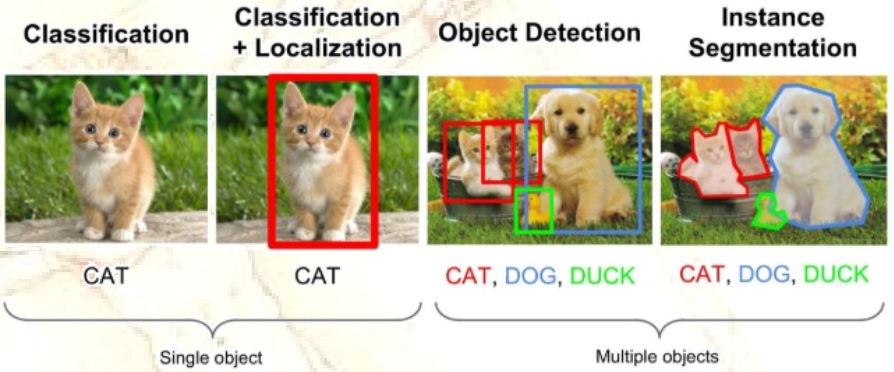
R-CNN
Fast R-CNN

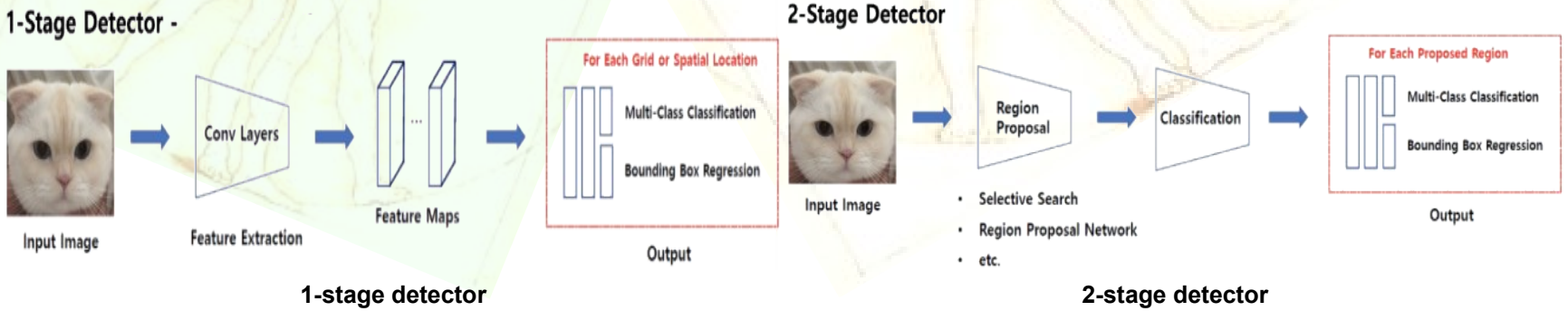
Pattern Recognition & Machine Learning Laboratory
Tae-jin Woo
Aug 04, 2021

Introduction

- **Computer vision**
  - **Types**
    - Classification
    - Classification + localization
    - Object detection
    - Instance segmentation
  - **Methods**
    - 1-stage detector
      - Simultaneous process of localization and classification
      - ex) Models of YOLO series
    - 2-stage detector
      - Sequential process of localization and classification
      - ex) Models of R-CNN series



Types of object detection

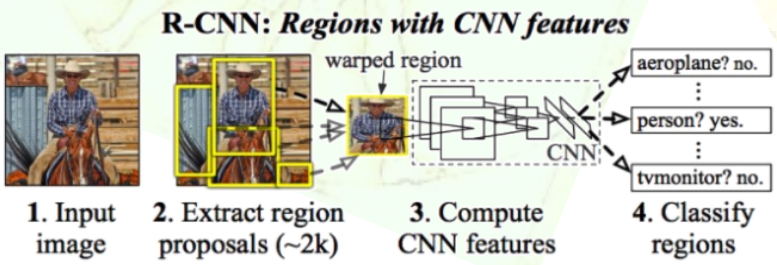


1-stage detector

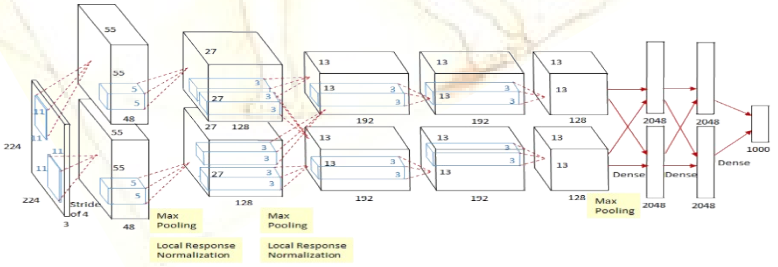
2-stage detector

R-CNN (1/5)

- **Introduction**
  - **Concept**
    - R-CNN: Regions based CNN for object detection
      - Combination of region proposals with CNN features
  - **Learning process**
    - Input an image
    - About 2k regional proposal outputs are extracted by selective search algorithm
      - Warped to make all extracted regional proposal outputs the same input size
      - Warping is used because input size of Fully Connected Layer (FCL) is fixed
    - Put 2k warped images each into CNN (AlexNet)
      - Supervised pre-training and domain-specific fine-tuning is used
    - Classification is performed to obtain the result for each CNN result
      - Binary SVM is used due to lack of training data
    - Linear regression is performed to adjust positions of each bounding box (bbox)



Object detection system overview

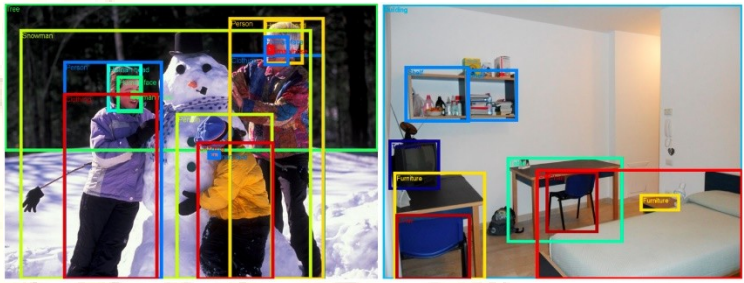


AlexNet structure for R-CNN

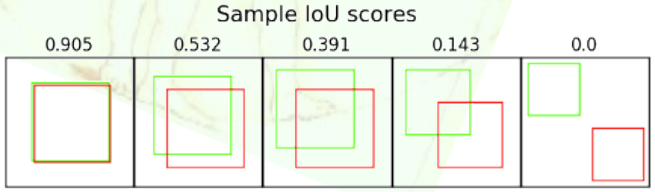
R-CNN (2/5)

- **Prior knowledge**
  - **Mean Average Precision (mAP)**
    - Average of precision corresponding to recall
      - Mean of AP values for all object classes
  - **Bounding box**
    - Rectangular box for detection display
      - Coordinate vector:
  - **Intersection over Union (IoU)**
    - Evaluation metrics for proposed region
      - Positive object criteria:
  - **Non Maximum Suppression (NMS)**
    - Remove overlapped bboxes for each object
      - Handle according to confidence and IoU
      - Drop if exceed IoU threshold with the highest confidence bbox

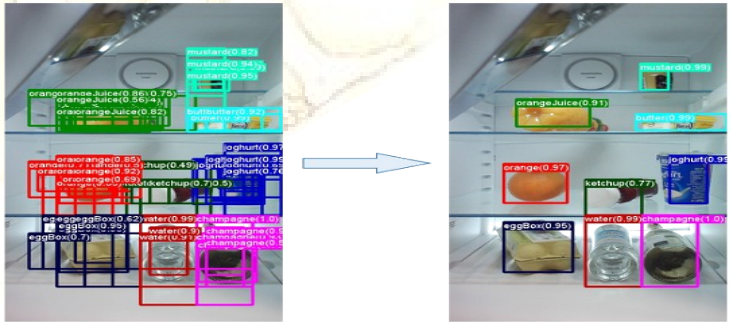
mAP



Bounding box



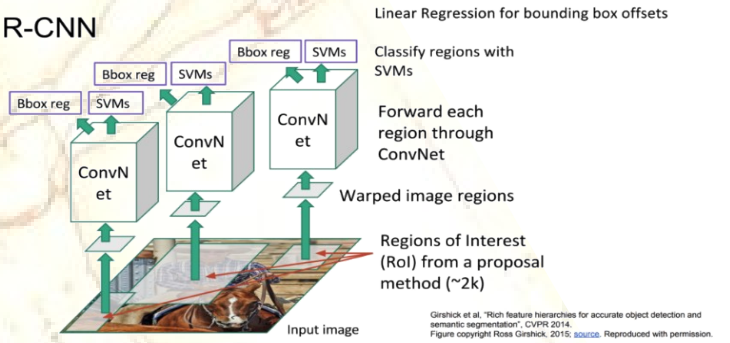
IoU



NMS

R-CNN (3/5)

- **Module design**
  - **Region proposals**
    - **Concept**
      - Find probable region of an object
      - They are called Regions of Interest (RoI)
    - **Method**
      - Selective search method is used
    - **Selective search**
      - Concept: Hierarchical grouping algorithm
      - Iteratively merges regions until becoming one region based on similarity
      - All created bounding boxes are warped to same size
  - **Feature extraction**
    - **Concept**
      - Extract features from each warped image
    - **Method**
      - Pre-trained AlexNet with slightly modified is applied
      - Fine-tuning of CNN is performed by new data set
      - Made to classify into N+1 classes (background)
    - **Advantage**
      - Achieved mAP improvement of 8%p

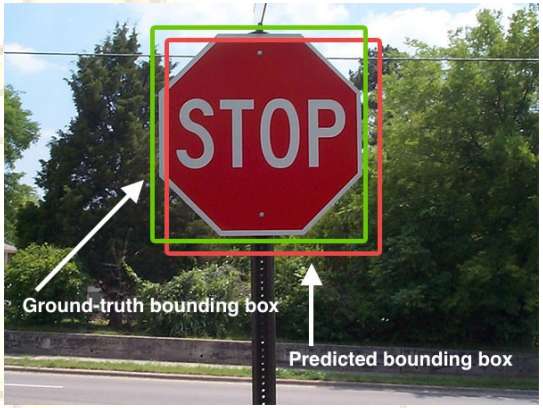


Overview of R-CNN

Selective search

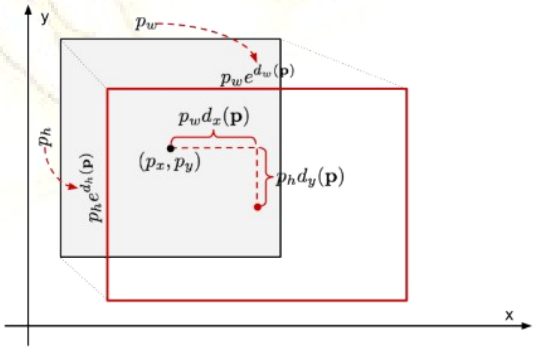
R-CNN (4/5)

- **Classification**
  - **Concept**
    - Classification via binary SVM
    - Use extracted features from CNN module
  - **Method**
    - Optimize SVM is trained independently per class
  - **Reason of SVM**
    - Better generalization due to lack of training data
    - Achieved mAP improvement of 4%p using SVM



Ground-truth bounding box

- **Bounding box regression**
  - **Concept**
    - Introduction of regression model to improve bboxes made by selective search
    - Use values of features from CNN of proposal
  - **Method**
    - Learnable parameters:
    - Regression:



Bounding box regression

    - Use regularization and IoU threshold to improve mAP

R-CNN (5/5)

- **Conclusion**
  - **Test result**
    - **PASCAL VOC 2010**
      - **Achieved state-of-the-art (SOTA) in 2013**
      - **Achieved improvement of mAP about 3%p by applying bbox regression**

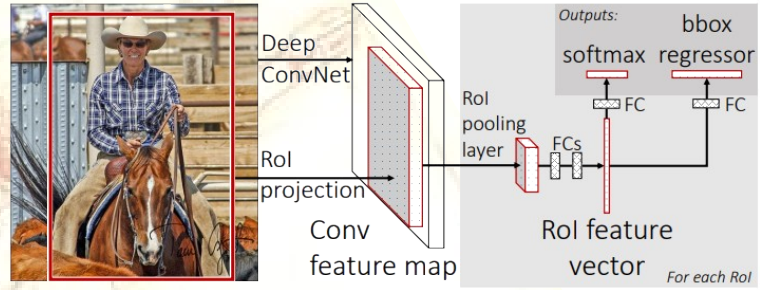
VOC 2010 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
DPM v5 [20]†	49.2	53.8	13.1	15.3	35.5	53.4	49.7	27.0	17.2	28.8	14.7	17.8	46.4	51.2	47.7	10.8	34.2	20.7	43.8	38.3	33.4
UVA [39]	56.2	42.4	15.3	12.6	21.8	49.3	36.8	46.1	12.9	32.1	30.0	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**

- **Appendix**
  - **Object proposal transformations**
    - **Difference according to RoI shape**
      - **Warping with improved mAP by 4%p**
  - **Positive-negative examples**
    - **Define different criteria for CNN and SVM**
      - **CNN: Positive if IoU is above 0.5**
        SVM: Positive only ground-truth boxes
      - **Different examples improved mAP**

Object proposal transformations

Fast R-CNN (1/2)

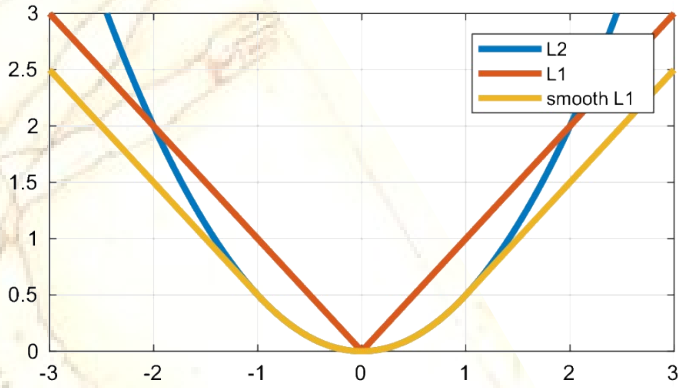
- Introduction
  - Concept
    - Fast R-CNN: Time efficient R-CNN
  - Cons of R-CNN
    - Slow speed
      - Perform all CNNs for 2k RoIs per image
    - Multi-stage pipelines
      - Impossible to update parameters of CNN through SVM and bbox regression
  - Improvements
    - RoI pooling layer
      - Find RoI using selective search with feature map after CNN layer
      - Need CNN only once per image
    - Single spatial bin
      - Avoid overfitting using only 7x7 spatial bin
    - Truncated SVD
      - Reduce learnable parameters of FC layer from to
    - Restricting of fine tuning layer
      - Reduce time by performing fine tuning only after conv3_1 of CNN architecture
    - Hierarchical sampling
      - Sample only 2 images per CNN, not 128 images



Object detection system overview

Fast R-CNN (2/2)

- **Training**
  - **Loss function**
    - **Multi-task loss function**
    - **Classification loss function**
    - **Localization loss function**
      - is ground-truth box coordinate vector for class u
        is predicted bounding box for class u
- **Conclusion**
  - **Test result**
    - **Validation with multiple models**
      - S: AlexNet, M, L: VGGNet
    - **Significant time reduction**
      - Maximum 18.3x faster in train time
      - Maximum 213x faster in test speedup
    - **Improvement of mAP**
      - 0.9%p higher in VOC07 mAP with L model



Regularization

	Fast R-CNN			R-CNN		
	S	**M**	**L**	**S**	**M**	**L**
train time (h)	**1.2**	2.0	9.5	22	28	84
train speedup	**18.3×**	14.0×	8.8×	1×	1×	1×
test rate (s/im)	0.10	0.15	0.32	9.8	12.1	47.0
▷ with SVD	**0.06**	0.08	0.22	-	-	-
test speedup	98×	80×	146×	1×	1×	1×
▷ with SVD	169×	150×	**213×**	-	-	-
VOC07 mAP	57.1	59.2	**66.9**	58.5	60.2	66.0
▷ with SVD	56.5	58.7	66.6	-	-	-

Fast R-CNN vs. R-CNN