Relationship between Convolutional Neural Networks and Convolutional Sparse Coding

José Rodríguez, Basanta Pahari

April 25, 2020

Goals

In this presentation :

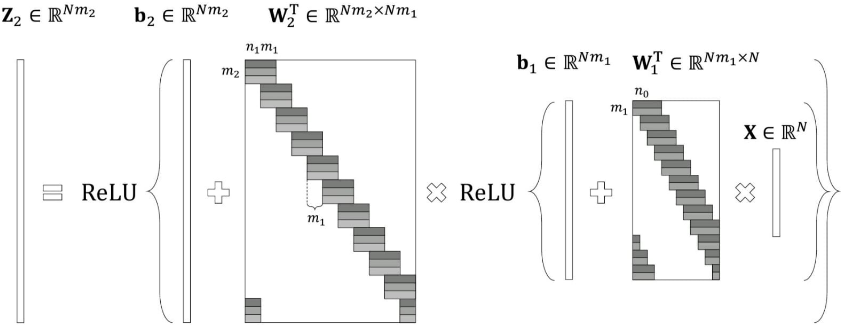
- Introduce Convolutional Neural Network (CNN) and Convolutional Sparse Coding (CSC).
- Present the relationship between the forward pass of a CNN and CSC.

Introduction

The forward pass of a CNN consists of

- Convolving an input signal $X \in \mathbb{R}^N$ with a collection of filters.
- Applying a point-wise non-linear function.
- Very frequently, a third operation pooling is applied. This is not analyzed in this work.

$$f(X, \{W_i\}_{i=1}^2, \{b_i\}_{i=1}^2) = Z_2 = ReLU\big(W_2^T(ReLU(W_1^T X + b_1) + b_2)\big)$$



where $W_1 \in \mathbb{R}^{N \times N\, m_1}$ $W_2 \in \mathbb{R}^{Nm_1 \times Nm_2}$ are convolutional matrices of m_1 filters of size n_0 and m_2 dilters of size n_1 respectively.

Sparse Representation

Given $X \in \mathbb{R}^N$, we want to write $X = D\Gamma$ where $D \in \mathbb{R}^{N \times M}$ and $\Gamma \in \mathbb{R}^M$.

We want to recover the sparsest representation.

- (P_0) $\min_\Gamma \|\Gamma\|_0 \; s.t. \; D\Gamma = X$
- (P_1) $\min_\Gamma \|\Gamma\|_1 \; s.t. \; D\Gamma = X$,

Here, $\|\Gamma\|_0$ counts the number of nonzero entries in Γ. P_1 is convex form of problem P_0.

Simple approaches to solving P_0 and P_1 include **hard and soft thresholding algorithms** respectively. These are equivalent to the following minimization problems:

- $\min_\Gamma \frac{1}{2}\|\Gamma - D^T X\|_2^2 + \beta\|\Gamma\|_0$ for the P_0

- $\min_\Gamma \frac{1}{2}\|\Gamma - D^T X\|_2^2 + \beta\|\Gamma\|_1$ for the P_1.

These problems admit a solution in the form $\mathbf{H}_\beta(D^T X)$ or $\mathbf{S}_\beta(D^T X)$ where

$$\mathbf{H}_\beta(z) = \begin{cases} z & z < -\beta \\ 0 & -\beta \leq z \leq \beta \\ z & \beta < z \end{cases}$$

$$\mathbf{S}_\beta(z) = \begin{cases} z + \beta & z < -\beta \\ 0 & -\beta \leq z \leq \beta \\ z - \beta & \beta < z \end{cases}$$

- $X = D\Gamma_P + (-D)(-\Gamma_N)$.
- Γ_p and $-\Gamma_N$ are all non-negative.
- X admits a non-negative sparse representation in $[D, -D]$ with the vector $[\Gamma_p, -\Gamma_N]^T$.]
- P_1 becomes $\min_\Gamma \frac{1}{2}\|\Gamma - D^T X\|_2^2 + \beta\|\Gamma\|_1 \; s.t. \Gamma \geq 0$.
- The solution becomes $\mathbf{S}_\beta^+(D^T X)$ where

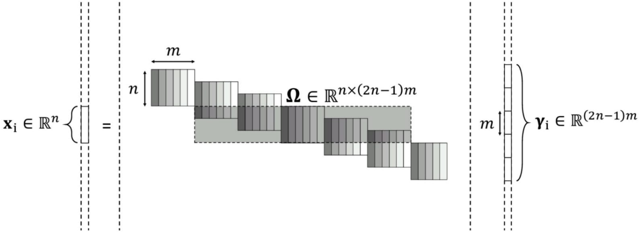
$$\mathbf{S}_\beta^+(z) = \max(z - \beta, 0) = ReLU(z - \beta).$$

Convolutional Sparse Coding

Dimension of $X \in \mathbb{R}^N$ may be too large, and the uniqueness of problem P_0 requires

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

where $\mu(D) := max_{i \neq j}|d_i^T d_j|$ is the mutual coherence of D.



So, measure sparsity of Γ in a localized way.

- $x_i = \Omega\gamma_i$ be $i-th$ $n-dimensional$ patch where Ω is extracted from the $i-th$ patch of D.
- γ_i contains all the coefficients of columns of Ω that contributes to x_i.
- $\|\Gamma\|_{0,\infty}^{\mathbf{s}} := \max_i \|\gamma_i\|_0$
- $(P_{0,\infty})$ $\min_\Gamma \|\Gamma\|_{0,\infty}^{\mathbf{s}}$ $s.t.$ $D\Gamma = X$.
- $(P_{0,\infty}^{\epsilon})$ $\min_\Gamma \|\Gamma\|_{0,\infty}^{\mathbf{s}}$ $s.t.$ $\|Y - D\Gamma\|_2^2 \leq \epsilon^2$ in case $Y = X + E$.

Definition

For a measured signal $Y = X + E$ and a set of convolutional dictionaries $\{D_i\}_{i=1}^K$ and vectors λ and ϵ, define the **deep coding problem** DCP_γ^ϵ as:

$$(DCP_\gamma^\epsilon): \qquad \begin{array}{l} \textit{find} \quad \{\Gamma_i\}_{i=1}^K \; s.t. \quad \|\Gamma_{i-1} - D_i\Gamma_i\| \leq \epsilon_{i-1}, \; \|\Gamma_i\|_{0,\infty}^{\mathbf{s}} \leq \lambda_i \\ \textit{for} \; i = 0, \ldots, K, \quad \Gamma_0 = Y \end{array}$$

Definition

For a set of global signals $\{X_j\}_j$ and their corresponding labels $\{h(X_j)\}_j$, a loss function ℓ and a vector λ, we define the **deep learning problem** DLP_λ as:

$$(DLP_\lambda): \min_{\{D_i\}_{i=1}^K, U} \sum_j \ell\left(h(X_j), U, DCP_\lambda^*(X, \{D_i\}_{i=1}^K)\right)$$

Layered Thresholding

$X = D_1\Gamma_1$, $\Gamma_1 = D_2\Gamma_2$ and so on. To solve this we have the following algorithm.

- Input: X, $\{D_i\}_{i=1}^K$. $\mathbf{P} \in \{\mathbf{H}, \mathbf{S}, \mathbf{S}^+\}$ (thresholding operator), $\{\beta_i\}_{i=1}^K$ (threshold).
- Process: $\hat{\Gamma}_0 \leftarrow X$, for $(i = 1 : K)$ do $\hat{\Gamma}_i \leftarrow \mathbf{P}_{\beta_i}(D_i^T \hat{\Gamma}_{i-1})$.
- Output: X, $\{\hat{\Gamma}_i\}_{i=1}^K$

For CSC:

$$\hat{\Gamma}_2 = \mathbf{P}_{\beta_2}\left(D_2^T \mathbf{P}_{\beta_1}(D_1^T X)\right).$$

For CNN:

$$f(X, \{W_i\}_{i=1}^2, \{b_i\}_{i=1}^2) = ReLU\big(W_2^T(ReLU(W_1^T X + b_1) + b_2)\big).$$

Theoretical Results

Theorem

(Uniqueness via mutual coherence): For a signal X satisfying the DCP_λ model, $\Gamma_{i-1} = D_i \Gamma_i$ where $\{D_i\}_{i=1}^K$ is a set of convolutional dictionaries and $\{\mu(D_i)\}_{i=1}^K$ are their corresponding mutual coherences. If for all $1 \leq i \leq K$ we have $\|\Gamma_i\|_{0,\infty}^{s} \leq \lambda_i < \frac{1}{2}\left(1 + \frac{1}{\mu(D_i)}\right)$, then $\{\Gamma_i\}_{i=1}^K$ is the unique solution to the DCP_λ.

Theorem

(Stability of the solution to the DCP_λ^ϵ problem): Suppose a signal X has a decomposition $\Gamma_{i-1} = D_i \Gamma_i$ and it is contaminated with noise E so that $Y = X + E$. Assume that we solve DCP_λ^ϵ problem for $\epsilon_0 = \|E\|_2$ and $\epsilon_i = 0$ for $i \geq 1$ obtaining solutions $\{\hat{\Gamma}_i\}_{i=1}^K$. If for all $1 \leq i \leq K$ $\|\Gamma_i\|_{0,\infty}^{s} \leq \lambda_i < \frac{1}{2}\left(1 + \frac{1}{\mu(D_i)}\right)$, then $\|\Gamma_i - \hat{\Gamma}_i\|_2^2 \leq 4\|E\|_2^2 \prod_{j=1}^{i} \frac{1}{1-(2\lambda_j - 1)\mu(D_j)}$.

Theorem

(Stability of the forward(layered soft thresholding) pass in the presence of noise): Let X *have a decomposition* $\Gamma_{i-1} = D_i \Gamma_i$ *and denote by* $Y = X + E$ *the corrupted signal with* $\|E\|_{2,\infty}^{P} \leq \epsilon_0$*. Denote by* $|\Gamma_i^{\min}|$ *and* $|\Gamma_i^{\max}|$ *the lowest and highest entries in absolute value of* Γ_i*, respectively. Let* $\hat{\Gamma}_i = \mathbf{S}_{\beta_1}(D_i^T \hat{\Gamma}_{i-1})$ *where* $\hat{\Gamma}_0 = Y$*. Assume that*

- $\|\Gamma_i\|_{0,\infty}^{S} < \frac{1}{2}\left(1 + \frac{1}{\mu(D_1)} \frac{|\Gamma_i^{\min}|}{|\Gamma_i^{\max}|}\right) - \frac{1}{\mu(D_1)} \frac{\epsilon_{i-1}}{|\Gamma_i^{\max}|}$ *and*
- $\|\Gamma_i\|_{0,\infty}^{S} \mu(D_i) |\Gamma_i^{\max}| + \epsilon_{i-1} < \beta_i < |\Gamma_i^{\min}| - \left(\|\Gamma_i\|_{0,\infty}^{S} - 1\right) \mu(D_i) |\Gamma_i^{\max}| - \epsilon_{i-1}$.

Then, the following holds

- *The support of* $\hat{\Gamma}_i$ *is equal to the support of* Γ_i *and*
- $\|\Gamma_i - \hat{\Gamma}_i\|_{2,\infty}^{P} \leq \sqrt{\|\Gamma_i\|_{0,\infty}^{P}} \left(\epsilon_{i-1} + \left(\|\Gamma_i\|_{0,\infty}^{S} - 1\right) \mu(D_i) |\Gamma_i^{\max}| + \beta_i\right)$.

References

- Convolutional Neural Networks Analyzed by Convolutional Sparse Coding, V. Papyan, Y. Romano, M. Elad., Journal of Machine Learning Research, 2017
- Sparse and Redundant Representations, M. Elad., Springer, 2010
- Working locally thinking globally: Theoretical guarantees for convolutional sparse coding., V. Papyan, J. SulamM. Elad., arXiv, 2016