Beyond Spatial Pyramids
Receptive Field Learning for Pooled Image Features

Yangqing Jia\textsuperscript{1} \quad Chang Huang\textsuperscript{2} \quad Trevor Darrell\textsuperscript{1}
\textsuperscript{1}UC Berkeley EECS \quad \textsuperscript{2}NEC Labs America
Goal

- Analysis of the pooling step in the image classification pipeline
- Evidence that spatial pyramids may be suboptimal
- A new method that learns receptive fields tailored to the classification tasks
Dense-coded Classification Pipeline

- dense feature extraction
- coding: encoding the image to K codebook activation maps
Codebook training / coding methods

- Not necessarily simple convolutions!
- Different types of dense features
  - SIFT (e.g. Caltech 101)
  - Raw / whitened pixel values (e.g. CIFAR)
- Sophistication in codebook learning and encoding
  - Vector quantization
  - Sparse coding (Olshausen et al. 1996)
  - LCC/LLC (Yu & Zhang, 2009; Wang et al. 2010)
Dense-coded Classification Pipeline

- Pooling: Compute statistics of the activations in specific spatial areas (receptive fields)
Dense-coded Classification Pipeline

- Classification: Adopting linear classifiers to predict the label

\[ f(x) = \text{"Bear"} \]
Existing Work on Pooling

- **Bag of Words**
- **Spatial Pyramids**
  - Lazebnik et al. 2006 (SPM), Yang et al. 2009 (ScSPM)
- **Better Pooling Operators**
  - Boureau et al. 2010
- **Grouping activation maps**
  - Boureau et al. 2011, Coates et al. 2011
- Relatively few work on the spatial receptive field designs!
Pooling is Task-Dependent
Pooling is Task-Dependent
Pooling is Task-Dependent

Solution: use overcomplete receptive fields!
Related Ideas

- **Boosted receptive fields**
  - Viola & Jones 2001 (Haar wavelets)
  - Shakhnarovich et al. 2003 (Region histograms)

- **Learning local descriptors**
  - Tola et al. 2008, Brown et al. 2010

- **Recent subcategory recognition works**
  - Zhang et al. 2012 (Pose pooling kernels)
  - Yao et al. 2012 (Fine-grained categorization)
Define a Pooled Feature

- Given $P$ receptive fields and $K$ coded activations

  \[ \mathcal{R} = \{ \mathbf{R}_1, \mathbf{R}_2, \cdots, \mathbf{R}_P \} \]

  \[ \mathcal{D} = \{ \mathbf{A}^1, \mathbf{A}^2, \cdots, \mathbf{A}^K \} \]

- $P \times K$ possible pooled features

  \[ x_{K \times p+k} = \text{op}(\mathbf{A}^k_{\mathbf{R}_p}) \]
Challenges & Solutions

Challenges

- A huge number of possible receptive fields
  - $2^{\#\text{pixels}}$ possible RFs
- Need to maintain reasonable prediction speed

Solutions

- Use reasonably over-complete RF candidates
- Select useful features via sparsity constraints
We propose to use rectangular bins built on small regular grids.

- Regular grids $(k \times k)$
- Spatial pyramid $(O(k^2)$ bins)
- Overcomplete RFs $(O(k^4)$ bins)
Structured Sparsity

- Find classifiers that use a subset of the features

\[
\min_{\mathbf{w}, \mathbf{b}} \frac{1}{N} \sum_{n=1}^{N} l(\mathbf{w}^\top \mathbf{x}_n + \mathbf{b}, \mathbf{y}_n) + \lambda_1 \|\mathbf{W}\|_2^2 + \lambda_2 \|\mathbf{W}\|_{1,\infty}
\]

- Classification Loss
- L2 regularization
- Structured Sparsity

(Feature computation: \( \mathbf{x}_{n,K \times p+k} = \text{op}(A_{n,R_p}^k) \))
Greedy Approximation to Structured Sparsity

- Incrementally select the feature with the largest gradient (Perkins et al. 2003)

\[
\text{score}(i) = \left\| \frac{\partial \mathcal{L}(W, b)}{\partial W_i} \right\|_F^2
\]

- Re-train classifier (fast!)

![Performance vs. Number of Features](image-url)
Experiment: CIFAR

- The CIFAR-10 dataset
  - 10 object classes
  - 50k training, 10k testing
- Coding strategy follows (Coates & Ng, 2011)

Does Spatial Pyramid suffice?

[kmeans (k=200) + triangular coding on 6x6 patches, CIFAR-10]
More codes vs. Smarter Pooling

Performance on CIFAR-10

Baseline
Ours, equal-dim
Ours, optimum-dim
More codes vs. Smarter Pooling

Performance on CIFAR-10

Higher accuracy with small codebook
More codes vs. Smarter Pooling

Performance on CIFAR-10

Consistent improvement when codebook size grows

Higher accuracy with small codebook
## Best Practice on CIFAR

<table>
<thead>
<tr>
<th>Method</th>
<th>Pooled Dim</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>ours, d=1600</td>
<td>6,400</td>
<td>80.17</td>
</tr>
<tr>
<td>ours, d=4000</td>
<td>16,000</td>
<td>82.04</td>
</tr>
<tr>
<td>ours, d=6000</td>
<td>24,000</td>
<td>83.11</td>
</tr>
<tr>
<td>Coates 2010 d=1600</td>
<td>6,400</td>
<td>77.9</td>
</tr>
<tr>
<td>Coates 2010 d=4000</td>
<td>16,000</td>
<td>79.6</td>
</tr>
<tr>
<td>Coates 2011 d=6000</td>
<td>48,000</td>
<td>81.5</td>
</tr>
<tr>
<td>Krizhevsky TR’10</td>
<td>N/A</td>
<td>78.9</td>
</tr>
<tr>
<td>Yu ICML’10</td>
<td>N/A</td>
<td>74.5</td>
</tr>
<tr>
<td>Ciresan Arxiv’11</td>
<td>N/A</td>
<td>80.49</td>
</tr>
<tr>
<td>Coates NIPS’11</td>
<td>N/A</td>
<td>82.0</td>
</tr>
</tbody>
</table>
Most useful Receptive Fields

- Cross-image bars
  - natural scene layout
- Whole-image pooling
  - hollistic statistics
- Small fields
  - local distinctiveness
- Corners
  - context matters
### Experiment: Caltech-101

- Better pooling increases performance over SPM (up to the implementation limit of the algorithm)

<table>
<thead>
<tr>
<th>Method</th>
<th>Codebook</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>ScSPM (Yang et al. 2009)</td>
<td>1024 (SC)</td>
<td>73.2±0.54</td>
</tr>
<tr>
<td>LCC+SPM (Wang et al. 2010)</td>
<td>1024 (LCC)</td>
<td>73.44</td>
</tr>
<tr>
<td>Our Method</td>
<td>1024 (SC)</td>
<td><strong>75.3±0.70</strong></td>
</tr>
</tbody>
</table>
Conclusion

- We proposed a new method that learns receptive fields tailored to the classification tasks.
- Showed consistent improvement over SPM on medium-sized codebooks.
- Future work
  - larger-scale feature learning with both overcomplete coding and overcomplete pooling
  - joint task-driven coding and pooling
Conclusion (cont’d)

\[ f(x) = “Bear” \]
Conclusion (cont’d)

- Agnostic to coding

- Multiple objectives
  - Better local descriptors?
  - Object-level HOG?
  - ...

- $f(x) = \text{"Bear"}$
Thank you!

Beyond Spatial Pyramids: Receptive Fields Learning for Pooled Image Features

Yangqing Jia, Chang Huang, Trevor Darrell
This page is intentionally left blank
## Experiment: MNIST

<table>
<thead>
<tr>
<th>Method</th>
<th>err%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Method</td>
<td>0.64</td>
</tr>
<tr>
<td>Coates ICML’11</td>
<td>1.02</td>
</tr>
<tr>
<td>Lauer PR’07</td>
<td>0.83</td>
</tr>
<tr>
<td>Labusch TNN’08</td>
<td>0.59</td>
</tr>
<tr>
<td>Ranzato CVPR’09</td>
<td>0.62</td>
</tr>
<tr>
<td>Jarrett ICCV’09</td>
<td>0.53</td>
</tr>
</tbody>
</table>
Thus spoke neuroscience

LGN (Whitening)

Complex Cells in V1 (spatial pooling)

Simple Cells in V1 (sparse coding ?)