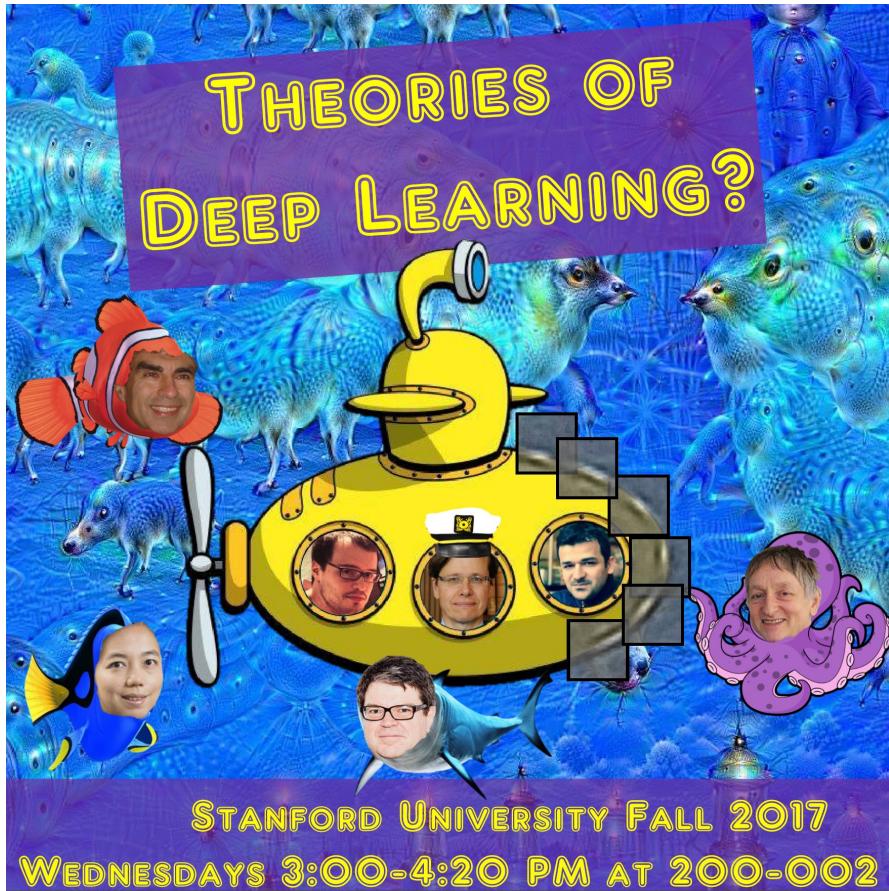


Sparsity in Convolutional Neural Networks

Speaker: Qingyun Sun
Math PhD @ Stanford

Acknowledgement:
Stats 385 @
Stanford
<https://stats385.github.io/>



The talk is based on:
Convolutional Neural Networks in View of Sparse Coding,
Vardan Petyan @ Stats 385, Stanford



Based on work of:
Vardan Petyan, Jeremias Sulam, Yaniv Romano,
Michael Elad



Sparsity: Central idea in Stats

Compressive Sensing:

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

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

Sparsity: Central idea in Stats

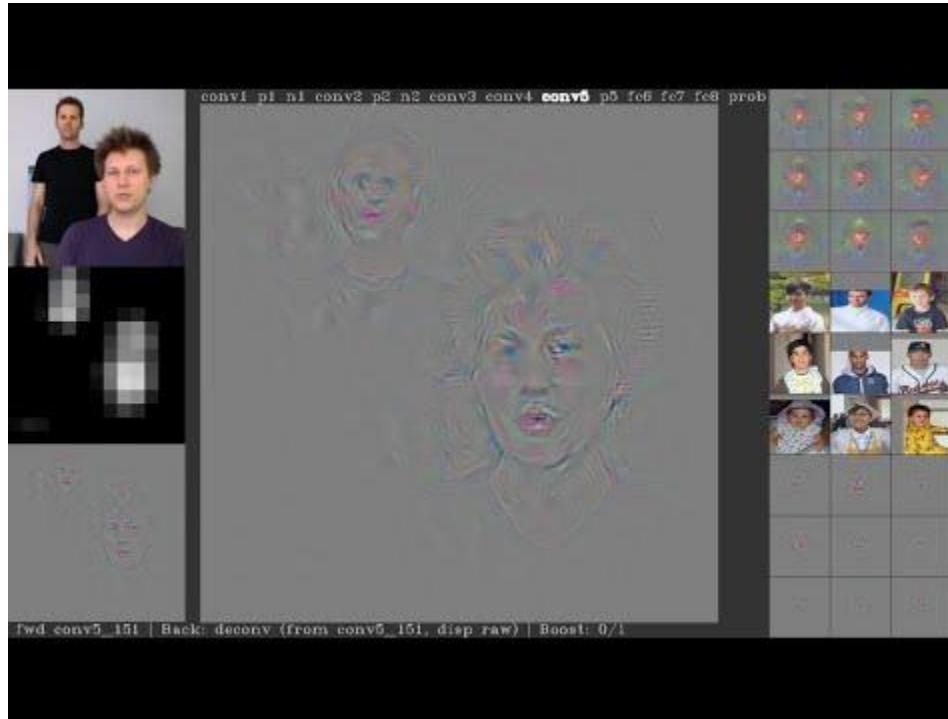
Lasso:

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

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

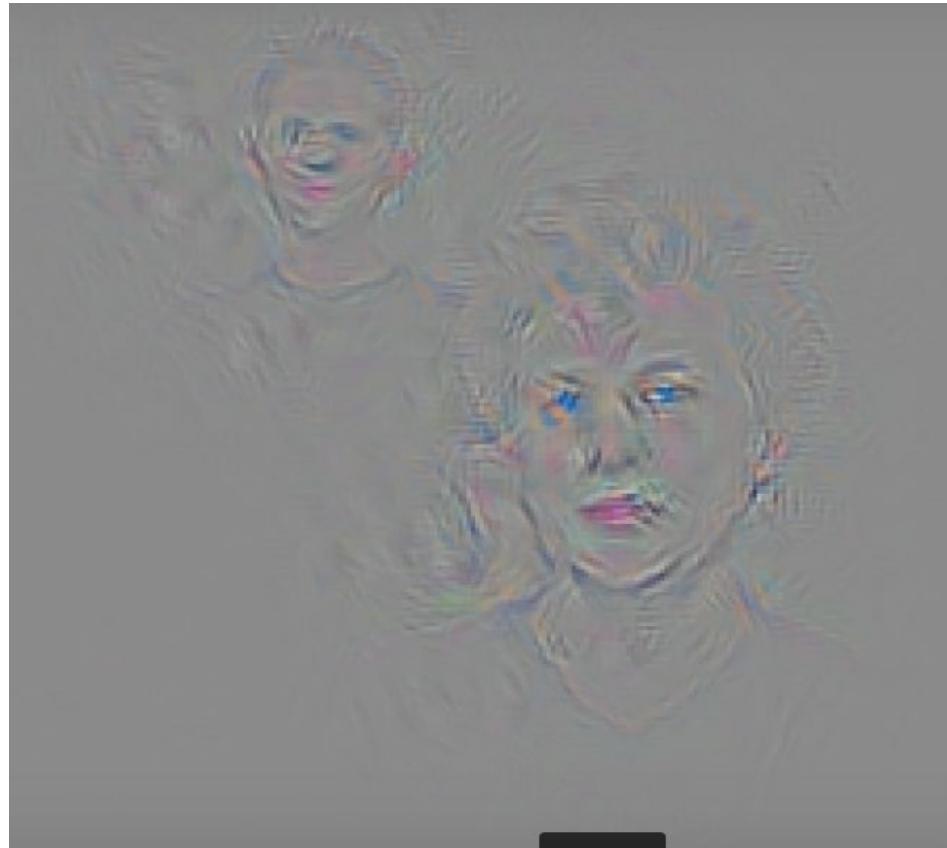
After Deep Revolution, is
sparsity still important?

Sparsity observed in CNN



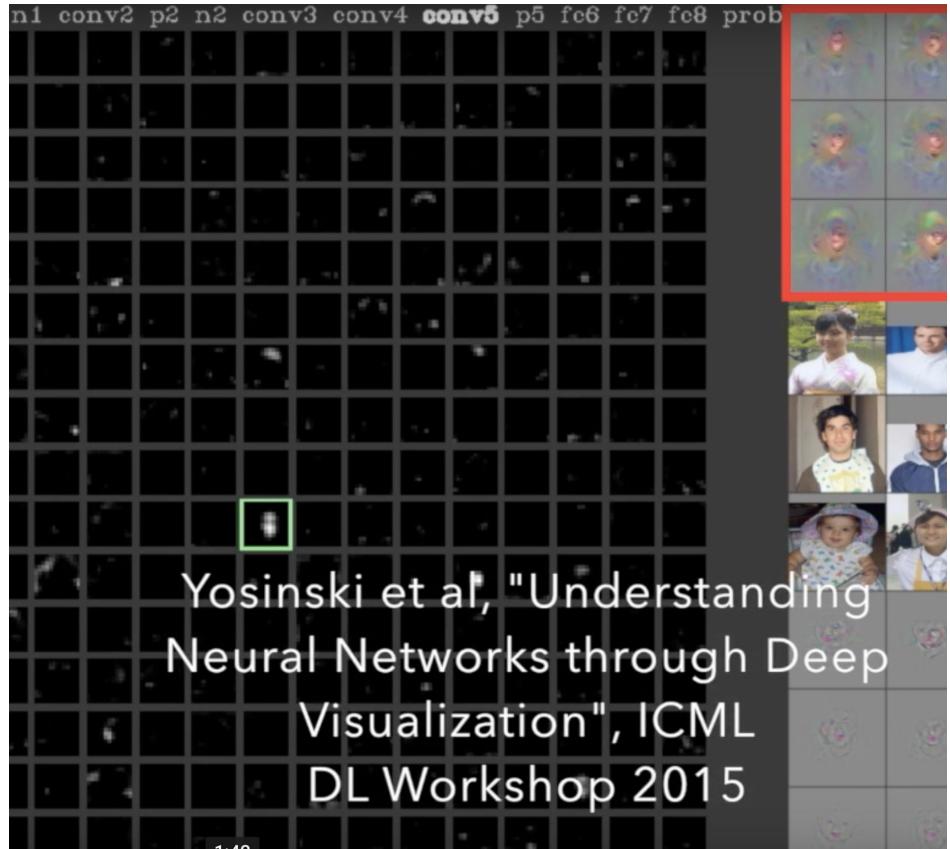
Sparsity in Practice

The activation of RELU layer is sparse.

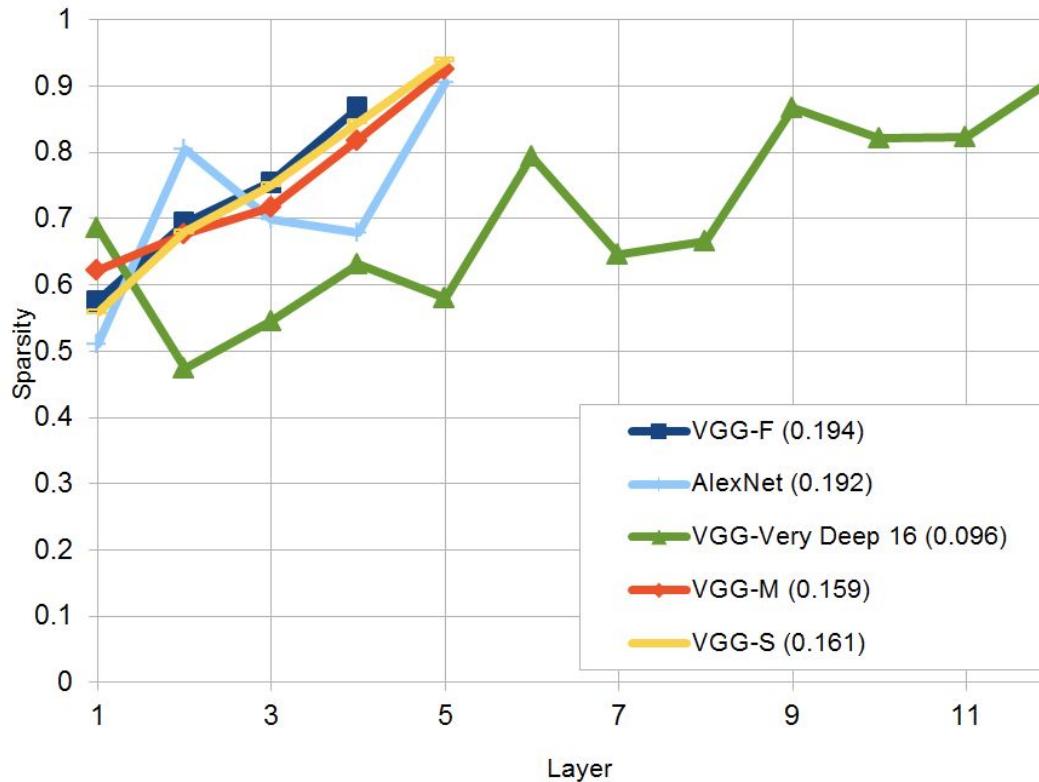


Sparsity in Practice

The activation of RELU layer is sparse.

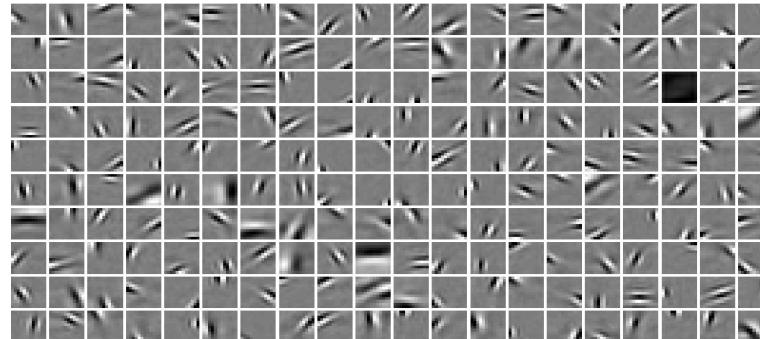


Sparsity in Practice



Olshausen & Field and AlexNet

Olshausen & Field



explicit sparsity

AlexNet

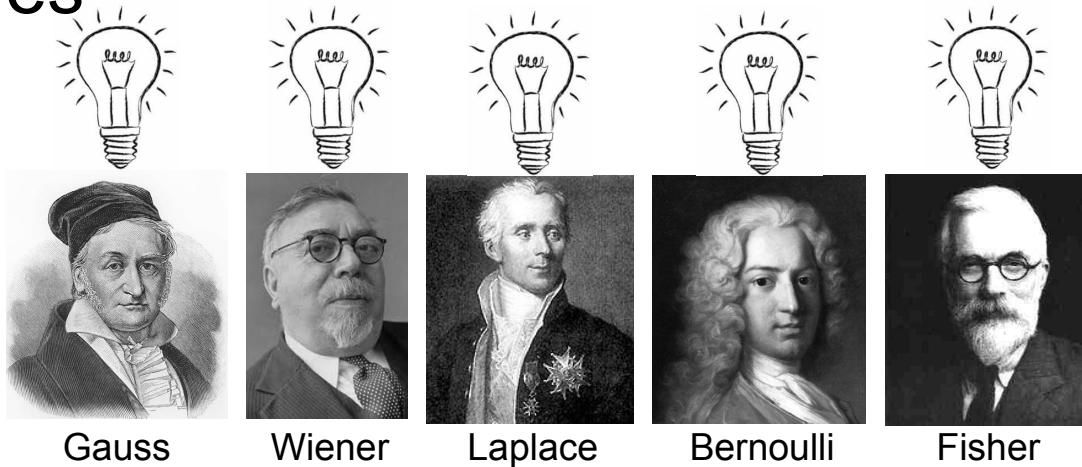


implicit sparsity

Theory of sparsity in CNN?

Breiman's “Two Cultures”

Generative modeling



Gauss

Wiener

Laplace

Bernoulli

Fisher

Predictive modeling



Generative modeling

Seeks to develop stochastic models which fit the data, and then make inferences about the data-generating mechanism based on the structure of those models.

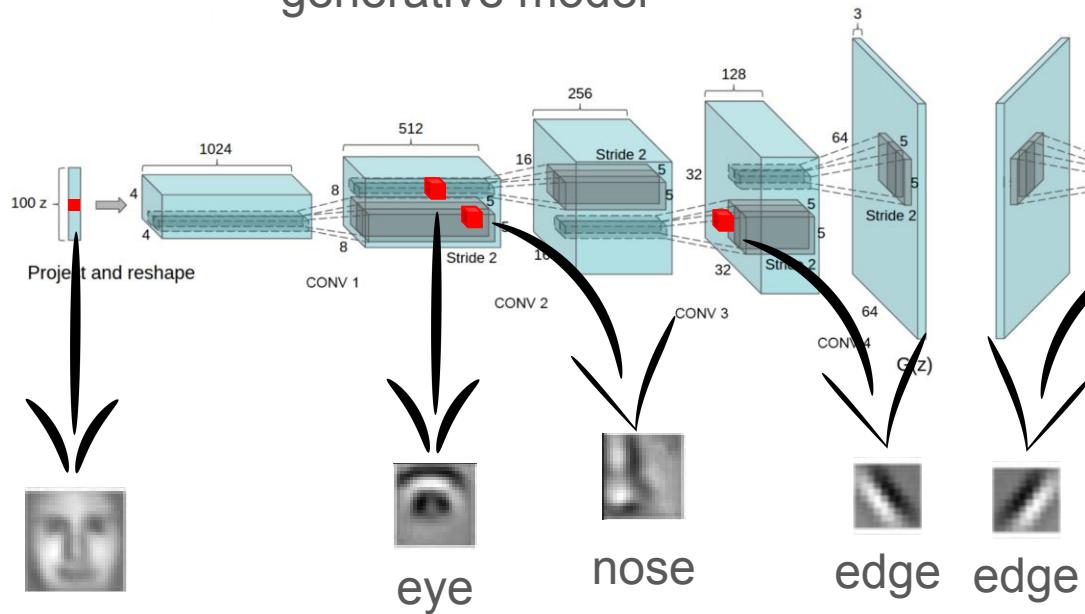
Predictive modeling

Predictive modeling is effectively silent about the underlying mechanism generating the data, and allows for many different predictive algorithms, preferring to discuss only accuracy of prediction made by different algorithm on various datasets.

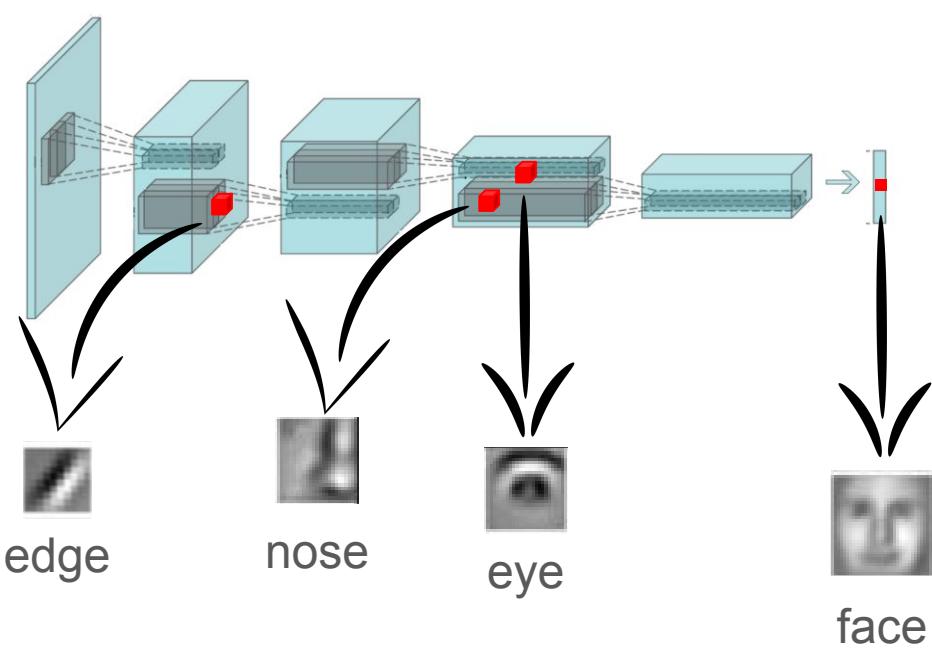
Generative Modeling



generative model



forward pass of CNN

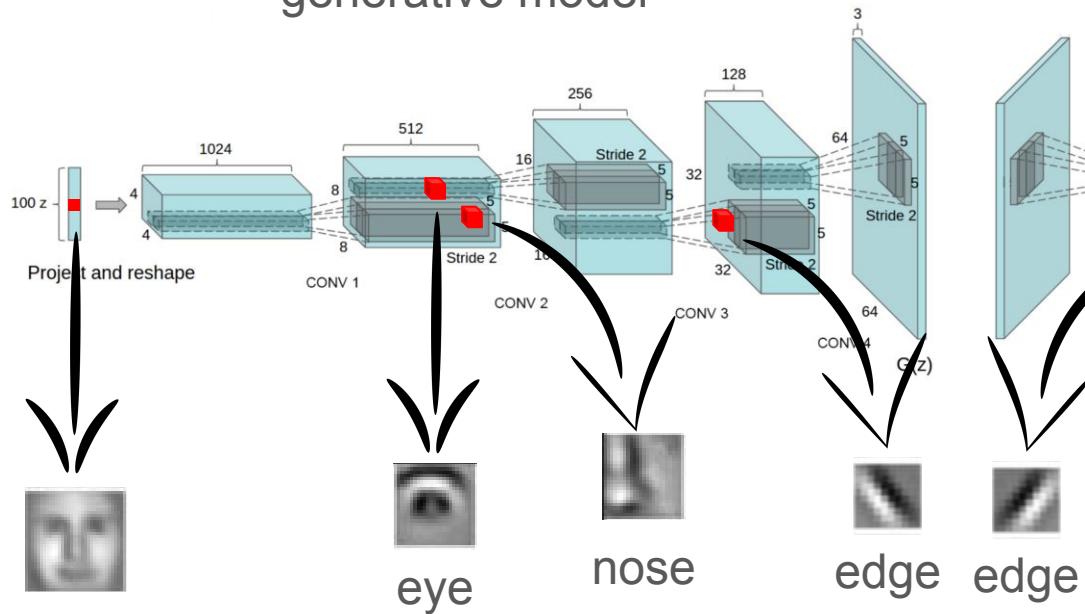




What generative model?

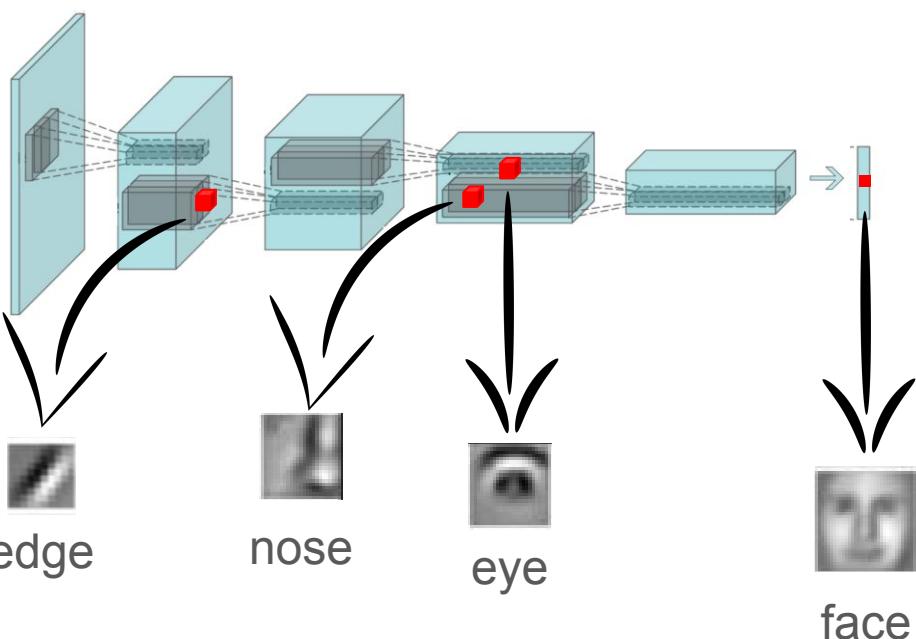


generative model



face

forward pass of CNN



face

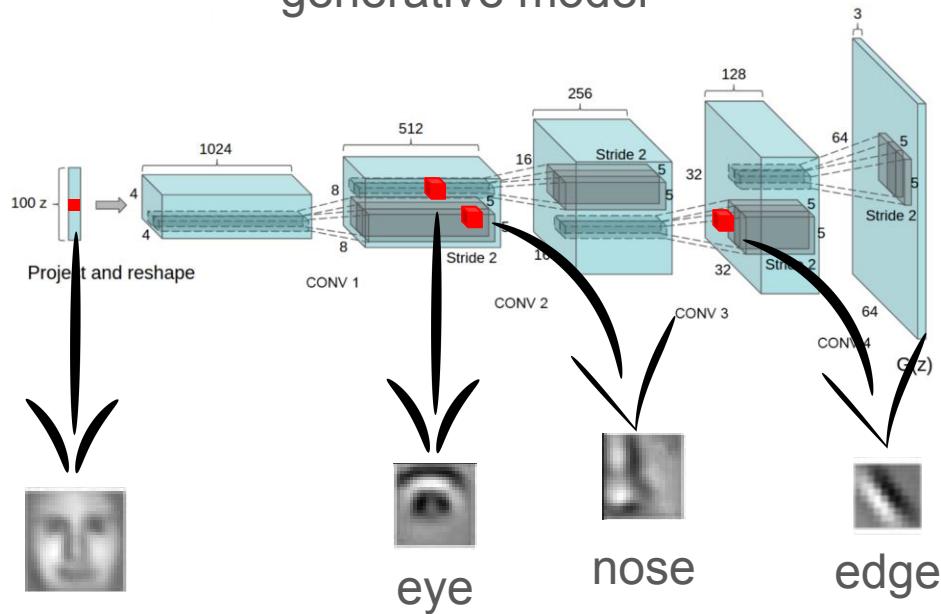


Properties of model?

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

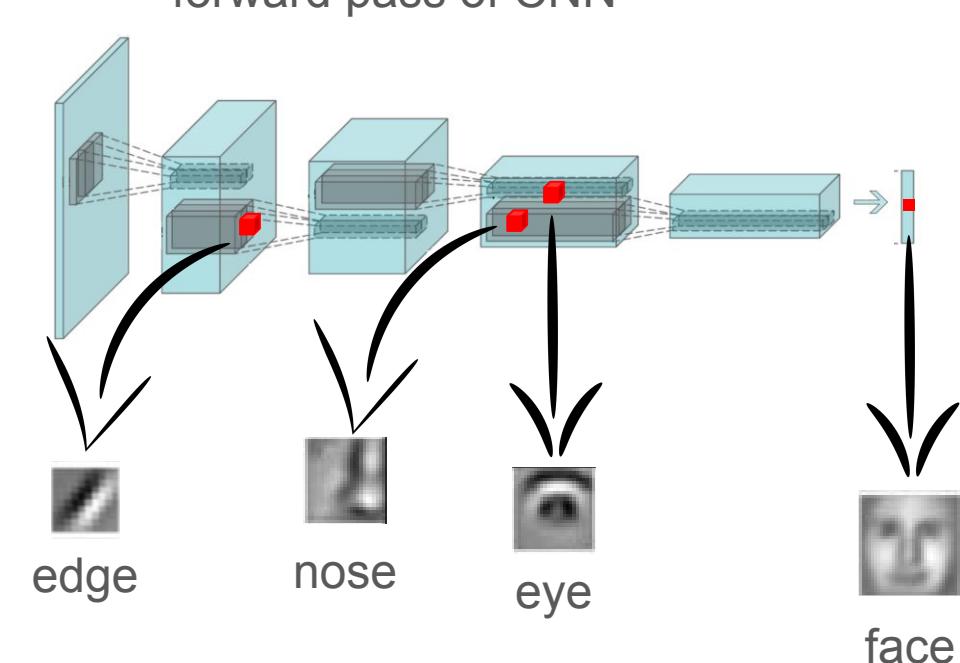


generative model



face

forward pass of CNN



face

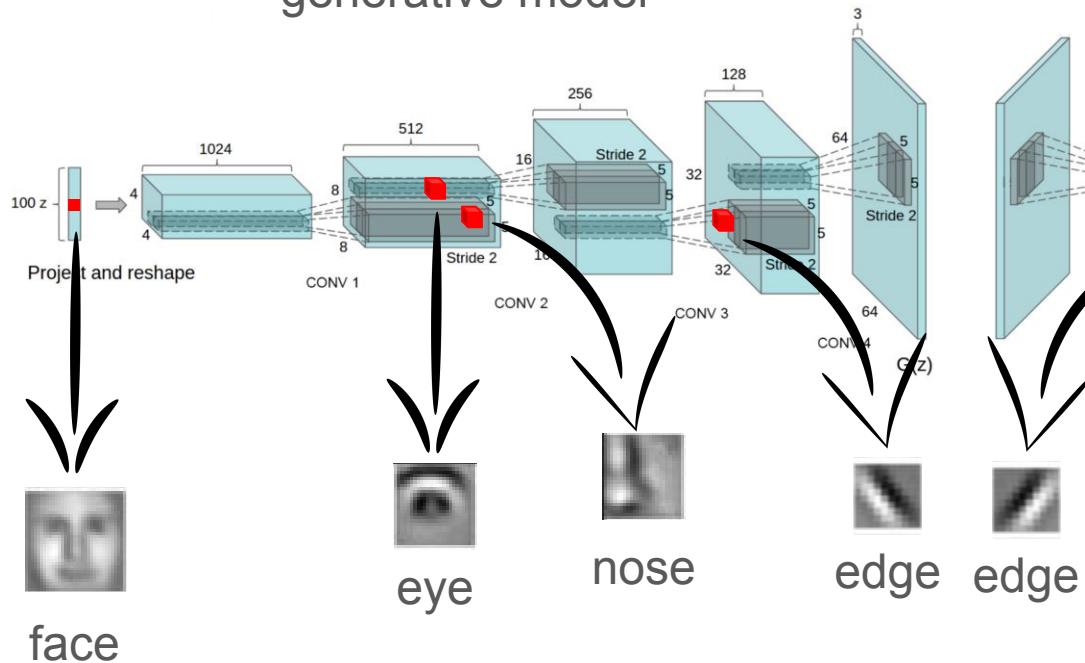


Properties of model?

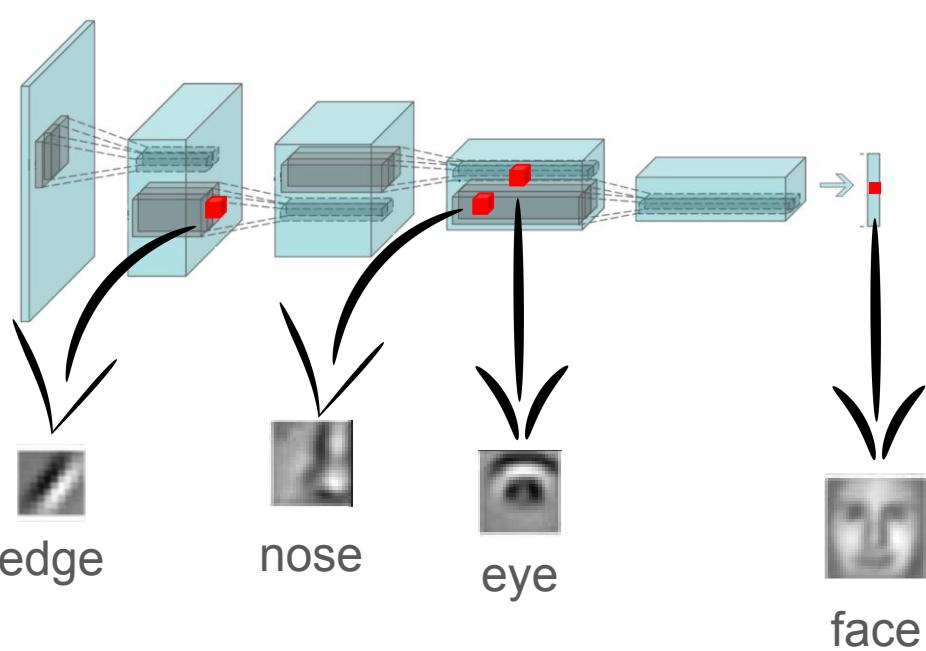
Weights and
pre-activations are
i.i.d Gaussian



generative model



forward pass of CNN



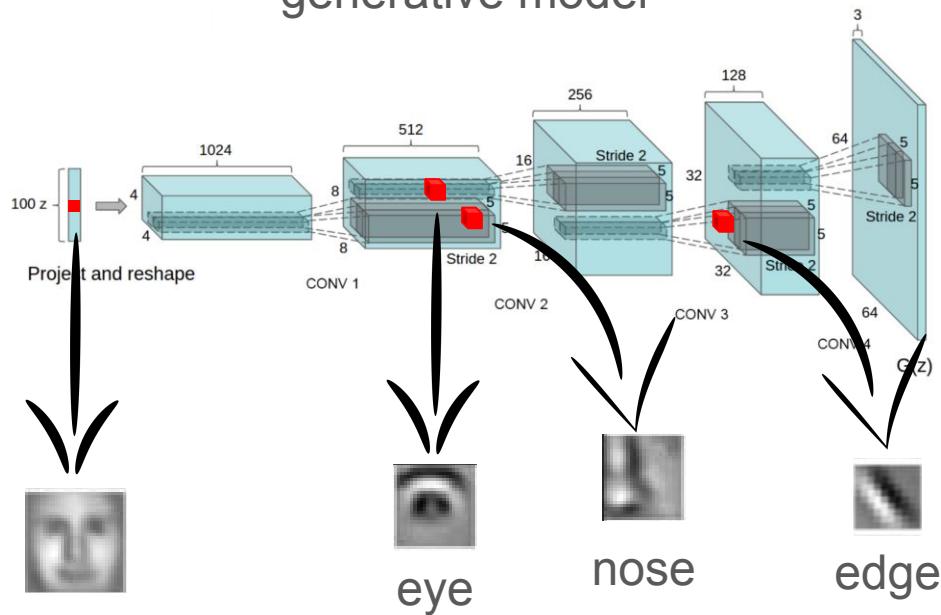


Properties of model?

Overparameterization is good for optimization



generative model



face

eye

nose

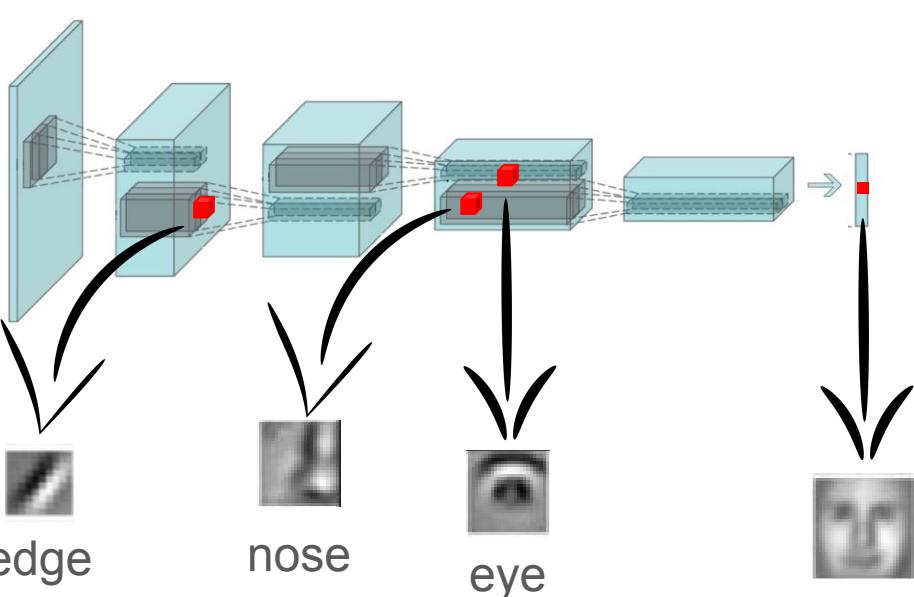
edge edge

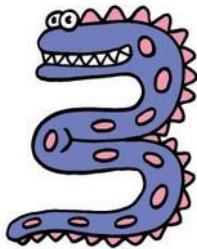
nose

eye

face

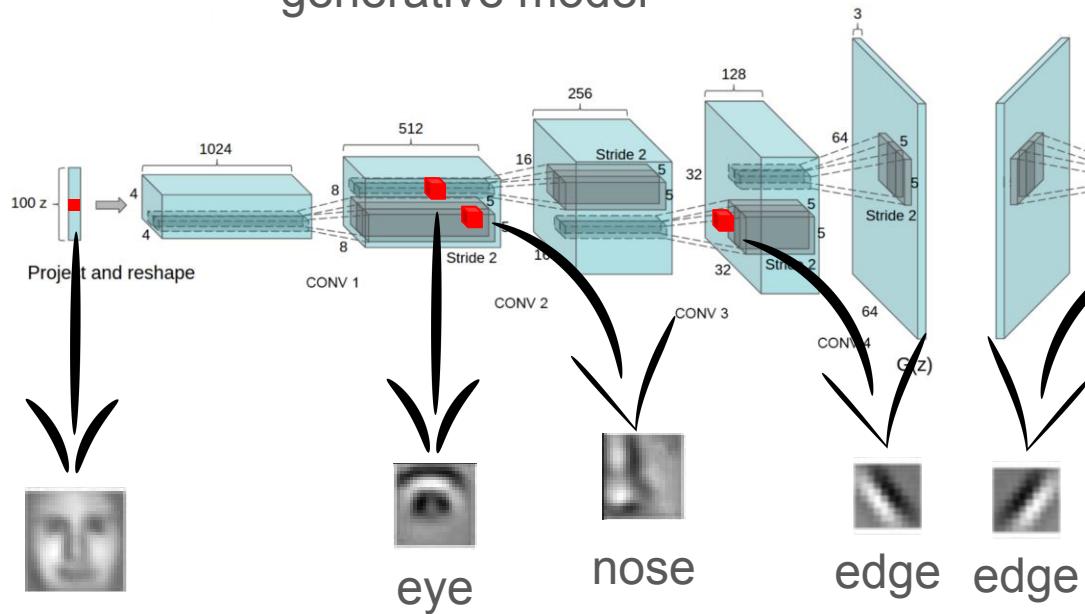
forward pass of CNN



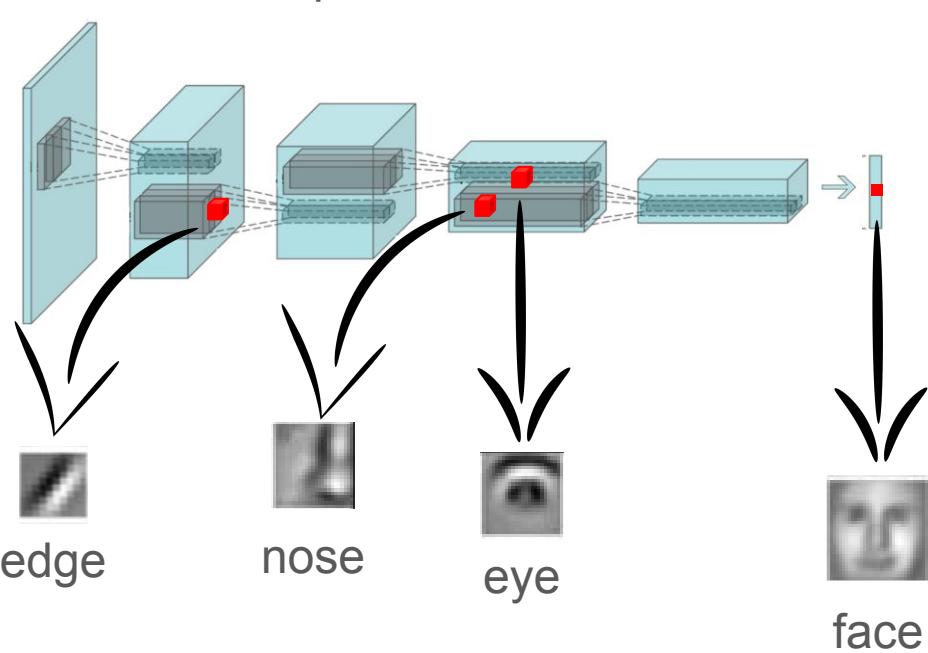


Success of inference?

generative model



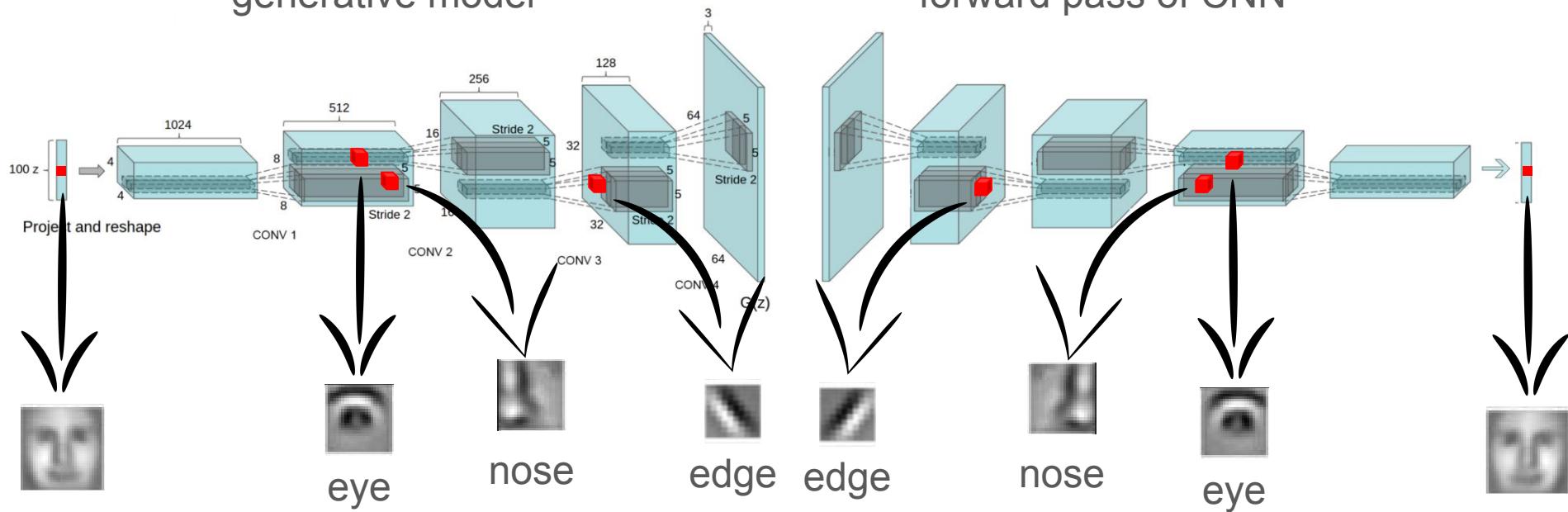
forward pass of CNN





Uniqueness of representation?

generative model



face

eye

nose

edge

edge

nose

eye

face

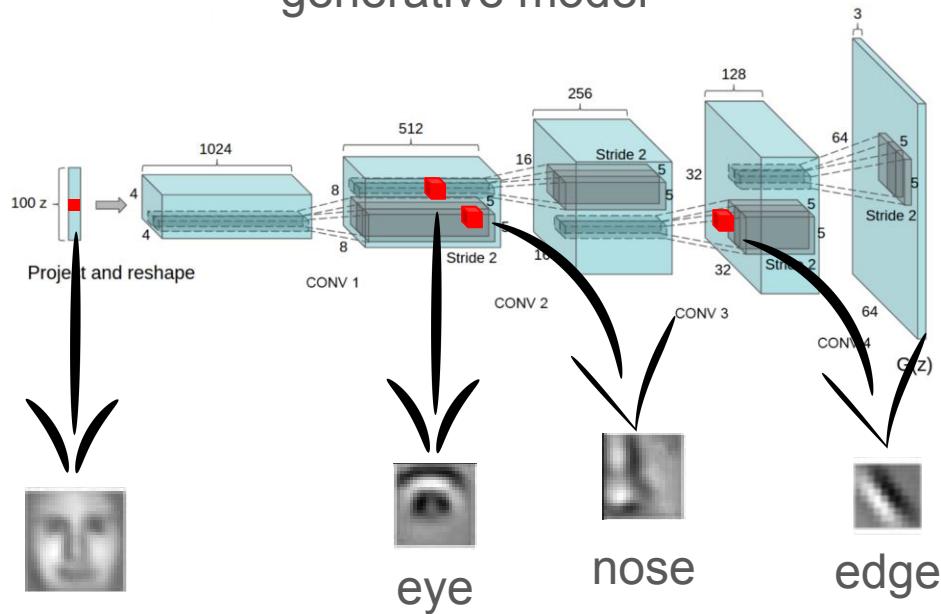


$$|||\Phi(F_\tau f) - \Phi(f)||| \leq C_c \|\tau\|_\infty^\alpha$$

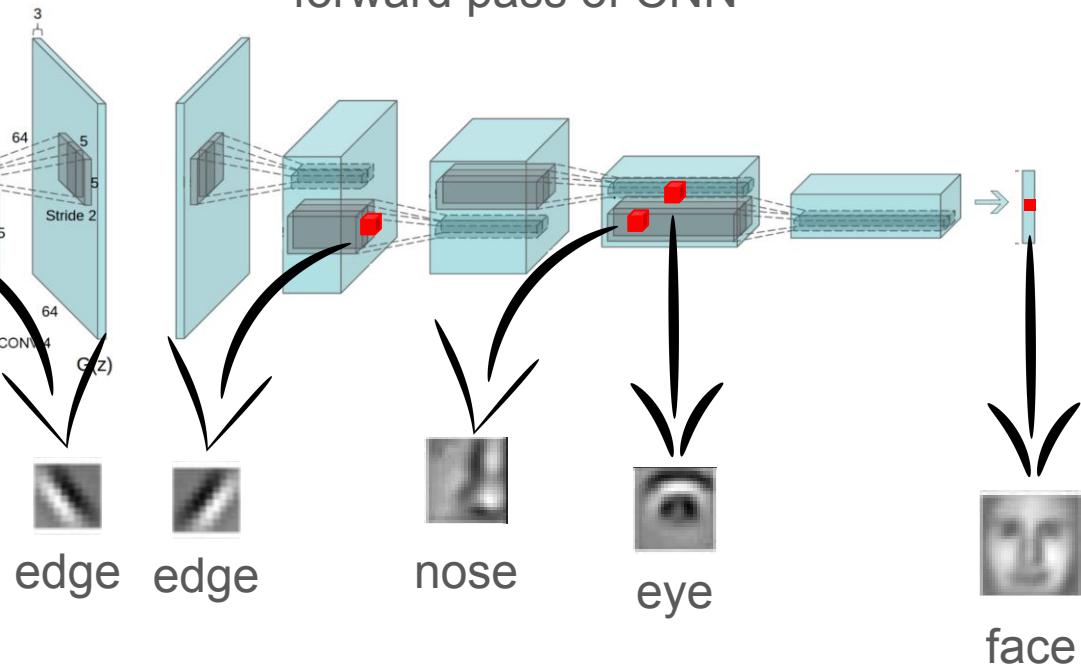


Stability to perturbations?

generative model



forward pass of CNN



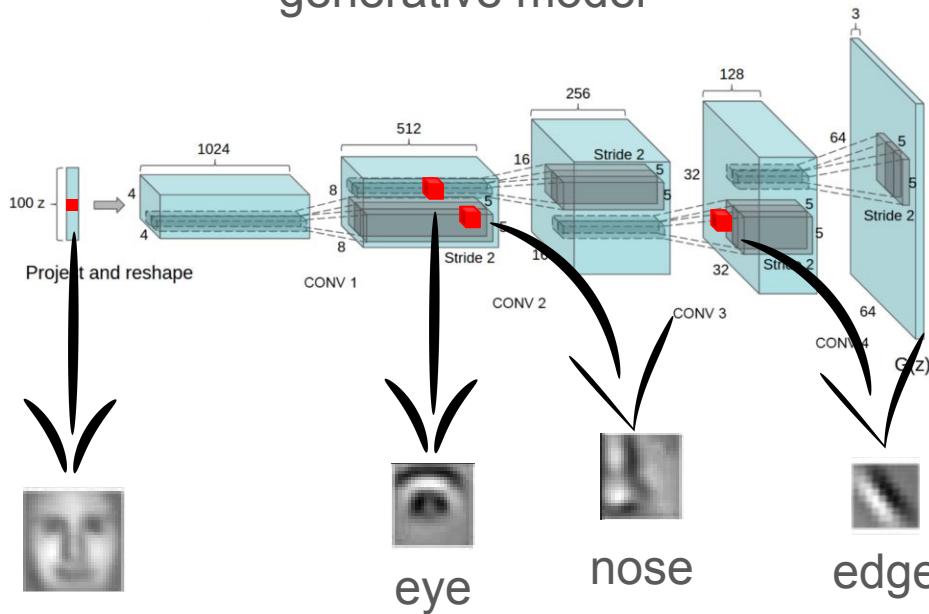


Better inference?

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



generative model



face

eye

nose

edge

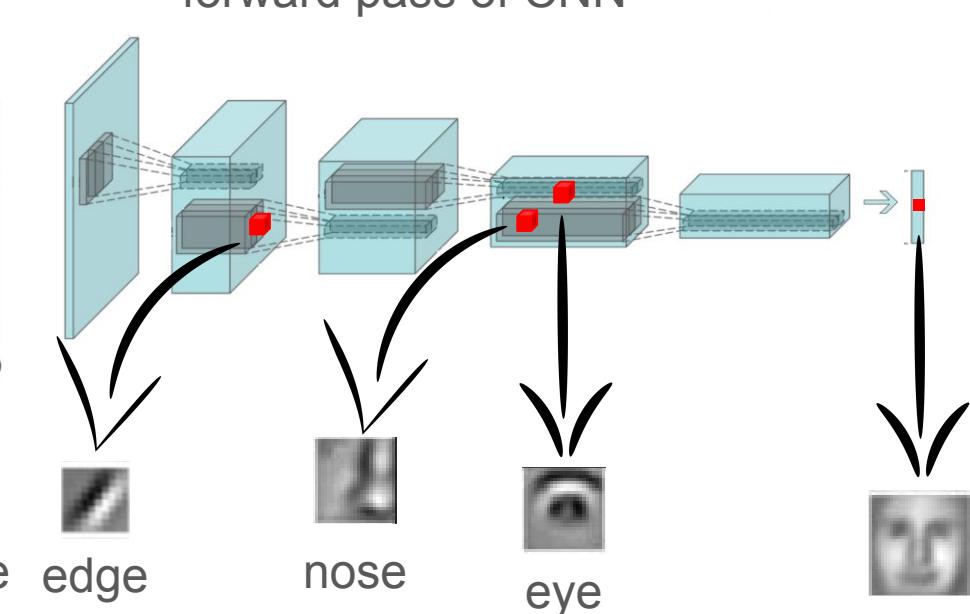
edge

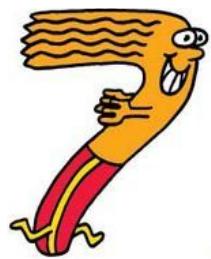
nose

eye

face

forward pass of CNN





Better training?

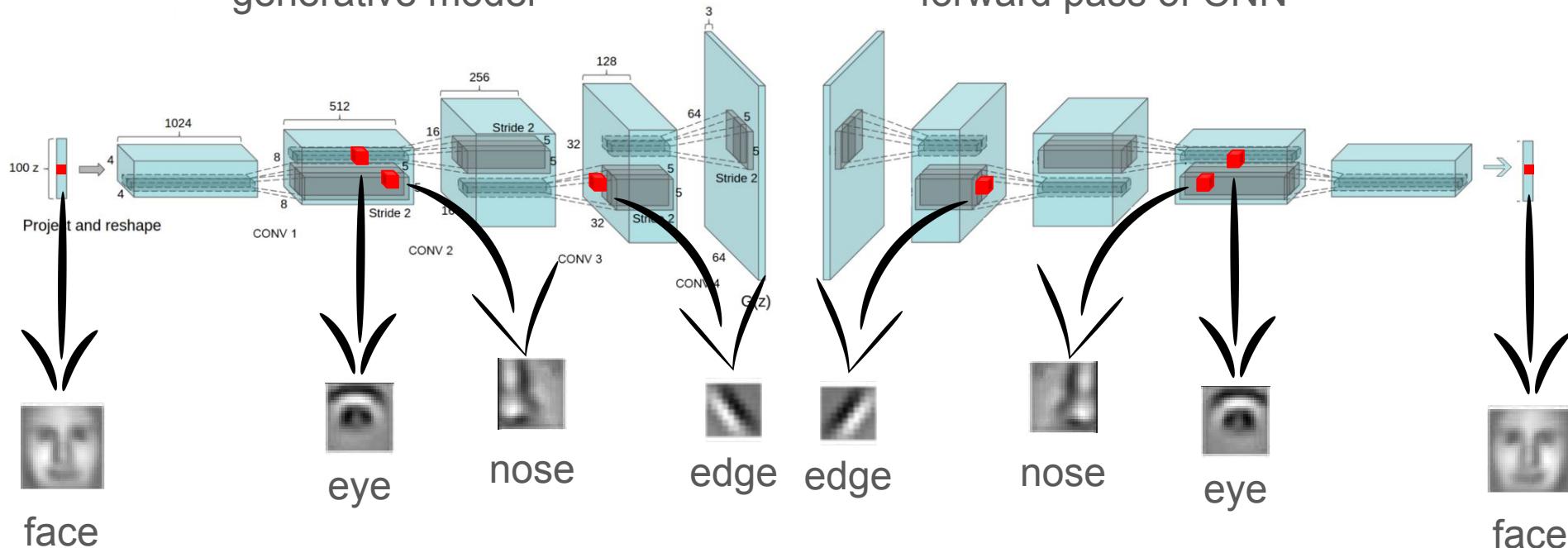


random features
k-means
matrix factorization



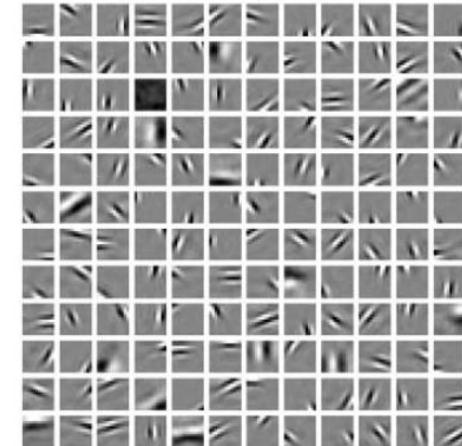
EM!

generative model



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

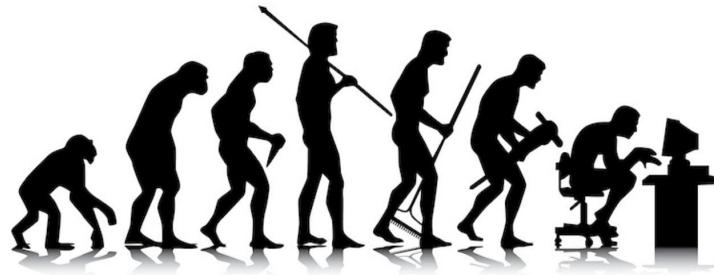
MULTI-LAYERED
CONVOLUTIONAL
NEURAL NETWORK



FIRST LAYER OF A
CONVOLUTIONAL
NEURAL NETWORK



FIRST LAYER OF A
NEURAL NETWORK



MULTI-LAYERED
CONVOLUTIONAL

SPARSE REPRESENTATION



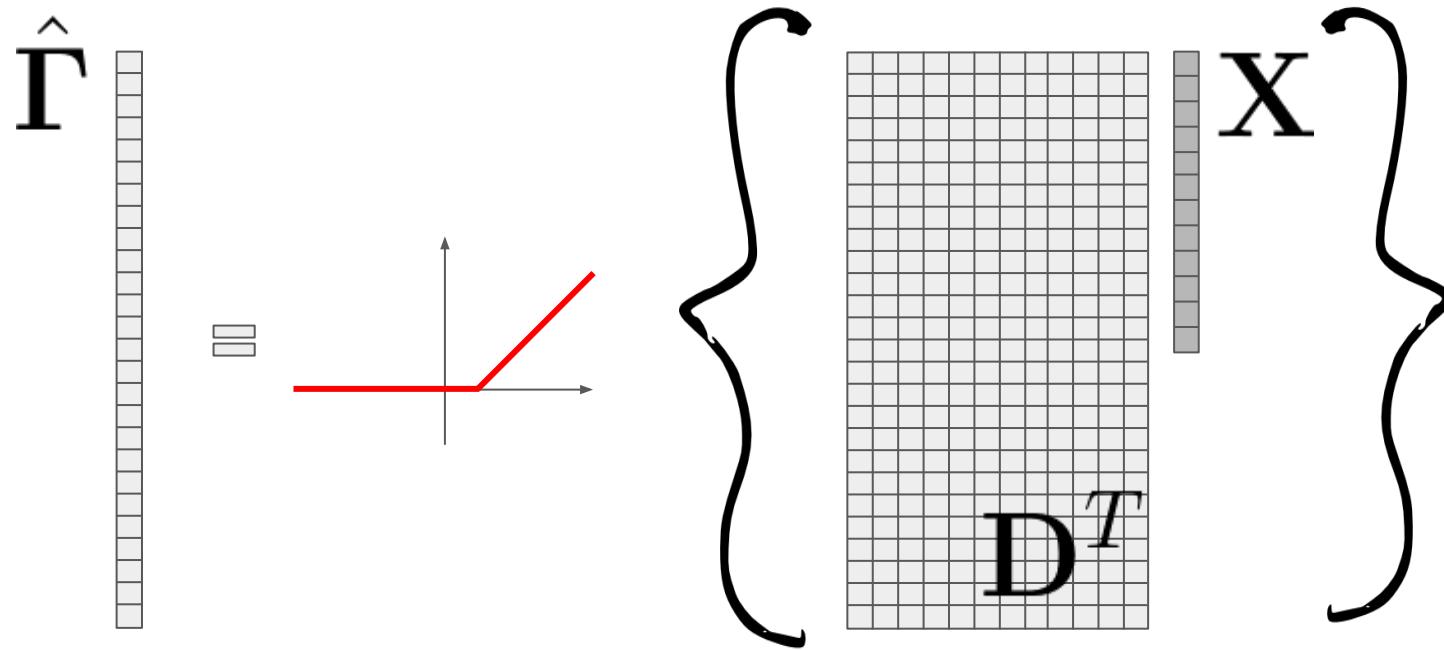
CONVOLUTIONAL

SPARSE REPRESENTATION



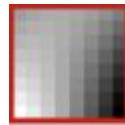
SPARSE REPRESENTATIONS

First Layer of a Neural Network



Sparse Modeling

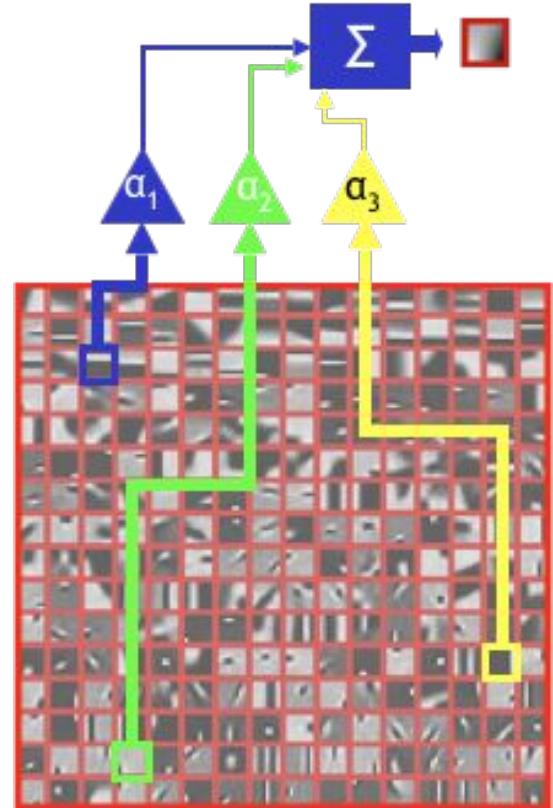
Task: model image patches of size 8x8 pixels



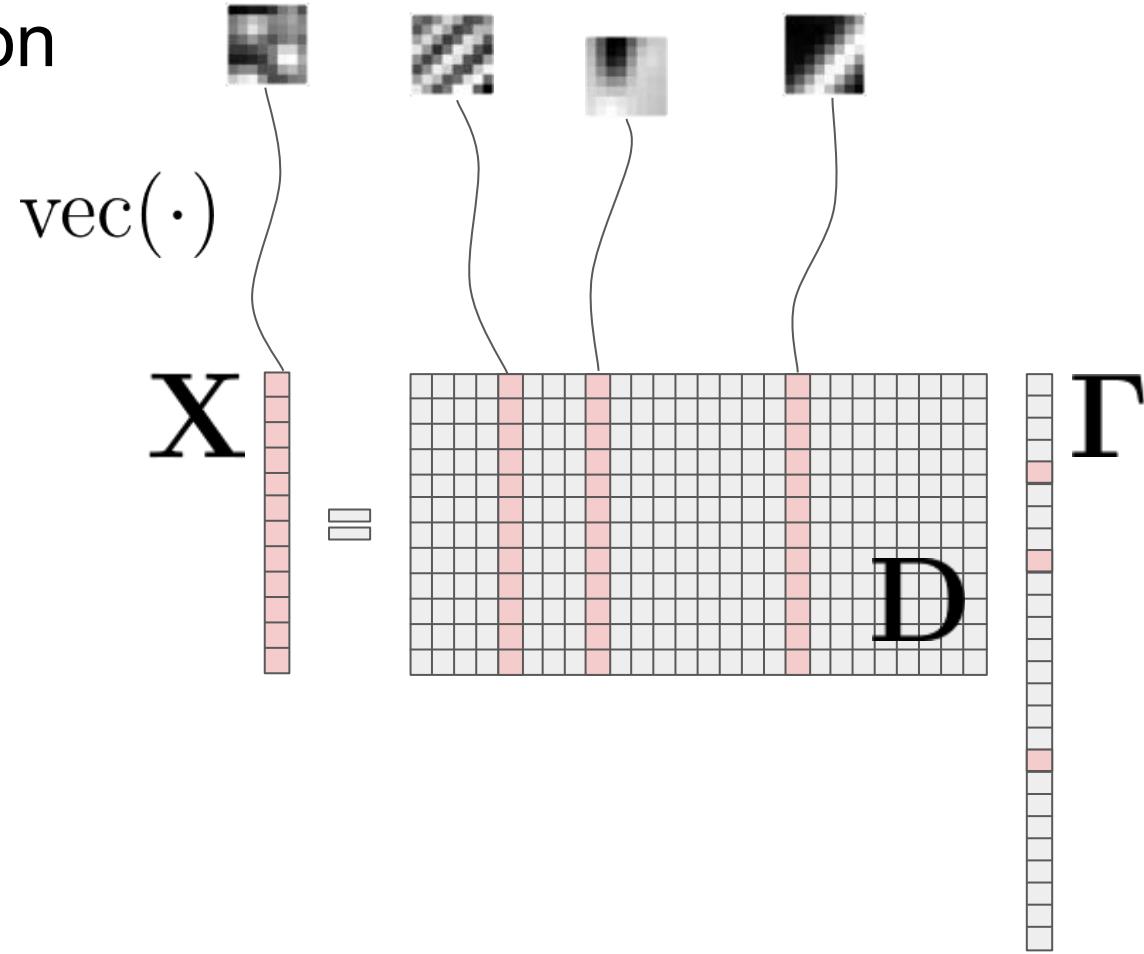
We assume a dictionary of such image patches is given, containing 256 atoms

Assumption: every patch can be described as a linear combination of a few atoms

Key properties: sparsity and redundancy



Matrix Notation



Sparse Coding

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

Convexify

$$\min_{\Gamma} \|\Gamma\|_0 \text{ s.t. } \mathbf{X} = \mathbf{D}\Gamma$$
$$\rightsquigarrow \min_{\Gamma} \|\Gamma\|_1 \text{ s.t. } \mathbf{X} = \mathbf{D}\Gamma$$

Sparse Coding

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

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

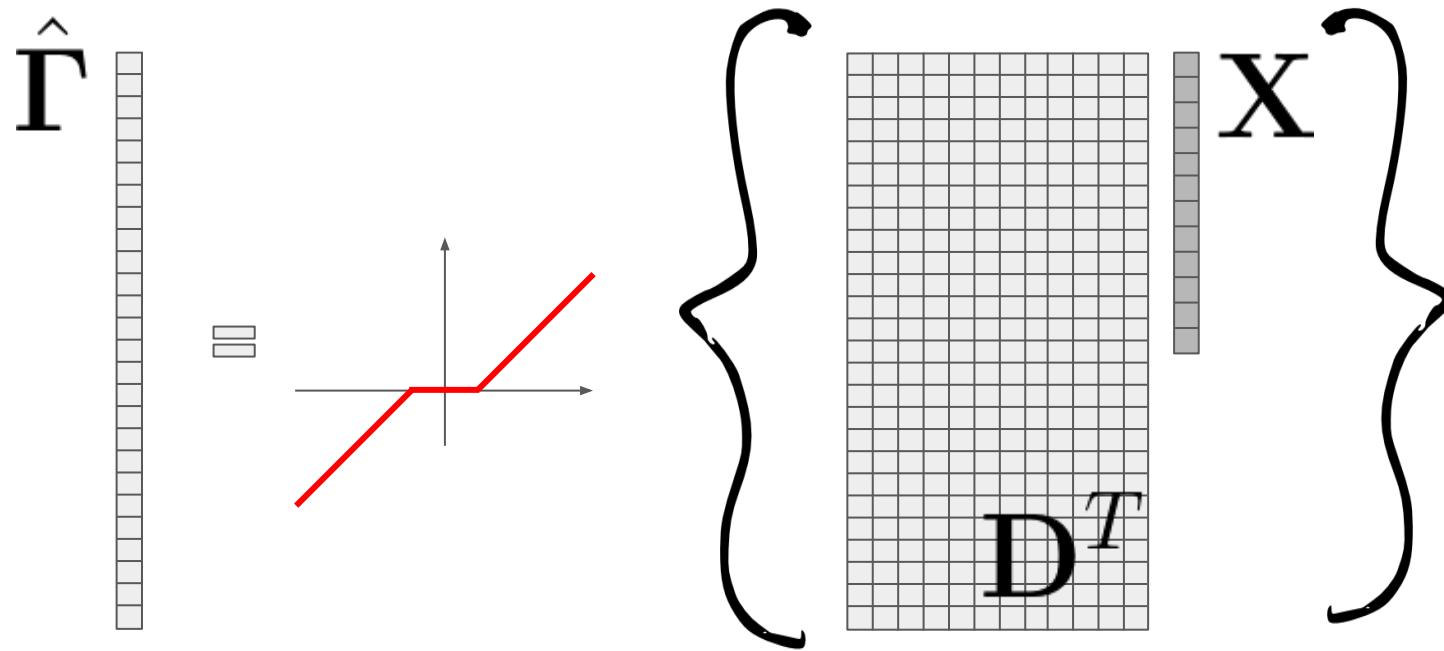
Convexify

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

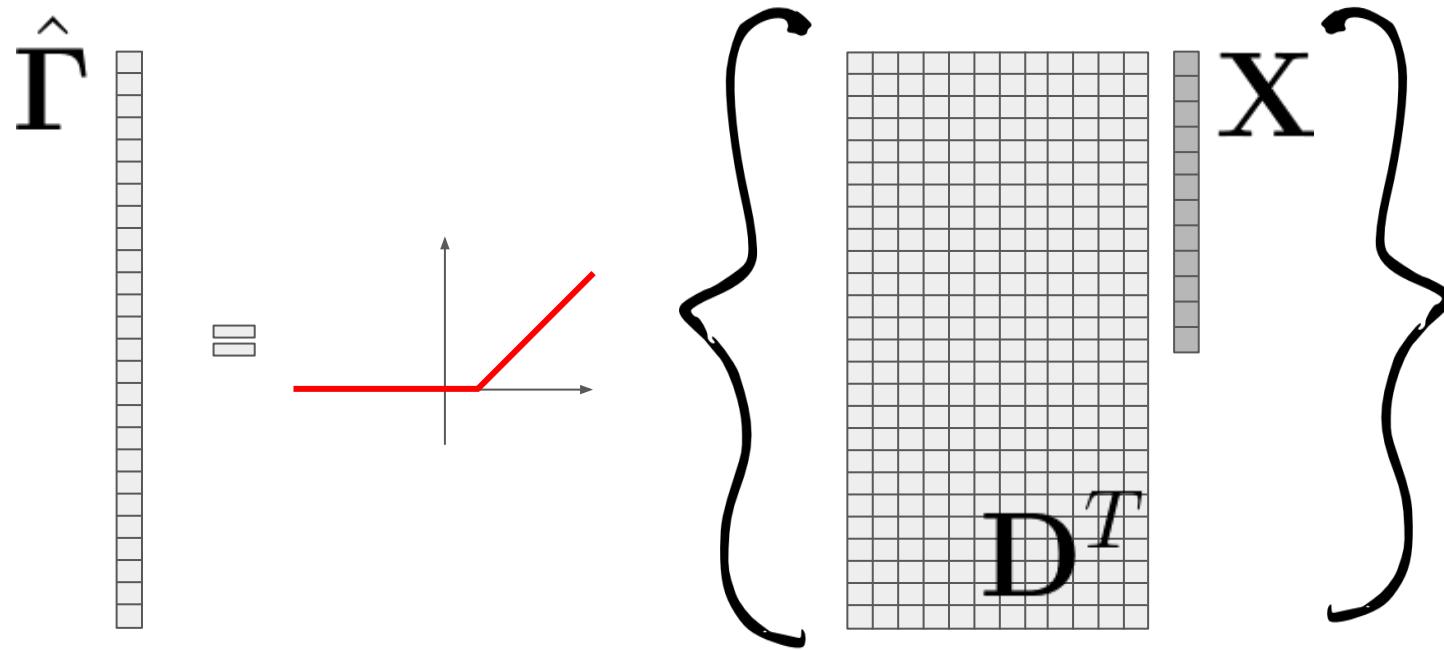
Crude
approximation

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

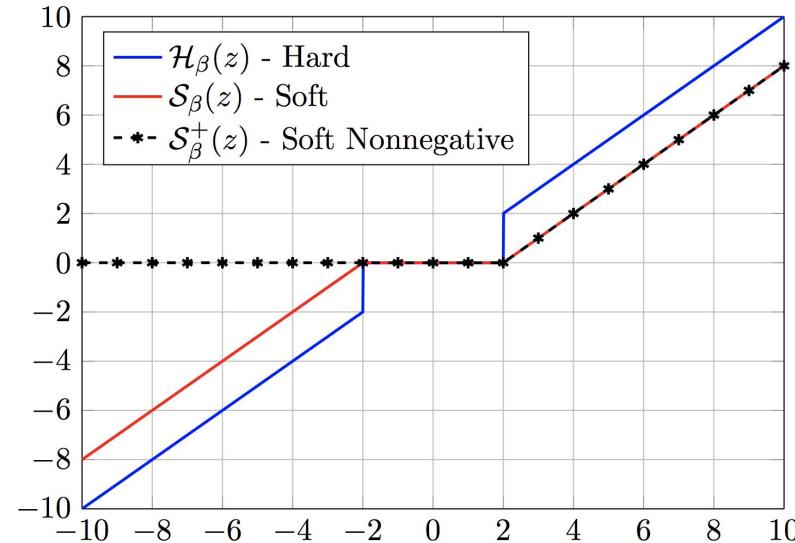
Thresholding Algorithm



First Layer of a Neural Network

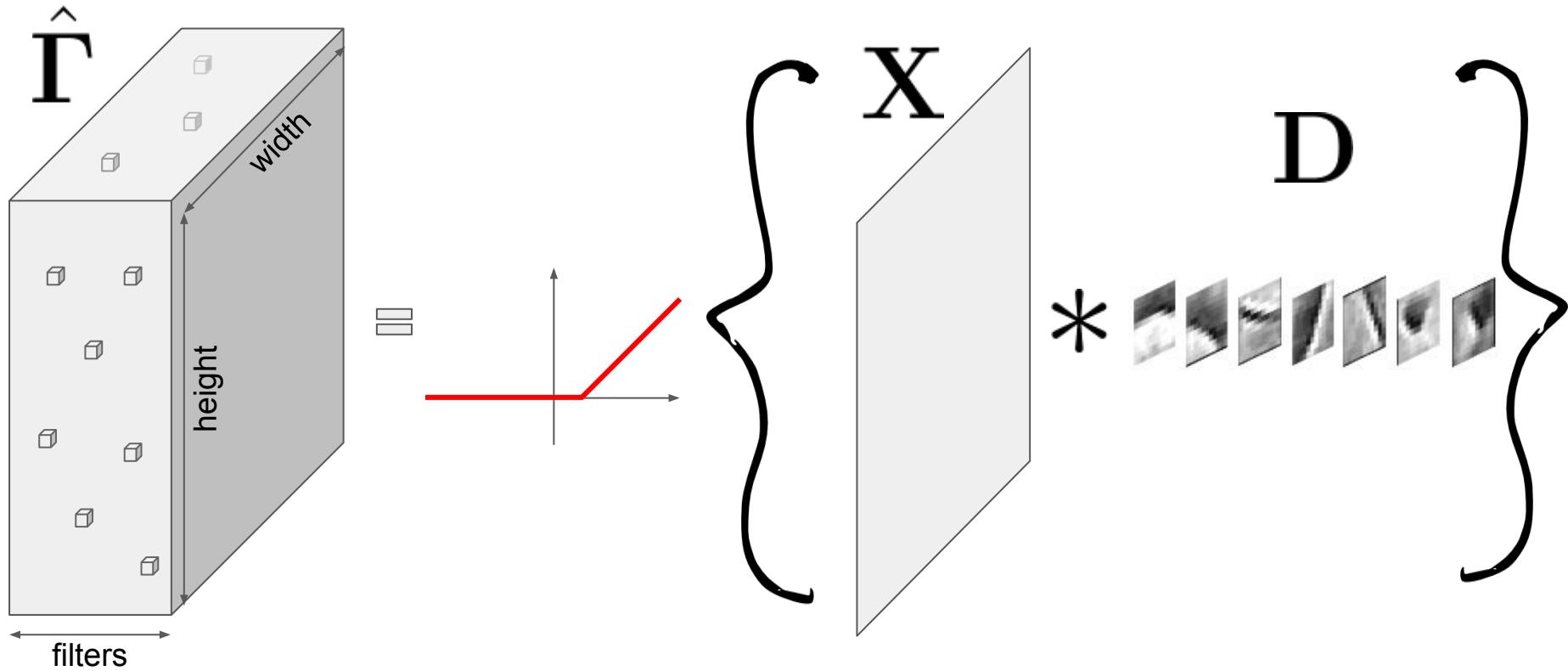


ReLU = Soft Nonnegative Thresholding



ReLU is equivalent to soft nonnegative thresholding

First layer of a Convolutional Neural Network

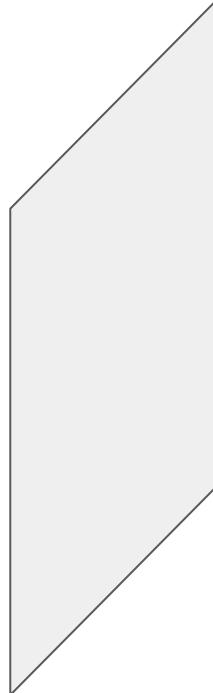


Convolutional Sparse Modeling

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

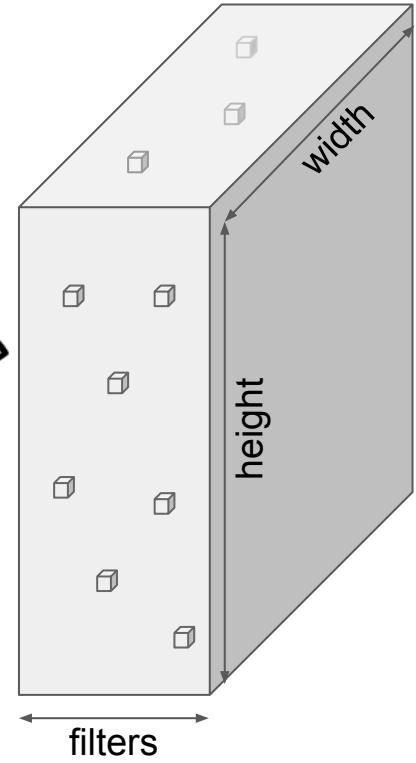
A diagram illustrating the convolutional sparse modeling equation. On the left, the matrix \mathbf{X} is shown as a grid with vertical lines on its left side. An equals sign follows. To the right of the equals sign is a large rectangular grid representing the product $\mathbf{D} \Gamma$. This grid is filled with small colored rectangles (purple, blue, green) forming a diagonal pattern that tapers to zero. The grid has a light gray background with a fine grid pattern. On the far right, there is a vertical column of dots, indicating that the matrix Γ is a tall column vector.

Convolutional Sparse Modeling

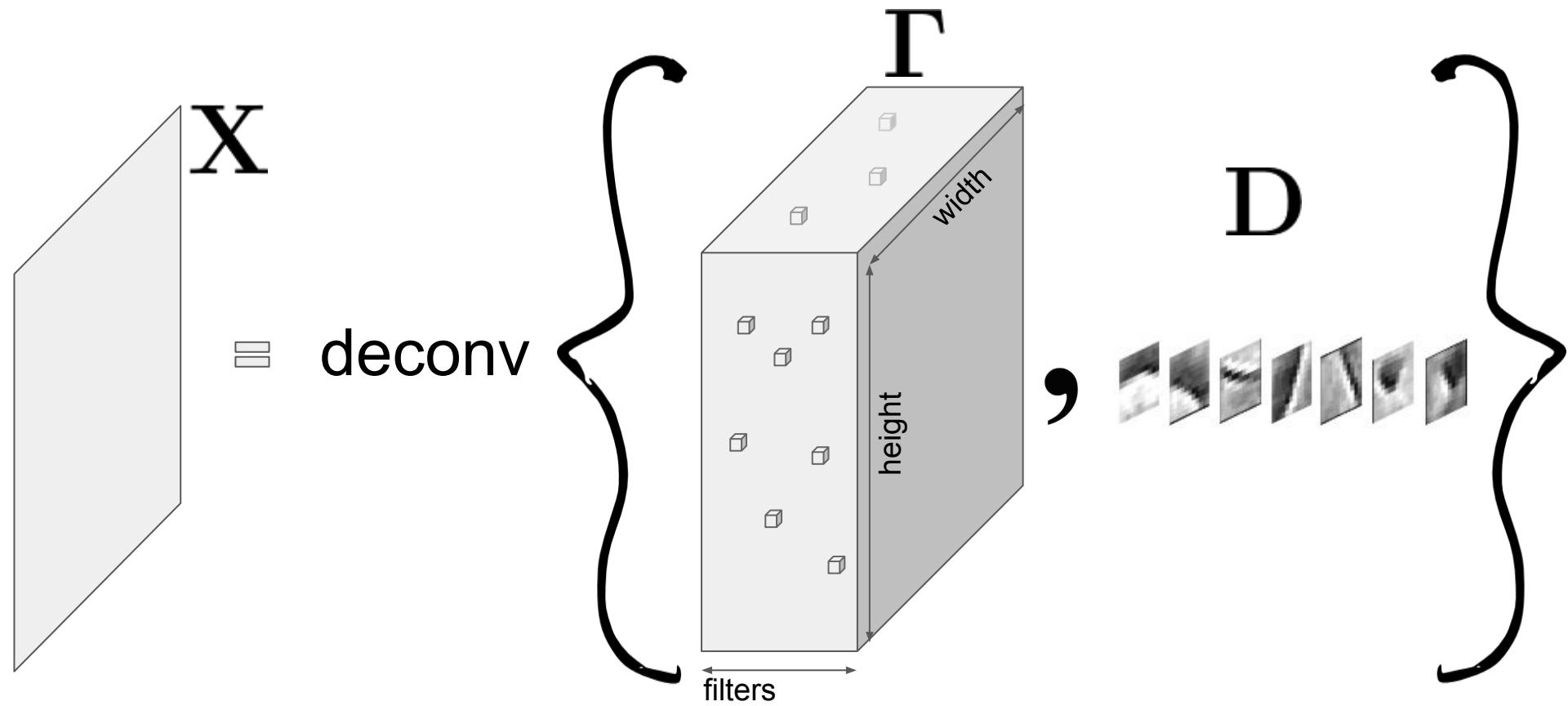


$$X = \sum_{k=1}^K d_k * z_k$$

A diagram illustrating the convolutional sparse modeling equation. A dashed arrow points from the left side of the equation to the input image X . Another dashed arrow points from the right side of the equation to a stack of filters below. A curved dashed arrow points from the term $d_k * z_k$ to the stack of filters, indicating that each filter d_k is applied to the input image X to produce a sparse representation z_k .

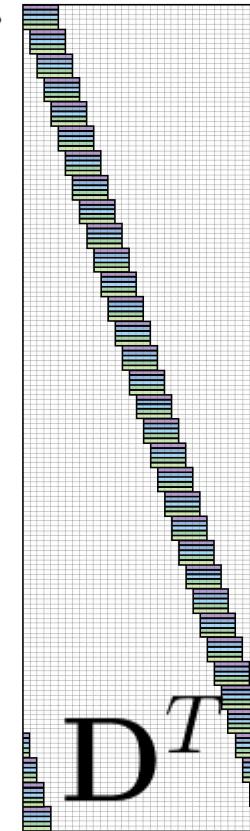
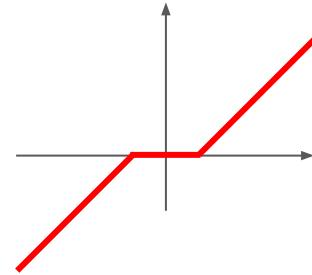


Convolutional Sparse Modeling



Thresholding Algorithm

$\hat{\Gamma}$

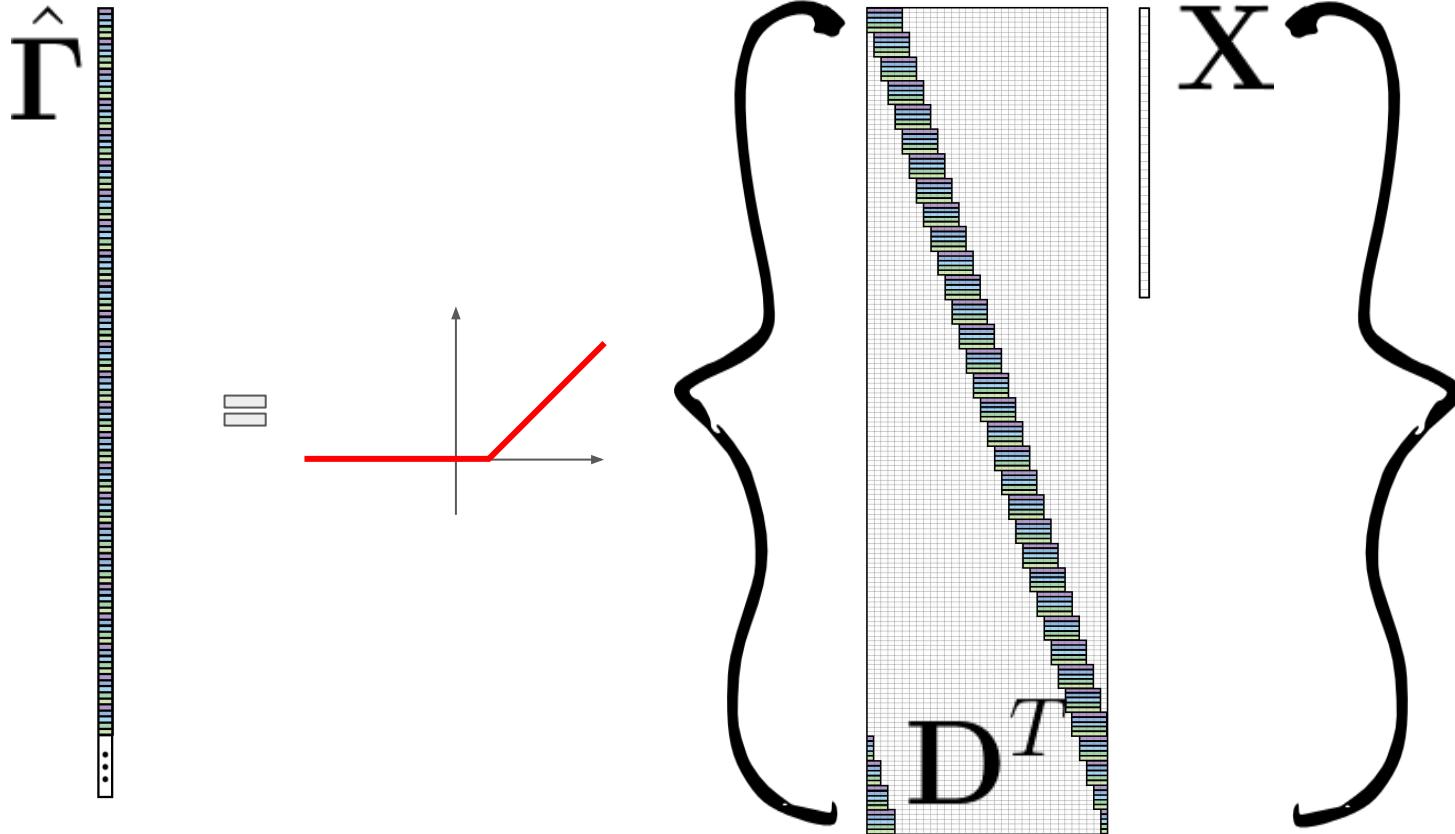


D^T

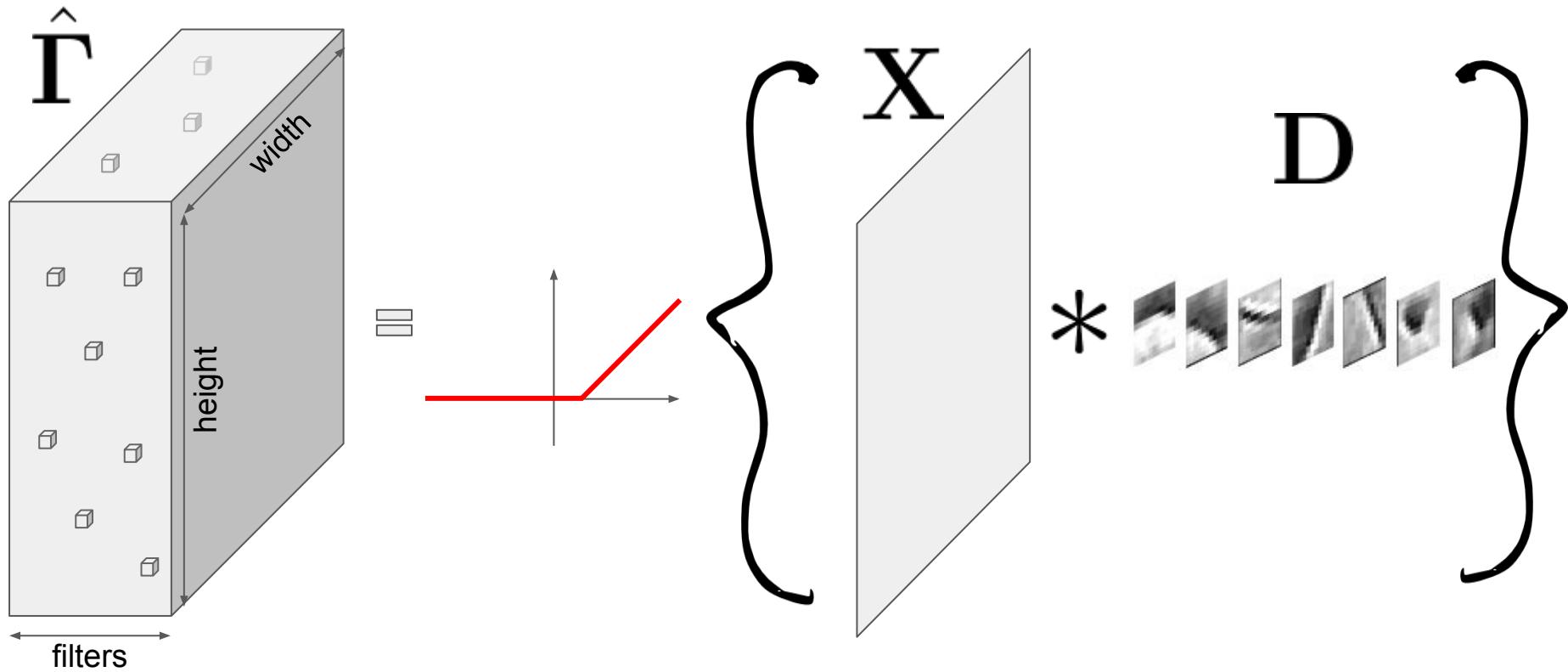
X



First layer of a Convolutional Neural Network

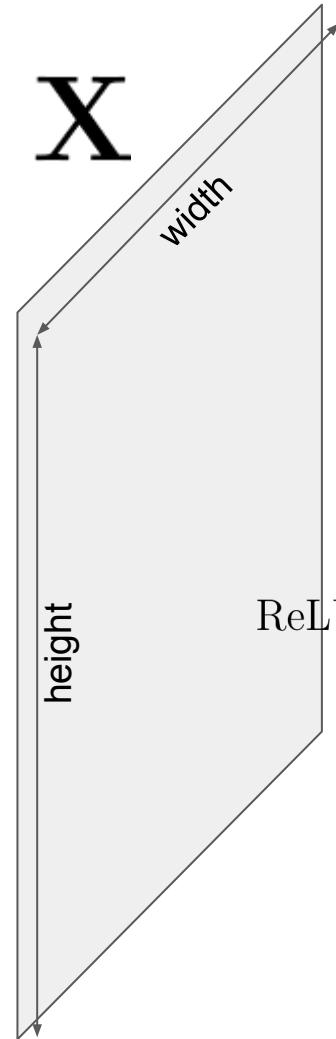


First layer of a Convolutional Neural Network

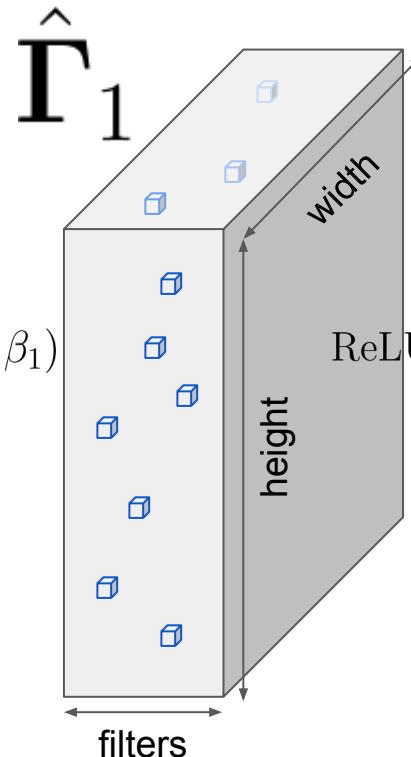


X

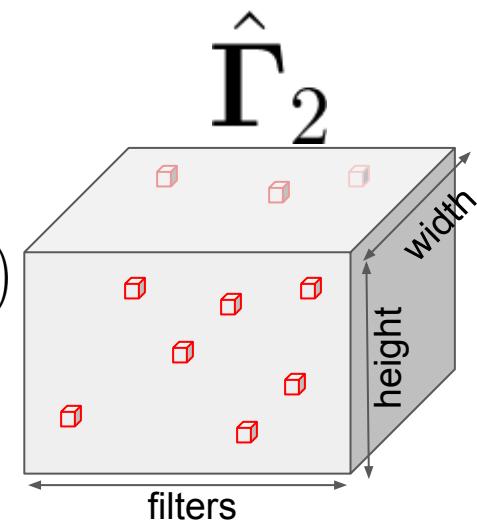
Convolutional Neural Network



$$\text{ReLU}(\text{conv}(X, D_1) + \beta_1)$$

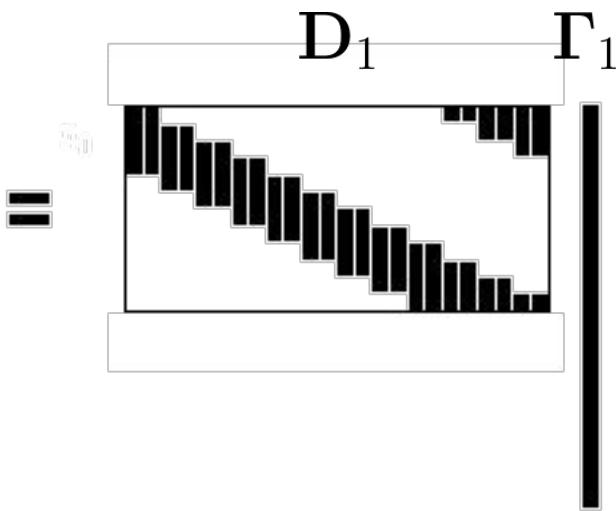


$$\text{ReLU}(\text{conv}(\hat{\Gamma}_1, D_2) + \beta_2)$$

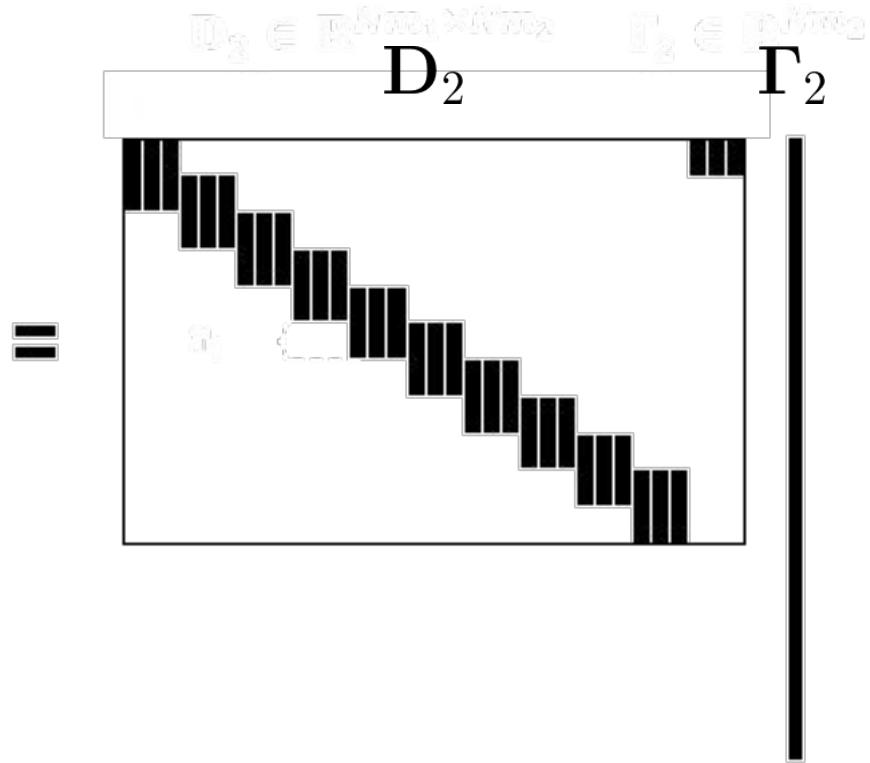


Multi-layered Convolutional Sparse Modeling

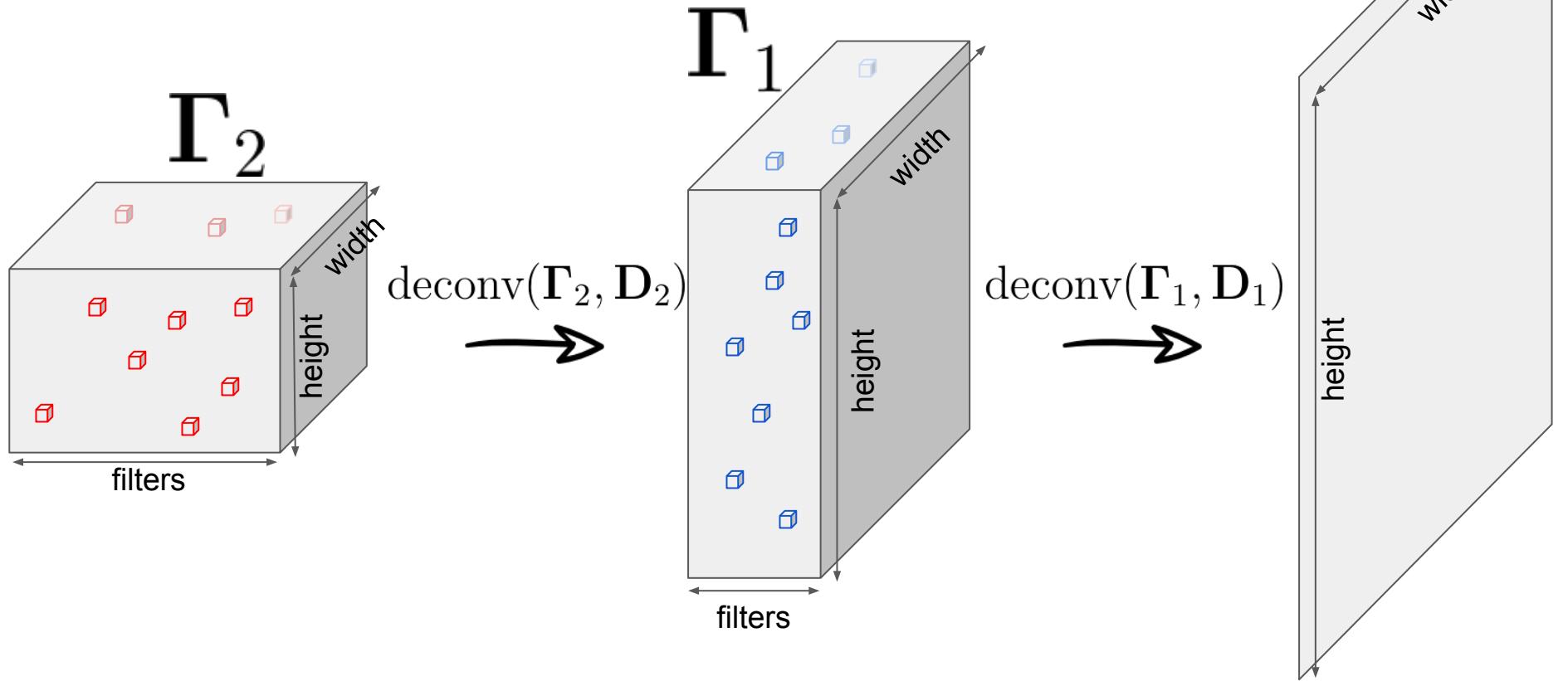
X



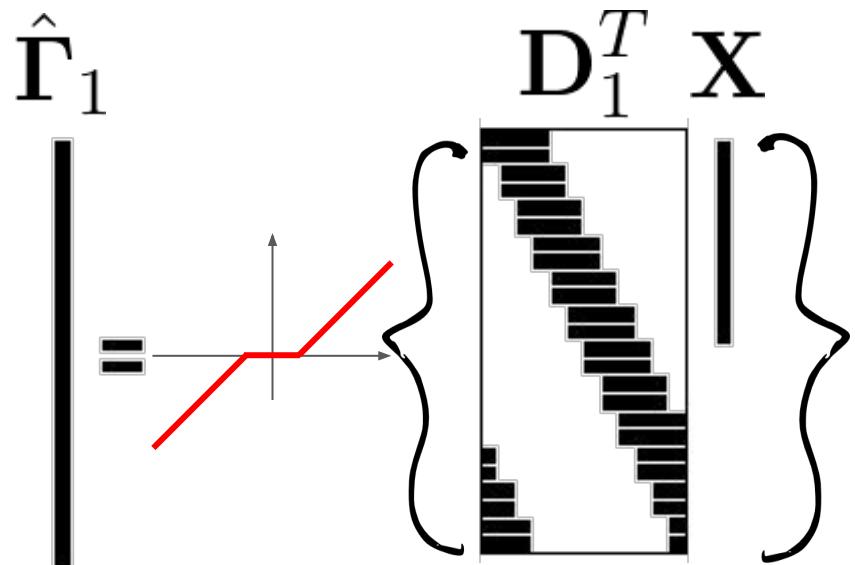
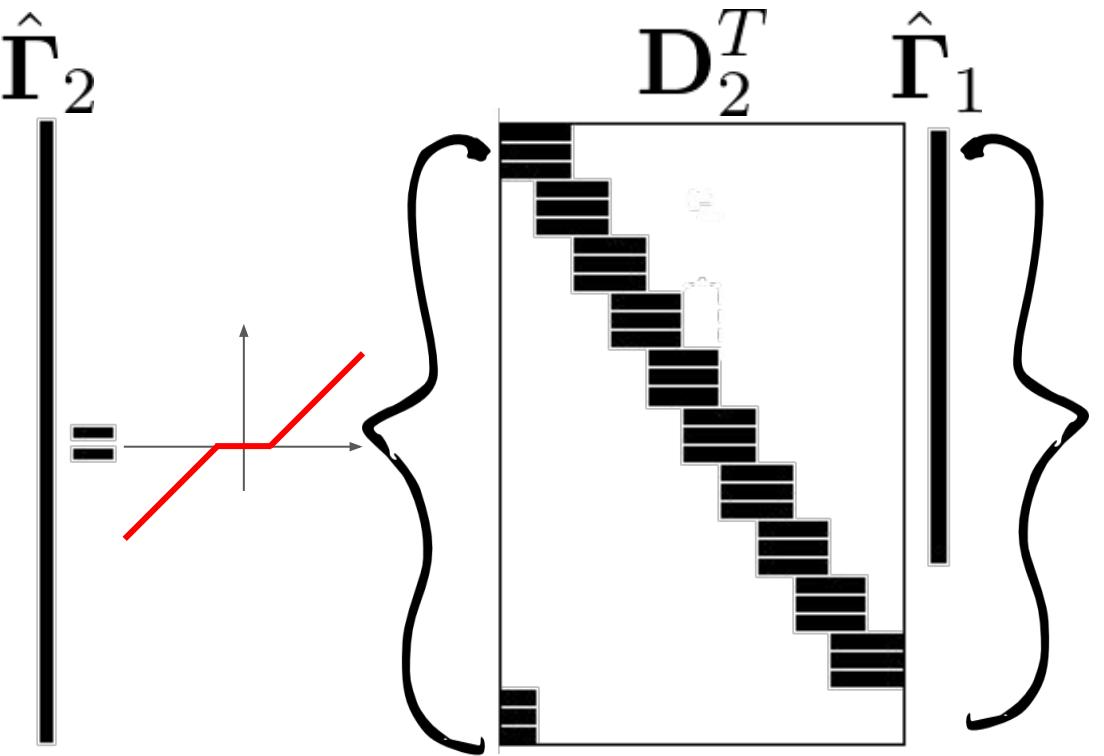
Γ_1



Multi-layered Convolutional Sparse Modeling



Layered Thresholding



$\hat{\Gamma}_2$

Convolutional Neural Network

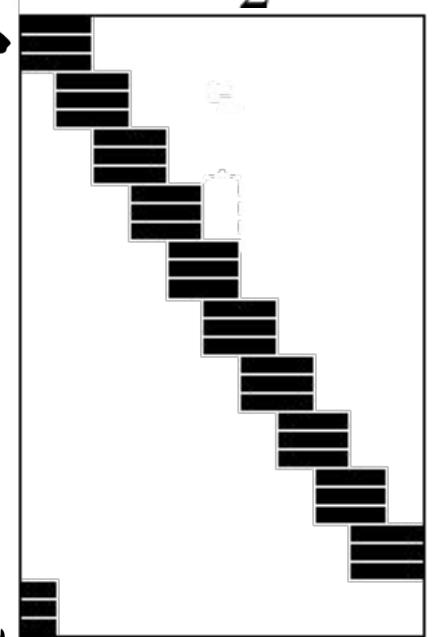
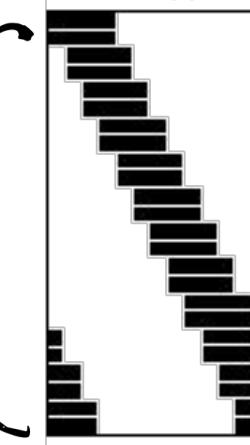
$\hat{\Gamma}_1$

$D_1^T \ X$

D_2^T

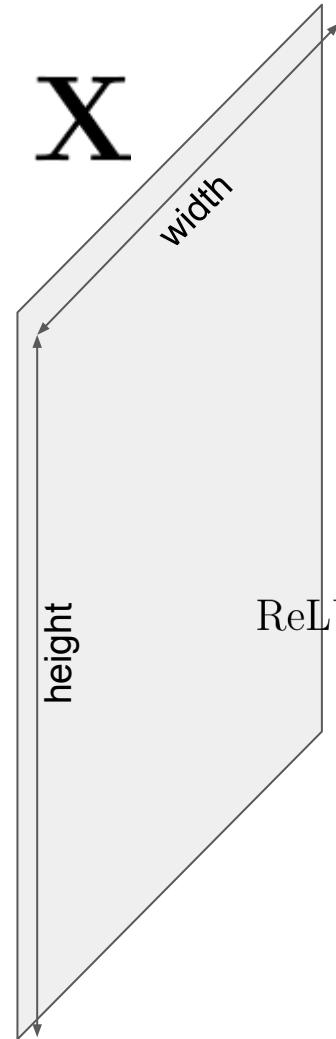
$\hat{\Gamma}_1$

=

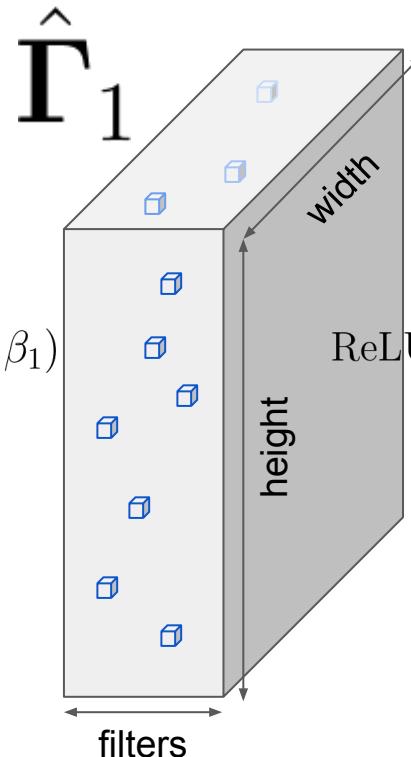


X

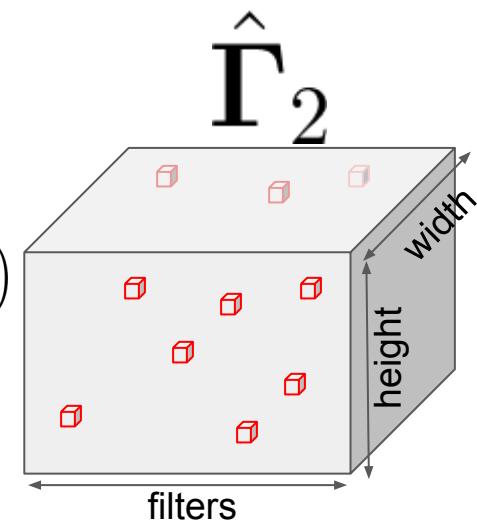
Convolutional Neural Network



$$\text{ReLU}(\text{conv}(X, D_1) + \beta_1)$$



$$\text{ReLU}(\text{conv}(\hat{\Gamma}_1, D_2) + \beta_2)$$



Theories of Deep Learning



Evolution of Models

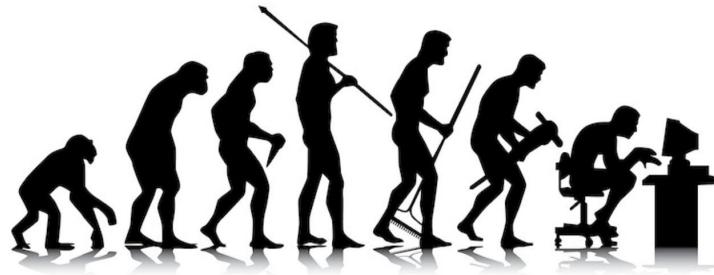
MULTI-LAYERED
CONVOLUTIONAL
NEURAL NETWORK



FIRST LAYER OF A
CONVOLUTIONAL
NEURAL NETWORK



FIRST LAYER OF A
NEURAL NETWORK



MULTI-LAYERED
CONVOLUTIONAL

SPARSE REPRESENTATION



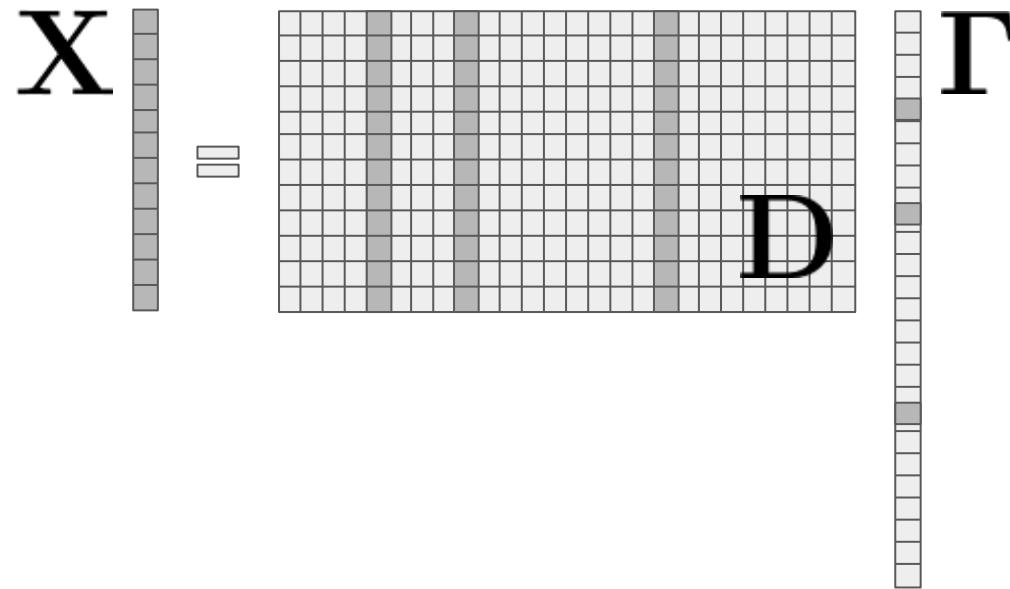
CONVOLUTIONAL

SPARSE REPRESENTATION



SPARSE REPRESENTATIONS

Sparse Modeling

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


The diagram illustrates the sparse modeling equation $\mathbf{X} = \mathbf{D}\Gamma$. On the left, a tall vertical vector \mathbf{X} is shown, consisting of a sequence of gray blocks. In the middle, the multiplication operator $=$ is represented by a horizontal double bar. To the right of the bar is a square matrix \mathbf{D} , which has a grid pattern with several vertical columns highlighted in gray. To the right of \mathbf{D} is a short vertical vector Γ , consisting of a sequence of gray blocks.

Classic Sparse Theory

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

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

Theorem: [Donoho and Elad, 2003]

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

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

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

$$\mathbf{D}^T$$

$$\mathbf{D}$$

=

$$\mathbf{D}^T \mathbf{D}$$

Convolutional Sparse Modeling

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

A diagram illustrating the convolutional sparse modeling equation. On the left, the matrix \mathbf{X} is shown as a grid with vertical lines on its left side. An equals sign follows. To the right of the equals sign is a large rectangular grid representing the product $\mathbf{D} \Gamma$. This grid is filled with small colored rectangles (purple, blue, green) forming a diagonal pattern that tapers to zero. The grid has a light gray background with a fine grid pattern. On the far right, there is a vertical column of dots, indicating that the matrix Γ is of infinite length.

Classic Sparse Theory for Convolutional Case

Theorem: [Donoho and Elad, 2003]

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

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

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

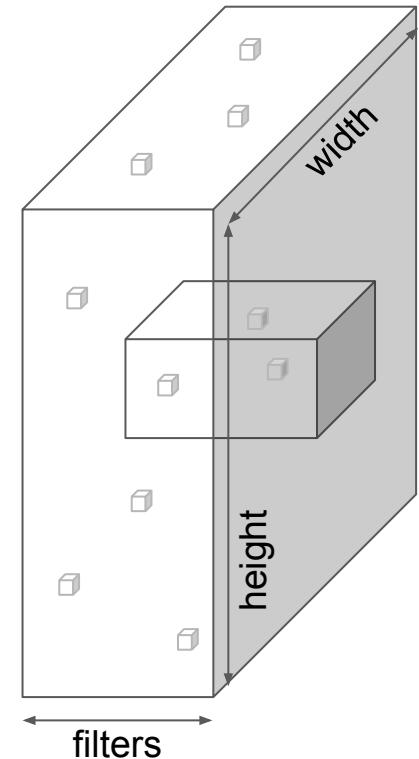
Success guaranteed when $\|\Gamma\|_0 < 8.43$

Very pessimistic!

Local Sparsity

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

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



Success of Basis Pursuit

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

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

Theorem: [Papyan, Sulam and Elad, 2016]

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

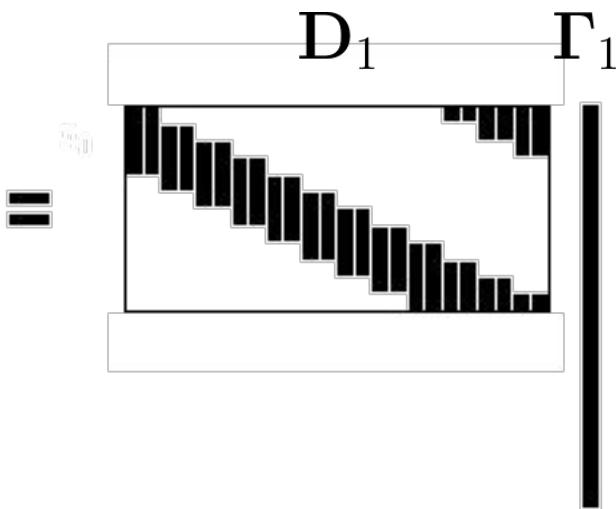
Then: $\|\hat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}\|_\infty \leq 7.5 \|\mathbf{E}\|_{2,\infty}$

Theoretical guarantee for:

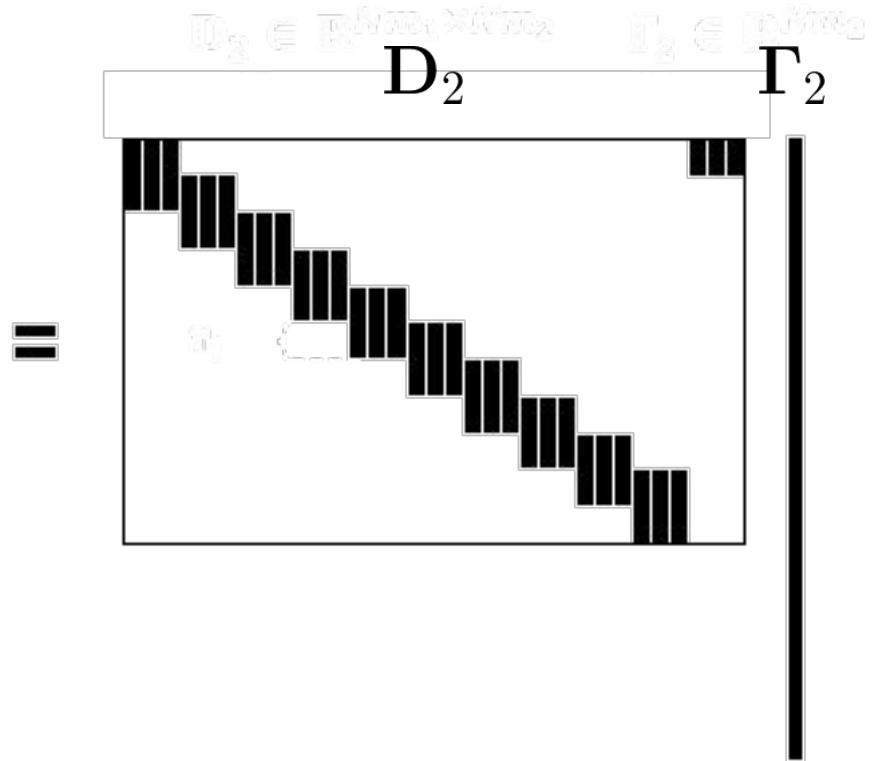
- [Zeiler et. al 2010]
- [Wohlberg 2013]
- [Bristow et. al 2013]
- [Fowlkes and Kong 2014]
- [Zhou et. al 2014]
- [Kong and Fowlkes 2014]
- [Zhu and Lucey 2015]
- [Heide et. al 2015]
- [Gu et. al 2015]
- [Wohlberg 2016]
- [Šorel and Šroubek 2016]
- [Serrano et. al 2016]
- [Papyan et. al 2017]
- [Garcia-Cardona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

Multi-layered Convolutional Sparse Modeling

X



Γ_1



Deep Coding Problem

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

$$\mathbf{X} = \mathbf{D}_1 \boldsymbol{\Gamma}_1, \quad \|\boldsymbol{\Gamma}_1\|_{0,\infty} \leq \lambda_1$$

$$\boldsymbol{\Gamma}_1 = \mathbf{D}_2 \boldsymbol{\Gamma}_2, \quad \|\boldsymbol{\Gamma}_2\|_{0,\infty} \leq \lambda_2$$

⋮

$$\boldsymbol{\Gamma}_{L-1} = \mathbf{D}_L \boldsymbol{\Gamma}_L, \quad \|\boldsymbol{\Gamma}_L\|_{0,\infty} \leq \lambda_L$$

Deep Coding Problem

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

$$\|\mathbf{Y} - \mathbf{D}_1\boldsymbol{\Gamma}_1\|_2 \leq \epsilon, \quad \|\boldsymbol{\Gamma}_1\|_{0,\infty} \leq \lambda_1$$

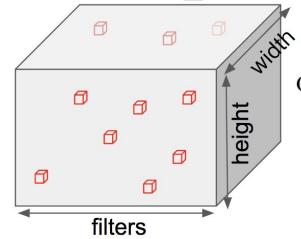
$$\boldsymbol{\Gamma}_1 = \mathbf{D}_2\boldsymbol{\Gamma}_2, \quad \|\boldsymbol{\Gamma}_2\|_{0,\infty} \leq \lambda_2$$

⋮

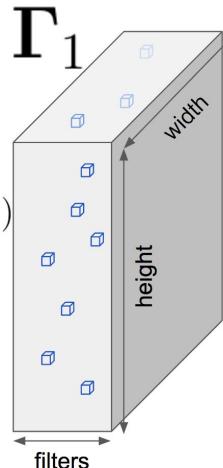
$$\boldsymbol{\Gamma}_{L-1} = \mathbf{D}_L\boldsymbol{\Gamma}_L, \quad \|\boldsymbol{\Gamma}_L\|_{0,\infty} \leq \lambda_L$$

Uniqueness

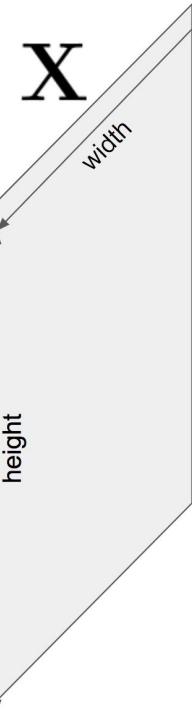
$$\Gamma_2$$



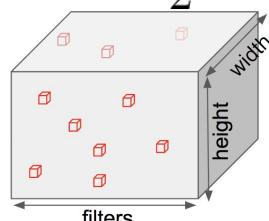
deconv(Γ_2, D_2)



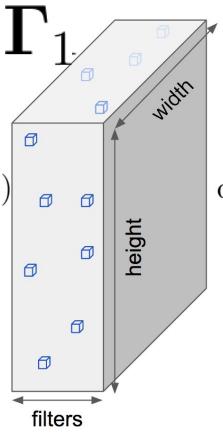
deconv(Γ_1, D_1)



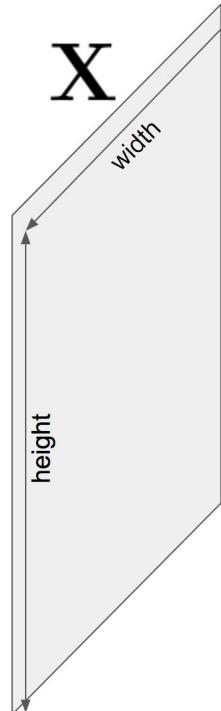
$$\hat{\Gamma}_2$$



deconv(Γ_2, D_2)



deconv(Γ_1, D_1)



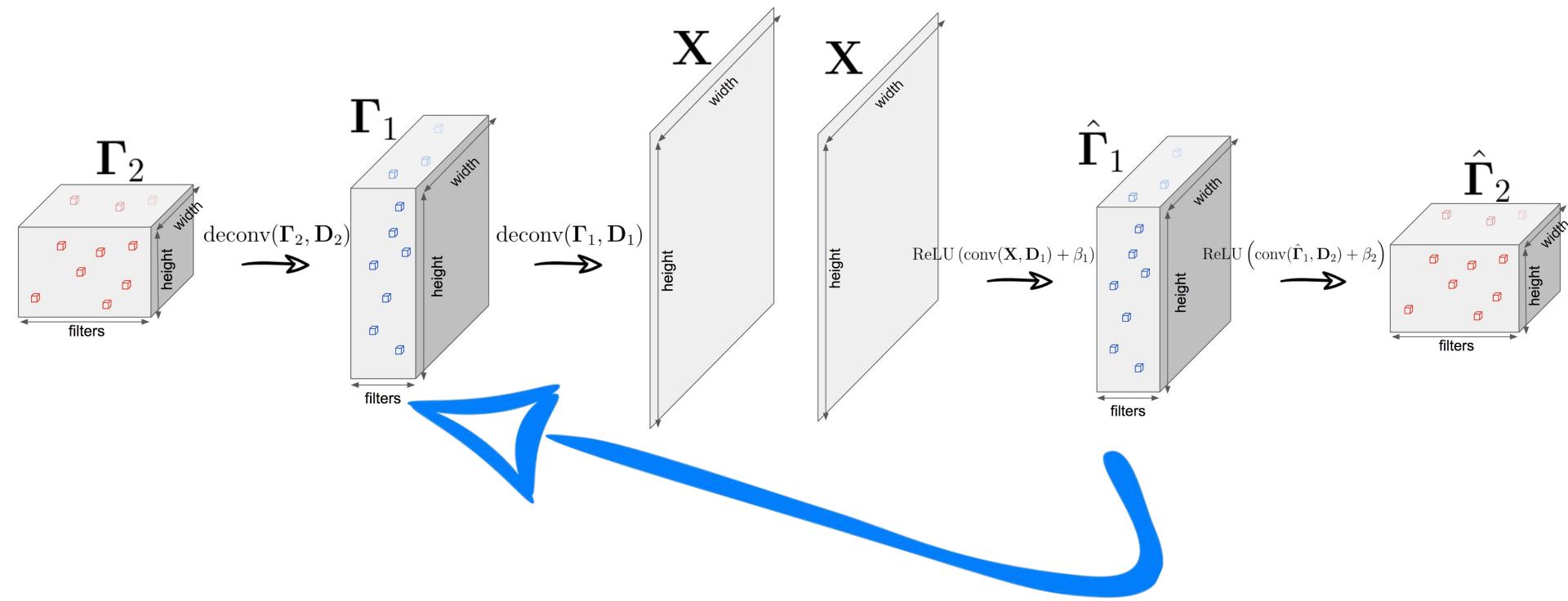
Uniqueness Theorem

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



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

Success of Forward Pass



Success of Forward Pass Theorem

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

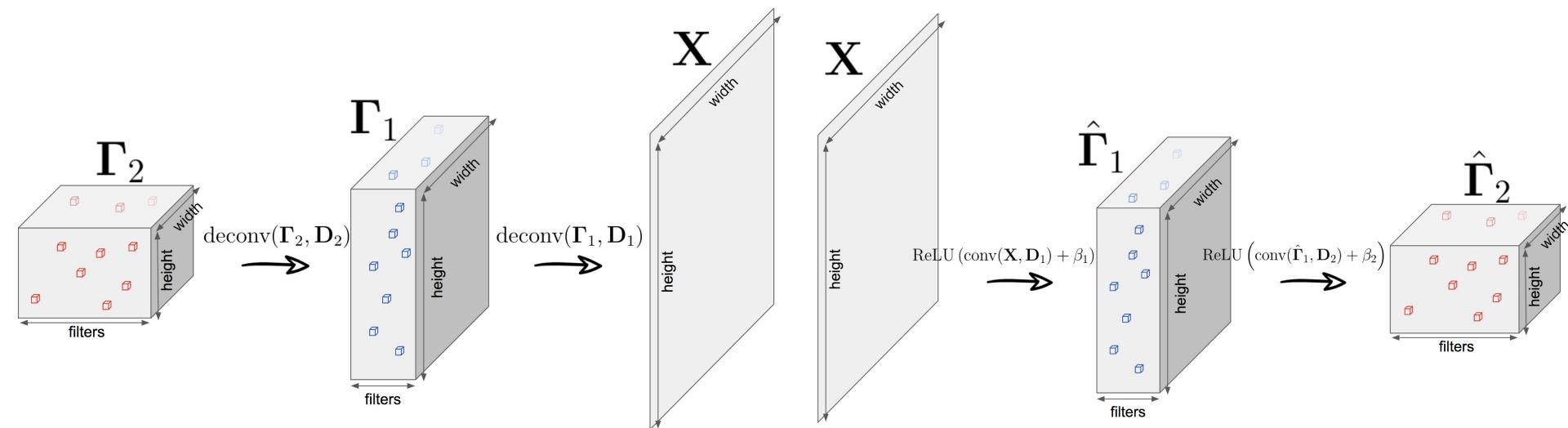


Layered thresholding guaranteed:

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

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

Generative Model and Crude Inference



Layered Lasso

# StatsDepartment

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

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

Success of Layered Lasso

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



Layered Lasso guaranteed:

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

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

Layered Iterative Thresholding

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

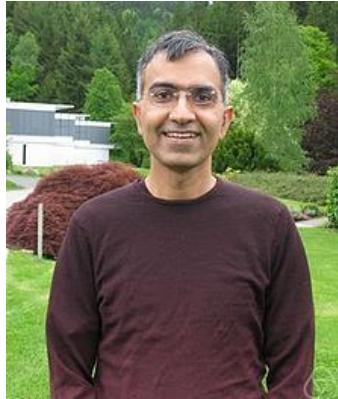
$$\boldsymbol{\Gamma}_2^t = \mathcal{S}_{\alpha_2} \left(\mathbf{D}_2^T \hat{\boldsymbol{\Gamma}}_1 + (\mathbf{I} - \mathbf{D}_2^T \mathbf{D}_2) \boldsymbol{\Gamma}_2^{t-1} \right)$$



Supervised Deep Sparse Coding Networks

[Sun et. al 2017]

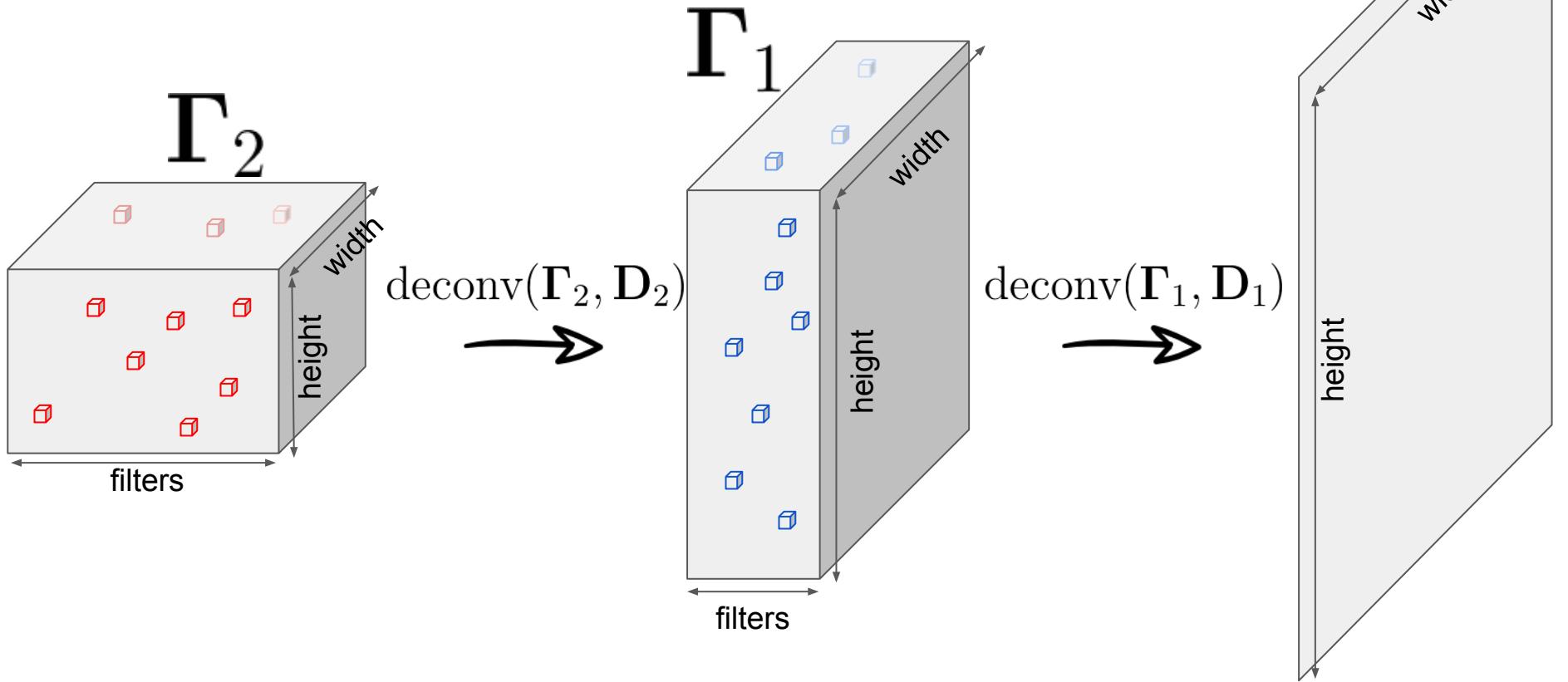
Method	# Params	# Layers	CIFAR-10	CIFAR-100
SCKN [34]	10.50M	10	10.20	-
OMP [18]	0.70M	2	18.50	-
PCANet [36]	0.28B	3	21.33	-
NOMP [7]	1.09B	4	18.60	39.92
NiN [32]	-	-	8.81	35.68
DSN [33]	1.34M	7	7.97	36.54
WRN [12]	36.5M	28	4.00	19.25
ResNet-110 [10]	0.85M	110	6.41	27.22
ResNet-1001 v2 [11]	10.2M	1001	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

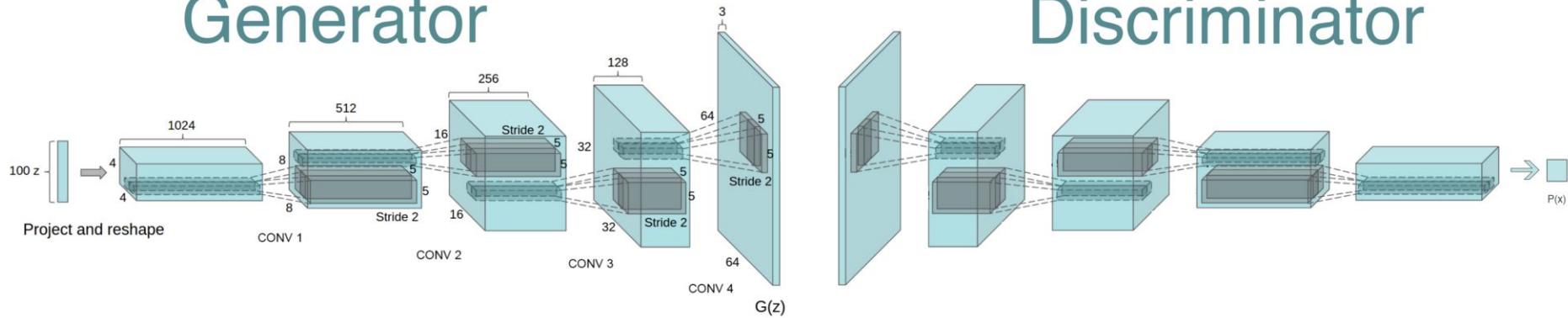


Multi-layered Convolutional Sparse Modeling



Generator in GANs [Goodfellow et. al 2014]

Generator



Discriminator

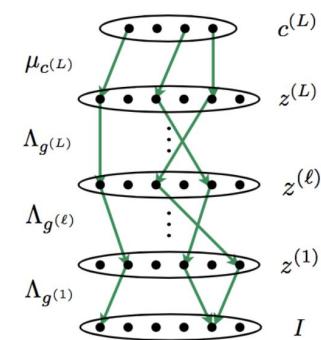
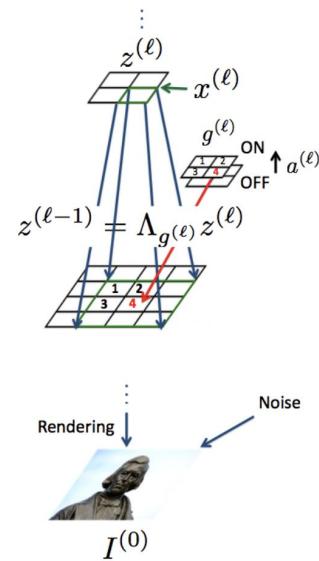
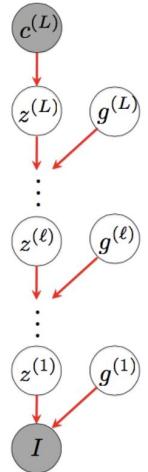
Sparsification of intermediate feature maps with **ReLU**

DRMM [Patel et. al]

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

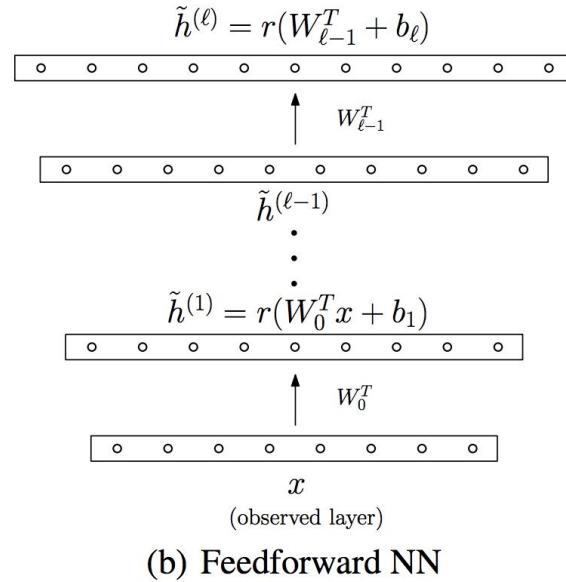
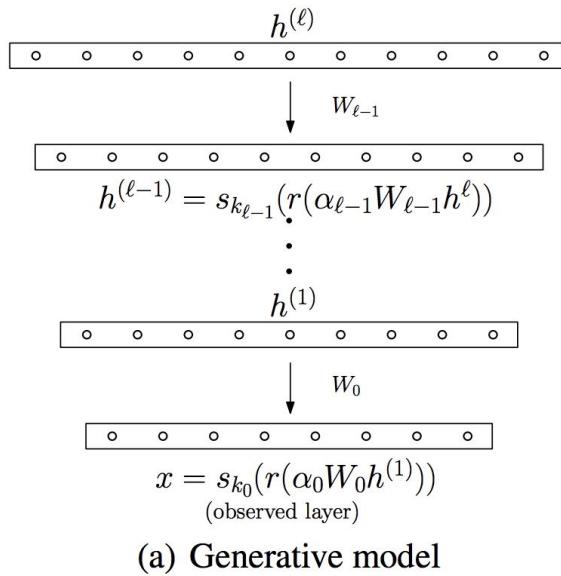
$$I \sim \mathcal{N}(\mu_{cg}, \sigma^2 \mathbf{1}_{D^{(0)}}),$$

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



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

[Arora et. al, 2015]



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

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

