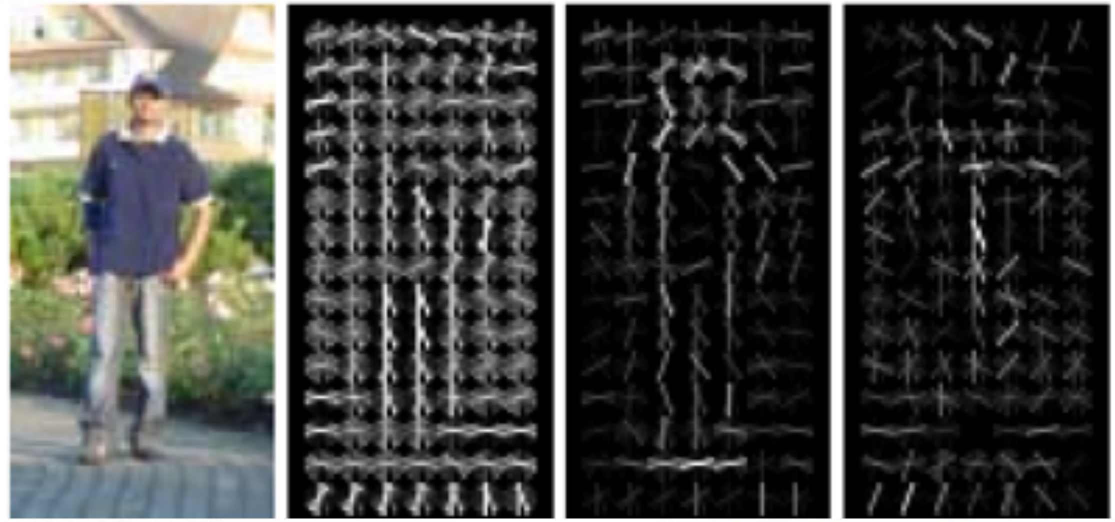
JHU vision lab

Globally Optimal Matrix Factorizations, Deep Learning and Beyond

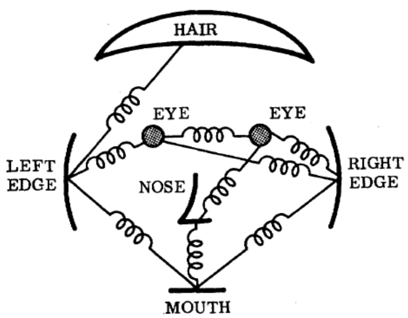
René Vidal
Center for Imaging Science
Institute for Computational Medicine

Object Recognition: 2000-2012

- Features
  - SIFT (Lowe 2004)
  - HoG (Dalal and Triggs 2005)



- Classifiers
  - Bag-of-visual-words
  - Deformable part model (Felzenszwalb et al. 2008)
- Databases
  - Caltech 101
  - PASCAL
  - ImageNet



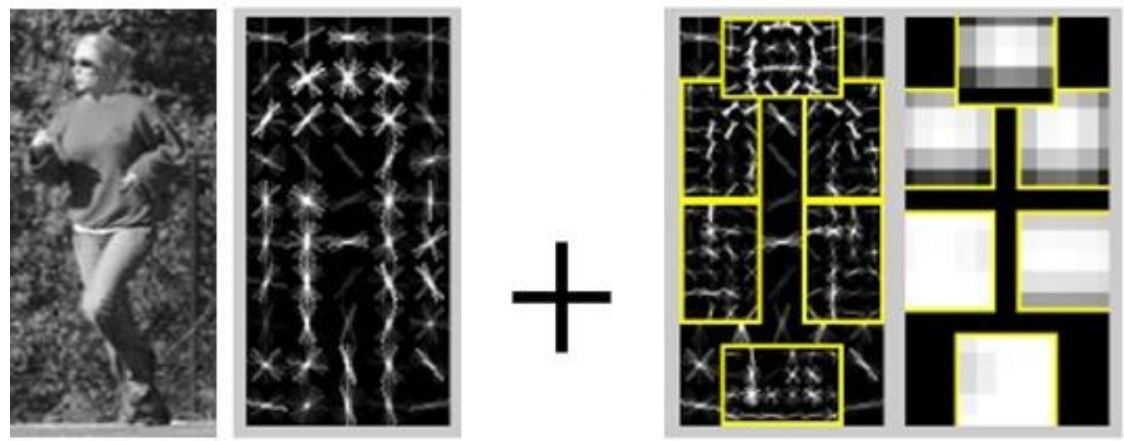
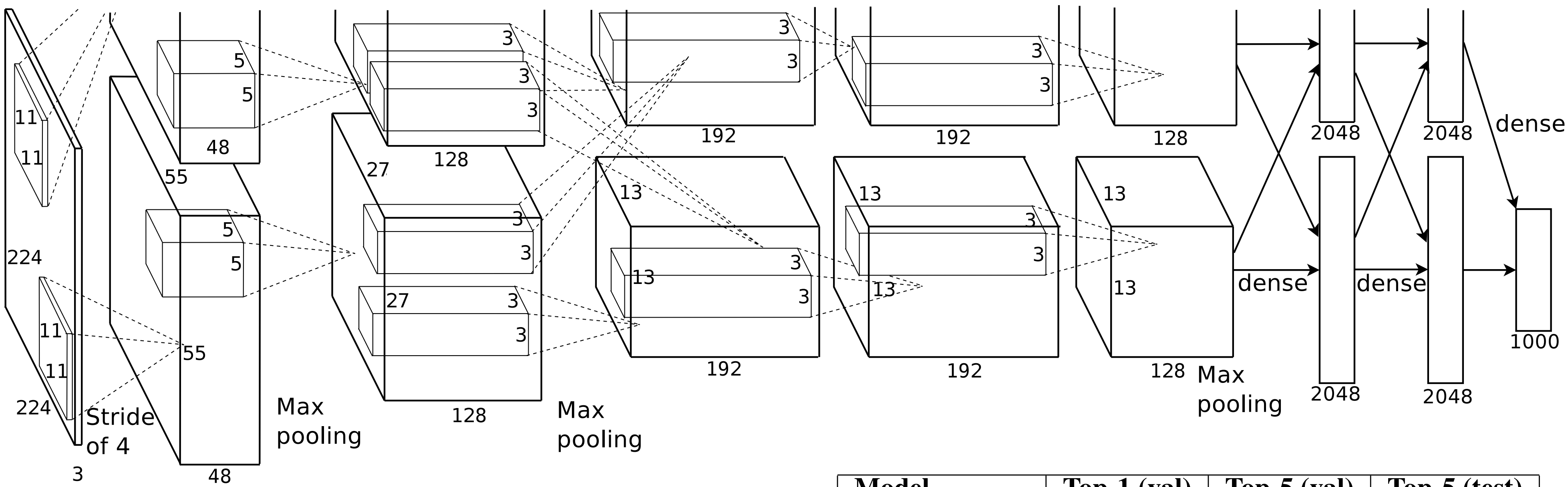


Image | Root filter (Coarse resolution) | Part filters (Fine resolution) | Deformation Models

- Performance on PASCAL VOC started to plateau 2010-2012

Learning Deep Image Feature Hierarchies

- Deep learning gives ~ 10% improvement on ImageNet
  - 1.2 million images, 1000 categories, 60 million parameters



Model	Top-1	Top-5
Sparse coding [2]	*47.1%*	*28.2%*
SIFT + FVs [24]	*45.7%*	*25.7%*
CNN	**37.5%**	**17.0%**

Table 1: Comparison of results on ILSVRC-2010 test set. In *italics* are best results achieved by others.

Model	Top-1 (val)	Top-5 (val)	Top-5 (test)
SIFT + FVs [7]	—	—	*26.2%*
1 CNN	40.7%	18.2%	—
5 CNNs	38.1%	16.4%	**16.4%**
1 CNN*	39.0%	16.6%	—
7 CNNs*	36.7%	15.4%	**15.3%**

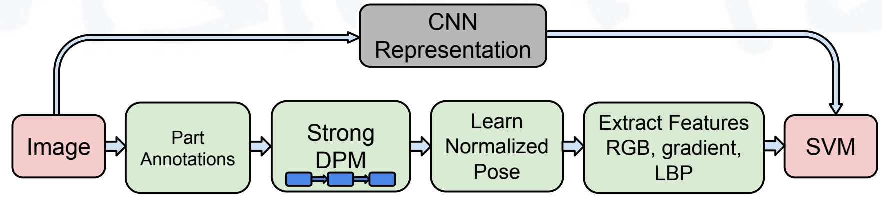
Table 2: Comparison of error rates on ILSVRC-2012 validation and test sets. In *italics* are best results achieved by others. Models with an asterisk* were "pre-trained" to classify the entire ImageNet 2011 Fall release. See Section 6 for details.

[1] Krizhevsky, Sutskever and Hinton. ImageNet classification with deep convolutional neural networks, NIPS'12.
[2] Sermanet, Eigen, Zhang, Mathieu, Fergus, LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. ICLR'14.
[3] Donahue, Jia, Vinyals, Hoffman, Zhang, Tzeng, Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. ICML'14.

Transfer from ImageNet to Smaller Datasets

- CNNs + SMVs [1]

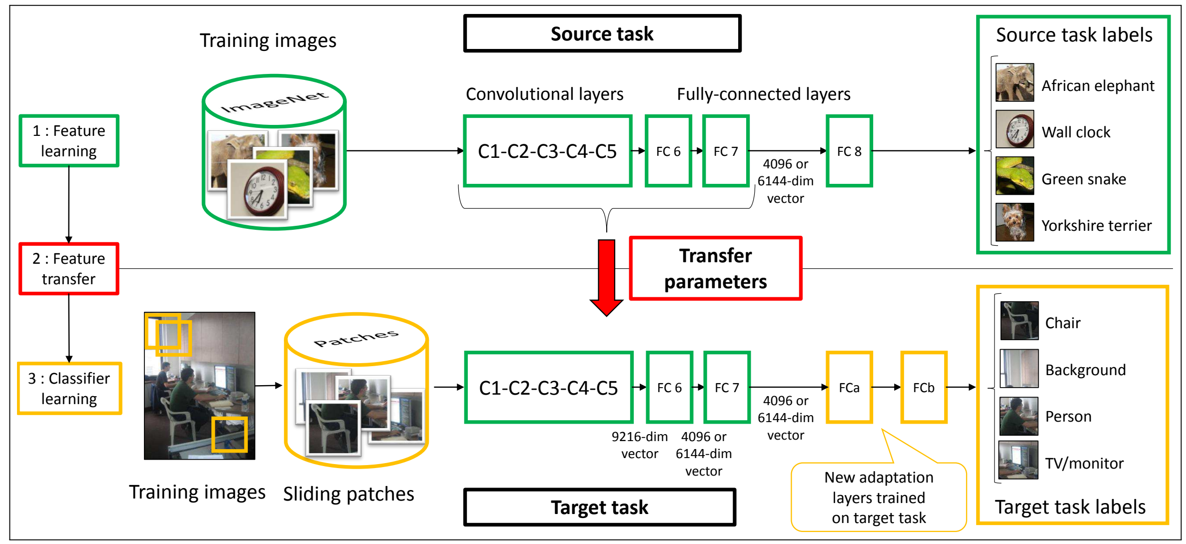
Pascal VOC 2007	mAP
GHM[8]	64.7
AGS[11]	71.1
NUS[39]	70.5
CNN-SVM	73.9
CNNaug-SVM	**77.2**



- Retrain top-layer [2]

Pascal VOC 2007	mAP
INRIA [32]	59.4
NUS-PSL [44]	70.5
PRE-1000C	**77.7**

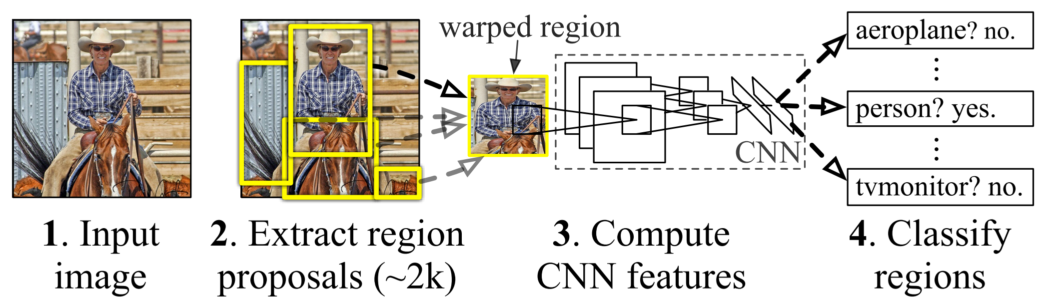
Pascal VOC 2012	mAP
NUS-PSL [49]	82.2
NO PRETRAIN	70.9
PRE-1000C	78.7
PRE-1000R	76.3
PRE-1512	**82.8**



- CNNs + SVMs for object detection [3]

VOC 2010 test	mAP
DPM v5 [20]†	33.4
UVA [39]	35.1
Regionlets [41]	39.7
SegDPM [18]†	40.4
R-CNN	50.2
R-CNN BB	**53.7**

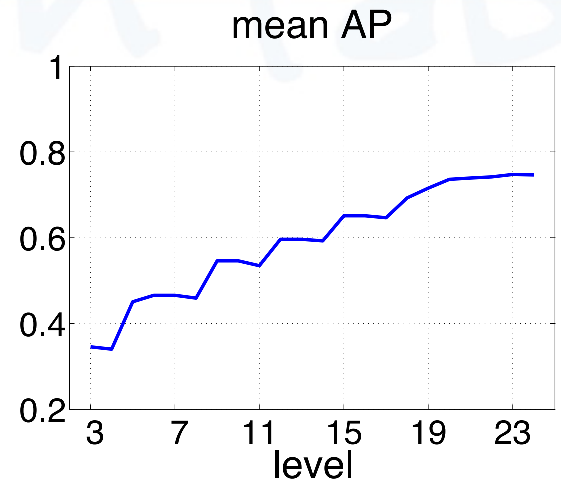
R-CNN: *Regions with CNN features*



[1] Razavian, Azizpour, Sullivan, Carlsson, CNN Features off-the-shelf: an Astounding Baseline for Recognition. CVPRW'14.
[2] Oquab, Bottou, Laptev, Sivic. Learning and transferring mid-level image representations using convolutional neural networks CVPR'14
[3] Girshick, Donahue, Darrell and Malik. Rich feature hierarchies for accurate object detection and semantic segmentation, CVPR'14

Center for IMAGING SCIENCE

Why this Performance?

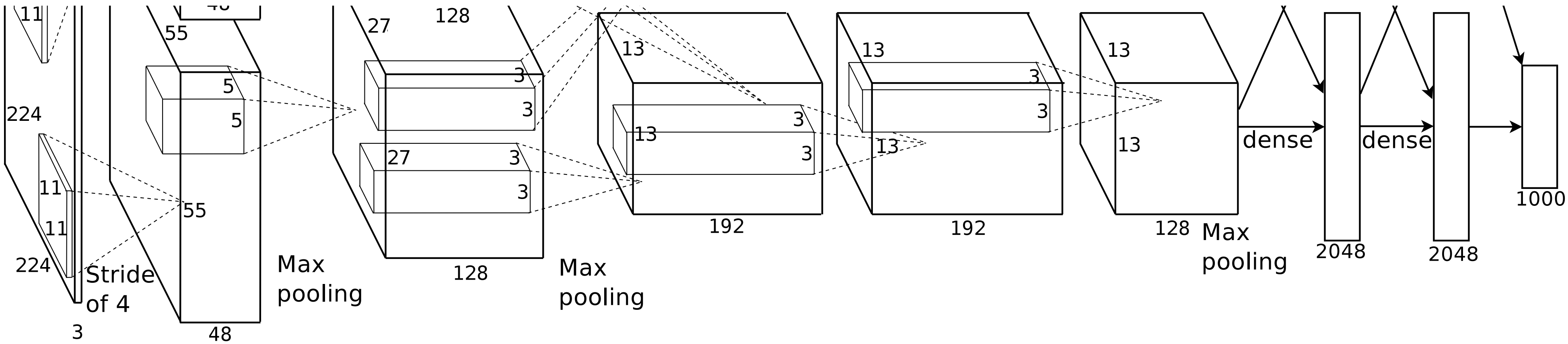


- More layers [1]
  - Multiple layers capture more invariances
  - Features are learned rather than hand-crafted
- More data
  - There is more data to train deeper networks
- More computing
  - GPUs go hand in hand with learning methods
- First attempt at a theoretical justification [2]
  - Theoretical support for invariance via scattering transform
  - Each layer must be a contraction to keep data volume bounded
  - Optimization issues are not discussed: stage-wise learning is used

[1] Razavian, Azizpour, Sullivan, Carlsson, CNN Features off-the-shelf: an Astounding Baseline for Recognition. CVPRW'14.
[2] Mallat and Waldspurger. Deep Learning by Scattering, arXiv 2013

What About Optimization?

- The learning problem is non-convex



$$\Phi(X^1,\ldots,X^K) = \psi_K(\cdots\psi_2(\psi_1(VX^1)X^2)\cdots X^K)$$

nonlinearity features weights

$$\min_{X^1,\ldots,X^K} \ell(Y,\Phi(X^1,\ldots,X^K)) + \lambda\Theta(X^1,\ldots,X^K)$$

loss labels regularizer

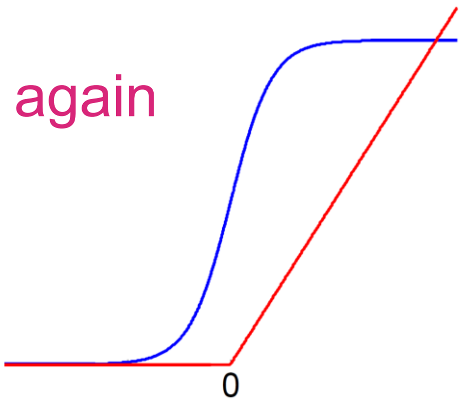
What About Optimization?

- The learning problem is non-convex

$$\min_{X^1,\ldots,X^K} \ell(Y,\Phi(X^1,\ldots,X^K)) + \lambda\Theta(X^1,\ldots,X^K)$$

  - Back-propagation, alternating minimization, descent method

- To get a good local minima
  - Random initialization
  - If training error does not decrease fast enough, start again
  - Repeat multiple times



- Mysteries
  - One can find many solutions with similar objective values
  - Rectified linear units work better than sigmoid/hyperbolic tangent

Contributions

$$\min_{X^1,\ldots,X^K} \ell(Y,\Phi(X^1,\ldots,X^K)) + \lambda\Theta(X^1,\ldots,X^K)$$

- **Assumptions:**
  - ℓ is convex and once differentiable
  - Φ and Θ are sums of positively homogeneous functions

$$f(\alpha X^1,\ldots,\alpha X^K) = \alpha^p f(X^1,\ldots,X^K) \quad \forall\alpha \geq 0$$

- **Theorem 1:** A local minimizer such that for some *i* and all k $X_i^k = 0$, then it is a global minimizer

- **Theorem 2:** If the size of the network is large enough, local descent can reach a global minimizer from any initialization

Outline

- Globally Optimal Low-Rank Matrix Factorizations [1,2]
  - PCA, Robust PCA, Matrix Completion
  - Nonnegative Matrix Factorization
  - Dictionary Learning
  - Structured Matrix Factorization

- Globally Optimal Positively Homogeneous Factorizations [2]
  - Tensor Factorization
  - Deep Learning

[1] Haeffele, Young, Vidal. Structured Low-Rank Matrix Factorization: Optimality, Algorithm, and Applications to Image Processing, ICML '14
[2] Haeffele, Vidal. Global Optimality in Tensor Factorization, Deep Learning and Beyond, arXiv, '15

JHU vision lab

Globally Optimal Matrix Factorizations

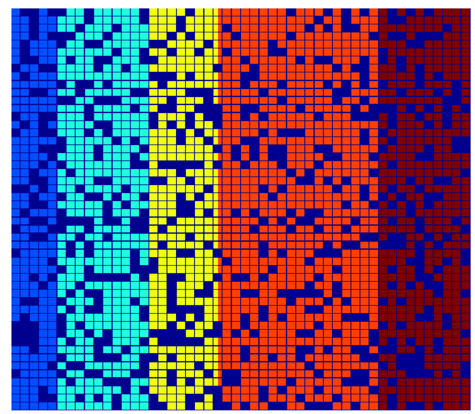
René Vidal
Center for Imaging Science
Institute for Computational Medicine

THE DEPARTMENT OF BIOMEDICAL ENGINEERING
The Whitaker Institute at Johns Hopkins

Center for IMAGING SCIENCE

Low Rank Modeling

- Models involving factorization are ubiquitous
  - PCA
  - Robust PCA
  - Matrix Completion
  - Nonnegative Matrix Factorization
  - Dictionary Learning

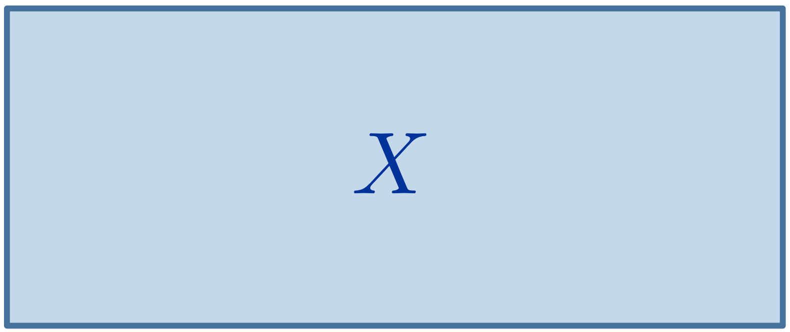


http://perception.csl.illinois.edu/matrix-rank/home.html

Typical Low-Rank Formulations

- Convex formulations

$$\min_{X} \ell(Y, X) + \lambda\Theta(X)$$

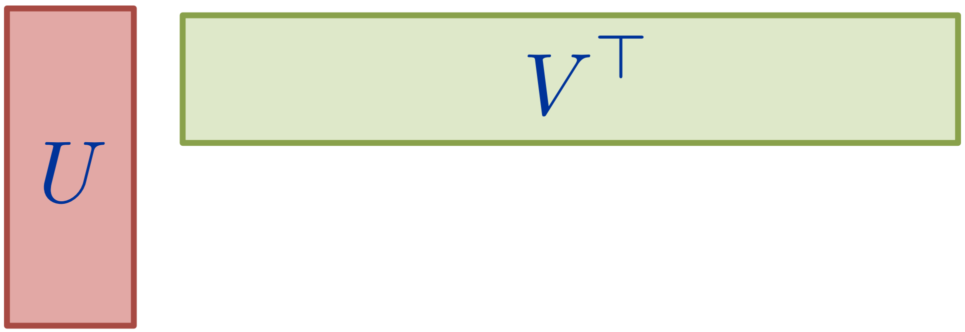


  – Robust PCA
  – Matrix Completion

- Convex
- Large problem size
- Unstructured factors

- Factorized formulations

$$\min_{U,V} \ell(Y, UV^\top) + \lambda\Theta(U, V)$$



  – Nonnegative matrix factorization
  – Dictionary learning

- Non-Convex
- Small problem size
- Structured factors

Why Do We Need Structure?

- Given a low-rank video $Y \in \mathbb{R}^{p \times t}$ $\quad \min_{X} \|Y - X\|_1 + \lambda \|X\|_*$

(a) Original frames

(b) Low-rank $\hat{L}$

(c) Sparse $\hat{S}$

$$\min_{U,V} \ell(Y, UV^\top) + \lambda \Theta(U, V)$$

- U: spatial basis
  - Low total-variation
  - Non-negative
- V: temporal basis
  - Sparse on particular basis set
  - Non-negative

[1] Candes, Li, Ma, Wright. Robust Principal Component Analysis? Journal of the ACM, 2011.

Need for Structured Factors

- Nonnegative matrix factorization

$$\min_{U,V} \|Y - UV^\top\|_F^2 \quad \text{s.t.} \quad U \geq 0, V \geq 0$$

- Sparse dictionary learning

$$\min_{U,V} \|Y - UV^\top\|_F^2 \quad \text{s.t.} \quad \|U_i\|_2 \leq 1, \|V_i\|_0 \leq r$$

- **Challenges to state-of-the-art methods**
  - Need to pick size of U and V a priori
  - Alternate between U and V, without guarantees of convergence to a global minimum

Tackling Non-Convexity: Nuclear Norm Case

- Convex problem

$$\min_{X} \ell(Y, X) + \lambda \|X\|_*$$

Factorized problem

$$\min_{U,V} \ell(Y, UV^\top) + \lambda \boxed{\Theta(U,V)}$$

- Variational form of the nuclear norm

$$\|X\|_* = \min_{U,V} \boxed{\sum_{i=1}^{r} |U_i|_2 |V_i|_2} \quad \text{s.t.} \quad UV^\top = X$$

- **Theorem:** Assume loss ℓ is convex and once differentiable. A local minimizer of the factorized problem such that for some i $U_i = V_i = 0$ is a global minimizer of the convex problem

- **Intuition:** regularizer Θ “comes from a convex function”

Tackling Non-Convexity: General Case

- A natural generalization is the projective tensor norm [1,2]

$$\|X\|_{u,v} = \min_{U,V} \sum_{i=1}^{r} \|U_i\|_u \|V_i\|_v \quad \text{s.t.} \quad UV^\top = X$$

- **Theorem [3,4]:** A local minimizer of the factorized problem

$$\min_{U,V} \ell(Y, UV^\top) + \lambda \sum_{i=1}^{r} \|U_i\|_u \|V_i\|_v$$

such that for some i $U_i = V_i = 0$, is a global minimizer of both the factorized problem and of the convex problem

$$\min_{X} \ell(Y, X) + \lambda \|X\|_{u,v}$$

[1] Bach, Mairal, Ponce, Convex sparse matrix factorizations, arXiv 2008.
[2] Bach. Convex relaxations of structured matrix factorizations, arXiv 2013.
[3] Haeffele, Young, Vidal. Structured Low-Rank Matrix Factorization: Optimality, Algorithm, and Applications to Image Processing, ICML '14
[4] Haeffele, Vidal. Global Optimality in Tensor Factorization, Deep Learning and Beyond, arXiv '15

Example: Nonnegative Matrix Factorization

- Original formulation

$$\min_{U,V} \|Y - UV^\top\|_F^2 \quad \text{s.t.} \quad U \geq 0, V \geq 0$$

- New factorized formulation

$$\min_{U,V} \|Y - UV^\top\|_F^2 + \lambda \sum_i |U_i|_2 |V_i|_2 \ \ \text{s.t.} \ \ U, V \geq 0$$

  - Note: regularization limits the number of columns in (U,V)

Example: Sparse Dictionary Learning

- Original formulation

$$\min_{U,V} \|Y - UV^\top\|_F^2 \quad \text{s.t.} \quad \|U_i\|_2 \leq 1, \|V_i\|_0 \leq r$$

- New factorized formulation

$$\min_{U,V} \|Y - UV^\top\|_F^2 + \lambda \sum_i |U_i|_2(|V_i|_2 + \gamma |V_i|_1)$$

Non Example: Robust PCA

- Original formulation [1]

$$\min_{X,E} \|E\|_1 + \lambda \|X\|_* \quad \text{s.t.} \quad Y = X + E$$

- Equivalent formulation

$$\min_{X} \|Y - X\|_1 + \lambda \|X\|_*$$

- New factorized formulation

$$\min_{U,V} \|Y - UV^\top\|_1 + \lambda \sum_i |U_i|_2 |V_i|_2$$

- Not an example because loss is not differentiable

[1] Candes, Li, Ma, Wright. Robust Principal Component Analysis? Journal of the ACM, 2011.

Optimization

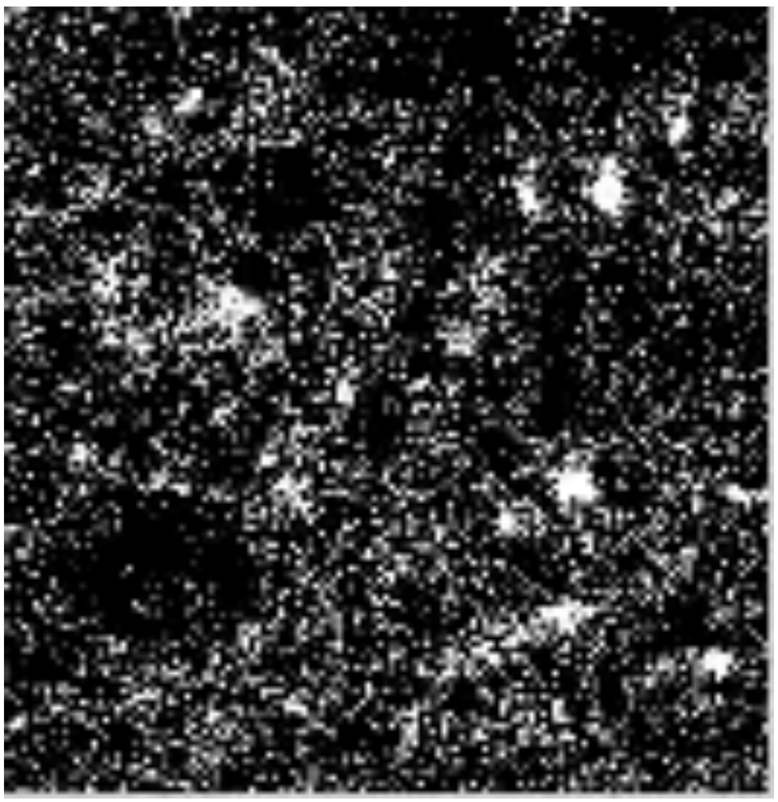
$$\min_{U,V} \ell(Y, UV^\top) + \lambda \sum_{i=1}^{r} \|U_i\|_u \|V_i\|_v$$

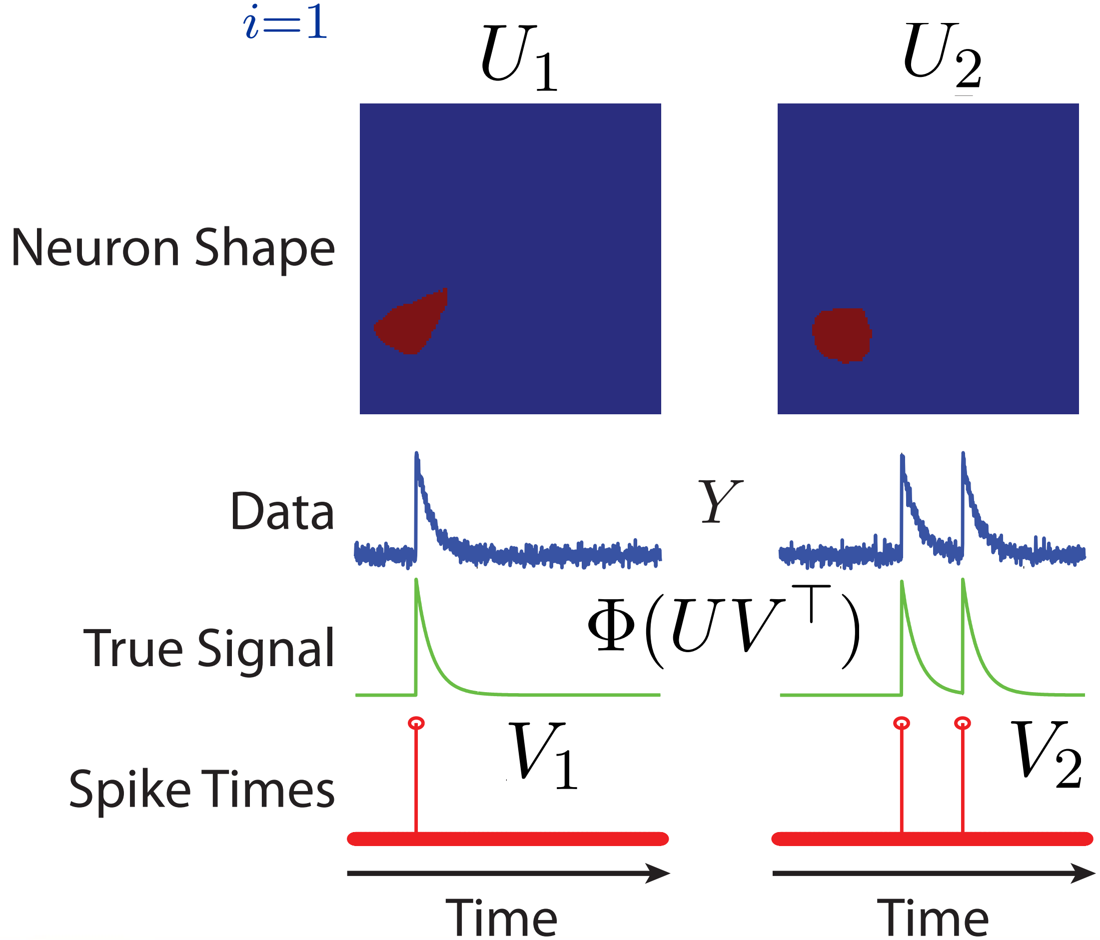
- Convex in U given V and vice versa
- Alternating proximal gradient descent
  - Calculate gradient of smooth term
  - Compute proximal operator
  - Acceleration via extrapolation
- Advantages
  - Easy to implement
  - Highly parallelizable
  - Guaranteed convergence to Nash equilibrium (may not be local min)

Neural Calcium Image Segmentation

- Find neuronal shapes and spike trains in calcium imaging

$$\min_{U,V} \|Y - \Phi(UV^\top)\|_F^2 + \lambda \sum_{i=1}^{r} \|U_i\|_u \|V_i\|_v$$



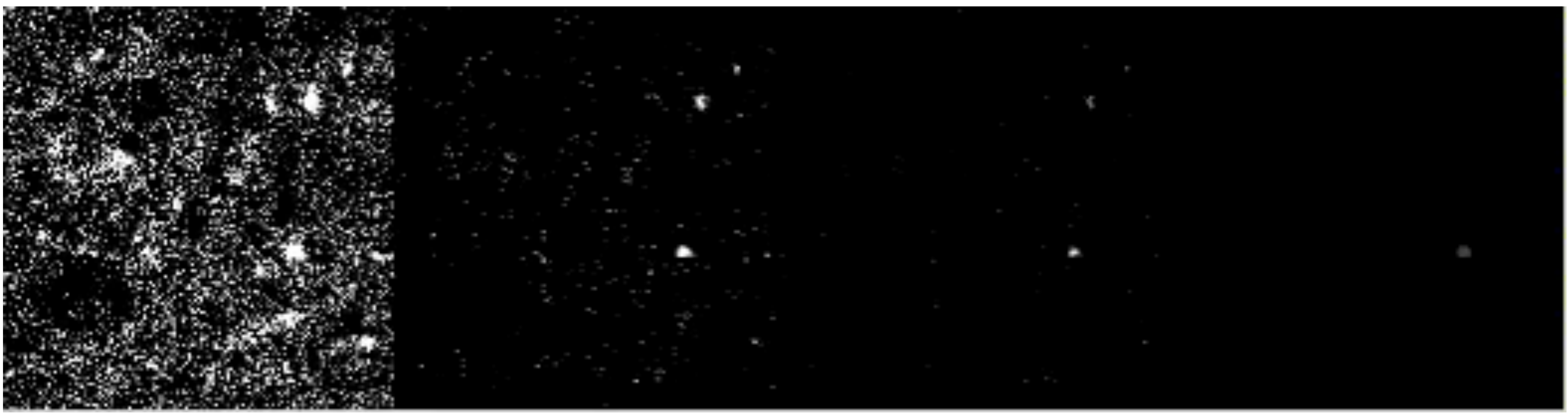


Neural Calcium Image Segmentation

$$\min_{U,V} \|Y - \Phi(UV^\top)\|_F^2 + \lambda \sum_{i=1}^{r} \|U_i\|_u \|V_i\|_v$$

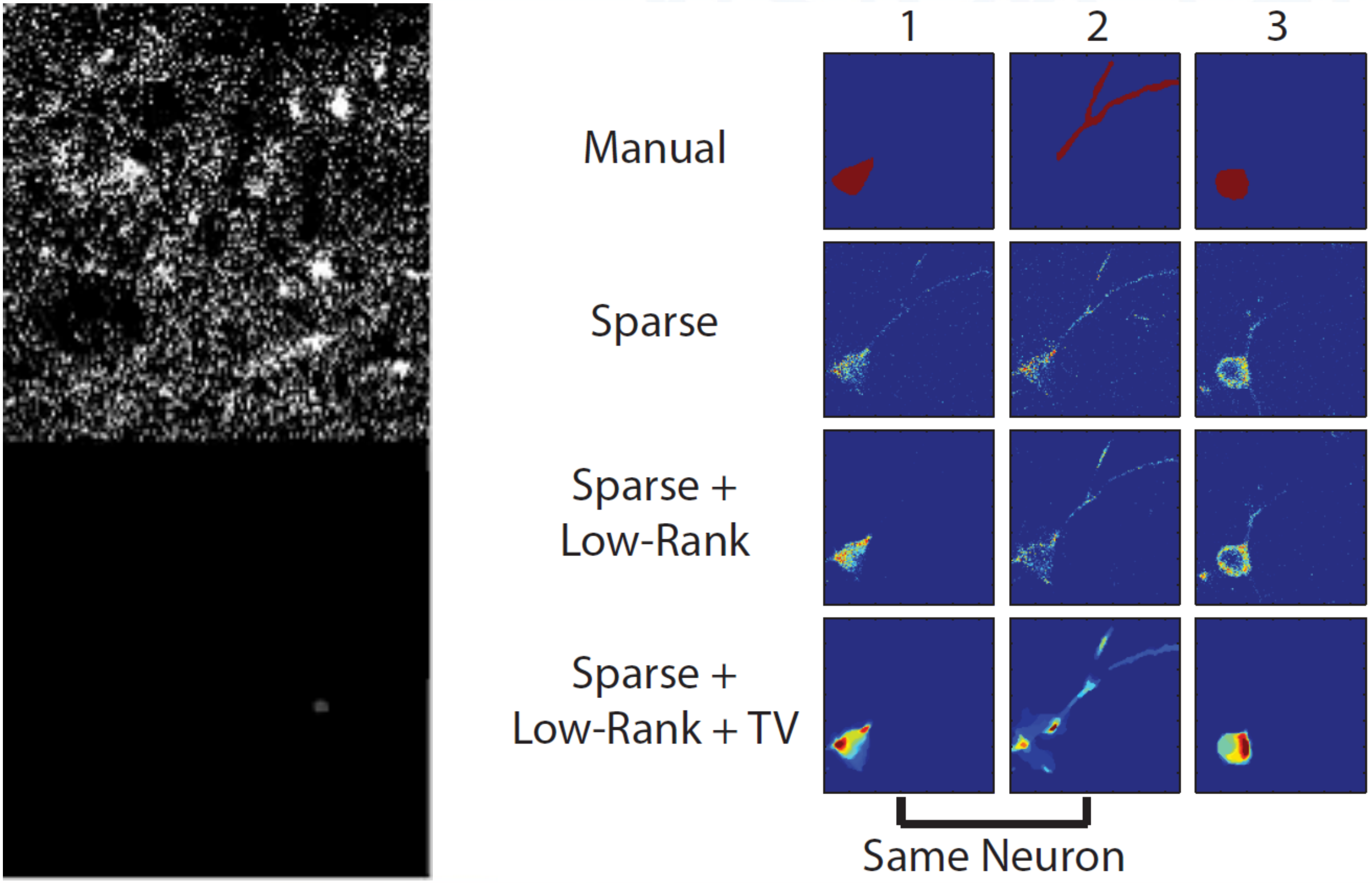
$$\|\cdot\|_u = \|\cdot\|_2 + \|\cdot\|_1 + \|\cdot\|_{TV}$$

$$\|\cdot\|_v = \|\cdot\|_2 + \|\cdot\|_1$$



Raw Data | Sparse | + Low Rank | +Total Variation

Neural Calcium Image Segmentation



Hyperspectral Compressed Recovery

- Prior method: NucTV (Golbabaee et al., 2012)

$$\min_{X} \|X\|_* + \lambda \sum_{i=1}^{t} \|X_i\|_{TV} \quad \text{s.t.} \quad \|Y - \Phi(X)\|_F^2 \leq \epsilon$$

- 180 Wavelengths
- 256 x 256 Images
- Computation per Iteration
  - SVT of whole image volume
  - 180 TV Proximal Operators
  - Projection onto Constraint Set

Center for IMAGING SCIENCE

Hyperspectral Compressed Recovery

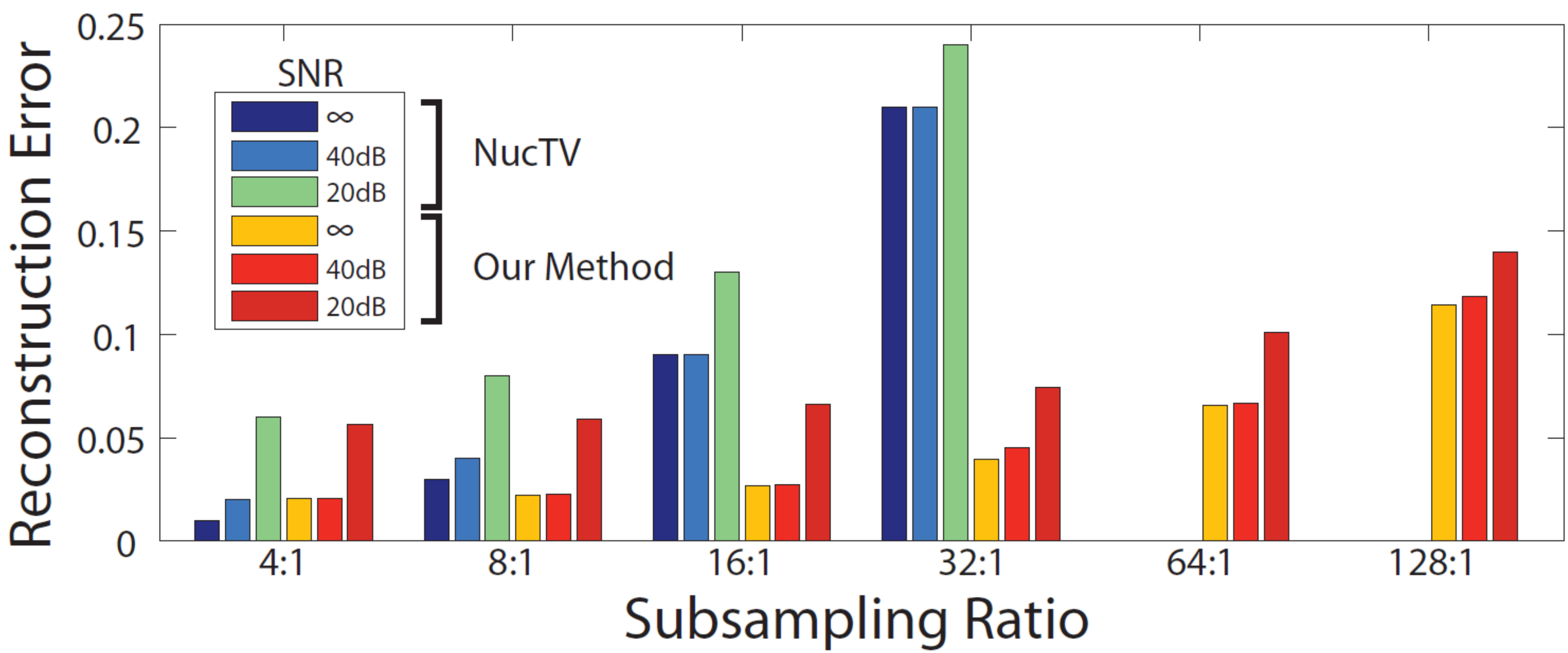
- Our method

$$\min_{U,V} \|Y - \Phi(UV^\top)\|_F^2 + \lambda \sum_{i=1}^{r} \|U_i\|_u \|V_i\|_v$$

- (U,V) have 15 columns
- Problem size reduced by 91.6%
- Computation per Iteration
  - Calculate gradient
  - 15 TV Proximal Operators
- Random Initializations

Hyperspectral Compressed Recovery

$$\frac{\|X_{true} - UV^\top\|_F}{\|X_{true}\|_F}$$



Center for Imaging Science

Conclussions

- Structured Low Rank Matrix Factorization
  - Structure on the factors captured by the Projective Tensor Norm
  - Efficient optimization for Large Scale Problems

- Local minima of the non-convex factorized form are global minima of the convex form

- Advantages in Applications
  - Neural calcium image segmentation
  - Compressed recovery of hyperspectral images

JHU vision lab

Globally Optimal Tensor Factorization, Deep Learning, and Beyond

René Vidal
Center for Imaging Science
Institute for Computational Medicine

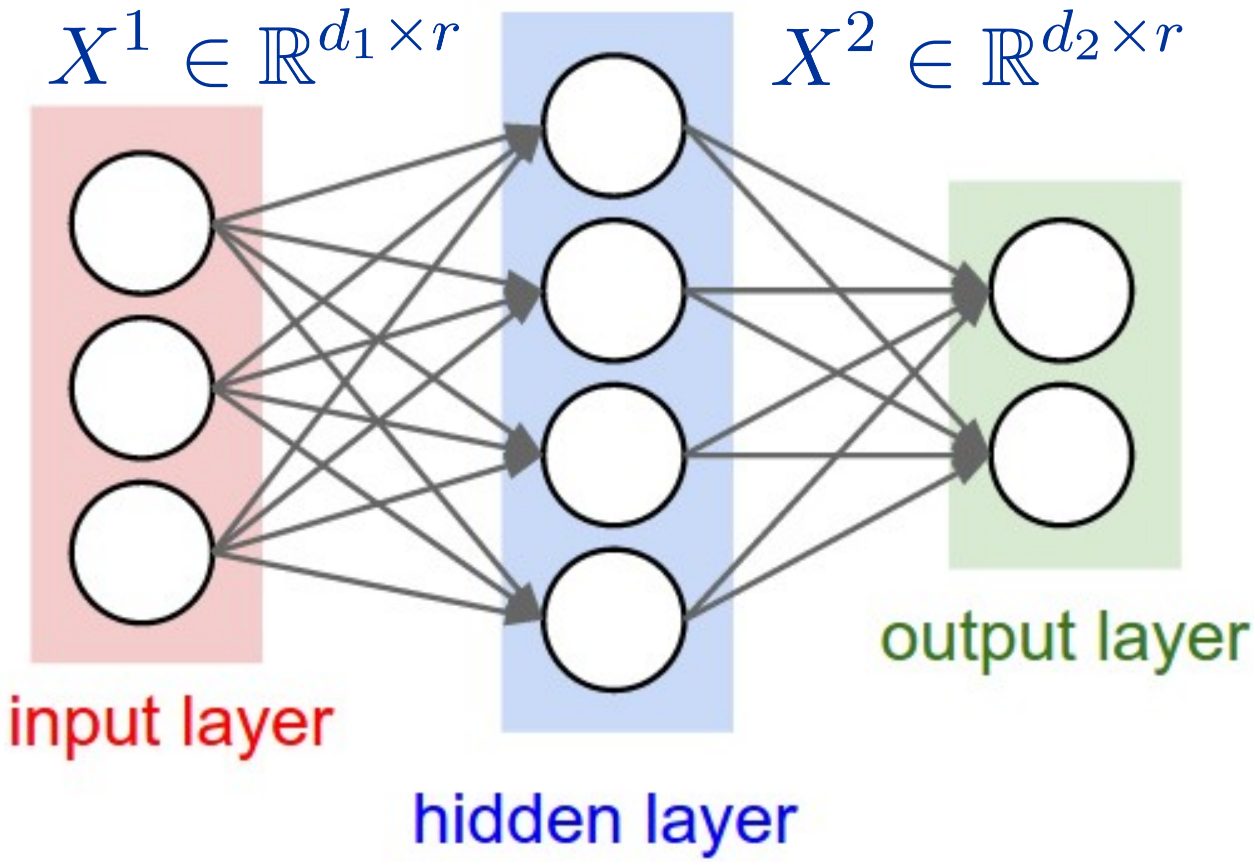
From Matrix Factorizations to Deep Learning

- Two-layer NN

$$\psi_1(x) = \max(x, 0)$$
$$\Phi(X^1, X^2) = \psi_1(VX^1)(X^2)^\top$$

$X^1 \in \mathbb{R}^{d_1\times r}$ $\quad$ $X^2 \in \mathbb{R}^{d_2\times r}$

$V \in \mathbb{R}^{N\times d_1}$



From Matrix Factorizations to Deep Learning

- Recall the generalized factorization problem

$$\min_{X^1,\ldots,X^K} \ell(Y, \Phi(X^1,\ldots,X^K)) + \lambda\Theta(X^1,\ldots,X^K)$$

- Matrix factorization is a particular case where K=2

$$\Phi(U,V) = \sum_{i=1}^{r} U_i V_i^\top, \;\; \Theta(U,V) = \sum_{i=1}^{r} \|U_i\|_u \|V_i\|_v$$

- Both Φ and Θ are sums of positively homogeneous functions

$$f(\alpha X^1,\ldots,\alpha X^K) = \alpha^p f(X^1,\ldots,X^K) \quad \forall \alpha \geq 0$$

- Other examples
  - Rectified linear unit + max pooling is pos. homogeneous of degree 1

"Matrix Multiplication" for K > 2

- In matrix factorization we have

$$\Phi(U,V) = UV^\top = \sum_{i=1}^{r} U_i V_i^\top$$

- By analogy we define

$$\Phi(X^1,\dots,X^K) = \sum_{i=1}^{r} \phi(X_i^1,\dots,X_i^K)$$

where X^k is a tensor, X_i^k is its i-th slice along its last dimension, and ϕ is a positively homogeneous function

- Examples
  - Matrix multiplication: $\phi(X^1, X^2) = X^1 {X^2}^\top$
  - Tensor product: $\phi(X^1,\dots,X^K) = X^1 \otimes \cdots \otimes X^K$
  - ReLU neural network: $\phi(X^1,\dots,X^K) = \psi_K(\cdots\psi_2(\psi_1(VX^1)X^2)\cdots X^K)$

"Projective Tensor Norm" for K > 2

- In matrix factorization we have

$$\|X\|_{u,v} = \min_{U,V} \sum_{i=1}^{r} \|U_i\|_u \|V_i\|_v \quad \text{s.t.} \quad UV^\top = X$$

- By analogy we define

$$\Omega_{\phi,\theta}(X) = \min_{\{X^k\}} \sum_{i=1}^{r} \theta(X_i^1, \dots, X_i^K) \text{ s.t. } \Phi(X^1, \dots, X^K) = X$$

where θ is a positively homogeneous function

- **Proposition:** $\Omega_{\phi,\theta}$ is convex

Main Result

- Theorem: A local minimizer of the factorized formulation

$$\min_{\{X^k\}} \ell\Big(Y, \sum_{i=1}^{r} \phi(X_i^1, \ldots, X_i^K)\Big) + \lambda \sum_{i=1}^{r} \theta(X_i^1, \ldots, X_i^K)$$

  such that for some i and for all k we have $X_i^k = 0$, gives a global minimizer for both the factorized formulation and the convex formulation

$$\min_{X} \ell(Y, X) + \lambda \Omega_{\phi,\theta}(X)$$

- Examples
  - Matrix factorization
  - Tensor factorization

Conclussions

- For many non-convex factorization problems, such as matrix factorization, tensor factorization, and deep learning, a local minimizer for the factors gives a global minimizer
- For matrix factorization, this
  - allows one to incorporate structure on the factors, and
  - gives efficient optimization method suitable for large problems
- For deep learning, this provides theoretical insights on why
  - many local minima give similar objective values
  - ReLU works better than sigmoidal functions
- While alternating minimization is efficient and guaranteed to converge, it is not guaranteed to converge to a local minimum

Acknowledgements

- PhD Students
  - Ben Haeffele, JHU

- Collaborators
  - Eric Young, JHU

- Grants
  - NIH DC00115
  - NIH DC00032
  - NSF 1218709
  - NSF 1447822

[1] Haeffele, Young, Vidal. Structured Low-Rank Matrix Factorization: Optimality, Algorithm, and Applications to Image Processing, ICML '14
[2] Haeffele, Vidal. Global Optimality in Tensor Factorization, Deep Learning and Beyond, arXiv, '15

Center for IMAGING SCIENCE

More Information,

Vision Lab @ Johns Hopkins University

http://www.vision.jhu.edu

Center for Imaging Science @ Johns Hopkins University

http://www.cis.jhu.edu

Thank You!