Feature Back-tracking with Sparse Deep Belief Networks

FSDM 3637

Chen Qiao, Jiajia Li, etc.
Xi'an Jiaotong University

Introduction

To fully explore valuable information contained in **high-dimensional data**, models with strong expression ability are needed. **Deep learning** can learn the layer-wise nonlinear expression or features by greedy layer-by-layer training, thus to find out the complex information of the data.

Searching for an approach to obtain the essential features of the data from the deep expression with **layered structure**, will be great helpful to obtain the essential features of the original data in a global and layer-wised way, and what is the most important, to obtain **the interpretability of the deep learning** framework.

We consider both **the architecture sparsity** and **the representation sparsity** of the neurons in deep learning. This is an effective attempt to simulate the sparse topology structure of the brain networks.

Based on the **sparse DBN** network architecture, the feature back-tracking approach based on sparse deep learning is proposed to obtain the **interpretability of the deep learning**.

Methods

➢ Unsupervised sparse learning of DBN

A DBN is a generative graphical model, composed of multiple layers of hidden units. DBN is composed of several **restricted Boltzmann machines (RBMs)**.

To learn the structured network of each RBM and improve the interpretability of the network, **the Kullback-Leibler (KL) divergence** on the hidden neurons is used to achieve the response sparsity, and **the L_1-norm** is used on the connection weights to obtain the connective sparsity. **The L_2-norm** is added on the connection weights to limit their increasing bounds.

The objective function of the sparse **RBM** is given by

$$\max_{\Theta} \mathcal{L}_{new}(\Theta) = \mathcal{L}(\Theta) - \frac{1}{2}\lambda_1 ||W||_2^2 - \lambda_2 \sum_{j=1}^{N_h} KL(\rho||p_j) - \lambda_3 ||W||_1$$

The parameters of the unsupervised sparse RBM can be updated by

$$\Delta W_{ij} = \epsilon \cdot (\langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{recon}) - \lambda_1 W_{ij} - \lambda_3 \cdot sign(W_{ij}) - \lambda_2 \cdot \frac{1}{N_s}\left(-\frac{\rho}{p_j} + \frac{1-\rho}{1-p_j}\right) \cdot \left(\sum_{q=1}^{N_s} \sigma_j^{(q)}\left(1-\sigma_j^{(q)}\right) v_i^{(q)}\right) \quad (1)$$

$$\Delta \alpha_i = \epsilon \cdot (\langle v_i \rangle_{data} - \langle v_i \rangle_{recon}) \quad (2)$$

$$\Delta \beta_j = \epsilon \cdot (\langle h_j \rangle_{data} - \langle h_j \rangle_{recon}) - \lambda_2 \cdot \frac{1}{N_s}\left(-\frac{\rho}{p_j} + \frac{1-\rho}{1-p_j}\right) \cdot \left(\sum_{q=1}^{N_s} \sigma_j^{(q)}\left(1-\sigma_j^{(q)}\right)\right) \quad (3)$$

➢ Features back-tracking approach based on sparse DBN

A stacked **DBN** with sparse architecture and sparse representation is trained by Eq. (1)-(3).

Algorithm 1. The FBT method on sparse DBN

Input: $L, w^{(l)}(l = L-1,\cdots,1), \alpha^{(l)}(l = L-1,\cdots,1), I^{(L)}$
Ensure: $I^{(1)}$
  For $l = L-1:-1:1$
    $I^{(l)} = i: |w_{ij}^{(l)}| \geq k^{(l)} \& |h_{j,1}^{(l+1)} - h_{j,2}^{(l+1)}| \geq r^{(l+1)}, \exists j \in I^{(l+1)}$
    $N^{(l)} = length(I^{(l)})$
  End For
Return: $I^{(1)}, N^{(l)}(l = L,\cdots,1)$

where L is the layer numbers of the DBN, $w^{(l)}$ is the connecting weight between the $(l+1)$-th layer and the l-th layer, $k^{(l)}$ and $r^{(l+1)}$ are thresholds and set $I^{(l)}$ contains the selected positions of the l-th layer. By back-tracking method, $I^{(1)}$ is achieved, which the essential features from the data which contribute most to the final prediction.

Results

◆ Results on the SNPs data of Schizophrenia

Table 1. Results before and after back-tracking of SNPs data

	Dim	ACA	SCR
Raw Testing Data	12513*3	0.9867	1.0000
Testing Data with Selected Risk Loci	2973	0.9856	0.0792

Table 2. 10 typical selected risk loci by the feature back-tracking method

SNPs	Genes	SNPs	Genes
rs10102965	NRG1	rs1110144	CNTNAP2
rs11607732	GRIK4	rs6586002	GRID1
rs11013103	PIP4K2A	rs2765993	PIP4K2A
rs2098469	GRIN2B	rs220573	GRIN2B
rs111888901	HAAO	rs6433777	CACNB4

The selected risk loci can get a space saving rate about **92%** with a classification accuracy around **98.56%**. Nearly, all of the distinguishable loci of schizophrenia from the raw data have been chosen correctly.

◆ Results on MNIST data

Table 3. Average performance with FBT

Types of digits	AVG- AR_o	AVG- AR_s	AVG-N_s	AVG-SR
2	0.9976	0.9893	61	0.92
3	0.9975	0.9807	131	0.83
10	0.9607	0.9563	289	0.63

The key pixels identifying different digits can be selected successfully. With only part of the pixels, high recognition accuracy rates can still be well kept. For some cases, the identification accuracy is even improved with the picked pixels.

Table 4. Results before and after pixel selection of digits 1

DIGIT	N_o	N_s	AR_o	AR_s	SR
1	784	146	0.9974	0.9522	0.81
2	784	31	0.9961	1	0.96

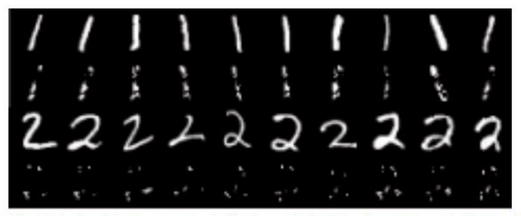

Figure 1. Original images and distinguishable images of digits 1 and 2

Table 5. Results before and after pixel selection of digits 0, 1 and 3

DIGIT	N_o	N_s	AR_o	AR_s	SR
0	784	124	0.9990	0.9801	0.84
1	784	126	0.9982	1	0.84
3	784	122	1	0.9814	0.84

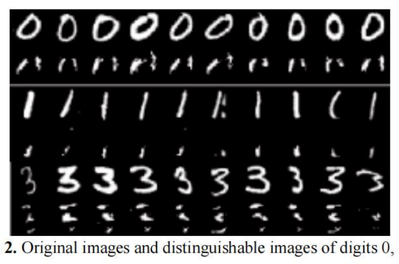

Figure 2. Original images and distinguishable images of digits 0, 1 and 3

The classification accuracy of digit 1 and digit 2 rises **from 96.61% to 100%** (Table 4 and Figure 1), the classification accuracy of digits 0, 1 and 3 increases to **100%** (Table 5 and Figure 2). The results show that the method can recognize the key positions to distinguish the digits.

Conclusions

- ❖ One approach of **interpretability** of deep learning framework is proposed here, which is quite meaningful in applications of deep learning.
- ❖ The corresponding regularization items on the hidden neurons and the connection weights are introduced in the network learning process, which can reduce the **complexity of networks** and enhance the **generalization ability**.
- ❖ This method has shown **quite well performance** of removing irrelevant features and reducing the difficulty and complexity of learning tasks, especially in searching for risk loci of schizophrenia and picking out the intrinsic pixels of different digits.