DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution and Fully Connected CRFs

Zhipeng Yan, Moyuan Huang, Hao Jiang

5/1/2017

Fig. 6: PASCAL VOC 2012 val results. Input image and our DeepLab results before/after CRF.
Outline

Background

- semantic segmentation
- Objective, Dataset, Basic Idea
- Predecessor: Fully Convolutional network
  - Fully Convolution, Deconvolution, Skip path

DeepLab

- Atrous Convolution
- ASPP

Fully Connected CRFs

Background: semantic segmentation
Semantic Segmentation

One of the most popular topics in Computer Vision

Like image classification, object detection...

The objective

Partition the image into segments according to different semantic groups

Semantic Segmentation

Example

Split the image according to object boundaries

Colors each object according to its semantic meanings
Semantic Segmentation - Dataset

- PASCAL VOC
- Cityscapes
- Microsoft COCO

Semantic Segmentation - Idea

- Convolutional Neural Network
- Feature extractor
- Post Processing
- Feature refinements
- Upsampling
- edge detection, smoothing
- CRF, local classification
- Multi-scale processing
- shift-interlace
Semantic Segmentation - Idea

Problems:

- Not end-to-end
- Fixed size input & output
- Small/restricted Field of View
- Low performance
- Time consuming

Solved at once by Fully Convolutional Network (J. Long, E. Shelhamer, T. Darrell)

Fully Convolutional Networks

Jonathan Long et al.
CVPR 2015 & PAMI 2016
Previous CNN

Previous CNN downsized the output layer-by-layer in order to reduce parameter.

Very harmful for dense classification tasks like image segmentation: Lot of information

Can design a network specifically for the Segmentation task?

Fully Convolutional Networks

Input size can be arbitrary.
   Remove fully connected layer
   Practical use

30% relative improvement on previous state-of-the-art
   Avoid fully connected layers in traditional CNNs which cause the loss of spatial information

5x efficiency
   Reduced parameters
   End-to-end training
Fully Convolutional Networks

3 Major innovations on network architecture

Removal of fully connected layers

Deconvolution

Skip path

Fully Convolutional Networks

Removal of fully connected layers

Dense output with relative size to the input

Replace with 1 x 1 convolutions to transform feature maps to class-wise predictions

Response map with “tabby cat” kernel
Fully Convolutional Networks

Deconvolution

Used for retrieving information, can be regarded as a reverse of convolution

Stride size so as to avoid overlap: though NNs can adjust the corresponding weight to avoid them, it’s really a struggle to avoid it completely.

Fully Convolutional Networks

Skip path

Concatenate low level features with high level features to handle multiscale objects

Provide options for different output sizes
Skip layer visualization
different layer provide us with different levels of information.

increasing locality of perceptual field

Different level of generality


---

Skip Layer

The skip layer technique is widely used in many popular deep networks such as ResNet, Inside-Outside Net, HED. The advantage as well as motivation is that it allows more lower level information to reach top level.

With skip layer, we can get a rather finer pixel output, instead of a small coarse one.

Upsampling is used to resolve the size incompatible problem between different layer. Combining is done by simple sum operation.
Skip Layer Model In detail
Skip layer on different scales

Segmentation is gradually changed to a finer level

Some Visualization

- [FCN architecture visualization with Caffe](#)
- [A distill dynamic paper](#)
Part II:  
DeepLab - Semantic Image Segmentation with DCNN, ASPP and Fully Connected FRFs

DeepLab

Unresolved challenges:

- Reduced feature resolution
  
  8 - 32 times downsample

- Poor prediction on multi-scale objects

- High uncertainty near object boundaries
  - Trucks sometimes can have size of half the image
  - Intrinsic problem of CNN
DeepLab: resolution

Why feature resolution reduced?

Do not downsample
Convolution on large images ⇒ Small FOV

Enlarge kernel size
Use **Atrous Convolution.**
O(n^2) more parameters ⇒ getting close to fully connected layer, slow training, overfitting...

- a.k.a Dilated convolution
- Large FOV with little parameters ⇒ Kill two birds with one stone!

DeepLab: Challenge 1

How to solve reduced resolution?

- Do not downsample
- Convolution on large images ⇒ Small FOV
- Enlarge kernel size
  - Use **Atrous Convolution.**
  - O(n^2) more parameters ⇒ getting close to fully connected layer, slow training, overfitting...
    - a.k.a Dilated convolution
    - Large FOV with little parameters ⇒ Kill two birds with one stone!
DeepLab: Challenge 1

Atrous Convolution

“Insert hole” into convolution kernel

Large receptive field with ‘sparse’ parameters
DeepLab: Challenge 2

Multi-scale object problem?

Objects of the same type can be hugely different in size

Previous CNN (AlexNet, VGGnet, FCN) models yield bad result

How to get features with different scale?

Multi scale training, or

Spatial Pyramid Pooling

Combined with Atrous Convolution, we have **Atrous Spatial Pyramid Pooling**

DeepLab: Challenge 2

**Atrous Spatial Pyramid Pooling**

Motivated by the spatial pyramid pooling

Filter maps with multiple scales controlled by **dilation rate**
DeepLab: Challenge 3

Poor performance near object boundaries

- Intrinsic CNN problem
- Convolution gives smooth outputs in small neighborhoods

How to improve test images near boundaries?

- Conditional Random Fields

DeepLab: Challenge 3

Conditional Random Field (CRF) is a graphical model where nodes are locally connected. It calculates the output probabilities at each node given neighborhood information w.r.t. some predefined energy function

\[ p(Y_u | X, Y_v, w \neq u) = p(Y_u | X, Y_v, w \sim u) \]

\( \sim \) means that \( u \) and \( v \) are connected in \( G \).

Local approximation - only information at adjacent nodes is taken into consideration

DCNN’s outputs probability vectors at each pixel, then CRF refines the output
DeepLab: Challenge 3

Training of CRF:

Energy definition (Identical to [Krahenbuhl et al])

$$\theta_{ij}(x_i, x_j) = \mu(x_i, x_j) \left[ w_1 \exp \left( -\frac{||p_i - p_j||^2}{2\sigma_1^2} - \frac{||I_i - I_j||^2}{2\sigma_2^2} \right) 
+ w_2 \exp \left( -\frac{||p_i - p_j||^2}{2\sigma_3^2} \right) \right]$$ (3)

Minimization through iterative compatibility transform

DeepLab: Challenge 3

Able to capture edge details and iteratively refine the prediction

Efficient CRF [Krahenbuhl et al] achieves 0.5 sec/image on PASCAL VOC.
DeepLab: Effectiveness of ASPP and CRF

Experiments with
- Different kernel sizes
- Before / After CRF

Results
- Larger dilation ⇒ higher performance, less parameters

<table>
<thead>
<tr>
<th>Kernel</th>
<th>Rate</th>
<th>FOV</th>
<th>Params</th>
<th>Speed</th>
<th>bef/aft CRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>7×7</td>
<td>4</td>
<td>224</td>
<td>134.3M</td>
<td>1.44</td>
<td>64.38 / 67.64</td>
</tr>
<tr>
<td>4×4</td>
<td>4</td>
<td>128</td>
<td>65.1M</td>
<td>2.90</td>
<td>59.80 / 63.74</td>
</tr>
<tr>
<td>4×4</td>
<td>8</td>
<td>224</td>
<td>65.1M</td>
<td>2.90</td>
<td>63.41 / 67.14</td>
</tr>
<tr>
<td>3×3</td>
<td>12</td>
<td>224</td>
<td>20.5M</td>
<td>4.84</td>
<td>62.25 / 67.64</td>
</tr>
</tbody>
</table>

DeepLab Whole Model

Fig. 1: Model Illustration. A Deep Convolutional Neural Network such as VGG-16 or ResNet-101 is employed in a fully convolutional fashion, using atrous convolution to reduce the degree of signal down-sampling (from 32x down 8x). A bilinear interpolation stage enlarges the feature maps to the original image resolution. A fully connected CRF is then applied to refine the segmentation result and better capture the object boundaries.
DeepLab Experiment Detail 1

“poly” learning rate policy

More effective way to change the learning rate

Yields better result

DeepLab Experiment Detail 2

Various training tricks
DeepLab: Results on PASCAL VOC 2012

Atrous Spatial Pyramid Pooling

<table>
<thead>
<tr>
<th>Method</th>
<th>before CRF</th>
<th>after CRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>LargeFOV</td>
<td>65.76</td>
<td>69.84</td>
</tr>
<tr>
<td>ASPP-S</td>
<td>66.98</td>
<td>69.73</td>
</tr>
<tr>
<td>ASPP-L</td>
<td>68.96</td>
<td>71.57</td>
</tr>
</tbody>
</table>

TABLE 3: Effect of ASPP on PASCAL VOC 2012 val set performance (mean IOU) for VGG-16 based DeepLab model. LargeFOV: single branch, $r = 12$. ASPP-S: four branches, $r = \{2, 4, 8, 12\}$. ASPP-L: four branches, $r = \{6, 12, 18, 24\}$.

Fig. 8: Qualitative segmentation results with ASPP compared to the baseline LargeFOV model. The ASPP-L model, employing multiple large FOVs can successfully capture objects as well as image context at multiple scales.
Performance on PASCAL VOC 2012 - Qualitative

Fig. 6: PASCAL VOC 2012 val results. Input image and our DeepLab results before/after CRF.

Performance on PASCAL VOC 2012 - Quantitative

<table>
<thead>
<tr>
<th>Method</th>
<th>mIOU</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepLab-CRF-LargeFOV-COCO [58]</td>
<td>72.7</td>
</tr>
<tr>
<td>MRRU_DEEP_GCRF [89]</td>
<td>73.2</td>
</tr>
<tr>
<td>CRF-RNN [59]</td>
<td>74.7</td>
</tr>
<tr>
<td>POSTECH_DeconvNet_CRF_VOC [61]</td>
<td>74.8</td>
</tr>
<tr>
<td>BoxSep [60]</td>
<td>75.2</td>
</tr>
<tr>
<td>Context + CRF-RNN [76]</td>
<td>75.3</td>
</tr>
<tr>
<td>QoI [66]</td>
<td>75.5</td>
</tr>
<tr>
<td>DeepLab-CRF-Attention [17]</td>
<td>75.7</td>
</tr>
<tr>
<td>CentralesSuperBoundaries++ [18]</td>
<td>76.0</td>
</tr>
<tr>
<td>DeepLab-CRF-Attention-DT [63]</td>
<td>76.3</td>
</tr>
<tr>
<td>H-ReNet + DenseCRF [90]</td>
<td>76.8</td>
</tr>
<tr>
<td>LRR_4s COCO [91]</td>
<td>76.8</td>
</tr>
<tr>
<td>DPN [62]</td>
<td>77.5</td>
</tr>
<tr>
<td>Adelaide_Context [40]</td>
<td>77.8</td>
</tr>
<tr>
<td>Oxford_TVC_HO_CRF [88]</td>
<td>77.9</td>
</tr>
<tr>
<td>Context CRF + Guidance CRF [92]</td>
<td>78.1</td>
</tr>
<tr>
<td>Adelaide_VeryDeep_FC2_VOC [93]</td>
<td>79.1</td>
</tr>
<tr>
<td>DeepLab-CRF (ResNet-101)</td>
<td>79.7</td>
</tr>
</tbody>
</table>

TABLE 5: Performance on PASCAL VOC 2012 test set. We have added some results from recent arXiv papers on top of the official leaderboard results.
Backbone network: VGG-16 vs. ResNet-101

DeepLab based on ResNet-101 delivers better segmentation results

Fig. 9: DeepLab results based on VGG-16 net or ResNet-101 before and after CRF. The CRF is critical for accurate prediction along object boundaries with VGG-16, whereas ResNet-101 has acceptable performance even before CRF.

DeepLab Experiment Results
Dataset: PASCAL-Context

59 classes, approx. 5000 images

Provides detailed semantic labels for the whole scene

Including object and background

Ex. ceiling

PASCAL-Context: Qualitative results

Fig. 11: PASCAL-Context results. Input image, ground-truth, and our DeepLab results before/after CRF.
PASCAL-Context: Quantitative results

Evaluation Results:
- ResNet-101 is better
- ASPP is more efficient than large FOV
- Using CRF improves the score

<table>
<thead>
<tr>
<th>Method</th>
<th>MSC</th>
<th>COCO Aug</th>
<th>LargeFOV</th>
<th>ASPP</th>
<th>CRF</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG-16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>37.6</td>
</tr>
<tr>
<td>DeepLab [38]</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>39.6</td>
</tr>
<tr>
<td>DeepLab [38]</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td>39.6</td>
</tr>
<tr>
<td>ResNet-101</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>39.6</td>
</tr>
<tr>
<td>DeepLab</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>41.4</td>
</tr>
<tr>
<td>DeepLab</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>42.9</td>
</tr>
<tr>
<td>DeepLab</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>43.5</td>
</tr>
<tr>
<td>DeepLab</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>44.7</td>
</tr>
<tr>
<td>DeepLab</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>45.7</td>
</tr>
<tr>
<td>O2P [45]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>18.1</td>
</tr>
<tr>
<td>CFM [51]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>34.4</td>
</tr>
<tr>
<td>FCN-4s [14]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>37.8</td>
</tr>
<tr>
<td>CRF-RNN [59]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>39.3</td>
</tr>
<tr>
<td>ParseNet [86]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>40.4</td>
</tr>
<tr>
<td>BoSUp [60]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>40.5</td>
</tr>
<tr>
<td>HO-CRF [88]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>41.3</td>
</tr>
<tr>
<td>Context [40]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>43.3</td>
</tr>
<tr>
<td>VeryDeep [93]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>44.5</td>
</tr>
</tbody>
</table>

TABLE 6: Comparison with other state-of-art methods on PASCAL-Context dataset.

Dataset: Cityscapes

- Street views of 50 different cities
- Large-scale dataset
- Image size 1024 x 2048
- Approx. 3000 training images and hundreds of validation images
- 20,000 augmented images with coarse label
Cityscapes qualitative result:

Cityscapes Quantitative results

3rd place on the leaderboard

63.1% and 64.8% on pre-release and official versions resp.
Cityscapes val set result:

![Table Image]

**Summary**

DeepLab’s model significantly advances the state-of-art in several challenging datasets

Successful combine the ideas from the deep convolutional neural networks and fully-connected conditional random field to produce accurate predictions and detailed segmentation maps
Reference:


Dataset: PASCAL-Person-Part

Contains detailed part annotations for each person, including eyes, nose.

Merge the annotations to be head, torso, upper/lower arms and upper. Lower legs

Resulting in 6 person part class and one background
Dataset: PASCAL-Context

Dataset 1: PASCAL VOC 2012

- 20 object classes, one background class
- Thousands of images
- Augmented by extra annotations

Fig. 12: PASCAL-Person-Part results. Input image, ground-truth, and our DeepLab results before/after CRF.