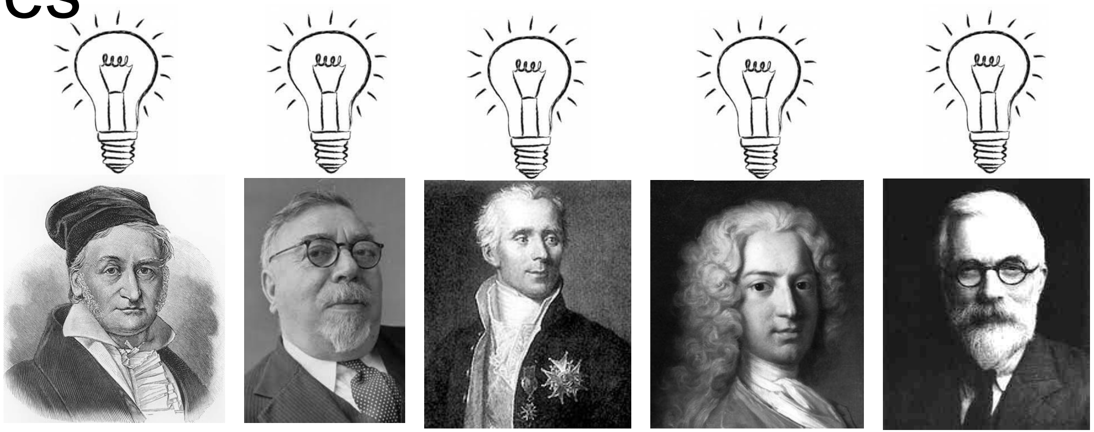
Convolutional Neural Networks in View of Sparse Coding

Joint work with:
Jeremias Sulam, Yaniv Romano, Michael Elad



Breiman’s “Two Cultures”

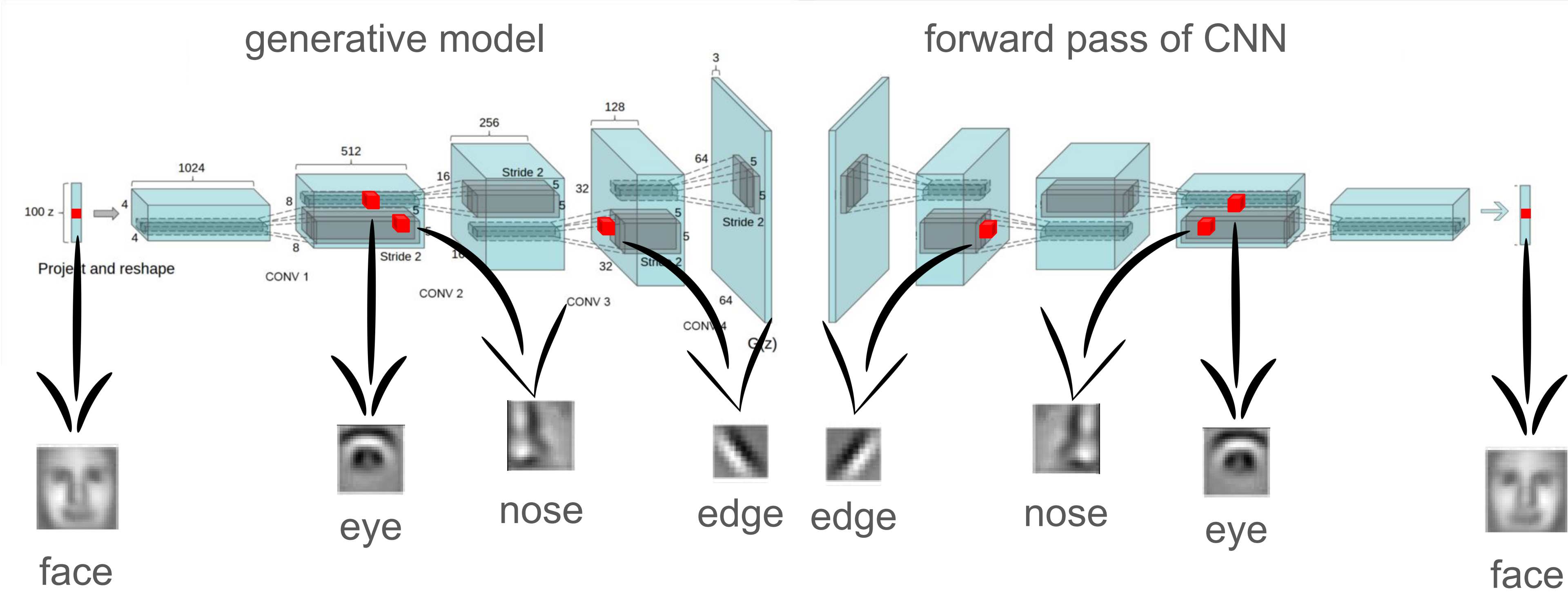
Generative modeling



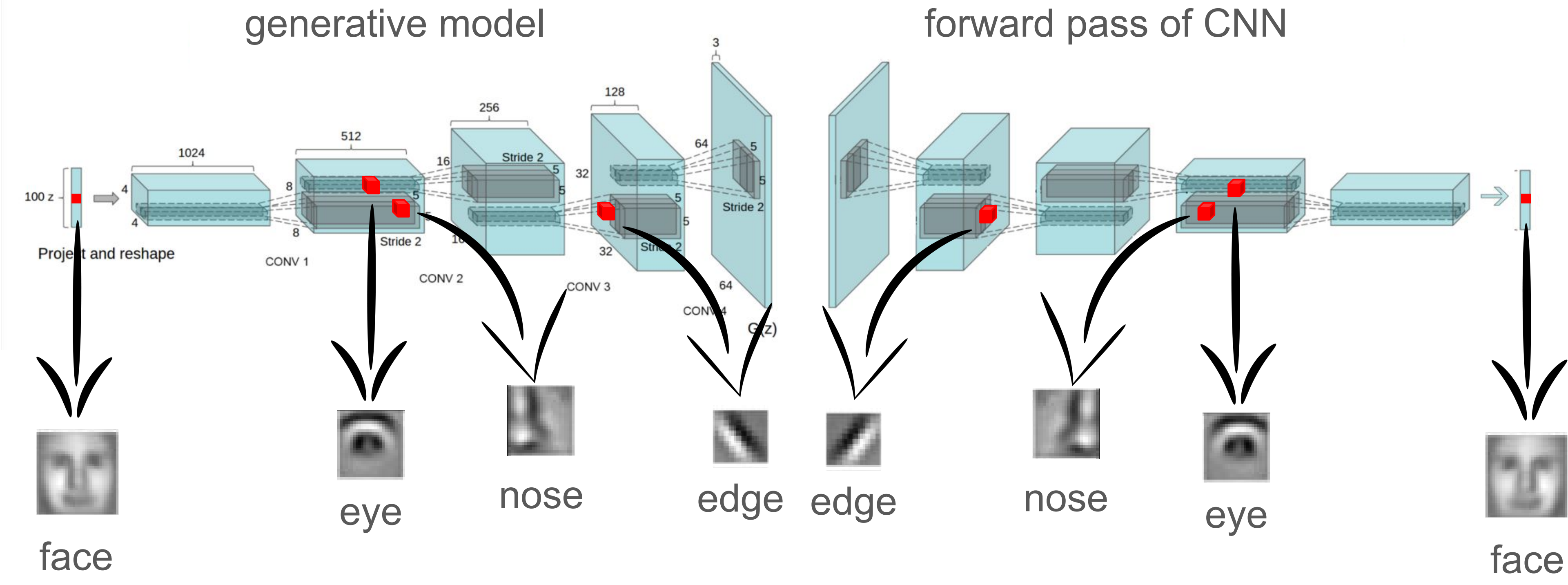
Gauss Wiener Laplace Bernoulli Fisher

Predictive modeling

Generative Modeling

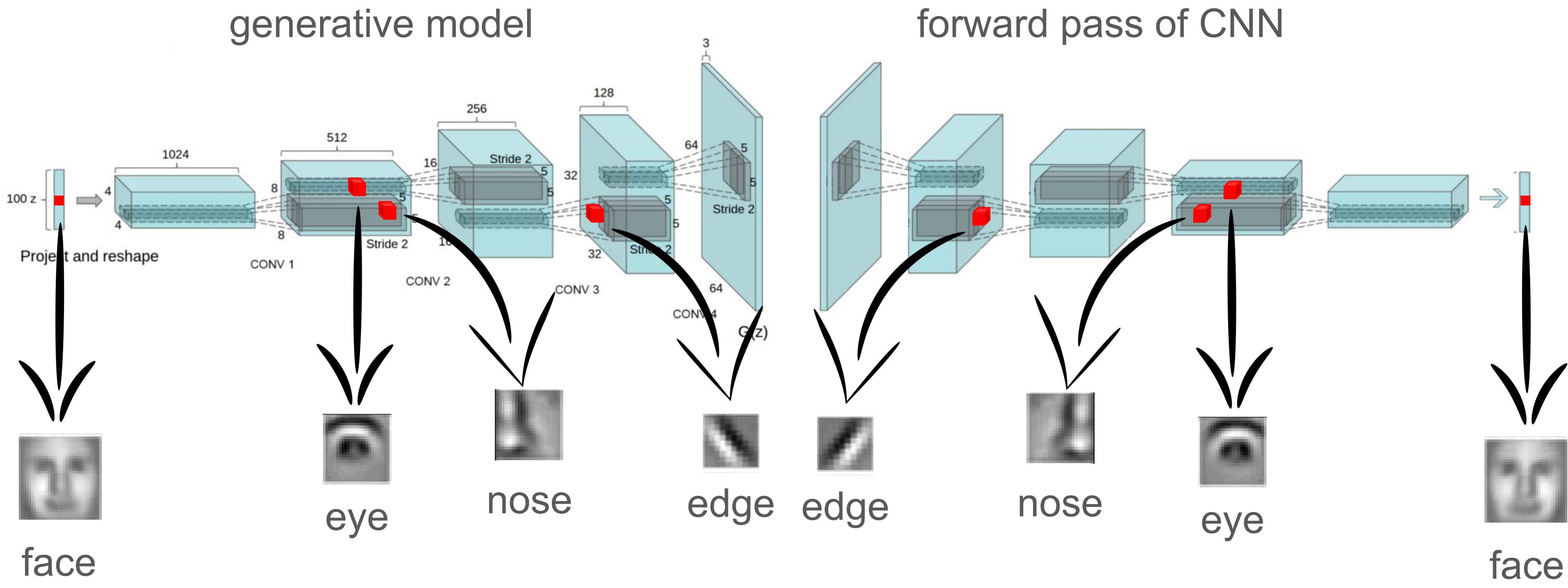


What generative model?



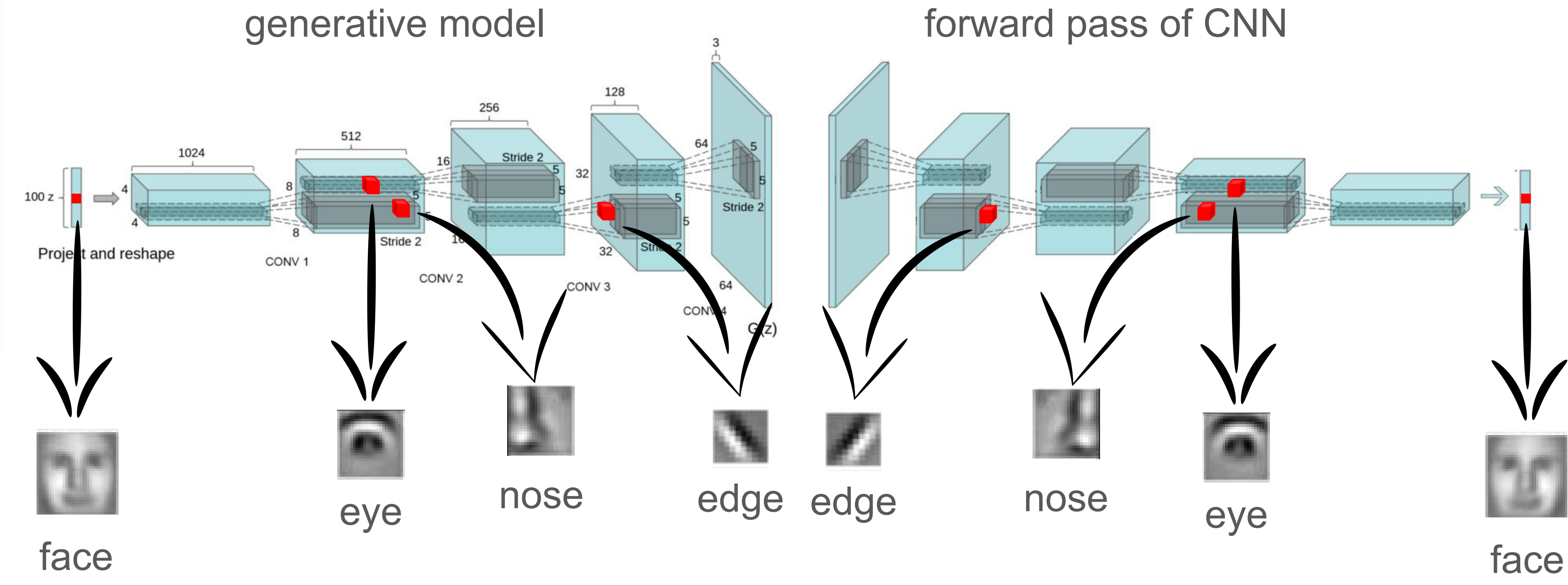
Properties of model?

for **hierarchical compositional** functions **deep** but not shallow networks avoid the curse of dimensionality because of **locality** of constituent functions



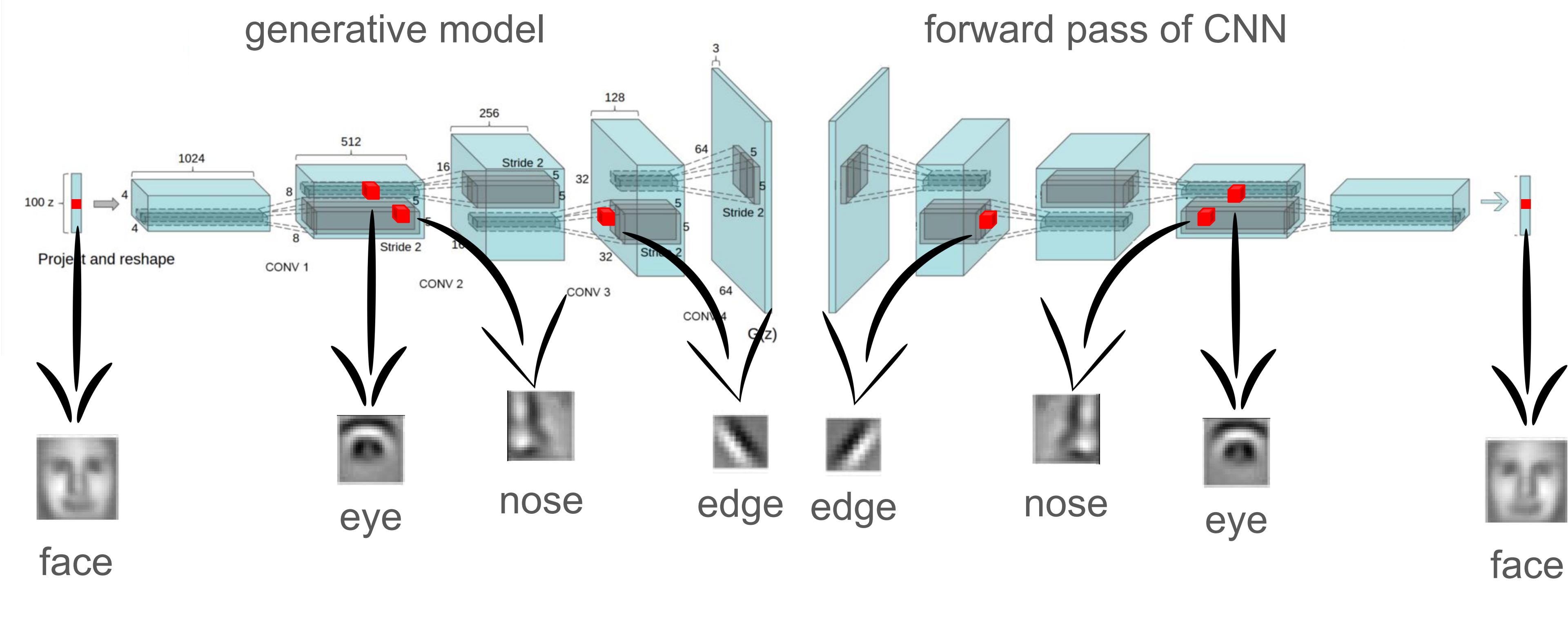
Properties of model?

Weights and pre-activations are i.i.d Gaussian

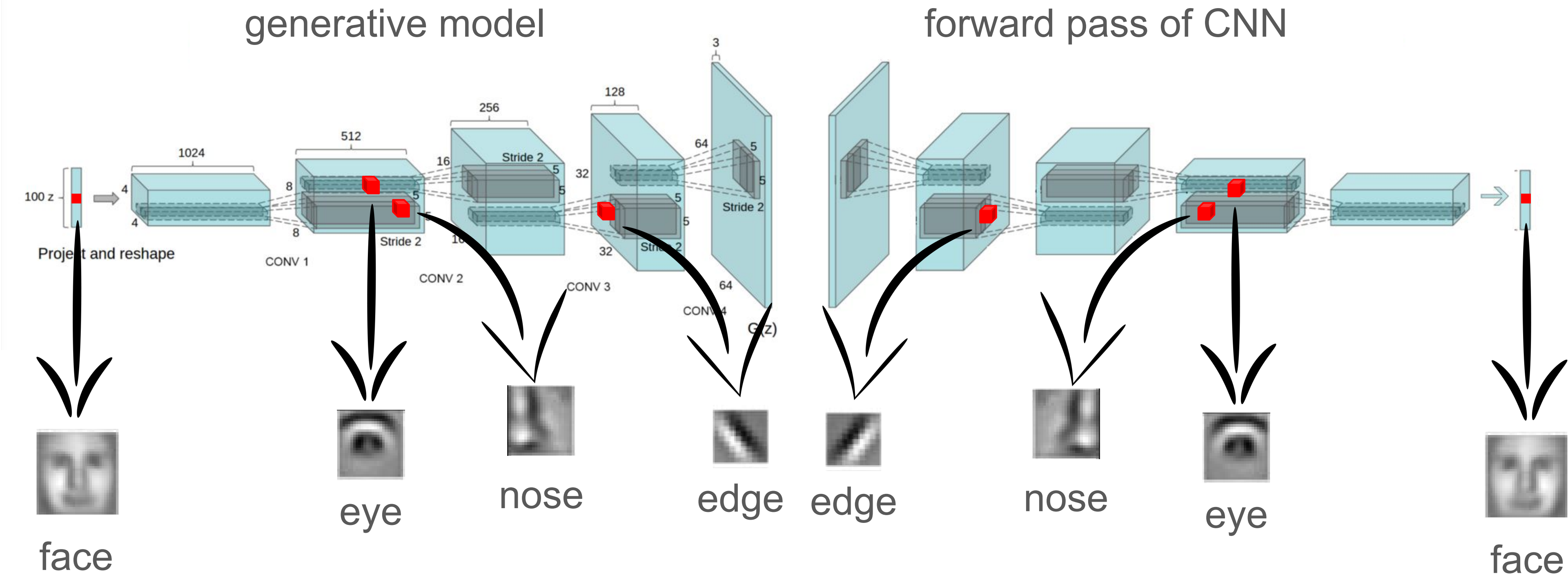


Properties of model?

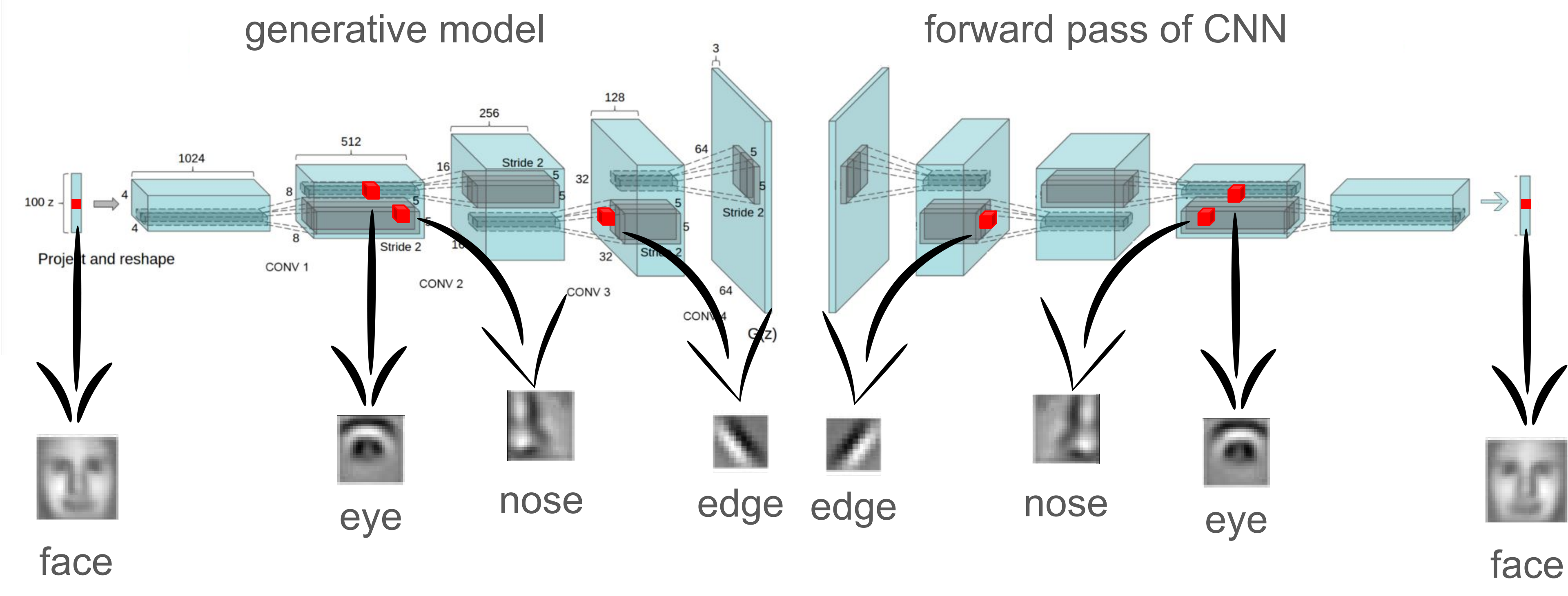
Overparameterization is good for optimization



Success of inference?

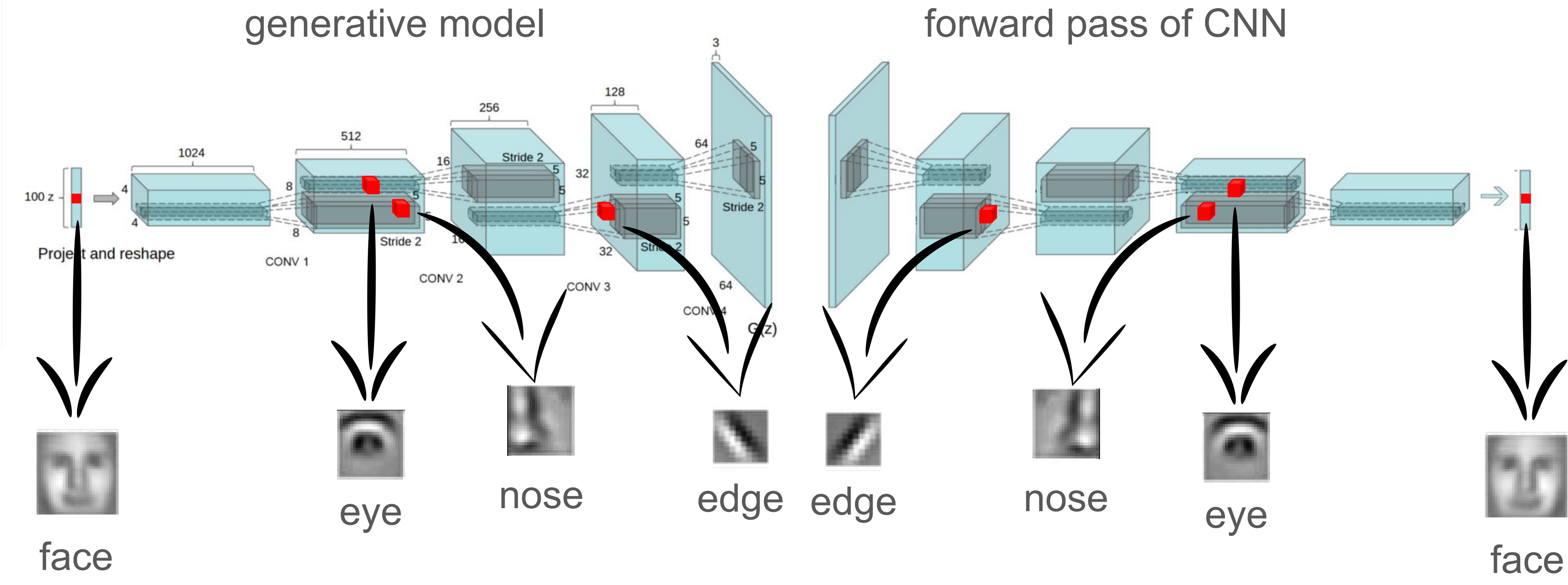


Uniqueness of representation?



Stability to perturbations?

$$|||\Phi(F_\tau f) - \Phi(f)||| \leq C_{\mathcal{C}} \|\tau\|_\infty^\alpha$$

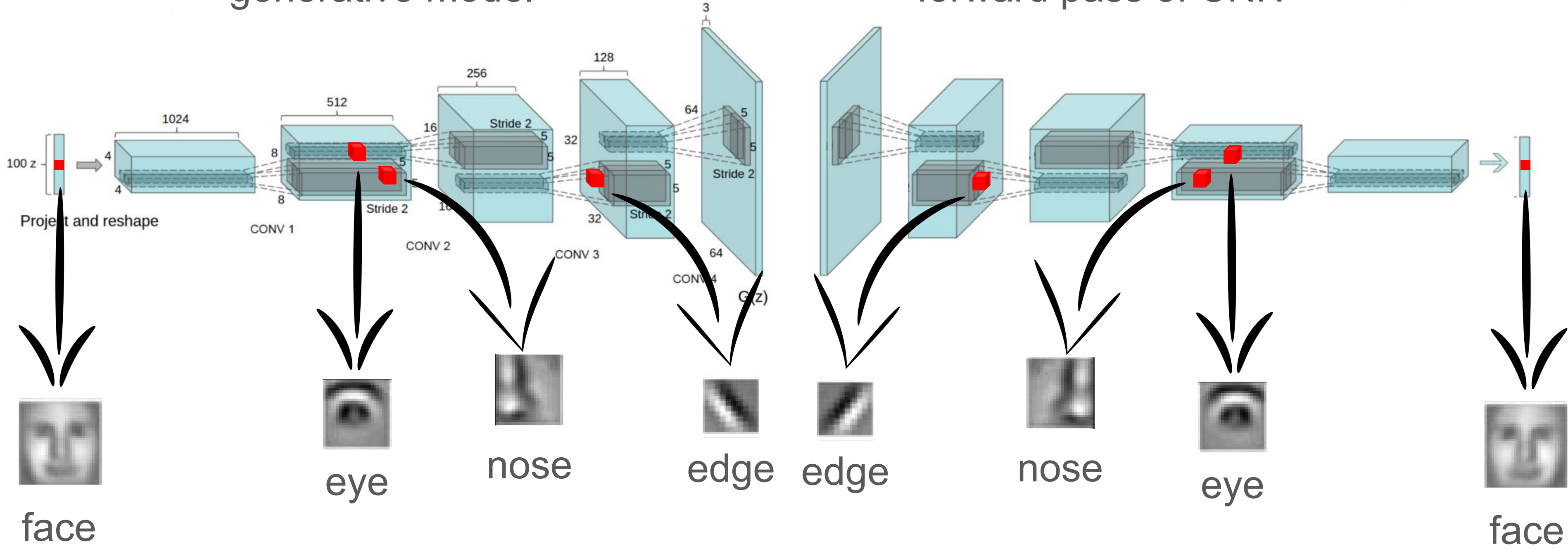


Better inference?

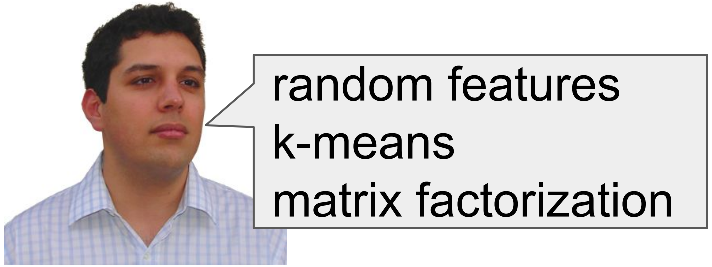
Information should propagate both within and between levels of representation in a bidirectional manner

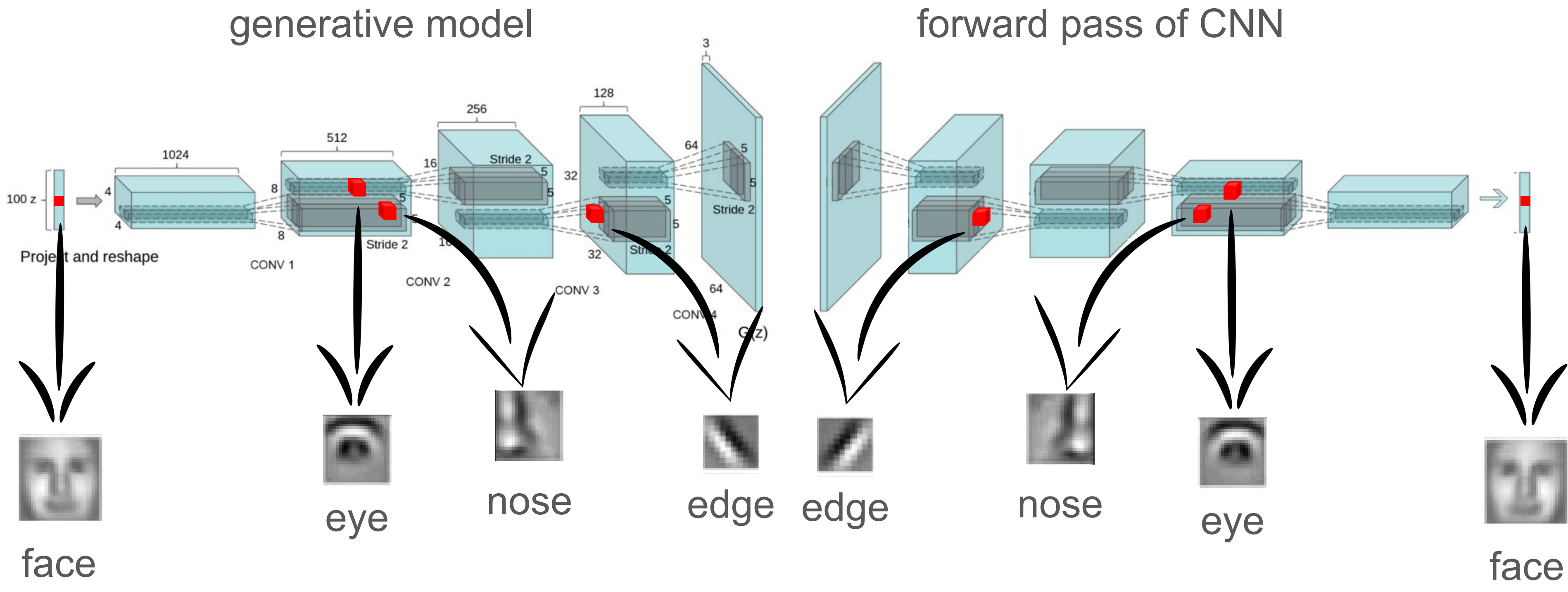
generative model

forward pass of CNN



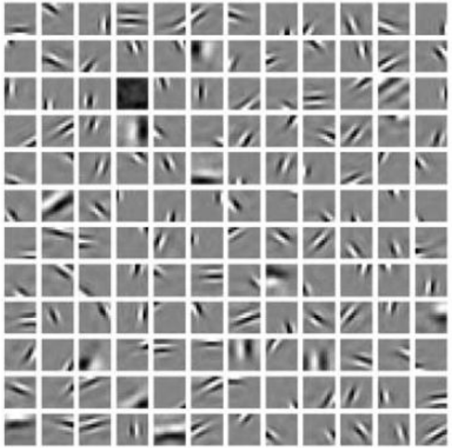
Better training?



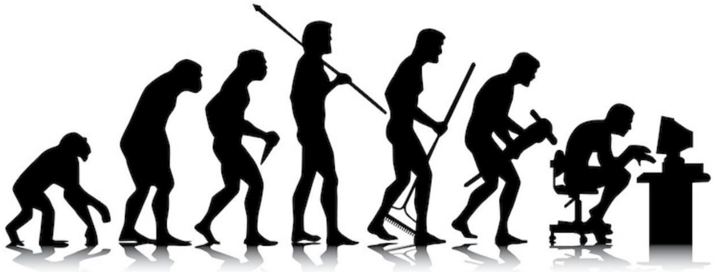


Sparse Representation Generative Model

- Receptive fields in visual cortex are spatially localized, oriented and bandpass
- Coding natural images while promoting sparse solutions results in a set of filters satisfying these properties [Olshausen and Field 1996]
- Two decades later…
  - vast theoretical study
  - different inference algorithms
  - different ways to train the model



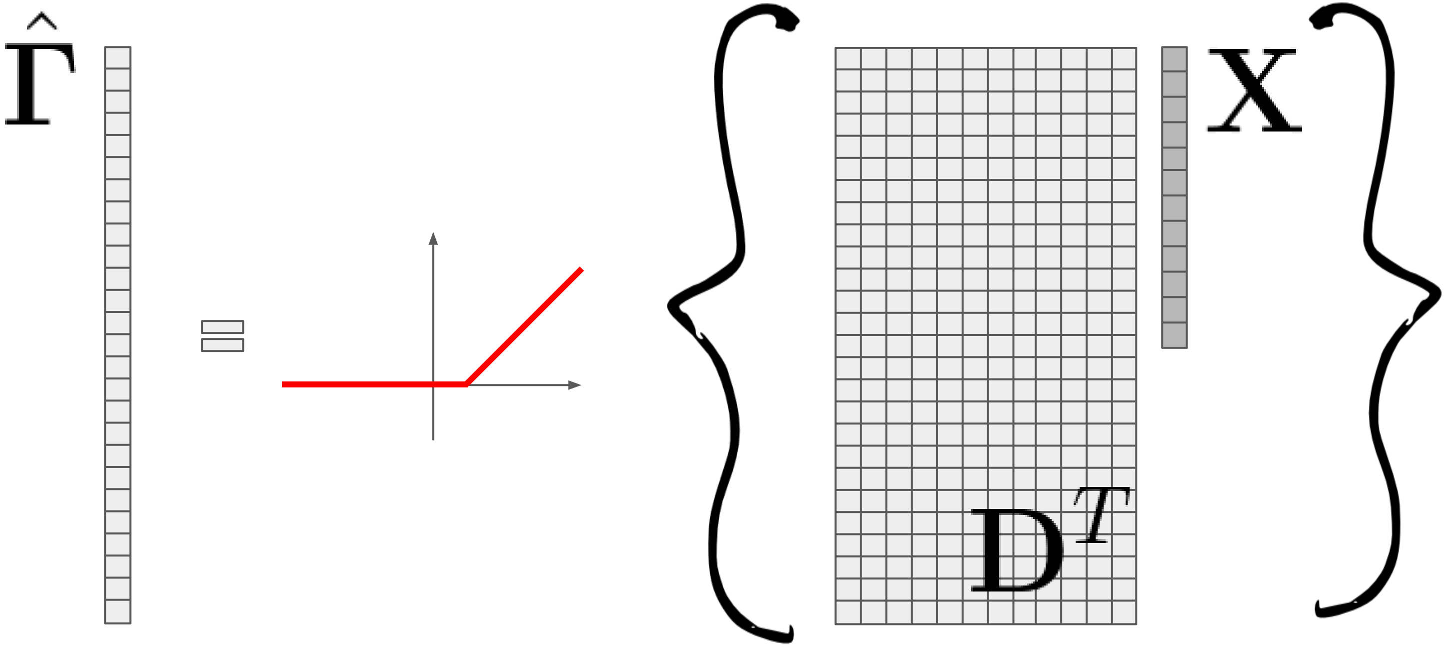
Evolution of Models



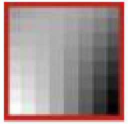
Multi-Layered Convolutional Neural Network → First Layer of a Convolutional Neural Network → First Layer of a Neural Network

Multi-Layered Convolutional Sparse Representation ← Convolutional sparse representation ← Sparse representations

First Layer of a Neural Network



Sparse Modeling

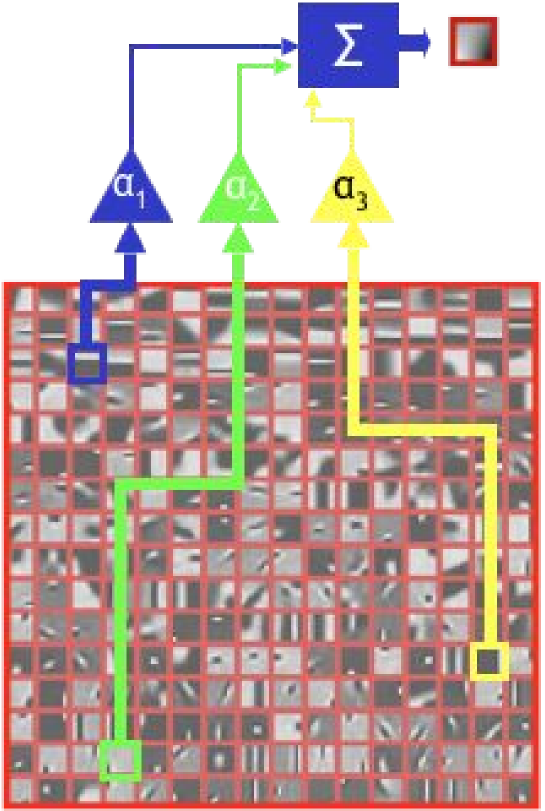


<u>Task:</u> model image patches of size 8x8 pixels

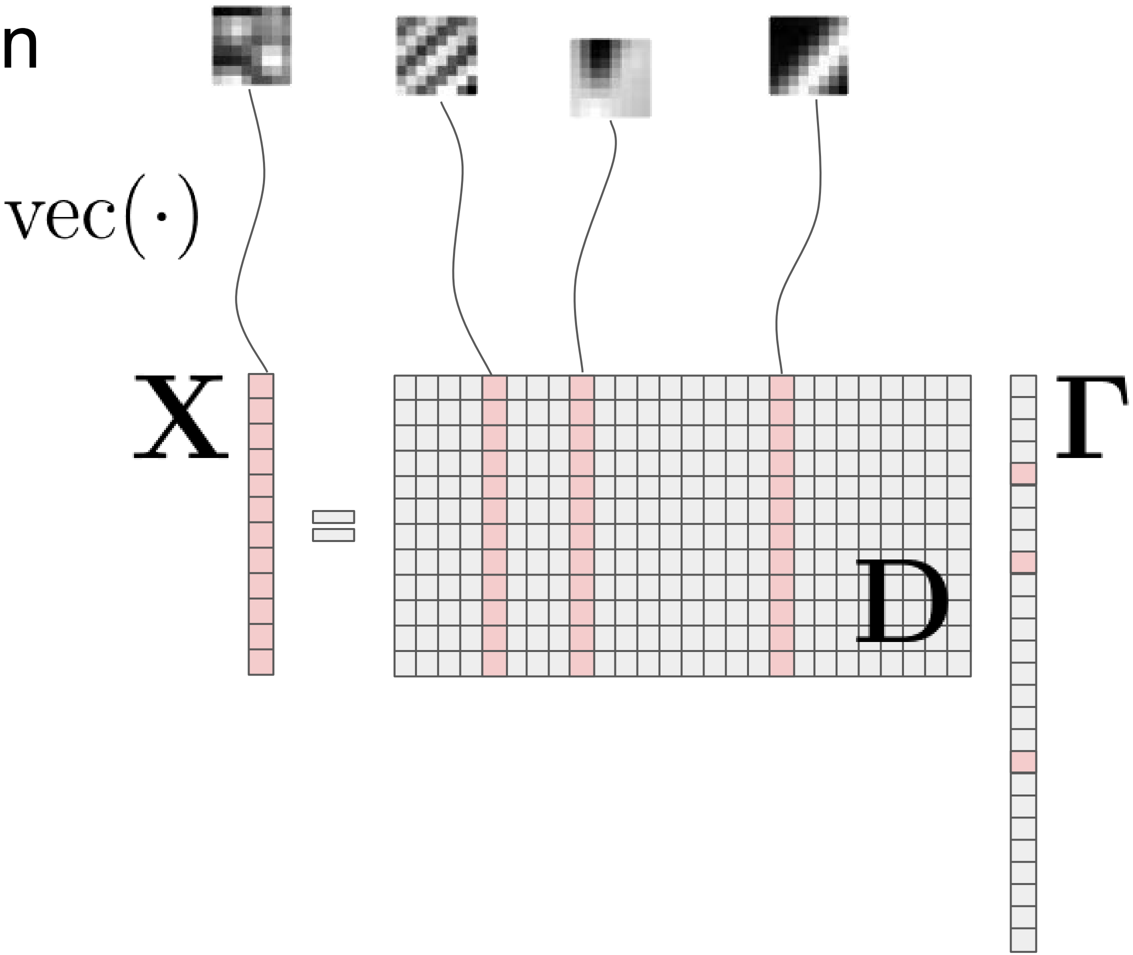
We assume a dictionary of such image patches is given, containing 256 atoms

<u>Assumption:</u> every patch can be described as a linear combination of a few atoms

<u>Key properties:</u> sparsity and redundancy



Matrix Notation



Sparse Coding

Given a signal, we would like to find its sparse representation

$$\min_{\mathbf{\Gamma}} \ \|\mathbf{\Gamma}\|_0 \ \text{ s.t. } \ \mathbf{X} = \mathbf{D}\mathbf{\Gamma}$$

Convexify

$$\min_{\mathbf{\Gamma}} \ \|\mathbf{\Gamma}\|_1 \ \text{ s.t. } \ \mathbf{X} = \mathbf{D}\mathbf{\Gamma}$$

Sparse Coding

Given a signal, we would like to find its sparse representation

$$\min_{\mathbf{\Gamma}} \ \|\mathbf{\Gamma}\|_0 \ \text{ s.t. } \ \mathbf{X} = \mathbf{D}\mathbf{\Gamma}$$

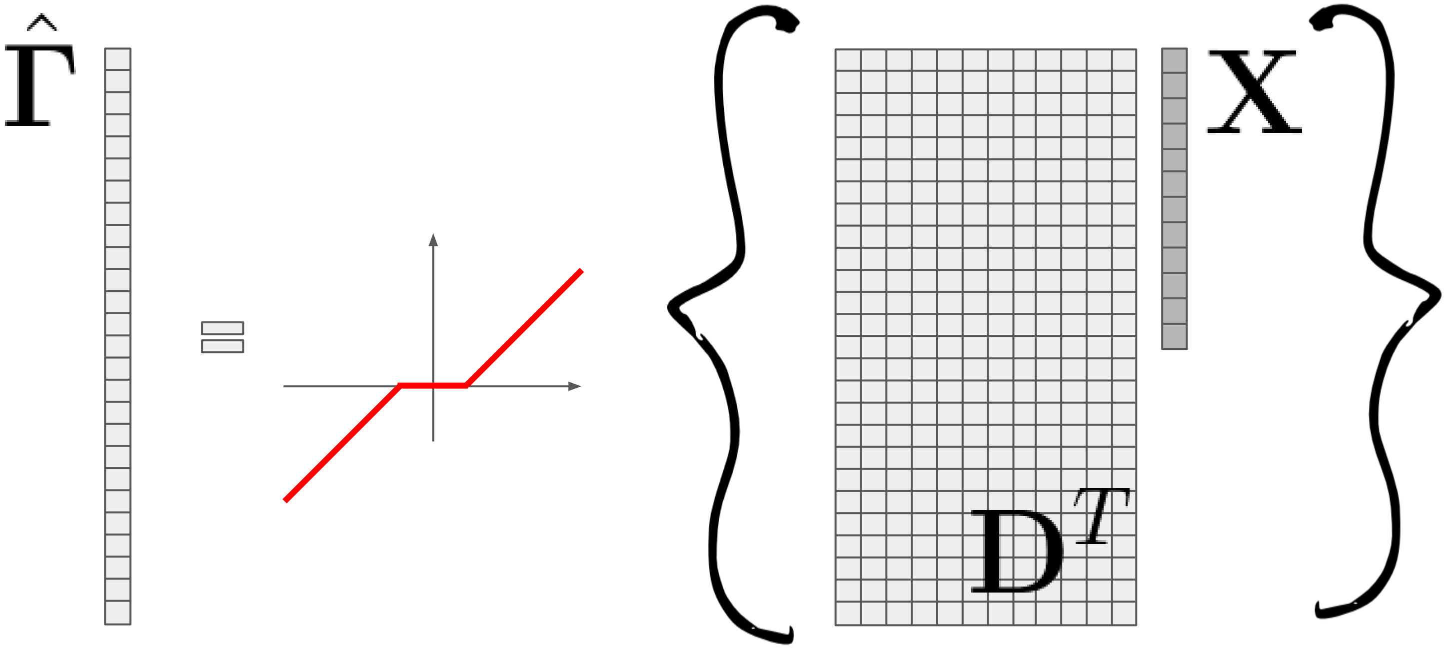
Convexify

$$\min_{\mathbf{\Gamma}} \ \|\mathbf{\Gamma}\|_1 \ \text{ s.t. } \ \mathbf{X} = \mathbf{D}\mathbf{\Gamma}$$

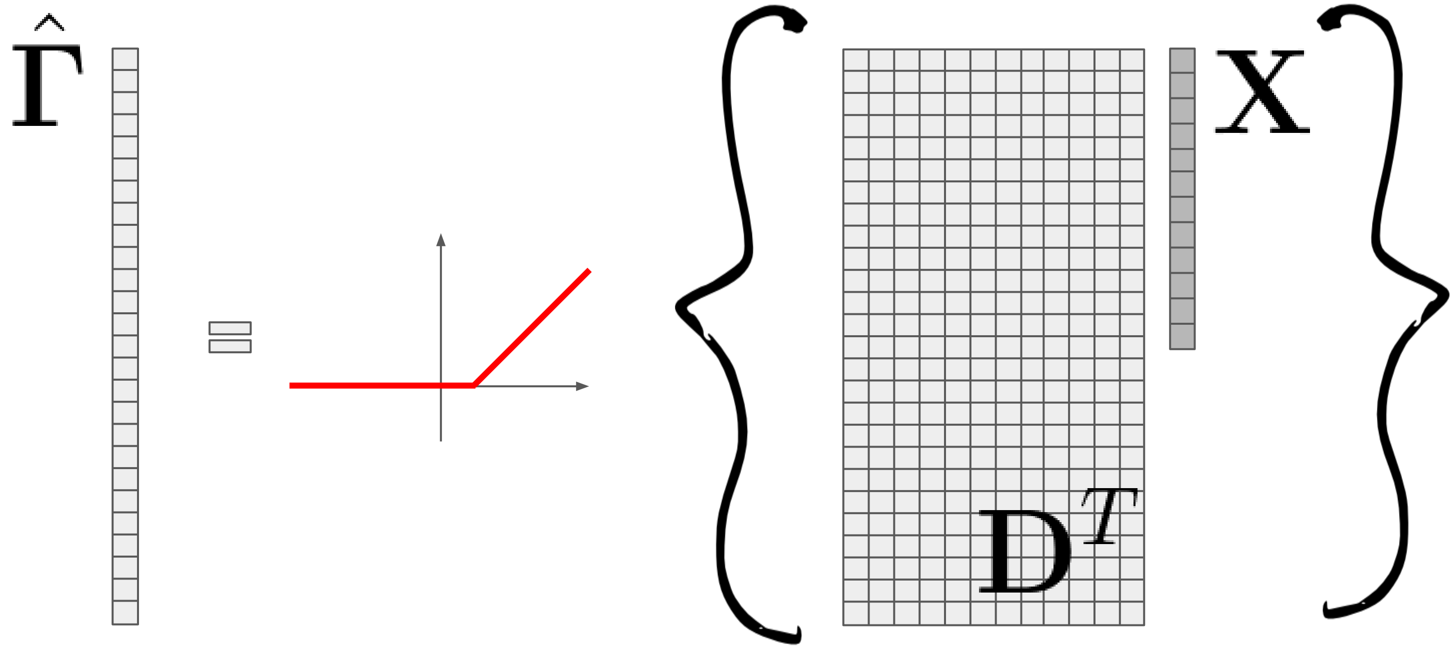
Crude approximation

$$\mathcal{S}_\beta\{\mathbf{D}^T\mathbf{X}\}$$

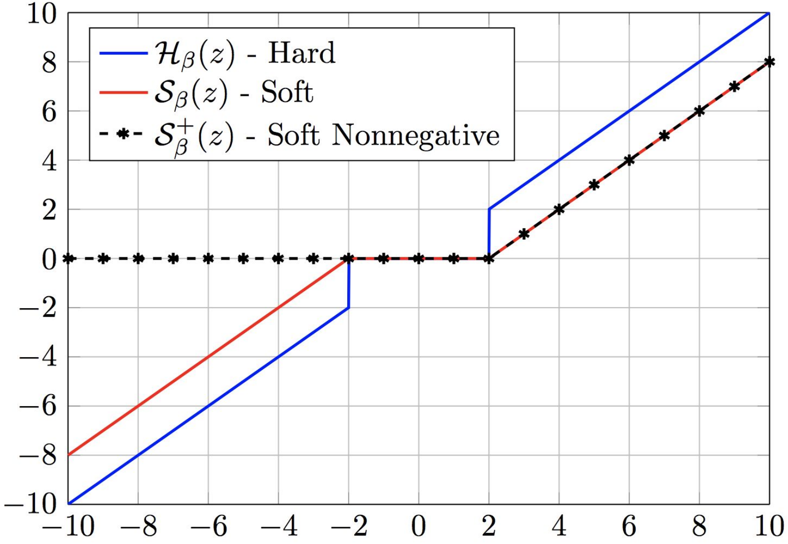
Thresholding Algorithm



First Layer of a Neural Network

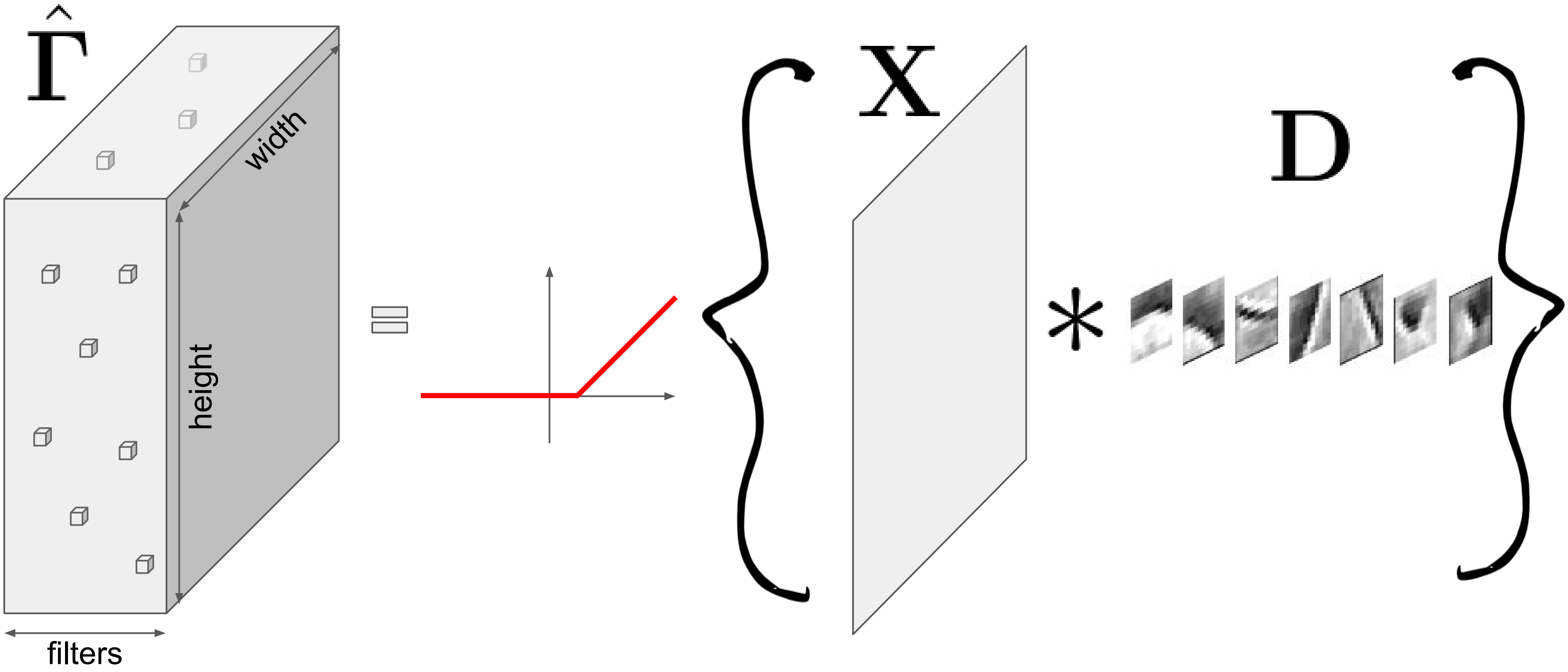


ReLU = Soft Nonnegative Thresholding

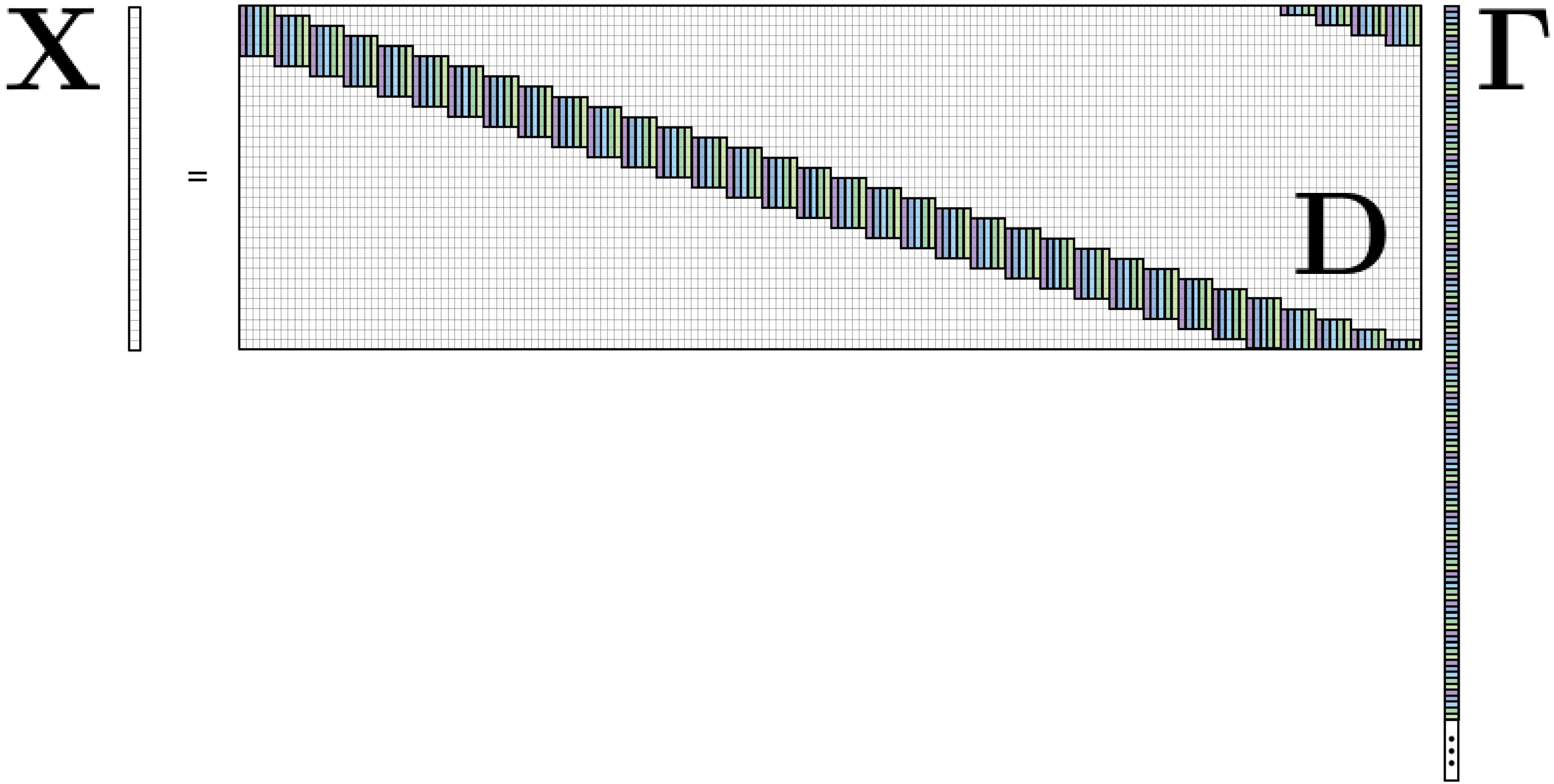


ReLU is equivalent to soft nonnegative thresholding

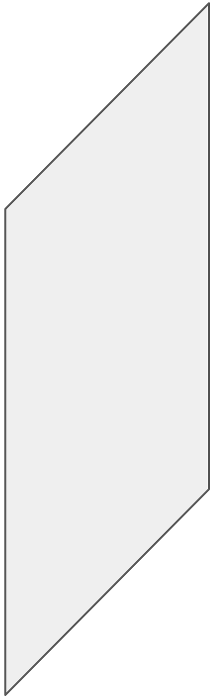
First layer of a **Convolutional** Neural Network



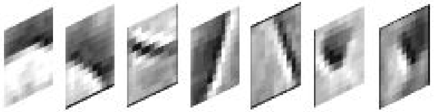
Convolutional Sparse Modeling

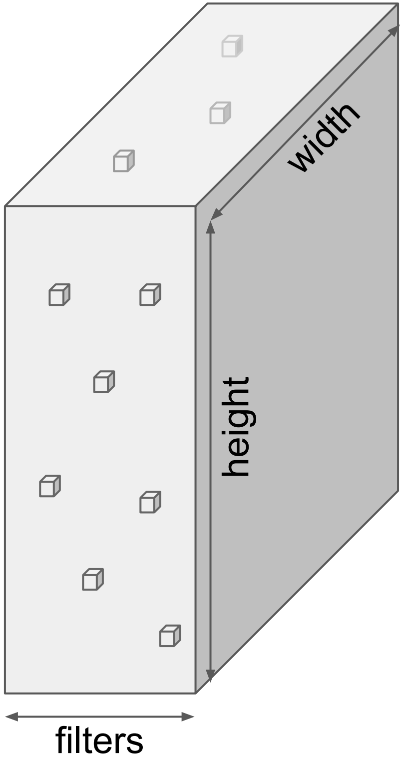


Convolutional Sparse Modeling

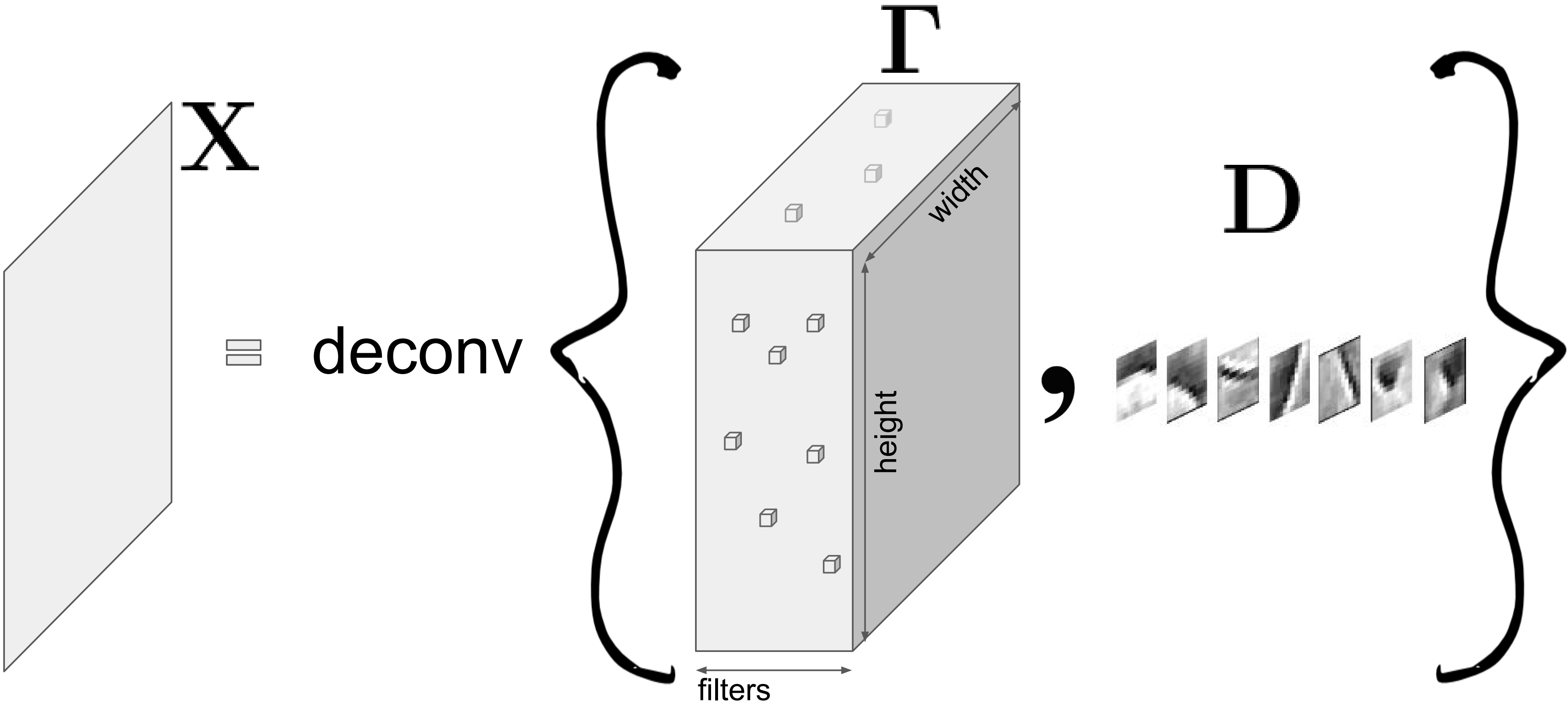


$$\mathbf{X} = \sum_{k=1}^{K} \mathbf{d}_k * \mathbf{z}_k$$

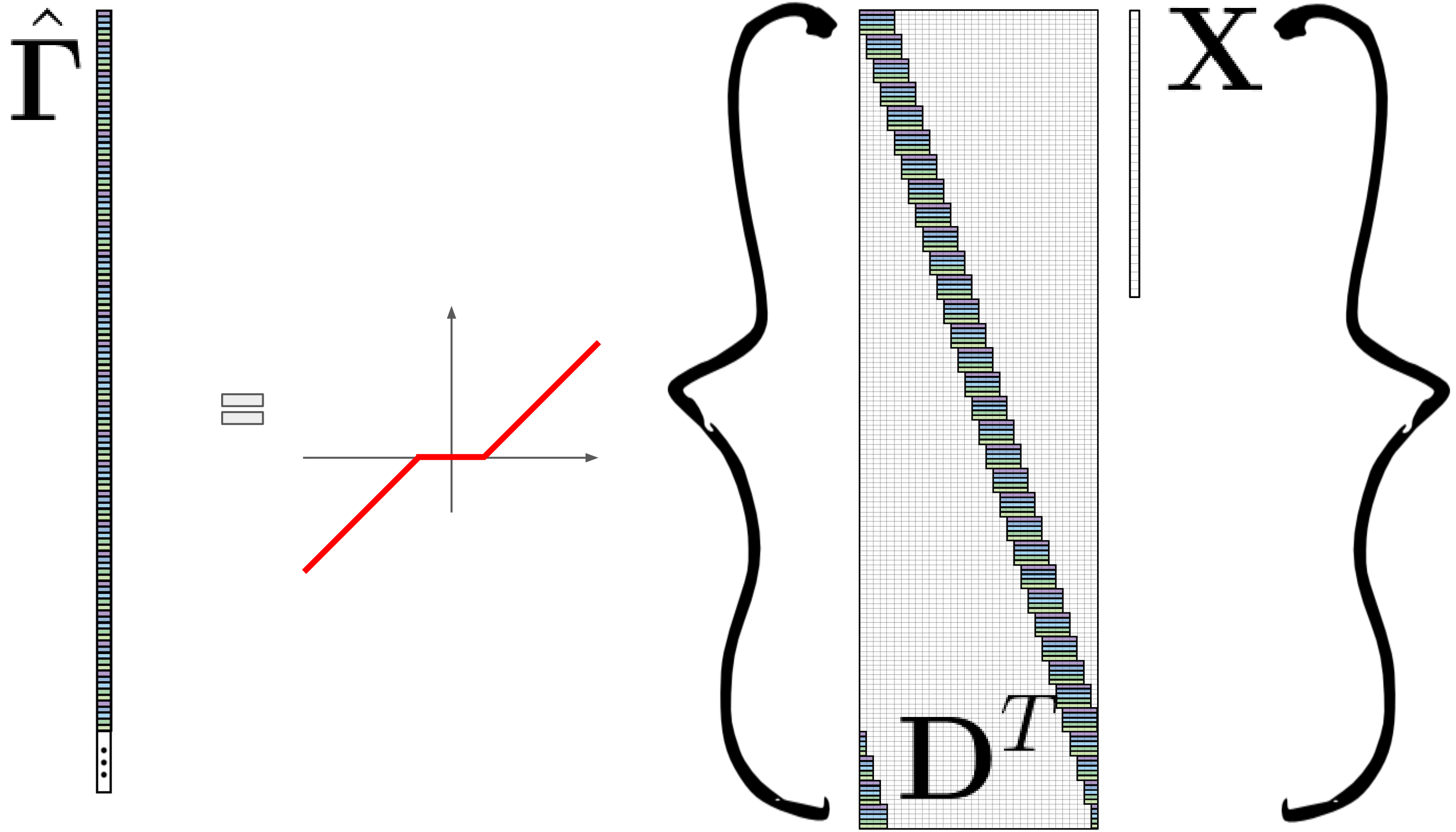




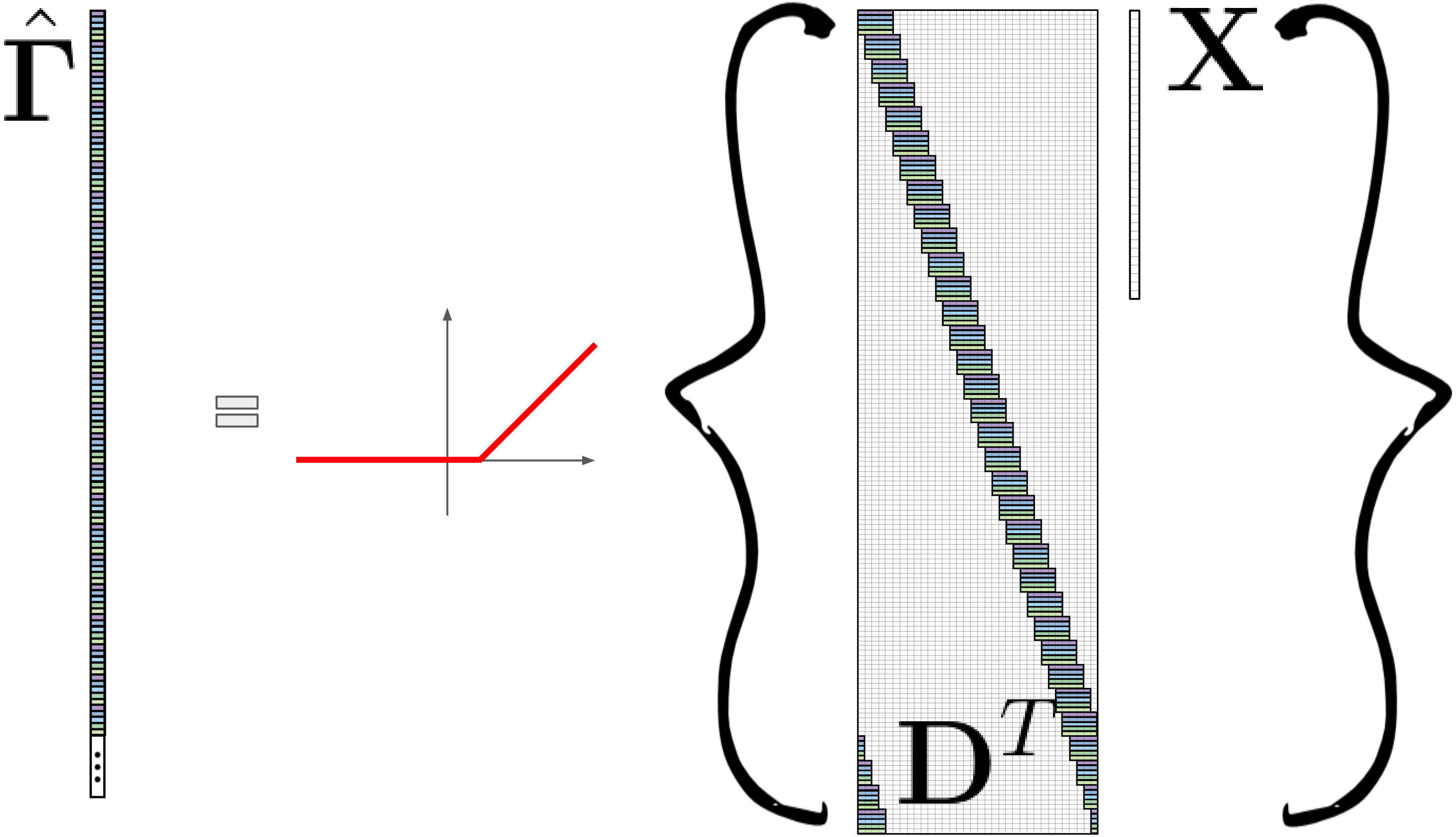
Convolutional Sparse Modeling



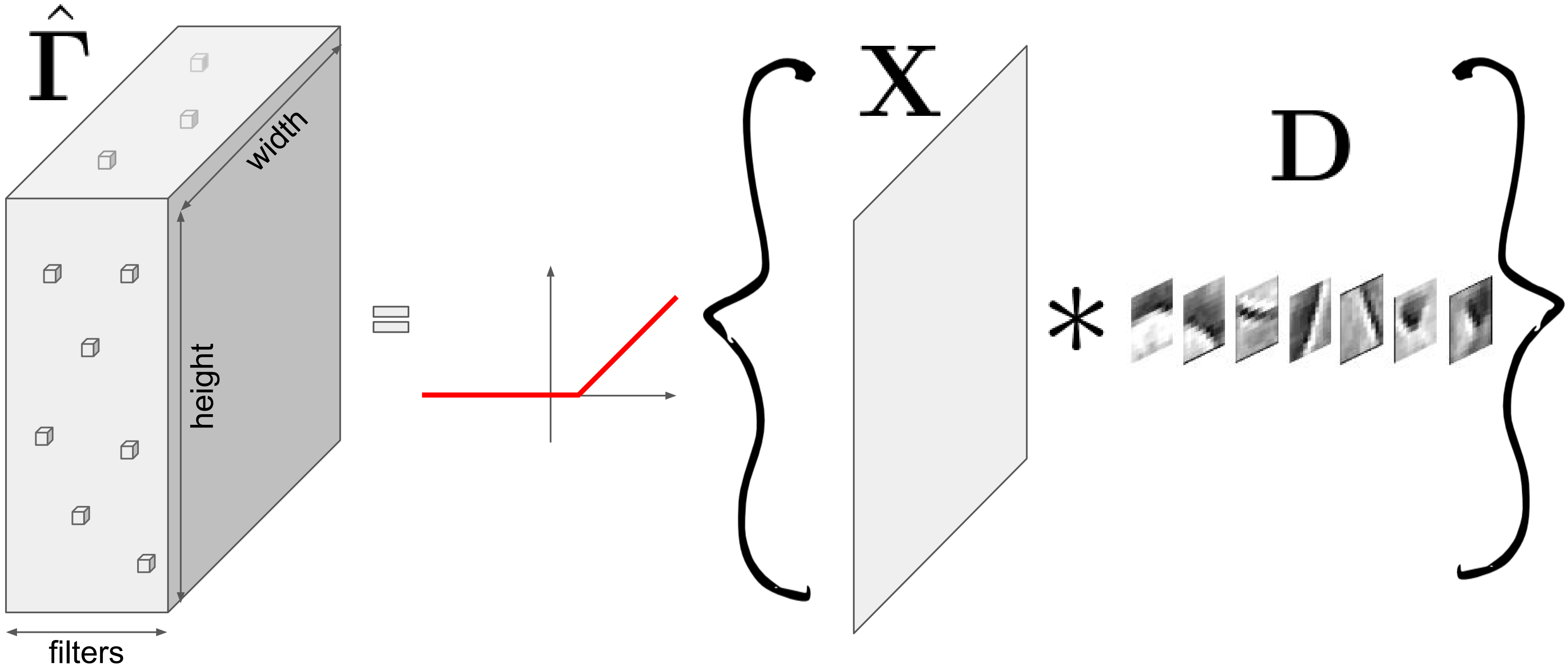
Thresholding Algorithm

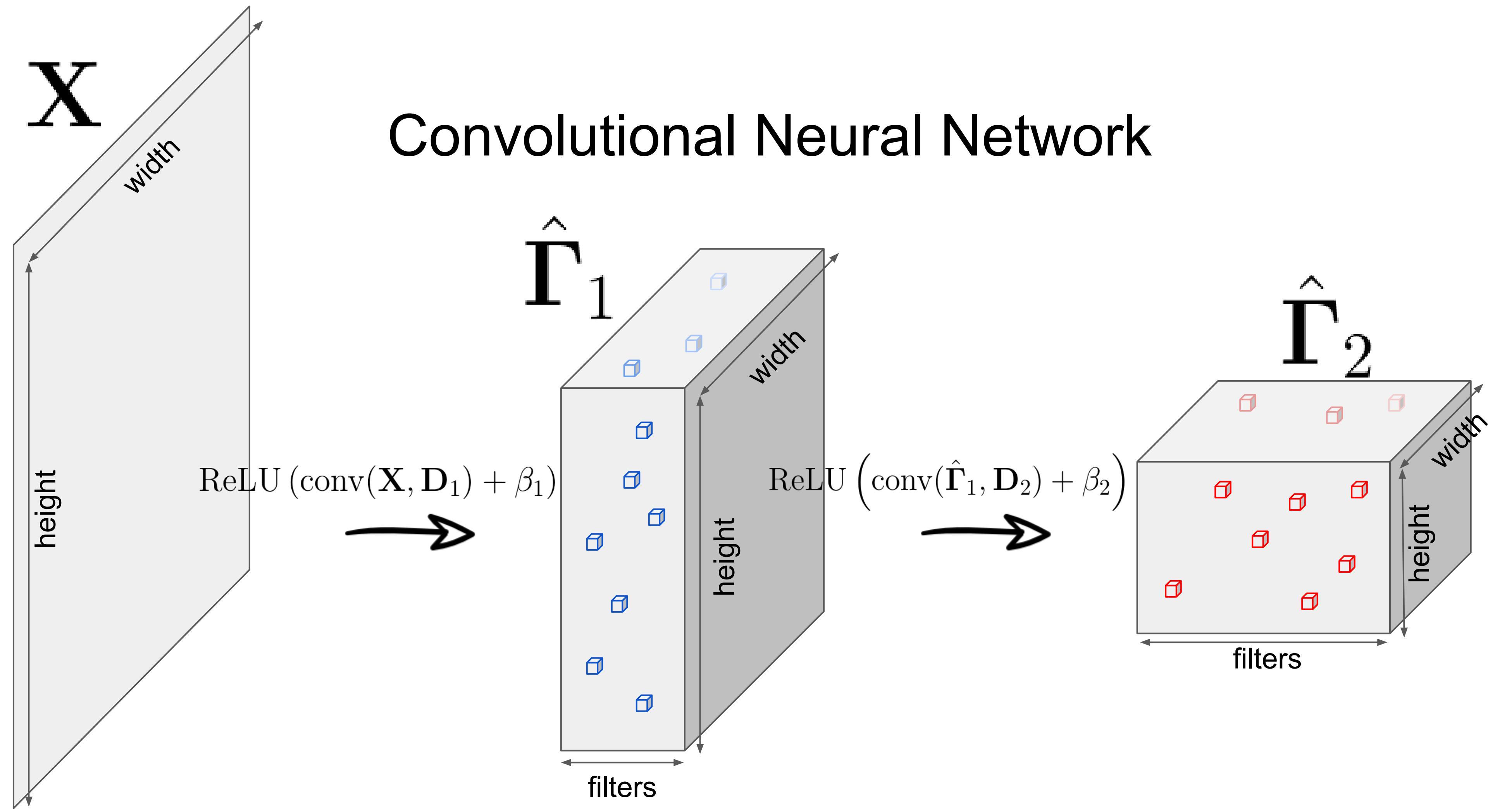


First layer of a Convolutional Neural Network

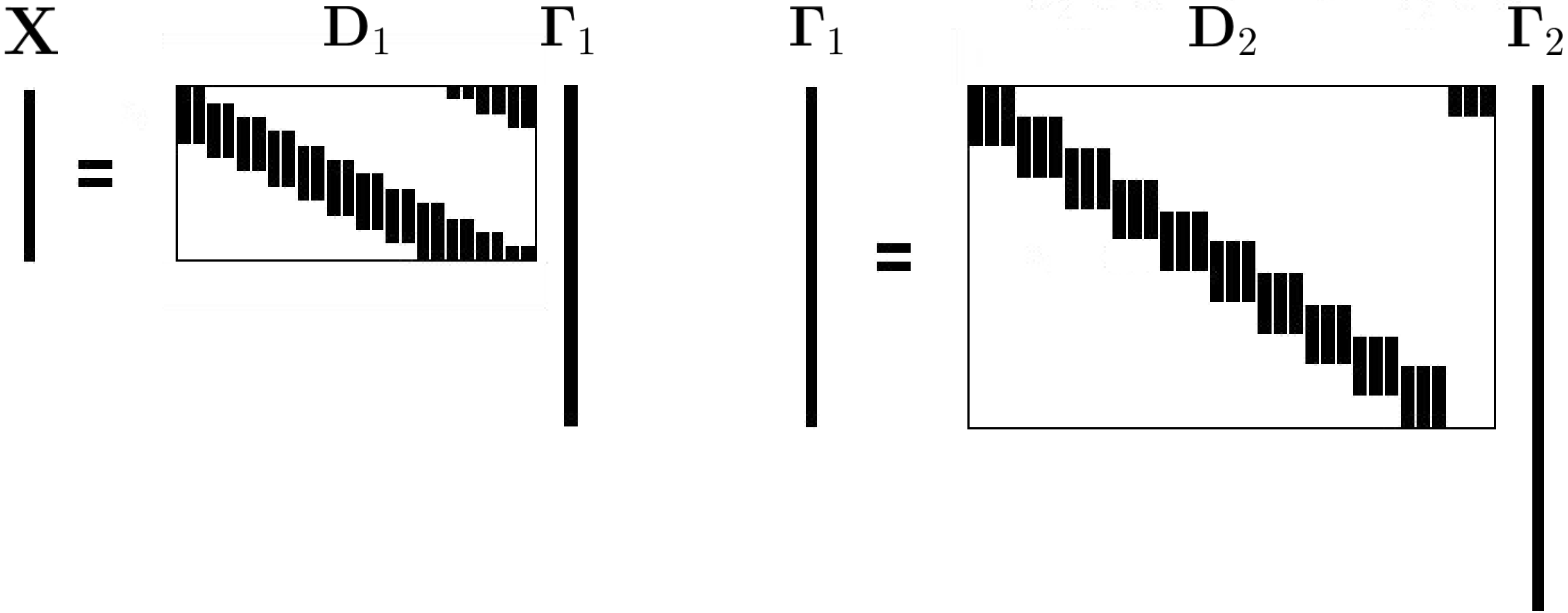


First layer of a Convolutional Neural Network

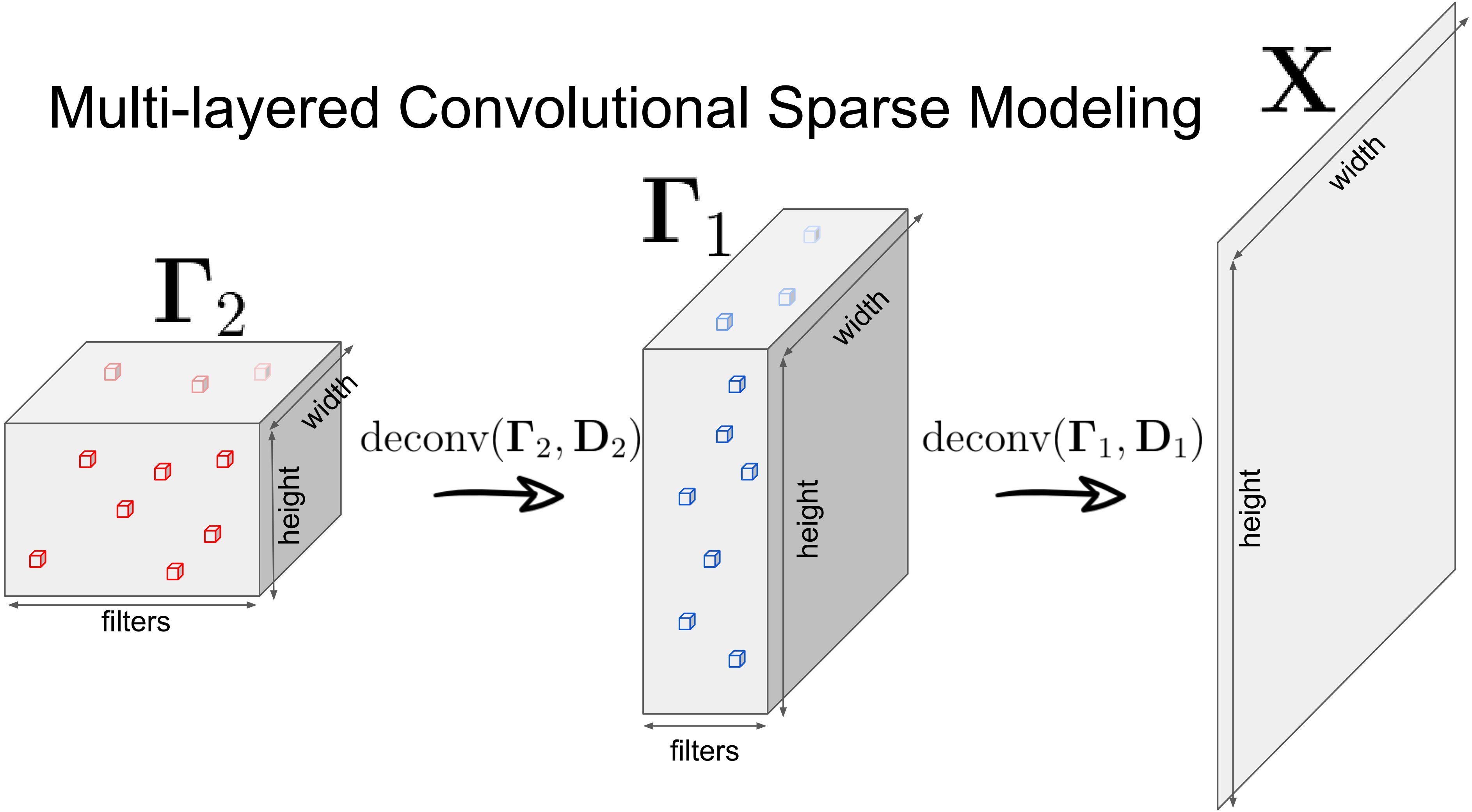



Convolutional Neural Network
$\mathbf{X}$
width
height
$\hat{\mathbf{\Gamma}}_1$
$\mathrm{ReLU}\left(\mathrm{conv}(\mathbf{X}, \mathbf{D}_1) + \beta_1\right)$
width
height
filters
$\hat{\mathbf{\Gamma}}_2$
$\mathrm{ReLU}\left(\mathrm{conv}(\hat{\mathbf{\Gamma}}_1, \mathbf{D}_2) + \beta_2\right)$
width
height
filters

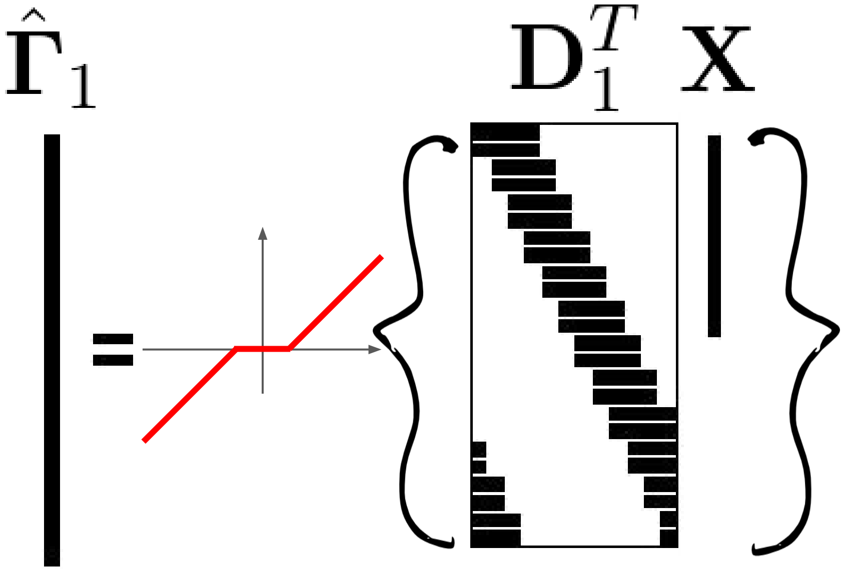
Multi-layered Convolutional Sparse Modeling

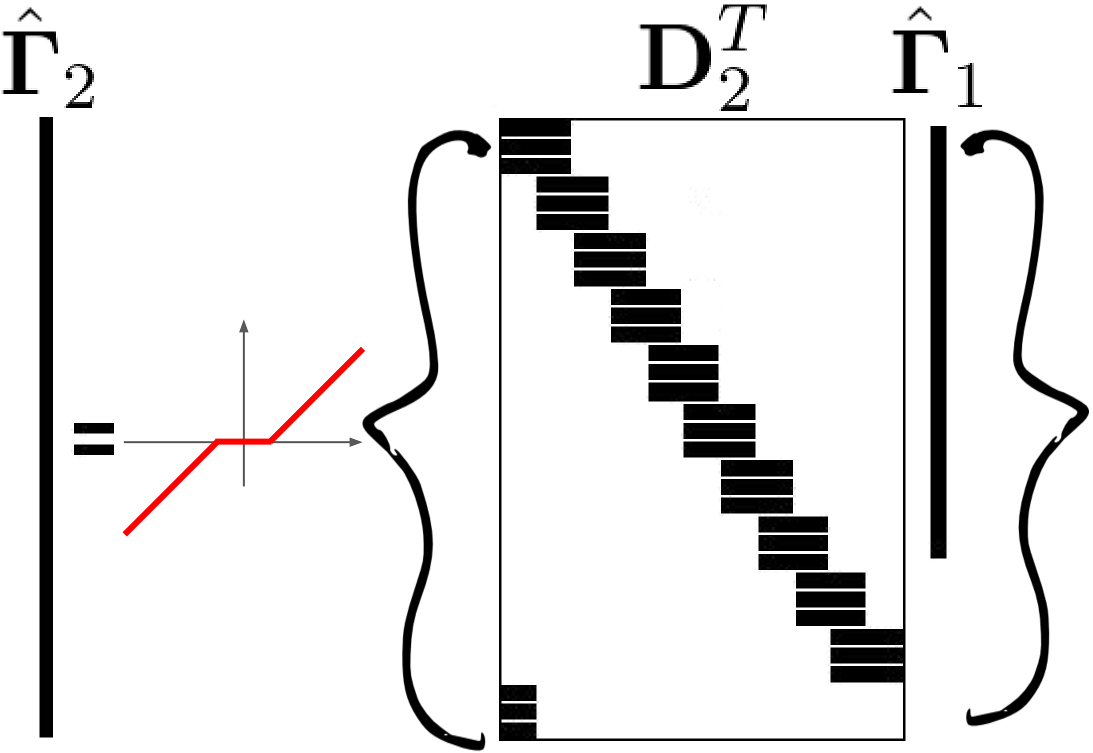


Multi-layered Convolutional Sparse Modeling

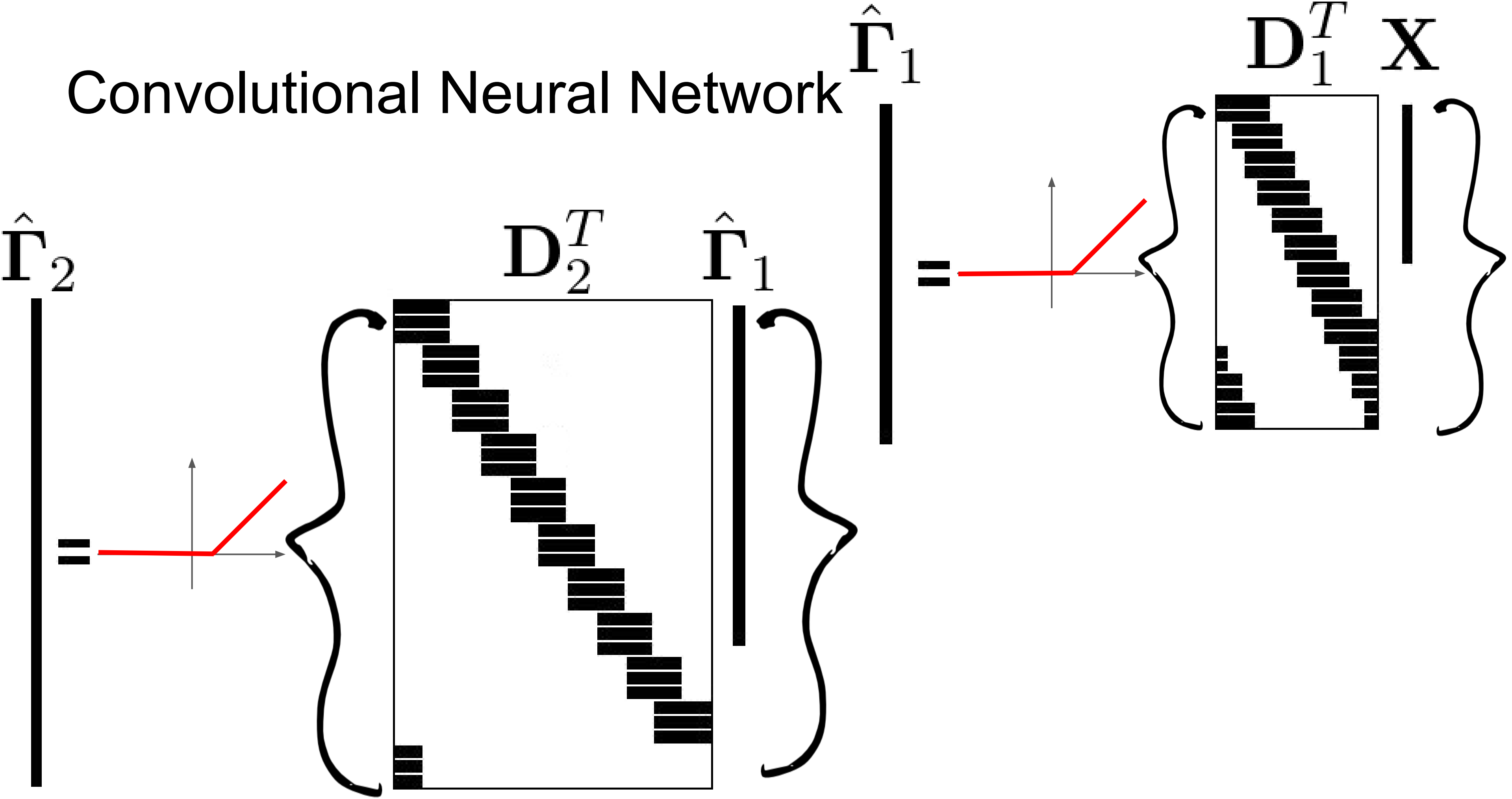


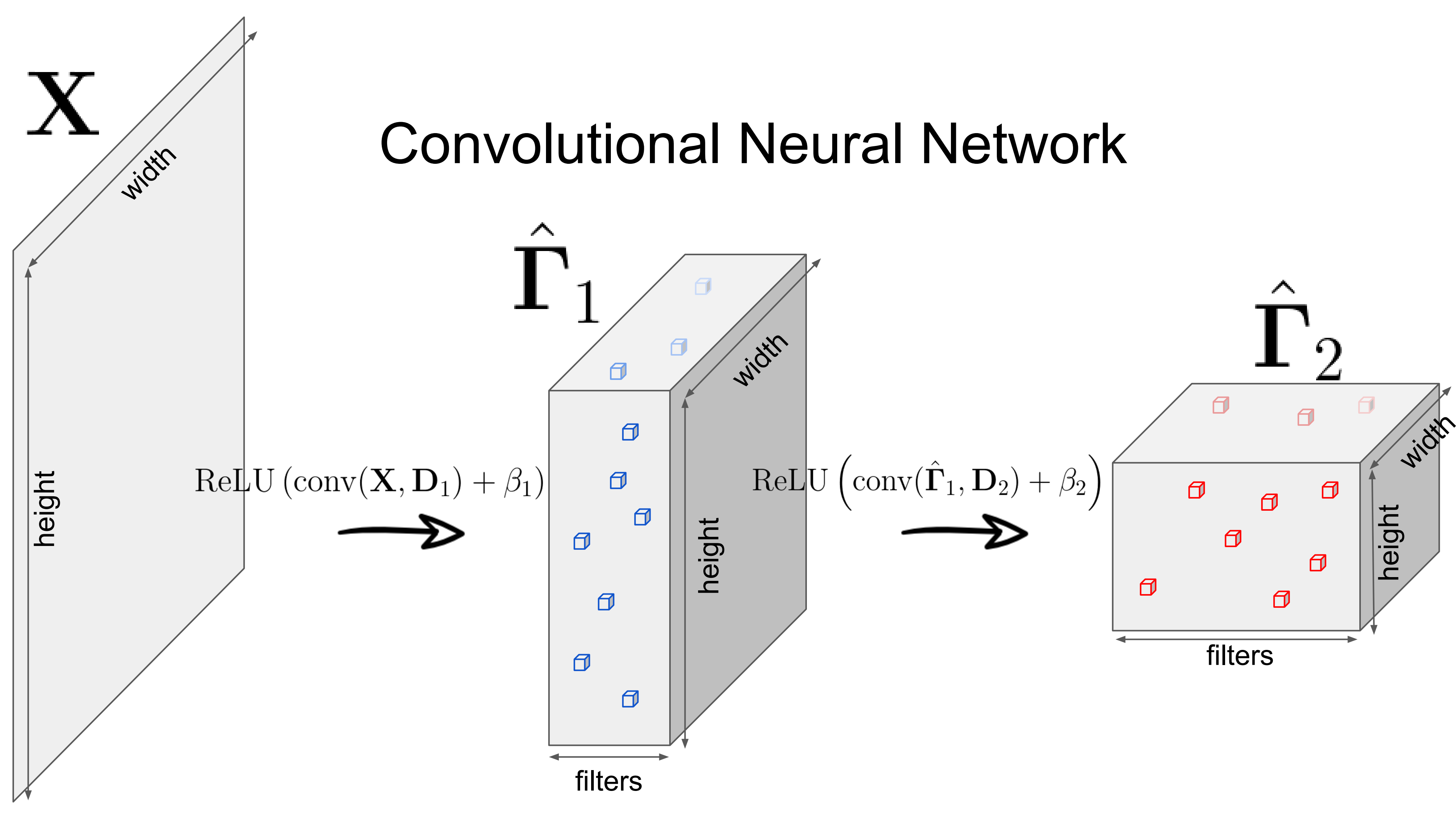
Layered Thresholding





Convolutional Neural Network

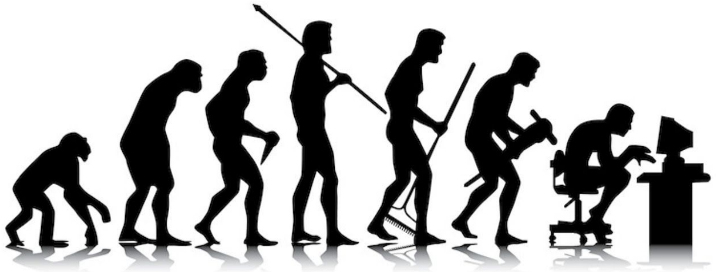




Convolutional Neural Network
$\mathbf{X}$
width
height
$\mathrm{ReLU}\left(\mathrm{conv}(\mathbf{X}, \mathbf{D}_1) + \beta_1\right)$
$\hat{\mathbf{\Gamma}}_1$
width
height
filters
$\mathrm{ReLU}\left(\mathrm{conv}(\hat{\mathbf{\Gamma}}_1, \mathbf{D}_2) + \beta_2\right)$
$\hat{\mathbf{\Gamma}}_2$
width
height
filters

Theories of Deep Learning

Evolution of Models



Multi-Layered Convolutional Neural Network

First Layer of a Convolutional Neural Network

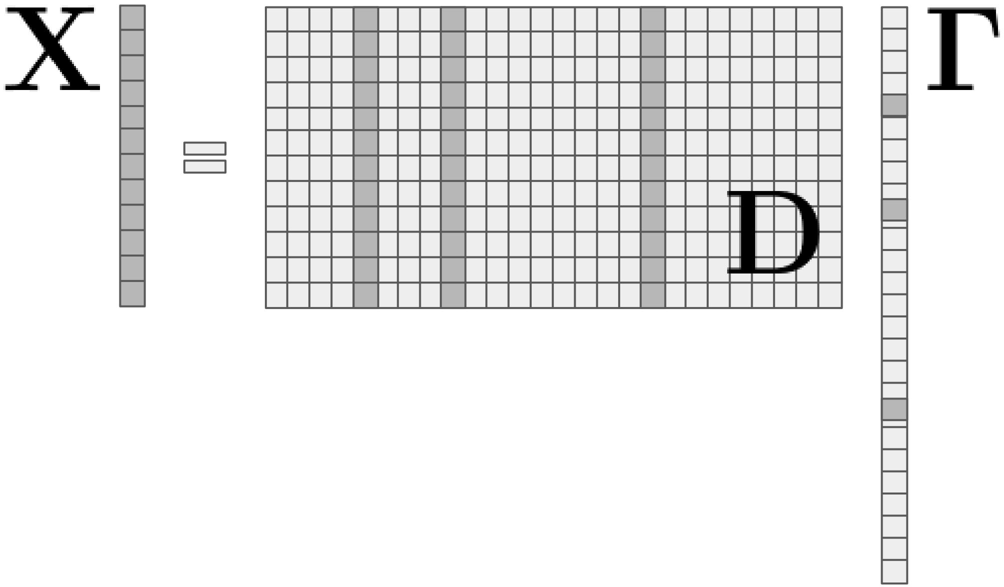
First Layer of a Neural Network

Multi-Layered Convolutional Sparse Representation

Convolutional sparse representation

Sparse representations

Sparse Modeling



Classic Sparse Theory

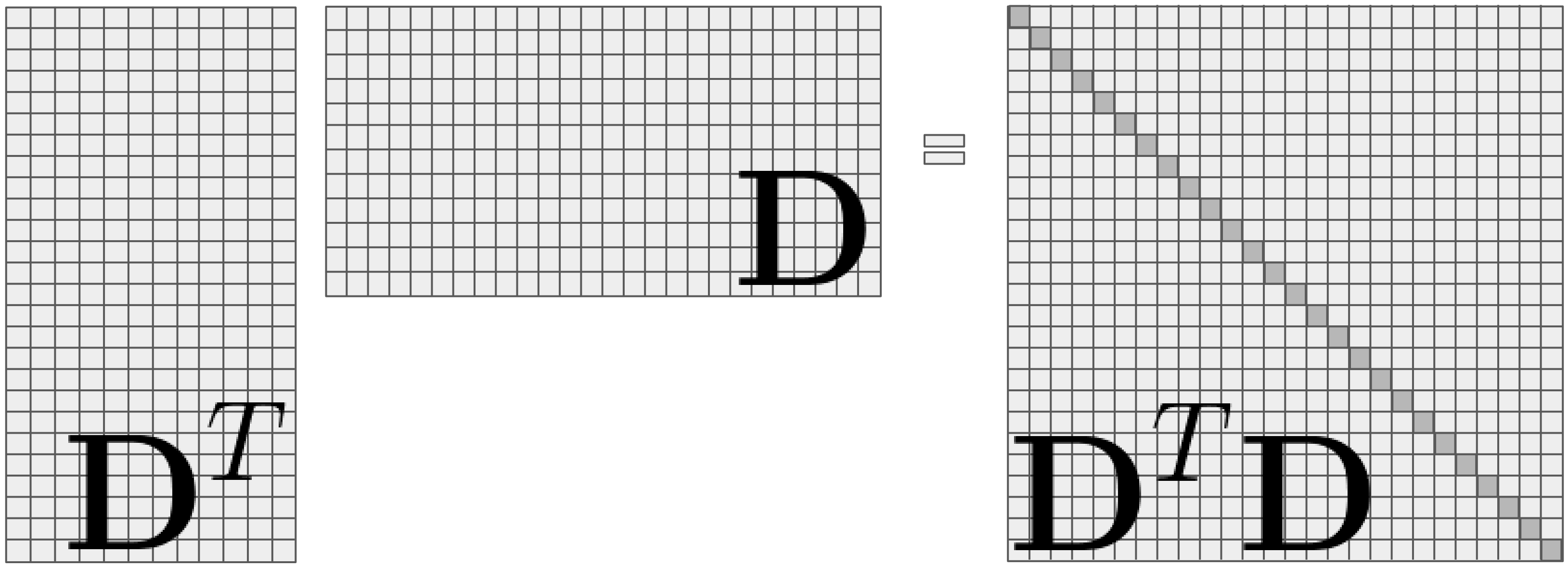
$$\mathbf{X} = \mathbf{D}\mathbf{\Gamma}$$

$$\hat{\mathbf{\Gamma}} = \arg\min_{\mathbf{\Gamma}} \ \|\mathbf{\Gamma}\|_1 \ \ \text{s.t.} \ \ \mathbf{X} = \mathbf{D}\mathbf{\Gamma}$$

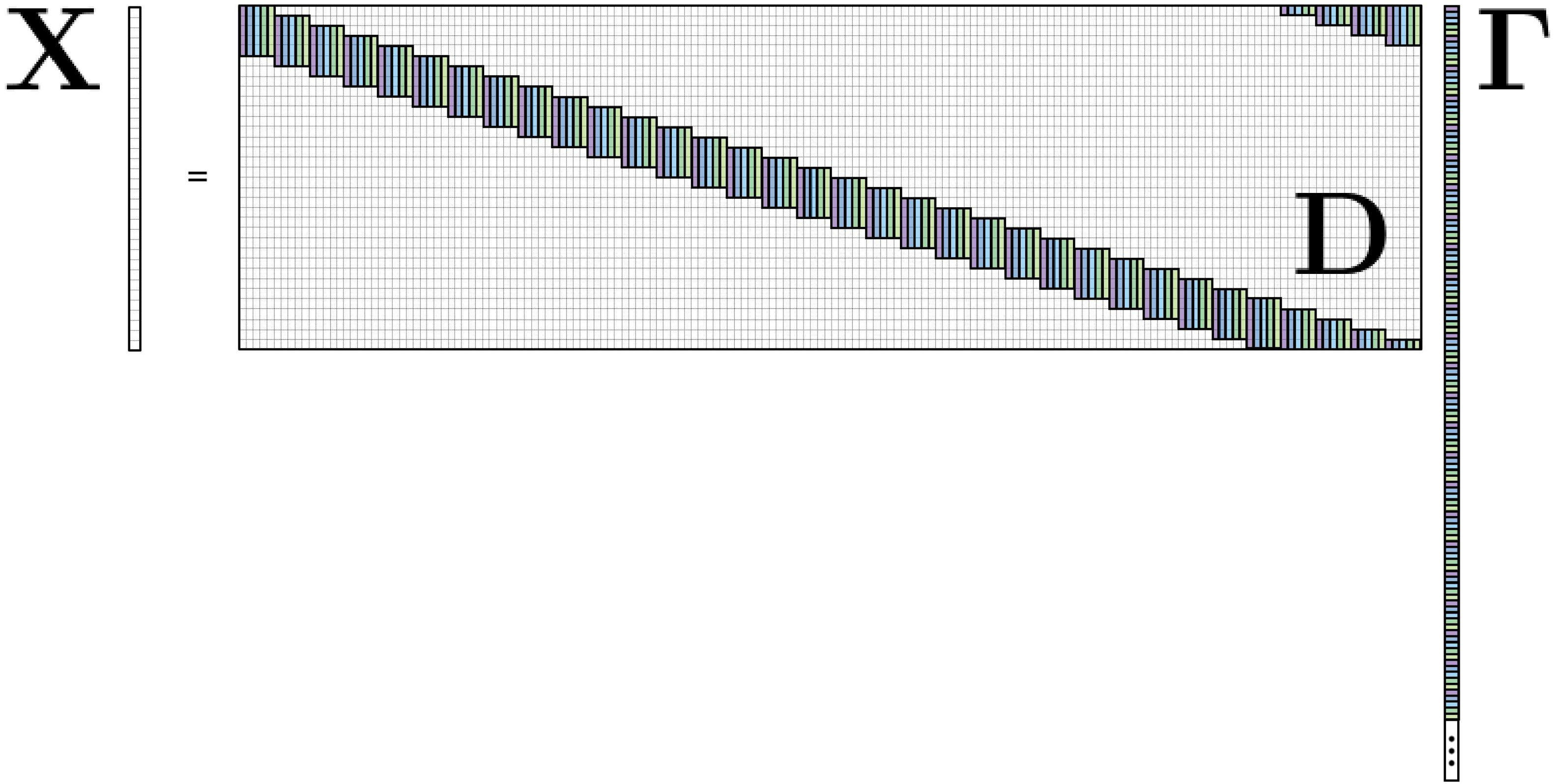
Theorem: [Donoho and Elad, 2003]
Basis pursuit is guaranteed to recover the true sparse vector assuming that

$$\|\mathbf{\Gamma}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right)$$

Mutual Coherence: $\mu\left(\mathbf{D}\right) = \max_{i \neq j} \left| (\mathbf{D}^T\mathbf{D})_{i,j} \right|$



Convolutional Sparse Modeling



Classic Sparse Theory for Convolutional Case

<u>Theorem:</u> [Donoho and Elad, 2003]

Basis pursuit is guaranteed to recover the true sparse vector assuming that

$$\|\mathbf{\Gamma}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right)$$

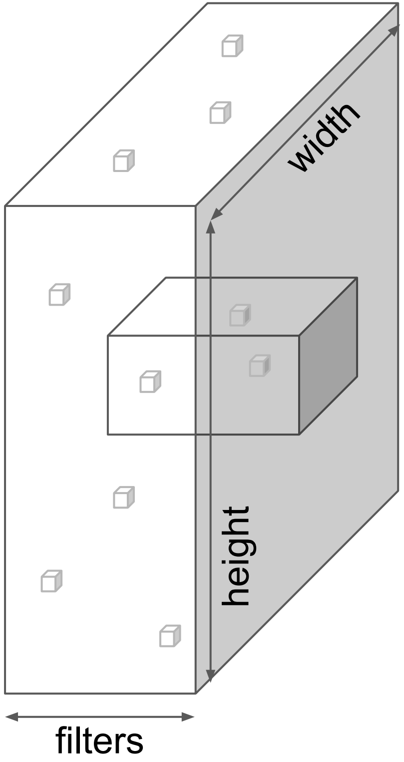
Assuming 2 atoms of length 64 $\quad \mu(\mathbf{D}) \geq 0.063 \quad$ [Welch, 1974]

Success guaranteed when $\quad \|\mathbf{\Gamma}\|_0 < 8.43$

Local Sparsity

$\|\mathbf{\Gamma}\|_{0,\infty}$ maximal number of non-zeroes in a local neighborhood

$$\min_{\mathbf{\Gamma}} \ \|\mathbf{\Gamma}\|_{0,\infty} \ \text{ s.t. } \ \mathbf{X} = \mathbf{D}\mathbf{\Gamma}$$



Success of Basis Pursuit

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

$$\hat{\mathbf{\Gamma}} = \arg\min_{\mathbf{\Gamma}} \ \frac{1}{2}\|\mathbf{Y} - \mathbf{D}\mathbf{\Gamma}\|_2^2 + \lambda\|\mathbf{\Gamma}\|_1$$

Theorem: [Papyan, Sulam and Elad, 2016]

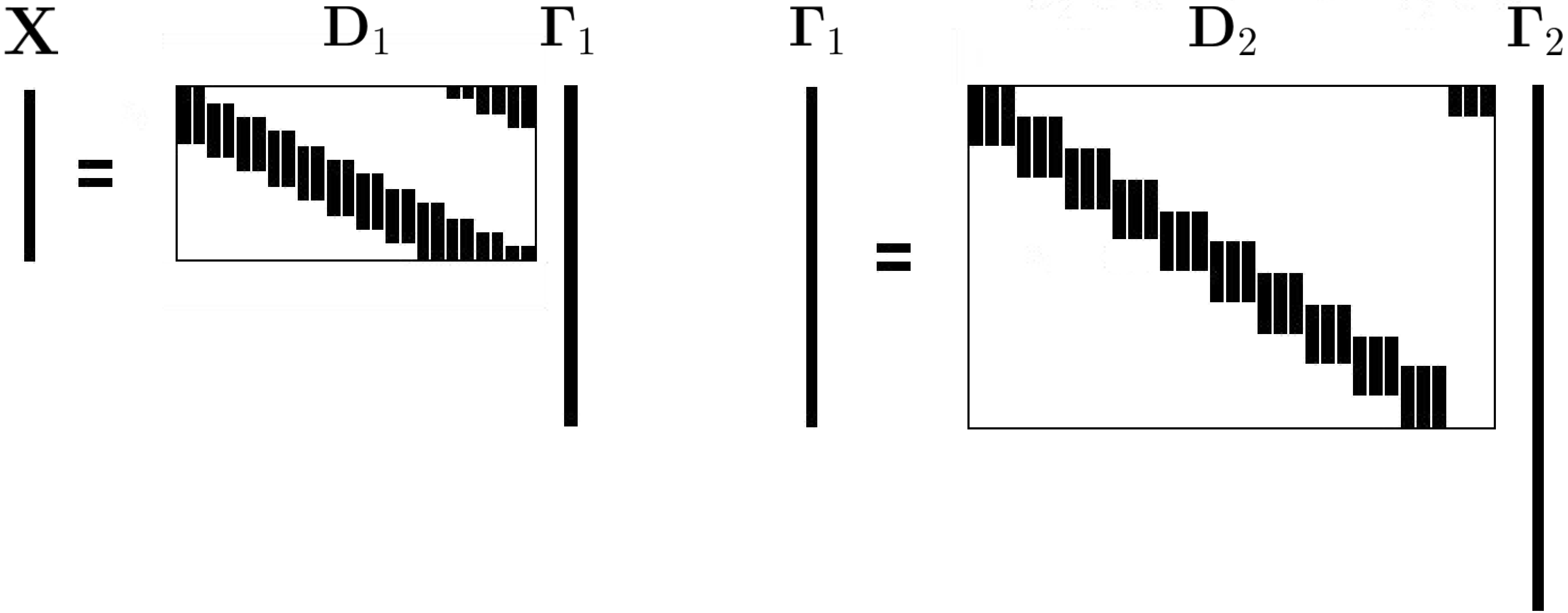
Assume: $\|\mathbf{\Gamma}\|_{0,\infty} < \frac{1}{3}\left(1 + \frac{1}{\mu(\mathbf{D})}\right)$

Then: $\|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_{\infty} \leq 7.5\|\mathbf{E}\|_{2,\infty}$

Theoretical guarantee for:

- [Zeiler et. al 2010]
- [Wohlberg 2013]
- [Bristow et. al 2013]
- [Fowlkes and Kong 2014]
- [Zhou et. al 2014]
- [Kong and Fowlkes 2014]
- [Zhu and Lucey 2015]
- [Heide et. al 2015]
- [Gu et. al 2015]
- [Wohlberg 2016]
- [Šorel and Šroubek 2016]
- [Serrano et. al 2016]
- [Papyan et. al 2017]
- [Garcia-Cardona and Wohlberg 2017]
- [Wohlberg and Rodriguez 2017]
- ...

Multi-layered Convolutional Sparse Modeling



Deep Coding Problem

Given $\mathbf{X}$, find a set of representations satisfying:

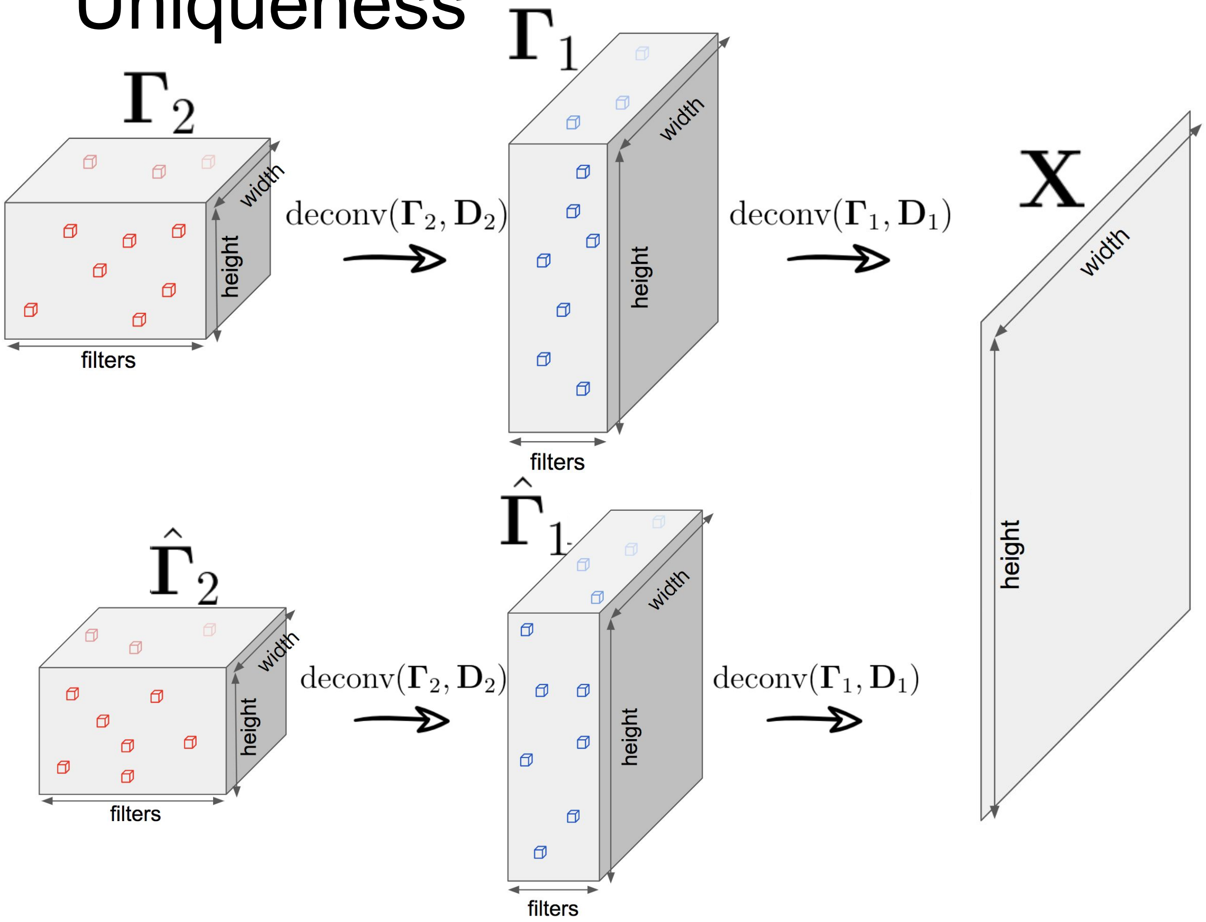
$$
\begin{aligned}
\mathbf{X} &= \mathbf{D}_1 \mathbf{\Gamma}_1, & \|\mathbf{\Gamma}_1\|_{0,\infty} &\leq \lambda_1 \\
\mathbf{\Gamma}_1 &= \mathbf{D}_2 \mathbf{\Gamma}_2, & \|\mathbf{\Gamma}_2\|_{0,\infty} &\leq \lambda_2 \\
&\vdots \\
\mathbf{\Gamma}_{L-1} &= \mathbf{D}_L \mathbf{\Gamma}_L, & \|\mathbf{\Gamma}_L\|_{0,\infty} &\leq \lambda_L
\end{aligned}
$$

Deep Coding Problem

Given $\mathbf{Y}$, find a set of representations satisfying:

$$\begin{aligned}
\|\mathbf{Y} - \mathbf{D}_1\mathbf{\Gamma}_1\|_2 \le \epsilon, &\qquad \|\mathbf{\Gamma}_1\|_{0,\infty} \le \lambda_1 \\
\mathbf{\Gamma}_1 = \mathbf{D}_2\mathbf{\Gamma}_2, &\qquad \|\mathbf{\Gamma}_2\|_{0,\infty} \le \lambda_2 \\
&\vdots \\
\mathbf{\Gamma}_{L-1} = \mathbf{D}_L\mathbf{\Gamma}_L, &\qquad \|\mathbf{\Gamma}_L\|_{0,\infty} \le \lambda_L
\end{aligned}$$

Uniqueness



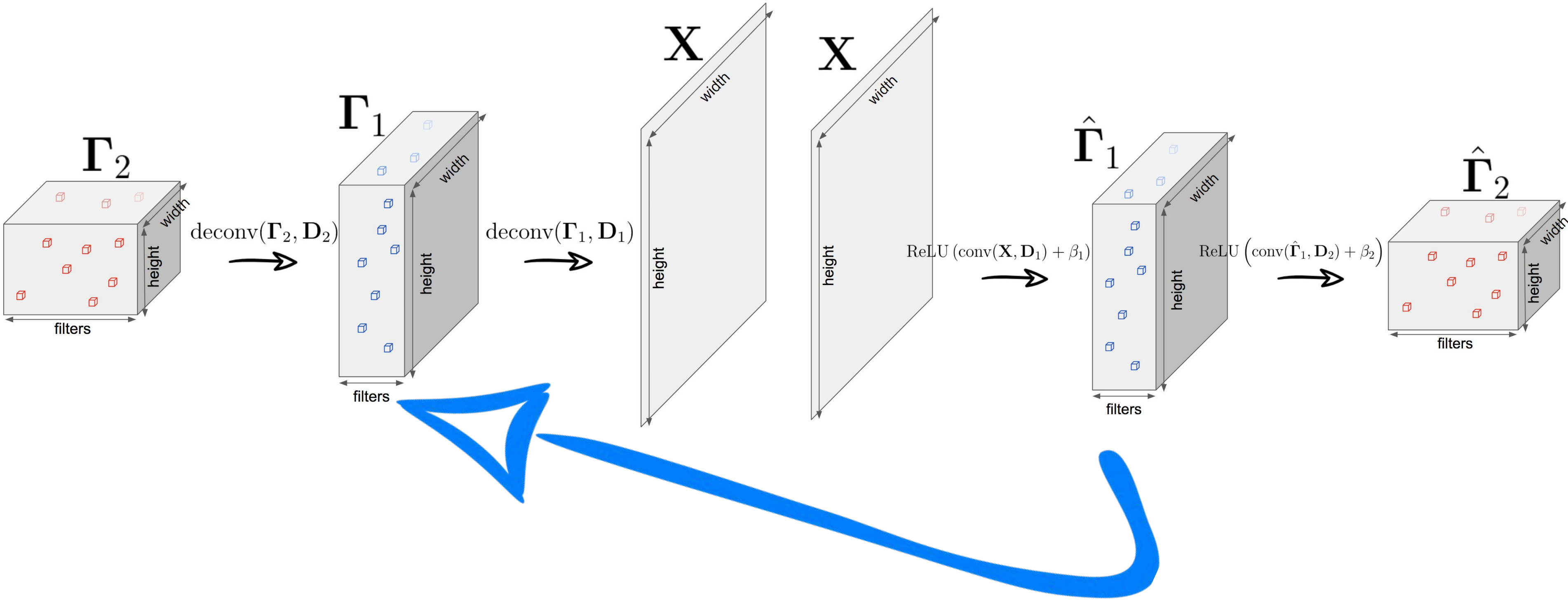
$\mathbf{\Gamma}_2$
width
height
filters
$\text{deconv}(\mathbf{\Gamma}_2, \mathbf{D}_2)$
$\mathbf{\Gamma}_1$
width
height
filters
$\text{deconv}(\mathbf{\Gamma}_1, \mathbf{D}_1)$
$\mathbf{X}$
width
height
$\hat{\mathbf{\Gamma}}_2$
width
height
filters
$\text{deconv}(\mathbf{\Gamma}_2, \mathbf{D}_2)$
$\hat{\mathbf{\Gamma}}_1$
width
height
filters
$\text{deconv}(\mathbf{\Gamma}_1, \mathbf{D}_1)$

Uniqueness Theorem

$$\|\mathbf{\Gamma}_l\|_{0,\infty} \leq \lambda_l < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D}_l)}\right)$$

$$\{\mathbf{\Gamma}_l\}_{l=1}^{L} \text{ are the unique feature maps of } \mathbf{X}$$

Success of Forward Pass



Success of Forward Pass Theorem

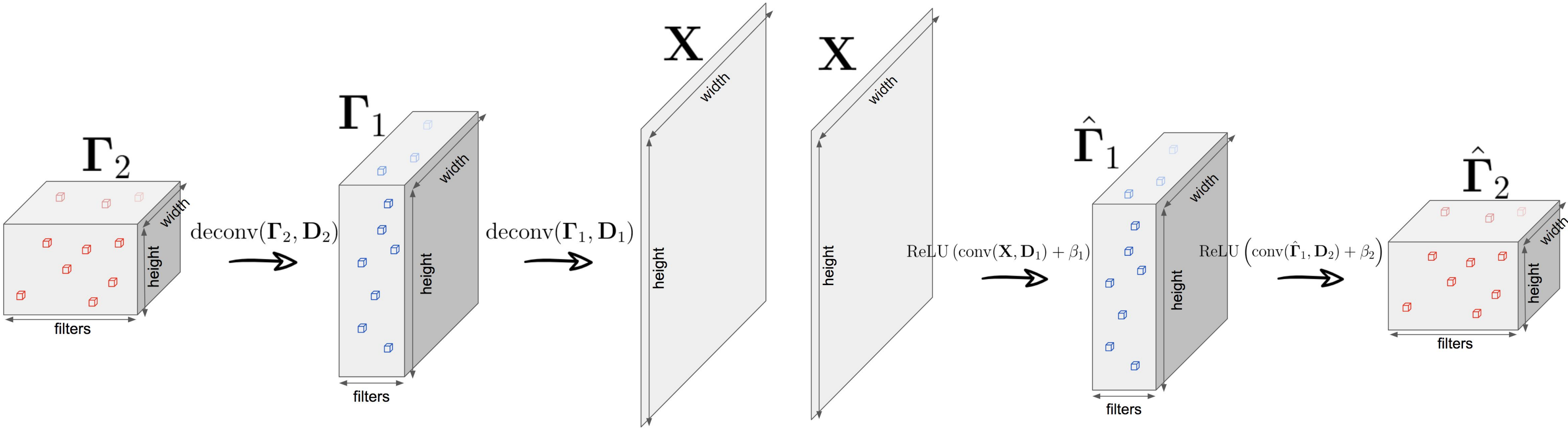
$$\|\mathbf{\Gamma}_l\|_{0,\infty} < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D}_l)}\frac{|\Gamma_l^{\min}|}{|\Gamma_l^{\max}|}\right) - \frac{1}{\mu(\mathbf{D}_l)}\frac{\epsilon_{l-1}}{|\Gamma_l^{\max}|}$$

Layered thresholding guaranteed:

1. Find correct places of nonzeros
2. $\|\hat{\mathbf{\Gamma}}_l - \mathbf{\Gamma}_l\|_{2,\infty} \leq \epsilon_l$

- Forward pass always fails at recovering representations exactly
- Success depends on ratio
- Distance increases with layer

Generative Model and Crude Inference



Layered Lasso

StatsDepartment

$$\hat{\mathbf{\Gamma}}_1 = \arg\min_{\mathbf{\Gamma}_1} \ \frac{1}{2}\|\mathbf{Y} - \mathbf{D}_1\mathbf{\Gamma}_1\|_2^2 + \alpha_1\|\mathbf{\Gamma}_1\|_1$$

$$\hat{\mathbf{\Gamma}}_2 = \arg\min_{\mathbf{\Gamma}_2} \ \frac{1}{2}\|\hat{\mathbf{\Gamma}}_1 - \mathbf{D}_2\mathbf{\Gamma}_2\|_2^2 + \alpha_2\|\mathbf{\Gamma}_2\|_1$$

Success of Layered Lasso

$$\|\mathbf{\Gamma}_l\|_{0,\infty} < \frac{1}{3}\left(1 + \frac{1}{\mu(\mathbf{D}_L)}\right)$$

Layered Lasso guaranteed:

1. Find only correct places of nonzeros
2. Find all coefficients that are big enough
3. $\|\hat{\mathbf{\Gamma}}_l - \mathbf{\Gamma}_l\|_{2,\infty} \leq \epsilon_l$

~~Forward pass always fails at recovering representations exactly~~

~~Success depends on ratio~~

Distance increases with layer

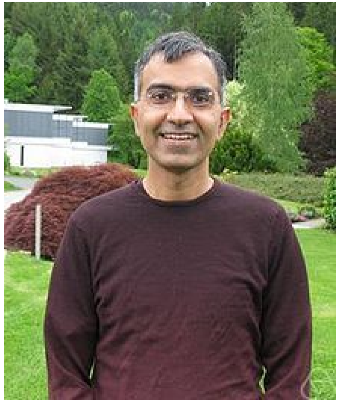
Layered Iterative Thresholding

$$\mathbf{\Gamma}_1^t = \mathcal{S}_{\alpha_1}\left(\mathbf{D}_1^T\mathbf{Y} + \left(\mathbf{I} - \mathbf{D}_1^T\mathbf{D}_1\right)\mathbf{\Gamma}_1^{t-1}\right)$$

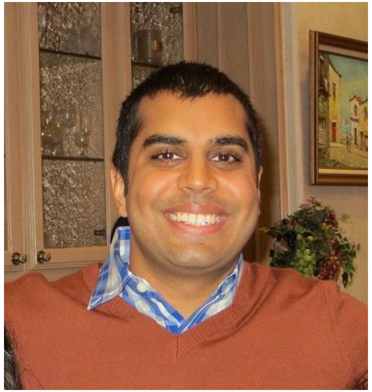
$$\mathbf{\Gamma}_2^t = \mathcal{S}_{\alpha_2}\left(\mathbf{D}_2^T\hat{\mathbf{\Gamma}}_1 + \left(\mathbf{I} - \mathbf{D}_2^T\mathbf{D}_2\right)\mathbf{\Gamma}_2^{t-1}\right)$$

Supervised Deep Sparse Coding Networks [Sun et. al 2017]

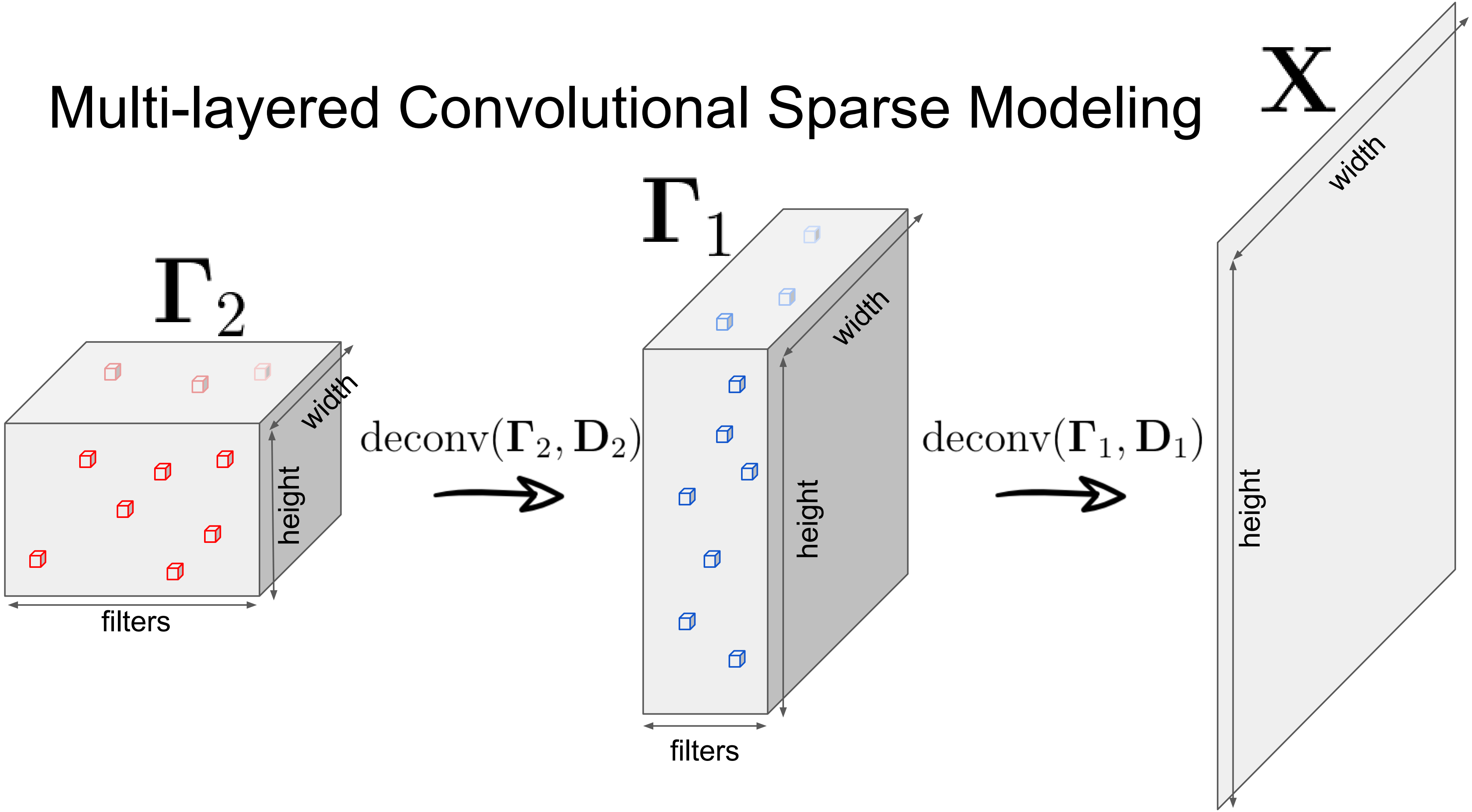
Method	# Params	# Layers	CIFAR-10	CIFAR-100
SCKN [34]	10.50M	10	10.20	-
OMP [18]	0.70M	2	18.50	-
PCANet [36]	0.28B	3	21.33	-
NOMP [7]	1.09B	4	18.60	39.92
NiN [32]	-	-	8.81	35.68
DSN [33]	1.34M	7	7.97	36.54
WRN [12]	36.5M	28	4.00	19.25
ResNet-110 [10]	0.85M	110	6.41	27.22
ResNet-1001 v2 [11]	10.2M	1001	4.92	27.21
ResNext-29 [14]	**68.10M**	**29**	**3.58**	**17.31**
SwapOut-20 [13]	1.10M	20	5.68	25.86
SwapOut-32 [13]	7.43M	32	4.76	22.72
SCN-1	0.17M	15	8.86	25.08
SCN-2	0.35M	15	7.18	22.17
SCN-4	**0.69M**	**15**	**5.81**	**19.93**



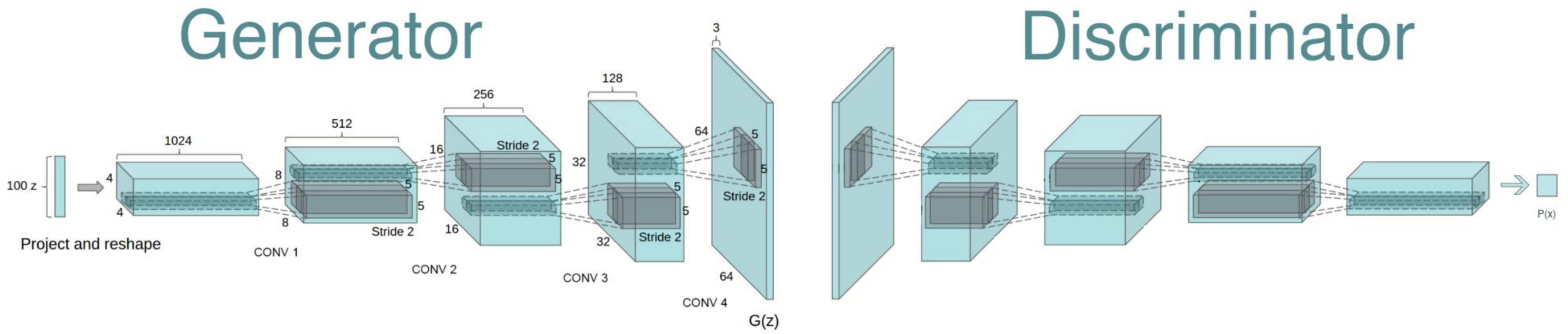
Relation to Other Generative Models



Multi-layered Convolutional Sparse Modeling



Generator in GANs [Goodfellow et. al 2014]



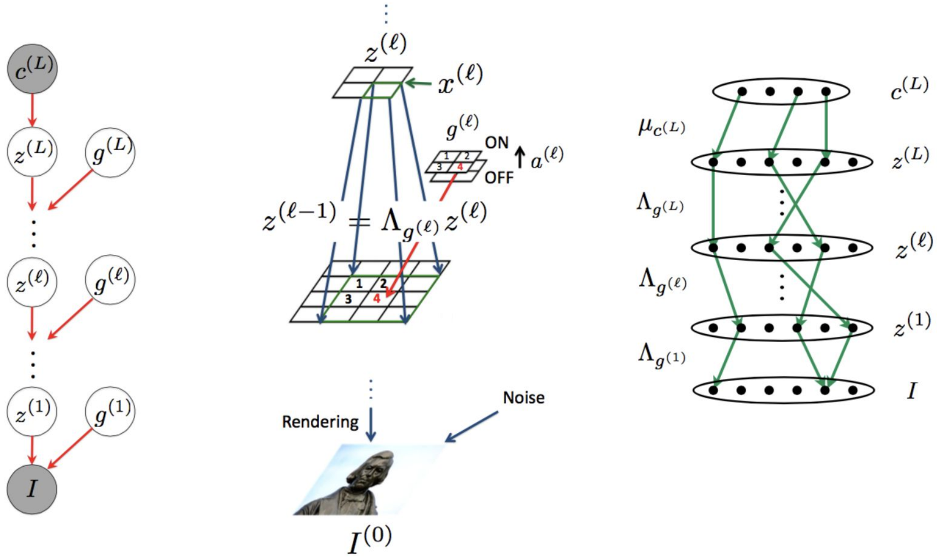
Sparsification of intermediate feature maps with **ReLU**

DRMM [Patel et. al]

$$\mu_{cg} \equiv \Lambda_g \mu_{c^{(L)}} \equiv \boxed{\Lambda^{(1)}_{g^{(1)}} \Lambda^{(2)}_{g^{(2)}} \dots \Lambda^{(L-1)}_{g^{(L-1)}} \Lambda^{(L)}_{g^{(L)}} \mu_{c^{(L)}}}$$

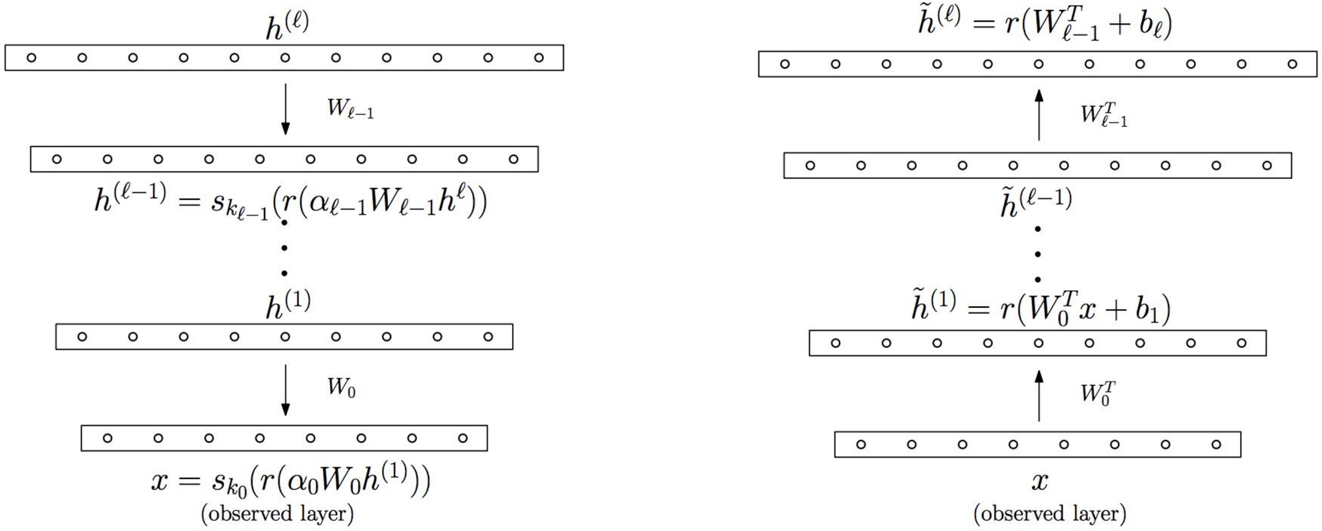
$$I \sim \mathcal{N}(\mu_{cg}, \sigma^2 \mathbf{1}_{D^{(0)}}),$$

$$\Lambda_{g^\ell} \equiv \boxed{\Gamma^{(\ell)}}\boxed{M_{a^{(\ell)}}}\mathcal{T}_{t^{(\ell)}}$$



Sparsification of intermediate feature maps with a **random mask**

[Arora et. al, 2015]



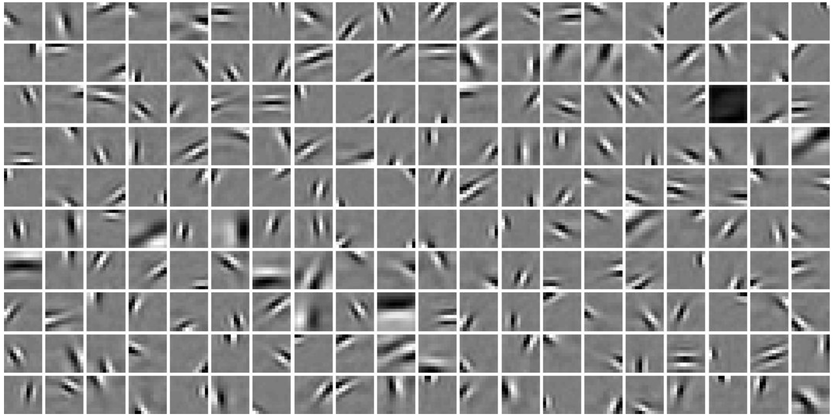
(a) Generative model (b) Feedforward NN

Sparsification of intermediate feature maps with a **random mask** and **ReLU**

Evidence

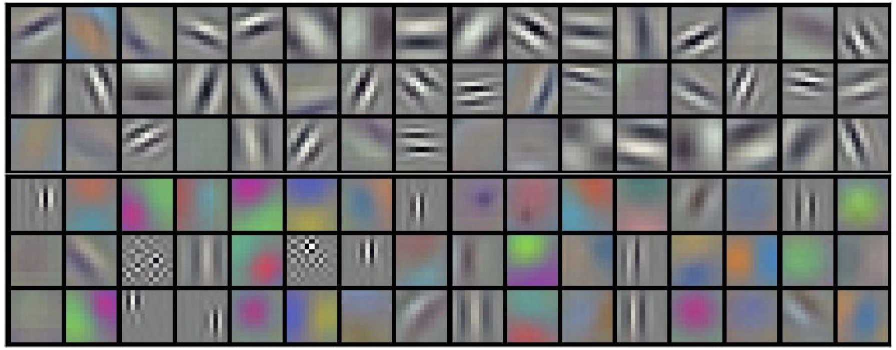
Olshausen & Field and AlexNet

Olshausen & Field



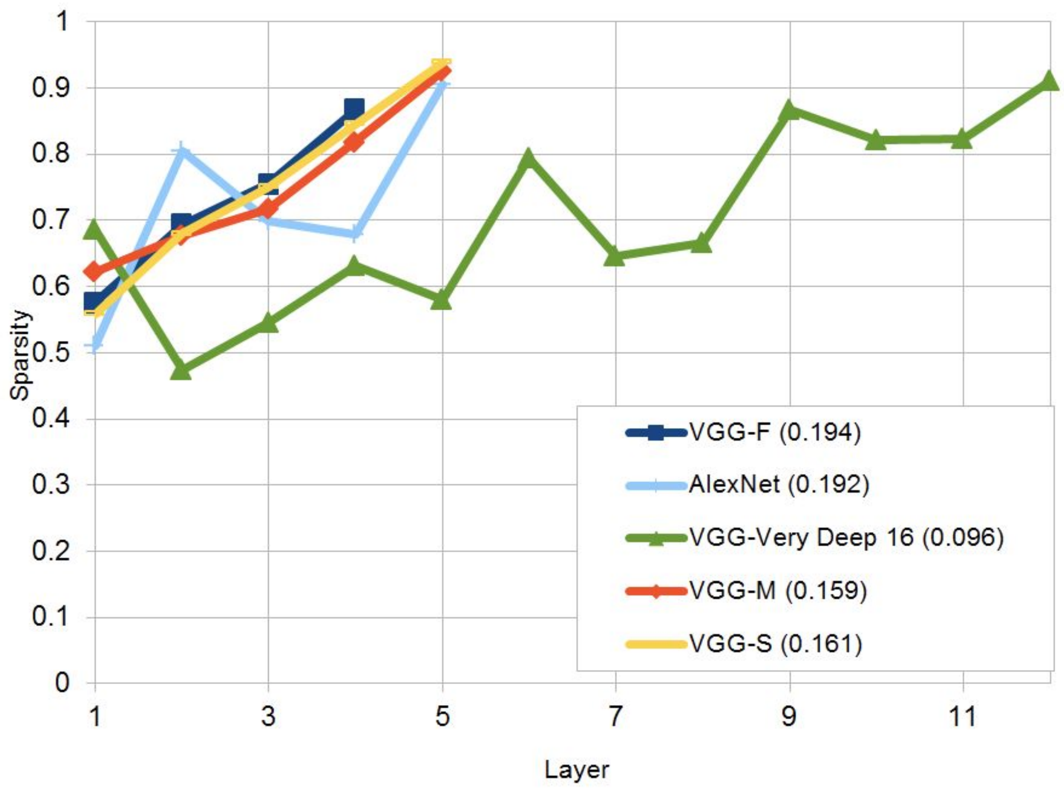
explicit sparsity

AlexNet

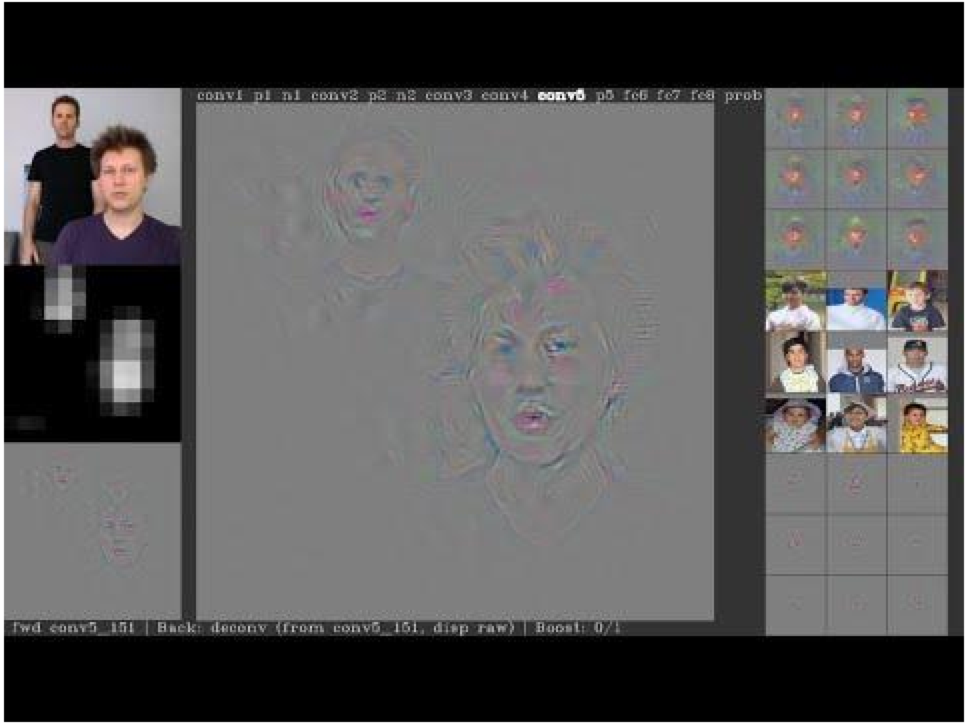


implicit sparsity

Sparsity in Practice



Sparsity in Practice



Mutual Coherence in Practice

[Shang 2015] measured the average mutual coherences of the different layers in the “all-conv” network:

Table 1: μ_{ij} for ImageNet All-Conv Model with relu

Layer Index	1	2	3	4	5	6	7	8	9
average $\mu_{ij,i\neq j}$	0.240	0.194	0.068	0.082	0.091	0.073	0.087	0.113	0.075
std	0.200	0.183	0.090	0.080	0.089	0.068	0.078	0.098	0.065

Regularizing Coherence

[Cisse et. al 2017] proposed the following regularization to improve the robustness of a network to adversarial examples:

$$\mathcal{R}(\mathbf{D}_l) = \|\mathbf{D}_l^T \mathbf{D}_l - \mathbf{I}\|_2^2$$

Local Sparsity

Do Deep Neural Networks Suffer from Crowding?

Anna Volokitin[1,3], **Gemma Roig**[1,2] **and Tomaso Poggio**[1,3]
1: Center for Brains, Minds, and Machines, Massachusetts Institute of Technology, Cambridge, MA, USA
2: Istituto Italiano di Tecnologia at Massachusetts Institute of Technology, Cambridge, MA
3: Computer Vision Laboratory, ETH Zurich, Switzerland

Abstract

Crowding is a visual effect suffered by humans, in which an object that can be recognized in isolation can no longer be recognized when other objects, called flankers, are placed close to it. In this work, we study the effect of crowding in artificial Deep Neural Networks for object recognition. We analyze both standard deep convolutional neural networks (DCNNs) as well as a new version of DCNNs which is 1) multi-scale and 2) with size of the convolution filters change depending on the eccentricity wrt to the center of fixation. Such networks, that we call eccentricity-dependent, are a computational model of the feedforward path of the primate visual cortex. Our results reveal that the eccentricity-dependent model, trained on target objects in isolation, can recognize such targets in the presence of flankers, if the targets are near the center of the image, whereas DCNNs cannot. Also, for all tested networks, when trained on targets in isolation, we find that recognition accuracy of the networks decreases the closer the flankers are to the target and the more flankers there are. We find that visual similarity between the target and flankers also plays a role and that pooling in early layers of the network leads to more crowding. Additionally, we show that incorporating the flankers into the images of the training set does not improve performance with crowding.

Summary

1 

Sparsity well established theoretically

2 

Sparsity is covertly exploited in practice: ReLU, dropout, stride, dilation, ...

3 

Sparsity is the secret sauce behind CNN

4 

Need to bring sparsity to the surface to better understand CNNs

5 

Andrej Karpathy agrees


PREACH!