Scaling Up Deep Learning

Yoshua Bengio

U. Montreal

June 25th, 2014

Workshop on Deep Learning Models for Emerging Big Data Applications

ICML'2014, Beijing, China

MIT Technology Review

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? →

Ado Mar

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

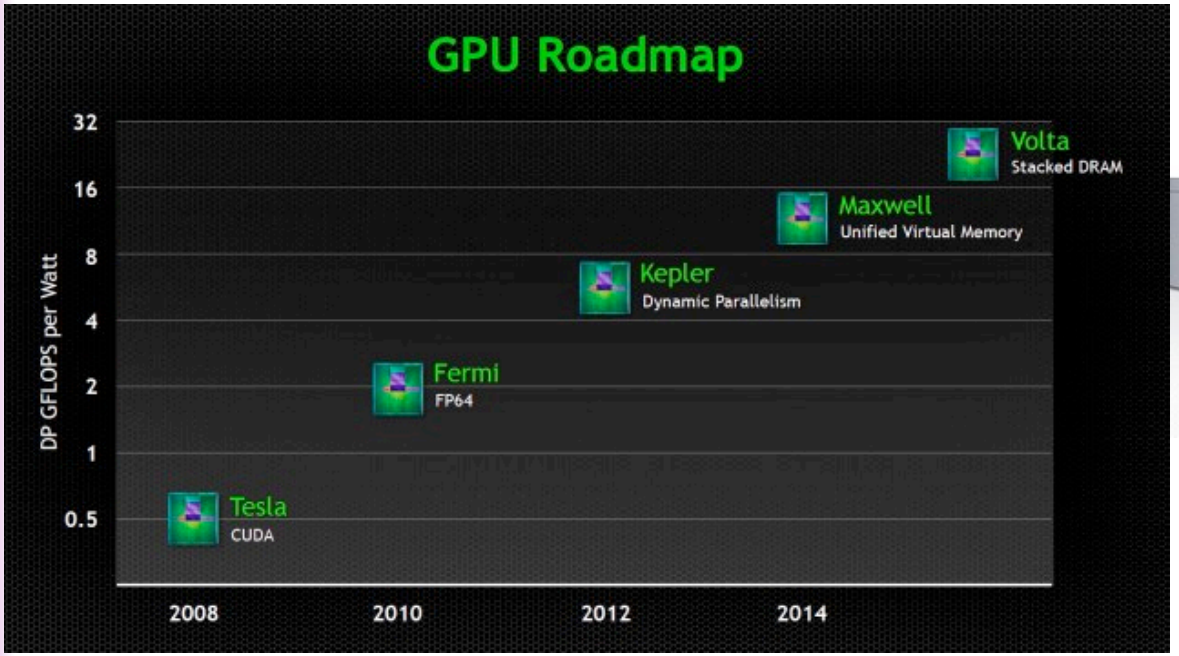
- Computational Scaling
- Optimization & Underfitting
- Intractable Marginalization, Approximate Inference & Sampling
- Disentangling Factors of Variation
- Reasoning & One-Shot Learning of Facts

Deep Learning Challenges
(Bengio, arxiv 1305.0445 Deep learning of representations: looking forward)

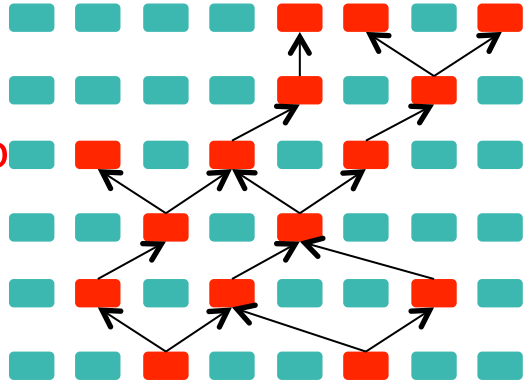
- Computational Scaling
- Optimization & Underfitting
- Intractable Marginalization, Approximate Inference & Sampling
- Disentangling Factors of Variation
- Reasoning & One-Shot Learning of Facts

Challenge: Computational Scaling

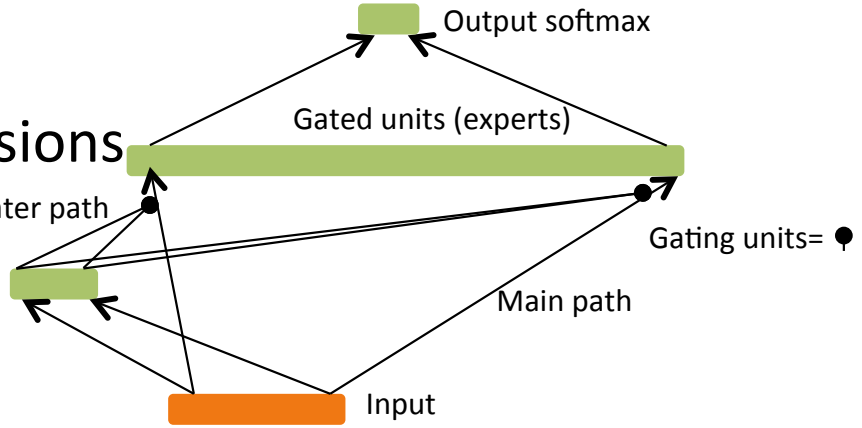
- Recent breakthroughs in speech, object recognition and NLP hinged on faster computing, GPUs, and large datasets
- A 100-fold speedup is possible without waiting another 10 yrs?
  - Challenge of distributed training
  - Challenge of conditional computation

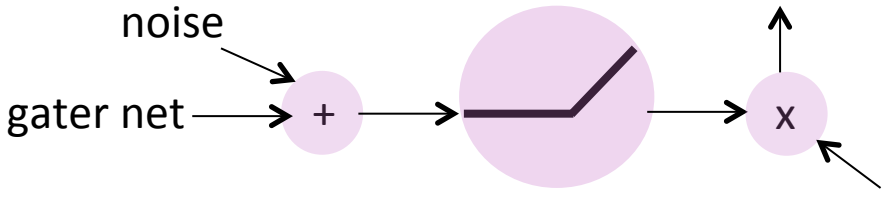


Conditional Computation: only visit a small fraction of parameters / example



- 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

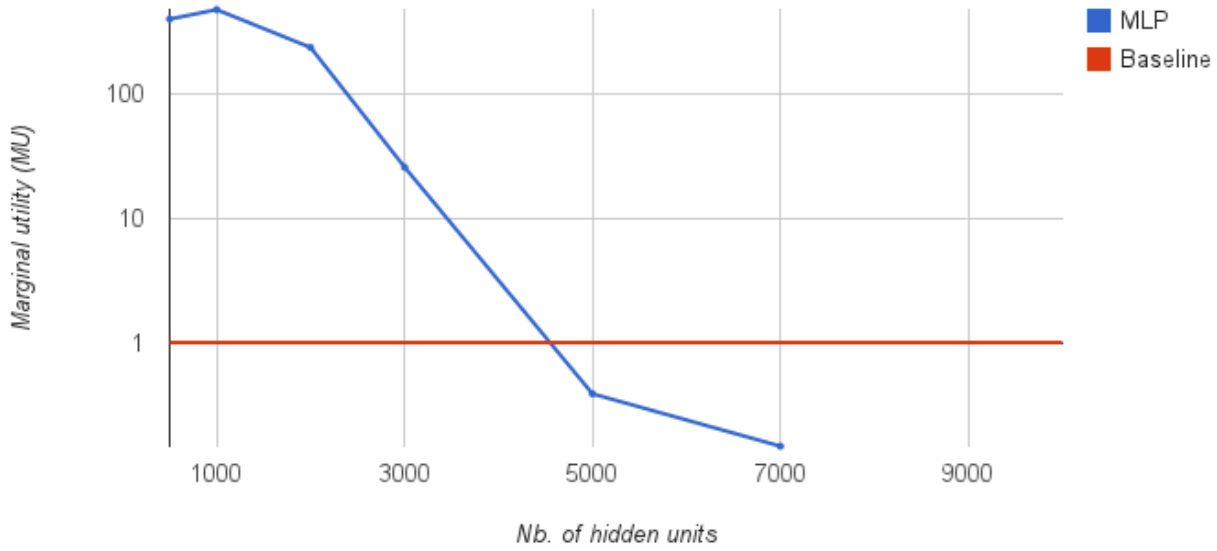
- Minibatches
- Large minibatches + 2nd order & natural gradient methods
- Asynchronous SGD (Bengio et al 2003, Le et al ICML 2012, Dean et al NIPS 2012)
  - Bottleneck: sharing weights/updates among nodes, to avoid node-models to move too far from each other
- Ideas forward:
  - Low-resolution sharing 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

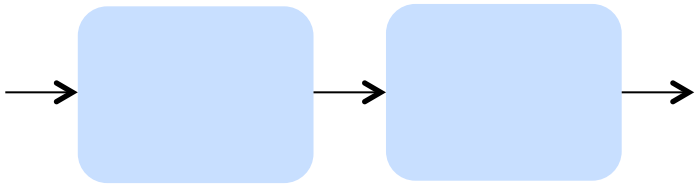
- On large datasets, major obstacle is underfitting
- **Marginal utility** of wider MLPs decreases quickly below memorization baseline



- 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

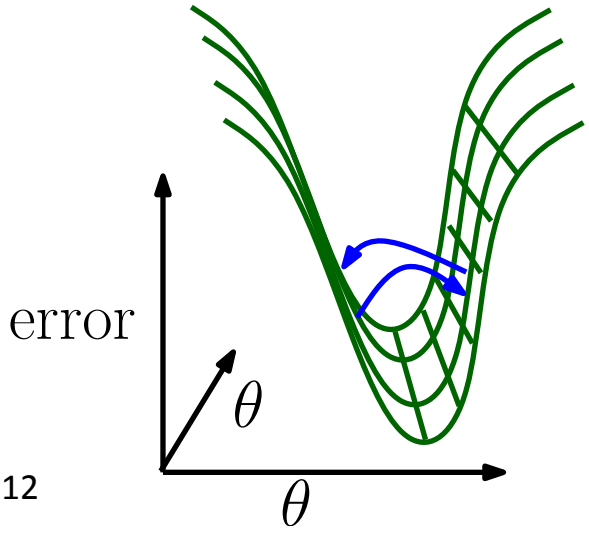
- ICASSP 2013 & ICML 2013 papers:
  - Putting together techniques to reduce the difficulty of training RNNs

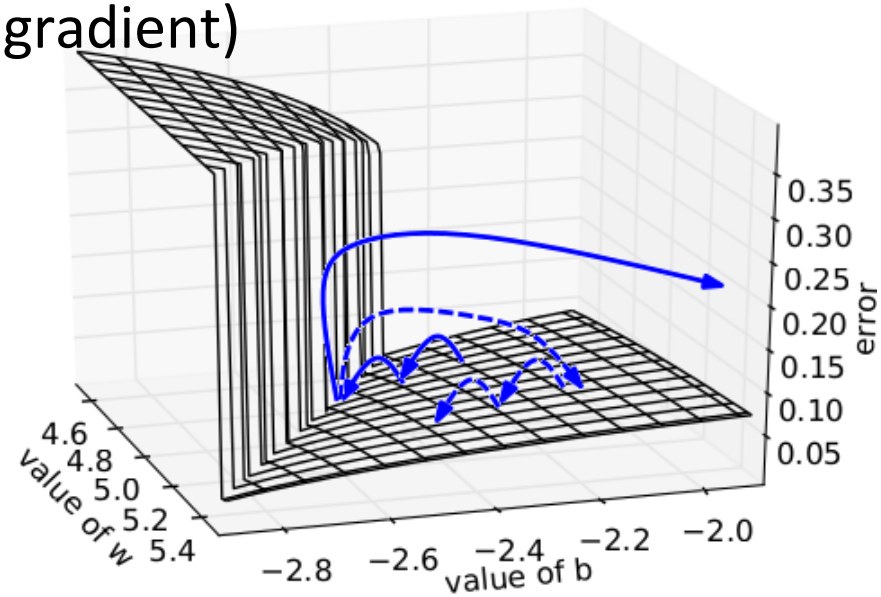
- ICLR 2014 paper: Deep Recurrent Nets
  - New architectures to boost capacity while maintaining traininability, by introducing more non-linearities as well as skip connections

RNN Training Tricks

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

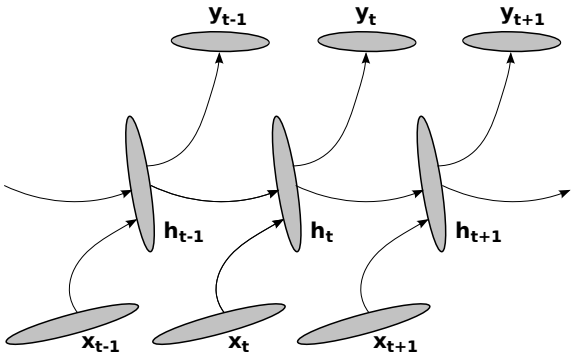
- Clipping gradients (avoid exploding gradients)
- Leaky integration (propagate long-term dependencies)
- Momentum (cheap 2nd order)
- Initialization (start in right ballpark avoids exploding/vanishing)
- Sparse Gradients (symmetry breaking)
- Gradient propagation regularizer (avoid vanishing gradient)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





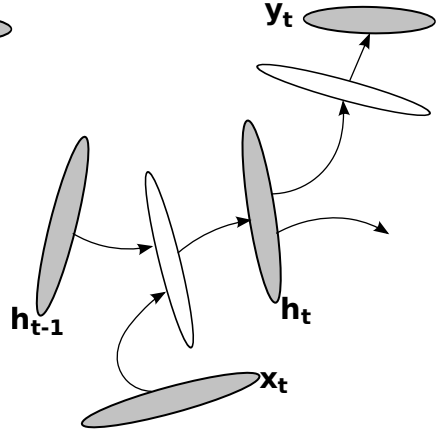
Increasing the Expressive Power of RNNs with more Depth

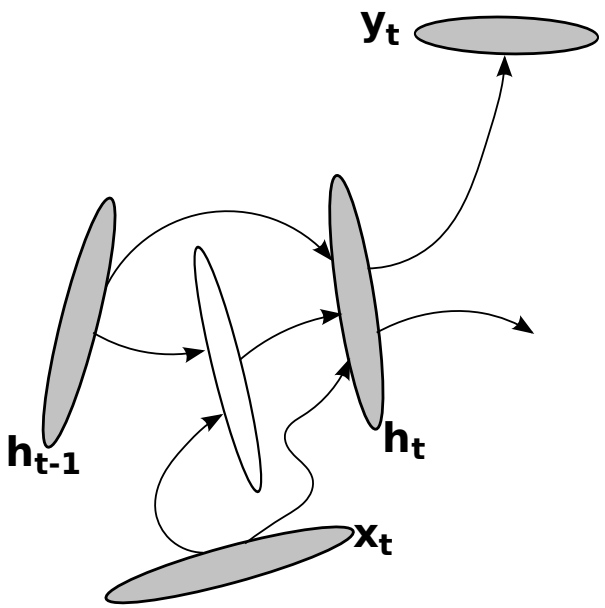
- ICLR 2014, *How to construct deep recurrent neural networks*



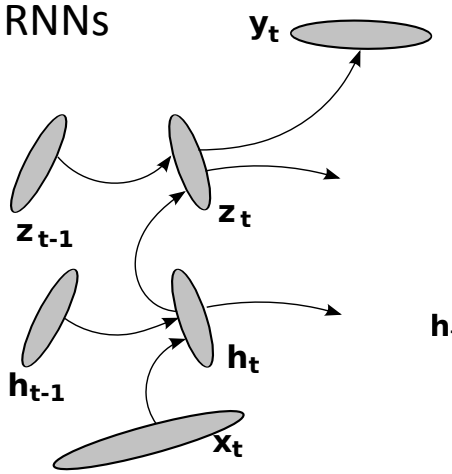
Ordinary RNNs

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





+ stacking

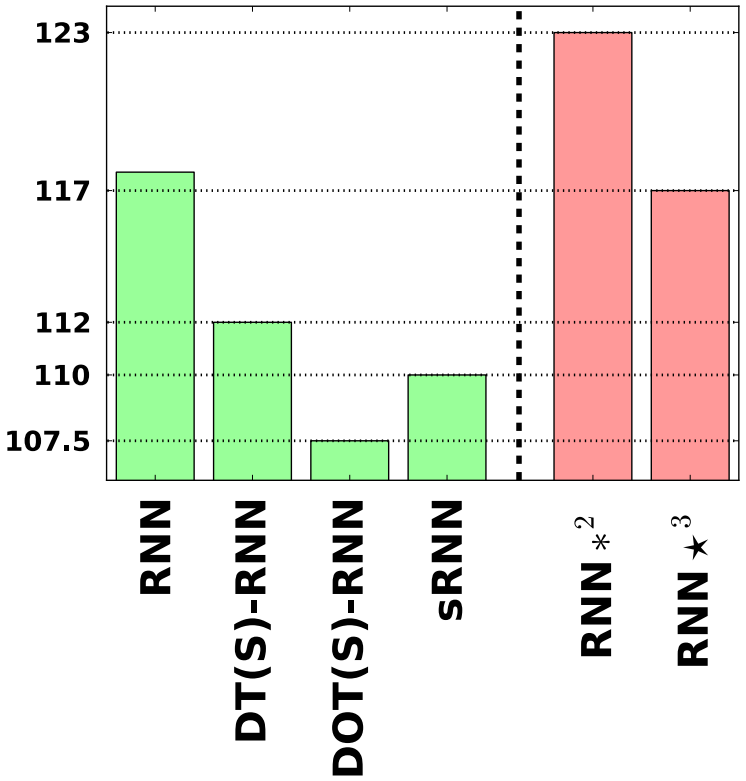


+ skip connections for creating shorter paths

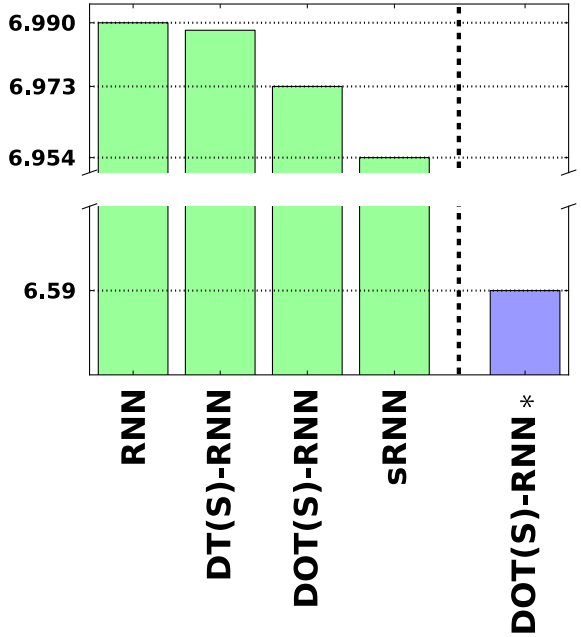
Deep RNN Results

- 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
- ∃ 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

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)

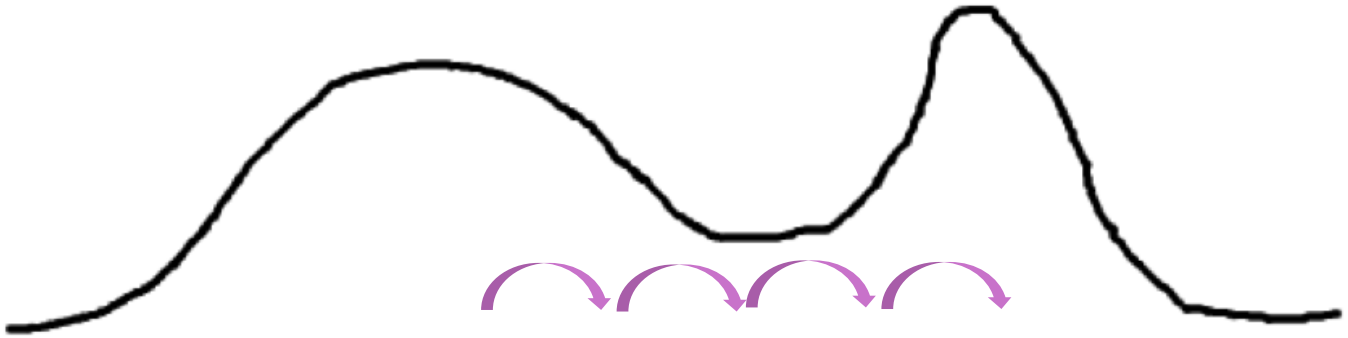
Two Fundamental Problems with Probabilistic Models with Many Random Variables

1. MCMC mixing between modes
   (manifold hypothesis)

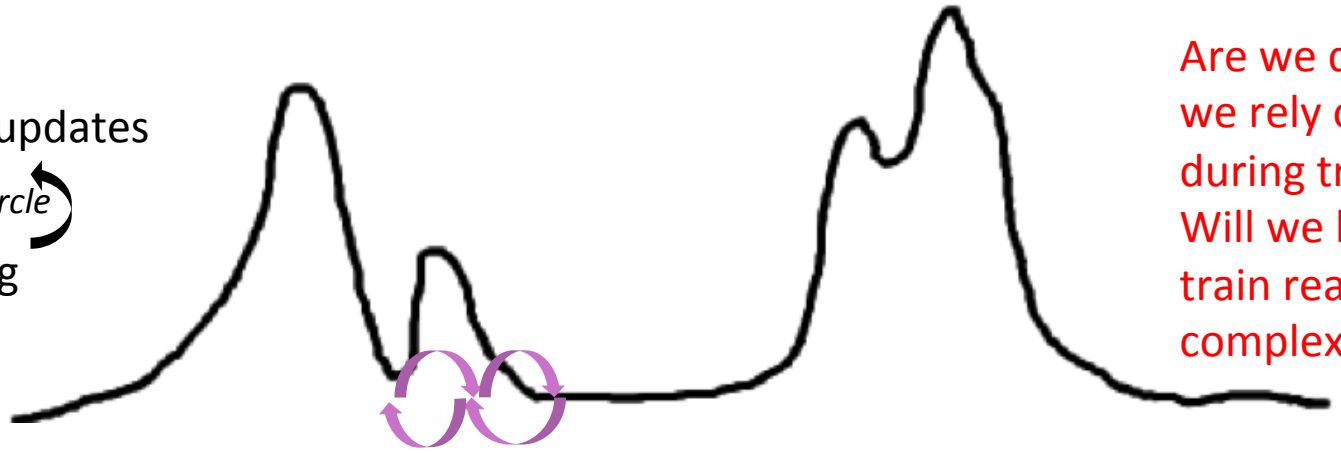
2. Many non-negligeable modes
   (both in posterior & joint distributions)

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

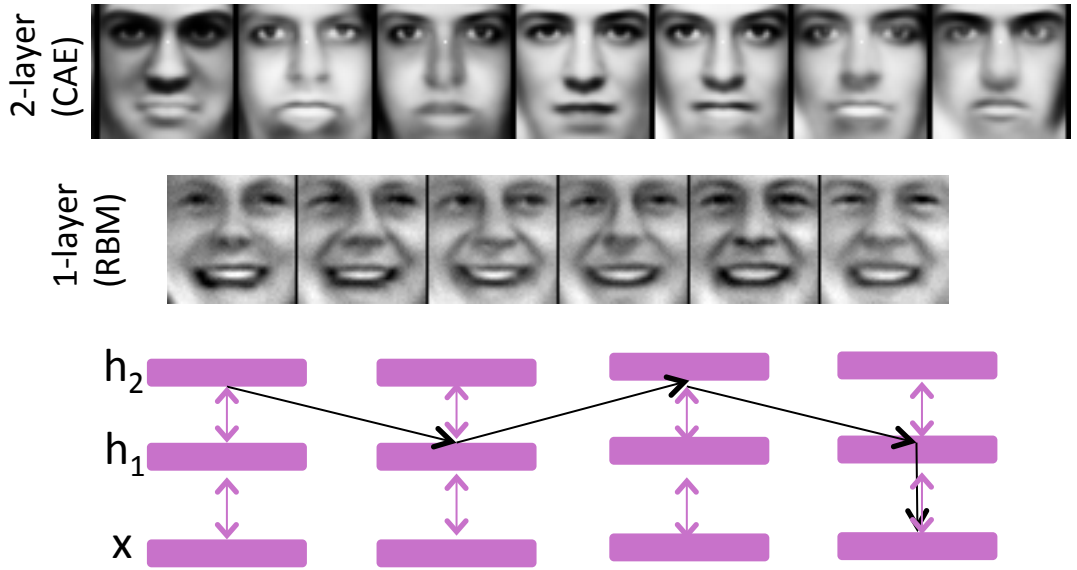


Are we doomed if we rely on MCMC during training? Will we be able to train really large & complex models?

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



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?