Learning Representations

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

- Paper assignments out
- Next weeks papers
  - Discriminative learning of deep convolutional feature point descriptors, Simo-Serra, E., Trulls, E., Ferraz, L., Kokkinos, I., Fua, P., & Moreno-Noguer, F., ICCV15
  - Data-Driven 3D Voxel Patterns for Object Category Recognition, Yu Xiang, Wongun Choi, Yuanqing Lin and Silvio Savarese., CVPR15.
  - Convolutional-recursive deep learning for 3d object classification, Socher, R., Huval, B., Bath, B., Manning, C. D., & Ng, A. Y., NIPS12
What we talked about so far...
My heart beats as if the world is dropping, you may not feel the love but i do its a heart breaking moment of your life. Enjoy the times that we have, it might not sound good but one thing it rhymes it might not be romantic but i think it is great, the best rhyme i’ve ever heard.
Macbeth was guilty.

I dared not speak of what I have done. Such treacherous thoughts overset my mind, and now I am sorry to say that I have done the deed.

I have murdered the King of Scotland, King Duncan.

He was my benefactor; this is how I repay him. I have betrayed him in the most unhonorable way possible, and I am afraid that I shall slip away now.

I was writing an essay for my Lady in search the bell that called me into the cell. But before she did, a cluster of the supernatural things that have happened in my life came to my mind. I heard the bell, and, I realized, I heard the bell in my mind.

But when she never came to me, I knew that I was doing the right thing.

As soon as Duncan fell, I knew that it was the bell summoning me.

I heard him pleading as the dagger pierced through his skin.

I dared not speak of what I have done. Such treacherous thoughts overset my mind, and now I am sorry to say that I have done the deed. I have murdered the King of Scotland, King Duncan.

He was my benefactor; this is how I repay him. I have betrayed him in the most unhonorable way possible, and I am afraid that I shall slip away now.

I was writing an essay for my Lady in search the bell that called me into the cell. But before she did, a cluster of the supernatural things that have happened in my life came to my mind. I heard the bell, and, I realized, I heard the bell in my mind.

But when she never came to me, I knew that I was doing the right thing.

As soon as Duncan fell, I knew that it was the bell summoning me.

I heard him pleading 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)
Ill-posedness

Linearity

With respect to: \{modeling parameters (decision), independent variables (representation)\}

Decision boundary

\[ ax + b - y = 0 \]

\[ ax^2 + bx + c - y = 0 \]

Not discussing kernels, reparametrization, etc.
Handcrafting...
Represent these cats for a cat detector!
Represent these cats for a cat detector! (II)
Represent these cats for a cat detector! (III)
Represent these cats for a cat detector! (IV)
Color Histograms

Deformable Part based Models (DPM)

Models based Shapes

Histogram of Gradients (HOG)

Felzenszwalb et al., 2010.
Dalal and Triggs, 2005.
Handcrafting...

- Was the only way for a long time.
- (almost) Worked for many great applications:
  - Image Retrieval, Structure-from-motion, Face detection, Identification, etc
- Why alternatives?
Handcrafting...

● Was the only way for a long time.
● (almost) Worked for many great applications:
  ○ Image Retrieval, Structure-from-motion, Face detection, Identification, etc
● Why alternatives?
  ○ Can’t quite find the discriminative signature for a problem.

Why a sad image?
Handcrafting...

- Was the only way for a long time.
- (almost) Worked for many great applications:
  - Image Retrieval, Structure-from-motion, Face detection, Identification, etc
- Why alternatives?
  - Can’t quite find the discriminative signature for a problem.
  - Discriminative signature can be found, but hard to approach programmatically.

Track the fish.
Handcrafting...

- Was the only way for a long time.
- (almost) Worked for many great applications:
  - Image Retrieval, Structure-from-motion, Face detection, Identification, etc
- Why alternatives?
  - Can’t quite find the discriminative signature for a problem.
  - Discriminative signature can be found, but hard to approach programmatically.
  - Too many contributing factors to the problem.
    - Fusion non-trivial. Rule-based fusion outruled.
    - Fusion of contributing factors itself a comparably complex representation problem.

Why a sad image?
Learning Representations
Two approaches to representation learning

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

- **Unsupervised**
  - Representation constrained on reconstruction.
Two approaches to **representation learning**

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

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

LeCun et al. 1998.
Hinton et al. 2006.
### Pros and Cons

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

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

---

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

- **Clarification:**
  - *Unsupervised* does not necessarily mean *Generic* and vice versa.
  - Many of the generic representations are actually trained discriminatively.
  - Lecture 12

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

Hinton et al. 2006.
Unsupervised representation learning

- Usually a reconstruction loss:
  - L2 pixel loss

Hinton et al. 2006.
Unsupervised representation learning

- Usually a reconstruction loss:
  - L2 pixel loss
  - Perceptual loss

Johnson et al. 2016.
Hinton et al. 2006.
Unsupervised representation learning methods

- Sparse Coding
- Basic Autoencoders
- Restricted Boltzmann Machine
- Generative Adversarial Network based feature learning
- ...

27
Sparse Coding

Observed Data
Subset of 25,000 characters

Subset of 1000 features
Sparse representations

New Image:

\[
\begin{align*}
\text{New Image:} & = 0.99 \times \text{New Image 1} + 0.97 \times \text{New Image 2} + 0.82 \times \text{New Image 3} \ldots
\end{align*}
\]
Sparse Coding

4 million *unlabelled images*

Learned features (out of 10,000)

\[
\text{New Image} = 0.9 \times \text{feature} + 0.8 \times \text{feature} + 0.6 \times \text{feature} \ldots
\]

Lee et al. 2006.
Salakhutdinov. 2016.
Autoencoder
Autoencoder

Hinton et al. 2006.
Autoencoder

Hinton et al. 2006.
Autoencoder: reconstruction results

Hinton et al. 2006.
Supervised Models
Neural Networks

Convolutional X

Neural Net

A Neuron

Stanford CS231n
Low-level matching

- SIFT/HOG/handcrafted local features counterpart
Brown, Hua Winder, local feature learning dataset
Low-level matching

- Learned vs l2 metric
Low-level matching: Some qualitative results

(a) true positives

(b) false negatives

(c) true negatives

(d) false positives

Zagoruyko & Komodakis. 2015.
Low-level matching: quantitative evaluation

- 1st layer filters
- Notice similarity to Gabor and sparse coding filters

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
<th>2ch-2stream</th>
<th>2ch-deep</th>
<th>2ch</th>
<th>siam</th>
<th>siam-$l_2$</th>
<th>pseudo-siam</th>
<th>pseudo-siam-$l_2$</th>
<th>siam-2stream</th>
<th>siam-2stream-$l_2$</th>
<th>[19]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yos</td>
<td>ND</td>
<td><strong>2.11</strong></td>
<td>2.52</td>
<td>3.05</td>
<td>5.75</td>
<td>8.38</td>
<td>5.44</td>
<td>8.95</td>
<td>5.29</td>
<td>5.58</td>
<td>6.82</td>
</tr>
<tr>
<td>Yos</td>
<td>Lib</td>
<td><strong>7.2</strong></td>
<td>7.4</td>
<td>8.59</td>
<td>13.48</td>
<td>17.25</td>
<td>10.35</td>
<td>18.37</td>
<td>11.51</td>
<td>12.84</td>
<td>14.58</td>
</tr>
<tr>
<td>ND</td>
<td>Yos</td>
<td><strong>4.1</strong></td>
<td>4.38</td>
<td>6.04</td>
<td>13.23</td>
<td>15.89</td>
<td>12.64</td>
<td>15.62</td>
<td>10.44</td>
<td>13.02</td>
<td>10.08</td>
</tr>
<tr>
<td>ND</td>
<td>Lib</td>
<td><strong>4.55</strong></td>
<td>6.05</td>
<td>8.77</td>
<td>13.24</td>
<td>12.87</td>
<td>16.58</td>
<td>6.45</td>
<td>8.79</td>
<td>12.42</td>
<td></td>
</tr>
<tr>
<td>Lib</td>
<td>Yos</td>
<td><strong>5</strong></td>
<td>7</td>
<td>14.89</td>
<td>19.91</td>
<td>12.5</td>
<td>17.83</td>
<td>9.02</td>
<td>13.24</td>
<td>11.18</td>
<td></td>
</tr>
<tr>
<td>Lib</td>
<td>ND</td>
<td><strong>1.9</strong></td>
<td>2.01</td>
<td>3.03</td>
<td>4.33</td>
<td>6.01</td>
<td>3.93</td>
<td>6.58</td>
<td>3.05</td>
<td>4.54</td>
<td>7.22</td>
</tr>
<tr>
<td>mean(1,4)</td>
<td></td>
<td><strong>4.56</strong></td>
<td>4.71</td>
<td>5.93</td>
<td>10.31</td>
<td>13.69</td>
<td>10.33</td>
<td>14.88</td>
<td>8.42</td>
<td>10.06</td>
<td>10.98</td>
</tr>
</tbody>
</table>

Zagoruyko & Komodakis. 2015.
## Dimensionality

<table>
<thead>
<tr>
<th>Training</th>
<th>Feature Dim.</th>
<th>Notre Dame</th>
<th>Yosemite</th>
<th>Liberty</th>
<th>Notre Dame</th>
<th>Yosemite</th>
<th>Liberty</th>
<th>Notre Dame</th>
<th>Yosemite</th>
</tr>
</thead>
<tbody>
<tr>
<td>nSIFT + L2 (no training)</td>
<td>128d</td>
<td>29.84</td>
<td>22.53</td>
<td>27.29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nSIFT squared diff. + linearSVM</td>
<td>128d</td>
<td>26.54</td>
<td>27.07</td>
<td>19.65</td>
<td>19.87</td>
<td>25.12</td>
<td>24.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nSIFT concat. + NNet (F=512)</td>
<td>256d</td>
<td>20.44</td>
<td>22.23</td>
<td>14.35</td>
<td>14.84</td>
<td>21.41</td>
<td>20.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simonyan et al (2014) discrim. proj.</td>
<td>&lt;64d</td>
<td>12.88</td>
<td>14.82</td>
<td>7.52</td>
<td>7.11</td>
<td>11.63</td>
<td>10.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MatchNet (F=1024, B=64)</td>
<td>64d</td>
<td>9.82</td>
<td>14.27</td>
<td>5.02</td>
<td>9.15</td>
<td>14.15</td>
<td>13.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MatchNet (F=512, B=128)</td>
<td>128d</td>
<td>9.48</td>
<td>15.40</td>
<td>5.18</td>
<td>8.27</td>
<td>14.40</td>
<td>12.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MatchNet (F=512, B=512)</td>
<td>512d</td>
<td>8.84</td>
<td>13.02</td>
<td>4.75</td>
<td>7.70</td>
<td>13.58</td>
<td>11.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MatchNet (F=512, w/o bottleneck)</td>
<td>4096d</td>
<td><strong>6.90</strong></td>
<td><strong>10.77</strong></td>
<td><strong>3.87</strong></td>
<td><strong>5.67</strong></td>
<td><strong>10.88</strong></td>
<td><strong>8.39</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Low-level matching: handcrafted vs learned features

Evaluation of matching

Wide Baseline Handling (matching)

Zamir et al. 2016.
Course webpage:
http://web.stanford.edu/class/cs331b/

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