Encoder-Decoder Networks for Semantic Segmentation

Sachin Mehta
Outline

- Overview of Semantic Segmentation
- Encoder-Decoder Networks
- Results
What is Semantic Segmentation?

Input: RGB Image

Output: A segmentation Mask
Encoder-Decoder Networks

Encoder
- Takes an input image and generates a high-dimensional feature vector
- Aggregate features at multiple levels

Decoder
- Takes a high-dimensional feature vector and generates a semantic segmentation mask
- Decode features aggregated by encoder at multiple levels
Building Blocks of CNNs

- Convolution
- Down-Sampling
- Up-Sampling
Convolution

Filter weights are learned from data
Down-Sampling

- Max-pooling
- Average Pooling
- Strided Convolution
Up-Sampling

> Un-pooling
> Deconvolution

- **Un-Pooling**

<table>
<thead>
<tr>
<th>a0</th>
<th>a1</th>
<th>b0</th>
<th>b1</th>
</tr>
</thead>
<tbody>
<tr>
<td>a2</td>
<td>a3</td>
<td>b2</td>
<td>b3</td>
</tr>
<tr>
<td>c0</td>
<td>c1</td>
<td>d0</td>
<td>d1</td>
</tr>
<tr>
<td>c2</td>
<td>c3</td>
<td>d2</td>
<td>d3</td>
</tr>
</tbody>
</table>

- **Max-pooling indices**

<table>
<thead>
<tr>
<th>a0</th>
<th>b1</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1</td>
<td>d3</td>
</tr>
</tbody>
</table>

- **Max-pooling**

- **Deconvolution**

<table>
<thead>
<tr>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>a0</td>
<td>b1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>c1</td>
<td>d3</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>x0</th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
</tr>
</thead>
<tbody>
<tr>
<td>x4</td>
<td>x5</td>
<td>x6</td>
<td>x7</td>
</tr>
<tr>
<td>x8</td>
<td>x9</td>
<td>x10</td>
<td>x11</td>
</tr>
<tr>
<td>x12</td>
<td>x13</td>
<td>x14</td>
<td>x15</td>
</tr>
</tbody>
</table>
Encoder-Decoder Networks
Encoder-Decoder Networks
Different Encoding Block Types

- VGG
- Inception
- ResNet
Encoder-Decoder Networks

Different Encoding Block Types

- VGG

Input

Max-Pool

Conv 3x3

Conv 3x3

Conv 3x3

Output
Encoder-Decoder Networks

Different Encoding Block Types

- **VGG**
- **Inception**
- **ResNet**

Max-Pool
Conv 1x1
Max-Pool
Conv 1x1
Conv 3x3
Conv 5x5
Concat
Conv 1x1

Input
Output

112 × 112 × 32
56 × 56 × 64
56 × 56 × 128
28 × 28 × 256
28 × 28 × 512
14 × 14 × 1024
7 × 7 × 2048
Encoder-Decoder Networks
Different Encoding Block Types

- VGG
- Inception
- ResNet

\[(224)^2 \times 3\]

Input

\[
\begin{array}{c}
112 \times 112 \\
\times 32 \\
\end{array}
\]

\[
\begin{array}{c}
56 \times 56 \\
\times 64 \\
\end{array}
\]

\[
\begin{array}{c}
56 \times 56 \\
\times 128 \\
\end{array}
\]

\[
\begin{array}{c}
28 \times 28 \\
\times 256 \\
\end{array}
\]

\[
\begin{array}{c}
28 \times 28 \\
\times 512 \\
\end{array}
\]

\[
\begin{array}{c}
14 \times 14 \\
\times 1024 \\
\end{array}
\]

\[
\begin{array}{c}
7 \times 7 \\
\times 2048 \\
\end{array}
\]

Output

- ResNet

Conv 3x3

Sum

Input

Conv 3x3
Different Encoding Block Types

Performance on the ImageNet 2012 Validation Dataset

- **Memory per image**
  - VGG
  - Inception
  - ResNet-18

- **Parameters**
  - VGG
  - Inception
  - ResNet-18

- **Inference Time**
  - VGG
  - Inception
  - ResNet-18

- **Classification Error**
  - VGG
  - Inception
  - ResNet-18
Encoder-Decoder Networks

Encoder

Decoder

(224)^2 \times 3

112 \times 112 \times 32

56 \times 56 \times 64

56 \times 56 \times 128

28 \times 28 \times 256

28 \times 28 \times 512

14 \times 14 \times 1024

7 \times 7 \times 2048

224 \times 224 \times C'

112 \times 112 \times 32

56 \times 56 \times 64

56 \times 56 \times 128

28 \times 28 \times 256

28 \times 28 \times 512

14 \times 14 \times 1024
Encoder-Decoder Networks
Encoder-Decoder Networks
Different Decoding Block Types

- VGG
- Inception
- ResNet
Encoder-Decoder Networks
Different Decoding Block Types

- VGG

Input

Un-Pool

Conv 3x3

Conv 3x3

Conv 3x3

Output
Encoder-Decoder Networks
Different Decoding Block Types

- VGG
- Inception
- ResNet
Encoder-Decoder Networks

Different Decoding Block Types

- VGG
- Inception
- ResNet

DeConv 3x3
Sum
Input
Output
Classification vs Segmentation

(a) ImageNet Classification Validation Set

(b) PASCAL VOC 2011 Validation Set

Our Work on Segmenting GigaPixel Breast Biopsy Images
Challenges with the dataset

- Limited computational resources

- Sliding window approach is promising but
  - Size of patch determines the context
  - Some biological structures may cover several patches
Challenges with the dataset

> Some biological structures are rare
  – Necrosis and Secretion have less than 1% of all the pixels

<table>
<thead>
<tr>
<th>Diagnosis Category</th>
<th>#ROIs (total)</th>
<th>#ROIs (train)</th>
<th>#ROIs (test)</th>
<th>Avg. ROI size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benign</td>
<td>9</td>
<td>4</td>
<td>5</td>
<td>$9K \times 9K$</td>
</tr>
<tr>
<td>Atypia</td>
<td>22</td>
<td>11</td>
<td>11</td>
<td>$6K \times 7K$</td>
</tr>
<tr>
<td>DCIS</td>
<td>22</td>
<td>12</td>
<td>10</td>
<td>$8K \times 10K$</td>
</tr>
<tr>
<td>Invasive</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>$38K \times 44K$</td>
</tr>
<tr>
<td>total</td>
<td>58</td>
<td>30</td>
<td>28</td>
<td>$10K \times 12K$</td>
</tr>
</tbody>
</table>
Training details

> Training Set: 30 ROIs
  – 25,992 patches of size 256x256 with augmentation
  – Split into training and validation set using 90:10 ratio

> Test Set: 28 ROIs

> Stochastic Gradient Descent for optimization

> Implemented in Torch
  – http://torch.ch/
Segmentation Results

RGB Image

Ground Truth Label

Legend:
- background
- epithelium
- stroma
- secretion
- necrosis
Segmentation Results

Encoder-Decoder Network with skip connection
Segmentation Results

Multi-Resolution Encoder-Decoder Network
Segmentation Results

RGB Image

Ground Truth

Plain

Multi-Resolution

Legend:
- background
- epithelium
- stroma
- secretion
- necrosis
Segmentation Results

F1-Score

![F1-Score Graph]

- **Background**: SVM + SP > Plain + MR > Proposed
- **Epithelium**: SVM + SP > Plain + MR > Proposed
- **Stroma**: SVM + SP > Plain + MR > Proposed
- **Secretion**: SVM + SP > Plain + MR > Proposed
- **Necrosis**: SVM + SP > Plain + MR > Proposed
Why Segmentation?

Results on Diagnosis

- Segmented whole dataset (428 ROIs) with the model trained on 30 ROIs
- Extracted histograms from segmentation masks and then trained different classifiers
- Weak classifiers are as good as strong classifiers

<table>
<thead>
<tr>
<th>classification task</th>
<th>Multilayer Perceptron</th>
<th>SVM with RBF kernel</th>
<th>Logistic Regression</th>
<th>Random Forest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Invasive vs others</td>
<td>.72</td>
<td>.78</td>
<td>.75</td>
<td>.80</td>
</tr>
<tr>
<td>Benign vs others</td>
<td>.65</td>
<td>.54</td>
<td>.60</td>
<td>.62</td>
</tr>
<tr>
<td>Atypia vs DCIS</td>
<td>.75</td>
<td>.73</td>
<td>.77</td>
<td>.75</td>
</tr>
</tbody>
</table>
Thank You!!
References

Two 3x3 filters are same as one 5x5 filter

Source: Rethinking the Inception Architecture for Computer Vision by Szegedy et al.