Structured Out Prediction (Semantic Segmentation)

Hyun Soo Park
CHALLENGES OF VISUAL RECOGNITION

• Appearance
  • DOF: texture, illumination, material, shading, ...
• Shape
  • DOF: object category, geometric pose, viewpoint, ...
CHALLENGES OF VISUAL RECOGNITION

• Appearance
  • DOF: texture, illumination, material, shading, …
• Shape
  • DOF: object category, geometric pose, viewpoint, …

\[ f(I) = l_{\text{human}} \]
Semantic Segmentation: Pixel Classification

0: Background/Unknown
1: Person
2: Purse
3: Plants/Grass
4: Sidewalk
5: Building/Structures
SEMANTIC SEGMENTATION FORMULATION

Unsupervised superpixel segmentation

Visual feature
- Color histogram
- BoW
- SIFT
- HOG

Classification

\[ L = \phi(x_i) \quad \text{e.g., } \phi(x_i) = w_{tree} \cdot x_i \]
**Semantic Segmentation Formulation**

Unsupervised superpixel segmentation

Visual feature
- Color histogram
- BoW
- SIFT
- HOG

\[ L = \phi(x_i) + \varphi(x_i, x_j) \]

CRF: Conditional Random Field
aka. joint classifications (structured pred.)
HOLISTIC PREDICTION VIA DEEP LEARNING

HxWx3 → HxWxD → Fully convolutional layers → Softmax → HxWxC

Computationally inefficient
**Holistic Prediction via Deep Learning**

HxWxD

Softmax

HxWxD

Bottleneck layer

Softmax

HxWxD

HxWx3

HxWx3

Upsampling?
**REVISITED: SPATIAL POOLING (DOWN-SAMPLING)**

Max-pooling (window size 2x2, stride 2)
**SPATIAL UNPOOLING (UP-SAMPLING)**

Max-unpooling (window size 2x2, stride 2)

Learnable parameter?
Can we learn upsampling?
LEARNING UPSAMPLING

Low-Resolution Image

High-Resolution Image
High-Resolution Super-Resolved Image
High-Resolution Reference Image
**Spatial Unpooling (Up-sampling)**

Max-unpooling (window size 2x2, stride 2)

---

Learnable parameter?
Can we learn upsampling?
### Revisited: Convolution

The convolution operation can be represented as:

\[ y_{11} = w_{11}x_{11} + w_{12}x_{12} + \cdots + w_{33}x_{33} \]
\[ y_{12} = w_{11}x_{12} + w_{12}x_{13} + \cdots + w_{33}x_{34} \]
\[ y_{21} = w_{11}x_{21} + w_{12}x_{21} + \cdots + w_{33}x_{43} \]
\[ y_{22} = w_{11}x_{22} + w_{12}x_{23} + \cdots + w_{33}x_{44} \]

The convolution can be visualized as:

\[ \begin{array}{c|c}
  x_{11} & x_{12} \\
  x_{21} & x_{22} \\
  \hline
  x_{31} & x_{32} \\
  x_{41} & x_{42} \\
\end{array} \quad \times \quad \begin{array}{c|c}
  w_{11} & w_{12} \\
  w_{21} & w_{22} \\
\end{array} \quad = \quad \begin{array}{c|c}
  y_{11} & y_{12} \\
  y_{21} & y_{22} \\
\end{array} \]
**Revisited: Convolution**

\[
\begin{pmatrix}
  x_{11} & x_{12} \\
  \cdots & \cdots \\
  x_{44}
\end{pmatrix}
\times
\begin{pmatrix}
  w_{11} & w_{12} \\
  \cdots & \cdots \\
  w_{13}
\end{pmatrix}
= 
\begin{pmatrix}
  y_{11} \\
  y_{12} \\
  \vdots \\
  y_{22}
\end{pmatrix}
\]

\[
\begin{pmatrix}
  y_{11} \\
  y_{12} \\
  y_{21} \\
  y_{22}
\end{pmatrix} =
\begin{pmatrix}
  w_{11} & w_{12} & w_{13} & 0 & w_{21} & w_{22} & w_{23} & 0 & w_{31} & w_{32} & w_{33} & 0 & 0 & 0 & 0 & 0 \\
  0 & w_{11} & w_{12} & w_{13} & 0 & w_{21} & w_{22} & w_{23} & 0 & w_{31} & w_{32} & w_{33} & 0 & 0 & 0 & 0 \\
  0 & 0 & 0 & 0 & w_{11} & w_{12} & w_{13} & 0 & w_{21} & w_{22} & w_{23} & 0 & w_{31} & w_{32} & w_{33} & 0 \\
  0 & 0 & 0 & 0 & 0 & w_{11} & w_{12} & w_{13} & 0 & w_{21} & w_{22} & w_{23} & 0 & w_{31} & w_{32} & w_{33} & 0 \\
  \vdots \\
  0 & 0 & 0 & 0 & 0 & 0 & w_{11} & w_{12} & w_{13} & 0 & w_{21} & w_{22} & w_{23} & 0 & w_{31} & w_{32} & w_{33}
\end{pmatrix}
\begin{pmatrix}
  x_{11} \\
  x_{12} \\
  \vdots \\
  x_{44}
\end{pmatrix}
\]
**Revisited: Convolution**

\[
\begin{array}{cc}
X_{11} & X_{12} \\
& \\
& X_{44}
\end{array}
\quad \otimes 
\quad \begin{array}{cc}
W_{11} & W_{12} \\
& W_{13}
\end{array}
= 
\begin{array}{cc}
y_{11} & y_{12} \\
y_{21} & y_{22}
\end{array}
\]

\[Y = WX\]

Non-zero entries of each row encode local connectivity.
**Inverse of Convolution**

\[
\begin{pmatrix}
X_{11} & X_{12} \\
X_{44} & \\
\end{pmatrix}
= 
\begin{pmatrix}
y_{11} & y_{12} \\
y_{22} & \\
\end{pmatrix}
\begin{pmatrix}
W_{11} & W_{12} \\
W_{13} & \\
\end{pmatrix}^{-1}
\]

\[
X = W^+ Y
\]

Inverse of W does not preserve local connectivity.
**Upconvolution ~ Inverse of Convolution**

\[
\begin{pmatrix}
  x_{11} & x_{12} \\
  \vdots & \vdots \\
  x_{44} & \\
\end{pmatrix}
\approx
\begin{pmatrix}
  y_{11} & y_{12} \\
  \vdots & \vdots \\
  y_{22} & \\
\end{pmatrix}
\otimes
\begin{pmatrix}
  z_{11} & z_{12} \\
  \vdots & \vdots \\
  z_{33} & \\
\end{pmatrix}

W = Z
\text{if } W^T W \approx I

Non-zero entries of column encode local connectivity.
**Upconvolution ~ Inverse of Convolution**

The diagram illustrates the process of upconvolution as the inverse of convolution. The upconvolution operation is represented by the transformation of a lower-dimensional input $X$ to a higher-dimensional output $Y$, denoted as $Z^T Y = X$. The diagram shows the learnable kernel and how non-zero entries of column encode local connectivity.
Transposed Convolution ~ Inverse of Convolution

Non-zero entries of column encode local connectivity.
**Transposed Convolution ~ Inverse of Convolution**

\[
\begin{align*}
X & = 
\begin{pmatrix}
X_{11} & X_{12} \\
& \\
X_{44} & 
\end{pmatrix} \\
Y & = 
\begin{pmatrix}
y_{11} & y_{12} \\
y_{22} & 
\end{pmatrix} \\
\bar{Z}^T & = 
\begin{pmatrix}
\bar{z}_{11} & \bar{z}_{12} \\
& \\
\bar{z}_{33} & 
\end{pmatrix}
\end{align*}
\]

Learnable kernel

\[
Z^T Y = X \implies y_{11} + y_{12} + y_{21} + y_{22} = X
\]

Non-zero entries of column encode local connectivity.
**Transposed Convolution as Upconvolution**

\[
\begin{align*}
\begin{array}{c}
\begin{array}{c}
X_{11} & X_{12} \\
\end{array}
\end{array}
\end{align*}
\begin{align*}
\begin{array}{c}
\begin{array}{c}
Y_{11} & Y_{12} \\
\end{array}
\end{array}
\end{align*}
\begin{align*}
\begin{array}{c}
\begin{array}{c}
X_{44} \\
\end{array}
\end{array}
\end{align*}
\begin{align*}
\begin{array}{c}
\begin{array}{c}
Z_{11} & Z_{12} & Z_{13} \\
Z_{21} & Z_{22} & Z_{23} \\
Z_{31} & Z_{32} & Z_{33} \\
\end{array}
\end{array}
\end{align*}
\begin{align*}
\begin{array}{c}
\begin{array}{c}
X_{11} & X_{12} \\
\end{array}
\end{array}
\end{align*}
\begin{align*}
\begin{array}{c}
\begin{array}{c}
W_{11} & W_{12} \\
W_{13} \\
\end{array}
\end{array}
\end{align*}
\begin{align*}
\begin{array}{c}
\begin{array}{c}
X_{44} \\
\end{array}
\end{array}
\end{align*}
\begin{align*}
\begin{array}{c}
\begin{array}{c}
Y_{11} & Y_{12} \\
\end{array}
\end{array}
\end{align*}
\begin{align*}
\begin{array}{c}
\begin{array}{c}
Y_{22} \\
\end{array}
\end{array}
\end{align*}
\end{align*}
\]

Upconvolution

Cf) convolution

\[
x_{11} = z_{11}y_{11}
\]

\[
x_{11} \text{ is dependent only on } y_{11}.
\]

\[
x_{11} \text{ contributes only to } y_{11}.
\]
**Transposed Convolution as Upconvolution**

\[ x_{11} \quad x_{12} \quad \Box \quad x_{44} \]

\[ \begin{array}{cc} y_{11} & y_{12} \\ y_{22} & \end{array} \]

\[ \begin{array}{ccc} z_{11} & z_{12} & z_{13} \\ z_{21} & z_{22} & z_{23} \\ z_{31} & z_{32} & z_{33} \end{array} \]

\[ \times \quad \uparrow \]

**Upconvolution**

\[ Z^T \]

\[ Y = X \]

\[ x_{11} = z_{11}y_{11} \]

\[ x_{12} = z_{12}y_{11} + z_{11}y_{12} \]
**Transposed Convolution as Upconvolution**

\[
\begin{bmatrix}
  x_{11} & x_{12} \\
  x_{21} & x_{22} \\
  x_{31} & x_{32} \\
  x_{41} & x_{42}
\end{bmatrix}
= \begin{bmatrix}
  y_{11} & y_{12} \\
  y_{21} & y_{22}
\end{bmatrix}
\otimes
\begin{bmatrix}
  z_{11} & z_{12} & z_{13} \\
  z_{21} & z_{22} & z_{23} \\
  z_{31} & z_{32} & z_{33}
\end{bmatrix}
\]

**Upconvolution**

\[
Z^T Y = X
\]

\[
x_{11} = z_{11} y_{11} \\
x_{12} = z_{12} y_{11} + z_{11} y_{12} \\
x_{13} = z_{13} y_{11} + z_{12} y_{12}
\]
**Transposed Convolution as Upconvolution**

\[
x_{11} x_{12} \quad y_{11} y_{12} \quad z_{11} z_{12} z_{13} \quad z_{21} z_{22} z_{23} \quad z_{31} z_{32} z_{33}
\]

\[
Y = X
\]

Upconvolution

\[
x_{11} = z_{11} y_{11}
\]
\[
x_{12} = z_{12} y_{11} + z_{11} y_{12}
\]
\[
x_{13} = z_{13} y_{11} + z_{12} y_{12}
\]
\[
x_{14} = z_{13} y_{12}
\]
TRANSPOSED CONVOLUTION AS UPCONVOLUTION

\[
x_{11} x_{12} = y_{11} y_{12} \quad \otimes \uparrow \quad \begin{bmatrix} z_{11} & z_{12} & z_{13} \\ z_{21} & z_{22} & z_{23} \\ z_{31} & z_{32} & z_{33} \end{bmatrix} \quad \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} =
\]

Upconvolution

\[
x_{11} = z_{11} y_{11}
\]

Convolution with zero padding

Flipped kernel

\[
X^T Y = X
\]
**Transposed Convolution as UpConvolution**

Upconvolution:

\[
x_{11} x_{12} 
\]

\[
y_{11} y_{12} \otimes \uparrow
\]

\[
z_{11} z_{12} z_{13}
\]

\[
z_{21} z_{22} z_{23}
\]

\[
z_{31} z_{32} z_{33}
\]

\[
\rightarrow
\]

Convolution with zero padding:

\[
x_{11} x_{12} 
\]

\[
y_{11} y_{12} \otimes 
\]

\[
x_{21} y_{21} y_{22} \otimes 
\]

\[
x_{31} z_{31} z_{32} z_{33} \otimes 
\]

\[
x_{41} z_{41} z_{42} z_{43} \otimes 
\]

\[
x_{42} z_{42} z_{43} \otimes 
\]

\[
x_{43} z_{43} \otimes 
\]

\[
x_{44} \otimes 
\]

\[
x_{11} = z_{11} y_{11}
\]

\[
x_{12} = z_{12} y_{11} + z_{11} y_{12}
\]

\[
Z^T
\]

\[
Y = X
\]
**Transposed Convolution as Upconvolution**

\[
\begin{align*}
X_{11} & \times X_{12} = Y_{11} \times Y_{12} \\
& \quad \uparrow \otimes \quad \downarrow \\
& \quad Z_{11} \times Z_{12} \times Z_{13} \\
& \quad Z_{21} \times Z_{22} \times Z_{23} \\
& \quad Z_{31} \times Z_{32} \times Z_{33}
\end{align*}
\]

**Upconvolution**

**Convolution with zero padding**

\[
\begin{align*}
X_{11} &= Z_{11}Y_{11} \\
X_{12} &= Z_{12}Y_{11} + Z_{11}Y_{12} \\
X_{13} &= Z_{13}Y_{11} + Z_{12}Y_{12}
\end{align*}
\]
**Transposed Convolution as Upconvolution**

\[
\begin{bmatrix}
  x_{11} & x_{12} \\
  x_{21} & x_{22}
\end{bmatrix}
\ast
\begin{bmatrix}
  y_{11} & y_{12} \\
  y_{21} & y_{22}
\end{bmatrix}
= \begin{bmatrix}
  z_{11} & z_{12} & z_{13} \\
  z_{21} & z_{22} & z_{23} \\
  z_{31} & z_{32} & z_{33}
\end{bmatrix}
\]

\[
\begin{bmatrix}
  x_{11} & x_{12} \\
  x_{21} & x_{22}
\end{bmatrix}
= \begin{bmatrix}
  y_{11} & y_{12} \\
  y_{21} & y_{22}
\end{bmatrix}
\ast
\begin{bmatrix}
  z_{33} & z_{32} & z_{31} \\
  z_{23} & z_{22} & z_{21} \\
  z_{13} & z_{12} & z_{11}
\end{bmatrix}
\]

**Upconvolution**

**Convolution with zero padding**

\[
x_{11} = z_{11}y_{11}
\]

\[
x_{12} = z_{12}y_{11} + z_{11}y_{12}
\]

\[
x_{13} = z_{13}y_{11} + z_{12}y_{12}
\]

\[
x_{14} = z_{13}y_{12}
\]
# Transposed Convolution as Upconvolution

| $x_{11}$ | $x_{12}$ |  | $y_{11}$ | $y_{12}$ |  | $z_{11}$ | $z_{12}$ | $z_{13}$ |  | $x_{11}$ | $x_{12}$ |
|---------|---------|  |---------|---------|  |---------|---------|---------|  |---------|---------|
|         |         |  |         |         |  |         |         |         |  |         |         |
|         |         |  |         |         |  |         |         |         |  |         |         |
| $x_{44}$|         |  | $y_{22}$|         |  | $z_{31}$| $z_{32}$| $z_{33}$|  | $x_{44}$|         |

Upconvolution

Convolution with zero padding

$$Z^T Y = X$$

- $x_{11} = z_{11} y_{11}$
- $x_{12} = z_{12} y_{11} + z_{11} y_{12}$
- $x_{13} = z_{13} y_{11} + z_{12} y_{12}$
- $x_{14} = z_{13} y_{12}$
- $x_{21} = z_{21} y_{11} + z_{11} y_{21}$
- $x_{22} = z_{22} y_{11} + z_{21} y_{12} + z_{12} y_{21} + z_{11} y_{22}$
- $x_{23} = z_{23} y_{11} + z_{22} y_{12} + z_{13} y_{21} + z_{12} y_{22}$
- $x_{24} = z_{23} y_{12} + z_{13} y_{22}$

Flipped kernel
**Transposed Convolution as Upconvolution**

\[
\begin{array}{c|c|c}
\; & y_{11} & y_{12} \\
\hline
x_{11} & x_{12} & y_{22} \\
\hline
\end{array}
\]

\[
\begin{array}{c|c|c}
\; & z_{11} & z_{12} & z_{13} \\
\hline
z_{21} & z_{22} & z_{23} \\
\hline
z_{31} & z_{32} & z_{33} \\
\end{array}
\]  

\[
\begin{array}{c|c|c}
\; & x_{11} & x_{12} \\
\hline
y_{11} & y_{12} & y_{22} \\
\hline
y_{21} & y_{22} & x_{44} \\
\end{array}
\]

**Upconvolution**

**Convolution with zero padding**

**Flipped kernel**

\[
\begin{array}{c|c|c}
\; & z_{33} & z_{32} & z_{31} \\
\hline
z_{23} & z_{22} & z_{21} \\
\hline
z_{13} & z_{12} & z_{11} \\
\end{array}
\]

**Upconvolution**
**Duality: Convolution vs. Upconvolution**

Convolution (3x3 kernel, no zero padding, 1 stride)

Upconvolution (3x3 kernel, 2x2 zero padding, 1 stride)
**Duality: Convolution vs. Upconvolution**

Convolution (3x3 kernel, no zero padding, 2 stride)

Upconvolution (3x3 kernel, 2x2 zero padding, 1 zero between input, 2 stride)
FULLY CONVOLUTIONAL NETWORK

227 x 227  55 x 55  27 x 27  13 x 13

“tabby cat”

227x227  10x10  Upsampling via deconv.

FULLY CONVOLUTIONAL NETWORK

DEEPER UPCONVOLUTIONAL LAYERS

Noh, Hong, and Han, “Learning Deconvolutional Network for Semantic Segmentation”, ICCV 2015
Receptive Field vs. Resolution

**Convolution network**

Desired properties:
- Larger receptive field (bigger spatial context) → reducing resolution
- Higher output resolution → reducing receptive field
**DEEPLAB: DILATED CONVOLUTION**

Enlarging receptive field without increasing reducing resolution

Chen, Papandreou, Kokkinos, Murphy, and Yuille, “Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs”, ICLR 2015
DEEPLAB: DILATED CONVOLUTION

224x224x3

28x28x1024

28x28x4096

1x1xN

final classification

1x1xN

coarse segmentation

28x28xN

final segmentation

bilinear upsampling (8x) to original resolution

dilated convolutions, stride 1, no pooling preserve spatial dims

1x1 convolutions, stride 1 preserve spatial dims
DeepLab: CRF Post-processing

Chen, Papandreou, Kokkinos, Murphy, and Yuille, “Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs”, ICLR 2015
DeepLab

Chen, Papandreou, Kokkinos, Murphy, and Yuille, “Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs”, ICLR 2015
Chen, Papandreou, Kokkinos, Murphy, and Yuille, “Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs”, ICLR 2015
CRF as Recurrent Neural Net (End-to-End)

CRF as Recurrent Neural Net (End-to-End)