Deep Learning with Attention Mechanisms

Yoshua Bengio

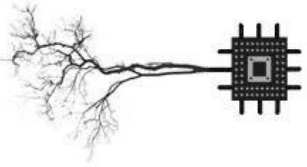
July 20th, 2015

Keynote speech at Russian Deep Learning Hackaton

Université de Montréal

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

LISA



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)?

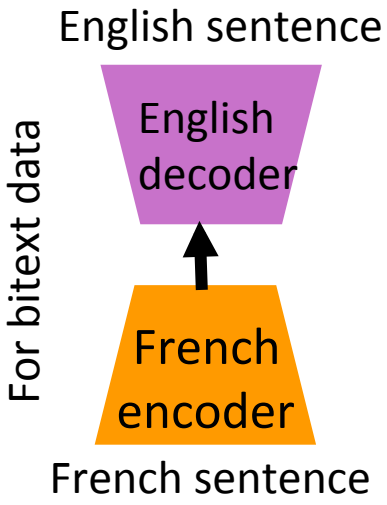
2014: The Year of Neural Machine Translation Breakthrough

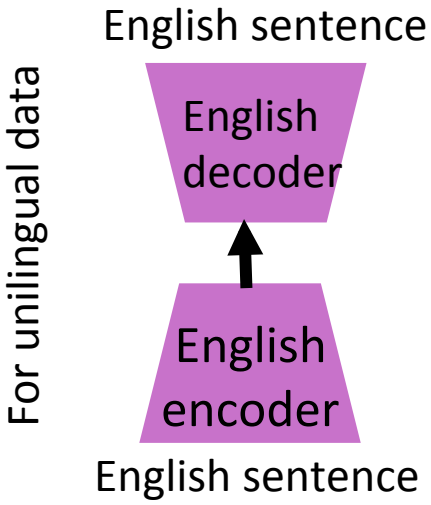
- (Devlin et al, ACL'2014)
- (Cho et al EMNLP'2014)
- (Bahdanau, Cho & Bengio, arXiv sept. 2014)
- (Jean, Cho, Memisevic & Bengio, arXiv dec. 2014)
- (Sutskever et al NIPS'2014)

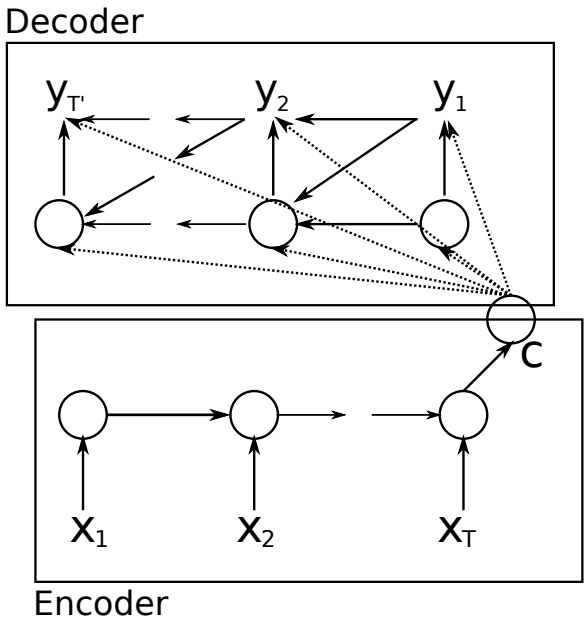
Earlier work: (Kalchbrenner & Blunsom et al 2013)

Encoder-Decoder Framework

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

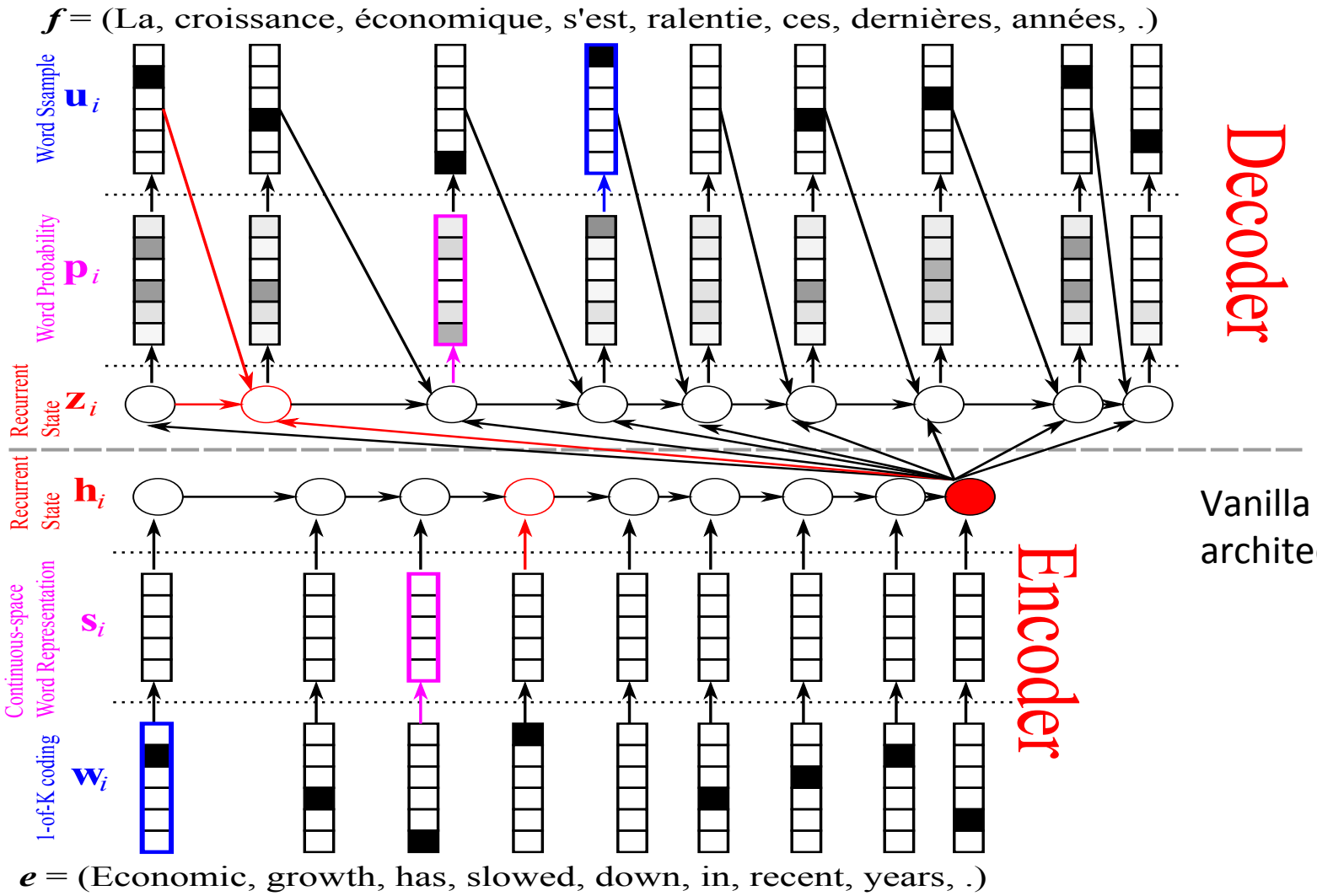






Encoder & Decoder RNN

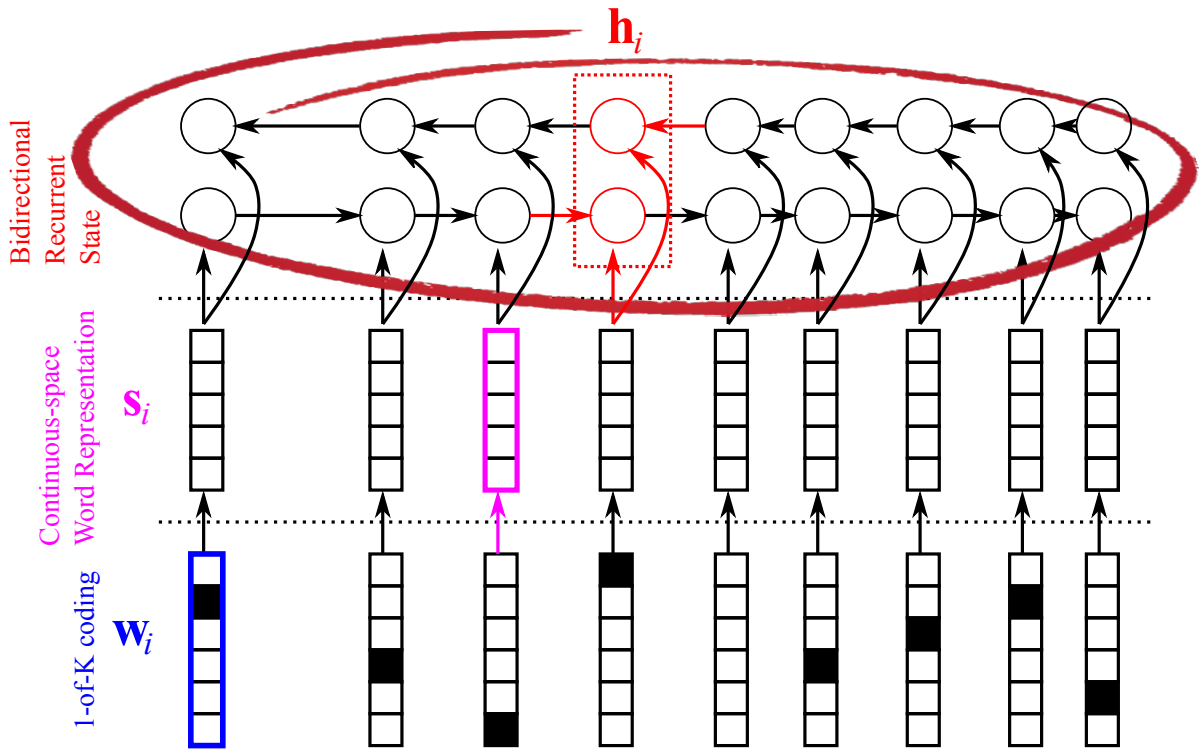
- Need to use gated RNN such as LSTM or GRU



Vanilla architecture

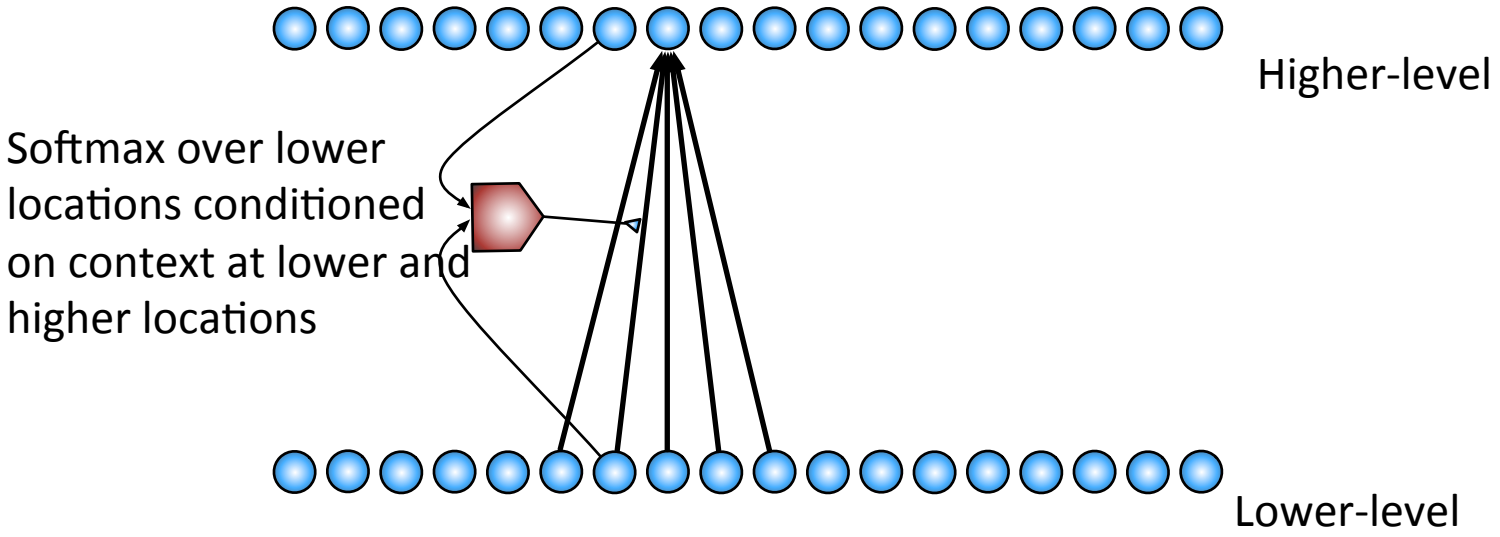
Bidirectional RNN for Input Side

- Following Alex Graves' work on handwriting



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



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

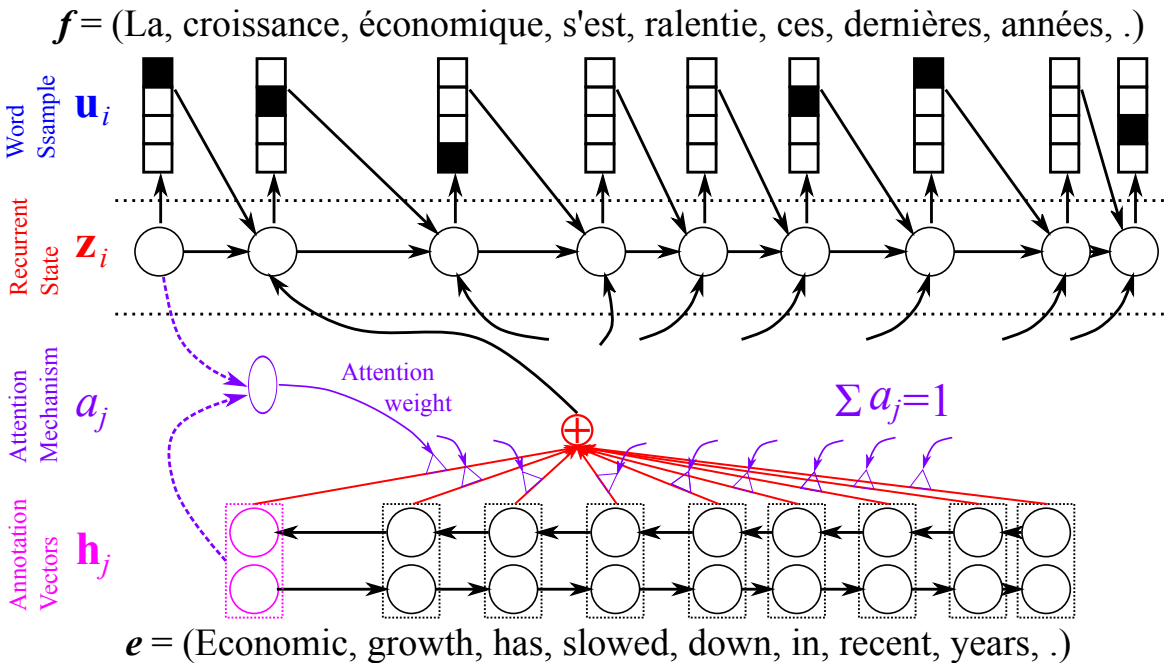
- With soft-attention: input fed to higher level at location **i** is a softmax-weighted sum of states at locations **j** at lower level
  - Train by back-prop
  - Fast training

- 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*

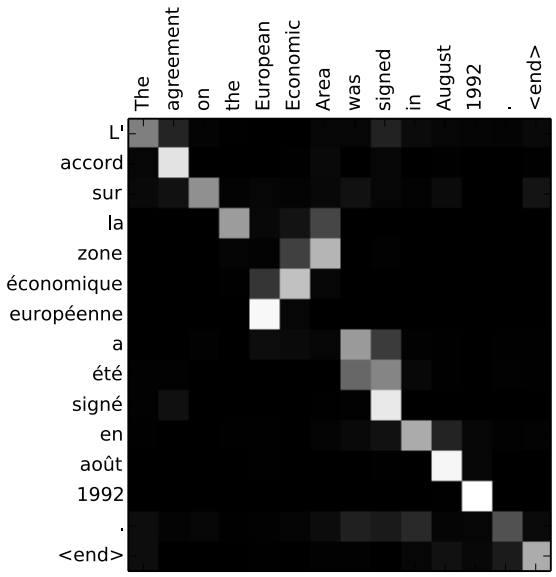
Attention-Based Neural Machine Translation

Related to earlier Graves 2013 for generating handwriting

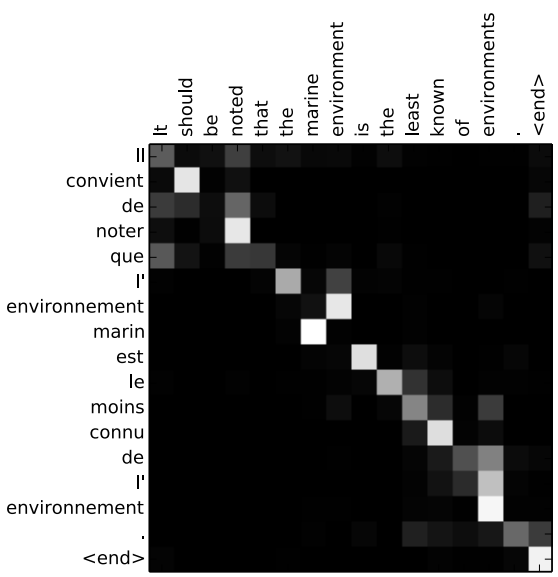
- (Bahdanau, Cho & Bengio, arXiv sept. 2014)
- (Jean, Cho, Memisevic & Bengio, arXiv dec. 2014)



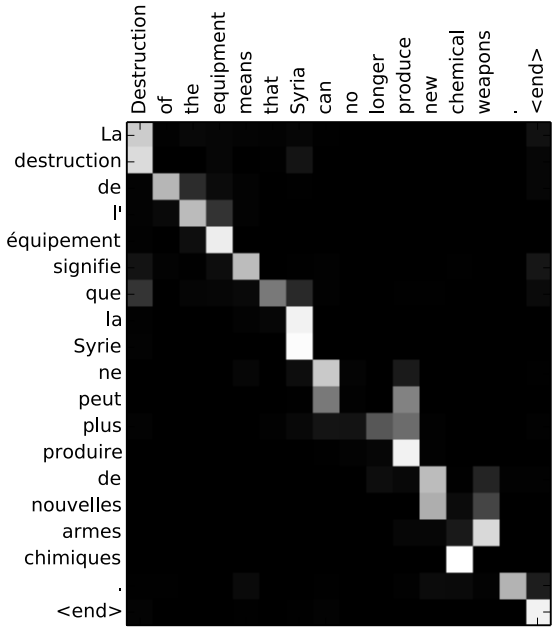
Predicted Alignments



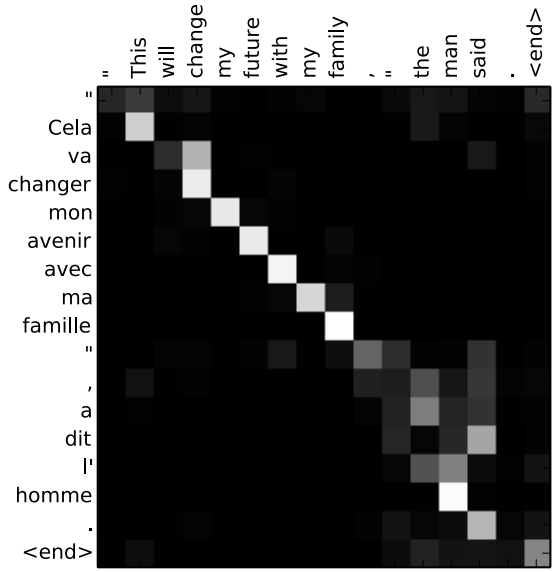
(a)



(b)

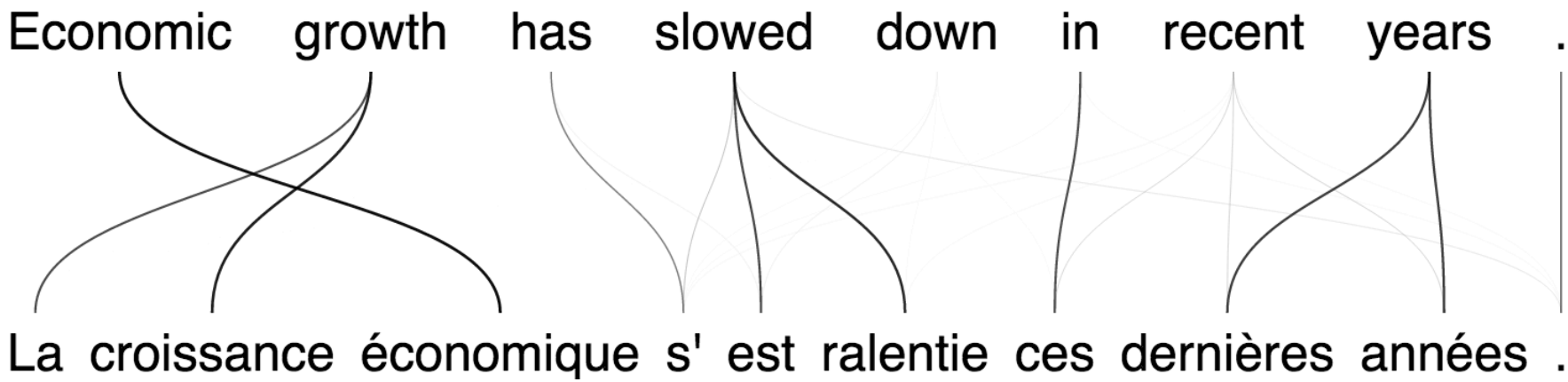


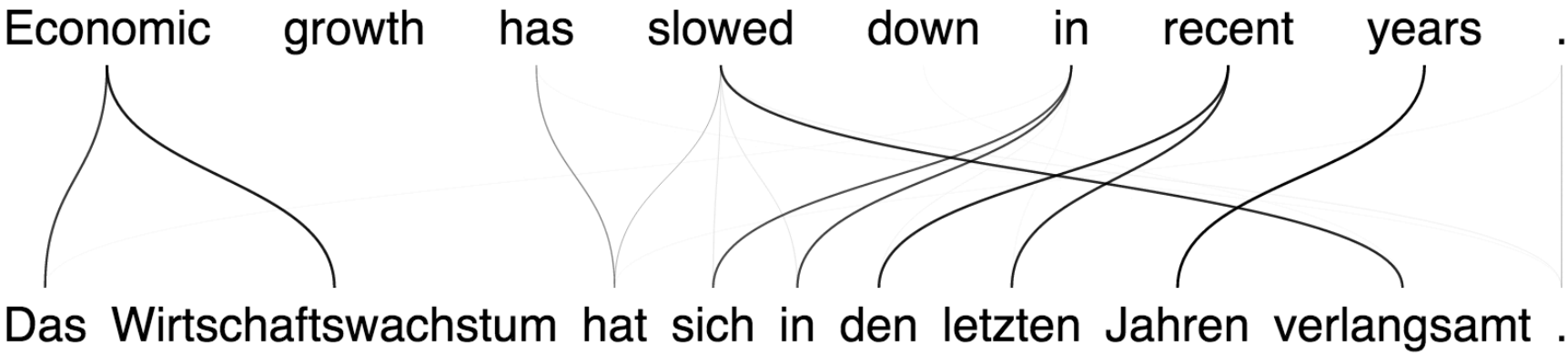
(c)



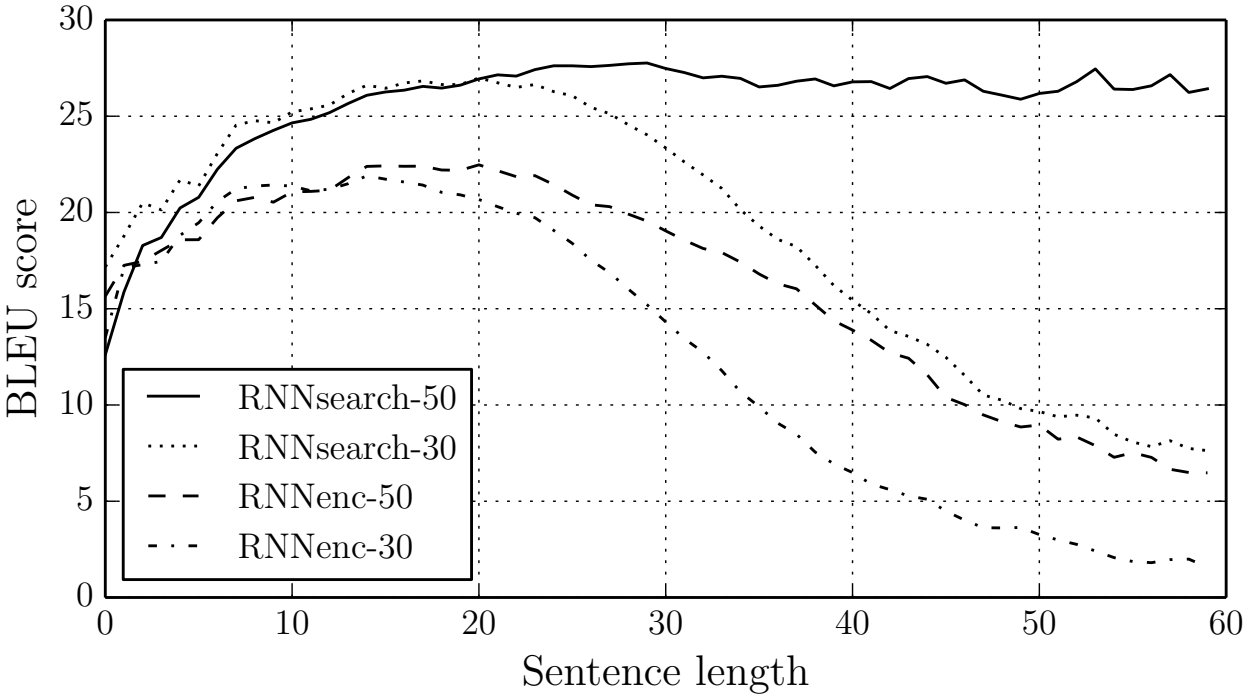
(d)

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)

Importance Sampling for Fast Training of Neural Language Models

(Bengio & Senecal 2008)

- IS: $E_p[f(x)] = \int p(x)f(x)dx = \int q(x)\frac{p(x)}{q(x)}f(x)dx = E_q[\frac{p(x)}{q(x)}f(x)]$
- During training of neural language model, the LL gradient is

$$\nabla \log p(y_t \mid y_{<t}, x)$$
$$= \nabla \mathcal{E}(y_t) - \sum_{k:y_k \in V} p(y_k \mid y_{<t}, x) \nabla \mathcal{E}(y_k)$$

- where $\mathcal{E}(y_j) = \mathbf{w}_j^\top \phi\left(y_{j-1}, z_j, c_j\right) + b_j$
- and the second term is an expectation that can be approximated by normalized importance sampling

$$\mathbb{E}_P\left[\nabla \mathcal{E}(y)\right] \approx \sum_{k:y_k \in V'} \frac{\omega_k}{\sum_{k':y_{k'} \in V'} \omega_{k'}} \nabla \mathcal{E}(y_k)$$

- with proposal distribution *Q* to sample *negative examples* V'

$$\omega_k = \exp\left\{\mathcal{E}(y_k) - \log Q(y_k)\right\}$$

Fast GPU Training with Large Vocabulary using Minibatch Importance Sampling *(Jean et al, arXiv 2015)*

- *(Bengio & Senecal 2008)* not adapted to the current GPU reality

- (Jean et al, arXiv 2015) uses the following scheme:
  - Proposal *Q* for a particular word y_t in a particular minibatch is uniform among the words present in the minibatch
  - Just optimize wrt following relative probability inside the minibatch, normalizing only over the words V' in minibatch:

$$p(y_t \mid y_{<t}, x) = \frac{\exp\left\{\mathbf{w}_t^\top \phi\left(y_{t-1}, z_t, c_t\right) + b_t\right\}}{\sum_{k:y_k \in V'} \exp\left\{\mathbf{w}_k^\top \phi\left(y_{t-1}, z_t, c_t\right) + b_k\right\}}$$

Out-of-Vocabulary Words

- During training the model is asked to generate UNK for OOV words
- At test time, when UNK is generated, we use a forced alignment to find the corresponding source word(s) and output them

- This is particularly important for proper nouns, numerical quantities, etc. and boosted our performance significantly (1.5 BLEU points)

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**	

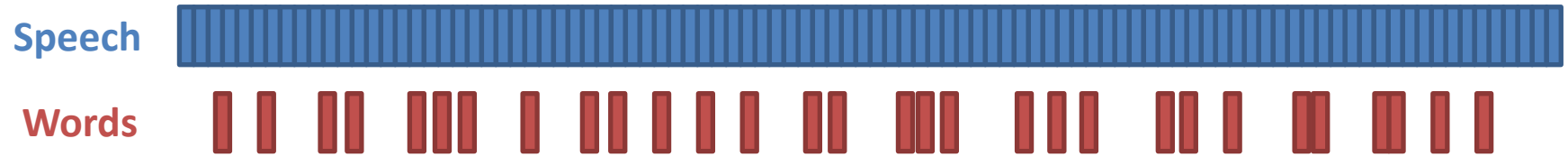
(b) **English→German**

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)

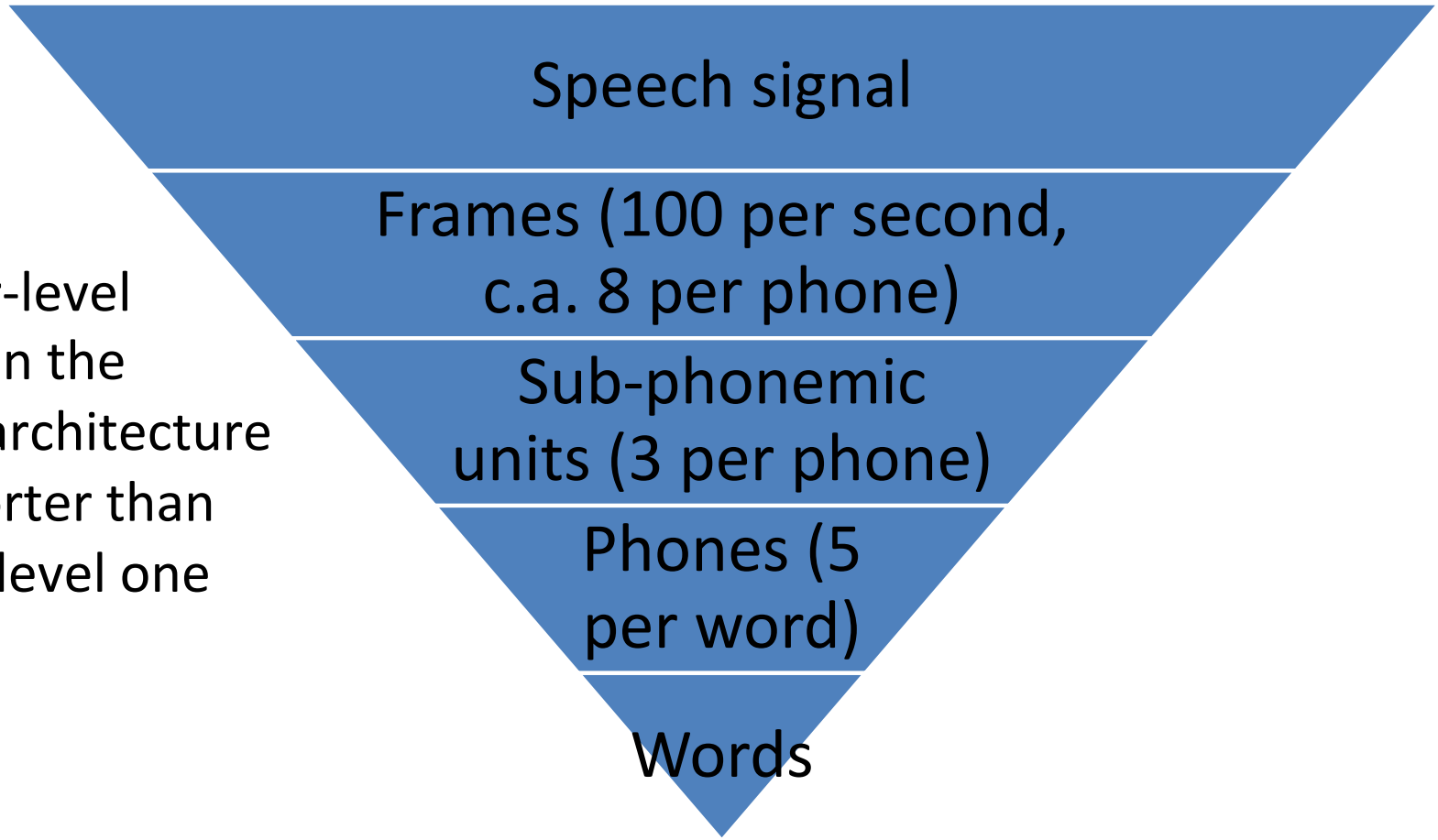
Translating from Other Sources?

- Speech
- Images
- Video

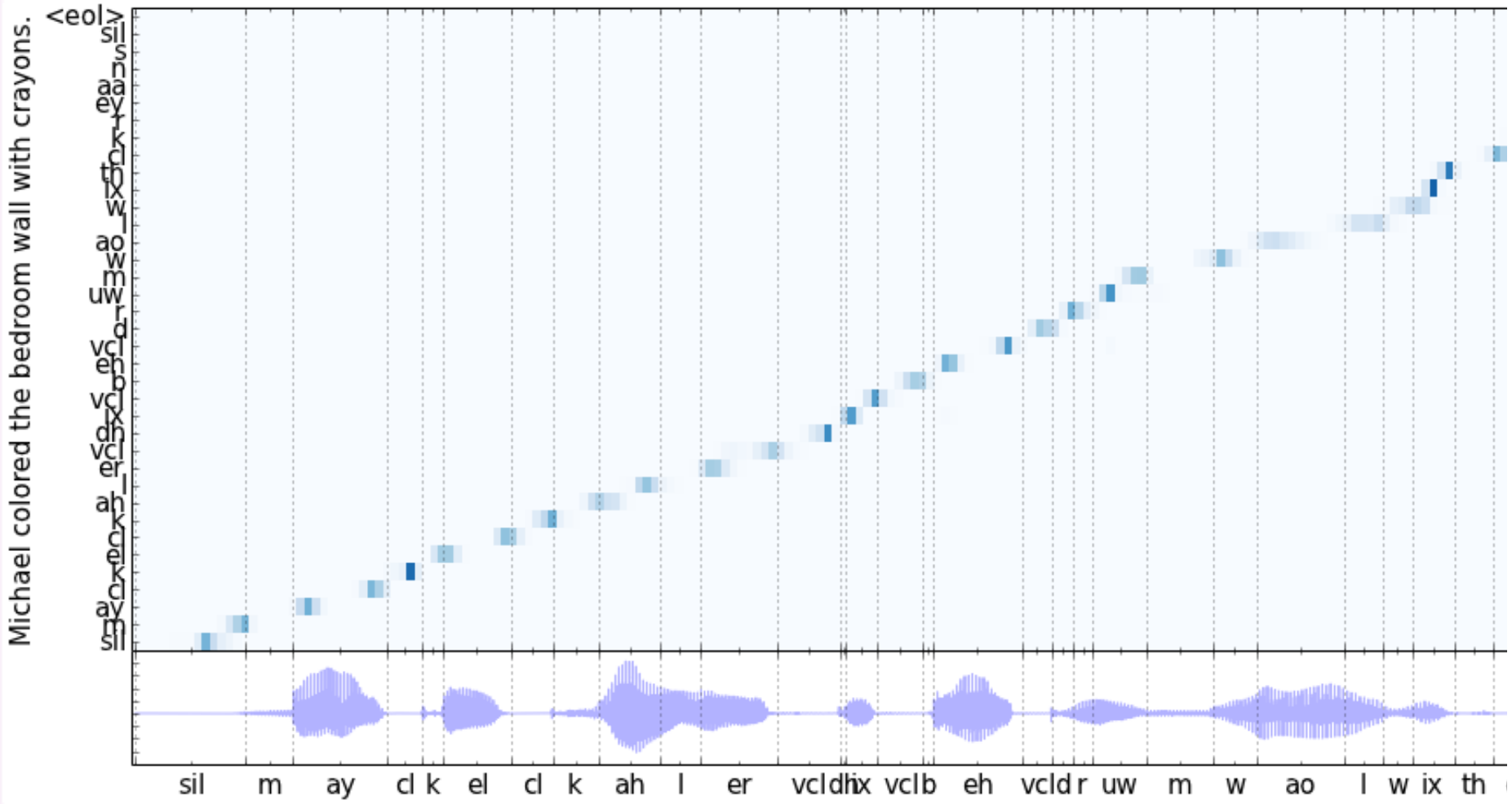
Multiple Time Scales in Speech



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

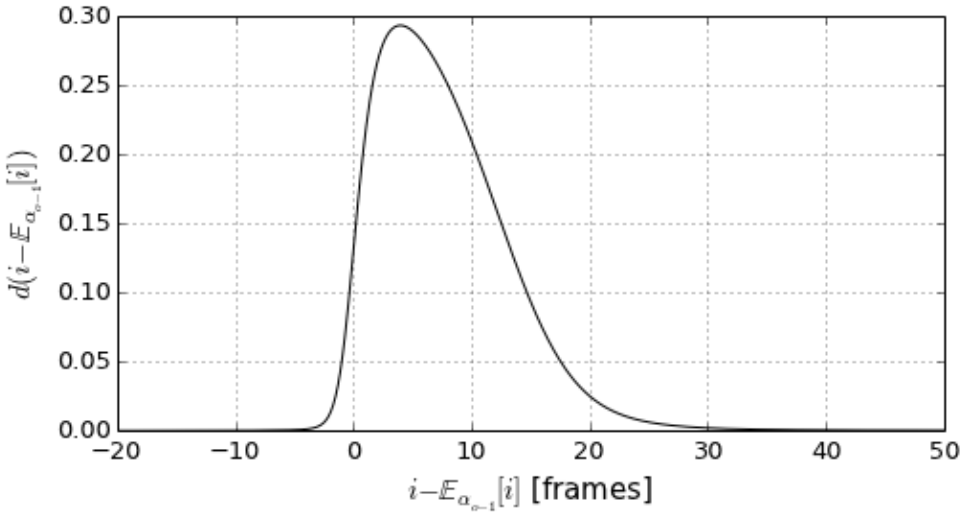


Acoustic-to-Phones Attention Alignment



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

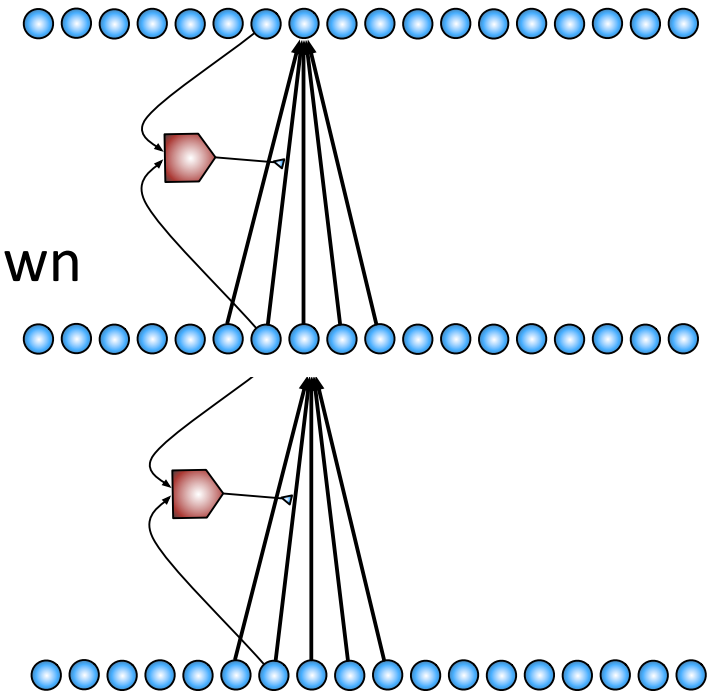
End-to-end Continuous Speech Recognition using Attention-based Recurrent NN: First Results

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

Model	DEV	TEST
Kaldi TIMIT s5 recipe with `basic` scorer		
Speaker independent triphone GMM-HMM	23.15%	24.32%
Speaker adapted triphone GMM-HMM	20.56%	21.65%
DBN-HMM with SMBR training	17.55%	18.79%
DBN-HMM with SMBR training and greedy search	32.02%	33.00%
Proposed model with the `basic` scorer		
Maxout network with per-frame training	18.41%	19.68%
RNN model with frozen acoustic layer	17.53%	18.68%
RNN model trained end-to-end	16.88%	18.57%
RNN model trained end-to-end with greedy search	17.06%	18.61%
References		
Deep RNN Transducer (Graves *et al.*, 2013b)	N/A	17.7%

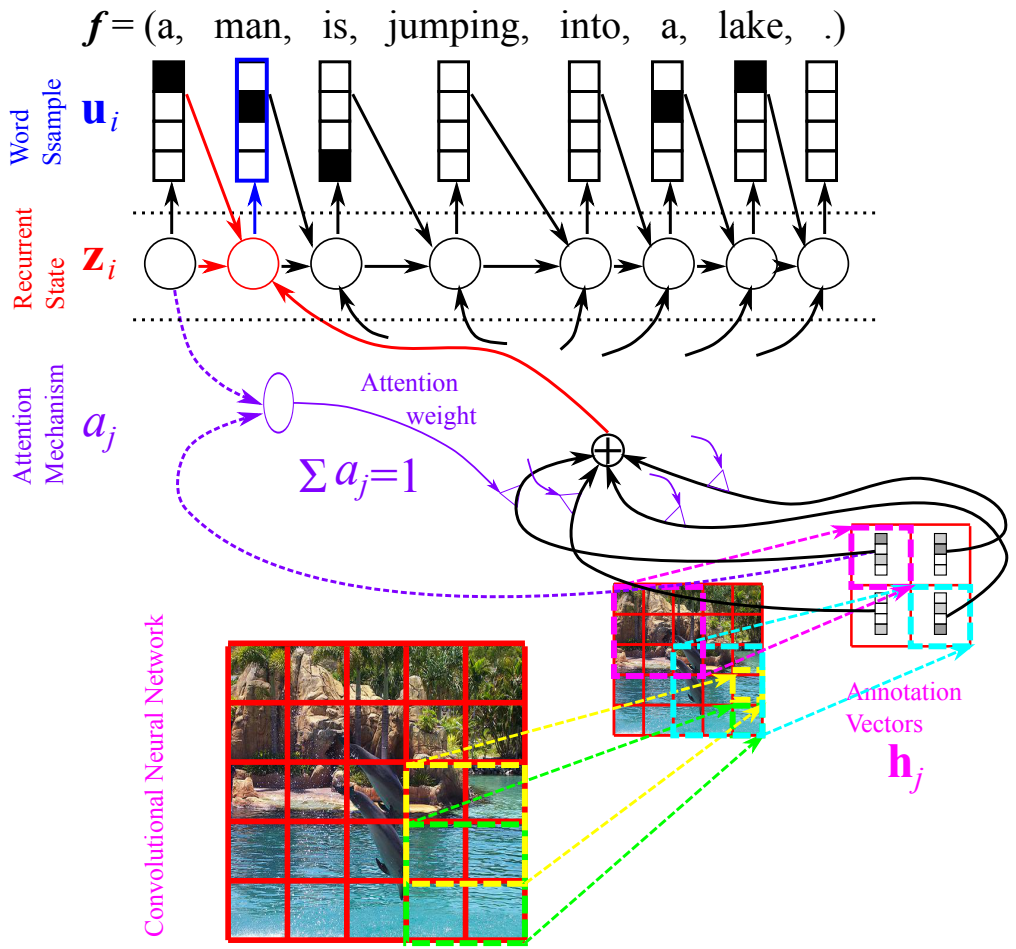
Ongoing Work: Multi-Level Attention

- Higher levels may represent slower time scale, asynchronously and adaptively reading from lower levels (automatic soft segmentation)

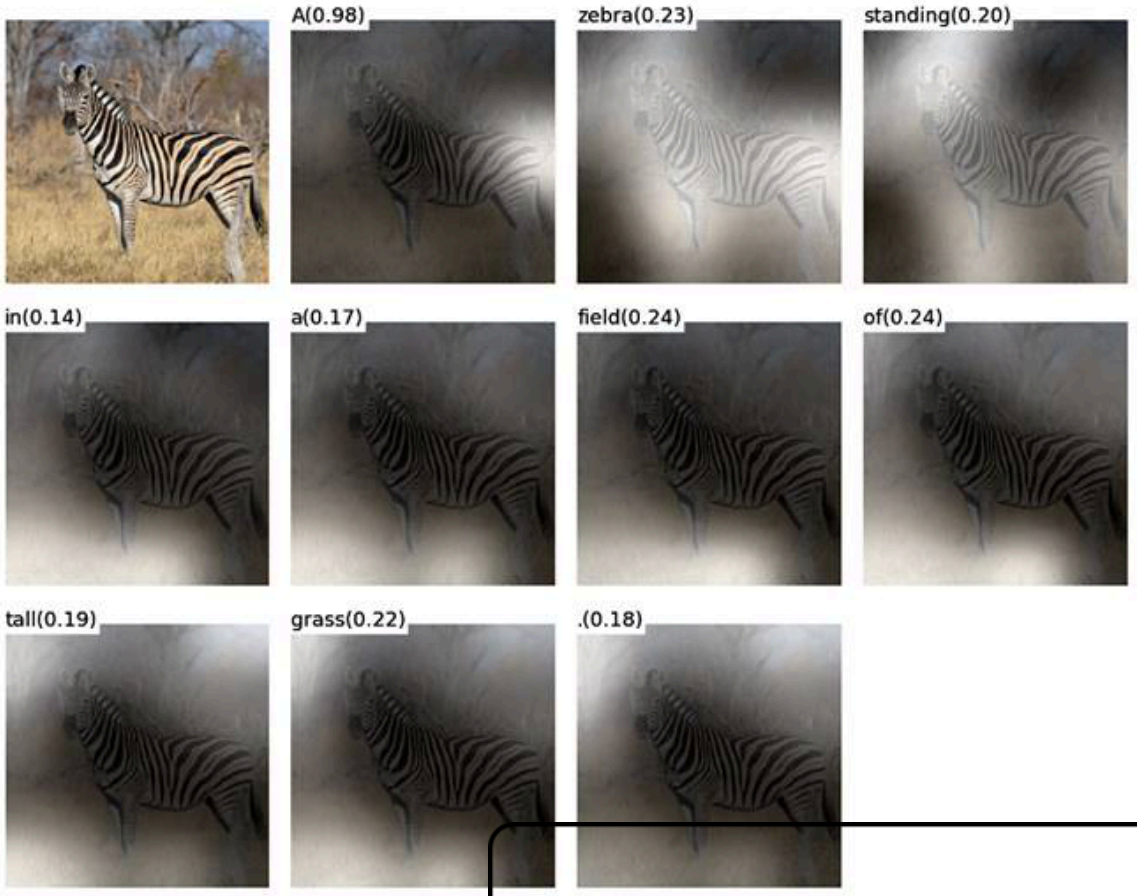


- Challenge: during training, the length of inner sequences is not known
  - predict stopping prob. at each time step
  - weigh upper attn weights in proportion to these prob.

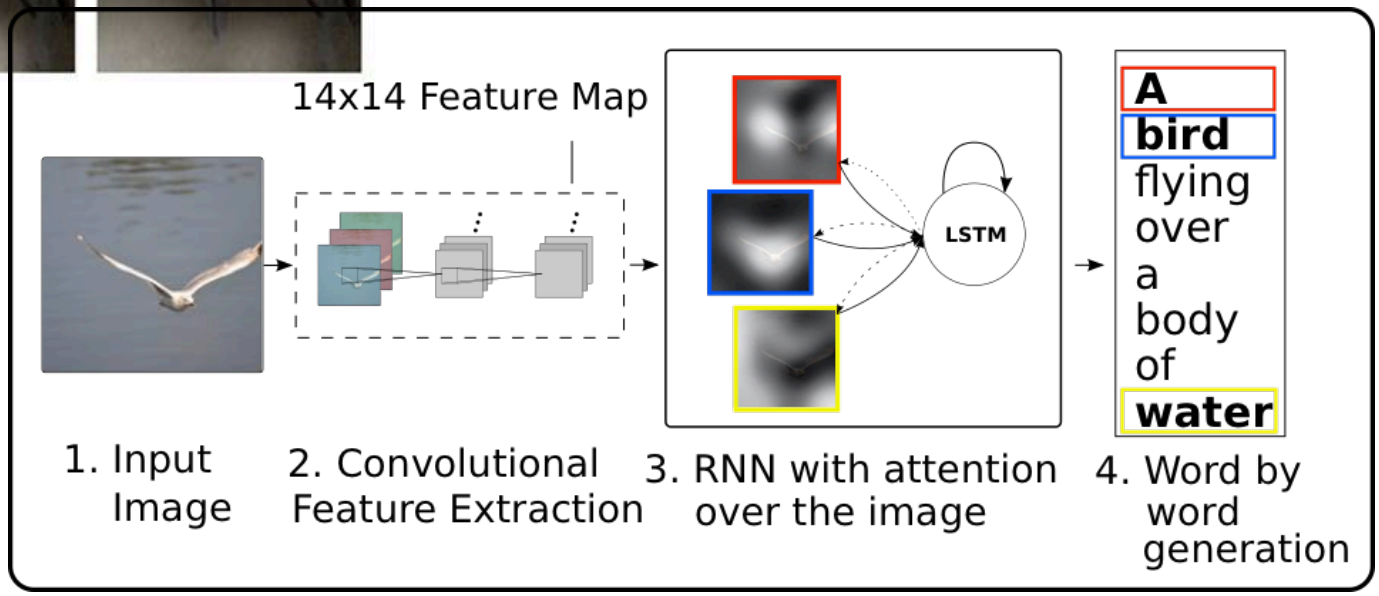
Image-to-Text: Caption Generation



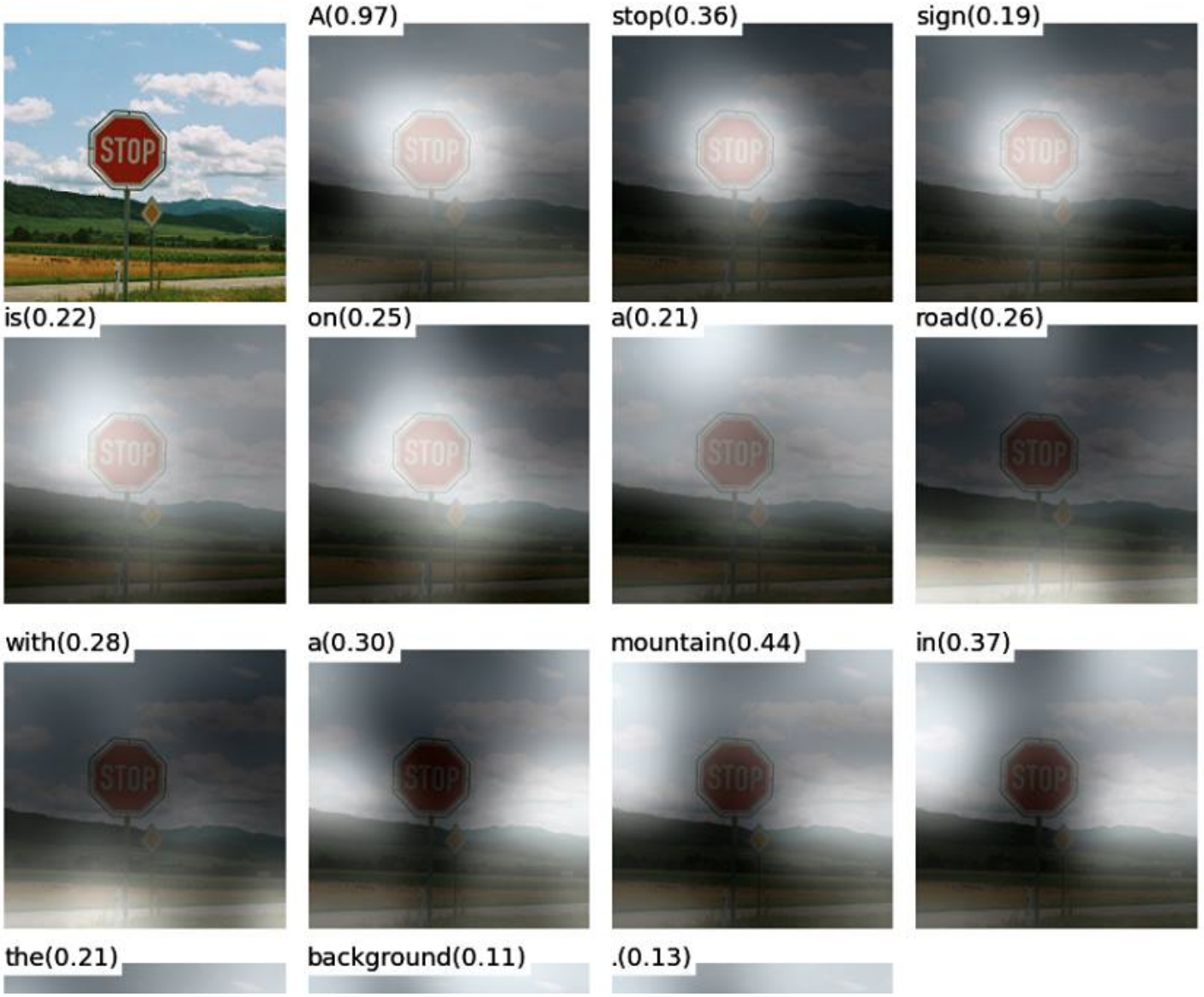
(Xu et al., 2015), (Yao et al., 2015)



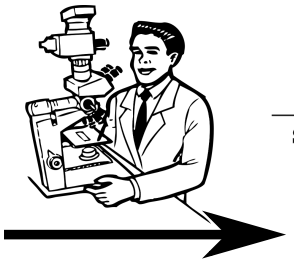
Paying Attention to Selected Parts of the Image While Uttering Words



Speaking about what one sees



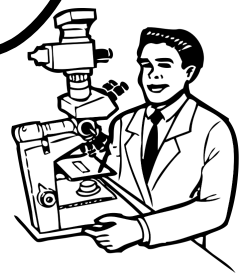
Let's go back 2.5 years back in time.. (Mitchell et al., 2012)



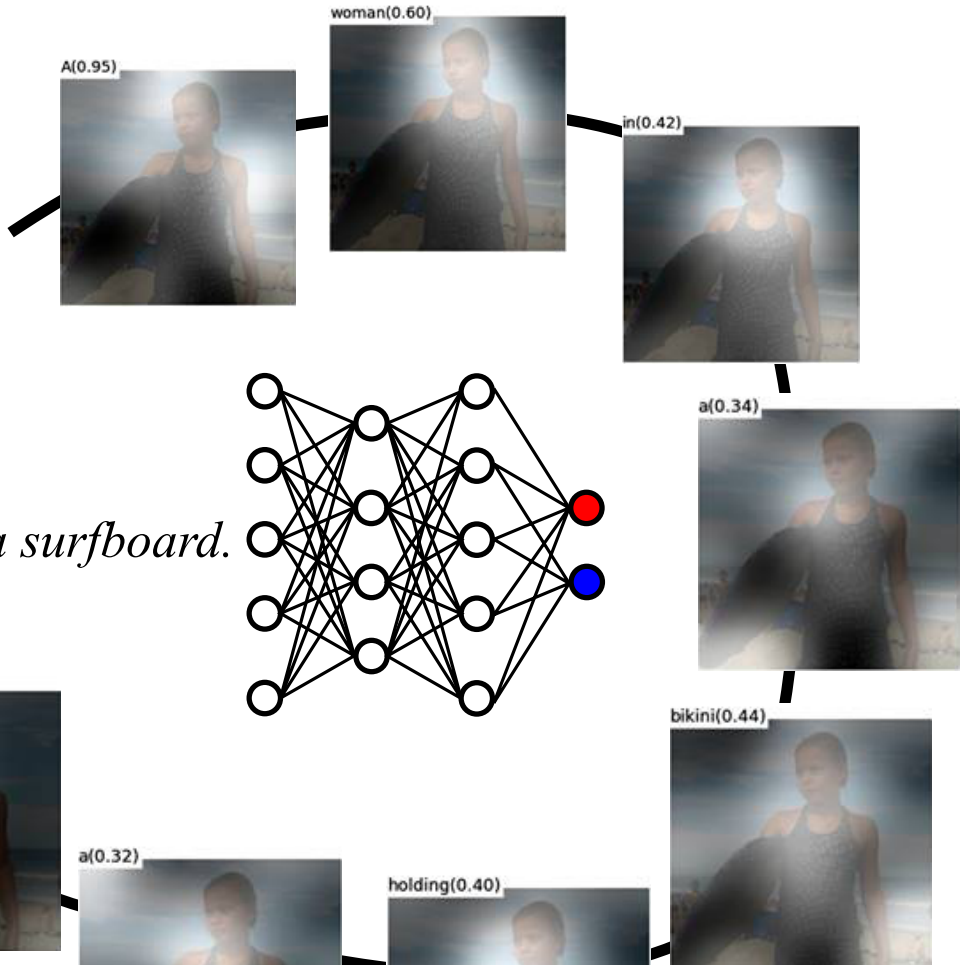
stuff:	*sky*	.999	
	id:	1	
	atts:	clear:0.432,	blue:0.945
		grey:0.853,	white:0.501 ...
	b. box:	(1,1 440,141)	
stuff:	*road*	.908	
	id:	2	
	atts:	wooden:0.722	clear:0.020 ...
	b. box:	(1,236 188,94)	
object:	*bus*	.307	
	id:	3	
	atts:	black:0.872,	red:0.244 ...
	b. box:	(38,38 366,293)	
preps:	id 1,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

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.

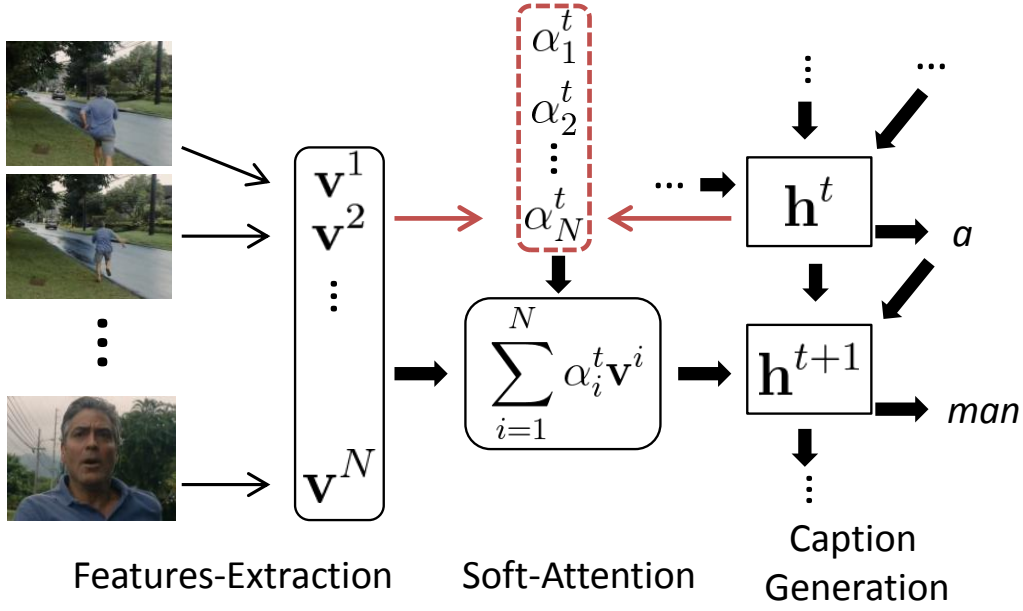
A person is standing on a beach
with a surfboard.

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

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

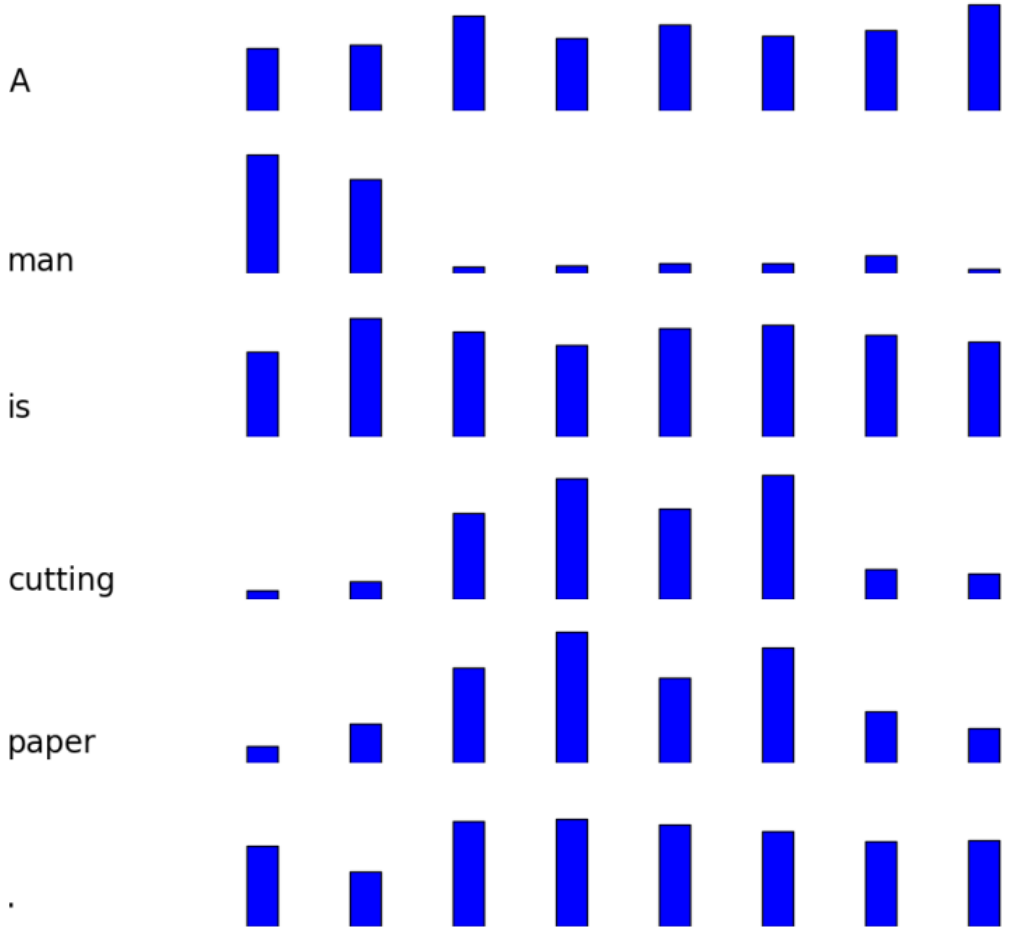
Attention through time for video caption generation

- (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.

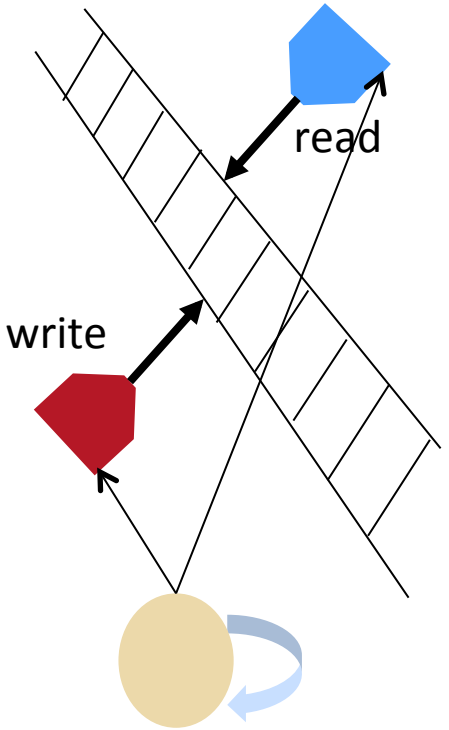
Corpus:
SOMEONE sits with his arm around SOMEONE.
He nuzzles her cheek, then kisses tenderly.
Test_sample:
SOMEONE sits beside SOMEONE.

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

Generated captions

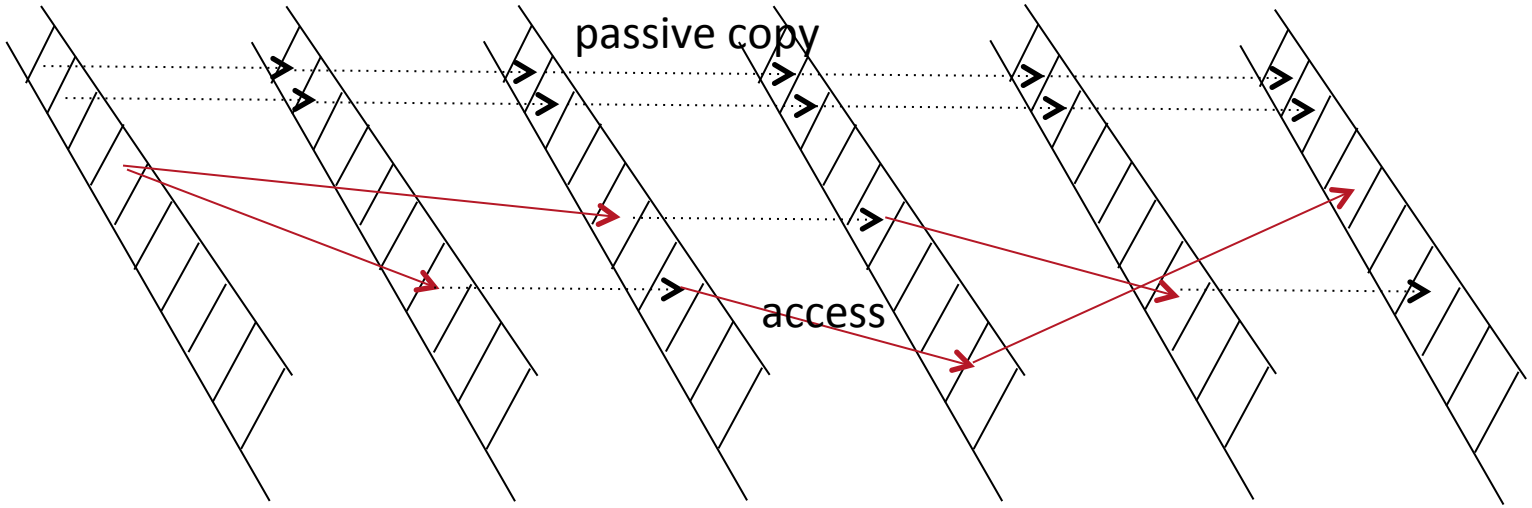
Attention Mechanisms for Memory Access

- 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

- 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, 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