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 Papyan @ Stats 385, Stanford

Based on work of:
Vardan Papyan, Jeremias Sulam, Yaniv Romano, Michael Elad
Sparsity: Central idea in Stats

Compressive Sensing:

\[ X = D\hat{\Gamma} \]

\[ \hat{\Gamma} = \arg \min_{\Gamma} \|\Gamma\|_1 \quad \text{s.t.} \quad X = D\Gamma \]
Sparsity: Central idea in Stats

Lasso:

\[ Y = D\Gamma + E \]

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

Yosinski et al., "Understanding Neural Networks through Deep Visualization", ICML DL Workshop 2015
Sparsity in Practice
Olshausen & Field and AlexNet

Olshausen & Field

AlexNet

explicit sparsity

implicit sparsity

Credit to: Vardan, Stats385@Stanford
Theory of sparsity in CNN?
Breiman’s “Two Cultures”

Generative modeling

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

face

eye

nose

edge

edge

nose

eye

face
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

Multi-Layered Convolutional Sparse Representation

First Layer of a Convolutional Neural Network

Convolutional Sparse representation

Sparse representations
First Layer of a Neural Network

\[ \hat{\Gamma} = D^T X \]
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

vec(·)
Sparse Coding

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

\[
\min_{\Gamma} \|\Gamma\|_0 \quad \text{s.t.} \quad X = D\Gamma
\]

Convexify

\[
\min_{\Gamma} \|\Gamma\|_1 \quad \text{s.t.} \quad X = D\Gamma
\]
Sparse Coding

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

\[
\min_{\Gamma} \| \Gamma \|_0 \quad \text{s.t.} \quad X = D\Gamma
\]

Convexify

\[
\min_{\Gamma} \| \Gamma \|_1 \quad \text{s.t.} \quad X = D\Gamma
\]

Crude approximation

\[
S_\beta\{D^T X\}
\]
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

\[ \mathbf{X} = \mathbf{D} \mathbf{\Gamma} \]
Convolutional Sparse Modeling

\[ X = \sum_{k=1}^{K} d_k \ast z_k \]
Convolutional Sparse Modeling

\[ X = \text{deconv} \]

\[ \Gamma \]

\[ \begin{array} {c}	ext{width} \\
\text{height} \\
\text{filters} \\
\end{array} \]

\[ D \]
Thresholding Algorithm

\[ \hat{\Gamma} \]

\[ X \]

\[ D^T \]
First layer of a Convolutional Neural Network
First layer of a Convolutional Neural Network
Convolutional Neural Network

\[
\text{ReLU} \left( \text{conv}(X, D_1) + \beta_1 \right) \rightarrow \text{ReLU} \left( \text{conv}(\hat{\Gamma}_1, D_2) + \beta_2 \right)
\]
Multi-layered Convolutional Sparse Modeling

$X = \Gamma_1 D_1 \Gamma_1 \quad \Gamma_1 = D_2 \Gamma_2$
Multi-layered Convolutional Sparse Modeling

\[ \text{deconv}(\Gamma_2, D_2) \rightarrow \Gamma_1 \rightarrow \text{deconv}(\Gamma_1, D_1) \rightarrow X \]
Layered Thresholding

\[ \hat{\Gamma}_2 \quad D_2^T \quad \hat{\Gamma}_1 \]

\[ \hat{\Gamma}_1 = \]

\[ D_1^T \quad X \]
Convolutional Neural Network
Convolutional Neural Network

\[
\text{ReLU} \left( \text{conv}(X, D_1) + \beta_1 \right) \rightarrow \text{ReLU} \left( \text{conv}(\hat{\Gamma}_1, D_2) + \beta_2 \right)
\]
Theories of Deep Learning
Evolution of Models

- Multi-Layered Convolutional Neural Network
- Multi-Layered Convolutional Sparse Representation
- First Layer of a Convolutional Neural Network
- First Layer of a Convolutional Sparse Representation
- Sparse representations
Sparse Modeling

\[ X \prod \Delta \Gamma \]
Classic Sparse Theory

\[ \mathbf{X} = \mathbf{D} \hat{\mathbf{\Gamma}} \]

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

**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: \[ \mu(D) = \max_{i \neq j} \left| (D^T D)_{i,j} \right| \]
Convolutional Sparse Modeling

\[ X = D \Gamma \]
Classic Sparse Theory for Convolutional Case

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

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

Assuming 2 atoms of length 64

$$\mu(D) \geq 0.063$$ [Welch, 1974]

Success guaranteed when

$$\| \Gamma \|_0 < 8.43$$
Local Sparsity

\[ \max_{\Gamma} \| \Gamma \|_{0, \infty} \text{ maximal number of non-zeroes in a local neighborhood} \]

\[ \min_{\Gamma} \| \Gamma \|_{0, \infty} \text{ s.t. } X = D \Gamma \]
Success of Basis Pursuit

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

**Theorem:** [Papyan, Sulam and Elad, 2016]

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

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

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:

$$
\mathbf{X} = \mathbf{D}_1 \Gamma_1, \quad \|\Gamma_1\|_{0,\infty} \leq \lambda_1
$$

$$
\Gamma_1 = \mathbf{D}_2 \Gamma_2, \quad \|\Gamma_2\|_{0,\infty} \leq \lambda_2
$$

$$
\vdots
$$

$$
\Gamma_{L-1} = \mathbf{D}_L \Gamma_L, \quad \|\Gamma_L\|_{0,\infty} \leq \lambda_L
$$
Deep Coding Problem

Given $Y$, find a set of representations satisfying:

$$
\|Y - D_1 \Gamma_1 \|_2 \leq \epsilon, \quad \|\Gamma_1\|_{0,\infty} \leq \lambda_1
$$

$$
\Gamma_1 = D_2 \Gamma_2, \quad \|\Gamma_2\|_{0,\infty} \leq \lambda_2
$$

$$
\vdots
$$

$$
\Gamma_{L-1} = D_L \Gamma_L, \quad \|\Gamma_L\|_{0,\infty} \leq \lambda_L
$$
Uniqueness

\[ \Gamma_2 \xrightarrow{\text{de conv}(\Gamma_2, D_2)} \Gamma_1 \xrightarrow{\text{de conv}(\Gamma_1, D_1)} X \]

\[ \hat{\Gamma}_2 \xrightarrow{\text{de conv}(\hat{\Gamma}_2, D_2)} \hat{\Gamma}_1 \xrightarrow{\text{de conv}(\hat{\Gamma}_1, D_1)} X \]
Uniqueness Theorem

\[ \| \Gamma_l \|_{0, \infty} \leq \lambda_l < \frac{1}{2} \left( 1 + \frac{1}{\mu(D_l)} \right) \]

\[ \{ \Gamma_l \}_{l=1}^{L} \] are the unique feature maps of \( X \)
Success of Forward Pass
Success of Forward Pass Theorem

\[ \|\Gamma_l\|_{0,\infty} < \frac{1}{2} \left( 1 + \frac{1}{\mu(D_l)} \frac{|\Gamma_{l\min}|}{|\Gamma_{l\max}|} \right) - \frac{1}{\mu(D_l)} \frac{\epsilon_{l-1}}{|\Gamma_{l\max}|} \]

Layered thresholding guaranteed:

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

- Forward pass always fails at recovering representations exactly
- Success depends on ratio
- Distance increases with layer
Generative Model and Crude Inference
\[
\hat{\Gamma}_1 = \arg \min_{\Gamma_1} \frac{1}{2} \| Y - D_1 \Gamma_1 \|_2^2 + \alpha_1 \| \Gamma_1 \|_1 \\
\hat{\Gamma}_2 = \arg \min_{\Gamma_2} \frac{1}{2} \| \hat{\Gamma}_1 - D_2 \Gamma_2 \|_2^2 + \alpha_2 \| \Gamma_2 \|_1
\]
Success of Layered Lasso

\[ \| \Gamma_l \|_{0,\infty} < \frac{1}{3} \left( 1 + \frac{1}{\mu(D_L)} \right) \]

Layered Lasso guaranteed:
1. Find only correct places of nonzeros
2. Find all coefficients that are big enough

\[ \| \hat{\Gamma}_l - \Gamma_l \|_{2,\infty} \leq \epsilon_l \]

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

\[ \Gamma_t^1 = S_{\alpha_1} \left( D_{1}^{T} Y + (I - D_{1}^{T} D_{1}) \Gamma_{1}^{t-1} \right) \]

\[ \Gamma_t^2 = S_{\alpha_2} \left( D_{2}^{T} \hat{\Gamma}_1 + (I - D_{2}^{T} D_{2}) \Gamma_{2}^{t-1} \right) \]
Multi-layered Convolutional Sparse Modeling

\[ \text{deconv}(\Gamma_2, D_2) \rightarrow \Gamma_1 \rightarrow \text{deconv}(\Gamma_1, D_1) \rightarrow X \]
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