A probabilistic model for recursive factorized image features.

Sergey Karayev
Mario Fritz
Sanja Fidler
Trevor Darrell

Outline

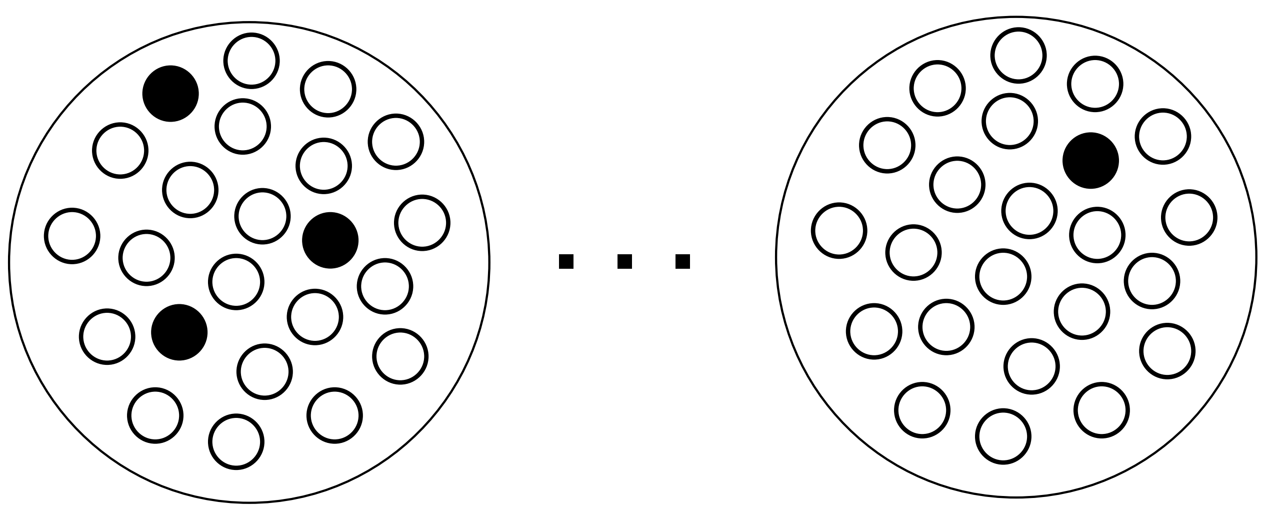
- Motivation:
  - ✳ Distributed coding of local image features
  - ✳ Hierarchical models
  - ✳ Bayesian inference
- Our model: Recursive LDA
- Evaluation

Local Features

- Gradient energy histograms by orientation and grid position in local patches.
- *Coded* and used in bag-of-words or spatial model classifiers.

Feature Coding

- Traditionally vector quantized as *visual words*.
- But coding as mixture of components, such as in sparse coding, is empirically better. (Yang et al. 2009)

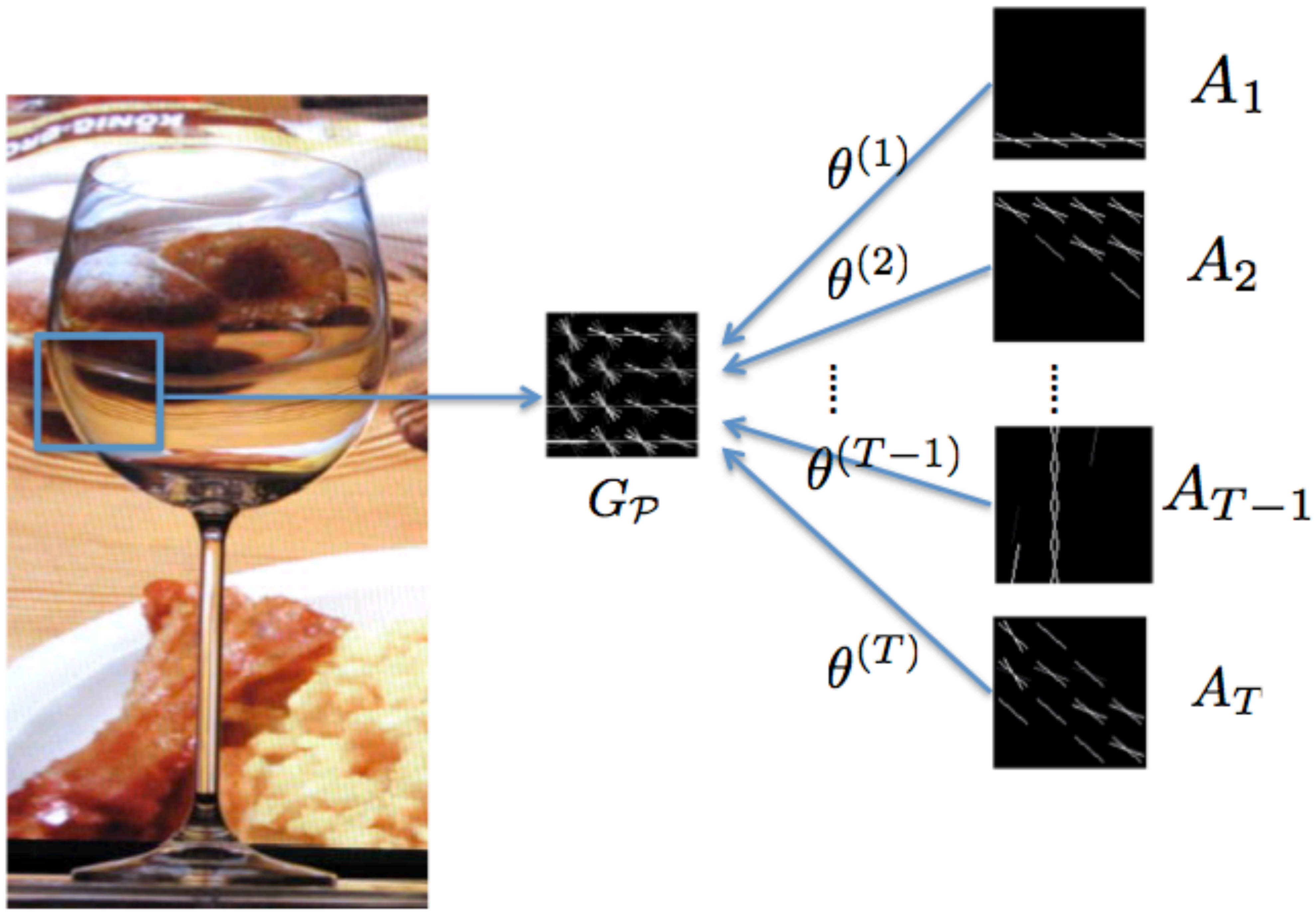


+ Decent combinatorial capacity ($\sim N^K$)

- Low combinatorial capacity (N)

picture from Bruno Olshausen

Another motivation: additive image formation



Fritz et al. An Additive Latent Feature Model for Transparent Object Recognition. NIPS (2009).

Outline

- Motivation:
  - ✶ Distributed coding of local image features
  - ✶ Hierarchical models
  - ✶ Bayesian inference
- Our model: Recursive LDA
- Evaluation

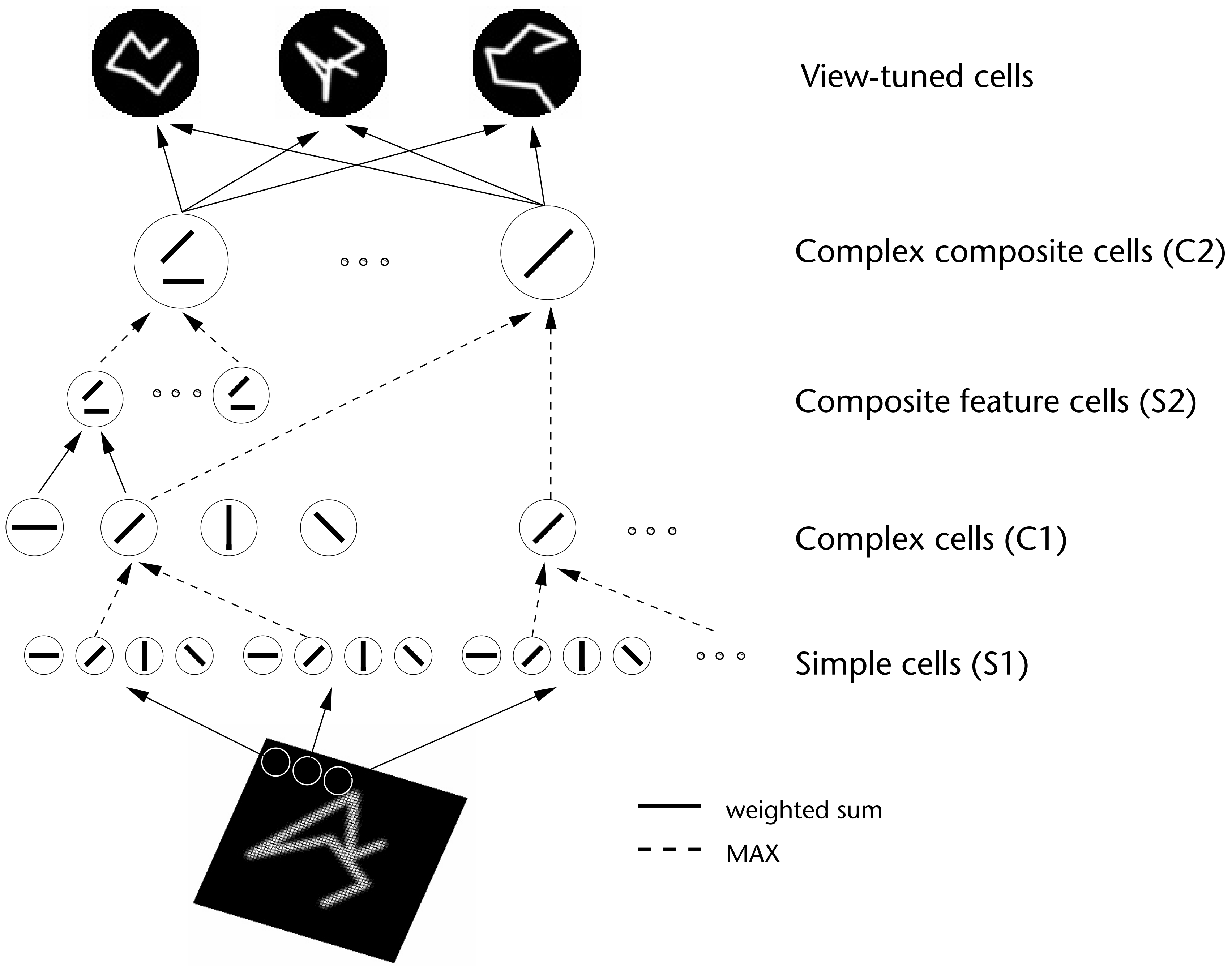
Hierarchies

- Biological evidence for increasing spatial support and complexity of visual pathway.
- Local features not robust to ambiguities. Higher layers can help resolve.
- Efficient parametrization possible due to sharing of lower-layer components.

Past Work

- HMAX models (Riesenhuber and Poggio 1999, Mutch and Lowe 2008)
- Convolutional networks (Ranzato et al. 2007, Ahmed et al. 2009)
- Deep Belief Nets (Hinton 2007, Lee et al. 2009)
- Hyperfeatures (Agarwal and Triggs 2008)
- Fragment-based hierarchies (Ullman 2007)
- Stochastic grammars (Zhu and Mumford 2006)
- Compositional object representations (Fidler and Leonardis 2007, Zhu et al. 2008)

HMAX



Riesenhuber and Poggio. Hierarchical models of object recognition in cortex. Nature Neuroscience (1999).

Convolutional Deep Belief Nets

Stacked layers, each consisting of feature extraction, transformation, and pooling.

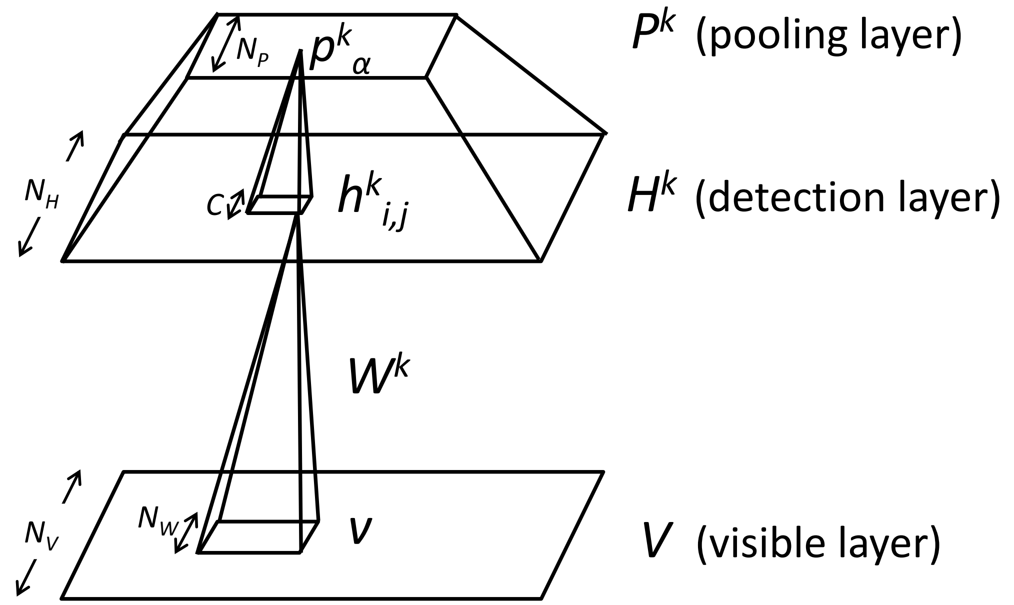
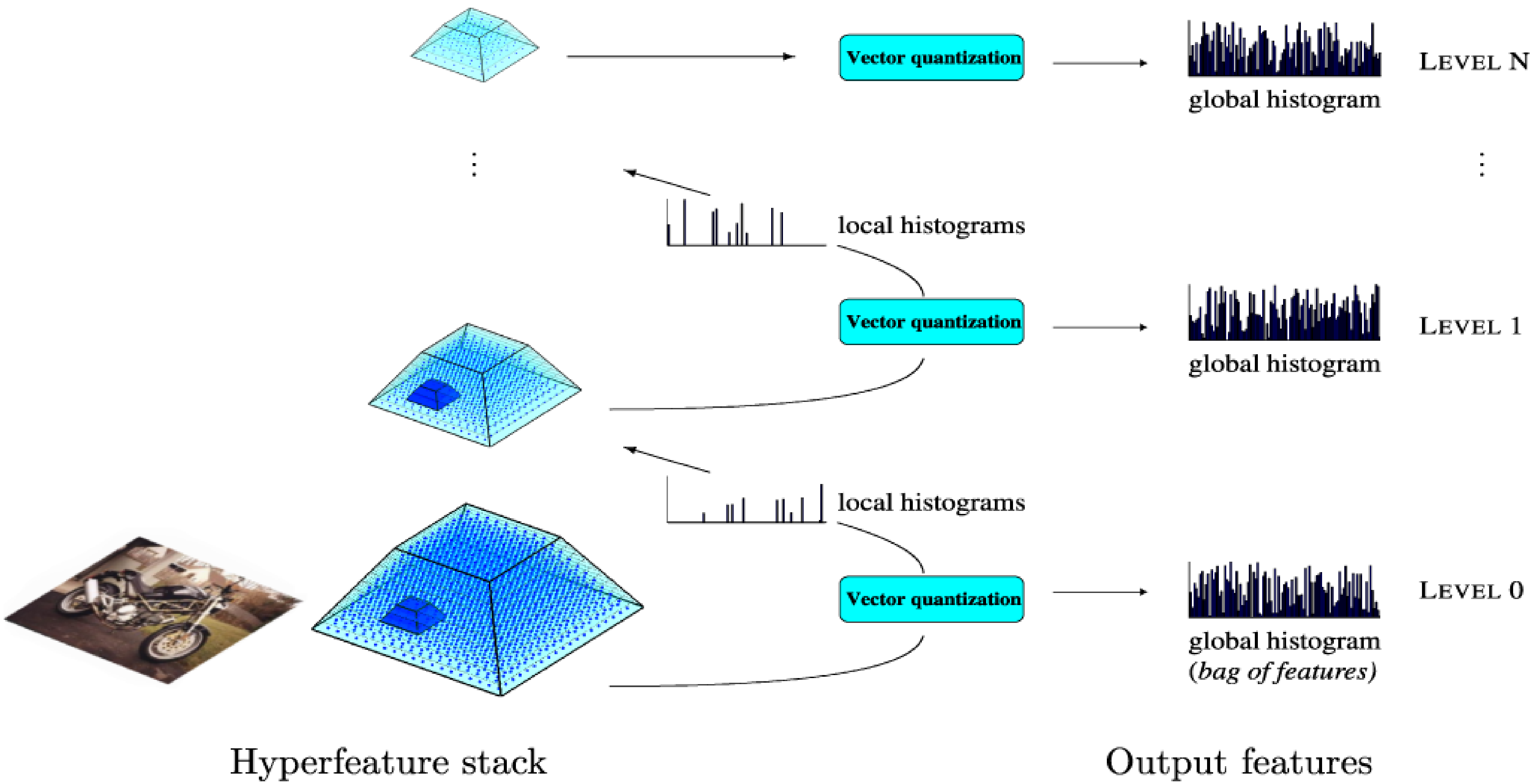


Figure 1. Convolutional RBM with probabilistic max-pooling. For simplicity, only group k of the detection layer and the pooing layer are shown. The basic CRBM corresponds to a simplified structure with only visible layer and detection (hidden) layer. See text for details.

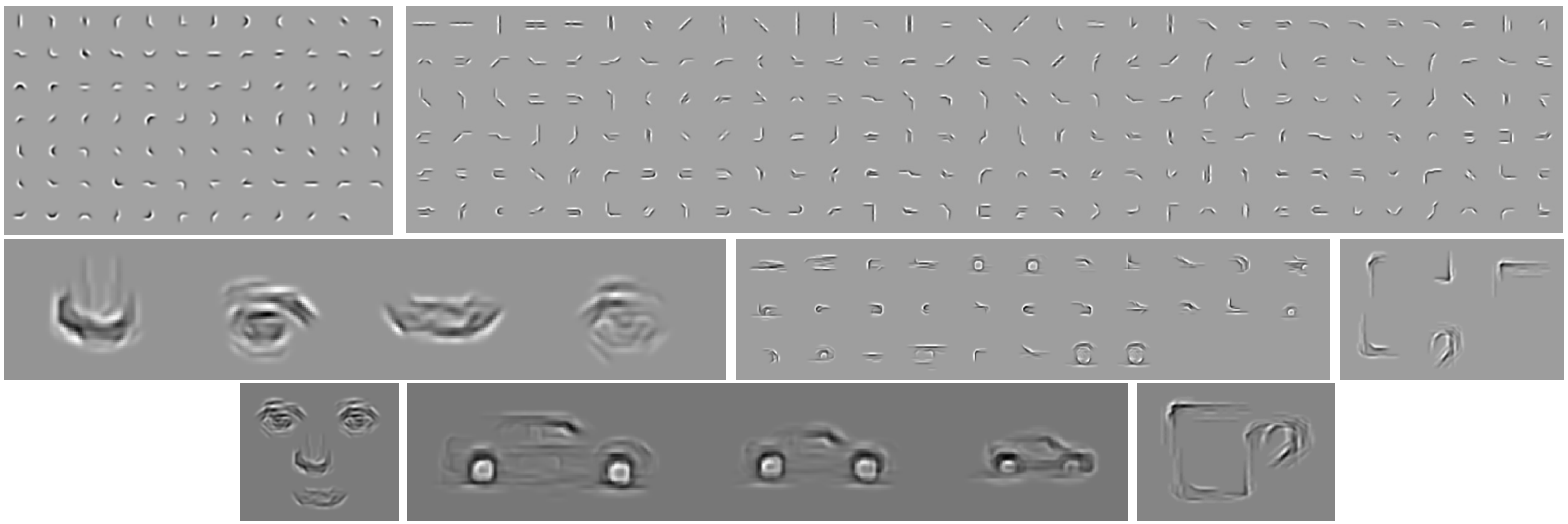
Lee et al. Convolutional deep belief networks for scalable unsupervised learning of hierarchical ….
Proceedings of the 26th Annual International Conference on Machine Learning (2009)

Hyperfeatures



Agarwal and Triggs. Multilevel Image Coding with Hyperfeatures. International Journal of Computer Vision (2008).

Compositional Representations



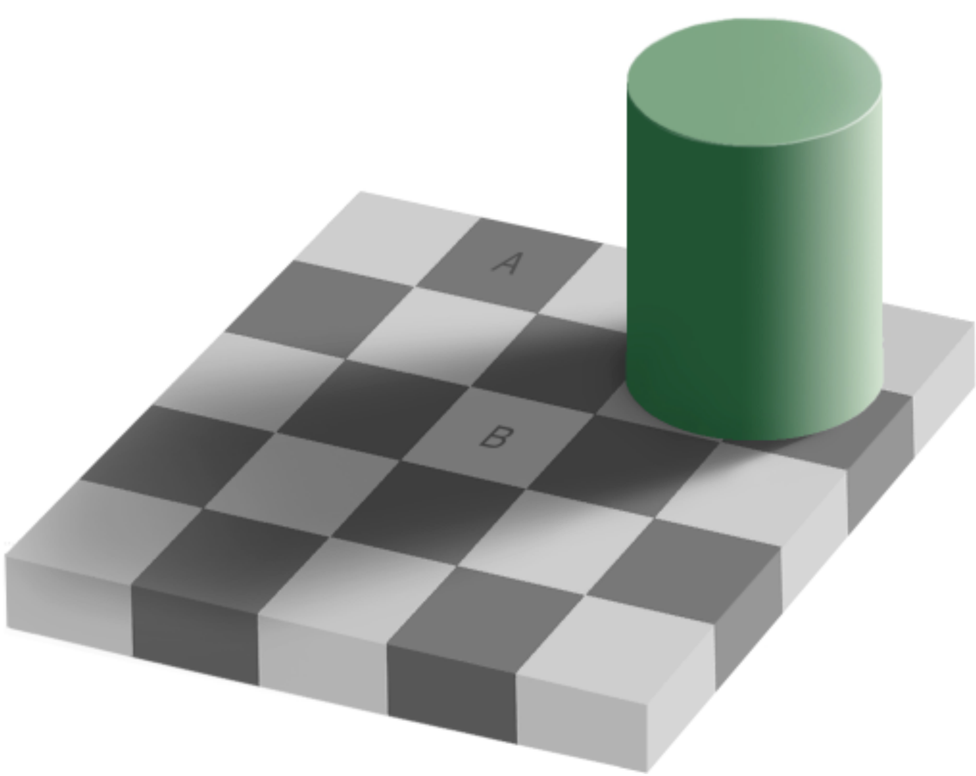
Fidler and Leonardis. Towards Scalable Representations of Object Categories: Learning a Hierarchy of Parts. CVPR (2007)

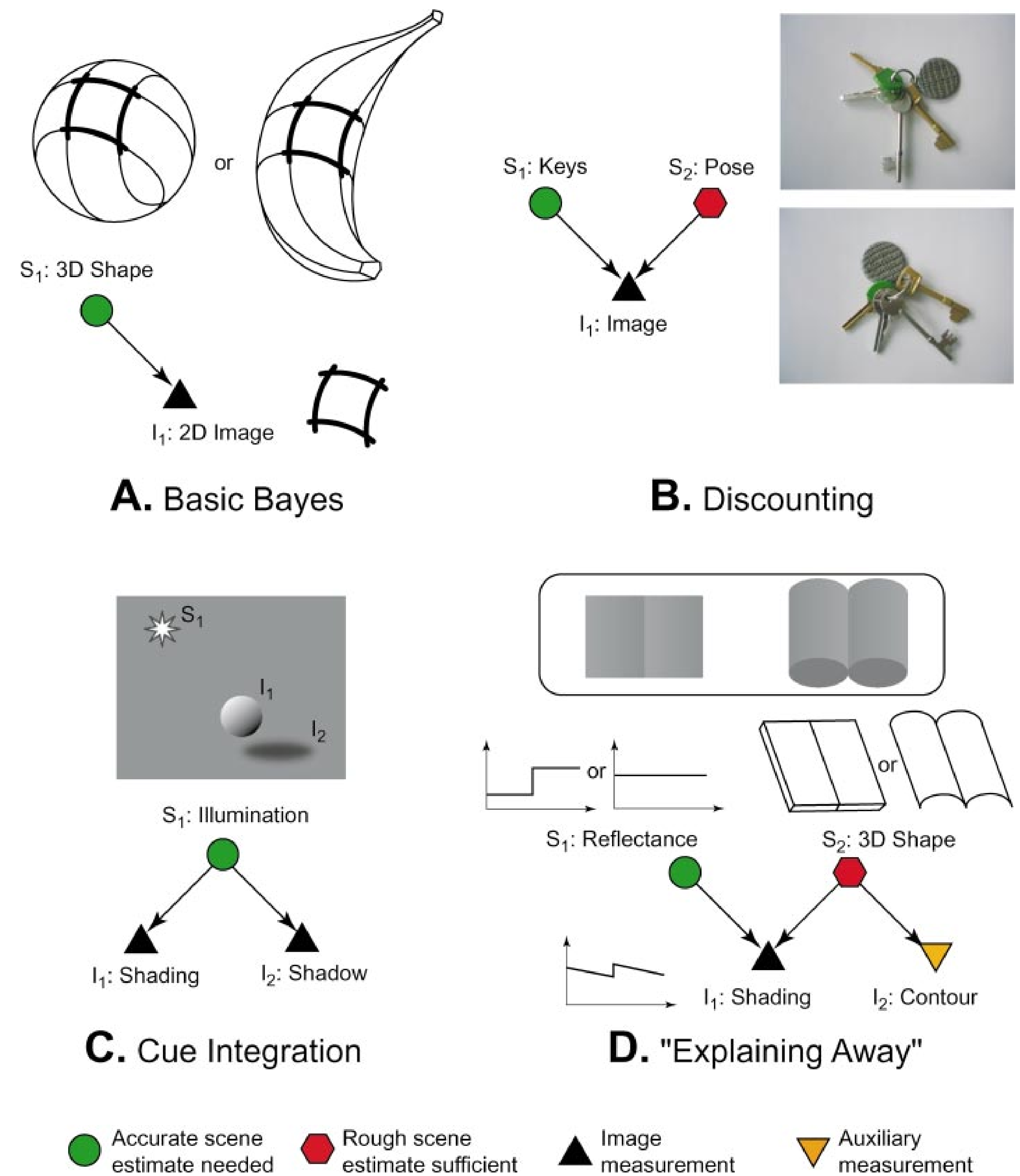
Outline

- Motivation:
  - ✶Distributed coding of local image features
  - ✶Hierarchical models
  - ✶Bayesian inference
- Our model: Recursive LDA
- Evaluation

Bayesian inference

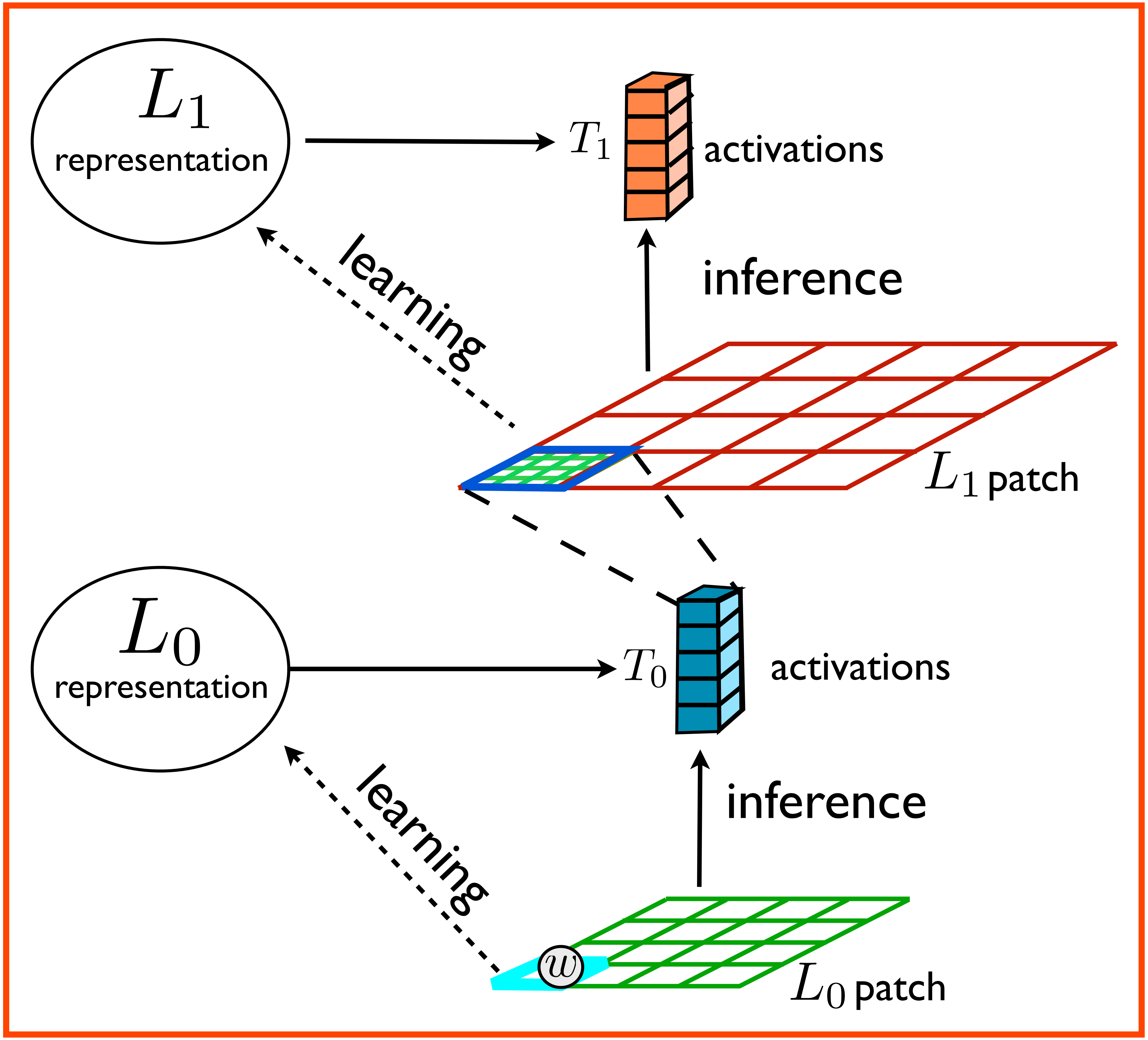
- The human visual cortex deals with inherently ambiguous data.
- Role of priors and inference (Lee and Mumford 2003).



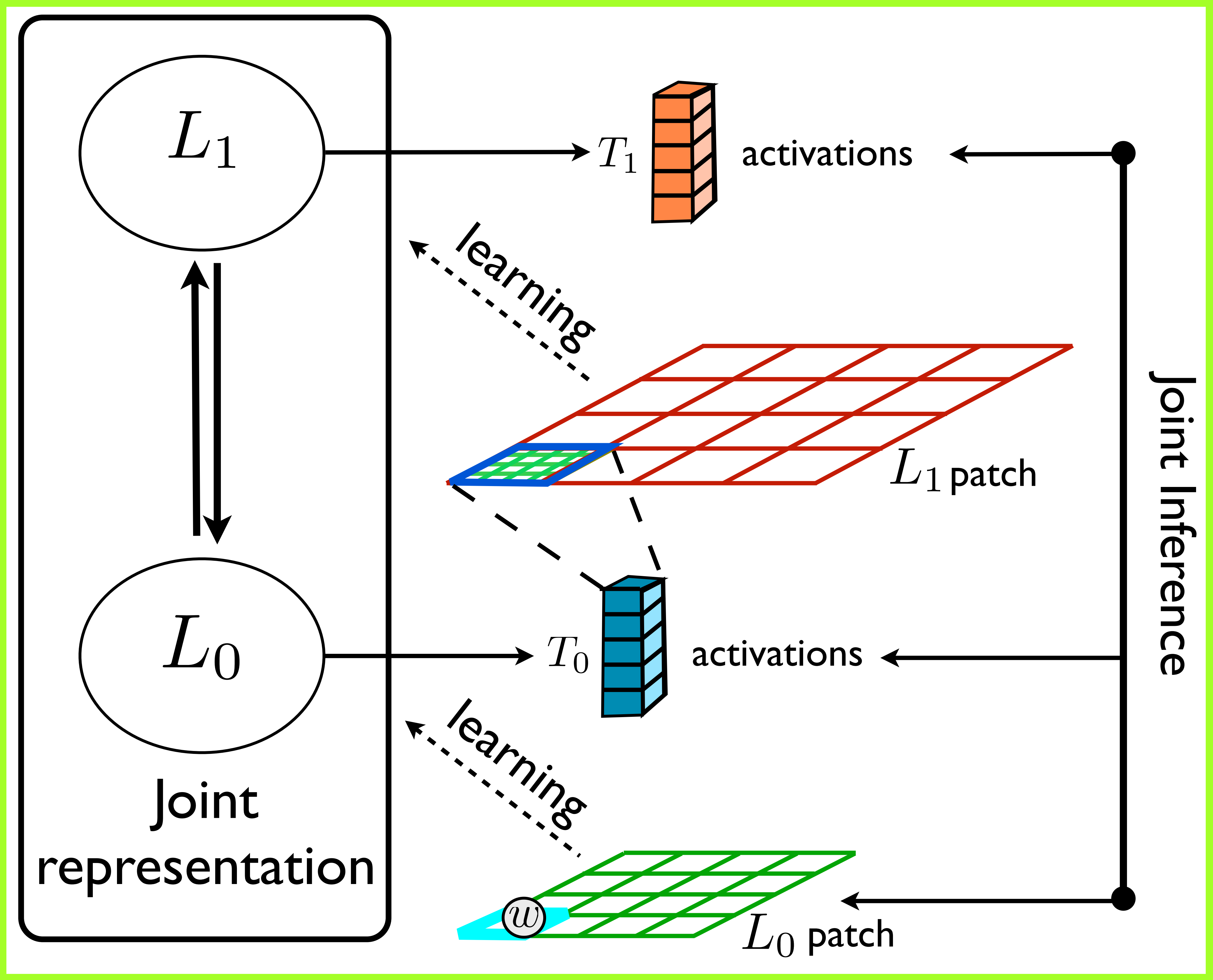


Kersten et al. Object perception as Bayesian inference. Annual Reviews (2004)

- But most hierarchical approaches do both learning and inference only from the bottom-up.



L_1
representation
T_1
activations
learning
inference
L_1 patch
L_0
representation
T_0
activations
learning
inference
w
L_0 patch



L_1
L_0
Joint representation
T_1
activations
T_0
activations
learning
learning
L_1 patch
L_0 patch
w
Joint Inference

What we would like

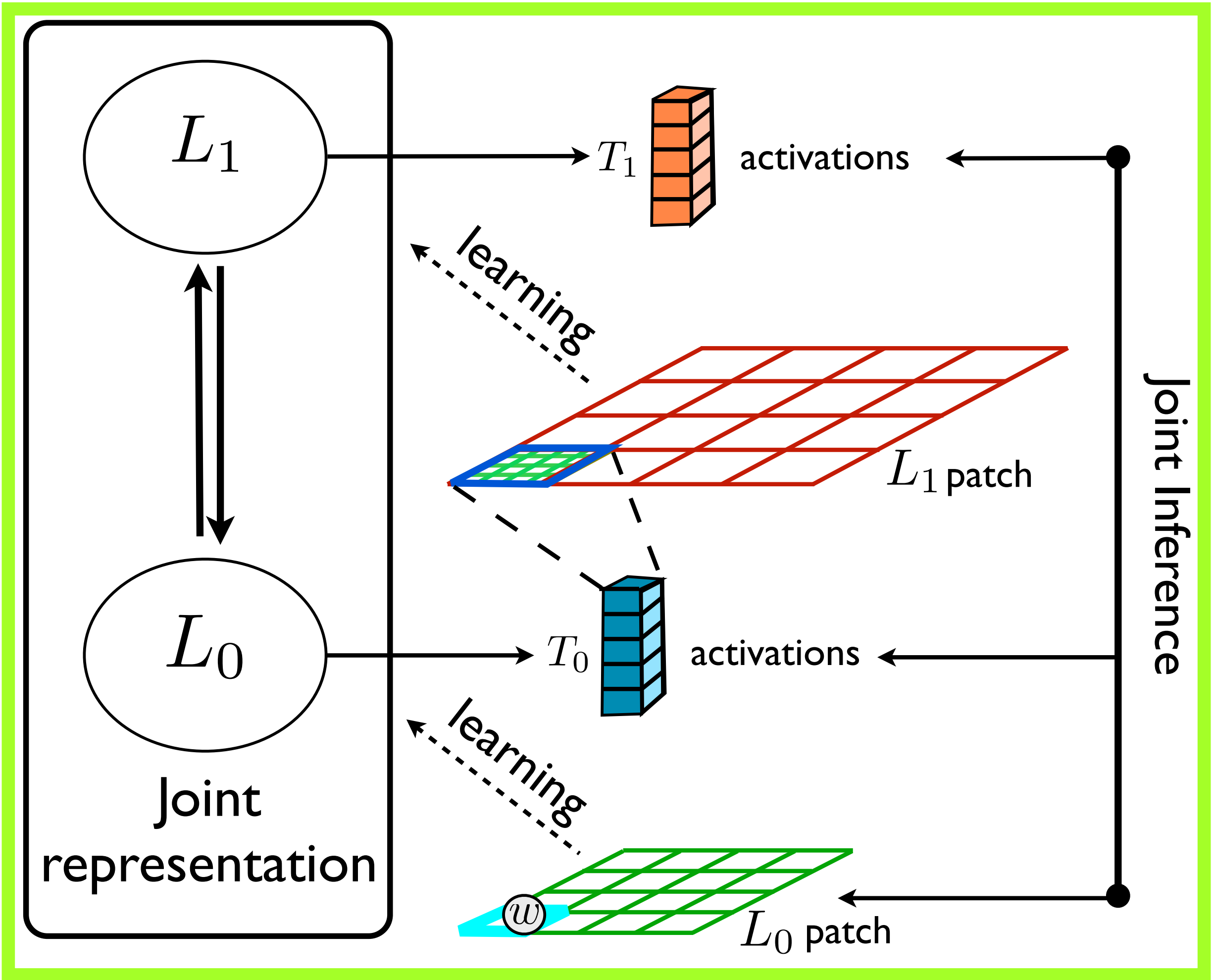
- Distributed coding of local features in a hierarchical model that would allow full inference.

Outline

- Motivation:
  - ✶ Distributed coding of local image features
  - ✶ Hierarchical models
  - ✶ Bayesian inference
- Our model: Recursive LDA
- Evaluation

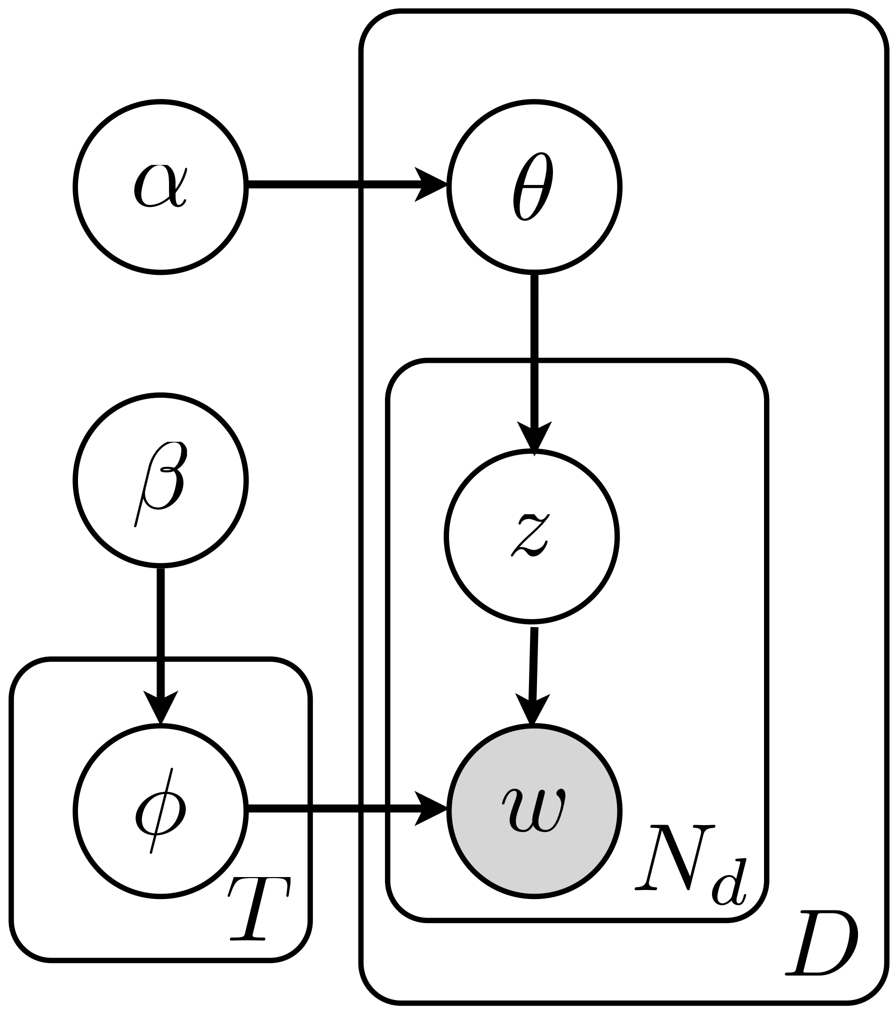
Our model: rLDA

- Based on Latent Dirichlet Allocation (LDA).
- Multiple layers, with increasing spatial support.
- Learns representation jointly across layers.



Latent Dirichlet Allocation

- Bayesian multinomial mixture model originally formulated for text analysis.



Blei et al. Latent Dirichlet allocation. Journal of Machine Learning Research (2003)

Latent Dirichlet Allocation

Corpus-wide, the multinomial distributions of words (topics) are sampled:

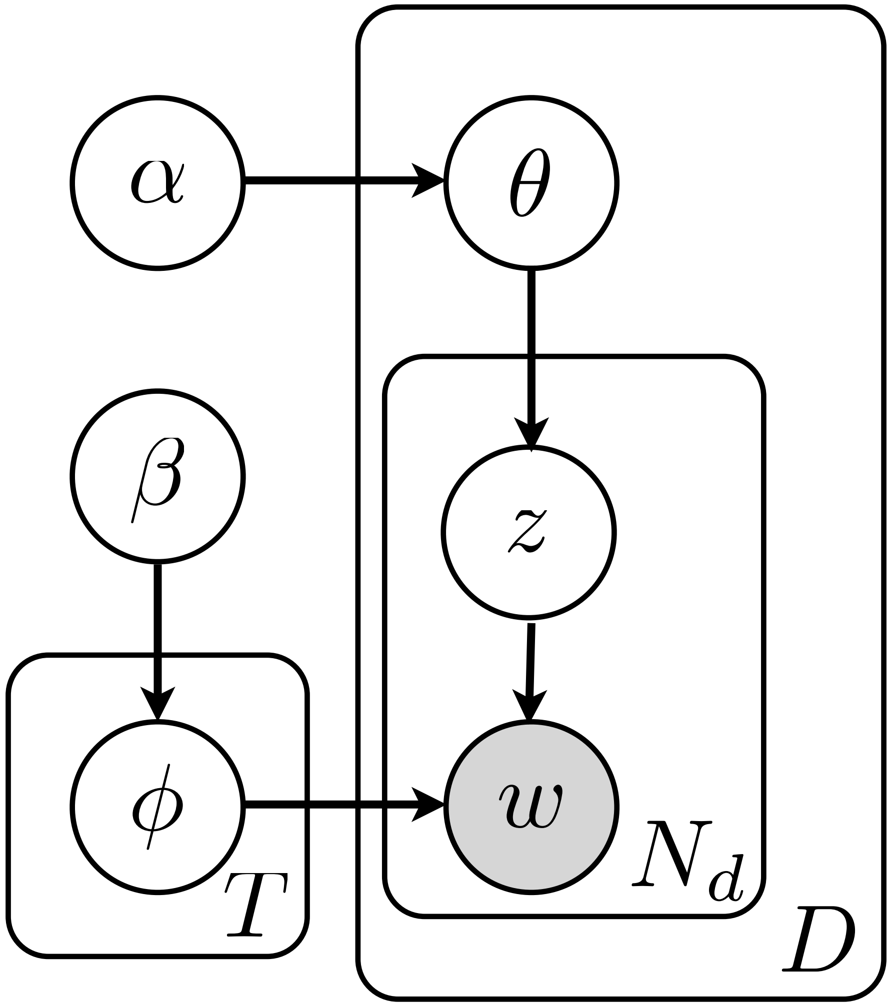
- $\phi \sim Dir(\beta)$

For each document, $d \in 1, \ldots, D$, mixing proportions $\theta^{(d)}$ are sampled according to:

- $\theta^{(d)} \sim Dir(\alpha)$

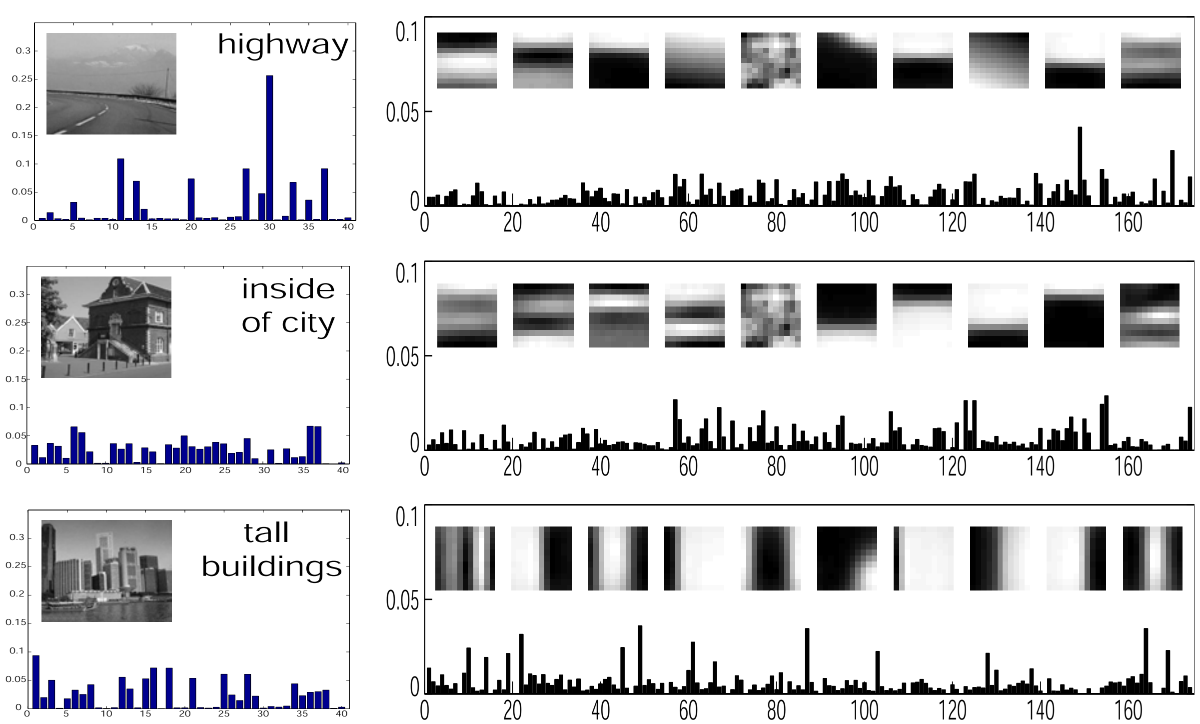
And N_d words w are sampled according to:

- $z \sim Mult(\theta^{(d)})$: sample topic given the document-topic mixing proportions
- $w \sim Mult(\phi^{(z)}$: sample word given the topic and the topic-word multinomials



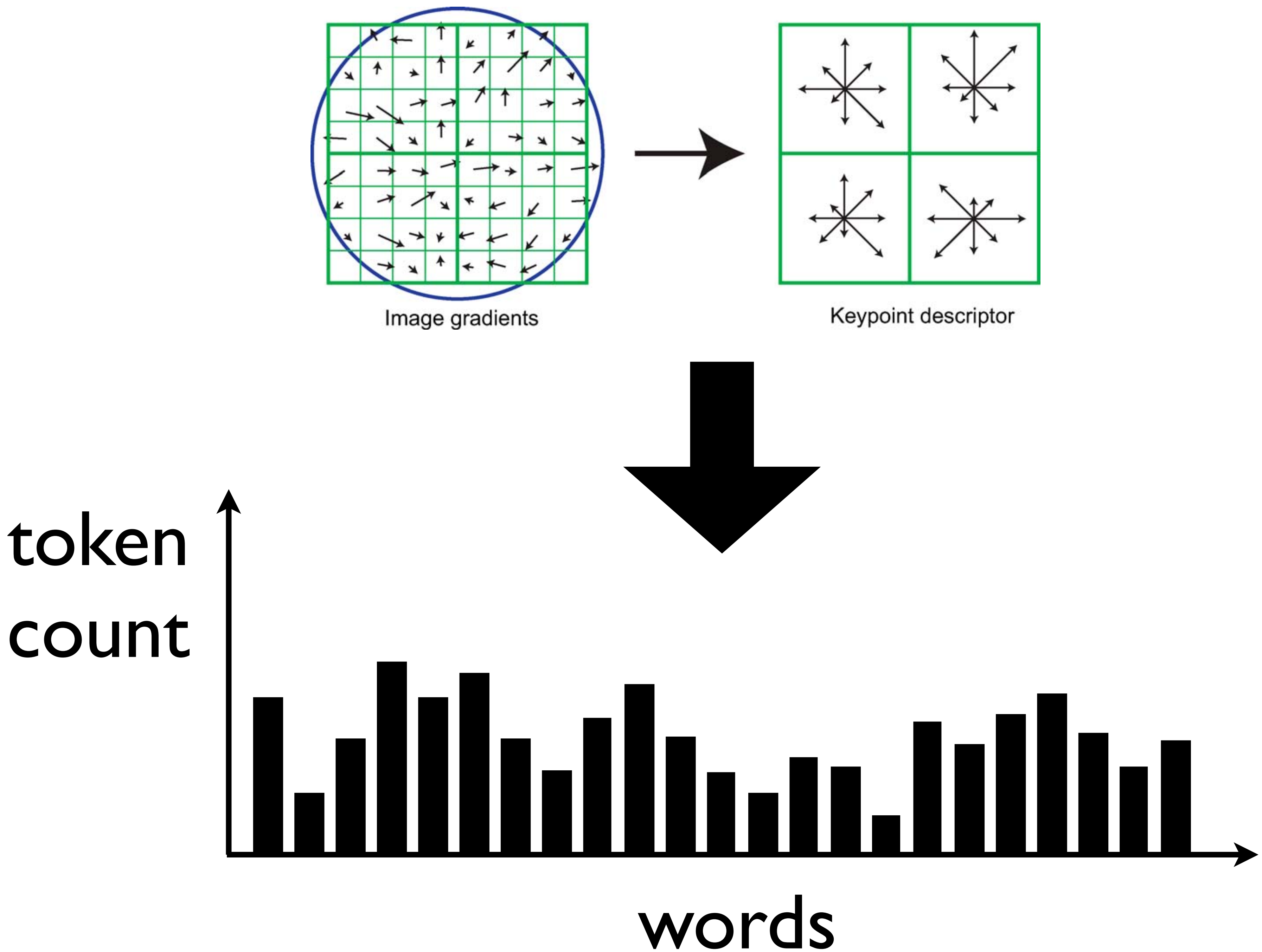
LDA in vision

- Past work has applied LDA to *visual words*, with topics being distributions over them.



Fei-Fei and Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories. CVPR (2005)

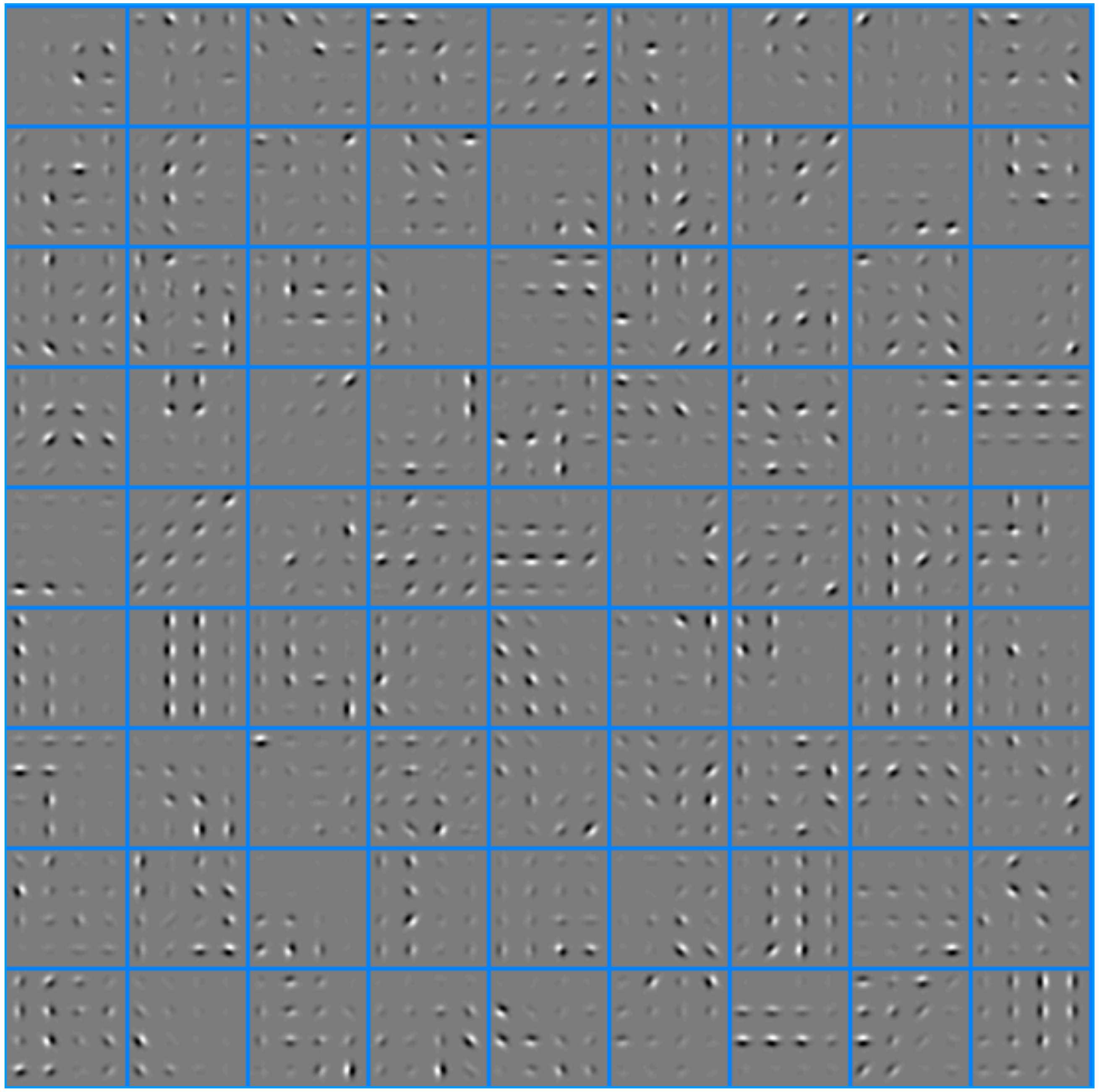
LDA-SIFT



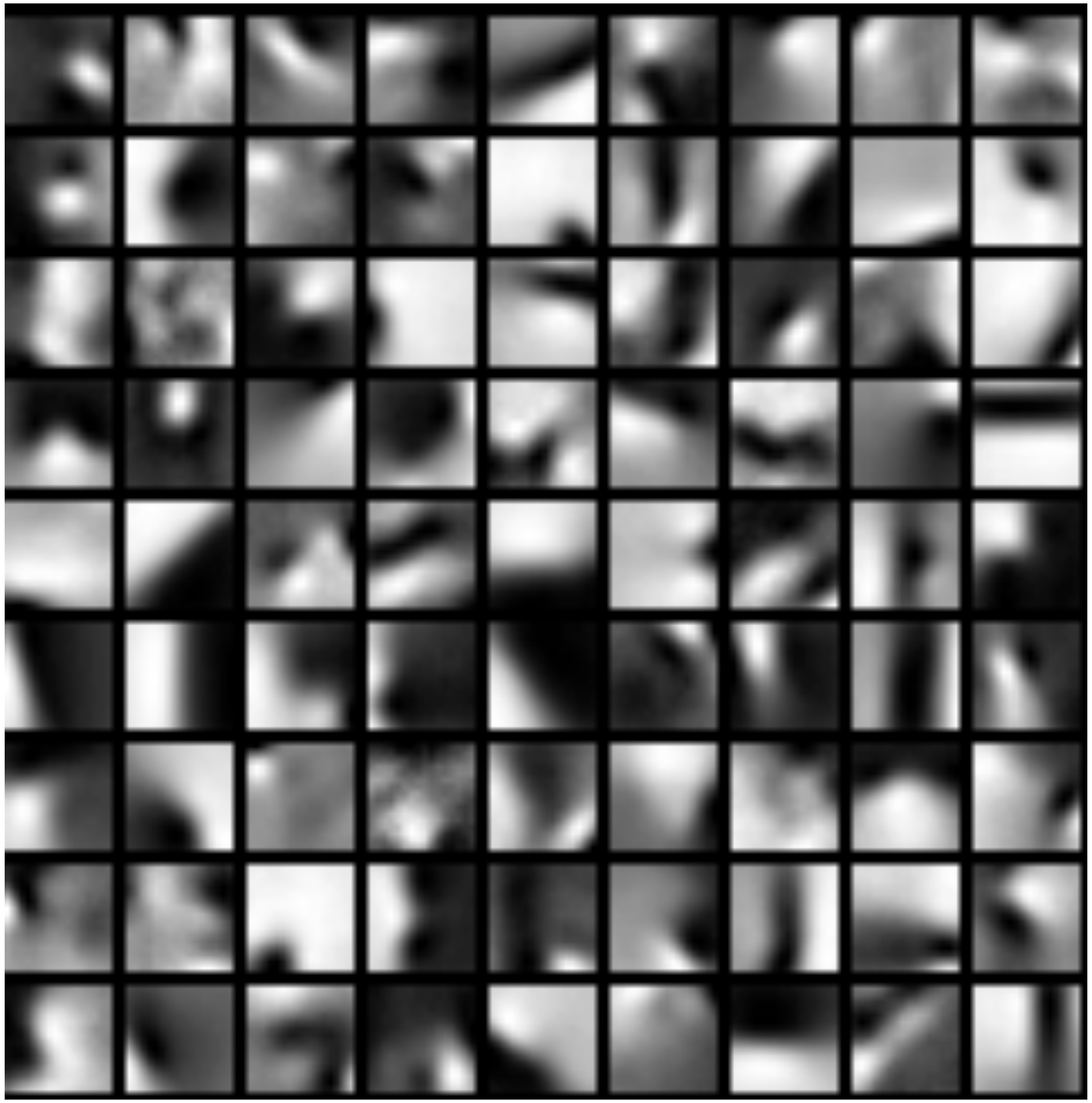
How training works

- (quantization, extracting patches, inference illustration)

Topics



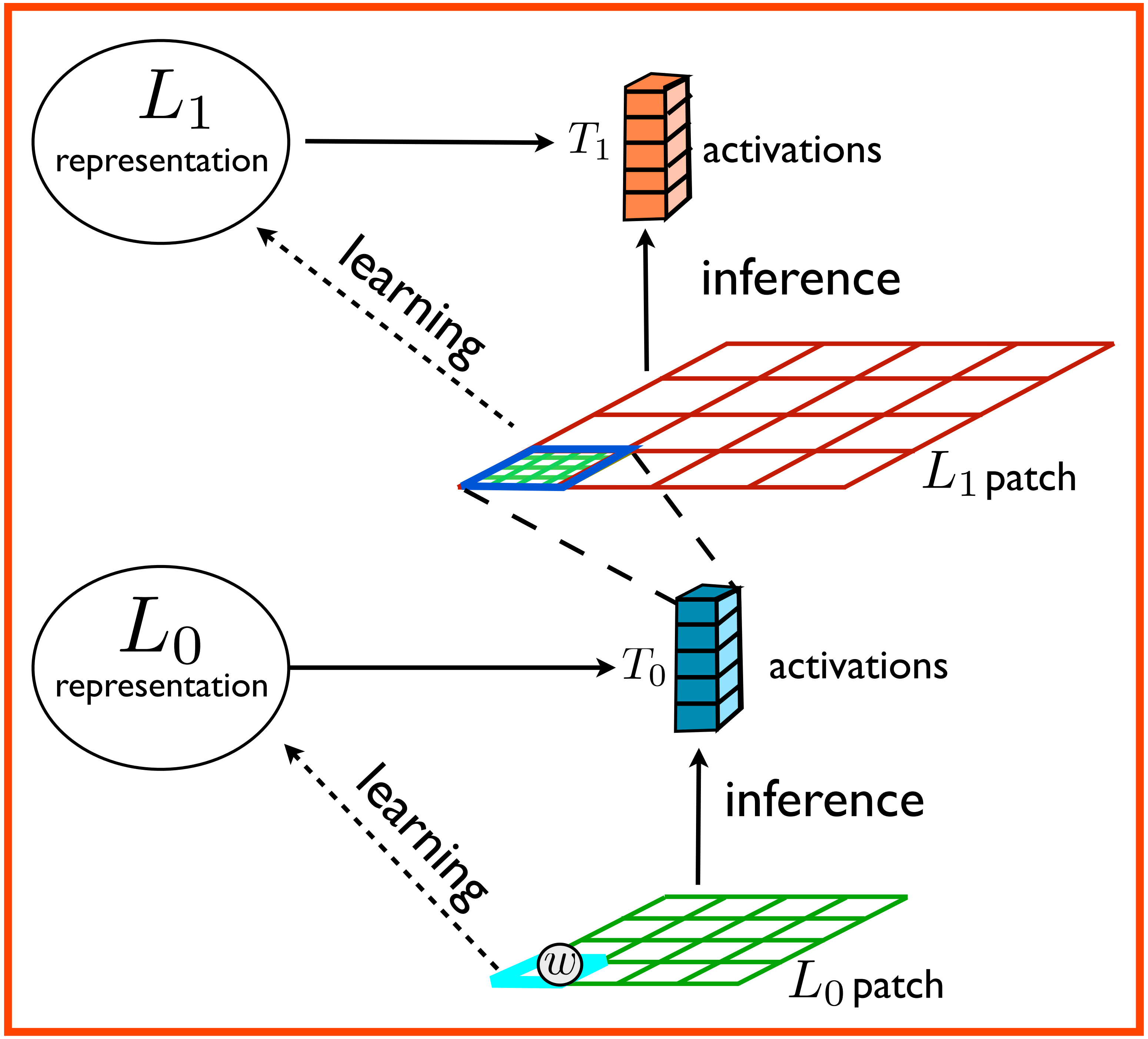
SIFT



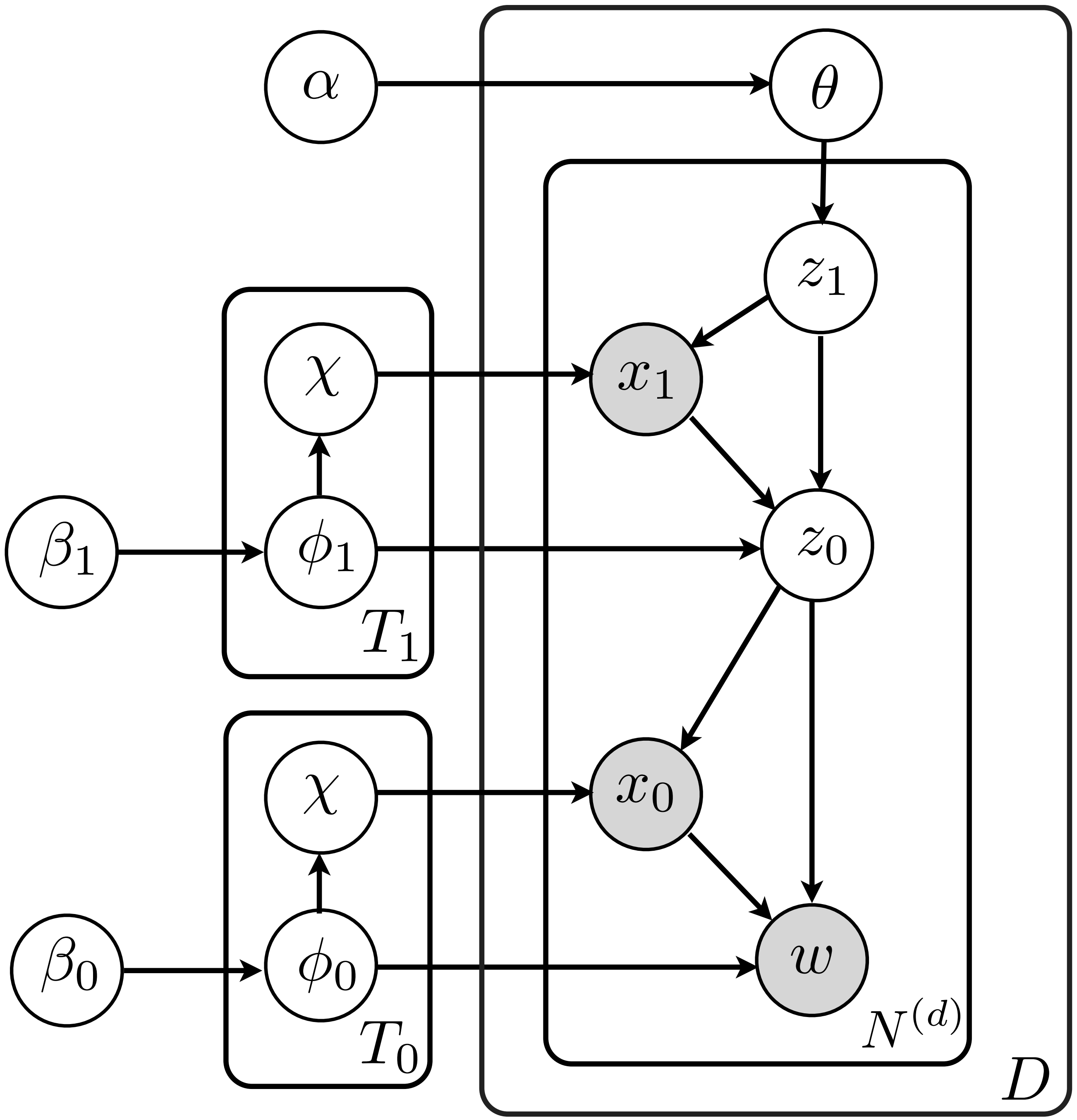
average image

(subset of 1024 topics)

Stacking two layers of LDA



Recursive LDA



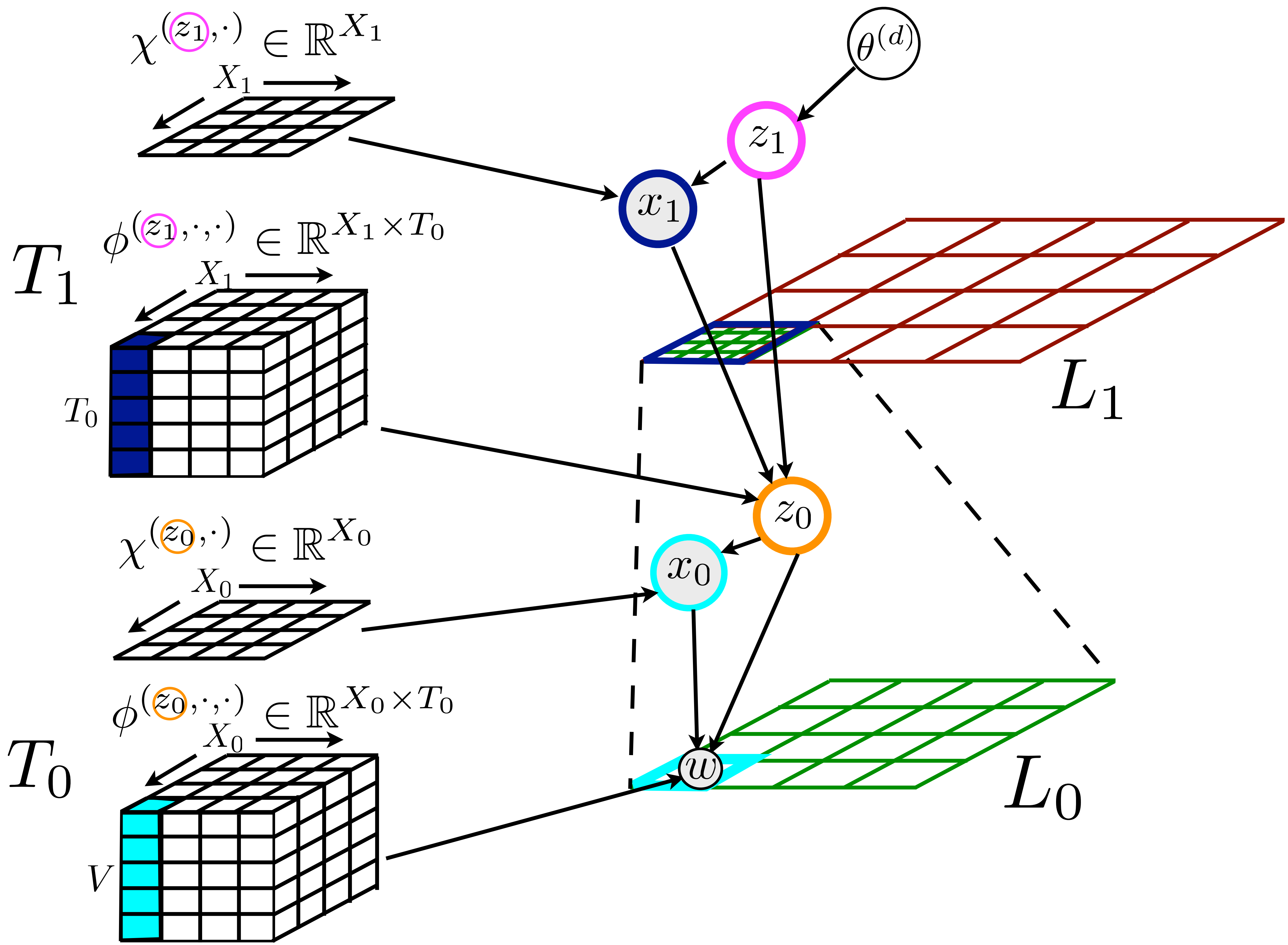
- $\phi_1 \sim \text{Dir}(\beta_1)$ and $\phi_0 \sim \text{Dir}(\beta_0)$: sample L_1 and L_0 multinomial parameters
- $\chi_1 \leftarrow \phi_1$ and $\chi_0 \leftarrow \phi_0$: compute spatial distributions from mixture distributions

For each document, $d \in \{1, \ldots, D\}$ top level mixing proportions $\theta^{(d)}$ are sampled according to:

- $\theta^{(d)} \sim \text{Dir}(\alpha)$: sample top level mixing proportions

For each document d, $N^{(d)}$ words w are sampled according to:

- $z_1 \sim \text{Mult}(\theta^{(d)})$: sample L_1 mixture distribution
- $x_1 \sim \text{Mult}(\chi_1^{(z_1,\cdot)})$: sample spatial position on L_1 given z_1
- $z_0 \sim \text{Mult}(\phi_1^{(z_1,x_1,\cdot)})$: sample L_0 mixture distribution given z_1 and x_1 from L_1
- $x_0 \sim \text{Mult}(\chi_0^{(z0,\cdot)})$: sample spatial position on L_0 given z_0
- $w \sim \text{Mult}(\phi_0^{(z_0,x_0,\cdot)})$: sample word given z_0 and x_0



$\chi^{(z_1,\cdot)} \in \mathbb{R}^{X_1}$
X_1
$\theta^{(d)}$
z_1
x_1
T_1
$\phi^{(z_1,\cdot,\cdot)} \in \mathbb{R}^{X_1 \times T_0}$
X_1
T_0
L_1
z_0
x_0
$\chi^{(z_0,\cdot)} \in \mathbb{R}^{X_0}$
X_0
$\phi^{(z_0,\cdot,\cdot)} \in \mathbb{R}^{X_0 \times T_0}$
T_0
X_0
V
w
L_0

Inference scheme

- Gibbs sampling: sequential updates of random variables with all others held constant.
- Linear topic response for initialization.

Outline

- Motivation:
  - ✴Distributed coding of local image features
  - ✴Hierarchical models
  - ✴Value of Bayesian inference
- Our model: Recursive LDA
- Evaluation

Evaluation

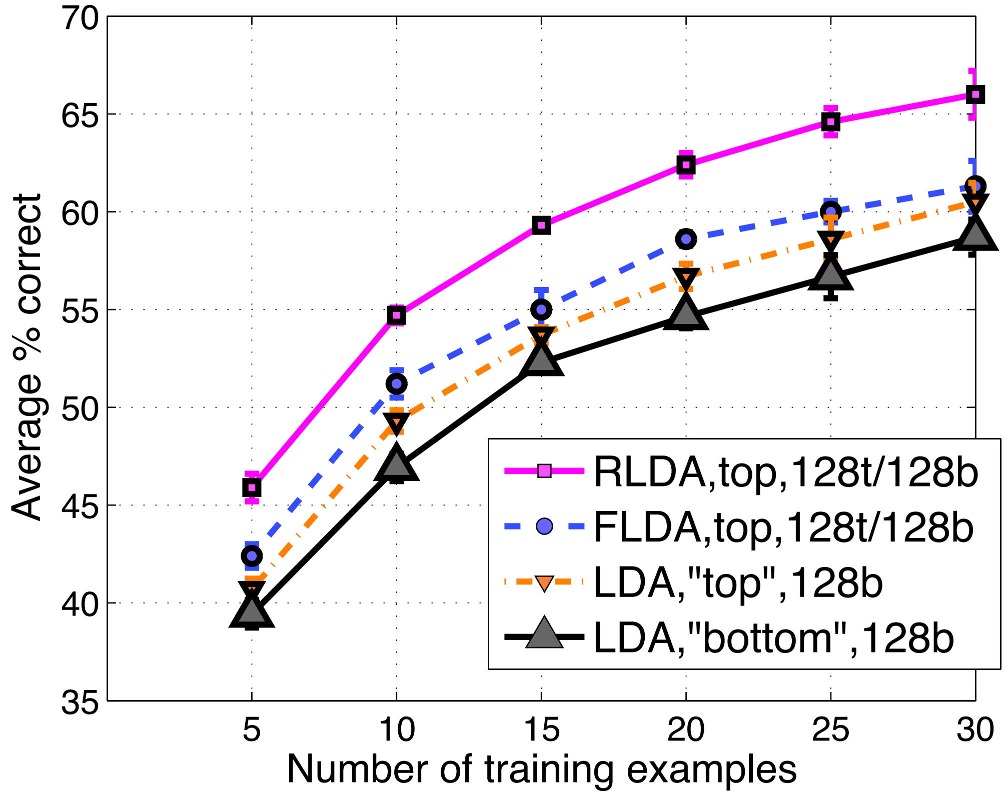
- 16px SIFT, extracted densely every 6px; max value normalized to 100 tokens
- Three conditions:
  - ✴ Single-layer LDA
  - ✴ Feed-forward two-layer LDA (FLDA)
  - ✴ Recursive two-layer LDA (RLDA)

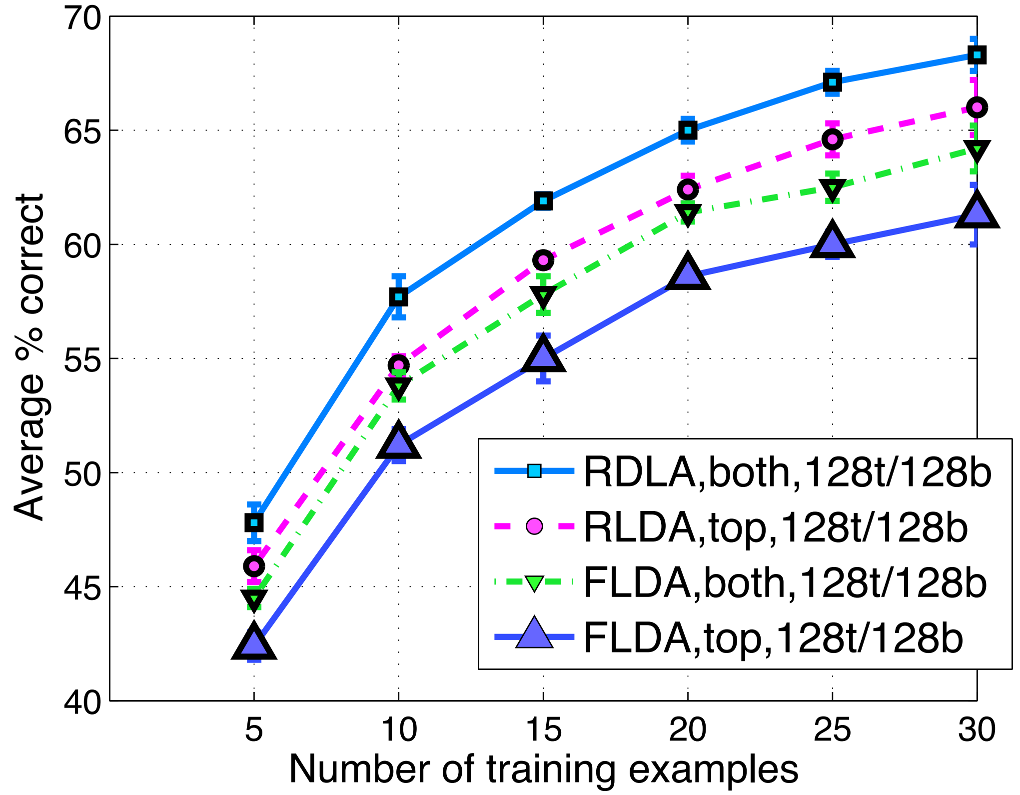
RLDA > FLDA > LDA

Approach				Caltech-101	
	Model	Basis size	Layer(s) used	**15**	**30**
128-dim models	LDA	128	"bottom"	$52.3 \pm 0.5\%$	$58.7 \pm 1.1\%$
	RLDA	128t/128b	bottom	$55.2 \pm 0.3\%$	$62.6 \pm 0.9\%$
	LDA	128	"top"	$53.7 \pm 0.4\%$	$60.5 \pm 1.0\%$
	FLDA	128t/128b	top	$55.4 \pm 0.5\%$	$61.3 \pm 1.3\%$
	RLDA	128t/128b	top	$59.3 \pm 0.3\%$	$66.0 \pm 1.2\%$
	FLDA	128t/128b	both	$57.8 \pm 0.8\%$	$64.2 \pm 1.0\%$
	RLDA	128t/128b	both	$61.9 \pm 0.3\%$	$68.3 \pm 0.7\%$

- additional layer increases performance
- full inference increases performance

RLDA > FLDA > LDA





- additional layer increases performance
- full inference increases performance
- using both layers increases performance

RLDA vs. other hierarchies

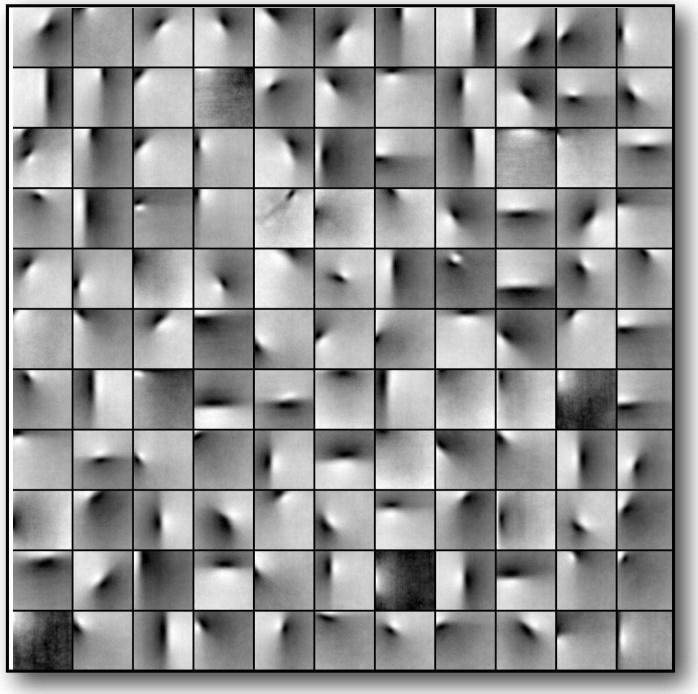
Approach			Caltech-101	
Model		Layer(s) used	**15**	**30**
Our Model	RLDA (1024t/128b)	bottom	$56.6 \pm 0.8\%$	$62.7 \pm 0.5\%$
	RLDA (1024t/128b)	top	$66.7 \pm 0.9\%$	$72.6 \pm 1.2\%$
	RLDA (1024t/128b)	both	$\mathbf{67.4 \pm 0.5}$	$\mathbf{73.7 \pm 0.8}\%$
Hierarchical Models	Sparse-HMAX [21]	top	51.0%	56.0%
	CNN [15]	bottom	–	$57.6 \pm 0.4\%$
	CNN [15]	top	–	$66.3 \pm 1.5\%$
	CNN + Transfer [2]	top	58.1%	67.2%
	CDBN [17]	bottom	$53.2 \pm 1.2\%$	$60.5 \pm 1.1\%$
	CDBN [17]	both	$57.7 \pm 1.5\%$	$65.4 \pm 0.4\%$
	Hierarchy-of-parts [8]	both	60.5%	66.5%
	Ommer and Buhmann [23]	top	–	$61.3 \pm 0.9\%$

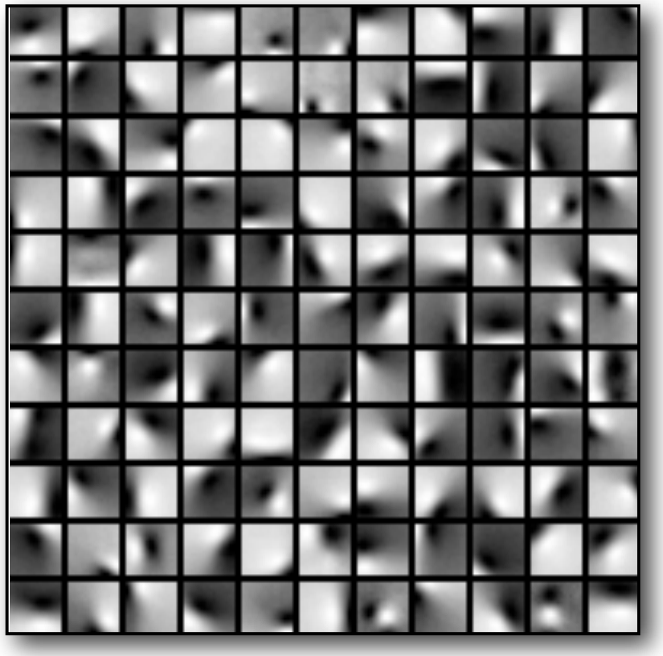
RLDA vs. single-feature state-of-the-art

- RLDA: 73.7%
- Sparse-Coding Spatial Pyramid Matching: 73.2% (Yang et al. CVPR 2009)
- SCSPM with “macrofeatures” and denser sampling: 75.7% (Bouerau et al. CVPR 2010)
- Locality-constrained Linear Coding: 73.4% (Wang et al. CVPR 2010)
- Saliency sampling + NBNN: 78.5% (Kanan and Cottrell, CVPR 2010)

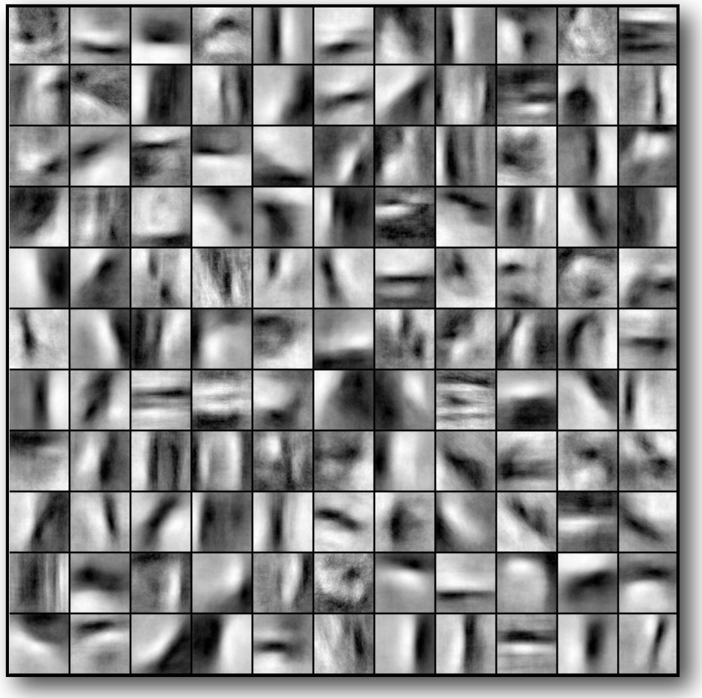
Bottom and top layers

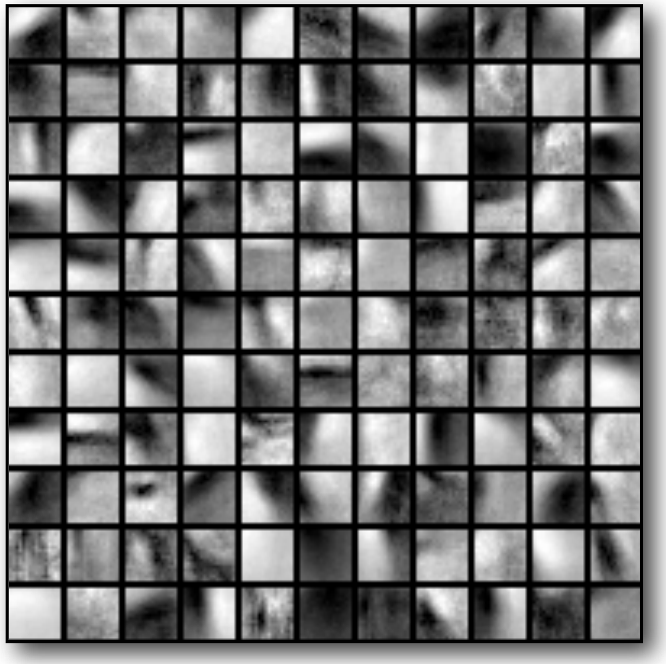
FLDA 128t/128b





RLDA 128t/128b





Top

Bottom

Conclusions

- Presented Bayesian hierarchical approach to modeling sparsely coded visual features of increasing complexity and spatial support.
- Showed value of full inference.

Future directions

- Extend hierarchy to object level.
- Direct discriminative component
- Non-parametrics
- Sparse Coding + LDA