Weakly Supervised Deep Detection Networks

Authors: Hakan Bilen, Andrea Vedaldi

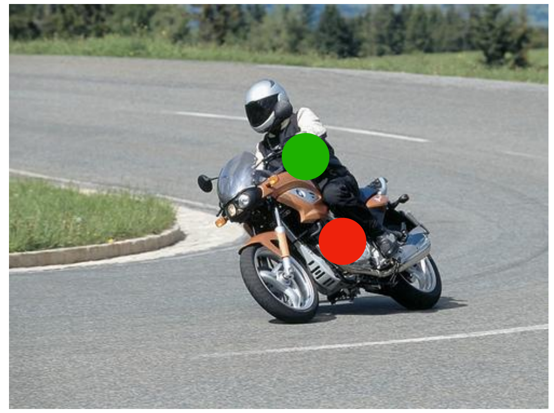
Atakan Serbes
27.03.2019

CNNs

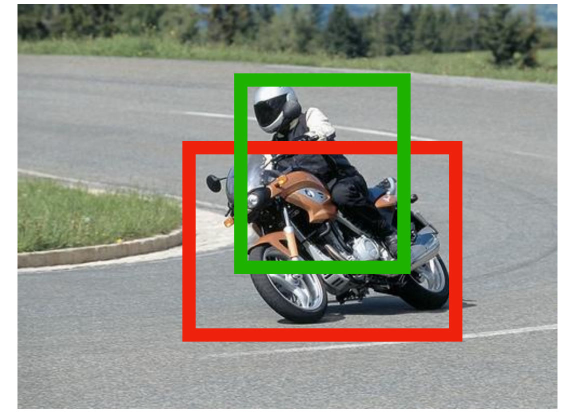
- **CNN**s have emerged as the new state-of-the-art for image recognition.
- Success comes from ability to learn from **large quantities of labelled data**

Manual Annotation

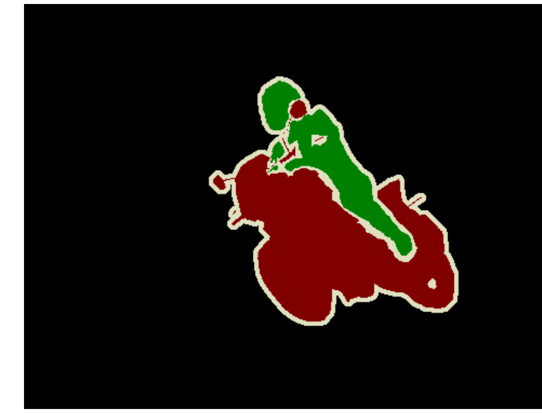
{motorbike,person}



{motorbike (point), person (point)}



{motorbike (b-box), person (b-box)}



{motorbike (pixel labels), person (pixel labels)}

1 sec per class

2.4 sec per class

10 sec per class

78 sec per class

annotation time

Weak supervision

Lower degree (or cheaper) annotation at train time than the required output at the test time

Weak Supervision

Problem of weak supervision is very important,

- Image understanding aims at learning **a growing body of complex visual concepts**

- CNN training is data-hungry and image labeling is tedious (thus WS can **reduce significantly the cost of data annotation** —such as image segmentation, image captioning, or object detection—)

For this paper, weakly supervised detection (WSD) is the problem of learning object detectors using only image-level labels

Motivation

- CNNs should contain meaningful representations of the data.
- There exists evidence that CNNs learn object and object parts in image classification [Zhou ICLR 15]
- Image level labels are plentiful

“Man”

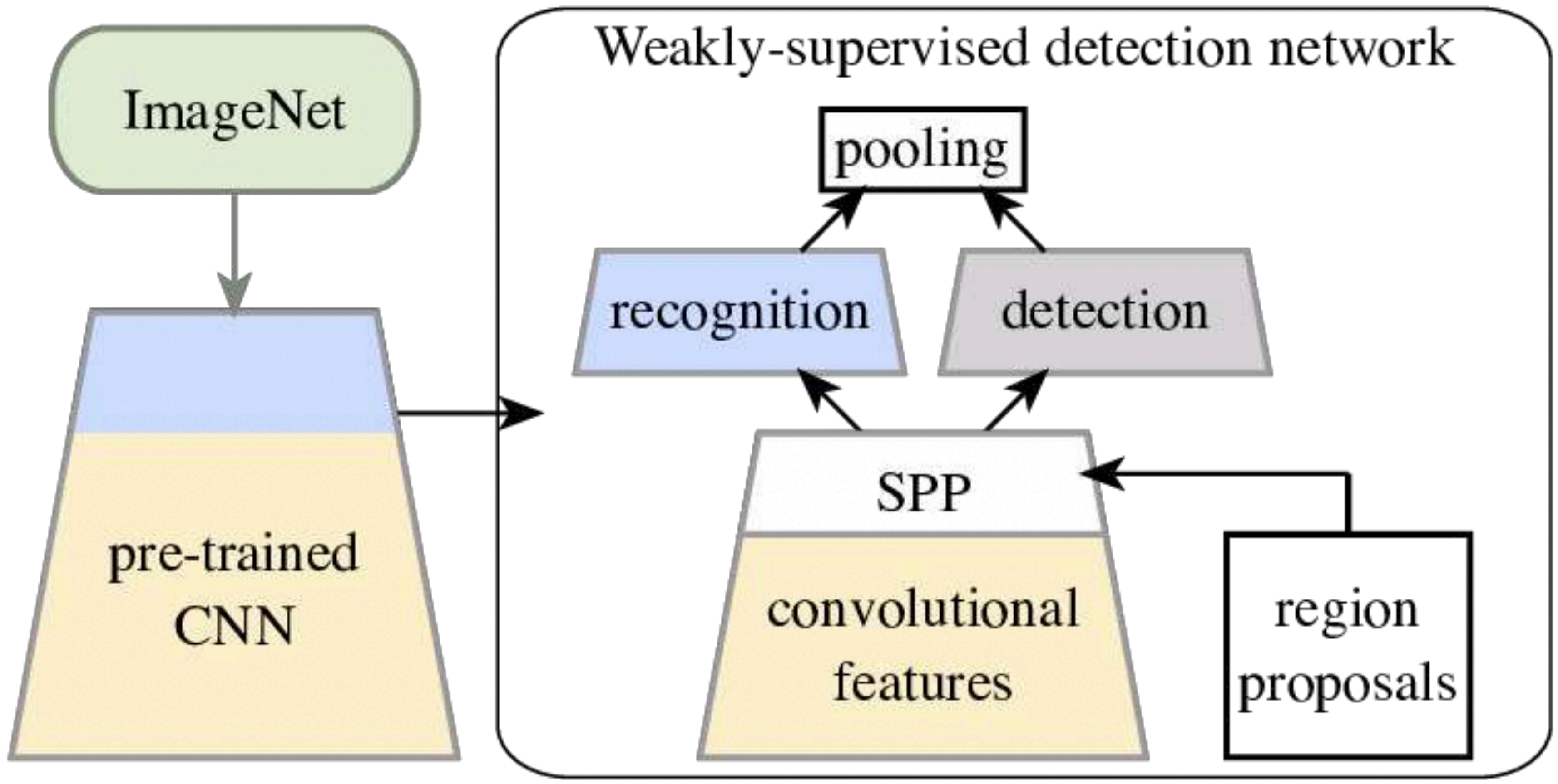
Motivation

Not the first to address the problem [Wang ECCV 14],

- Uses a pre-trained CNN to describe image regions

Comprises several components beyond the CNN and requires significant fine-tuning

Method



- A novel end-to-end method for weakly supervised object detection (WSOD) using pre-trained CNNs is proposed.
- Weakly supervised deep detection network (WSDDN)

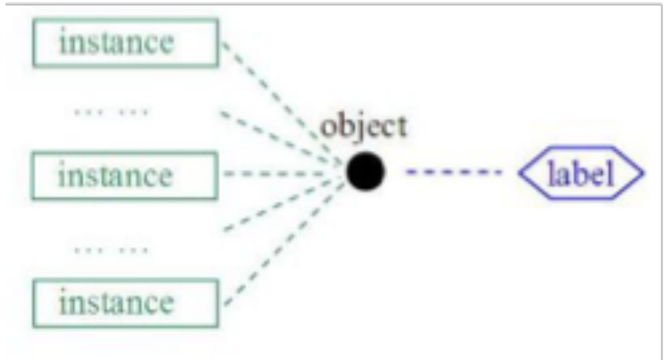
Fig: [Bilen CVPR 16]

Related Work

- Formulating **WSD as multiple instance learning (MIL)** where image is interpreted as a bag of regions.

- **Identifying similarity** between image parts [Song et al ICML 14]

- **CNN based related works**, [Cinbis TPAMI 17] combine multi-fold MIL with CNN features



(b) Multi-instance learning

Method

1. CNN pre-trained on a large-scale image classification task such as ImageNet ILSVRC 2012 data [Russakovsky IJCV 2015] (no bounding-box annotation)

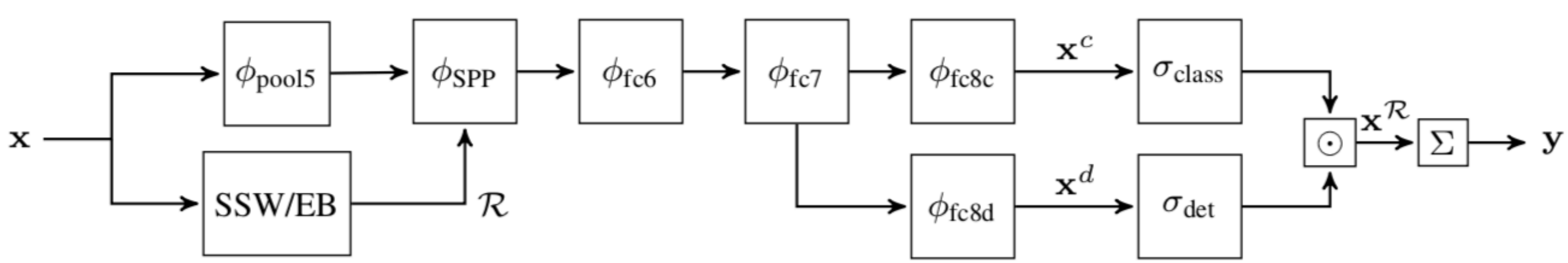
2. Construct WSDDN as an architectural modification of this CNN

3. Train / Fine-tune the WSDDN on a target dataset using only image-level annotations

Method

Modifications to pre-trained CNN

- Replace last pooling layer with a spatial pyramid pooling [He ECCV 14, Lazebnik CVPR 16]
- Add a parallel branch to the classification one that contains a fc layer followed by a soft-max layer
- Combine the classification and detection streams by element-wise product of two feature vectors



Method - Spatial Pyramid Pooling

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Replace last pooling layer with a spatial pyramid pooling [He ECCV 14, Lazebnik CVPR 16]

- Regions proposals are in different scales, SPP configures them to be compatible with the first fully-connected layer

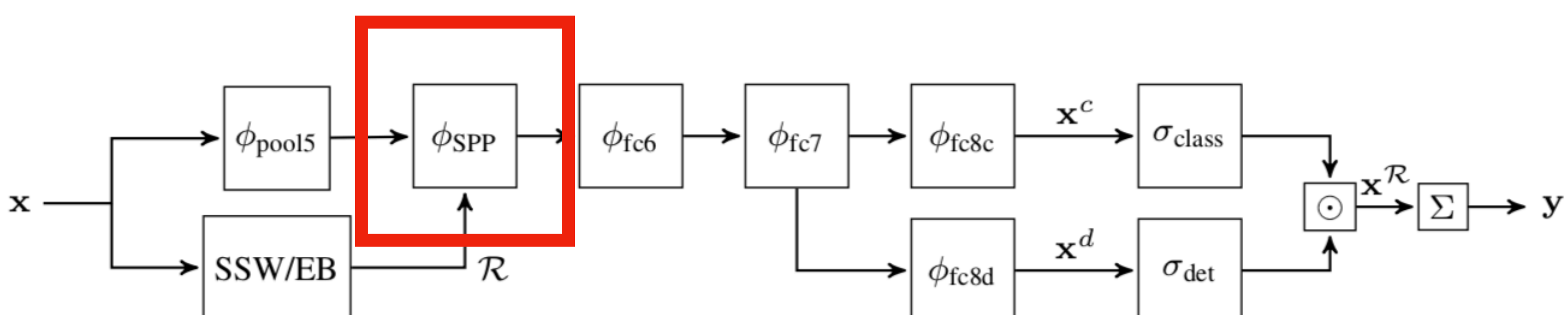
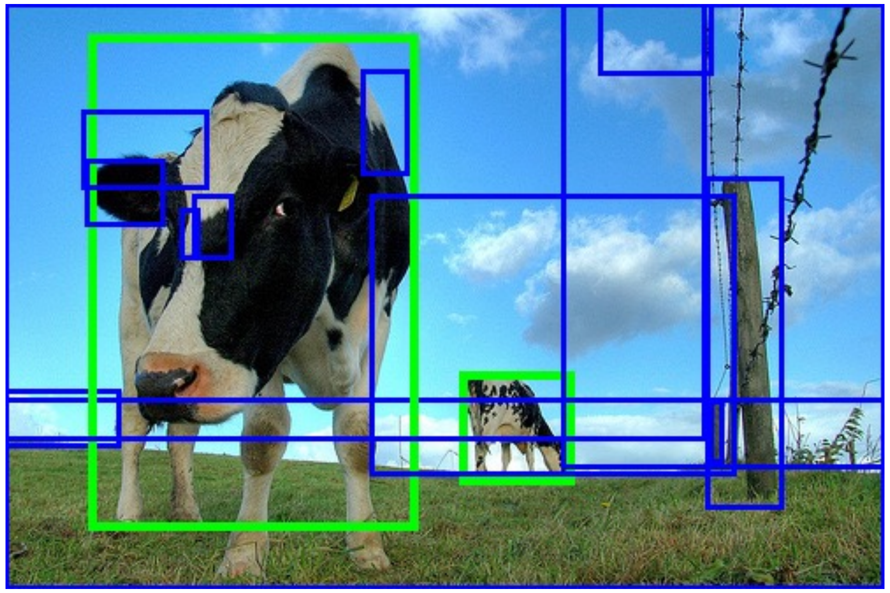
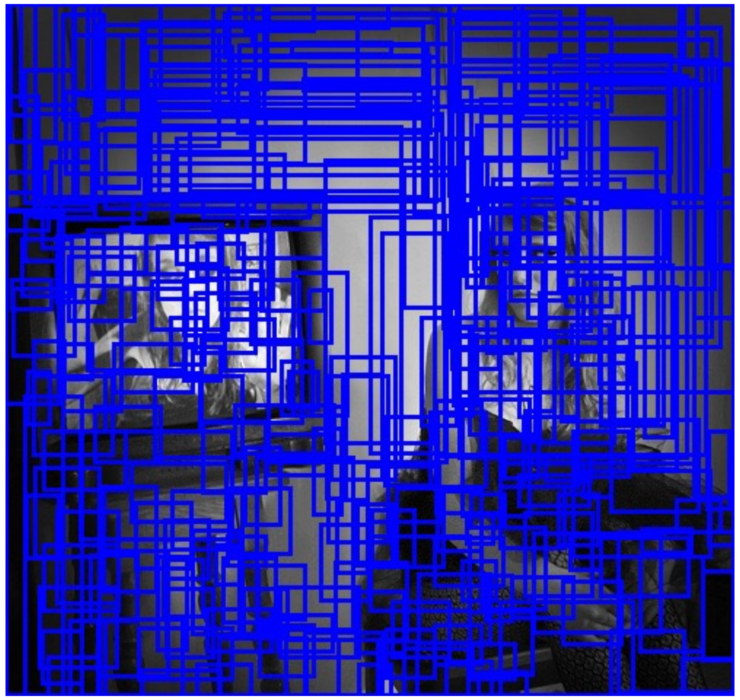


Fig: [Simonyan ICLR 15]

Method - Region Proposals





Region Proposals

- Given an image x, candidate object regions R are obtained by a region proposal mechanism
- Selective Search Windows (SSW) [Sande ICCV 11] and Edge Boxes (EB) [Zitnick ECCV 14] are used.

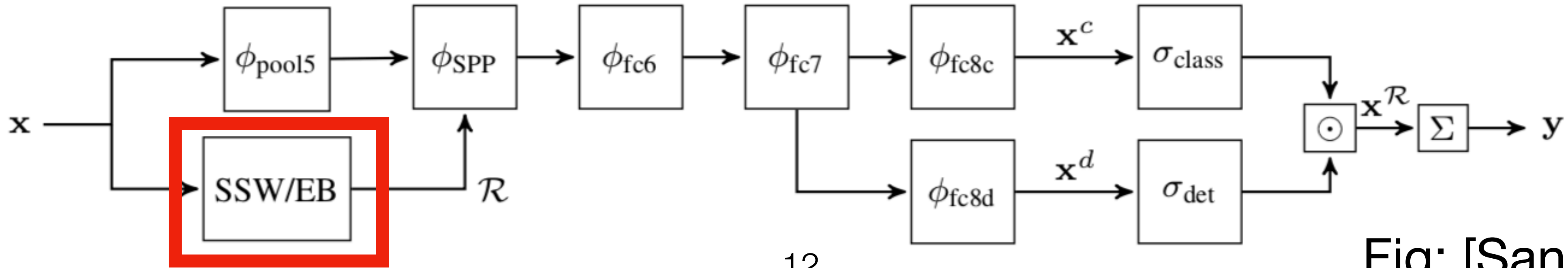
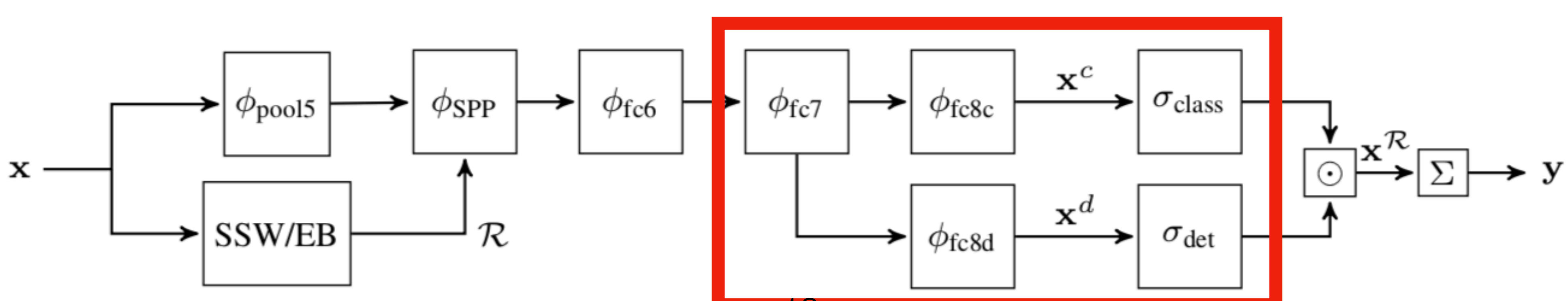


Fig: [Sande ICCV 14]

Method - Two Stream Architecture

Divide object detection into two sub-tasks with a two stream architecture,

- **Classification stream:** assign each region to a class
- **Detection stream:** picks most promising windows in an image given a class

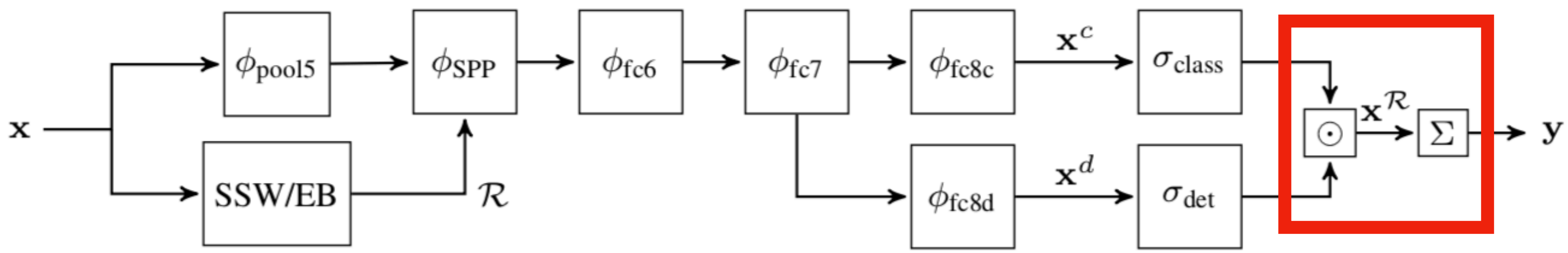


Method - Element-wise product

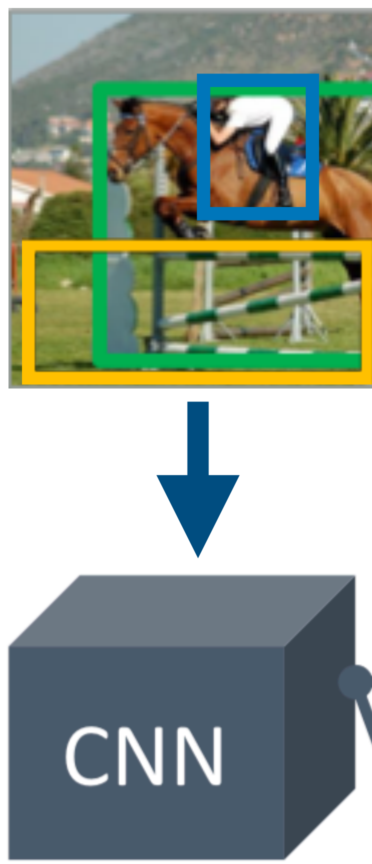
- Element-wise product
- Summation over regions to get an image-level class score
- It is a sum of element-wise product of soft-max normalized scores over R regions

    > thus in range of (0, 1)

$$y_c = \sum_{r=1}^{|\mathcal{R}|} x_{cr}^{\mathcal{R}}.$$



Method



Classification stream

	R_1	R_2	R_3	R_4
horse	0.52	0.47	0.04	0.93
person	0.48	0.53	0.96	0.07

Normalize over classes

Detection stream

	R_1	R_2	R_3	R_4
horse	0.04	0.01	0.07	0.88
person	0.02	0.03	0.91	0.04

Normalize over regions

	R_1	R_2	R_3	R_4
horse	0.02	0.05	0.00	0.82
person	0.01	0.02	0.87	0.00

horse	**0.89**
person	**0.90**

$$\sigma_{\text{class}}(\mathbf{x}_r^c) \odot \sigma_{\text{det}}(\mathbf{x}_r^d)$$

Fig: [Bilen CVPR 16]

Experimental Setup

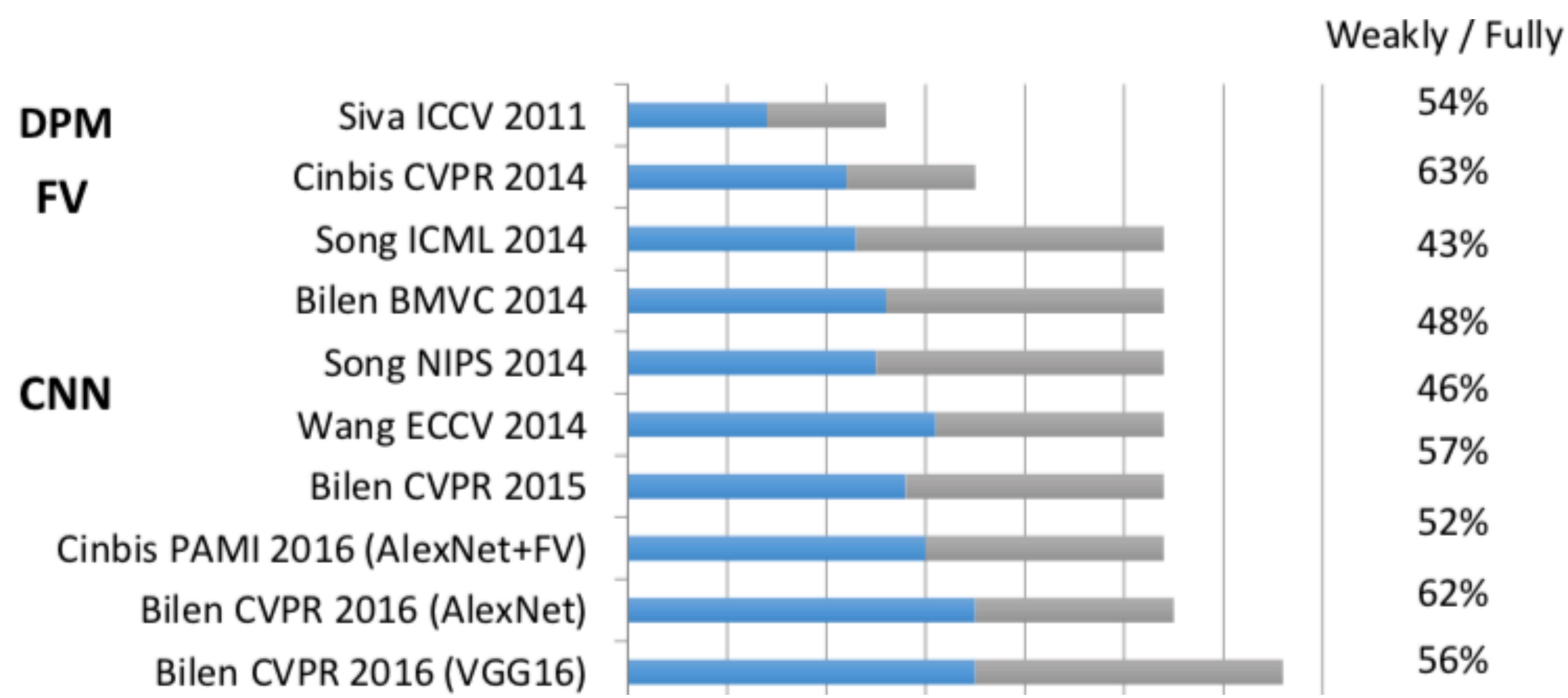
Method is evaluated with three pre-trained CNN models as in [Girshick ICCV 2015]

- **S** (Small) : VGG-CNN-F which is similar to AlexNet with reduced # of conv. filters. [Chatfield BMVC 14]
- **M** (Medium) : VGG-CNN-M-1024 with same depth as **S** but has smaller stride in first conv. layer
- **L** (Large) : VGG-VD16 [Simonyan ICLR 15]

These models are modified to become WSDDNs, are then trained on the PASCAL VOC datasets [Everingham IJCV 10]

Conclusion

WSL on PASCAL 07, Performance at test time



Fully supervised detection level is still very far away

Conclusion

End-to-end learning + No custom deep learning layers

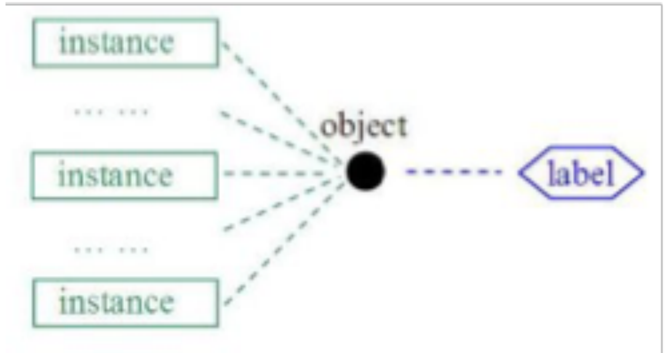
- State-of-the-art results with AlexNet (62% of supervised)
- Does not work well with deeper networks because it focuses on smaller regions with deeper networks. An object part (e.g. person face) is detected instead as the object as a whole.

Related Work

Most approaches formulate WSD as multiple instance learning (MIL) where image is interpreted as a bag of regions.

- (+) images = contain object of interest
  (-) images = no regions contain the object

- Results in a non-convex optimization problem, solvers tend to get stuck in local optima, soln. depends on the initialization.



(b) Multi-instance learning

Related Work

Another line of research in WSD is based on the idea of identifying similarity between image parts [Song et al ICML 14]

- Constructs a graph to find initial boxes which are relevant and discriminative

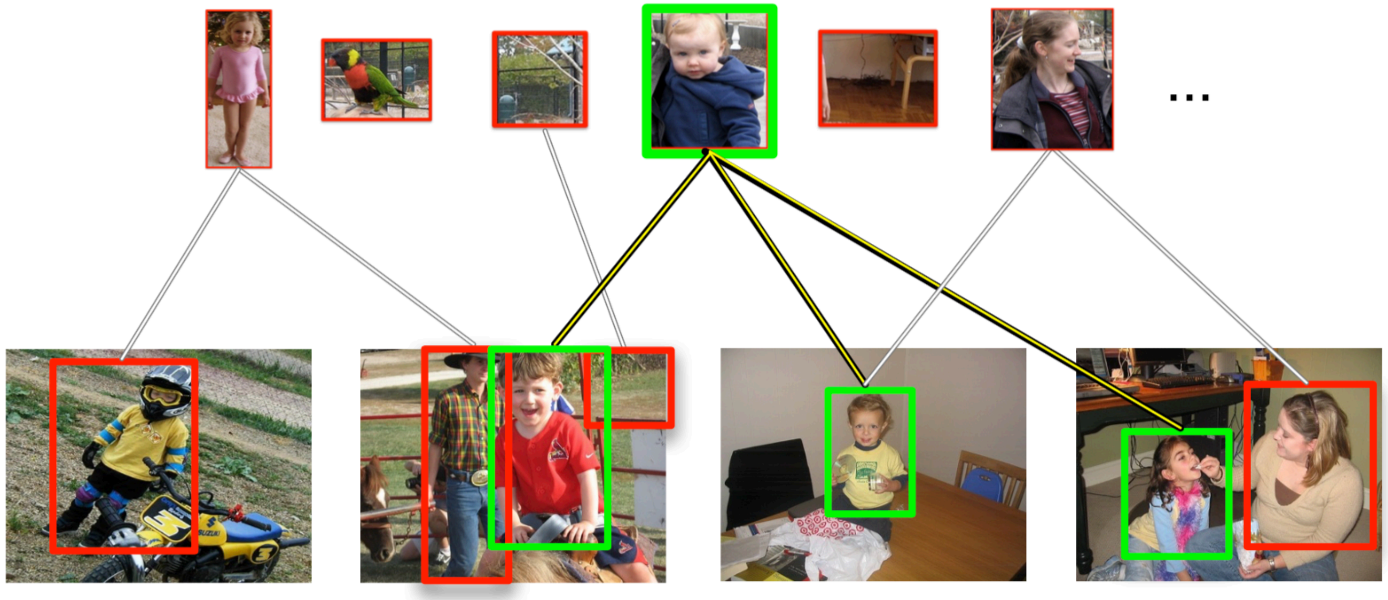


Fig: [Song et al ICML 14]

Related Work

There are also CNN based related works,

- [Cinbis TPAMI 17] combine multi-fold MIL with CNN features

- [Wang ECCV 14] develop a semantic clustering method on top of pre-trained CNN features

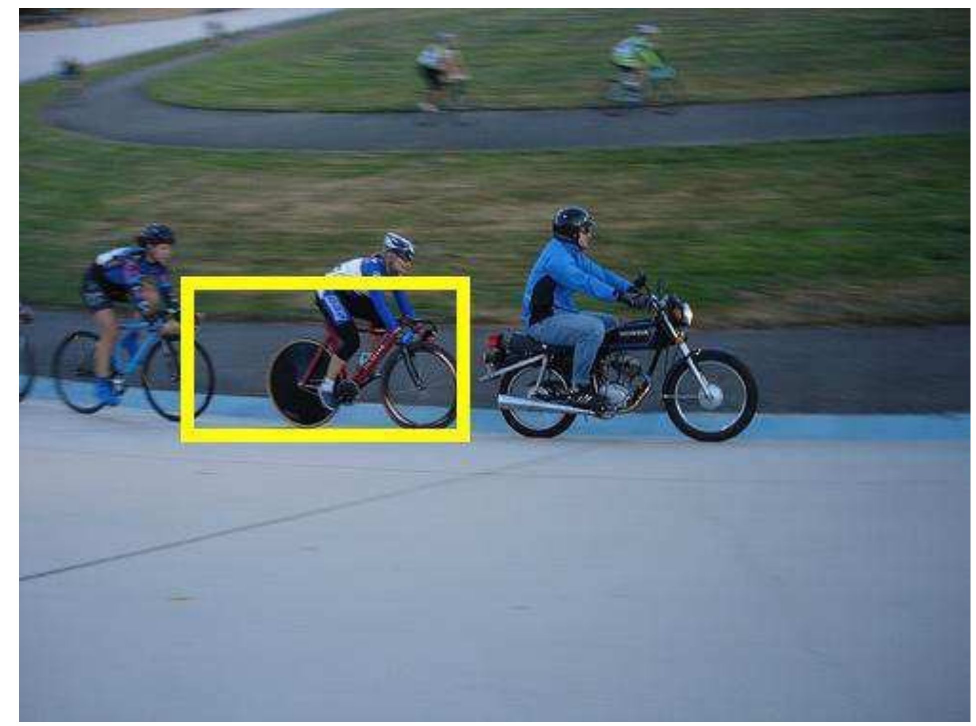


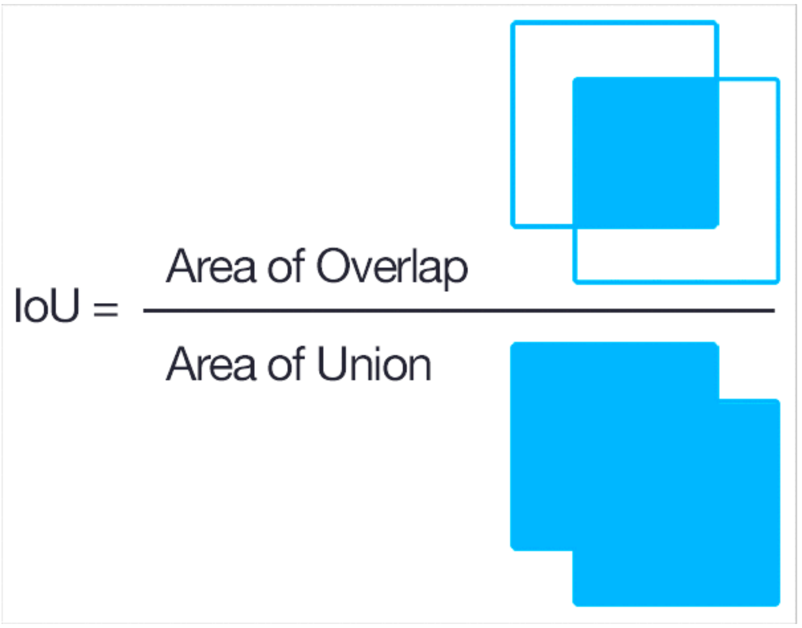
Fig: [Cinbis TPAMI 17]

Performance measures

For detection two performance measures are used,

1- **Standard (PASCAL) object evaluation criterion**

Avg. Precision at intersection over union (IoU) 50%



Performance measures

Other performance measure,

2- **Correct Localization (CorLoc)** [Alex IJCV 12]:

the percentage of positive training images is correctly localized at IoU 50%