Outline

1. Recap MLP
2. Convolutional Neural Networks
3. Large deep convnets
4. Beyond ImageNet
   1. Fully Convolutional Networks (FCNs)
   2. Transfer
From ImageNet to complex scenes

- ImageNet: huge dataset (1.2M training images) with labels ... but centered objects

- How to apply/adapt/modify learning strategies to deal with:

  - ImageNet
  - VOC 2012
  - MS COCO
From ImageNet to complex scenes?

- Working on datasets with complex scenes (large and cluttered background), not centered objects, variable size, ...

- Select relevant regions $\rightarrow$ better prediction

- Full annotations expensive $\Rightarrow$ training with weak supervision
How to adapt VGG16 archi. for large/complex images?

Main work of my PhD student Thibaut Durand

Thibaut Durand, Nicolas Thome, and Matthieu Cord

WELDON: Weakly supervised learning of deep convolutional neural networks
In CVPR, 2016.

Thibaut Durand, Taylor Mordan, Nicolas Thome, and Matthieu Cord

WILDCAT: Weakly Supervised Learning of Deep ConvNets for Image Classification, Pointwise Localization and Segmentation.
In CVPR, 2017.
Naïve approach: brut transfer (next Section)

- Resize the image
Sliding window ⇒ convolutional layers
Sliding window $\Rightarrow$ convolutional layers

- Fully connected as convolutional layer (here 4096 conv. filters $7 \times 7 \times 512$)
Sliding window $\Rightarrow$ convolutional layers

$h' \times w' \times 3 \rightarrow h' \times w' \times 64 \rightarrow \frac{h'}{2} \times \frac{w'}{2} \times 128 \rightarrow \frac{h'}{4} \times \frac{w'}{4} \times 256 \rightarrow \frac{h'}{8} \times \frac{w'}{8} \times 512 \rightarrow \frac{h'}{16} \times \frac{w'}{16} \times 512 \rightarrow h \times w \times 4096$

$h = \frac{h'}{32} - 6 \quad w = \frac{w'}{32} - 6$

- convolution + ReLU
- max pooling
Transfer – Pooling – Classification

Feature extraction network

\[ \text{d} \times \text{h} \times \text{w} \rightarrow \text{C} \times 1 \times 1 \]
Image-based strategy

- Global Average Pooling (GoogLeNet, ResNet)

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba
Learning Deep Features for Discriminative Localization.
In *CVPR*, 2016.
Region-based strategy

- **Deep MIL**

  Maxime Oquab, Léon Bottou, Ivan Laptev and Josef Sivic
  In *CVPR*, 2015.

- **WELDON and ProNet [Sun, CVPR16]**

  Thibaut Durand, Nicolas Thome, and Matthieu Cord
  WELDON: Weakly Supervised Learning of Deep ConvNets.
  In *CVPR*, 2016.
Pixel contribution to the classification
Pixel contribution to the classification

Class Activation Mapping

$W_1 \ast + W_2 \ast + \ldots + W_n \ast = \text{Class Activation Map (Australian terrier)}$
Pooling schemes

- Max [Oquab, CVPR15]
  \[ y^c = \max_{i,j} z^c_{ij} \]

- GAP [Zhou, CVPR16]
  \[ y^c = \frac{1}{N} \sum_{i,j} z^c_{ij} \]

  \[ y^c = \frac{1}{\beta} \log \left( \frac{1}{N} \sum_{i,j} \exp(\beta \cdot z^c_{ij}) \right) \]
Max pooling limitation

- Classifying only with the max scoring region

- Loss of contextual information
Max pooling limitation

- Classifying only with the max scoring region
- Loss of contextual information
WELDON: max+min pooling

- $h^+$: presence of the class $\rightarrow$ high $h^+$
- $h^-$: localized evidence of the absence of class

original image  bedroom

airport inside  dining room
Region-based strategy

\[ h' \times w' \times 3 \quad h' \times w' \times 64 \]

\[ h = \frac{h'}{32} - 6 \quad w = \frac{w'}{32} - 6 \]

- convolution + ReLU
- max pooling
- convolution
- \( k\text{-max} + k\text{-min} \) pooling

bedroom

dining room
- Generalization to $K$ models per class
- Catch multiple class-related modalities
WILDCAT Architecture
Thibaut Durand, Taylor Mordan, Nicolas Thome, and Matthieu Cord
WILDCAT: Weakly Supervised Learning of Deep ConvNets for Image
Classification, Pointwise Localization and Segmentation.
In CVPR, 2017.
Class activation maps

bus
cat
horse
aeroplane
bottle
bicycle
Class activation maps

bicycle

bird

motorbike

person

sheep

bird
Class activation maps

cow  motorbike  horse

person  car  person
Visual recognition task: localization

<table>
<thead>
<tr>
<th>Method</th>
<th>VOC 2012</th>
<th>MS COCO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep MIL</td>
<td>74.5</td>
<td>41.2</td>
</tr>
<tr>
<td>ProNet</td>
<td>77.7</td>
<td>46.4</td>
</tr>
<tr>
<td>WSLocalization</td>
<td>79.7</td>
<td>49.2</td>
</tr>
</tbody>
</table>
In preview Segmentation

- WSL segmentation framework
  - Learning with image-level labels (presence/absence of the class)
  - Difficult task: no information about location and extend of objects
- Localized features in spatial maps
- Deep + fully connected CRFs
In preview Segmentation