Multi-Layer Convolutional Sparse Coding

Shihua Zhang

November 10, 2021

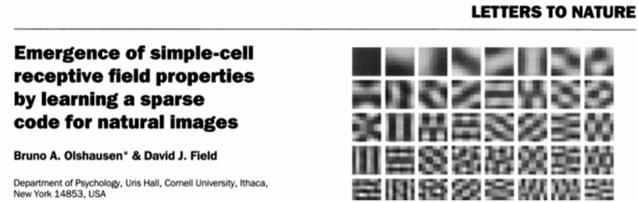
Outline

1. Sparse Coding
2. Convolutional Sparse Coding (CSC)
3. Connection between CSC and CNN
4. Towards to Understand ResNet and MSDNet

Sparse Coding: Birth

- Inspired by signal transform and visual cortex studies, sparse coding of natural images was developed [Olshausen and Field, 1996].

Bruno Olshausen David Field
Department of Psychology at Cornell University



LETTERS TO NATURE

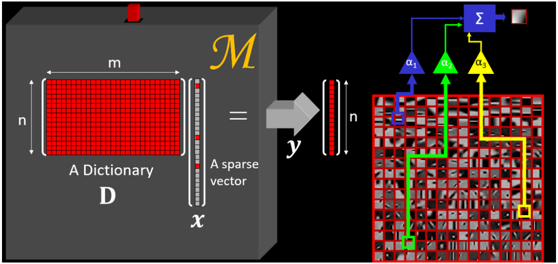
Emergence of simple-cell receptive field properties by learning a sparse code for natural images

Bruno A. Olshausen* & David J. Field

Department of Psychology, Uris Hall, Cornell University, Ithaca, New York 14853, USA

Sparse Coding: Model

- Task: model image patches
- Assumption: every patch can be described as a linear combination of a few atoms, where the atoms are learned from data.



- Assume $D \in \mathbb{R}^{n \times m}$ is an overcomplete dictionary $(m \gg n)$, $y \in \mathbb{R}^n$ is an input signal, $x \in \mathbb{R}^m$ is a sparse representation of y based on D:

$$y = Dx$$

Sparse Coding

- Let $P(\cdot)$ be a regularization term to ensure sparseness, then the problem can be rewritten as follows:

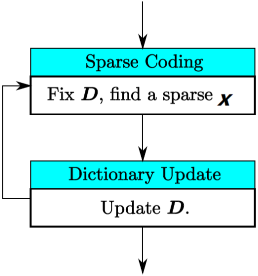
$$\min_{x,D} \frac{1}{2}\|y - Dx\|_2^2 + \lambda P(x)$$

Sparse Coding

- Let $P(\cdot)$ be a regularization term to ensure sparseness, then the problem can be rewritten as follows:

$$\min_{x,D} \frac{1}{2}\|y - Dx\|_2^2 + \lambda P(x)$$

- It can be splitted to two subproblems:
  - Sparse coding: Given y, fix D, find a sparse x
  - Dictionary learning: Given a family of y, find a suitable dictionary D.



Iterative Shrinkage Thresholding Algorithm (ISTA)

- The origin problem can be rewritten as:

$$\min_x \frac{1}{2}\|y - Dx\|_2^2 + \lambda\|x\|_1 \qquad (P_1)$$

It is a traditional problem called Basis Pursuit (BP).

Iterative Shrinkage Thresholding Algorithm (ISTA)

- The origin problem can be rewritten as:

$$\min_x \frac{1}{2}\|y - Dx\|_2^2 + \lambda\|x\|_1 \qquad (P_1)$$

  It is a traditional problem called Basis Pursuit (BP).

- ISTA updates:

$$x_{k+1} = S_{\frac{\lambda}{L}}\left(x_k - \frac{1}{L}D^T(Dx_k - y)\right)$$

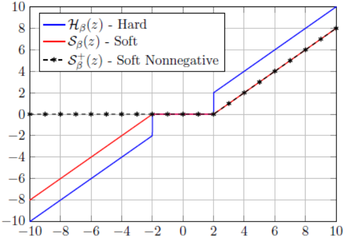


Figure 3: The thresholding operators for a constant $\beta = 2$.

Theoretical Guarantee

- Can ISTA find the unique solution?
  ——The answer is YES under certain circumstances

Definition 1

Assume d_i is the column vector of D, $\hat{d}_i = \frac{\hat{d}_i}{\|\hat{d}_i\|_2}$, the mutual coherence $\mu(D)$ of dictionary D is defined as: $\mu(D) = \max_{i \neq j} |\hat{d}_i^T \hat{d}_j|$

Theorem 2

The convex relaxation approaches above can recover the true solution x^* if $\|x^*\|_0 < \frac{1}{2}(1 + \frac{1}{\mu(D)})$ [Donoho et al., 2005]

Approximation Algorithm

- There is a simplest approximation algorithm [Papyan et al., 2017a]:
  - Compute the inner products between signal y and all atoms in D.
  - Choose the atoms corresponding to the highest responses.

Approximation Algorithm

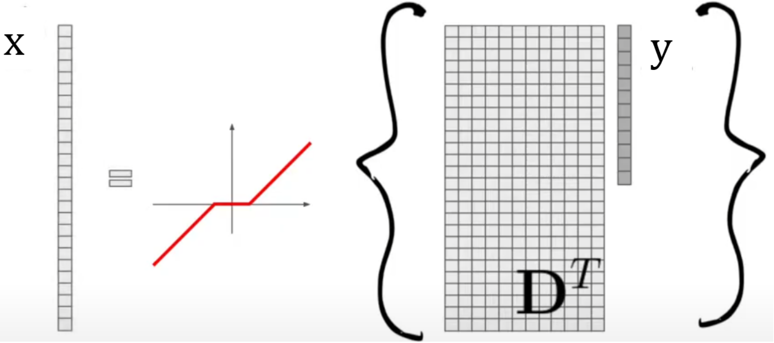
- There is a simplest approximation algorithm [Papyan et al., 2017a]:
  - Compute the inner products between signal y and all atoms in D.
  - Choose the atoms corresponding to the highest responses.

- The approximation problem can be written as:

$$\min_x \frac{1}{2}\|x - D^T y\|_2^2 + \beta\|x\|_1$$

- The solution to the above form is simple: $x = S_\beta(D^T y)$.
- The theoretical guarantee of this method is weaker than ISTA.

Approximation Algorithm



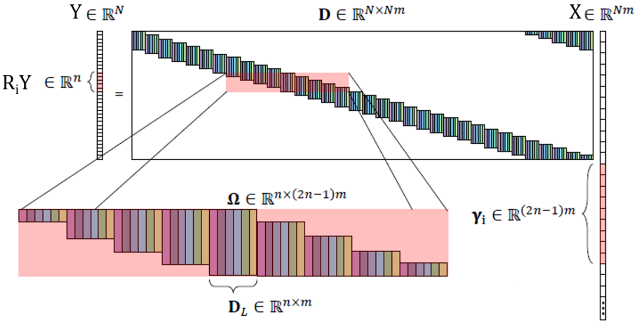
- This is very similar with a one layer hidden neural network!!

Outline

1. Sparse Coding
2. Convolutional Sparse Coding (CSC)
3. Connection between CSC and CNN
4. Towards to Understand ResNet and MSDNet

Convolutional Sparse Coding (CSC)

- Sparse coding suffers from the curse of dimensionality.
- Solution 1: train a local model for patches extracted from Y and process them independently.
- Solution 2: adopt convolutional dictionary built from shifted versions of a local matrix D_L [Sulam and Elad, 2015].



Convolution Sparse Coding (CSC)

- Why convolutional dictionary?
- Convolutional model can train the local patches naturally.
  • Assume the patch size is n, R_i is a extract operator, $a_i = R_i Y \in \mathbb{R}^n$ is a local patch extracted from Y and begin at the i-th entry of Y.
  • For convolution model, $a_i = R_i Y = R_i DX = \Omega\gamma_i$, γ_i is the corresponding patches in X.
  • Convolutional dictionary decrease the parameters significantly.

Convolution Sparse Coding (CSC)

- Why convolutional dictionary?

- Convolutional model can train the local patches naturally.
  • Assume the patch size is n, R_i is a extract operator, $a_i = R_i Y \in \mathbb{R}^n$ is a local patch extracted from Y and begin at the i-th entry of Y.
  • For convolution model, $a_i = R_i Y = R_i DX = \Omega\gamma_i$, γ_i is the corresponding patches in X.
  • Convolutional dictionary decrease the parameters significantly.

- Advantage: For a large value of m, $\mu(D) \approx \frac{1}{\sqrt{2n}}$. Classical sparse coding results would allow merely $O(\sqrt{n})$ non-zeros in all X while convolution model allow $O(\sqrt{n})$ non-zeros in n-length patches.

Convolution Sparse Coding (CSC)

- Convolutional model has a better theoretical guarantee.

Definition 3

Define the pseudo-norm $\mathcal{L}_{0,\infty}$ of a global sparse vector X as:

$$\|X\|_{0,\infty} = \max_i \|\gamma_i\|_0$$

Theorem 4

Given the system of linear equations $Y = DX$, if a solution X exists satisfying

$$\|X\|_{0,\infty} < \frac{1}{2}\left(1 + \frac{1}{\mu(D)}\right)$$

then BP and OMP is guaranteed to recover it [Papyan et al., 2017b].

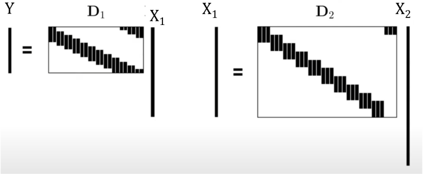
Multi-Layer CSC

- Double sparsity attempts to benefit from both the computational efficiency of analytically defined matrices and the adaptability of data driven dictionaries (Rubinstein et al., 2010).

$$Y = D_1 D_2 X_2$$

Here D_1 is an analytic dictionary and D_2 is a trained sparse one.

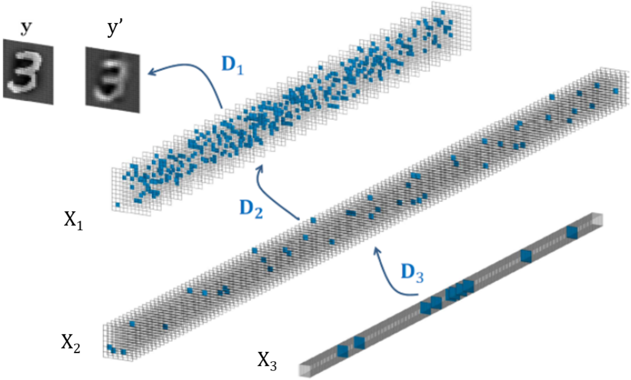
- Since both D_2 and X_2 are sparse, we expect $X_1 = D_2 X_2$ is sparse.



- In CSC, further regard the representation X_1 as a signal and learn its sparse representation X_2 [Papyan et al., 2017a].

$$Y = D_1 X_1, \quad X_1 = D_2 X_2$$

Multi-Layer CSC



- Intuitively, $Y = D_1X_1$ assumes that the signal Y is a superposition of atoms taken from D_1. While $Y = D_1D_2X_2$ views the signal as a superposition of more complex entities (molecules) taken from D_1D_2.

- Clearly, the construction can be extended to more than two layers.

Definition 5

For a global signal Y, a set of convolutional dictionaries $\{D_i\}_{i=1}^K$, and a vector λ, define the deep coding problem DCP_λ as:

$$(DCP_\lambda): \quad \text{find} \quad \{X_i\}_{i=1}^K \qquad \text{s.t.} \quad \begin{aligned} Y &= D_1 X_1, & \|X_1\|_{0,\infty} &\leqslant \lambda_1 \\ X_1 &= D_2 X_2, & \|X_2\|_{0,\infty} &\leqslant \lambda_2 \\ & & &\vdots \\ X_{K-1} &= D_K X_K, & \|X_K\|_{0,\infty} &\leqslant \lambda_K \end{aligned}$$

where the scalar λ_i is the i-th entry of λ.

Multi-Layer CSC

- The DCP_λ problem can be extended to a noisy regime.

Definition 6

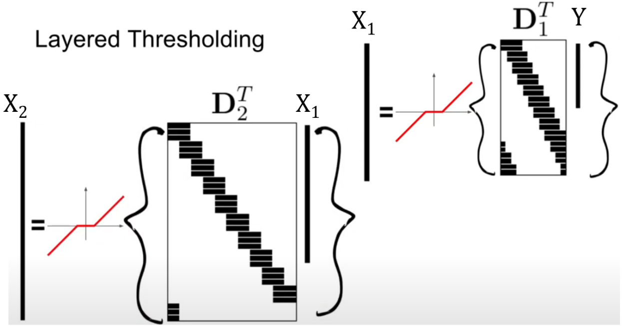
For a global signal Y, a set of convolutional dictionaries $\{D_i\}_{i=1}^K$, and a vector λ and ϵ, define the deep coding problem DCP_λ^ϵ as:

$$(DCP_\lambda^\epsilon) : \text{find} \quad \{X_i\}_{i=1}^K \quad \text{s.t.} \quad \begin{aligned} \|Y - D_1 X_1\|_2 \leqslant \epsilon_0, &\quad \|X_1\|_{0,\infty} \leqslant \lambda_1 \\ \|X_1 - D_2 X_2\|_2 \leqslant \epsilon_1, &\quad \|X_2\|_{0,\infty} \leqslant \lambda_2 \\ &\quad \vdots \\ \|X_{K-1} - D_K X_K\|_2 \leqslant \epsilon_{K-1}, &\quad \|X_K\|_{0,\infty} \leqslant \lambda_K \end{aligned}$$

where the scalar λ_i and ϵ_i is the i-th entry of λ and ϵ.

- For DCP_λ problem, we can use the layered thresholding method

$$X_i = S_{\beta_i}(D_i^T X_{i-1})$$



Theoretical Guarantee

Theorem 7

Suppose a signal Z has a decomposition $Z = D_1 X_1, \ldots, X_{K-1} = D_K X_K$ and that it is contaminated with noise E to create the signal $Y = Z + E$, such that $\|E\|_{0,\infty} \leqslant \epsilon_0$. Denote by $|X_i^{min}|$ and $|X_i^{max}|$ the lowest and highest entries in absolute value in the vector X_i, respectively. Let $\{X_i'\}_{i=1}^K$ be the set of solutions obtained by running the layered soft thresholding algorithm with thresholds $\{\beta_i\}_{i=1}^K$,i.e.$X_i' = S_{\beta_i}(D_i^T X_{i-1}')$ where $X_0' = Y$. Assuming that $\forall 1 \leqslant i \leqslant K$

a. $\|X_i\|_{0,\infty} < \frac{1}{2}\left(1 + \frac{1}{\mu(D_i)} \frac{|X_i^{\min}|}{|X_i^{\max}|}\right) - \frac{1}{\mu(D_i)} \frac{\epsilon_{i-1}}{|X_i^{\max}|}$

b. The threshold β_i is chosen according to

$$\left|X_i^{\min}\right| - \left(\|X_i\|_{0,\infty} - 1\right) \mu\left(D_i\right) |X_i^{\max}| - \epsilon_{i-1} > \beta_i > \|X_i\|_{0,\infty} \mu\left(D_i\right) |X_i^{\max}| + \epsilon_{i-1}$$

then 1. The support of the solution X_i' is equal to that of X_i;

2. $\|X_i' - X_i\|_{2,\infty} \leqslant \epsilon_i$,

where $\epsilon_i = \sqrt{\|X_i\|^{\mathrm{P}}_{0,\infty}} \left(\epsilon_{i-1} + \mu\left(D_i\right)\left(\|X_i\|_{0,\infty} - 1\right)\left|X_i^{\max}\right| + \beta_i\right)$

Layered ISTA

- For DCP_λ problem, we can also use the layered ISTA

$$x_{k+1}^{l} = S_{\frac{\lambda}{L}}\left(x_k^l - \frac{1}{L}(D^l)^T(D^l x_k^l - x^{l-1})\right)$$

Theorem 8

For DCP_λ problem, the layered ISTA is guaranteed to recover the true representation $\{X_i\}$, if $\forall 1 \leqslant i \leqslant K$

$$\|X_i\|_{0,\infty} < \frac{1}{2}\left(1 + \frac{1}{\mu(D_i)}\right)$$

Outline

1 Sparse Coding

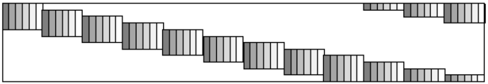
2 Convolutional Sparse Coding (CSC)

3 Connection between CSC and CNN

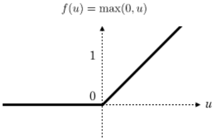
4 Towards to Understand ResNet and MSDNet

Convolution Neural Network (CNN)

- Convolution operator can be expressed as a convolutional matrix multipier.



- ReLU is the commonly used nonlinear activation:



- The output of the i-th layer is

$$X_i = \text{ReLU}(W_i X_{i-1} + b_i)$$

Connections of ML-CSC and NN

- Convolutional Sparse Coding (CSC) [Zeiler et al., 2011]
  - Why Convolutional? Local interactions!
  - Dictionary can be learned via local processing
- Multi-Layered CSC (ML-CSC) [Papyan et al., 2017a]
  - Why Deep? Learn more complex filters!
  - Related closely with CNN
  - Sparse dictionaries assumption

Connections of ML-CSC and NN

- Convolutional Sparse Coding (CSC) [Zeiler et al., 2011]
  - Why Convolutional? Local interactions!
  - Dictionary can be learned via local processing
- Multi-Layered CSC (ML-CSC) [Papyan et al., 2017a]
  - Why Deep? Learn more complex filters!
  - Related closely with CNN
  - Sparse dictionaries assumption
- In CSC, using layered threshold algorithm, the update of X_i is:

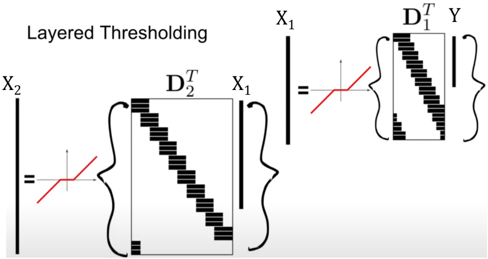
$$X_i = S_{\beta_i}(D_i^T X_{i-1})$$

where the thresholding operator $S_{\beta_i}(\cdot)$ is very similar to ReLU$(\cdot)$

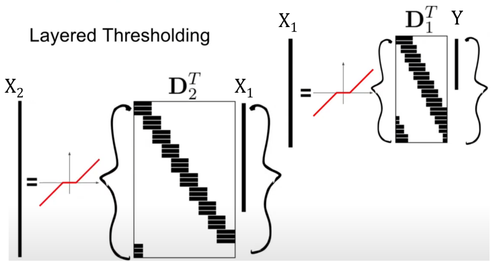
- It is trival that the update form of X_i is same to the the update of features X in the forward propagation of CNN

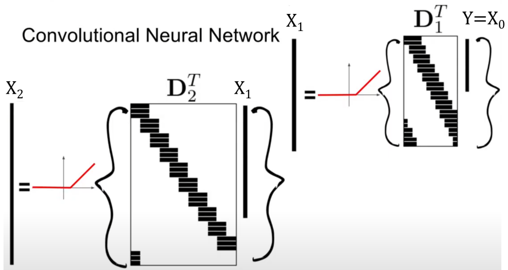
$$X_i = \mathrm{ReLU}(W_i X_{i-1} + b_i)$$

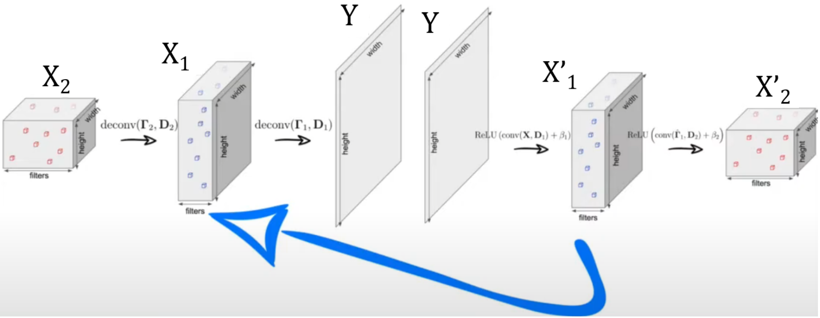
Connection of ML-CSC and CNN



Connection of ML-CSC and CNN







X2
X1
Y
Y
X'1
X'2
deconv(Γ2, D2)
deconv(Γ1, D1)
ReLU (conv(X, D1) + β1)
ReLU (conv(Γ̂1, D2) + β2)
filters
height
width

Success of Forward Pass

- If $\|X_i\|_{0,\infty} < \frac{1}{2}\left(1 + \frac{1}{\mu(D_i)} \frac{|X_i^{\min}|}{|X_i^{\max}|}\right) - \frac{1}{\mu(D_i)} \frac{\epsilon_{i-1}}{|X_i^{\max}|}$
  Layered thresholding guarantees:
  • Find correct places of nonzeros.
  • $\|X_i' - X_i\|_{2,\infty} \leqslant \epsilon_i$,
  where $\epsilon_i = \sqrt{\|X_i\|_{0,\infty}^{\mathrm{P}}}\left(\epsilon_{i-1} + \mu(D_i)\left(\|X_i\|_{0,\infty} - 1\right)|X_i^{\max}| + \beta_i\right)$

- Limits:
  ⋆ Forward pass always fail at recovering representations exactly.
  ⋆ Success depends on ratio.
  ⋆ Distance increases with layer.

Another view of connection

- In ISTA, the code is updated as follows:

$$x_{k+1} = S_{\frac{\lambda}{L}} \left(k_i - \frac{1}{L} D^T (Dx_k - y) \right)$$

- Let the initial code $x_0 = 0$. Then we have

$$x_1 = S_{\frac{\lambda}{L}} \left(\frac{1}{L} D^T y \right)$$

- Multi-Layer CSC with initial code $x_0^i = 0$

$$x^{i+1} = S_{\frac{\lambda}{L}} \left(\frac{1}{L} (D_i)^T x^i \right)$$

Deep CNN:

$$X^{l+1} = \mathrm{ReLU}((W^l)^T X^l + b^l)$$

Success of Layered ISTA

- If $\|X_i\|_{0,\infty} < \frac{1}{3}\left(1 + \frac{1}{\mu(D_i)}\right)$
  Layered ISTA guarantees:
  • Find only correct places of nonzeros.
  • Find all coefficients that are big enough.
  • $\|X_i' - X_i\|_{2,\infty} \leqslant \epsilon_i$,
  where $\epsilon_i = \|E\|_{2,\infty}^{\mathrm{P}} 7.5^i \prod_{j=1}^{i} \sqrt{\|X_j\|_{0,\infty}^{\mathrm{P}}}$

- Limits:
  ⋆ Distance increases with layer.

Conclusion

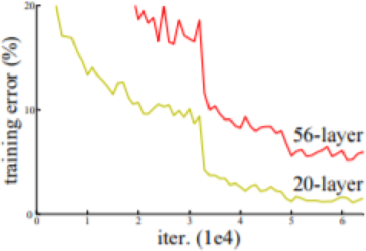
- Sparsity was well established theoretically.
- Sparsity is covertly exploited in practice: ReLU, dropout, stride, dilation...
- Sparsity is the secret sauce behind CNN.
- Need to bring sparsity to the surface to better understand CNNs.

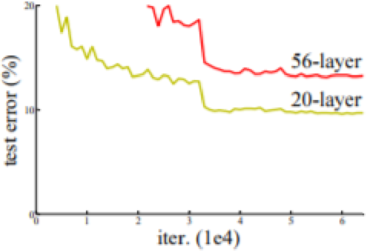
Outline

1. Sparse Coding
2. Convolutional Sparse Coding (CSC)
3. Connection between CSC and CNN
4. Towards to Understand ResNet and MSDNet

Residual Neural Network (ResNet)

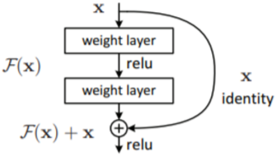
- For plain networks, there are two serious problems [He et al., 2016]:
  • Vanishing gradients
  • Degradation problem: with the network depth increasing, accuracy gets saturated and then degrades rapidly.





Residual Neural Network (ResNet)

- To avoid these problems, skip connections were introduced.

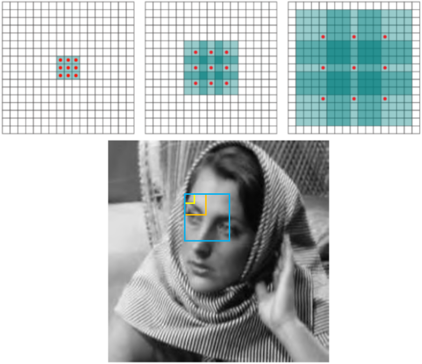


- The output of the i-th layer is

$$X_i = \sigma(W_{i-1,i}X_{i-1} + b_i + W_{i-2,i}X_{i-2})$$

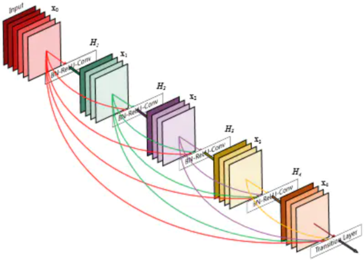
Mixed-Scale Dense CNN (MSDNet)

- MSDNet is another model of deep learning.
- Two special structures: dialated convolution and dense connection.
- Dialated convolution: it can capture the features in different scale with the same amount of parameters [Pelt and Sethian, 2018].



Mixed-Scale Dense CNN (MSDNet)

- Dense connection: the iuput of current layer is the concatenation of the output of all the previous layers.



- Dense connection can maximize the utilization of data and features captured by the shallow layers.

Towards to Understand Skip-Connection DNN

Can we generalize ML-CSC for those advanced NNs?

ResNet, DenseNet, MSDNet, ...

Towards to Understand Skip-Connection DNN

Can we generalize ML-CSC for those advanced NNs?
ResNet, DenseNet, MSDNet, ...

Three factors in each layer of ML-CSC
affect their performance [Zhang and Zhang, 2021]

- The initialization (Res-CSC)
- The dictionary design (MSD-CSC)
- The number of iterations (Optimization)

Res-CSC

Here we denote X as the signal and Γ as the sparse code. ISTA update:

$$\Gamma^{k+1} = S_{\frac{\beta}{L}}\left(\Gamma^k - \frac{1}{L}\left(-D^T X + D^T D \Gamma^k\right)\right) \tag{1}$$

Its first step when set $\Gamma^0 = 0$

$$\Gamma^1 = S_{\frac{\beta}{L}}\left(\frac{1}{L}\left(D^T X\right)\right) \tag{2}$$

Res-CSC

Here we denote X as the signal and Γ as the sparse code. ISTA update:

$$\Gamma^{k+1} = S_{\frac{\beta}{L}}\left(\Gamma^k - \frac{1}{L}\left(-D^T X + D^T D \Gamma^k\right)\right) \tag{1}$$

Its first step when set $\Gamma^0 = 0$

$$\Gamma^1 = S_{\frac{\beta}{L}}\left(\frac{1}{L}\left(D^T X\right)\right) \tag{2}$$

Layer-Initialization is the key

$$\Gamma^1 = S_{\frac{\beta}{L}}\left(\frac{1}{L} D^T X + X_{-1} - \frac{1}{L} D^T D X_{-1}\right) \tag{3}$$

Res-CSC

Layer-Initialization is the key!

$$\Gamma^1 = S_{\frac{\beta}{L}}\left(\frac{1}{L}D^T X + X_{-1} - \frac{1}{L}D^T D X_{-1}\right) \tag{4}$$

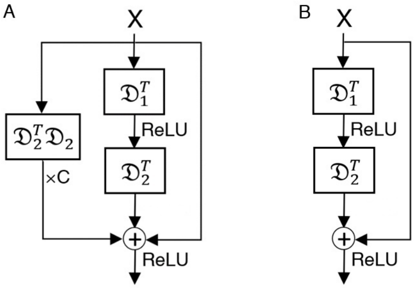


Figure: Res-CSC (A) and its variant (B)

Matrix-vector Multiplication Form

B

$$\begin{array}{|c|c|}\hline 1 & 2 \\ \hline 3 & 4 \\ \hline \end{array} \otimes \begin{array}{|c|c|c|c|}\hline 1 & 3 & 5 & 4 \\ \hline 7 & 9 & 2 & 6 \\ \hline 4 & 6 & 8 & 3 \\ \hline 1 & 7 & 2 & 5 \\ \hline \end{array}$$

$$= \begin{bmatrix} 1&2&0&0&3&4&0&0&0&0&0&0&0&0&0&0 \\ 0&1&2&0&0&3&4&0&0&0&0&0&0&0&0&0 \\ 0&0&1&2&0&0&3&4&0&0&0&0&0&0&0&0 \\ 0&0&0&0&1&2&0&0&3&4&0&0&0&0&0&0 \\ 0&0&0&0&0&1&2&0&0&3&4&0&0&0&0&0 \\ 0&0&0&0&0&0&1&2&0&0&3&4&0&0&0&0 \\ 0&0&0&0&0&0&0&0&1&2&0&0&3&4&0&0 \\ 0&0&0&0&0&0&0&0&0&1&2&0&0&3&4&0 \\ 0&0&0&0&0&0&0&0&0&0&1&2&0&0&3&4 \end{bmatrix} \cdot \begin{bmatrix} 1\\3\\5\\4\\7\\9\\2\\6\\4\\6\\8\\3\\1\\7\\2\\5 \end{bmatrix}$$

Figure: Dilation convolution ($s = 1$)

Matrix-vector Multiplication Form

C

$$\begin{array}{|c|c|c|}\hline 1 & 0 & 2 \\ \hline 0 & 0 & 0 \\ \hline 3 & 0 & 4 \\ \hline \end{array} \otimes \begin{array}{|c|c|c|c|}\hline 1 & 3 & 5 & 4 \\ \hline 7 & 9 & 2 & 6 \\ \hline 4 & 6 & 8 & 3 \\ \hline 1 & 7 & 2 & 5 \\ \hline \end{array}$$

$$= \begin{bmatrix} 1 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 4 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 3 \\ 5 \\ 4 \\ 7 \\ 9 \\ 2 \\ 6 \\ 4 \\ 6 \\ 8 \\ 3 \\ 1 \\ 7 \\ 2 \\ 5 \end{bmatrix}$$

Figure: Dilation convolution ($s = 2$)

Sparse Coding Scheme of MSD-CSC

Dictionary design $D_i^{s_i} = \begin{bmatrix} \mathrm{I} & \left(F_i^{s_i}\right)^T \end{bmatrix}$

$$\begin{bmatrix} \Gamma_{i-1}^{(1)} \\ \Gamma_{i-1}^{(2)} \\ \Gamma_{i-1}^{(3)} \\ \vdots \\ \Gamma_{i-1}^{(j)} \\ \vdots \\ \Gamma_{i-1}^{(n)} \end{bmatrix} \approx \left[\begin{array}{ccccccc|c} 1 & 0 & 0 & \cdots & 0 & 0 & 0 & \\ 0 & 1 & 0 & \cdots & 0 & 0 & 0 & \\ 0 & 0 & 1 & \cdots & 0 & 0 & 0 & \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \left(F_i^{s_i}\right)^T \\ 0 & 0 & 0 & \cdots & 1 & 0 & 0 & \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 & \\ 0 & 0 & 0 & \cdots & 0 & 0 & 1 & \end{array}\right] \cdot \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ \Gamma_{i-1}^{(j)} - \xi_j \\ 0 \\ \vdots \\ 0 \\ \hline \Gamma_i^{(1)} \\ \Gamma_i^{(2)} \\ \Gamma_i^{(3)} \\ \vdots \\ \Gamma_i^{(j)} \\ \vdots \\ \Gamma_i^{(n)} \end{bmatrix}$$

Sparse Coding Scheme of MSD-CSC

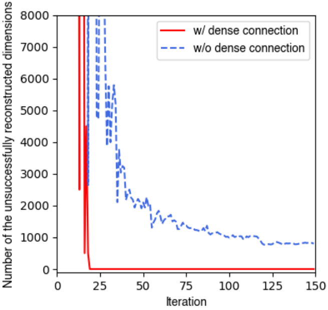
Dictionary design $D_i^{s_i} = \begin{bmatrix} \mathrm{I} & \left(F_i^{s_i}\right)^T \end{bmatrix}$

$$\begin{bmatrix} \Gamma_{i-1}^{(1)} \\ \Gamma_{i-1}^{(2)} \\ \Gamma_{i-1}^{(3)} \\ \vdots \\ \Gamma_{i-1}^{(j)} \\ \vdots \\ \Gamma_{i-1}^{(n)} \end{bmatrix} \approx \left[\begin{array}{ccccccc|c} 1 & 0 & 0 & \cdots & 0 & 0 & 0 & \\ 0 & 1 & 0 & \cdots & 0 & 0 & 0 & \\ 0 & 0 & 1 & \cdots & 0 & 0 & 0 & \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \left(F_i^{s_i}\right)^T \\ 0 & 0 & 0 & \cdots & 1 & 0 & 0 & \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 & \\ 0 & 0 & 0 & \cdots & 0 & 0 & 1 & \end{array}\right] \cdot \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ \Gamma_{i-1}^{(j)} - \xi_j \\ 0 \\ \vdots \\ 0 \\ \hline \Gamma_i^{(1)} \\ \Gamma_i^{(2)} \\ \Gamma_i^{(3)} \\ \vdots \\ \Gamma_i^{(j)} \\ \vdots \\ \Gamma_i^{(n)} \end{bmatrix}$$

Proposition 1: For a given MSDNet, there exists a MSD-CSC model, which is equivalent to MSDNet when propagates with the layered thresholding algorithm.

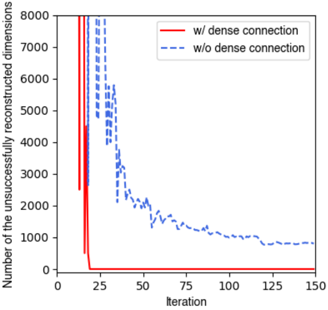
MSD-CSC and ML-CSC

Simulation study demonstrates that MSD-CSC
shows better reconstruction ability than ML-CSC



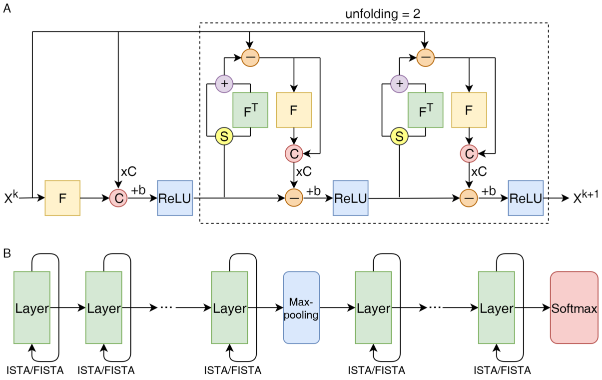
MSD-CSC and ML-CSC

Simulation study demonstrates that MSD-CSC
shows better reconstruction ability than ML-CSC



Theorem 1: For the Lasso problem in each layer, the performance of MSD-CSC is better than that of ML-CSC.

Unfold the iteration for MSD-CSC



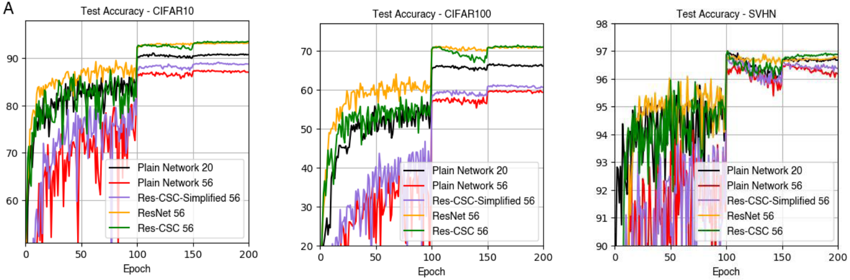
Comparison of CNN vs CSC

Table 1. The relationship between the generalized CNN and generalized CSC model.

CNN	CSC
The ith convolution with dilation scale s_i	The convolutional matrix $D_i^{s_i}$
Bias term	The balance coefficient β and $\lambda_{\max}(D^\top D)$
ReLU	Soft non-negative thresholding operator $S_\beta^+(\cdot)$
Feed-forward algorithm	$\Gamma^0 = 0$ in the update formula and iterate once
ResNet	$\Gamma^0 = X_{-1}$ in the update formula and iterate Equations (3) and (5) alternately
Dense connection	The identity matrix in $D_i^{s_i}$

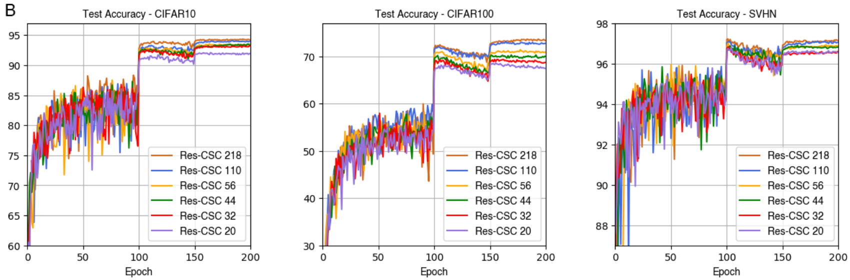
Performance of Res-CSC

Res-CSC indeed show equivalent performance



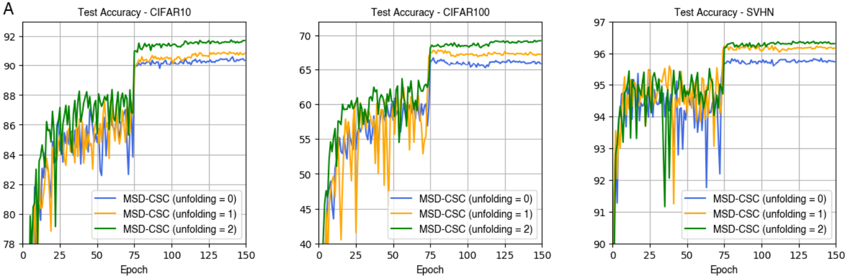
Performance of Res-CSC

Res-CSC can alleviate the degradation phenomenon



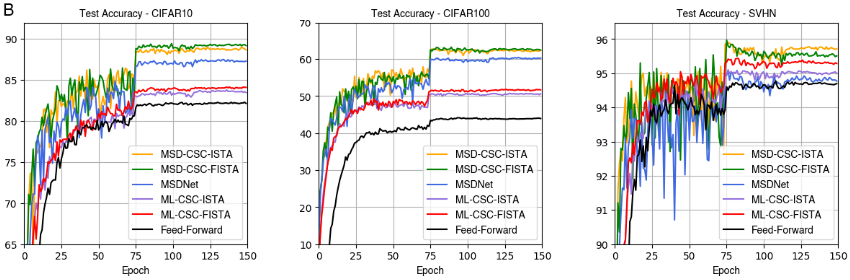
Performance of MSD-CSC

Unfolding indeed improve the performance of MSD-CSC



Performance of MSD-CSC

MSD-CSC shows better performance than MSDNet



Summary

- CNN lead to remarkable results in many fields.
- ResNet and MSDNet have even more superior performance.
  - Clear and profound theoretical understanding is still lacking.

- Sparse coding is a powerful model
  - Enjoys from a vast theoretical study, supporting its success.
  - CSC and ML-CSC have been proposed recently.

Summary

- CNN lead to remarkable results in many fields.
- ResNet and MSDNet have even more superior performance.
  - Clear and profound theoretical understanding is still lacking.

- Sparse coding is a powerful model
  - Enjoys from a vast theoretical study, supporting its success.
  - CSC and ML-CSC have been proposed recently.

- Res-CSC and MSD-CSC have been proposed here!!
  - Res-CSC/MSD-CSC can be equivalent with ResNet/MSDNet.
  - All Residual, Dilation and Dense operations can be explained.
  - Optimization in each layer can be improved with unfolding.

References I

Daubechies, I., Defrise, M., and De Mol, C. (2004).
An iterative thresholding algorithm for linear inverse problems with a sparsity constraint.
Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences, 57(11):1413–1457.

Donoho, D. L., Elad, M., and Temlyakov, V. N. (2005).
Stable recovery of sparse overcomplete representations in the presence of noise.
IEEE Transactions on information theory, 52(1):6–18.

He, K., Zhang, X., Ren, S., and Sun, J. (2016).
Deep residual learning for image recognition.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Olshausen, B. A. and Field, D. J. (1996).
Emergence of simple-cell receptive field properties by learning a sparse code for natural images.
Nature, 381(6583):607–609.

Papyan, V., Romano, Y., and Elad, M. (2017a).
Convolutional neural networks analyzed via convolutional sparse coding.
The Journal of Machine Learning Research, 18(1):2887–2938.

Papyan, V., Sulam, J., and Elad, M. (2017b).
Working locally thinking globally: Theoretical guarantees for convolutional sparse coding.
IEEE Transactions on Signal Processing, 65(21):5687–5701.

Pelt, D. M. and Sethian, J. A. (2018).
A mixed-scale dense convolutional neural network for image analysis.
Proceedings of the National Academy of Sciences, 115(2):254–259.

References II

Sulam, J. and Elad, M. (2015).
Expected patch log likelihood with a sparse prior.
In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 99–111. Springer.

Zeiler, M., Krishnan, D., Taylor, G., and Fergus, R. (2011).
Deconvolutional networks for feature learning.
In *Comput. Vis. Pattern Recognit.(CVPR), 2010 IEEE Conf*, pages 2528–2535. Citeseer.

Zhang, Z. and Zhang, S. (2021).
Towards understanding residual and dilated dense neural networks via convolutional sparse coding.
National Science Review, 8(3):nwaa159.