Semantic Segmentation
and
Image Processing

with Convnets
Overview

• Methods where output is also an image
  • Fully Convolutional Nets [Long et al., CVPR 2015]
  • Depth, normals and semantic labels from a single image [Eigen ICCV 2015]

• Image processing with Convnets
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

end-to-end learning

“tabby cat”

1000-dim vector
lots of pixels, little time?
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

convolution

H × W  H/4 × W/4  H/8 × W/8  H/16 × W/16  H/32 × W/32
upsampling output
end-to-end, pixels-to-pixels network
end-to-end, pixels-to-pixels network
spectrum of deep features

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

(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
no skips

stride 16
1 skip

stride 8
2 skips

ground truth
skip FCN computation

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
SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation

Max pooling indices transferred to decoder to improve output resolution

https://arxiv.org/abs/1511.00561
How to do the Upsampling?

Also known as Deconvolution
See https://distill.pub/2016/deconv-checkerboard/

Avoid artifacts by doing bilinear interpolation
UNet: Convolutional Networks for Biomedical Image Segmentation

Segmentation of a 512x512 image takes less than a second on a recent GPU
Dilated / Atrous Convolutions

[Multi-Scale Context Aggregation by Dilated Convolutions, Yu and Koltun, 2015]

- No pooling operations
- Constant resolution feature maps
- Integrate increasing spatial context by special kind of dilated convolution

<table>
<thead>
<tr>
<th>Layer</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convolution</td>
<td>3×3</td>
<td>3×3</td>
<td>3×3</td>
<td>3×3</td>
<td>3×3</td>
<td>3×3</td>
<td>1×1</td>
<td></td>
</tr>
<tr>
<td>Dilation</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Truncation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Receptive field</td>
<td>3×3</td>
<td>5×5</td>
<td>9×9</td>
<td>17×17</td>
<td>33×33</td>
<td>65×65</td>
<td>67×67</td>
<td>67×67</td>
</tr>
<tr>
<td>Output channels</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>Large</td>
<td>2C</td>
<td>2C</td>
<td>4C</td>
<td>8C</td>
<td>16C</td>
<td>32C</td>
<td>32C</td>
<td>C</td>
</tr>
</tbody>
</table>

- Constant 64x64 spatial resolution throughout
Dilated / Atrous Convolutions

[Multi-Scale Context Aggregation by Dilated Convolutions, Yu and Koltun, 2015]
Further Resources

http://blog.qure.ai/notes/semantic-segmentation-deep-learning-review
Overview

• Methods where output is now an image
  • Fully Convolutional Nets [Long et al., CVPR 2015]
  • Depth, normals and semantic labels from a single image [Eigen ICCV 2015]

• Image processing with Convnets
  • Image colorization [Zhang et al. ECCV 2016]
Beyond Object Classification with Convolutional Networks

David Eigen (NYU -> Clarifai)
Rob Fergus (Facebook / NYU)
Motivation

- Understand input scene
  - Semantic
  - Geometric
Motivation

- Understand input scene
  - Semantic
  - Geometric

Input Image

Semantic Map

Depth
Motivation

- Understand input scene
  - Semantic
  - Geometric
• **Predict Pixel Maps from a Single Image**
Architecture

Input: 320x240

Output 1: 19x14
Architecture

Input: 320x240

Output 2: 75x55

upsample
Architecture

Input: 320x240

Output: 147x109
Architecture

Input: 320x240

conv+pool \hspace{1cm} \textbf{concat} \hspace{1cm} \textbf{convolutions}
Architecture

Input: 320x240

conv+pool  concat  convolutions
Losses

Depth:
\[ d = D - D^* \]

\[ L_{depth}(D, D^*) = \frac{1}{n} \sum_i d_i^2 - \frac{1}{2n^2} \left( \sum_i d_i \right)^2 + \frac{1}{n} \sum_i [(\nabla_x d_i)^2 + (\nabla_y d_i)^2] \]

Norm

Labels
Training

- Pre-train Alexnet/VGGnet scale 1 with Imagenet
- Scale 2 & 3 random initialization
- Joint train layers 1 & 2 for each task
  - Loss on output of layer 2
- Fix layers 1 & 2, train layer 3
- For depth & normals task, share scale 1
  - But separate scale 2 & 3’s
  - 1.6x speedup
Evaluation

• NYU Depth dataset
  – RGB, Depth and per-pixel labels
  – Indoor scenes

• Supervised training of models

• Compare to range of other methods
  – Also on SIFTFlow and PASCAL VOC’11
Depths Comparison

Eigen NIPS’14 (2 scales)  Ours  Ground Truth
Depth Comparison

- m3d = Make3D [Saxena & Ng 2006]
Surface Normals

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
<td><img src="image5" alt="Image" /></td>
</tr>
<tr>
<td></td>
<td><img src="image6" alt="Image" /></td>
<td><img src="image7" alt="Image" /></td>
<td><img src="image8" alt="Image" /></td>
<td><img src="image9" alt="Image" /></td>
<td><img src="image10" alt="Image" /></td>
</tr>
<tr>
<td></td>
<td><img src="image11" alt="Image" /></td>
<td><img src="image12" alt="Image" /></td>
<td><img src="image13" alt="Image" /></td>
<td><img src="image14" alt="Image" /></td>
<td><img src="image15" alt="Image" /></td>
</tr>
<tr>
<td></td>
<td><img src="image16" alt="Image" /></td>
<td><img src="image17" alt="Image" /></td>
<td><img src="image18" alt="Image" /></td>
<td><img src="image19" alt="Image" /></td>
<td><img src="image20" alt="Image" /></td>
</tr>
</tbody>
</table>
### Surface Normal Estimation (GT [6])

<table>
<thead>
<tr>
<th>Method</th>
<th>Angle Distance Mean</th>
<th>Angle Distance Median</th>
<th>Within 11.25° Deg.</th>
<th>Within 22.5° Deg.</th>
<th>Within 30° Deg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3DP [6]</td>
<td>34.2</td>
<td>30.0</td>
<td>18.5</td>
<td>38.6</td>
<td>50.0</td>
</tr>
<tr>
<td>Ladicky &amp; al [16]</td>
<td>32.5</td>
<td>22.3</td>
<td>27.4</td>
<td>50.2</td>
<td>60.1</td>
</tr>
<tr>
<td>Fouhey &amp; al [7]</td>
<td>35.1</td>
<td>19.2</td>
<td>37.6</td>
<td>53.3</td>
<td>58.9</td>
</tr>
<tr>
<td>Wang &amp; al [33]</td>
<td>26.6</td>
<td>15.3</td>
<td>40.1</td>
<td>61.4</td>
<td>69.0</td>
</tr>
<tr>
<td>Ours (AlexNet)</td>
<td>23.1</td>
<td>15.1</td>
<td>39.4</td>
<td>63.6</td>
<td>72.7</td>
</tr>
<tr>
<td>Ours (VGG)</td>
<td>20.5</td>
<td>13.2</td>
<td>44.0</td>
<td>68.5</td>
<td>77.2</td>
</tr>
</tbody>
</table>

### Surface Normal Estimation (GT [27])

<table>
<thead>
<tr>
<th>Method</th>
<th>Angle Distance Mean</th>
<th>Angle Distance Median</th>
<th>Within 11.25° Deg.</th>
<th>Within 22.5° Deg.</th>
<th>Within 30° Deg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3DP [6]</td>
<td>37.7</td>
<td>34.1</td>
<td>14.0</td>
<td>32.7</td>
<td>44.1</td>
</tr>
<tr>
<td>Ladicky &amp; al [16]</td>
<td>35.5</td>
<td>25.5</td>
<td>24.0</td>
<td>45.6</td>
<td>55.9</td>
</tr>
<tr>
<td>Wang &amp; al [33]</td>
<td>28.8</td>
<td>17.9</td>
<td>35.2</td>
<td>57.1</td>
<td>65.5</td>
</tr>
<tr>
<td>Ours (AlexNet)</td>
<td>25.9</td>
<td>18.2</td>
<td>33.2</td>
<td>57.5</td>
<td>67.7</td>
</tr>
<tr>
<td>Ours (VGG)</td>
<td>22.2</td>
<td>15.3</td>
<td>38.6</td>
<td>64.0</td>
<td>73.9</td>
</tr>
</tbody>
</table>
Results: Normals

Angle from Ground Truth

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours+VGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wang &amp; al</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fouhey &amp; al '14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ladicky &amp; al</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3DP</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Output from each scale

input

depth

coarse  fine

normals
Semantic Labels: NYUD
Results: NYUD 40 Classes

- Use RGB + ground truth depth & normals as inputs

**Per-Pixel Acc.**

- Ours (VGG)
- Long & al
- Ours
- Gupta & al '14
- Gupta & al '13

**Per-Class Acc.**

- 20. 27. 34. 41. 48.
Results: NYUD Labels

- Use RGB + ground truth depth & normals as inputs

4 Classes: Pixel Acc.

- Ours
- Gupta & al
- Mueller & al
- Stuckler & al
- Khan & al
- Couprie & al

13 Classes: Pixel Acc.

- Ours
- Khan
- Hermans
- Wang
- Couprie
### Pascal VOC Semantic Segmentation

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Long et al. [19]</td>
<td>90.3</td>
<td>75.9</td>
<td>83.2</td>
<td>62.7</td>
</tr>
<tr>
<td>Ours (VGG)</td>
<td>90.3</td>
<td>72.4</td>
<td>82.9</td>
<td>62.2</td>
</tr>
</tbody>
</table>

Semantic Labels: Pascal VOC’11
Contribution from different scales

- On NYU Depth

<table>
<thead>
<tr>
<th>Contributions of Scales</th>
<th>Depth</th>
<th>Normals</th>
<th>4-Class</th>
<th>13-Class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>RGB+D+N</td>
<td>RGB</td>
</tr>
<tr>
<td></td>
<td>Pixelwise Error</td>
<td>Pixelwise Accuracy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scale 1 only</td>
<td>0.218</td>
<td>29.7</td>
<td>71.5</td>
<td>58.1</td>
</tr>
<tr>
<td>Scale 2 only</td>
<td>0.290</td>
<td>31.8</td>
<td>77.4</td>
<td>65.1</td>
</tr>
<tr>
<td>Scales 1 + 2</td>
<td>0.216</td>
<td>26.1</td>
<td>80.1</td>
<td>69.8</td>
</tr>
<tr>
<td>Scales 1 + 2 + 3</td>
<td>0.198</td>
<td>25.9</td>
<td>80.6</td>
<td>70.5</td>
</tr>
</tbody>
</table>

- Depth & normals: scale 1 most important
- Semantic labels: scale 2 most important
  (if D & N are available)
Using Predicted Depths

- Use *predicted* depth/normals as input?

![Graph showing Per-Pixel and Per-Class Accuracies for different scales and input types]

- NYU Depth 13-class

  - RGB only
  - RGB + Pred D&N
  - RGB + GT D&N
Panoptic Segmentation: Unifying Semantic and Instance Segmentation

Why solve it?

Semantic Segmentation
- per-pixel annotation
- simple accuracy measure
- instances indistinguishable

Panoptic Segmentation

Object Detection/Seg
- each object detected and segmented separately
- “stuff” is not segmented

FCN 8s, Dilation8, DeepLab, PSPNet, RefineNet, U-Net, etc.

Fast/er R-CNN, DeepMask, SharpMask, Mask R-CNN, FCIS, YOLO, RetinaNet, FPN, etc.
Panoptic Segmentation: Unifying Semantic and Instance Segmentation

Why solve it?

Mask R-CNN

instances

PSPNet

semantic scores

panoptic prediction

[https://arxiv.org/abs/1801.00868]
Panoptic Segmentation: Unifying Semantic and Instance Segmentation  
[https://arxiv.org/abs/1801.00868]

Panoptic COCO  
Panoptic CityScapes
Overview

• Methods where output is also an image
  – Fully Convolutional Nets [Long et al., CVPR 2015]
  – Depth, normals and semantic labels from a single image [Eigen ICCV 2015]

• Image processing with Convnets
Denoising with ConvNets

- Burger et al. “Can plain NNs compete with BM3D?” CVPR 2012
Deblurring with Convnets

- Blind deconvolution
  - Learning to Deblur, Schuler et al., arXiv 1406.7444, 2014
Inpainting with Convnets

- Image Denoising and Inpainting with Deep Neural Networks, Xie et al. NIPS 2012.
- Mask-specific inpainting with deep neural networks, Köhler et al., Pattern Recognition 2014
Removing Local Corruption

- Restoring An Image Taken Through a Window Covered with Dirt or Rain, Eigen et al., ICCV 2013.
Removing Local Corruption

Restoring An Image Taken Through a Window Covered with Dirt or Rain

Rain Sequence
Each frame processed independently

David Eigen, Dilip Krishnan and Rob Fergus
ICCV 2013
Learning to See in the Dark

[Chen et al., arXiv 1805.01934]
Learning to See in the Dark
[Chen et al., arXiv 1805.01934]
Class Project Admin

- Showcase
- Report
- Deadline is Friday Dec 21\textsuperscript{st} midnight
  - Feel free to turn in earlier
  - Will try to grade them and compute final grades by Christmas
- Will post all of this to Piazza
Poster Presentation session

• Thursday, December 20th at 7:00-9:00 pm (405 Silver).

• Poster presentation of your project
  – Will be part of grading

• Pizza & drinks will be served!
Project Expectations

• Grading (45% of total grade for class)
  • Novelty / Technical difficulty of problem [10%]
  • Quality of Results [10%]
  • Quality of implementation [10%]
  • Quality of writeup [10%]
  • Presentation at CS Showcase [5%]
  • How many people in your group
Project Expectations

• Report
  • 4-8 page conference paper style report on your project
  • Intro (with refs to related work)
  • Method (be sure to cite any code/pre-trained models)
  • Experiments (must have plots/results figures; also should have baselines; ideally some kind of ablation experiments too)
  • Discuss (brief)
  • See examples: http://openaccess.thecvf.com/CVPR2018.py

• Zip of source code or link to Github (please ensure you give access to robfergus)

• For CS Showcase poster:
  • no need to print big poster
  • just make PPT slides and print out (8-12 is enough)
Project Expectations

• Generalities
  • Please make sure you have *something* working, even if you don’t achieve overall goal
  • Even a small part of an ambitious project can be OK
  • So please have a safe plan B option in mind
  • Expect all projects to train something, i.e. must use b-prop at some point
  • Just evaluating existing models is NOT OK.
  • Cluster gets busy -- please don’t leave it all to last moment.