Deep Learning: Progress in Theory and Attention Mechanisms

Yoshua Bengio

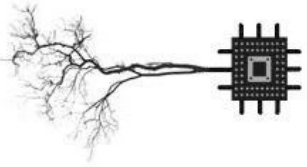
June 11, 2015

Deep Vision CVPR Workshop, Boston, USA

Université de Montréal

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

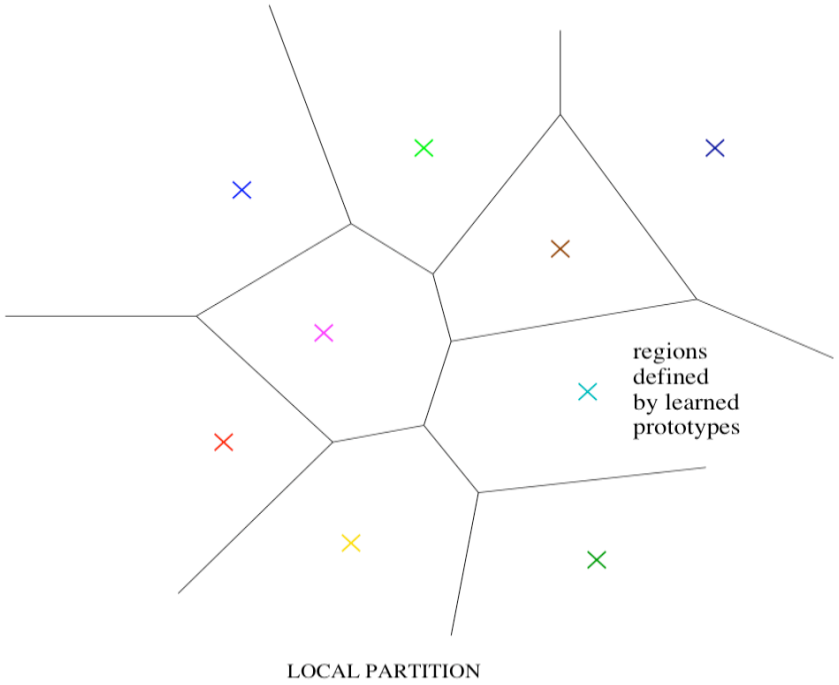
LISA

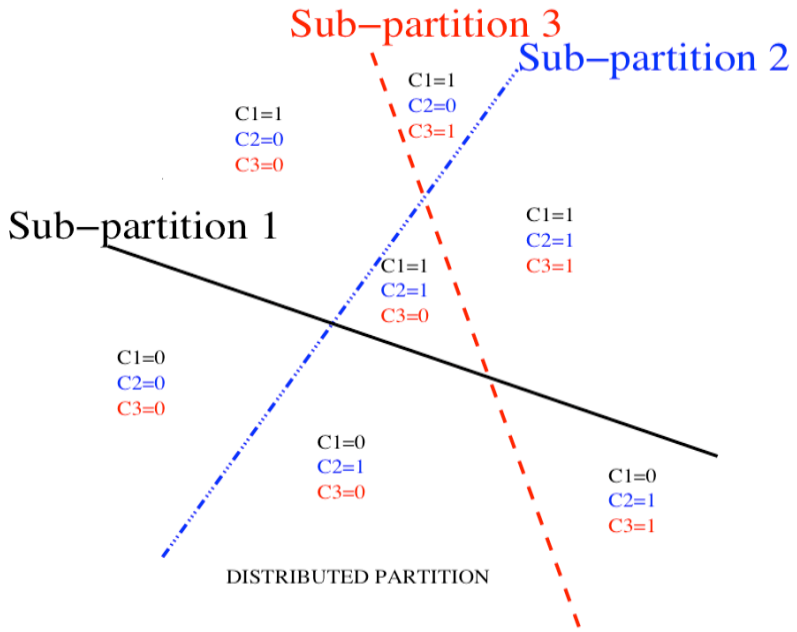


Progress in Deep Learning Theory

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

Exponential advantage of distributed representations





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

Exponential advantage of distributed representations

- *Bengio 2009* (Learning Deep Architectures for AI, Foundations & Trends 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***, Bengio, Goodfellow & Courville, to appear, MIT Press.

Exponential advantage of depth

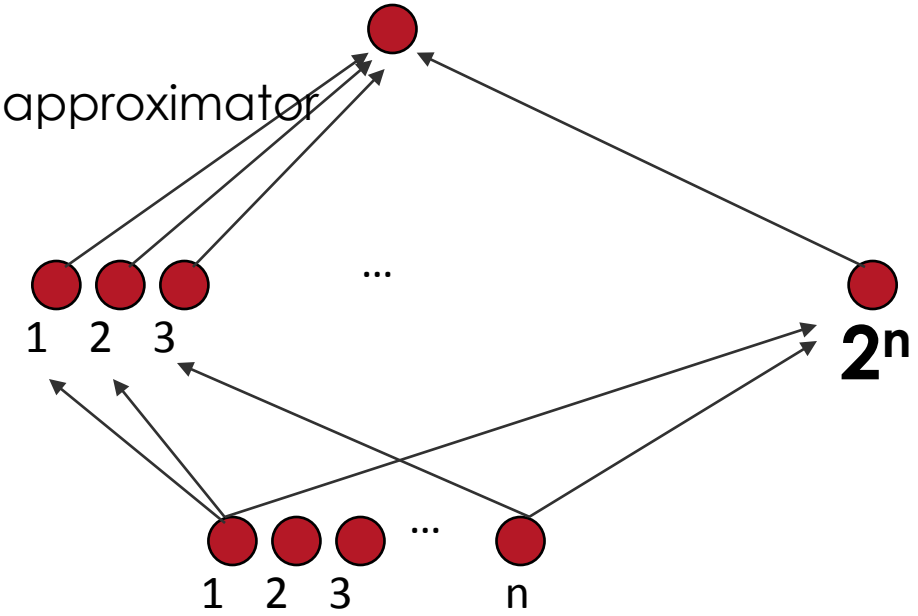
Theoretical arguments:

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

RBMs & auto-encoders = universal approximator

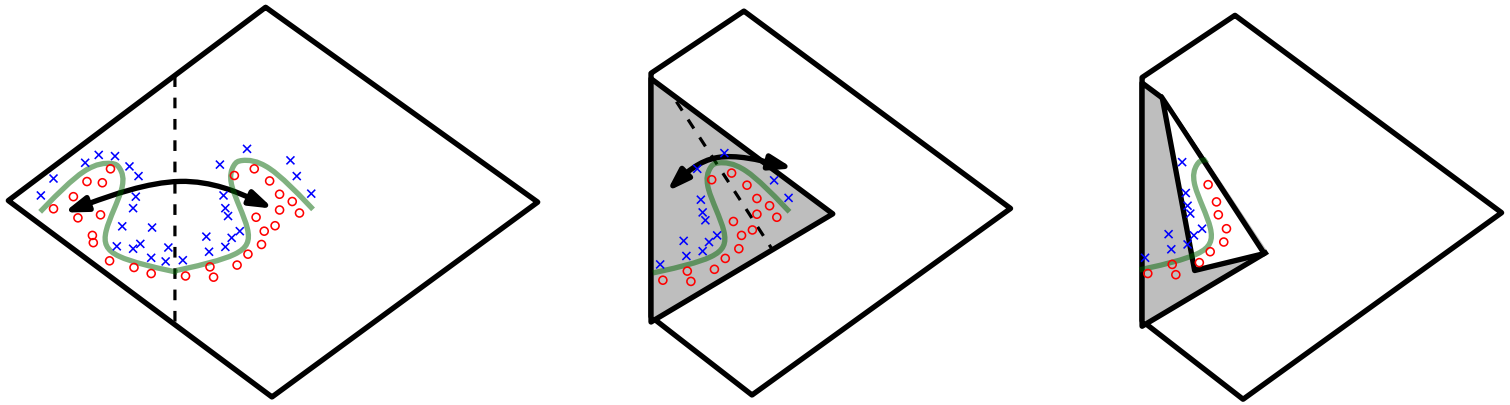
Theorems on advantage of depth:
(Hastad et al 86 & 91, Bengio et al 2007, Bengio & Delalleau 2011, Braverman 2011, Pascanu et al 2014, Montufar et al **NIPS 2014**)

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



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:

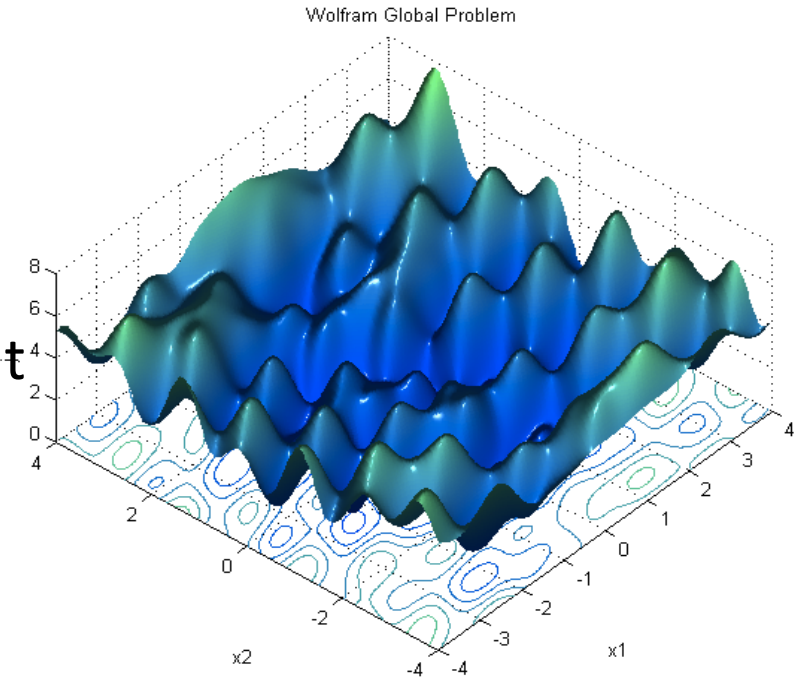


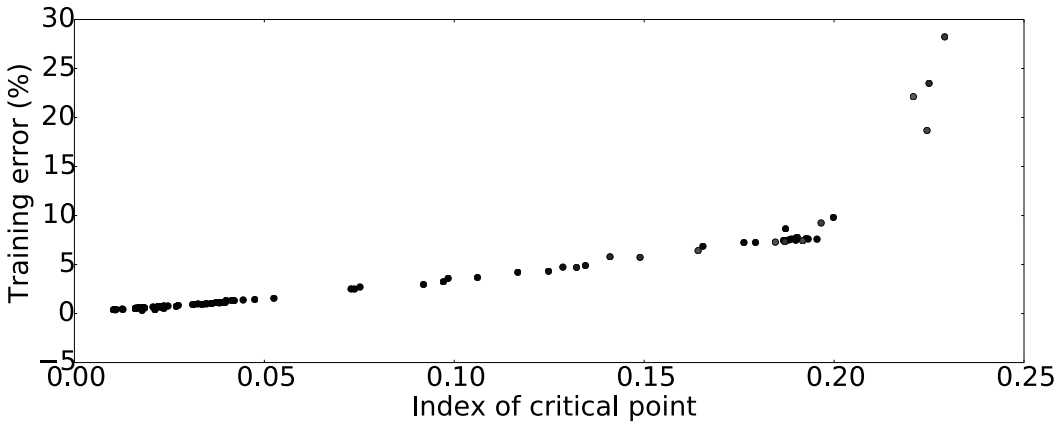
A Myth is Being Debunked: Local Minima in Neural Nets → Convexity is not needed

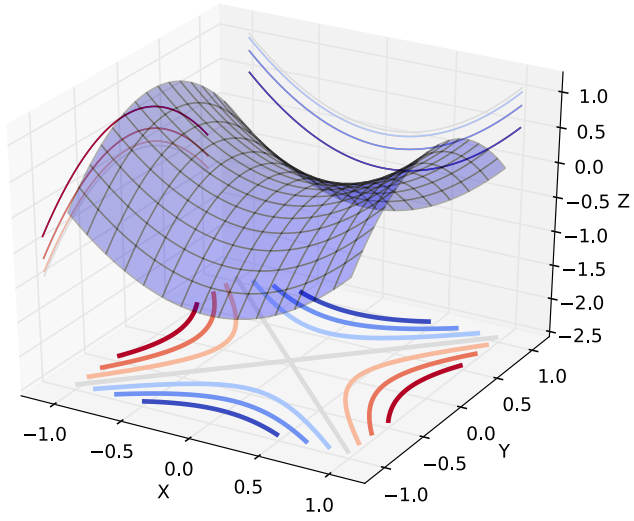
- (Pascanu, Dauphin, Ganguli, Bengio, arXiv May 2014): *On the saddle point problem for non-convex optimization*
- (Dauphin, Pascanu, Gulcehre, Cho, Ganguli, Bengio, NIPS' 2014): *Identifying and attacking the saddle point problem in high-dimensional non-convex optimization*
- (Choromanska, Henaff, Mathieu, 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)

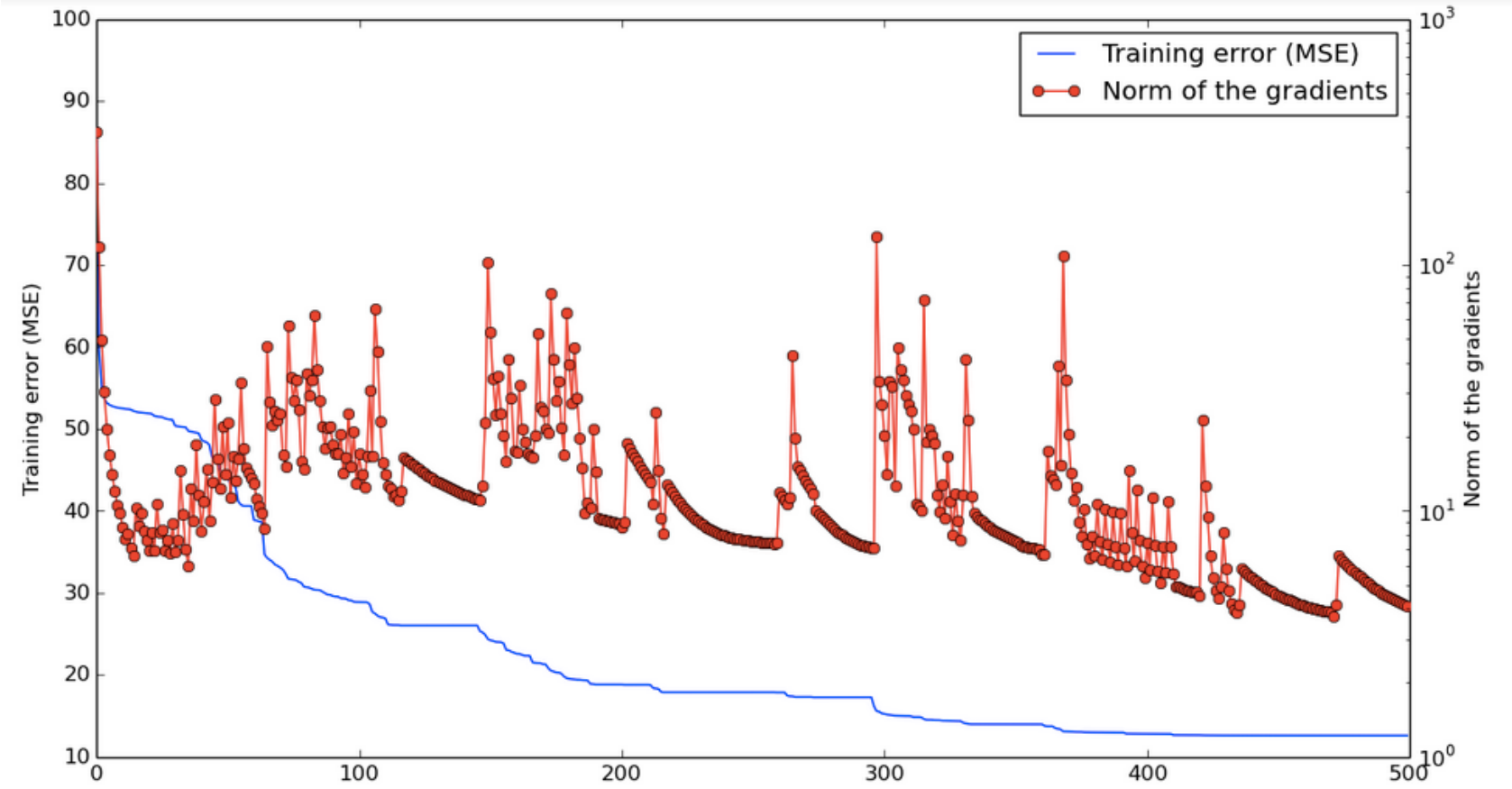






Saddle Points During Training

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

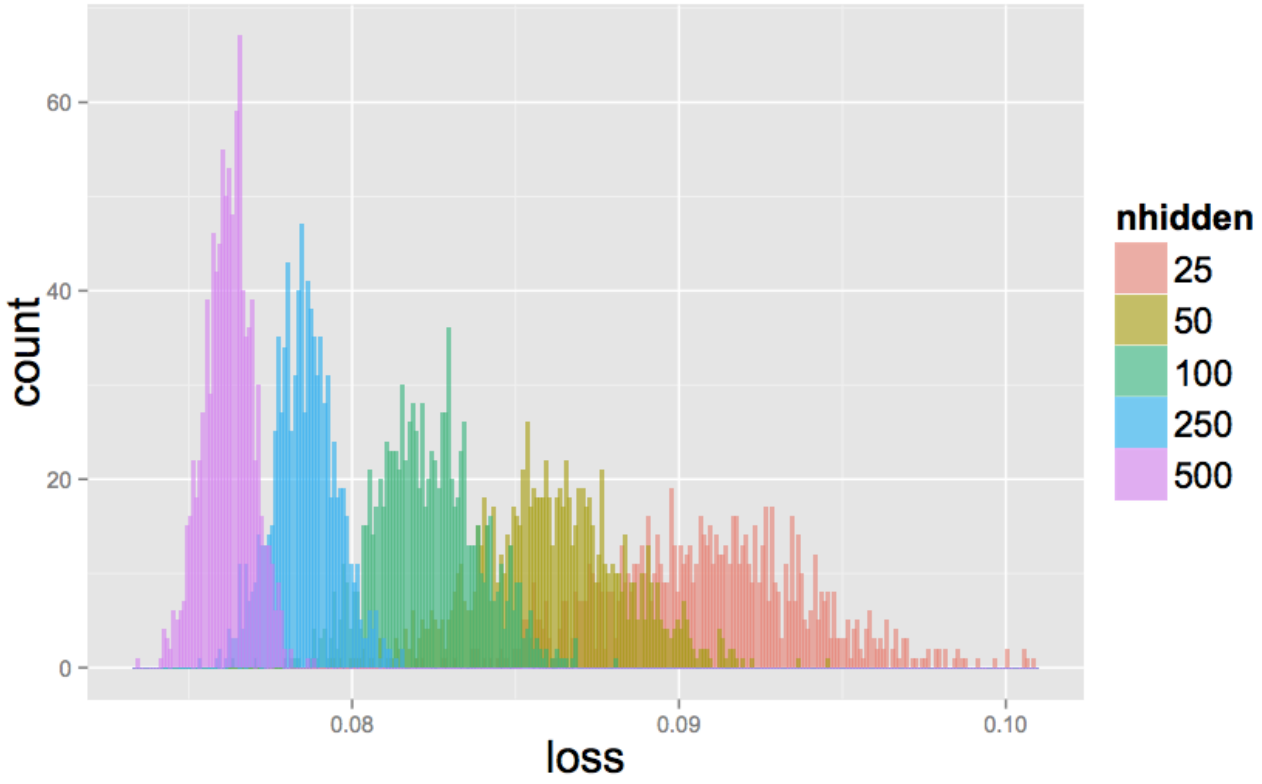


Low Index Critical Points

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)

  - Variational auto-encoder breakthrough

(Kingma & Welling ICLR 2014)

(Gregor et al arXiv 2015)

Denoising Auto-Encoder

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

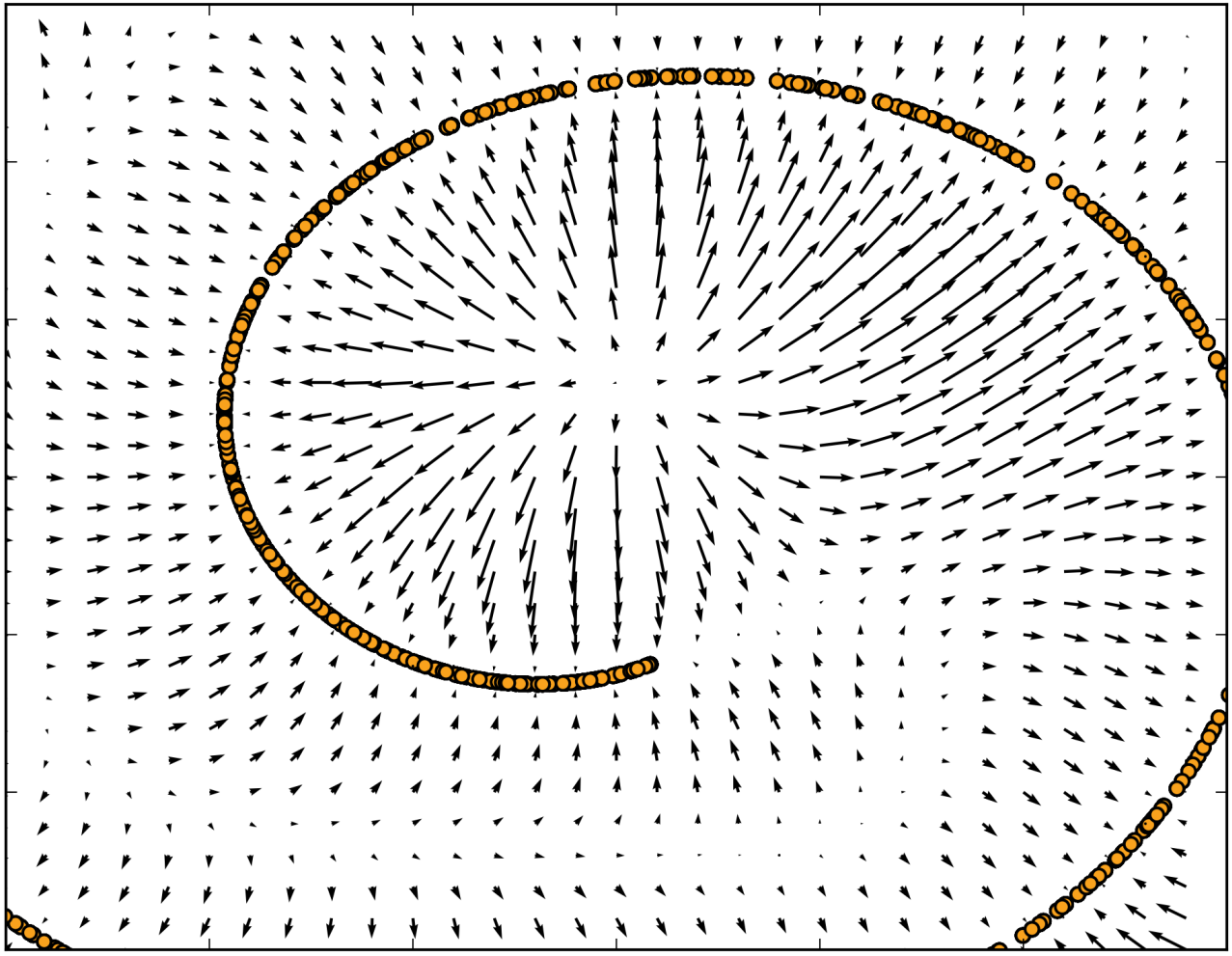
$$\text{reconstruction}(x) - x \quad \rightarrow \quad \sigma^2 \frac{\partial \log p(x)}{\partial x}$$

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

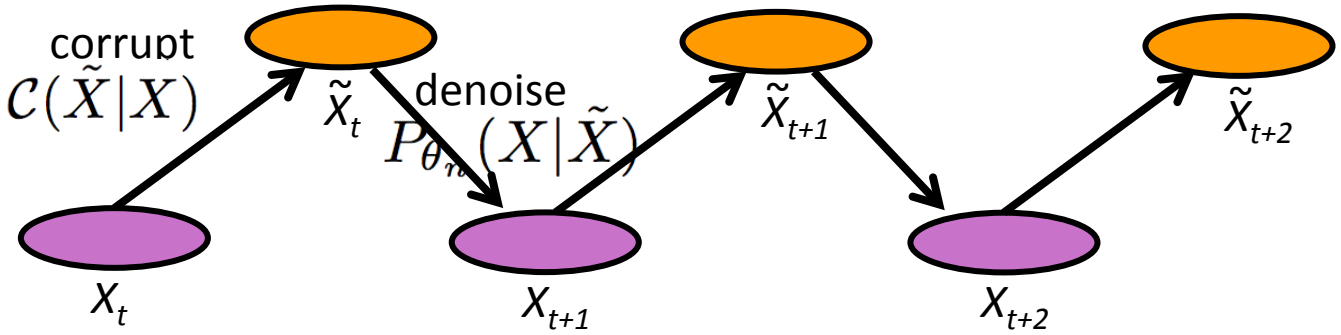
  [equivalent when noise→0]

prior: examples concentrate near a lower dimensional "manifold"

Regularized Auto-Encoders Learn a Vector Field that Estimates a Gradient Field (Alain & Bengio ICLR 2013)



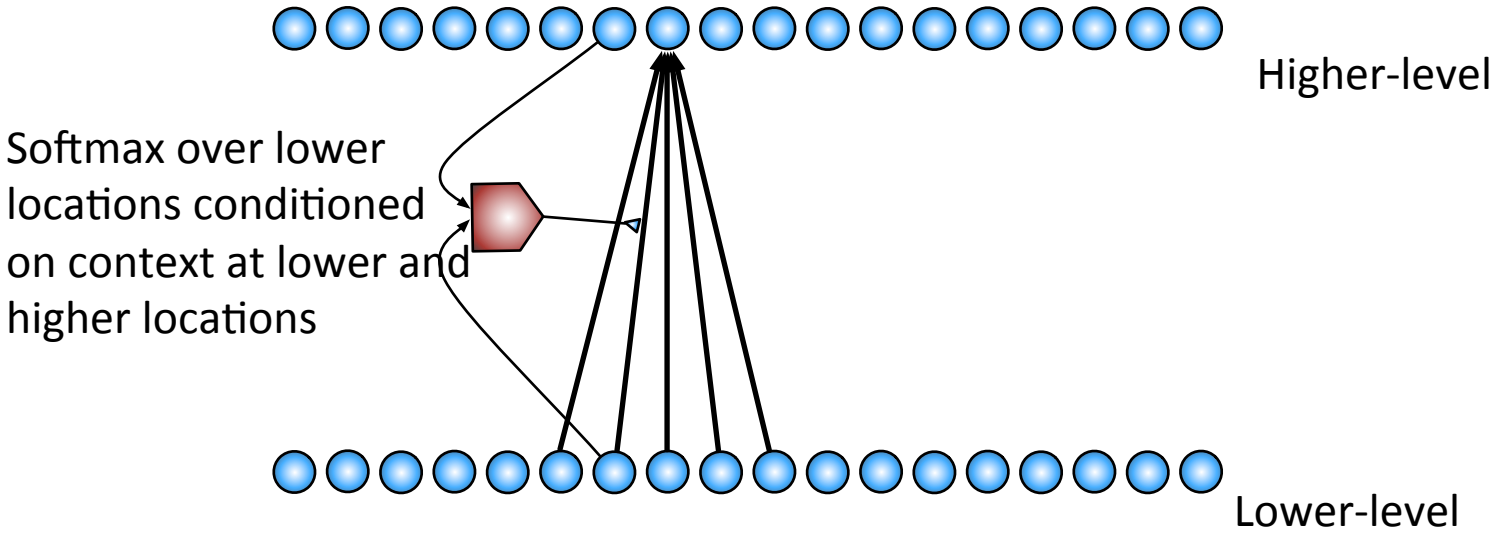
Denoising Auto-Encoder Markov Chain



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

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



Applying an attention mechanism to

- Translation
- Speech
- Images
- Video
- Memory

End-to-End Machine Translation

- Classical Machine Translation: several models separately trained by max. likelihood, brought together with logistic regression on top, based on n-grams
- Neural language models already shown to outperform n-gram models in terms of generalization power
- Why not train a neural translation model end-to-end to estimate P(target sentence | source sentence)?

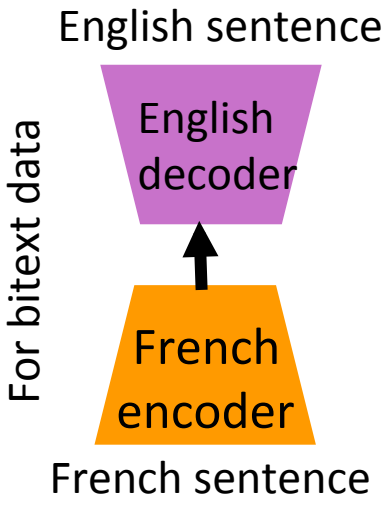
2014: The Year of Neural Machine Translation Breakthrough

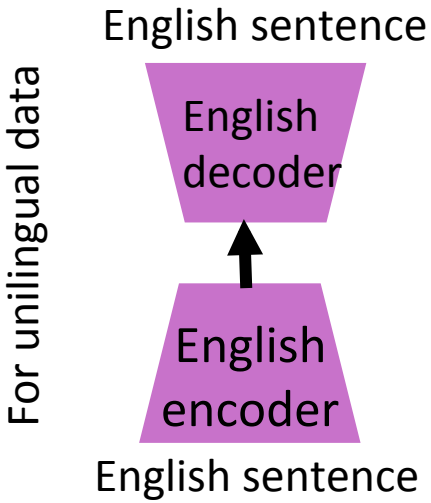
- (Devlin et al, ACL'2014)
- (Cho et al EMNLP'2014)
- (Bahdanau, Cho & Bengio, arXiv sept. 2014)
- (Jean, Cho, Memisevic & Bengio, arXiv dec. 2014)
- (Sutskever et al NIPS'2014)

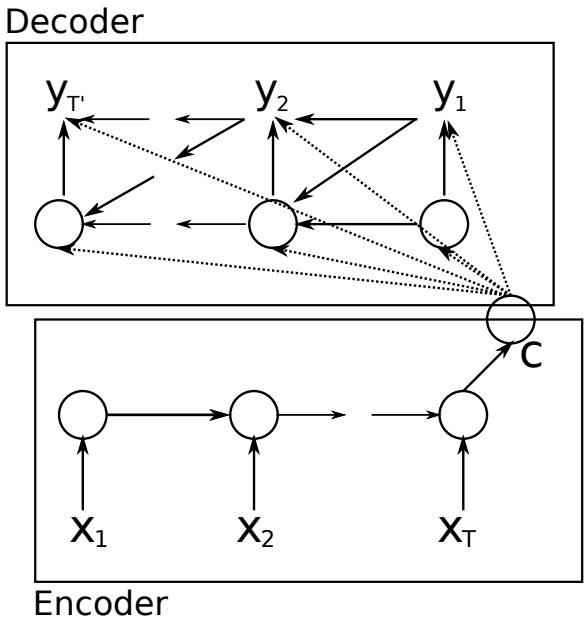
Earlier work: (Kalchbrenner & Blunsom et al 2013)

Encoder-Decoder Framework

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

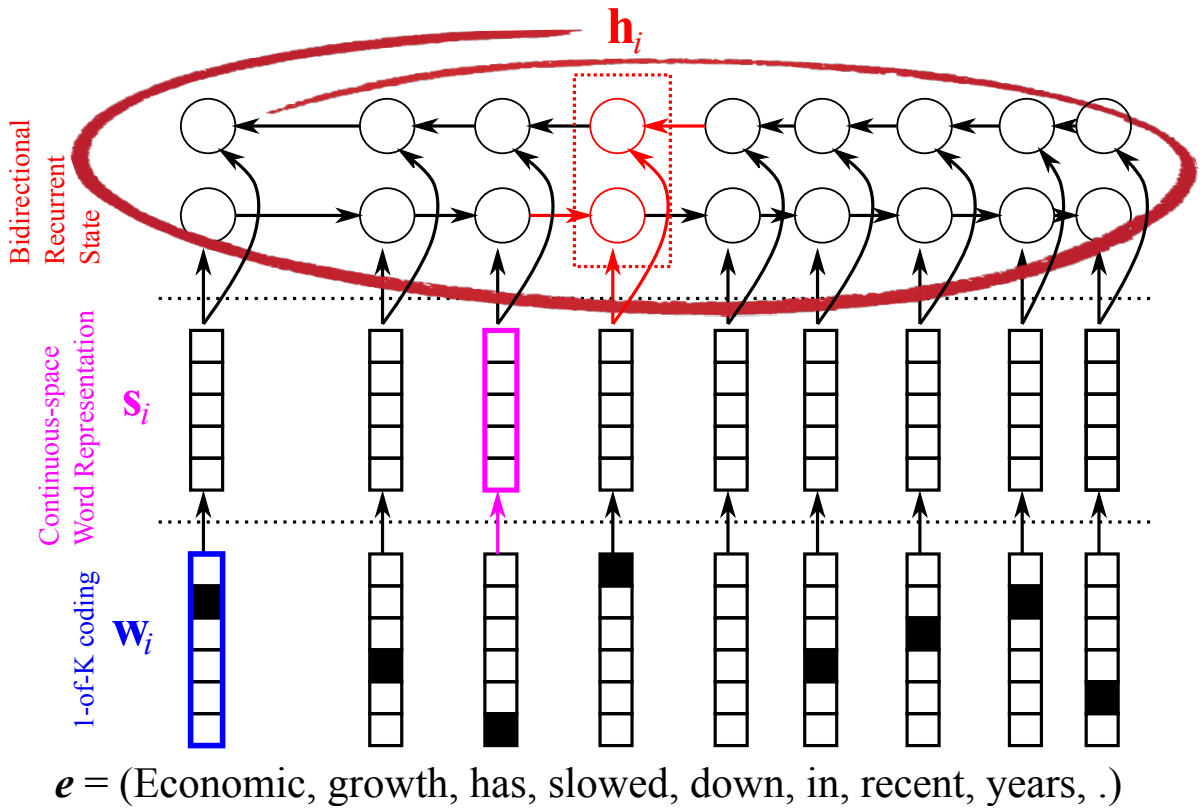






Bidirectional RNN for Input Side

- Following Alex Graves' work on handwriting



Attention: Many Recent Papers

- (Xu et al 2015, caption generation, U. Montreal + U. Toronto)
- (Ba et al 2014, Mnih et al 2014, visual attention, Google DeepMind)
- (Chorowski et al 2014, speech recognition, U. Montreal)
- (Bahdanau et al 2014, machine translation, U. Montreal)

And Older Papers

- (Larochelle & Hinton 2010, MNIST, U. Toronto)
- (Graves 2013, handwriting generation)
- (Denil et al 2014, visual tracking)
- (Tang et al 2014, generative models of images)

Soft-Attention vs Stochastic Hard-Attention

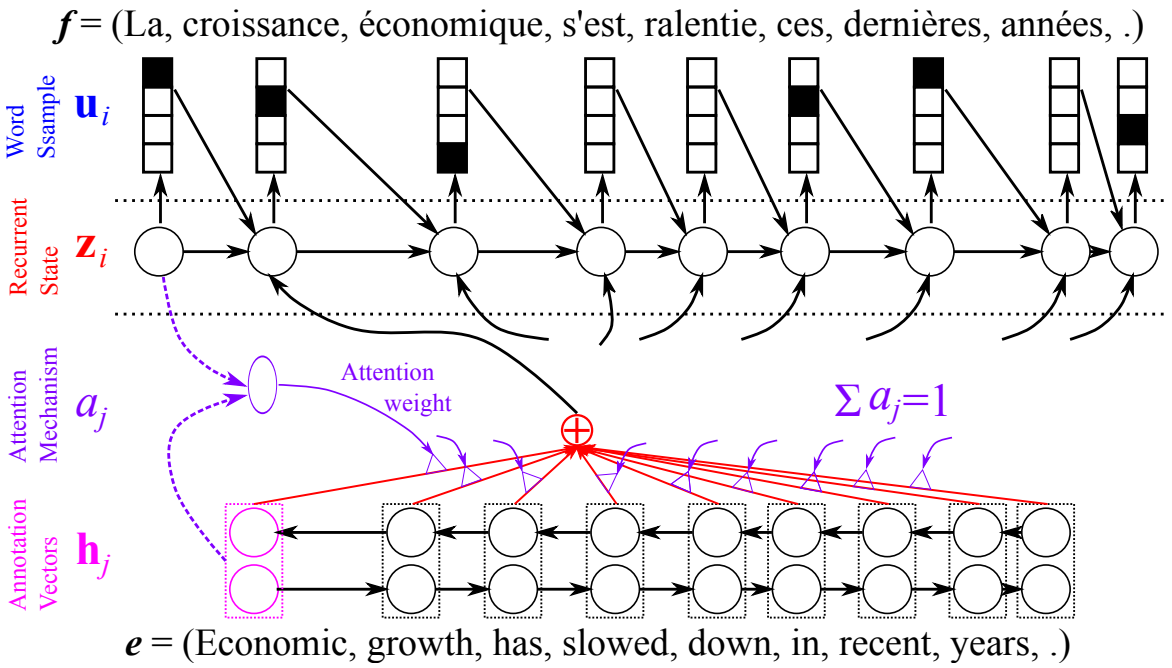
- With soft-attention: input fed to higher level at location **i** is a softmax-weighted sum of states at locations **j** at lower level
  - Train by back-prop
  - Fast training

- With stochastic hard-attention: sample an input location according to the softmax output
  - Get a gradient on the decisions via REINFORCE - baseline
  - Noisy gradient, slower training but works
  - *Symmetry breaking*

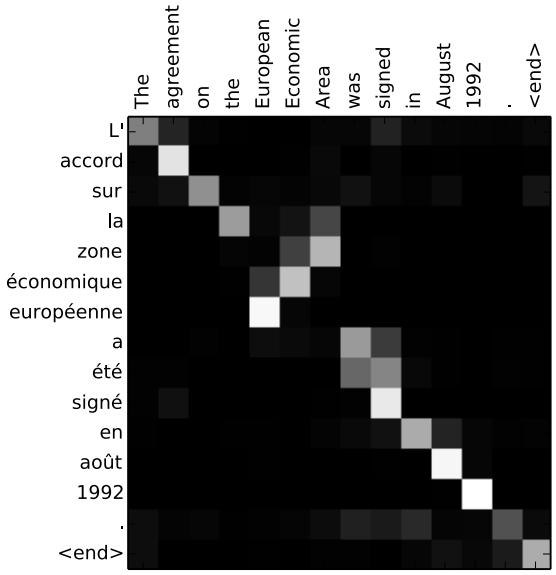
Attention-Based Neural Machine Translation

Related to earlier Graves 2013 for generating handwriting

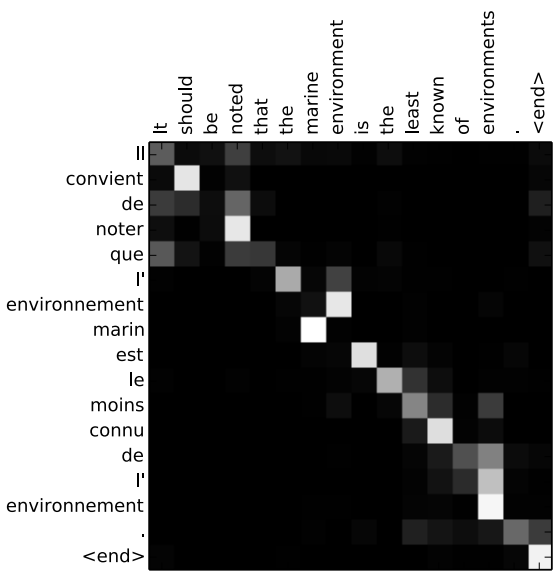
- (Bahdanau, Cho & Bengio, arXiv sept. 2014)
- (Jean, Cho, Memisevic & Bengio, arXiv dec. 2014)



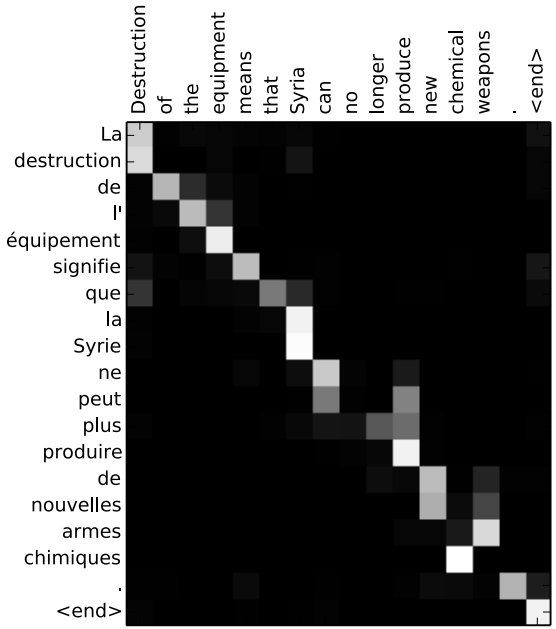
Predicted Alignments



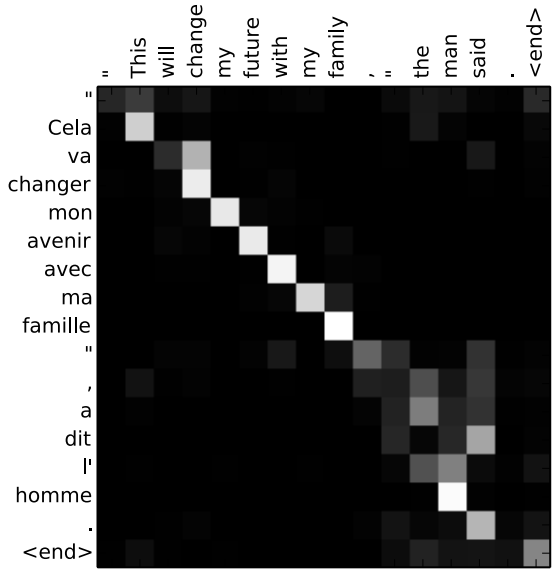
(a)



(b)

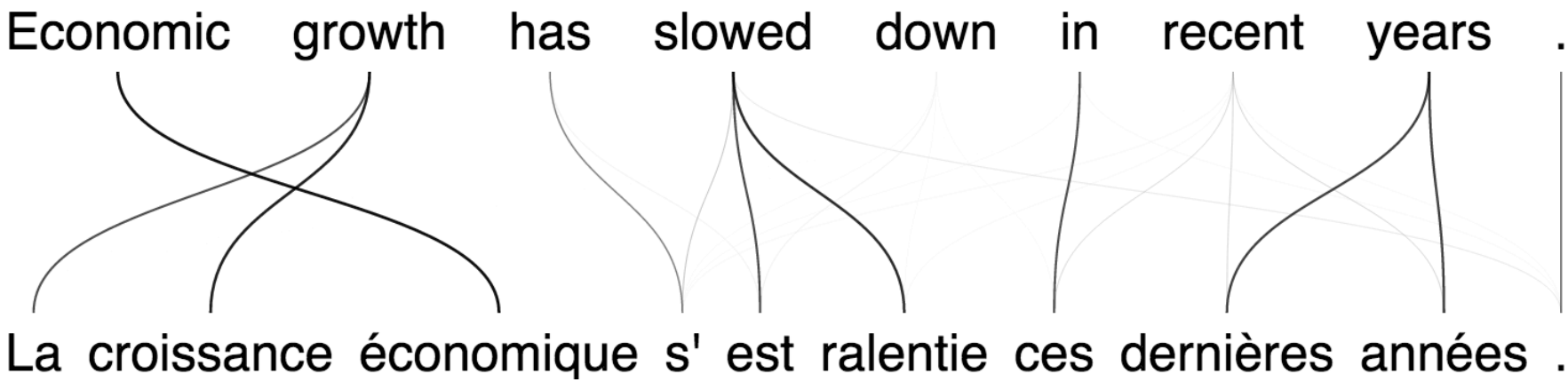


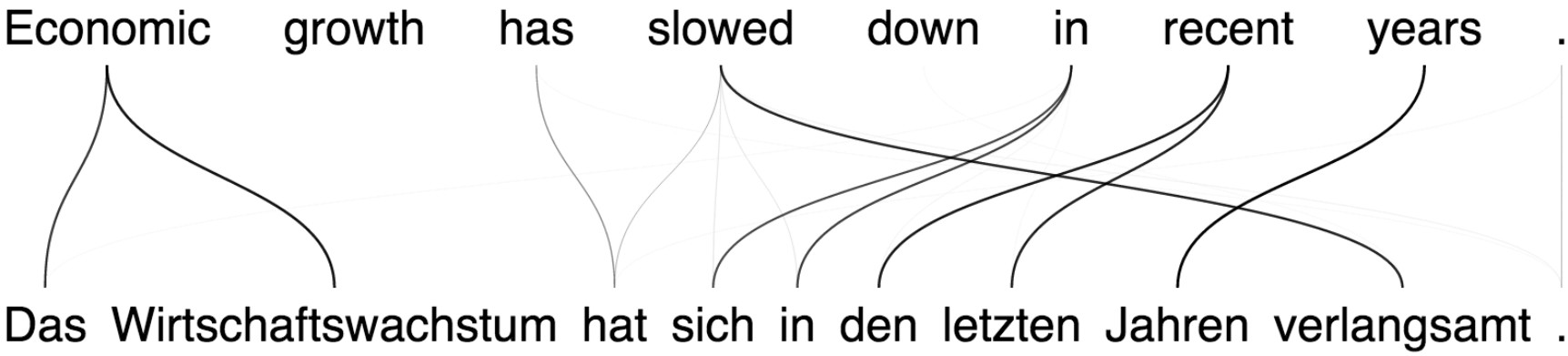
(c)



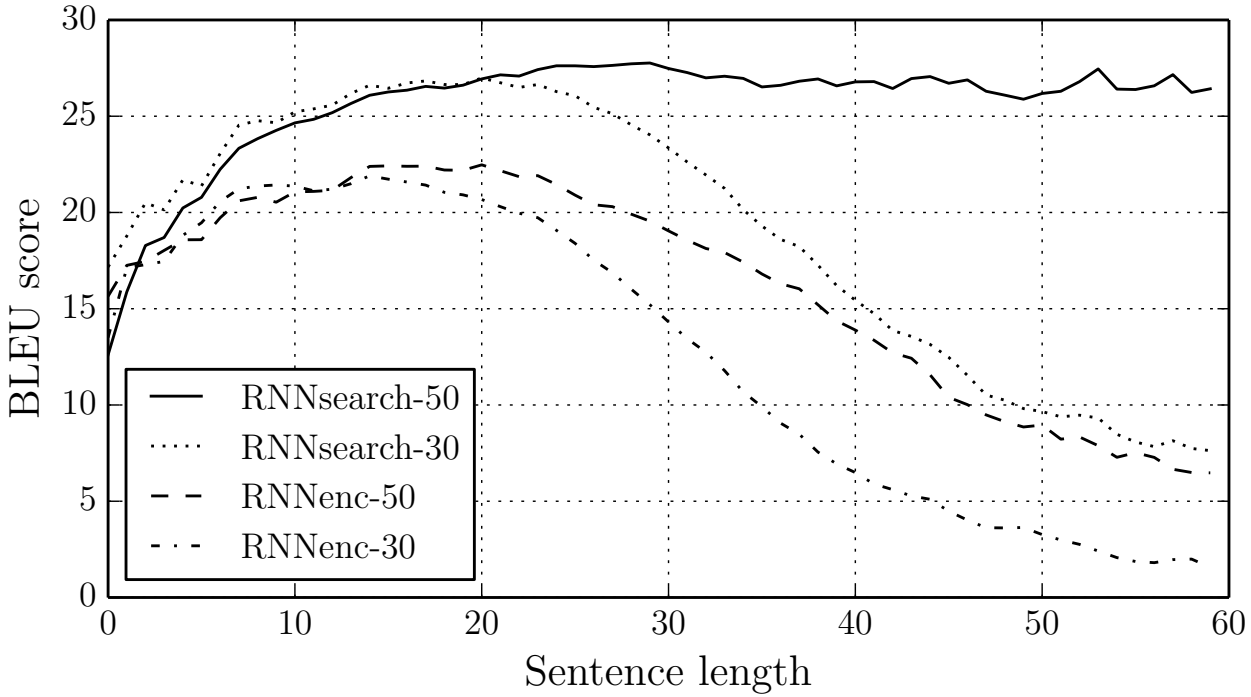
(d)

En-Fr & En-De Alignments





Improvements over Pure AE Model



- RNNenc: encode whole sentence
- RNNsearch: predict alignment
- BLEU score on full test set (including UNK)

And, the rest is history..

	NMT(A)	NMT(A)-LV	Google	P-SMT	
Basic NMT	29.48	32.68	30.6*	33.3*	**37.03•**
+Candidate List	–	33.28	–		
+UNK Replace	32.49	33.99	32.7°		
+Ensemble	–	**36.71**	**36.9°**		

(a) **English→French**

	NMT(A)	NMT(A)-LV	P-SMT
Basic NMT	16.02	16.95	**20.67◇**
+Candidate List	–	17.51	
+UNK Replace	18.27	18.87	
+Ensemble	–	**20.98**	

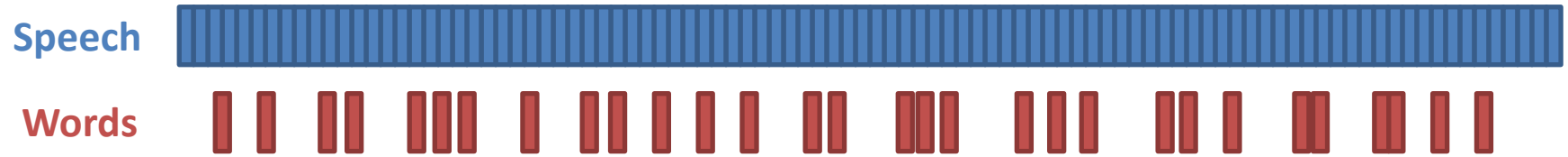
(b) **English→German**

NMT(A): **(Bahdanau et al., 2014)**, NMT(A)-LV: **(Jean et al., 2014)**,
(⋆): (Sutskever et al., 2014), (○): (Luong et al., 2014),
(•): (Durrani et al., 2014), (∗): (Cho et al., 2014), (◇): (Buck et al., 2014)

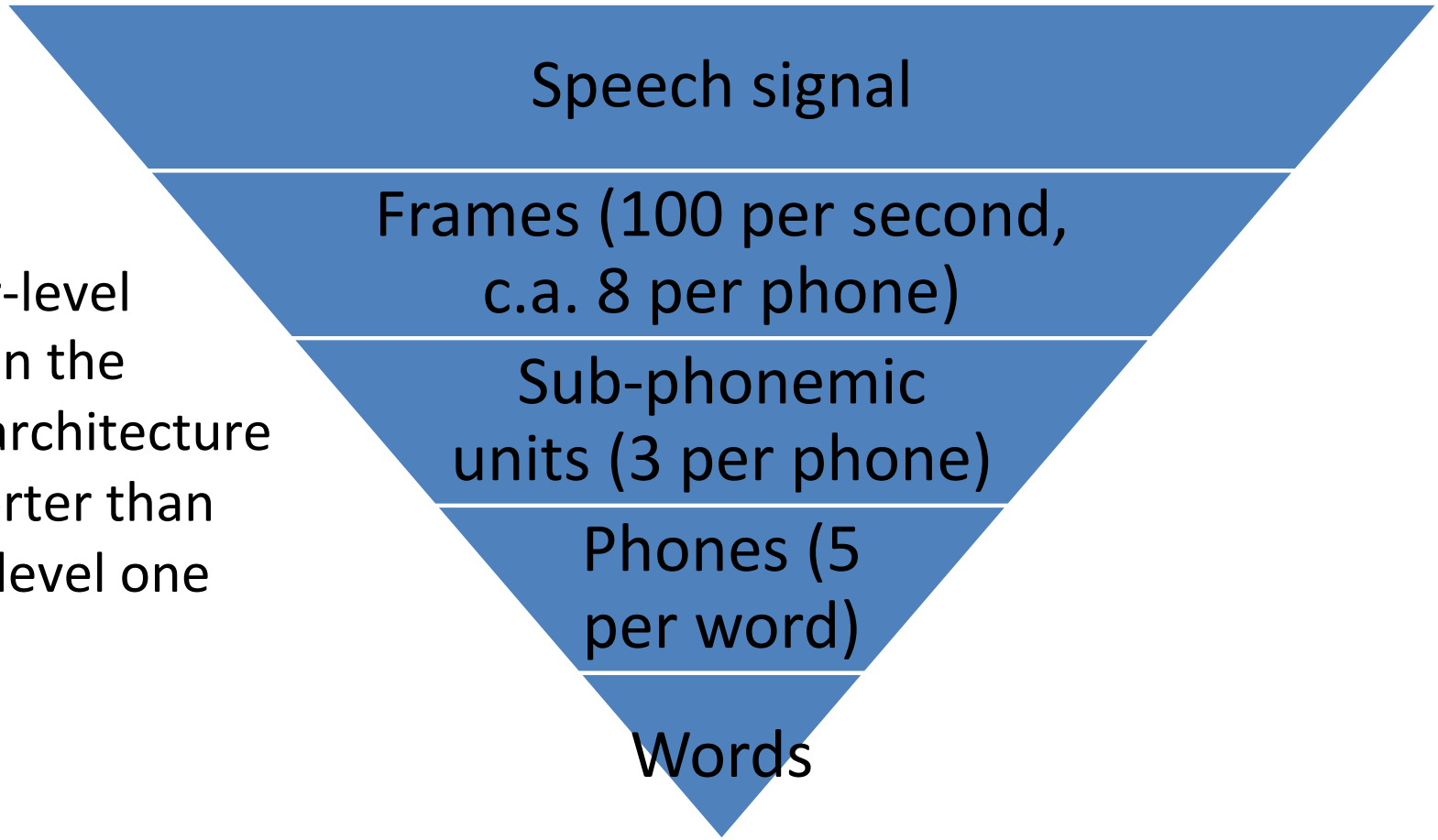
Translating from Other Sources?

- Speech
- Images
- Video

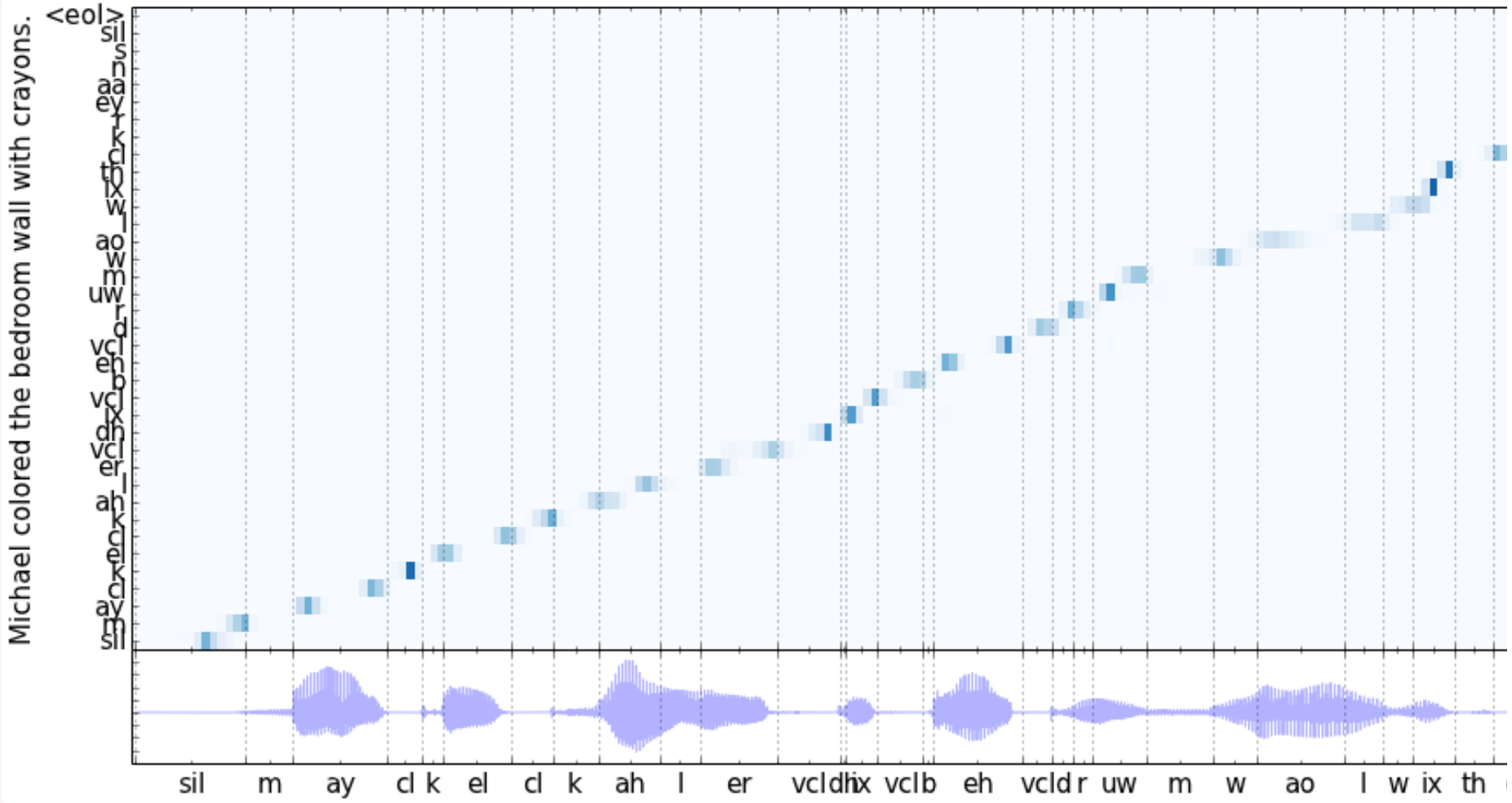
Multiple Time Scales in Speech



The higher-level sequence in the attention architecture can be shorter than the lower-level one



Acoustic-to-Phones Attention Alignment



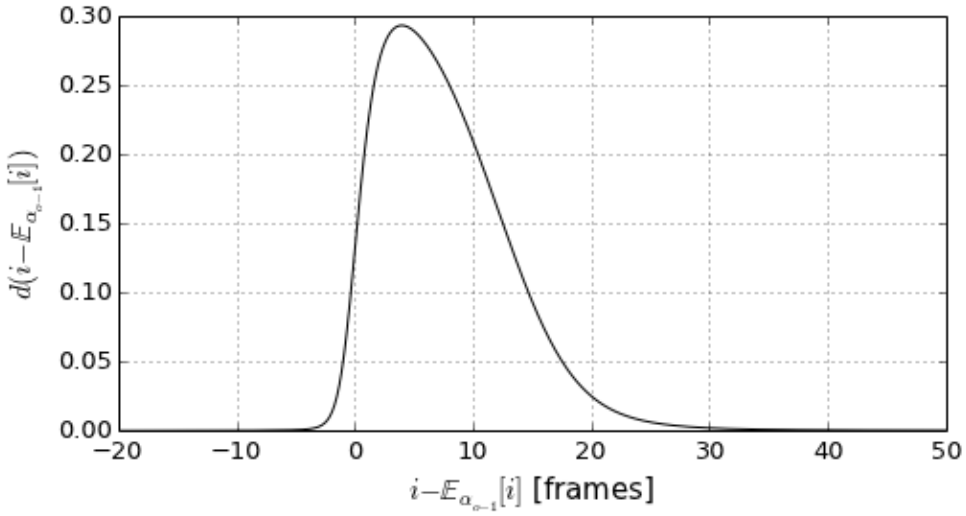
3 Inputs to Attention Mechanism

- Higher-level RNN state at current output location
- Lower-level RNN state at all input locations
- Previous attention pattern (for previous output location)

- All three were necessary to apply attention-based models to speech recognition (Chorowski et al, 2014, 2015)

Left-to-Right Soft Constraint

- Whereas with translation the word order can change a lot, the acoustic→phonetic mapping is mostly left-to-right.
- The strength of that prior can be learned by structuring the attention location probability distribution:



$$c_o = \sum_{t=1}^{T} \alpha_{ot} f_t$$

$$\alpha_{ot} \propto d\left(t - \mathbb{E}_{\alpha_{o-1}}[t]\right) \exp\left(m(s_{o-1}, h_t)\right)$$

Prior: relative to last selection

Match between previous state and each input

End-to-end Continuous Speech Recognition using Attention-based Recurrent NN: First Results

(Chorowski, Bahdanau, Cho & Bengio, arXiv Dec. 2014)

(Chorowski et al, NIPS submission, 2015)

Model	Dev	Test
Baseline Model	15.9%	18.7%
Baseline + Conv. Features	16.1%	18.0%
Baseline + Conv. Features + Smooth Focus	15.8%	**17.6%**
RNN Transducer [17]	N/A	17.7%
HMM over Time and Frequency Convolutional Net [26]	13.9%	16.7%

TIMIT Phoneme Error Rate

FDHC0_SX209: Michael colored the bedroom wall with crayons.

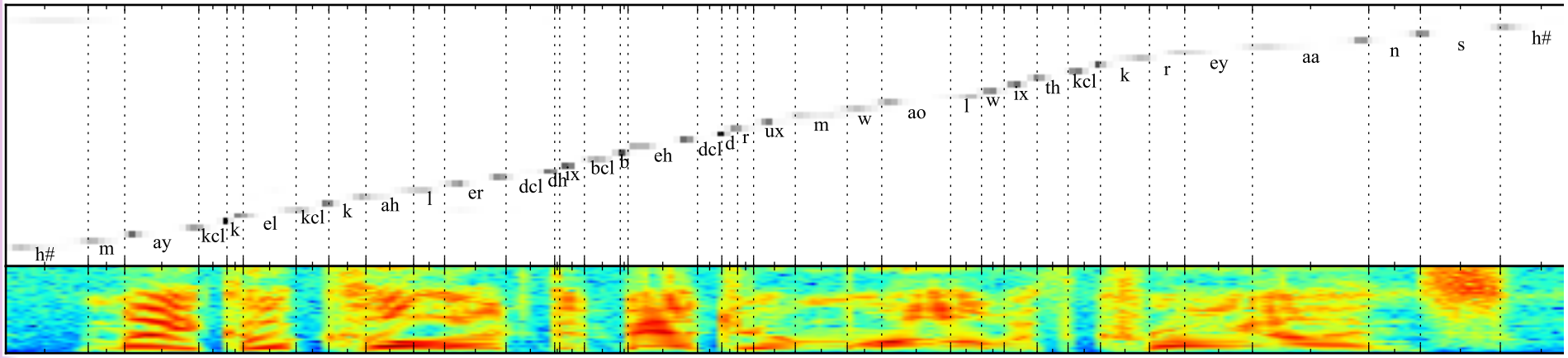
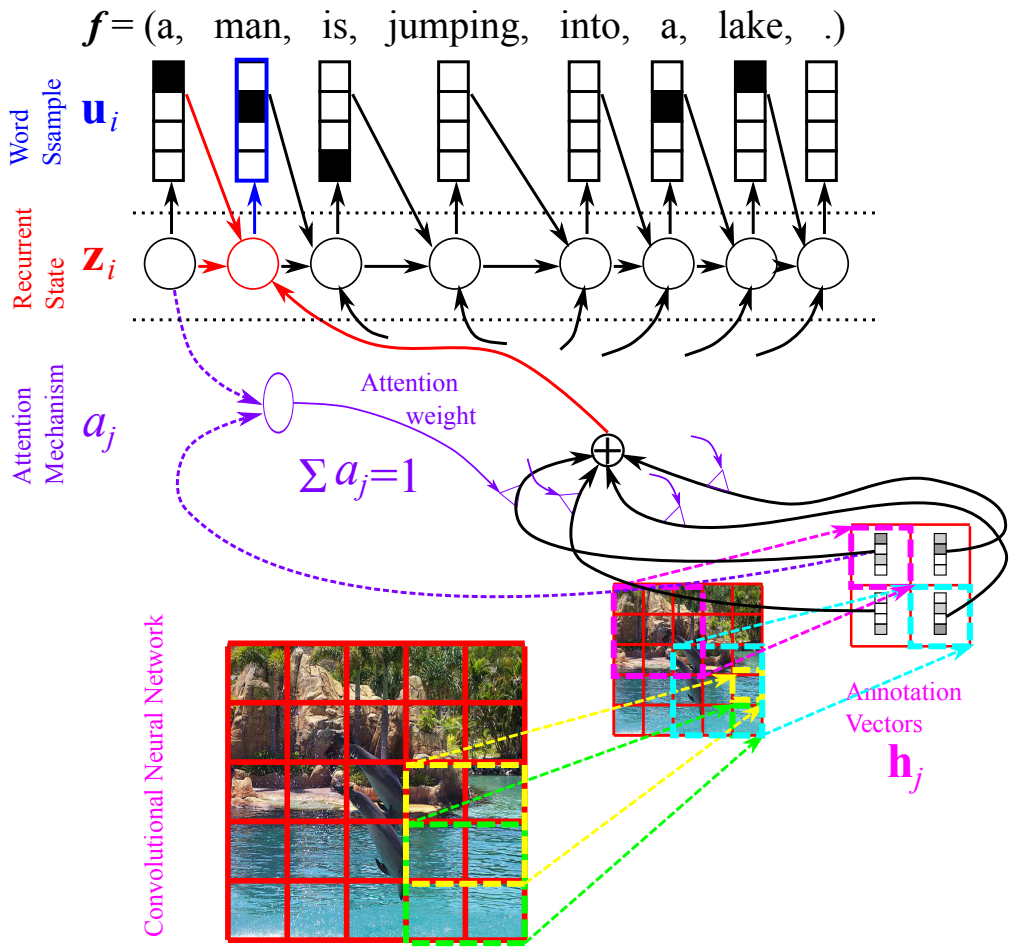
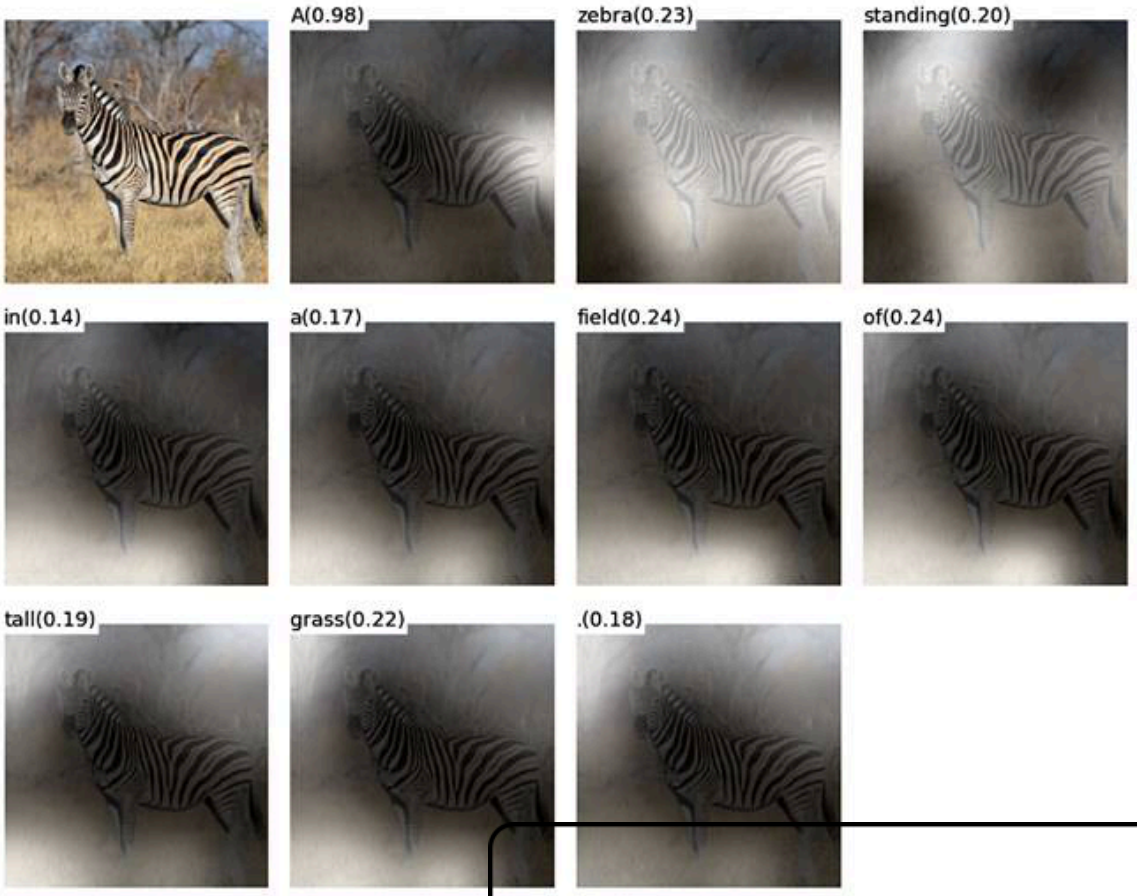


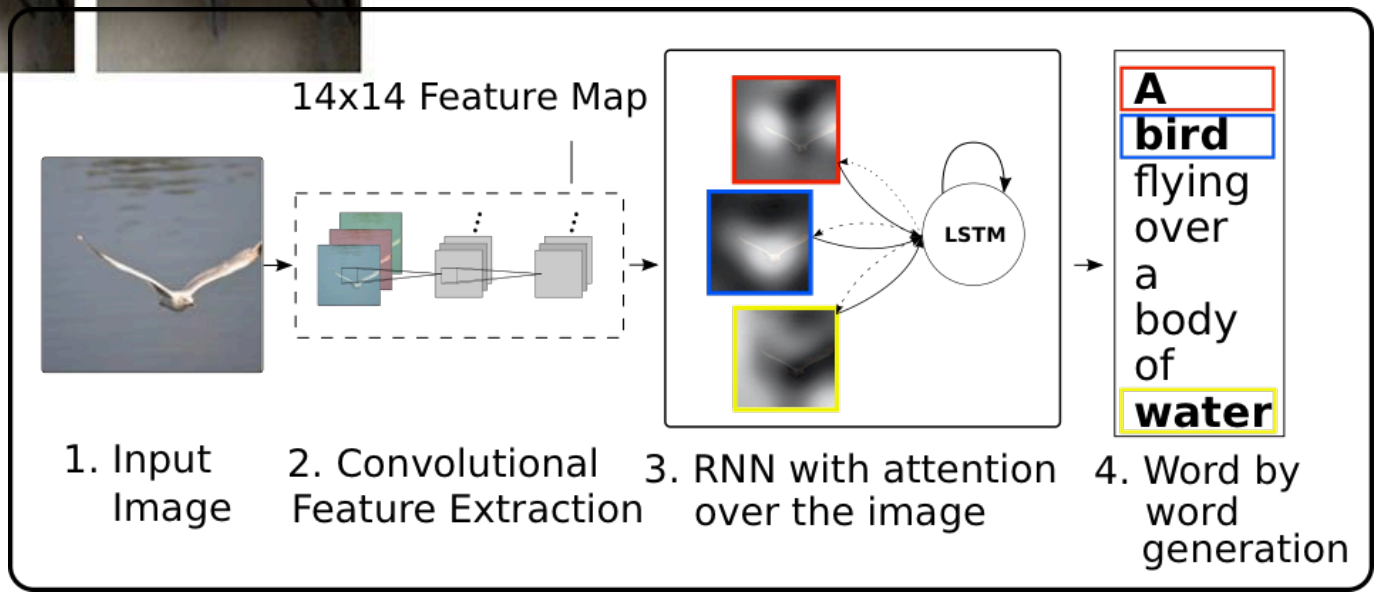
Image-to-Text: Caption Generation



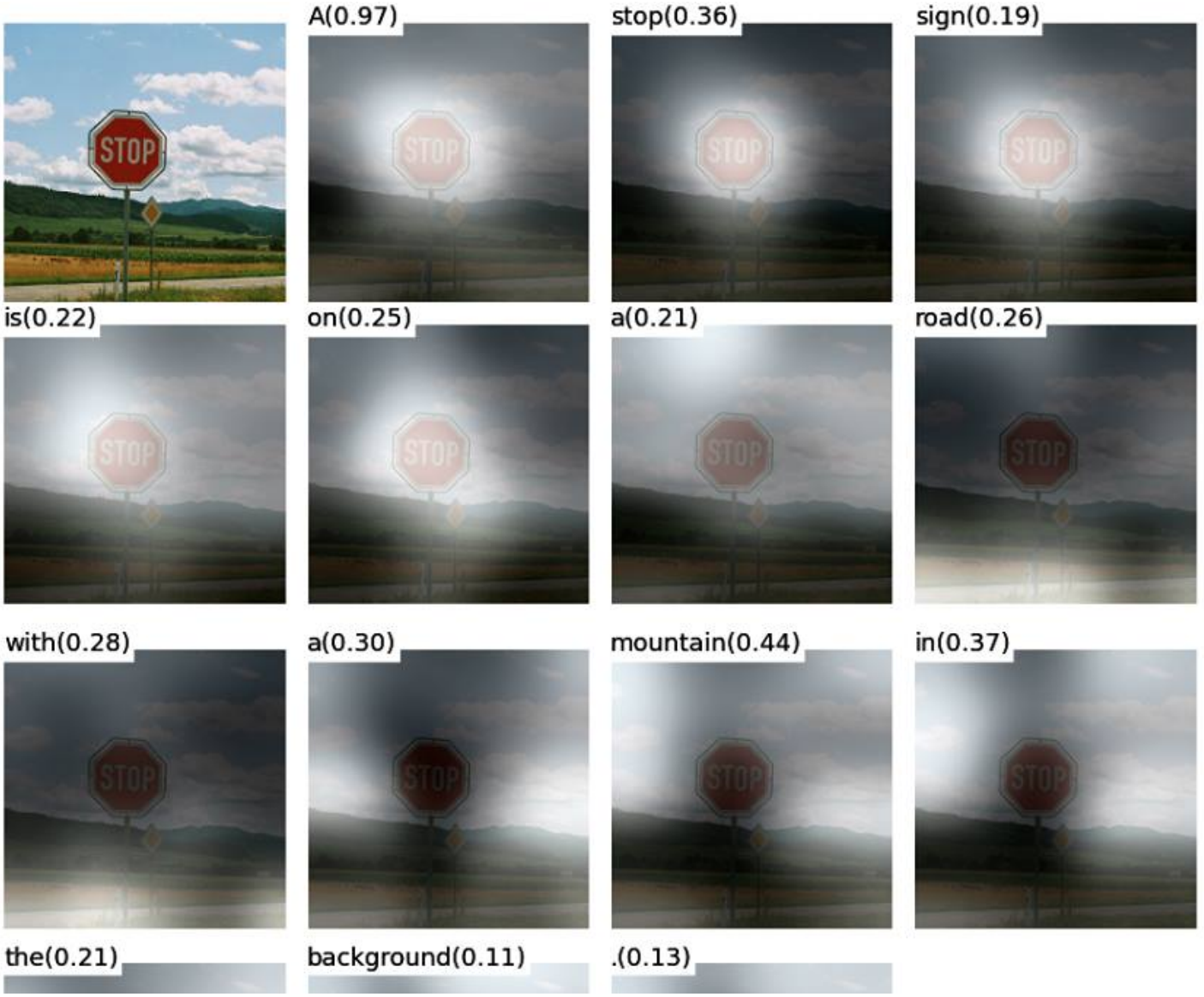
(Xu et al., 2015), (Yao et al., 2015)



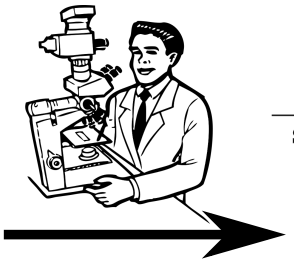
Paying Attention to Selected Parts of the Image While Uttering Words



Speaking about what one sees



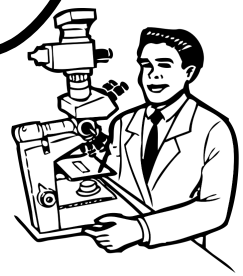
Let's go back 2.5 years back in time.. (Mitchell et al., 2012)



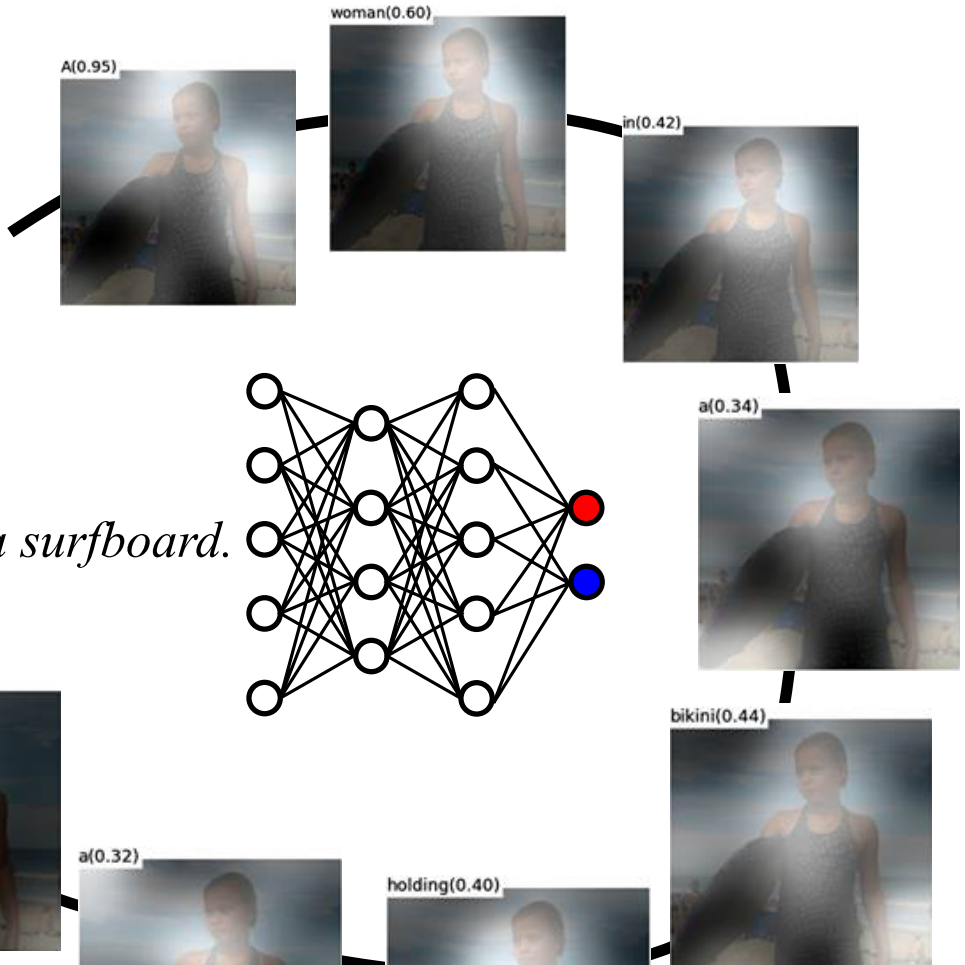
stuff:	*sky*	.999	
	id:	1	
	atts:	clear:0.432,	blue:0.945
		grey:0.853,	white:0.501 ...
	b. box:	(1,1 440,141)	
stuff:	*road*	.908	
	id:	2	
	atts:	wooden:0.722	clear:0.020 ...
	b. box:	(1,236 188,94)	
object:	*bus*	.307	
	id:	3	
	atts:	black:0.872,	red:0.244 ...
	b. box:	(38,38 366,293)	
preps:	id 1, id 2: by	id 1, id 3: by	id 2, id 3: below

The bus by the road with a clear blue sky

1. Group the Nouns
2. Order the Nouns
3. Filter Incorrect Attributes
4. Group Plurals
5. Gather Local Sub-(parse) trees
6. Create Full Trees
7. Get Final Tree, Clear Mark-Up
8. Prenominal Modifier Ordering



And in 2015... End-to-End Neural Net



A woman in a bikini holding a surfboard.

The neural nets successfully learned to

- map a phrase in one language to that in another language
- extract semantics [and syntax] of a sentence
- separate different objects in an image
- separate the background from foreground objects
- create a syntactically and semantically correct sentence

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.

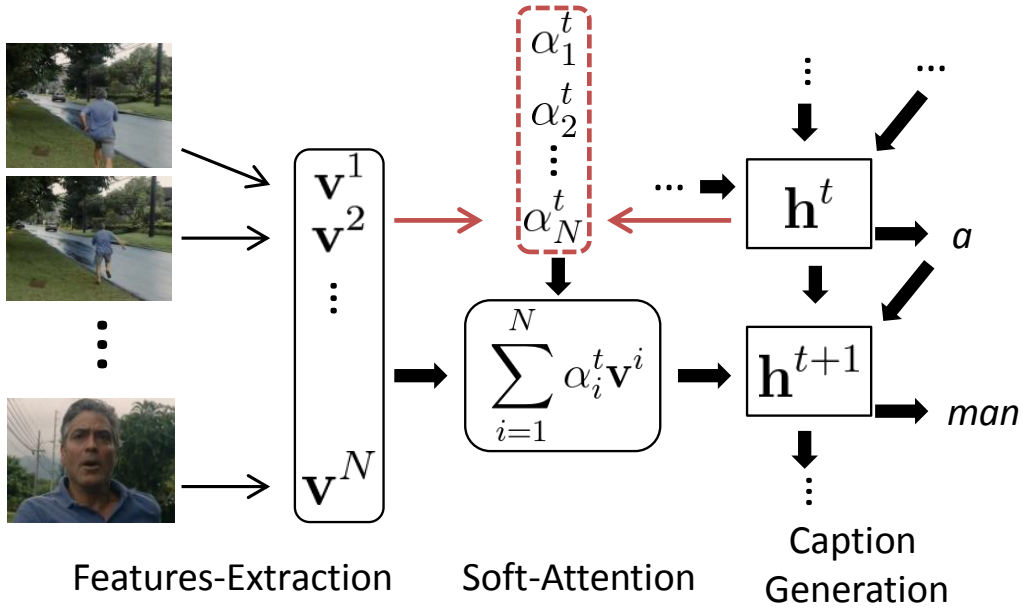
A person is standing on a beach
with a surfboard.

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

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

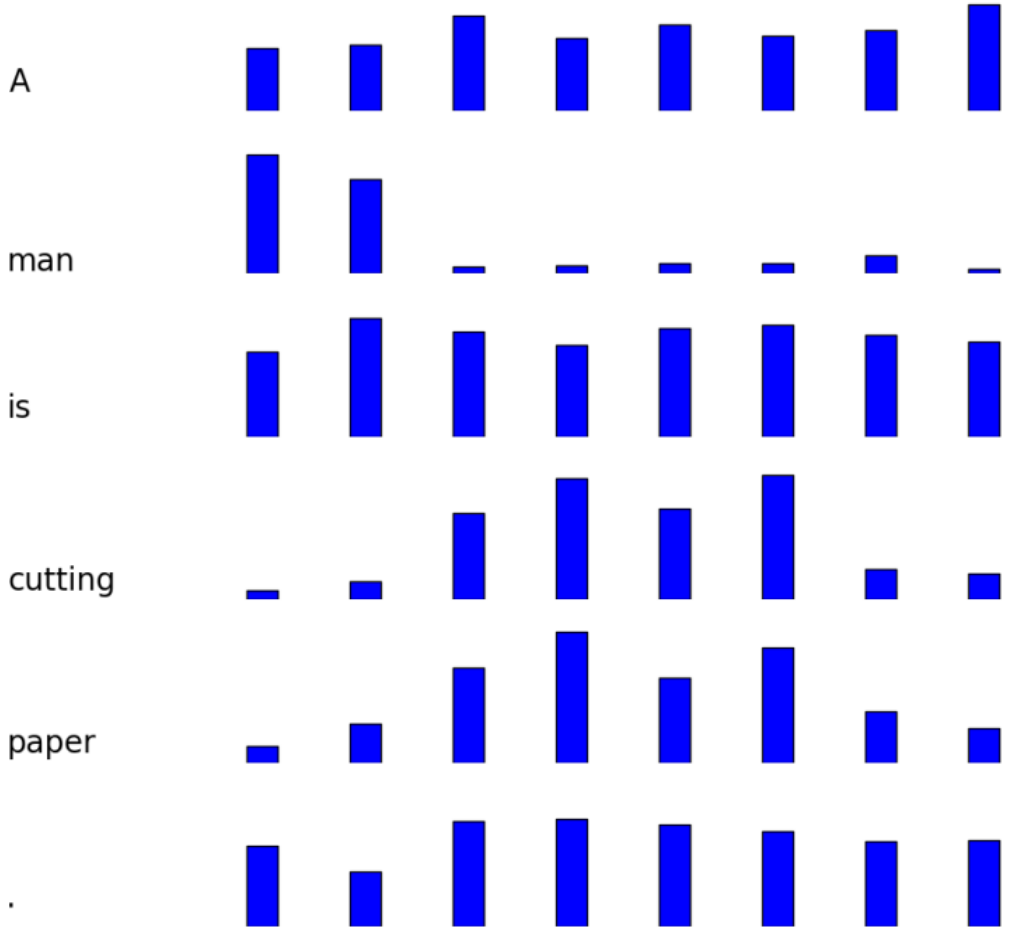
Attention through time for video caption generation

- (Yao et al arXiv 1502.08029, 2015) *Video Description Generation Incorporating Spatio-Temporal Features and a Soft-Attention Mechanism*
- Attention can be focused temporally, i.e., selecting input frames



Attention through time for video caption generation (Yao et al 2015)

- Attention is focused at appropriate frames depending on which word is generated.



Attention through time for video caption generation (Yao et al 2015)

- Soft-attention worked best in this setting

Model	Feature	Bleu					Meteor	Perplexity
		1	2	3	4	mb		
non-attention	GNet	32.0	9.2	3.4	1.2	0.3	4.43	88.28
	$\text{GNet+3DConv}_{\text{non-att}}$	33.6	10.4	4.3	1.8	0.7	5.73	84.41
soft-attention	GNet	31.0	7.7	3.0	1.2	0.3	4.05	66.63
	$\text{GNet+3DConv}_{\text{att}}$	28.2	8.2	3.1	1.3	0.7	5.6	**65.44**

Corpus:
She rushes out.
Test_sample:
The woman turns away.

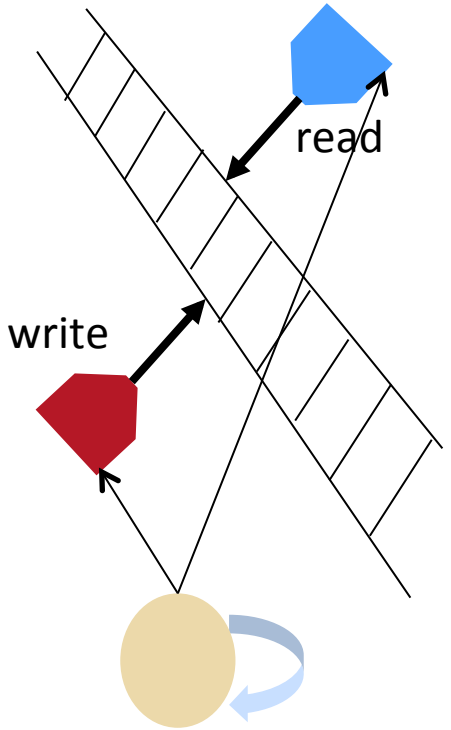
Corpus:
SOMEONE sits with his arm around SOMEONE.
He nuzzles her cheek, then kisses tenderly.
Test_sample:
SOMEONE sits beside SOMEONE.

Corpus:
SOMEONE shuts the door.
Test_sample:
as he turns on his way to the door , SOMEONE turns away.

Generated captions

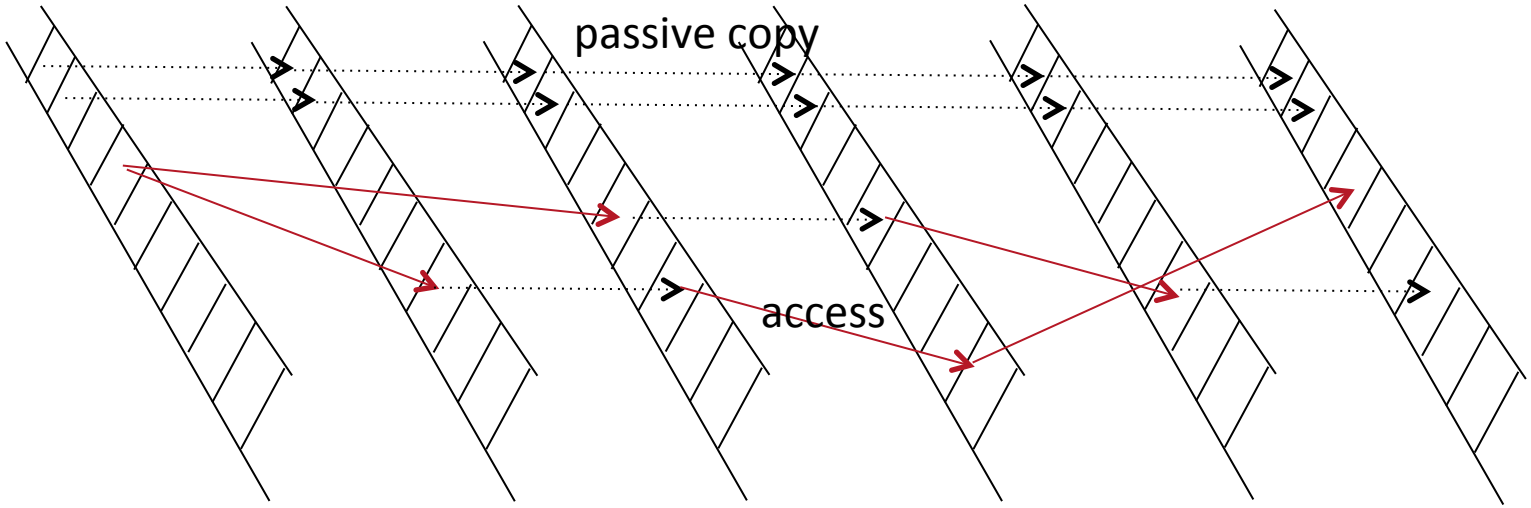
Attention Mechanisms for Memory Access

- Neural Turing Machines (Graves et al 2014)
- and Memory Networks (Weston et al 2014)
- Use a form of attention mechanism to control the read and write access into a memory
- The attention mechanism outputs a softmax over memory locations
- For efficiency, the softmax should be sparse (mostly 0's), e.g. maybe using a hash-table formulation.



Sparse Access Memory for Long-Term Dependencies

- Whereas LSTM memories always decay exponentially (even if slowly), a mental state stored in an external memory can stay for arbitrarily long durations, until evoked for read or write.
- Need to replace the soft gater or softmax attention by hard one that is 0 most of the time, and yet for which training works (again, may use noisy decisions and/or REINFORCE).
- Different « threads » can run in parallel if we view the memory as an associative one.



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 unsupervised learning.
- Attention mechanisms allow the learner to make a selection, soft or hard
- They have been extremely successful for machine translation and caption generation
- They could be interesting for speech recognition and video, especially if we used them to capture multiple time scales
- They could be used to help deal with long-term dependencies, allowing some states to last for arbitrarily long

MILA: Montreal Institute for Learning Algorithms