Why Sparse Coding Works

Mathematical Challenges in Deep Learning

Shaowei Lin (UC Berkeley)

shaowei@math.berkeley.edu

10 Aug 2011

Deep Learning Kickoff Meeting

What is Sparse Coding?

- There are many formulations and algorithms for sparse coding in the literature. We isolate the basic mathematical idea in order to study its properties.
- Let $y \in \mathbb{R}^n$ be a data point in a data set Y. Assume that y can be expressed as

$$y = Aa$$

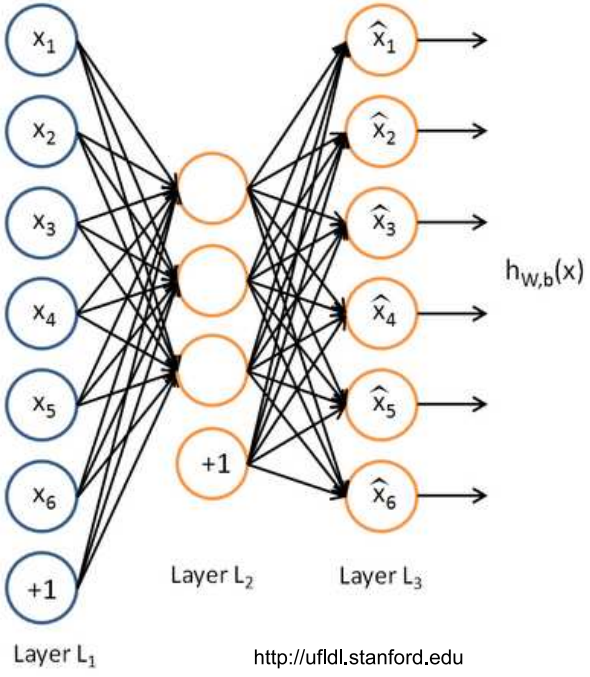
  where $A \in \mathbb{R}^{n \times m}$ is a *dictionary* matrix and $a \in \mathbb{R}^m$ is a sparse vector. (Ignore noise in the data y for now.)
- Assume a has at most k nonzero entries (*k-sparse*).
- Usually, the dictionary A is *overcomplete*: there are fewer rows than columns ($n < m$). The columns of A are the *atoms* of the dictionary.

What is Sparse Coding?

$$y = Aa$$

- e.g. y small image patch, A dictionary of features, a sparse representation of y.
- In *compressed sensing*, we attempt to recover a, assuming we know the dictionary A.
- In *sparse coding*, we attempt to find a dictionary A so that each $y \in Y$ has a sparse representation.
- When A is overcomplete, both problems are ill-posed without the sparsity condition.

Sparse Coding in Deep Learning



- Implement sparse coding on each layer of the autoencoder. Each layer is a Restricted Boltzmann Machine (RBM).

- Without sparse coding, the autoencoder learns a low-dim representation similar to Principal Component Analysis.

Sparse Coding in Deep Learning

- **Aim**: Find code map $b(y)$ and dictionary B such that

$$\hat{y} = Bb(y)$$

  is as close to y as possible.

- **Method**: Relax the problem and optimize:

$$\min_{B,b(y)} \left[\sum_{y} \| y - Bb(y) \|^2 + \beta S(b(y)) \right] + \lambda W(B)$$

  $S(b)$ sparsity penalty,
  $W(B)$ weight penalty to limit growth of B,
  penalties controlled by parameters β, λ.

Sparse Coding in Deep Learning

$$\min_{B,b(y)} \left[\sum_{y} \| y - Bb(y) \|_2 + \beta S(b(y)) \right] + \lambda W(B)$$

- Solve using gradient descent methods.
- Leads to local update rules for B such as Hebb's learning rule, Oja's learning rule, etc.

QN: ***Why does sparse coding work?***

What do we mean by *Why*?

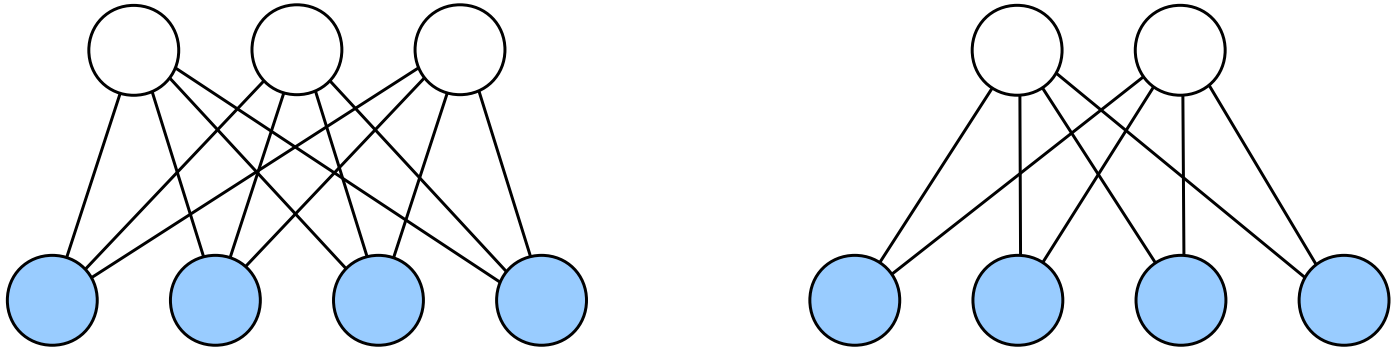
Two ways we want to understand sparse coding:

- *"Sparse coding only applies if the data actually have sparse structure." [Olshausen and Field, 1997]*

  How can we explain this rigorously using statistical learning theory and model selection?

- The brain evolved to extract sparse structure in data effectively and efficiently. How does the brain do this? We want to mimic what goes on in the brain for many real-world applications.

Sparse Structure in Data



In statistical learning theory, when the sample size is large and all given models attain the true distribution of the data, model selection can be accomplished using Watanabe's *Generalized Bayesian Information Criterion*:

$$N\mathcal{S} - \lambda \log N + (\theta - 1) \log \log N$$

N sample size, $\mathcal{S}$ entropy of true distribution, λ, θ *learning coefficient* and its multiplicity.

Sparse Structure in Data

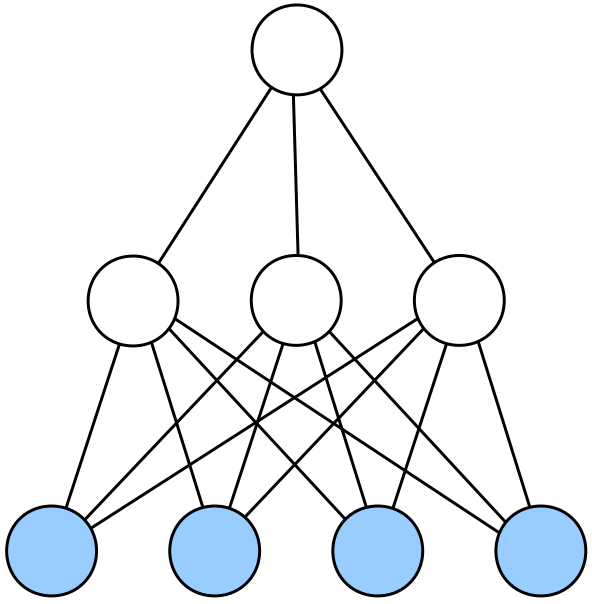
$$N\mathcal{S} - \lambda \log N + (\theta - 1) \log \log N$$

- Conventional BIC: $\lambda = \frac{1}{2}(\text{model dimension}), \theta = 1$.

  model dimension $\propto$ number of hidden factors

  This may explain why sparsity is a good candidate as a model selection criterion.
- See [Geiger and Rusakov, 2005] for computation of λ for binary RBMs with one hidden node.
- *What is the learning coefficient of a general RBM?*

Sparse Structure in Data



- *As hidden features start having dependencies among themselves, is sparsity still effective for model selection?*
- Perhaps we can discover new learning criteria from this exploration.

Uniqueness in Sparse Coding

- In principle, optimal recovery in sparse coding is NP-hard and not necessarily unique.
- But there are many suboptimal methods for approx a solution (e.g. optimization by gradient descent).
- In fact, given some mild assumptions about the dictionary and code sparsity, one can show that solutions to the sparse coding problem are unique. [Hillar and Sommer, 2011]
- We say a dictionary A is *incoherent* if

$$Aa_1 = Aa_2 \text{ for } k\text{-sparse } a_1, a_2 \in \mathbb{R}^m \;\Rightarrow\; a_1 = a_2$$

Uniqueness in Sparse Coding

- **Reconstructing representation a:**

  Given $A \in \mathbb{R}^{n \times m}$ and $y = Aa$ for some k-sparse a.
  If $n < Ck \log(m/k)$, then it is impossible to recover a from y. (C is some constant.)

- **Reconstructing dictionary A:**

  ACS Reconstruction Theorem
  Suppose $A \in \mathbb{R}^{n \times m}$ is incoherent.
  There exist k-sparse $a_1, \ldots, a_N \in \mathbb{R}^m$ such that:
  if $B \in \mathbb{R}^{n \times m}$, k-sparse $b_1, \ldots, b_N \in \mathbb{R}^m$ satisfy $Aa_i = Bb_i$,
  then $A = BPD$ for some permutation matrix $P \in \mathbb{R}^{m \times m}$
  and invertible diagonal matrix $D \in \mathbb{R}^{m \times m}$.
  [Hillar and Sommer, 2011]

Mimicking the Brain

- Proof of the ACS Reconstruction Theorem requires delicate combinatorial tools from Ramsey Theory.
- The theorem assures us that if there is sparse structure in the data, then this structure can be recovered if we can solve the sparse coding problem optimally.
- *Are there also mild assumptions on the suboptimal methods to ensure they find this optimal solution?*
- This may show us the extent to which learning in the brain depends on the formulation of the learning rule.
- *Can we implement these methods on the neuron level? Build parallel computers for deep learning?*

Summary

- Sparse coding and model selection
- Correct recovery in sparse coding
- Learning group symmetries

  How do our brains account for group symmetries in the data (rotations, translations, etc.)? Do these symmetries arise naturally from temporal coherence (i.e. videos instead of still images)?

References

1. A. Cueto, J. Morton and B. Sturmfels: Geometry of the restricted Boltzmann machine, in Algebraic Methods in Statistics and Probability, (eds. M. Viana and H. Wynn), AMS, *Contemporary Mathematics* **516** (2010) 135–153.
2. D. Geiger and D. Rusakov: Asymptotic model selection for naive Bayesian networks, *J. Mach. Learn. Res.* **6** (2005) 1–35.
3. S. Gleichman: Blind compressed sensing, preprint (2010).
4. C. J. Hillar and F. T. Sommer: Ramsey theory reveals the conditions when sparse dictionary learning on subsampled data is unique, preprint (2011).
5. S. Lin: Algebraic methods for evaluating integrals in Bayesian statistics, PhD dissertation, Dept. Mathematics, UC Berkeley (2011).
6. A. Ng, J. Ngiam, C. Y. Foo, Y. Mai, C. Suen: Unsupervised feature learning and deep learning tutorial: `ufldl.stanford.edu` (2011).
7. B. Olshausen and D. Field: Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research* **37** (1997) 3311–3325.
8. S. Watanabe: *Algebraic Geometry and Statistical Learning Theory*, Cambridge Monographs on Applied and Computational Mathematics **25** (2009).