Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, Alan L. Yuille
ICLR 2015

Carlos Feres & Mark Weber
Outline

● Context

● Dense labeling challenge
  ○ Increasing resolution
  ○ Controlling the receptive field size

● Localization challenge
  ○ Fully-connected CRFs
  ○ Multi-scale prediction

● Experiments

● Comments
Can we use a CNN for this task?

Context

- CNNs excel at classification/detection
  - Spatial invariance helps build hierarchical abstractions of data

- CNNs perform poorly on low-level tasks
  - Needs precise location and details

- Reduction of signal resolution
  - Information is not “dense”

- Spatial insensitivity (invariance)
  - Exact outline of objects is lost
Dense labeling challenge

- Want CNNs with **high spatial resolution & wide receptive field**

- Increase spatial resolution
  - *Hole algorithm*

- Adapt classification CNN for segmentation
  - Redefine the CNN architecture
  - Control the receptive field size
Increase spatial resolution

- How do CNNs lose resolution?

  - **Convolution**
    - 3x3 filters, stride = 2
    - loses “in-between” features

  - **Max-pooling**
    - 2x2 filters, stride = 2
    - loses 75% of activations

- Striding & max-pooling !!
Increase spatial resolution

- Possible solution: stride = 1 & remove pooling
  - Requires larger kernels for same receptive field

- Computationally expensive & possible overfitting
Increase spatial resolution

- **Hole algorithm (atrous, dilated convolution)**
  - Large, sparse kernel (zeros in between values) = upsampled small kernel
  - Complexity ~ # nonzero elements
  - Arbitrary size receptive field

![Diagram showing convolution and kernel sizes](image-url)
Increase spatial resolution

- Example in 2D [2]:

Adapting CNN for segmentation

VGG16

Proposed fully conv. CNN

Based on slides from deepsystems.io
Controlling receptive field size

Further adjustments in 1st FC layer:
- **DeepLab-CRF**
  - kernel = 4,
  - rate = 4,
  - 4096 filters
- **DeepLab-CRF-4x4**
  - kernel = 4,
  - rate = 8,
  - 4096 filters
- **DeepLab-CRF-LargeFOV**
  - kernel = 3,
  - rate = 12,
  - 1024 filters
  - same performance as 7x7!
Dense labeling summary
Localization challenge

Score map (before softmax)

Belief map (after softmax)

Object is detected
Outline is inaccurate
Conditional Random Fields

- A set of variables $X$ (one per pixel)
- Image $I$
- We look for maximum a posteriori (MAP):

$$x^* = \arg \max_x P(x | I)$$

- Gibbs distribution:

$$P(X|I) \approx \exp\{-E(X|I)\}$$

Fully-Connected CRFs

\[ P(X|I) \approx \exp\{-E(X|I)\} \]

- (Gibbs) Energy: \[ E(x) = \sum_i \phi(x_i) + \sum_{i<j} \psi(x_i, x_j) \]
Gibbs Energy

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

- Unary term \( \phi(x_i) \)
  - CNN output

- Pairwise term \( \psi(x_i, x_j) \)
  - Gaussian kernel for appearance
    \[ k^{(1)}(f_i, f_j) \]
  - Gaussian kernel for smoothness
    \[ k^{(2)}(f_i, f_j) \]
Mean Field Approximation

- Computing $P(X)$ is hard
- Choose family of distribution such that:

$$Q(X) = \prod_{i} Q_i(X_i)$$

- Find $Q(X)$ that approximates $P(X)$ best

$\Rightarrow$ Minimize KL-Divergence

$$KL(Q\|P)$$
Update rule

- Derived from KL-Divergence:

\[ Q_i(x_i = l) = \frac{1}{Z_i} \exp\{-\phi(x_i) - \sum_{l' \neq l} \sum_{m=1}^{2} w^{(m)} \sum_{i \neq j} k^{(m)}(f_i, f_j) Q_j(l')\} \]

\[ [G^{(m)} \otimes Q(l')](f_i) - Q_i(l') \]
Gaussian Approximation

\[ Q_i = \ldots [G^{(m)} \otimes Q(l')](f_i) - Q_i(l') \]

Issue:

- Kernel size == Image size

\[ O(|I|^2) \]

New runtime:

- Kernel Size == constant

\[ O(|I|) \]
CRF Summary

- Post-processing step
- Fully connected CRF consists of
  - Unary terms (CNN outputs)
  - Pairwise terms (encode global context)
- Mean Field Approximation for $Q(X)$
- Iterate until convergence
CRF Results

Image/G.T.  DCNN output  CRF Iteration 1  CRF Iteration 2  CRF Iteration 10
Complete model (DeepLab v1)

Input → adapted for segmentation → Deep Convolutional Neural Network → Aeroplane Coarse Score map → Bi-linear Interpolation → upsampling image to original resolution

Input

Deep Convolutional Neural Network

Aeroplane Coarse Score map

Bi-linear Interpolation

upsample image to original resolution

Final Output

refine segmentation

Fully Connected CRF
Extension: Multiscale-prediction
Experiments

<table>
<thead>
<tr>
<th>Method</th>
<th>mean IOU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepLab</td>
<td>59.80</td>
</tr>
<tr>
<td>DeepLab-CRF</td>
<td>63.74</td>
</tr>
<tr>
<td>DeepLab-MSc</td>
<td>61.30</td>
</tr>
<tr>
<td>DeepLab-MSc-CRF</td>
<td>65.21</td>
</tr>
<tr>
<td>DeepLab-7x7</td>
<td>64.38</td>
</tr>
<tr>
<td>DeepLab-CRF-7x7</td>
<td>67.64</td>
</tr>
<tr>
<td>DeepLab-LargeFOV</td>
<td>62.25</td>
</tr>
<tr>
<td>DeepLab-CRF-LargeFOV</td>
<td>67.64</td>
</tr>
<tr>
<td>DeepLab-MSc-LargeFOV</td>
<td>64.21</td>
</tr>
<tr>
<td>DeepLab-MSc-CRF-LargeFOV</td>
<td>68.70</td>
</tr>
</tbody>
</table>

PASCAL VOC 2012 'val' (trained in augmented 'train')

<table>
<thead>
<tr>
<th>Method</th>
<th>mean IOU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSRA-CFM</td>
<td>61.8</td>
</tr>
<tr>
<td>FCN-8s</td>
<td>62.2</td>
</tr>
<tr>
<td>TTI-Zoomout-16</td>
<td>64.4</td>
</tr>
<tr>
<td>DeepLab-CRF</td>
<td>66.4</td>
</tr>
<tr>
<td>DeepLab-MSc-CRF</td>
<td>67.1</td>
</tr>
<tr>
<td>DeepLab-CRF-7x7</td>
<td>70.3</td>
</tr>
<tr>
<td>DeepLab-CRF-LargeFOV</td>
<td>70.3</td>
</tr>
<tr>
<td>DeepLab-MSc-CRF-LargeFOV</td>
<td>71.6</td>
</tr>
</tbody>
</table>

PASCAL VOC 2012 'test' (trained in augmented 'trainval')

Experiments

- Field of view
  - Modifications to first FC layer

<table>
<thead>
<tr>
<th>Method</th>
<th>kernel size</th>
<th>input stride</th>
<th>receptive field</th>
<th># parameters</th>
<th>mean IOU (%)</th>
<th>Training speed (img/sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepLab-CRF-7x7</td>
<td>7 × 7</td>
<td>4</td>
<td>224</td>
<td>134.3M</td>
<td>67.64</td>
<td>1.44</td>
</tr>
<tr>
<td>DeepLab-CRF</td>
<td>4 × 4</td>
<td>4</td>
<td>128</td>
<td>65.1M</td>
<td>63.74</td>
<td>2.90</td>
</tr>
<tr>
<td>DeepLab-CRF-4x4</td>
<td>4 × 4</td>
<td>8</td>
<td>224</td>
<td>65.1M</td>
<td>67.14</td>
<td>2.90</td>
</tr>
<tr>
<td>DeepLab-CRF-LargeFOV</td>
<td>3 × 3</td>
<td>12</td>
<td>224</td>
<td>20.5M</td>
<td>67.64</td>
<td>4.84</td>
</tr>
</tbody>
</table>

Training in PASCAL VOC 2012 ‘val’
Experiments

- CRF

image

ground-truth

DeepLab-CRF
Experiments

- CRF: Success
Experiments

- CRF: Success
Experiments

- CRF: Failures
Experiments

- Labeling IoU

<table>
<thead>
<tr>
<th>Method</th>
<th>bkg</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>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSRA-CF/M</td>
<td>-</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>51.5</td>
<td>69.1</td>
<td>68.1</td>
<td>71.7</td>
<td>67.5</td>
<td>50.4</td>
<td>66.5</td>
<td>44.4</td>
<td>58.9</td>
<td>53.5</td>
<td>61.8</td>
</tr>
<tr>
<td>FCN-8s</td>
<td>-</td>
<td>76.8</td>
<td>34.2</td>
<td>68.9</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>73.9</td>
<td>45.2</td>
<td>72.4</td>
<td>37.4</td>
<td>70.9</td>
<td>55.1</td>
<td>62.2</td>
</tr>
<tr>
<td>TTI-Zoomout-16</td>
<td>89.8</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>68.8</td>
<td>44.3</td>
<td>70.2</td>
<td>40.2</td>
<td>68.9</td>
<td>55.3</td>
<td>64.4</td>
</tr>
<tr>
<td>DeepLab-CRF</td>
<td>92.1</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>78.6</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>66.4</td>
</tr>
<tr>
<td>DeepLab-MSc-CRF</td>
<td>92.6</td>
<td>80.4</td>
<td>36.8</td>
<td>77.4</td>
<td>55.2</td>
<td>66.4</td>
<td>81.5</td>
<td>77.5</td>
<td>78.9</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>67.1</td>
</tr>
<tr>
<td>DeepLab-CRF-7x7</td>
<td>92.8</td>
<td>83.9</td>
<td>36.6</td>
<td>77.5</td>
<td>58.4</td>
<td>68.0</td>
<td>84.6</td>
<td>79.7</td>
<td>83.1</td>
<td>29.5</td>
<td>74.6</td>
<td>59.3</td>
<td>78.9</td>
<td>76.0</td>
<td>82.1</td>
<td>80.6</td>
<td>60.3</td>
<td>81.7</td>
<td>49.2</td>
<td>78.0</td>
<td>60.7</td>
<td>70.3</td>
</tr>
<tr>
<td>DeepLab-CRF-LargeFOV</td>
<td>92.6</td>
<td>83.5</td>
<td>36.6</td>
<td>82.5</td>
<td>62.3</td>
<td>66.5</td>
<td>85.4</td>
<td>78.5</td>
<td>83.7</td>
<td>30.4</td>
<td>72.9</td>
<td>60.4</td>
<td>78.5</td>
<td>75.5</td>
<td>82.1</td>
<td>79.7</td>
<td>58.2</td>
<td>82.0</td>
<td>48.8</td>
<td>73.7</td>
<td>63.3</td>
<td>70.3</td>
</tr>
<tr>
<td>DeepLab-MSc-CRF-LargeFOV</td>
<td>93.1</td>
<td>84.4</td>
<td>54.5</td>
<td>81.5</td>
<td>63.6</td>
<td>65.9</td>
<td>85.1</td>
<td>79.1</td>
<td>83.4</td>
<td>30.7</td>
<td>74.1</td>
<td>59.8</td>
<td>79.0</td>
<td>76.1</td>
<td>83.2</td>
<td>80.8</td>
<td>59.7</td>
<td>82.2</td>
<td>50.4</td>
<td>73.1</td>
<td>63.7</td>
<td>71.6</td>
</tr>
</tbody>
</table>

PASCAL VOC 2012 ‘test’ (trained with ‘trainval’)
Comments

● **Strengths**
  ○ Template to adapt classification CNNs for image segmentation
  ○ Successful incorporation of methods from other disciplines
  ○ Experiments with each individual modification allows to compare impact

● **Weaknesses**
  ○ Two processes: CNN + postprocessing
  ○ Paper is not self contained (need references a lot)
  ○ Key concepts are not well explained (better in Deeplab v2 [2])

Thanks!

Questions?
Appendix: Experiments

- **Object boundaries**
  - Void label (regularly seen in boundaries)
  - Mean IoU of pixels in narrow band (trimap)

![image](image.png) ![ground-truth](ground-truth.png)

- trimap 2px
- trimap 10px

![Pixelwise Accuracy](accuracy.png)
![mean IOU](iou.png)

- DL-MSc-CRF
- DeepLab-CRF
- DeepLab-MSc
- DeepLab
Appendix: Experiments

- Multi-scale features
  - Modify kernel size of first FC layer and rate