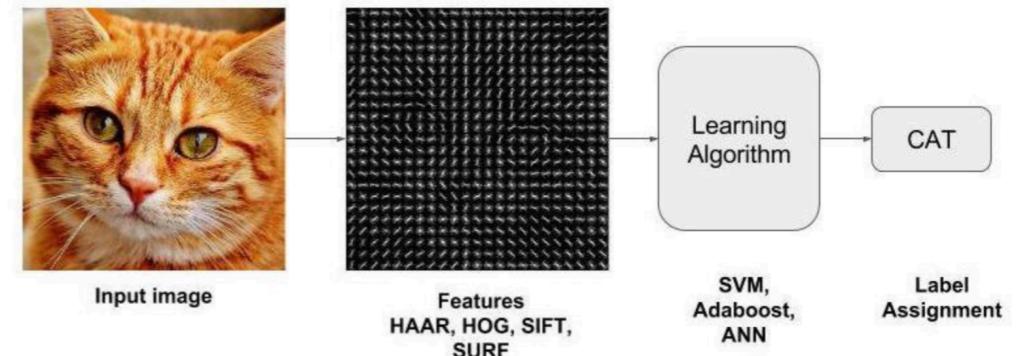
2. 物体検出・識別について

- 検出=入力に対し物体らしいものを抽出
- 識別=検出した物体らしいものを分類する



3. R-CNN以前の方法と問題

- SIFT/HOG + 線形SVMを用いたシステム
- 輝度ベクトルを特徴量とする
==> 輪郭が取り出されるイメージ
- 計算コストは低いが一定以上の精度が向上しなかった
==> PASCAL VOCにおいて2010年からわずかな識別精度の改善のみ

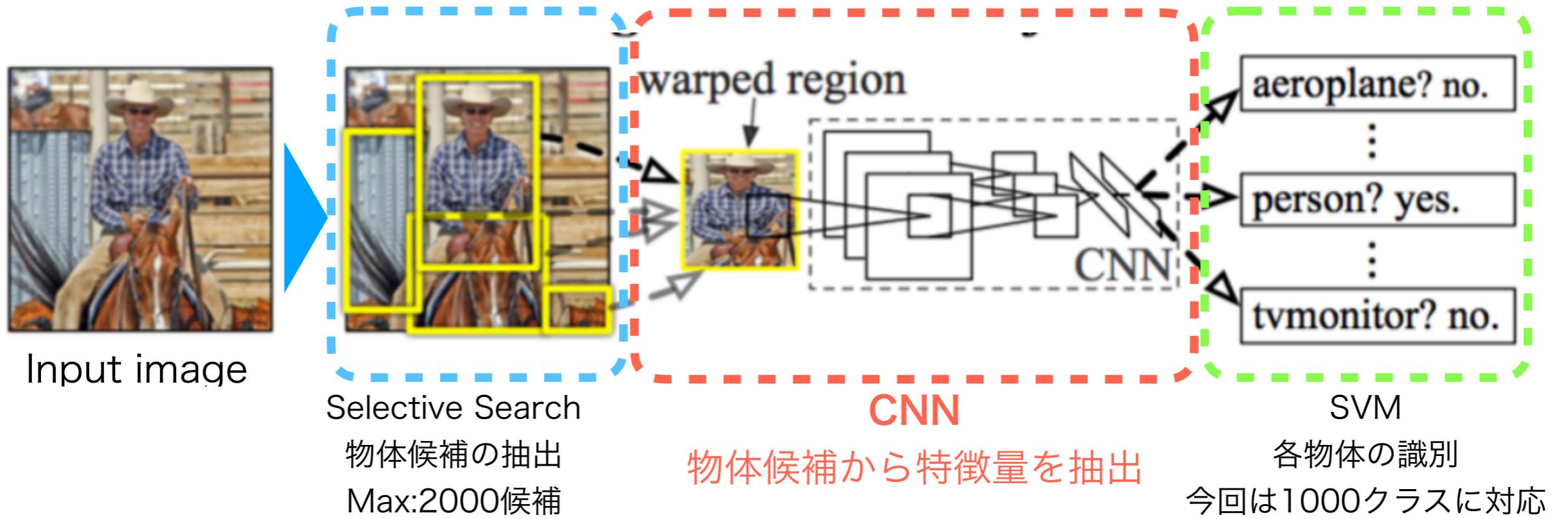


4. なぜCNNを導入したか？

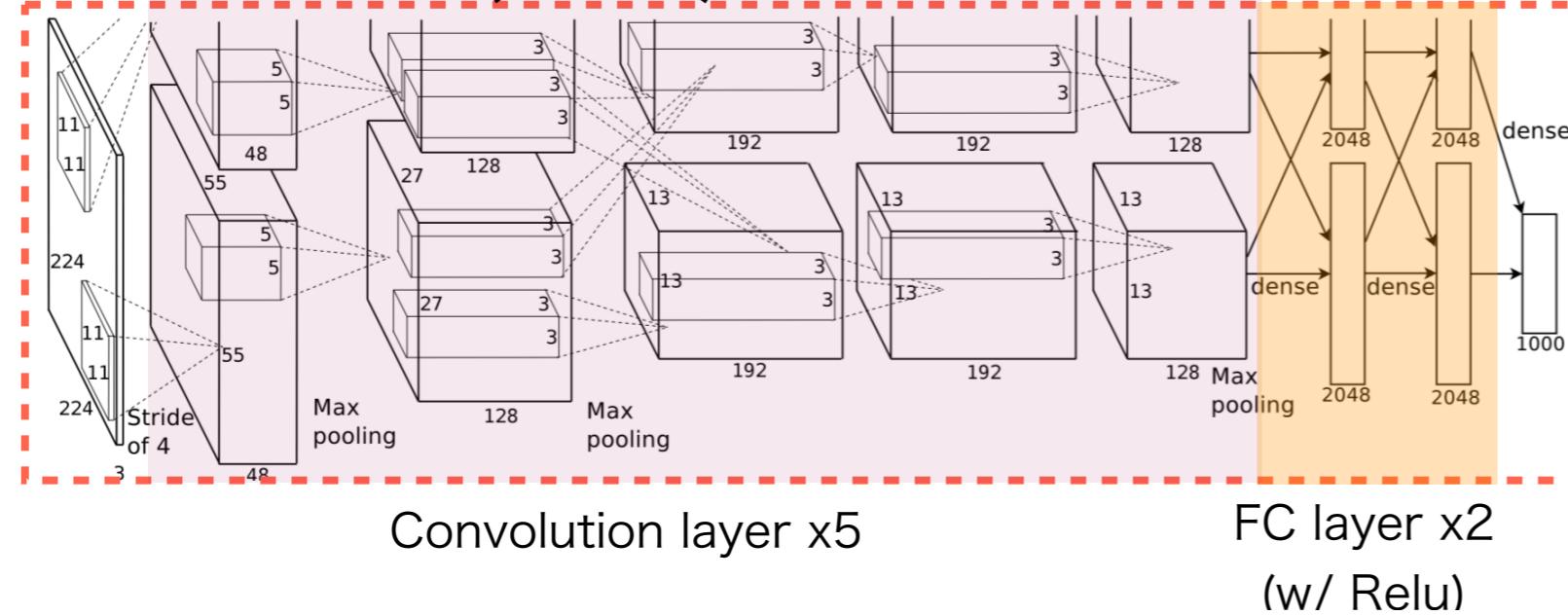
- CNNによって物体候補からより有効な特徴量の抽出を行い識別精度の改善を試みた

R-CNN Design

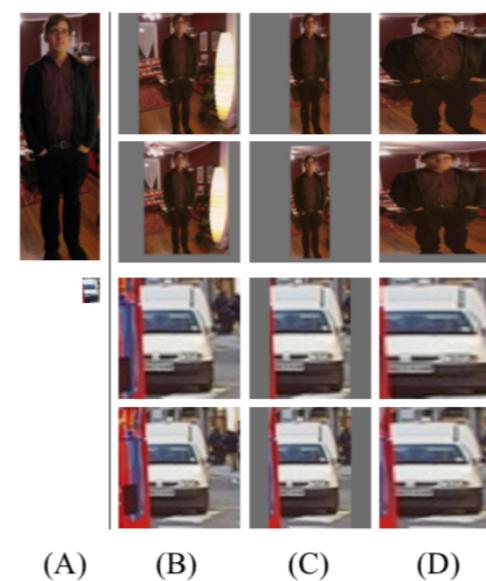
R-CNN design



CNN architecture (AlexNet)



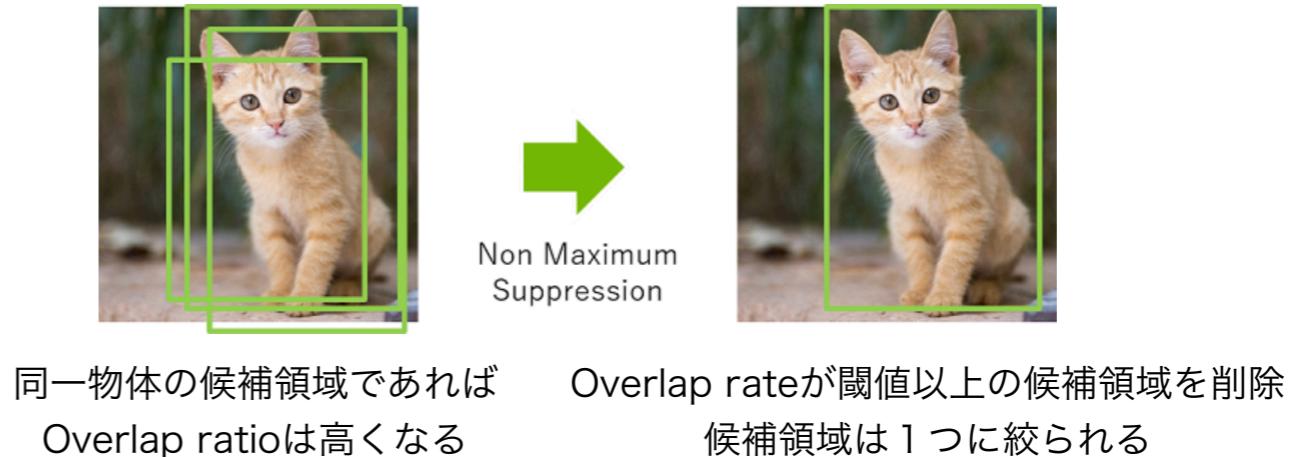
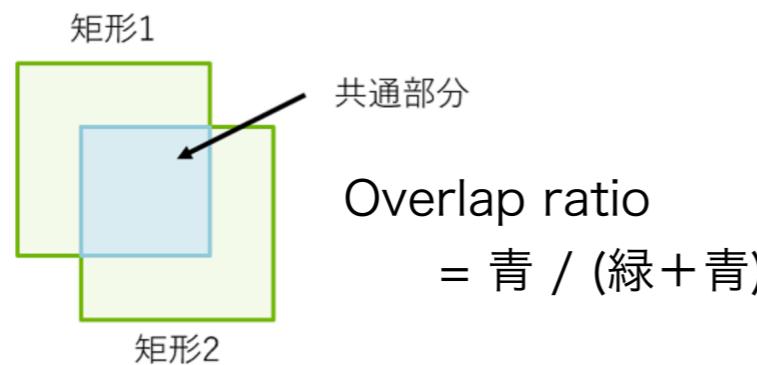
Region warping (CNN入力サイズの調整)



R-CNN Design

1. 物体候補の選択(non-maximum suppression)

- 同一の物体に対し抽出されてしまった複数の物体候補を1つに絞る
- ある物体候補領域を基準にそれ以外の物体候補領域のOverlap ratioを計算、計算結果が閾値以上である場合は基準と同一物体の候補領域と判断し削除する



2. BoundingBoxの回帰

- BoundingBoxの位置精度の向上のため、CNN特徴量を使ってBoundingBoxのパラメータ(x, y, h, w)に対し回帰を行う
- 回帰パラメータの学習は物体のクラス別に行う(物体毎にBoundingBoxの形状が大きく異なるため)

Training method

1. CNNの学習

- Pre-training (AlexNetのオリジナル論文の通りに)
 - ILSVR2012で画像分類を学習(Bounding boxは含まれない)
 - CNN Convolution layerはPre-trainingでほぼ学習が完了しているらしい
- Fine-tuning(実際のデータに対応するため)
 - PASCAL VOC、ILSVR2013より抽出された物体候補(Warped image)を使って学習
 - 学習データ : IoU > 0.5をPositive それ以外をNegativeとして学習
 - OptimizerはSGD($\text{lr}=0.001$)
 - CNN FC layerはここでより学習されるらしい

2. SVMの学習

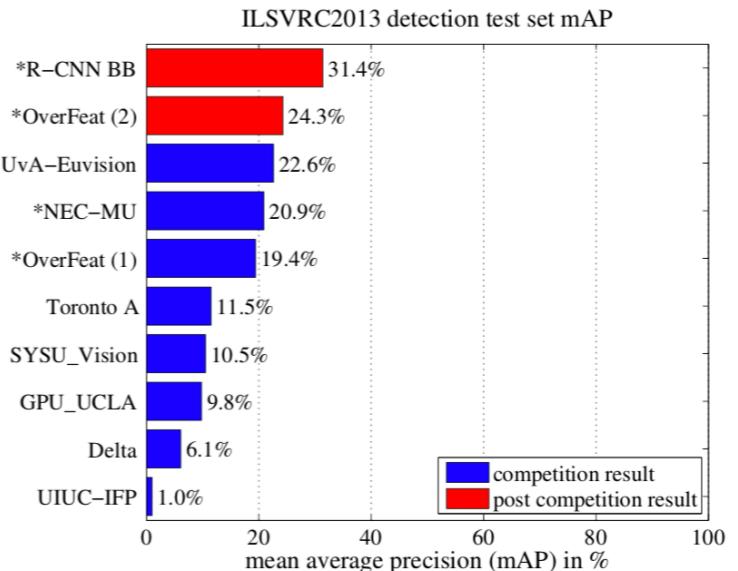
- 学習データ : IoU > 0.3をPositive それ以外をNegativeとする (Grid Searchの結果)
- Negativeデータセットを作る方法としてHard negative mining method を用いた
 - 識別できなかったデータを学習データセットに残し正しく識別できたものを外すことで識別の難しいデータのみが残る

Test result

- PASCAL VOC 2010 & ILSVR 2013

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 [20] [†]	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 [39]	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 [41]	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 [18] [†]	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

Table 1: Detection average precision (%) on VOC 2010 test. R-CNN is most directly comparable to UVA and Regionlets since all methods use selective search region proposals. Bounding-box regression (BB) is described in Section C. At publication time, SegDPM was the top-performer on the PASCAL VOC leaderboard. [†]DPM and SegDPM use context rescoring not used by the other methods.



- CNN Designによる違い

VOC 2007 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
R-CNN T-Net	64.2	69.7	50.0	41.9	32.0	62.6	71.0	60.7	32.7	58.5	46.5	56.1	60.6	66.8	54.2	31.5	52.8	48.9	57.9	64.7	54.2
R-CNN T-Net BB	68.1	72.8	56.8	43.0	36.8	66.3	74.2	67.6	34.4	63.5	54.5	61.2	69.1	68.6	58.7	33.4	62.9	51.1	62.5	64.8	58.5
R-CNN O-Net	71.6	73.5	58.1	42.2	39.4	70.7	76.0	74.5	38.7	71.0	56.9	74.5	67.9	69.6	59.3	35.7	62.1	64.0	66.5	71.2	62.2
R-CNN O-Net BB	73.4	77.0	63.4	45.4	44.6	75.1	78.1	79.8	40.5	73.7	62.2	79.4	78.1	73.1	64.2	35.6	66.8	67.2	70.4	71.1	66.0

Table 3: Detection average precision (%) on VOC 2007 test for two different CNN architectures. The first two rows are results from Table 2 using Krizhevsky et al.'s architecture (T-Net). Rows three and four use the recently proposed 16-layer architecture from Simonyan and Zisserman (O-Net) [43].

O-Net : Convolution layer x13, MaxPooling layer x5, FC x3

Subjects

- 学習をCNN Pre-training, CNN Fine-tuning, SVMについて個別に行う必要がある。(ad-hocな学習である)
- 計算コストが高い
 - 全ての物体候補に対しCNNで特徴量を抽出しているため

Appendix

Selective Search