Why are deep nets reversible: A simple theory, with implications for training

Yanyao Yi & Muxuan Liang

University of Wisconsin - Madison
Department of Statistics

yanyao.yi@wisc.edu

November 10, 2016

Overview

1. Introduction
   - Abstract
   - Deep nets are reversible?
   - Two Properites and One Assumption
   - Achievement–SHADOW Method
2. Single Layer Generative Model
   - Single Layer Generative Model
   - Theorem
3. Full Multilayer Model
   - Full Multilayer Model
   - Theorem
4. Experiments
   - Verify Random-like nets
   - Shadow method

Abstract

This paper presents a simple generative model for RELU deep nets, with the following characteristics:

i. The generative model is just the reverse of the feedforward net
ii. Its correctness can be proven under a clean theoretical assumption: Random-like nets hypothesis

Deep nets are reversible?

- Restricted Boltzmann Machine (RBM) :
  sample a **reconstruction** of the visible units
- Denoising Autoencoders:
  train the autoencoder to **reconstruct** the input from a corrupted version of it
- Deep Boltzmann Machine: it is related to RBM.
- ......

Two Properites and One Assumption

Let **x** denote the data/input to the deep net and **h** denote the hidden representation. The generative model has to satisfy the following two Properties:

- **Property(a):** Specify a joint distribution of x and h, or at least $p(x|h)$
- **Property(b):** A proof that the deep net itself is a method of computing $p(h|x)$

Two Properties and One Assumption

Random-like nets hypothesis:

which says that real-life deep nets - even those obtained from standard supervised learning – are "random -like", meaning their edge weights behave like random numbers.

Notice, this is distinct from saying that edge weights actually are randomly generated or uncorrelated. **Instead,** we mean that the weighted graph has **bulk properties** similar to those of random weighted graphs.

Why random-like nets hypothesis?

If a deep net is random-like, it can been shown that it has an associated simple generative model $p(x|h)$ that we call the **shadow distribution** (property(a)), and for which Property (b) also automatically holds in an approximate sense.

Achievement – SHADOW Method

The deep net being sought is random-like can be used to improve training. Namely, take a labeled data point x, and use the current feedfoward net to compute its label h. Now use the shadow distribution $p(x|h)$ to compute a synthetic data point $\tilde{x}$, label it with h, and add it to the training set for the next iteration. We call this the **SHADOW method**.

Single Layer Generative Model

Let's start with a single layer neural net $h = r(W^T x + b)$, where r is the rectifier linear function, $x \in R^n, h \in R^m$.

We define a shadow distribution $P_{W,\rho}$ for $x|h$, such that a random sample $\tilde{x}$ from this distribution satisfies Property (b), i.e., $r(W^T \tilde{x} + b) \approx h$ where $\approx$ denotes approximate equality of vectors.

Shadow Distribution $P_{W,\rho}$ for $x|h$

Given h, sampling x from the distribution $P_{W,\rho}(x|h)$ consist of first computing $r(\alpha Wh)$ for a scalar α and then randomly zeroing-out each coordinate with probability $1-\rho$.(We refer to this noise model as "dropout noise") Here ρ can be reduced to make x as sparse as needed; typically ρ will be small. More formally we have (with $\odot$ denoting entry-wise product of two vectors):

$$x = r(\alpha Wh) \odot n_{drop} \tag{1}$$

where $\alpha = 2/(\rho n)$ is a scaling factor, and $n_{drop} \in \{0,1\}^n$ is a binary random vector with following probability distribution where ($||n_{drop}||_0$ denotes the number of non-zeros of n_{drop}),

$$Pr[n_{drop}] = \rho^{||n_{drop}||_0} \tag{2}$$

Theorem

Theorem 1 (Reversibility)

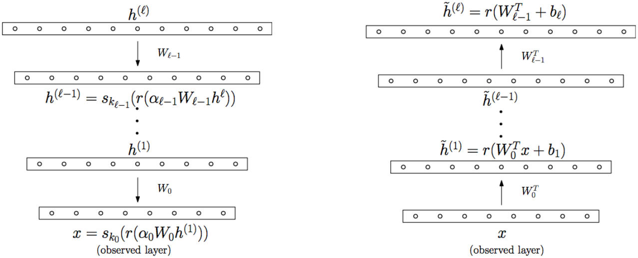
Let $t = \rho n$ be the expected number of non-zeros in the vector n_{drop} and k satisfy that $|h|_0 < k$ and $k < t < k^2$. The random-like net hypothesis here means that the entries of W independently satisfy $W_{ij} \sim N(0, 1)$.
For $(1 - n^{-5})$ measure of $W's$, there exists offset vector b, such that the following holds: when $h \sim D_h$ and $Pr[x|h]$ is specified by model (1), then with high probability over the choice of $(h, \tilde{x})$,

$$||r(W^T \tilde{x} + b) - h||^2 \leq \tilde{O}(k/t) \cdot ||h||^2 \quad (3)$$

Theorem

Theorem 2 (Informal version of Theorem 1)

If entries of W are drawn from i.i.d Gaussian prior, then for $\tilde{x}$ that is generated from model (1), there exists a threshold $\theta \in R$ such that $r(W^T\tilde{x} + \theta 1) \approx h$, where 1 is the all-1's vector.



(a) Generative model (b) Feedforward NN

Figure: Generative model vs. Feedforward NN

The feedforward net is shown in above Figure 1(a). The j^{th} layer has n_j nodes, while the observable layer has n_0 nodes. The corresponding generative model is in Figure 1(b). The number of variables at each layer, and the edge weights match exactly in the two models.

Full Multilayer Model

The generative model uses the hidden variable at layer j to produce the hidden variables at $j-1$ using exactly the single-layer generative analog of the corresponding layer of the deep net. It starts with a value $h^{(l)}$ of the top layer, which is generative from some arbitrary distribution D_l over the set of k_l-sparse vectors in R^{n_l}. Namely, apply a random sampling function $s_{k_{l-1}}(\cdot)$ on the vector $r(\alpha_{l-1}W_{l-1}h^{(l)})$, where k_{l-1} is the target sparsity of $h^{(l-1)}$, and $\alpha_{l-1} = 2/k_{l-1}$ is a scaling constant. Repeating the same stochastic process, we generate x at the bottom layer. In formula, we have

$$x = s_{k_0}(r(\alpha_0 W_0 s_{k_1}(r(\alpha_1 W_1 \cdots)))) \tag{4}$$

Theorem

Theorem 3.1(2-Layer Reversibility)

For $l = 2$, and $k_2 < k_1 < k_0 < k_2^2$, for 0.9 measure of the weights (W_0, W_1), the following holds: There exists constant offset vector b_0, b_1 such that when $h^{(2)} \sim D_2$ and $Pr[x|h^{(2)}]$ is specified as model (4), then network has reversibility in the sense that the feedforward calculation gives $\tilde{h}^{(2)}$ satisfying

$$\forall i \in [n_2], \quad E[|\tilde{h}_i^{(2)} - h_i^{(2)}|^2] < \epsilon \tau^2 \tag{5}$$

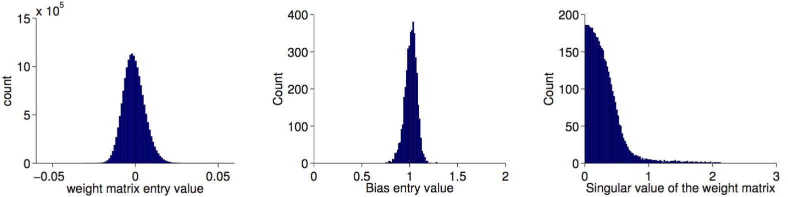
where $\tau = \frac{1}{k_2} \sum_i h_i^{(2)}$ is the average of the non-zero entries of $h^{(2)}$ and $\epsilon = \tilde{O}(k_2/k_1)$.

Theorem 3.2(3-Layer Reversibility, informally stated)

For $l = 3$, when $k_3 < k_2 < k_1 < k_0 < k_3^2$ and $\sqrt{k_3}k_2 < k_0$, the 3-layer generative model has the same type of reversibility properties as in Theorem3.1.

Alexnet

- Verify gaussain distribution on weights
- Verify the distribution of singular values of a gaussian random matrix



(a) Histogram of the entries in the weight matrix

(b) Histogram of the entries in the bias vector

(c) Singular values of the weight matrix

Figure: Verify Random-like Nets

Shadow method

- Reconstruction by Shadow method
- Performance on some datasets

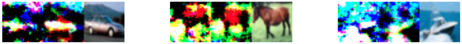


Figure: Reconstruction by Random weights, Training after 10^6 iteration compared with Original picture

Shadow method

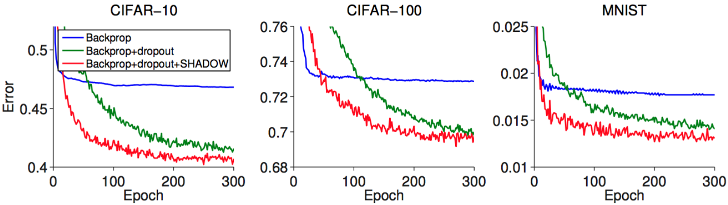


Figure: Performance by using dropout and shadow method

Thank You