

LA-UR-16-28813

Approved for public release; distribution is unlimited.

Title: Generative Modeling for Machine Learning on the D-Wave

Author(s): Thulasidasan, Sunil

Intended for: Report

Issued: 2016-11-15

Disclaimer:

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the Los Alamos National Security, LLC for the National Nuclear Security Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.



Generative Modeling for Machine Learning on the D-Wave

Sunil Thulasidasan

Information Sciences (CCS-3)
Los Alamos National Laboratory
sunil@lanl.gov

UNCLASSIFIED

Generative Models

- Two approaches in machine learning:
 - Discriminative: Learn $P(y|x)$
 - Generative: Learn $P(y,x)$
- Discriminative models are easier to train, but generative models are more powerful because in some sense it “understands” the world better.

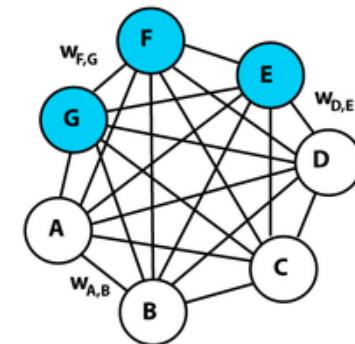
UNCLASSIFIED

Boltzmann Machines: A Generative Model

- Energy based model. Assign a scalar energy value to configurations of interest
- Associate lower energy with plausible configurations
- Probability given by

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

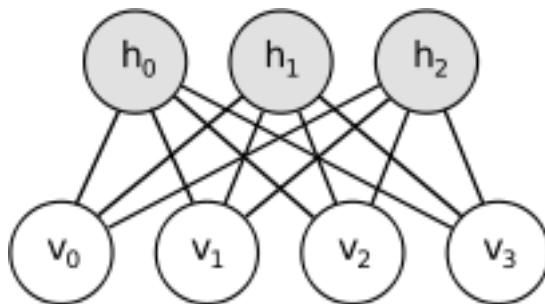
- Consists of **visible units** (data) and **hidden units** (capture dependencies between data)
General Boltzmann machines have arbitrary connectivity. Hard to train.



UNCLASSIFIED

Restricted Boltzmann Machines

- Restrict connections to occur only between pairs of visible and hidden units. No connections among visible units or hidden units.

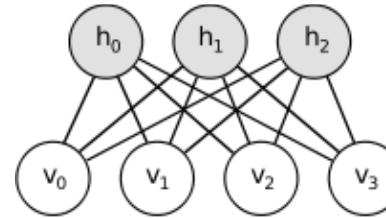


- h 's are independent given v and v 's are independent given h (markov property)

UNCLASSIFIED

Restricted Boltzmann Machines

- Energy given by



$$E(v, h) = -b'v - c'h - h'Wv$$

- Conditional independence implies:

$$p(h|v) = \prod_i p(h_i|v)$$

$$p(v|h) = \prod_j p(v_j|h)$$

- Once we know the parameters (b, c, W) generating data is easy

UNCLASSIFIED

Learning Parameters: RBM Training

- Learn parameters that maximize log-likelihood of data. Assuming data independence, we have

$$\arg \max_{(w,b,c)} \ell(w, b, c) = \sum_{t=1}^n \log P(v^t)$$

- The gradient is given by

$$\begin{aligned} \nabla_{\theta} \ell(\theta) &= \sum_{t=1}^n \mathbb{E}_{p(h|v)} \left[\nabla_{\theta} (-E(v^t, h)) \right] \\ &\quad - n \mathbb{E}_{p(v,h)} (\nabla_{\theta} (-E(v^t, h))) \end{aligned}$$

UNCLASSIFIED

RBM Training

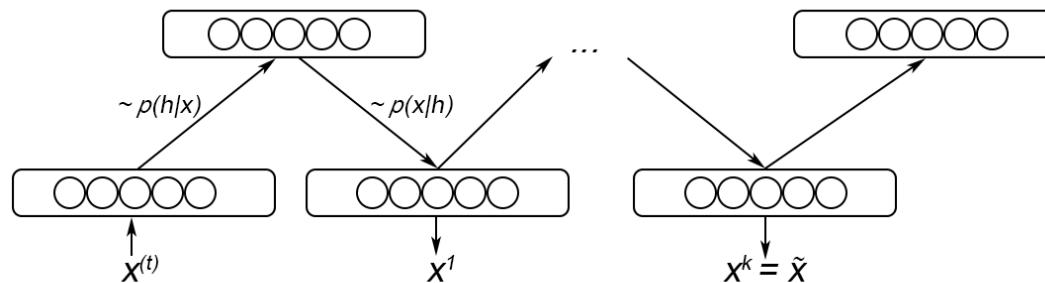
$$\begin{aligned}\nabla_{\theta} \ell(\theta) &= \sum_{t=1}^n \mathbb{E}_{p(h|v)} [\nabla_{\theta} (-E(v^t, h))] \\ &\quad - n \mathbb{E}_{p(v,h)} (\nabla_{\theta} (-E(v^t, h)))\end{aligned}$$

- Gradient depends on joint distribution
- Intractable since it involves the partition function Z
- To avoid this, use Gibb's sampling to sample from joint (Boltzmann distribution). Involves running a Markov chain to convergence (Markov Chain Monte Carlo or MCMC)

UNCLASSIFIED

Practical Ways to Train RBM

- Instead of running MCMC to convergence, run it for just a few (k) steps. Sample from this distribution (Contrastive Divergence)
- In practice, k (number of steps) is < 100 . Some times even 1 step works well !



UNCLASSIFIED

D-Wave as a Boltzmann Sampler

- D-Wave is a physical Boltzmann machine
- In theory, should give samples from a Boltzmann distribution (parameterized by some *effective* temperature) after annealing
- Approach: Instead of Gibbs's sampling, map RBM onto D-Wave and sample from solution states

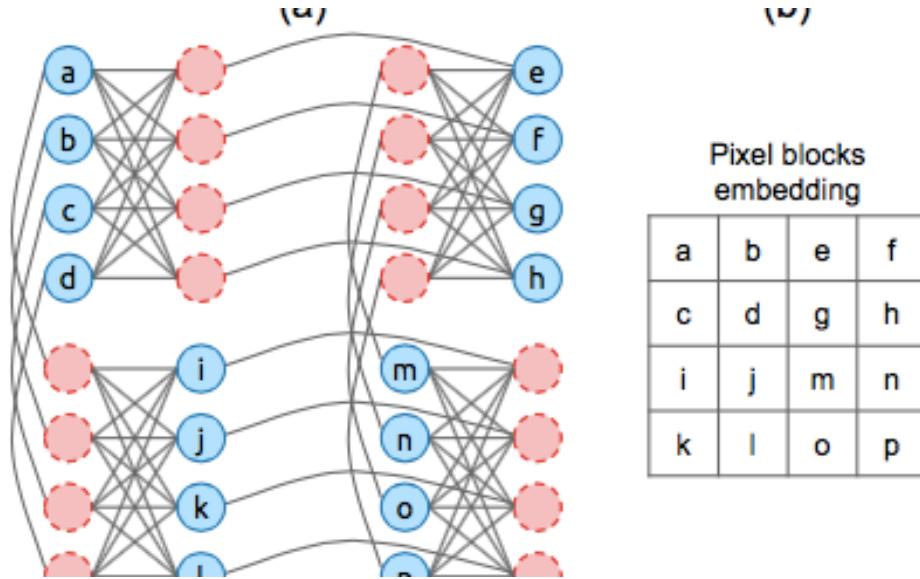
UNCLASSIFIED

Mapping RBM onto the D-Wave

- RBM's are full bipartite graphs. D-Wave has sparse connectivity.
- Using logical qubits, can implement up to 48x48 bipartite graph. Lots of qubits lost
- For this work, no qubit chaining. Map each pixel of the training image directly onto a qubit

UNCLASSIFIED

Chimera Restricted RBM



Same embedding as in Benedetti et al
(2015) and Doulin et al (2014)

UNCLASSIFIED

Mapping binary RBM to Ising Model

- RBM's are binary $\{0,1\}$ units.
- To map this to Ising model, where units are in $\{+1,-1\}$ we use the following transformation described in Domoulin (2014)

$$W' = \frac{W}{4}$$

$$b'_i = \frac{1}{2}b_i + \frac{1}{4} \sum_j W_{ij}$$

$$c'_i = \frac{1}{2}c_i + \frac{1}{4} \sum_j W_{ji}$$

UNCLASSIFIED

Experiments

- Basic Outline (classical side):
 - Initialize visible units and hidden units
 - Clamp visible units to a training sample
 - Run few steps of contrastive divergence for gradient
 - Update parameters
 - Run till convergence
- On the D-Wave, same process except we do not run contrastive divergence, but sample from solution states

UNCLASSIFIED

Data

- MNIST (handwritten digits 0-9)
- Train on 1000 digits and learn features.
- And then see if the model can generate its own representations.



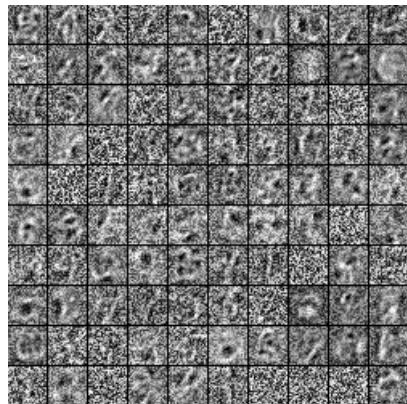
UNCLASSIFIED

D-Wave Effective Temperature, Parameter Noise etc

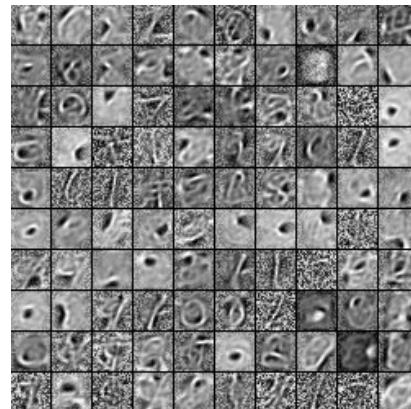
- D-Wave *effective temperature* is different from *physical temperature*. Estimate this via sampling and then find a best fit
- Did not do any corrections for weight and bias noise.
- Effective temperature also fluctuates during training (Benedetti et al 2015). Did not correct for this.

UNCLASSIFIED

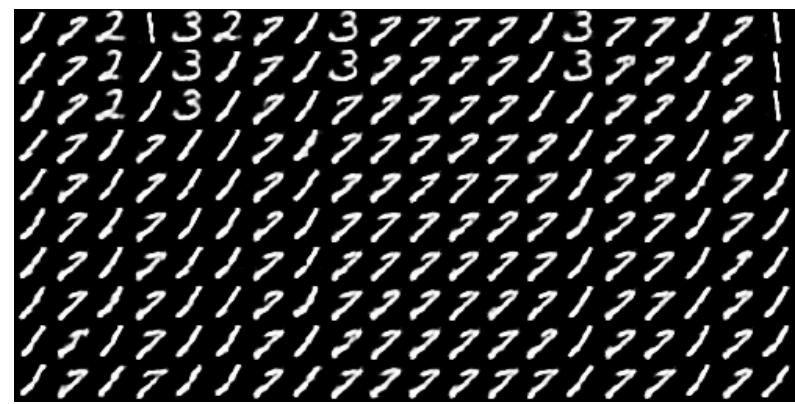
Experiments: Contrastive Divergence (CD) 1 Step



Filters learned
after epoch 1



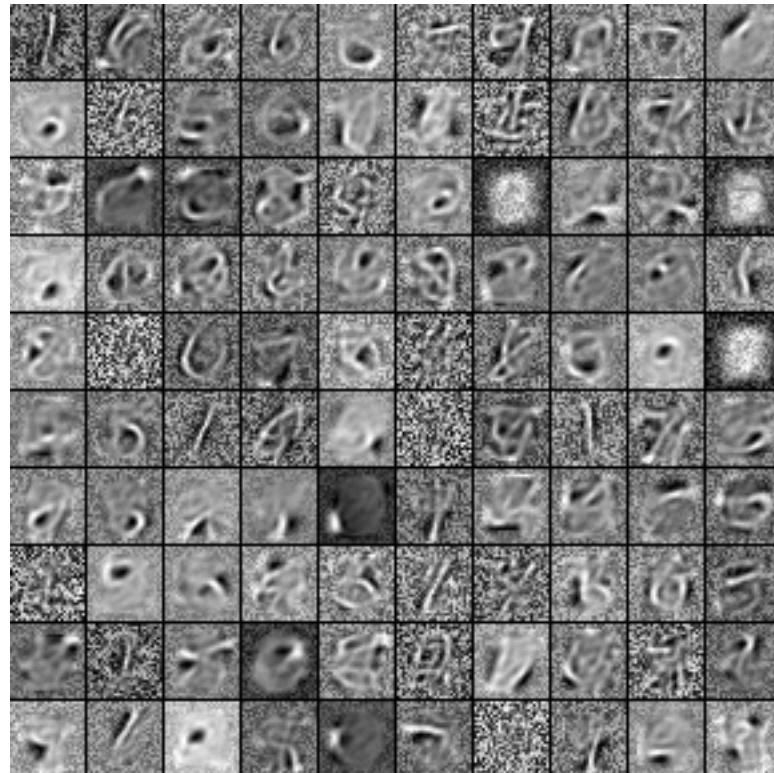
Filters learned
after epoch 15



Generated Images

UNCLASSIFIED

After 50 Steps of CD



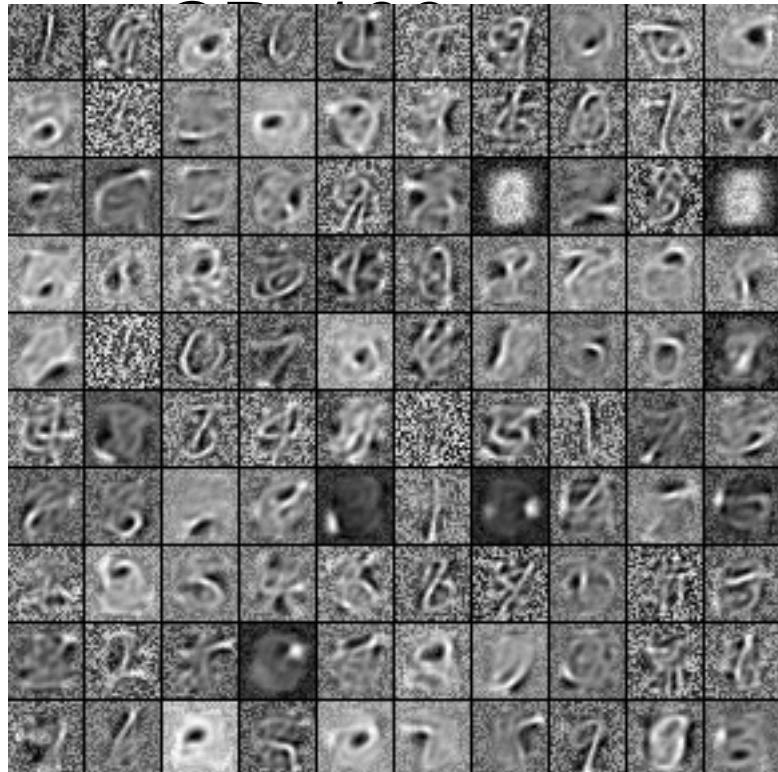
Filters

6	9	9	4	6	1	6	9	4	4	6	4	9	9	4	7	6	9	6	1
6	9	9	9	6	1	6	9	9	7	6	6	7	9	9	9	6	9	6	1
6	9	9	9	6	1	6	9	9	9	6	6	9	7	9	9	6	9	6	1
6	9	9	9	6	1	6	6	9	9	9	6	9	4	9	9	6	9	6	1
6	9	9	9	6	1	6	6	9	9	6	6	9	6	9	9	6	9	6	1
6	9	9	9	6	1	6	6	9	9	6	6	9	6	9	9	6	9	6	1
6	9	9	9	6	1	6	6	9	9	6	6	9	6	9	9	6	9	6	1
6	9	9	9	6	1	6	6	9	9	6	6	9	6	9	9	6	9	6	1
6	9	9	9	6	1	6	6	9	9	6	6	9	6	9	9	6	9	6	1
6	9	9	9	6	1	6	6	9	9	6	6	9	6	9	9	6	9	6	1

Generated Images

UNCLASSIFIED

After 100 Steps of CD



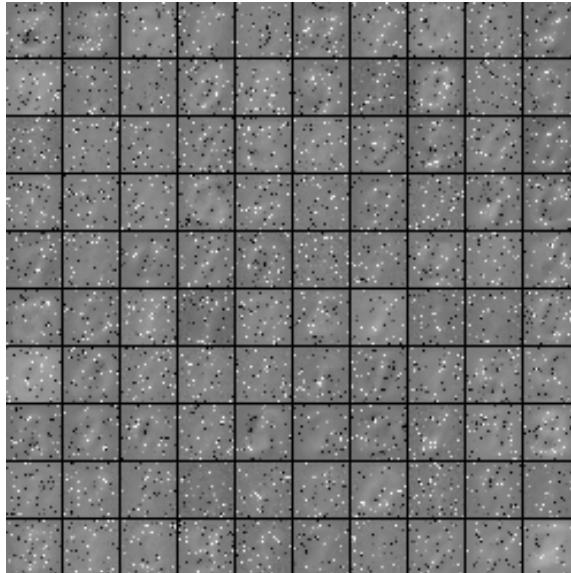
Filters



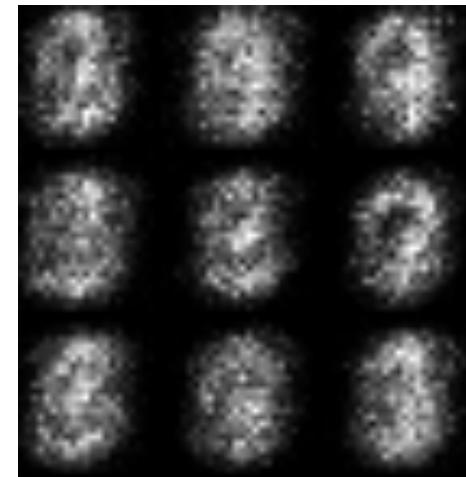
Generated Images

UNCLASSIFIED

D-Wave (Experiment 1)



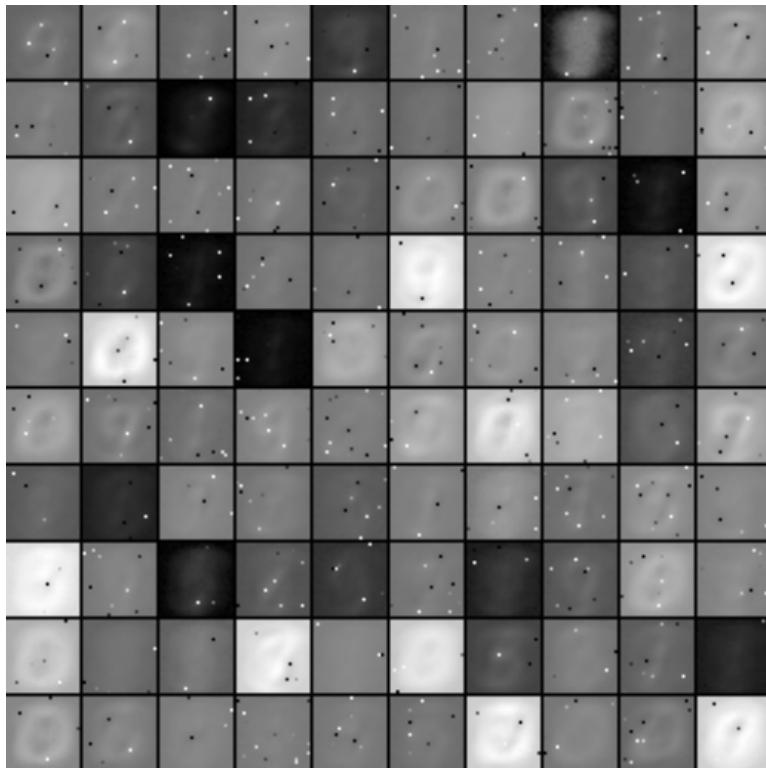
Filters learnt are sparse due to sparse connectivity graph



Generated images are noisy and largely indistinguishable from one another

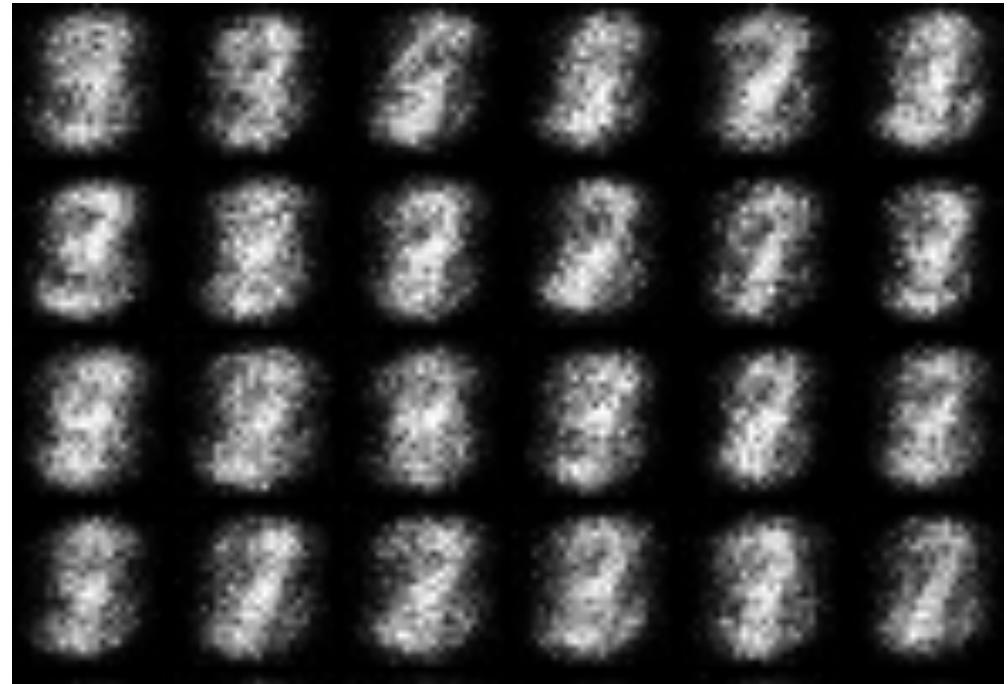
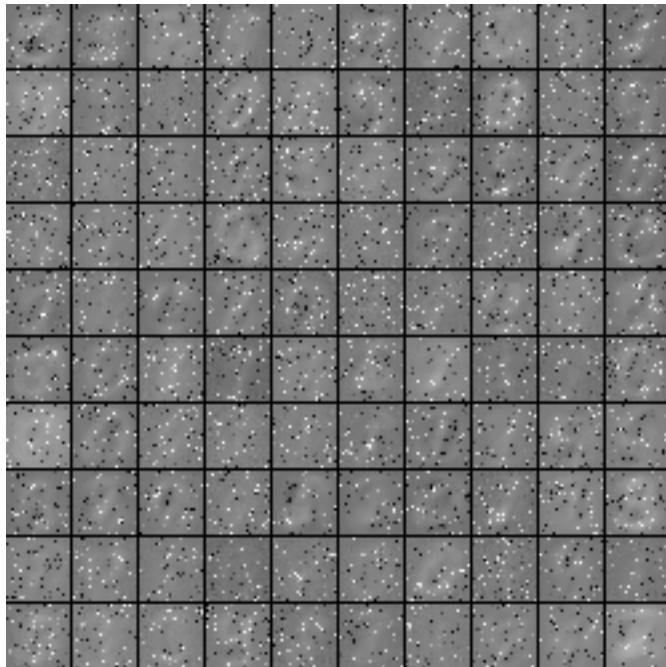
UNCLASSIFIED

D-Wave (Experiment 2)



UNCLASSIFIED

D-Wave (Experiment 3)



UNCLASSIFIED

D-Wave Observations

- Effective temperature and parameter noise affect modeling
- However, limited connectivity is a *much* bigger problem
 - RBM's are robust to limited connections. But the D-Wave has less than 1% of connections of a complete bipartite graph.
 - Qubit chaining can overcomes connectivity issues, but then image has to be significantly down-sampled.

UNCLASSIFIED

References

- **Dumoulin et al 2014.** On the Challenges of Physical Implementations of RBMs. *In Proceedings of AAAI 2014*
- **S. Adachi, M. Henderson, 2015.** Application of Quantum Annealing to Training of Deep Neural Networks
- **Benedetti et al. 2015.** Estimation of Effective Temperatures in Quantum Annealers for Sampling Applications: A Case Study with Possible Applications in Deep Learning

UNCLASSIFIED