A Fuller Understanding of Fully Convolutional Networks

Evan Shelhamer*  Jonathan Long*  Trevor Darrell

UC Berkeley  in CVPR'15, PAMI'16
pixels in, pixels out

semantic segmentation

monocular depth + normals Eigen & Fergus 2015

optical flow Fischer et al. 2015

boundary prediction Xie & Tu 2015

colorization
Zhang et al. 2016
convnets perform classification

< 1 millisecond

"tabby cat"

1000-dim vector

end-to-end learning
lots of pixels, little time?

~1/10 second

end-to-end learning
a classification network

convolution

fully connected

227 × 227  55 × 55  27 × 27  13 × 13

“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
end-to-end, pixels-to-pixels network

conv, pool, nonlinearity

upsampling

pixelwise output + loss
spectrum of deep features

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

image

intermediate layers

fuse features into **deep jet**

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

end-to-end, joint learning of semantics and location

interp + sum

dense output
skip layer refinement

input image  stride 32  stride 16  stride 8  ground truth
no skips 1 skip 2 skips
A multi-stream network that fuses features/predictions across layers
Relative to prior state-of-the-art SDS:

- 30% relative improvement for mean IoU
- 286× faster

*Simultaneous Detection and Segmentation Hariharan et al. ECCV14
== segmentation with Caffe ==
care and feeding of fully convolutional networks
usage

- train full image at a time without sampling

- reshape network to take input of any size

- forward time is ~100ms for 500 x 500 x 21 output (on M. Titan X)
image-to-image optimization

<table>
<thead>
<tr>
<th></th>
<th>batch size</th>
<th>mom.</th>
<th>pixel acc.</th>
<th>mean acc.</th>
<th>mean IU</th>
<th>f.w. IU</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN-accum</td>
<td>20</td>
<td>0.9</td>
<td>86.0</td>
<td>66.5</td>
<td>51.9</td>
<td>76.5</td>
</tr>
<tr>
<td>FCN-online</td>
<td>1</td>
<td>0.9</td>
<td>89.3</td>
<td>76.2</td>
<td>60.7</td>
<td>81.8</td>
</tr>
<tr>
<td>FCN-heavy</td>
<td>1</td>
<td>0.99</td>
<td>90.5</td>
<td>76.5</td>
<td>63.6</td>
<td>83.5</td>
</tr>
</tbody>
</table>
momentum and batch size

\[ p^{(1/k)} = p'(1/k') \]

\[ g_t = -\eta \sum_{i=0}^{k-1} \nabla_{\theta} \ell(x_{kt+i}; \theta_{t-1}) + pg_{t-1} \]

\[ g_t = -\eta \sum_{s=0}^{\infty} \sum_{i=0}^{k-1} p^s \nabla_{\theta} \ell(x_{k(t-s)+i}; \theta_{t-s}) \]
sampling images?

no need! no improvement from sampling across images
sampling pixels?

no need! no improvement from (partially) decorrelating pixels

uniform

poisson
context?

- do FCNs incorporate contextual cues?

- loses 3-4 % points when the background is masked

- can learn from BG/shape alone if forced to!
  - Standard  85 IU
  - BG alone  38 IU
  - Shape      29 IU
past and future history of fully convolutional networks
history

Shape Displacement Network
Matan & LeCun 1992

Convolutional Locator Network
Wolf & Platt 1994
The scale pyramid is a classic multi-resolution representation.

Fusing multi-resolution network layers is a learned, nonlinear counterpart.

Scale Pyramid, *Burt & Adelson ‘83*
jets

The local jet collects the partial derivatives at a point for a rich local description

The deep jet collects layer compositions for a rich, learned description

Jet, Koenderink & Van Doorn ‘87
extensions

- detection + instances
- structured output
- weak supervision
detection: fully conv. proposals

Fast R-CNN, *Girshick* ICCV’15

Faster R-CNN, *Ren et al.* NIPS’15

end-to-end detection by proposal FCN RoI classification
fully conv. nets + structured output

Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs.
Chen* & Papandreou* et al. ICLR 2015.
fully conv. nets + structured output

<table>
<thead>
<tr>
<th>Method</th>
<th>Without COCO</th>
<th>With COCO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plain FCN-8s</td>
<td>61.3</td>
<td>68.3</td>
</tr>
<tr>
<td>FCN-8s and CRF disconnected</td>
<td>63.7</td>
<td>69.5</td>
</tr>
<tr>
<td>End-to-end training of CRF-RNN</td>
<td>69.6</td>
<td>72.9</td>
</tr>
</tbody>
</table>

dilation for structured output

- enlarge effective receptive field for same no. params
- raise resolution
- convolutional context model: similar accuracy to CRF but non-probabilistic

Multi-Scale Context Aggregation by Dilated Convolutions. Yu & Koltun. ICLR 2016
DeepLab: Chen* & Papandreou* et al. ICLR 2015.  
CRF-RNN: Zheng* & Jayasumana* et al. ICCV 2015
fully conv. nets + weak supervision

FCNs expose a spatial loss map to guide learning: segment from tags by MIL or pixelwise constraints

fully conv. nets + weak supervision

FCNs expose a spatial loss map to guide learning:
mine boxes + feedback to refine masks

BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation. 
Dai et al. 2015.
fully conv. nets + weak supervision

FCNs can learn from sparse annotations == sampling the loss

Levels of supervision

- Full
- Image-level
- Point-level
- Objectness prior

fully convolutional networks are fast, end-to-end models for pixelwise problems

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

[DNA structure](https://www.google.com)

[github.com/BVLC/caffe](https://github.com/BVLC/caffe)

[caffe.berkeleyvision.org](https://caffe.berkeleyvision.org)

[fcn.berkeleyvision.org](https://fcn.berkeleyvision.org)

[model example](https://model.example)
[inference example](https://inference.example)
[solving example](https://solving.example)