Beyond Spatial Pyramids

Receptive Field Learning for Pooled Image Features

Yangqing Jia[1] Chang Huang[2] Trevor Darrell[1]
[1]UC Berkeley EECS [2] NEC Labs America

NEC Laboratories America
Relentless passion for innovation

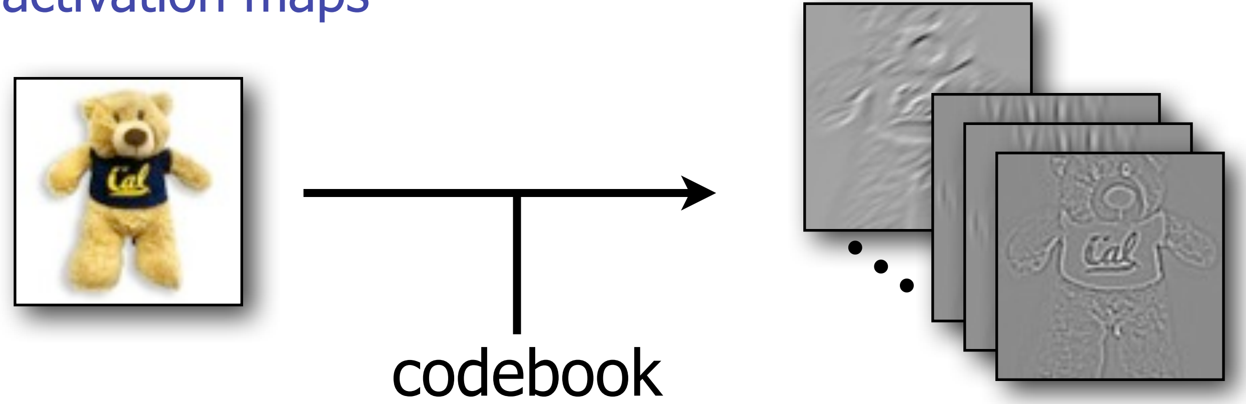
Goal

 → coding → pooling → "Bear"

- 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

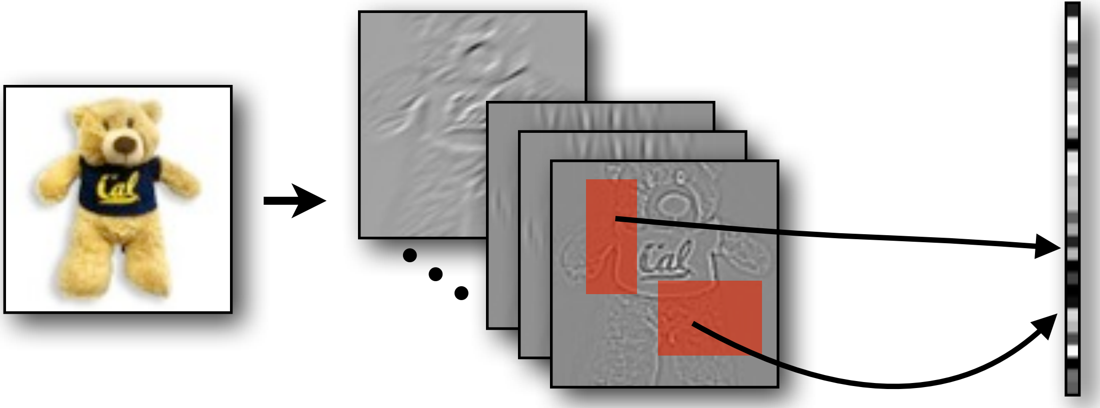


Codebook training / coding methods

- 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)

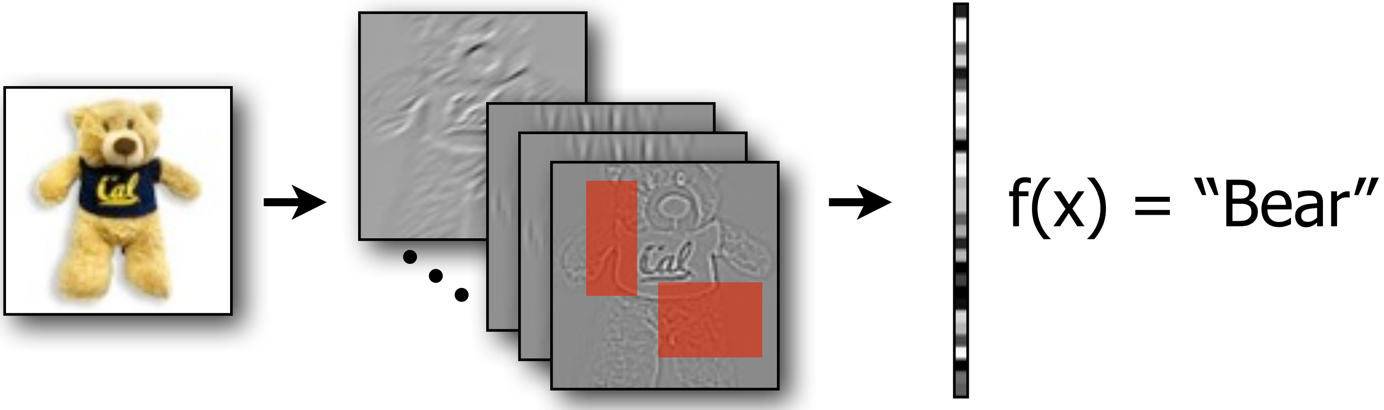
Dense-coded Classification Pipeline

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



Dense-coded Classification Pipeline

- Classification: Adopting linear classifiers to predict the label

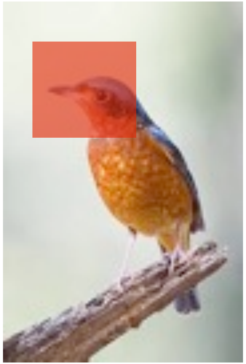


Existing Work on Pooling

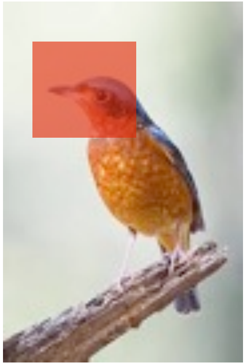
- Bag of Words
- Spatial Pyramids
  - Lazebnik et al. 2006 (SPM), Yang et al. 2009 (ScSPM)
- Better Pooling Operators
  - Boureau et al. 2010
- Grouping activation maps
  - Boureau et al. 2011, Coates et al. 2011
- Relatively few work on the spatial receptive field designs!

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

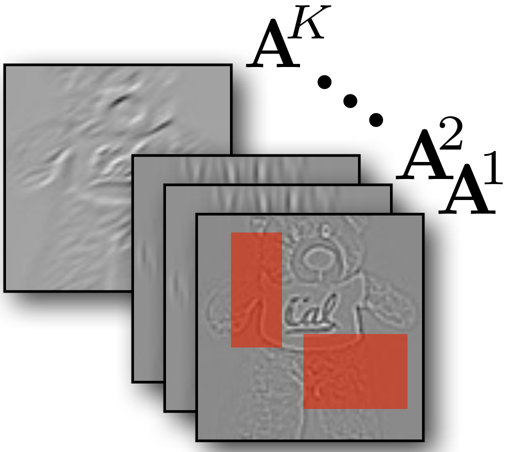
- Given P receptive fields and K coded activations

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

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

- P x K possible pooled features

$$x_{K \times p + k} = op(\mathbf{A}^k_{\mathbf{R_p}})$$



Challenges & Solutions

Challenges

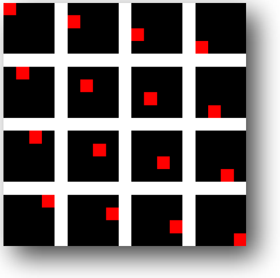
- A huge number of possible receptive fields
  - $2^{\#pixels}$ possible RFs
- Need to maintain reasonable prediction speed

Solutions

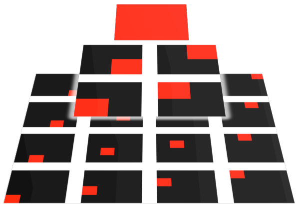
- Use reasonably over-complete RF candidates
- Select useful features via sparsity constraints

Overcomplete Receptive Fields

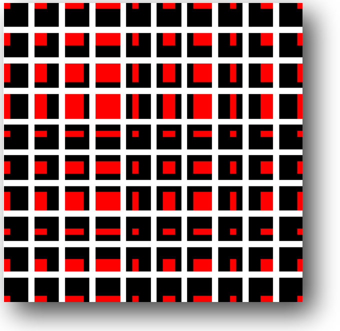
- We propose to use rectangular bins built on small regular grids



Regular grids
(k x k)



Spatial pyramid
($O(k^2)$ bins)



Overcomplete RFs
($O(k^4)$ bins)

Structured Sparsity

- Find classifiers that use a subset of the features

$$\min_{\mathbf{W},\mathbf{b}} \quad \underbrace{\frac{1}{N}\sum_{n=1}^{N} l(\mathbf{W}^\top \mathbf{x}_n + \mathbf{b}, \mathbf{y}_n)}_{\text{Classification Loss}} + \underbrace{\lambda_1 \|\mathbf{W}\|_{\mathrm{Fro}}^2}_{\text{L2 regularization}} + \underbrace{\lambda_2 \|\mathbf{W}\|_{1,\infty}}_{\text{Structured Sparsity}}$$

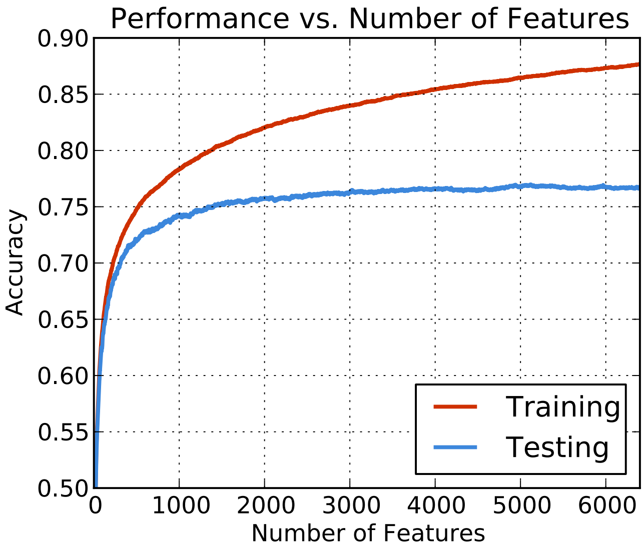
(Feature computation: $x_{n,K\times p+k} = \mathrm{op}(\mathbf{A}^k_{n,\mathbf{R}_p})$)

Greedy Approximation to Structured Sparsity

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

$$\mathrm{score}(i) = \left\| \frac{\partial \mathcal{L}(\mathbf{W}, \mathbf{b})}{\partial \mathbf{W}_{i,\cdot}} \right\|_{\mathrm{Fro}}^2$$

- Re-train classifier (fast!)

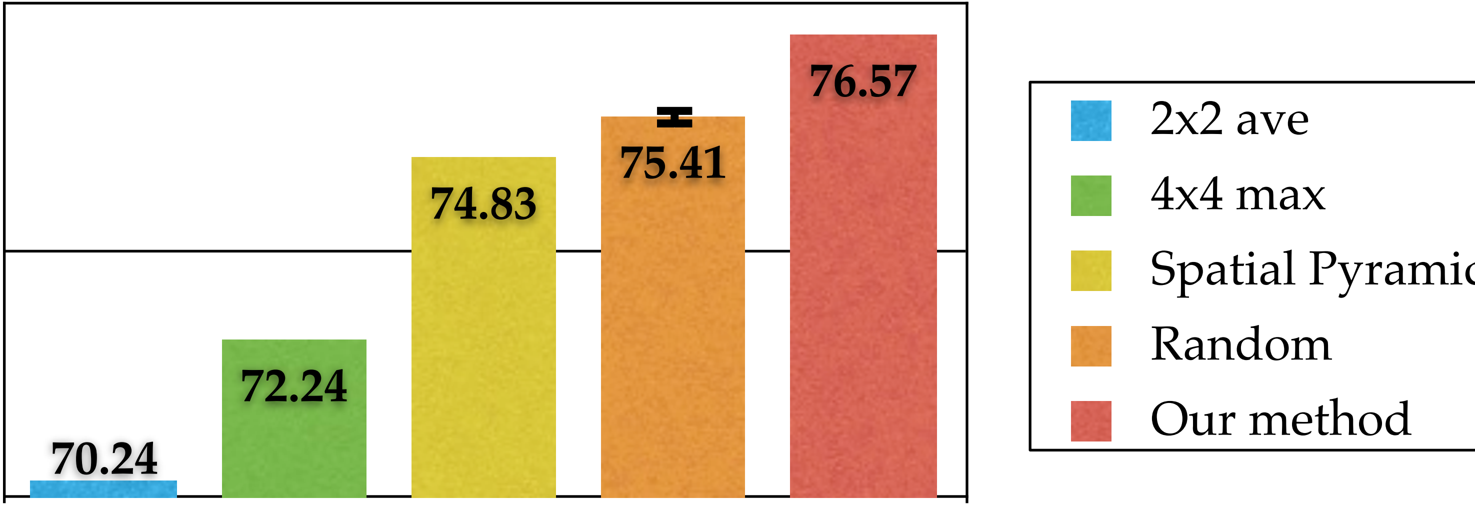


Experiment: CIFAR

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

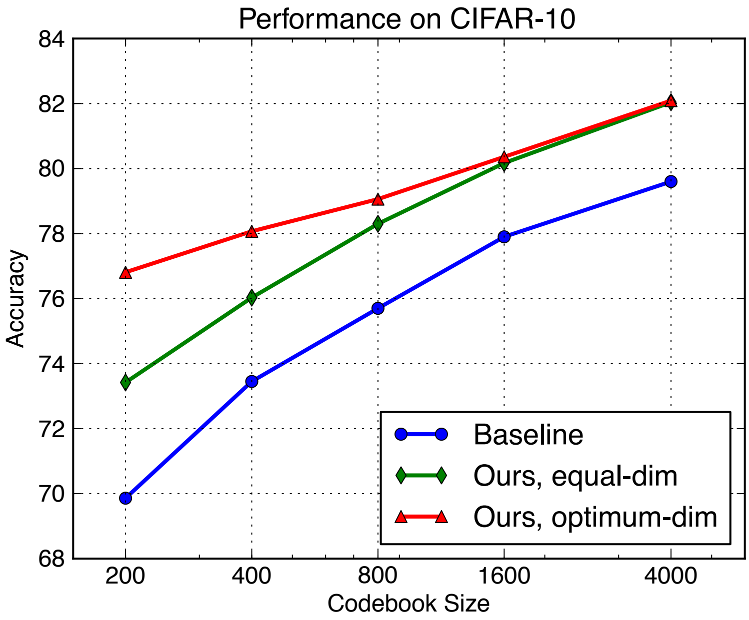
(Image courtesy of Alex Krizhevsky, http://www.cs.toronto.edu/~kriz/cifar.html)

Does Spatial Pyramid suffice?



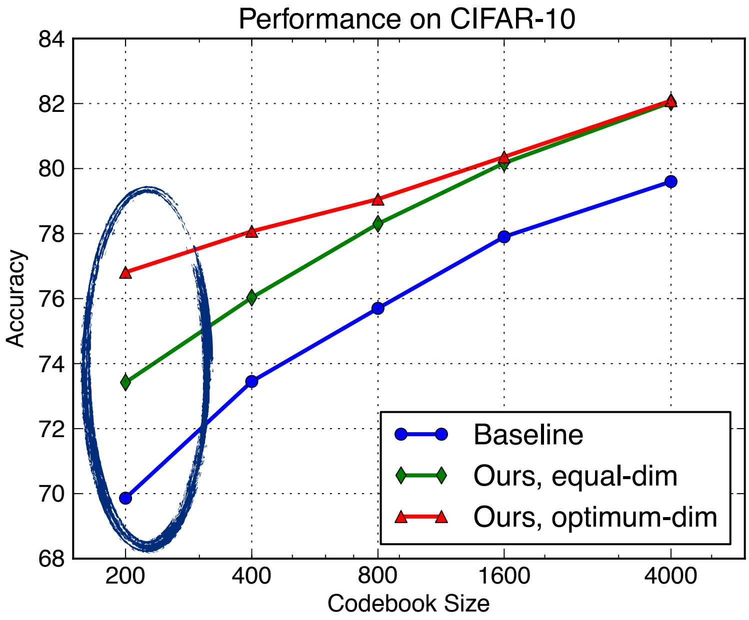
[kmeans (k=200) + triangular coding on 6x6 patches, CIFAR-10]

More codes vs. Smarter Pooling

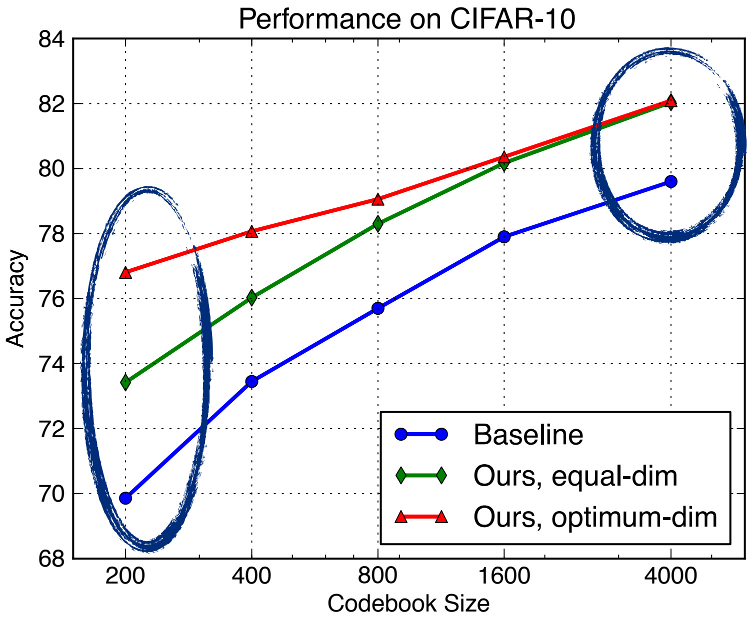


More codes vs. Smarter Pooling

Higher accuracy with small codebook



More codes vs. Smarter Pooling



Higher accuracy with small codebook

Consistent improvement when codebook size grows

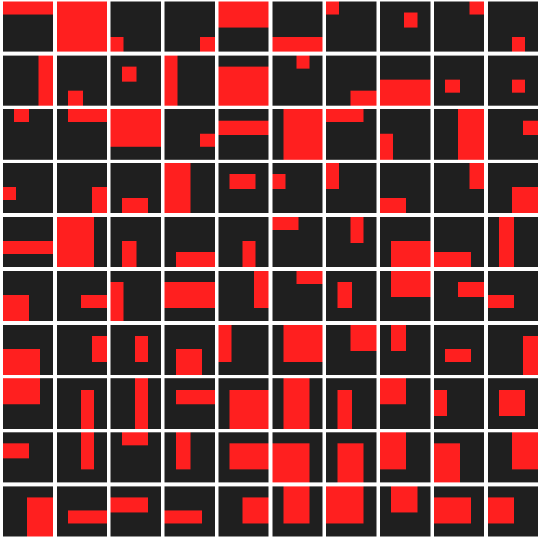
Best Practice on CIFAR

Method	Pooled Dim	Accuracy
ours, d=1600	6,400	80.17
ours, d=4000	16,000	82.04
ours, d=6000	24,000	**83.11**
Coates 2010 d=1600	6,400	77.9
Coates 2010 d=4000	16,000	79.6
Coates 2011 d=6000	48,000	81.5
Krizhevsky TR'10	N/A	78.9
Yu ICML'10	N/A	74.5
Ciresan Arxiv'11	N/A	80.49
Coates NIPS'11	N/A	82.0

Most useful Receptive Fields

- Cross-image bars
  - natural scene layout
- Whole-image pooling
  - hollistic statistics
- Small fields
  - local distinctiveness
- Corners
  - context matters

most useful ←



→ least useful

Experiment: Caltech-101

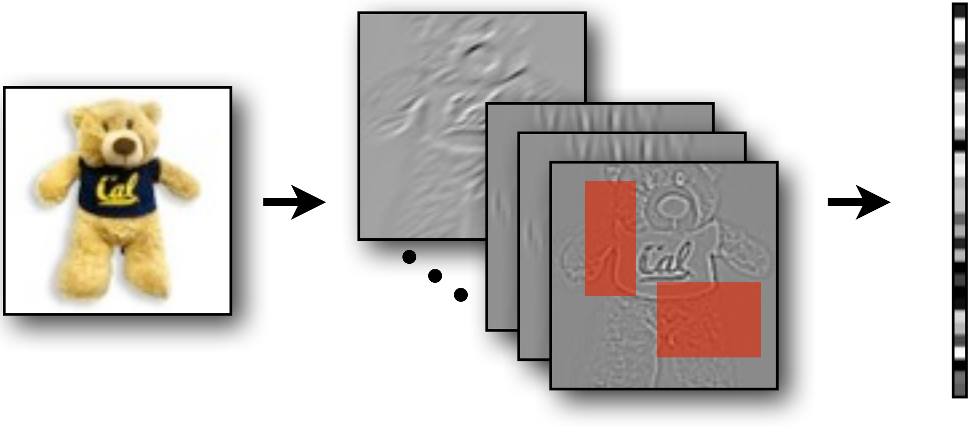
- Better pooling increases performance over SPM (up to the implementation limit of the algorithm)

Method	Codebook	Performance
ScSPM (Yang et al. 2009)	1024 (SC)	73.2±0.54
LCC+SPM (Wang et al. 2010)	1024 (LCC)	73.44
Our Method	1024 (SC)	**75.3±0.70**

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

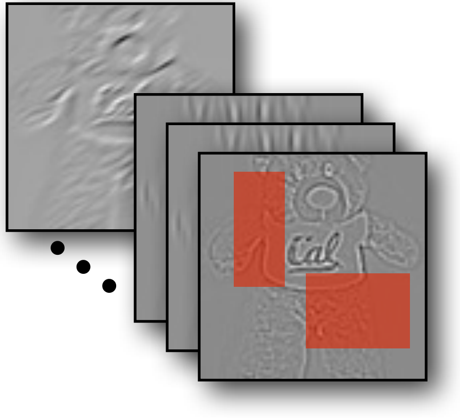
Conclusion (cont'd)



$f(x) =$ "Bear"

Conclusion (cont'd)

- Agnostic to coding



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

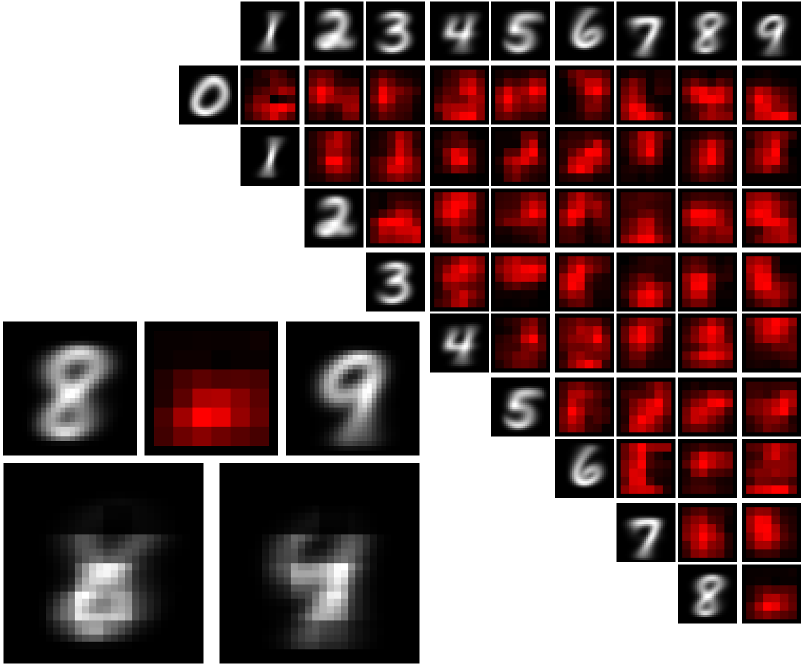
Thank you!

Beyond Spatial Pyramids: Receptive Fields Learning for Pooled Image Features

Yangqing Jia, Chang Huang, Trevor Darrell

- This page is intentionally left blank

Experiment: MNIST



Method	err%
Our Method	0.64
Coates ICML'11	1.02
Lauer PR'07	0.83
Labusch TNN'08	0.59
Ranzato CVPR'09	0.62
Jarrett ICCV'09	0.53

Thus spoke neuroscience