A Probabilistic Model for Recursive Factorized Image Features

Sergey Karayev[1], Mario Fritz[2], Sanja Fidler[3], Trevor Darrell[1]
[1]UC Berkeley, [2]MPI Informatics, [3]University of Toronto

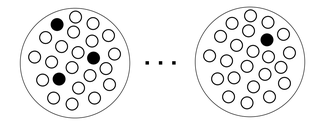
Motivation

Our goal: **distributed coding** of local features in a **hierarchical model** that would allow **full inference**.

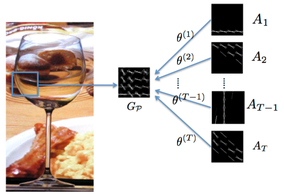
Distributed Coding

Considering local patch gradient energy histograms by orientation and grid location.

- Traditionally, vector quantized as *visual words*.
- Coding as a mixture of components has been shown to be empirically better.
- May represent additive image formation.



picture from Bruno Olshausen

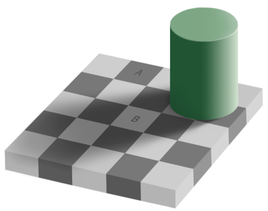


Hierarchical Models

- Biological evidence for increasing spatial support and complexity of visual pathway.
- Higher layers can help resolve local ambiguities.
- Possibly fewer parameters due to sharing of lower-layer components.
- Lots of past work in vision: HMAX, Convolutional Deep Belief Nets, Hyperfeatures, Hierarchies of Parts.

Bayesian Inference

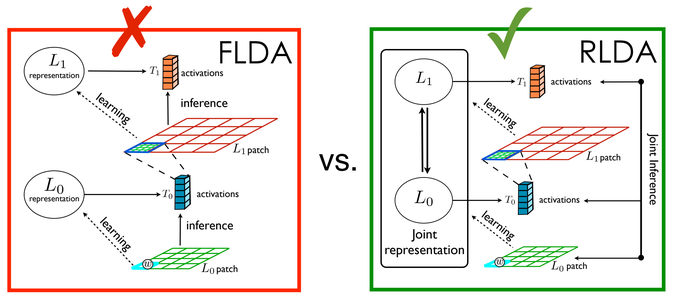
- Visual systems deal with inherently ambiguous data, which Bayesian inference helps disambiguate.
- Hierarchical priors allow unsupervised learning of complex visual structures.

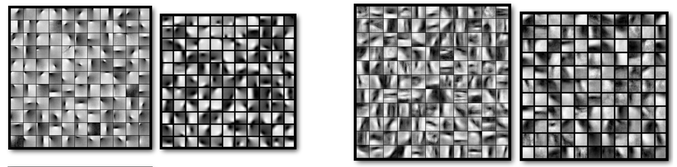


Model

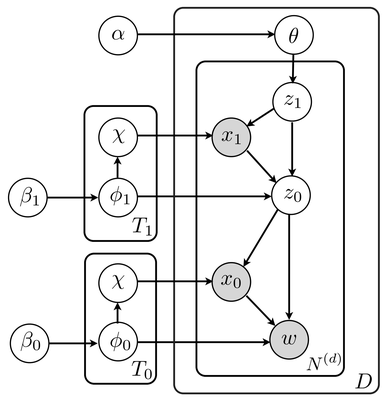
We formulate the Recursive LDA toward these goals.

- Based on Latent Dirichlet Allocation for distributed coding.
- Multiple layers in the hierarchy, with increasing spatial support.
- Unlike most hierarchical models, learns and does inference jointly across layers.





Visualization of learned components. RLDA learns more complex features at both layers.



- $\phi_1 \sim \text{Dir}(\beta_1)$ 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
- $x_0 \sim \text{Mult}(\chi_0^{(z_0,\cdot)})$: 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

Results

Setting in line with previous experiments on image coding:

- Evaluation on Caltech101
- Spatial Max Pooling
- Linear SVM classification

Three conditions:

- Single-layer LDA, on small and large patches ("bottom" and "top")
- Feed-forward two-layer LDA (FLDA)
- **Recursive two-layer LDA (RLDA)**

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	$61.9 \pm 0.3\%$	$68.3 \pm 0.7\%$

Results:

- Additional layer increases performance (FLDA > LDA).
- Full inference increases performance (RLDA > FLDA).
- Using both layers increases performance.
- Using only RLDA's lower layer is still better than LDA.

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 [22]	top	–	$61.3 \pm 0.9\%$

Acknowledgements

Various co-authors of this work were supported in part by the Air Force Office of Scientific Research, National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a; a Feodor Lynen Fellowship granted by the Alexander von Humboldt Foundation; by awards from the US DOD and DARPA, including contract W911NF-10-2-0059; by NSF awards IIS-0905647 and IIS-0819984; by EU FP7-215843 project POETICON; and by Toyota and Google.