Sparse Modeling
in
Deep Learning

Michael Elad

Computer Science Department
The Technion - Israel Institute of Technology
Haifa 32000, Israel

ICML Workshop: July 14th, 2018

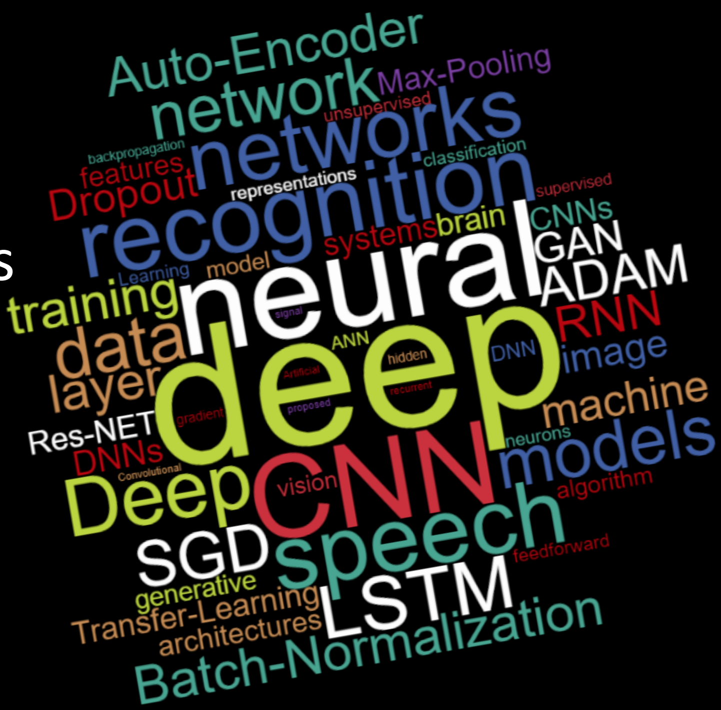
The research leading to these results has been received funding from the European union's Seventh Framework Program (FP/2007-2013) ERC grant Agreement ERC-SPARSE- 320649

This Lecture is About …

A Proposed Theory for Deep-Learning (DL)

Explanation:

- DL has been extremely successful in solving a variety of learning problems
- DL is an empirical field, with numerous tricks and know-how, but almost no theoretical foundations
- A theory for DL has become the holy-grail of current research in Machine-Learning and related fields



Who Needs Theory ?

We All Do !!

… because … A theory

- … could bring the next rounds of ideas to this field, breaking existing barriers and opening new opportunities
- … could map clearly the limitations of existing DL solutions, and point to key features that control their performance
- … could remove the feeling with many of us that DL is a "dark magic", turning it into a solid scientific discipline

Ali Rahimi
NIPS 2017
Test-of-Time
Award

"Machine learning has become alchemy"

Yan LeCun

Understanding is a good thing … but another goal is inventing methods. In the history of science and technology, engineering preceded theoretical understanding:

- Lens & telescope → Optics
- Steam engine → Thermodynamics
- Airplane → Aerodynamics
- Radio & Comm. → Info. Theory
- Computer → Computer Science

A Theory for DL ?

Raja Giryes (TAU): Studied the architecture of DNN in the context of their ability to give distance-preserving embedding of signals

Gitta Kutyniok (TU) & Helmut Bolcskei (ETH): Studied the ability of DNN architectures to approximate families of functions

Stephane Mallat (ENS) & Joan Bruna (NYU): Proposed the scattering transform (wavelet-based) and emphasized the treatment of invariances in the input data

Richard Baraniuk & Ankit Patel (RICE): Offered a generative probabilistic model for the data, showing how classic architectures and learning algorithms relate to it

Data

Architecture

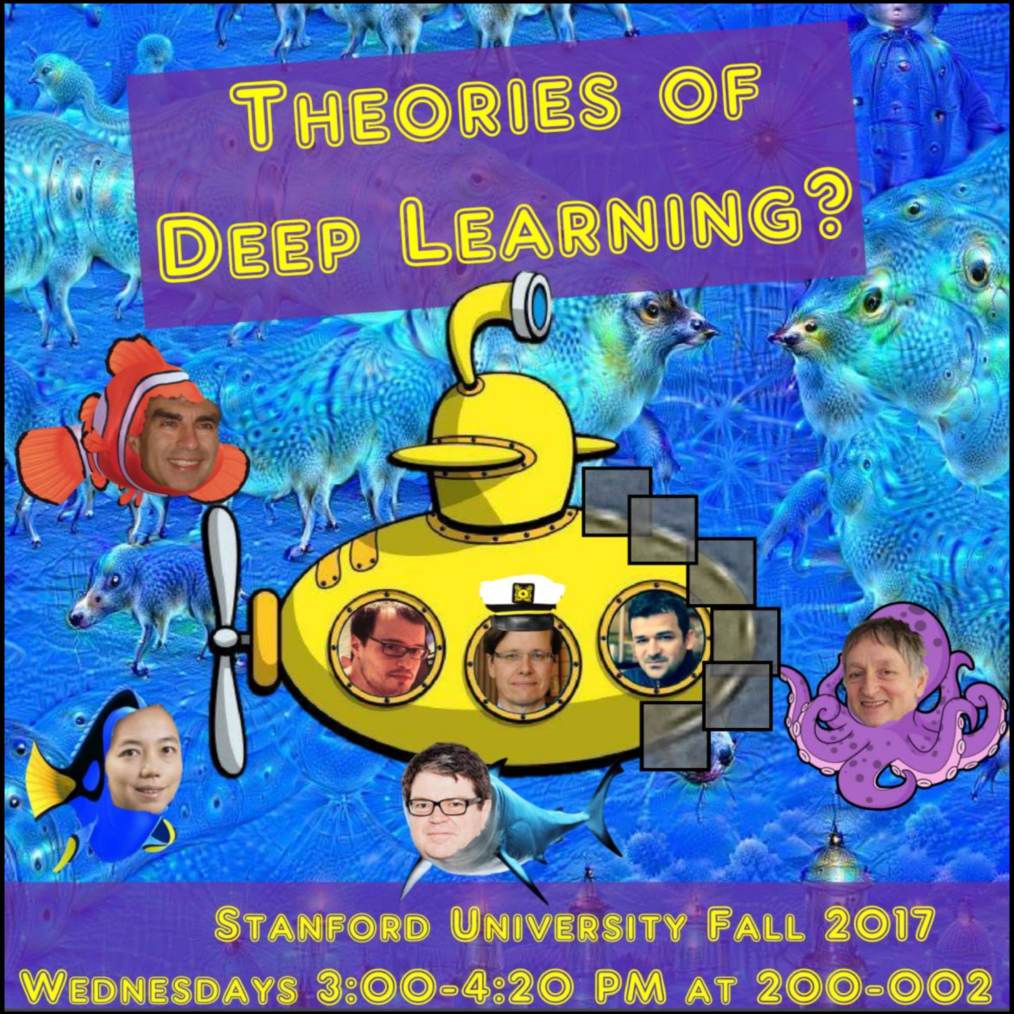
Algorithms

Rene Vidal (JHU): Explained the ability to optimize the typical non-convex objective and yet get to a global minima

Naftali Tishby (HUJI): Introduced the Information Bottleneck (IB) concept and demonstrated its relevance to deep learning

Stefano Soatto's team (UCLA): Analyzed the Stochastic Gradient Descent (SGD) algorithm, connecting it to the IB objective

So, is there a Theory for DL ?



The answer is tricky:

There are already various such attempts, and some of them are truly impressive

… but …

none of them is complete

Interesting Observations

- Theory origins: Signal Proc., Control Theory, Info. Theory, Harmonic Analysis, Sparse Represen., Quantum Physics, PDE, Machine learning …

Ron Kimmel: "*DL is a dark monster covered with mirrors. Everyone sees his reflection in it …*"

David Donoho: "*… these mirrors are taken from Cinderella's story, telling each that he is the most beautiful*"

- Today's talk is on our proposed theory:


Yaniv Romano


Vardan Papyan


Jeremias Sulam


Aviad Aberdam

√ Data

√ Architecture

Algorithms

… and yes, our theory is the best 

Our Story: More Specifically

Sparseland
Sparse Representation Theory

CSC
Convolutional Sparse Coding

ML-CSC
Multi-Layered Convolutional Sparse Coding

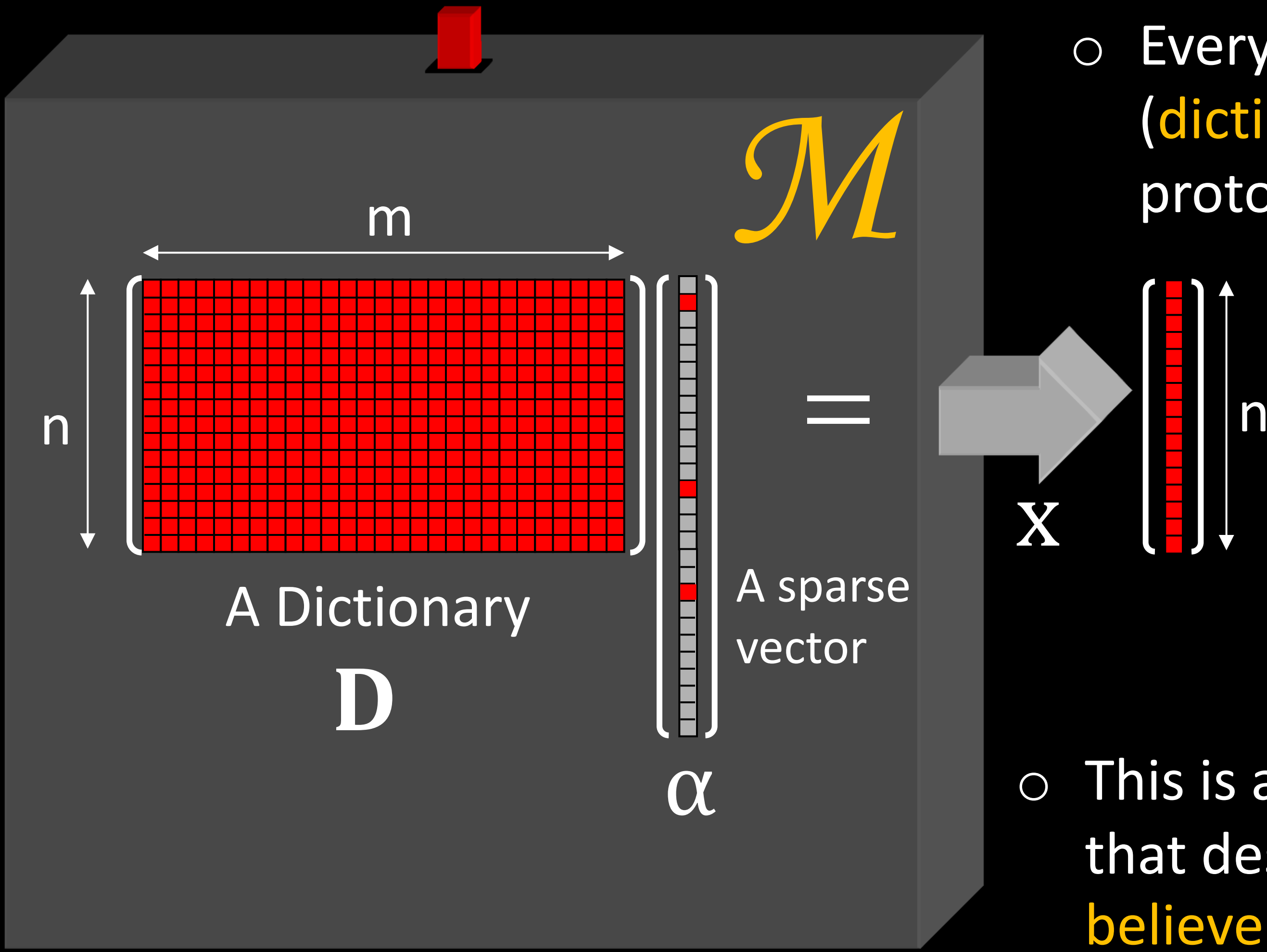
Sparsity-Inspired Models → Deep-Learning

- In this talk we shall start with a brief overview of the first two models, and then step directly to the ML-CSC model and its connection to deep-learning
- If you feel that you are missing key information, you can complement this by viewing my YouTube IPAM talk from February 2018

Michael Elad
The Computer-Science Department
The Technion

Brief Background on Sparse Modeling

Sparseland : A Formal Description



- Every column in **D** (dictionary) is a prototype signal (atom)
  - The vector $\underline{\alpha}$ is generated with few non-zeros at arbitrary locations and values
- This is a generative model that describes how (we believe) signals are created

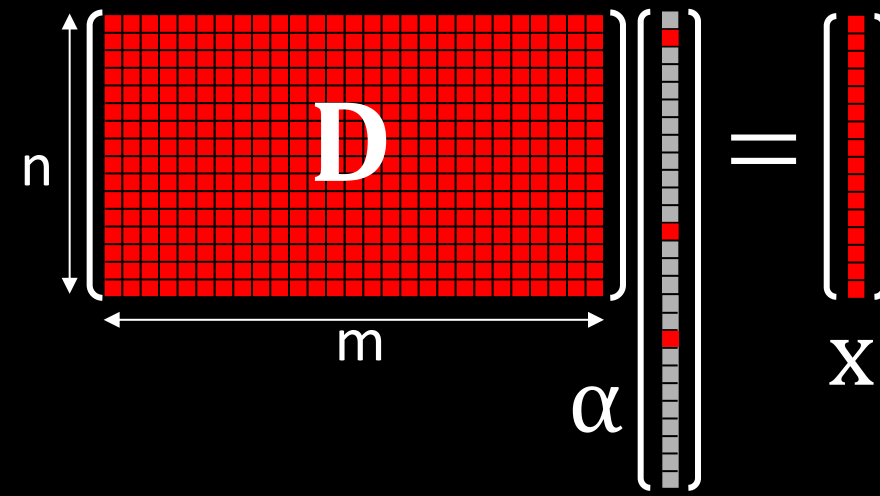
Michael Elad
The Computer-Science Department
The Technion

Atom Decomposition

$$\min_\alpha \|\alpha\|_0 \ \text{ s.t. } \ x = \mathbf{D}\alpha$$

$$\min_\alpha \|\alpha\|_0 \ \text{ s.t. } \ \|\mathbf{D}\alpha - y\|_2 \le \varepsilon$$

Approximation Algorithms



Relaxation methods

Basis-Pursuit

Greedy methods

Thresholding/OMP

- L_0 – counting number of non-zeros in the vector
- This is a projection onto the *Sparseland* model
- These problems are known to be NP-Hard problem

Pursuit Algorithms

$$\min_\alpha \|\alpha\|_0 \ \text{ s.t. } \ \|\mathbf{D}\alpha - y\|_2 \le \varepsilon$$

Approximation Algorithms

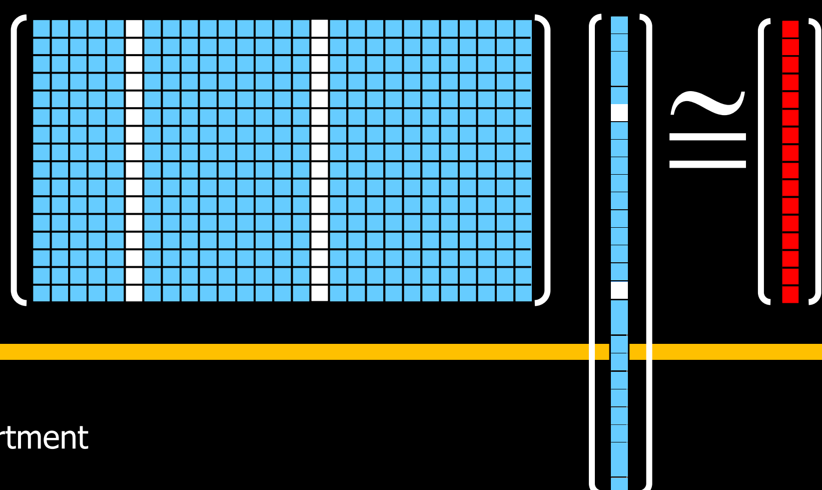
Basis Pursuit

Change the L_0 into L_1 and then the problem becomes convex and manageable

$$\min_\alpha \|\alpha\|_1 \quad \text{s.t.} \quad \|\mathbf{D}\alpha - y\|_2 \le \varepsilon$$

Matching Pursuit

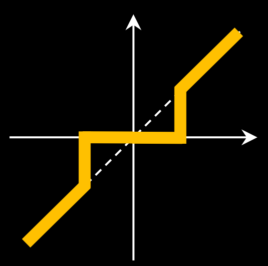
Find the support greedily, one element at a time



Thresholding

Multiply y by $\mathbf{D}^T$ and apply shrinkage:

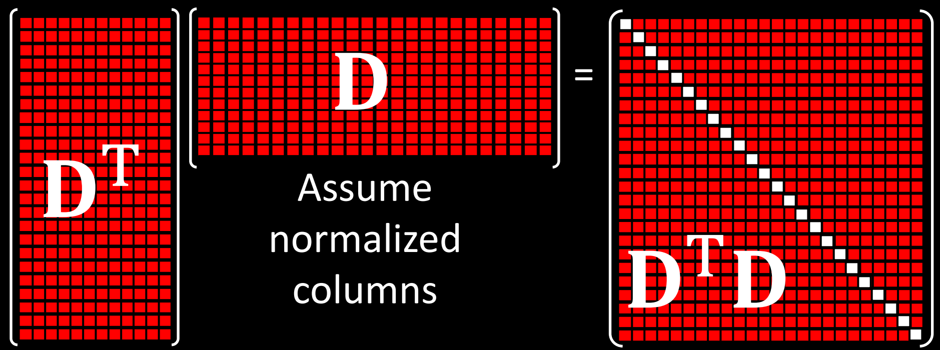
$$\hat{\alpha} = \mathcal{P}_\beta\{\mathbf{D}^T y\}$$



Michael Elad
The Computer-Science Department
The Technion

The Mutual Coherence

- Compute



- The Mutual Coherence $\mu(\mathbf{D})$ is the largest off-diagonal entry in absolute value
- We will pose all the theoretical results in this talk using this property, due to its simplicity
- You may have heard of other ways to characterize the dictionary (Restricted Isometry Property - RIP, Exact Recovery Condition - ERC, Babel function, Spark, …)

Basis-Pursuit Success

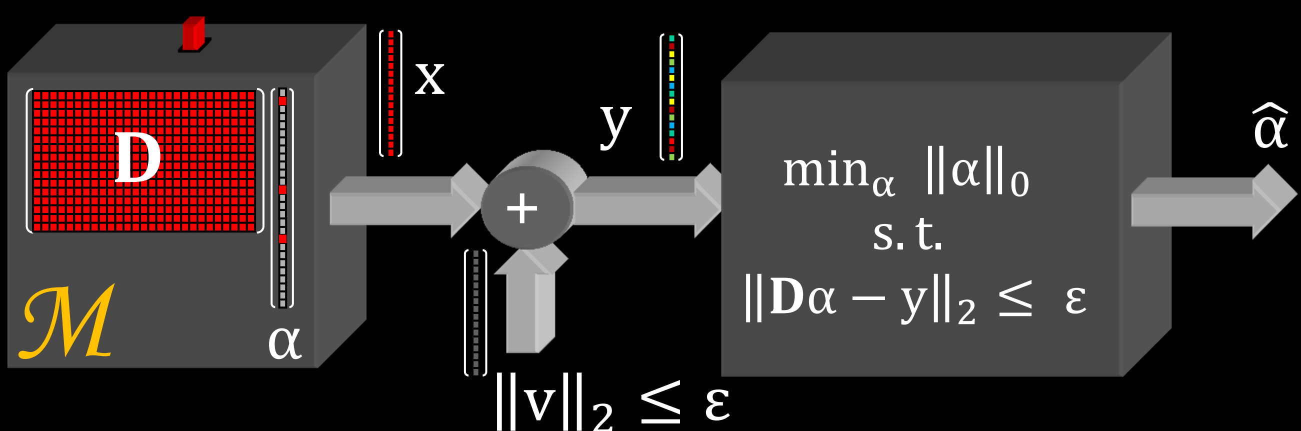
Theorem: Given a noisy signal $y = \mathbf{D}\alpha + v$ where $\|v\|_2 \le \varepsilon$ and α is sufficiently sparse,

$$\|\alpha\|_0 < \frac{1}{4}\left(1 + \frac{1}{\mu}\right)$$

then Basis-Pursuit: $\min_\alpha \ \|\alpha\|_1 \ \ \text{s.t.} \ \ \|\mathbf{D}\alpha - y\|_2 \le \varepsilon$

leads to a stable result: $\|\hat{\alpha} - \alpha\|_2^2 \le \dfrac{4\varepsilon^2}{1 - \mu(4\|\alpha\|_0 - 1)}$

Donoho, Elad & Temlyakov ('06)



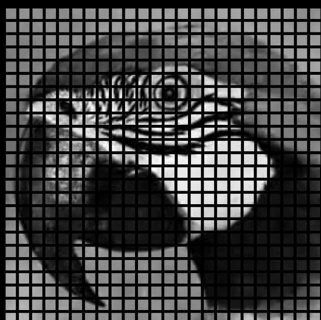
Comments:

- If ε=0 → $\hat{\alpha} = \alpha$
- This is a worst-case analysis – better bounds exist
- Similar theorems exist for many other pursuit algorithms

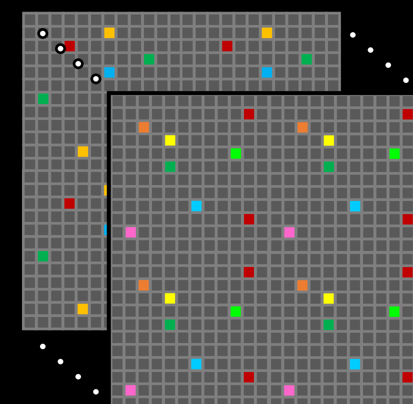
Convolutional Sparse Coding (CSC)

m filters convolved with their sparse representations

i-th feature-map: An image of the same size as $\mathbf{X}$ holding the sparse representation related to the i-filter



$$[\mathbf{X}] = \sum_{\mathrm{i}=1}^{m} \mathrm{d_i} * [\Gamma_\mathrm{i}]$$



An image with N pixels

The i-th filter of small size n

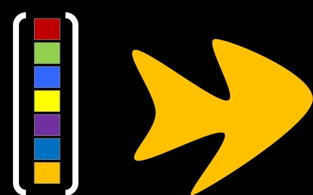
This model emerged in 2005-2010, developed and advocated by Yan LeCun and others. It serves as the foundation of Convolutional Neural Networks

CSC in Matrix Form

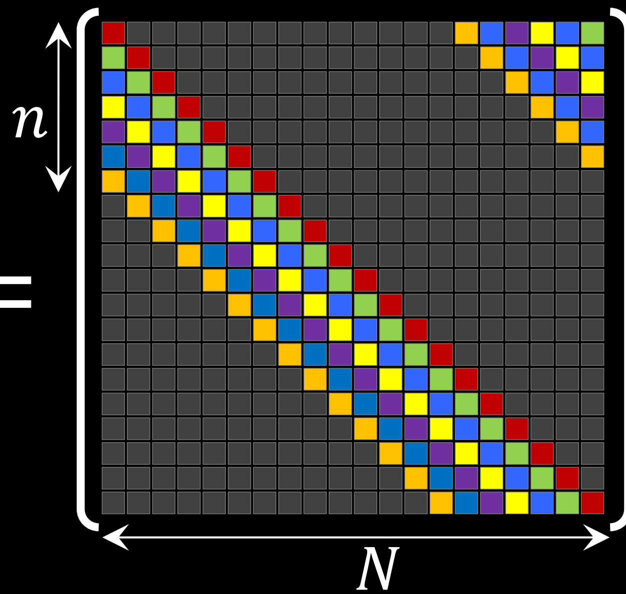
- Here is an alternative global sparsity-based model formulation

$$\mathbf{X} = \sum_{\mathrm{i}=1}^{m} \mathbf{C}^{\mathrm{i}}\mathbf{\Gamma}^{\mathrm{i}} = \begin{bmatrix}\mathbf{C}^1 & \cdots & \mathbf{C}^m\end{bmatrix}\begin{bmatrix}\mathbf{\Gamma}^1 \\ \vdots \\ \mathbf{\Gamma}^m\end{bmatrix} = \mathbf{D\Gamma}$$

- $\mathbf{C}^{\mathrm{i}} \in \mathbb{R}^{N \times N}$ is a banded and Circulant matrix containing a single atom with all of its shifts



$$\mathbf{C}^{\mathrm{i}} =$$

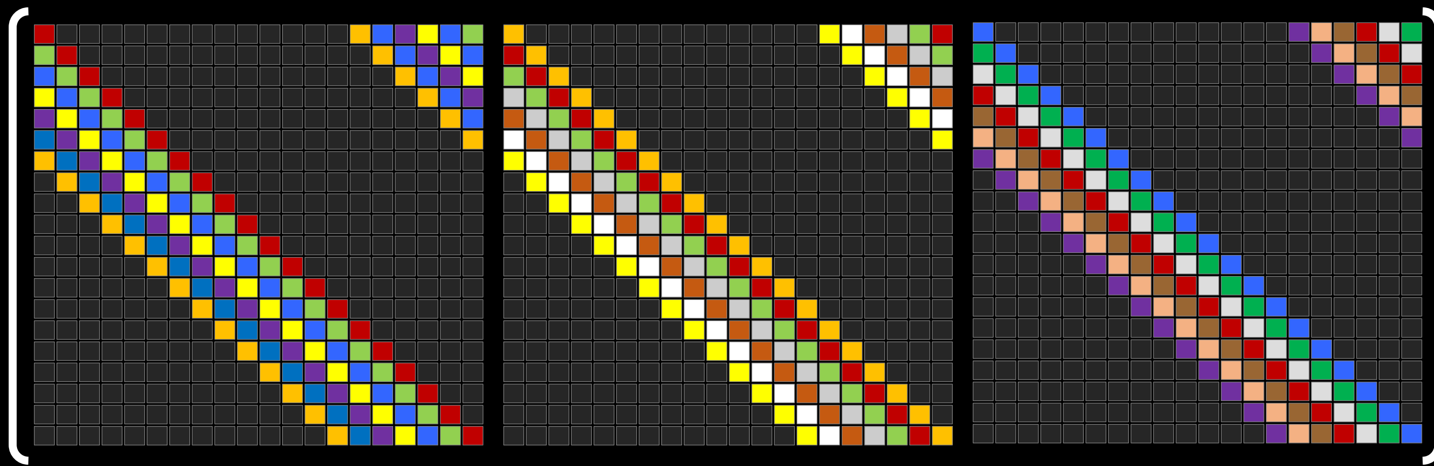


- $\mathbf{\Gamma}^{\mathrm{i}} \in \mathbb{R}^{N}$ are the corresponding coefficients ordered as column vectors

Michael Elad
The Computer-Science Department
The Technion

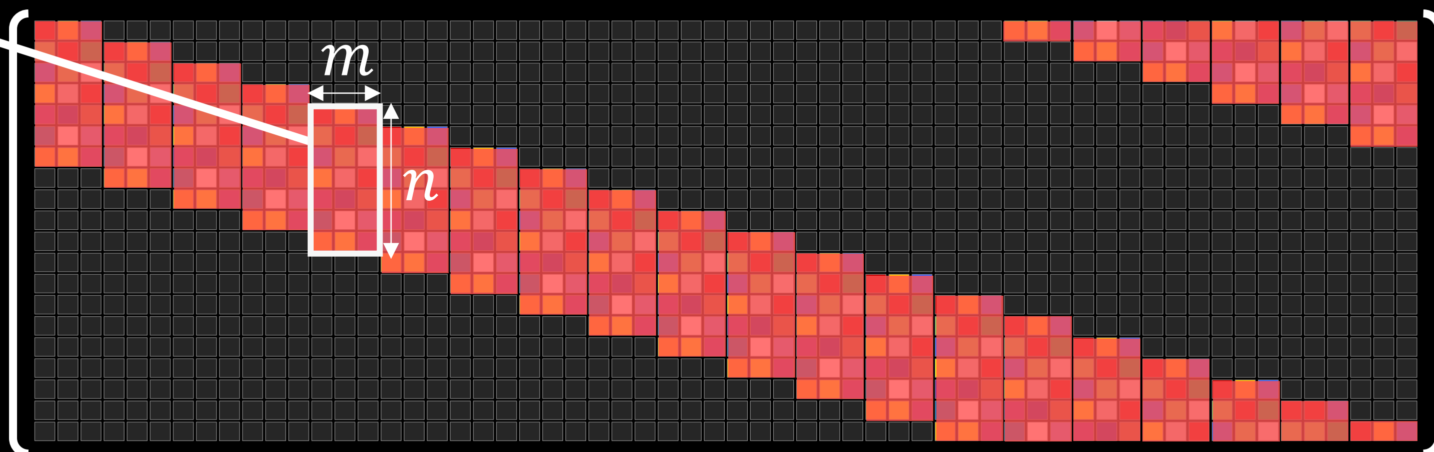
The CSC Dictionary

$[\mathbf{C}^1\ \mathbf{C}^2\ \mathbf{C}^3] =$



$\mathbf{D}_{\mathrm{L}}$

$\mathbf{D} =$

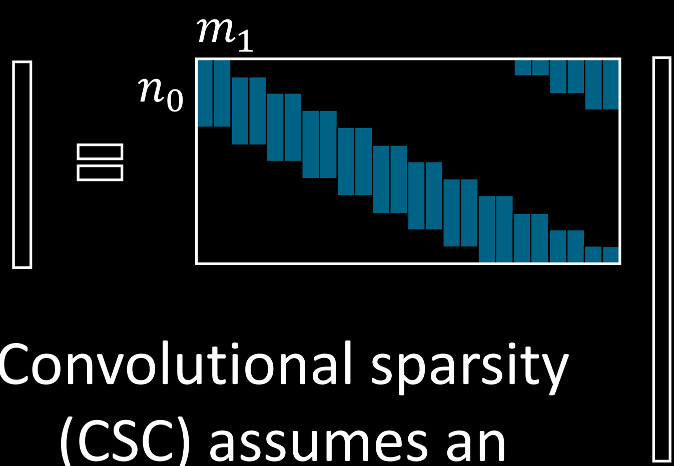


Michael Elad
The Computer-Science Department
The Technion

Multi-Layered Convolutional Sparse Modeling

From CSC to Multi-Layered CSC

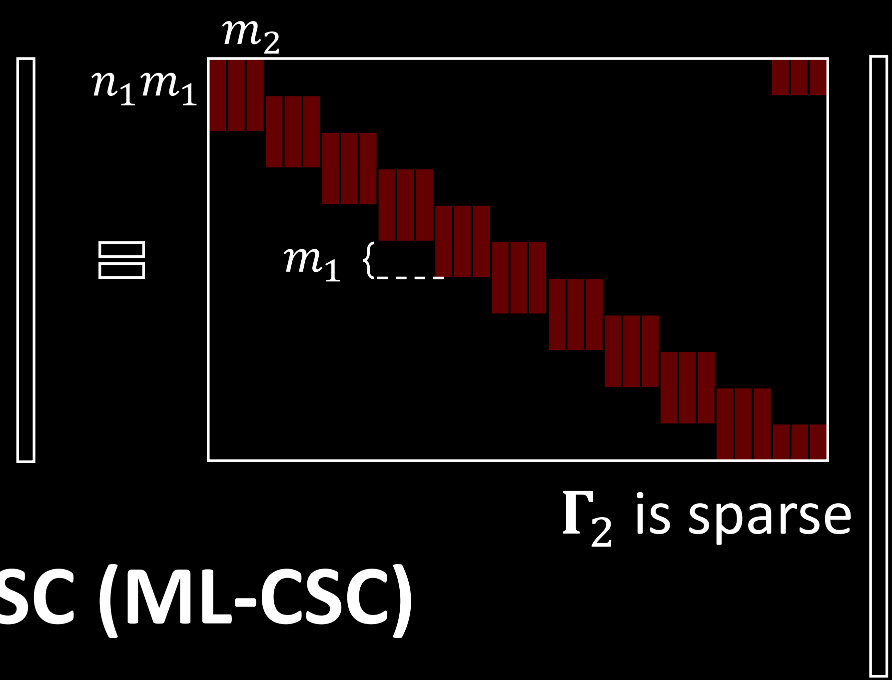
$\mathbf{X} \in \mathbb{R}^{N}$ $\quad \mathbf{D}_1 \in \mathbb{R}^{N \times N m_1}$ $\quad \mathbf{\Gamma}_1 \in \mathbb{R}^{N m_1}$



Convolutional sparsity (CSC) assumes an inherent structure is present in natural signals - $\mathbf{\Gamma}_1$ is sparse

We propose to impose the same structure on the representations themselves

$\mathbf{\Gamma}_1 \in \mathbb{R}^{N m_1}$ $\quad \mathbf{D}_2 \in \mathbb{R}^{N m_1 \times N m_2}$ $\quad \mathbf{\Gamma}_2 \in \mathbb{R}^{N m_2}$



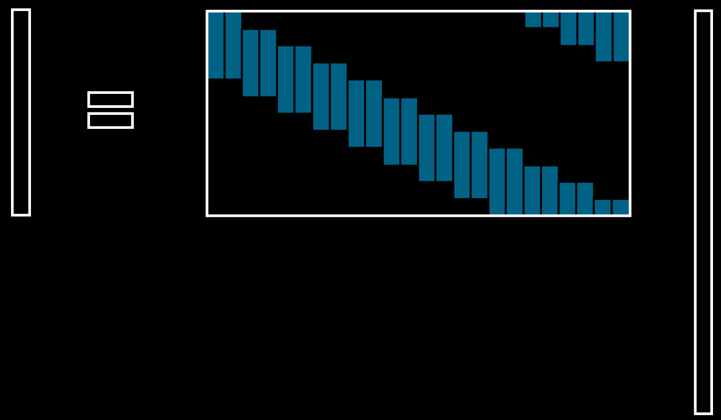
$\mathbf{\Gamma}_2$ is sparse

Multi-Layer CSC (ML-CSC)

Michael Elad
The Computer-Science Department
The Technion

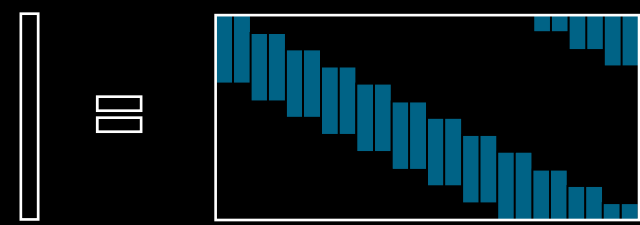
Intuition: From Atoms to Molecules

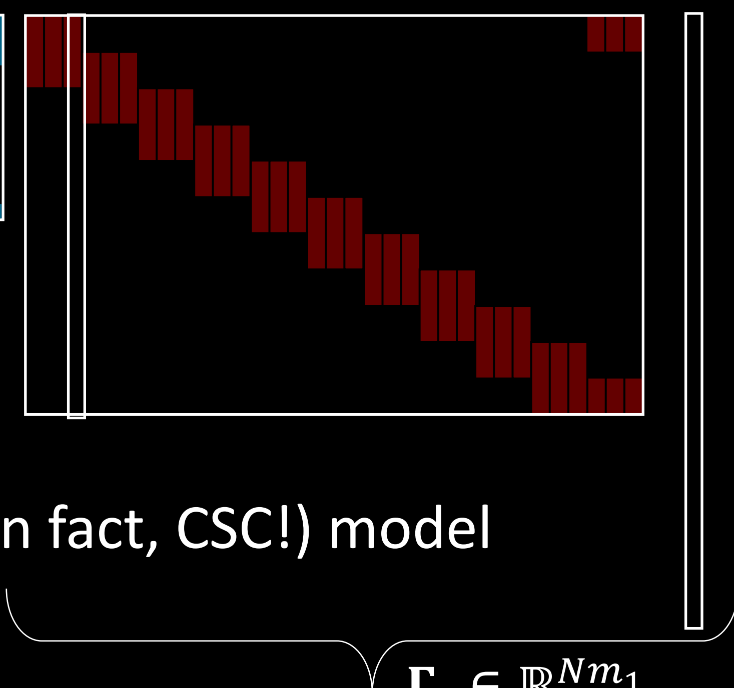
$\mathbf{X} \in \mathbb{R}^{N}$ $\quad$ $\mathbf{D}_1 \in \mathbb{R}^{N \times N m_1}$ $\quad$ $\mathbf{\Gamma}_1 \in \mathbb{R}^{N m_1}$



Michael Elad
The Computer-Science Department
The Technion

Intuition: From Atoms to Molecules





- We can chain the all the dictionaries into one effective dictionary $\mathbf{D}_{\text{eff}} = \mathbf{D}_1\mathbf{D}_2\mathbf{D}_3 \cdots \mathbf{D}_{\text{K}} \rightarrow \mathbf{x} = \mathbf{D}_{\text{eff}}\,\boldsymbol{\Gamma}_{\text{K}}$
- This is regular Sparse Representation (in fact, CSC!) model
- However:
  - A key property: $\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2, \boldsymbol{\Gamma}_3, \ldots, \boldsymbol{\Gamma}_{\text{K}}$ are all sparse
  - We get a series of descriptions of x with different abstraction levels:

$$\mathbf{x} = \mathbf{D}_1\boldsymbol{\Gamma}_1 = \mathbf{D}_1\mathbf{D}_2\boldsymbol{\Gamma}_2 = \mathbf{D}_1\mathbf{D}_2\mathbf{D}_3\boldsymbol{\Gamma}_3 = \cdots = \mathbf{D}_{\text{eff}}\,\boldsymbol{\Gamma}_{\text{K}}$$

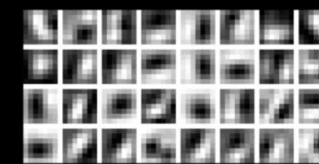
Michael Elad
The Computer-Science Department
The Technion

A Small Taste: Model Training (MNIST)

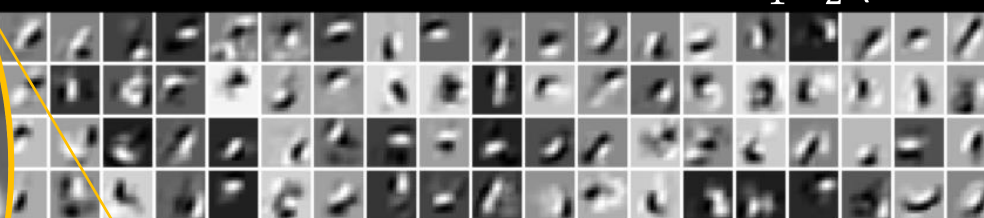
MNIST Dictionary:

- D_1: 32 filters o … f 2 (dense)
- D_2: 128 fil … f 1 - 99.09 % sparse
- D3: 102 … parse

$\mathbf{D}_1$ (7×7)



$\mathbf{D}_1\mathbf{D}_2$ (15×15)



$\mathbf{D}_1\mathbf{D}_2\mathbf{D}_3$ (28×28)

ML-CSC: Pursuit

- Deep–Coding Problem $(\mathbf{DCP_\lambda})$ (dictionaries are known):

$$\left\{\begin{array}{ll} \mathbf{X} = \mathbf{D}_1\mathbf{\Gamma}_1 & \|\mathbf{\Gamma}_1\|_0 \leq \lambda_1 \\ \mathbf{\Gamma}_1 = \mathbf{D}_2\mathbf{\Gamma}_2 & \|\mathbf{\Gamma}_2\|_0 \leq \lambda_2 \\ \vdots & \vdots \\ \mathbf{\Gamma}_{K-1} = \mathbf{D}_K\mathbf{\Gamma}_K & \|\mathbf{\Gamma}_K\|_0 \leq \lambda_K \end{array}\right\}$$

- Or, more realistically for noisy signals,

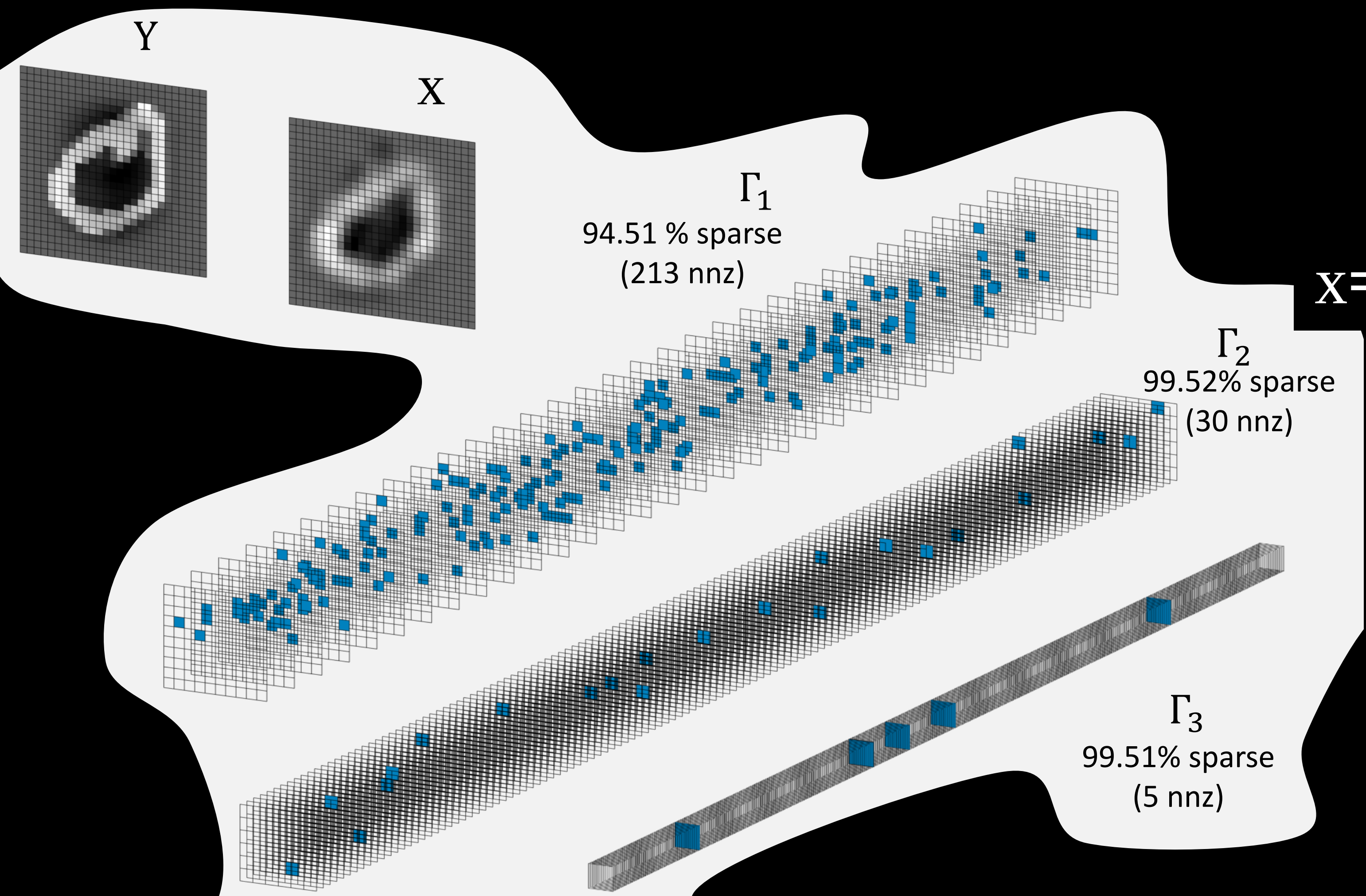
$$\text{Find } \left\{\mathbf{\Gamma}_j\right\}_{j=1}^{K} \quad s.t. \quad \left\{\begin{array}{ll} \|\mathbf{Y} - \mathbf{D}_1\mathbf{\Gamma}_1\|_2 \leq \mathcal{E} & \|\mathbf{\Gamma}_1\|_0 \leq \lambda_1 \\ \mathbf{\Gamma}_1 = \mathbf{D}_2\mathbf{\Gamma}_2 & \|\mathbf{\Gamma}_2\|_0 \leq \lambda_2 \\ \vdots & \vdots \\ \mathbf{\Gamma}_{K-1} = \mathbf{D}_K\mathbf{\Gamma}_K & \|\mathbf{\Gamma}_K\|_0 \leq \lambda_K \end{array}\right\}$$

A Small Taste: Pursuit

$$x = \mathbf{D}_1 \Gamma_1$$

$$x = \mathbf{D}_1 \mathbf{D}_2 \Gamma_2$$

$$x = \mathbf{D}_1 \mathbf{D}_2 \mathbf{D}_3 \Gamma_3$$



Michael Elad
The Computer-Science Department
The Technion

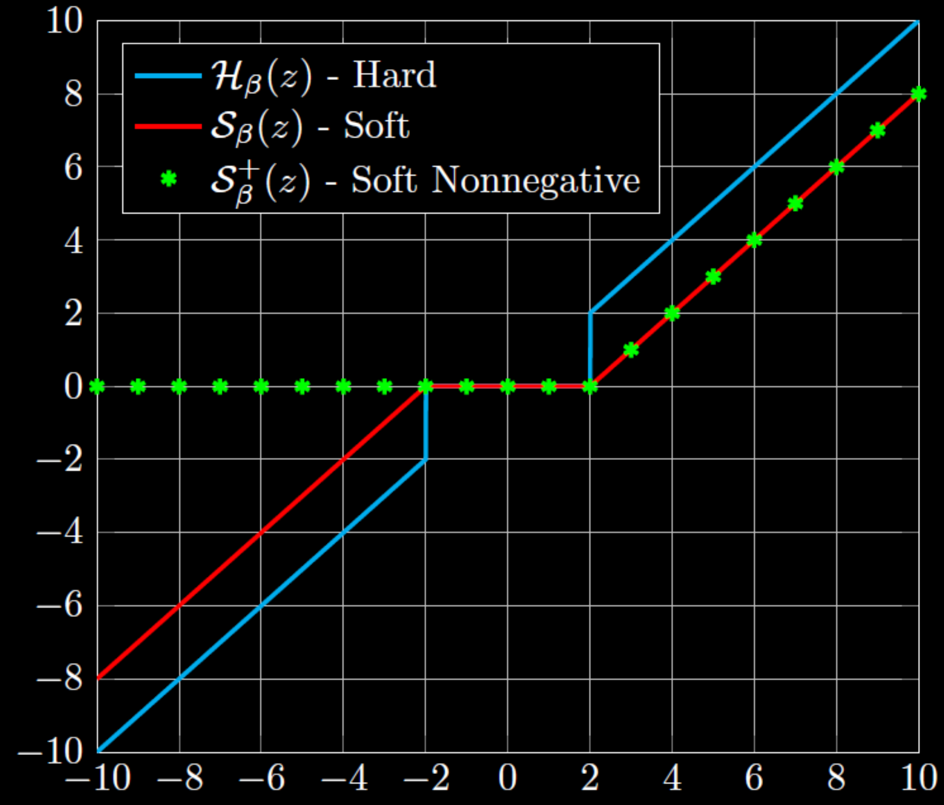
ML-CSC: The Simplest Pursuit

The simplest pursuit algorithm (single-layer case) is the THR algorithm, which operates on a given input signal $\mathbf{Y}$ by:

$$\mathbf{Y} = \mathbf{D}\mathbf{\Gamma} + \mathbf{E}$$

and $\mathbf{\Gamma}$ is sparse

$$\hat{\mathbf{\Gamma}} = \mathcal{P}_\beta\left(\mathbf{D}^\mathrm{T}\mathbf{Y}\right)$$



Consider this for Solving the DCP

- Layered Thresholding (LT):

$$\hat{\boldsymbol{\Gamma}}_2 = \underbrace{\mathcal{P}_{\beta_2}\left(\mathbf{D}_2^T \overbrace{\mathcal{P}_{\beta_1}\left(\mathbf{D}_1^T\mathbf{Y}\right)}^{\text{Estimate } \boldsymbol{\Gamma}_1 \text{ via the THR algorithm}}\right)}_{\text{Estimate } \boldsymbol{\Gamma}_2 \text{ via the THR algorithm}}$$

$$\left(\mathbf{DCP}_\lambda^\mathcal{E}\right): \text{ Find } \left\{\boldsymbol{\Gamma}_j\right\}_{j=1}^K \quad s.t.$$

$$\begin{cases} \|\mathbf{Y}-\mathbf{D}_1\boldsymbol{\Gamma}_1\|_2 \le \mathcal{E} & \|\boldsymbol{\Gamma}_1\|_0 \le \lambda_1 \\ \boldsymbol{\Gamma}_1 = \mathbf{D}_2\boldsymbol{\Gamma}_2 & \|\boldsymbol{\Gamma}_2\|_0 \le \lambda_2 \\ \vdots & \vdots \\ \boldsymbol{\Gamma}_{K-1} = \mathbf{D}_K\boldsymbol{\Gamma}_K & \|\boldsymbol{\Gamma}_K\|_0 \le \lambda_K \end{cases}$$

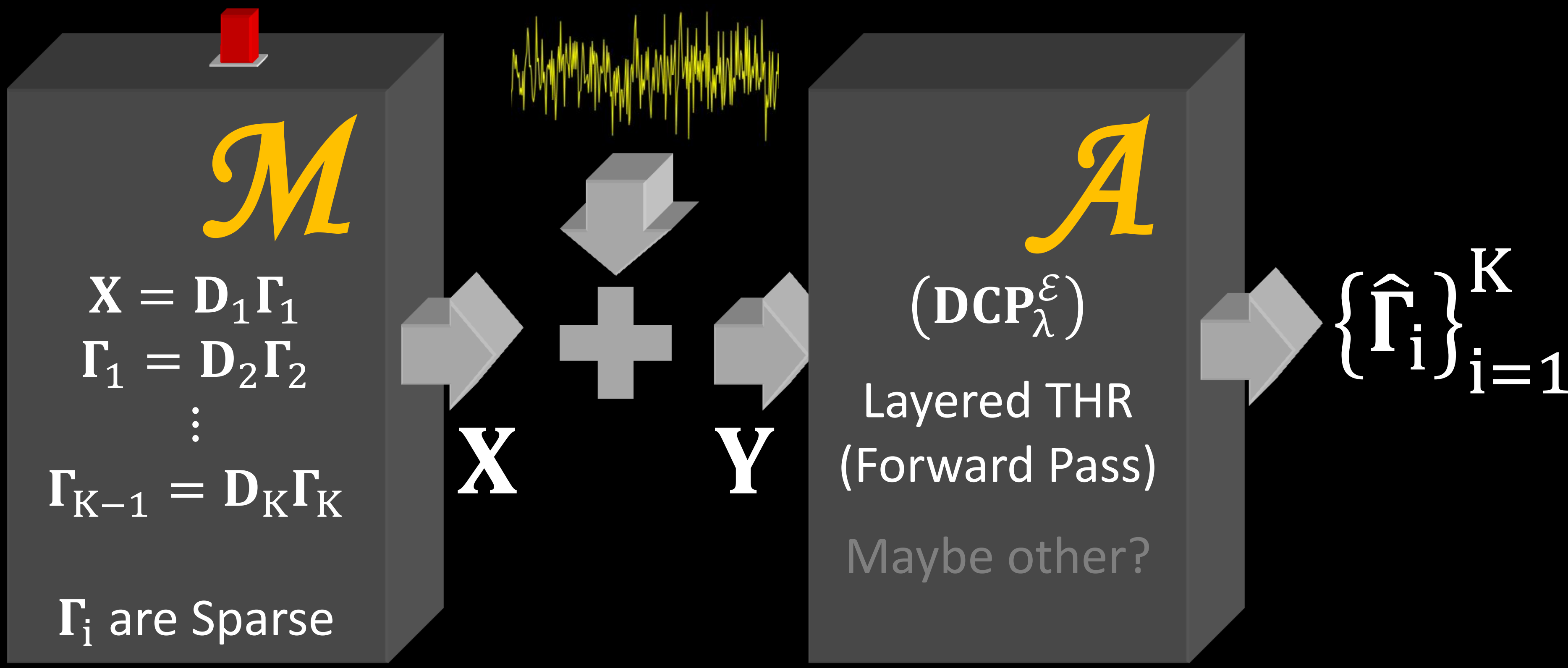
- Now let's take a look at how Conv. Neural Network operates:

$$f(\mathbf{Y}) = \text{ReLU}\left(\mathbf{b}_2 + \mathbf{W}_2^T \, \text{ReLU}\left(\mathbf{b}_1 + \mathbf{W}_1^T\mathbf{Y}\right)\right)$$

The layered (soft nonnegative) thresholding and the CNN forward pass algorithm are the very same thing !!!

Michael Elad
The Computer-Science Department
The Technion

Theoretical Path



Armed with this view of a generative source model, we may ask new and daring theoretical questions

Michael Elad
The Computer-Science Department
The Technion

Success of the Layered-THR

Theorem: If $\|\boldsymbol{\Gamma}_i\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D}_i)} \cdot \frac{|\boldsymbol{\Gamma}_i^{\min}|}{|\boldsymbol{\Gamma}_i^{\max}|}\right) - \frac{1}{\mu(\mathbf{D}_i)} \cdot \frac{\varepsilon_L^{i-1}}{|\boldsymbol{\Gamma}_i^{\max}|}$

then the **Layered Hard THR** (with the proper thresholds) **finds the correct supports** and $\left\|\boldsymbol{\Gamma}_i^{LT} - \boldsymbol{\Gamma}_i\right\|_{2,\infty}^{p} \le \varepsilon_L^i$, where we have defined $\varepsilon_L^0 = \|\mathbf{E}\|_2$ and

$$\varepsilon_L^i = \sqrt{\|\boldsymbol{\Gamma}_i\|_0} \cdot \left(\varepsilon_L^{i-1} + \mu(\mathbf{D}_i)(\|\boldsymbol{\Gamma}_i\|_0 - 1)|\boldsymbol{\Gamma}_i^{\max}|\right)$$

Papyan, Romano & Elad ('17)

The stability of the forward pass is guaranteed if the underlying representations are sparse and the noise is bounded

Problems:

1. Contrast
2. Error growth
3. Error even if no noise

Layered Basis Pursuit (BP)

- We chose the Thresholding algorithm due to its simplicity, but we do know that there are better pursuit methods – how about using them?
- Lets use the Basis Pursuit instead …

$$\left(\mathbf{DCP}_\lambda^\mathcal{E}\right):\ \text{Find}\ \ \{\mathbf{\Gamma}_\mathrm{j}\}_{\mathrm{j}=1}^{\mathrm{K}}\quad s.t.$$

$$\begin{Bmatrix} \|\mathbf{Y}-\mathbf{D}_1\mathbf{\Gamma}_1\|_2 \le \mathcal{E} & \|\mathbf{\Gamma}_1\|_0 \le \lambda_1 \\ \mathbf{\Gamma}_1 = \mathbf{D}_2\mathbf{\Gamma}_2 & \|\mathbf{\Gamma}_2\|_0 \le \lambda_2 \\ \vdots & \vdots \\ \mathbf{\Gamma}_{\mathrm{K}-1} = \mathbf{D}_\mathrm{K}\mathbf{\Gamma}_\mathrm{K} & \|\mathbf{\Gamma}_\mathrm{K}\|_0 \le \lambda_\mathrm{K} \end{Bmatrix}$$

$$\mathbf{\Gamma}_1^{\mathrm{LBP}} = \min_{\mathbf{\Gamma}_1} \frac{1}{2}\|\mathbf{Y}-\mathbf{D}_1\mathbf{\Gamma}_1\|_2^2 + \lambda_1\|\mathbf{\Gamma}_1\|_1$$

$$\downarrow$$

$$\mathbf{\Gamma}_2^{\mathrm{LBP}} = \min_{\mathbf{\Gamma}_2} \frac{1}{2}\left\|\mathbf{\Gamma}_1^{\mathrm{LBP}}-\mathbf{D}_2\mathbf{\Gamma}_2\right\|_2^2 + \lambda_2\|\mathbf{\Gamma}_2\|_1$$

$$\downarrow$$

$$\vdots$$

Deconvolutional networks
[Zeiler, Krishnan, Taylor & Fergus ‘10]

Success of the Layered BP

Theorem: **Assuming that** $\|\boldsymbol{\Gamma}_i\|_0 < \frac{1}{3}\left(1 + \frac{1}{\mu(\mathbf{D}_i)}\right)$ then the Layered Basis Pursuit performs very well:

1. The support of $\boldsymbol{\Gamma}_i^{\mathrm{LBP}}$ is contained in that of $\boldsymbol{\Gamma}_i$
2. The error is bounded: $\left\|\boldsymbol{\Gamma}_i^{\mathrm{LBP}} - \boldsymbol{\Gamma}_i\right\|_2 \le \varepsilon_L^i$, where
$$\varepsilon_L^i = 7.5^i \|\mathbf{E}\|_2 \prod_{j=1}^{i} \sqrt{\left\|\boldsymbol{\Gamma}_j\right\|_0}$$
3. Every entry in $\boldsymbol{\Gamma}_i$ greater than $\varepsilon_L^i / \sqrt{\|\boldsymbol{\Gamma}_i\|_0}$ will be found

Papyan, Romano & Elad ('17)

Problems:
1. ~~Contrast~~
2. Error growth
3. ~~Error even if no noise~~

Layered Iterative Thresholding

Layered BP: $\Gamma_j^{LBP} = \min_{\Gamma_j} \frac{1}{2} \|\Gamma_{j-1}^{LBP} - \mathbf{D}_j \Gamma_j\|_2^2 + \xi_j \|\Gamma_j\|_1$ j

Layered Iterative Soft-Thresholding:

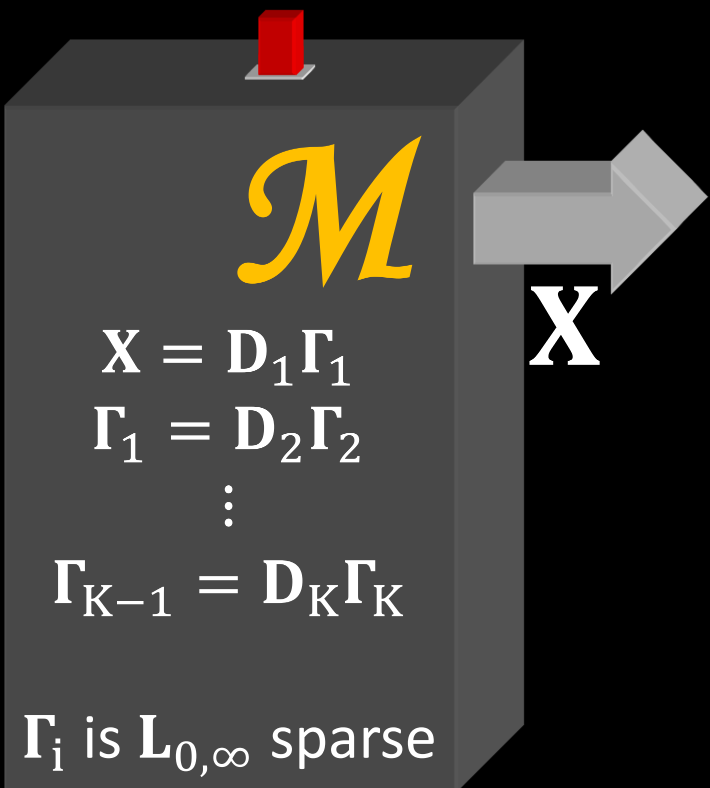
t $\Gamma_j^t = \mathcal{S}_{\xi_j / c_j}\left(\Gamma_j^{t-1} + \mathbf{D}_j^T\left(\hat{\Gamma}_{j-1} - \mathbf{D}_j \Gamma_j^{t-1}\right)\right)$ j

Note that our suggestion implies that groups of layers share the same dictionaries

Can be seen as a very deep recurrent neural network [Gregor & LeCun '10]

Reflections and Recent Results

Where are the Labels?

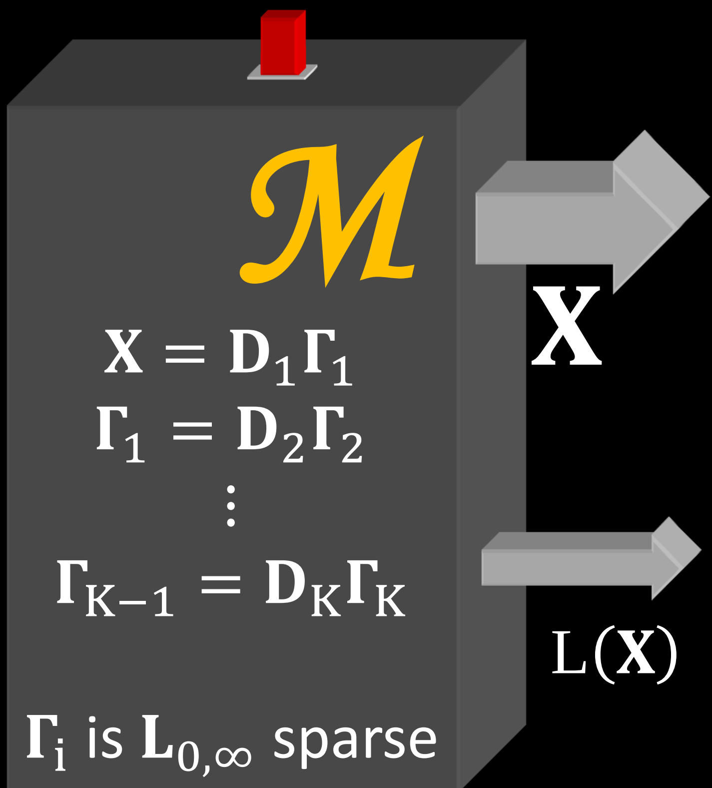


We presented the ML-CSC as a machine that produces signals $\mathbf{X}$

Answer 1:

- We do not need labels because everything we show refer to the **unsupervised** case, in which we operate on signals, not necessarily in the context of recognition

Where are the Labels?



We presented the ML-CSC as a machine that produces signals $\mathbf{X}$

Answer 2:

- In fact, this model could be augmented by a synthesis of the corresponding label by:

$$\mathrm{L}(\mathbf{X}) = sign\left\{c + \sum_{j=1}^{K} w_j^T \Gamma_j\right\}$$

- This assumes that knowing the representations suffices for classification → **supervised** mode
- Thus, a successful pursuit algorithm can lead to an accurate recognition if the network is augmented by a FC classification layer
- In fact, we can analyze theoretically the classification accuracy and the sensitivity to adversarial noise – see later

What About Learning?

Sparseland
Sparse Representation Theory

→

CSC
Convolutional Sparse Coding

→

ML-CSC
Multi-Layered Convolutional Sparse Coding

All these models rely on proper **Dictionary Learning Algorithms** to fulfil their mission:

- Sparseland: We have unsupervised and supervised such algorithms, and a beginning of theory to explain how these work
- CSC: We have few and only unsupervised methods, and even these are not fully stable/clear
- ML-CSC: Two algorithms were proposed – unsupervised (to appear in IEEE-TSP) and supervised (submitted to NIPS 2018) – both available on ArxiV

Michael Elad
The Computer-Science Department
The Technion

Fresh from the Oven (1)

Main Focus:

- Better pursuit &
- Dictionary learning

Contributions:

- Proposed a projection based pursuit (i.e. Verifying that the obtained signal obeys the synthesis equations), accompanied by better theoretical guarantees
- Proposes the first dictionary learning algorithm for the ML-CSC model for an unsupervised mode of work (as an auto-encoder, and trading representations' sparsities by dictionary sparsity)

Multilayer Convolutional Sparse Modeling: Pursuit and Dictionary Learning

Jeremias Sulam, *Member, IEEE*, Vardan Papyan, Yaniv Romano, and Michael Elad, *Fellow, IEEE*

***Abstract*—The recently proposed multilayer convolutional sparse coding (ML-CSC) model, consisting of a cascade of convolutional sparse layers, provides a new interpretation of convolutional neural networks (CNNs). Under this framework, the forward pass in a CNN is equivalent to a pursuit algorithm aiming to estimate the nested sparse representation vectors from a given input signal. Despite having served as a pivotal connection between CNNs and sparse modeling, a deeper understanding of the ML-CSC is still lacking. In this paper, we propose a sound pursuit algorithm for the ML-CSC model by adopting a projection approach. We provide new and improved bounds on the stability of the solution of such pursuit and we analyze different practical alternatives to implement this in practice. We show that the training of the filters is essential to allow for nontrivial signals in the model, and we derive an online algorithm to learn the dictionaries from real**

as atoms [1]. Backed by elegant theoretical results, this model led to a series of works dealing either with the problem of the pursuit of such decompositions, or with the design and learning of better atoms from real data [2]. The latter problem, termed dictionary learning, empowered sparse enforcing methods to achieve remarkable results in many different fields from signal and image processing [3]–[5] to machine learning [6]–[8].

Neural networks, on the other hand, were introduced around forty years ago and were shown to provide powerful classification algorithms through a series of function compositions [9], [10]. It was not until the last half-decade, however, that through a series of incremental modifications these methods

To appear in IEEE-TSP

Fresh from the Oven (2)

Main Focus:

- Holistic pursuit &
- Relation to the Co-Sparse analysis model

Contributions:

- Proposed a systematic way to synthesize signals from the ML-CSC model
- Develop performance bounds for the oracle in various pursuit strategies
- Constructs the first provable holistic pursuit that mixes greedy-analysis and relaxation-synthesis pursuit algorithms

MULTI LAYER SPARSE CODING: THE HOLISTIC WAY

AVIAD ABERDAM*, JEREMIAS SULAM†, AND MICHAEL ELAD‡

Abstract. The recently proposed multi-layer sparse model has raised insightful connections between sparse representations and convolutional neural networks (CNN). In its original conception, this model was restricted to a cascade of *convolutional synthesis* representations. In this paper, we start by addressing a more general model, revealing interesting ties to fully connected networks. We then show that this multi-layer construction admits a brand new interpretation in a unique symbiosis between synthesis and analysis models: while the deepest layer indeed provides a synthesis representation, the mid-layers decompositions provide an analysis counterpart. This new perspective exposes the suboptimality of previously proposed pursuit approaches, as they do not fully leverage all the information comprised in the model constraints. Armed with this understanding, we address fundamental theoretical issues, revisiting previous analysis and expanding it. Motivated by the limitations of previous algorithms, we then propose an integrated – *holistic* – alternative that estimates all representations in the model simultaneously, and analyze all these different schemes under stochastic noise assumptions. Inspired by the synthesis-analysis duality, we further present a Holistic Pursuit algorithm, which alternates between synthesis and analysis sparse coding steps, eventually solving for the entire model as a whole, with provable improved performance. Finally, we present numerical results that demonstrate the practical advantages of our approach.

Submitted to SIMODS

Fresh from the Oven (3)

Main Focus:

- Take the labels into account
- Analyze classification performance and sensitivity to adversarial noise

Contributions:

- Develop bounds on the maximal adversarial noise that guarantees a proper classification
- Expose the higher sensitivity of poor pursuit methods (Layered-THR) over better ones (Layered-BP)

Classification Stability for Sparse-Modeled Signals

Yaniv Romano
Department of Statistics
Stanford University
yromano@stanford.edu

Michael Elad
Department of Computer Science
Technion – Israel Institute of Technology
elad@cs.technion.ac.il

Abstract

Despite their impressive performance, deep convolutional neural networks (CNNs) have been shown to be sensitive to small adversarial perturbations. These nuisances, which one can barely notice, are powerful enough to fool sophisticated and well

Submitted to NIPS 2018

Fresh from the Oven (4)

Main Focus:

- Better and provable ISTA-like pursuit algorithm
- Examine the effect of the number of iterations in the unfolded architecture

Contributions:

- Develop a novel ISTA-like algorithms for the ML-CSC model, with proper mathematical justifications
- Demonstrate the architecture obtained when unfolding this algorithm
- Show that for the same number of parameters, more iterations lead to better classification

On Multi-Layer Basis Pursuit, Efficient Algorithms and Convolutional Neural Networks

Jeremias Sulam*
Computer Science Department
Technion – Israel Institute of Technology
jsulam@cs.technion.ac.il

Aviad Aberdam*
Electrical Engineering Department
Technion – Israel Institute of Technology
aaberdam@campus.technion.ac.il

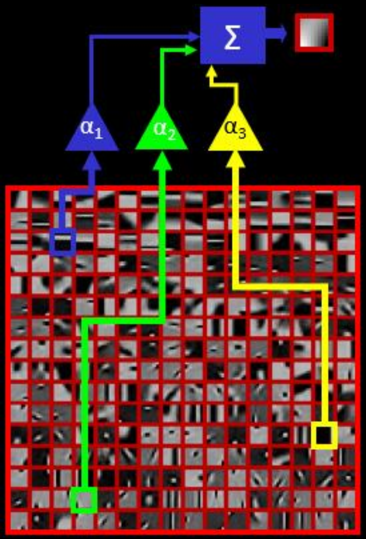
Michael Elad
Computer Science Department
Technion – Israel Institute of Technology
elad@cs.technion.ac.il

Submitted to NIPS 2018

Time to Conclude

This Talk

Sparseland ← The desire to model data



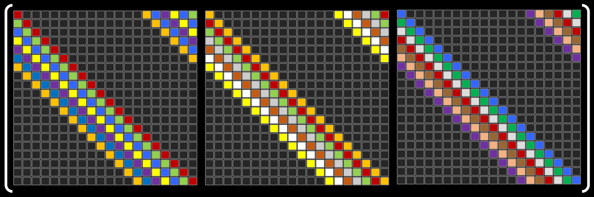
This entire talk is based on the *Sparseland* model, constructing variations of it

This Talk

Sparseland

The desire to model data

Novel View of Convolutional Sparse Coding



We rely on our in-depth theoretical study of the CSC model (not presented!)

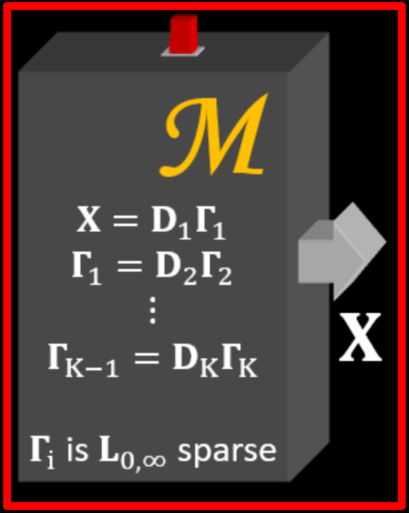
This Talk

Sparseland

The desire to model data

Novel View of Convolutional Sparse Coding

Multi-Layer Convolutional Sparse Coding



We propose a multi-layer extension of CSC, shown to be tightly connected to CNN

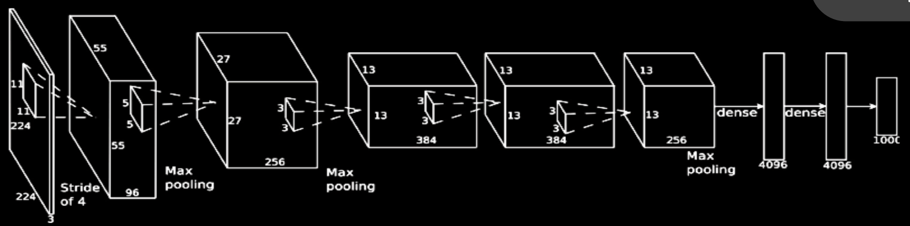
Michael Elad
The Computer-Science Department
The Technion

This Talk

Sparseland

The desire to model data

Novel View of Convolutional Sparse Coding



Multi-Layer Convolutional Sparse Coding

A novel interpretation and theoretical understanding of CNN

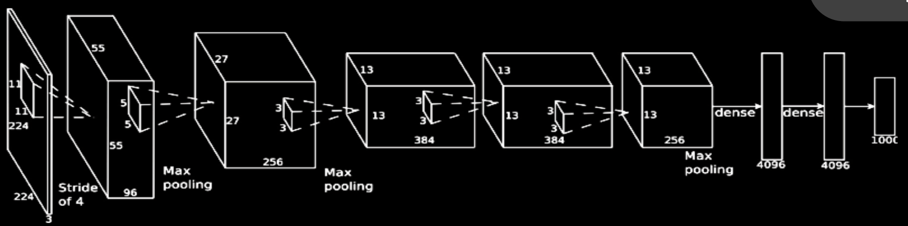
The ML-CSC was shown to enable a theoretical study of CNN, along with new insights

This Talk

Sparseland

The desire to model data

Novel View of Convolutional Sparse Coding



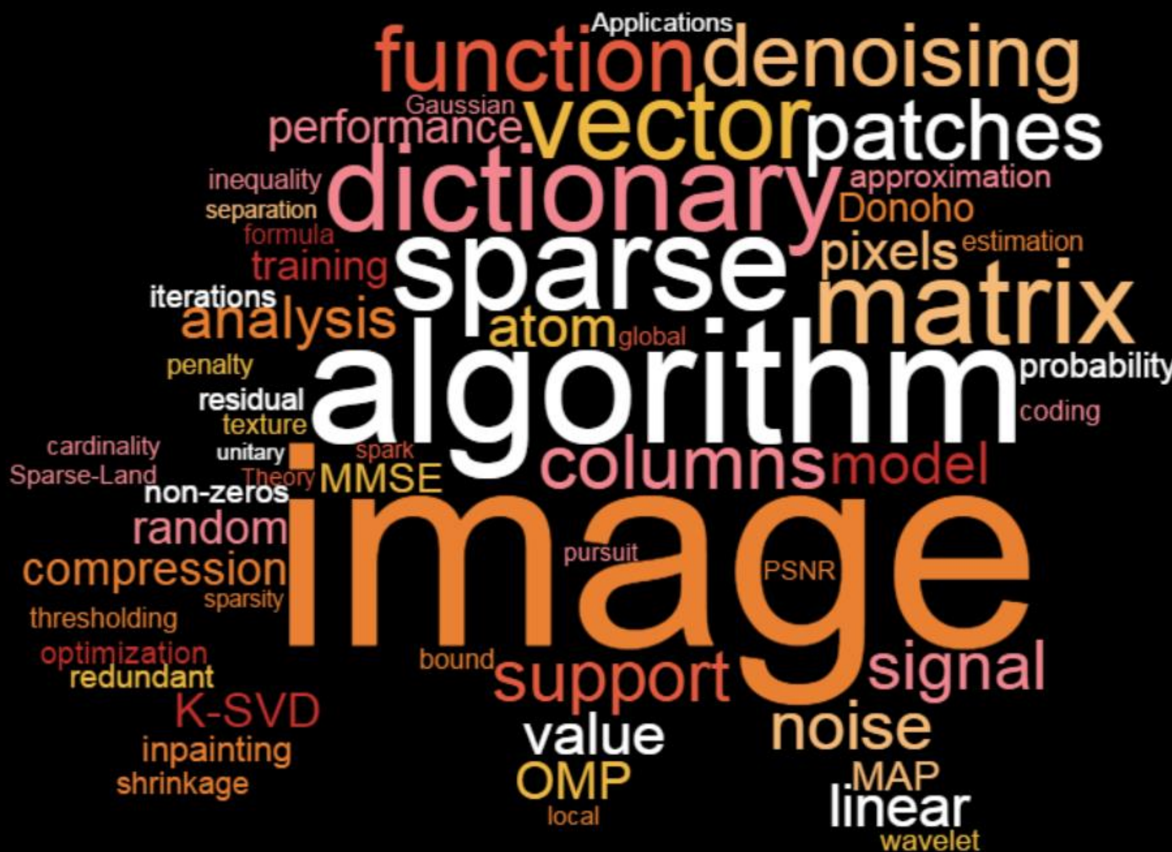
Multi-Layer Convolutional Sparse Coding

A novel interpretation and theoretical understanding of CNN

Take Home Message

The Multi-Layer Convolutional Sparse Coding model could be a new platform for theoretically understanding deep learning, and developing it further

Michael Elad
The Computer-Science Department
The Technion



More on these (including these slides and the relevant papers) can be found in http://www.cs.technion.ac.il/~elad

Michael Elad
The Computer-Science Department
The Technion