Fully Convolutional Networks for Semantic Segmentation [1]
Jonathan Long, Evan Shelhamer and Trevor Darrell

ECE 289G: Paper Presentation
Philipp Gysel
Analyse Genome of *C Elegans*

1) Nucleus
2) Nucleus membrane
3) Cytoplasm
4) Cell wall
5) External medium

[2] and http://www.godandscience.org
PASCAL VOC 2011

Person

Motorbike

Chair
Accuracy Metric

- Mean intersection over union (IU): \( \frac{1}{n_{cl}} \sum_i n_{ii} / \left( t_i + \sum_j n_{ji} - n_{ii} \right) \)
- Pixel Accuracy: \( \frac{1}{n_{cl}} \sum_i n_{ii} / t_i \)
FCN: Fully Connected CNN

CNN:

FCN:

“Cat”

“Dog”
FCN Speedup

- Keep kernel sizes and strides
- Replace dense layer with convolution

Runtime of FCN vs naïve CNNs

Inference [ms]
Upsampling: Backwards strided convolution

In-network upsampling

pixelwise loss
Deep jet

<table>
<thead>
<tr>
<th></th>
<th>pixel acc.</th>
<th>mean IU</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN-32s-fixed</td>
<td>83.0</td>
<td>45.4</td>
</tr>
<tr>
<td>FCN-32s</td>
<td>89.1</td>
<td>59.4</td>
</tr>
<tr>
<td>FCN-16s</td>
<td>90.0</td>
<td>62.4</td>
</tr>
<tr>
<td>FCN-8s</td>
<td><strong>90.3</strong></td>
<td><strong>62.7</strong></td>
</tr>
</tbody>
</table>

Ground truth

16x upsampled prediction (FCN-16s)

8x upsampled prediction (FCN-8s)

2x conv7

4x conv7

2x pool4

pool3

prediction (FCN-32s)
Experiment setup

- Pre-trained networks: **AlexNet**, **VGG** and **GoogLeNet**
- Convert dense to convolutional layer
- Discard final classifier
- Add deconvolution layer for up-sampling
- Fine-tuning end-to-end on
  - PASCAL VOC
  - NYUDv2
  - SIFT
Experiment #1: PASCAL VOC 2011

- 20 classes (e.g. airplane, boat, bicycle, person, cat)
- Relative margin of 20% to previous state-of-art
- Inference time is reduced 114x and 286x respectively

<table>
<thead>
<tr>
<th></th>
<th>mean IU VOC2011 test</th>
<th>mean IU VOC2012 test</th>
<th>inference time</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-CNN [12]</td>
<td>47.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SDS [17]</td>
<td>52.6</td>
<td>51.6</td>
<td>~ 50 s</td>
</tr>
<tr>
<td>FCN-8s</td>
<td>62.7</td>
<td>62.2</td>
<td>~ 175 ms</td>
</tr>
</tbody>
</table>
Experiment #2: NYUDv2

- 40 classes from indoor scenes
- Densely labeled RGB and depth images

<table>
<thead>
<tr>
<th></th>
<th>pixel acc.</th>
<th>mean IU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gupta et al. [15]</td>
<td>60.3</td>
<td>28.6</td>
</tr>
<tr>
<td>FCN-32s RGB</td>
<td>60.0</td>
<td>29.2</td>
</tr>
<tr>
<td>FCN-32s RGBD</td>
<td>61.5</td>
<td>30.5</td>
</tr>
<tr>
<td>FCN-32s HHA</td>
<td>57.1</td>
<td>24.2</td>
</tr>
<tr>
<td>FCN-32s RGB-HHA</td>
<td>64.3</td>
<td>32.8</td>
</tr>
<tr>
<td>FCN-16s RGB-HHA</td>
<td>65.4</td>
<td>34.0</td>
</tr>
</tbody>
</table>

Experiment #3: SIFT Flow

- 33 semantic categories (e.g. bridge, mountain)
- 3 geometric categories (horizontal, vertical)
- Two-headed FCN

<table>
<thead>
<tr>
<th></th>
<th>pixel acc.</th>
<th>geom. acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liu et al. [25]</td>
<td>76.7</td>
<td>-</td>
</tr>
<tr>
<td>Tighe et al. [36]</td>
<td>-</td>
<td>90.8</td>
</tr>
<tr>
<td>Tighe et al. [37]</td>
<td>75.6</td>
<td>-</td>
</tr>
<tr>
<td>Tighe et al. [37]</td>
<td>78.6</td>
<td>-</td>
</tr>
<tr>
<td>Farabet et al. [9]</td>
<td>72.3</td>
<td>-</td>
</tr>
<tr>
<td>Farabet et al. [9]</td>
<td>78.5</td>
<td>-</td>
</tr>
<tr>
<td>Pinheiro et al. [31]</td>
<td>77.7</td>
<td>-</td>
</tr>
<tr>
<td>FCN-16s</td>
<td>85.2</td>
<td>94.3</td>
</tr>
</tbody>
</table>
Questions?
References


