Beyond Spatial Pyramids: Receptive Field Learning for Pooled Image Features

Yangqing Jia[1] Chang Huang[2] Trevor Darrell[1]

[1]UC Berkeley EECS & ICSI [2]NEC Labs America

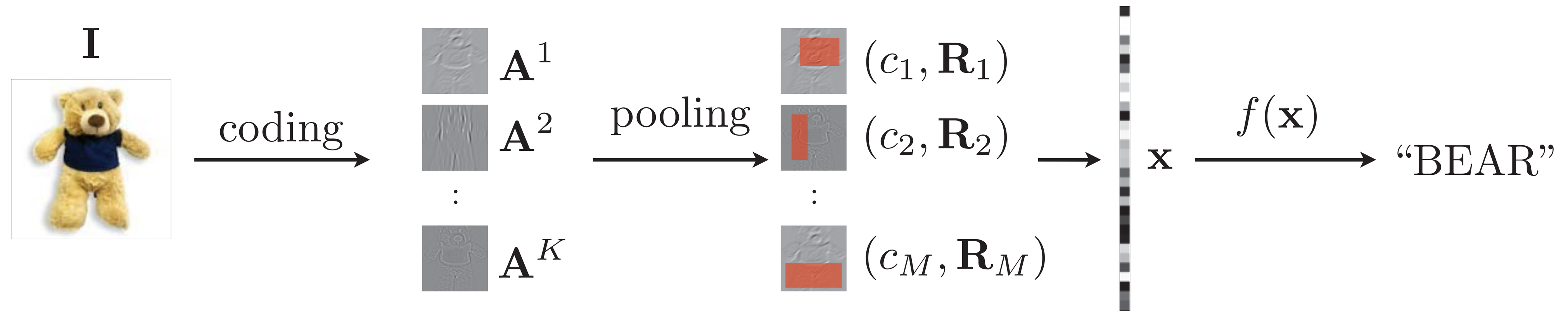
{jiayq,trevor}@berkeley.edu chuang@sv.nec-labs.com

1. Contributions

The key contributions of our work are:

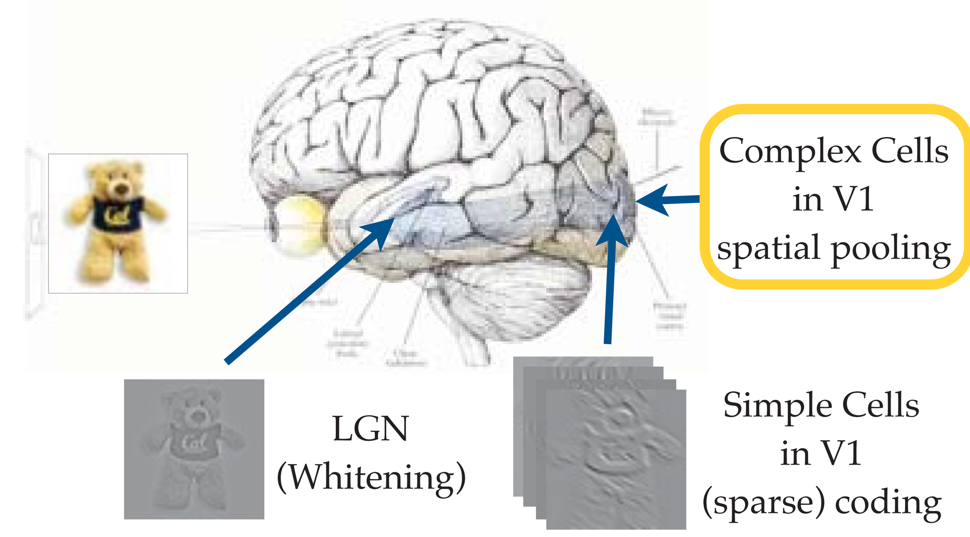
- Analysis of the spatial receptive field (RF) designs for pooled features.
- Evidence that spatial pyramids may be suboptimal in feature generation.
- An algorithm that jointly learns adaptive RF and the classifiers, with an efficient implementation using over-completeness and structured sparsity.

2. The Pipeline



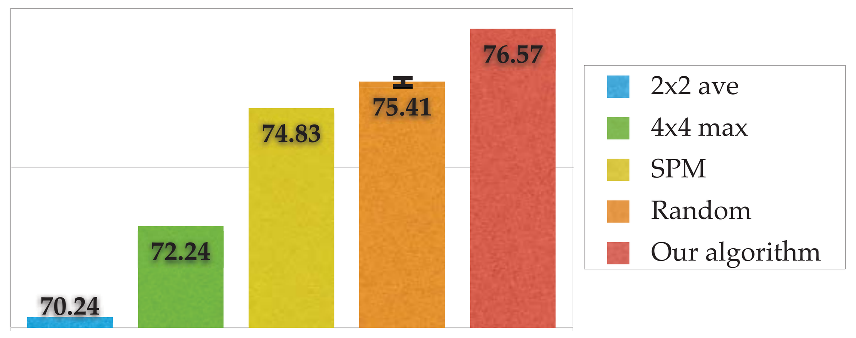
State-of-the-art classification algorithms take a two-layer pipeline: the coding layer learns activations from local image patches, and the pooling layer aggregates activations in multiple spatial regions. Linear classifiers are learned from the pooled features.

3. Neuroscience Inspiration



4. Spatial Pooling Revisited

- Much work has been done on the coding part, while the spatial pooling methods are often hand-crafted.
- Sample performances on CIFAR-10 with different receptive field designs:



(with a dictionary of size 200)

Note the suboptimality of SPM - random selection from an overcomplete set of spatially pooled features consistently outperforms SPM.

- We propose to learn the spatial receptive fields as well as the codes and the classifier.

5. Notations

- $\mathbf{I}$: image input.
- $\mathbf{A}^1, \cdots, \mathbf{A}^K$: code activation as matrices, with $\mathbf{A}^k_{ij}$: activation of code k at position (i, j).
- $\mathbf{R}_i$: RF of the i-th pooled feature.
- $\mathrm{op}(\cdot)$: pooling operator, such as $\max(\cdot)$.
- $f(\mathbf{x}, \boldsymbol{\theta})$: the classifier based on pooled features $\mathbf{x}$.
- A pooled feature x_i is defined by choosing a code indexed by c_i and a spatial RF $\mathbf{R}_i$:

$$x_i = \mathrm{op}(\mathbf{A}^{c_i}_{\mathbf{R}_i})$$

The vector of pooled features $\mathbf{x}$ is then determined by the set of parameters $\mathcal{C} = \{c_1, \cdots, c_M\}$ and $\mathcal{R} = \{\mathbf{R}_1, \cdots, \mathbf{R}_M\}$.

6. The Learning Problem

- Given a set of training data $\{(\mathbf{I}_n, \mathbf{y}_n)\}_{n=1}^N$, we jointly learn the classifier and the pooled features as (assuming that coding is done in an unsupervised way):

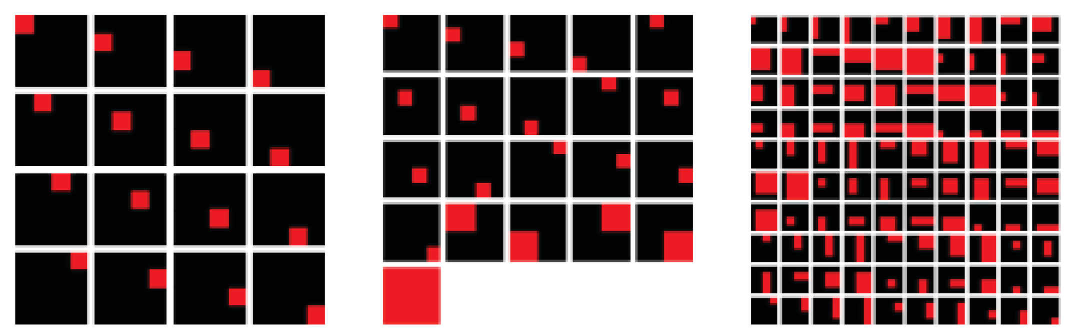
$$\min_{\mathcal{C},\mathcal{R},\boldsymbol{\theta}} \quad \frac{1}{N}\sum_{n=1}^{N} l(f(\mathbf{x}_n; \boldsymbol{\theta}), \mathbf{y}_n) + \lambda\,\mathrm{Reg}(\boldsymbol{\theta})$$

where $\quad x_{ni} = \mathrm{op}(\mathbf{A}^{c_i}_{n,\mathbf{R}_i})$

- Advantage: pooled features are tailored towards the classification task (also reduces redundancy).
- Disadvantage: may be intractable - an exponential number of possible receptive fields.
- Solution: reasonably overcomplete receptive field candidates + sparsity constraints to control the number of final features.

7. Overcomplete RF

- We propose to use overcomplete receptive field candidates based on regular grids:



(a) Base (b) SPM (c) Ours

- The structured sparsity regularization is adopted to select only a subset of features for classification:

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

where $\|\mathbf{W}\|_{1,\infty} = \sum_{i=1}^{M} \max_{j \in \{1,\cdots,L\}} |W_{ij}|$.

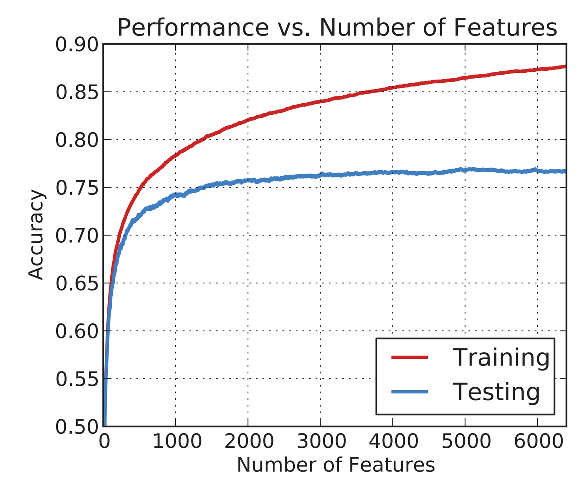
8. Greedy Feature Selection

- Directly perform optimization is still time and memory consuming.
- Following [Perkins JMLR03], We adopted an incremental, greedy approach to select features based on their scores:

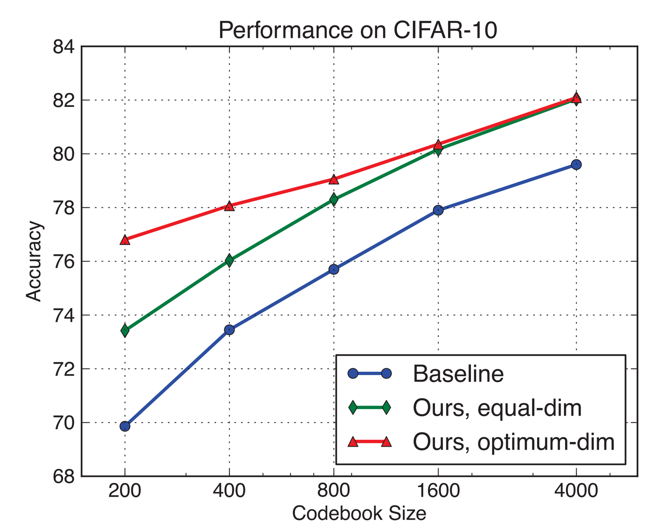
$$\mathrm{score}(x_i) = \left\| \frac{\partial \mathcal{L}(\mathbf{W}, \mathbf{b})}{\partial \mathbf{W}_{i,\cdot}} \right\|^2_{\mathrm{Fro}}$$

- After each increment, the model is retrained only with respect to an active subset of selected features to ensure fast re-training:

$$\mathbf{W}^{(t+1)}_{\mathcal{S}_A,\cdot}, \mathbf{b} = \arg\min_{\mathbf{W}_{\mathcal{S}_A,\cdot}, \mathbf{b}} \mathcal{L}(\mathbf{W}, \mathbf{b})$$



- Benefit of overcompleteness in spatial pooling + feature selection: higher performance with smaller codebooks and lower feature dimensions.



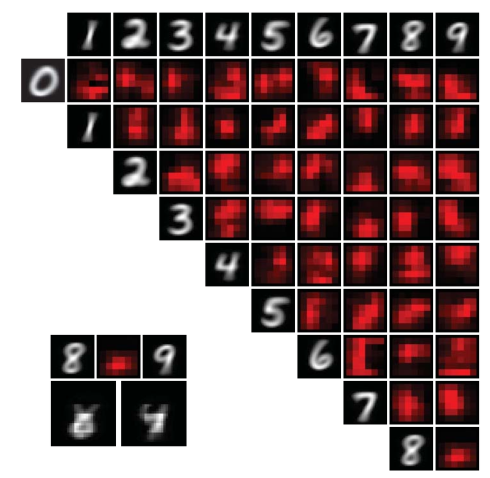
9. Results

- Performance comparison on CIFAR-10 with state-of-the-art approaches:

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

- Result on MNIST and the 1-vs-1 saliency map obtained from our algorithm:

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



10. References

- A Coates and AY Ng. The importance of encoding versus training with sparse coding and vector quantization. ICML 2011.
- S Perkins, K Lacker, and J Theiler. Grafting: fast, incremental feature selection by gradient descent in function space. JMLR, 3:1333–1356, 2003.
- DH Hubel and TN Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. of Physiology, 160(1):106–154, 1962.