Fully Convolutional Networks for Semantic Segmentation

Jonathan Long* Evan Shelhamer* Trevor Darrell
UC Berkeley
Semantic Segmentation

- what kind of thing is each pixel part of?
- what kind of stuff is each pixel?

Challenges
- tension between recognition and localization
- amount of computation
Segmentation: PASCAL VOC

**Leaderboard**

<table>
<thead>
<tr>
<th>Model</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSRA_BoxSup</td>
<td>75.2</td>
</tr>
<tr>
<td>Oxford_TVCGRNN_COCO</td>
<td>74.7</td>
</tr>
<tr>
<td>DeepLab-MSC-CRF-LargeFOV-COCO-CrossJoint</td>
<td>73.9</td>
</tr>
<tr>
<td>Adelaide_Context_CNN_CRF_VOC</td>
<td>72.9</td>
</tr>
<tr>
<td>DeepLab-CRF-COCO-LargeFOV</td>
<td>72.7</td>
</tr>
<tr>
<td>POSTECH_EDconvNet_CRF_VOC</td>
<td>72.5</td>
</tr>
<tr>
<td>Oxford_TVCGRNN_VOC</td>
<td>72.0</td>
</tr>
<tr>
<td>DeepLab-MSC-CRF-LargeFOV</td>
<td>71.6</td>
</tr>
<tr>
<td>MSRA_BoxSup</td>
<td>71.0</td>
</tr>
<tr>
<td>DeepLab-CRF-COCO-Strong</td>
<td>70.4</td>
</tr>
<tr>
<td>DeepLab-CRF-LargeFOV</td>
<td>70.3</td>
</tr>
<tr>
<td>TTL_zoomout_v2</td>
<td>69.6</td>
</tr>
<tr>
<td>DeepLab-CRF-MSC</td>
<td>67.1</td>
</tr>
<tr>
<td>DeepLab-CRF</td>
<td>66.4</td>
</tr>
<tr>
<td>CRF_RNN</td>
<td>65.2</td>
</tr>
<tr>
<td>TTL_zoomout_16</td>
<td>64.4</td>
</tr>
<tr>
<td>Hypercolumn</td>
<td>62.8</td>
</tr>
<tr>
<td>FCN-8s</td>
<td>62.2</td>
</tr>
<tr>
<td>MSRA_CFM</td>
<td>61.8</td>
</tr>
<tr>
<td>TTL_zoomout</td>
<td>58.4</td>
</tr>
<tr>
<td>SDS</td>
<td>51.6</td>
</tr>
<tr>
<td>NUS_UDS</td>
<td>50.0</td>
</tr>
<tr>
<td>TTYC-divmbest-rerank</td>
<td>48.1</td>
</tr>
<tr>
<td>Bonn_O2PCPMC_FGT_SEGM</td>
<td>47.8</td>
</tr>
<tr>
<td>Bonn_O2PCPMC_FGT_SEGM</td>
<td>47.5</td>
</tr>
<tr>
<td>BONNGC_O2P_CPMC_CSI</td>
<td>46.8</td>
</tr>
<tr>
<td>Bonn_CMBC_O2P_CPMC_LIN</td>
<td>46.7</td>
</tr>
</tbody>
</table>

**FCN:**
- pixelwise convnet
- state-of-the-art, in Caffe

Deep learning with Caffe: end-to-end networks lead to 50% relative improvement or 30 points absolute and >100x speedup in 1 year!
convnets perform classification

~1 millisecond

end-to-end learning

“tabby cat”

1000-dim vector

“tabby cat”
convnets perform segmentation?

~100 ms

end-to-end learning
a classification network

convolution

fully connected

“tabby cat”
becoming fully convolutional

convolution

227 × 227 55 × 55 27 × 27 13 × 13 1 × 1
becoming fully convolutional
upsampling output

convolution

H × W  H/4 × W/4  H/8 × W/8  H/16 × W/16  H/32 × W/32  H × W
end-to-end, pixels-to-pixels network
Relative to prior state-of-the-art SDS:

- 30% relative improvement in accuracy (67.2% on VOC 2012)
- 286× faster

*Simultaneous Detection and Segmentation
Hariharan et al. ECCV14
spectrum of deep features

combine *where* (local, shallow) with *what* (global, deep)

(future visualizations)

image

intermediate layers

fuse features into deep jet

(cf. Hariharan et al. CVPR15 “hypercolumn”)
skip layer refinement

input image  stride 32  stride 16  stride 8  ground truth

no skips  1 skip  2 skips
graphical model refinement

Input Image | FCN-8s | DeepLab | CRF-RNN | Ground Truth

DeepLab: Chen* & Papandreou* et al. ICLR 2015.

[ comparison credit: CRF as RNN, Zheng* & Jayasumana* et al. ICCV 2015 ]
nets for many pixelwise tasks

monocular depth estimation (Eigen & Fergus 2015)

semantic segmentation

boundary prediction (Xie & Tu 2015)

optical flow Fischer et al. 2015
fully convolutional networks are fast, end-to-end models for pixelwise problems

- **code** in Caffe master
- **models** for PASCAL VOC, NYUDv2, SIFT Flow, PASCAL-Context

[link to fcn.berkeleyvision.org]
[link to github.com/BVLC/caffe]

model example
inference example
solving example