Announcement

This week’s office hour will be:

Tuesday: 4-6pm by Ziyan
Thursday: 12:30-2:30 by Prof. Vicente
Friday: 2-4pm by Ziyan
What is image segmentation?
Previous Image Segmentation Methods

- K-means clustering
- Graph-based methods (graph-cut…)
- Histogram-based image segmentation
- Thresholding
- etc.

Deep Learning!
Data

• PASCAL VOC 2007 & 2012
  - trainval: 5011/11540
  - test: 4952/?
• MS COCO
  - trainval: 123,287 images
Evaluation

- Pixel accuracy
- Intersection over union
- Precision
Pixel accuracy

pixel accuracy = 95%
Intersection over union

If two classes

\[ IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}} \]

= 50%
Precision

- A **true positive (TP)** is observed when a prediction-target mask pair has an IoU score which exceeds some predefined threshold.

- A **false positive (FP)** indicates a predicted object mask had no associated ground truth object mask.

**Precision = TP/(TP+FP)**

**Example**
Threshold: 0.5

- True positive
  - ground truth
  - prediction
  - IoU = 0.8

- False negative
  - ground truth

- False positive
  - IoU = 0.1
Deep learning models

- Fully convolutional network (FCN)
- U-Net
- Dilated convolution
- Mask R-CNN
Fully Convolutional Network (FCN)

Original paper: https://people.eecs.berkeley.edu/~jonlong/long_shelhamer_fcn.pdf
From Image Classification to Semantic Segmentation

If we change fully connected layers to 1x1 convolutional layer
Upsampling

\[ \text{upsampling} = \text{deconvolution} = \text{transposed convolution} \]

In a sense, upsampling with factor \( f \) is convolution with a fractional input stride of \( 1/f \)
Upsampling

Convolution is a process getting the output size smaller. Thus, the name, deconvolution, is coming from when we want to have upsampling to get the output size larger.

Blue: Input, Green: Output
Fuse the outputs
FCN - results
FCN - results

PASCAL VOC 2012 result: 62.2% mIoU score
U-Net

Original paper: https://arxiv.org/abs/1505.04597

Based on fully convolutional networks
U-Net structure

Fig. 1. U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.
U-Net architecture

Two consecutive convolution layers

For Up-sampling, Transposed Convolutional layers are used.

The parameters for each Transpose Convolution are such that the height and width of the image are doubled while the depth (no. of channels) is halved.
Dilated convolution

In FCN, pooling layers are used to increase the reception field. However…it will also decrease the size of feature map. This is why we have deconvolution layer to upsample the feature.

Can we have a new structure to combine these two steps?

Yes!
Dilated Convolution: add holes to convolution map to increase reception field
Comparison

Deconvolution

Dilated convolution
More details

Original paper: https://arxiv.org/abs/1511.07122
Results

PASCAL VOC 2012 result

<table>
<thead>
<tr>
<th></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>mibike</th>
<th>person</th>
<th>plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>lv</th>
<th>mean IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN-8s</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>DeepLab</td>
<td>72</td>
<td>31</td>
<td>71.2</td>
<td>53.7</td>
<td>60.5</td>
<td>77</td>
<td>71.9</td>
<td>73.1</td>
<td>25.2</td>
<td>62.6</td>
<td>49.1</td>
<td>68.7</td>
<td>63.3</td>
<td>73.9</td>
<td>73.6</td>
<td>50.8</td>
<td>72.3</td>
<td>42.1</td>
<td>67.9</td>
<td>52.6</td>
<td>62.1</td>
</tr>
<tr>
<td>DeepLab-Msc</td>
<td>74.9</td>
<td>34.1</td>
<td>72.6</td>
<td>52.9</td>
<td>61.0</td>
<td>77.9</td>
<td>73.0</td>
<td>73.7</td>
<td>26.4</td>
<td>62.2</td>
<td>49.3</td>
<td>68.4</td>
<td>64.1</td>
<td>74.0</td>
<td>75.0</td>
<td>51.7</td>
<td>72.7</td>
<td>42.5</td>
<td>67.2</td>
<td>55.7</td>
<td>62.9</td>
</tr>
<tr>
<td>Our front end</td>
<td>82.2</td>
<td>37.4</td>
<td>72.7</td>
<td>57.1</td>
<td>62.7</td>
<td>82.8</td>
<td>77.8</td>
<td>78.9</td>
<td>28</td>
<td>70</td>
<td>51.6</td>
<td>73.1</td>
<td>72.8</td>
<td>81.5</td>
<td>79.1</td>
<td>56.6</td>
<td>77.1</td>
<td>49.9</td>
<td>75.3</td>
<td>60.9</td>
<td>67.6</td>
</tr>
</tbody>
</table>
Mask R-CNN


Do you remember Faster R-CNN?

It is trained end-to-end, region proposal method is actually an internal deep network and the ROIs are derived from the feature maps.

Mask RCNN = Faster RCNN + FCN
Mask R-CNN structure
More details

Use interpolation

<table>
<thead>
<tr>
<th></th>
<th>AP</th>
<th>AP_{50}</th>
<th>AP_{75}</th>
<th>AP_{bb}</th>
<th>AP_{bb_50}</th>
<th>AP_{bb_75}</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoIPool</td>
<td>23.6</td>
<td>46.5</td>
<td>21.6</td>
<td>28.2</td>
<td>52.7</td>
<td>26.9</td>
</tr>
<tr>
<td>RoIAlign</td>
<td>30.9</td>
<td>51.8</td>
<td>32.1</td>
<td>34.0</td>
<td>55.3</td>
<td>36.4</td>
</tr>
<tr>
<td></td>
<td>+7.3</td>
<td>+5.3</td>
<td>+10.5</td>
<td>+5.8</td>
<td>+2.6</td>
<td>+9.5</td>
</tr>
</tbody>
</table>
Results

Samples from COCO dataset

It can get mAP ~75.9% on PASCAL VOC 2012 Test set
Thank you!


Review

For RPN. Each bounding box has 4 coordinates and has/does not has object tags
ROI pooling