Category-level recognition

1. Classification
   - Contains a car? [yes/no]
   - List categories present
   - Which city is this from?

2. Detection
   - Localize horses (if present)
   - Segment people (if present)
   - Parse objects into parts
Semantic Image Segmentation
Datasets

- **PASCAL VOC segmentation**
  - O(10K) images, 20 classes + bgnd
  - Also bounding box annotations

- **MS COCO**
  - O(100K) images, 80 classes + bgnd
  - Also 5 text captions / image
Applications

- Fine-grained image recognition
  - Explicit localization
  - More natural description of “stuff”

- Image manipulation and editing
System Overview

Input

Deep Convolutional Neural Network

Aeroplane Coarse Score map

Final Output

Fully Connected CRF

Bi-linear Interpolation
Basic Ingredients: (1) Conv Nets

- Train convnet to predict label of center pixel
- Apply in sliding window fashion

See also: J Long, E Shelhamer, T Darrell: Fully Convolutional Networks for Semantic Segmentation (arXiv)
The accuracy/localization tradeoff

- Large CNN receptive field
  → poor performance near boundaries
Explicit control of receptive field size

- Reduce RF size by conv layer manipulation
- In VGG: Subsample first FC layer 7x7 → 3x3
Explicit control of response density

● Decrease score map stride: 32 → 8
● Efficient implementation with “atrous” algorithm
Accurate Boundary Recovery w. CRF

Raw score maps

After dense CRF
CRFs in a nutshell

- a set of i.i.d. samples $D = \{(x^n, y^n)\}_{n=1,...,N}$, $(x^n, y^n) \sim d(x, y)$
- feature functions $(\phi_1(x, y), \ldots, \phi_D(x, y)) \equiv \phi(x, y)$
- parametrized family $p(y|x, w) = \frac{1}{Z(x,w)} \exp(\langle w, \phi(x, y) \rangle)$

$$E(x) = \sum_i \underbrace{\psi_u(x_i)}_{\text{unary term}} + \sum_i \sum_{j \in N_i} \underbrace{\psi_p(x_i, x_j)}_{\text{pairwise term}}$$

- **Unary term**
  - From classifier
  - TextonBoost [Shotton et al. 09]

- **Pairwise term**
  - Consistent labeling
Grid CRF

\[ E(x) = \sum_i \psi_u(x_i) + \sum_i \sum_{j \in N_i} \psi_p(x_i, x_j) \]

- Efficient inference
  - 1 second for 50,000 variables
- Limited expressive power
- Only local interactions
- Excessive smoothing of object boundaries
  - Shrinking bias
Grid CRF limitations

\[ E(x) = \sum_i \psi_u(x_i) + \sum_i \sum_{j \in N_i} \psi_p(x_i, x_j) \]

- Efficient inference
  - 1 second for 50,000 variables
- Limited expressive power
- Only local interactions
- Excessive smoothing of object boundaries
  - Shrinking bias
Grid CRF limitations

\[ E(x) = \sum_i \psi_u(x_i) + \sum_i \sum_{j \in N_i} \psi_p(x_i, x_j) \]

- Efficient inference
  - 1 second for 50,000 variables
- Limited expressive power
- Only local interactions
- Excessive smoothing of object boundaries
  - Shrinking bias
2011: Fully-connected CRF (Krahnebühl & Koltun)

\[ E(x) = \sum_i \psi_u(x_i) + \sum_i \sum_{j<i} \psi_p(x_i, x_j) \]

Every node is connected to every other node
- Connections weighted differently

P Krähenbühl and V Koltun, Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials, NIPS 2011
Fully-connected CRF

\[ E(x) = \sum_i \psi_u(x_i) + \sum_i \sum_{j>i} \psi_p(x_i, x_j) \]

- Long-range interactions
- No more shrinking bias
Fully-connected CRF

\[ E(x) = \sum_i \psi_u(x_i) + \sum_i \sum_{j > i} \psi_p(x_i, x_j) \]

- Long-range interactions
- No more shrinking bias
Fully-connected CRF

\[ E(x) = \sum_i \psi_u(x_i) + \sum_i \sum_{j>i} \psi_p(x_i, x_j) \]

- Long-range interactions
- No more shrinking bias
Fully-connected CRF: FAST

- Inference in 0.2 seconds
  - 50’000 variables
  - MCMC inference: 36 hrs
- Pairwise potentials: linear combinations of Gaussians

**MSRC dataset**
- 591 images
- 21 classes

<table>
<thead>
<tr>
<th></th>
<th>Time</th>
<th>Global</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unary</td>
<td>-</td>
<td>84.0</td>
<td>76.6</td>
</tr>
<tr>
<td>Grid CRF</td>
<td>1s</td>
<td>84.6</td>
<td>77.2</td>
</tr>
<tr>
<td>FC CRF</td>
<td>0.2s</td>
<td>86.0</td>
<td>78.3</td>
</tr>
</tbody>
</table>

How? Mean Field + some tricks
Trick: Pairwise Term

\[ \theta_{ij}(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^{K} w_m \cdot k^m(f_i, f_j) \]

Pots model

\[ \mu(x_i, x_j) = 1 \text{ if } x_i \neq x_j \]

\[ w_1 \exp \left( - \frac{||p_i - p_j||^2}{2\sigma^2_A} - \frac{||I_i - I_j||^2}{2\sigma^2_B} \right) + w_2 \exp \left( - \frac{||p_i - p_j||^2}{2\sigma^2_Y} \right) \]

Gaussian kernels

\[ Q_i(x_i = l) = \frac{1}{Z_i} \exp \left\{ -\psi_u(x_i) - \sum_{l' \in \mathcal{L}} \mu(l, l') \sum_{m=1}^{K} \sum_{j \neq i} w^{(m)} k^{(m)}(f_i, f_j) Q_j(l') \right\} \]

Fast summation through separable convolution

- Initialize \( Q_i(x_i) \leftarrow \frac{1}{Z_i} \exp\{-\phi_u(x_i)\} \)
- while not converged
  - Message passing: \( \tilde{Q}_i^{(m)}(l) \leftarrow \sum_{j \neq i} k^{(m)}(f_i, f_j) Q_j(l) \)
  - Compatibility transform: \( \hat{Q}_i(x_i) \leftarrow \sum_{l \in \mathcal{L}} \mu^{(m)}(x_i, l) \sum_m w^{(m)} \tilde{Q}_i^{(m)}(l) \)
  - Local update: \( Q_i(x_i) \leftarrow \exp\{-\psi_u(x_i) - \hat{Q}_i(x_i)\} \)
  - Normalize: \( Q_i(x_i) \)

P Krähenbühl, V Koltun, NIPS 2011
2014: Fully connected CRFs + Deep Classifiers

L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. Yuille
Evolution from mean field updates

Figure 1: Score map (input before softmax function) and belief map (output of softmax function) for Aeroplane. We show the score (1st row) and belief (2nd row) maps after each mean field iteration. The output of the last DCNN layer is used as input to the mean field inference method.
Our Results (input, DCNN, CRF-DCNN)
Our Results (input, DCNN, CRF-DCNN)
Comparisons to other techniques on VOC test

<table>
<thead>
<tr>
<th>Technique</th>
<th>mean</th>
<th>aero</th>
<th>plane</th>
<th>bicycle</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>dining</th>
<th>table</th>
<th>dog</th>
<th>horse</th>
<th>motor</th>
<th>bike</th>
<th>person</th>
<th>potted</th>
<th>plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tv/monitor</th>
<th>submission date</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepLab-CRF-MSc</td>
<td>67.1</td>
<td>60.4</td>
<td>36.8</td>
<td>77.4</td>
<td>55.2</td>
<td>66.4</td>
<td>81.5</td>
<td>77.8</td>
<td>78.0</td>
<td>27.1</td>
<td>68.2</td>
<td>52.7</td>
<td>74.3</td>
<td>69.6</td>
<td>79.4</td>
<td>79.0</td>
<td>56.9</td>
<td>78.8</td>
<td>45.2</td>
<td>72.7</td>
<td>59.3</td>
<td>30-Dec-2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DeepLab-CRF</td>
<td>60.4</td>
<td>78.4</td>
<td>33.1</td>
<td>78.2</td>
<td>55.6</td>
<td>65.3</td>
<td>81.3</td>
<td>75.5</td>
<td>78.0</td>
<td>25.3</td>
<td>69.2</td>
<td>52.7</td>
<td>75.2</td>
<td>69.0</td>
<td>79.1</td>
<td>77.6</td>
<td>54.7</td>
<td>78.3</td>
<td>45.1</td>
<td>73.3</td>
<td>56.2</td>
<td>23-Dec-2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTI_zoomout_16</td>
<td>64.4</td>
<td>81.9</td>
<td>35.1</td>
<td>78.2</td>
<td>57.4</td>
<td>56.5</td>
<td>80.5</td>
<td>74.0</td>
<td>79.8</td>
<td>22.4</td>
<td>69.6</td>
<td>53.7</td>
<td>74.0</td>
<td>76.0</td>
<td>76.6</td>
<td>76.8</td>
<td>54.3</td>
<td>70.2</td>
<td>40.2</td>
<td>68.9</td>
<td>55.3</td>
<td>24-Nov-2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FCN-8s</td>
<td>62.2</td>
<td>76.8</td>
<td>34.2</td>
<td>69.0</td>
<td>49.4</td>
<td>60.3</td>
<td>75.3</td>
<td>74.7</td>
<td>77.6</td>
<td>21.4</td>
<td>62.5</td>
<td>46.8</td>
<td>71.8</td>
<td>63.9</td>
<td>76.5</td>
<td>73.0</td>
<td>54.2</td>
<td>72.4</td>
<td>37.4</td>
<td>70.9</td>
<td>55.1</td>
<td>12-Nov-2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSRA_CFM</td>
<td>61.8</td>
<td>75.7</td>
<td>26.7</td>
<td>69.5</td>
<td>48.8</td>
<td>65.6</td>
<td>81.0</td>
<td>69.2</td>
<td>73.3</td>
<td>30.0</td>
<td>68.7</td>
<td>51.3</td>
<td>69.1</td>
<td>68.1</td>
<td>71.7</td>
<td>67.5</td>
<td>50.4</td>
<td>66.3</td>
<td>44.4</td>
<td>58.9</td>
<td>53.3</td>
<td>17-Dec-2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTI_zoomout</td>
<td>58.4</td>
<td>70.3</td>
<td>31.9</td>
<td>68.3</td>
<td>46.4</td>
<td>52.1</td>
<td>75.3</td>
<td>68.4</td>
<td>75.3</td>
<td>19.2</td>
<td>58.4</td>
<td>49.9</td>
<td>69.6</td>
<td>63.0</td>
<td>70.1</td>
<td>67.6</td>
<td>41.5</td>
<td>64.0</td>
<td>34.9</td>
<td>64.2</td>
<td>47.3</td>
<td>17-Nov-2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SDS</td>
<td>51.6</td>
<td>63.3</td>
<td>25.7</td>
<td>63.0</td>
<td>39.8</td>
<td>59.2</td>
<td>70.9</td>
<td>61.4</td>
<td>54.9</td>
<td>16.8</td>
<td>45.0</td>
<td>48.2</td>
<td>50.5</td>
<td>51.0</td>
<td>57.7</td>
<td>63.3</td>
<td>31.8</td>
<td>58.7</td>
<td>31.2</td>
<td>55.7</td>
<td>48.5</td>
<td>21-Jul-2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NUS_UDS</td>
<td>50.0</td>
<td>67.0</td>
<td>24.5</td>
<td>47.2</td>
<td>45.0</td>
<td>47.9</td>
<td>65.3</td>
<td>60.6</td>
<td>58.5</td>
<td>15.5</td>
<td>58.8</td>
<td>37.4</td>
<td>45.8</td>
<td>59.9</td>
<td>62.0</td>
<td>52.7</td>
<td>40.8</td>
<td>48.2</td>
<td>36.8</td>
<td>53.1</td>
<td>45.6</td>
<td>29-Oct-2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTC_nbest-rerank</td>
<td>48.1</td>
<td>62.7</td>
<td>25.6</td>
<td>46.9</td>
<td>43.0</td>
<td>34.8</td>
<td>58.4</td>
<td>58.8</td>
<td>55.0</td>
<td>14.6</td>
<td>47.5</td>
<td>31.2</td>
<td>44.7</td>
<td>51.0</td>
<td>60.9</td>
<td>53.3</td>
<td>36.6</td>
<td>50.9</td>
<td>30.1</td>
<td>50.2</td>
<td>46.8</td>
<td>15-Nov-2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BONN_CPVCPCMC</td>
<td>47.8</td>
<td>64.0</td>
<td>27.3</td>
<td>54.1</td>
<td>30.2</td>
<td>48.7</td>
<td>56.6</td>
<td>57.7</td>
<td>52.5</td>
<td>14.2</td>
<td>54.8</td>
<td>29.6</td>
<td>42.2</td>
<td>58.0</td>
<td>54.8</td>
<td>50.2</td>
<td>36.6</td>
<td>58.6</td>
<td>31.6</td>
<td>48.4</td>
<td>38.6</td>
<td>08-Aug-2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BONN_cpvcpcmc</td>
<td>47.5</td>
<td>63.4</td>
<td>27.3</td>
<td>56.1</td>
<td>37.7</td>
<td>47.2</td>
<td>57.9</td>
<td>59.3</td>
<td>55.0</td>
<td>11.5</td>
<td>50.8</td>
<td>30.5</td>
<td>45.0</td>
<td>58.4</td>
<td>57.4</td>
<td>48.6</td>
<td>34.6</td>
<td>53.3</td>
<td>32.4</td>
<td>47.6</td>
<td>39.2</td>
<td>23-Sep-2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BONN_GCPVCPC</td>
<td>46.8</td>
<td>63.6</td>
<td>26.8</td>
<td>45.6</td>
<td>41.7</td>
<td>47.1</td>
<td>54.3</td>
<td>58.6</td>
<td>55.1</td>
<td>14.5</td>
<td>49.0</td>
<td>20.9</td>
<td>46.1</td>
<td>52.6</td>
<td>58.2</td>
<td>53.4</td>
<td>32.0</td>
<td>44.5</td>
<td>34.6</td>
<td>45.3</td>
<td>43.1</td>
<td>23-Sep-2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BONN_CMPCPC</td>
<td>46.7</td>
<td>63.9</td>
<td>23.8</td>
<td>44.6</td>
<td>40.3</td>
<td>45.5</td>
<td>59.6</td>
<td>58.7</td>
<td>57.1</td>
<td>11.7</td>
<td>45.9</td>
<td>34.9</td>
<td>43.0</td>
<td>54.9</td>
<td>58.0</td>
<td>51.5</td>
<td>34.6</td>
<td>44.1</td>
<td>29.9</td>
<td>50.5</td>
<td>44.5</td>
<td>23-Sep-2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Pre-CNN: Up to 50%
CNN: 60-64%
CNN + CRF: >67%
More data helps

- Pre-train on MS-COCO, refine in PASCAL:
  - Pascal Train: 67%
  - Coco + Pascal: 71%

- Preliminary eval on COCO: ~40% mean IoU
Comparisons to previous state-of-the-art

(b) TTI-Zoomout-16 vs. DeepLab-CRF

(a) FCN-8s vs. DeepLab-CRF
Towards Weaker Annotations

G. Papandreou, L.-C. Chen, K. Murphy and A. Yuille
Weaker Annotations: Bounding Boxes
Weaker Annotations: Image Level

Expectation-Maximization during net training
Weaker Annotations: Hybrid Approach
Weak Annotation Pascal Results

Image
39%

BBox (Base/Grab)
54.2% / 60.4%

Hybrid (1.5K / 3K)
63.5% / 66.4%

Strong (10K)
66.4%