Natural image statistics & neural representation learning

Matthias Bethge

Werner Reichardt Centre for Integrative Neuroscience & Bernstein Centre for Computational Neuroscience & Max Planck Institute for Biological Cybernetics Tübingen
Perceptual inference

object detection/scene parsing

3D configuration, shape

dynamics
David Marr: isolating different inference mechanisms
The robustness of vision
Robust vision ultimately denotes the task-independent ability to transform image information into a new representation which makes the different physical/causal sources explicit that underly the pixel variations across different images.

As such it allows vision systems to generalize across varying conditions and to achieve stable interpretations across time.
Visual inference (object recognition/detection, scene parsing)

Google, Facebook AI, Microsoft, Baidu, Clarifai, Deep Mind, Madbits,…
Visual inference (object recognition/detection, scene parsing)

Google, Facebook AI, Microsoft, Baidu, Clarifai, Deep Mind, Madbits,…
Neocognitron (Fukushima, 1980)/H-Max
(stacking S- and C-cells)

There is no simple trick that will miraculously solve the vision problem.

It’s not important which “religion” you pick but how much you can contribute to the specific problems in your area.

(figure from scholarpedia)
just pattern recognition?

**Figure 3. The Ventral Visual Pathway**

The ventral visual stream has been parsed into distinct visual areas based on anatomical connectivity patterns. Each area has a specific role in visual processing. For example, the lateral occipital cortex (LOC) shows strong commonality in the population representation of object categories. Lesion studies in monkeys and humans have established a likely homology with the human cortex in and around the human inferior temporo-occipital (IT) cortex. Invariance is required for object recognition in these areas.

- **(task)**
  - **sensory signals**
  - **classification**

**Key Areas**

- V1 (primary visual cortex)
- V2
- V4
- IT (inferior temporo-occipital cortex)
- PIT
- AIP
- LGN (lateral geniculate nucleus)
- Retina

**Neuronal Densities**

Neuronal densities are shown above each area. The approximate dimensionality (number of projection neurons) is also indicated. The central visual field is represented as a pie chart, showing the approximate median response latency for each area.
just pattern recognition?

sensory signals → feature space transformations → classification/regression → classification
just pattern recognition?

sensory signals —> feature space transformations —> classification/regression —> classification

large number of classes
just pattern recognition?

- infinite datasets
- sensory signals
- feature space transformations
- classification/regression
- large number of classes
- classification
just pattern recognition?

combinatorial algebra of tasks

task

infinite datasets

sensory signals

feature space transformations

classification/regression

large number of classes

classification
just pattern recognition?

combinatorial algebra of tasks

task

infinite datasets

sensory signals

feature space transformations

classification/regression

large number of classes

classification

unsupervised representation learning
Redundancy Reduction

Atick & Redlich, Neural Comput., 1990

Ganglion cells: Whitening

Barlow’s redundancy reduction hypothesis (1961)
Redundancy Reduction

Ganglion cells: Whitening
Atick&Redlich, Neural Comput., 1990

Simple cells: Independent Component Analysis (ICA)
Bell&Sejnowski, Vision Res., 1997

Barlow’s redundancy reduction hypothesis (1961)
Redundancy Reduction

Ganglion cells: Whitening
Atick&Redlich, Neural Comput., 1990

Simple cells: Independent Component Analysis (ICA)
Bell&Sejnowski, Vision Res., 1997

Complex cells: Subspace ICA
Hyvarinen&Hoyer, Neural Comput., 2000
Hyvarinen&Köster, Network, 2007

Retina
LGN
V1

bandpass filtering
orientation selectivity
phase invariance
Redundancy Reduction

Barlow’s redundancy reduction hypothesis (1961)
Barlow’s redundancy reduction hypothesis (1961)
Barlow’s redundancy reduction hypothesis (1961)
Density estimation via redundancy reduction

Redundancy reduction can be interpreted as a flexible density estimation method
Statistical models of natural image patches
I. Linear models

(receptive field modeling)
Linear models of natural images

Pixel basis
Linear models of natural images

\[ = s_1 \cdot \begin{array} {c} \text{Pixel basis} \\ \end{array} + s_2 \cdot \begin{array} {c} \text{Pixel basis} \\ \end{array} + s_3 \cdot \begin{array} {c} \text{Pixel basis} \\ \end{array} + \ldots \]
Linear models of natural images

\[ = s_1 \cdot + s_2 \cdot + s_3 \cdot + \ldots \]

Pixel basis
Linear models of natural images

\[ = s_1 \cdot \begin{array}{c} \includegraphics{image_1} \end{array} + s_2 \cdot \begin{array}{c} \includegraphics{image_2} \end{array} + s_3 \cdot \begin{array}{c} \includegraphics{image_3} \end{array} + \ldots \]
Linear models of natural images

\[ s_1 \cdot \quad + s_2 \cdot \quad + s_3 \cdot \quad + \ldots \]

Fast Fourier basis
Linear models of natural images

\[ = s_1 \cdot \begin{array} \text{ } \end{array} + s_2 \cdot \begin{array} \text{ } \end{array} + s_3 \cdot \begin{array} \text{ } \end{array} + \ldots \]

Discrete cosine basis
Linear models of natural images

\[ s_1 \cdot \text{Image} + s_2 \cdot \text{Pattern} + s_3 \cdot \text{Component} + ... \]

Discrete cosine basis
Linear models of natural images

\[
= s_1 \cdot \begin{array}{c}
\text{[Image]}
\end{array}
+ s_2 \cdot \begin{array}{c}
\text{[Image]}
\end{array}
+ s_3 \cdot \begin{array}{c}
\text{[Image]}
\end{array}
+ \ldots
\]

Haar wavelet basis
Linear models of natural images

\[ = s_1 \cdot \square + s_2 \cdot \square + s_3 \cdot \square + \ldots \]

Haar wavelet basis
Model comparison

Original data

independent Fourier model

independent pixel model
How to start with natural image statistics
How to start with natural image statistics
How to start with natural image statistics
Independent Fourier model

Redundancy Reduction

\[ \mathcal{F} \]
Independent Fourier model

Redundancy Reduction

$\mathcal{F}$
Independent Fourier model

Redundancy Reduction

\( \mathcal{F} \)

shuffle
Independent Fourier model

Redundancy Reduction $\mathcal{F}$

Generative model $\mathcal{F}^{-1}$

shuffle
Independent pixel model

shuffle
Linear models of natural images

Generative model:

\[ p_x(x), \quad x = As, \quad p_s(s) = \prod_k p_k(s_k) \]
Linear models of natural images

Generative model:

\[ p_x(x), \quad x = A s, \quad p_s(s) = \prod_k p_k(s_k) \]

\[
X = \begin{pmatrix}
x_{11} & x_{12} & \cdots \\
x_{21} & x_{22} & \cdots \\
\vdots & \vdots & \ddots \\
x_{n1} & x_{n2} & \cdots 
\end{pmatrix}
\]

\[ X = A S \]

\[
S = \begin{pmatrix}
s_{11} & s_{12} & \cdots \\
s_{21} & s_{22} & \cdots \\
\vdots & \vdots & \ddots \\
s_{n1} & s_{n2} & \cdots 
\end{pmatrix}
\]
Linear models of natural images

Generative model:

\[ p_x(x), \quad x = As, \quad p_s(s) = \prod_k p_k(s_k) \]

\[
X = \begin{pmatrix}
  x_{11} & x_{12} & \cdots \\
  x_{21} & x_{22} & \cdots \\
  \vdots & \vdots & \ddots \\
  x_{n1} & x_{n2} & \cdots
\end{pmatrix}
\]

\[
S = \begin{pmatrix}
  s_{11} & s_{12} & \cdots \\
  s_{21} & s_{22} & \cdots \\
  \vdots & \vdots & \ddots \\
  s_{n1} & s_{n2} & \cdots
\end{pmatrix}
\]
Linear models of natural images

Generative model:

\[ p_x(x), \quad x = As, \quad p_s(s) = \prod_k p_k(s_k) \]

\[ X = \begin{pmatrix} x_{11} & x_{12} & \cdots \\ x_{21} & x_{22} & \cdots \\ \vdots & \vdots & \ddots \\ x_{n1} & x_{n2} & \cdots \end{pmatrix} \quad \xrightarrow{\text{X = AS}} \quad S = \begin{pmatrix} s_{11} & s_{12} & \cdots \\ s_{21} & s_{22} & \cdots \\ \vdots & \vdots & \ddots \\ s_{n1} & s_{n2} & \cdots \end{pmatrix} \]
Linear models of natural images

Generative model:

\[ p_x(x), \quad x = As, \quad p_s(s) = \prod_k p_k(s_k) \]

\[
X = \begin{pmatrix} x_{11} & x_{12} & \cdots \\ x_{21} & x_{22} & \cdots \\ \vdots & \vdots & \ddots \\ x_{n1} & x_{n2} & \cdots \end{pmatrix}
\]

\[
S = A^{-1}X
\]

“Synthesis”

“Analysis”

Data

“Sources/ Causes…”
2nd-order redundancy reduction I: Principal Component Analysis (PCA)
2nd-order redundancy reduction I: Principal Component Analysis (PCA)
2nd-order redundancy reduction I: Principal Component Analysis (PCA)
2nd-order redundancy reduction I: Principal Component Analysis (PCA)
2nd-order redundancy reduction I: Principal Component Analysis (PCA)
2nd-order redundancy reduction I: Principal Component Analysis (PCA)
2nd-order redundancy reduction I: Principal Component Analysis (PCA)
2nd-order redundancy reduction I: Principal Component Analysis (PCA)
Nonorthogonal decorrelation via whitening (variance equalization)
Nonorthogonal decorrelation via whitening (variance equalization)
Nonorthogonal decorrelation via whitening (variance equalization)
Nonorthogonal decorrelation via whitening (variance equalization)

\[ C_x = I \quad s = Ux \]
\[ \Rightarrow C_s = UC_x U^\top = UU^\top = I \]
Symmetric Whitening

Redundancy Reduction

\[ \mathcal{F} \]

Generative model

\[ \mathcal{F}^{-1} \]

shuffle
Symmetric Whitening

Redundancy Reduction

\[ \mathcal{F} \]

Generative model

\[ \mathcal{F}^{-1} \]

shuffle
Higher-order redundancy reduction: Independent Component Analysis (ICA)

Find the most non-Gaussian directions:
Higher-order redundancy reduction: Independent Component Analysis (ICA)

Find the most non-Gaussian directions:
Higher-order redundancy reduction: Independent Component Analysis (ICA)

Find the most non-Gaussian directions:
Higher-order redundancy reduction: Independent Component Analysis (ICA)

Find the most non-Gaussian directions:
Optimal linear redundancy reduction: Independent Component Analysis (ICA)
II. Nonlinear ($\nu$-spherical) models

(contrast gain control modeling)
Different types of sparsity: factorial vs spherical

identical kurtotice (“sparse”) marginals

factorial density

spherical density

Nonlinear redundancy reduction of spherical data

Radial Gaussianization
(Lyu & Simoncelli, 2008)
(Sinz & Bethge, 2008)
Radial Factorization

Cartesian but not factorial

ground truth

factorial density

spherical density

⇒ Use Lp-spherical distributions

[Sinz & Bethge, NIPS, 2008.]
Family of $L_p$-spherical distributions

$p$-generalized Normal distributions

factorial distributions

$p(x) = \prod_{k} p_k(x_k)$

$p$-spherical distributions

$p(x) = p(||Wx||_p)$

Gaussian scale mixture (GSM) models

Normal distribution

$p \rightarrow \infty$

$p = 2$

$p = 1$

$p = 0.5$

ICA vs radial factorization

ICA coefficients

Radially factorized coefficients
Objective function?

How effective are (current) neural model representations in capturing the higher-order correlations of natural images?
Unsupervised Learning

Set of images with specified pixel histograms

Set of natural images
Unsupervised Learning

Set of images with specified pixel histograms

Set of natural images
Unsupervised Learning

Set of images with specified pixel histograms

Set of images with specified second-order statistics

Set of natural images
Unsupervised Learning

- Set of images with specified pixel histograms
- Set of images with specified second-order statistics
- Set of images with some specified higher-order statistics
- Set of natural images
Unsupervised Learning

Set of images with specified pixel histograms

Set of images with specified second-order statistics

Set of images with some specified higher-order statistics

Set of natural images

synthesis models
Unsupervised Learning

- Set of images with specified pixel histograms
- Set of images with specified second-order statistics
- Set of images with some specified higher-order statistics
- Set of natural images
- Look-up tables (flickr, facebook,...)
- Synthesis models
Unsupervised Learning

- Set of images with specified pixel histograms
- Set of natural images
- Look-up tables (flickr, facebook, ...)
- Evaluation: $E[- \log \hat{p}(x)]$
Unsupervised Learning

Set of images with specified pixel histograms

Set of images with specified second-order statistics

Set of images with some specified higher-order statistics

Set of natural images

Evaluate $E[- \log \hat{p}(x)]$

Synthesis models

Look-up tables (flickr, facebook,...)
Unsupervised Learning

Evaluate $E[-\log \hat{p}(x)]$

or equivalently:

$I(\hat{F}(x)) = \sum_k h[y_k] - h[y]$
Unsupervised Learning

Set of images with specified pixel histograms

- bandpass filtering specifies second-order statistics
- orientation selectivity can specify some higher-order statistics
- complex cell pooling, contrast gain control, hierarchical architectures,...

Set of natural images
Unsupervised Learning

Set of images with specified pixel histograms

Linear

Set of natural images

Complex cell pooling, contrast gain control, hierarchical architectures,...

Orientation selectivity can specify some higher-order statistics

Bandpass filtering specifies second-order statistics
Unsupervised Learning

Set of images with specified pixel histograms

Set of natural images

Linear
- bandpass filtering specifies second-order statistics
- orientation selectivity can specify some higher-order statistics

Non-linear
- complex cell pooling, contrast gain control, hierarchical architectures,...
How effective are hierarchical representations in capturing the higher-order correlations of natural images?
Lp-nested symmetric distributions

\[ p(x) = p(\|Wx\|_p) \]

\[ p(x) = p(\nu(Wx)) \]
Lp-nested symmetric distributions

\[ p(x) = p(\|Wx\|_p) \]

\[ p(x) = p(\nu(Wx)) \]

\[ \nu(y) = \left( (|y_1|^{p_1} + |y_2|^{p_1})^{\frac{p_0}{p_1}} + |y_3|^{p_0} \right)^{\frac{1}{p_0}} \]

[Sinz & Bethge (2010). JMLR, 3409-3451.]
Lp-nested symmetric distributions

[Sinz et al, NIPS 2009]
Multi-layer ICA

search for filters with non-Gaussian histograms

“Gaussianize”

Multi-layer ICA

search for filters with non-Gaussian histograms

“Gaussianize”

Multi-layer ICA

search for filters with non-Gaussian histograms

“Gaussianize”

Hierarchical Models of Natural Images

Given the ensemble of data, it is called whitenning matrix. Let $X_1$ be orthogonal and $P$ be positive definite. $\hat{X}_1$ is a whitened version of $X_1$.

Quantitative evaluation

$H(\text{inspired Approach}) = p_1 + \cdots + p_k + D_i X_i$.

Independent component Analysis

Conclusion

{\image}

Multi-layer ICA

$\Delta I [\text{bits/pixel}]$ vs. #layers

Lp-spherical distribution


References
Quantitative model comparison

Higher-order redundancy reduction [%]

# Layers

1 2 3 4 5

- mixture of GSMs
- $L_p$-spherical/nested model
- multilayer ICA
- whitening
Quantitative model comparison

Lucas Theis

Higher-order redundancy reduction [%]

mixture of GSMs

$L_p$-spherical/nested model

multilayer ICA

Deep Belief Net

whitening

Likelihood comparison

- PCA
- DBN
- ICA
- GSM
- OICA
- HICA
- \(L_p\)-elliptical
- ISA
- MoGaussian
- PoT
- HISA
- MoGSM
- MCG
- MCGSM
- MSMCGSM

Log-likelihood [bit/px]
MCGSM: a directed mixture of experts model of natural images

Key advantages:

1.) Built-in translation invariance

2.) Model if you can and ignore if not
Mixture of conditional GSMs (MCGSM)

\[
p(y|x) = \sum_{c=1}^{m} \sum_{s=1}^{n} p(c|x) p(s|c, x) p(y|c, s, x)
\]

Gating:
\[
p(c, s|x) \propto \exp \left( -\frac{\lambda_{c,s}}{2} x^\top K_c x \right)
\]

Prediction:
\[
p(y|c, s, x) \propto \exp \left( -\frac{1}{2} \frac{(y - w_c^\top x)^2}{\sigma_{s,c}^2} \right)
\]

Likelihood: MCGSM >> MoGSM
Likelihood: MCGSM >> MoGSM
Synthesized images from the MCGSM
Synthesized textures from the MCGSM

Original

MCGSM
Synthesized textures from the MCGSM

Original

MCGSM
Synthesized textures from the MCGSM

Original

MCGSM
Synthesized textures from the MCGSM
class "horizontal"
class “diagonal downwards”
class “diagonal upwards”
class “vertical”
class “flat/sky”
class “trunks & brunches”
class “leaves”
class “contours”
Stacking MCGSMs?
What information is encoded in the class labels?
What information is encoded in the class labels?
What information is encoded in the class labels?
What information is encoded in the class labels?
Multi-scale MCGSM

MCGSM trained on natural images (van Hateren)

Unsupervised Learning → Likelihood

- PCA
- DBN
- ICA
- GSM
- OICA
- HICA
- $L_p$-elliptical
- ISA
- MoGaussian
- PoT
- HISA
- MoGSM
- MCG
- MCGSM
- MSMCGSM

Log-likelihood [bit/px]
Why care about likelihoods? What does it take to build useful image representations?
Why care about likelihoods? What does it take to build useful image representations?

Ultimately, we seek for universal image representations that are not overfit to specific tasks or datasets and thus robust under changing conditions.
Bag of features
Bag of features

• simplest example:
  sum over white pixels
Bag of features

- simplest example: sum over white pixels
- classification of MNIST digits 1 against 2 yields an average error rate of 6%
Bag of features

- simplest example: sum over white pixels

- classification of MNIST digits 1 against 2 yields an average error rate of 6%

- highly invariant
Bag of features

- simplest example: sum over white pixels

- classification of MNIST digits 1 against 2 yields an average error rate of 6%

- highly invariant

- but not robust (relies on an epiphenomenon)
## Texture Classification

### Table of Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bruna &amp; Mallat, 2011</td>
<td>0.09%</td>
</tr>
<tr>
<td><strong>MCGSM</strong></td>
<td><strong>0.29%</strong></td>
</tr>
<tr>
<td>Broadhurst, 2005</td>
<td>0.78%</td>
</tr>
<tr>
<td>Crosier &amp; Griffin, 2008</td>
<td>1.4%</td>
</tr>
<tr>
<td>Hayman et al., 2004</td>
<td>1.54%</td>
</tr>
<tr>
<td>Varma &amp; Zisserman, 2009</td>
<td>1.97%</td>
</tr>
<tr>
<td>Zhang et al., 2006</td>
<td>4.7%</td>
</tr>
</tbody>
</table>

![Texture Classification Diagram](image-url)
Filling-in

\[ Q_\theta(\text{“missing pixels”} \mid \text{“known pixels”}) = \frac{Q_\theta(\text{“all pixels”})}{Q_\theta(\text{“known pixels”})} \]
Filling-in

$$Q_\theta("missing \ pixels" \mid "known \ pixels") = \frac{Q_\theta("all \ pixels")}{Q_\theta("known \ pixels")}$$
Acknowledgments

Fabian Sinz  Reshad Hosseini

Lucas Theis  Sebastian Gerwinn
Thanks!