Learning Deconvolution Network for Semantic Segmentation

Hyeonwoo Noh, Seunghoon Hong, Bohyung Han

Mehmet Günel
What is this paper about?

- A novel semantic segmentation algorithm
- Convolution & Deconvolution layers
- Fully convolutional network integrated with deep deconvolution network and makes proposal-wise prediction
- Identifies detailed structures and handles objects in multiple scales naturally
Overview - What is and what is not

• Semantic segmentation
  – Scene labeling
  – Pixel-wise classification

Semantically meaningful parts + classify each part into predetermined classes

Classify each pixel!
Problem: Background

• Semantic segmentation algorithms are often formulated to solve structured pixel-wise labeling problems based on CNN

• Conditional random field (CRF) is optionally applied to the output map for fine segmentation

• Network accepts a whole image as an input and performs fast and accurate inference
Problem: Limitations

- Fixed-size receptive field
  
  The object that is substantially larger or smaller than the receptive field may be fragmented or mislabeled.

  Small objects are often ignored and classified as background.
Problem: Limitations

(a) Inconsistent labels due to large object size
Problem: Limitations

(b) Missing labels due to small object size
Related Work


Object proposals
Contributions

- A multi-layer deconvolution network, which is composed of deconvolution, unpooling, and rectified linear unit (ReLU) layers

- Free from scale issues found in FCN-based methods and identifies finer details of an object

- PASCAL VOC 2012 dataset best accuracy with FCN
Network Model

Approximately 252M parameters in total
Pooling & Unpooling

Example specific
Convolution & Deconvolution

Convolution

Deconvolution

Class specific
Training Stage

- **Batch Normalization**
  - Internal covariate shift problem

- **Two-stage Training**
  - crop object instances using ground-truth annotations
  - utilize object proposals to construct more challenging examples
Segmentation Maps Integration

Formula

\[ P(x, y, c) = \max_i G_i(x, y, c), \quad \forall i, \quad (1) \]

\[ P(x, y, c) = \sum_i G_i(x, y, c), \quad \forall i. \quad (2) \]
Experimental Setup

- PASCAL VOC 2012 segmentation dataset
- All training and validation images are used to train
- They used augmented segmentation annotations
  - Extend the bbox 1.2 times larger to include local context around the object
  - Object & background labeling
  - 250 × 250 input image randomly cropped to 224 × 224 with optional horizontal + flipping
  - The number of training examples is 0.2M and 2.7M in the first and the second stage
Experimental Setup

- Caffe framework
- Stochastic gradient descent with momentum
- Initial learning rate, momentum and weight; 0.01, 0.9 and 0.0005
- VGG 16-layer net pre-trained on ILSVRC
- Network converges after approximately 20K and 40K SGD iterations with mini-batch of 64 samples
- Training takes 6 days (2 days for the first stage and 4 days for the second stage)
- Nvidia GTX Titan X GPU with 12G memory
Inference

• For each testing image, we generate approximately 2000 object proposals, and select top 50 proposals based on their objectness scores

• Compute pixel-wise maximum to aggregate proposal-wise predictions
Evaluation Metrics

- *comp6* evaluation protocol;
  - intersection over Union (IoU) between ground truth and predicted segmentations
Visualization of activations
Visualization of activations
Visualization of activations
Visualization of activations
Visualization of activations
Visualization of activations
Visualization of activations
Visualization of activations
Visualization of activations
Visualization of activations
Results

- CRF increase approximately 1% point

- Ensemble with FCN-8s improves mean IoU about 10.3% and 3.1% point with respect to FCN-8s and DeconvNet
## Results - Comparisons

Evaluation results on PASCAL VOC 2012 test set. (algorithms trained without additional data)

<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>mbk</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>EDeconvNet+CRF</td>
<td>93.1</td>
<td>89.9</td>
<td>39.3</td>
<td>79.7</td>
<td>63.9</td>
<td>68.2</td>
<td>87.4</td>
<td>81.2</td>
<td>86.1</td>
<td>28.5</td>
<td>77.0</td>
<td>62.0</td>
<td>79.0</td>
<td>80.3</td>
<td>83.6</td>
<td>80.2</td>
<td>58.8</td>
<td>83.4</td>
<td>54.3</td>
<td>80.7</td>
<td>65.0</td>
<td>72.5</td>
</tr>
<tr>
<td>DeepLab-CRF</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>
<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>FCN8s</td>
<td>91.2</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>MSRA-CFM</td>
<td>87.7</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>Hypercolumn</td>
<td>88.9</td>
<td>68.4</td>
<td>27.2</td>
<td>68.2</td>
<td>47.6</td>
<td>61.7</td>
<td>76.9</td>
<td>72.1</td>
<td>71.1</td>
<td>24.3</td>
<td>59.3</td>
<td>44.8</td>
<td>62.7</td>
<td>59.4</td>
<td>73.5</td>
<td>70.6</td>
<td>52.0</td>
<td>63.0</td>
<td>38.1</td>
<td>60.0</td>
<td>54.1</td>
<td>59.2</td>
</tr>
</tbody>
</table>
Results
Results - Strengths

Better results
Results - Strengths

(a) Input image  
(b) FCN-8s  
(c) Ours
Results - Weakness

Worse than FCN results
Results

Ensemble results
Conclusions & Future Directions

- A novel semantic segmentation algorithm by learning a deconvolution network
- Elimination of fixed-size receptive field limit in the fully convolutional network
- Ensemble approach of FCN + CRF
- State-of-the-art performance in PASCAL VOC 2012 without external data
- A bigger network with better proposals