Learning Representations II
(objects, scenes, videos, and mixtures)

CS331B: Representation Learning in Computer Vision
Amir R. Zamir
Silvio Savarese
(class logistics)

- Project proposal feedbacks
What we talked about so far...
I do not speak of what I have done. Such trivial thoughts crowd to my mind and now I am sorry to say that I have done the deed. I have murdered the King of Scotland. King Duncan.

I have destroyed his life. I have betrayed him in the most unacceptable way. I have proved my cowardice. I have brought my name into dishonour and to my shame. I have blood on my hands. I have done the deed.

As I stepped close to Duncan’s room, I thought that I would panic and freeze, but when I got nearer, a sickening thought made me feel like I was doing the right thing. As soon as Duncan died, I knew that it was the right thing to do. I found him standing as the dagger pierced through his skin.
Some basics concepts related to representations

- Ill-posedness
- Readout Non-linearity
- Dimensionality
- Computational Complexity
- Encoding power (i.e., performance)
- Narrowness of application domain (vertical vs horizontal representations)
Handcrafting Representations

Color Histograms

Deformable Part based Models (DPM)

Models based Shapes

Histogram of Gradients (HOG)

Felzenszwalb et al., 2010.
Dalal and Triggs, 2005.
Learning Representations

- **Supervised**
  - Representation constrained on task(s).

- **Unsupervised**
  - Representation constrained on reconstruction.

LeCun et al. 1998.
Hinton et al. 2006.
Unsupervised representation learning

- Sparse Coding
- Basic Autoencoders

New Image:

\[
\begin{align*}
\text{New Image:} & = 0.99 \times \text{Sparse representations} + 0.97 \times \text{Sparse representations} + 0.82 \times \text{Sparse representations} + \ldots
\end{align*}
\]
Supervised representation learning

Convolutional X

Neural Net

A Neuron
Supervised Low-level Matching

Zagoruyko & Komodakis. 2015.
Lectures 4&6

- 3D, activities, segmentations, layout, BoW, etc.
Learning Representations II

- objects
- scenes
- videos
- recurrent models
- methods of mixing representations
- GANs
Objects
Objects (ImageNet)

11*11*3 Convs

Query | Nearest Neighbors

Krizhevsky et al. 2012.
Deng et al. 2009.
Objects (ImageNet)

- How the end results looked like:

<table>
<thead>
<tr>
<th>Model</th>
<th>Top-1</th>
<th>Top-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparse coding [2]</td>
<td>47.1%</td>
<td>28.2%</td>
</tr>
<tr>
<td>SIFT + FVs [24]</td>
<td>45.7%</td>
<td>25.7%</td>
</tr>
<tr>
<td>CNN</td>
<td>37.5%</td>
<td>17.0%</td>
</tr>
</tbody>
</table>

Krizhevsky et al. 2012.
Deng et al. 2009.
Under the hood of object-based representation
Parts appear in object representation

- Top 9 activations in feature maps in validation data, projected down to pixels using a deconvolutional network.
- For each feature map the corresponding image patches are also shown.
Attributes appear in object representation

Escorcia et al. 2015.
Attributes appear in object representation

- How to form an attribute detector, given neuron activations per sample.
Attributes appear in object representation

LEAST and MOST affected images, by the ATTRIBUTE neurons.

Layer distribution of attribute neuron.
ImageNet features (empirically) transfer!

- (Details and discussions in lecture 10)
Scenes
Scenes (MIT Places) - 205 categories, 2.5m images.

Figure 1: Image samples from the scene categories grouped by their queried adjectives.

Figure 2: Comparison of the number of images per scene category in three databases.

Zhou et al. 2014.
Minimal Image Representation (simplified image)

- Minimal Image Representation ≡ what matters in the image.
Receptive Fields

- Receptive Field ≡ what matters to a neuron/representation unit.
Receptive Fields

- Theoretical vs empirical (actual size) RFs
- Theoretical RF ≥ empirical RF

<table>
<thead>
<tr>
<th></th>
<th>pool1</th>
<th>pool2</th>
<th>conv3</th>
<th>conv4</th>
<th>pool5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theoretic size</td>
<td>19</td>
<td>67</td>
<td>99</td>
<td>131</td>
<td>195</td>
</tr>
<tr>
<td>Places-CNN actual size</td>
<td>17.8±1.6</td>
<td>37.4±5.9</td>
<td>52.1±10.6</td>
<td>60.0±13.7</td>
<td>72.0±20.0</td>
</tr>
<tr>
<td>ImageNet-CNN actual size</td>
<td>17.9±1.6</td>
<td>36.7±5.4</td>
<td>51.1±9.9</td>
<td>60.4±16.0</td>
<td>70.3±21.6</td>
</tr>
</tbody>
</table>

Semantics of Units/Neurons

- How to assign semantics to a neuron:
Semantics of Units/Neurons

- Assigned semantics and their precision
- Objects appear in Scene representation
Semantics of Units/Neurons

- **Objects** appear in **Scene** representation
- counts of neurons discovering objects:

In Places-CNN

In ImageNet-CNN

Demo

Connections and activations in Places

- Each unit shows one convolutional neuron.
- Wires show strongest connections for the neuron.
- For the selected neuron, four images that most strongly activate each the neuron are shown.

Zhou et al. 2014. 2015.
Videos
(\textit{~static}) video processing - two stream CNNs

Simonyan & Zisserman. 2014.
### (~static) video processing - two stream CNNs

<table>
<thead>
<tr>
<th>Method</th>
<th>UCF-101</th>
<th>HMDB-51</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improved dense trajectories (IDT) [26, 27]</td>
<td>85.9%</td>
<td>57.2%</td>
</tr>
<tr>
<td>IDT with higher-dimensional encodings [20]</td>
<td>87.9%</td>
<td>61.1%</td>
</tr>
<tr>
<td>IDT with stacked Fisher encoding [21] (based on Deep Fisher Net [23])</td>
<td>-</td>
<td>66.8%</td>
</tr>
<tr>
<td>Spatio-temporal HMAX network [11, 16]</td>
<td>-</td>
<td>22.8%</td>
</tr>
<tr>
<td>“Slow fusion” spatio-temporal ConvNet [14]</td>
<td>65.4%</td>
<td>-</td>
</tr>
<tr>
<td>Spatial stream ConvNet</td>
<td>73.0%</td>
<td>40.5%</td>
</tr>
<tr>
<td>Temporal stream ConvNet</td>
<td>83.7%</td>
<td>54.6%</td>
</tr>
<tr>
<td>Two-stream model (fusion by averaging)</td>
<td>86.9%</td>
<td>58.0%</td>
</tr>
<tr>
<td>Two-stream model (fusion by SVM)</td>
<td><strong>88.0%</strong></td>
<td><strong>59.4%</strong></td>
</tr>
</tbody>
</table>

Simonyan & Zisserman. 2014.
(~static) video processing - frame fusion

Karpathy et al. 2015.
(~static) video processing - frame fusion

Karpathy et al. 2015.
(~static) video processing - frame fusion

<table>
<thead>
<tr>
<th>Model</th>
<th>Clip Hit@1</th>
<th>Video Hit@1</th>
<th>Video Hit@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature Histograms + Neural Net</td>
<td>-</td>
<td>55.3</td>
<td>-</td>
</tr>
<tr>
<td>Single-Frame</td>
<td>41.1</td>
<td>59.3</td>
<td>77.7</td>
</tr>
<tr>
<td>Single-Frame + Multires</td>
<td>42.4</td>
<td>60.0</td>
<td>78.5</td>
</tr>
<tr>
<td>Single-Frame Fovea Only</td>
<td>30.0</td>
<td>49.9</td>
<td>72.8</td>
</tr>
<tr>
<td>Single-Frame Context Only</td>
<td>38.1</td>
<td>56.0</td>
<td>77.2</td>
</tr>
<tr>
<td>Early Fusion</td>
<td>38.9</td>
<td>57.7</td>
<td>76.8</td>
</tr>
<tr>
<td>Late Fusion</td>
<td>40.7</td>
<td>59.3</td>
<td>78.7</td>
</tr>
<tr>
<td>Slow Fusion</td>
<td>41.9</td>
<td>60.9</td>
<td>80.2</td>
</tr>
<tr>
<td>CNN Average (Single+Early+Late+Slow)</td>
<td>41.4</td>
<td>63.9</td>
<td>82.4</td>
</tr>
</tbody>
</table>

Karpathy et al. 2015.
Recurrent Modeling
Recurrent models
Recurrent models
Structure Prediction + Recurrent Models (Structural-RNN)
Structural-RNN
In CVPR 16 (Best Student Paper Award)

Click here to see the slides

Ashesh Jain
Amir R. Zamir
Silvio Savarese
Ashutosh Saxena
Methods of Mixing Representations
Mixing Representations

- Sometimes you wish to mix two/multiple tasks/representations
  - To transfer information or labeled data across tasks
  - To form a multi-task representation
  - To form a better single-task representation
Mixing Representations - LwF

(a) Original Model

(b) Fine-tuning

(c) Feature Extraction

(d) Joint Training

(e) Learning without Forgetting

ECCV’16
Mixing Representations

- Pros and cons of each method:

<table>
<thead>
<tr>
<th></th>
<th>Fine Tuning</th>
<th>Duplicating and Fine Tuning</th>
<th>Feature Extraction</th>
<th>Joint Training</th>
</tr>
</thead>
<tbody>
<tr>
<td>new task performance</td>
<td>good</td>
<td>good</td>
<td>× medium</td>
<td>best</td>
</tr>
<tr>
<td>original task performance</td>
<td>× bad</td>
<td>good</td>
<td>good</td>
<td>good</td>
</tr>
<tr>
<td>training efficiency</td>
<td>fast</td>
<td>fast</td>
<td>fast</td>
<td>× slow</td>
</tr>
<tr>
<td>testing efficiency</td>
<td>fast</td>
<td>× slow</td>
<td>fast</td>
<td>fast</td>
</tr>
<tr>
<td>storage requirement</td>
<td>medium</td>
<td>× large</td>
<td>medium</td>
<td>× large</td>
</tr>
<tr>
<td>requires previous task data</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>× yes</td>
</tr>
</tbody>
</table>
To be continued...
Course webpage:
http://web.stanford.edu/class/cs331b/

http://www.cs.stanford.edu/~amirz/
http://cvgl.stanford.edu/silvio/