Generative Modeling for Machine Learning on the D-Wave

Sunil Thulasidasan
CCS-3, sunil@lanl.gov

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.

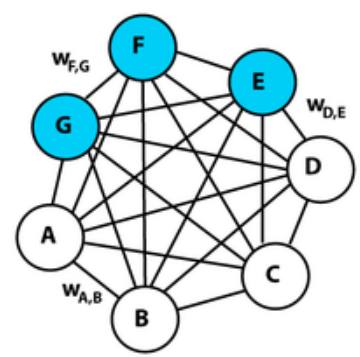
Boltzmann Machines: A generative Model

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

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

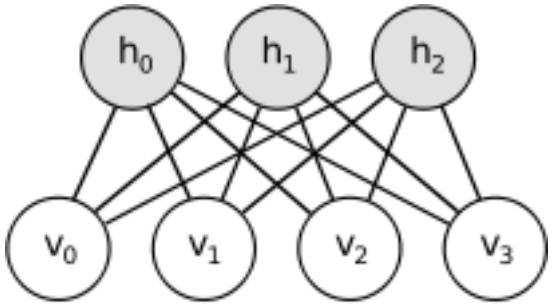
- Consists of **visible units** (data) and **hidden units** (capture dependencies between data)

General Boltzmann machines have arbitrary connectivity. Hard to train.



Restricted Boltzmann Machines

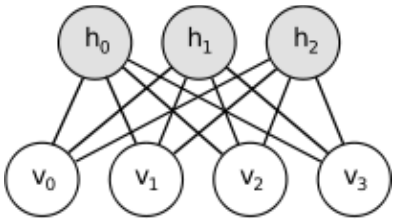
- Restrict connections to occur only between pairs of visible and hidden units. No connections among visible units or hidden units.



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

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

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

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

RBM Training

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

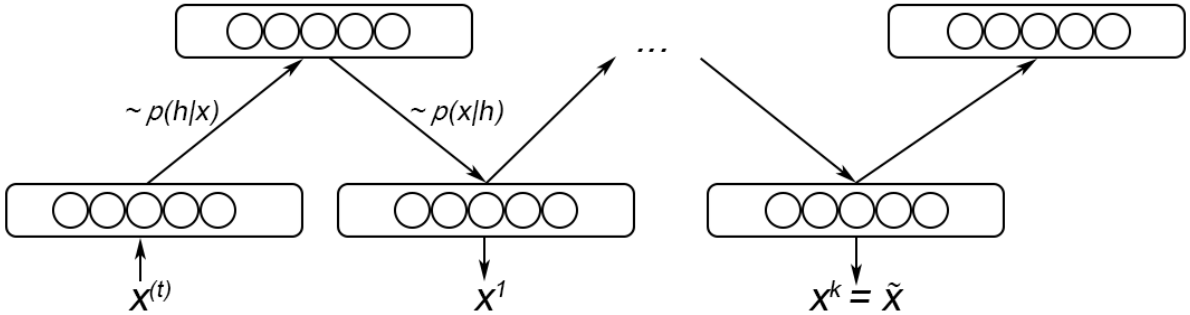
Gradient depends on joint distribution

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

Practical Ways to Train RBM

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



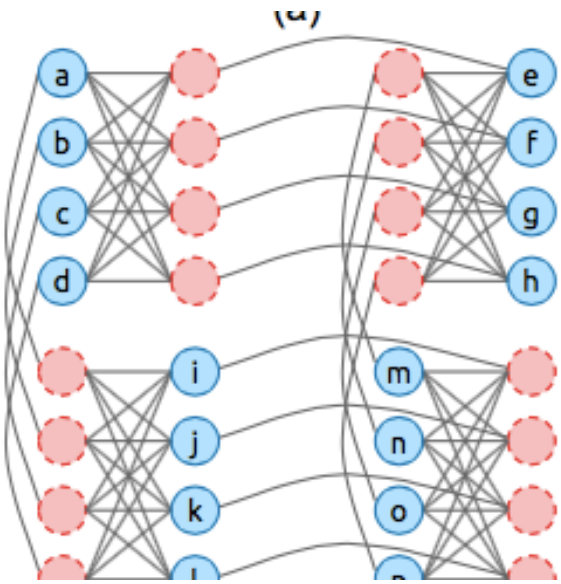
D-Wave as a Boltzmann Sampler

- D-Wave is a physical boltzmann machine
- In theory, should give samples from a boltzmann (parametrized by some *effective* temperature) distribution after annealing
- Idea: instead of Gibbs's sampling map RBM into D-Wave and sample from solution states

Mapping RBM into D-Wave

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

Chimera Restricted RBM



(b)

Pixel blocks embedding

a	b	e	f
c	d	g	h
i	j	m	n
k	l	o	p

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

Mapping binary RBM to Ising

- 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)

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

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

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

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

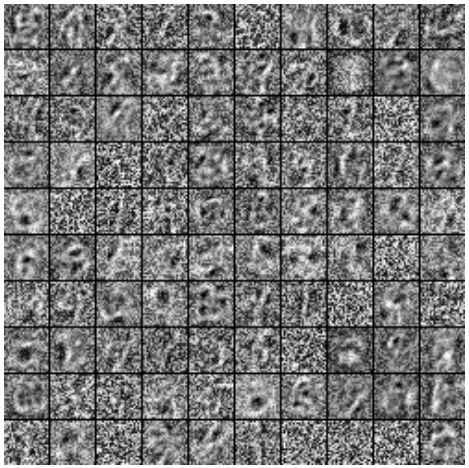
Data

- MNIST (handwritten digits 0-9)
- Train on 1000 digits. Learn features. And then see if the model can generate it's own representations.

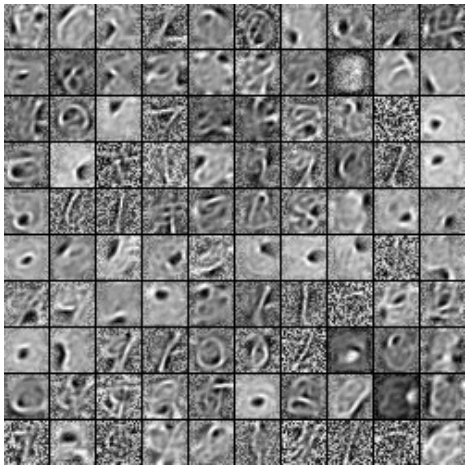
D-Wave effective temperature, parameter noise etc

- 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 corret for this.

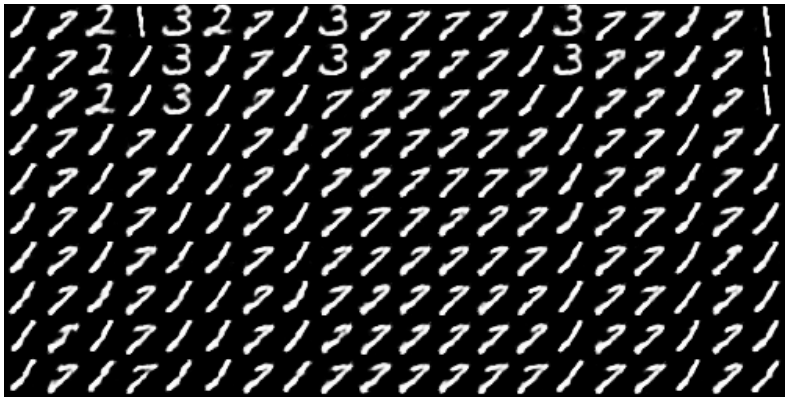
Experiments: Contrastive Divergence (CD) 1 Step



Epoch 1 filters

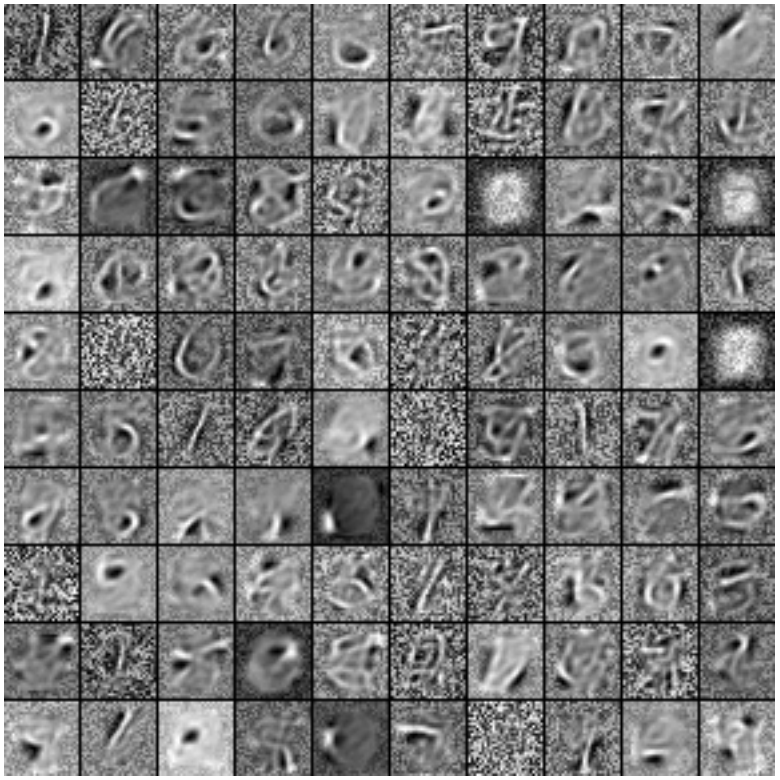


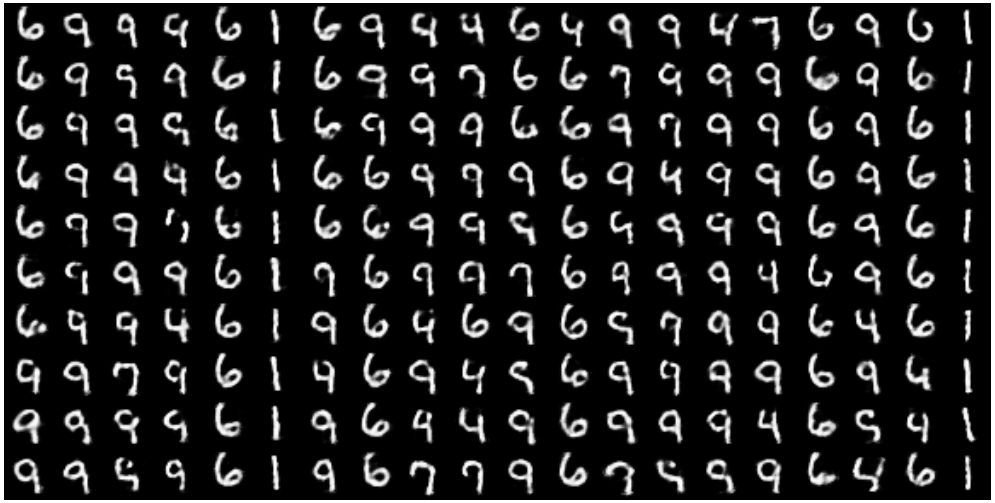
Epoch 15 filters



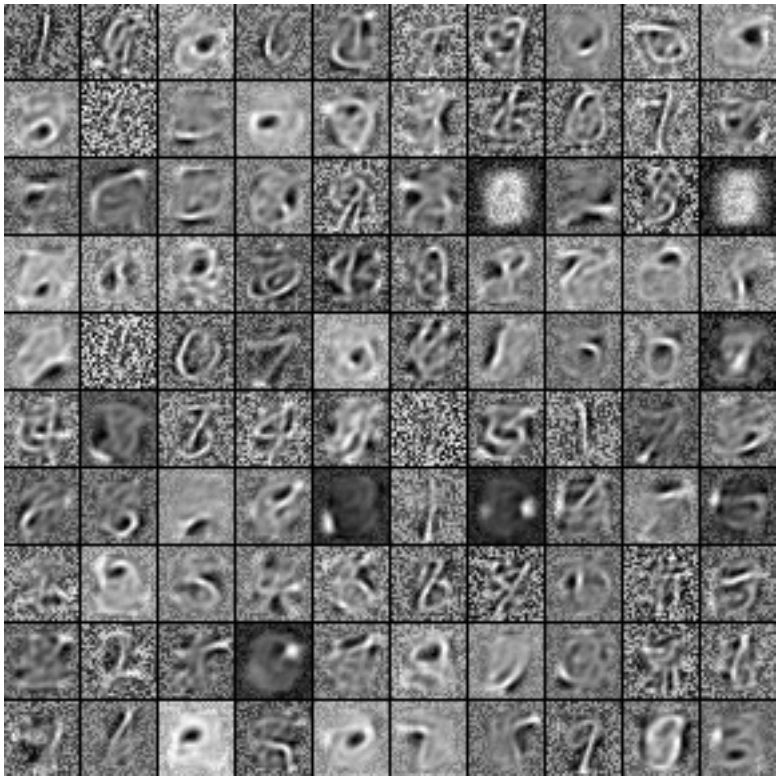
Generated Images

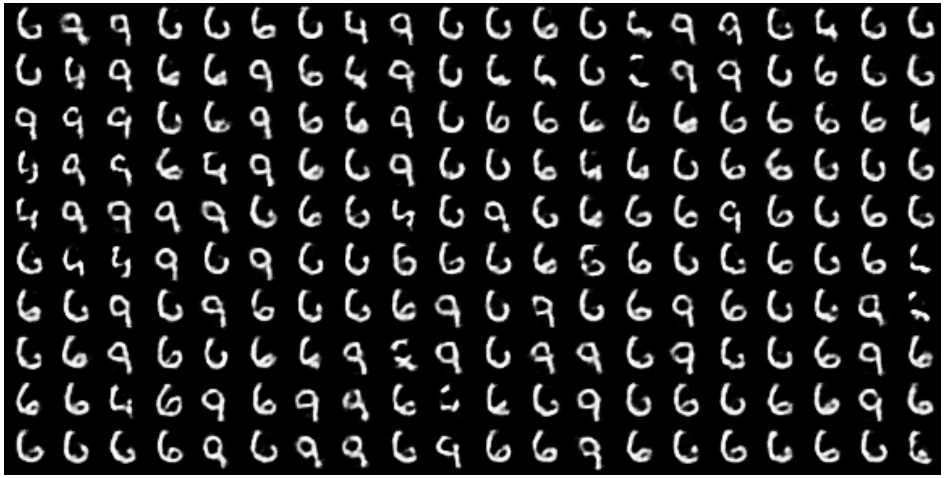
CD-50



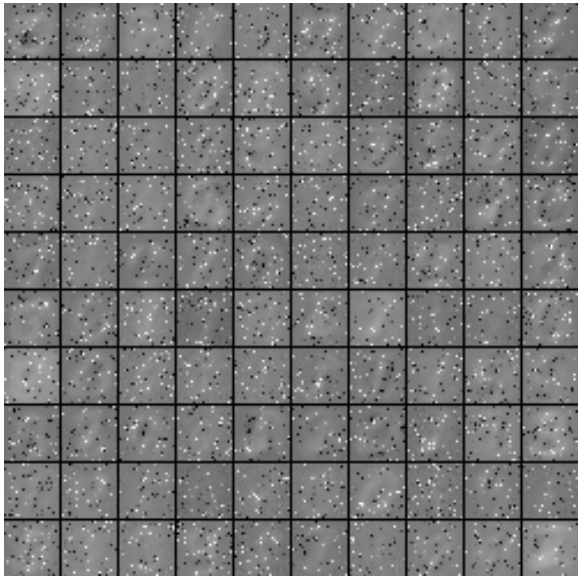


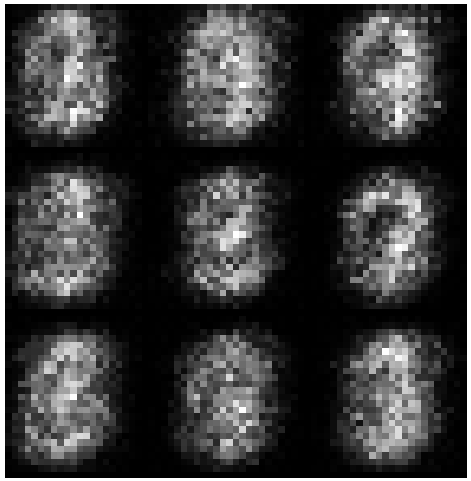
CD-100





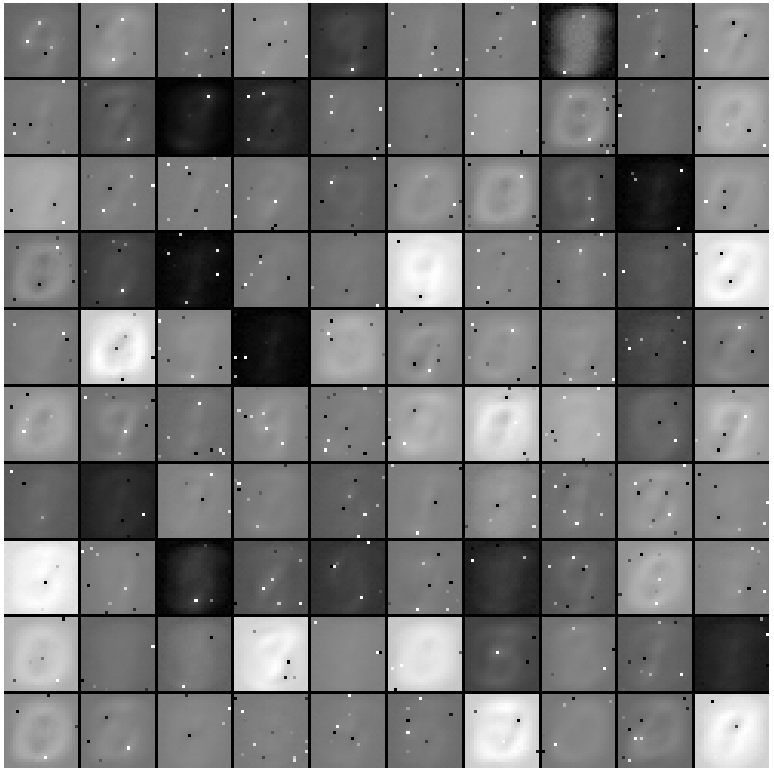
D-Wave (expt-1)

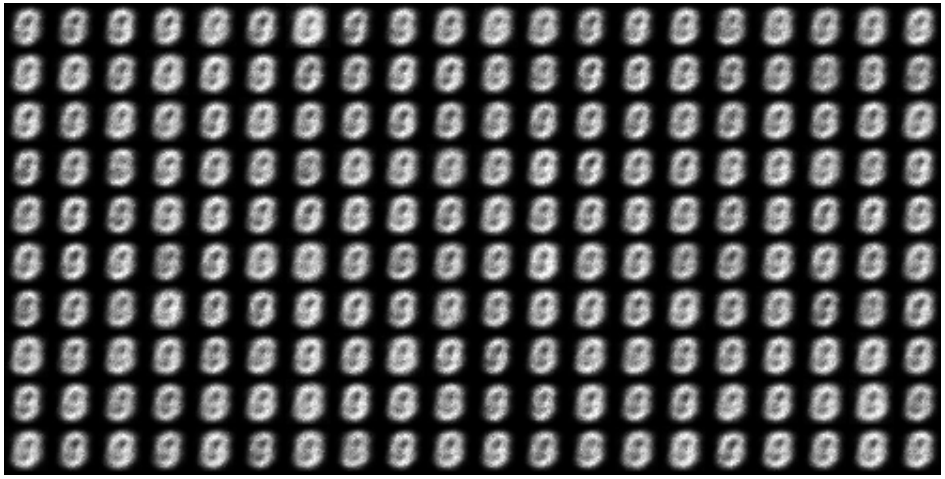




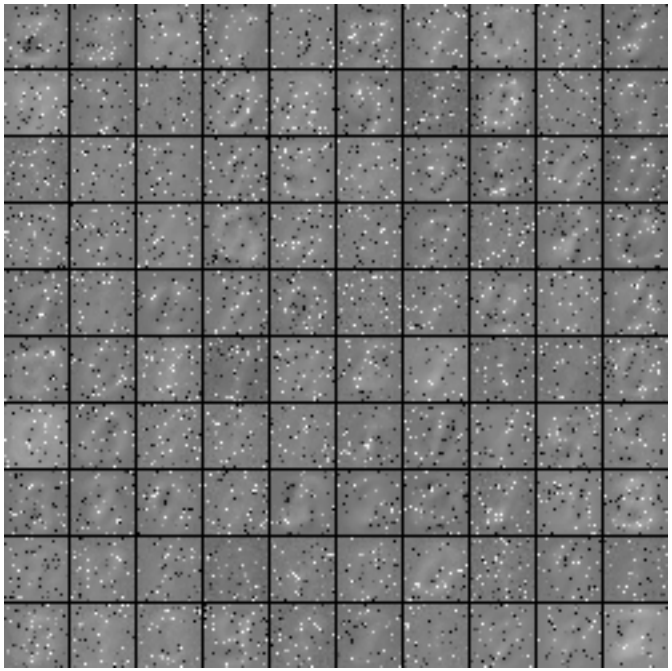
Filters are sparse due to sparse connectivity graph

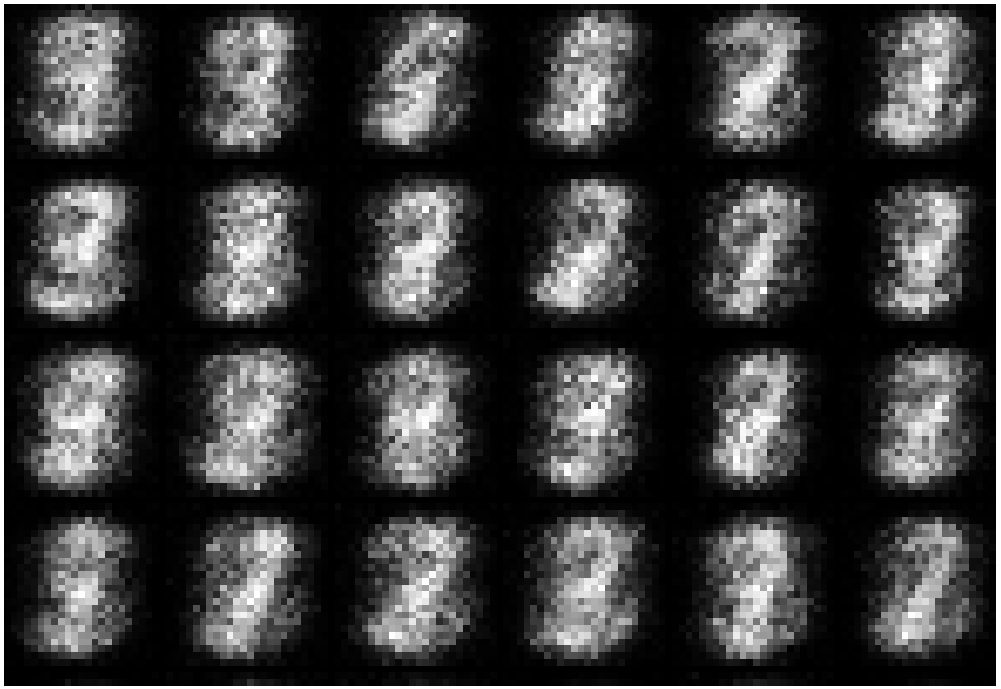
D-wave (expt-2)





D-Wave Expt 3





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 possible connections on a complete bipartite graph.
  - Qubit chaining can overcomes connectivity issues, but then image has to be significantly down-sampled.