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, ...

\[ f(I) = l_{human} \]

Human

Semantic segmentation

Human

Grass
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

\[ L = \phi(x_i) \quad \text{e.g.,} \quad \phi(x_i) = w_{\text{tree}} \cdot x_i \]
SEMANTIC SEGMENTATION FORMULATION

Unsupervised superpixel segmentation

$\mathbf{x}_i$ Visual feature
- Color histogram
- BoW
- SIFT
- HOG

$\mathbf{x}_j$

Classification

$L = \phi(\mathbf{x}_i) + \varphi(\mathbf{x}_i, \mathbf{x}_j)$

Context

CRF: Conditional Random Field
aka. joint classifications (structured pred.)
**Holistic Prediction via Deep Learning**

HxWx3 $\rightarrow$ HxWxD $\rightarrow$ Fully convolutional layers $\rightarrow$ Softmax $\rightarrow$ HxWxCi

Computationally inefficient
**Holistic Prediction via Deep Learning**

- **Input:** $H \times W \times 3$
- **Output:** $H \times W \times 1$

**Upper Pathway:**
- $H \times W \times D$
- Softmax
- $H \times W \times 1$

**Lower Pathway:**
- $H \times W \times D$
- Bottleneck layer
- Softmax
- $H \times W \times 1$
FULLY CONVOLUTIONAL NETWORK

“tabby cat”

FULLY CONVOLUTIONAL NETWORK

HOLISTIC PREDICTION VIA DEEP LEARNING

HxWx3 → HxWxD → BOTTLENECK LAYER → SOFTMAX → HxWxC1

HxWx3 → HxWxD → SOFTMAX → HxWxC1
Holistic Prediction via Deep Learning

HxWxD → Softmax → HxWxCl

HxWx3 → HxWxD → Bottleneck layer → Upsampling? → Softmax → HxWxCl
**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**

\[
\begin{array}{ccc}
\times_{11} & \times_{12} & \\
\times_{21} & \times_{22} & \times_{44}
\end{array}
\otimes
\begin{array}{ccc}
\times_{11} & \times_{12} & \\
\times_{21} & \times_{22} & \times_{44}
\end{array}
\]

\[
y_{11} = w_{11}x_{11} + w_{12}x_{12} + \cdots + w_{33}x_{33}
\]

\[
y_{12} = w_{11}x_{12} + w_{12}x_{13} + \cdots + w_{33}x_{34}
\]

\[
y_{21} = w_{11}x_{21} + w_{12}x_{21} + \cdots + w_{33}x_{43}
\]

\[
y_{22} = w_{11}x_{22} + w_{12}x_{23} + \cdots + w_{33}x_{44}
\]
**Revisited: Convolution**

\[
\begin{bmatrix}
  x_{11} & x_{12} \\
  x_{21} & x_{22} \\
  x_{31} & x_{32} \\
  x_{41} & x_{42}
\end{bmatrix}
\times
\begin{bmatrix}
  w_{11} & w_{12} & w_{13} \\
  & & \\
  w_{21} & w_{22} & w_{23} \\
  & & \\
  w_{31} & w_{32} & w_{33}
\end{bmatrix}
= 
\begin{bmatrix}
  y_{11} & y_{12} \\
  y_{21} & y_{22} \\
  y_{31} & y_{32} \\
  y_{41} & y_{42}
\end{bmatrix}
\]
**Revisited: Convolution**

\[
\begin{array}{cc}
X_{11} & X_{12} \\
X_{21} & X_{22} \\
X_{31} & X_{32} \\
X_{41} & X_{42} \\
\end{array}
\quad \otimes \quad
\begin{array}{cc}
W_{11} & W_{12} \\
W_{21} & W_{22} \\
W_{31} & W_{32} \\
W_{41} & W_{42} \\
\end{array}
= \begin{array}{cc}
y_{11} & y_{12} \\
y_{21} & y_{22} \\
\end{array}
\]

\[
Y = \begin{array}{cc}
W_{11} & W_{12} \\
W_{21} & W_{22} \\
W_{31} & W_{32} \\
W_{41} & W_{42} \\
\end{array}
\quad X
\]

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

\[
\begin{bmatrix}
  x_{11} & x_{12} \\
  x_{21} & x_{22} \\
  x_{31} & x_{32} \\
  x_{41} & x_{42}
\end{bmatrix}
= 
\begin{bmatrix}
  y_{11} & y_{12} \\
  y_{21} & y_{22}
\end{bmatrix}
\otimes
\begin{bmatrix}
  w_{11} & w_{12} \\
  w_{21} & w_{22}
\end{bmatrix}^{-1}
\]

\[
X = W^+ Y
\]

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

Non-zero entries of column encode local connectivity.

\[ W = Z \]
if \( W^T W \approx I \)
**Upconvolution ~ Inverse of Convolution**

![Upconvolution Diagram]

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

\[ Z^T Y \approx X \]

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

\[ Z^T Y = X \]

Non-zero entries of column encode local connectivity.
TRANSPOSED CONVOLUTION AS UPCONVOLUTION

\[
\begin{align*}
X_{11} & \times X_{12} \\
X_{13} & \times X_{14} \\
X_{41} & \times X_{42} \\
X_{43} & \times X_{44}
\end{align*}
\]

\[
\begin{align*}
Y_{11} & \times Y_{12} \\
Y_{13} & \times Y_{14} \\
Y_{21} & \times Y_{22} \\
Y_{23} & \times Y_{24}
\end{align*}
\]

\[
\begin{align*}
Z_{11} & \times Z_{12} \times Z_{13} \\
Z_{21} & \times Z_{22} \times Z_{23} \\
Z_{31} & \times Z_{32} \times Z_{33}
\end{align*}
\]

\[
\begin{align*}
W_{11} & \times W_{12} \\
W_{13} & \times W_{14} \\
W_{21} & \times W_{22} \\
W_{23} & \times W_{24}
\end{align*}
\]

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

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

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

\[
y_{11} \quad y_{12} \\
\downarrow \quad \downarrow \\
y_{22}
\]

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

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

\[
\begin{align*}
    x_{11} & \quad x_{12} \\
    x_{11} & \quad x_{12} \\
    x_{11} & \quad x_{12} \\
    x_{44} & = \begin{bmatrix} y_{11} & y_{12} \\ y_{22} \end{bmatrix} \times \begin{bmatrix} z_{11} & z_{12} & z_{13} \\ z_{21} & z_{22} & z_{23} \\ z_{31} & z_{32} & z_{33} \end{bmatrix} \\
    \text{Upconvolution} \\
    Z^T & \quad Y = X \\
    \begin{bmatrix} x_{11} \end{bmatrix} & = z_{11}y_{11} \\
    \begin{bmatrix} x_{12} \end{bmatrix} & = z_{12}y_{11} + z_{11}y_{12} \\
    \begin{bmatrix} x_{13} \end{bmatrix} & = z_{13}y_{11} + z_{12}y_{12} \\ 
\end{align*}
\]
**Transposed Convolution as Upconvolution**

\[
\begin{array}{cccc}
 x_{11} & x_{12} & \cdots & x_{44} \\
 \vdots & \vdots & \ddots & \vdots \\
 x_{11} & x_{12} & \cdots & x_{44} \\
\end{array}
\]

\[
\begin{array}{cccc}
 y_{11} & y_{12} & \cdots & y_{22} \\
 \vdots & \vdots & \ddots & \vdots \\
 y_{11} & y_{12} & \cdots & y_{22} \\
\end{array}
\]

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

**Upconvolution**

\[
Y = X
\]

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

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

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

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

\[
\begin{bmatrix}
  x_{11} & x_{12} \\
  & \phantom{x} & \phantom{x} \\
  & \phantom{x} & \phantom{x} \\
  x_{44} & \phantom{x} & \phantom{x}
\end{bmatrix}
\otimes
\begin{bmatrix}
  y_{11} & y_{12} & y_{22} \\
  & \phantom{y} & \phantom{y} & \phantom{y} \\
  & \phantom{y} & \phantom{y} & \phantom{y} \\
  & \phantom{y} & \phantom{y} & \phantom{y}
\end{bmatrix}
\uparrow
\Rightarrow
\begin{bmatrix}
  z_{11} & z_{12} & z_{13} \\
  z_{21} & z_{22} & z_{23} \\
  z_{31} & z_{32} & z_{33}
\end{bmatrix}
\]

**Upconvolution**

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

**Convolution with zero padding**

\[
\begin{bmatrix}
  y_{11} & y_{12} \\
  y_{21} & y_{22} \\
  & \phantom{y} & \phantom{y} & \phantom{y} \\
  & \phantom{y} & \phantom{y} & \phantom{y}
\end{bmatrix}
\otimes
\begin{bmatrix}
  z_{33} & z_{32} & z_{31} \\
  z_{23} & z_{22} & z_{21} \\
  z_{13} & z_{12} & z_{11}
\end{bmatrix}
\]

**Flipped kernel**
**Transposed Convolution as Upconvolution**

\[
x_{11} x_{12} = y_{11} y_{12} \quad \otimes \quad \uparrow \\
x_{11} x_{12} = y_{11} y_{12} \quad \otimes \quad \uparrow \\
\]

Upconvolution

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

Convolution with zero padding

\[
Z^T Y = X
\]

Flipped kernel

\[
\begin{bmatrix}
z_{33} & z_{32} & z_{31} \\
z_{23} & z_{22} & z_{21} \\
z_{13} & z_{12} & z_{11}
\end{bmatrix}
\]
**Transposed Convolution as Upconvolution**

\[ \begin{align*}
X_{11} & \quad X_{12} \\
X_{21} & \quad X_{22} \\
X_{31} & \quad X_{32} \\
X_{41} & \quad X_{42}
\end{align*} \]

\[ \begin{align*}
Y_{11} & \quad Y_{12} \\
Y_{21} & \quad Y_{22}
\end{align*} \]

\[ \begin{align*}
Z_{11} & \quad Z_{12} \quad Z_{13} \\
Z_{21} & \quad Z_{22} \quad Z_{23} \\
Z_{31} & \quad Z_{32} \quad Z_{33}
\end{align*} \]

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

Upconvolution

Convolution with zero padding

Flipped kernel
**Transposed Convolution as Upconvolution**

\[
\begin{align*}
\begin{pmatrix}
x_{11} & x_{12} \\
x_{21} & x_{22} \\
x_{31} & x_{32} \\
x_{41} & x_{42} \\
x_{43} & x_{44} \\
\end{pmatrix}
&= \begin{pmatrix}
y_{11} & y_{12} \\
y_{21} & y_{22} \\
y_{31} & y_{32} \\
y_{41} & y_{42} \\
y_{43} & y_{44} \\
\end{pmatrix} \otimes \begin{pmatrix}
z_{11} & z_{12} & z_{13} \\
z_{21} & z_{22} & z_{23} \\
z_{31} & z_{32} & z_{33} \\
\end{pmatrix}
\end{align*}
\]

\[
\begin{align*}
x_{11} &= z_{11}y_{11} \\
x_{12} &= z_{12}y_{11} + z_{11}y_{12} \\
x_{13} &= z_{13}y_{11} + z_{12}y_{12} \\
x_{14} &= z_{13}y_{12}
\end{align*}
\]
**Transposed Convolution as UpConvolution**

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

**Upconvolution**

\[ X \]
= \[ Y \]
\[ \otimes \]
\[ Z^T \]

**Convolution with zero padding**

\[ x_{11} = z_{11}y_{11} \]
\[ x_{12} = z_{12}y_{11} + z_{11}y_{12} \]
\[ x_{13} = z_{13}y_{11} + z_{12}y_{12} \]
\[ x_{14} = z_{13}y_{12} \]
\[ x_{21} = z_{21}y_{11} + z_{11}y_{21} \]
\[ x_{22} = z_{22}y_{11} + z_{21}y_{12} + z_{12}y_{21} + z_{11}y_{22} \]
\[ x_{23} = z_{23}y_{11} + z_{22}y_{12} + z_{13}y_{21} + z_{12}y_{22} \]
\[ x_{24} = z_{23}y_{12} + z_{13}y_{22} \]
**Transposed Convolution as Upconvolution**

\[
\begin{array}{c|c|c|c}
 x_{11} & x_{12} & ~ & ~ \\
 ~ & ~ & ~ & ~ \\
 ~ & ~ & ~ & ~ \\
 x_{44} & & & ~
\end{array}
\]

\[
\begin{array}{c|c|c|c}
 y_{11} & y_{12} & ~ & ~ \\
 ~ & ~ & ~ & ~ \\
 ~ & ~ & ~ & ~ \\
 y_{22} & & & ~
\end{array}
\]

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

\[
\begin{array}{c|c|c|c|c|c}
 x_{11} & x_{12} & ~ & ~ & ~ & ~ \\
 ~ & ~ & ~ & ~ & ~ & ~ \\
 ~ & ~ & ~ & ~ & ~ & ~ \\
 x_{44} & & & ~ & ~ & ~
\end{array}
\]

\[
\begin{array}{c|c|c|c|c|c}
 y_{11} & y_{12} & ~ & ~ & ~ & ~ \\
 y_{21} & y_{22} & ~ & ~ & ~ & ~ \\
 ~ & ~ & ~ & ~ & ~ & ~ \\
 ~ & ~ & ~ & ~ & ~ & ~ \\
 ~ & ~ & ~ & ~ & ~ & ~ \\
 ~ & ~ & ~ & ~ & ~ & ~
\end{array}
\]

**Upconvolution**

**Convolution with zero padding**

**Flipped kernel**
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)
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

coarse segmentation 28x28N

bilinear upsampling (8x) to original resolution

final segmentation
DeepLab: CRF Post-processing

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
DEEPLAB

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)**