Outline

1. Neural Nets
2. Deep Convolutional Neural Networks
3. Modern Deep Architectures
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

What is the next step for deeps?
- Are the ConvNets providing generic features?
  - transferring deep architectures to new tasks
- Have a look to the data first
From ImageNet to complex scenes?

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

VOC07/12  MIT67  15 Scene  COCO  VOC12 Action

- Select relevant regions $\rightarrow$ better prediction

- Full annotations expensive $\Rightarrow$ training with weak supervision
How to adapt VGG16 archi. for large/complex images?
Naïve approach: brut transfer (next Section)

- Resize the image
Sliding window $\Rightarrow$ 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

$h' \times w' \times 3$  \hspace{1cm} $h' \times w' \times 64$

$\frac{h'}{2} \times \frac{w'}{2} \times 128$

$\frac{h'}{4} \times \frac{w'}{4} \times 256$

$\frac{h'}{8} \times \frac{w'}{8} \times 512$

$\frac{h'}{16} \times \frac{w'}{16} \times 512$

$h = \frac{h'}{32} - 6$  \hspace{1cm} $w = \frac{w'}{32} - 6$

$h \times w \times 4096$

- Convolution + ReLU
- Max pooling
Transfer – Pooling – Classification
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**

  ![Diagram showing Deep MIL process](image)

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

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

  ![Diagram showing WELDON and ProNet process](image)

  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]
  \[ y^c = \max_{i,j} z_{ij}^c \]

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

  \[ y^c = \frac{1}{\beta} \log \left( \frac{1}{N} \sum_{i,j} \exp(\beta \cdot z_{ij}^c) \right) \]
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
Region-based strategy

- CAM for WELDON
- 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
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 extent of objects
- Localized features in spatial maps
- Deep + fully connected CRFs
In preview Segmentation

image | ground truth | heatmap1 | heatmap2 | prediction
Outline

1. Neural Nets
2. Deep Convolutional Neural Networks
3. Beyond ImageNet
   1. Fully Convolutional Networks (FCNs)
   2. Transfer
Transfer from ImageNet

**Transfer as generic features**
Brut Deep features (learned from ImageNet)

**Transfer learning**
Frozen features + SVM => solution to small datasets
Frozen features + Deep
Fine tuning not easy in that case (small datasets)
Transfer from ImageNet

Source: ImageNet => AlexNet trained

Target: Chopped AlexNet (layer i) + SVM trained and test on Cal-101 and Cal-256:

=> Results better than SoA CV methods on Cal-101!
Transfer: fine-tuning of a deep model

Train a deep (AlexNet) on source (ImageNet)
Keep the deep params. for target and complete with a small deep on top (fully trained on target)
Fine-tune the whole model on target data

Challenge: only limited target data, careful about overfitting
Solution: Freeze the gradient’s update for AlexNet part

![Diagram showing the transfer learning process](image)
Transfer: fine-tuning of a deep model

Train a deep (AlexNet) on source (ImageNet)
Keep the deep params. for target and complete with a small deep on top (fully trained on target)
Fine-tune the whole model on target data
  Challenge: only limited target data, careful about overfitting
  Solution: Freeze the gradient’s update for AlexNet part
Other solution: use smaller gradient’s update for AlexNet part
Layer Transfer

Which layer can be transferred (copied)?

- Speech: usually copy the last few layers
- Image: usually copy the first few layers
Transfer: fine-tuning of a deep model

• Task description
  • Source data: \((x^s, y^s)\)  
  A large amount
  • Target data: \((x^t, y^t)\)  
  Very little

Rq: One-shot learning: only a few examples in target domain

Many different contexts:

In vision: from ImageNet to small datasets

In speech: (supervised) speaker adaption
  • Source data: audio data and transcriptions from many speakers
  • Target data: audio data and its transcriptions of specific user
More on transfer framework

<table>
<thead>
<tr>
<th>Target Data</th>
<th>Source Data (ImageNet)</th>
</tr>
</thead>
<tbody>
<tr>
<td>labelled</td>
<td>labelled</td>
</tr>
<tr>
<td>Frozen or fine-tuning</td>
<td></td>
</tr>
<tr>
<td>unlabeled</td>
<td></td>
</tr>
</tbody>
</table>
General Framework for Transfer Learning

Dog/Cat Classifier

Data *not directly related to* the task considered

ImageNet: Similar domain, different tasks (1000 classes)
General Framework for Transfer Learning

Dog/Cat Classifier

Data *not directly related to* the task considered

Similar domain, completely different tasks

Different domains, same task
General Framework for Transfer Learning

<table>
<thead>
<tr>
<th>Target Data</th>
<th>Source Data (not directly related to the task)</th>
</tr>
</thead>
<tbody>
<tr>
<td>labelled</td>
<td>labelled</td>
</tr>
<tr>
<td></td>
<td>Fine-tuning</td>
</tr>
<tr>
<td></td>
<td>Multitask Learning</td>
</tr>
<tr>
<td>labelled</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Domain-adversarial training</td>
</tr>
<tr>
<td></td>
<td>Zero-shot learning</td>
</tr>
<tr>
<td>labelled</td>
<td>labeled</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Self-taught learning</td>
</tr>
<tr>
<td></td>
<td>Not considered here</td>
</tr>
<tr>
<td>labelled</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Self-taught Clustering</td>
</tr>
<tr>
<td></td>
<td>Not considered here</td>
</tr>
<tr>
<td>unlabeled</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>