CVPR12 Tutorial on Deep Learning

Sparse Coding

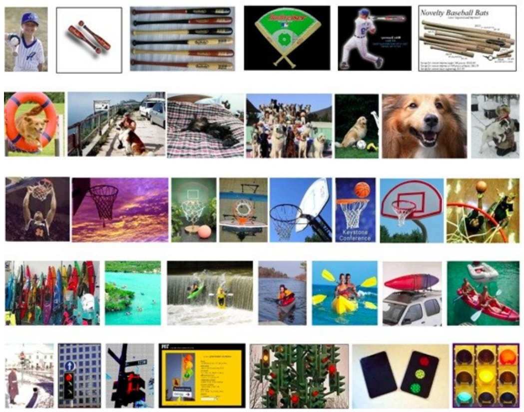
Kai Yu

yukai@baidu.com

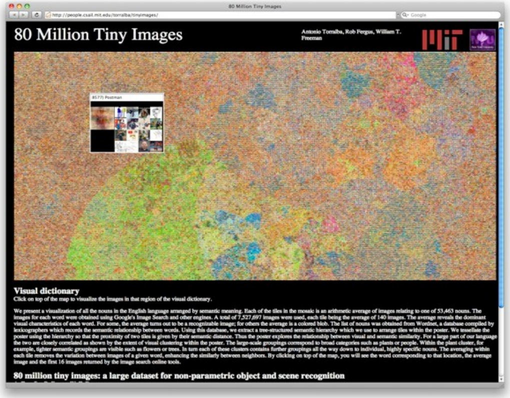
Department of Multimedia, Baidu

Relentless research on visual recognition

Caltech 101



80 Million Tiny Images

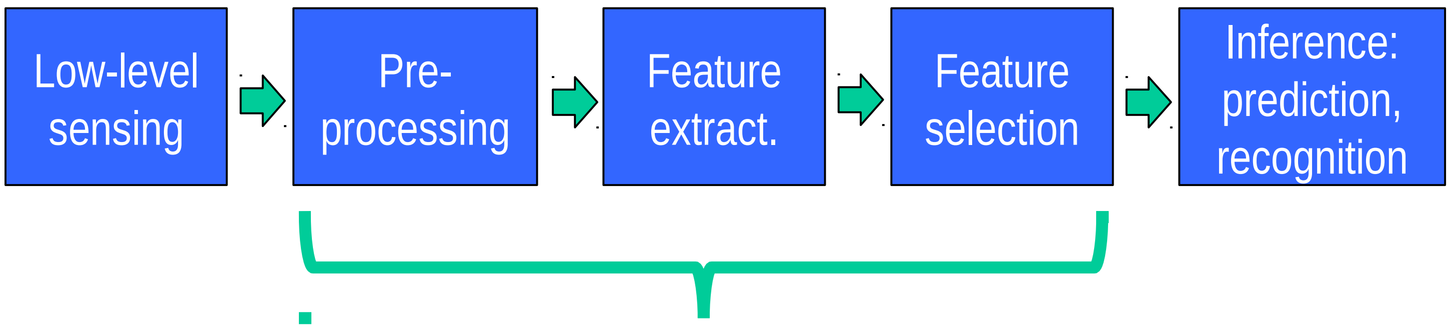


PASCAL VOC

ImageNet

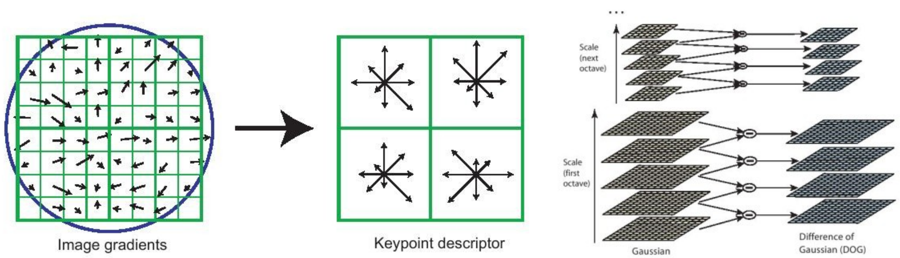
The pipeline of machine visual perception

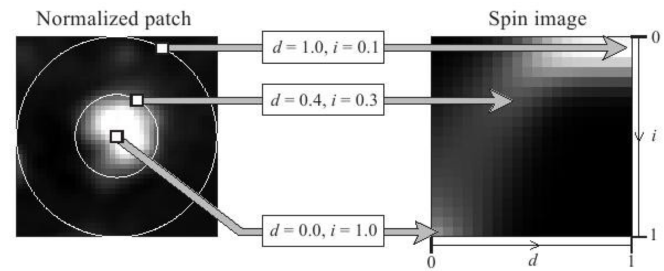
Most Efforts in Machine Learning

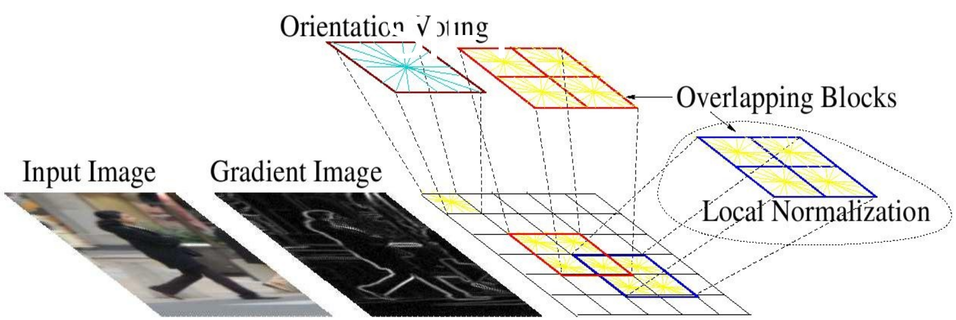


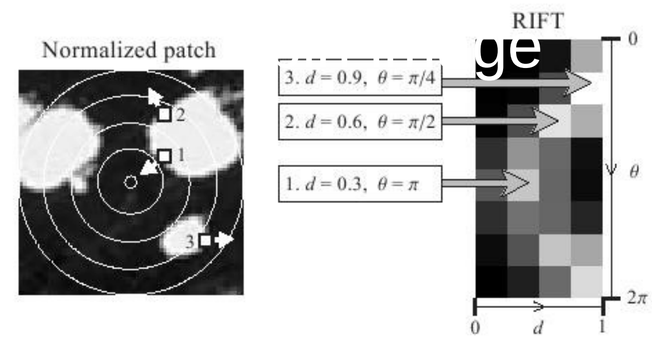
- Most critical for accuracy
- Account for most of the computation for testing
- Most time-consuming in development cycle
- Often hand-craft in practice

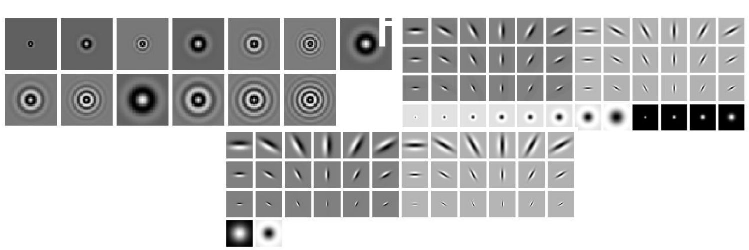
Computer vision features

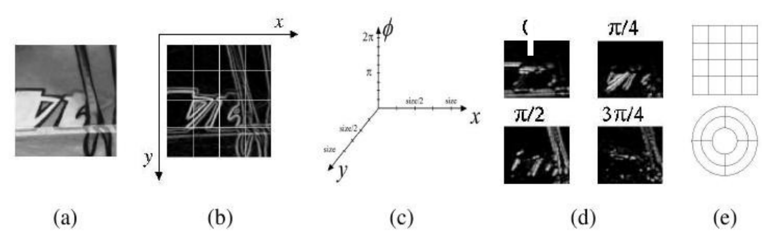






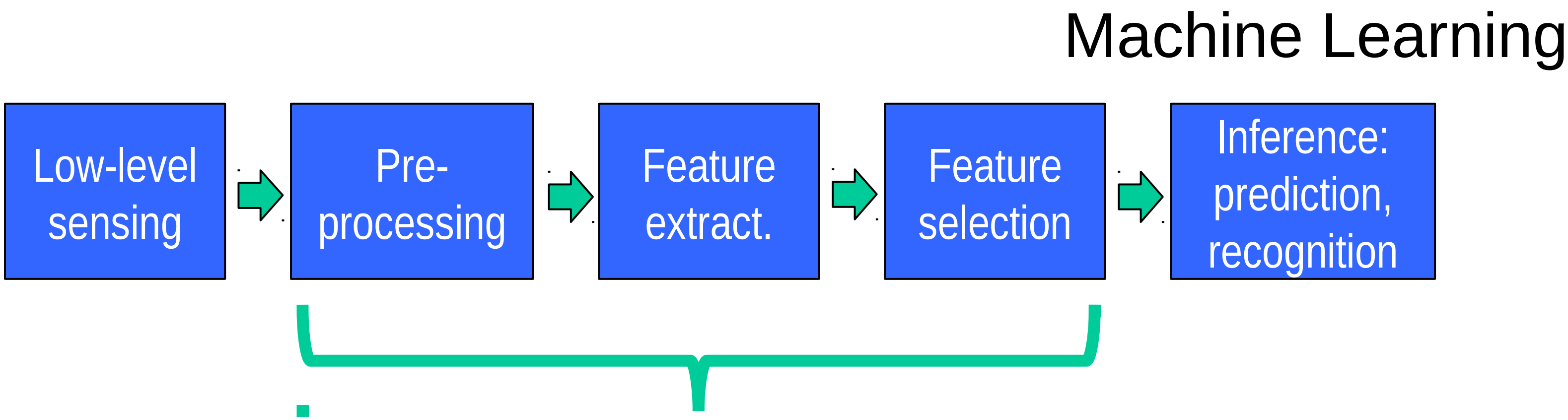






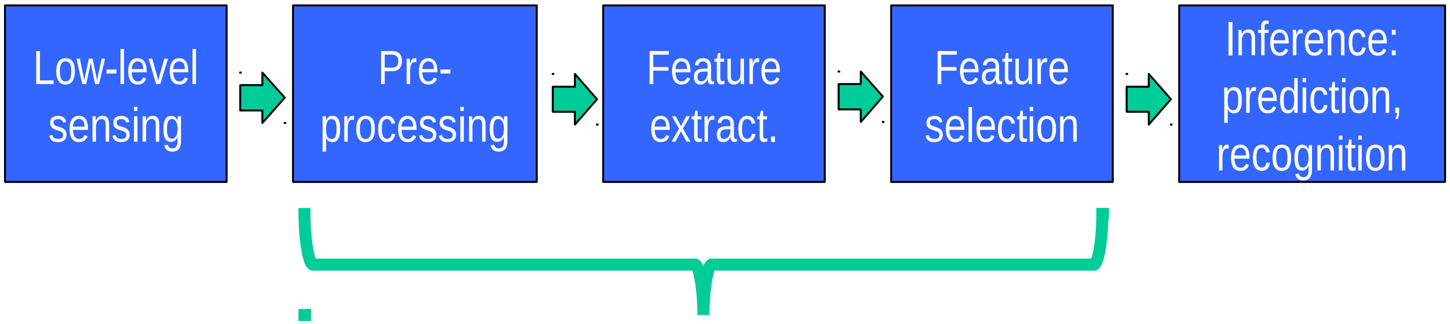
Slide Credit: Andrew Ng

Learning features from data



Feature Learning:
instead of design features,
let’s design feature learners

Learning features from data via sparse coding



Sparse coding offers an effective building block to learn useful features

Outline

1. Sparse coding for image classification
2. Understanding sparse coding
3. Hierarchical sparse coding
4. Other topics: e.g. structured model, scale-up, discriminative training
5. Summary

"BoW representation + SPM" Paradigm - I

Bag-of-visual-words representation
(BoW) based on VQ coding

 Figure credit: Fei-Fei Li

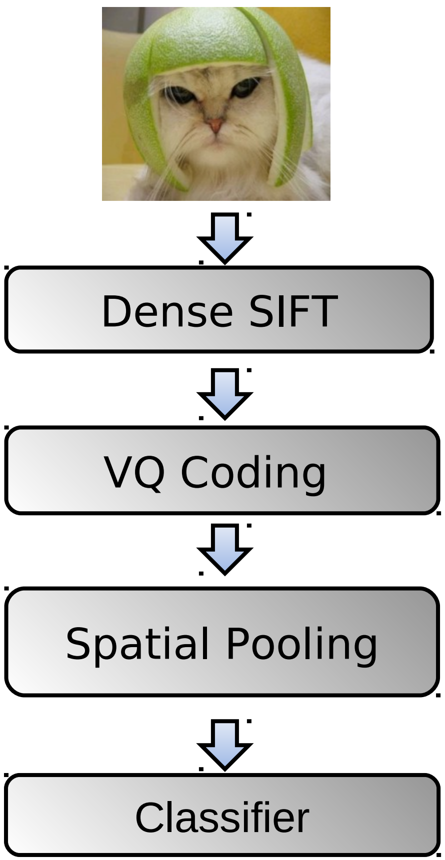
"BoW representation + SPM" Paradigm - II

Spatial pyramid matching:
pooling in different scales and locations

Figure credit: Svetlana Lazebnik

Image Classification using “BoW + SPM”

Image Classification

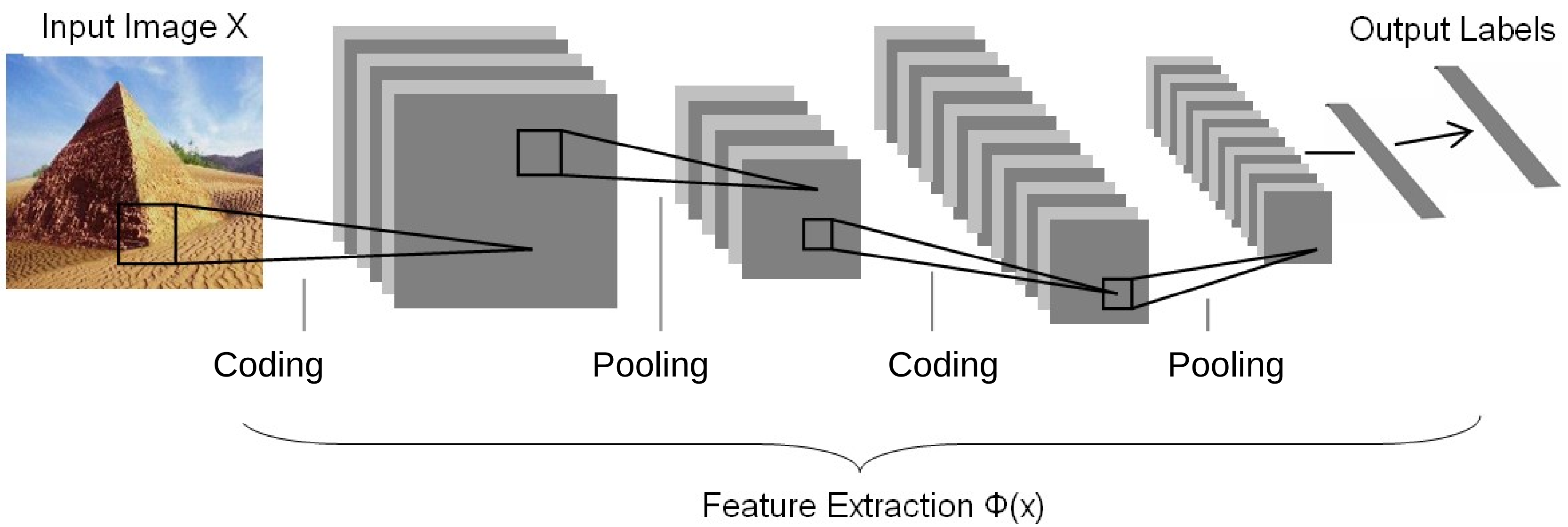


$$\left[x^{(1)}, x^{(2)}, \ldots, x^{(m)}\right] \in \mathbb{R}^{128}$$

$$\left[a^{(1)}, a^{(2)}, \ldots, a^{(m)}\right] \in \mathbb{R}^{k}$$

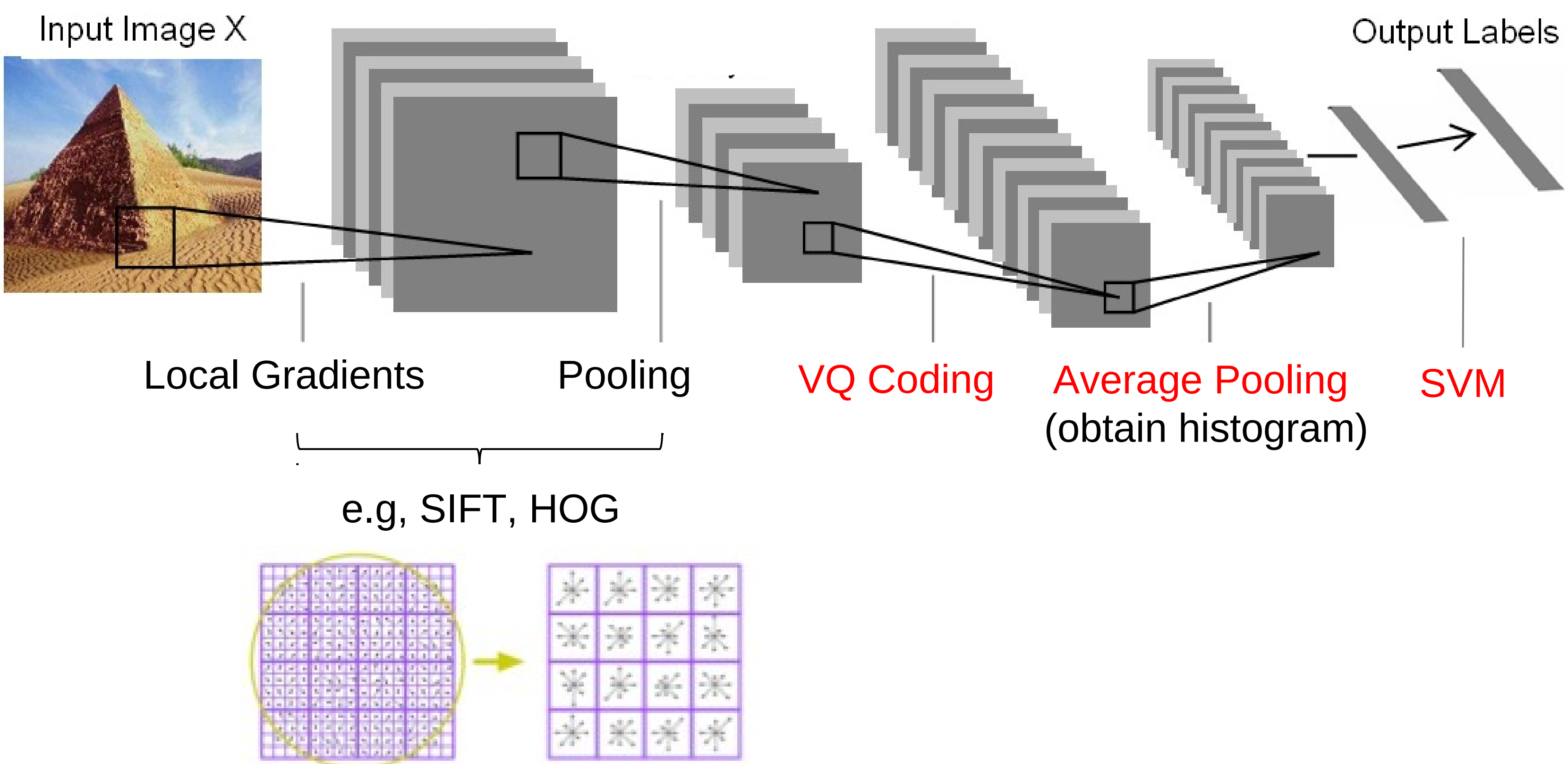
$$a = \sum_{i=1}^{m} v_i a^{(i)}$$

The Architecture of “Coding + Pooling”



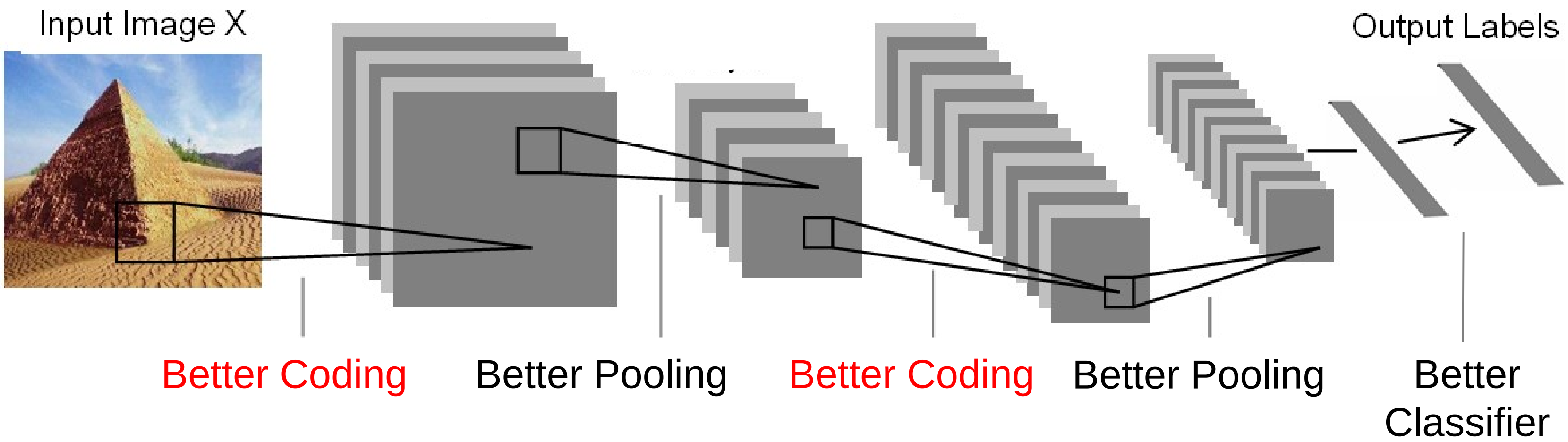
- e.g., convolutional neural net, HMAX, BoW, …

“BoW+SPM” has two coding+pooling layers



SIFT feature itself follows a coding+pooling operation

Develop better coding methods



- Coding: nonlinear mapping data into another feature space
- Better coding methods: sparse coding, RBMs, auto-encoders

What is sparse coding

Sparse coding (Olshausen & Field,1996). Originally developed to explain early visual processing in the brain (edge detection).

Training: given a set of random patches x, learning a dictionary of bases [$\Phi_{1,}$ $\Phi_{2,}$ …]

Coding: for data vector x, solve LASSO to find the sparse coefficient vector a

$$\min_{a,\phi} \sum_{i=1}^{m} \left\| x_i - \sum_{j=1}^{k} a_{i,j}\phi_j \right\|^2 + \lambda \sum_{i=1}^{m}\sum_{j=1}^{k} |a_{i,j}|$$

Sparse coding: training time

Input: Images $x_1, x_2, \ldots, x_m \Leftarrow\rceil\dashv\rfloor\langle\rangle\backslash\mathcal{R}^d$)

Learn: Dictionary of bases $\phi_1, \phi_2, \ldots, \phi_k \Leftarrow\dashv\Updownarrow\int\wr\mathcal{R}^d$).

$$\min_{a,\phi} \sum_{i=1}^{m} \left\| x_i - \sum_{j=1}^{k} a_{i,j}\phi_j \right\|^2 + \lambda \sum_{i=1}^{m}\sum_{j=1}^{k} |a_{i,j}|$$

Alternating optimization:

1.Fix dictionary $\phi_1, \phi_2, \ldots, \phi_k$, optimize a (a standard LASSO problem)

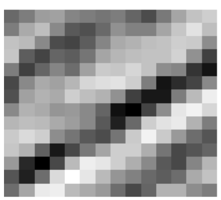
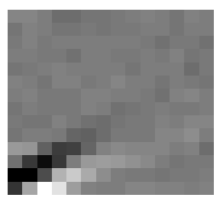
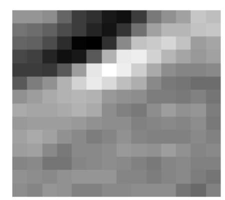
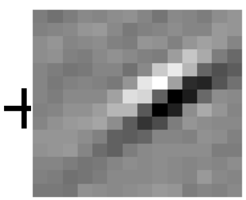
2.Fix activations a, optimize dictionary $\phi_1, \phi_2, \ldots, \phi_k$ (a convex QP problem)

Sparse coding: testing time

Input: Novel image patch x $\Leftarrow\rangle\backslash\mathcal{R}^{d}$) and previously learned ϕ_i's

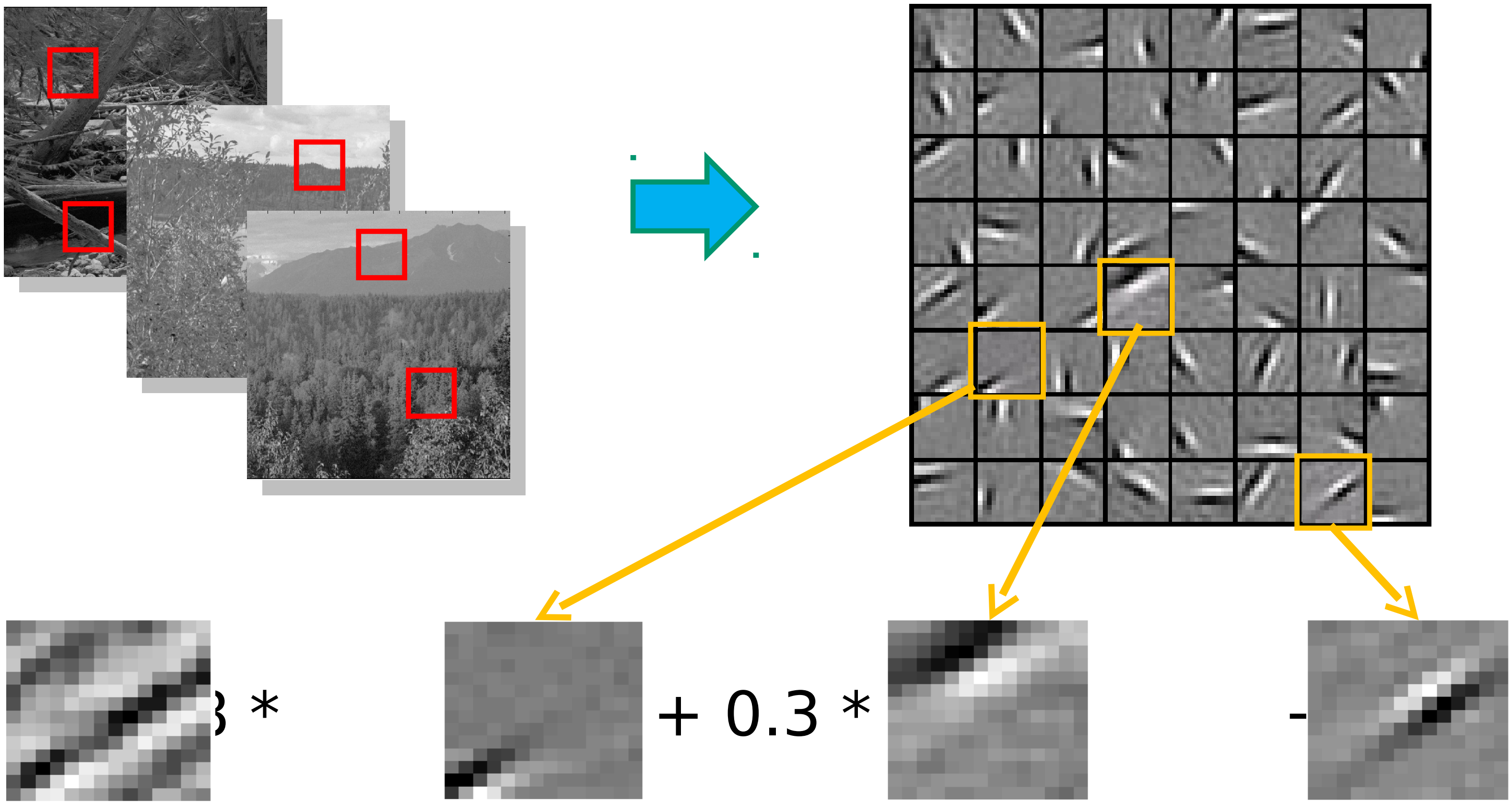
Output: Representation [$a_{i,1}$, $a_{i,2}$, ..., $a_{i,\kappa}$] of image patch x_i.

$$\min_a \sum_{i=1}^{m} \left\| x_i - \sum_{j=1}^{k} a_{i,j}\phi_j \right\|^2 + \lambda \sum_{i=1}^{m} \sum_{j=1}^{k} |a_{i,j}|$$

 3 *  + 0.3 *  + 

Represent x_i as: a_i = [0, 0, ..., 0, **0.8**, 0, ..., 0, **0.3**, 0, ..., 0, **0.5**, ...]

Sparse coding illustration



$[a_1, \ldots, a_{64}] = [0, 0, \ldots, 0, \mathbf{0.8}, 0, \ldots, 0, \mathbf{0.3}, 0, \ldots, 0, \mathbf{0.5}, 0]$

(feature representation)

Slide credit: Andrew Ng

Compact & easily interpretable

Self-taught Learning

[Raina, Lee, Battle, Packer & Ng, ICML 07]

Motorcycles

Not motorcycles

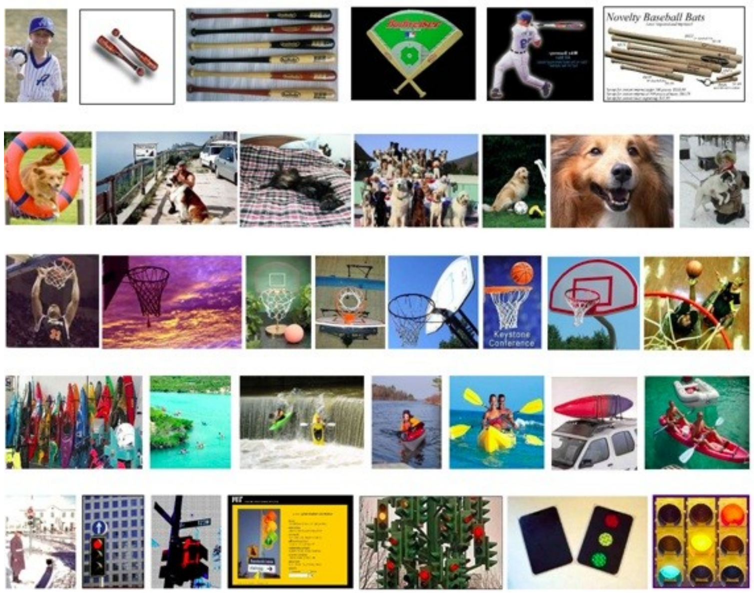
Unlabeled images

Testing:
What is this

Slide credit: Andrew Ng

Classification Result on Caltech 101

9K images, 101 classes



64%

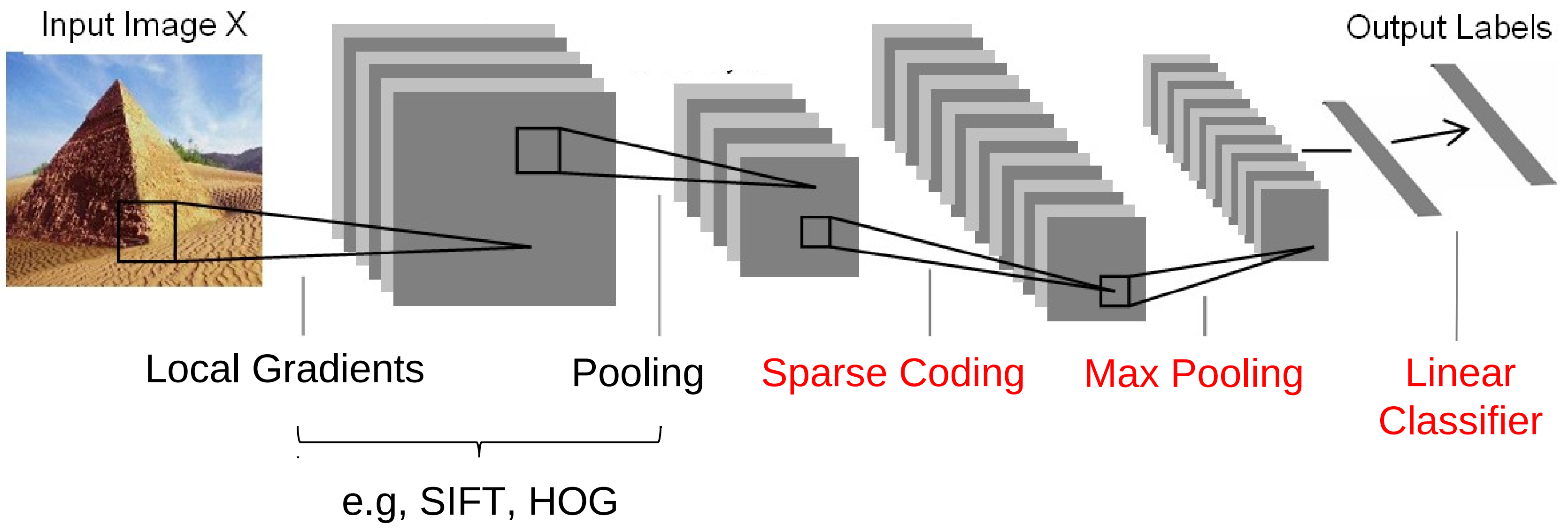
SIFT VQ + Nonlinear SVM

~50%

Pixel Sparse Coding + Linear SVM

Sparse Coding on SIFT – ScSPM algorithm

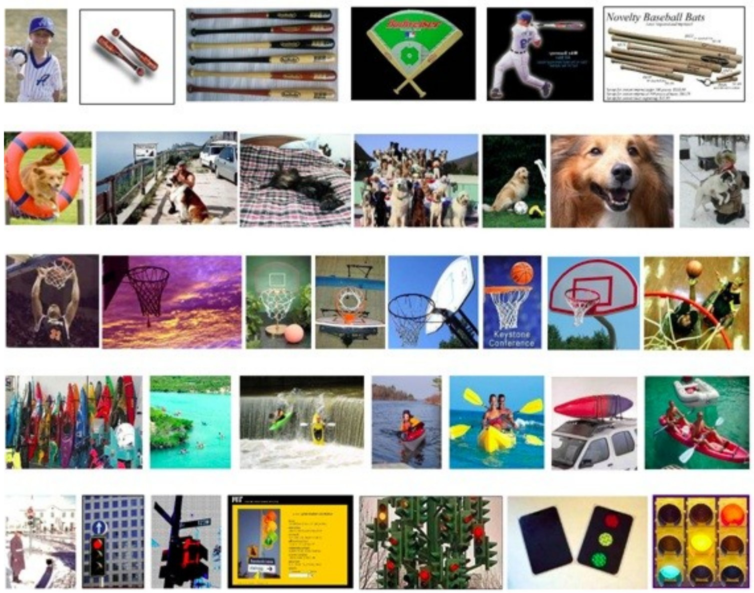
[Yang, Yu, Gong & Huang, CVPR09]



Sparse Coding on SIFT – the ScSPM algorithm

[Yang, Yu, Gong & Huang, CVPR09]

Caltech-101



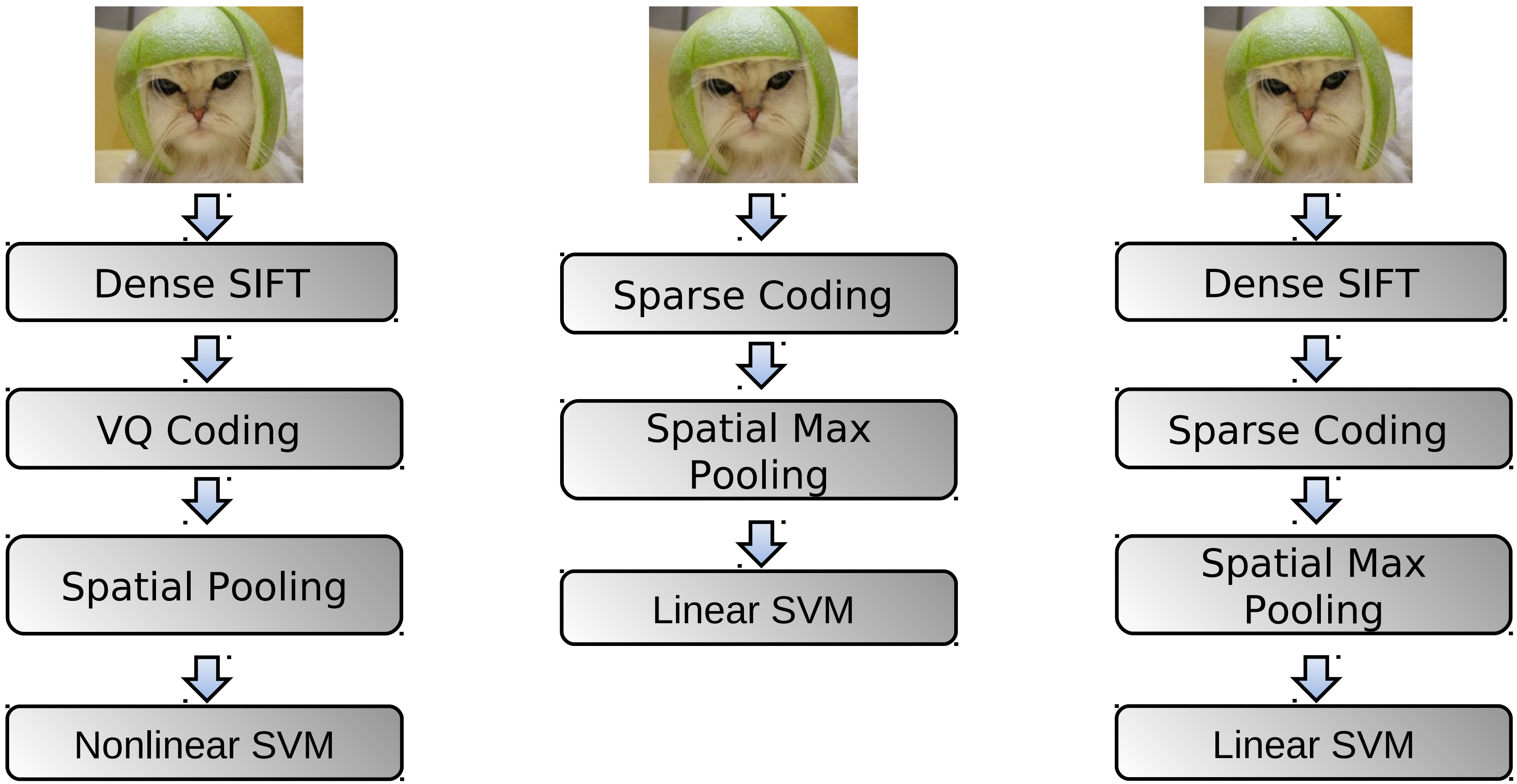
64%

SIFT VQ + Nonlinear SVM

73%

SIFT Sparse Coding + Linear SVM (ScSPM)

Summary: Accuracies on Caltech 101



Key message:
- Deep models are preferred
- Sparse coding is a better building block

Outline

1. Sparse coding for image classification
2. Understanding sparse coding
3. Hierarchical sparse coding
4. Other topics: e.g. structured model, scale-up, discriminative training
5. Summary

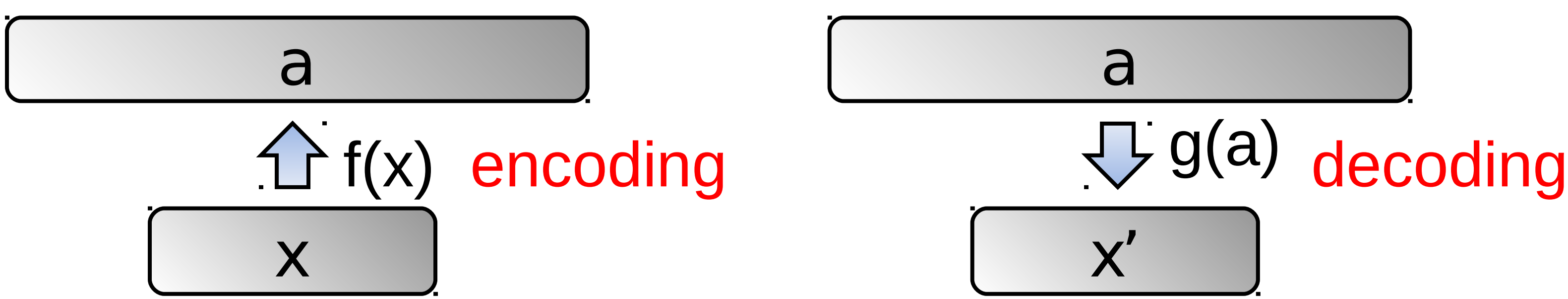
Outline

1. Sparse coding for image classification
2. Understanding sparse coding
   - Connections to RBMs, autoencoders, …
   - Sparse activations vs. sparse models, ..
   - Sparsity vs. locality
   - local sparse coding methods
3. Hierarchical sparse coding
4. Other topics: e.g. structured model, scale-up, discriminative training
5. Summary

Classical sparse coding

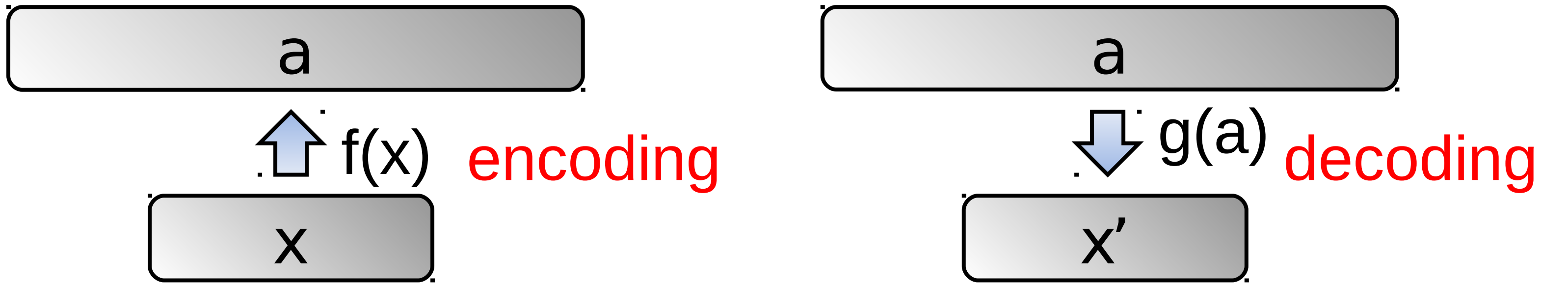
$$\min_{a} \sum_{i=1}^{m} \left\| x_i - \sum_{j=1}^{k} a_{i,j} \phi_j \right\|^2 + \lambda \sum_{i=1}^{m} \sum_{j=1}^{k} |a_{i,j}|$$

- a is sparse
- a is often higher dimension than x
- Activation a=f(x) is nonlinear implicit function of x
- reconstruction x'=g(a) is linear & explicit



RBM & autoencoders

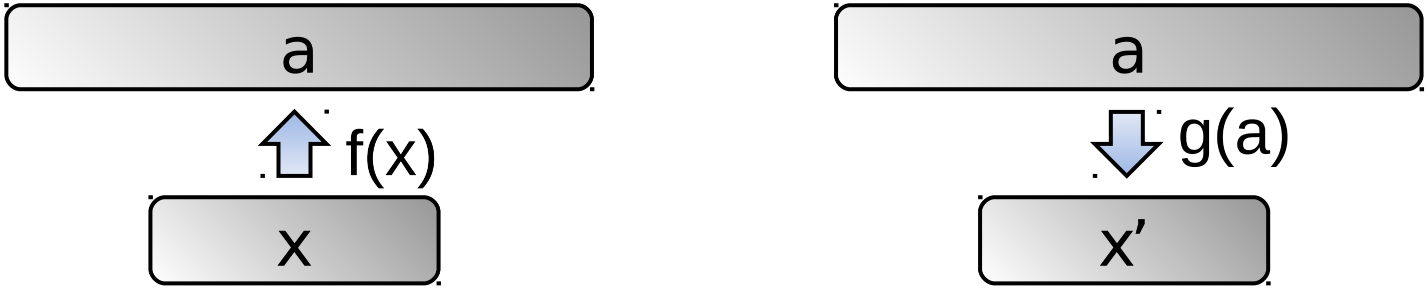
- also involve activation and reconstruction
- but have explicit f(x)
- not necessarily enforce sparsity on a
- but if put sparsity on a, often get improved results [e.g. sparse RBM, Lee et al. NIPS08]



Sparse coding: A broader view

Any feature mapping from x to a, i.e. a = f(x), where

- a is sparse (and often higher dim. than x)
- f(x) is nonlinear
- reconstruction x'=g(a) , such that x'≈x



Therefore, sparse RBMs, sparse auto-encoder, even VQ can be viewed as a form of sparse coding.

Outline

1. Sparse coding for image classification
2. Understanding sparse coding
   - Connections to RBMs, autoencoders, …
   - Sparse activations vs. sparse models, …
   - Sparsity vs. locality
   - local sparse coding methods
3. Hierarchical sparse coding
4. Other topics: e.g. structured model, scale-up, discriminative training
5. Summary

Sparse activations vs. sparse models

For a general function learning problem a = f(x):

1. sparse model: f(x)'s parameters are sparse
   - example: LASSO f(x)=<w,x>, w is sparse
   - the goal is feature selection: all data selects a common subset of features
   - hot topic in machine learning

2. sparse activations: f(x)'s outputs are sparse
   - example: sparse coding a=f(x), a is sparse
   - the goal is feature learning: different data points activate different feature subsets

Example of sparse models

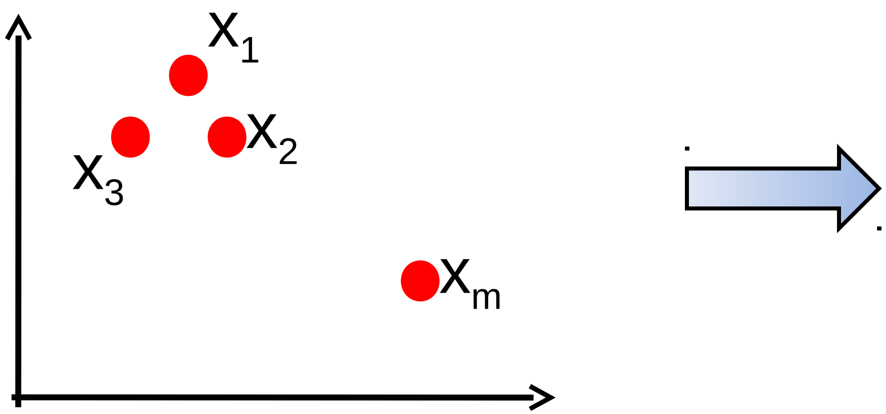
f(x)=<w,x>, where
w=[0, 0.2, 0, 0.1, 0, 0]

x_1 [| | | | | |]
x_2 [| | | | | |]
x_3 [| | | | | |]
⋮
x_m [| | | | | |]

[0 | 0 | 0 0]
[0 | 0 | 0 0]
[0 | 0 | 0 0]
⋮
[0 | 0 | 0 0]

- because the 2nd and 4th elements of w are non-zero, these are the two selected features in x
- globally-aligned sparse representation

Example of sparse activations (sparse coding)



a_1 [0 | | 0 0 … 0]

a_2 [| | 0 0 0 … 0]

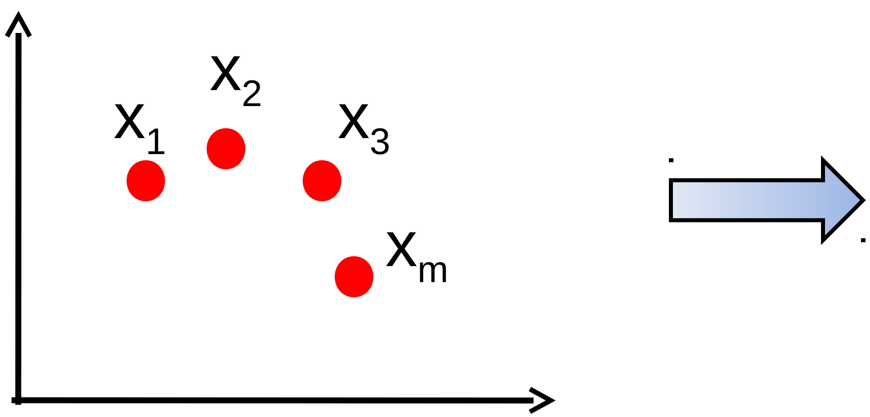
a_3 [| 0 | 0 0 … 0]

⋮

a_m [0 0 0 | | … 0]

- different x has different dimensions activated
- locally-shared sparse representation: similar x’s tend to have similar non-zero dimensions

Example of sparse activations (sparse coding)



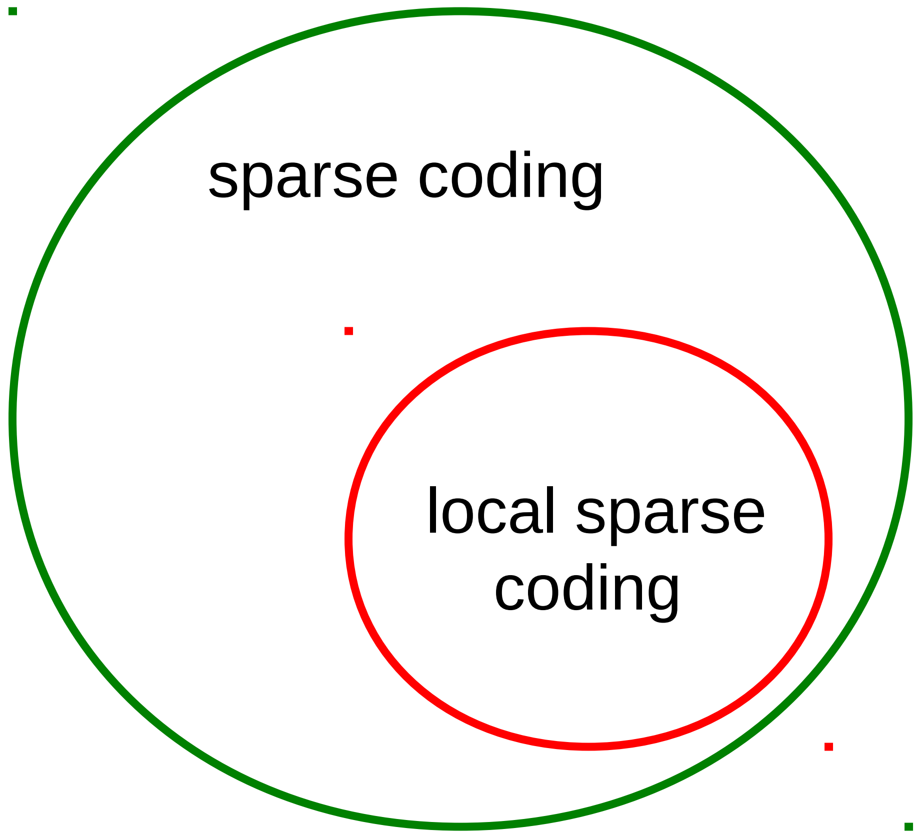
a_1 [| | 0 0 0 … 0]
a_2 [0 | | 0 0 … 0]
a_3 [0 0 | | 0 … 0]
⋮
a_m [0 0 0 | | … 0]

- another example: preserving manifold structure
- more informative in highlighting richer data structures, i.e. clusters, manifolds,

Outline

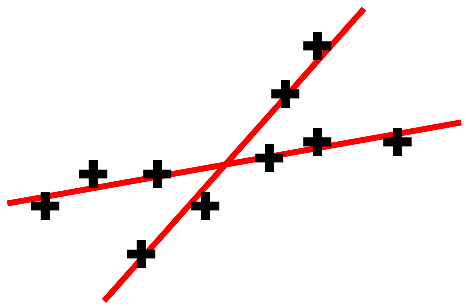
1. Sparse coding for image classification
2. Understanding sparse coding
   - Connections to RBMs, autoencoders, …
   - Sparse activations vs. sparse models, …
   - Sparsity vs. locality
   - Local sparse coding methods
3. Hierarchical sparse coding
4. Other topics: e.g. structured model, scale-up, discriminative training
5. Summary

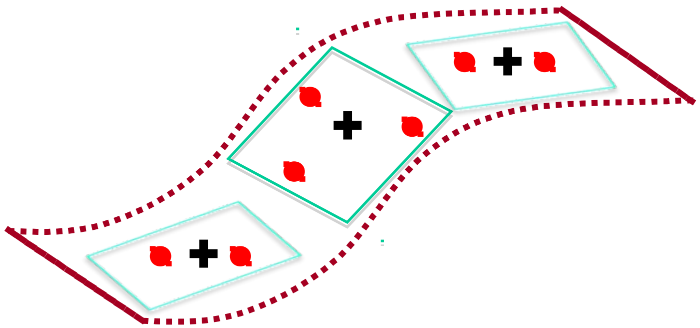
Sparsity vs. Locality



- Intuition: similar data should get similar activated features
- Local sparse coding:
  - data in the same neighborhood tend to have shared activated features;
  - data in different neighborhoods tend to have different features activated.

Sparse coding is not always local: example





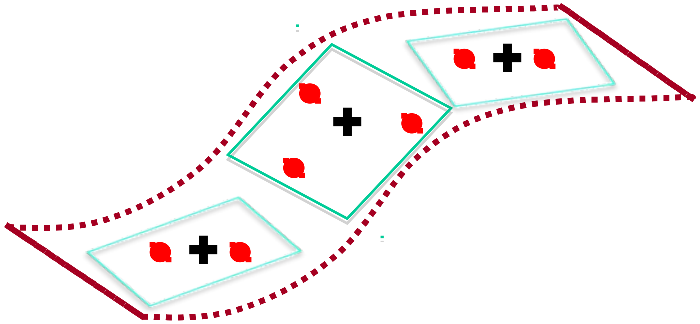
Case 1
independent subspaces

Case 2
data manifold (or clusters)

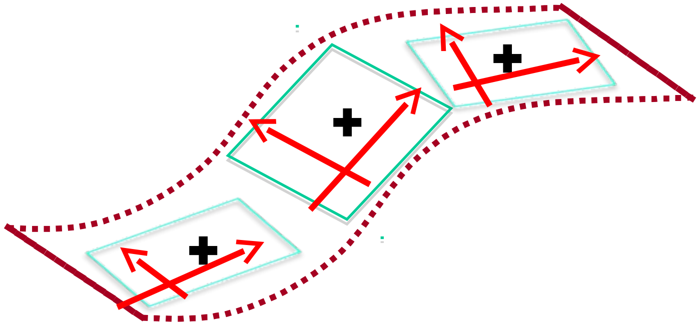
- Each basis is a "direction"
- Sparsity: each datum is a linear combination of only several bases.

- Each basis an "anchor point"
- Sparsity: each datum is a linear combination of neighbor anchors.
- Sparsity is caused by locality.

Two approaches to local sparse coding



Approach 1
Coding via local anchor points



Approach 2
Coding via local subspaces

Classical sparse coding is empirically local

$$\min_{a} \sum_{i=1}^{m} \left\| x_i - \sum_{j=1}^{k} a_{i,j} \phi_j \right\|^2 + \lambda \sum_{i=1}^{m} \sum_{j=1}^{k} |a_{i,j}|$$

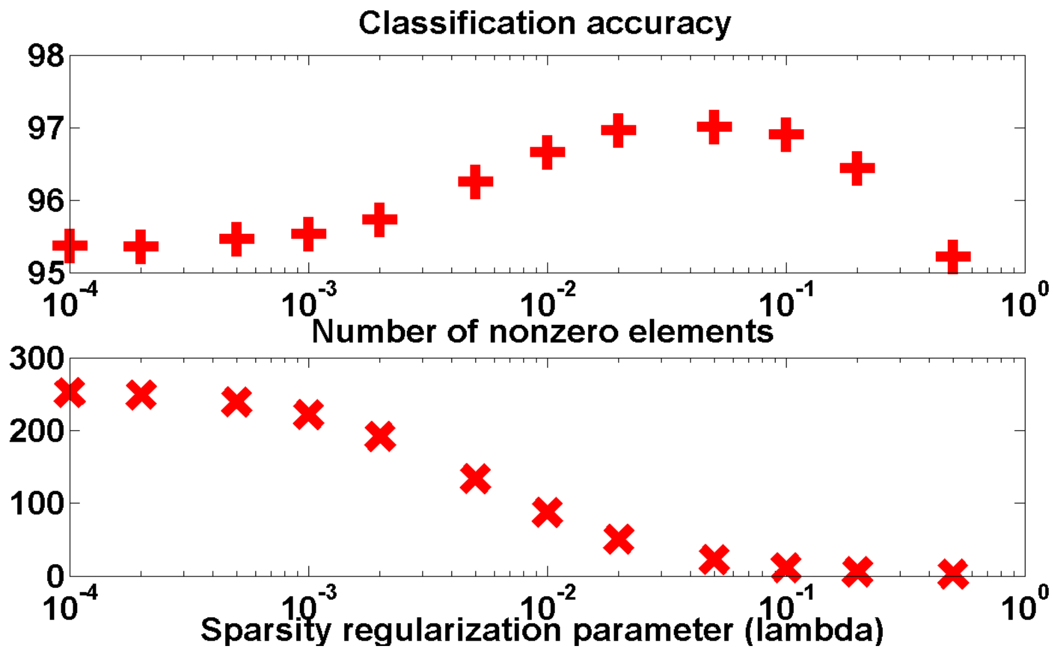
- When it works best for classification, the codes are often found local.

- It's preferred to let similar data have similar non-zero dimensions in their codes.

MNIST Experiment: Classification using SC

$$\min_{a,\phi} \sum_{i=1}^{m} \left\| x_i - \sum_{j=1}^{k} a_{i,j}\phi_j \right\|^2 + \lambda \sum_{i=1}^{m}\sum_{j=1}^{k} |a_{i,j}|$$

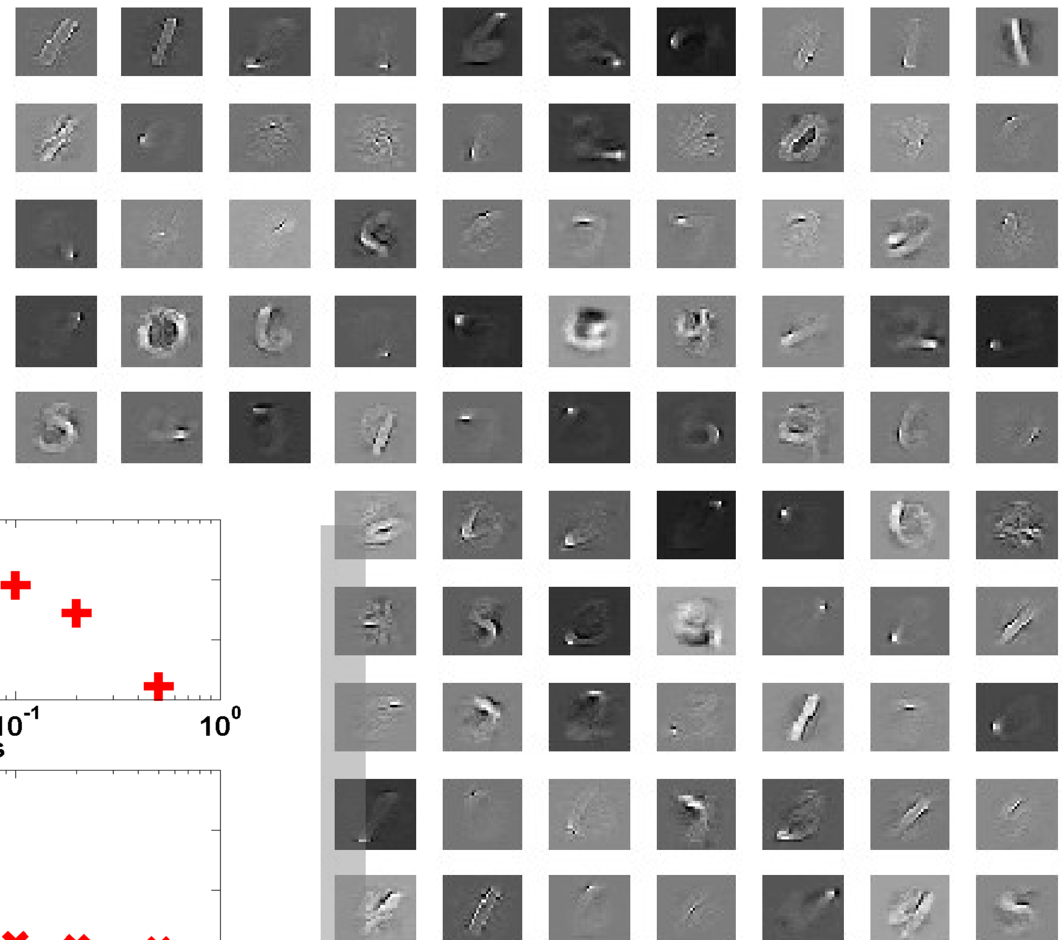
Try different values

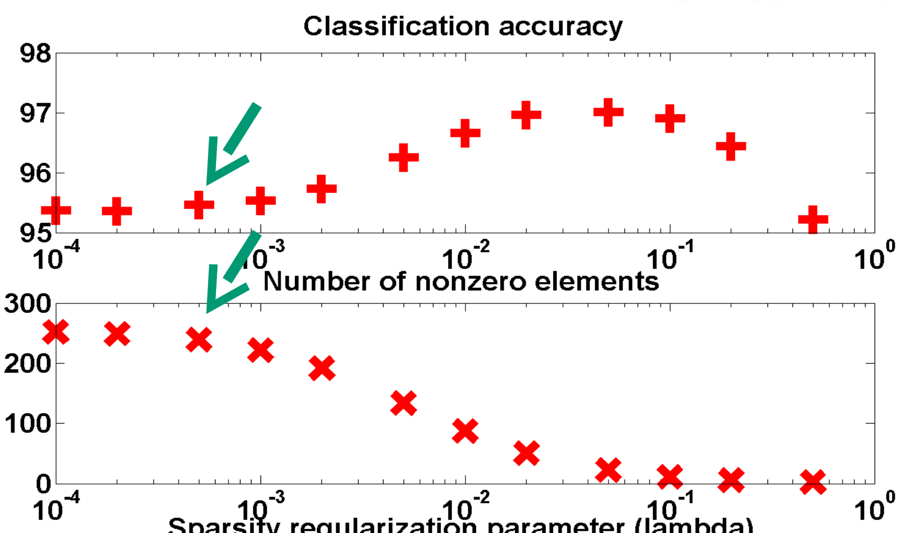


- 60K training, 10K for test
- Let k=512
- Linear SVM on sparse codes

MNIST Experiment: Lambda = 0.0005

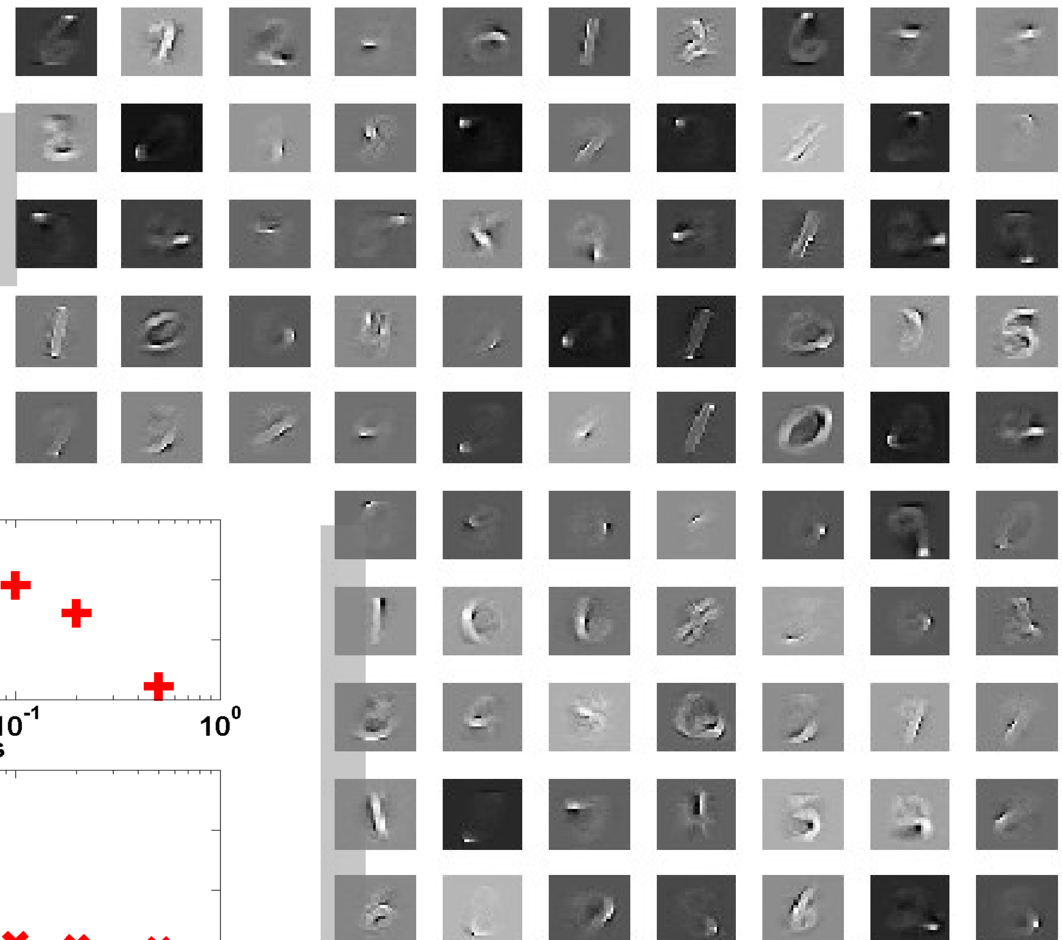
Each basis is like a part or direction.

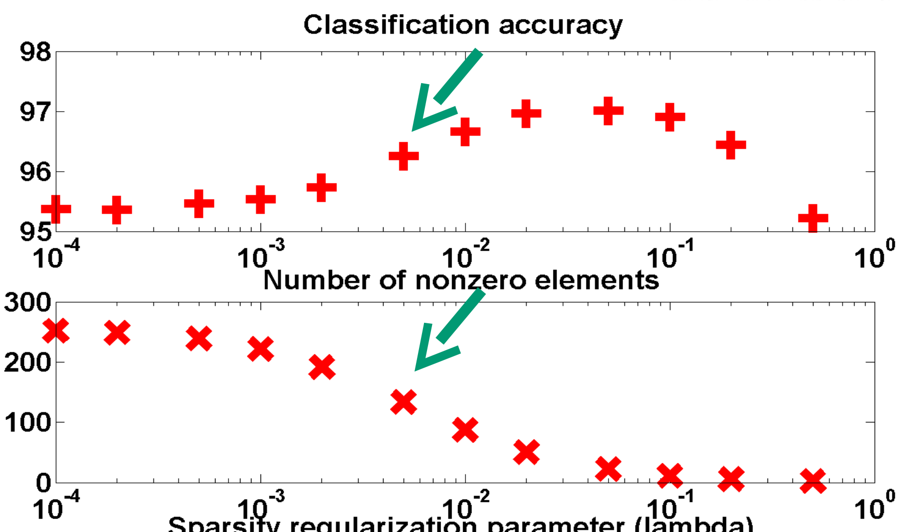




MNIST Experiment: Lambda = 0.005

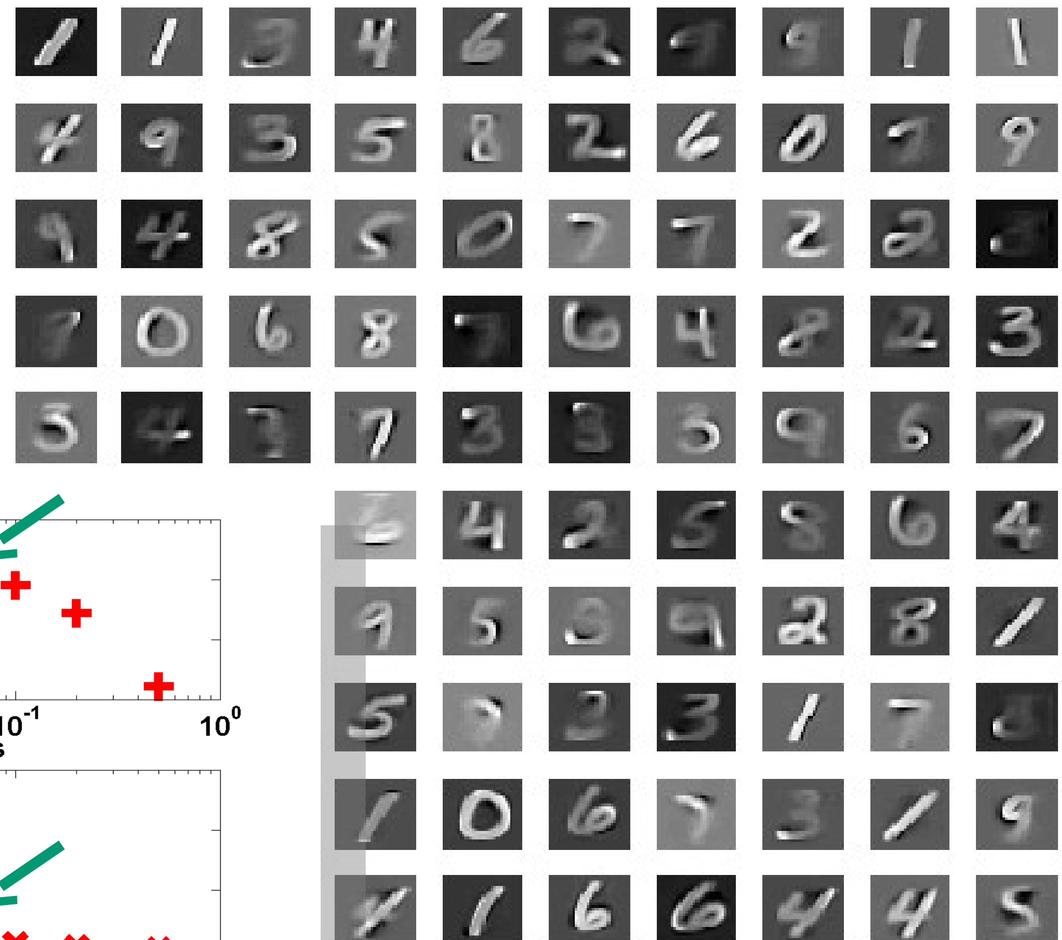
Again, each basis is like a part or direction.

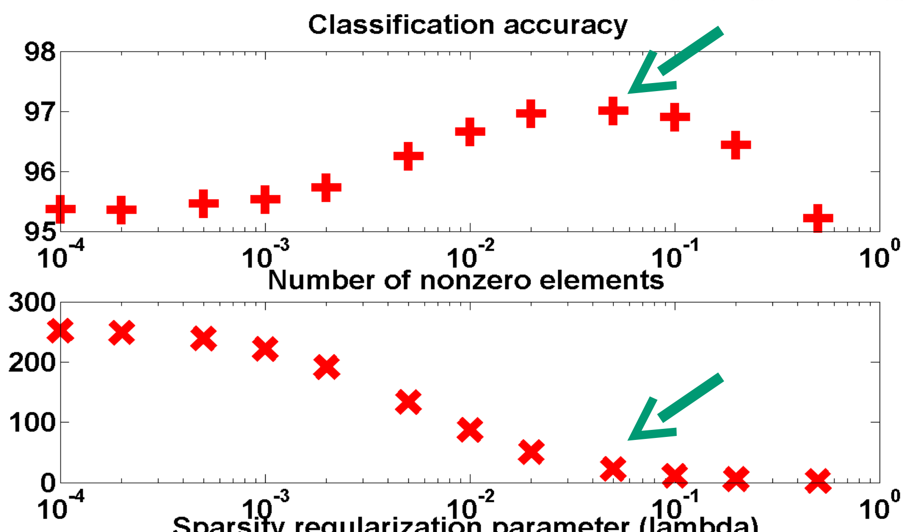




MNIST Experiment: Lambda = 0.05

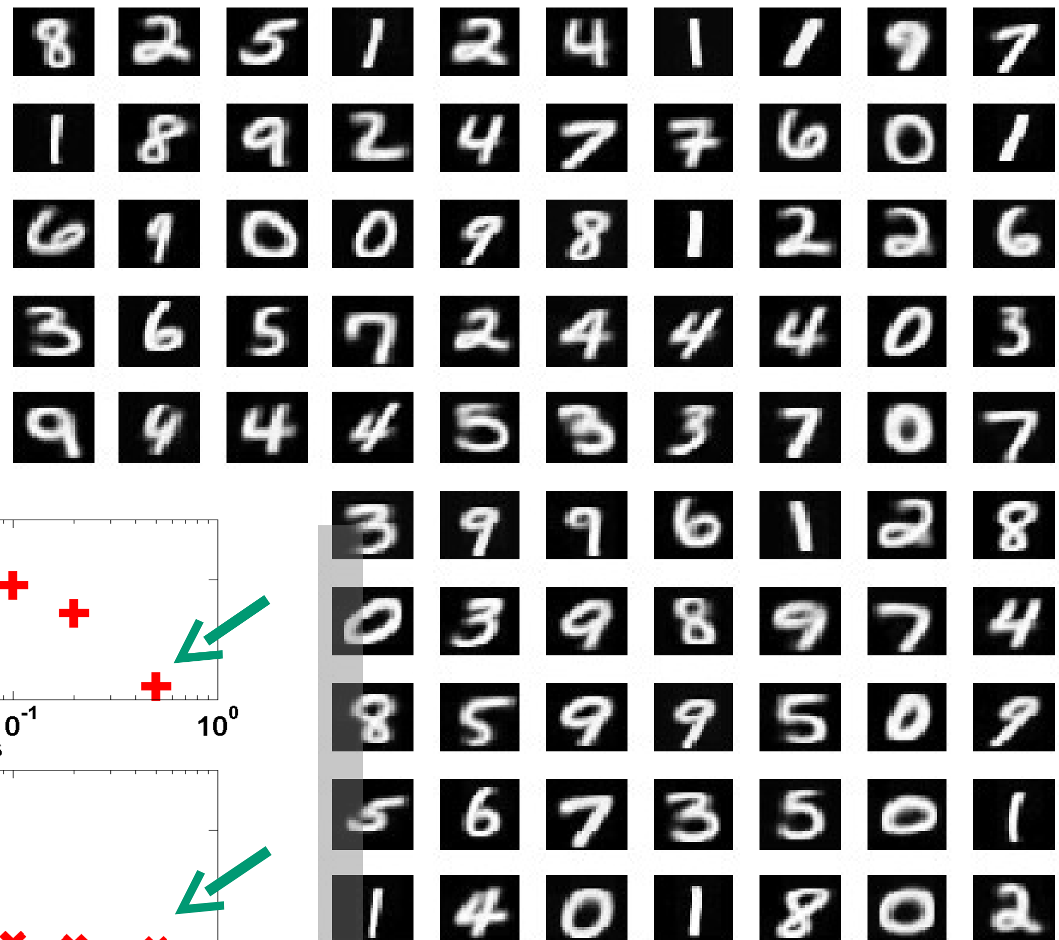
Now, each basis is more like a digit !

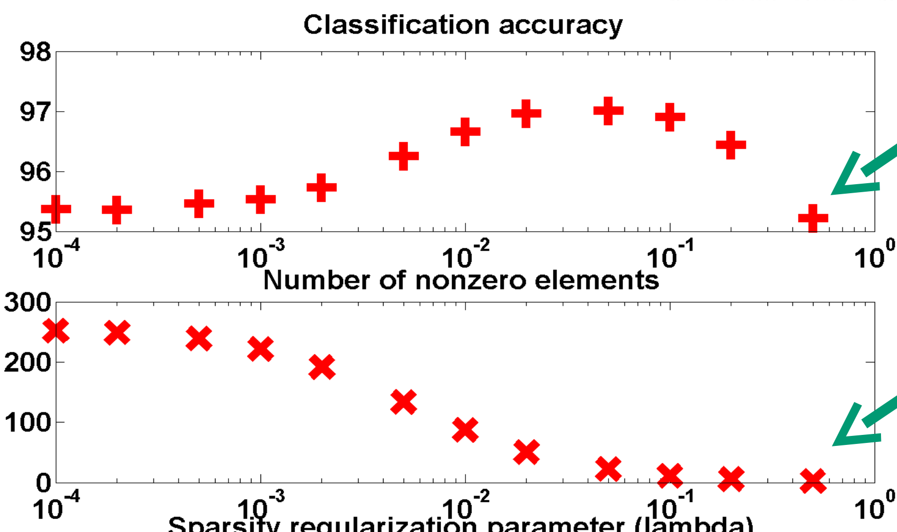




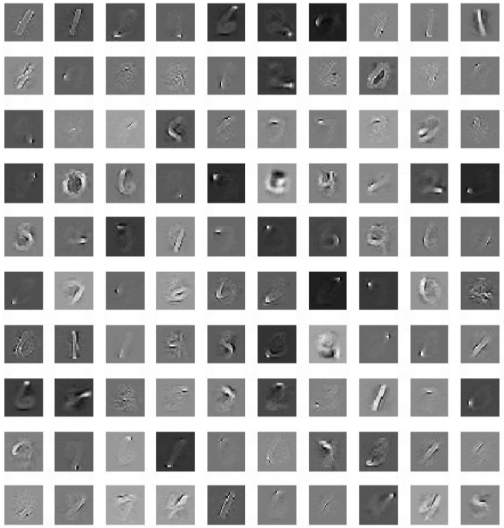
MNIST Experiment: Lambda = 0.5

Like VQ now!

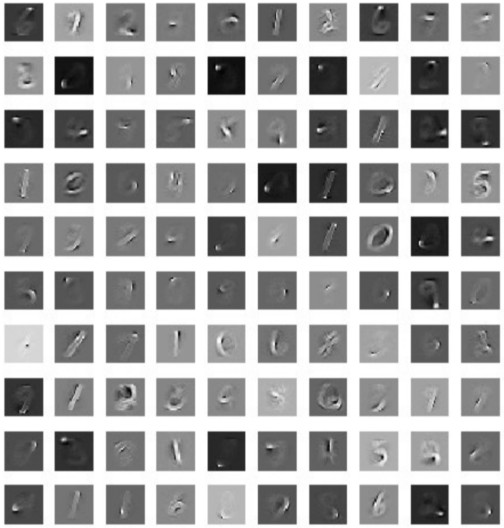




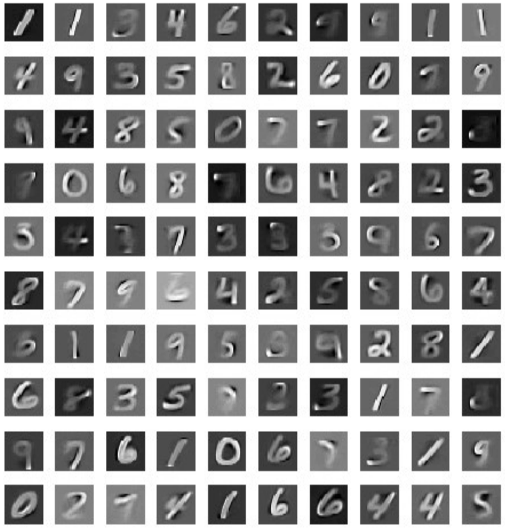
Geometric view of sparse coding



Error: 4.54%



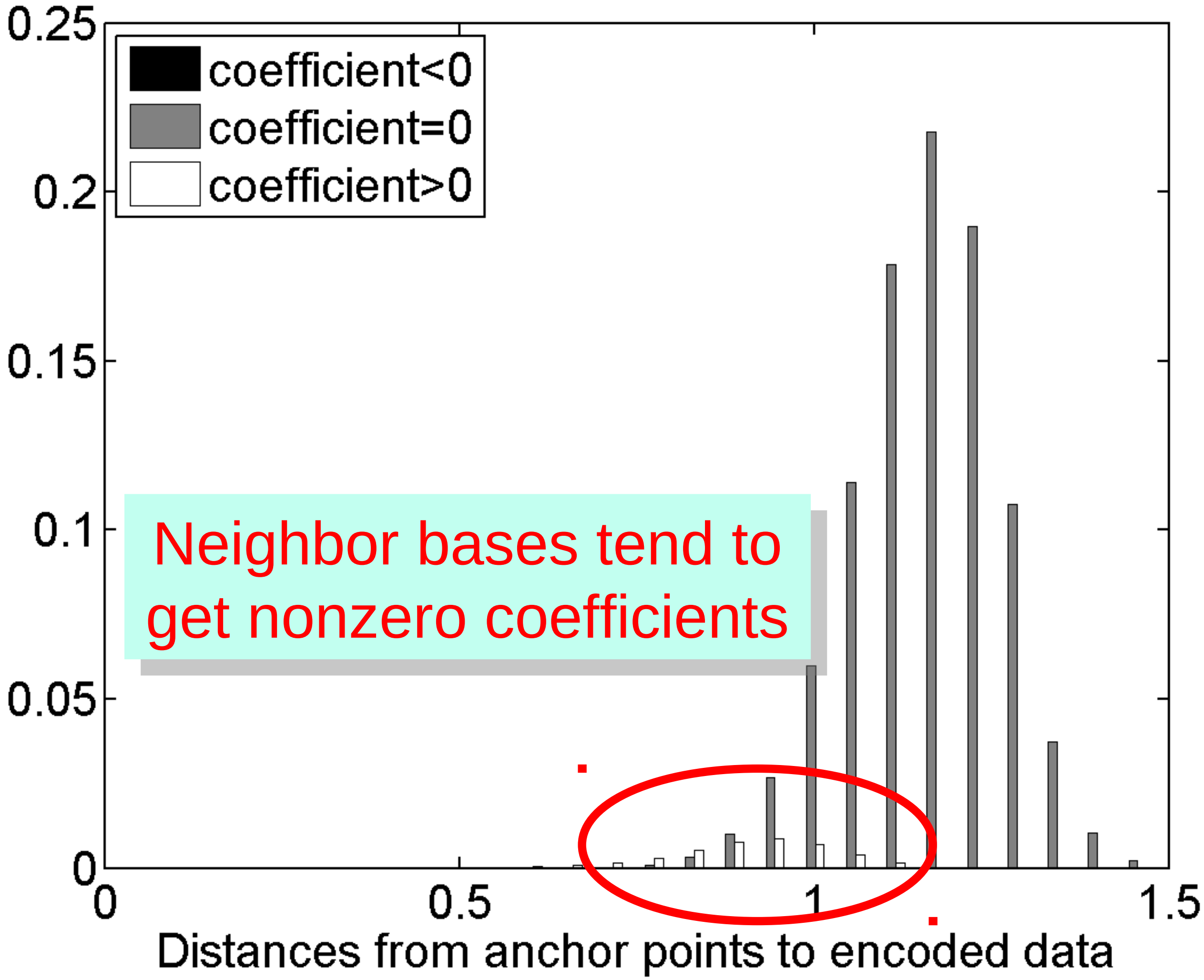
Error: 3.75%



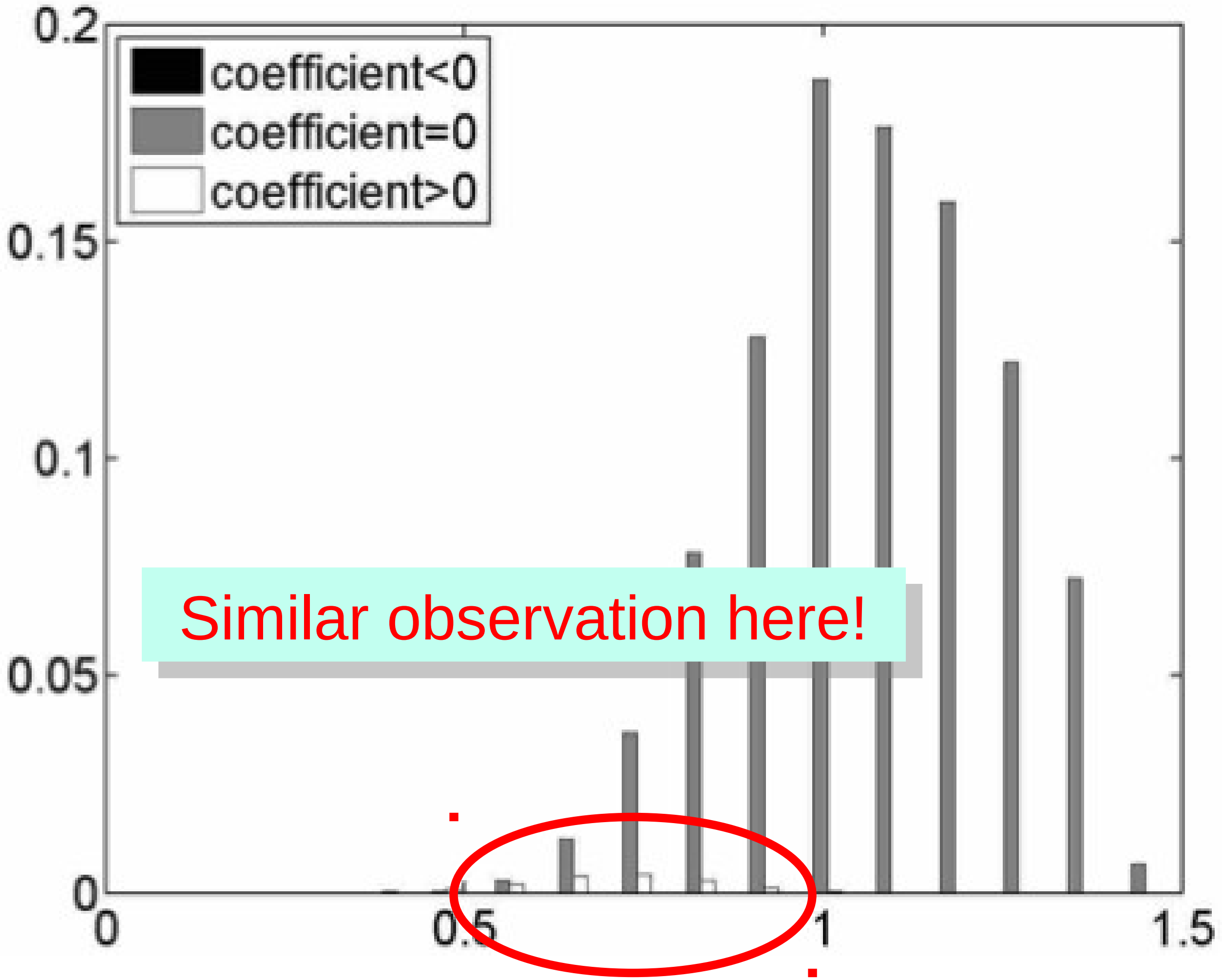
Error: 2.64%

- When sparse coding achieves the best classification accuracy, the learned bases are like digits – each basis has a clear local class association.

Distribution of coefficients (MNIST)



Distribution of coefficient (SIFT, Caltech101)



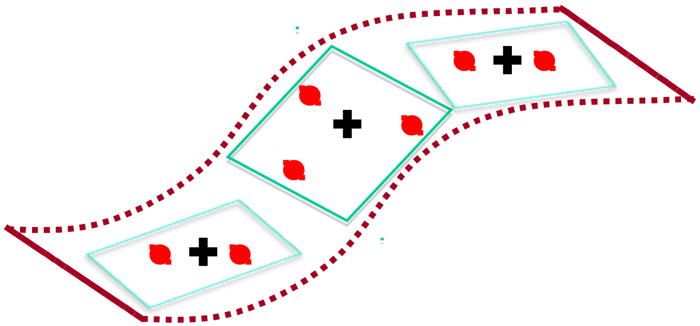
Outline

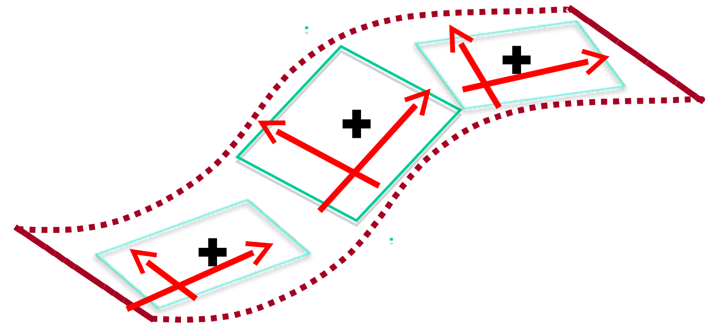
1. Sparse coding for image classification
2. Understanding sparse coding
   - Connections to RBMs, autoencoders, …
   - Sparse activations vs. sparse models, …
   - Sparsity vs. locality
   - Local sparse coding methods
3. Other topics: e.g. structured model, scale-up, discriminative training
4. Summary

Why develop local sparse coding methods

- Since locality is a preferred property in sparse coding, let's explicitly ensure the locality.
- The new algorithms can be well theoretically justified
- The new algorithms will have computational advantages over classical sparse coding

Two approaches to local sparse coding





Approach 1
Coding via local anchor points

Approach 2
Coding via local subspaces

Local coordinate coding

Learning locality-constrained linear coding for image classification, Jingjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang. In **CVPR 2010**.

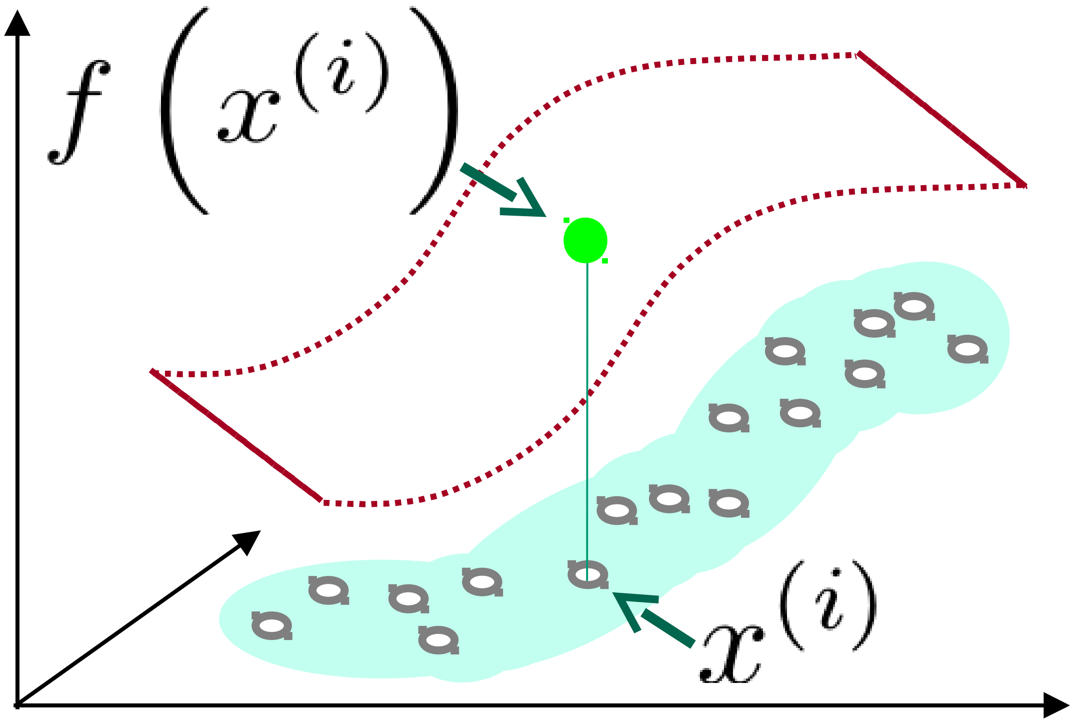
Nonlinear learning using local coordinate coding, Kai Yu, Tong Zhang, and Yihong Gong. In **NIPS 2009**.

Super-vector coding

Image Classification using Super-Vector Coding of Local Image Descriptors, Xi Zhou, Kai Yu, Tong Zhang, and Thomas Huang. In **ECCV 2010**.

Large-scale Image Classification: Fast Feature Extraction and SVM Training, Yuanqing Lin, Fengjun Lv, Shenghuo Zhu, Ming Yang, Timothee Cour, Kai Yu, LiangLiang Cao, Thomas Huang. In **CVPR 2011**

A function approximation framework to understand coding



- Assumption: image patches x follow a nonlinear manifold, and f(x) is smooth on the manifold.

- Coding: nonlinear mapping

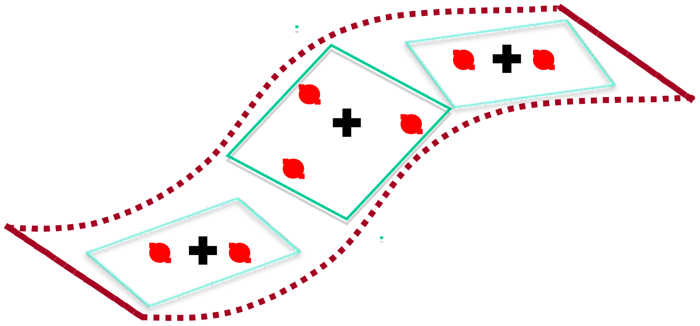
  x → a

  typically, a is high-dim & sparse

- Nonlinear Learning:

  f(x) = <w, a>

Local sparse coding

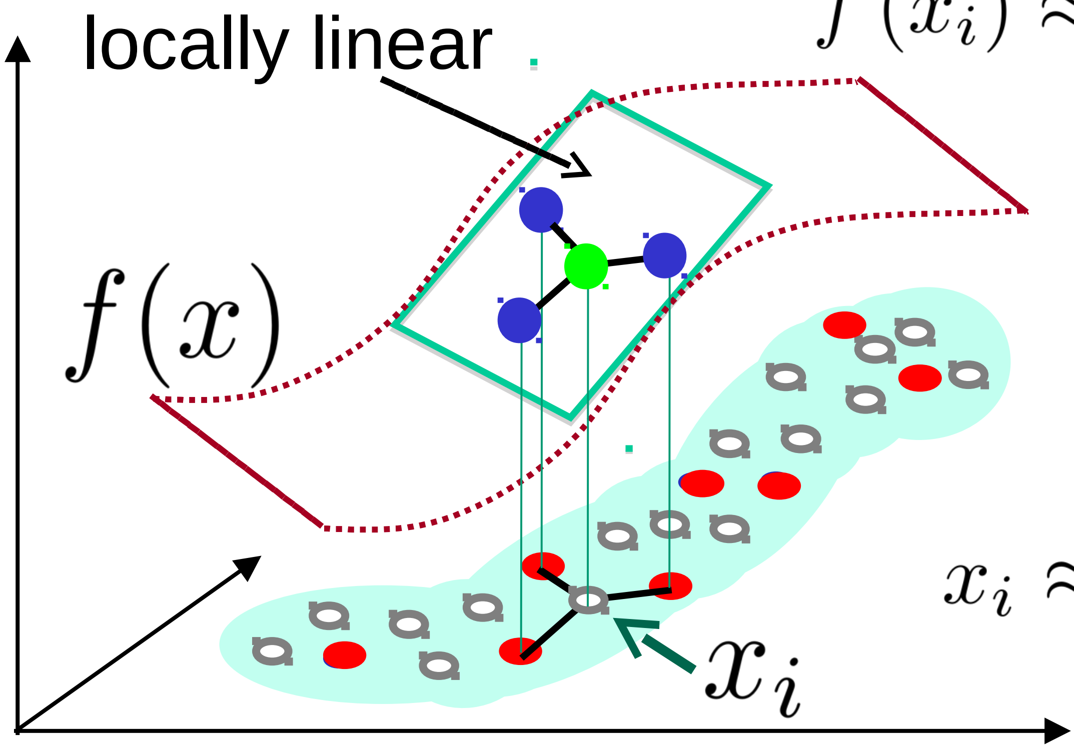


Approach 1
Local coordinate coding

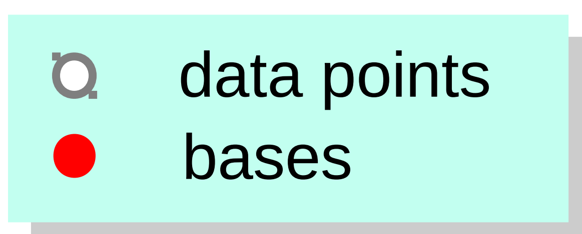
Function Interpolation based on LCC

Yu, Zhang, Gong, NIPS 10

$$f\left(x_i\right) \approx \sum_{j=1}^{k} a_{i,j} f(\phi_j)$$



$$x_i \approx \sum_{j=1}^{k} a_{i,j} \phi_j$$



Local Coordinate Coding (LCC): connect coding to nonlinear function learning

If f(x) is (alpha, beta) Lipschitz smooth

$$\left| f(x_i) - \sum_{j=1}^{k} \cdots \right. \quad \cdots_{i,j}| \|x_i - \phi_j\|^2$$

The key message:

A good coding scheme should
1. have a small coding error,
2. and also be sufficiently local

Function approximation error

Coding error

Locality term

Local Coordinate Coding (LCC)

Yu, Zhang & Gong, NIPS 09
Wang, Yang, Yu, Lv, Huang CVPR 10

- Dictionary Learning: k-means (or hierarchical k-means)

- Coding for x, to obtain its sparse representation *a*

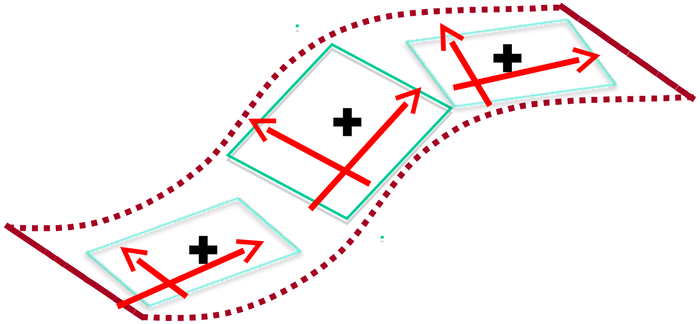
Step 1 – ensure locality: find the K nearest bases

$$[\phi_j]_{j \in J(x)}$$

Step 2 – ensure low coding error:

$$\min_a \left\| x - \sum_{j \in J(x)} a_{i,j} \phi_j \right\|^2, \quad \text{s.t.} \sum_{j \in J(x)} a_{i,j} = 1$$

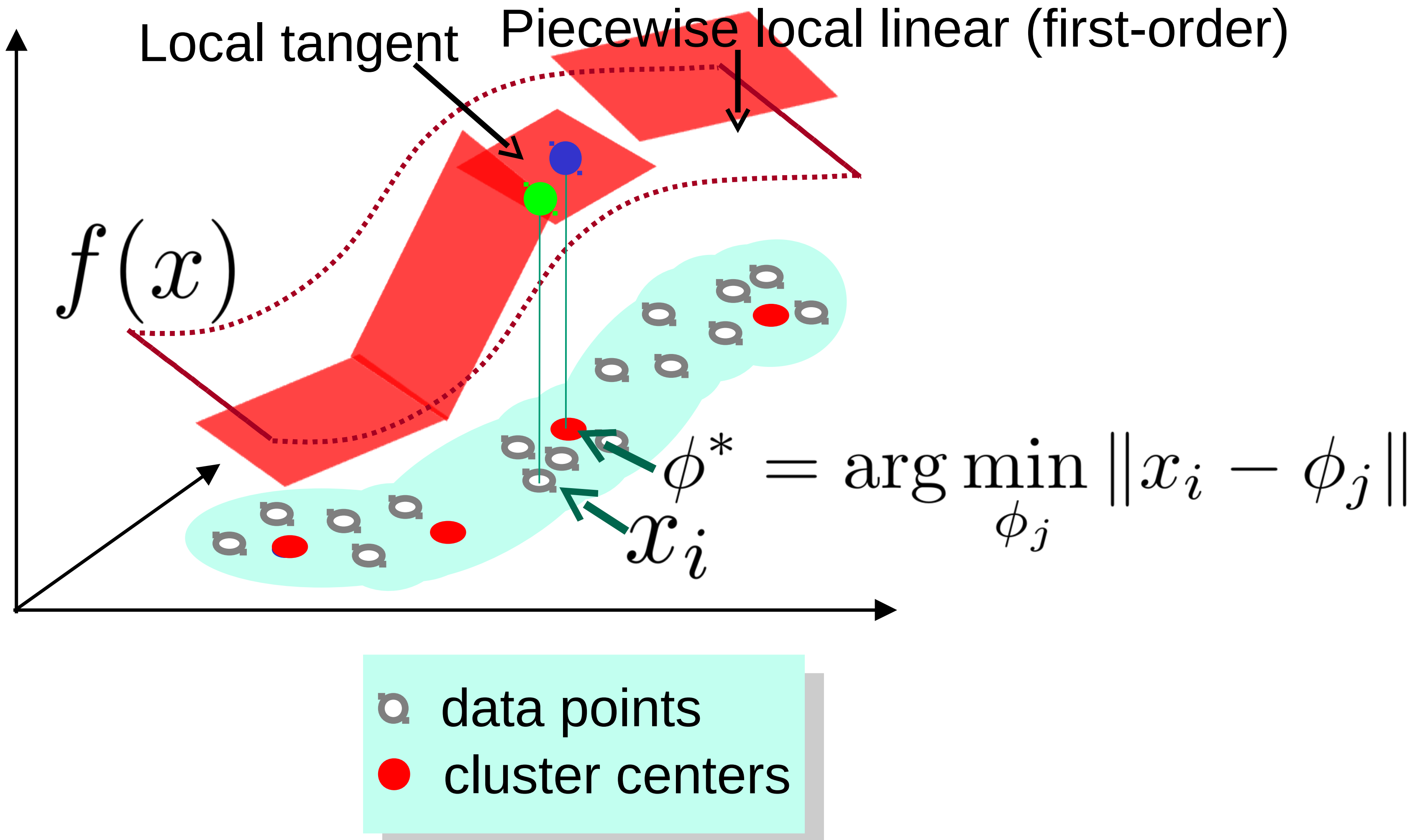
Local sparse coding



Approach 2
Super-vector coding

Function approximation via super-vector coding:

Zhou, Yu, Zhang, and Huang, ECCV 10



Super-vector coding: Justification

If f(x) is beta-Lipschitz smooth, and $\phi^* = \arg\min_{\phi_j} \|x - \phi_j\|$

Local tangent

$$\underbrace{\left| f(x) - f\left(\phi^*\right) - \nabla f\left(\phi^*\right)^\top \left(x - \phi^*\right) \right|}_{\text{Function approximation error}} \leq \frac{\beta}{2} \underbrace{\left\| x - \phi^* \right\|^2}_{\text{Quantization error}}$$

Super-Vector Coding (SVC)

Zhou, Yu, Zhang, and Huang, ECCV 10

- Dictionary Learning: k-means (or hierarchical k-means)

- Coding for x, to obtain its sparse representation *a*

Step 1 – find the nearest basis of x, obtain its VQ coding

e.g. [0, 0, 1, 0, …]

Step 2 – form super vector coding:

$$\text{e.g. } [\underbrace{0,\ 0,\ 1,\ 0,\ \ldots,}_{\text{Zero-order}}\ \underbrace{0,\ 0,\ (x-m_3)\ ,\ \ 0\ ,}_{\text{Local tangent}}\ \ldots\ ]$$

Results on ImageNet Challenge Dataset

ImageNet Challenge:
1.4 million images, 1000 classes

40%

VQ + Intersection Kernel

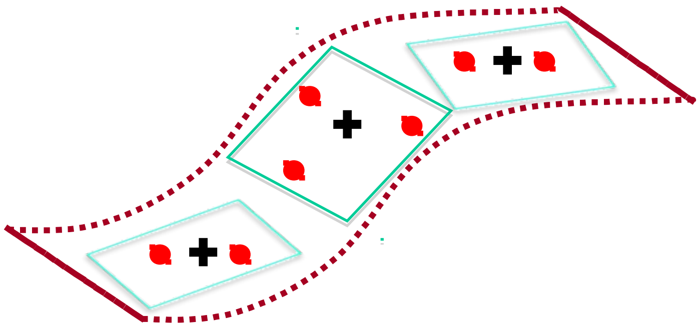
62%

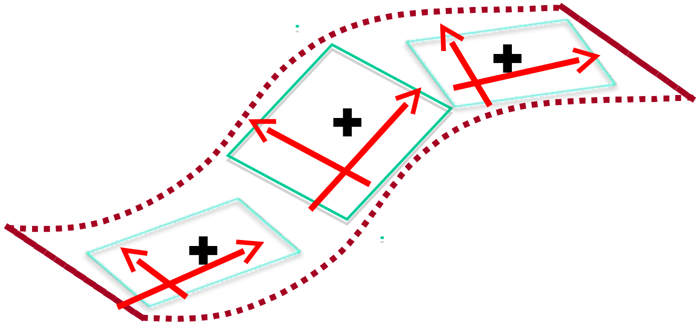
LCC + Linear SVM

65%

SVC + Linear SVM

Summary: local sparse coding





Approach 1
Local coordinate coding

Approach 2
Super-vector coding

- Sparsity achieved by explicitly ensuring locality
- Sound theoretical justifications
- Much simpler to implement and compute
- Strong empirical success

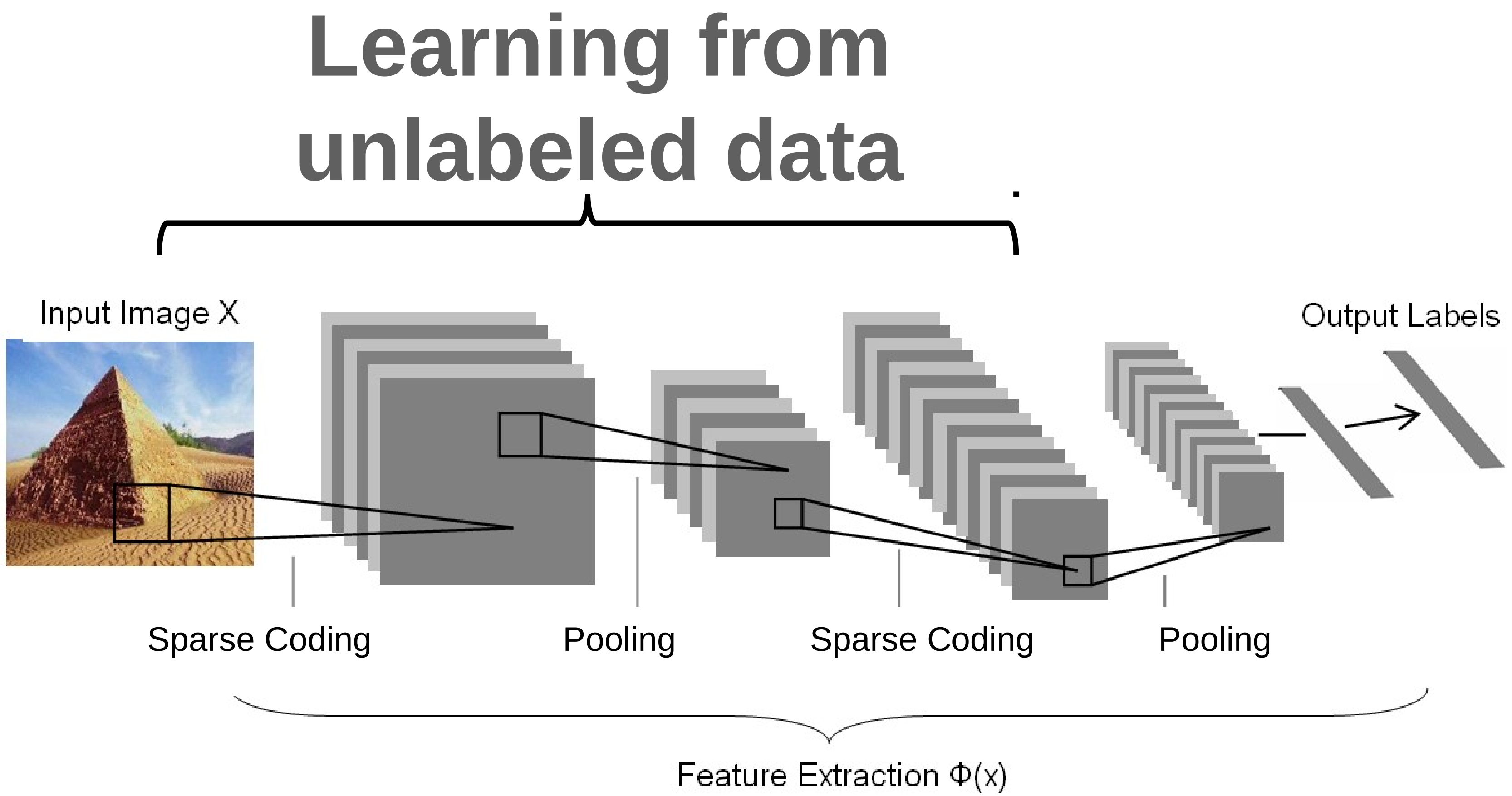
Outline

1. Sparse coding for image classification
2. Understanding sparse coding
3. Hierarchical sparse coding
4. Other topics: e.g. structured model, scale-up, discriminative training
5. Summary

Hierarchical sparse coding

Yu, Lin, & Lafferty, CVPR 11

Matthew D. Zeiler, Graham W. Taylor, and Rob Fergus, ICCV 11



A two-layer sparse coding formulation

Yu, Lin, & Lafferty, CVPR 11

$$(\widehat{W}, \widehat{\alpha}) = \arg\min_{W,\alpha} \quad L(W, \alpha) + \frac{\lambda_1}{n}\|W\|_1 + \gamma\|\alpha\|_1$$

$$\text{subject to} \quad \alpha \succeq 0,$$

$$L(W, \alpha) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{2}\left\|x_i - Bw_i\right\|^2 + \lambda_2 w_i^\top \Omega(\alpha) w_i\right\}$$

$$\Omega(\alpha) \equiv \left(\sum_{k=1}^{q} \alpha_k \operatorname{diag}(\varphi_k)\right)^{-1}$$

MNIST Results - classification

Yu, Lin, & Lafferty, CVPR 11

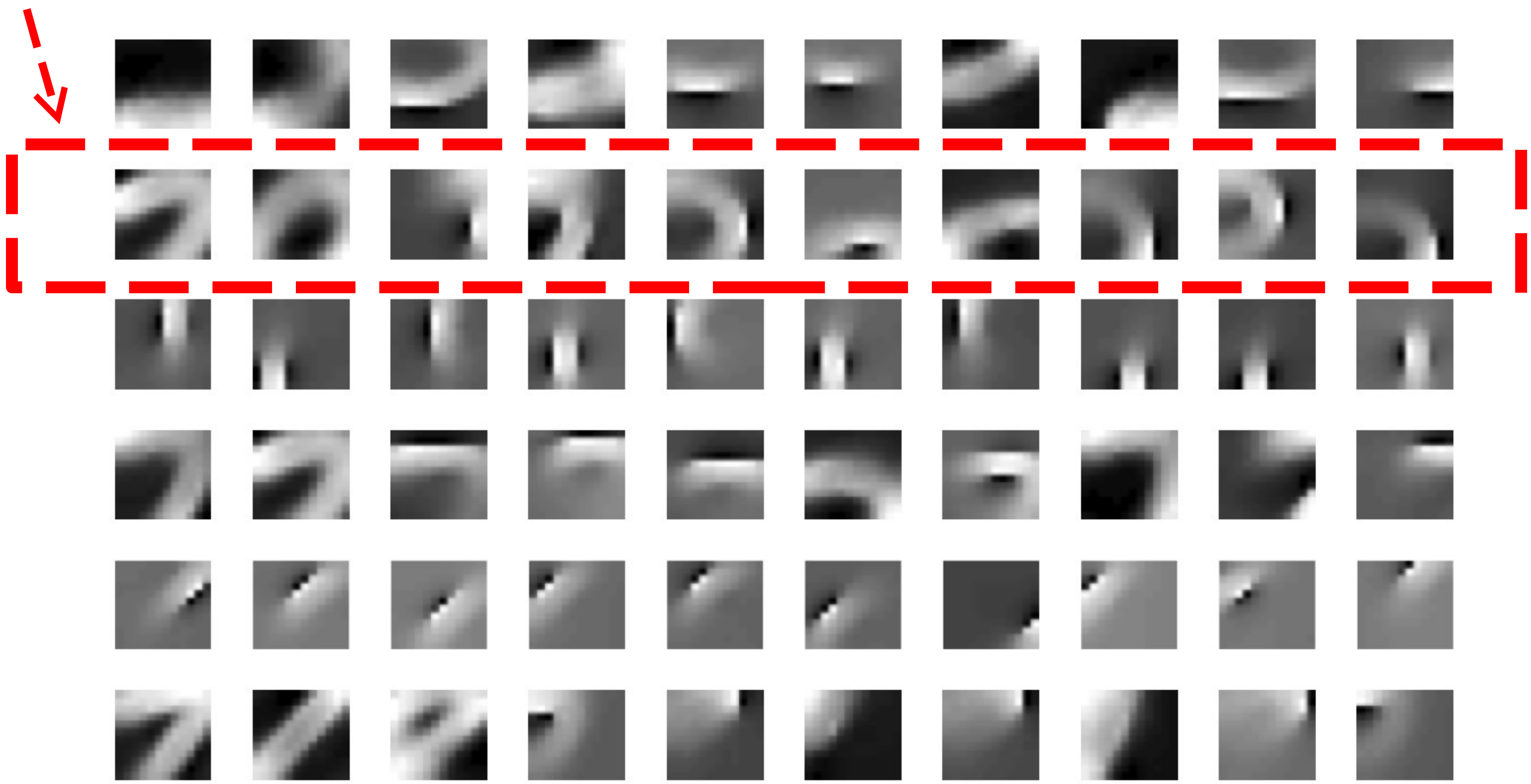
Methods	Error rate (%)
Sparse coding (unsupervised)	2.10
Local coordinate coding (unsupervised) [21]	1.90
Extended local coordinate coding (unsupervised) [21]	1.64
Differentiable sparse coding (supervised) [5]	1.30
Discriminative sparse coding (supervised) [15]	1.05
One-layer sparse coding (unsupervised)	0.98
Convolutional neural network (supervised) [11]	0.82
Hierarchical sparse coding (unsupervised)	**0.77**

- **HSC vs. CNN:** HSC provide even better performance than CNN ☺☺☺ more amazingly, HSC learns features in **unsupervised** manner!

MNIST results -- learned dictionary

Yu, Lin, & Lafferty, CVPR 11

A hidden unit in the second layer is connected to a unit group in the 1st layer: invariance to translation, rotation, and deformation



Caltech101 results - classification

Yu, Lin, & Lafferty, CVPR 11

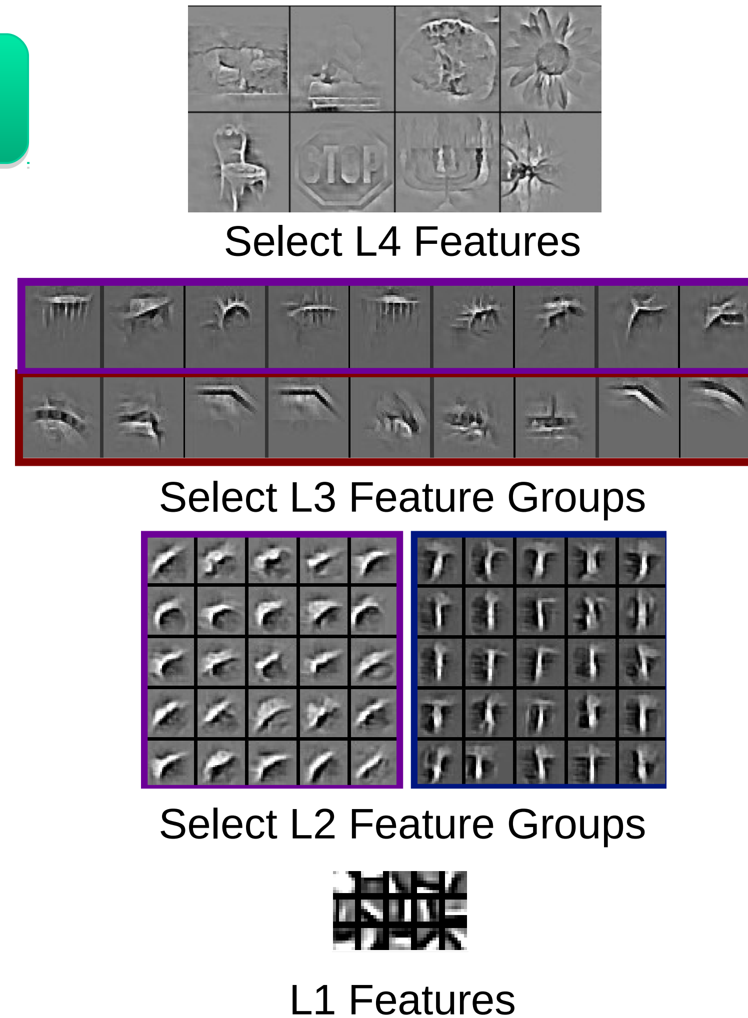
Methods	Accuracy (%)
VQ coding on SIFT (nonlinear SVM) [10]	64.4
Sparse coding on SIFT [20]	73.2
One-layer sparse coding on pixels [18]	46.6
One-layer convolution deep belief network on pixels [13]	60.5
Two-layer convolution deep belief network on pixels [13]	65.4
Two-layer convolutional neural network on pixels [9]	66.3
Hierarchical sparse coding on pixels - architecture I	**70.8**
Hierarchical sparse coding on pixels - architecture II	**74.0**

- **Learned descriptor:** performs slightly better than SIFT + SC

Adaptive Deconvolutional Networks for Mid and High Level Feature Learning

Matthew D. Zeiler, Graham W. Taylor, and Rob Fergus, ICCV 2011

- Hierarchical Convolutional Sparse Coding.
- Trained with respect to image from all layers (L1-L4).
- Pooling both spatially and amongst features.
- Learns invariant mid-level features.



Outline

1. Sparse coding for image classification
2. Understanding sparse coding
3. Hierarchical sparse coding
4. Other topics: e.g. structured model, scale-up, discriminative training
5. Summary

Other topics of sparse coding

- Structured sparse coding, for example
  - Group sparse coding [Bengio et al, NIPS 09]
  - Learning hierarchical dictionary [Jenatton, Mairal et al, 2010]

- Scale-up sparse coding, for example
  - Feature-sign algorithm [Lee et al, NIPS 07]
  - Feed-forward approximation [Gregor & LeCun, ICML 10]
  - Online dictionary learning [Mairal et al, ICML 2009]

- Discriminative training, for example
  - Backprop algorithms [Bradley & Jbagnell, NIPS 08; Yang et al. CVPR 10]
  - Supervised dictionary training [Mairal et al, NIPS08]

Summary of Sparse Coding

- Sparse coding is an effect way for (unsupervised) feature learning
- A building block for deep models
- Sparse coding and its local variants (LCC, SVC) have pushed the boundary of accuracies on Caltech101, PASCAL VOC, ImageNet, …
- Challenge: discriminative training is not straightforward