IE685: MSc.Phd. Research Project I

End Sem Presentation

Title : Weakly Supervised Deep Detection Network

Rupesh Yadav | 21i190004

Guide : Prof. P. Balamurugan

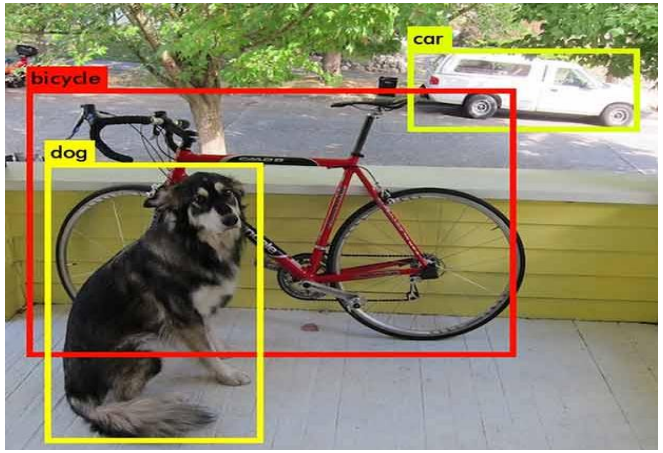
Object Detection

- Deep Learning based methods are widely used for this purpose
  - Fully Supervised methods
  - **Weakly supervised methods**

- In Weakly supervised methods only image level labels are used (no bounding boxes)

Training Image:-

Fully Supervised setting



Weakly Supervised setting

dog, bicycle, car

Image Source: Google Images

Architecture

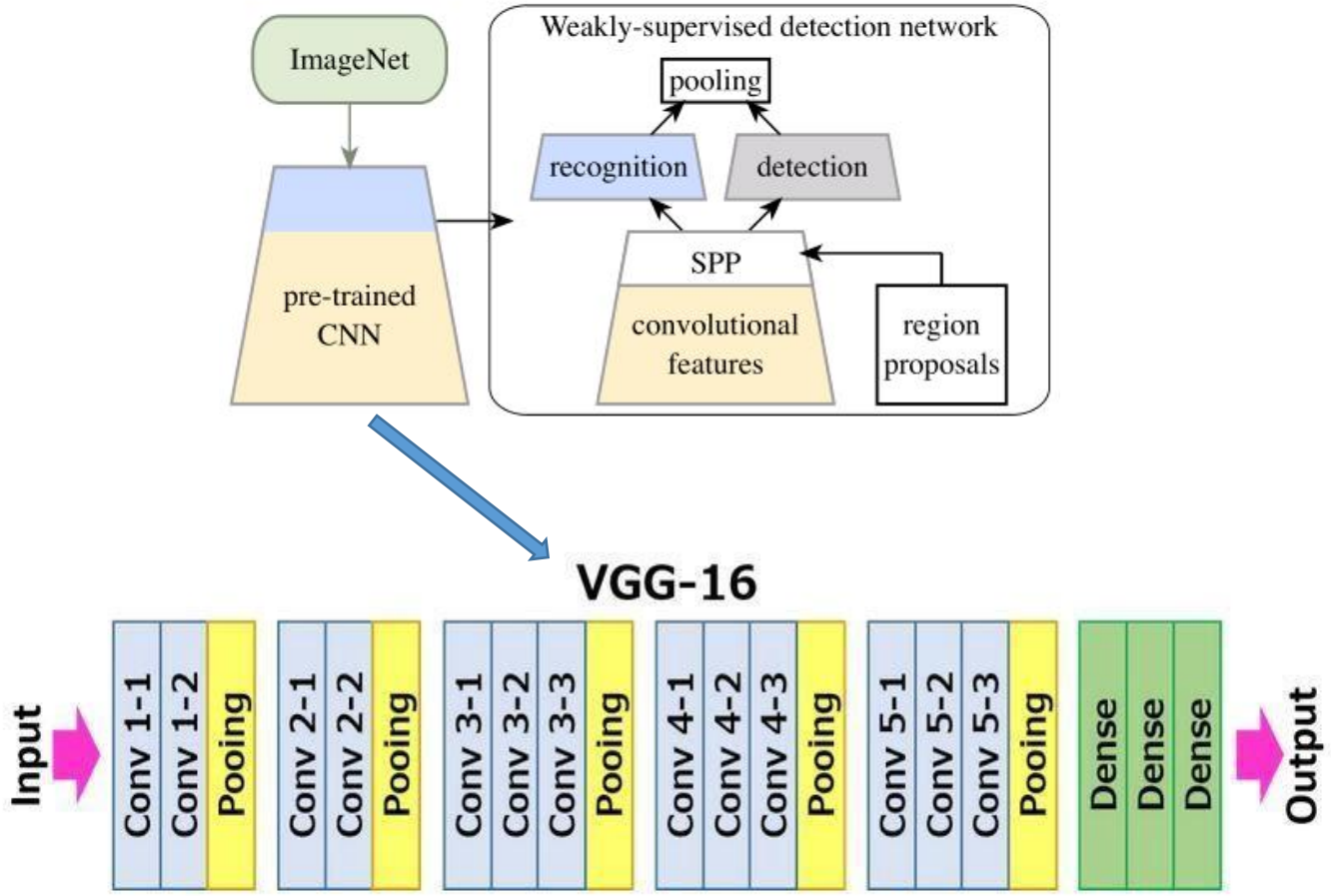


Image Sources:

1) Bilen, Hakan, and Andrea Vedaldi. "Weakly supervised deep detection networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

2) Simonyan, Karen and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition.” CoRR abs/1409.1556 (2015): n. pag.

Making of WSDDN

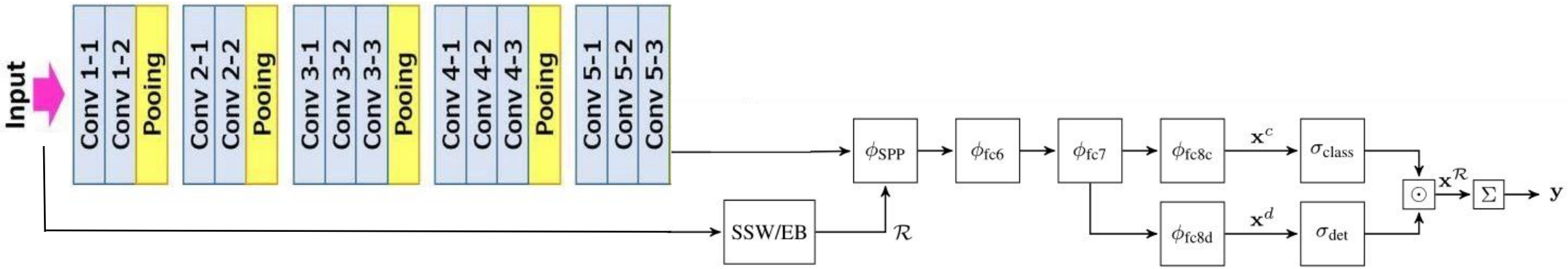


Figure 2. **Weakly-supervised deep detection network.** The figure illustrates the architecture of WSDDN

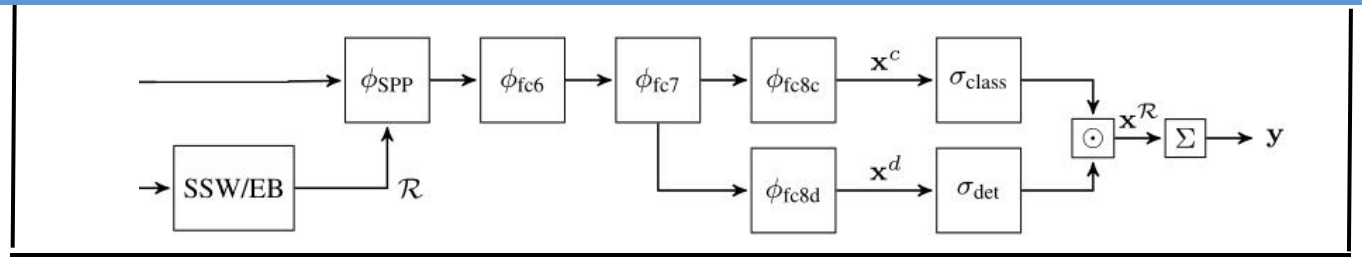
Modifications in Pre trained Network

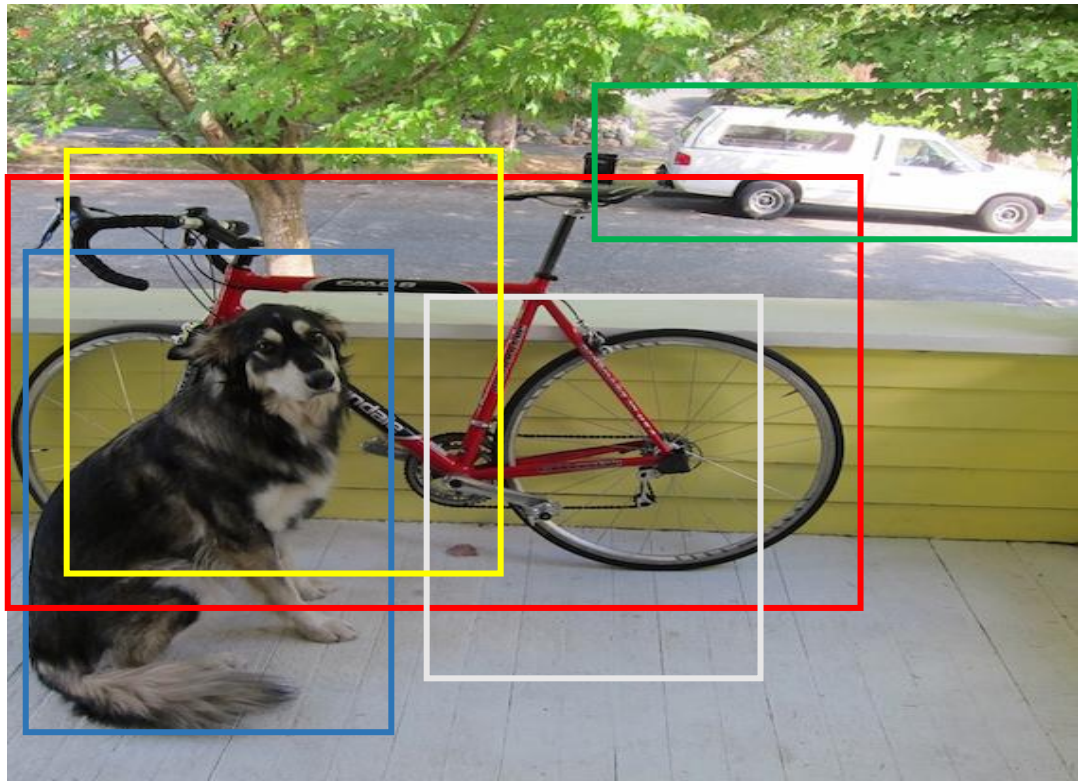
1. Pooling layer in last convolutional block is replaced by a layer implimenting Spatial Pyramid Pooling (SPP).
   - SPP is a pooling layer that removes the fixed-size constraint of the network
2. A parallel detection branch is added to the classification one that contains a fully-connected layer followed by a soft-max layer.
3. The classification and detection streams are combined by element-wise product followed by summing scores across regions.

Image Sources:

- Bilen, Hakan, and Andrea Vedaldi. "Weakly supervised deep detection networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
- Simonyan, Karen and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition.” CoRR abs/1409.1556 (2015): n. pag.

Classification and detection data streams





$[\sigma_{class}(\mathbf{x}^c)]$ =

regions →

classes ↓

	r	g	b	w	y
dog	0.3	0	0.9	0.8	0.5
bicycle	0.6	0.1	0.1	0.2	0.5
car	0.1	0.9	0	0	0

$[\sigma_{det}(\mathbf{x}^d)]$ =

regions →

classes ↓

	r	g	b	w	y
dog	0.3	0	0.5	0	0.2
bicycle	0.5	0	0.1	0.2	0.2
car	0.2	0.8	0	0	0

Combining region scores & Detection

- The final score of each region is obtained by taking the element-wise (Hadamard) product of the two scoring matrices as

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

$\mathbf{x}^{\mathcal{R}}$ =

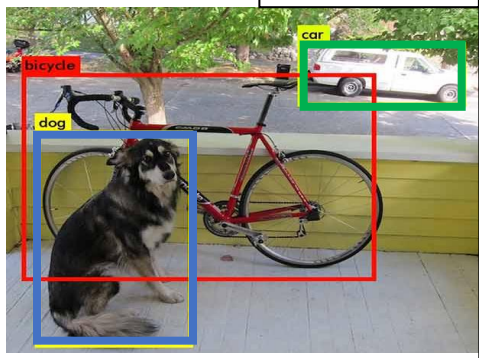
regions →

classes ↓

	r	g	b	w	y
dog	0.09	0	0.45	0	0.1
bicycle	0.30	0	0.01	0.04	0.1
car	0.02	0.72	0	0	0

- The region scores are then used to rank image regions by likelihood of centring an object (for each class independently)
  - **eg.** for dog -> *blue, yellow, red* *(now* using 0.1 as threshold) -> *blue, yellow*

After final detection



- For detection, standard non-maxima suppression is then performed (by iteratively removing regions with IoU larger than 40% with regions already selected)

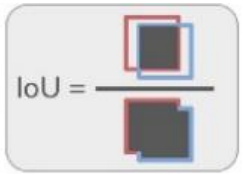


Image Source: https://towardsdatascience.com/what-is-average-precision-in-object-detection-localization-algorithms-and-how-to-calculate-it-3f330efe697b

Image Source: https://cvexplained.files.wordpress.com/2020/07/image-17.png

Training and Loss function

- **Image level classification scores**

$$\phi^{\mathbf{y}}(\mathbf{x}\,|\mathbf{w}) = \sum_{r=1}^{|\mathcal{R}|} x_{cr}^{\mathcal{R}}.$$

$\phi^{\mathbf{y}}(\mathbf{x}\,|\mathbf{w})$ =

classes ↓

dog	0.64
bicycle	0.45
car	0.74

Ground truth: $\mathbf{y}$ =

dog	1
bicycle	1
car	1

, $\mathbf{y}_i \in \{-1, 1\}^C$

- Loss/Error Function:

$$E(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n}\sum_{k=1}^{C} \log(y_{ki}(\phi_k^{\mathbf{y}}(\mathbf{x}_i|\mathbf{w}) - \frac{1}{2}) + \frac{1}{2})$$

Performance metrics

- Average Precision (AP)
- CorLoc

explaination of loss fn

y_k	$\phi_k^{\mathbf{y}}(\mathbf{x}_i\|\mathbf{w})$	$\log(y_{ki}(\phi_k^{\mathbf{y}}(\mathbf{x}_i\|\mathbf{w}) - \frac{1}{2}) + \frac{1}{2})$
-1	0.99	highly -ve
1	0.99	≈ 0
1	0.01	highly -ve
-1	0.01	≈ 0

Experiment Setup

- I have taken following 3 different base networks to make three detection models
  - **AlexNet** -> **A** (name of the corrs. detection model)
  - **VGG16** -> **B**
  - **VGG19** -> **C**

- **PASCAL VOC 2007** dataset was used to perform training and testing
  - It contains 5011 images in **train+validation** set and 4952 in **test** set
  - Its contains images containing objects from 20 different classes

PASCAL VOC annotation format

```
▼<annotation>
  <folder>VOC2007</folder>
  <filename>000001.jpg</filename>
 ▼<source>
    <database>The VOC2007 Database</database>
    <annotation>PASCAL VOC2007</annotation>
    <image>flickr</image>
    <flickrid>341012865</flickrid>
  </source>
 ▼<owner>
    <flickrid>Fried Camels</flickrid>
    <name>Jinky the Fruit Bat</name>
  </owner>
 ▼<size>
    <width>353</width>
    <height>500</height>
    <depth>3</depth>
  </size>
  <segmented>0</segmented>
 ▼<object>
    <name>dog</name>
    <pose>Left</pose>
    <truncated>1</truncated>
    <difficult>0</difficult>
   ▼<bndbox>
      <xmin>48</xmin>
      <ymin>240</ymin>
      <xmax>195</xmax>
      <ymax>371</ymax>
    </bndbox>
  </object>
```

Experiment-1

- Time taken (approximate) by models to train for 1 epoch
  - **A** -> 15 minutes
  - **B** -> 40 minutes
  - **C** -> 40 minutes

- I have trained all the models for 20 epochs with following learning rates

Model ↓ *Epochs* →	0 – 10	10-20
A	10^{-5}	10^{-6}
B	50^{-6}	10^{-6}
C	50^{-6}	50^{-7}

- I have used the region proposals obtained from **Edge Box** method for the training purpose
  - For each image in the training set it contains around **2000-3000** regions/boxes
  - It also provides a **score** for each region proposals, which is also used while training

Experiment-1 & Results

- For each image in test set the average score for 10 **augmented** images were taken
  - Five scales [80, 576, 688, 864, 1200] and their horizontal flips are used
- **Testing time** was around **3 hours** for all the models
- The **ensemble** of all three models showed some improved but with increase in computation time
- **Average Precision (AP)** obtained from all models:
  - **A** -> 26.7%
  - **B** -> 28.2%
  - **C** -> 30.6%
  - ensemble -> 32.6%

Spatial Regulariser

- As WSDDN is optimised for image-level class labels, it does not guarantee any spatial smoothness.
  - **Spatial smoothness:** If a region obtains a high score for an object class, the neighbouring regions with high overlap will also have high scores.

Spatial Regulariser

- As this method does not have access to ground truth boxes, a soft regularisation strategy is followed that penalises the feature map discrepancies between the highest scoring region and the regions with at least 60% IoU during training.

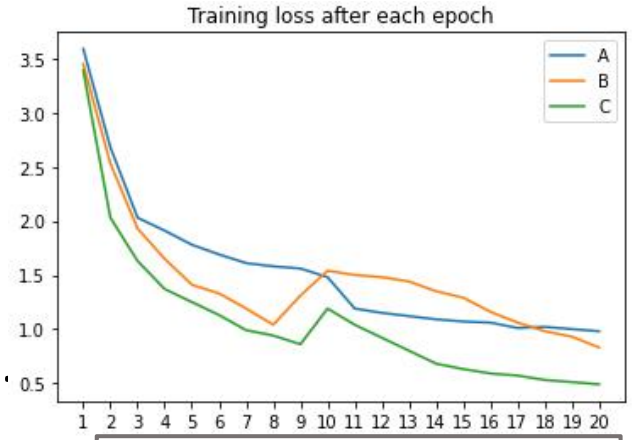
- The following regulariser term is added to the loss function while training.

$$\frac{1}{nC}\sum_{k=1}^{C}\sum_{i=1}^{n_k^+}\frac{1}{2}\phi_k^y(x_i|w)(\phi_{kp}^{fc7}-\phi_{ki}^{fc7})^T(\phi_{kp}^{fc7}-\phi_{ki}^{fc7})$$

where n_k^+ is the number of positive images for class k and $kp = argmax_j \phi_{kj}^y$ is the highest scoring region in image i for class k.

Experiments-2 & Results

- Trained the models with the regularisation component
  - Its code implementation was done by me
- Training and testing time was same as that of un-regularised models.
- Using this regularisation strategy showed improvement in the models performance.



For un-regularized models

Table 4.3: my results

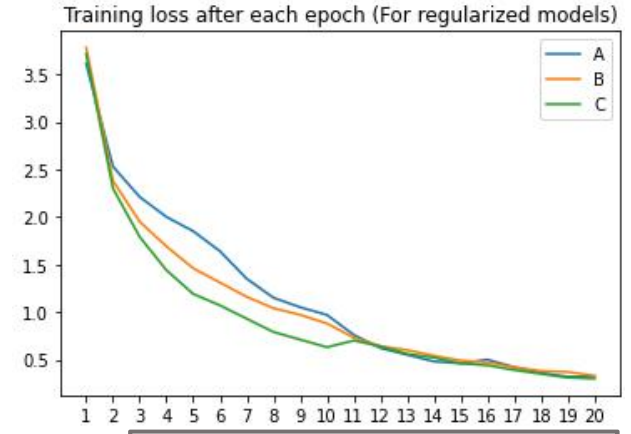
	A	B	C	ens
EB+Box score	26.7	28.2	30.6	32.6
EB+Box sc.+sp. reg	27.8	32.1	32.2	34.7

Table 4.4: results in original paper

	S	M	L	ens
EB+Box score	33.4	32.7	30.4	36.7
EB+Box sc.+sp. reg	34.5	34.9	34.8	39.3

Table 4.5: VOC 2007 test detection mean average precision (%). The ensemble network is denoted as Ens.

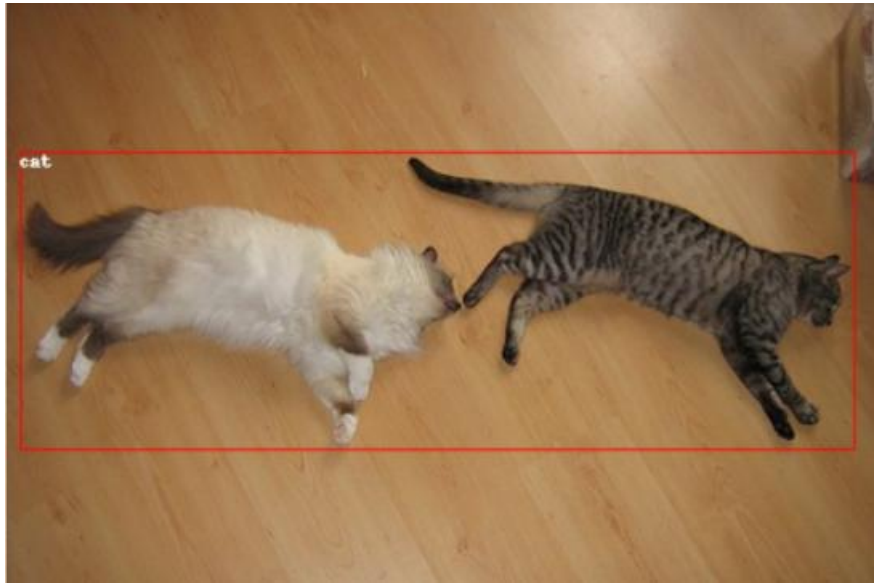
For regularised models

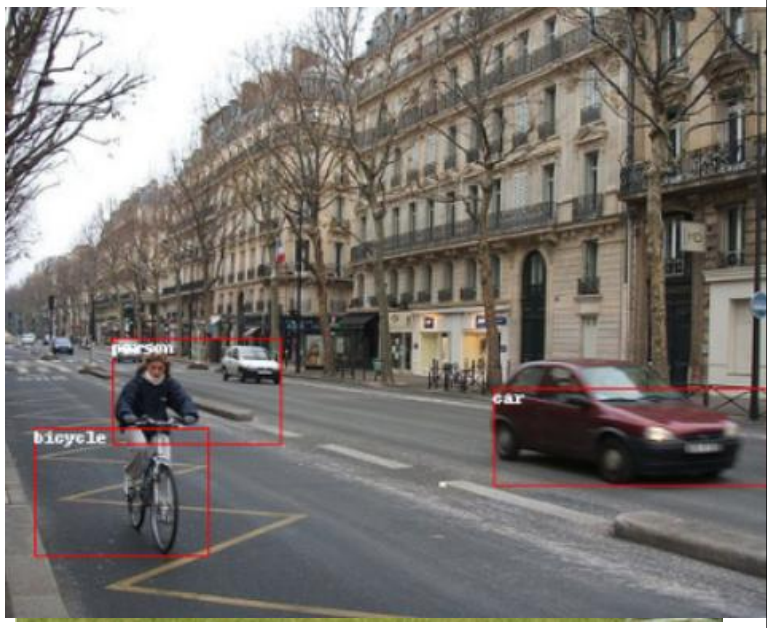


For regularized models

Detection Results

- I have also calculated **CorLoc** (with criterion IoU 0.2) on the PASCAL VOC 2007 positive test images and got 44.0%, 51.6% and 50.8% by model A, B and C respectively





Experiment-3

Using the model for classification

- Instead of just detection task I have also tested this model for multi-class image classification task on same PASCAL VOC test set.
  - I have got 81.7%, 89.8% and 89.9% Average Precision (AP) by A, B and C respectively.
  - In contrast the max AP reported in the original paper is 89.7%

Experiment-4

- As the training and testing time was too long
- I was able to reduce the time by taking only half of the region proposals based on the EB scores for training and testing
    - Both training and testing time was reduced by a factor of 2 for all three models
    - But with a drop of around only 2-3% in the performance of the models

Conclusion

- In this project I have trained and experimented with the network proposed by *Bilen et. al. [1]*.
- I was able to reduce the training and testing time by a factor of 2 by selecting only top 50% regions based on the EB scores.
- Apart from object detection I have also tested my models on the task of image classification on PASCAL VOC test data and got my best AP as 0.2% more than the best AP of paper.
- I used this model for detection on random images but it was taking around 2 sec for an image most of this time was taken by region proposal.

Types of error in majority of false detection,

- It grouped multiple object instances with a single bounding box
- Focus on parts (e.g. “faces”) rather than whole object

[1] Bilen Hakan, and Andrea Vedaldi. "Weakly supervised deep detection networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

Future Work

- The real time performance can be improved in future by using more efficient region proposal mechanism.
- As the performance of my ensemble model is not that good as authors. Hence, better ensemble method can be used to improve performance.
- Instead of taking simple average of the scores from all 10 augmentations of an image other criterian (e.g. some voting strategy) can be used to further improve the overall performance

References

- In this presentation I have presented the work of **Hakan Bilen and Andrea Vedaldi** published in the paper mentioned below
  - Bilen, Hakan, and Andrea Vedaldi. "Weakly supervised deep detection networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

- Simonyan, Karen and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition.” CoRR abs/1409.1556 (2015).
- https://in.mathworks.com/discovery/object-detection.html
- https://towardsdatascience.com/what-is-average-precision-in-object-detection-localization-algorithms-and-how-to-calculate-it-3f330efe697b
- Solovyev, Roman and Wang, Weimin and Gabruseva, Tatiana ”Weighted boxes fusion: Ensembling boxes from different object detection models”, Image and Vision Computing, page 1-6, year 2021, publisher Elsevie

Thank You!