Deep Learning Theory

Yoshua Bengio

November 20, 2015

Toyota Technological Institute at Chicago

Université de Montréal

PLUG: **Deep Learning**, MIT Press book in preparation, draft chapters online for feedback

Progress in Deep Learning Theory

- Exponential advantage of distributed representations
- Exponential advantage of depth
- Myth-busting : non-convexity & local minima
- Probabilistic interpretations of auto-encoders

Machine Learning, AI & No Free Lunch

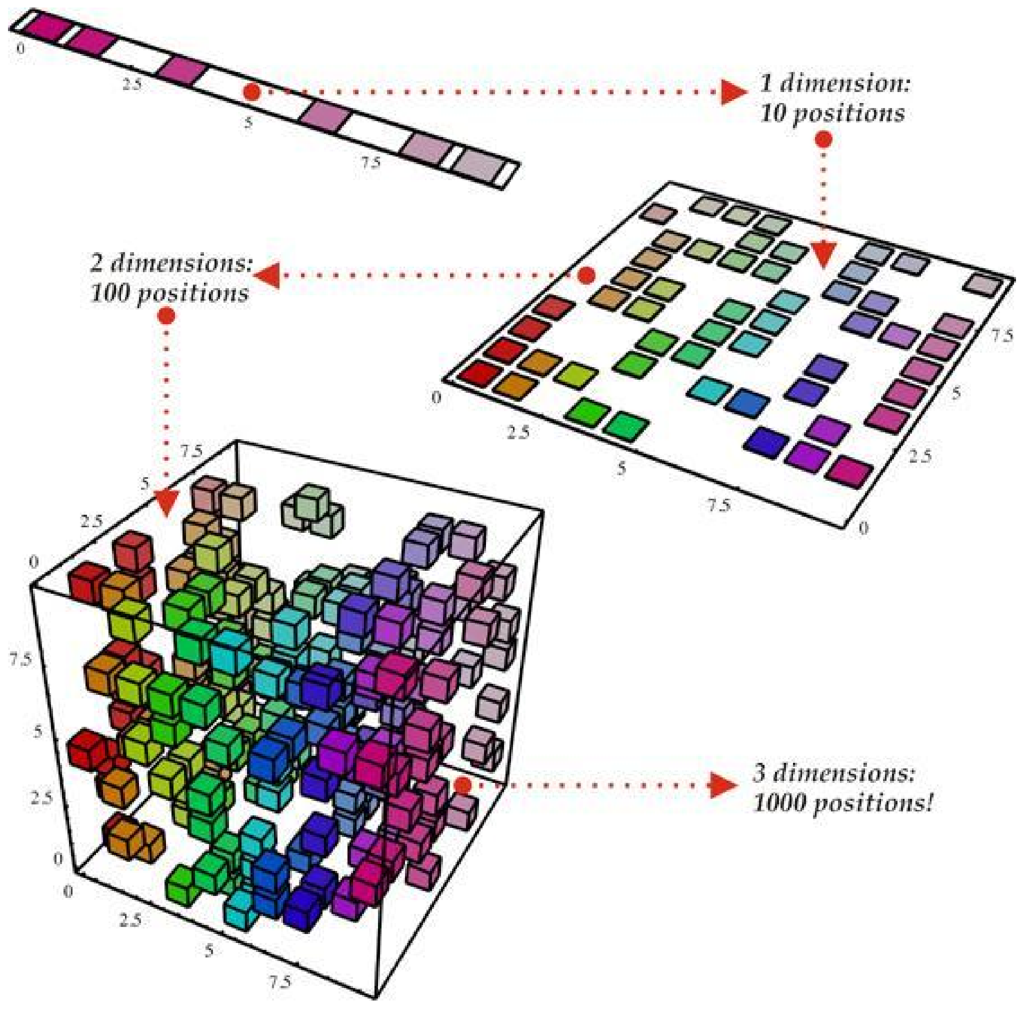
- Four key ingredients for ML towards AI

1. Lots & lots of data
2. Very flexible models
3. Enough computing power
4. Powerful priors that can defeat the curse of dimensionality

ML 101. What We Are Fighting Against: The Curse of Dimensionality

To generalize locally, need representative examples for all relevant variations!

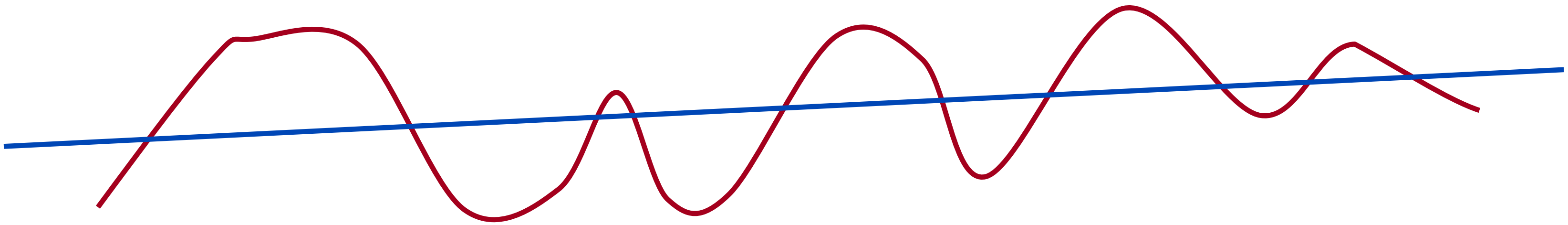
Classical solution: hope for a smooth enough target function, or make it smooth by handcrafting good features / kernel



Not Dimensionality so much as Number of Variations

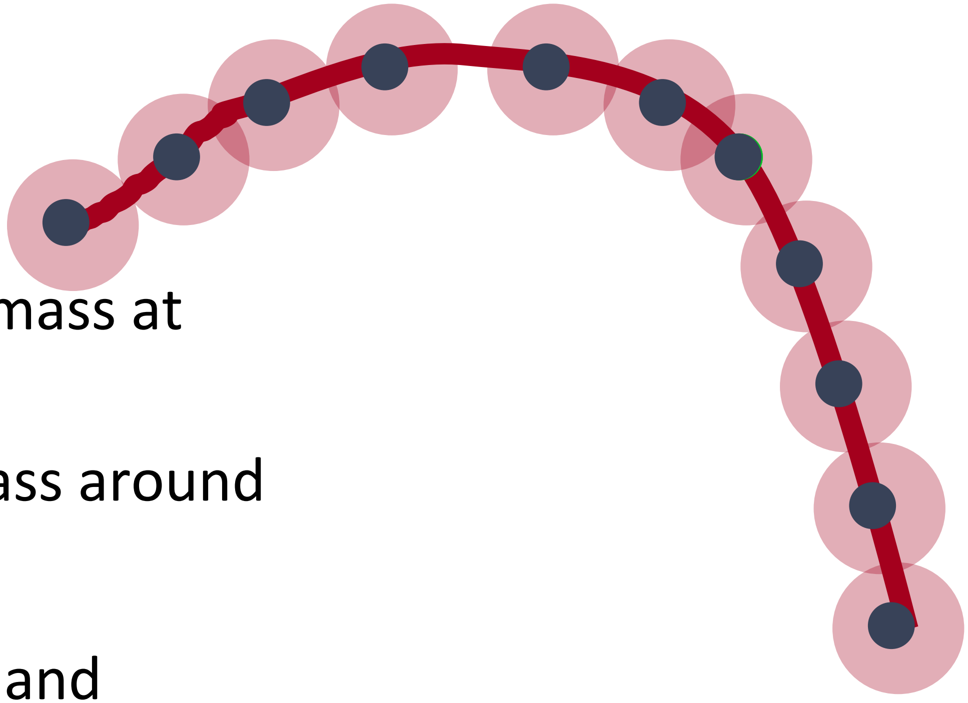
(Bengio, Dellalleau & Le Roux 2007)

- Theorem: Gaussian kernel machines need at least *k* examples to learn a function that has *2k* zero-crossings along some line



- Theorem: For a Gaussian kernel machine to learn some maximally varying functions over *d* inputs requires $O(2^d)$ examples

Putting Probability Mass where Structure is Plausible



- Empirical distribution: mass at training examples
- Smoothness: spread mass around
- Insufficient
- Guess some ‘structure’ and generalize accordingly

Bypassing the curse of dimensionality

We need to build compositionality into our ML models

Just as human languages exploit compositionality to give representations and meanings to complex ideas

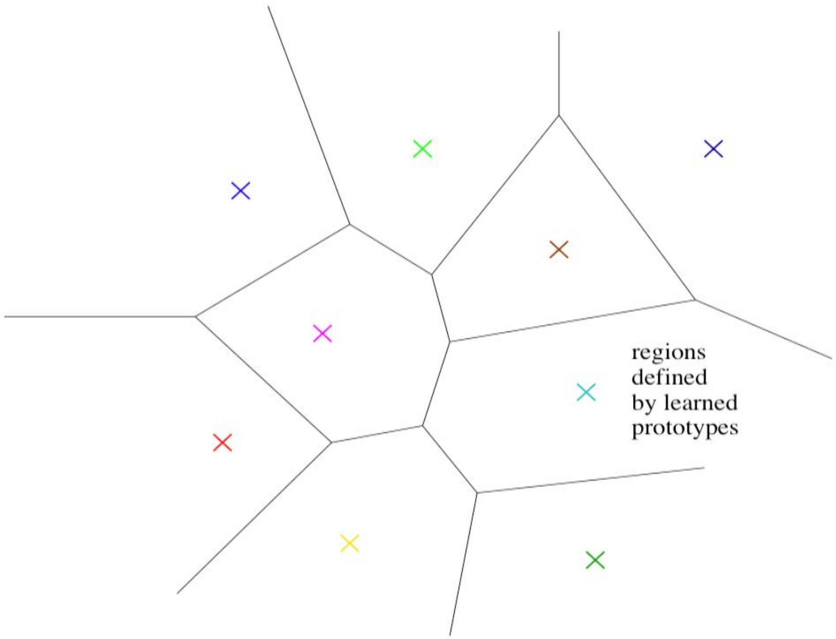
Exploiting compositionality gives an exponential gain in representational power

(1) Distributed representations / embeddings: feature learning

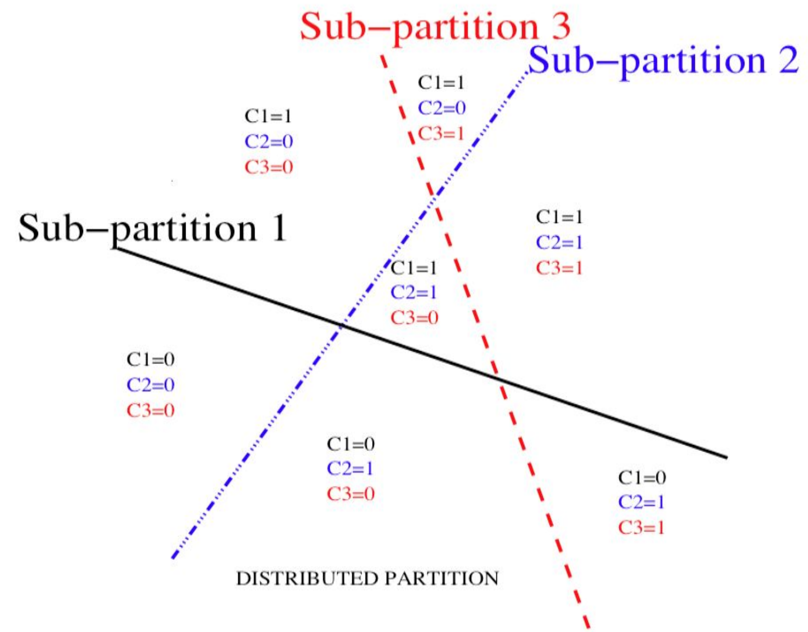
(2) Deep architecture: multiple levels of feature learning

Additional prior: compositionality is useful to describe the world around us efficiently

Exponential advantage of distributed representations



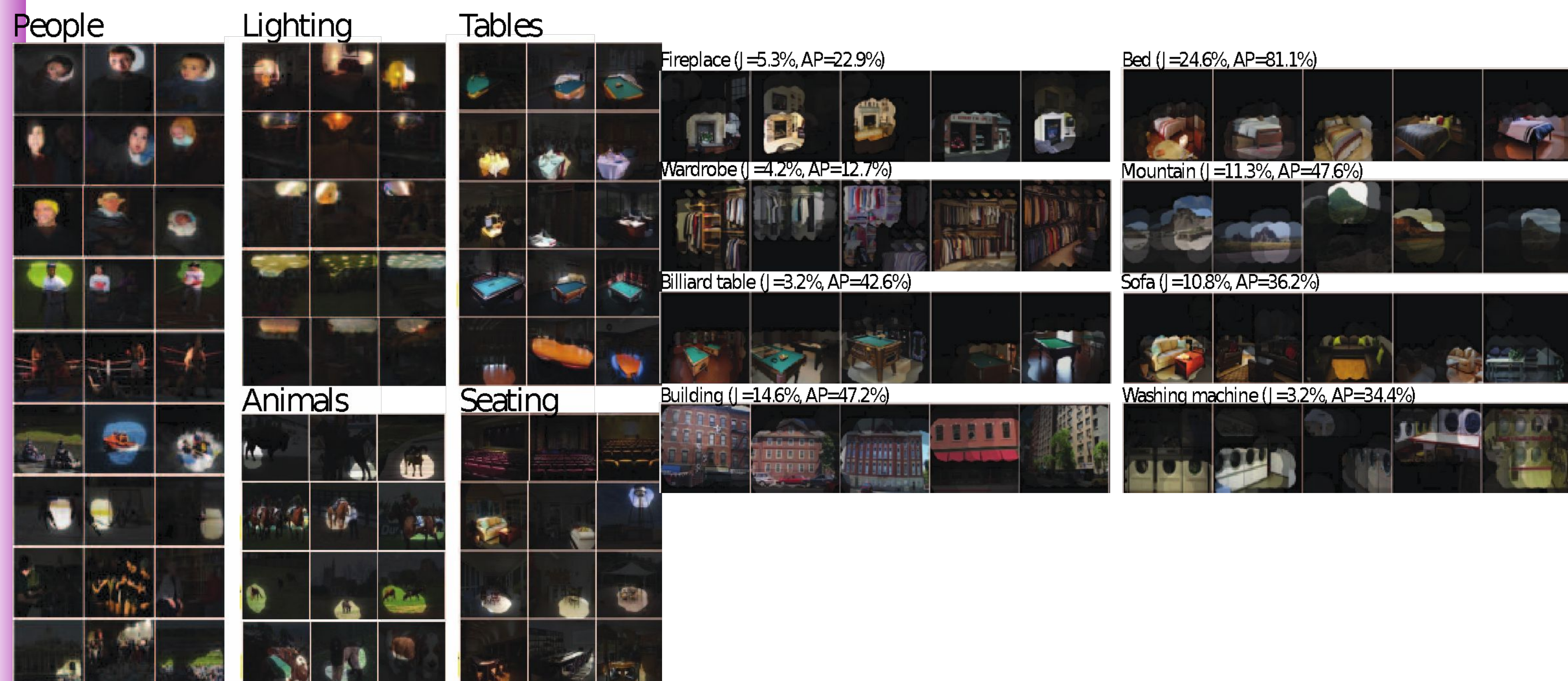
LOCAL PARTITION



Learning a **set of parametric features** that are not mutually exclusive can be exponentially more statistically efficient than having nearest-neighbor-like or clustering-like models

Hidden Units Discover Semantically Meaningful Concepts

- Zhou et al & Torralba, arXiv1412.6856 submitted to ICLR 2015
- Network trained to recognize places, not objects



Each feature can be discovered without the need for seeing the exponentially large number of configurations of the other features

- Consider a network whose hidden units discover the following features:
  - Person wears glasses
  - Person is female
  - Person is a child
  - Etc.

If each of n feature requires $O(k)$ parameters, need $O(nk)$ examples

Non-parametric methods would require $O(2^n)$ examples

Exponential advantage of distributed representations

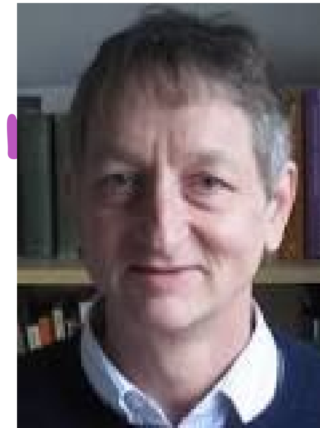
- *Bengio 2009* (Learning Deep Architectures for AI, F & T in ML)
- *Montufar & Morton 2014* (When does a mixture of products contain a product of mixtures? SIAM J. Discr. Math)

- Longer discussion and relations to the notion of priors: ***Deep Learning***, to appear, MIT Press.

- Prop. 2 of ***Pascanu, Montufar & Bengio ICLR'2014***: number of pieces distinguished by 1-hidden-layer rectifier net with *n* units and *d* inputs (i.e. *O(nd)* parameters) is

$$\sum_{j=0}^{d} \binom{n}{j} = O(n^d)$$

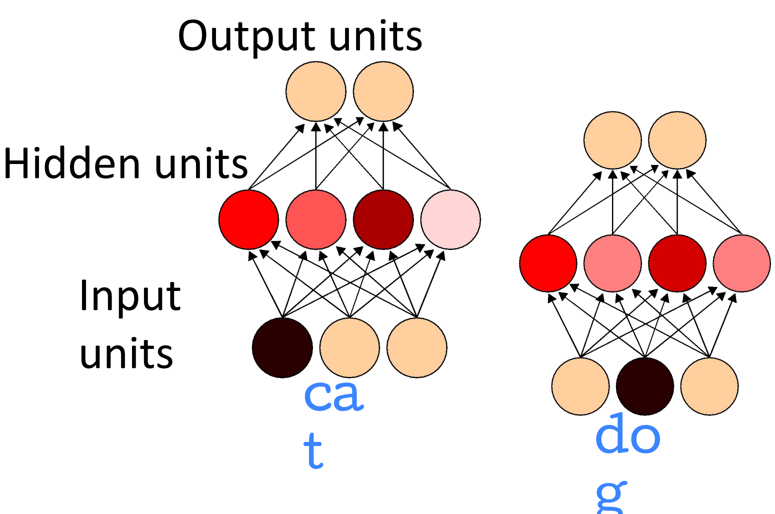
Classical Symbolic AI vs Represen Learning

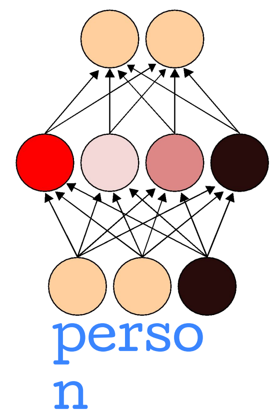


Geoffrey Hinton

- Two symbols are equally far from each other
- Concepts are not represented by symbols in our brain, but by patterns of activation

(Connectionism, 1980's)

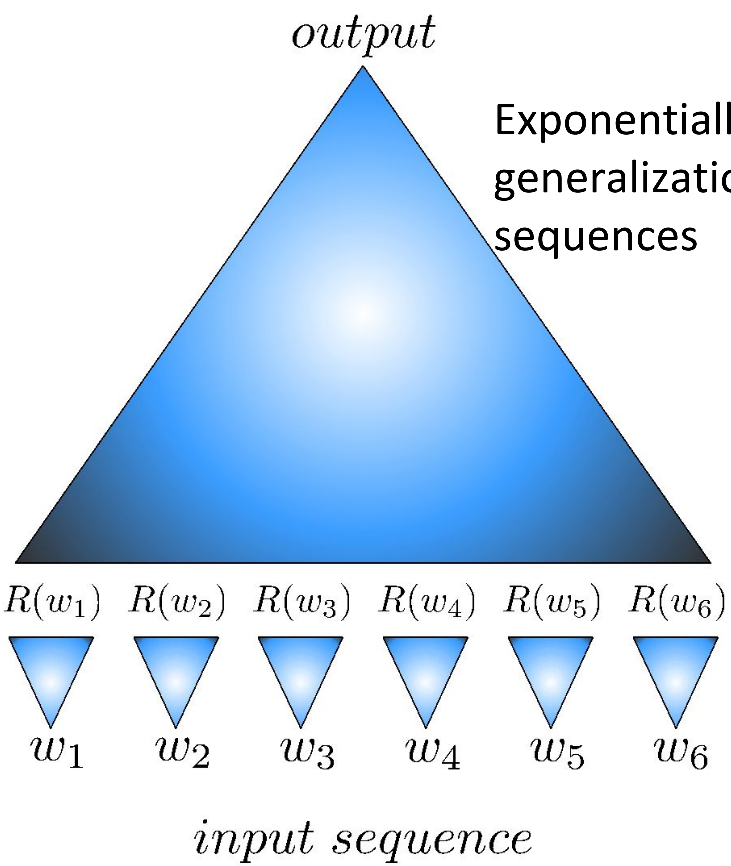




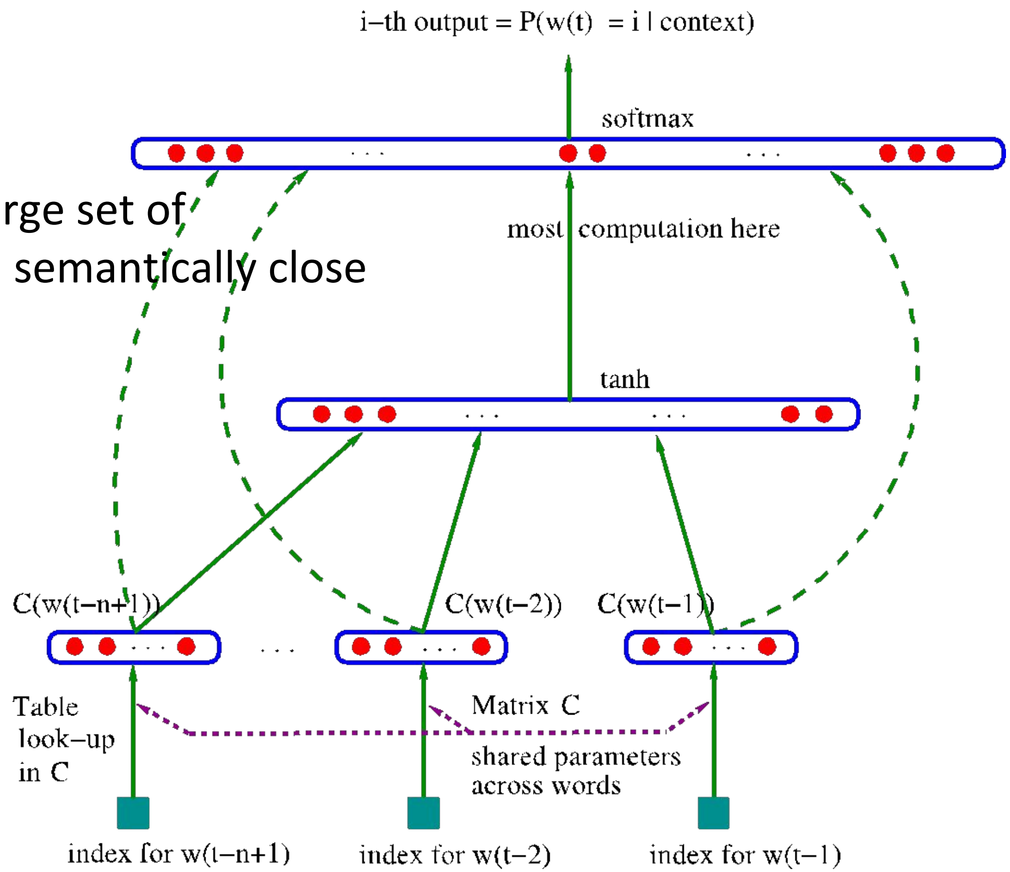
David Rumelhart

Neural Language Models: fighting one exponential by another one!

- (Bengio et al NIPS'2000)



Exponentially large set of generalizations: semantically close sequences

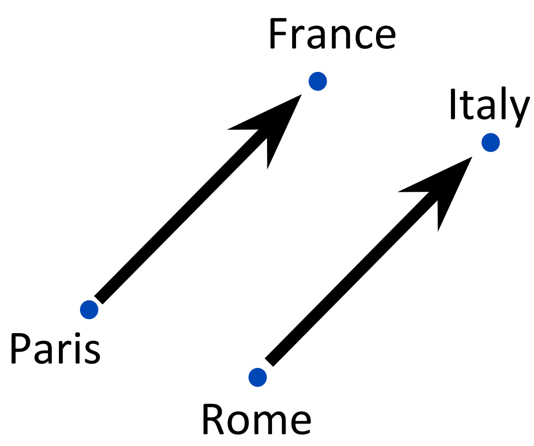


Exponentially large set of possible contexts

Neural word embeddings: visualization directions = Learned Attributes

Analogical Representations for Free (Mikolov et al, ICLR 2013)

- Semantic relations appear as linear relationships in the space of learned representations
- King – Queen $\approx$ Man – Woman
- Paris – France + Italy $\approx$ Rome

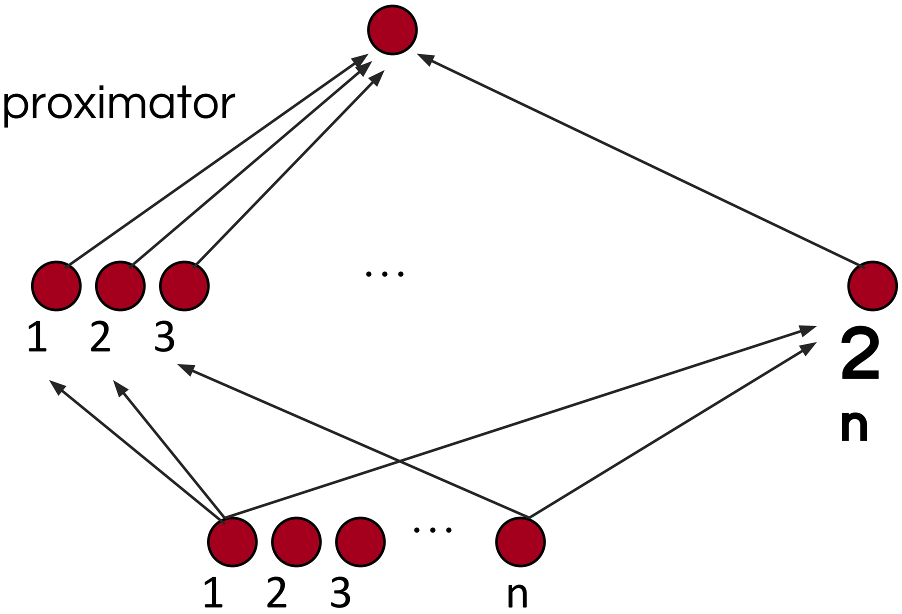


Exponential advantage of depth

Theoretical arguments:

2 layers of { Logic gates / Formal neurons / RBF units } = universal approximator

RBMs & auto-encoders = universal approximator



Theorems on advantage of depth:

(Hastad et al 86 & 91, Bengio et al 2007, Bengio & Delalleau 2011, Martens et al 2013, Pascanu et al 2014, Montufar et al **NIPS 2014**)

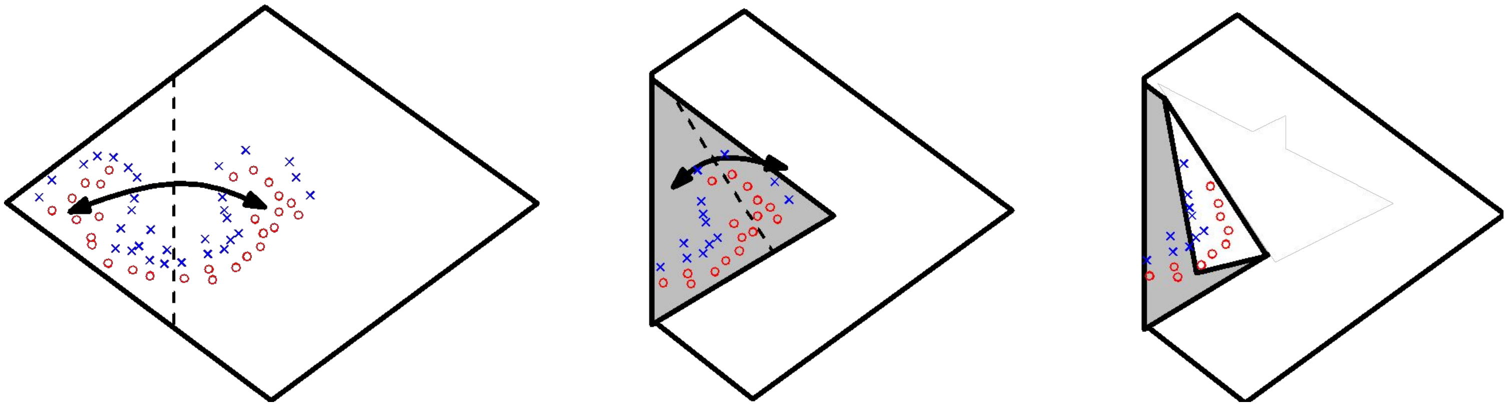
Some functions compactly represented with k layers may require exponential size with 2 layers

Why does it work? No Free Lunch

- It only works because we are making some assumptions about the data generating distribution
- Worse-case distributions still require exponential data
- But the world has structure and we can get an exponential gain by exploiting some of it

Exponential advantage of depth

- Expressiveness of deep networks with piecewise linear activation functions: exponential advantage for depth *(Montufar et al, NIPS 2014)*
- 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:

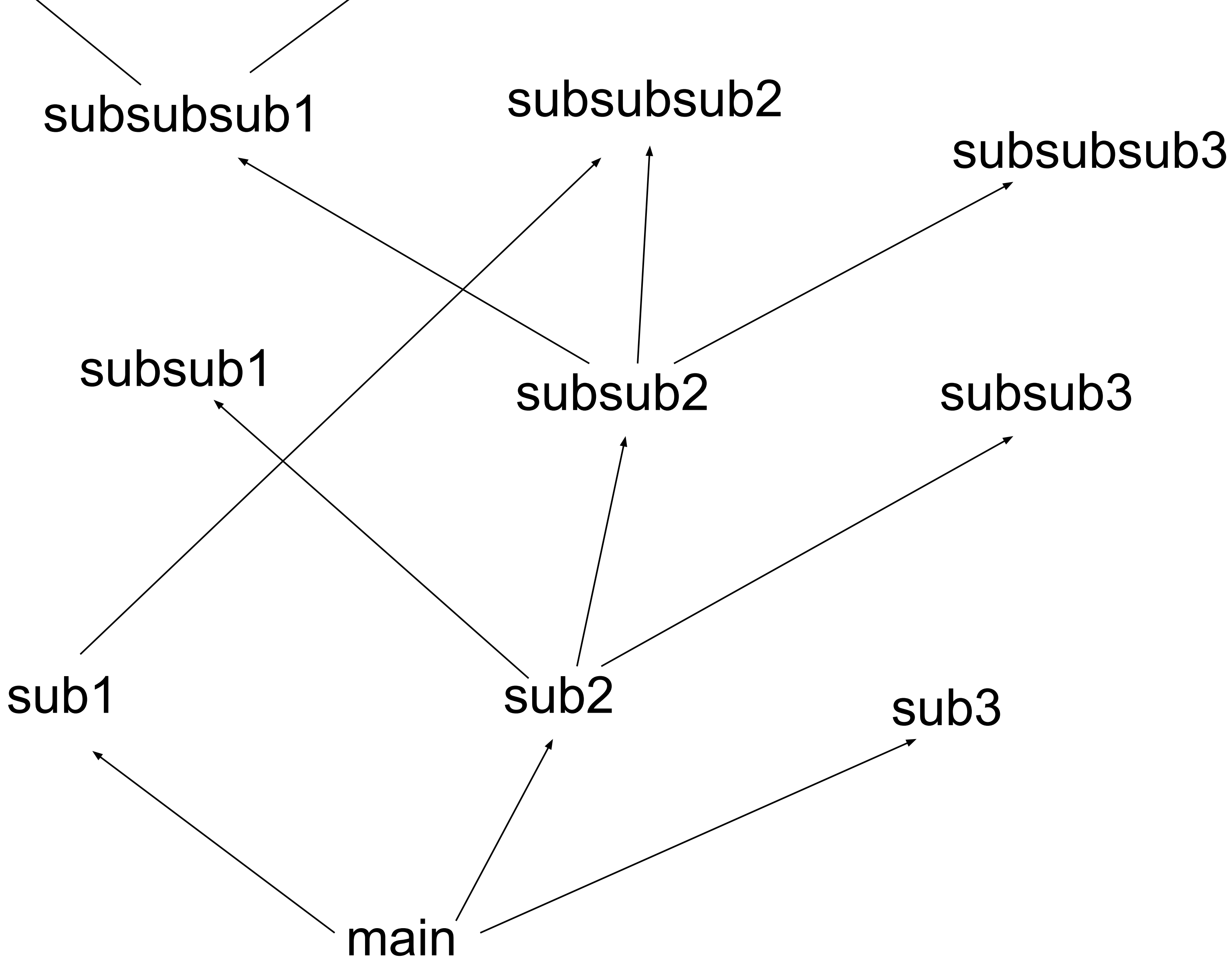


subroutine1 includes
subsub1 code and
subsub2 code and
subsubsub1 code

subroutine2 includes
subsub2 code and
subsub3 code and
subsubsub3 code and …

main

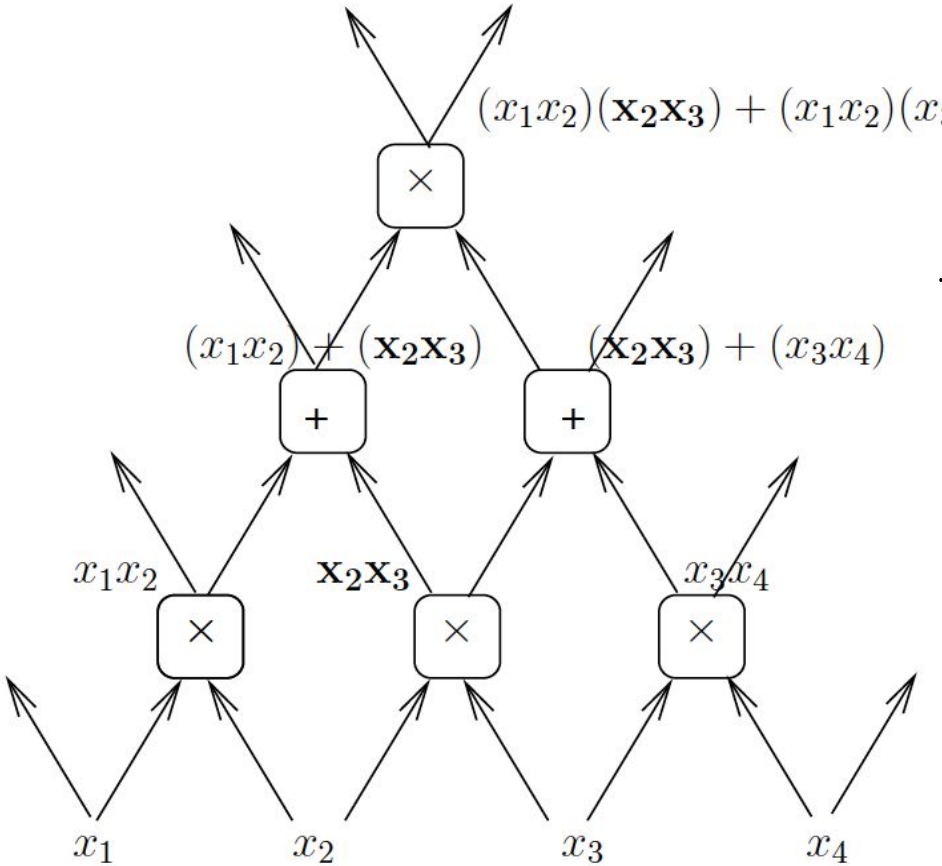
“Shallow” computer program



“Deep” computer program

Sharing Components in a Deep Architecture

Polynomial expressed with shared components: advantage of depth may grow exponentially



Sum-product networks (SPNs)

Theorems in (Bengio & Delalleau, ALT 2011; Delalleau & Bengio NIPS 2011)

However, the expressive power of SPNs used to represent distributions may be fundamentally limited (Martens & Medabalini 2014)

Exponential advantage of depth

- Expressiveness of deep networks with piecewise linear activation functions: exponential advantage for depth *(Montufar et al, NIPS 2014)*
- Number of pieces distinguished for a network with depth L and n_i units per layer is at least

$$\left(\prod_{i=1}^{L-1} \left\lfloor \frac{n_i}{n_0} \right\rfloor^{n_0}\right) \sum_{j=0}^{n_0} \binom{n_L}{j}$$

or, if hidden layers have width n and input has size n_0

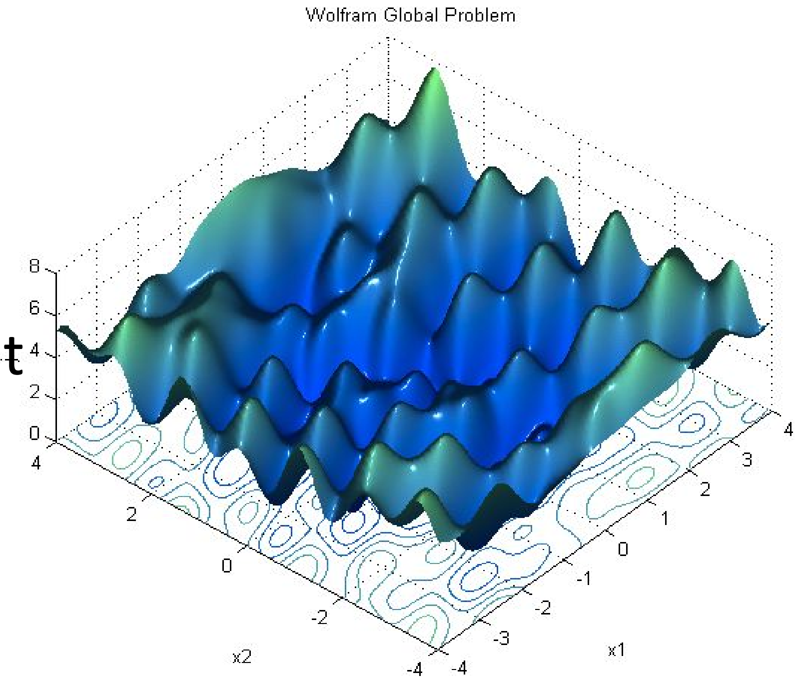
$$\Omega\left(\left({}^{n}/_{n_0} \right)^{(L-1)n_0} n^{n_0} \right)$$

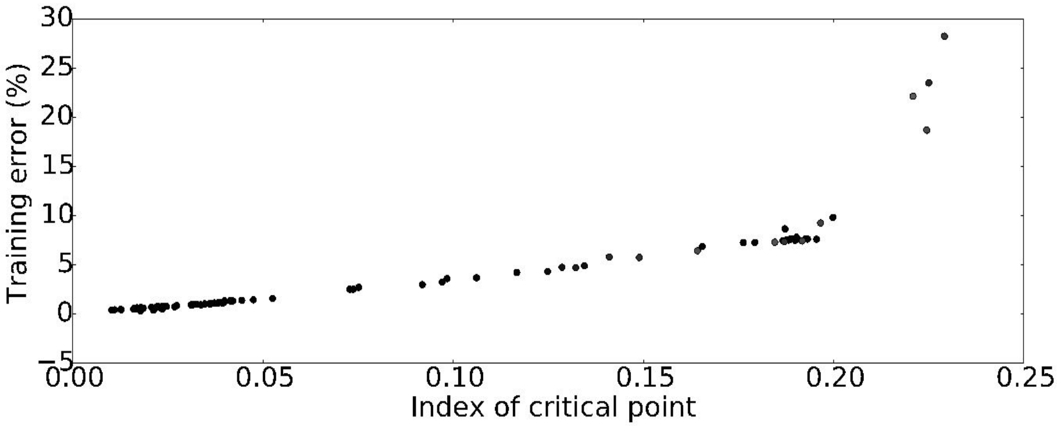
A Myth is Being Debunked: Local Minima in Neural Nets
→ Convexity is not needed

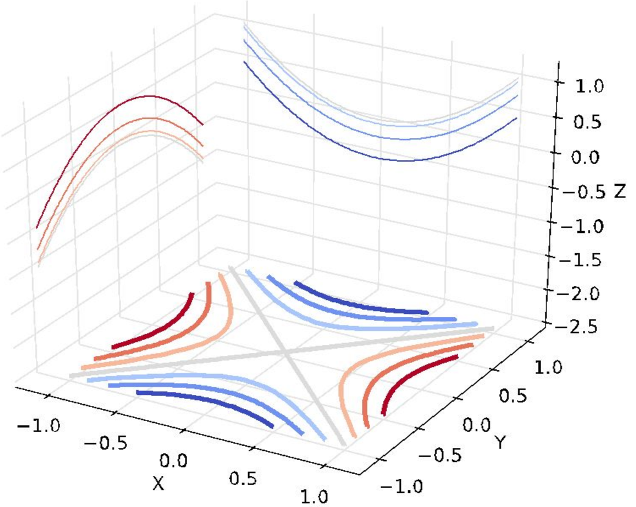
- (Pascanu, Dauphin, Ganguli, Bengio, arXiv May 2014): *On the saddle point problem for non-convex optimization*
- (Dauphin, Pascanu, Gulcehre, Cho, Ganguli, Bengio, NIPS’ 2014): *Identifying and attacking the saddle point problem in high-dimensional non-convex optimization*
- (Choromanska, Henaff, Mathieu, Ben Arous & LeCun, AISTATS’ 2015): *The Loss Surface of Multilayer Nets*

Saddle Points

- Local minima dominate in low-D, but saddle points dominate in high-D
- Most local minima are close to the bottom (global minimum error)

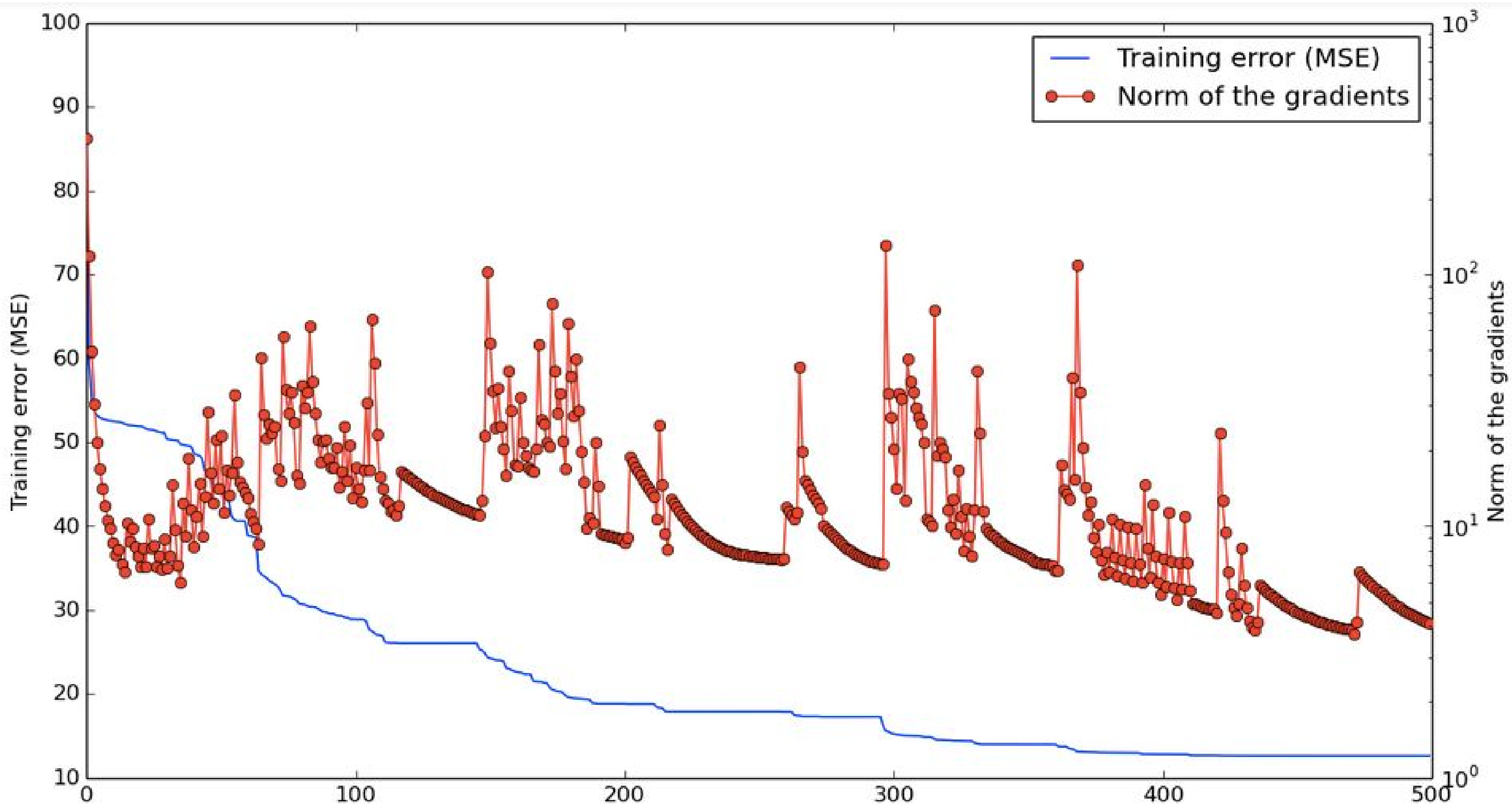






Saddle Points During Training

- Oscillating between two behaviors:
  - Slowly approaching a saddle point
  - Escaping it

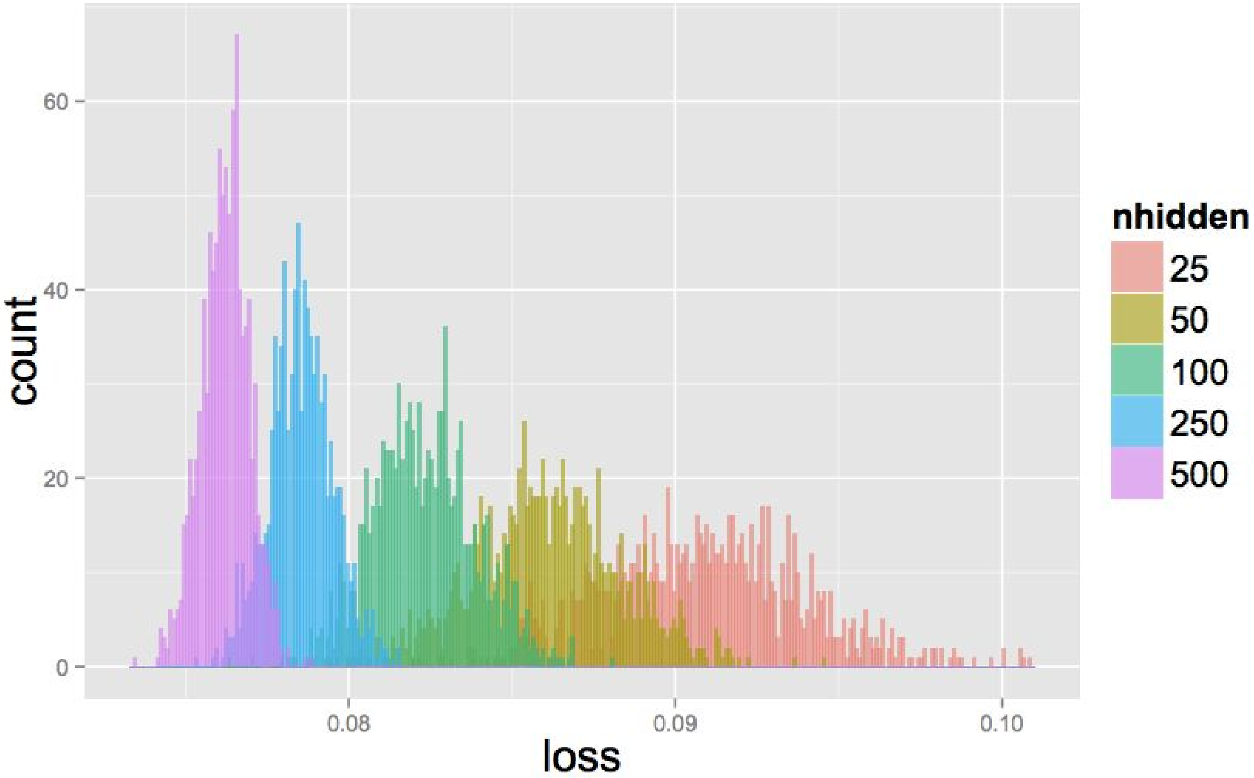


Low Index Critical Points

Choromanska et al & LeCun 2014, 'The Loss Surface of Multilayer Nets'

Shows that deep rectifier nets are analogous to spherical spin-glass models

The low-index critical points of large models concentrate in a band just above the global minimum



The Next Challenge: Unsupervised Learning

- Recent progress mostly in supervised DL
- Real technical 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

Why Latent Factors & Unsupervised Representation Learning? Because of *Causality*.

- If Ys of interest are among the causal factors of X, then

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

is tied to P(X) and P(X|Y), and P(X) is defined in terms of P(X|Y), i.e.

- The best possible model of X (unsupervised learning) MUST involve Y as a latent factor, implicitly or explicitly.
- Representation learning SEEKS the latent variables H that explain the variations of X, making it likely to also uncover Y.

Probabilistic interpretation of auto-encoders

- Manifold & probabilistic interpretations of auto-encoders
  - Denoising Score Matching as inductive principle

    (Vincent 2011)
  - Estimating the gradient of the energy function

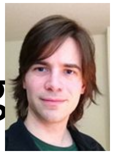
    (Alain & Bengio ICLR 2013)
  - Sampling via Markov chain

    (Bengio et al NIPS 2013; Sohl-Dickstein et al ICML 2015)
  - Variational auto-encoders

    (Kingma & Welling ICLR 2014)

    (Gregor et al arXiv 2015)

Denoising Auto-Encoder

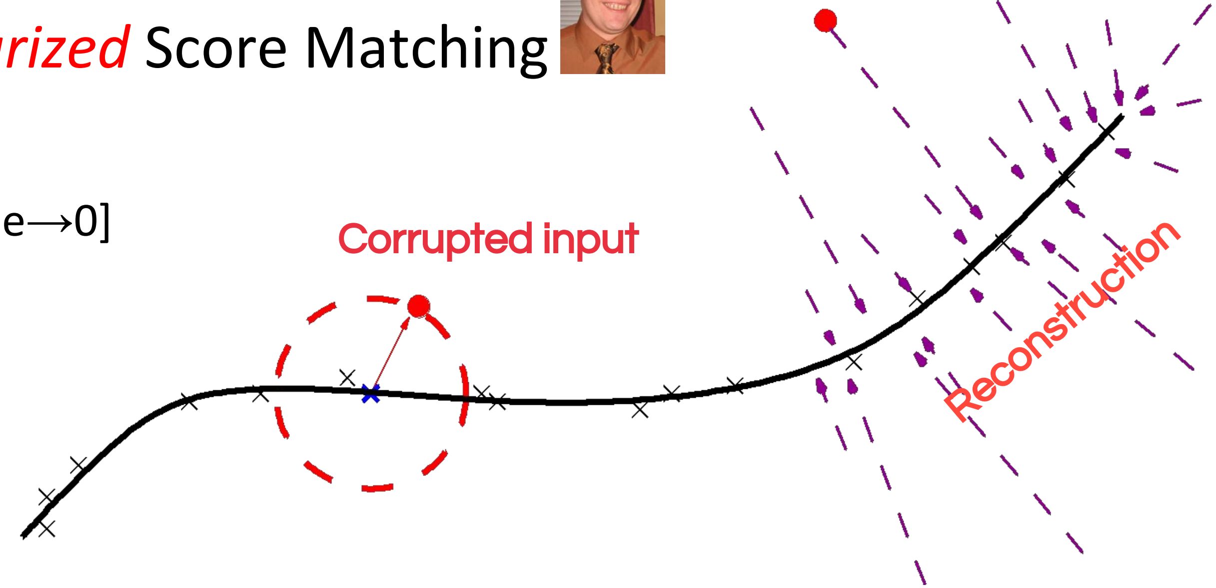
- Learns a vector field pointing towards hig probability direction (Alain & Bengio 2013)

$$\text{reconstruction}(x) - x \quad \rightarrow \quad \sigma^2 \frac{\partial \log p(x)}{\partial x}$$

- Some DAEs correspond to a kind of Gaan RBM with *regularized* Score Matching (Vincent 2011)

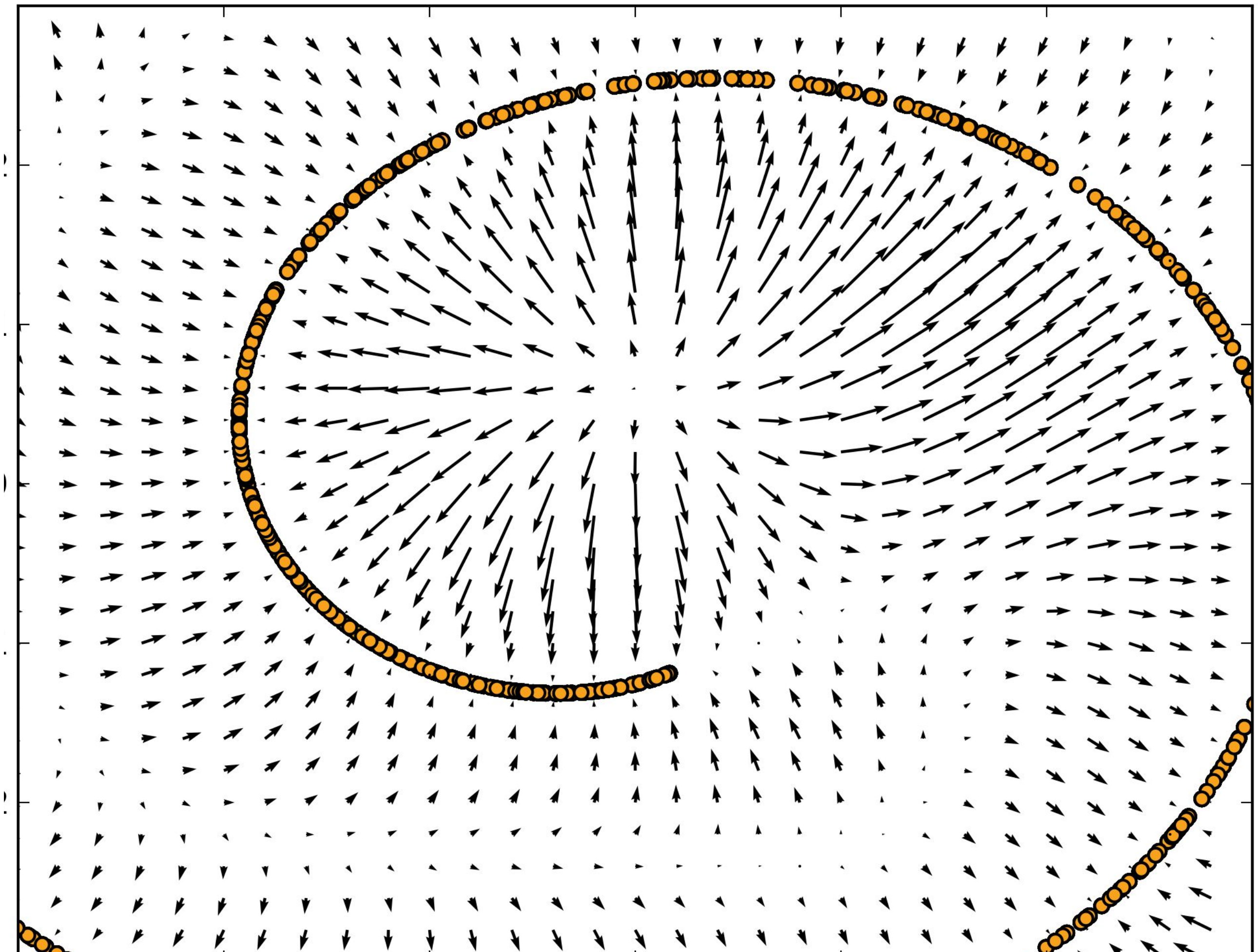
  [equivalent when noise→0]

prior: examples concentrate near a lower dimensional "manifold"

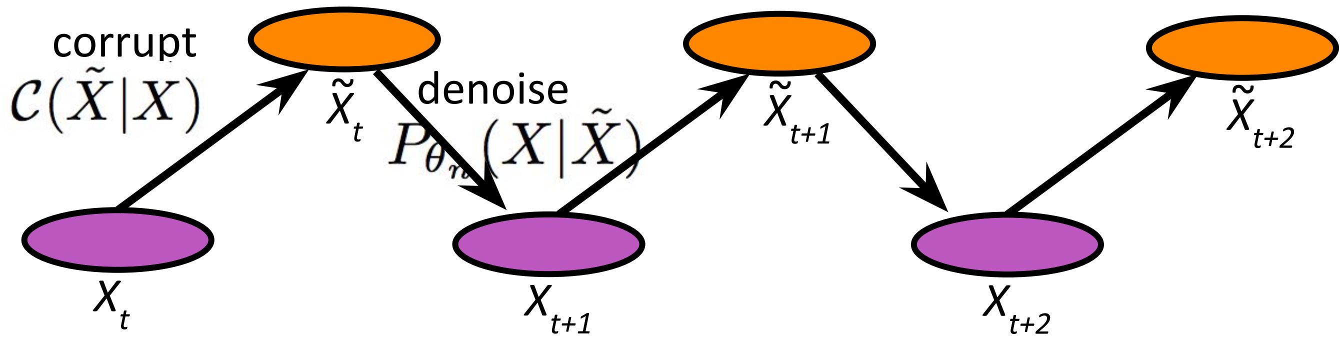


Regularized Auto-Encoders Learn a Vector Field that Estimates a Gradient Field

(Alain & Bengio ICLR 2013)



Denoising Auto-Encoder Markov Chain



The corrupt-encode-decode-sample Markov chain associated with a DAE samples from a consistent estimator of the data generating distribution

Variational Auto-Encoders (VAEs)

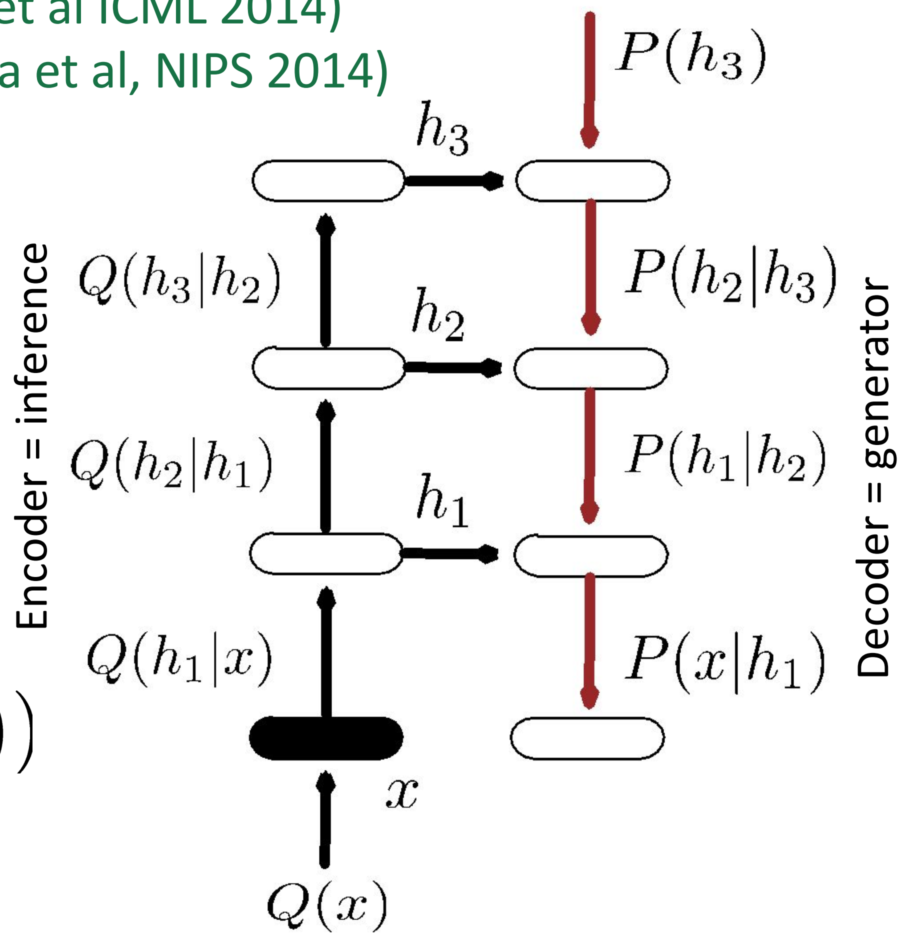
(Kingma & Welling 2013, ICLR 2014)
(Gregor et al ICML 2014; Rezende et al ICML 2014)
(Mnih & Gregor ICML 2014; Kingma et al, NIPS 2014)

- Parametric approximate inference
- Successor of Helmholtz machine (Hinton et al '95)
- Maximize variational lower bound on log-likelihood:

$$\min KL(Q(x,h)||P(x,h))$$

where $Q(x)$ = data distr.

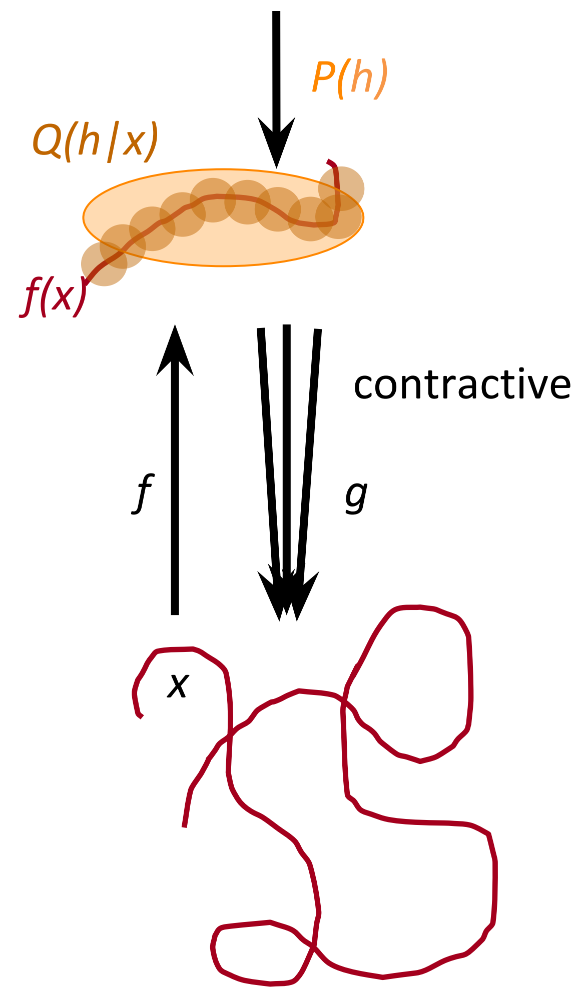
or equivalently



$$\max \sum_x Q(h|x) \log \frac{P(x,h)}{Q(h|x)} = \max \sum_x Q(h|x) \log P(x|h) + KL(Q(h|x)||P(h))$$

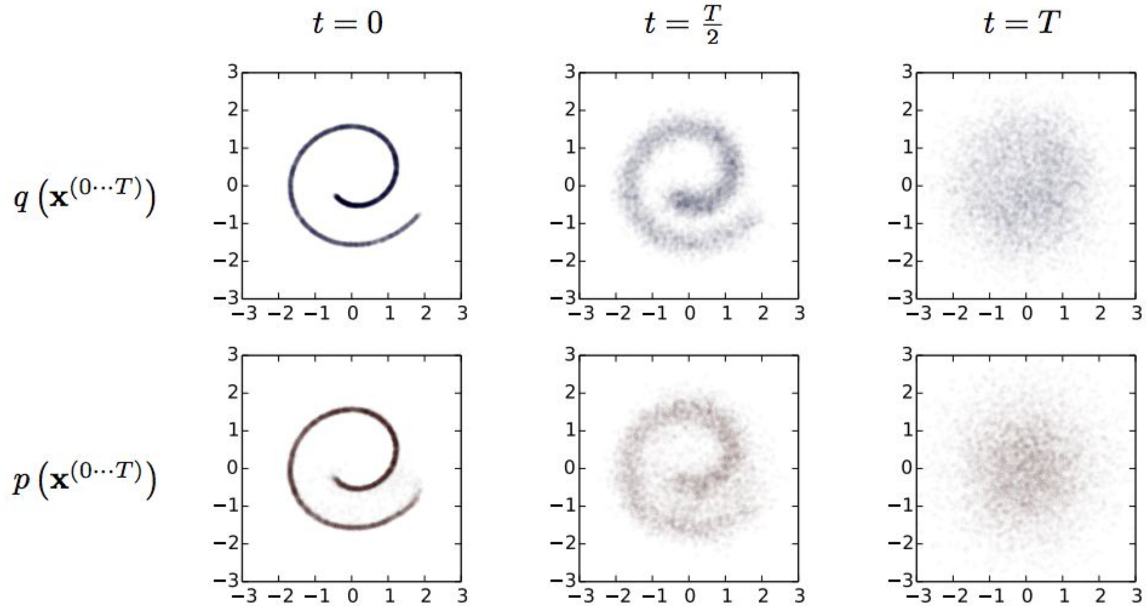
Geometric Interpretation of VAEs

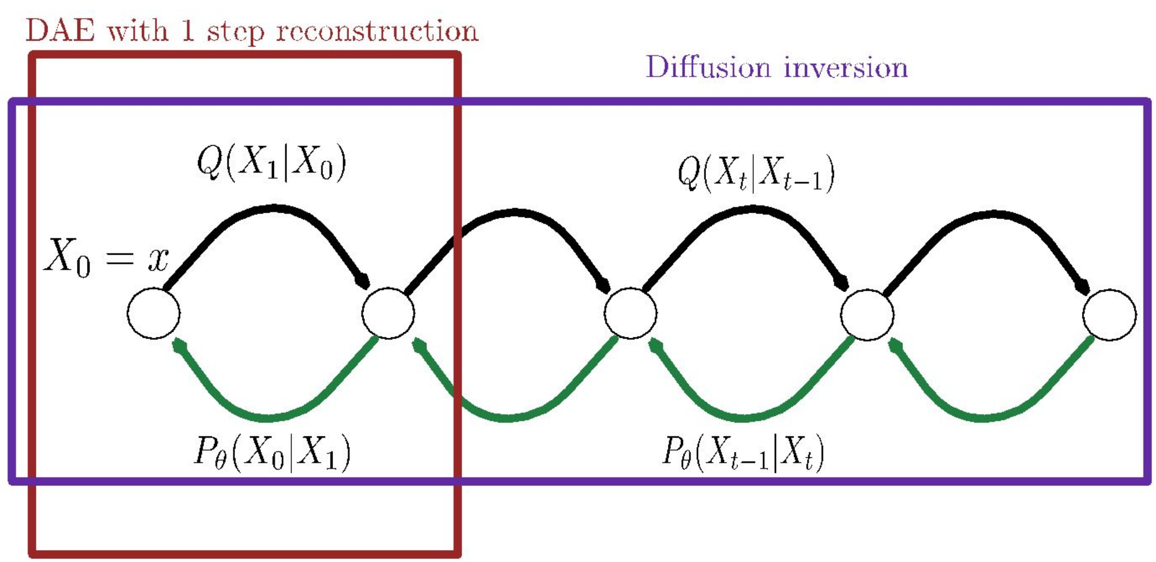
- Encoder: map input to a new space where the data has a simpler distribution
- Add noise between encoder output and decoder input: train the decoder to be robust to mismatch between encoder output and prior output.



Denoising Auto-Encoder vs Diffusion Inverter *(Sohl-Dickstein et al ICML 2015)*

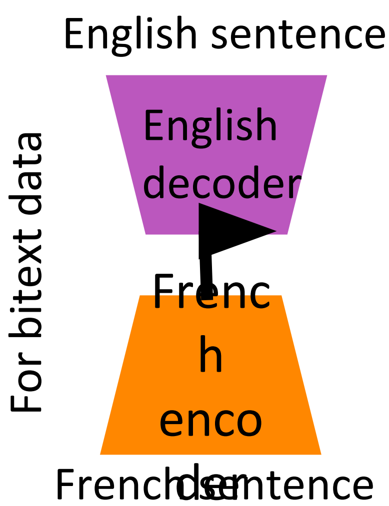
- DAE: after 1 step of diffusion (adding noise, Q), try to reconstruct the clean original (with P).
- Diffusion inverter: after each step of diffusion, try to stochastically undo the effect of diffusion.

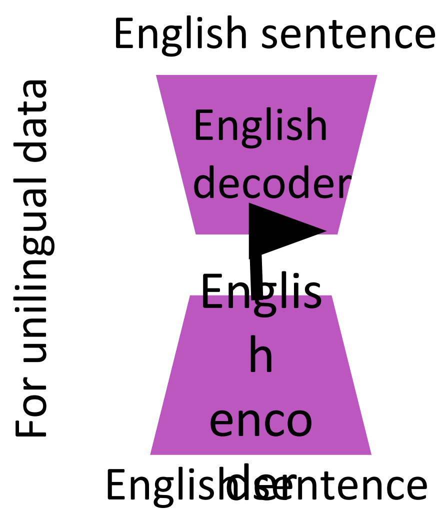


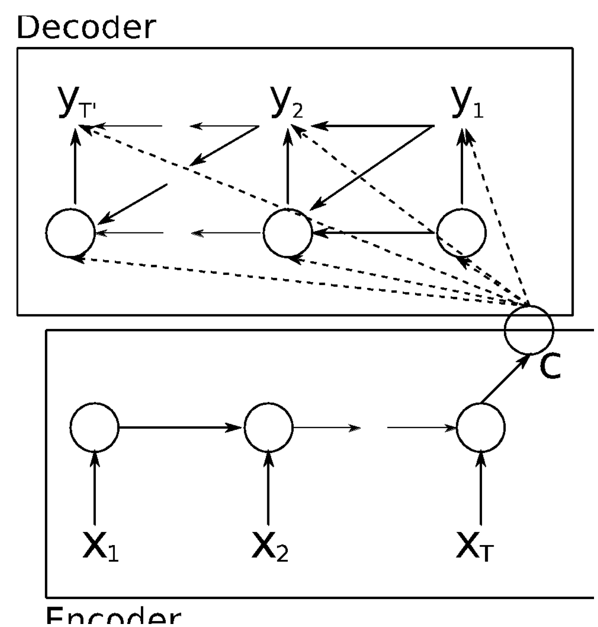


Encoder-Decoder Framework

- Intermediate representation of meaning
  = ‘universal representation’
- Encoder: from word sequence to sentence representation
- Decoder: from representation to word sequence distribution

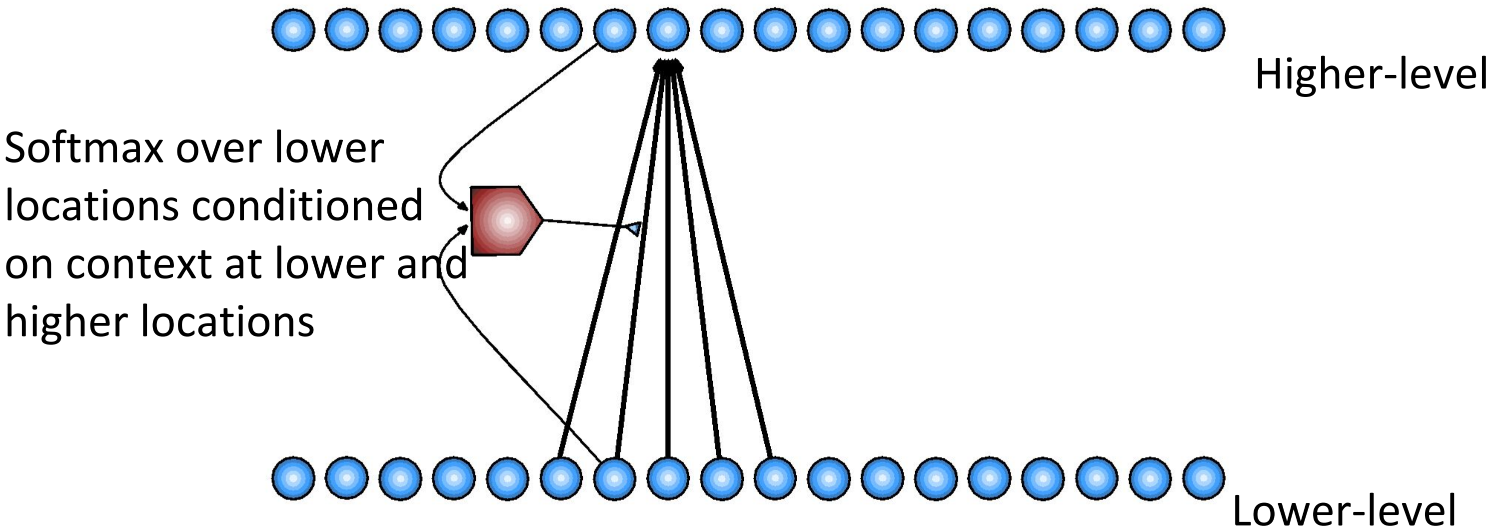






Attention Mechanism for Deep Learning

- Consider an input (or intermediate) sequence or image
- Consider an upper level representation, which can choose « where to look », by assigning a weight or probability to each input position, as produced by an MLP, applied at each position



End-to-End Machine Translation with Recurrent Nets and Attention Mechanism

- Reached the state-of-the-art in one year, from scratch

(a) **English→French (WMT-14)**

	NMT(A)	Google	P-SMT
NMT	32.68	30.6*	**37.03•**
+Cand	33.28	–	
+UNK	33.99	32.7°	
+Ens	**36.71**	**36.9°**	

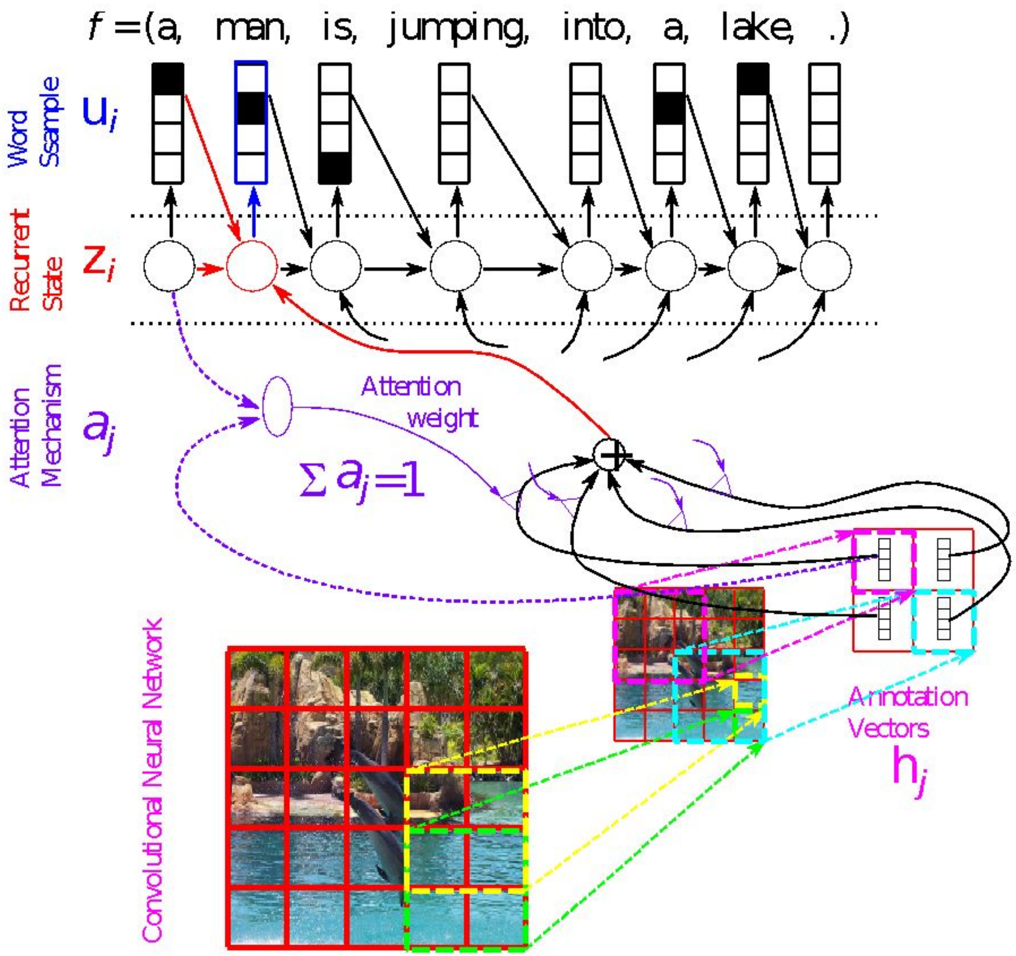
(b) **English→German (WMT-15)**

Model	Note
24.8	Neural MT
24.0	U.Edinburgh, Syntactic SMT
23.6	LIMSI/KIT
22.8	U.Edinburgh, Phrase SMT
22.7	KIT, Phrase SMT

(c) **English→Czech (WMT-15)**

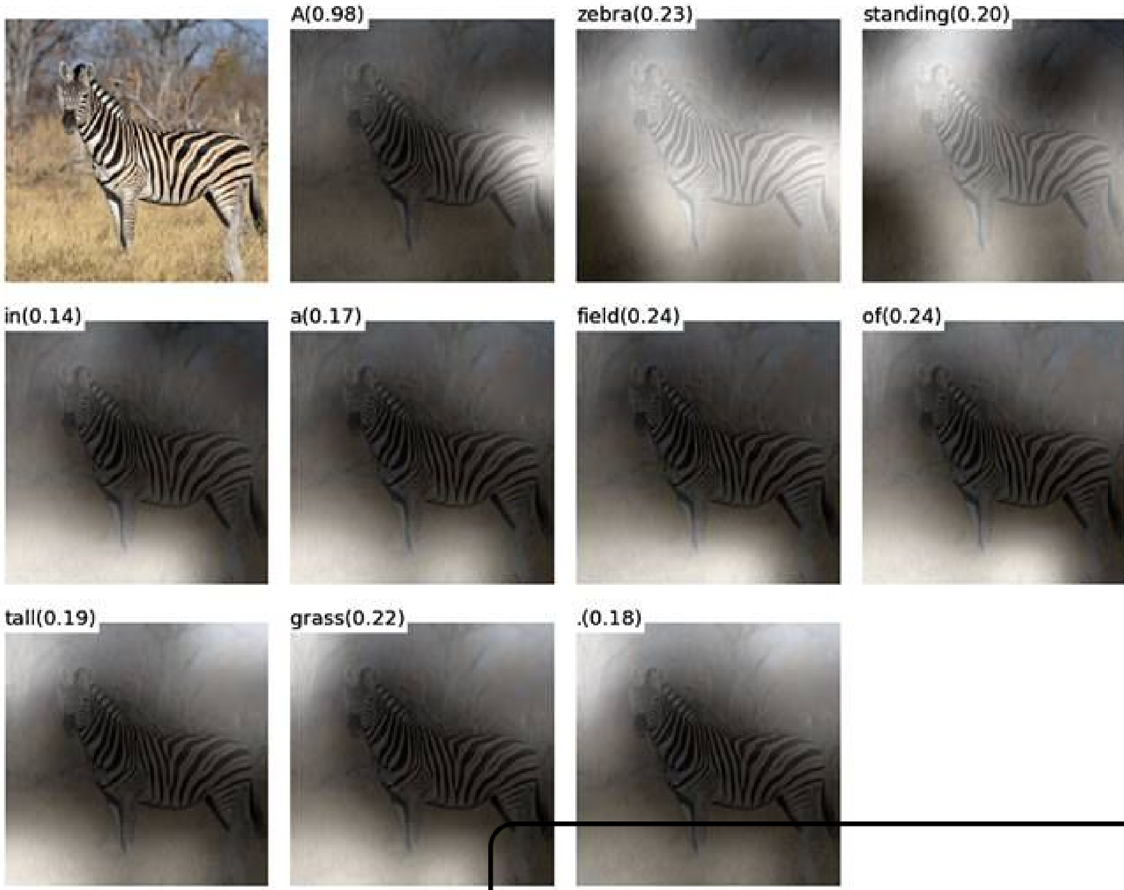
Model	Note
18.3	Neural MT
18.2	JHU, SMT+LM+OSM+Sparse
17.6	CU, Phrase SMT
17.4	U.Edinburgh, Phrase SMT
16.1	U.Edinburgh, Syntactic SMT

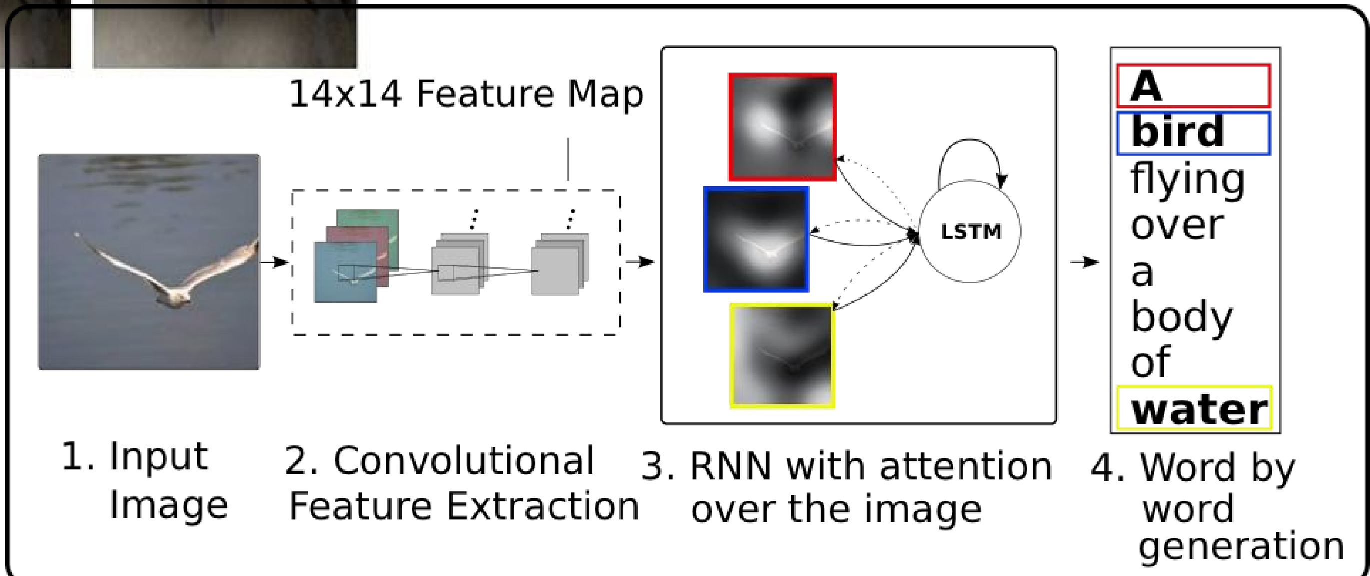
Image-to-Text: Caption Generation with Attention



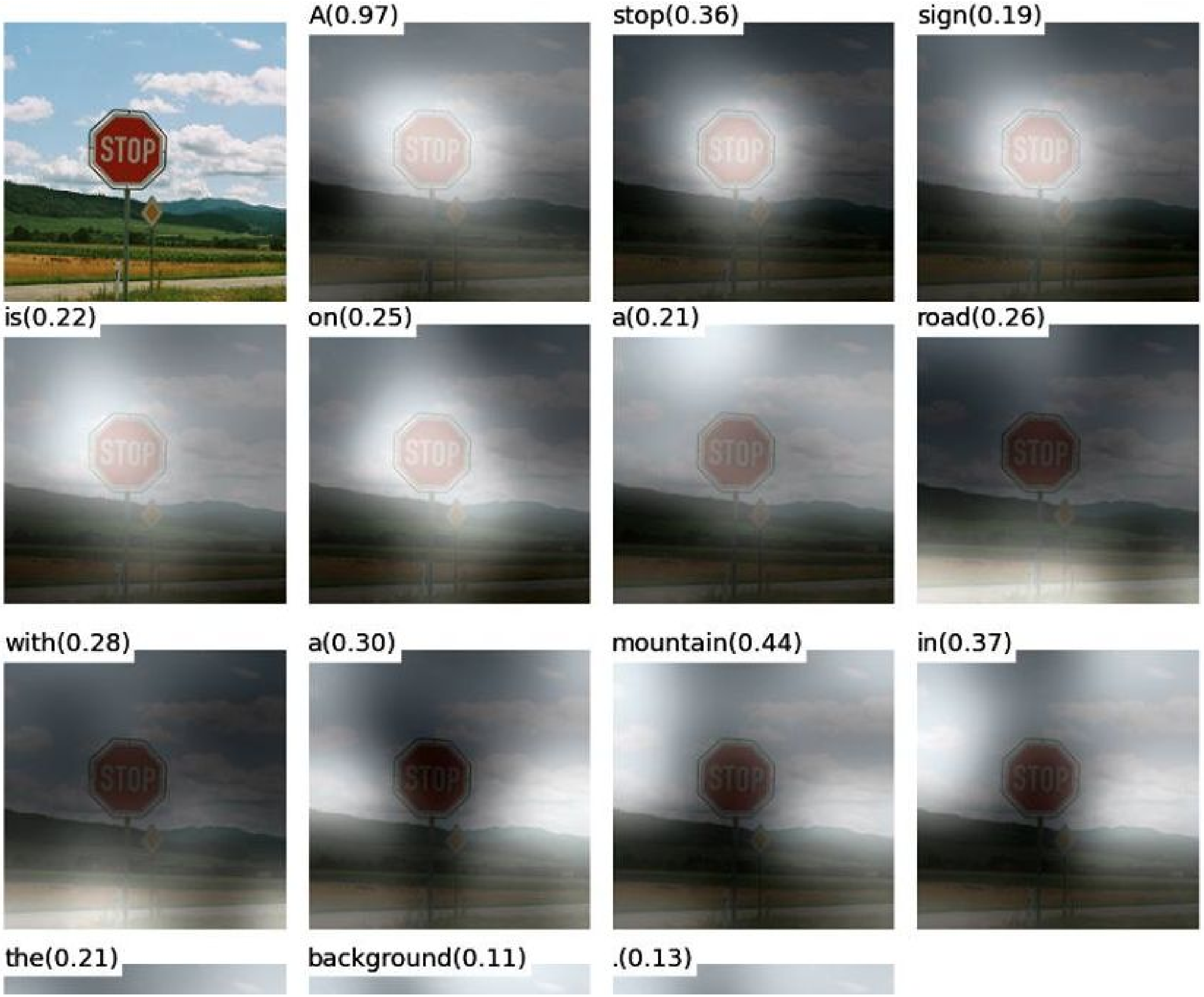
(Xu et al., 2015), (Yao et al., 2015)

Paying Attention to Selected Parts of the Image While Uttering Words





Speaking about what one sees



Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Results from (Xu et al, arXiv Jan. 2015, ICML 2015)

Table 1. BLEU-1,2,3,4/METEOR metrics compared to other methods, † indicates a different split, (—) indicates an unknown metric, ∘ indicates the authors kindly provided missing metrics by personal communication, Σ indicates an ensemble, *a* indicates using AlexNet

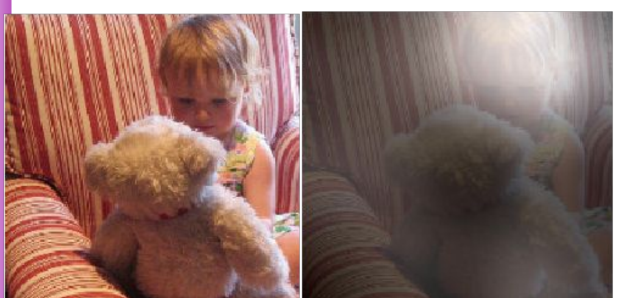
		BLEU				
Dataset	Model	B-1	B-2	B-3	B-4	METEOR
Flickr8k	Google NIC(Vinyals et al., 2014)$^{\dagger\Sigma}$	63	41	27	—	—
	Log Bilinear (Kiros et al., 2014a)$^{\circ}$	65.6	42.4	27.7	17.7	17.31
	Soft-Attention	**67**	44.8	29.9	19.5	18.93
	Hard-Attention	**67**	**45.7**	**31.4**	**21.3**	**20.30**
Flickr30k	Google NIC$^{\dagger\circ\Sigma}$	66.3	42.3	27.7	18.3	—
	Log Bilinear	60.0	38	25.4	17.1	16.88
	Soft-Attention	66.7	43.4	28.8	19.1	**18.49**
	Hard-Attention	**66.9**	**43.9**	**29.6**	**19.9**	18.46
COCO	CMU/MS Research (Chen & Zitnick, 2014)^{a}	—	—	—	—	20.41
	MS Research (Fang et al., 2014)$^{\dagger a}$	—	—	—	—	20.71
	BRNN (Karpathy & Li, 2014)$^{\circ}$	64.2	45.1	30.4	20.3	—
	Google NIC$^{\dagger\circ\Sigma}$	66.6	46.1	32.9	24.6	—
	Log Bilinear$^{\circ}$	70.8	48.9	34.4	24.3	20.03
	Soft-Attention	70.7	49.2	34.4	24.3	**23.90**
	Hard-Attention	**71.8**	**50.4**	**35.7**	**25.0**	23.04

The Good

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

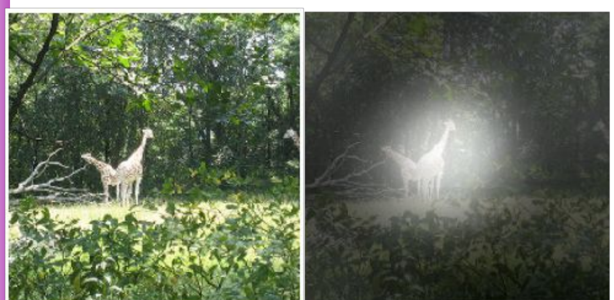


A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

And the Bad



A large white bird standing in a forest.

A woman holding a clock in her hand.

A man wearing a hat and
a hat on a skateboard.

A person is standing on a beach
with a surfboard.

A woman is sitting at a table
with a large pizza.

A man is talking on his cell phone
while another man watches.

The Next Frontier: Reasoning and Question Answering

- Currently working on artificial tasks, with memory networks:

Joe went to the garden then Fred picked up the milk; Joe moved to the bathroom and Fred dropped the milk, and then Dan moved to the living_room.
Where is Dan? A: living room I believe
Where is Joe? A: the bathroom
Fred moved to the bedroom and Joe went to the kitchen then Joe took the milk there and Dan journeyed to the bedroom; Joe discarded the milk.
Where is the milk now ? A: the milk is in the kitchen
Where is Dan now? A: I think he is in the bedroom
Joe took the milk there, after that Mike travelled to the office, then Joe went to the living_room, next Dan went back to the kitchen and Joe travelled to the office.
Where is Joe now? A: I think Joe is in the office

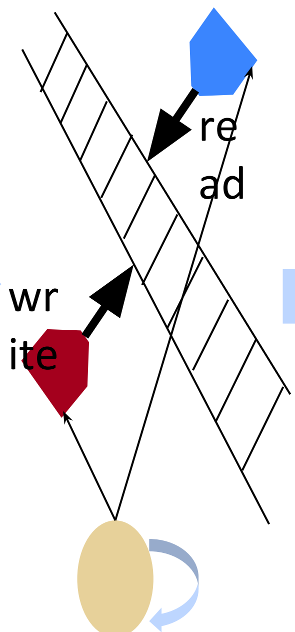
From "Memory Networks", Weston et al. ICLR 2015;
"End-to-end memory networks", Sukhbatar et al. NIPS'2015

Sam walks into the kitchen. Sam picks up an apple. Sam walks into the bedroom. Sam drops the apple. Q: Where is the apple? A. Bedroom	Brian is a lion. Julius is a lion. Julius is white. Bernhard is green. Q: What color is Brian? A. White	Mary journeyed to the den. Mary went back to the kitchen. John journeyed to the bedroom. Mary discarded the milk. Q: Where was the milk before the den? A. Hallway

Ongoing Project: Knowledge Extraction

- Learn to fill the memory network from natural language descriptions of facts
- Force the neural net to understand language
- Extract knowledge from documents into a usable form

WIKIPEDIA
The Free Encyclopedia



Conclusions

- Theory for deep learning has progressed substantially on several fronts:
  - why it generalizes better,
  - why local minima are not the issue people thought, and
  - the probabilistic interpretation of deep auto-encoders

- But more theory would be great! Many things remain mysterious…

MILA: Montreal Institute for Learning Algorithms