Restricted Boltzmann Machine

Energy-based models

$$p(x) = \frac{e^{-E(x)}}{Z}$$

Intuitively (or exactly) speaking, the probability of a certain state is inversely proportional to its energy.

This probability distribution is often called a Boltzmann distribution.

This is where the name restricted **Boltzmann** machine comes from.

Boltzmann machine

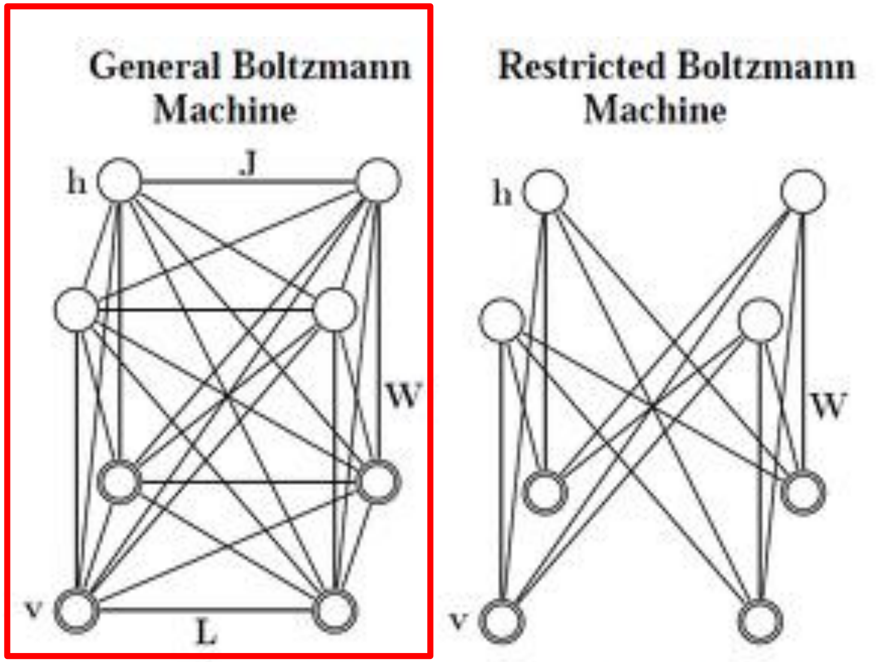


Figure 1: **Left**: A general Boltzmann machine. The top layer represents a vector of stochastic binary "hidden" features and the bottom layer represents a vector of stochastic binary "visible" variables. **Right**: A restricted Boltzmann machine with no hidden-to-hidden and no visible-to-visible connections.

A Boltzmann machine is a network whose elements consist of 0 and 1 (binary units).

Restricted Boltzmann Machine

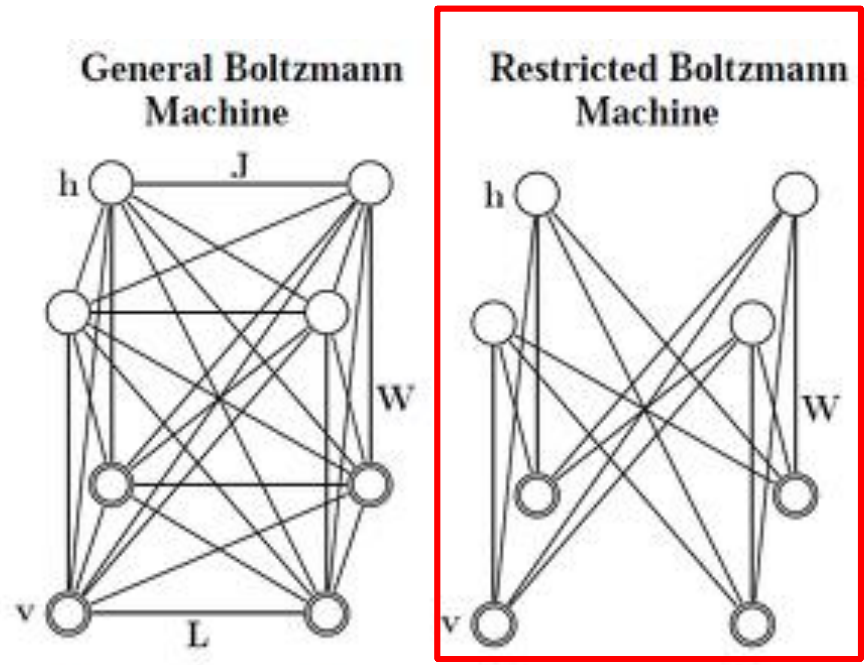
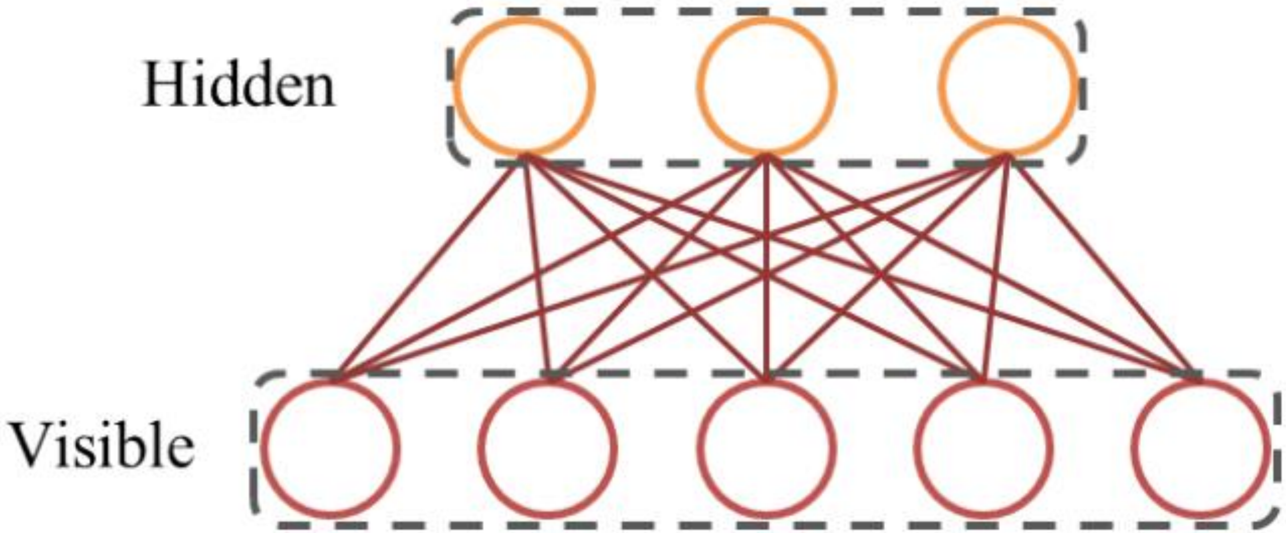


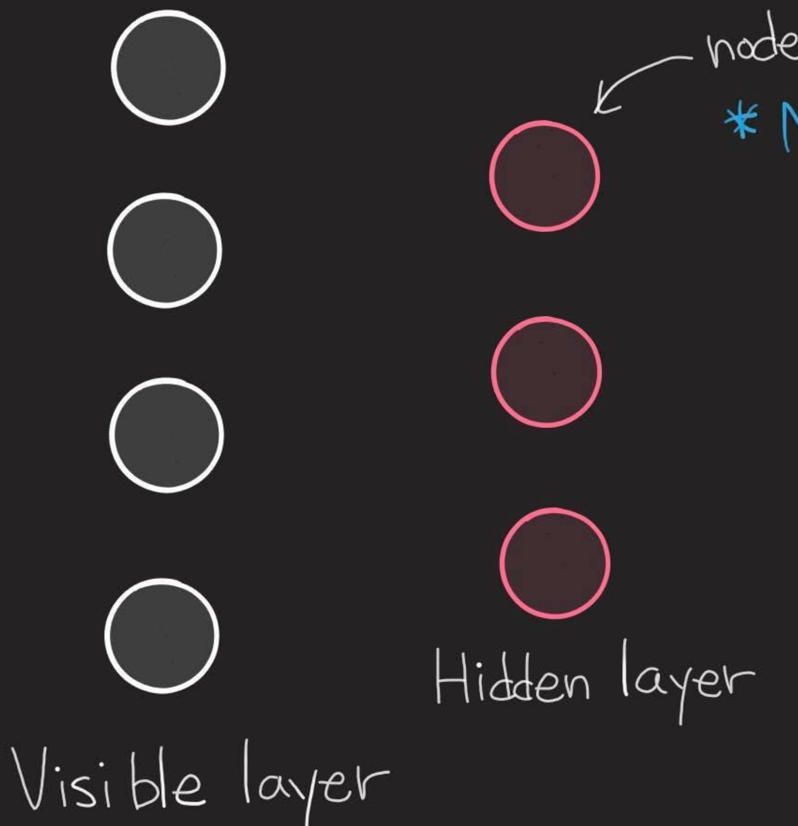
Figure 1: **Left**: A general Boltzmann machine. The top layer represents a vector of stochastic binary "hidden" features and the bottom layer represents a vector of stochastic binary "visible" variables. **Right**: A restricted Boltzmann machine with no hidden-to-hidden and no visible-to-visible connections.

A restricted Boltzmann machine **restricts** connections between visible and hidden nodes.

Restricted Boltzmann Machine



Restricted Boltzmann Machine (RBM)



* Nodes are stachastically determined.

Energy

$$E(v,h\mid\theta) = \boxed{-\sum_{i=1}^{D}\sum_{j=1}^{F} w_{ij} v_i h_j - \sum_{i=1}^{D} b_i v_i - \sum_{j=1}^{F} a_j h_j}$$

Probability

$$P(v,h\mid\theta) = \boxed{\frac{1}{Z(\theta)} \exp\left(-E(v,h\mid\theta)\right)}$$

normalizing constant

$$Z(\theta) = \sum_{v,h} \exp\left(-E(v,h\mid\theta)\right)$$

Conditional distributions

$$P(h_j = 1 \mid v) = \boxed{g\left(\sum_i w_{ij} v_i + a_j\right)}, \quad P(v_i = 1 \mid h) = g\left(\sum_j w_{ij} h_j + b_i\right)$$

$$g(x) = 1/(1+\exp(-x))$$

$$\hat{\Theta} = \arg\max_{\theta} \boxed{\log P(v \mid \theta)}$$

$$* \ \frac{\partial}{\partial\theta}\log f(\theta) = \frac{1}{f(\theta)}\frac{\partial}{\partial\theta} f(\theta)$$

$$\Rightarrow \frac{\partial}{\partial\theta}\left(-\log P(v)\right) = \frac{\partial}{\partial\theta}\left(-\log \boxed{\sum_h P(v,h)}\right)$$

sum rule

$$= \frac{\partial}{\partial\theta}\left(-\log \underbrace{\boxed{\sum_h \frac{1}{Z}\exp(-E(v,h))}}_{=f(\theta)}\right)$$

$$= \boxed{-\frac{Z}{\sum_h \exp(-E(v,h))}\left(\sum_h \frac{1}{Z}\frac{\partial}{\partial\theta}\exp(-E(v,h)) - \sum_h \frac{1}{Z^2}\frac{\partial Z}{\partial\theta}\exp(-E(v,h))\right)}$$

$$= \boxed{-\frac{\cancel{Z}}{\cancel{\sum_h \exp(-E(v,h))}}\left(\frac{\sum_h \frac{1}{\cancel{Z}}\frac{\partial}{\partial\theta}\exp(-E(v,h))}{\sum_h \exp(-E(v,h))} - \sum_h \frac{1}{Z^{2}}\frac{\partial Z}{\partial\theta}\cancel{\exp(-E(v,h))}\right)}$$

$$\frac{\partial}{\partial\theta}(-\log p(v)) = \sum_h \left(\underbrace{\boxed{\frac{\exp(-E(v,h))}{\sum_{h'} \exp(-E(v,h'))}}}_{= p(v|h)} \cdot \frac{\partial}{\partial\theta} E(v,h) \right) + \frac{1}{Z}\frac{\partial Z}{\partial\theta}$$

$$* \; Z(\theta) = \sum_{v,h} \exp(-E(v,h))$$

$$= \sum_h p(v|h) \frac{\partial}{\partial\theta} E(v,h) + \frac{1}{Z}\frac{\partial Z}{\partial\theta}$$

$$= \sum_h p(v|h) \frac{\partial}{\partial\theta} E(v,h) - \underbrace{\frac{1}{Z} \sum_{v,h} \exp(-E(v,h))}_{= p(v,h)} \frac{\partial}{\partial\theta} E(v,h)$$

$$= E\left[\frac{\partial}{\partial\theta} E(v,h) \,\middle|\, v\right] - E\left[\frac{\partial}{\partial\theta} E(v,h)\right]$$

positive phase

: Try to lower the energy of observed v.

negative phase

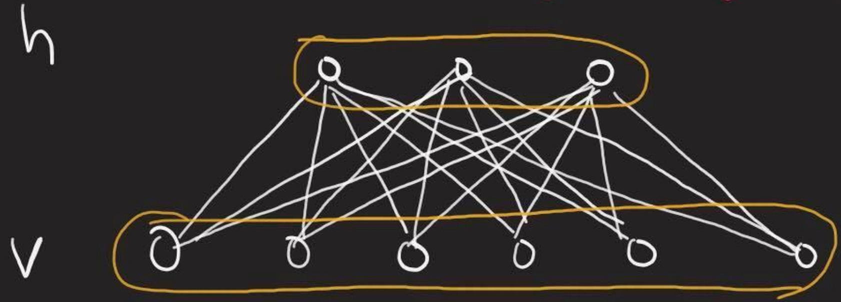
: Try to increase the energy of all (v,h) we don't know h

$$\frac{\partial}{\partial\theta}(-\log p(v)) = \boxed{E\left[\frac{\partial}{\partial\theta}E(v,h) \mid v\right] - E\left[\frac{\partial}{\partial\theta}E(v,h)\right]}$$

This what we know

We don't know both of v and h.

$$*\ E(v,h|\theta) = v^T \underbrace{W}_{\theta} h + \underbrace{b}_{\theta}^T v + \underbrace{a}_{\theta}^T h$$

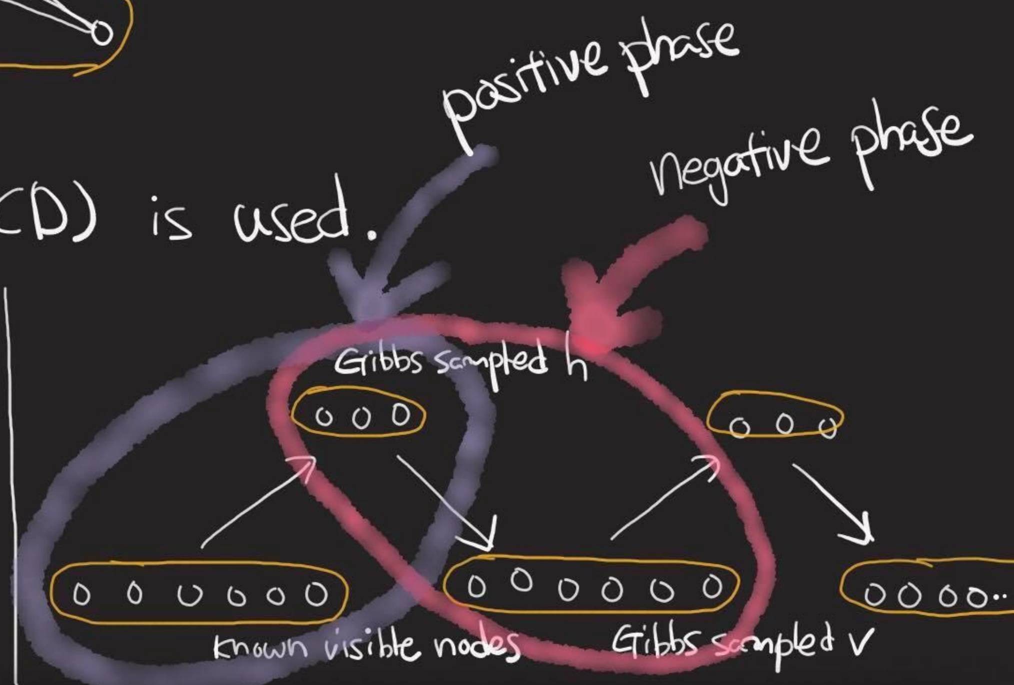


Contrastive Divergence (CD) is used.

$$\frac{\partial}{\partial w_{ij}}E(v,h|\theta) = v_i h_j$$

$$\frac{\partial}{\partial b_i}E(v,h|\theta) = v_i$$

$$\frac{\partial}{\partial a_j}E(v,h|\theta) = h_j$$



Deep Belief Network (DBN)

A DBN simply stacks RBM, level by level.

Training is done in a layer-wise manner.

It is an unsupervised learning method.

DBN had been widely used as a pre-training method until 2014..

More information

http://enginius.tistory.com/315

Boltzmann Machine은 [0,1]의 값을 갖는 binary unit들로 이루어진 network를 의미한다.

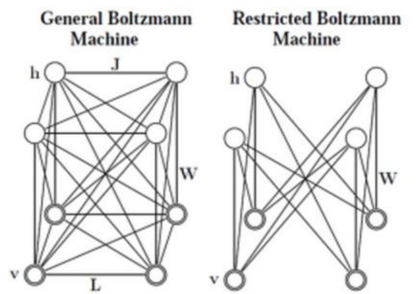


Figure 1: **Left**: A general Boltzmann machine. The top layer represents a vector of stochastic binary "hidden" features and the bottom layer represents a vector of stochastic binary "visible" variables. **Right**: A restricted Boltzmann machine with no hidden-to-hidden and no visible-to-visible connections.

위의 Figure1은 Boltzmann Machine을 안다면 누구나 한번쯤 봤을 그림이다. 먼저 왼쪽의 모형이 general BM이다. 이 BM의 특징은 full-connectivity에 있다. 그리고 오른쪽에 모형이 restricted BM이다. 이 모형은 visible node와 hidden node를 분리시켰다. 이것이 BM과 RBM의 차이이다. 이 간단한 차이로 RBM은 실제 구현이 가능하고, BM은 구현이 매우 어렵다.

앞서 설명하였듯이 BM에서 node는 0또는 1의 binary한 값을 갖는다. 그리고 각 node사이에는 symmetric하게 link가 있는데, 이 link에는 weight가 존재한다. 이 weight의 값은 굳이 양수일 필요 없이 모든 값을 가질 수 있다. RBM의 경우 각 node를 visible과 hidden으로 나눠 놓았고, 여기서 visible node는 우리의 data가 들어가는 곳을 의미하고, hidden의 경우 우리는 각 node가 1이 될 확률만을 알게된다.

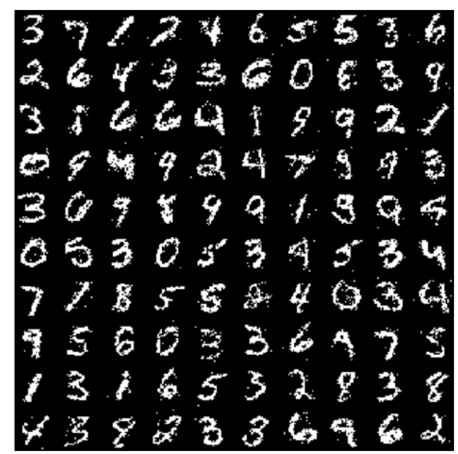
BM의 상태는 에너지를 통해서 설명될 수 있는데 엔트로피와 마찬가지로 에너지가 높을 수록 그 존재 확률이 낮아지게 된다. 먼저 특정 상태의 에너지는 다음과 같이 표시된다.

The energy of the state $\{\mathbf{v}, \mathbf{h}\}$ is defined as:

$$E(\mathbf{v}, \mathbf{h}; \theta) = -\mathbf{v}^\top \mathbf{W}^1 \mathbf{h}^1 - \mathbf{h}^{1\top} \mathbf{W}^2 \mathbf{h}^2, \qquad (1)$$

그리고 이때 해당 상태의 확률은 다음과 같다.

$$P(\mathbf{v}; \theta) = \frac{P^*(\mathbf{v}; \theta)}{\mathcal{Z}(\theta)} = \frac{1}{\mathcal{Z}(\theta)} \sum \exp\left(-E(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2; \theta)\right).$$



Applications

My first conference paper

International Conference

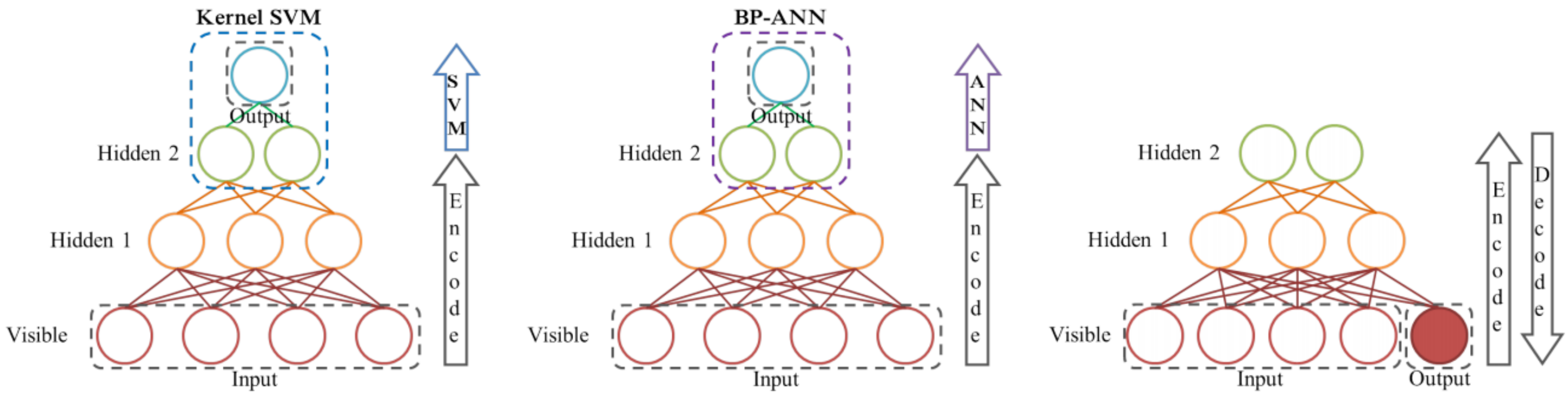
- **Sungjoon Choi**, Kyungjae Lee, Songhwai Oh, "**Robust Learning From Demonstration Using Leveraged Gaussian Processes and Sparse Constrained Opimization**", in IEEE Conference on Robotics and Automation (ICRA), 2016
- **Sungjoon Choi**, Eunwoo Kim, Kyungjae Lee, Songhwai Oh, "**Leveraged Non-Stationary Gaussian Process Regression for Autonomous Robot Navigation**", in IEEE Conference on Robotics and Automation (ICRA), 2015

Sungjoon Choi, Eunwoo Kim, Songhwai Oh, "**Human Behavior Prediction for Smart Homes Using Deep Leering**", ROMAN, 2013

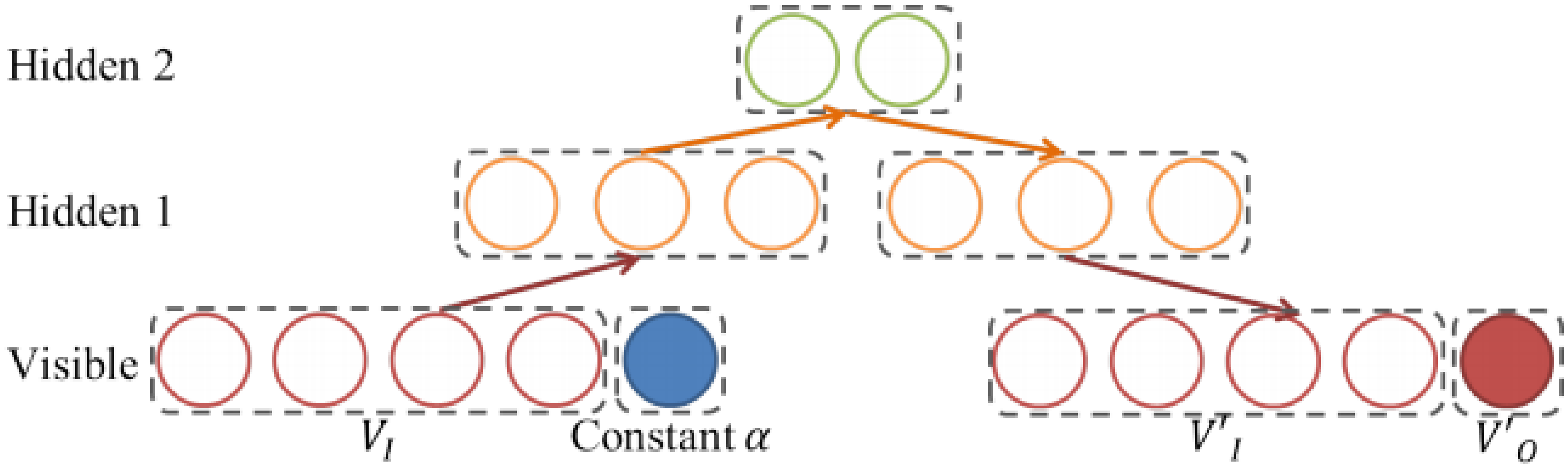
Process Motion Model Using l1-Norm Based Low-Rank Kernel Matrix Approximation" in Proc. of the IEEE International Conference on Intelligent Robots and Systems (IROS), 2014.

- **Sungjoon Choi**, Eunwoo Kim, Songhwai Oh, "**Real-Time Navigation in Crowded Dynamic Environments Using Gaussian Process Motion control**", in IEEE Conference on Robotics and Automation (ICRA), 2014
- **Sungjoon Choi**, Mahdi Jadaliha, Jongeun Choi, Songhwai Oh, "**Distributed Gaussian Process Regression for Mobile Sensor Networks Under Localization Uncertainty**", in IEEE Conference on Decision and Control (CDC), 2013
- **Sungjoon Choi**, Eunwoo Kim, Songhwai. Oh, "**Human Behavior Prediction for Smart Homes Using Deep Learning**", in IEEE International Symposium on Robot and Human Interactive Communications (ROMAN), 2013

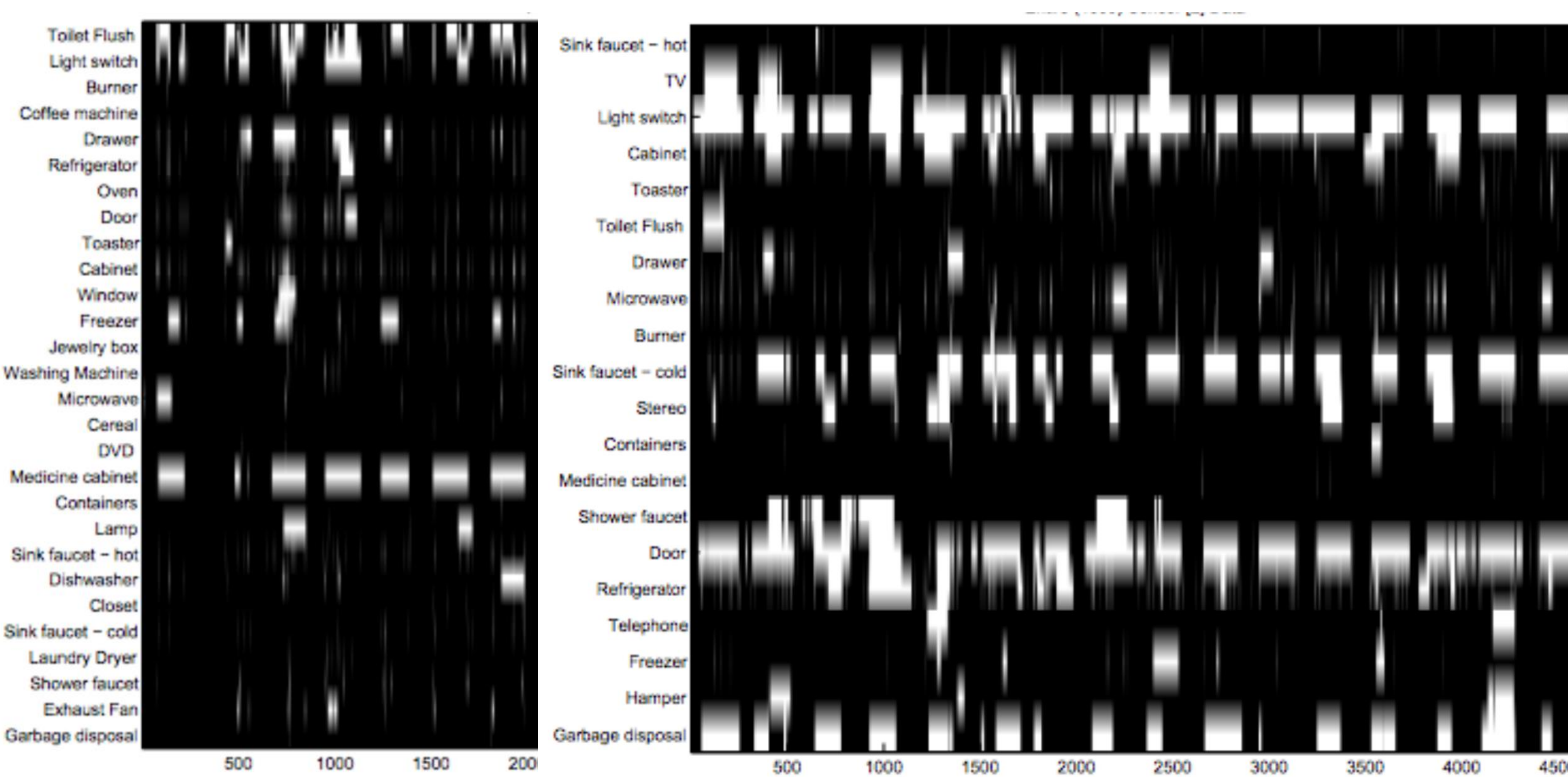
DBN-Reconstruct



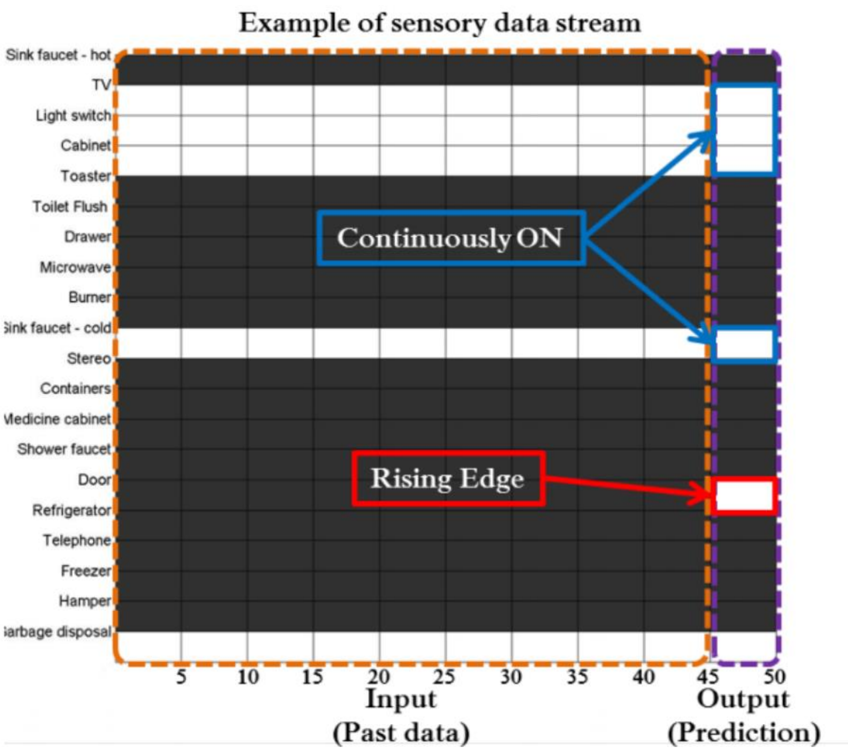
DBN-Reconstruct

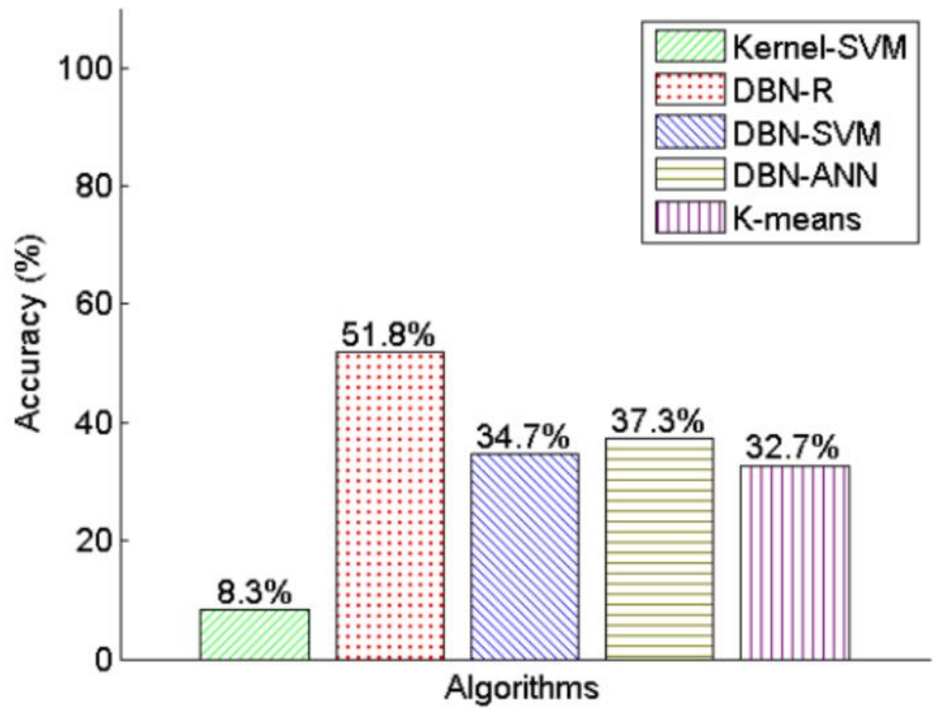


Use to predict Home Dataset

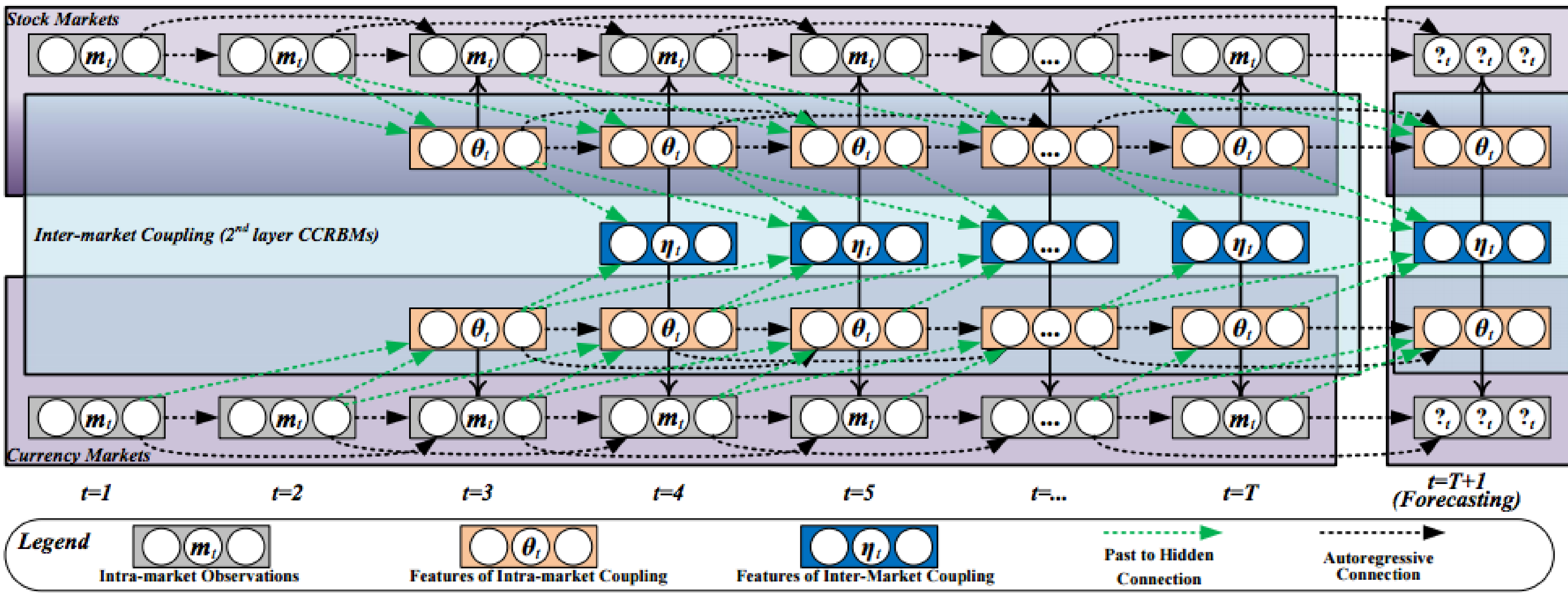


Result


Example of sensory data stream
Sink faucet - hot
TV
Light switch
Cabinet
Toaster
Toilet Flush
Drawer
Microwave
Burner
Sink faucet - cold
Stereo
Containers
Medicine cabinet
Shower faucet
Door
Refrigerator
Telephone
Freezer
Hamper
Garbage disposal
Continuously ON
Rising Edge
5 10 15 20 25 30 35 40 45 50
Input
(Past data)
Output
(Prediction)


Accuracy (%)
0 20 40 60 80 100
Algorithms
Kernel-SVM
DBN-R
DBN-SVM
DBN-ANN
K-means
8.3%
51.8%
34.7%
37.3%
32.7%

Finance forecasting



Conditional RBM

Conditional Restricted Boltzmann Machines

In order to model temporal coupling, we need to use CRBM (Taylor 2009) instead of RBM. The CRBM assign a probability to any joint setting of the visible units $\mathbf{v}$ and hidden units $\mathbf{h}$ conditional on $\mathbf{u}$ by

$$P(\mathbf{v}, \mathbf{h} \mid \mathbf{u}) = exp(-E(\mathbf{v}, \mathbf{h}, \mathbf{u}))/Z \quad (5)$$

where Z is a normalization constant and $E(\mathbf{v}, \mathbf{h}, \mathbf{u})$ is an energy function:

$$E(\mathbf{v}, \mathbf{h}, \mathbf{u}) = -\mathbf{v}^{\mathrm{T}}\mathbf{W}\mathbf{h} - \mathbf{u}^{\mathrm{T}}\mathbf{A}\mathbf{v} - \mathbf{u}^{\mathrm{T}}\mathbf{B}\mathbf{h} - \mathbf{a}^{\mathrm{T}}\mathbf{v} - \mathbf{b}^{\mathrm{T}}\mathbf{h} \quad (6)$$

where $\mathbf{v} \in \{0,1\}^D$ is a vector of binary visible units, $\mathbf{h} \in \{0,1\}^F$ is a vector of binary hidden units and $\mathbf{u} \in \{0,1\}^D$ is a vector of binary visible units. $\mathbf{W} \in \mathbb{R}^{D\times F}$ encodes the interactions between $\mathbf{v}$ and $\mathbf{h}$, $\mathbf{A} \in \mathbb{R}^{D\times D}$ encodes the interactions between $\mathbf{u}$ and $\mathbf{v}$, $\mathbf{B} \in \mathbb{R}^{D\times F}$ encodes the interactions between $\mathbf{u}$ and $\mathbf{h}$. $\mathbf{a} \in \mathbb{R}^D$ and $\mathbf{b} \in \mathbb{R}^F$ denote the biases of $\mathbf{v}$ and $\mathbf{h}$ separately. Hence, $\Omega = \{\mathbf{W}, \mathbf{A}, \mathbf{B}, \mathbf{a}, \mathbf{b}\}$ are the model parameters that need to learn.