Semantic Segmentation

CS 6384 Computer Vision
Professor Yu Xiang
The University of Texas at Dallas
Semantic Understanding

Object Detection  Semantic Segmentation  Instance Segmentation

Conditional Random Fields Meet Deep Neural Networks for Semantic Segmentation. Arnab et al., IEEE SIGNAL PROCESSING MAGAZINE, 2018
Semantic Segmentation

• Label pixels into semantic classes

• Naïve method
  • Classify each pixel independently

• Better idea
  • Using context of pixels
Conditional Random Fields (CRFs)

- Pixel labeling problem

\[ \mathcal{G} = (\mathcal{V}, \mathcal{E}) \]

2D grid for images

where \( X_1 \in \{ \text{bg, cat, dog, person} \} \)
Conditional Random Fields (CRFs)

• Model the conditional probability distribution

\[ P(X|I) = \frac{1}{Z(I)} \exp(- \sum_{c \in C_G} \phi_c(X_c|I)) \]

- Partition function (normalization factor)
- Clique
- Potential function

\[ G = (V, E) \]

2D grid for images

\[ X_i \in \{bg, cat, dog, person\} \]
Conditional Random Fields (CRFs)

\[ P(X|I) = \frac{1}{Z(I)} \exp\left( - \sum_{c \in \mathcal{C}_g} \phi_c(X_c|I) \right) \]

- Energy function
\[ E(x|I) = \sum_{c \in \mathcal{C}_g} \phi_c(x_c|I) \quad x \in \mathcal{L}^N \]

\[ P(x|I) = \frac{1}{Z(I)} \exp(-E(x|I)) \quad Z(I) = \sum_x \exp(-E(x|I)) \]

- Maximum a posteriori (MAP) labeling
\[ x^* = \arg \max_{x \in \mathcal{L}^N} P(x|I) \]
Conditional Random Fields (CRFs)

• Unary potential and pairwise potential

\[ E(x, I) := \sum_{u \in V} \psi_u(X_u = x_u | I) + \sum_{\{u, v\} \in E} \psi_{u,v}(X_u = x_u, X_v = x_v | I) \]

E.g., classifier output  
E.g., smoothing pairwise potential \[ [x_u \neq x_v] \]

• Energy minimization problem
  • NP-hard
  • Exact and approximate algorithms exist to obtain acceptable solutions

Conditional Random Fields (CRFs)

Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. Krähenbühl & Koltun, NeurIPS, 2011

Conditional Random Fields Meet Deep Neural Networks for Semantic Segmentation. Arnab et al., IEEE SIGNAL PROCESSING MAGAZINE, 2018

\[ E(x) = \sum_i \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j) \]
Combining Neural Networks with CRFs

• Utilize neural networks to compute unary potentials

Better classifier

Semantic image segmentation with deep convolutional nets and fully connected CRFs. Chen et al., ICLR, 2015.

Conditional Random Fields Meet Deep Neural Networks for Semantic Segmentation. Arnab et al., IEEE SIGNAL PROCESSING MAGAZINE, 2018
DeepLab

Atrous convolution

(a) Sparse feature extraction

(b) Dense feature extraction

\[ y[i] = \sum_{k=1}^{K} x[i + r \cdot k]w[k] \]

Fully Convolutional Networks

• Adapt classification networks for dense prediction

Treat FC layers as convolutions with kernels that cover the entire input regions

Fully Convolutional Networks for Semantic Segmentation. Long et al., CVPR, 2015
Fully Convolutional Networks

• Convert AlexNet

[224x224x3] INPUT
[55x55x96] **CONV1**: 96 11x11 filters at stride 4, pad 0
[27x27x96] **MAX POOL1**: 3x3 filters at stride 2
[27x27x96] **NORM1**: Normalization layer
[27x27x256] **CONV2**: 256 5x5 filters at stride 1, pad 2
[13x13x256] **MAX POOL2**: 3x3 filters at stride 2
[13x13x256] **NORM2**: Normalization layer
[13x13x384] **CONV3**: 384 3x3 filters at stride 1, pad 1
[13x13x384] **CONV4**: 384 3x3 filters at stride 1, pad 1
[13x13x256] **CONV5**: 256 3x3 filters at stride 1, pad 1
[6x6x256] **MAX POOL3**: 3x3 filters at stride 2
[4096] **FC6**: 4096 neurons
[4096] **FC7**: 4096 neurons
[1000] **FC8**: 1000 neurons (class scores)

Fully Convolutional Networks for Semantic Segmentation. Long et al., CVPR, 2015
Fully Convolutional Networks

• Deconvolution for up-sampling

Input: 2 x 2

Output: 4 x 4

3 x 3 “deconvolution”, stride 2, pad 1

Input gives weight for filter

Sum where output overlaps

---

Fully Convolutional Networks for Semantic Segmentation. Long et al., CVPR, 2015
Fully Convolutional Networks

• Combine predictions with different resolutions

<table>
<thead>
<tr>
<th>Fully Convolutional Networks for Semantic Segmentation. Long et al., CVPR, 2015</th>
<th>pixel acc.</th>
<th>mean acc.</th>
<th>mean IU</th>
<th>f.w. IU</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN-32s-fix</td>
<td>83.0</td>
<td>59.7</td>
<td>45.4</td>
<td>72.0</td>
</tr>
<tr>
<td>FCN-32s</td>
<td>89.1</td>
<td>73.3</td>
<td>59.4</td>
<td>81.4</td>
</tr>
<tr>
<td>FCN-16s</td>
<td>90.0</td>
<td>75.7</td>
<td>62.4</td>
<td>83.0</td>
</tr>
<tr>
<td>FCN-8s</td>
<td><strong>90.3</strong></td>
<td><strong>75.9</strong></td>
<td><strong>62.7</strong></td>
<td><strong>83.2</strong></td>
</tr>
</tbody>
</table>
U-Net

Convolution Deconvolution

U-Net: Convolutional Networks for Biomedical Image Segmentation, Ronneberger et al., MICCAI 2015
Instance Segmentation

• Separate object instances in the same class
• Detection + segmentation

https://ai-pool.com/d/could-you-explain-me-how-instance-segmentation-works
Mask R-CNN

Mask R-CNN. He et al., ICCV, 2017
RoI Pooling vs. RoI Align

RoI

$(x, y, h, w)$

RoI mapping to feature map

$s \times (x, y, h, w)$

$s = \frac{1}{16}$

Rol Pooling

Rol Align
## Mask R-CNN

<table>
<thead>
<tr>
<th></th>
<th>align?</th>
<th>bilinear?</th>
<th>aggr.</th>
<th>AP</th>
<th>AP(_{50})</th>
<th>AP(_{75})</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>RoIPool</em> [12]</td>
<td></td>
<td></td>
<td>max</td>
<td>26.9</td>
<td>48.8</td>
<td>26.4</td>
</tr>
<tr>
<td><em>RoIWarps</em> [10]</td>
<td>✓</td>
<td></td>
<td>max</td>
<td>27.2</td>
<td>49.2</td>
<td>27.1</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td></td>
<td>ave</td>
<td>27.1</td>
<td>48.9</td>
<td>27.1</td>
</tr>
<tr>
<td><em>RoIAlign</em></td>
<td>✓</td>
<td>✓</td>
<td>max</td>
<td>30.2</td>
<td>51.0</td>
<td>31.8</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>ave</td>
<td>30.3</td>
<td>51.2</td>
<td>31.5</td>
</tr>
</tbody>
</table>

Mask R-CNN. He et al., ICCV, 2017
Unseen Object Instance Segmentation

• Can we train a model to segment objects that are in the training set?
Unseen Clustering Network

Learning RGB-D Feature Embeddings for Unseen Object Instance Segmentation. Xiang et al., CoRL, 2020
Unseen Clustering Network

- Intra-cluster loss function

\[
\mu^k = \frac{\sum_{i=1}^{N} x_i^k}{\| \sum_{i=1}^{N} x_i^k \|}
\]

Spherical mean

\[
d(\mu^k, x_i^k) = \frac{1}{2} (1 - \mu^k \cdot x_i^k)
\]

Cosine distance

\[
\ell_{\text{intra}} = \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{N} \frac{1 \left\{ d(\mu^k, x_i^k) - \alpha \geq 0 \right\} d^2(\mu^k, x_i^k)}{\sum_{i=1}^{N} 1 \left\{ d(\mu^k, x_i^k) - \alpha \geq 0 \right\}}
\]

- Inter-cluster loss function

\[
\ell_{\text{inter}} = \frac{2}{K(K-1)} \sum_{k<k'} \left[ \delta - d(\mu^k, \mu^{k'}) \right]^2_{+}
\]

Learning RGB-D Feature Embeddings for Unseen Object Instance Segmentation. Xiang et al., CoRL, 2020
Unseen Clustering Network

Learning RGB-D Feature Embeddings for Unseen Object Instance Segmentation. Xiang et al., CoRL, 2020
Summary

• Semantic segmentation
  • Label pixels into object classes
  • Traditional methods: conditional random fields
  • Deep learning methods: deconvolution, atrous convolution

• Instance segmentation
  • Separate object instances in the same class
  • Detection + segmentation inside each box

• Unseen object instance segmentation
  • Clustering-based methods to group pixels into objects
Further Reading

• Fully-connect CRFs, 2011 https://arxiv.org/abs/1210.5644
• FCN, 2015 https://arxiv.org/abs/1411.4038
• Unet, 2015 https://arxiv.org/abs/1505.04597