

Deep Learning Challenges

Yoshua Bengio

U. Montreal

April 11th, 2014

Google Research, Mountain View, CA



10 BREAKTHROUGH TECHNOLOGIES 2013

Intr

Deep Learning

With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart. →

Temporary Social Media

Messages that quickly self-destruct could enhance the privacy of online communications and make people freer to be spontaneous. →

Prenatal DNA Sequencing

Reading the DNA of fetuses will be the next frontier of the genomic revolution. But do you really want to know about the genetic problems or musical aptitude of your unborn child? →

Adv Man

Ske
prin
wor
mar
the
tech
jet p

Memory Implants

A maverick neuroscientist believes he has deciphered the code by which the brain

Smart Watches

Ultra-Efficient Solar Power

Doubling the efficiency of a solar cell would completely

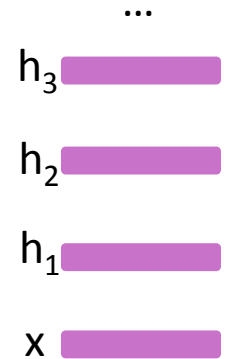
Big Pho

Coll
ana
from
pho

Deep Representation Learning


Learn multiple levels of representation of increasing complexity/abstraction

- theory: exponential gain
- brains are deep
- cognition is compositional
- Better mixing (Bengio et al, ICML 2013)
- **They work! SOTA on industrial-scale AI tasks (object recognition, speech recognition, language modeling, music modeling)**



Montreal Deep Nets Win Emotion Recognition in the Wild Challenge

Predict emotional expression from video (using images + audio)



National University

Results!

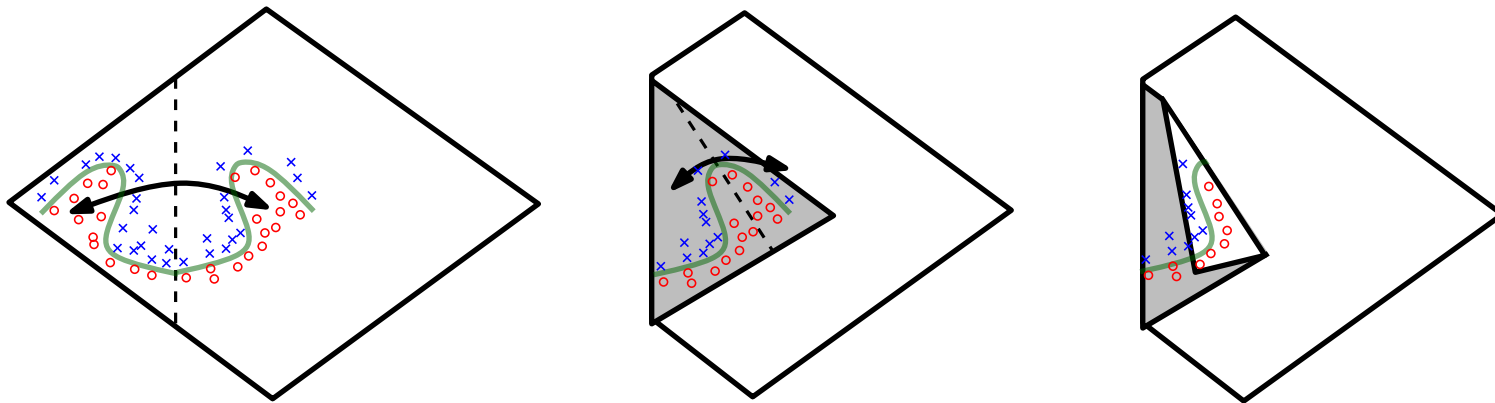
Team Name	Classification accuracy	
Audio baseline	22.4 %	
Video baseline	22.7 %	
Fusion	27.5 %	
Nottingham	24.7 %	
Oulu	21.5 %	
KIT	29.8 %	
UCSD	37.1 %	2nd
ICT@CAS	35.9 %	3rd
York	27.6 %	
LNMIIT	20.5 %	
Montreal	41.0 %	1st
Ulm	27.2 %	

Dec. 9, 2013

New theoretical result: Expressiveness of deep nets with piecewise-linear activation fns

(Pascanu, Montufar, Cho & Bengio; ICLR 2014)

Deeper nets with rectifier/maxout units are exponentially more expressive than shallow ones (1 hidden layer) because they can split the input space in many more (not-independent) linear regions, with constraints, e.g., with abs units, each unit creates mirror responses, folding the input space:



Deep Learning Challenges

(Bengio, arxiv 1305.0445 Deep Learning of representations: Looking forward)

- Computational Scaling
- Optimization & Underfitting
- Intractable Marginalization, Approximate Inference & Sampling
- Disentangling Factors of Variation
- Reasoning & One-Shot Learning of Facts

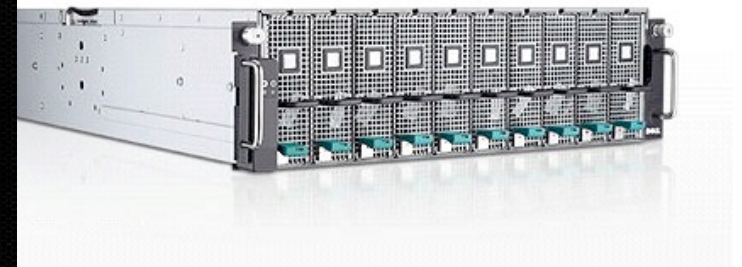
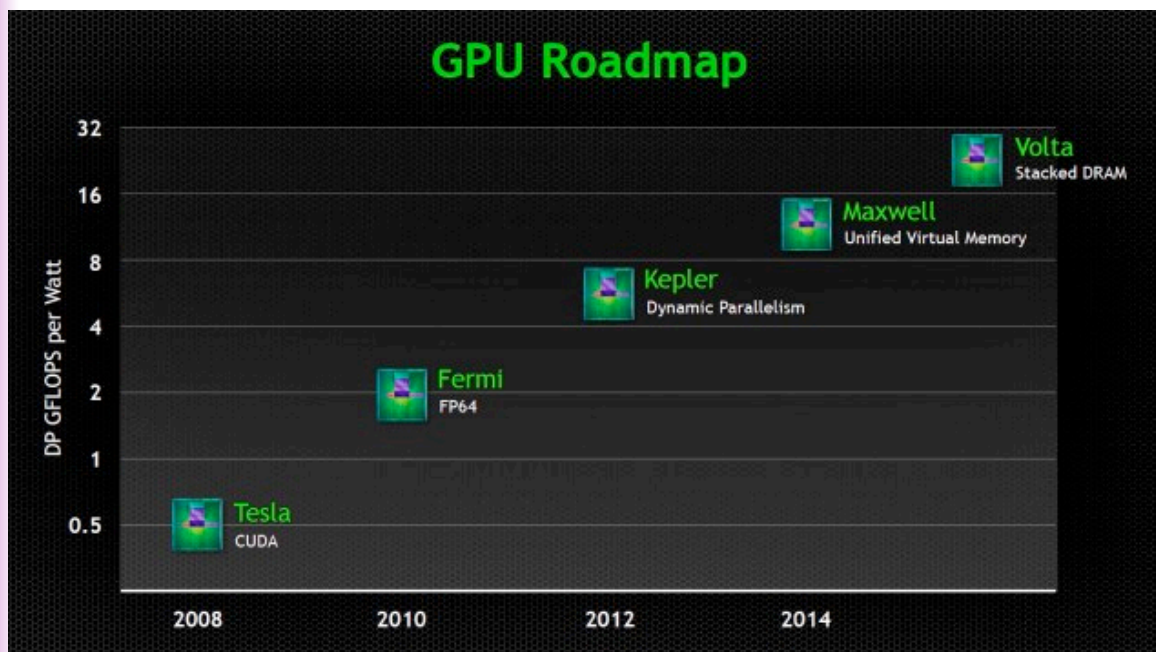
Deep Learning Challenges

(Bengio, arxiv 1305.0445 Deep Learning of representations: Looking forward)

- Computational Scaling
- Optimization & Underfitting
- Intractable Marginalization, Approximate Inference & Sampling
- Disentangling Factors of Variation
- Reasoning & One-Shot Learning of Facts

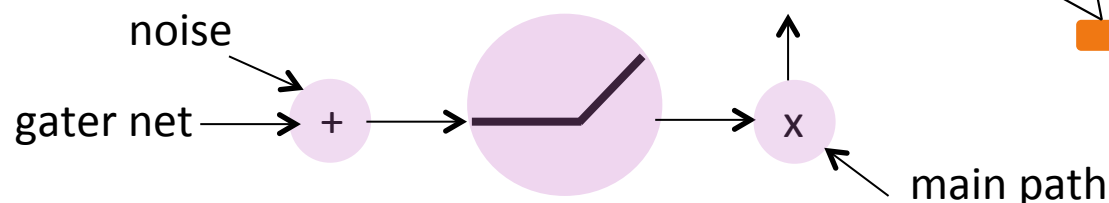
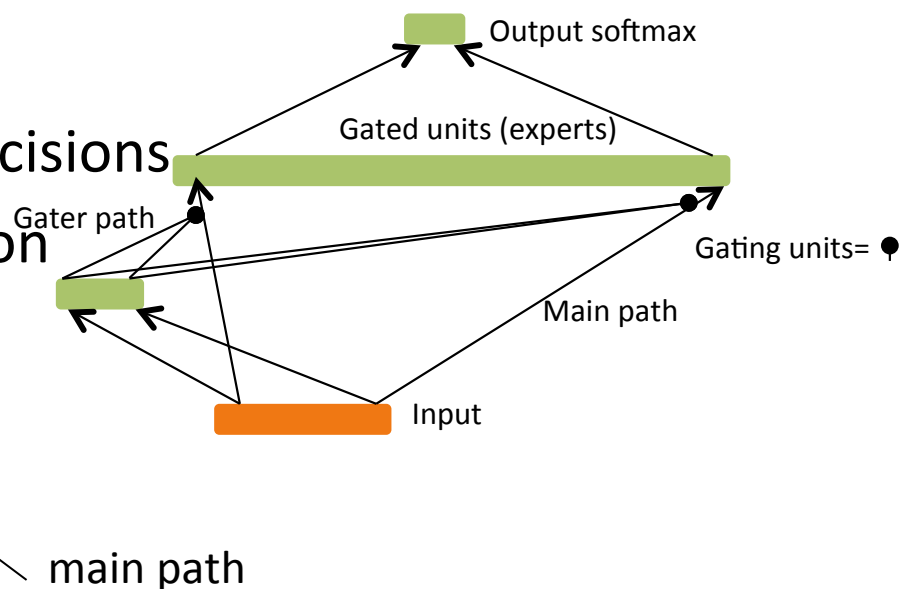
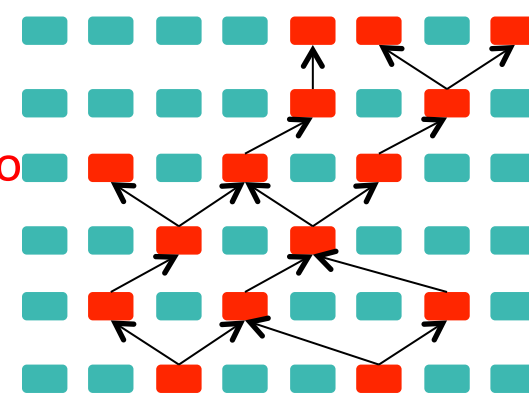
Challenge: Computational Scaling

- Recent breakthroughs in speech, object recognition and NLP hinged on faster computing, GPUs, and large datasets
- A 100-fold speedup is possible without waiting another 10 yrs?
 - Challenge of distributed training
 - Challenge of conditional computation



Conditional Computation: only visit a small fraction of parameters / example

- Deep nets vs decision trees
- Hard mixtures of experts (Collobert, Bengio & Bengio 2002)
- Conditional computation for deep nets: sparse distributed gaters selecting combinatorial subsets of a deep net
- Challenges:
 - Credit assignment for hard decisions
 - Gated architectures exploration
- **Noisy rectifiers** work well



Distributed Training

- Minibatches
- Large minibatches + 2nd order & natural gradient methods
- Asynchronous SGD (Bengio et al 2003, Le et al ICML 2012, Dean et al NIPS 2012)
 - Bottleneck: sharing weights/updates among nodes, to avoid node-models to move too far from each other
- Ideas forward:
 - Low-resolution sharing only where needed
 - Specialized conditional computation (each computer specializes in updates to some cluster of gated experts, and prefers examples which trigger these experts)

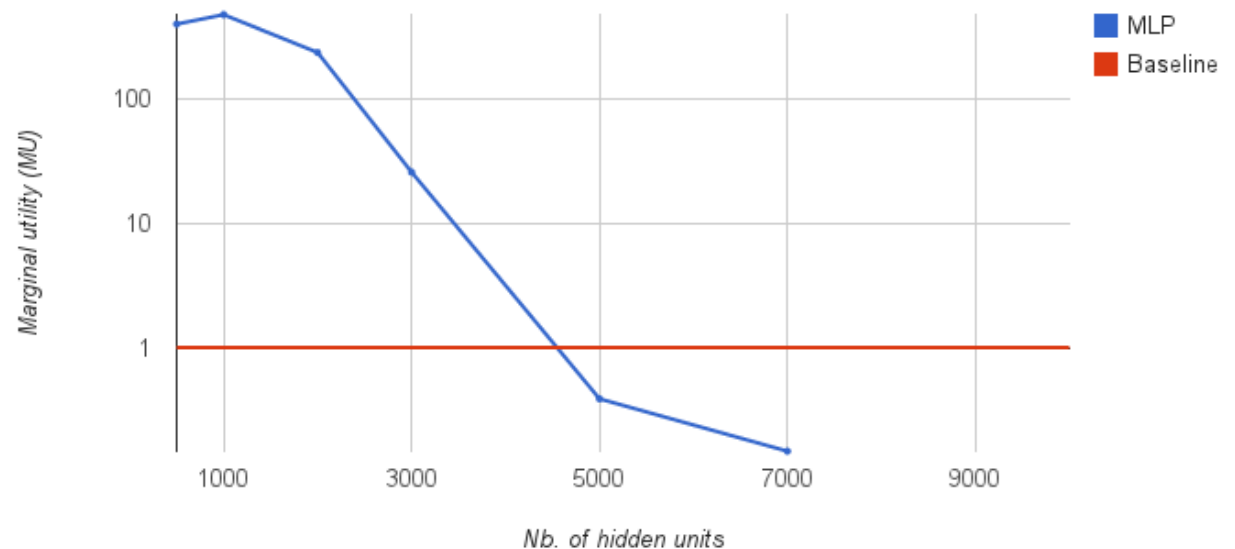
Deep Learning Challenges

(Bengio, arxiv 1305.0445 Deep Learning of representations: Looking forward)

- Computational Scaling
- Optimization & Underfitting
- Intractable Marginalization, Approximate Inference & Sampling
- Disentangling Factors of Variation
- Reasoning & One-Shot Learning of Facts

Optimization & Underfitting

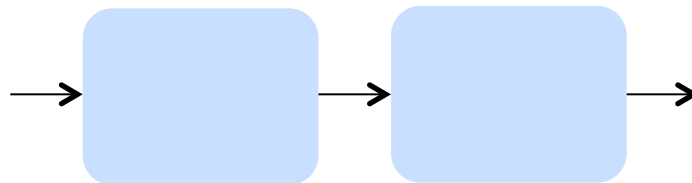
- On large datasets, major obstacle is underfitting
- **Marginal utility** of wider MLPs decreases quickly below memorization baseline



- Current limitations: local minima, ill-conditioning or else?

Guided Training, Intermediate Concepts

- In (Gulcehre & Bengio ICLR'2013) we set up a task that seems almost impossible to learn by shallow nets, deep nets, SVMs, trees, boosting etc
- Breaking the problem in two sub-problems and pre-training each module separately, then fine-tuning, nails it
- *Need prior knowledge to decompose the task*
- **Guided pre-training** allows to find much better solutions, escape effective local minima



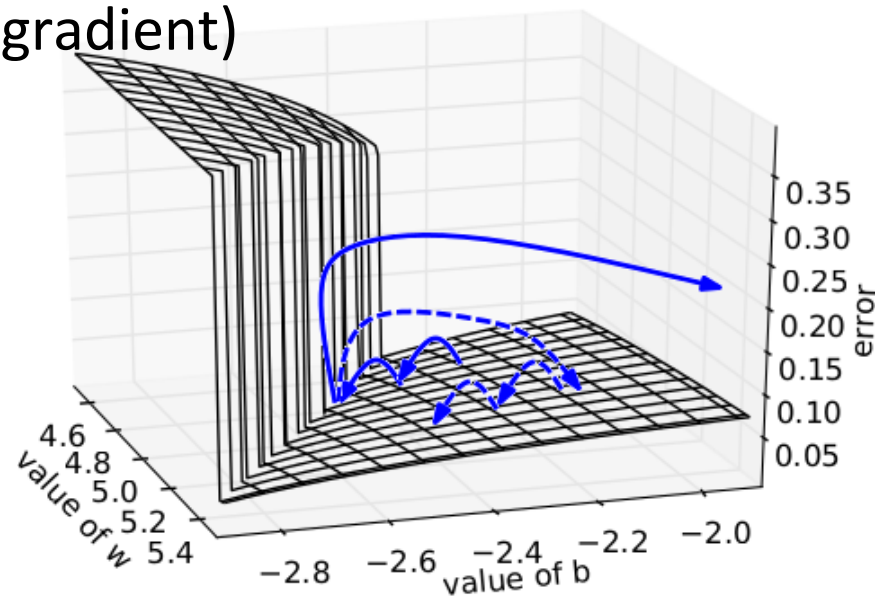
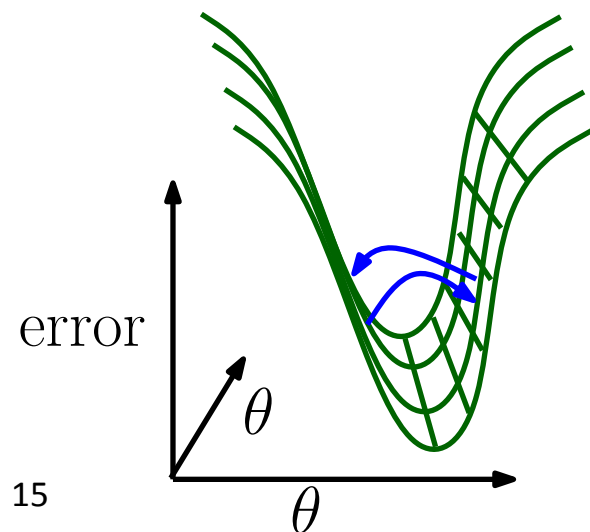
On the difficulty of training RNNs

- ICASSP 2013 & ICML 2013 papers:
 - Putting together techniques to reduce the difficulty of training RNNs
- ICLR 2014 paper: Deep Recurrent Nets
 - New architectures to boost capacity while maintaining trainability, by introducing more non-linearities as well as skip connections

RNN Training Tricks

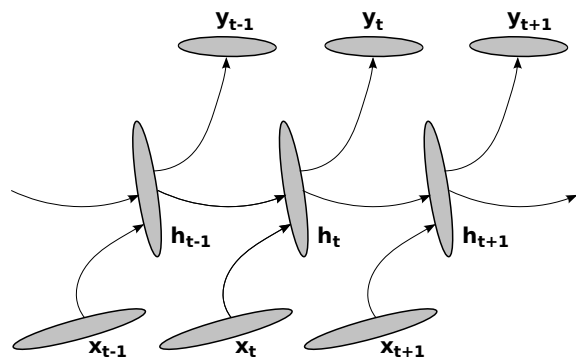
(Pascanu, Mikolov, Bengio, ICML 2013; Bengio, Boulanger & Pascanu, ICASSP 2013)

- Clipping gradients (avoid exploding gradients)
- Leaky integration (propagate long-term dependencies)
- Momentum (cheap 2nd order)
- Initialization (start in right ballpark avoids exploding/vanishing)
- Sparse Gradients (symmetry breaking)
- Gradient propagation regularizer (avoid vanishing gradient)
- LSTM self-loops (avoid vanishing gradient)



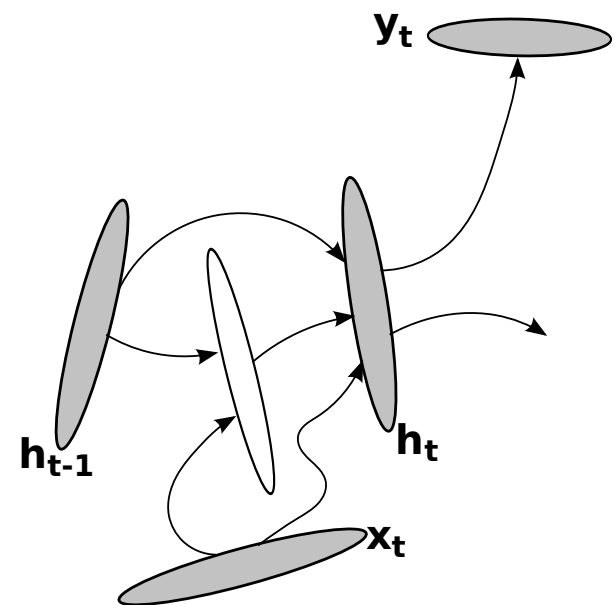
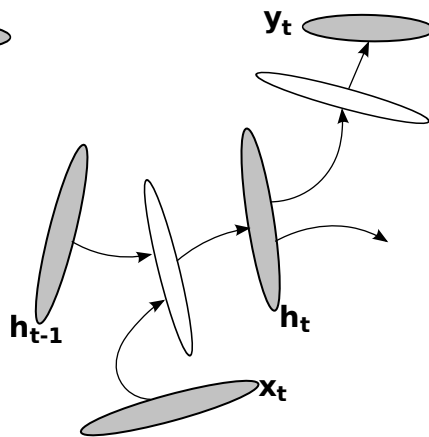
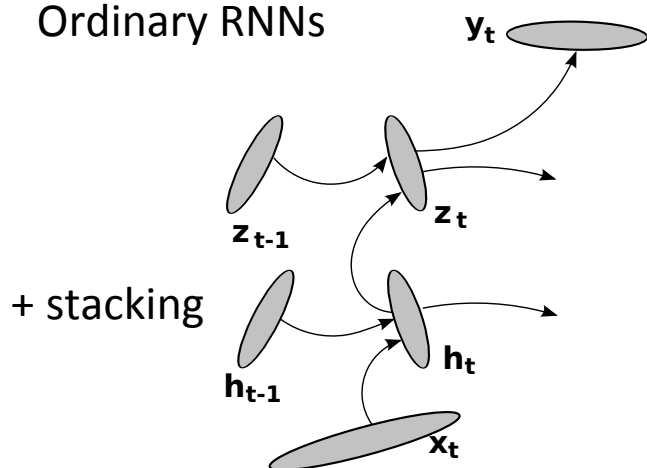
Increasing the Expressive Power of RNNs with more Depth

- ICLR 2014, *How to construct deep recurrent neural networks*



Ordinary RNNs

+ deep hid-to-out
+ deep hid-to-hid
+ deep in-to-hid

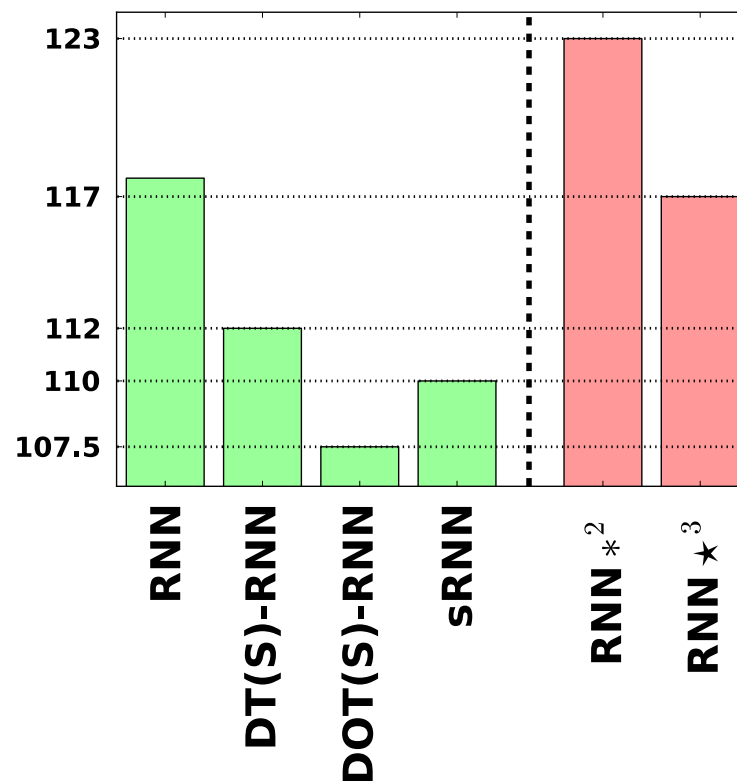
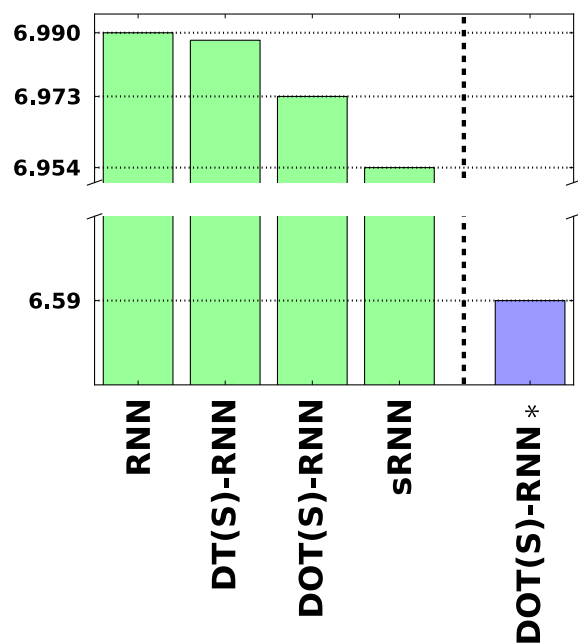


+ skip connections for
creating shorter paths

Deep RNN Results

- Language modeling
(Penn Treebank perplexity)

- Music modeling (Muse, NLL)



More results in the ICLR 2014 paper

Deep Learning Challenges

(Bengio, arxiv 1305.0445 Deep Learning of representations: Looking forward)

- Computational Scaling
- Optimization & Underfitting
- Intractable Marginalization, Approximate Inference & Sampling
- Disentangling Factors of Variation
- Reasoning & One-Shot Learning of Facts

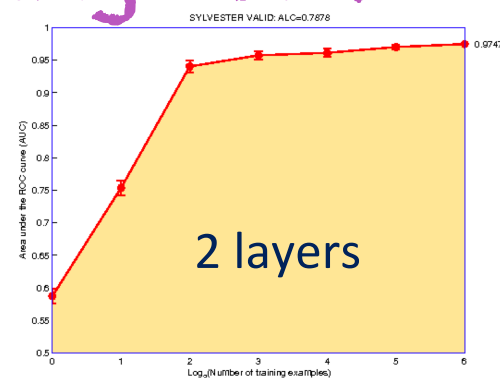
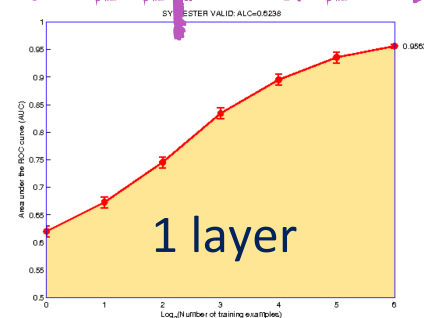
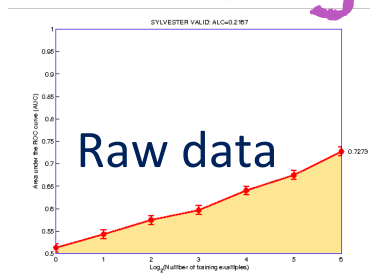
Why Unsupervised Learning?

- Recent progress mostly in supervised DL
- \exists real challenges for unsupervised DL
- Potential benefits:
 - Exploit tons of unlabeled data
 - Answer new questions about the variables observed
 - Regularizer – transfer learning – domain adaptation
 - Easier optimization (local training signal)
 - Structured outputs

How do humans generalize from very few examples?

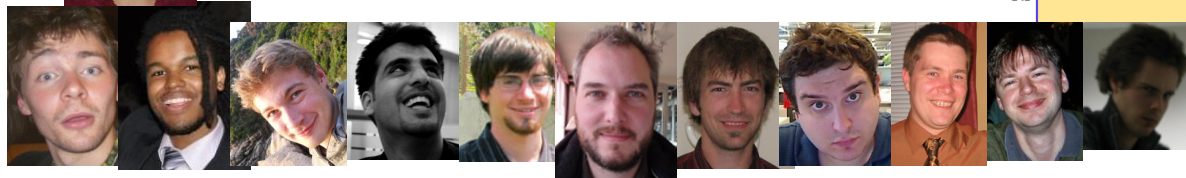
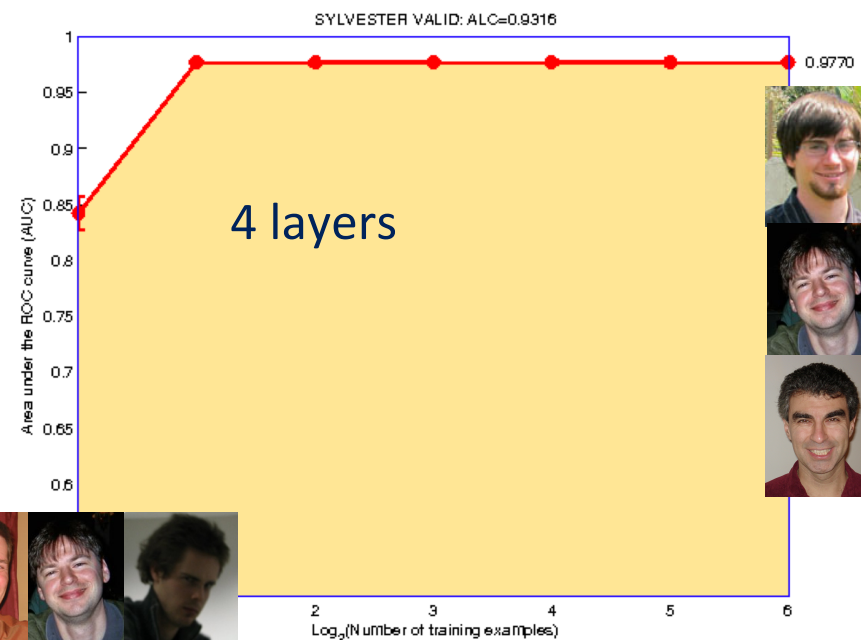
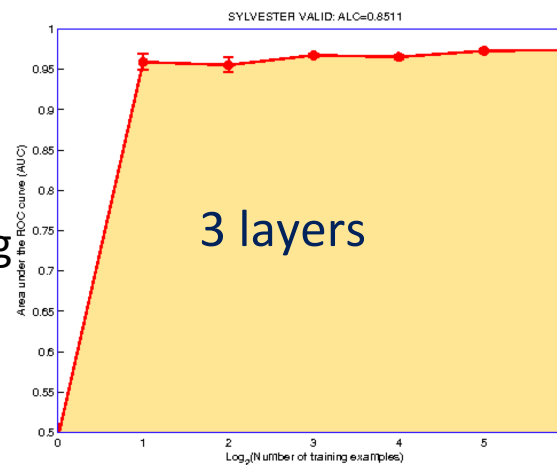
- They **transfer** knowledge from previous learning:
 - **Abstract** (i.e. deep) representations
 - Explanatory factors
- Previous learning from: unlabeled data
 - + labels for other tasks

Unsupervised and Transfer Learning Challenge + Transfer Learning Challenge: Deep Learning 1st Place



NIPS'2011
Transfer
Learning
Challenge
Paper:
ICML'2012

ICML'2011
workshop on
Unsup. &
Transfer Learning



Basic Challenge with Probabilistic Models: marginalization

- Joint and marginal likelihoods involve intractable sums over configurations of random variables (inputs x , latent h , outputs y) e.g.

$$P(x) = \sum_h P(x,h)$$

$$P(x,h) = e^{-\text{energy}(x,h)} / Z$$

$$Z = \sum_{x,h} e^{-\text{energy}(x,h)}$$

- MCMC methods can be used for these sums, by sampling from a chain of x 's (or of (x,h) pairs) approximately from $P(x,h)$

Two Fundamental Problems with Probabilistic Models with Many Random Variables

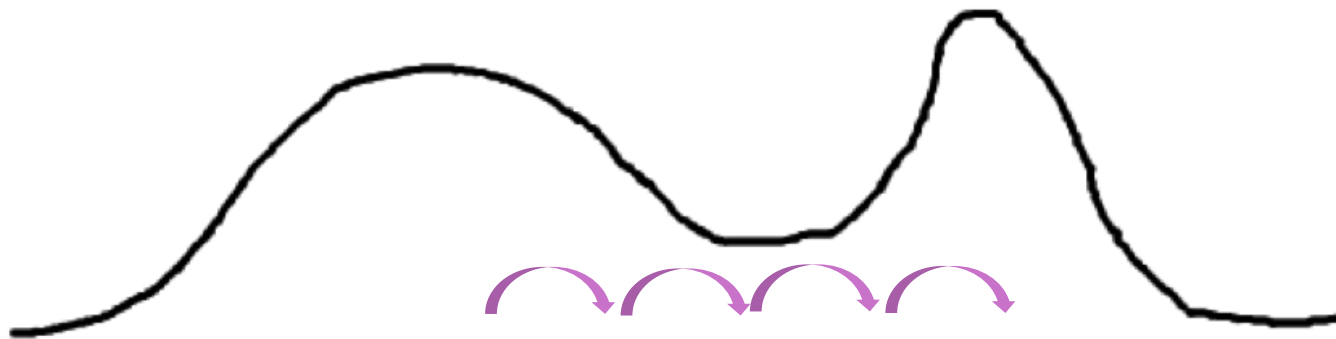
1. MCMC mixing between modes (manifold hypothesis)



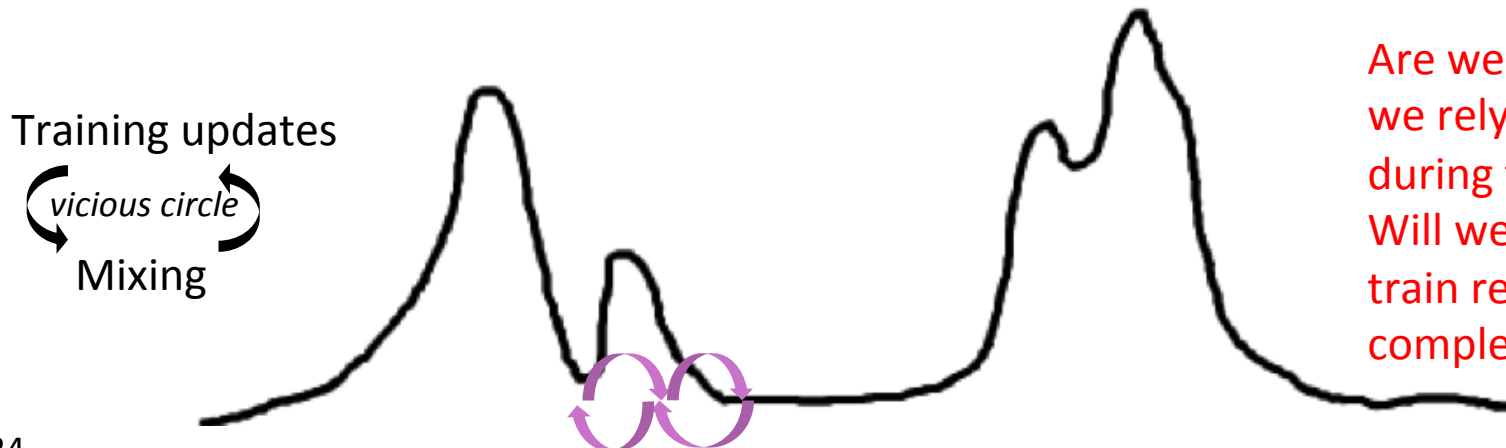
2. Many non-negligible modes (both in posterior & joint distributions)

For gradient & inference: More difficult to mix with better trained models

- Early during training, density smeared out, mode bumps overlap



- Later on, hard to cross empty voids between modes

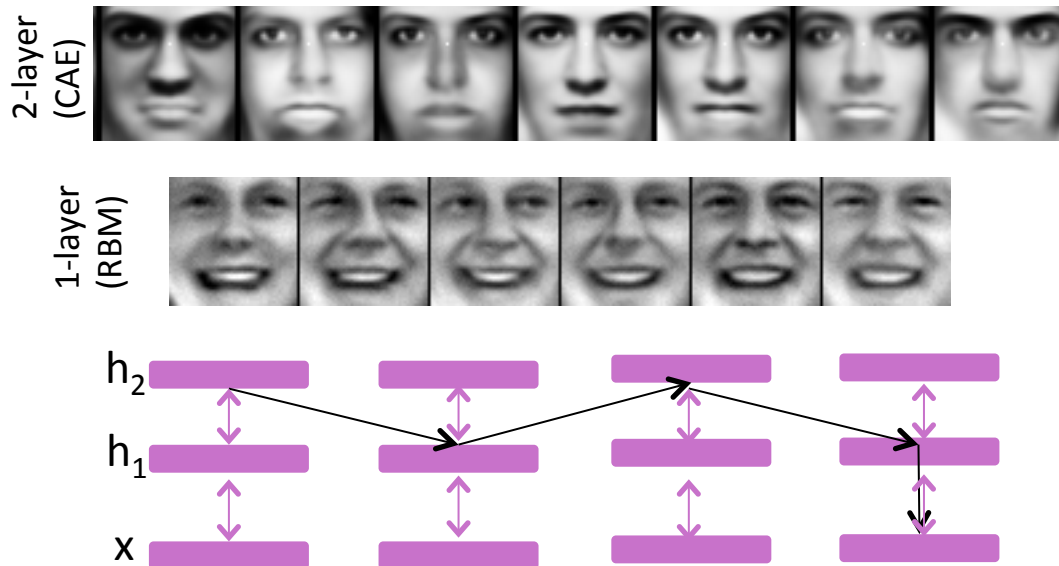


Are we doomed if
we rely on MCMC
during training?
Will we be able to
train really large &
complex models?

Poor Mixing: Depth to the Rescue

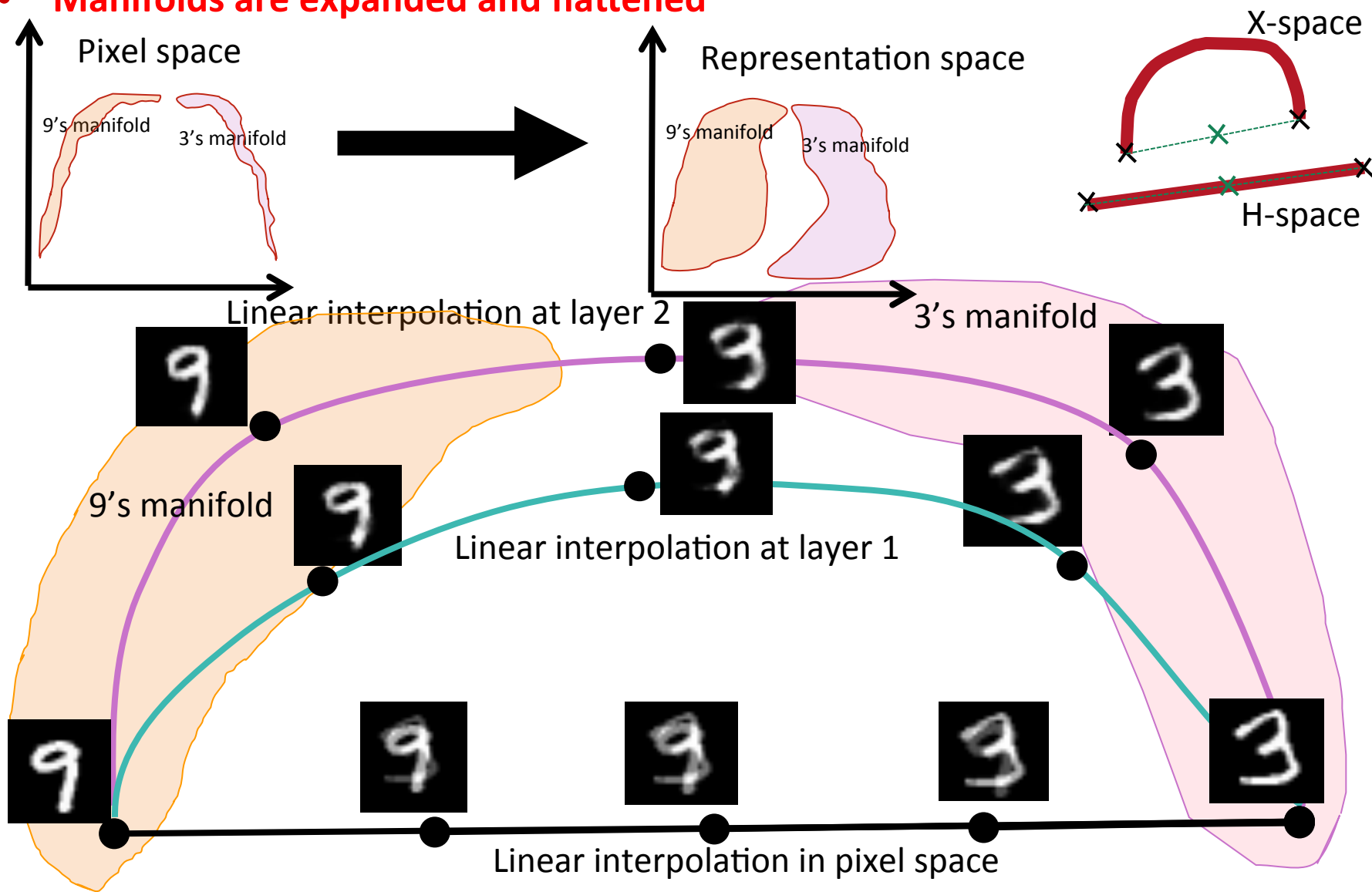
(Bengio et al ICML 2013)

- Sampling from DBNs and stacked Contractive Auto-Encoders:
 1. MCMC sampling from top layer model
 2. Propagate top-level representations to input-level repr.
- Deeper nets visit more modes (classes) faster



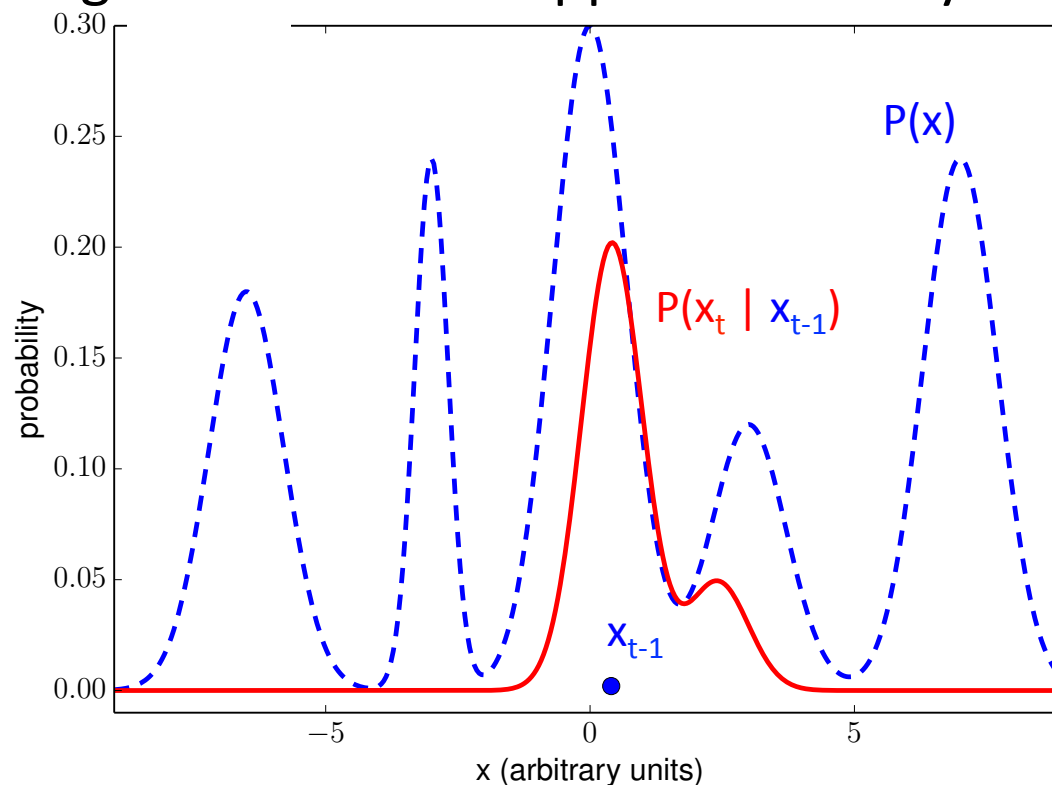
Space-Filling in Representation-Space

- Deeper representations \rightarrow abstractions \rightarrow disentangling
- Manifolds are expanded and flattened



Many Modes Challenge: Instead of Learning $P(x)$ directly, Learn Markov chain operator $P(x_t | x_{t-1})$

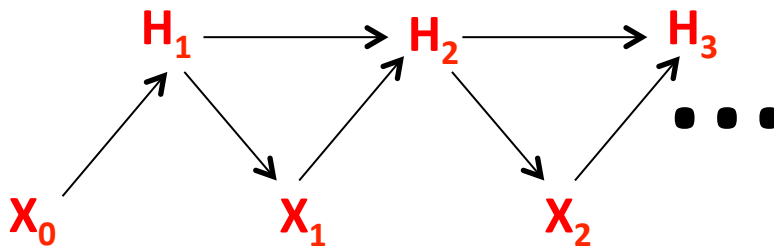
- $P(x)$ may have many modes, making the normalization constant intractable, and MCMC approximations poor
- Partition fn of $P(x_t | x_{t-1})$ much simpler because most of the time a local move, might even be well approximated by unimodal



Generative Stochastic Networks

- Generalizes the denoising auto-encoder training scheme
 - Introduce latent variables in the Markov chain (over X, H)
 - Instead of a fixed corruption process, have a deterministic function with parameters θ_1 and a noise source Z as input

$$H_{t+1} = f_{\theta_1}(X_t, Z_t, H_t)$$



$$H_{t+1} \sim P_{\theta_1}(H|H_t, X_t)$$

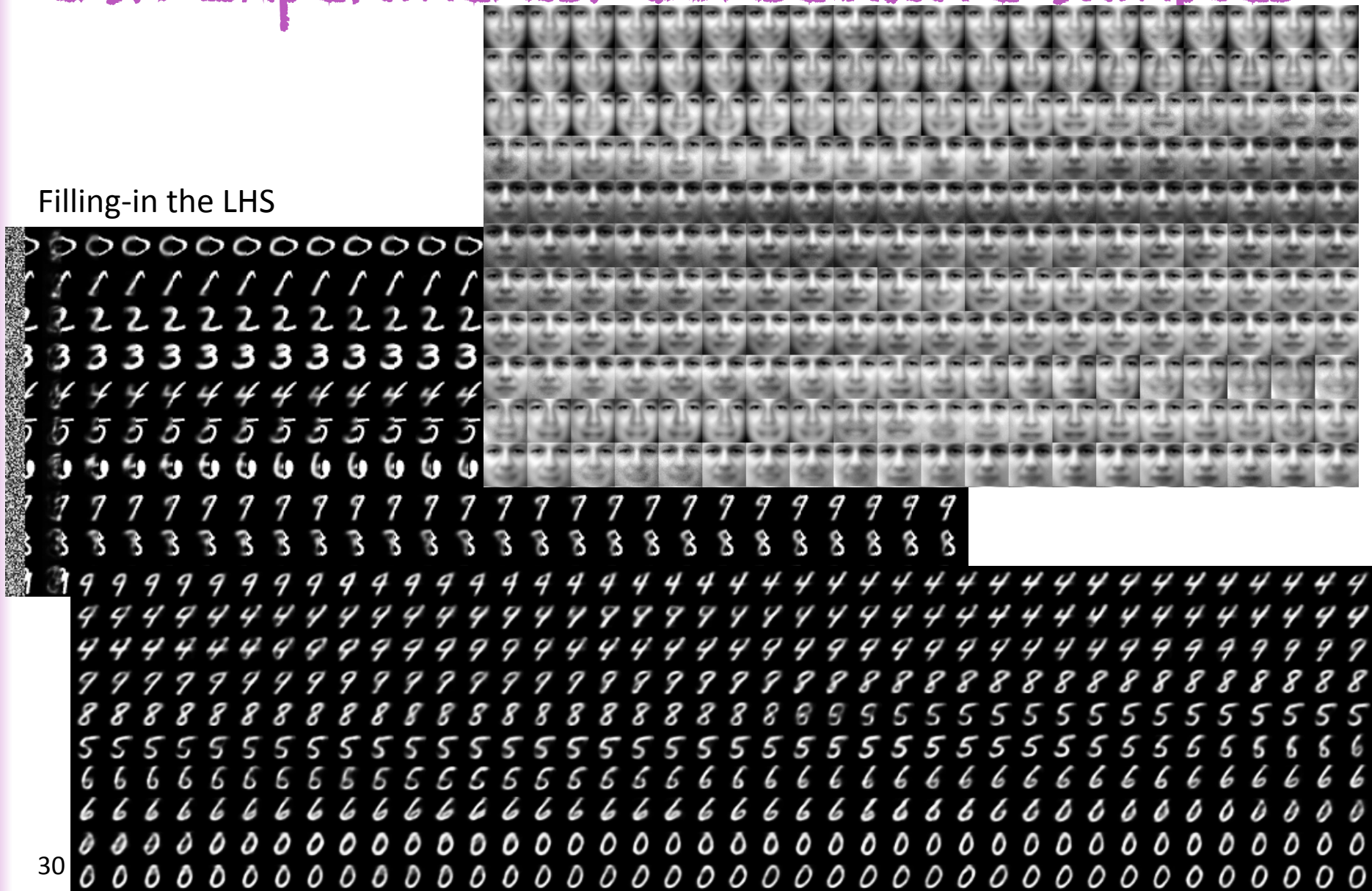
$$X_{t+1} \sim P_{\theta_2}(X|H_{t+1})$$

Consistent Estimator Theorem

If the parametrization is rich enough to have $P(X|H)$ a consistent estimator and the Markov chain is ergodic, then maximizing the expected log of $P_{\theta_2}(X|f_{\theta_1}(X, Z_{t-1}, H_{t-1}))$ makes the stationary distribution of the Markov chain a consistent estimator of the true data generating distribution.

GSN Experiments: Consecutive Samples

Filling-in the LHS



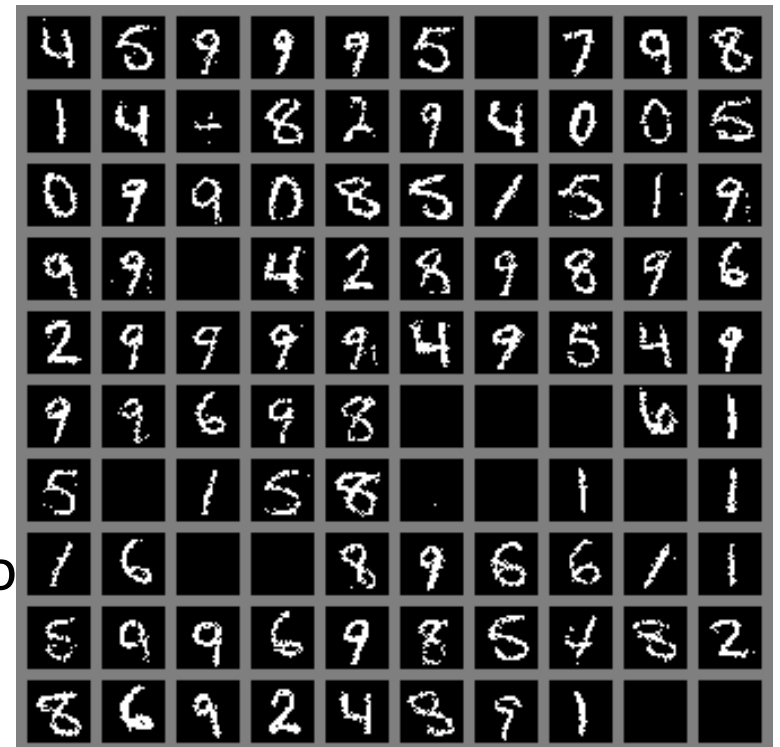
A Proper Generative Model for Dependency Networks, MP-DBMs, and efficient deep NADE sampling

- Dependency nets (Heckerman et al 2000) estimate $P_{\theta_i}(X_i | X_{-i})$ not guaranteed to be conditionals of a unique joint
- Heckerman et al's sampling iterates over i : not ergodic?
- Randomly choosing i : proper GSN
- Defines a unique joint distribution = stationary distr. of chain (which averages out over resampling orders)
- Generalized to estimators of $P(\text{subset}(X) | X \setminus \text{subset}(X))$ and justify efficient sampling schemes for MP-DBMs and deep NADE.

4	5	9	9	7	5		7	9	8
1	4	-	8	2	9	4	0	0	5
0	7	9	0	8	5	/	5	1	9
9	9		4	2	8	9	8	9	6
2	9	9	9	9	4	9	5	4	9
9	9	6	9	8				6	1
5		/	5	8			1		1
/	6			8	9	6	6	/	1
5	9	9	6	9	8	5	4	8	2
8	6	9	2	4	8	9	1		

MP-DBM Results

- Single model of (X,Y) vs multiple stages of training DBM + fine-tuning
- SOTA on permutation-invariant MNIST (at time of submission):
 - 0.88% error
- Salakhutdinov & Hinton's DBM: 0.95%
- NORB: 10.6% (vs 10.8% with S&H's DBM)
- DBM (Gibbs) samples of trained MP-DBM are ugly, while GSN sampling works because it better corresponds to the training criterion:



Reparametrizing latent variables

- Insight from (Bengio et al 2013, arxiv 1306.1091 & 1308.3432) papers on GSNs and stochastic neurons:
 - Sampling from continuous latent variables (given some ancestors) can be rewritten as a deterministic function of other variables and of independent noise sources: $h = f(x; \eta)$
 - This enables rewriting the gradient log-likelihood as back-prop, averaged over samples of the noise sources

$$P(y|x) = \int_h P(y|h, x)P(h|x)dh = \int_{\eta} P(y|f(x; \eta), x)P(\eta)d\eta$$
$$\frac{\partial P(y|x)}{\partial \theta} = \int_{\eta} \frac{\partial P(y|f(x; \eta), x)}{\partial \theta} P(\eta)d\eta$$

- A deeper formal analysis of this approach:
 - Kingma & Welling 2014, arxiv 1402.0480; see also Wierstra et al 2014, arxiv 1401.4082.

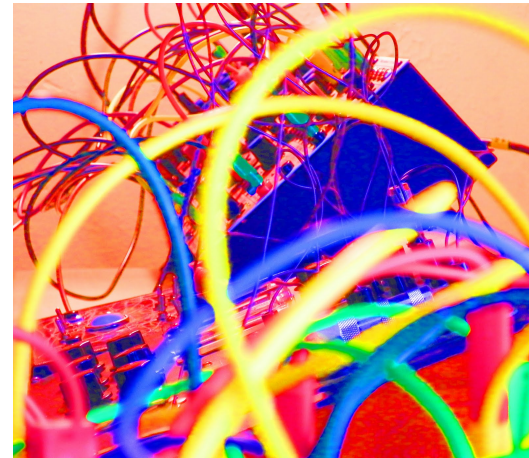
Deep Learning Challenges

(Bengio, arxiv 1305.0445 Deep Learning of representations: Looking forward)

- Computational Scaling
- Optimization & Underfitting
- Intractable Marginalization, Approximate Inference & Sampling
- Disentangling Factors of Variation
- Reasoning & One-Shot Learning of Facts

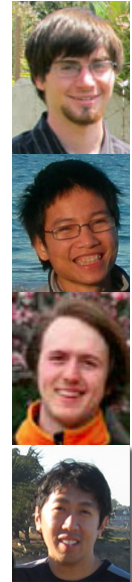
Invariance and Disentangling

- Invariant features
- Which invariances?
- Alternative: learning to disentangle factors
- Good disentangling →
avoid the curse of dimensionality



Emergence of Disentangling

- (Goodfellow et al. 2009): sparse auto-encoders trained on images
 - some higher-level features more invariant to geometric factors of variation
- (Glorot et al. 2011): sparse rectified denoising auto-encoders trained on bags of words for sentiment analysis
 - different features specialize on different aspects (domain, sentiment)



WHY?

Broad Priors as Hints to Disentangle the Factors of Variation

- *Multiple factors*: distributed representations
- Multiple levels of abstraction: *depth*
- *Semi-supervised* learning: Y is one of the factors explaining X
- *Multi-task* learning: different tasks share some factors
- *Manifold* hypothesis: probability mass concentration
- Natural *clustering*: class = manifold, well-separated manifolds
- Temporal and spatial *coherence*
- *Sparsity*: most factors irrelevant for particular X
- *Simplicity* of factor dependencies (in the right representation)

Conclusions

- Deep Learning has matured
 - Int. Conf. on Learning Representation 2013 a huge success!
- Industrial applications (Google, Microsoft, Baidu, Facebook, ...)
- Room for improvement:
 - Scaling computation
 - Optimization
 - Bypass intractable marginalizations
 - More disentangled abstractions
 - Reason from incrementally added facts

LISA team: **Merci! Questions?**

