From Recurrent Models to the advent of Attention: a recap

Giuseppe Attanasio

Data Science Lab: process and methods

Research Bites | December 18, 2020

Politecnico di Torino

From Recurrent Models to the advent of Attention: a recap

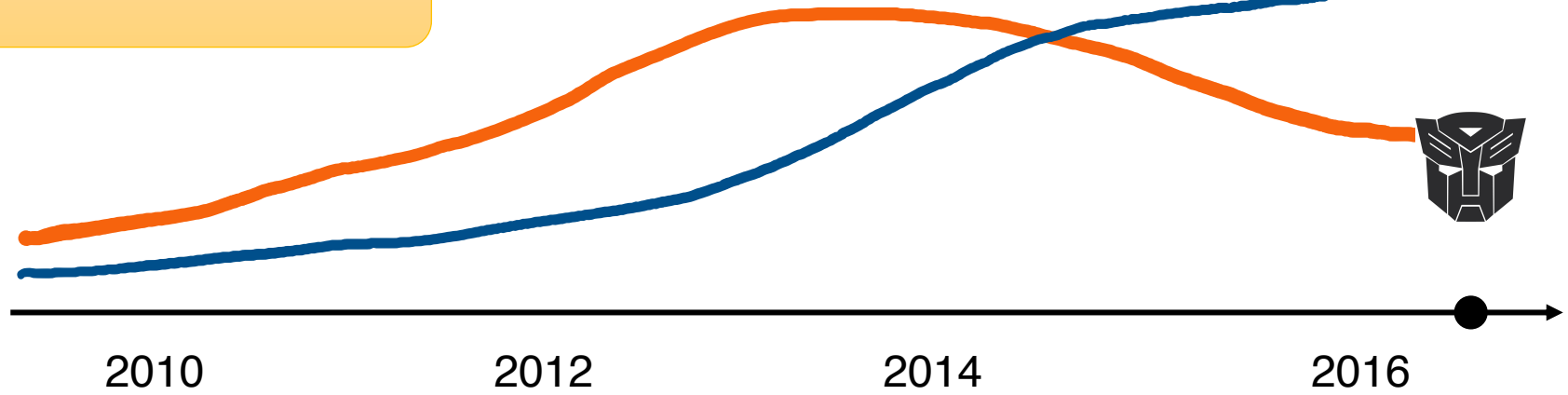
Disc. #1: we'll focus on intuitions. Many further technicalities are left aside.

1

From Recurrent Models to the advent of Attention: a recap

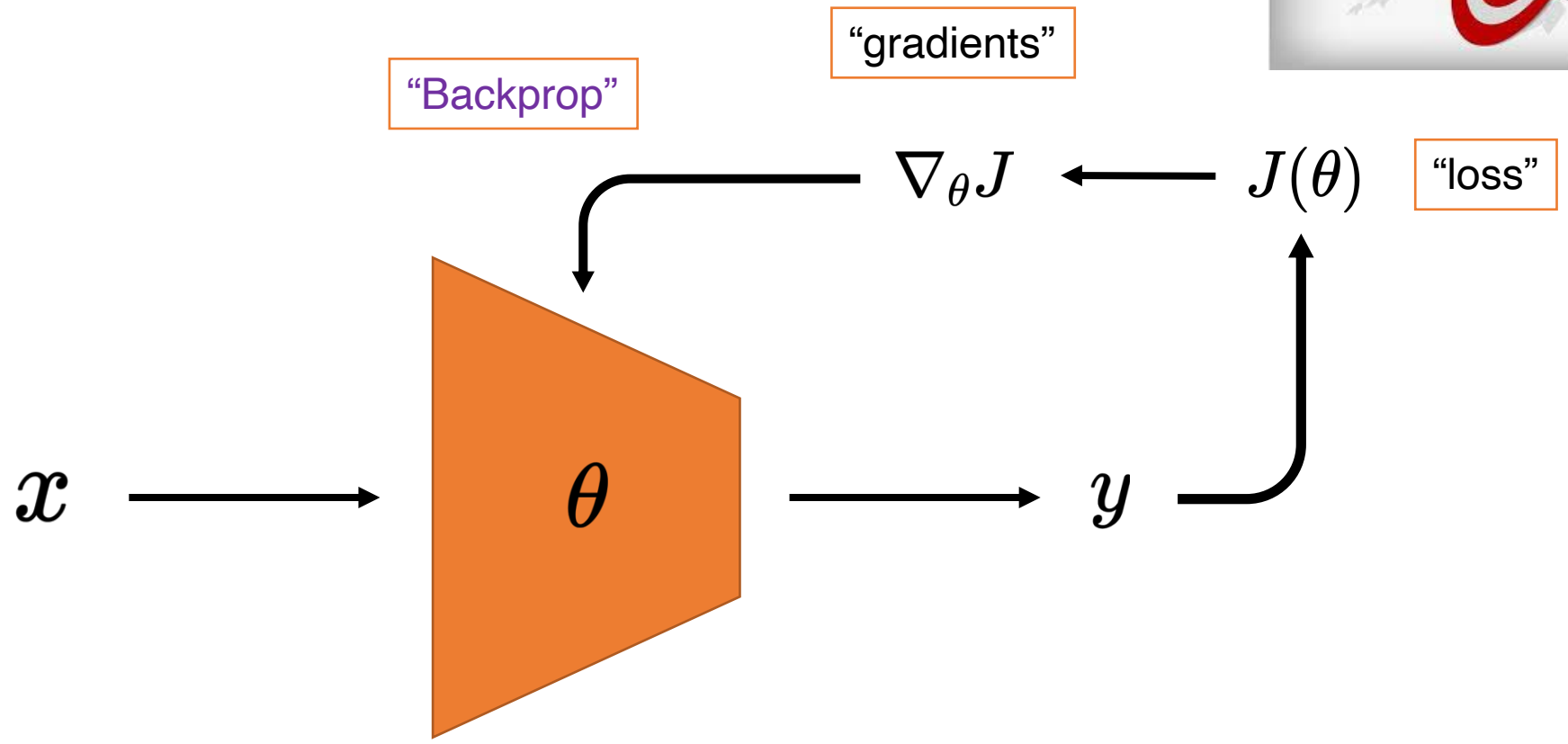
2

Disc #2: A NLP historical walkthrough

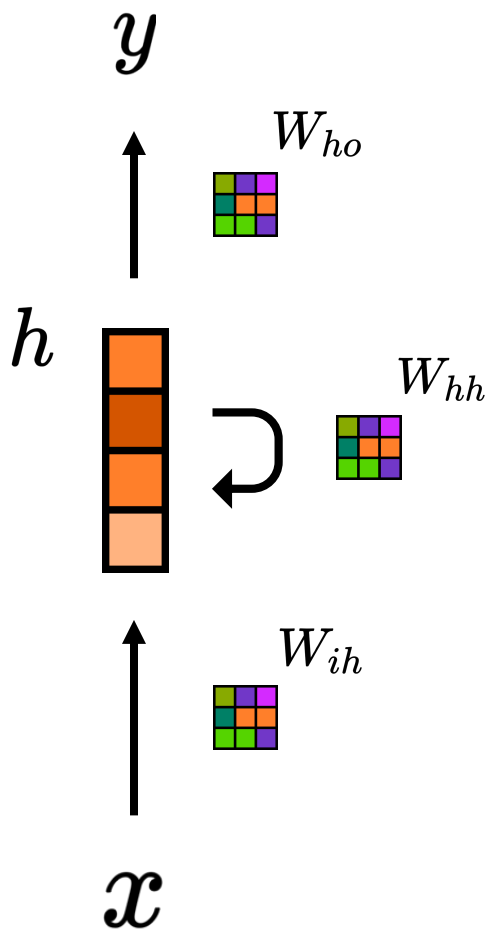


Attention is all you need. Waswani et al.

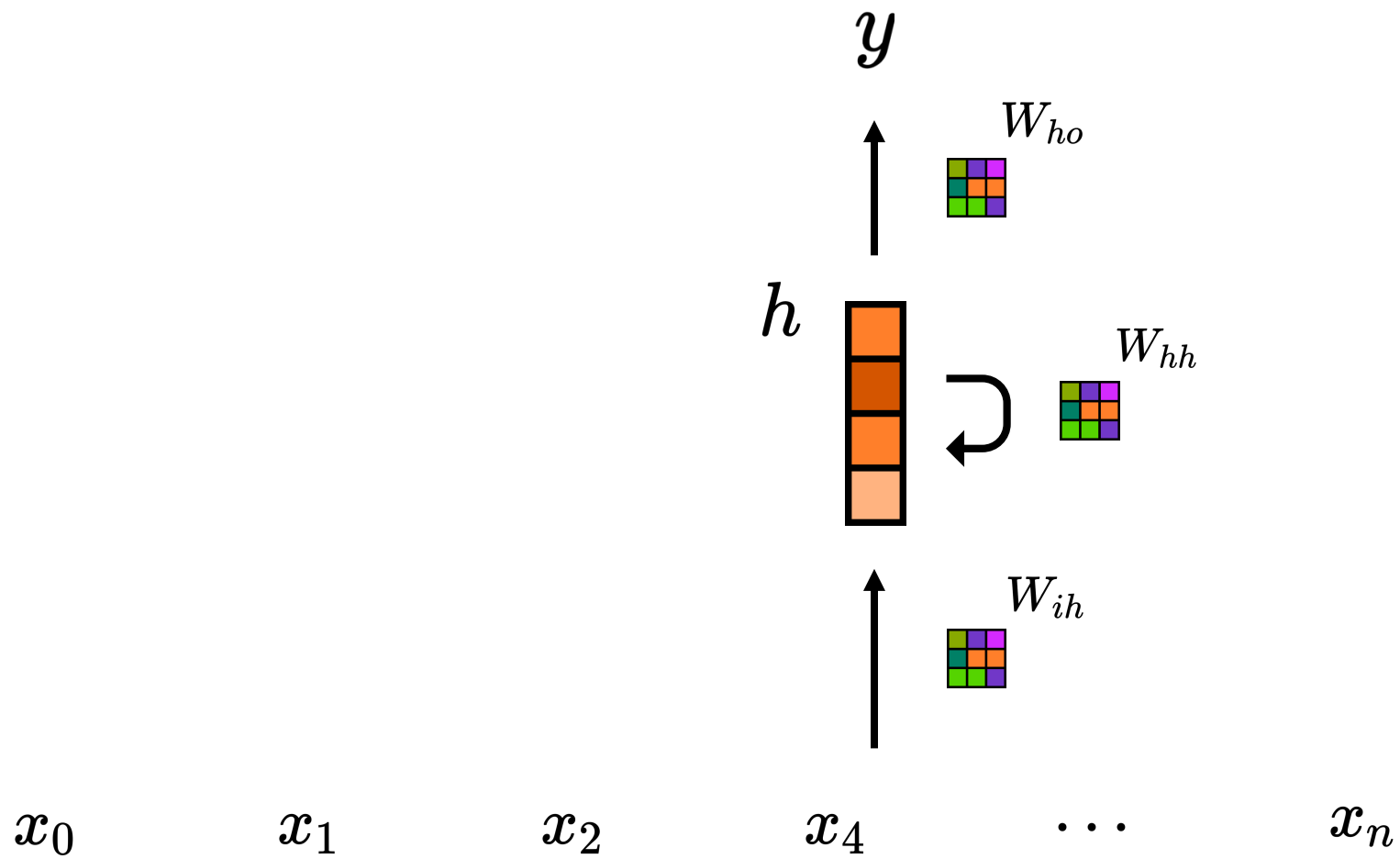
Neural networks: a primer



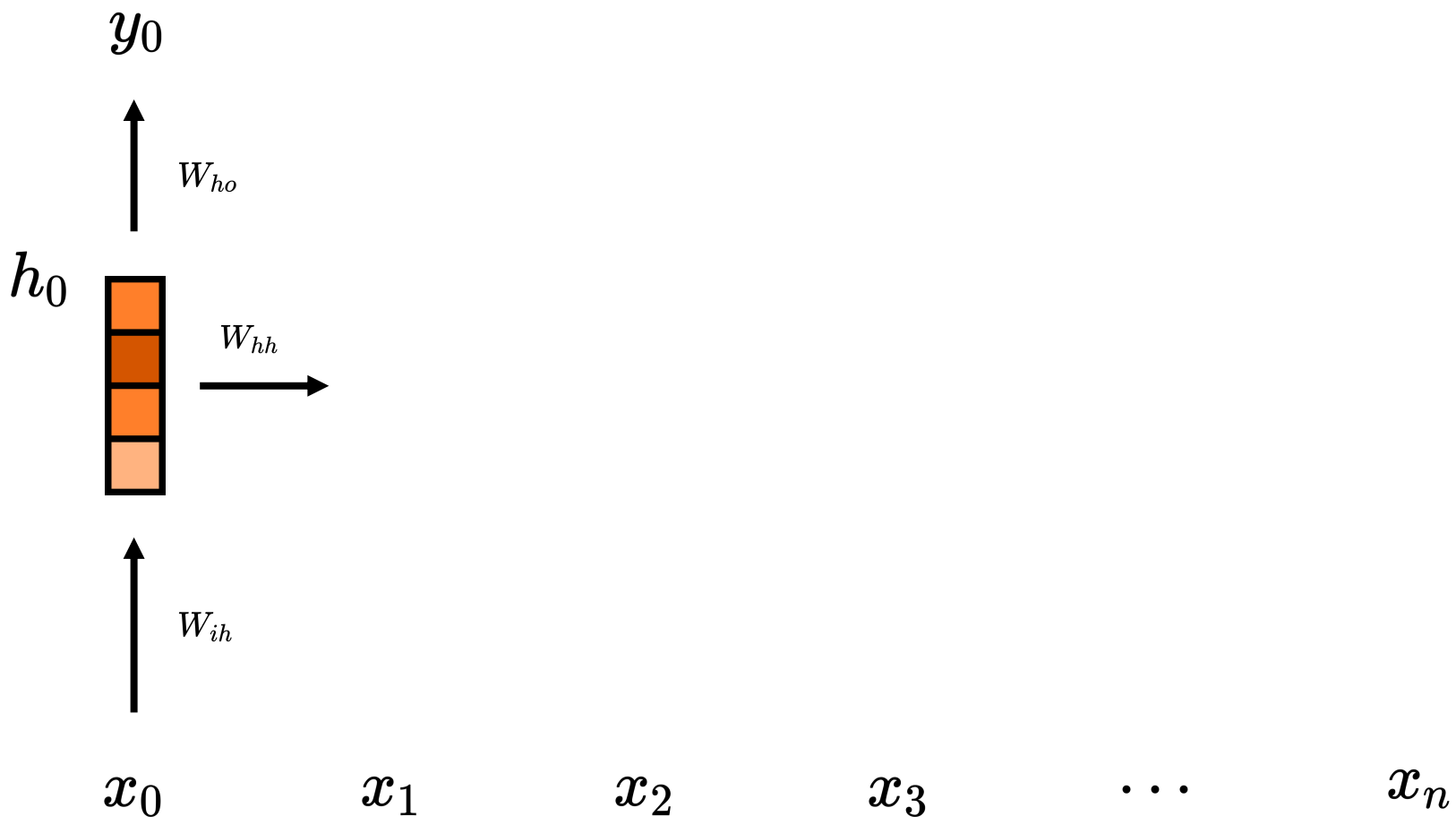
Recurrent Neural Networks



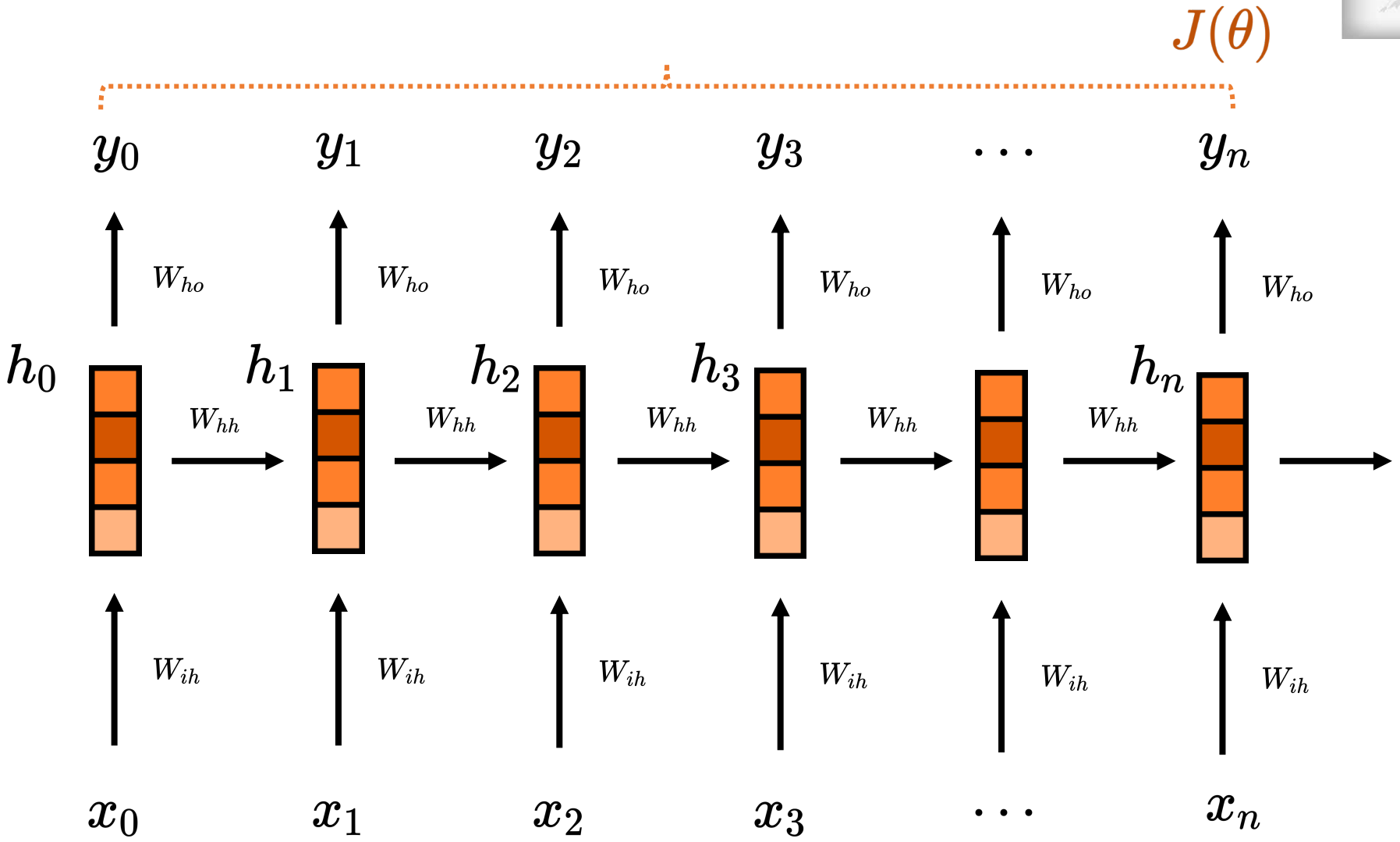
Recurrent Neural Networks



Recurrent Neural Networks



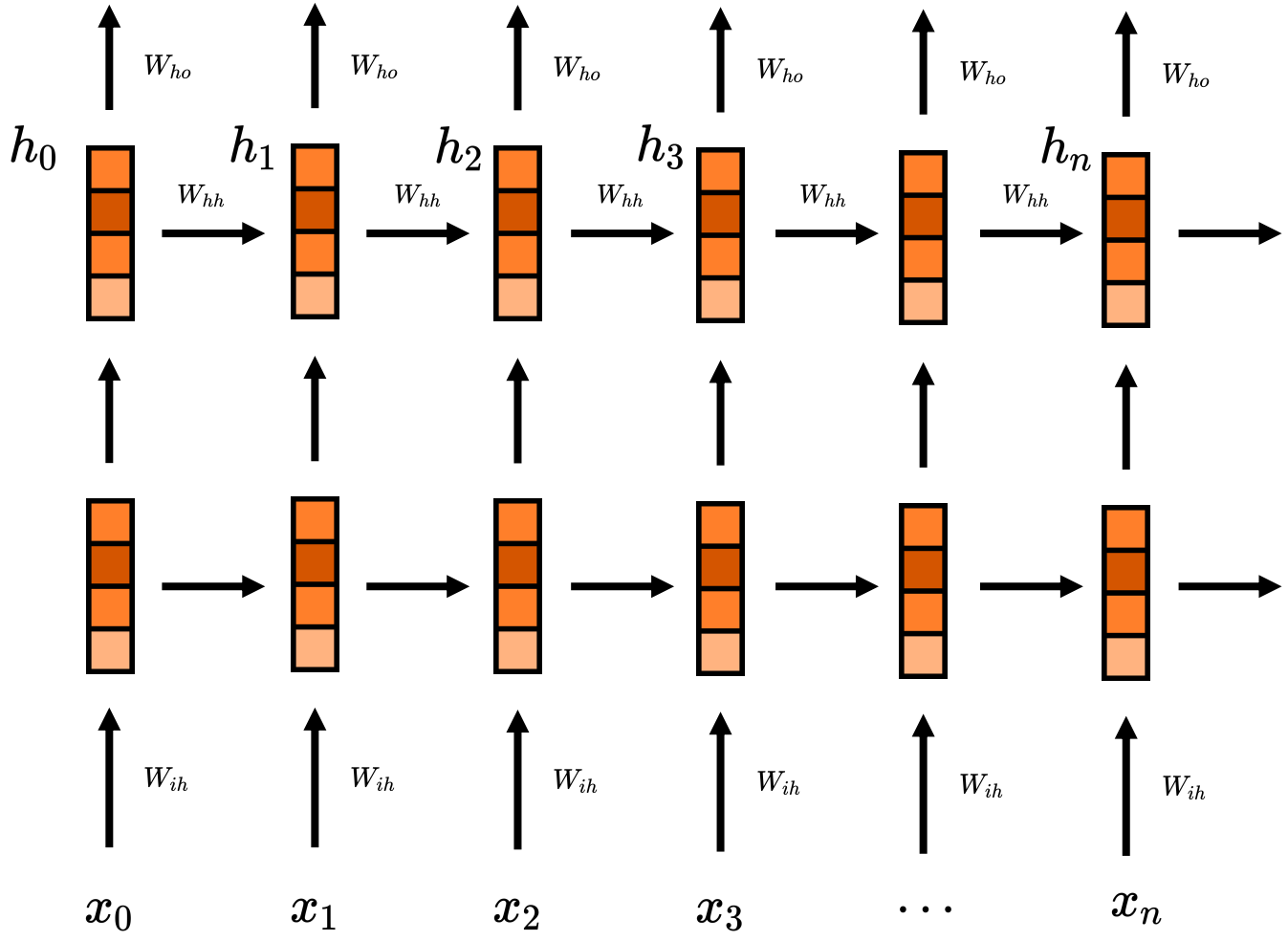
Recurrent Neural Networks



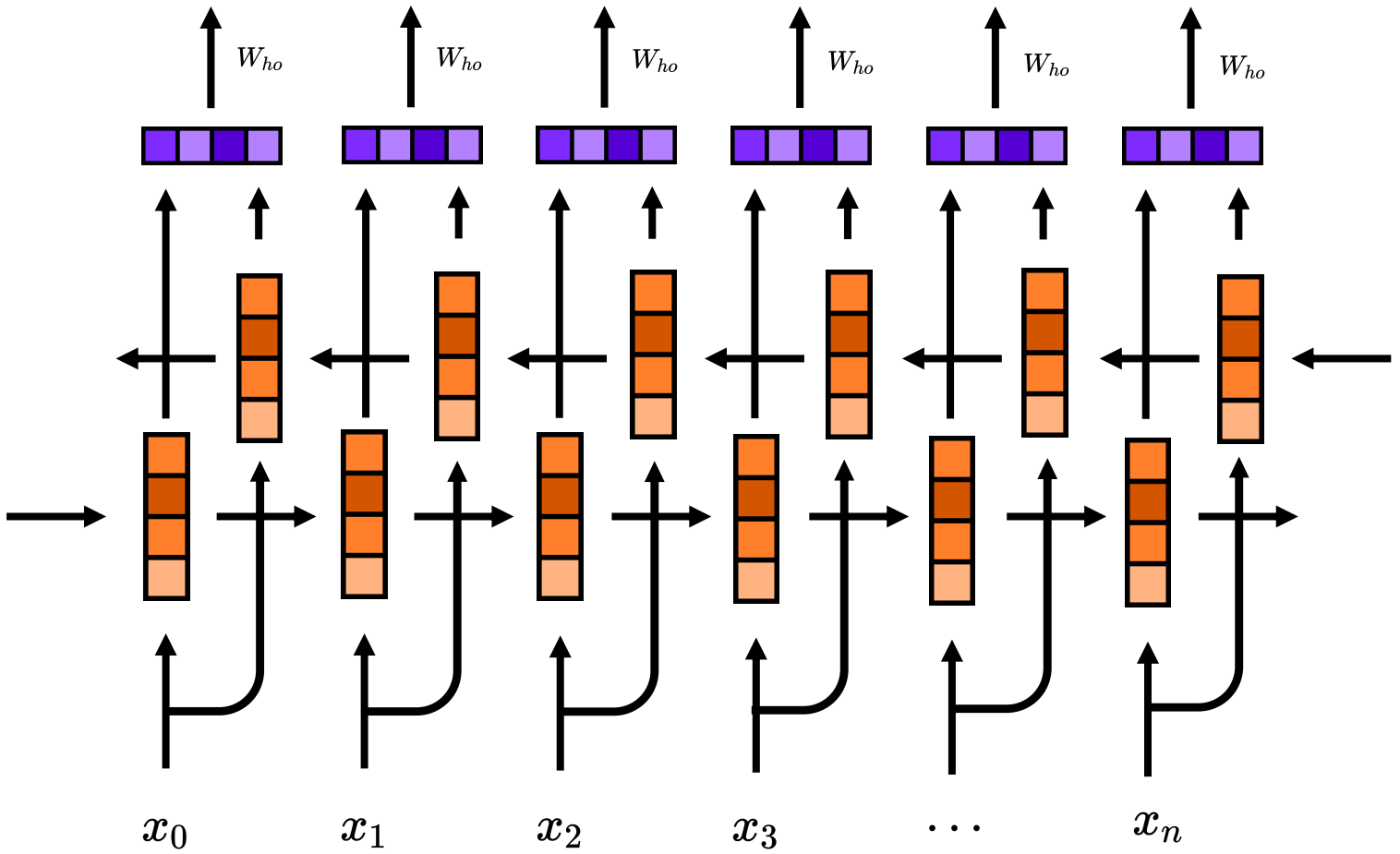
"Backpropagation through time"

A. Karpathy. The Unreasonable Effectiveness of Recurrent Neural Networks . https://karpathy.github.io/2015/05/21/rnn-effectiveness/

Stacked layers in RNNs



Right-to-left units in RNNs



Pros & Cons of RNNs

* Spoiler: we (almost) fix that with Gated RNNs

- Weights are shared across time
  - the number of parameters is low (3 matrices in Vanilla RNN)
  - all inputs get equal “treatment”
- They can handle sequences of arbitrary length
  - theoretically, each input “influences” all the future outputs no matter of the distance
- The architecture is flexible
  - We can stack layers or add a right-to-left flow
- Recurrence inhibits parallelization
- Although it’s there, the information flow gets cut by vanishing gradients*

Pascanu, R., Mikolov, T. and Bengio, Y., 2013, February. On the difficulty of training recurrent neural networks. In International conference on machine learning

Language modeling

- Model language entails predicting the next item (word or character), given a context.

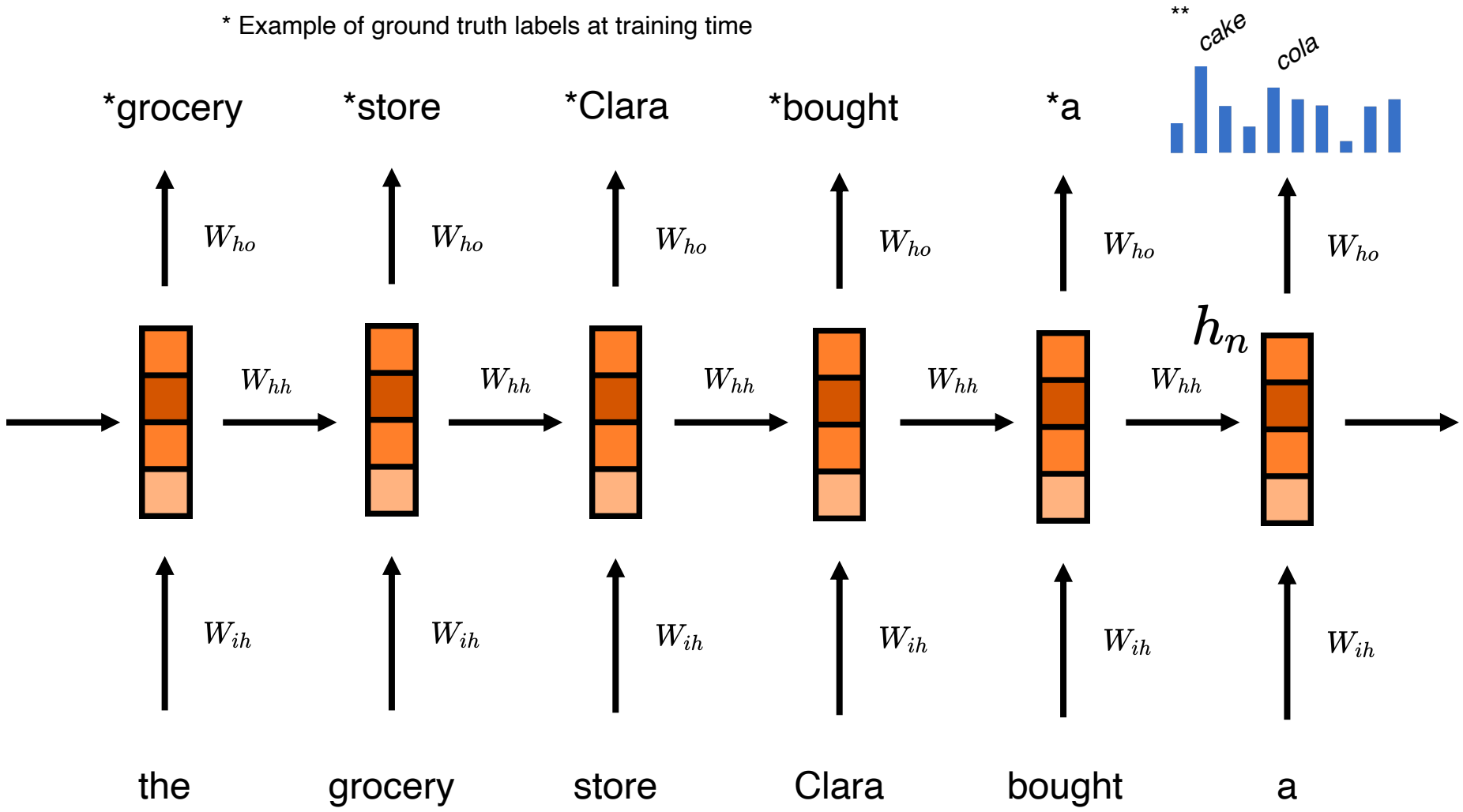
Back at the grocery store, Clara bought a _____

- "Grocery store"-related stuff should be more likely: we are modeling a probability!

RNNs for Language Modeling

** Example of PDF at inference time

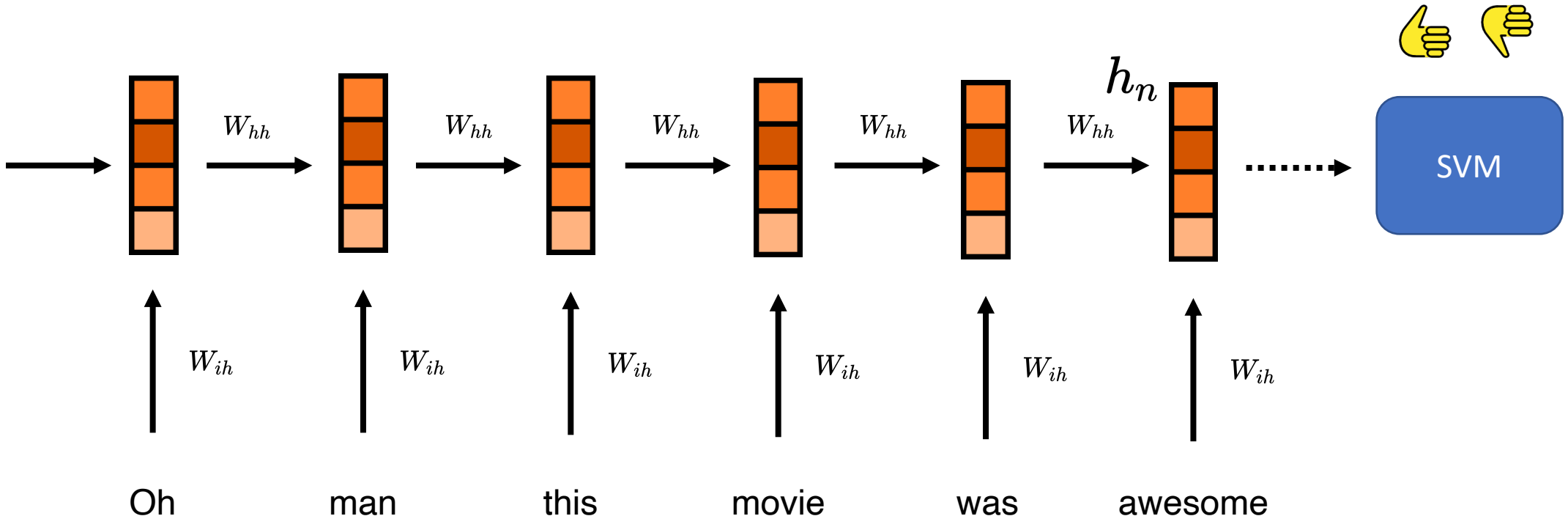
* Example of ground truth labels at training time



A. Karpathy. Minimal NumPy implementation. https://gist.github.com/karpathy/d4dee566867f8291f086

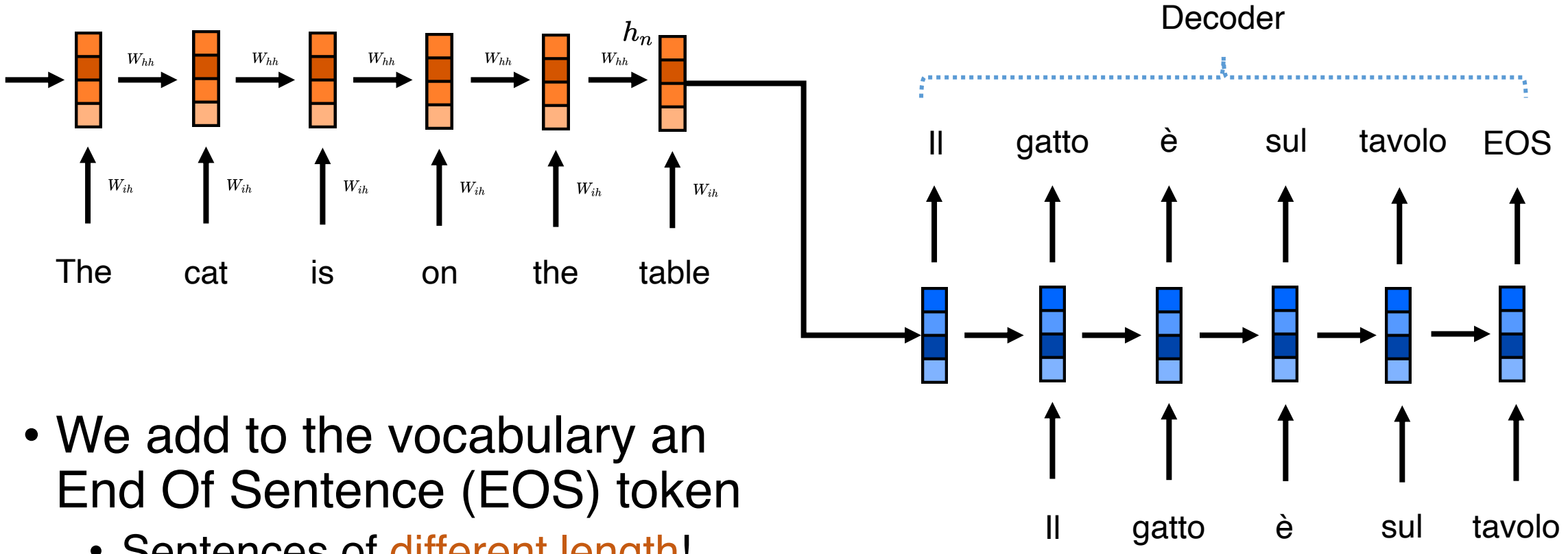
RNNs for Sentiment Analysis

- Generally, we can use the network as an "encoder" for further downstream tasks.



Actually, you can use *all* the hidden states (e.g. concatenating them).

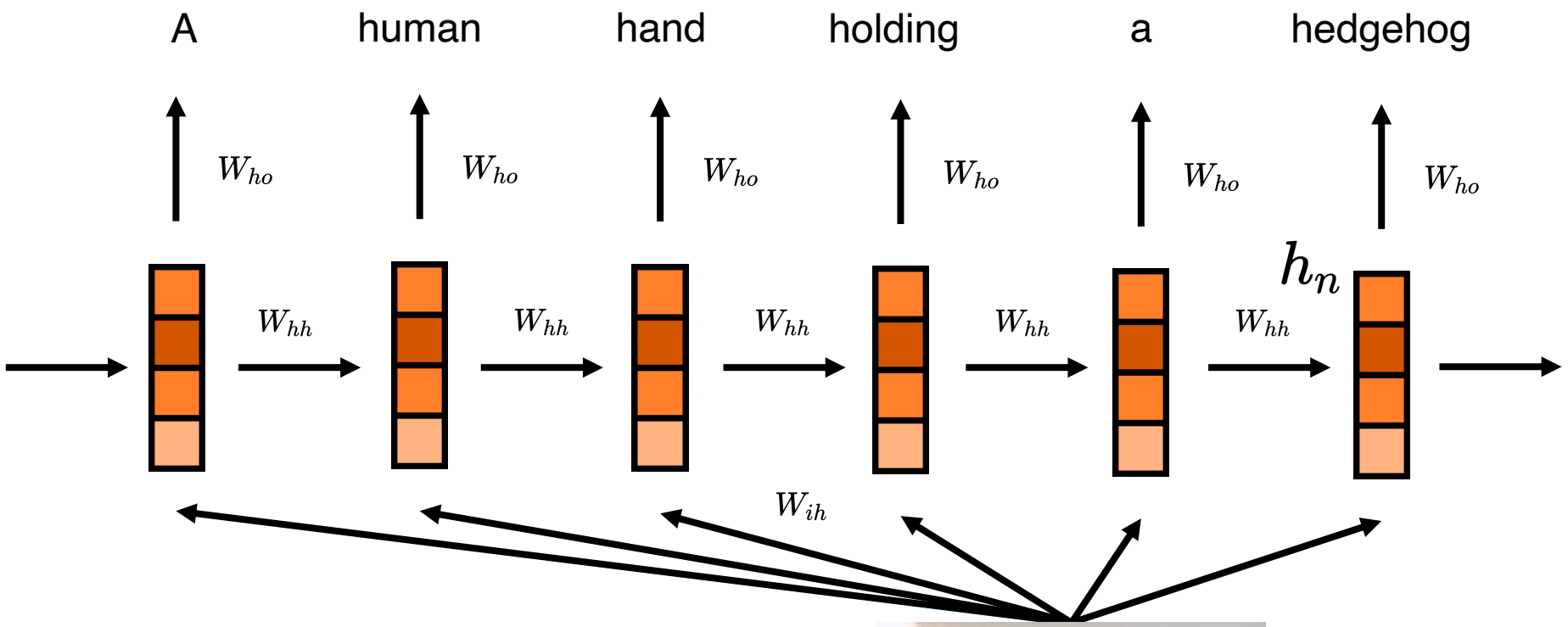
RNNs for Neural Machine Translation



- We add to the vocabulary an End Of Sentence (EOS) token
  - Sentences of different length!

Sutskever, I., Vinyals, O. and Le, Q.V., 2014. Sequence to sequence learning with neural networks. Advances in neural information processing systems

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.



RNNs for Image Captioning

https://www.reddit.com/r/aww/comments/ketvt3/may_i_offer_you_this_cute_hedgehog_in_these/

Generating Stories about Images

Generated story about image
Model: Romantic Novels

"He was a shirtless man in the back of his mind, and I let out a curse as he leaned over to kiss me on the shoulder.

He wanted to strangle me, considering the be-atiful boy I'd become wearing his boxers."

https://medium.com/@samim/generating-stories-about-images-d163ba41e4ed

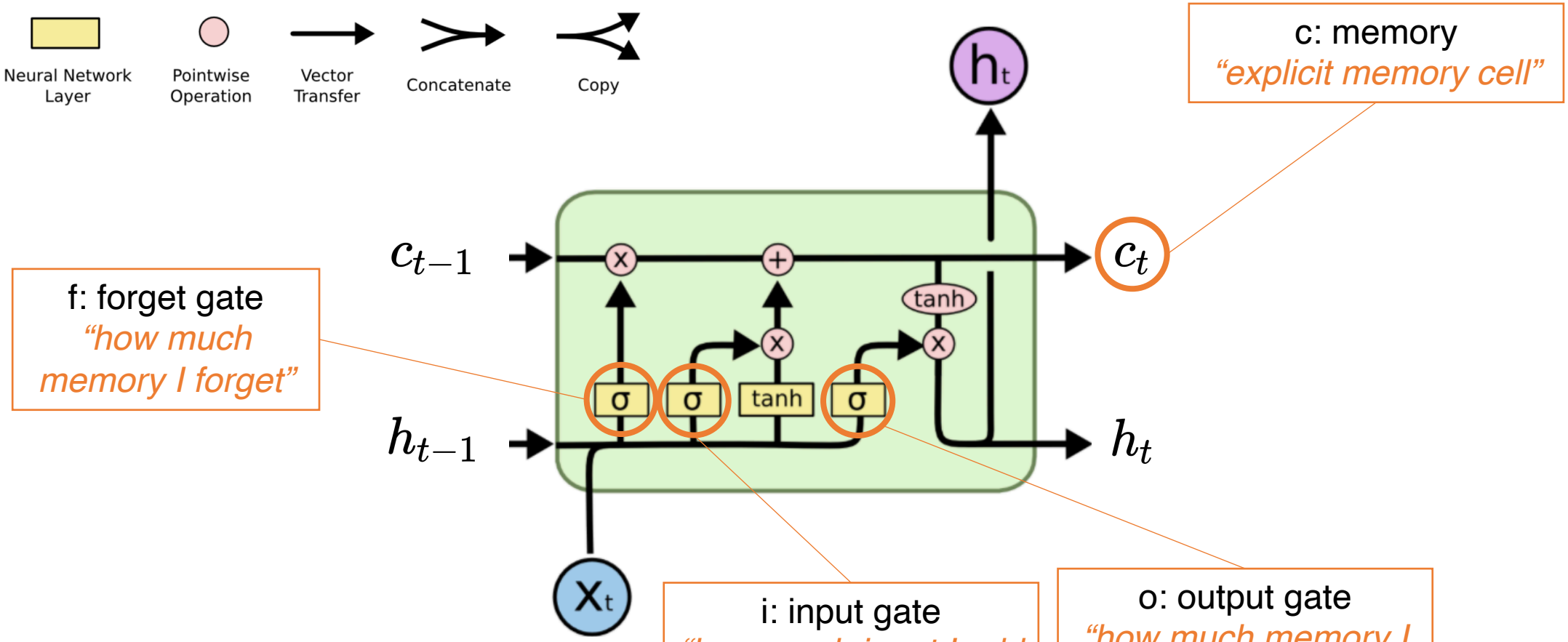
Gated RNNs

Yesterday I visited my grandma, and I brought there a bunch of stuff. Also, I installed that Alexa device as you asked. I have strong doubts that it will work, but when you're ready, we can try to video-call ____

- If the information flow gets cut by vanishing gradient
  - Add explicit memory
  - Let the network learn how to use it (i.e., read from / write to it)

- They idea of explicit memory and learned gates is dated 1997! Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. Neural computation, 9(8), pp.1735-1780.

Gated RNNs: LSTM



https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Photo by Samantha Gades on Unsplash

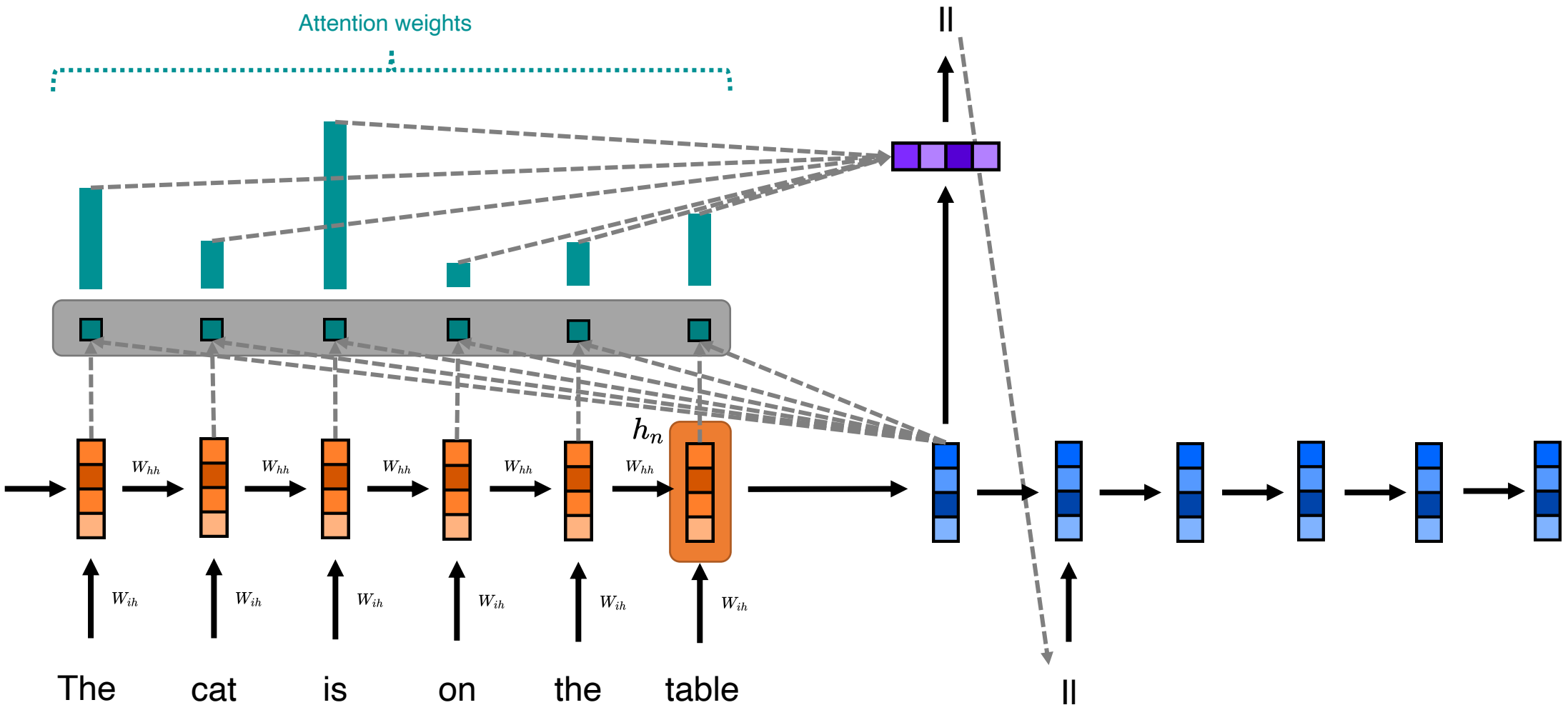
Attention [~2014-2016]

- Motivated by the human ability to focus on salient information and discard the rest
  - … or the Cocktail party problem

- A groundbreaking innovation in sequence modeling
- Innovative to the extent that temporal constraints get loose, if not discarded at all
- Core idea:

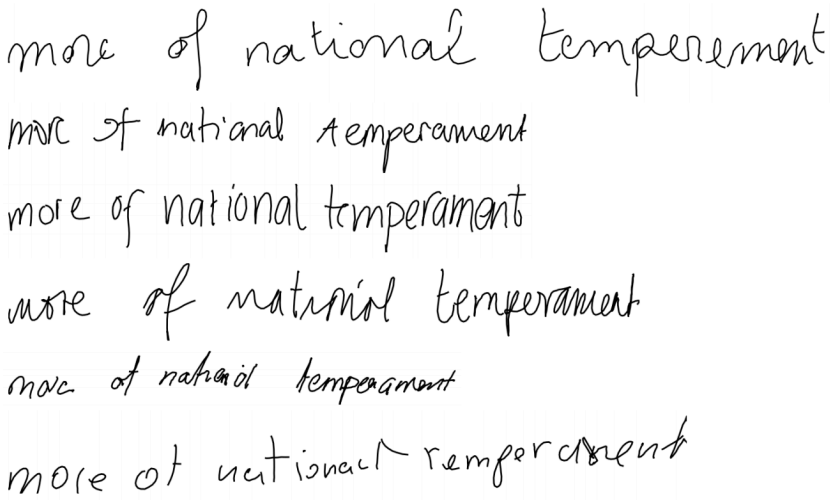
  we let the network learn how to discard information

RNNs for Neural Machine Translation (2)

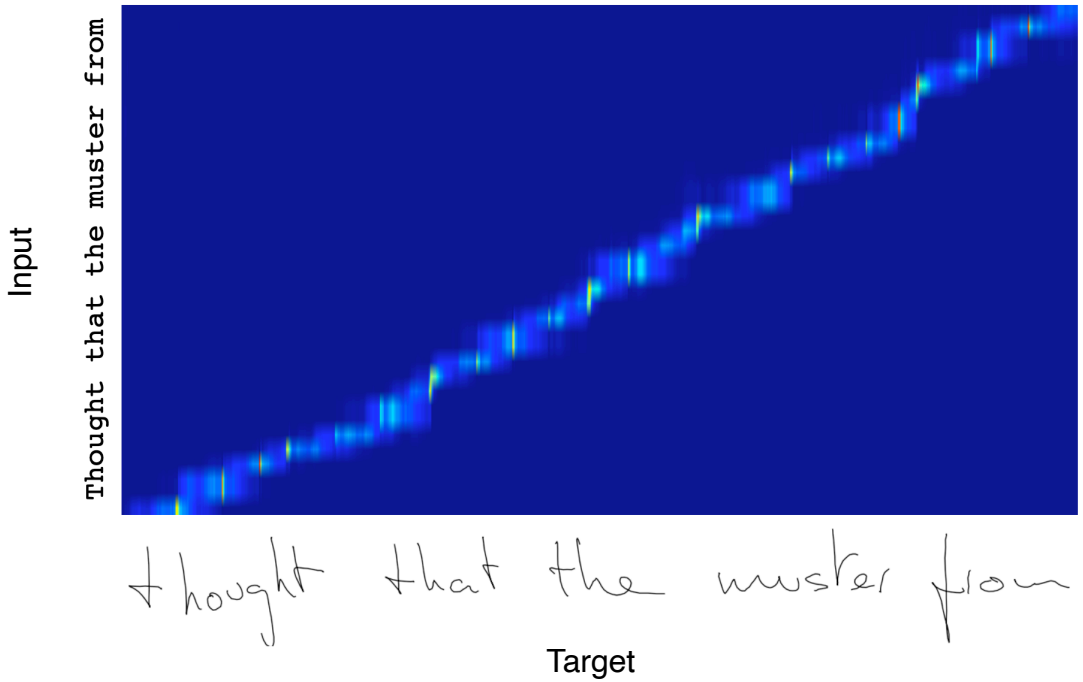


Generating sequences with RNNs

- Architecture: (custom) encoder-decoder Stacked LSTMs
- Task: generate handwriting corresponding to input text



The top line is real, the rest are samples from the decoder network



Graves, A., 2013. Generating sequences with recurrent neural networks.

Attention [2016-today)

- Attention is all you need. Waswani et al.
- The paper introduces the Transformer
  - No more recurrent units
  - Encoder-decoder architecture
  - A clever associative attention is used
- Building block of renowed language models
  - BERT, XLNet, GPT-(1|2|3), T5, and more
- Sadly, there's no room for that in the talk

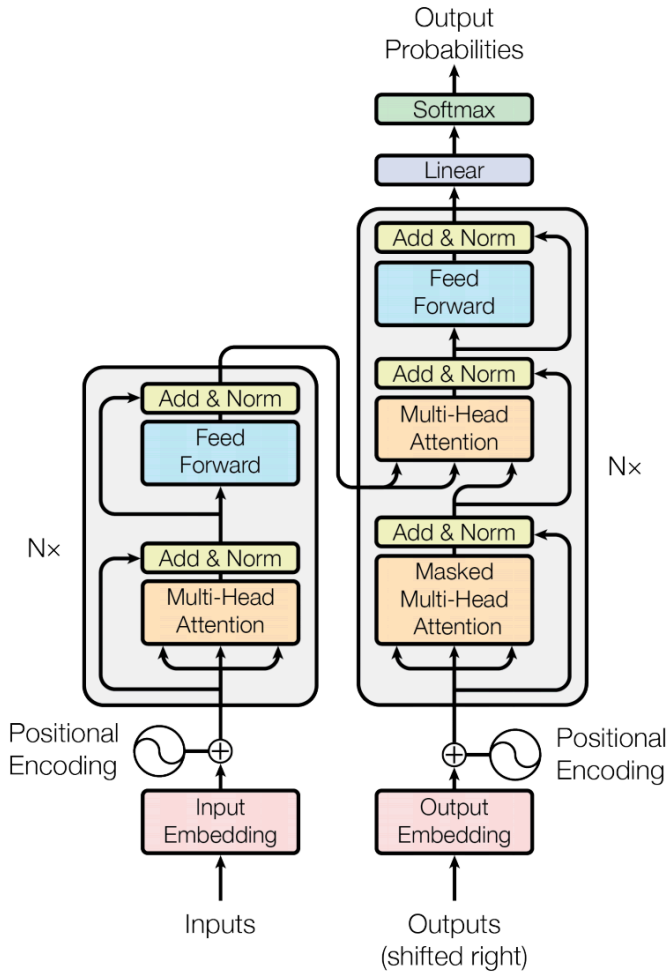


Figure 1: The Transformer - model architecture.

There’s more

- Attention as explanation
  - There is a huge debate on that
- Transformer as enabler for new neural architectures
  - Lambda Networks (under peer review)
  - Visual Transformers: Token-based Image Representation and Processing for Computer Vision
  - An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (under peer review)

Thank for your attention!

 giuseppe.attanasio@polito.it

 gattanasio.cc

 @peppeatta