Semiconducting Bilinear Deep Learning for Incomplete Image Recognition

Sheng-hua Zhong[1], Yan Liu[1], Fu-lai Chung[1], Gangshan Wu[2]

[1] Department of Computing, The Hong Kong Polytechnic University, Hong Kong, P. R. China

[2] Department of Computer Science & Technology, Nanjing University, Nanjing, P. R. China

{csshzhong@comp.polyu.edu.hk, csyliu@comp.polyu.edu.hk, cskchung@comp.polyu.edu.hk, gswu@nju.edu.hk

ABSTRACT

Image recognition with incomplete data is a well-known hard problem in multimedia content analysis. This paper proposes a novel deep learning technique called semiconducting bilinear deep belief networks (SBDBN) by referencing human's visual cortex and intelligent perception. Inheriting from deep models, SBDBN simulates the laminar structure of human's cerebral cortex and the neural loop in human's visual areas. To address the special difficulties of image recognition with incomplete data, we design a novel second-order deep architecture with semiconducting restricted boltzmann machines. Moreover, two peaks activation of human's perception is implemented by three learning stages of semiconducting bilinear discriminant initialization, greedy layer-wise reconstruction, and global fine-tuning. Owing to exploiting the embedding information according to the reliable features rather than any completion of missing features, the proposed SBDBN has demonstrated outstanding recognition ability on two standard datasets and one constructed dataset, comparing with both incomplete image recognition techniques and existing deep learning models.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and scene understanding – *Representation, data structures, and transforms*;

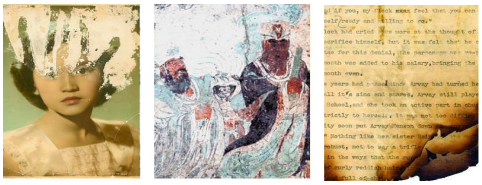
I.2.6 [**Artificial Intelligence**]: Learning –*connection and neural nets*

General Terms: Algorithm

Keywords: Deep learning, semiconducting bilinear discriminant initialization, semiconducting RBM, image recognition, missing features.

1. INTRODUCTION

Incomplete data, data values are partially observed [1], exists in a wide range of fields, including social sciences, computer vision, and remote sensing [2]. In general, features missing in real-world data are resulted from measurement noise, corruption or occlusion [3]. Figure 1 shows some real incomplete data examples. Figure 1 (a) demonstrates some example images with missing features due to noise and corruption, including: old picture, ancient fresco, and a burned paper with some available handwriting. Obviously, it is more difficult for computer to recognize meaningful patterns with the incomplete data. If the image distortion is very serious, even human beings can't recognize the images correctly. Figure 1 (b) provides more general cases of incomplete data in our daily life. David Beckham is one of the most iconic athletes and most fans have no difficulty to recognize him from these four images. But it is not an easy task for many face recognition models because some key facial features to identify persons, such as characters of eyes and mouth, are not observable.



(a)Incomplete images due to noise and corruption

(b) Incomplete face images due to the occlusion in the important facial features regions

Figure 1. The examples of incomplete images due to noise, corruption or occlusion.

Current works on incomplete data can be roughly categorized into three groups based on the modeling of the missing values [4]. The first kind of techniques fills the missing values based on the modeling of the available information, and then learns the decision function in a general way. Williams et al. developed a logistic regression classification algorithm for incomplete data. Conditional density functions were estimated using a Gaussian mixture model, with parameter estimation performed using both expectation maximization (LRCEM) [5] and Variational Bayesian (LRCVBEM) [2]. [6] proposed a novel second order cone programming formulation (SOCP) for designing robust classifiers which can handle uncertainty in observations. The second kind of techniques seeks the final decision boundary by estimating the missing value and constructing predictive model jointly. [1] proposed a statistical model names Quadratically Gated Mixture

ICMR'12, June 5-8, Hong Kong, China

of Experts (QGME) for multi-class nonlinear recognition. [4] derived a generic joint optimization weighted infinite imputations (WII) method, which learned the decision function and the distribution of imputations dependently. The third kind of techniques doesn't intend to estimate the missing values. They learn the decision function only based on the visible input, which can avoid the additional error introduced by estimating the unknown values. [7] [3] recognized the incomplete data directly without any completion of the missing features using a max-margin learning framework. For each sample, the margin is rescaled according to the visible attributes.

This paper simulates a new thought to solve the problem of image recognition with incomplete data by referencing human's visual cortex and intelligent perception. Deep learning, which models the learning tasks using deep architectures composed of multiple layers of parameterized nonlinear modules, is selected in this paper. Deep architectures simulate the laminar structure of human's cerebral cortex and the information delivery between multiple layers reproduces the neural loop in human's visual areas. Therefore, deep learning has demonstrated distinguished ability of information abstraction and robust performance of data classification in various visual data analysis tasks [8]. To address the difficulties caused by incomplete data, we propose a novel deep learning technique called semiconducting bilinear deep belief networks (SBDBN) based on the representative deep model called deep belief networks (DBN). Compared with existing deep models, the proposed SDBN has several attractive characters:

1) Most deep models, such as DBN, unfold the image, a natural second-order tensor, to one-dimensional vector as the input to the deep architecture. Such kind of vectorization will cause the high computational cost, and more importantly, the loss of spatial information. The incomplete data problem always suffers from insufficient information; therefore, we try to preserve the existing information included in the data as much as we can. So the proposed SBDBN utilizes a novel deep architecture constructed by a set of second-order planes instead of first-order vector. Such kind of design is also consistent with human's visual perception. In the primary visual cortex, all the way through the optic tract to a nerve position is a direct correspondence from an angular position in the field of view of the eye, just like a matrix.

2) Thanks to the flexibility of human's visual perception procedure, most fans can recognize David Beckham from the images shown in Figure 1 (b), although some key facial features for person identification are covered. Humans can automatically adjust their attention to the available features and emphasize the contributions from them consciously and, actually sometimes unconsciously. The proposed SBDBN borrows this idea by designing a set of Semiconducting Restricted Boltzmann Machine (SRBM) to construct the deep architecture. Restricted Boltzmann Machine (RBM) is a two-layer recurrent neural network in which stochastic binary inputs are connected to stochastic binary outputs using symmetrically weighted connections, which has been widely used as the building blocks in many deep models. Different with the densely connected RBM, SRBM sets a semiconductor switch on each connection between the lower layer, i.e. the input layer, and the upper layer, i.e. the first hidden layer. In the training stage, if the feature value of this training data is missing, the semiconductor switch is off and the weight on this linkage will keep unchanged. Otherwise, the semiconductor switch is on, and the weight on this linkage will be updated according to the learning objective function. Similarly, in the test stage, we can determine the category of the data only based on the available features by controlling the semiconductor switch.

3) Based on the proposed new deep architecture, we present a novel deep learning algorithm with three stages: semiconducting bilinear discriminant initialization, greedy layer-wise reconstruction, and global fine-tuning. The rationale of three-stage learning comes from the phenomenon of two peaks activation in visual cortex areas. With regard to object recognition, the early peak is related to the activation of an "initial guess" based on the acquired discriminative knowledge, while the late peak reflects the post-recognition activation of conceptual knowledge related to the recognized object. In most existing deep models, "post activation" is modeled by fine-tuning stage, but the "initial guess" process is neglected. However, initial guess plays an important role in human perception, especially under the case that the data is incomplete. In our model, we propose semiconducting bilinear discriminant initialization to realize "initial guess" under insufficient information.

The remainder of this paper is organized as follows. Related work on deep learning is reviewed in Section 2. A novel deep architecture and a new deep learning algorithm are introduced in Section 3. Section 4 shows the performance of the proposed techniques in real image recognition and retrieval tasks and Section 5 concludes this paper.

2. RELATED WORK ON DEEP LEARNING

Different from shallow learning models, such as support vector machine (SVM), deep learning is about learning multiple levels of representation and abstraction that helps to make sense of data. Some theoretical analyses from machine learning also provide support for the argument that deep models are more compact and expressive than shallow models in representing most learning functions, especially highly variable ones. For example, to model the d-dimensional parity function, Gaussian SVM uses $O(d2^d)$ parameters while deep learning only needs $O(d^2)$ parameters with $O(log_2 d)$ hidden layers [9]. The effectiveness of a deep model makes it promising for use in solving hard learning problems, for example, in semantically identifying the class of images from low-level visual features.

The performance of deep learning has been notable, especially after the introduction of the deep belief networks (DBN) model. The learning procedure of DBN can be divided into two stages: abstracting information layer by layer and fine-tuning the whole deep network to the ultimate learning target [8]. Figure 2 shows a DBN with one input layer H^1, three hidden layers H^2, H^3, H^4, while x is the unfolding vector of input data, and y is the learning target. In the first stage, DBN pairs each feed-forward layer with a feed-back layer that attempts to reconstruct the input of the layer from the output. In Figure 2, the layer-wise reconstruction happens between H^1 and H^2, H^2 and H^3, H^3 and H^4, which is implemented by a family of Restricted Boltzmann Machines (RBMs) [10]. After a greedy unsupervised learning of each pair of layers, the lower-level features are progressively combined into more compact high-level representations. The whole deep network is then refined using a contrastive version of the "wake-sleep" algorithm via a global gradient-based optimization strategy.

Furthermore, empirical validations in various real-world applications have shown that DBN performs impressively in analyses of visual data, such as in image classification [11], image

annotation [12], and image retrieval [13].

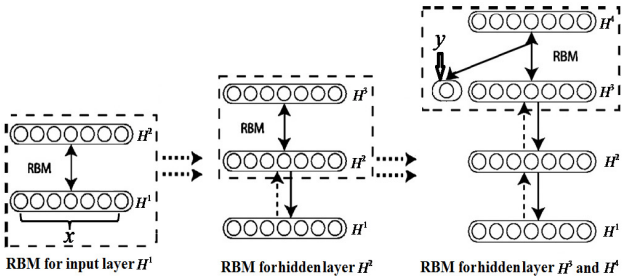


Figure 2. Structure of the deep belief network (DBN).

3. SEMICONDUCTING DEEP BELIEF NETWORKS

In this section, we propose a novel deep learning architecture based on semiconducting bilinear deep belief network (SBDBN). Our semiconducting bilinear deep belief network, which is aimed at the task of incomplete image recognition, is described in Section 3.1. The semiconducting bilinear discriminant initialization stage is discussed in Section 3.2. Section 3.3 contains details of the greedy layer-wise reconstruction by semiconducting RBM. The global fine-tuning process of the whole deep network is described in Section 3.4. We provide the procedure of SBDBN in Section 3.5.

3.1 Framework of Semiconducting Bilinear Deep Belief Network

Let X be a set of incomplete data samples as shown below:

$$X = [\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_k, ..., \mathbf{X}_K] \quad (1)$$

where $\mathbf{X}_k$ is a sample datum with missing features in the image space $\mathbb{R}^{I \times J}$ and K is the number of sample data. Let F_k denote the set of missing features of the sample $\mathbf{X}_k$, $(\mathbf{X}_k)_{ij}$ is missing if $(\mathbf{X}_k)_{ij} \in F_k$. Let Y be a set of labels corresponding to X, which can be seen as:

$$Y = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_k, ..., \mathbf{y}_K] \quad (2)$$

And $\mathbf{y}_k$ is the label vector of $\mathbf{X}_k$ in $\mathbb{R}^C$, where C is the number of classes.

$$y_k^c = \begin{cases} 1 & \text{if } \mathbf{X}_k \in c\text{th class} \\ 0 & \text{if } \mathbf{X}_k \notin c\text{th class} \end{cases} \quad (3)$$

Based on the given training set, the goal in image recognition is to learn a mapping function from the image set X to the label set Y, and then recognize the new coming data points according to the learned mapping function.

To address the problem of incomplete image recognition, we propose a novel semiconducting bilinear deep learning technique SBDBN. Figure 3 shows the architecture of SBDBN. A fully interconnected directed belief network includes the semiconducting input layer H^1, hidden layer H^2,..., H^N, and one label layer La at the top. The semiconducting input layer H^1 has $I \times J$ units, and this size is equal to the dimension of the input features. In our model, we use the pixel values of sample datum $\mathbf{X}_k$ as the original input features. In the top, the label layer has C units, which is equal to the number of classes. The search of the mapping function from X to Y is transformed to the problem of finding the optimum parameter space θ^*.

The procedure under supervised or semi-supervised learning framework of our SBDBN is listed below:

1. The strategy of semiconducting bilinear discriminant projection is utilized to construct a projection to map the original data into a discriminant bilinear subspace based on the reliable features.
2. The initial symmetrically weighted connections are constructed between adjacent layers according to the "initial guess" based on the discriminant information. The size of the deep architecture is determined automatically based on the optimum dimension to retain the discriminant information.
3. After the architecture of the next layer is determined, the parameter space is refined by the greedy layer-wise information reconstruction using semiconducting RBMs as building blocks.
4. Repeat the first to third stages until the parameter space θ in all N layers is constructed.
5. In the "post activation" stage, the whole deep model is fine-tuned to minimize the recognition error based on backpropagation.

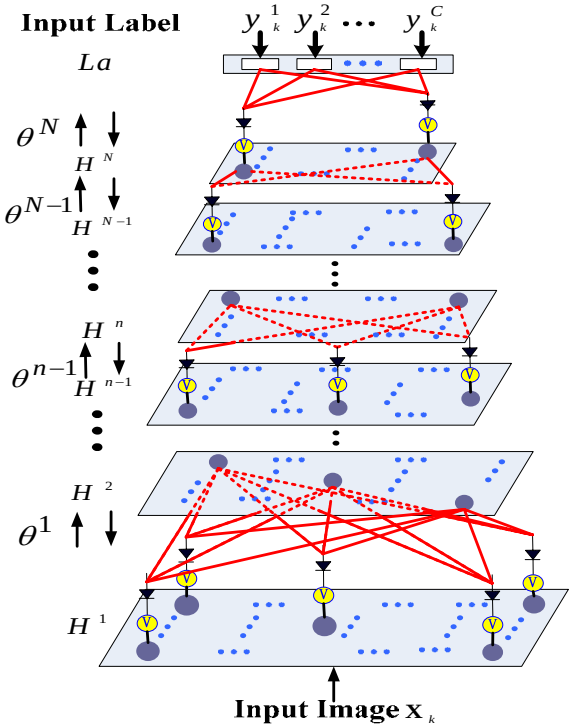


Figure 3. Architecture of SBDBN.

3.2 Semiconducting Bilinear Discriminant Initialization

In this subsection, we introduce the semiconducting bilinear discriminant projection (SBDP), which is utilized to extract the discriminant information from the image datasets with incomplete features.

Given the training data points $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_K \in \mathbb{R}^{I\times J}$ with missing features set F_k, $(\mathbf{X}_k)_{ij}$ is missing if $(\mathbf{X}_k)_{ij} \in F_k$. SBDP aims to find two projection matrices $\mathbf{U} \in \mathbb{R}^{I\times P}$ and $\mathbf{V} \in \mathbb{R}^{J\times Q}$ such that the latent representation $\mathbf{TX}_1, \mathbf{TX}_2, ..., \mathbf{TX}_K \in \mathbb{R}^{I\times J}$ can be obtained by $\mathbf{TX}_s = \mathbf{U}^T\mathbf{X}_s\mathbf{V}$ (s = 1, ..., K) from reliable features.

In order to preserve the discriminant information from reliable features in the learning procedure, the objective function of SBDP could be represented as follows:

$$\begin{aligned} &\arg\max_{\mathbf{U},\mathbf{V}} J(\mathbf{U},\mathbf{V}) = \sum_{s,t=1}^{K} \| \mathbf{U}^T(\mathbf{X}_s \cap \mathbf{Z}_{st}^{\mathrm{X}} - \mathbf{X}_t \cap \mathbf{Z}_{st}^{\mathrm{X}})\mathbf{V} \|^2 (\alpha \mathbf{B}_{st} - (1-\alpha)\mathbf{W}_{st}) \\ &s.t.\ (Z_{st}^{\mathrm{X}})_{ij} = \begin{cases} 0, & \text{if } (\mathbf{X}_s)_{ij} \in F_s \text{ or } (\mathbf{X}_t)_{ij} \in F, \\ 1, & \text{else,} \end{cases}, \mathbf{U}^T\mathbf{U} = \mathbf{I}_P,\ \mathbf{V}^T\mathbf{V} = \mathbf{I}_Q \end{aligned} \tag{4}$$

Different with the bilinear discriminant projection (BDP) in [11], we extract the discriminant information based on the reliable features. In Equation (4), $\alpha \in [0,1]$ is the parameter used to balance the between-class weights $\mathbf{B}_{st}$ and the within class weights $\mathbf{W}_{st}$, which are defined as follows [11]:

$$\mathbf{B}_{st} = \begin{cases} \frac{1}{n_d} - \frac{1}{n_c}, & \text{if } \mathbf{y}_s^c = \mathbf{y}_t^c = 1, \\ \frac{1}{n_d}, & \text{else,} \end{cases}, \mathbf{W}_{st} = \begin{cases} \frac{1}{n_c}, & \text{if } \mathbf{y}_s^c = \mathbf{y}_t^c = 1, \\ 0, & \text{else,} \end{cases} \tag{5}$$

where $\mathbf{y}_s^c$ denotes the class label of datum point $\mathbf{X}_s$, n_d is the number of data points in all class and n_c is the number of data points in class c, where $c \in \{1, ..., C\}$.

By simultaneously maximizing the distances between data points from different classes and minimizing the distance between data points from the same class, the discriminant information is preserved at the greatest extent in the projected feature space. Optimizing $J(\mathbf{U},\mathbf{V})$ by solving $\mathbf{U}$ (or $\mathbf{V}$) with fixed $\mathbf{V}$ (or $\mathbf{U}$) is a convex optimization problem. Let $\mathbf{E}_{st} = \alpha\mathbf{B}_{st} - (1-\alpha)\mathbf{W}_{st}$, with the fixed $\mathbf{V}$. The optimal $\mathbf{U}$ is composed of the first P eigenvectors of the following eigendecomposition problem:

$$\mathbf{D}_{\mathbf{V}}\mathbf{u} = \lambda\mathbf{u} \tag{6}$$

where $\mathbf{D}_{\mathbf{V}} = \sum_{st}\mathbf{E}_{st}(\mathbf{X}_s \cap \mathbf{Z}_{st}^{\mathrm{X}} - \mathbf{X}_t \cap \mathbf{Z}_{st}^{\mathrm{X}})\mathbf{V}\mathbf{V}^T(\mathbf{X}_s \cap \mathbf{Z}_{st}^{\mathrm{X}} - \mathbf{X}_t \cap \mathbf{Z}_{st}^{\mathrm{X}})^T$. Similarly, with the fixed $\mathbf{U}$, the optimal $\mathbf{V}$ is composed of the first Q eigenvectors of the following eigendecomposition problem:

$$\mathbf{D}_{\mathbf{U}}\mathbf{v} = \lambda\mathbf{v} \tag{7}$$

where $\mathbf{D}_{\mathbf{U}} = \sum_{st}\mathbf{E}_{st}(\mathbf{X}_s \cap \mathbf{Z}_{st}^{\mathrm{X}} - \mathbf{X}_t \cap \mathbf{Z}_{st}^{\mathrm{X}})^T\mathbf{U}\mathbf{U}^T(\mathbf{X}_s \cap \mathbf{Z}_{st}^{\mathrm{X}} - \mathbf{X}_t \cap \mathbf{Z}_{st}^{\mathrm{X}})$.

Therefore, we can alternately optimize $\mathbf{U}$ (with a fixed $\mathbf{V}$) and $\mathbf{V}$ (with a fixed $\mathbf{U}$). The above steps monotonically increase $J(\mathbf{U},\mathbf{V})$ and since the function is upper bounded, it will converge to a critical point with transformation matrices $\mathbf{U}$ and $\mathbf{V}$.

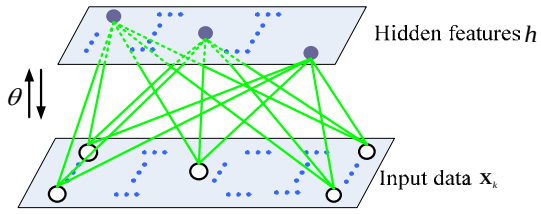
In SBDP, the sizes of P and Q are determined by the number of positive eigenvalues in $\mathbf{D}_{\mathbf{V}}$ and $\mathbf{D}_{\mathbf{U}}$, respectively, since adding the eigenvectors corresponding to the nonpositive eigenvalues will not increase $J(\mathbf{U},\mathbf{V})$ in Equation (4). As a result, the original dimension $I \times J$ is automatically reduced into $P \times Q$ after the SBDP procedure.

3.3 Greedy Layer-Wise Reconstruction by Semiconducting RBM

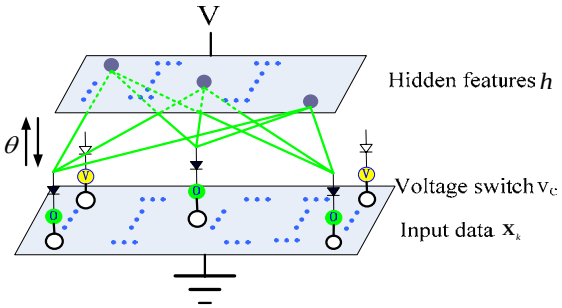
RBM in deep belief network helps us to abstract the embedding information by layer-wise reconstruction. Unfortunately, RBM cannot work when some features are missing, and the corresponding units of the networks are empty.

Inspired from electronic circuits [14], in proposed SBDBN, we design a semiconducting RBM instead of original RBM. In the original RBM, every feature is input to the RBM just as Figure 4 (a). Different with original RBM, in the semiconducting RBM which is shown in Figure 4 (b), we add a semiconductor switch to control the connection with the current layer to the higher hidden layer. When the input feature $(\mathbf{X}_k)_{ij} \in F_k$, set $(\mathrm{v_c})_{ij}$ = V, and the switch is off.

V is the "voltage" of the hidden layer, and the "voltage" of the visual data is $\mathrm{v_c}$. In this case, the corresponding parameters $(\theta)_{ij}$ related to $(\mathbf{X}_k)_{ij}$ will not be tuned. Otherwise, if the input feature $(\mathbf{X}_k)_{ij} \notin F_k$, set $(\mathrm{v_c})_{ij}$ = 0, and the switch is on. In this turn, the corresponding parameters $(\theta)_{ij}$ will be updated.



(a) The operation principle of RBM



(a) The operation principle of semiconducting RBM

Figure 4. The operation principle of RBM and semiconducting RBM. (a) The operation principle of RBM, every feature of $\mathbf{X}_k$ is input. (b) The operation principle of semiconducting RBM. When $\mathbf{X}_k$ come in, the switch turns ON if $(\mathbf{X}_k)_{ij} \notin F_k$.

With the semiconducting RBM, greedy layer-wise reconstruction is different from the stage in DBN. The sample data $\mathbf{X}_k$ with incomplete features F_k is input to the deep architecture as the state of the input layer H^1 to construct an RBM with the first hidden layer H^2. The energy of the state $(\mathbf{h}^1, \mathbf{h}^2)$ in the first semiconducting RBM is:

$$E\left(\mathbf{h}^1,\mathbf{h}^2;\theta^1\right)$$
$$=-\sum_{i=1,j=1}^{i\le I,j\le J}\sum_{p=1,q=1}^{p\le P^2,q\le Q^2} h_{ij}^1(A_{ij,pq}^1\cap Z_{ij,pq}^{A,1})h_{pq}^2-\sum_{i=1,j=1}^{i\le I,j\le J}(b_{ij}^1\cap Z_{ij}^{b,1})h_{ij}^1-\sum_{p=1,q=1}^{p\le P^2,q\le Q^2}c_{pq}^1h_{pq}^2 \quad (8)$$

where $I\times J$ is the number of units in H^1, while $P^2\times Q^2$ is the number of units in H^2. $\theta^1=\left(\mathbf{A}^1,\mathbf{b}^1,\mathbf{c}^1\right)$ are the model parameters between the input layer H^1 and first hidden layer H^2. $A_{ij,pq}^1$ is the symmetric interaction term between the input unit (i,j) in H^1 and the hidden unit (p,q) in H^2. b_{ij}^1 is the $(i,j)^{th}$ bias of layer H^1 and c_{pq}^1 is the $(p,q)^{th}$ bias of layer H^2. $Z_{ij,pq}^{A,1}$ and $Z_{ij}^{b,1}$ are the switch parameters to control the corresponding parameters $(\theta)_{ij}$ related to $(\mathbf{X}_k)_{ij}$ will or will not be tuned.

$$Z_{ij,pq}^{A,1}=\begin{cases}0, & \text{if } (\mathbf{X}_k)_{ij}\in F_k,\\ 1, & \text{else,}\end{cases},\quad Z_{ij}^{b,1}=\begin{cases}0, & \text{if } (\mathbf{X}_k)_{ij}\in F_k,\\ 1, & \text{else,}\end{cases} \quad (9)$$

Therefore the first RBM has the following joint distribution:

$$P\left(\mathbf{h}^1,\mathbf{h}^2;\theta^1\right)=\frac{e^{-E\left(\mathbf{h}^1,\mathbf{h}^2;\theta^1\right)}}{\sum_{\mathbf{h}^1}\sum_{\mathbf{h}^2}e^{-E\left(\mathbf{h}^1,\mathbf{h}^2;\theta^1\right)}} \quad (10)$$

The log probability of the model assigned to $\mathbf{h}^1$ in H^1 is:

$$\log P\left(\mathbf{h}^1\right)=\log\sum_{\mathbf{h}^2}e^{-E\left(\mathbf{h}^1,\mathbf{h}^2;\theta^1\right)}-\log\sum_{\mathbf{h}^1}\sum_{\mathbf{h}^2}e^{-E\left(\mathbf{h}^1,\mathbf{h}^2;\theta^1\right)} \quad (11)$$

Similar with existing deep learning models, we utilize the stochastic steepest ascent in the log probability of the training data to update the parameter space $\theta^1=\left(\mathbf{A}^1,\mathbf{b}^1,\mathbf{c}^1\right)$.

$$A_{ij,pq}^1=\vartheta A_{ij,pq}^1+\triangle A_{ij,pq}^1\cap Z_{ij,pq}^{A,1} \quad (12)$$

$$\triangle A_{ij,pq}^1=\varepsilon_{\mathbf{A}}(<h_{ij}^1(0)h_{pq}^2(0)>_{data}-<h_{ij}^1(1)h_{pq}^2(1)>_{recon}) \quad (13)$$

Where $\langle\cdot\rangle_{data}$ denotes an expectation with respect to the data distribution and $\langle\cdot\rangle_{recon}$ denotes the "reconstruction" distribution of data after one step. Other parameters in θ^1 update function can be calculated in a similar manner.

$$b_{ij}^1=\vartheta b_{ij}^1+\triangle b_{ij}^1\cap Z_{ij}^{b,1}=\vartheta b_{ij}^1+\varepsilon_{\mathbf{b}}(h_{ij}^1(0)-h_{ij}^1(1))\cap Z_{ij}^{b,1} \quad (14)$$

$$c_{pq}^1=\vartheta c_{pq}^1+\triangle c_{pq}^1=\vartheta c_{pq}^1+\varepsilon_{\mathbf{c}}(h_{pq}^2(0)-h_{pq}^2(1)) \quad (15)$$

where ϑ is the momentum and $\varepsilon_{\mathbf{A}}$, $\varepsilon_{\mathbf{b}}$, $\varepsilon_{\mathbf{c}}$ are the learning rate of model parameters $\mathbf{A}$, $\mathbf{b}$ and $\mathbf{c}$.

As we described before, we find a semiconducting bilinear projection based on the reliable features that can automatically reduce the original dimension $I\times J$ to $P\times Q$ through the transformation matrices $\mathbf{U}^1$ and $\mathbf{V}^1$. As a result, in our model, the number of neurons in layer H^2 is determined by the row and column size of the transformation matrices $\mathbf{U}^1$ and $\mathbf{V}^1$.

$$P^2=row(\mathbf{U}^1),\ Q^2=column(\mathbf{V}^1) \quad (16)$$

We set the discriminative transformation parameters obtained from the semiconducting bilinear discriminant projection as the initial symmetrically connection weights by Equation (17).

$$A_{ij,pq}^1(0)=(\mathbf{U}_{ip}^1)^T\mathbf{V}_{jq}^1 \quad (17)$$

The above discussion is the greedy layer-wise abstraction for the first semiconducting RBM. Similar operations can be performed on the higher layer pairs.

3.4 Global Fine-Tuning

Above, we use the greedy layer-by-layer reconstruction algorithm by semiconducting RBM to learn a deep model. In this section, we use backpropagation through the whole deep model to fine-tune the parameters $\theta=\mathbf{[A,b,c]}$ for optimal reconstruction.

In the greedy layer-by-layer information abstraction stage, a global search has been performed for a sensible and good region in the whole parameter space. Therefore, before proceeding to the process of fine-tuning, we have already constructed a good data concept extraction model. In our model, backpropagation is utilized to adjust the entire deep network to find good local optimum parameters $\theta^*=[\mathbf{A}^*,\mathbf{b}^*,\mathbf{c}^*]$ to effectively recognize the data. In this stage, the learning algorithm is used to minimize the recognition error $[-\sum_k\mathbf{y}_k\log\hat{\mathbf{y}}_k]$, where $\mathbf{y}_k$ and $\hat{\mathbf{y}}_k$ are the correct label and the output label value of sample datum $\mathbf{X}_k$.

3.5 Semiconducting Bilinear Deep Learning Algorithm

In this section, the detailed procedure of the SBDBN is described in Algorithm 1.

Algorithm 1: Semiconducting Bilinear Deep Belief Network

Input: Training data set X, Corresponding labels set Y
Missing features set F_k in $\mathbf{X}_k$
Number of layers N, Number of epochs E
Switch parameters $\mathbf{Z}_{st}^{\mathrm{X}}$, $\mathbf{Z}^A$ and $\mathbf{Z}^b$
Between-class weights $\mathbf{B}_{st}$, Within class weights $\mathbf{W}_{st}$
Initial bias parameters $\mathbf{b}$ and $\mathbf{c}$
Momentum ϑ, Parameter α

Output: Optimal parameter space $\theta^*=[\mathbf{A}^*,\mathbf{b}^*,\mathbf{c}^*]$

for n = 1,…, N **do**
  for e = 1,…, E **do**
    if n = 1
      $T^n=X$
    else
      for k = 1,…, K **do**
        $\mathbf{T}_k^n=\sigma(\mathbf{T}_k^{n-1}A^{n-1}+c^{n-1})$
      end for
    end if
    while not convergent **do**
      $\mathbf{D}_V=\sum_{st}\mathbf{E}_{st}(\mathbf{T}_s^n\cap\mathbf{Z}_{st}^{\mathrm{Xn}}-\mathbf{T}_t^n\cap\mathbf{Z}_{st}^{\mathrm{Xn}})\mathbf{W}^T(\mathbf{T}_s^n\cap\mathbf{Z}_{st}^{\mathrm{Xn}}-\mathbf{T}_t^n\cap\mathbf{Z}_{st}^{\mathrm{Xn}})^T$
      $\mathbf{D}_U=\sum_{st}\mathbf{E}_{st}(\mathbf{T}_s^n\cap\mathbf{Z}_{st}^{\mathrm{Xn}}-\mathbf{T}_t^n\cap\mathbf{Z}_{st}^{\mathrm{Xn}})^T\mathbf{U}\mathbf{U}^T(\mathbf{T}_s^n\cap\mathbf{Z}_{st}^{\mathrm{Xn}}-\mathbf{T}_t^n\cap\mathbf{Z}_{st}^{\mathrm{Xn}})$
      Fix $\mathbf{V}$, compute $\mathbf{U}$ by solving $\mathbf{D}_V\mathbf{u}=\boldsymbol{\lambda}\mathbf{u}$

14. Fix $\mathbf{U}$, compute $\mathbf{V}$ by solving $\mathbf{D}_U\mathbf{v} = \boldsymbol{\lambda}\mathbf{v}$
15. **end while**
16. Determine the size of next layer
$P^{n+1} = row(\mathbf{U}^n)$, $Q^{n+1} = column(\mathbf{V}^n)$
17. Compute initial connection weights
$A_{ij,pq}^n(0) = (\mathbf{U}_{ip}^n)^T \mathbf{V}_{jq}^n$
18. The energy in the current semiconducting RBM
$E(\mathbf{h}^n, \mathbf{h}^{n+1}; \theta^n)$
$$= -\sum_{i=1,j=1}^{i\le P^n, j\le Q^n} \sum_{p=1,q=1}^{p\le P^{n+1}, q\le Q^{n+1}} h_{ij}^n (A_{ij,pq}^n \cap Z_{ij,pq}^{An}) h_{pq}^{n+1} - \sum_{i=1,j=1}^{i\le P^n, j\le Q^n} (b_{ij}^n \cap Z_{ij}^{bn}) h_{ij}^n - \sum_{p=1,q=1}^{p\le P^{n+1}, q\le Q^{n+1}} c_{pq}^n h_{pq}^{n+1}$$
19. Update the weights and biases
$$A_{ij,pq}^n = \vartheta A_{ij,pq}^n + \triangle A_{ij,pq}^n \cap Z_{ij,pq}^{An}$$
$$b_{ij}^n = \vartheta b_{ij}^n + \triangle b_{ij}^n \cap Z_{ij}^{b,n}$$
$$c_{pq}^n = \vartheta c_{pq}^n + \triangle c_{pq}^n$$
20. **end for**
21. **end for**
22. Calculate optimal parameter space $\theta^* = \arg\min_{\theta}[-\sum_k \mathbf{y}_k \log \hat{\mathbf{y}}_k]$

4. EXPERIMENTS AND RESULTS

In this section, three datasets with different kinds of visual data are used to demonstrate the performance of the proposed SBDBN. The first dataset is the MNIST, a standard large database of hand written digits containing 70,000 images with 10 classes [15]. The second standard dataset is the BioID face dataset consists of 1521 face images collected contains 23 subjects [16]. The face images in BioID are under a large variety of illumination and background. The third dataset StarFace is collected and constructed by our group, including 120 face images with 4 superstars.

In the experiments setting, we set the parameters follows the setting of previous work on bilinear deep belief network [8] and other general setting of deep learning. For example, the balance weight α is set as 0.5. In greedy layer wise learning, the number of epochs is fixed at 50 and the learning rate η is equal to 0.1.The initial momentum ϑ is 0.5. In the fine-tuning stage, the method of conjugate gradients is utilized and three line searches are performed in each epoch until convergence.

We compare the performance of SBDBN with other representative incomplete image recognition models and deep learning models, including *k*-nearest neighbor estimation (*k*-NNE), support vector machines (SVM) [17], logistic regression classification expectation maximization (LRCEM) [5], maximize geometric margin (GEOM) [3], deep belief networks (DBN) [8], and bilinear deep belief networks (BDBN) [11]. In *k*-NNE, the missing features were set with the mean value obtained from the nearest neighbors' instances. Neighborhood was measured using a Euclidean distance in the subspace relevant to each pair of samples. The number of neighbors was varied across 3,5,10, 15, 20, and the best result of these four (on test data) is shown. In EM, a Gaussian mixture model is learned by iterating between (1) learning a GMM model of the filled data and (2) re-filling missing values using cluster means, weighted by the posterior probability that a cluster generated the sample. The number of clusters was varied across 3, 5, 10, 15, 20, and the best result is reported.

4.1 Experiments on Handwriting Dataset MNIST

In this section, we explore performance of SBDBN on image dataset of handwritten digits MNIST [15] when features are missing at random. MNIST is a standard large database of hand written digits containing 70,000 images with 10 classes. MNIST is often used to compare deep learning performance [18] [19]. Figure 5 shows sample images of MNIST, some of which are difficult to recognize.

Figure 5. Some sample images in MNIST.

The first experiment in this dataset is used to demonstrate the effectiveness of SBDBN for recognition on incomplete images with fixed missing ratio. We follow the same experimental setting of [3]. 1200 images including 600 images of the digits 5 and 600 images of digit 6 are randomly selected from MNIST. These images are partitioned to 1000 training data and 200 test data. We removed a square patch of pixels from each image that covered 25% of the total number of pixels. The location of the patch was uniformly sampled for each image, and typical examples are given in Figure 6.

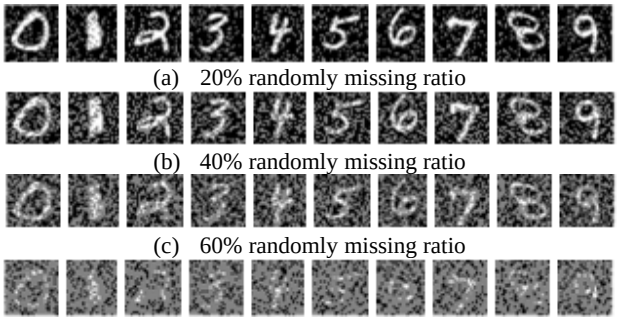
Figure 6. Examples of MNIST images of the digits '5' and '6' after fixed missing ratio pixels are removed with random centers.

We perform 5 random splits and report the average results over the 5 trials. The recognition performance of SBDBN with other incomplete image recognition models is shown in Table 1. "Zero" means that the missing values were set to zero. "Mean" means that the missing values were set to the average value of the feature over all data. From Table 1, it can easily be seen that, compared with other representative incomplete image recognition models, the deep learning models achieved better performance. This proves that the deep learning models have great recognition ability. Owing to the semiconducting bilinear discriminant initialization and semiconducting RBM of SBDBN, the recognition ability is promoted by fully exploiting the embedding information according to the reliable features rather than partially influenced by the forecasting accuracy of the missing features. Therefore, our proposed SBDBN achieved best performance in deep learning models.

Table 1. Recognition accuracy rate on test data.

Deep Model	Acc.	Other Model for Incomplete Data	Acc.
SBDBN	**98.5**	SVM (Zero)	95
BDBN (Zero)	97	SVM (Mean)	95
BDBN (Mean)	97.5	*k*-NNE	94
DBN (Zero)	96	LRCEM	95
DBN (Mean)	96.5	GEOM	95

In the second experiment, we demonstrate the incomplete image recognition when features are missing at random under different missing ratio. 10,000 images from MNIST are utilized as the training data, and the remaining 60,000 images are utilized for test. Some sample images with different missing ratios are shown in Figure 7. We perform 5 random missing trails and report the average results over the 5 trials. Although the original image samples selected in Figure 7 are not difficult to recognize, when the missing ratio becomes larger, even human cannot recognize these handwritten digits easily.



(a) 20% randomly missing ratio

(b) 40% randomly missing ratio

(c) 60% randomly missing ratio

(d) 80% randomly missing ratio

Figure 7. Examples of images for different percent of pixels missing randomly.

Table 2 shows the performance comparison under different missing ratios. Obviously, SBDBN shows higher incomplete image recognition accuracy rate. When 80% features missing, although the recognition by human is adequately hard, SBDBN also demonstrates the acceptable performance.

Table 2. Recognition accuracy rate on test data with different missing ratio.

Missing Ratio	SBDBN	DBN (Zero)	DBN (Mean)	SVM (Zero)	SVM (Mean)
20%	**96.04**	95.99	95.80	84.84	84.25
40%	**95.29**	93.62	93.75	81.10	83.52
60%	**92.41**	89.70	91.62	77.96	79.70
80%	**78.78**	77.68	76.85	60.47	73.69

4.2 Experiments on Face Image Dataset BioID

In this section, we explore performance of SBDBN for face recognition on dataset of BioID [16] when important facial features are missing.

Over the last ten years or so, face recognition has become a popular area of research in image analysis and understanding. Because the nature of the problem, not only computer science researchers are interested in it, but neuroscientists and psychologists also. Although much progress has been made in these years, face recognition remains a research area far from maturity, and its applications are still limited in controllable environments [20]. Face recognition with incomplete features remains a well-known challenge and little work is proposed to solve it. Although several datasets provide face images with glass [21], head cover [22], sunglass, masks, or wigs [23], the categories of these datasets are too limited and most of them are thermal infrared face images [22][23]. Fortunately, the important facial feature points are provided in BioID face dataset which helps us to generate incomplete face images datasets with missing important facial feature regions.

BioID face dataset consists of 1521 face images collected contains 23 subjects. The number of images in every category of BioID is varied, from 35 to 118. Therefore, firstly, we choose the categories with more than 50 face images as the subset we work on. This subset includes 1208 images in 14 categories. Then, just like the procedure on face datasets, the original images are normalized (in scale and orientation) so that the two eyes are aligned at the same position. Finally, the facial areas are cropped and downsampled into the final images. The size of each final image in all of the experiments is 32×32 pixels, with 256 gray levels per pixel. Some sample images are shown in Figure 8.

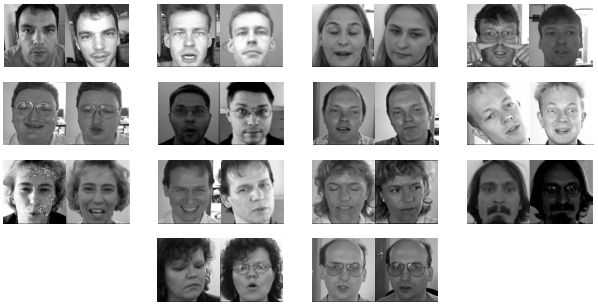


Figure 8. Sample images in BioID.

The experiment in this dataset is used to demonstrate the face recognition effectiveness of SBDBN when important facial features are missing. To every image, we removed a rectangle region of pixels and generate five kinds of facial regions missing images. The locations of missing regions are related with important facial feature regions, including forehead, eye, nose, mouth, and chin. Sample images with important facial regions missing are given in Figure 9.

Figure 9. Sample images with important facial regions missing.

In the above experiments, deep learning models demonstrated a better performance than other existing recognition models. Therefore, in this experiment, we compare proposed SBDBN with other deep learning models. For this dataset, 250 images with different missing regions are randomly selected for each person to form the training set and the rest to form the test set. We perform 5 random splits and report the average results over the 5 trials. Table 3 shows the face recognition accuracy rate of the test dataset. Although the recognition accuracy of DBN and BDBN are both higher than 90%, the recognition accuracy of SBDBN is the highest. This phenomenon is due to DBN and BDBN both trusts on the reliable features and unreliable features. The missing features located in the important facial regions will influence the recognition accuracy.

Table 3. Recognition accuracy rate on test data.

Deep Model	SBDBN	DBN (Zero)	DBN (Mean)	BDBN (Zero)	BDBN (Mean)
Acc.	**97**	94.21	93.93	95.36	94.93

4.3 Experiments on Face Image Dataset StarFace

To further prove the effectiveness of proposed SBDBN in real natural images, we collect and construct a new dataset StarFace from Google, including 120 face images of David Beckham, Victoria Beckham, Tom Cruise, and Julia Roberts. Figure 10

shows some samples images utilized in the retrieval experiment. From these sample images, it is obviously that some important facial feature regions have been occluded. To every occlusion region, we mark them as missing feature regions.

We compare proposed SBDBN with DBN on face retrieval under unsupervised learning framework. Under the unsupervised learning framework, there are only two stages in our model. The semiconducting RBM is utilized in the greedy layer-wise reconstruction stage. Then the parameter space is fine-tuned by minimizing the discrepancy between the original data and its reconstruction. To every category, we randomly select one image with a kind of important facial region missing as the query image.

(a)David Beckham (b) Victoria Beckham

(c) Tom Cruise (d) Julia Roberts

Figure 10. Sample images after preprocessing from StarFace.

The mean value of normalized discounted cumulative gain (NDCG) is utilized to evaluate the retrieval results. From the NDCG scores in Table 4, the SBDBN has better retrieval performance. From Figure 10, it is obvious that our algorithm is effective in image retrieval although some important facial features are missing.

Table 4. Comparison of NDCG scores on StarFace.

	SBDBN	DBN (Zero)	DBN (Mean)
NDCG@10	**0.4829**	0.4388	0.4268
NDCG@20	**0.4240**	0.3762	0.3659

Figure 10. A query image with first ten images which are retrieved out.

5. CONCLUSION AND FUTURE WORK

In this paper, we propose a novel deep learning model, SBDBN for the well-known challenging multimedia content analysis tasks, image recognition with incomplete data. Owing to the semiconducting bilinear discriminant initialization and semiconducting RBM of SBDBN, the recognition ability is promoted by fully exploiting the embedding information according to the reliable features rather than any completion of missing features. In our experiments on real-world image recognition and retrieval tasks, SBDBN shows the distinguishing and robust recognition ability on the incomplete data. In future, we will utilize semiconducting bilinear deep belief network for multimedia content analysis in a large scale dataset with noisy labels.

6. ACKNOWLEDGMENTS

This research was supported by HK PolyU 5183/11E.

7. REFERENCES

[1] Liao, X. Li, H. and Carin, L.. 2007. Quadratically gated mixture of experts for incomplete data classification, In *Proceedings of the International Conference on Machine learning*.

[2] Williams, D., Liao, X. J., Xue, Y., *et al.*. 2007. On classification with incomplete data, *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[3] Chechik, G., Heitz, Elidan, G. G., *et al.*. 2008. Max-margin classification of data with absent features, *Journal of Machine Learning Research*.

[4] Dick, U., Haider, P. and Scheffer, T.. 2008. Learning from incomplete data with infinite imputations, In *Proceedings of the International Conference on Machine learning*.

[5] Williams, D., Liao, X., Xue, Y., *et al.*. 2005. Incomplete-data classification using logistic regression, *Proceedings of the International Conference on Machine learning*.

[6] Shivaswamy, P. K., Bhattacharyya, C. and Smola, A. J.. 2006. Second order cone programming approaches for handling missing and uncertain data. *Journal of Machine Learning Research*.

[7] Chechik, G., Heitz, G., Elidan, G., *et al.*. 2006. Max-margin classification of incomplete data, *Advances in Neural Information Processing Systems*.

[8] Hinton, G.E., and Salakhutdinov, R.R. 2006. Reducing the dimensionality of data with neural networks. In *Science*.

[9] Larochelle, H., Erhan, D., Courville, A., Bergstra, J. and Bengio, Y.. 2007. An empirical evaluation of deep architectures on problems with many factors of variation. In *ICML*.

[10] Smolensky, P.. Information processing in dynamical systems: foundations of harmony theory. 1986. In *Parallel Distributed Processing: Explorations in The Microstructure of Cognition*, vol. 1: Foundations, MIT Press.

[11] Zhong, S.H., Liu, Y., Liu, Y.. 2011. Bilinear deep learning for image classification. In *Proceedings of the 19th ACM International Conference on Multimedia*.

[12] Wang, Z. Xia, D. Chang, E.Y.. 2010. A deep-learning model-based and data-driven hybrid architecture for image annotation. In *VLS-MCMR, ACM*.

[13] Hörster, E. and Lienhart, R.. 2008. Deep networks for image retrieval on large-scale databases, In *ACMMM*.

[14] Malik, N. R. 1995. *Electronic Circuits: Analysis, Simulation, and Design*: Prentice Hall.

[15] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P.. 1998. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*.

[16] Jesorsky, O., Kirchberg, K., Frischholz, R.. 2001. Robust face detection using the hausdorff distance. In *Proceedings of the 3th International Conference on Audio- and Video-based Biometric Person Authentication*.

[17] Boser, B. E., Guyon,I. M. and Vapnik,V.N.. 1992. A training algorithm for optimal margin classifiers. In *COLT*,.

[18] Salakhutdinov, R.R., Hinton,G.E.. 2007. Learning a nonlinear embedding by preserving class neighbourhood structure, In *AISTATS*.

[19] Weston, J., Ratle, F., Collobert, R.. 2008. Deep learning via semi-supervised embedding, In *ICML*.

[20] Wang, C.X., Qing, L.Y., Miao, J., Fang, F., and Chen, X.L.. 2011. Attention Driven Face Recognition: A Combination of Spatial Variant Fixations and Glance, In *Proceedings of IEEE International conference on Automatic Face and Gesture Recognition*.

[21] Yale face database. DOI= http://cvc.yale.edu/projects/yalefaces/yalefaces.html.

[22] IEEE OTCBVS WS Series Bench; Roland Miezianko, Terravic Research Infrared Database.

[23] IEEE OTCBVS WS Series Bench: DOE University Research Program in Robotics under grant DOE-DE-FG02-86NE37968; DOD/TACOM/NAC/ARC Program under grant R01-1344-18; FAA/NSSA grant R01-1344-48/49; Office of Naval Research under grant#N000143010022.