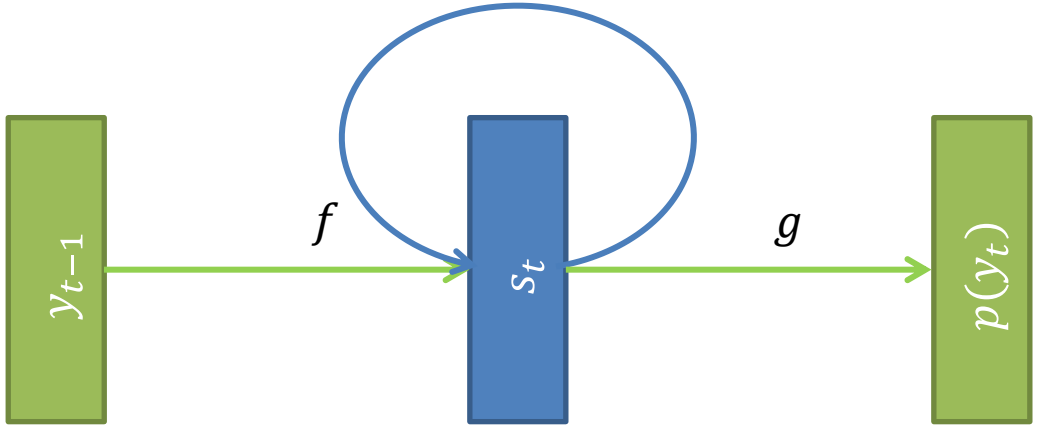
Neural nets for NLP: Attention Mechanism

Neural Networks

RNNs Learn $p(Y)$



Decompose

$$p(Y) = \prod p(y_t | y_{t-1}, y_{t-2}, \cdots, y_1)$$

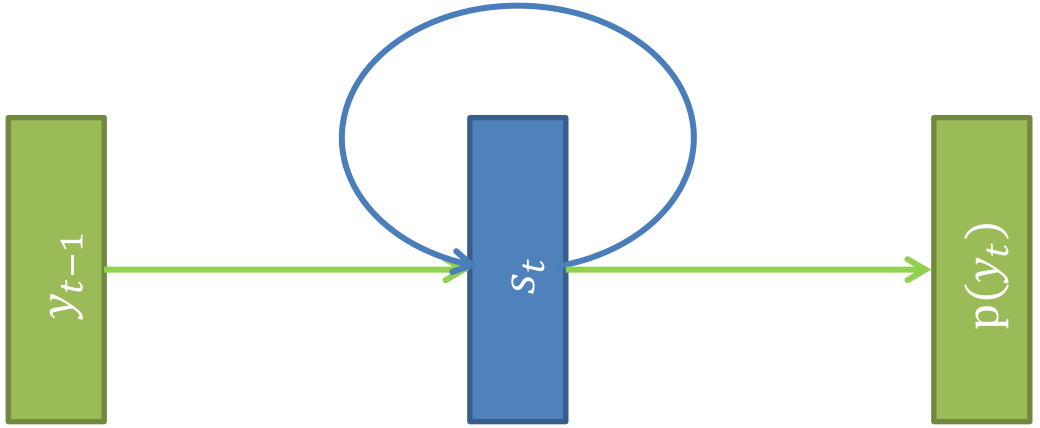
Model the probabilities using a recurrent relation

$$p(y_t | y_{t-1}, y_{t-2}, \cdots, y_1) = g(s_t)$$

$$s_t = f(s_{t-1}, y_{t-1})$$

$g()$, $f()$ are implemented using neural networks, i.e. they are flexibly parameterized, smooth functions.

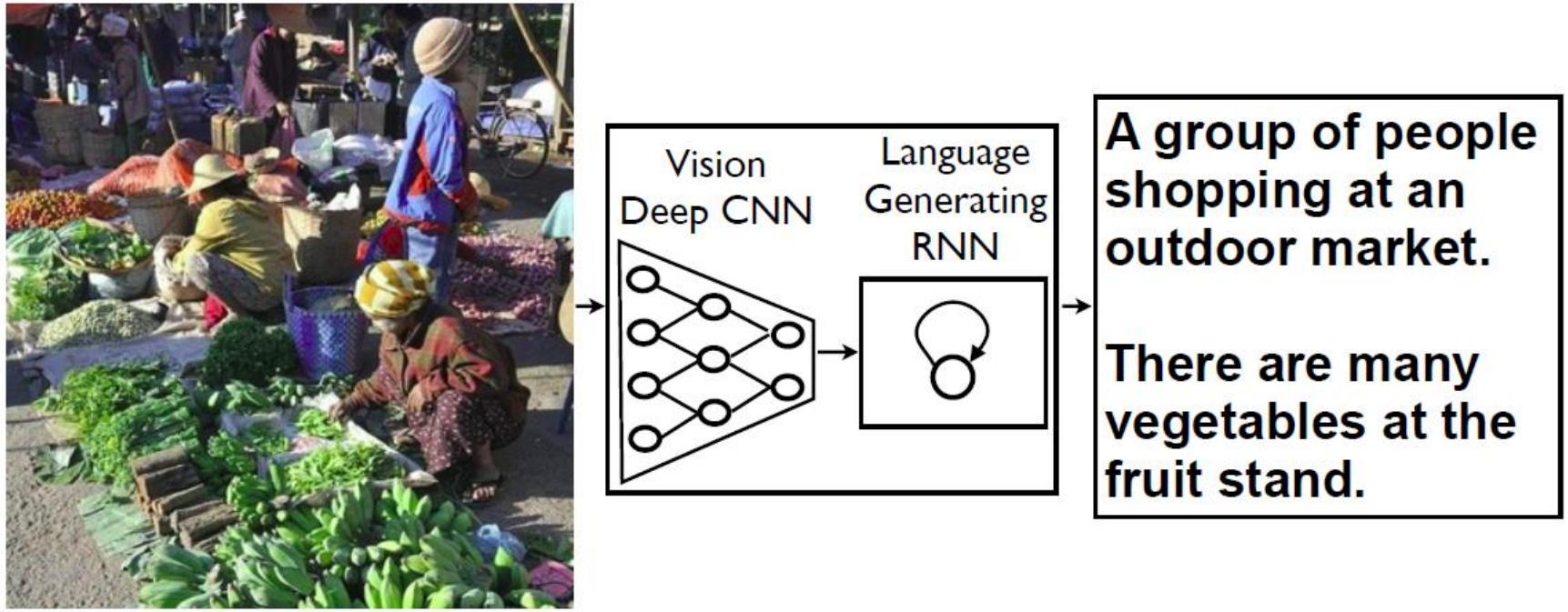
How to condition an RNN?



RNN gives us $p(Y)$ but we want $p(Y|X)$

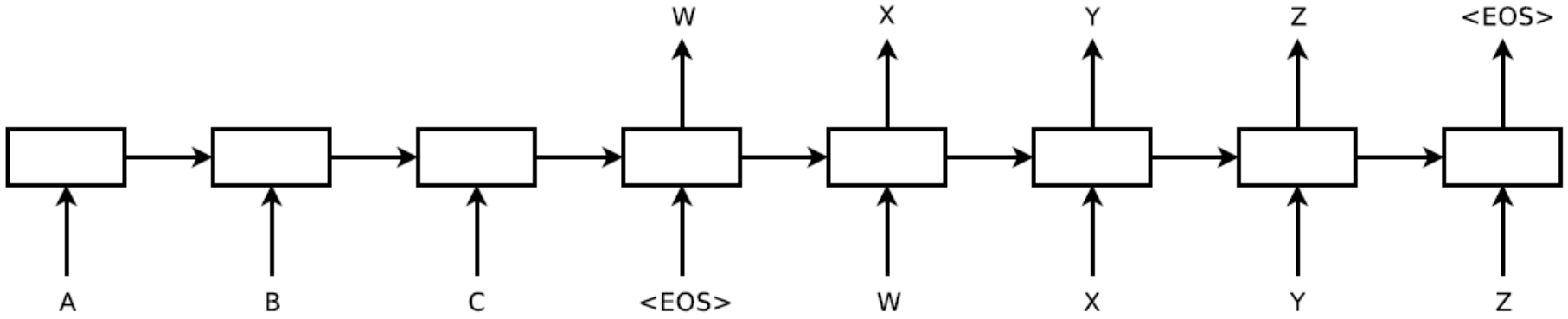
- Idea #1: conditioned through the first hidden state
- Idea #2: condition separately on every step

Idea #1
condition through the 1st hidden state



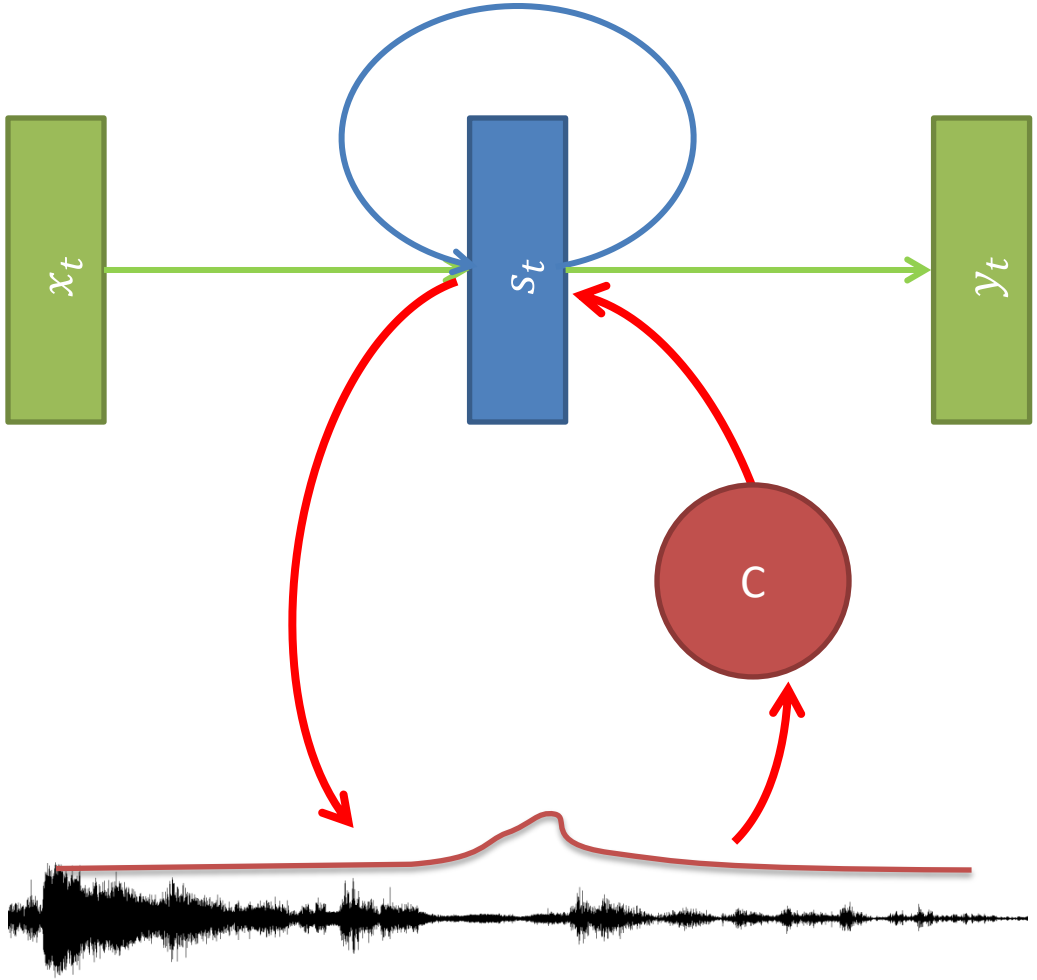
O. Vinyals et al, “Show and Tell: a Neural Image Caption Generator https://arxiv.org/abs/1411.4555

Idea #1
condition through the 1st hidden state



http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf

Idea #2: Attention



1. Choose relevant frames

$$e_f = \text{score}(x_f, s_{t-1})$$
$$\alpha_f = \text{SoftMax}(e)_f$$

2. Summarize into context

$$c = \sum_f \alpha_f x_f$$

3. Compute next state

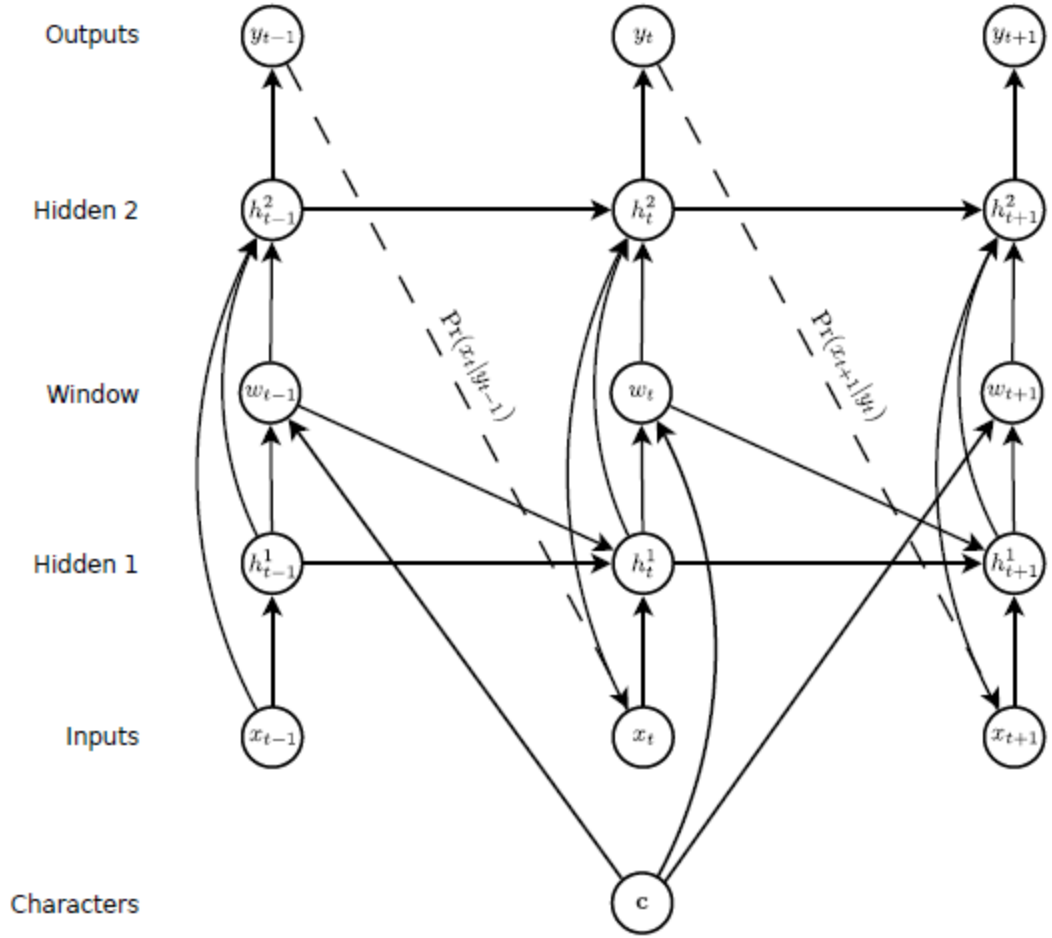
$$s_t = f(s_{t-1}, y_{t-1}, c)$$

Soft attention originally proposed for the Montreal neural translation project [Bahdanau2014]

Attention mechanism in RNNs

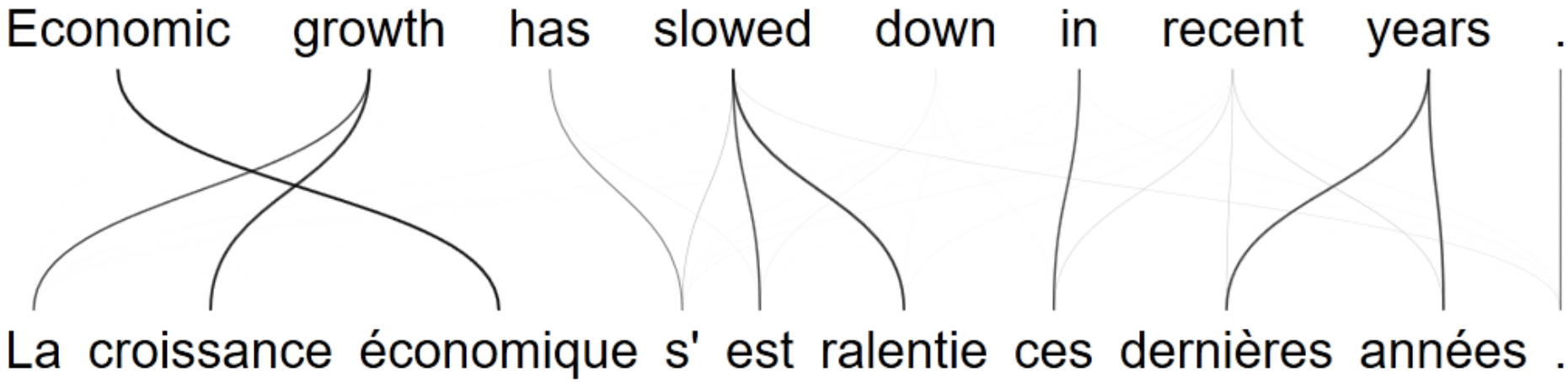
from his travels it might have been
from his travels it might have been
from his travels it might have been

- This is a network to generate handwriting
- At each step the network looks at a *context* $\boldsymbol{c}$
- $\boldsymbol{c}$ is a summarization of a small fragment of the input sequence

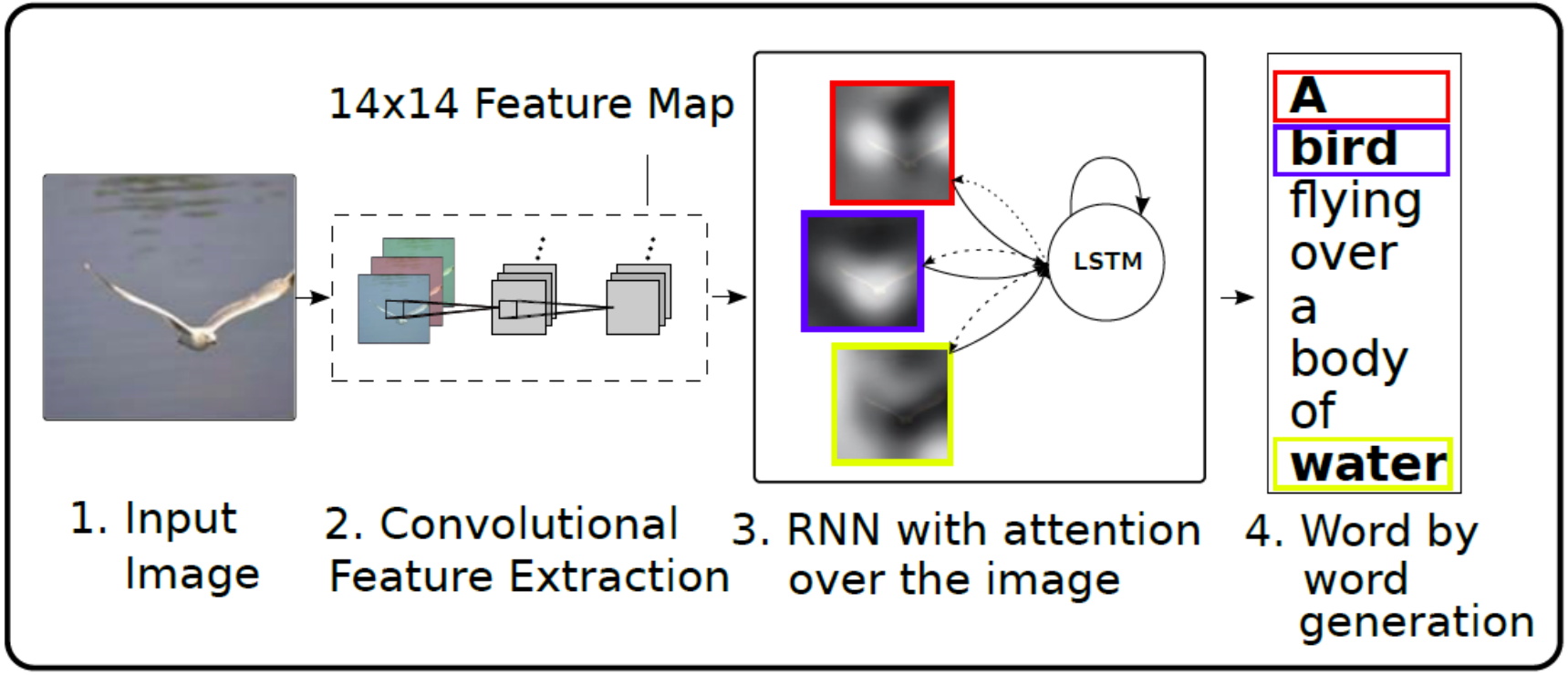


Graves, A., 2013. Generating Sequences With Recurrent Neural Networks. arXiv:1308.0850 [cs]

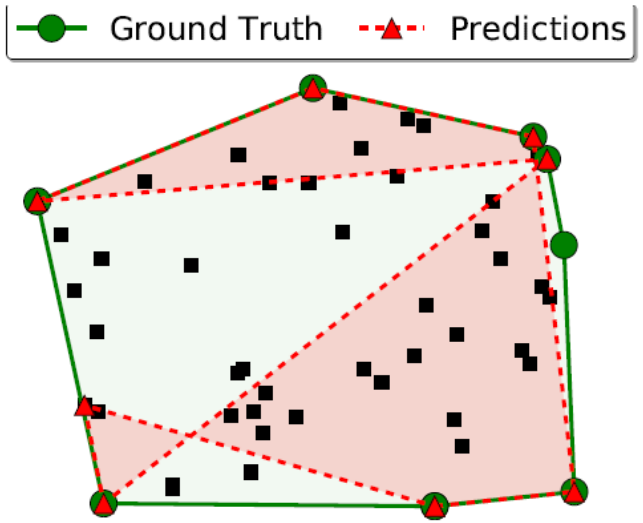
Attention mechanism in translation



Attention mechanism for captioning

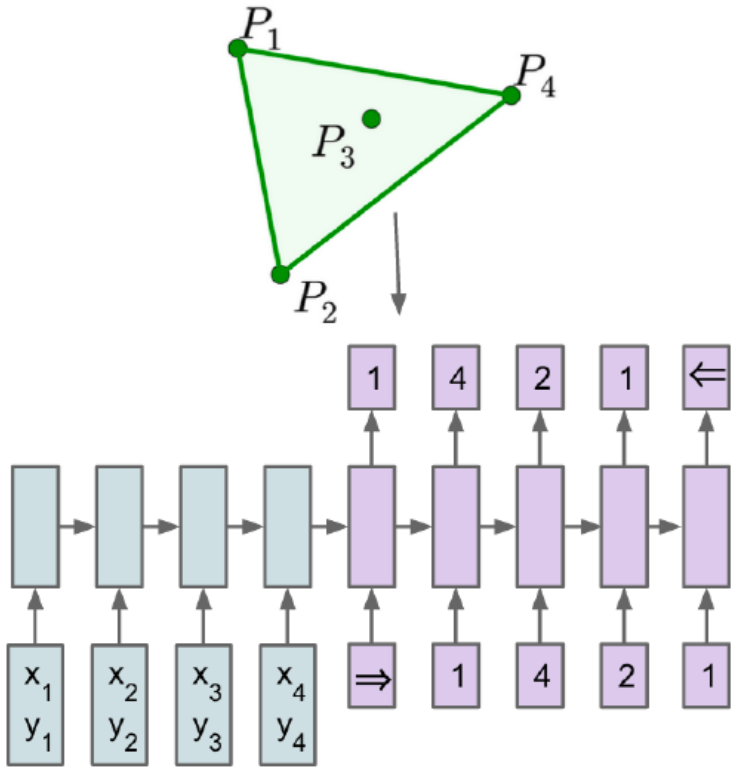


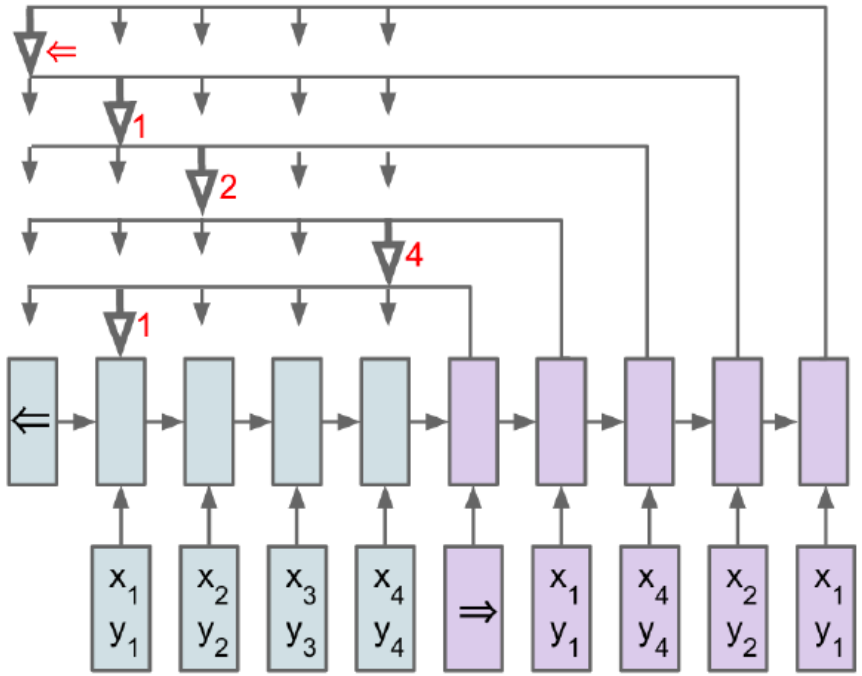
http://arxiv.org/pdf/1502.03044.pdf



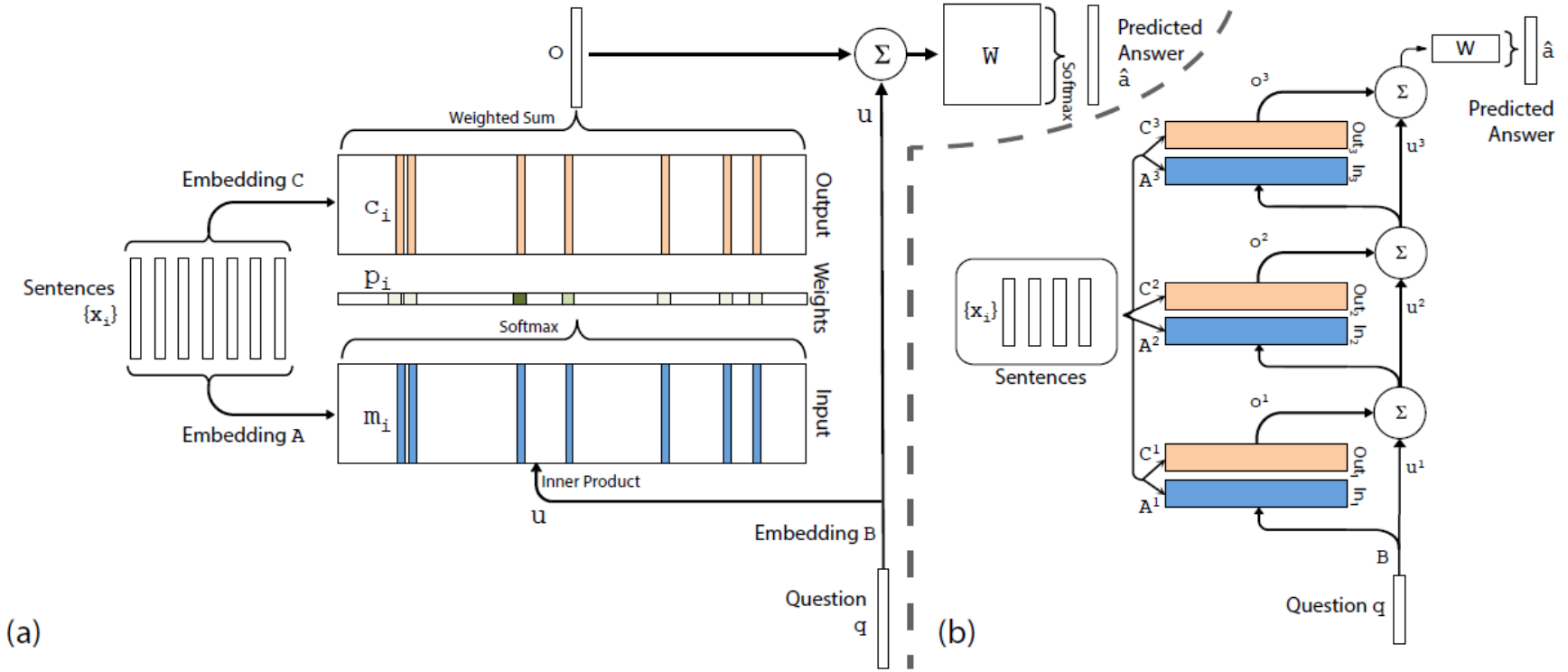
Convex Hulls & TSP

http://papers.nips.cc/paper/5866-pointer-networks.pdf





Reasoning – facts in memory



Story (16: basic induction)	Support	Hop 1	Hop 2	Hop 3
Brian is a frog.	yes	0.00	0.98	0.00
Lily is gray.		0.07	0.00	0.00
Brian is yellow.	yes	0.07	0.00	1.00
Julius is green.		0.06	0.00	0.00
Greg is a frog.	yes	0.76	0.02	0.00
What color is Greg? Answer: yellow Prediction: yellow				

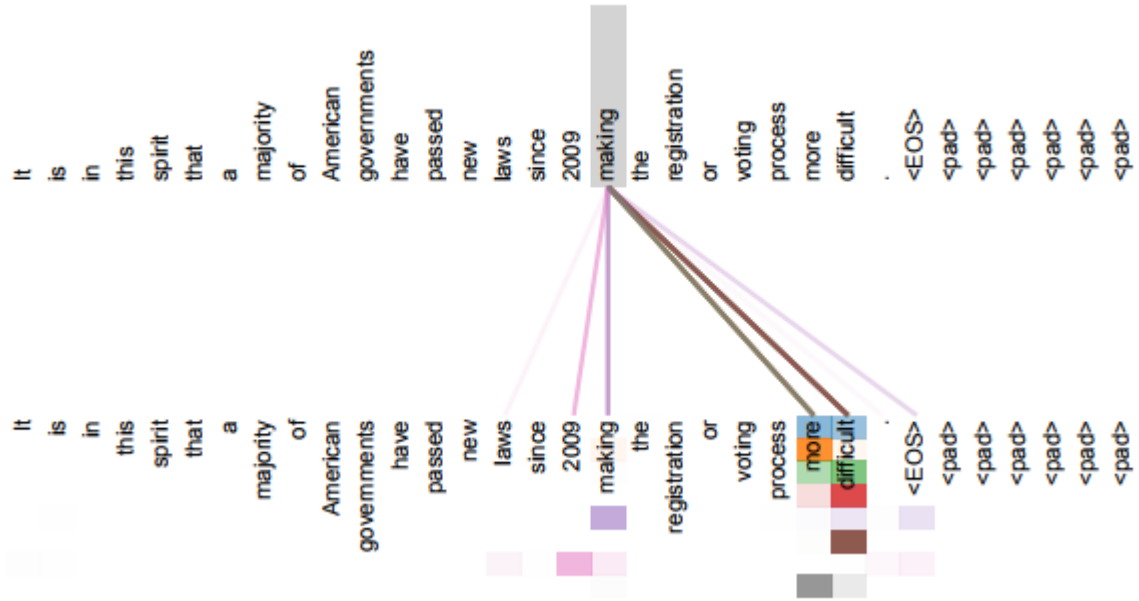
http://papers.nips.cc/paper/5846-end-to-end-memory-networks.pdf

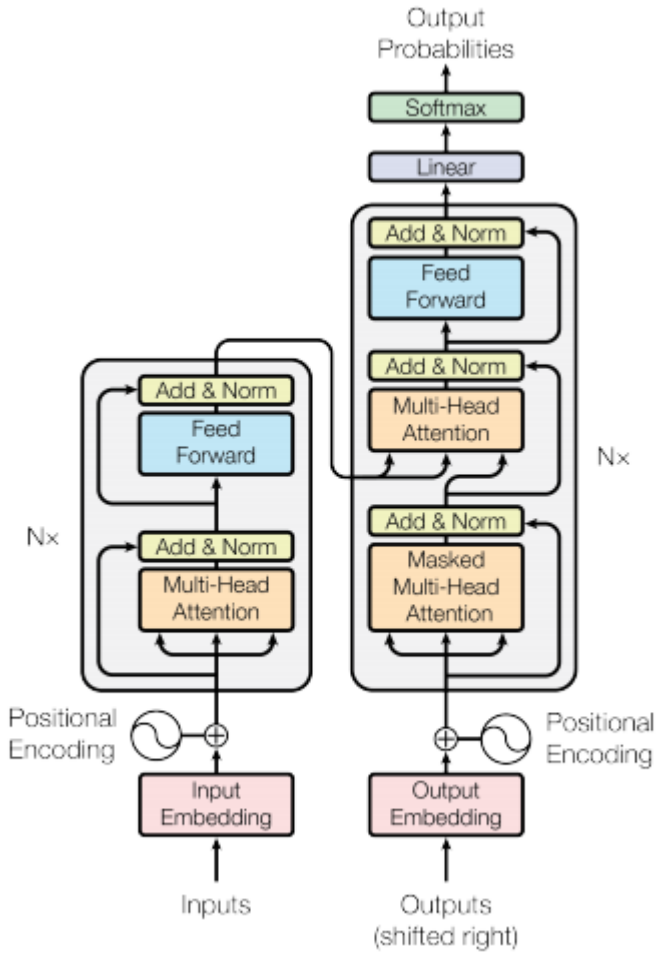
Current NLP King: Transformer AKA Attention is All You Need

RNN: compress history into the state vector

UniRNN: attention over history!

BiRNN: attention over whole sequence





A. Vasvani et al. https://arxiv.org/pdf/1706.03762.pdf