Factoring movies into 'what' and 'where'

Jack Culpepper Advisor: Bruno Olshausen

bjc@cs.berkeley.edu

CIAR Summer School University of Toronto

August 15, 2006

Sparsenet is wasteful because it 'copies' basis functions



Why? Because sparsenet codes aren't translation invariant

- Ok, actually they are, *sort of*: the activations of low spatial frequency basis functions don't change much for small shifts (they don't need to, because they are so big and boring)
- That's why there are more 'copies' of high frequency BFs
- ...but we'd like a code that gives the same output when the input slightly translated, scaled, rotated, or occluded.
- I.e., we'd like to account for slight transformations of the image with another set of variables.

Sparsenet BFs ordered by spatial frequency



What happens as we slide our window?







a_i(y)

Morphable basis functions increase representational power

- We don't have to 'copy' each orientation and frequency to each position
- There are a small number of useful morphs, each of which can be applied to each template basis function
- We can use M basis functions and N morphs to gain the expressiveness of $M \times N$ basis functions

Morphable BFs can be used to separate 'what' from 'where'

- Modeling this way allows us to read out the two factors as separate variables
- The model then captures the separate functions of the ventral ('what') and dorsal ('where') processing streams in mamallian visual cortex
 - Strong support for this idea of two streams comes from lesion studies in monkey, in which damage to dorsal or ventral cortical regions impairs spatial abilities or object identification, respectively

Image transformations

- Consider a 1-dimensional, discretely sampled image of a 2dimensional world.
- We can express arbitrary transformations from I(x) to I'(x'), $x \in \mathcal{X}, x' \in \mathcal{X}'$ as follows.

$$I(x) = \sum_{x'} T(x, x') I'(x')$$

• Or, in vector-matrix notation,

$$I = TI'$$

Visualizing 1-d image transformations



Basis function expansions of image transformations

• Transformations between sequential frames of natural movies are 2-d and not necessarily affine.

- Add dimensions
$$y \in \mathcal{Y}$$
 and $y' \in \mathcal{Y}'$:

$$I(x,y) = \sum_{x'} \sum_{y'} T(x,y,x',y') I'(x',y')$$

- We can capture the space of all possible transforms by expressing T(x, y, x', y') as a basis function expansion:

$$T(x, y, x', y') = \sum_{j} c_{j} \psi_{j}(x, y, x', y')$$

Sparse coding in transform space

- The transformations lie inside a subspace of $\mathcal{X} \times \mathcal{Y} \times \mathcal{X}' \times \mathcal{Y}'$ (not all possible transforms are viable).
- We expect that for a small image patch, the correspondence between most sequential movie frames can be 'explained' by a small number of underlying causes – such as object motion or observer self-motion.
- Thus, it makes sense to learn them via sparse coding!

									COMPANY AND A DESCRIPTION OF		and the second diversion of			All Division in which the			Long to Long							
				100 100		a the little	1 A		State of the local division in which the local division in the loc			100	1.00											
	ter The Part	10 Mar 1997			-		-	Sec. 19.	- 18 18- I	10. 10. 1	ميا الإليا الا	Ref. Ball	-	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	182 18	10.10	10 10 10	1000						
1000	100	Ser Land Ser	State of State	100	8 8 3		100 100	10 100	100	A DECK	6. M. M.	And Address of	St. March			and the second	1000							
Sec. 1	See No. Oak	The Local Division in which the							All Street				-		14, 15		1000	and the second second						
	1 10 M	100 100				-		-	and the second	100					1000									
	and the state	1 m 1 m	1000					10 at 1	And Designation				10 Mar 1		The lot of		10.04							
	C. D. D.C.					States and states	1.00		Co-Co-	1000	100	100	325	1.00	And Designation	6 mil 1		and the second						
	State of the local division of the local div	and the second second					20.00					Contraction of the local division of the loc	and the state of the	-	1		and the second							
	and Street Property	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1				101.00					100	1000				1. 1.			1.00		1000		
	all, Par, Par,	The Party name				State of the local division of the local div											and in the second							
1000	and strength of the local division of the lo	And Designed Street											State of the local division of the local div											
	the state of the	And in case of	1000				1.00	1.1	100	See St. 1			1000			1.0								
		1000	1000			1000					1000		20 CON 10	-		1000								
and the second	12 10 20	100			化化化化	States and the	100		1.10	10 8 10 1	a play the	a brings Mirely		1000		4 4								
	and the second	10.00	100			All Property in which the		and the Party of the Party	-		10 10 10	1.00					100							
Sec. 2	1000	THE R OWNER WHEN	1 T	100	1	and the second	City states in	· · ·	1. S		10 all 10 all 10	City Manuel	the little H	10.00	1000	A 150 -	-			-		-		
		24 34 30	1000	R. 18 3	1000	and the second second	Con States		and the second second	100			-	1000	100	100	100	A DESCRIPTION OF						
100	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 S. 21	-	Section 2		Section 2											-					1000		
				1000			10.00			State of the local division of the local div														
			Sec. Sec.			States in the		1. 1.	10 IN				1. 1. 1.	-		1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1		and the second				-		
	ALC: NO. OF CO.	And Design of the local division of the loca	-					A 14 12					1. 1. 1. 1.			1000	(P 1) P	the second second						
		1990 - Barris Barris		ACCESSION OF TAXABLE		and the second	100 100 100							12 Total Total			and the second							1000
	1 m 1 m	a of Balls Inco	1.00			1.11	1 m	- 3- 3h	1. 1. 1. 1.	1000		and the second	-		1. Car	a Property	100					_		_
	A PROPERTY.	States and					10 10 1	NE 48 (1	100			1000			102 10	1000	1. 20.							
								100											1. A 1. A 1.					
			1000	A DECEMBER OF		and the state of the	100	1.000		100			and the second second		See Se							1000		1.00
		1 1 1 1	1 I I I I I	1							10 10 10 Inc.			1.000	1000									
	The other Design	100	1.00	1.00	And Division of the	No. of Concession, name			Constanting of the local division of the loc	14 10 2 1	1.00			10 10 10 10 10 10 10 10 10 10 10 10 10 1	10.00	100								
				1 a. 1 a. 1		and the second second	C (18)		State of the local division of the local div	100			200		101	1.1	4 4							
		100	1.00	1. C	90 JAN 3	10 CT	100 100	100 100	T 104 EC.144	16	and the state	1 In 1 In 1	L. C	1000	Salar and			1000						
100	1000			1.00		State of the			1000	THE REAL PROPERTY NAME	and the last	100	1000	10 A. 10 A.		8 8 -	3-3							
	the second se		the second se				Contraction of the local distance of the loc		1	ALC: NOT THE OWNER WATER OF THE OWNER OWNER OF THE OWNER	the second second second	a the state of the	the state of the s	the second second	1.000	1000	and the second se							
100		1.11.11.11.11	1.11.14.1	1.10		- C. C.	5 I (10	100 100 100	South States	100 Barris 6		State State	100	The other Distances	1000		and the second							
			1 1 1 1																		1.0	1		
				24					30	2 3						-			10					
														100				43					E.	
																		8	8				E.	
																			100			1	100	
																						10.00		
																							1.1.1.1	
																							1.1.1.1.1.1.1	
																		No. of States				1.	0.0	
																						1.1		
									해 번 것 수 는 는 는 한 한 은 한 이 가 다. 해 번 한 한 한 한 한 한 한 한 이 이 이 이														10 (1) (1) (1) (1)	
									3 4 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2													1 1 1	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	In the later
									10. 10. 10. 10. 10. 10. 10. 10. 10. 10.													1 1 1 1	10 10 10 10 10 10 10 10 10 10 10 10 10 1	a a a a a a
																						1 1 1 1	10 10 10 10 10 10 10 10 10 10 10 10 10 1	
																							10 10 10 10 10 10 10 10 10 10 10 10 10 1	The Indiana and an and
									30 20 20 20 20 20 20 20 20 20 20 20 20 20													1 1 1 1 1	1 1 1 1 1	a to the second se
									3 1 2 2 2 3 2 4 4 4 4 4 4 1 1 1 1 1 1 1 1 1 1 1 1 1													1 1 1 1 1 1 1 1	10.000	
																						1 1 1 1 1		
									4 12 12 12 12 12 12 12 12 12 12 12 12 12										As as he had be as he			1 1 1 1 1 1	THE R. LEWIS CO., NAMES	
																						1 1 1 1 1 1 1		
									에 해 해 해 해 해 해 해 해 해 하는 해 해 해 해 해 해 해 해 해 해 해 해 해 해 해 해 해 해 해															

Putting it all together

• Our complete model is now:

$$I(x,y) = \sum_{x',y'} T(x,y,x',y') I'(x',y')$$

where

$$T(x, y, x', y') = \sum_{j} \psi_j(x, y, x', y')c_j$$

and

$$I'(x',y') = \sum_{i} \phi_i(x',y')a_i$$

A bilinear model

• Rearranging the terms

$$I(x,y) = \sum_{x',y',i,j} \psi_j(x,y,x',y')\phi_i(x',y')a_ic_j$$

and precomputing the sums over x^\prime and y^\prime

$$\Gamma_{ij}(x,y) = \sum_{x',y'} \psi_j(x,y,x',y')\phi_i(x',y')$$

$$I(x,y) = \sum_{i,j} \Gamma_{ij}(x,y) a_i c_j$$

(Grimes & Rao, 2005)



Transforms in coefficient space

- Sparse coding is capable of learning transformations between the pixels in sequential movie frames
- It turns out that it is also capable of learning transformations between the sparse codes for representing sequential movie frames, and this is even easier



Modeling slow underlying causes

 Edges don't pop into existence in one position, disappear, then pop into existence in a new position. They appear, then translate. We capture this using a latent variable model to describe the frame by frame coefficients:

$$I(x, y, t) = \sum_{i} \phi_{i}(x, y)a_{i}(t) + \nu_{I}$$
$$a_{i}(t) = \sum_{j} T_{ij}(t)b_{j}(t)$$
$$T_{ij} = \sum_{k} \Gamma_{ijk}c_{k}(t) + \nu_{a}$$

A bilinear model of slow underlying causes

• This gives a bilinear model for the frame by frame coefficients:

$$I(x, y, t) = \sum_{i} \phi_i(x, y) a_i(t) + \nu_I$$
$$a_i(t) = \sum_{j,k} \Gamma_{ijk} b_j(t) c_k(t) + \nu_a$$

- We'd like to learn Γ that produces sparse c and b, and slowly varying b



Probabilistic framework

 Assuming that reconstruction errors in a are due to i.i.d. Gaussian noise, the probability of a given the coefficients and model parameters is a Gaussian distribution:

$$P(\mathbf{a}|\mathbf{b},\mathbf{c},\mathbf{\Gamma}) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\{-\frac{1}{2\sigma_n^2} \sum_{i,t} [e(i,t)]^2\}$$

where

$$e(i,t) = a_i(t) - \hat{a}_i(t)$$
$$\hat{a}_i(t) = \sum_{jk} \Gamma_{ijk} b_j(t) c_k(t)$$

 $Z_n = \sqrt{2\pi}\sigma_n$, and σ_n^2 is the variance of the noise.

Prior knowledge

$$P(\mathbf{b}|\mathbf{\Gamma}) = \prod_{j,t} \frac{1}{Z_S} \exp\{-S(b_j(t))\} \prod_j \frac{1}{Z_R} \exp\{-R(b_j(t))\}$$
$$P(\mathbf{c}|\mathbf{\Gamma}) = \prod_{k,t} \frac{1}{Z_S} \exp\{-S(c_k(t))\}$$

Wish to adapt our model parameters so that (1) the coefficients follow distributions that are peaked at zero with heavy tails, because we believe that our images can be 'explained' by a small number of independent, additive features, and (2) the 'what' coefficients vary slowly in time.

Slow prior

• The time-varying model imposes a smoothness constraint over time on the 'what' variables using the following cost function:

$$R(b_j(t)) = \frac{1}{2} \sum_{t}^{\tau-1} [b_j(t) - b_j(t+1)]^2$$

$$\frac{\partial R}{\partial b_l(u)} = 2b_l(u) - b_l(u+1) - b_l(u-1)$$

Probabilistic bilinear sparse coding

• Using Bayes' rule, we can compute the probability of different explanations for an image's sparse code:

 $P(\mathbf{b}, \mathbf{c}|\mathbf{a}, \Gamma) = P(\mathbf{a}|\mathbf{b}, \mathbf{c}, \Gamma)P(\mathbf{b}, \mathbf{c}|\Gamma)/P(\mathbf{a}, \Gamma)$

• Given an image's code a and a set of basis functions Γ , we select the most likely coefficients for describing it by computing the maximum a-posteriori (MAP) estimates b^* and $c^*.$

 $P(\mathbf{b}, \mathbf{c}|\mathbf{a}, \Gamma) \propto P(\mathbf{a}|\mathbf{b}, \mathbf{c}, \Gamma) P(\mathbf{b}|\Gamma) P(\mathbf{c}|\Gamma)$

Objective function

• From this, we derive our objective function, minimization of which is equivalent to finding the MAP estimates:

$$\mathcal{L} = -\log P(\mathbf{b}, \mathbf{c} | \mathbf{a}, \Gamma)$$

$$\propto \frac{1}{2\sigma_n^2} \sum_{i,t} [e(i,t)]^2 + \sum_{j,t} S(b_j(t)) + \sum_{k,t} S(c_k(t)) + \sum_j R(b_j(t))$$