Rich feature hierarchies for accurate object detection and semantic segmentation

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik

Speaker: Yucong Shen

4/5/2018
Develop of Object Detection

1. DPM (Deformable parts models)
2. R-CNN
3. Fast R-CNN
4. Faster R-CNN
5. Mask R-CNN
1. Region Proposal
2. Feature extraction
3. Classification

**R-CNN: Regions with CNN features**

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

*Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik*  
Rich feature hierarchies for accurate object detection and semantic segmentation
Region Proposal

1. Take an input image, extract 2000 bottom-up region proposals with selective search.
2. Selective search:
   1. Segment input image into 1K to 2K regions
   2. Compute the similarity of neighboring regions
   3. Merge the similarity neighboring regions
3. Resize all the proposal regions into $227 \times 227$
Selective Search: Color Similarity and Texture Similarity

1. **Color Similarity** For every region, we will get a color histogram, there would be 25 (75 if RGB) intervals:
   \[ C_i = \{ c_1^i, c_2^i, ... c_n^i \} \]
   \[ S_{\text{color}}(r_i, r_j) = \sum_{k=1}^{n} \min(c_k^i, c_k^j) \]

2. **Texture Similarity** For every region, calculate Guassian derivative for every channel’s 8 different direction with variance 1. Obtain the histogram of them: \[ T_i = \{ t_1^i, t_2^i, ..., t_n^i \} \]
   \[ S_{\text{texture}}(t_i, t_j) = \sum_{k=1}^{n} \min(t_k^i, t_k^j) \]
1 **Network Structure** Train the
   1. AlexNet
   2. VGG-16

to extract a 4096-dimension feature vector

2 **Supervised Pre-training** pre-train CNN network on large auxiliary dataset (ImageNet)

3 **Domain-specific fine-tuning** Continue SGD training of CNN using only warped regions proposals. The CNN architecture is unchanged except replacing the 1000-way classification layer with the (N+1)-way classification layer. In each iteration, the batch size is 128, 32 positive samples, 96 negative samples.

4 **Object category classification** Treat all region proposals with $\geq 0.5$ with a ground-truth box as positives for that box's class and the rest as negatives. Once features are extracted and training labels are applied, we optimize one linear SVM per class.
1 **Training** SVM can only deal with 2-way classification problem. Labeling the data as negative when IoU less than 0.3. Once features are extracted, train N linear SVM classifier. After that, training a linear regression predict a new detection window given the 4096-dimension vectors features for a selective search region proposal.

2 **Testing** Extract 2000 proposal regions with selective search, normalization all the regions to $227 \times 227$, forward propogate in CNN, extract the 4096-dimension feature vector, testing with the SVM classifier, removing the redundant bounding box with NMS, obtain the bounding box after bounding box regression.
Result on PASCAL VOC and ILSIRC

Rich feature hierarchies for accurate object detection and semantic segmentation
Result on PASCAL VOC and ILSIIRC

<table>
<thead>
<tr>
<th>VOC 2010 test</th>
<th>aero</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>table</th>
<th>dog</th>
<th>horse</th>
<th>mbike</th>
<th>person</th>
<th>plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tv</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPM v5 [20]†</td>
<td>49.2</td>
<td>53.8</td>
<td>13.1</td>
<td>15.3</td>
<td>35.5</td>
<td>53.4</td>
<td>49.7</td>
<td>27.0</td>
<td>17.2</td>
<td>28.8</td>
<td>14.7</td>
<td>17.8</td>
<td>46.4</td>
<td>51.2</td>
<td>47.7</td>
<td>10.8</td>
<td>34.2</td>
<td>20.7</td>
<td>43.8</td>
<td>38.3</td>
<td>33.4</td>
</tr>
<tr>
<td>UVA [39]</td>
<td>56.2</td>
<td>42.4</td>
<td>15.3</td>
<td>12.6</td>
<td>21.8</td>
<td>49.3</td>
<td>36.8</td>
<td>46.1</td>
<td>12.9</td>
<td>32.1</td>
<td>30.0</td>
<td>36.5</td>
<td>43.5</td>
<td>52.9</td>
<td>32.9</td>
<td>15.3</td>
<td>41.1</td>
<td>31.8</td>
<td>47.0</td>
<td>44.8</td>
<td>35.1</td>
</tr>
<tr>
<td>Regionlets [41]</td>
<td>65.0</td>
<td>48.9</td>
<td>25.9</td>
<td>24.6</td>
<td>24.5</td>
<td>56.1</td>
<td>54.5</td>
<td>51.2</td>
<td>17.0</td>
<td>28.9</td>
<td>30.2</td>
<td>35.8</td>
<td>40.2</td>
<td>55.7</td>
<td>43.5</td>
<td>14.3</td>
<td>43.9</td>
<td>32.6</td>
<td>54.0</td>
<td>45.9</td>
<td>39.7</td>
</tr>
<tr>
<td>SegDPM [18]†</td>
<td>61.4</td>
<td>53.4</td>
<td>25.6</td>
<td>25.2</td>
<td>35.5</td>
<td>51.7</td>
<td>50.6</td>
<td>50.8</td>
<td>19.3</td>
<td>33.8</td>
<td>26.8</td>
<td>40.4</td>
<td>48.3</td>
<td>54.4</td>
<td>47.1</td>
<td>14.8</td>
<td>38.7</td>
<td>35.0</td>
<td>52.8</td>
<td>43.1</td>
<td>40.4</td>
</tr>
<tr>
<td>R-CNN</td>
<td>67.1</td>
<td>64.1</td>
<td>46.7</td>
<td>32.0</td>
<td>30.5</td>
<td>56.4</td>
<td>57.2</td>
<td>65.9</td>
<td>27.0</td>
<td>47.3</td>
<td>40.9</td>
<td>66.6</td>
<td>57.8</td>
<td>65.9</td>
<td>53.6</td>
<td>26.7</td>
<td>56.5</td>
<td>38.1</td>
<td>52.8</td>
<td>50.2</td>
<td>50.2</td>
</tr>
<tr>
<td>R-CNN BB</td>
<td><strong>71.8</strong></td>
<td><strong>65.8</strong></td>
<td><strong>53.0</strong></td>
<td><strong>36.8</strong></td>
<td><strong>35.9</strong></td>
<td><strong>59.7</strong></td>
<td><strong>60.0</strong></td>
<td><strong>69.9</strong></td>
<td><strong>27.9</strong></td>
<td><strong>50.6</strong></td>
<td><strong>41.4</strong></td>
<td><strong>70.0</strong></td>
<td><strong>62.0</strong></td>
<td><strong>69.0</strong></td>
<td><strong>58.1</strong></td>
<td><strong>29.5</strong></td>
<td><strong>59.4</strong></td>
<td><strong>39.3</strong></td>
<td><strong>61.2</strong></td>
<td><strong>52.4</strong></td>
<td><strong>53.7</strong></td>
</tr>
</tbody>
</table>

---

**ILSVRC2013 detection test set mAP**

- *R-CNN BB: 31.4%
- *OverFeat (2): 24.3%
- Uva-Euvision: 22.6%
- *NEC-MU: 20.9%
- *OverFeat (1): 19.4%
- Toronto A: 11.5%
- SYSU_Vision: 10.5%
- GPU_UCLA: 9.8%
- Delta: 6.1%
- UIUC-IFP: 1.0%

---

**ILSVRC2013 detection test set class AP box plots**

- Average precision (AP) in %
Visulization of Pool5

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik

Rich feature hierarchies for accurate object detection and semantic segmentation
## Result on PASCAL ILSIRC

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

<table>
<thead>
<tr>
<th>VOC 2007 test</th>
<th>aero</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>table</th>
<th>dog</th>
<th>horse</th>
<th>mbike</th>
<th>person</th>
<th>plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tv</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-CNN pool5</td>
<td>51.8</td>
<td>60.2</td>
<td>36.4</td>
<td>27.8</td>
<td>23.2</td>
<td>52.8</td>
<td>60.6</td>
<td>49.2</td>
<td>18.3</td>
<td>47.8</td>
<td>44.3</td>
<td>40.8</td>
<td>56.6</td>
<td>58.7</td>
<td>42.4</td>
<td>23.4</td>
<td>46.1</td>
<td>36.7</td>
<td>51.3</td>
<td>55.7</td>
<td>44.2</td>
</tr>
<tr>
<td>R-CNN fc6</td>
<td>59.3</td>
<td>61.8</td>
<td>43.1</td>
<td>34.0</td>
<td>25.1</td>
<td>53.1</td>
<td>60.6</td>
<td>52.8</td>
<td>21.7</td>
<td>47.8</td>
<td>42.7</td>
<td>47.8</td>
<td>52.5</td>
<td>58.5</td>
<td>44.6</td>
<td>25.6</td>
<td>48.3</td>
<td>34.0</td>
<td>53.1</td>
<td>58.0</td>
<td>46.2</td>
</tr>
<tr>
<td>R-CNN fc7</td>
<td>57.6</td>
<td>57.9</td>
<td>38.5</td>
<td>31.8</td>
<td>23.7</td>
<td>51.2</td>
<td>58.9</td>
<td>51.4</td>
<td>20.0</td>
<td>50.5</td>
<td>40.9</td>
<td>46.0</td>
<td>51.6</td>
<td>55.9</td>
<td>43.3</td>
<td>23.3</td>
<td>48.1</td>
<td>35.3</td>
<td>51.0</td>
<td>57.4</td>
<td>44.7</td>
</tr>
<tr>
<td>R-CNN FT pool6</td>
<td>58.2</td>
<td>63.3</td>
<td>37.9</td>
<td>27.6</td>
<td>26.1</td>
<td>54.1</td>
<td>66.9</td>
<td>51.4</td>
<td>26.7</td>
<td>55.5</td>
<td>43.4</td>
<td>43.1</td>
<td>57.7</td>
<td>59.0</td>
<td>45.8</td>
<td>28.1</td>
<td>50.8</td>
<td>40.6</td>
<td>53.1</td>
<td>56.4</td>
<td>47.3</td>
</tr>
<tr>
<td>R-CNN FT fc6</td>
<td>63.5</td>
<td>66.0</td>
<td>47.9</td>
<td>37.7</td>
<td>29.9</td>
<td>62.5</td>
<td>70.2</td>
<td>60.2</td>
<td>32.0</td>
<td>57.9</td>
<td>47.0</td>
<td>53.5</td>
<td>60.1</td>
<td>64.2</td>
<td>52.2</td>
<td>31.3</td>
<td>55.0</td>
<td>50.0</td>
<td>57.7</td>
<td>63.0</td>
<td>53.1</td>
</tr>
<tr>
<td>R-CNN FT fc7</td>
<td>64.2</td>
<td>69.7</td>
<td>50.0</td>
<td>41.9</td>
<td>32.0</td>
<td>62.6</td>
<td>71.0</td>
<td>60.7</td>
<td>32.7</td>
<td>58.5</td>
<td>46.5</td>
<td>56.1</td>
<td>60.6</td>
<td>66.8</td>
<td>54.2</td>
<td>31.5</td>
<td>52.8</td>
<td>48.9</td>
<td>57.9</td>
<td>64.7</td>
<td>54.2</td>
</tr>
<tr>
<td>R-CNN FT fc7 BB</td>
<td>68.1</td>
<td>72.8</td>
<td>56.8</td>
<td>43.0</td>
<td>36.8</td>
<td>66.3</td>
<td>74.2</td>
<td>67.6</td>
<td>34.4</td>
<td>63.5</td>
<td>54.5</td>
<td>61.2</td>
<td>69.1</td>
<td>68.6</td>
<td>58.7</td>
<td>33.4</td>
<td>62.9</td>
<td>51.1</td>
<td>62.5</td>
<td>64.8</td>
<td>58.5</td>
</tr>
<tr>
<td>DPM v5 [20]</td>
<td>33.2</td>
<td>60.3</td>
<td>10.2</td>
<td>16.1</td>
<td>27.3</td>
<td>54.3</td>
<td>58.2</td>
<td>23.0</td>
<td>20.0</td>
<td>24.1</td>
<td>26.7</td>
<td>12.7</td>
<td>58.1</td>
<td>48.2</td>
<td>43.2</td>
<td>12.0</td>
<td>21.1</td>
<td>36.1</td>
<td>46.0</td>
<td>43.5</td>
<td>33.7</td>
</tr>
<tr>
<td>DPM ST [28]</td>
<td>23.8</td>
<td>58.2</td>
<td>10.5</td>
<td>8.5</td>
<td>27.1</td>
<td>50.4</td>
<td>52.0</td>
<td>7.3</td>
<td>19.2</td>
<td>22.8</td>
<td>18.1</td>
<td>8.0</td>
<td>55.9</td>
<td>44.8</td>
<td>32.4</td>
<td>13.3</td>
<td>15.9</td>
<td>22.8</td>
<td>46.2</td>
<td>44.9</td>
<td>29.1</td>
</tr>
<tr>
<td>DPM HSC [31]</td>
<td>32.2</td>
<td>58.3</td>
<td>11.5</td>
<td>16.3</td>
<td>30.6</td>
<td>49.9</td>
<td>54.8</td>
<td>23.5</td>
<td>21.5</td>
<td>27.7</td>
<td>34.0</td>
<td>13.7</td>
<td>58.1</td>
<td>51.6</td>
<td>39.9</td>
<td>12.4</td>
<td>23.5</td>
<td>34.4</td>
<td>47.4</td>
<td>45.2</td>
<td>34.3</td>
</tr>
</tbody>
</table>
1. The ILSVRC2013 detection dataset is split into three sets: train (395,918), val (20,121), and test (40,152) in 200 classes.

2. The training images are not exhaustively annotated, so train images cannot be used for hard negative mining, so use val for both training and validation.

3. Training data is required for three procedures in R-CNN: (1) CNN fine-tuning, (2) detector SVM training (3) bounding-box regressor training.

<table>
<thead>
<tr>
<th>test set</th>
<th>val₂</th>
<th>val₂</th>
<th>val₂</th>
<th>val₂</th>
<th>val₂</th>
<th>val₂</th>
<th>val₂</th>
<th>test</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM training set</td>
<td>val₁</td>
<td>val₁+train₁₅k</td>
<td>val₁+train₁k</td>
<td>val₁+train₁k</td>
<td>val₁+train₁k</td>
<td>val₁+train₁k</td>
<td>val₁+train₁k</td>
<td>test</td>
<td>test</td>
</tr>
<tr>
<td>CNN fine-tuning set</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>val₁+train₁k</td>
<td>val₁+train₁k</td>
</tr>
<tr>
<td>bbox reg set</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>val</td>
</tr>
<tr>
<td>CNN feature layer</td>
<td>fc₆</td>
<td>fc₆</td>
<td>fc₆</td>
<td>fc₇</td>
<td>fc₇</td>
<td>fc₇</td>
<td>fc₇</td>
<td>fc₇</td>
<td>fc₇</td>
</tr>
<tr>
<td>mAP</td>
<td>20.9</td>
<td>24.1</td>
<td>24.1</td>
<td>26.5</td>
<td>29.7</td>
<td>31.0</td>
<td>30.2</td>
<td>31.4</td>
<td></td>
</tr>
<tr>
<td>median AP</td>
<td>17.7</td>
<td>21.0</td>
<td>21.4</td>
<td>24.8</td>
<td>29.2</td>
<td>29.6</td>
<td>29.0</td>
<td>30.3</td>
<td></td>
</tr>
</tbody>
</table>
(A) the original object proposal at its actual scale relative to the transformed CNN inputs; (B) tightest square with context; (C) tightest square without context; (D) warp.

Warping with context padding (p = 16 pixels) outperformed the warped by a large margin (3-5 mAP points).
1 **Purpose** improve localization performance of the bounding-box.

2 **Approach** Predict a new bounding box for the detection using a class-specific bounding-box regressor after scoring each selective search proposal with a class-specific detection SVM.

3 **Algorithm** The input is a set of N training pairs \( \{(P^i, G^i)\} \) \(i=1,2,...,N\), where \( P^i = (P^i_x, P^i_y, P^i_\omega, P^i_h) \) specifies the pixel coordinates of the center of proposal \( P^i \)’s bounding box together with \( P^i \)’s width and height in pixels. Each ground-truth bounding box \( G \) is specified as: \( G = (G_x, G_y, G_\omega, G_h) \). The goal is to learn a transformation that maps a proposed box \( P \) to a ground-truth box \( G \).
Bounding-box Regression

Set the transformation as: \( d_x(P), d_y(P), d_\omega(P), d_h(P) \). First specify translation of the center of \( P \):

\[
\hat{G}_x = P_\omega d_x(P) + P_x \\
\hat{G}_y = P_h d_y(P) + P_x
\]

Then specify log-space translations of the width and height of \( P \)'s bounding box:

\[
\hat{G}_\omega = P_\omega \exp(d_\omega) \\
\hat{G}_h = P_h \exp(d_h)
\]

Each function \( d_*(P) \) (where * is one of \( x,y,h,w \)) can be obtained by a linear function from the pool5 features of proposal \( \phi_5(P) \):

\[
d_*(P) = w_*^T \phi_5(P), \text{ where } w_* \text{ is a vector of learnable model parameters.} 
\]
We learn \( w_* \) by optimizing the regularized least squares objective (ridge regression):

\[
    w_* = \text{argmin}_{\hat{w}_*} \sum_i^N (t_i^* - \hat{w}_* \phi_5(P_i))^2 + \lambda ||\hat{w}_*||^2
\]

The regression targets \( t_* \) are defined as:

\[
    t_x = (G_x - P_x)/P_\omega \\
    t_y = (G_y - P_y)/P_h \\
    t_\omega = \log(G_\omega/P_\omega) \\
    t_h = \log(G_h/P_h)
\]