

From Recurrent Models to the advent of Attention: a recap

Giuseppe Attanasio



**Politecnico
di Torino**

Data Science Lab: process and methods

Research Bites | January 13, 2022

From Recurrent Models to the advent of Attention: a recap



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

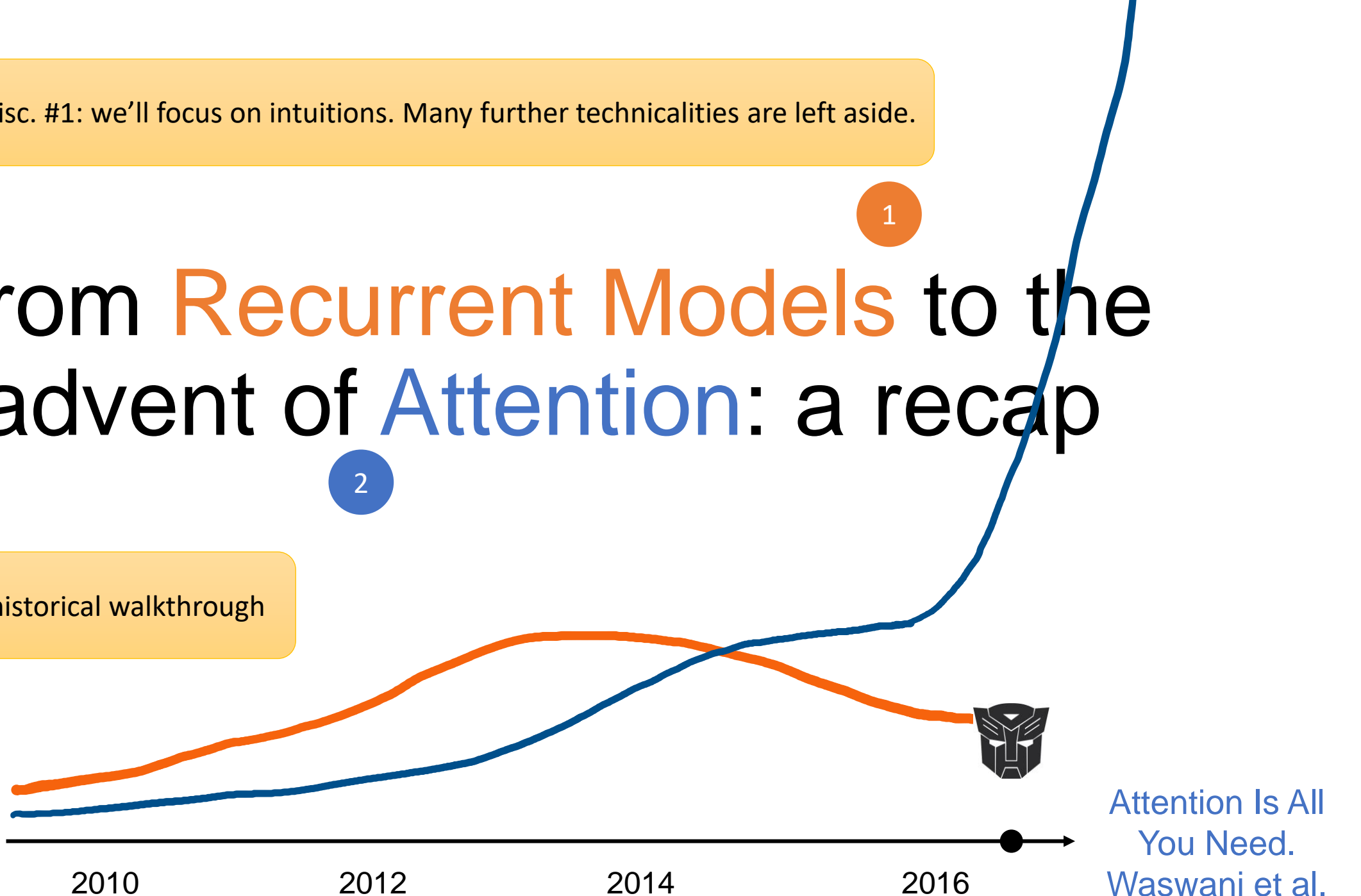
1

From Recurrent Models to the advent of Attention: a recap

2

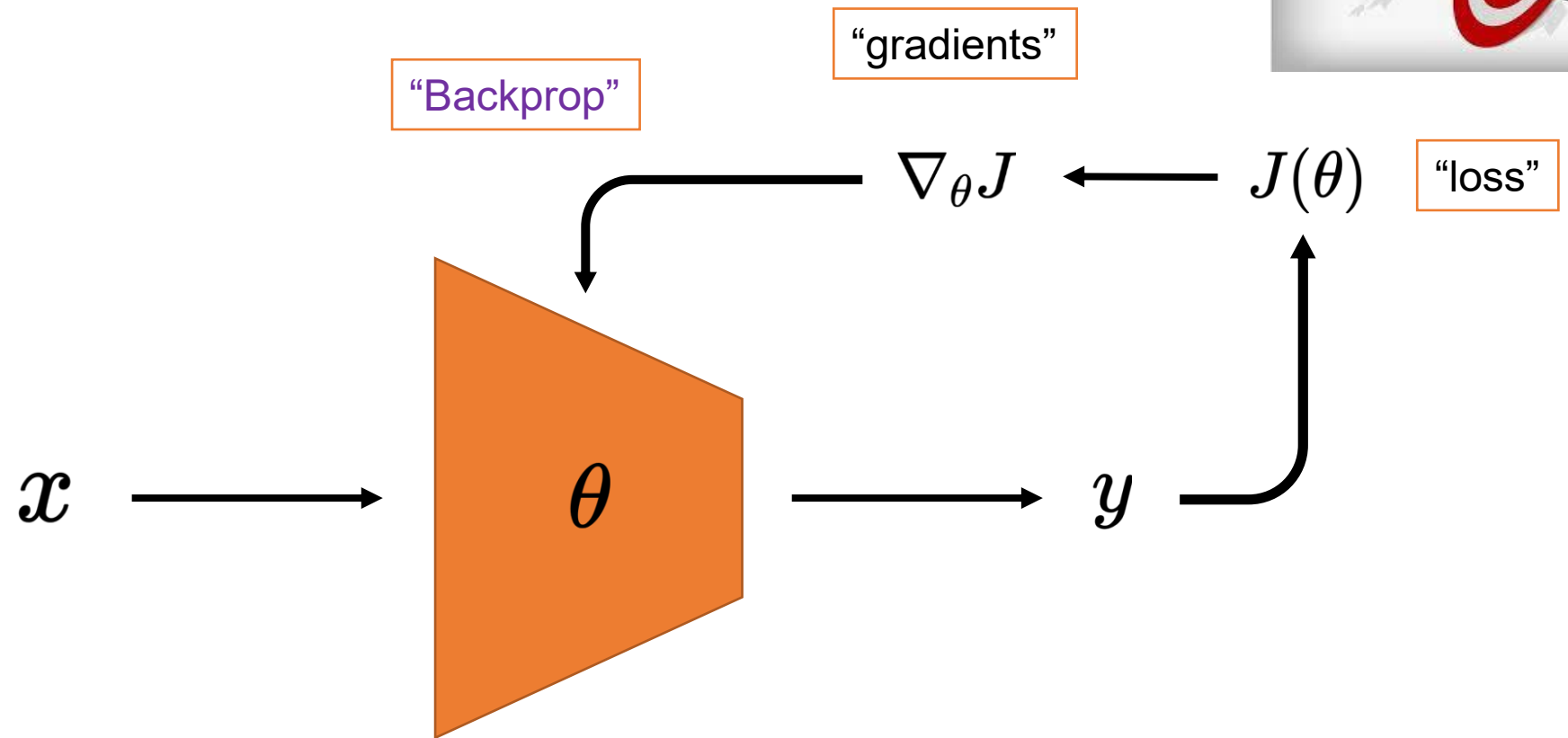


Disc #2: an NLP historical walkthrough

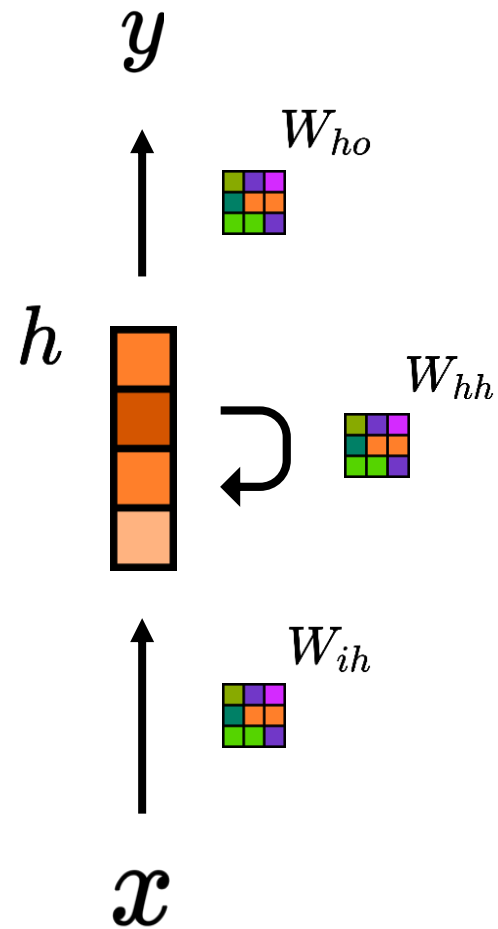


Attention Is All
You Need.
Waswani et al.

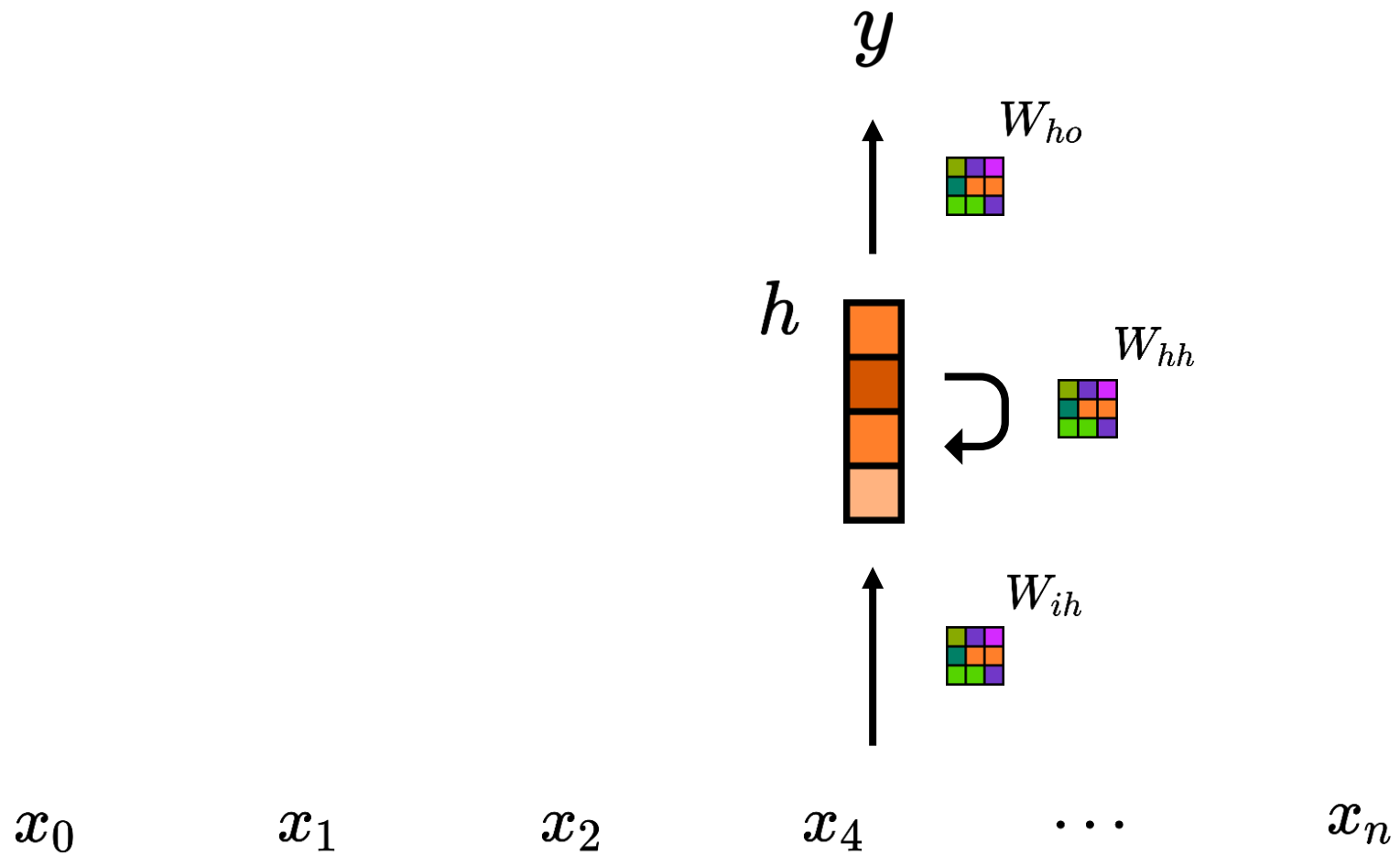
Neural networks: a primer



Recurrent Neural Networks



Recurrent Neural Networks



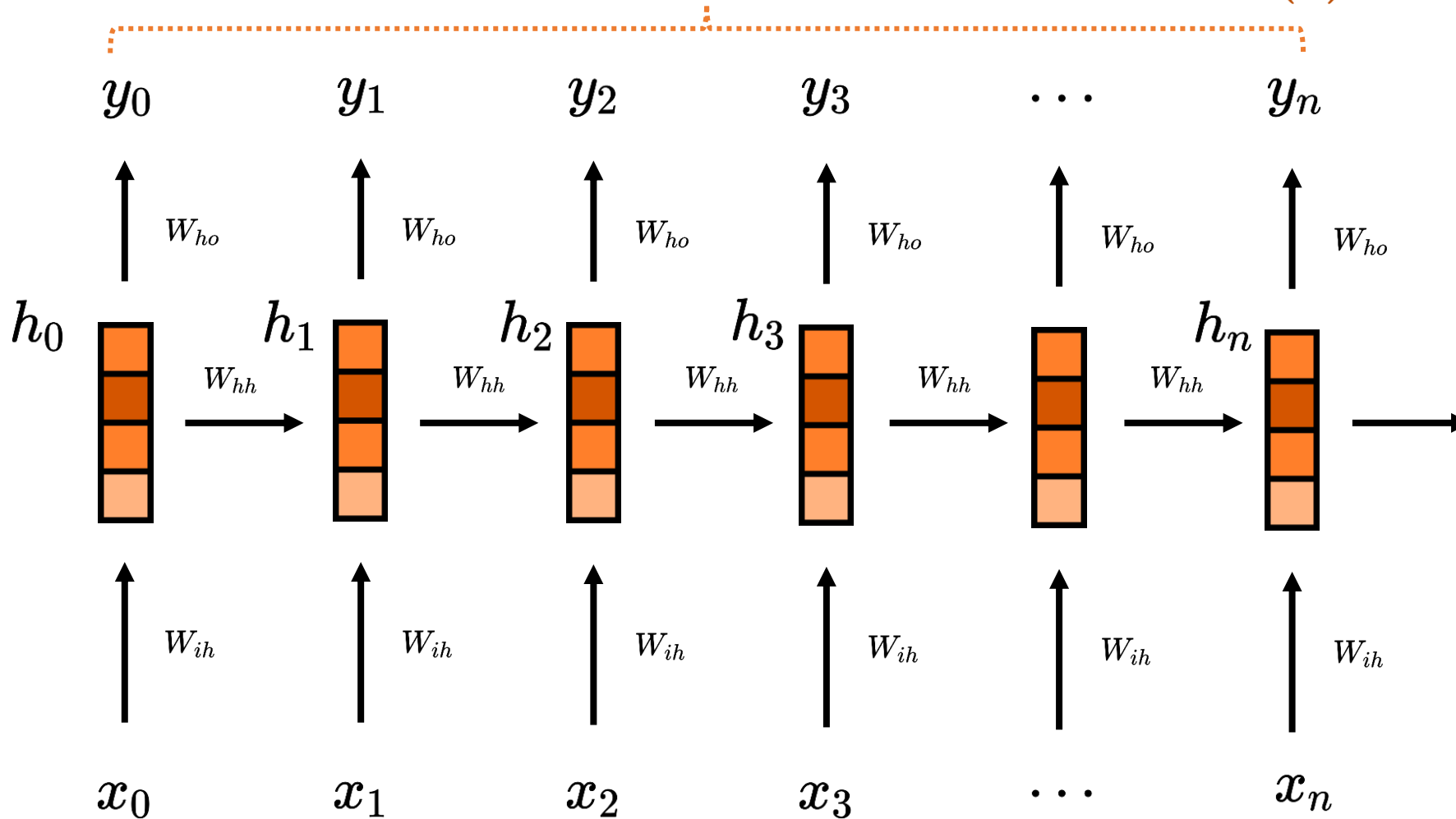
Recurrent Neural Networks



Recurrent Neural Networks

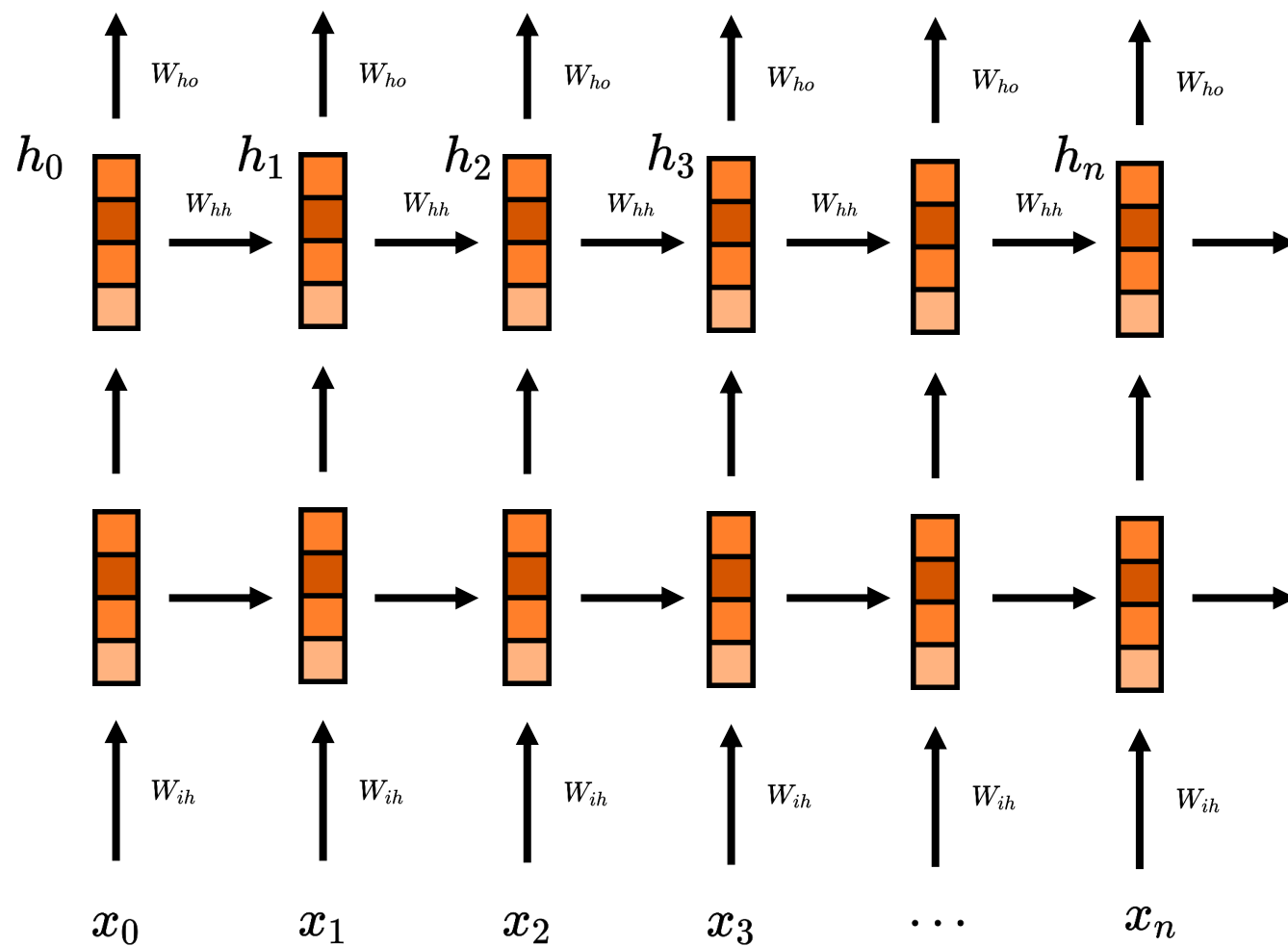


$J(\theta)$

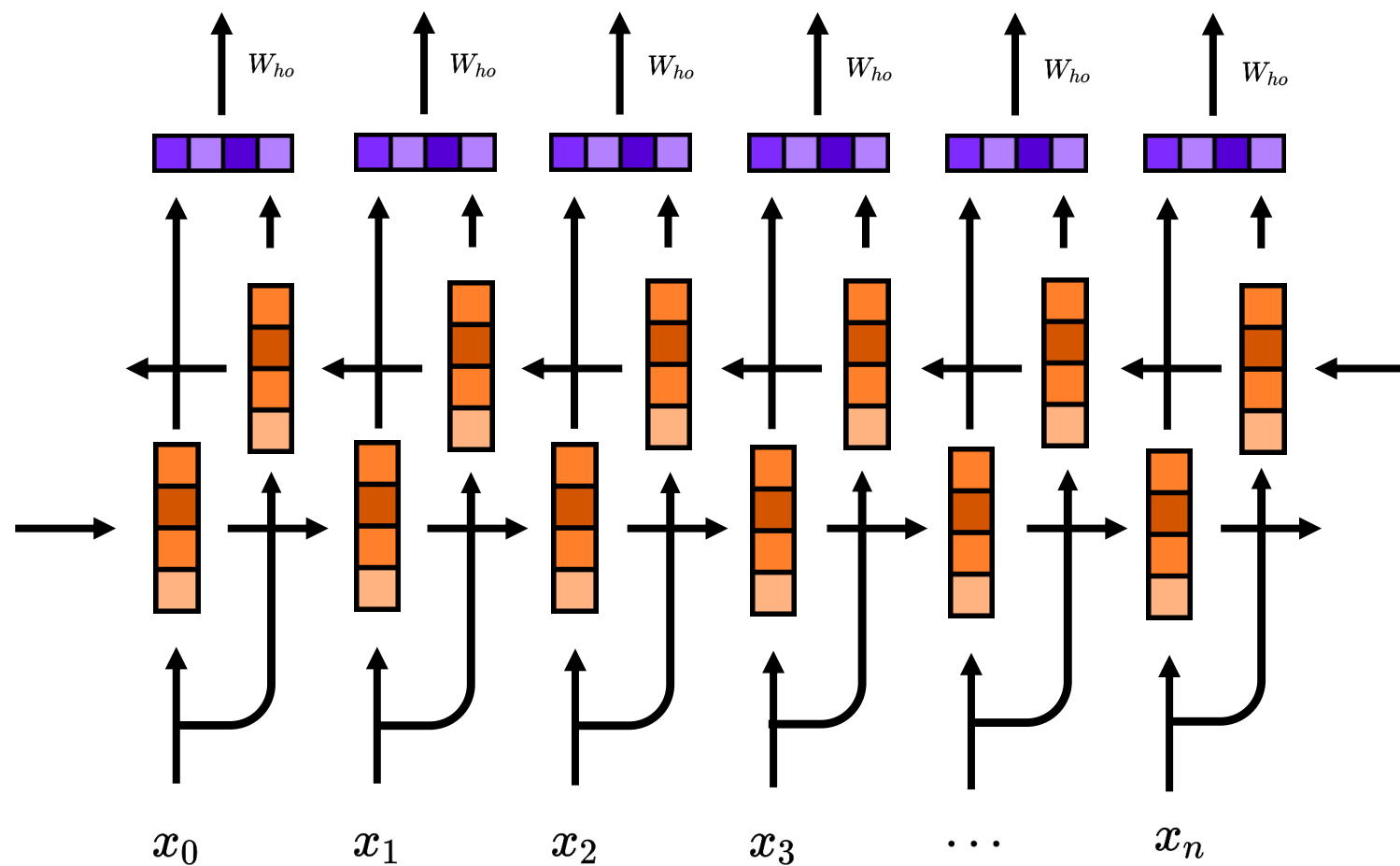


"Backpropagation through time"

Stacked layers in RNNs



Right-to-left units in RNNs



```
import torch
```

```
input_size = 8
```

```
hidden_size = 16
```

```
num_layers = 2
```

```
rnn = torch.nn.RNN(input_size=input_size, hidden_size=hidden_size, num_layers=num_layers)
```

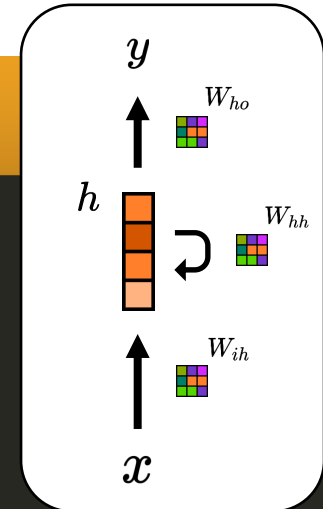
```
# Define input and initial hidden
```

```
in_seq = torch.randn((5, 1, input_size)) # sequence of 5 items
```

```
h0 = torch.randn((num_layers, 1, hidden_size)) # one initial hidden per layer
```

```
# Compute "one step"
```

```
yn, hn = rnn(in_seq, h0)
```



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***

Language modeling

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

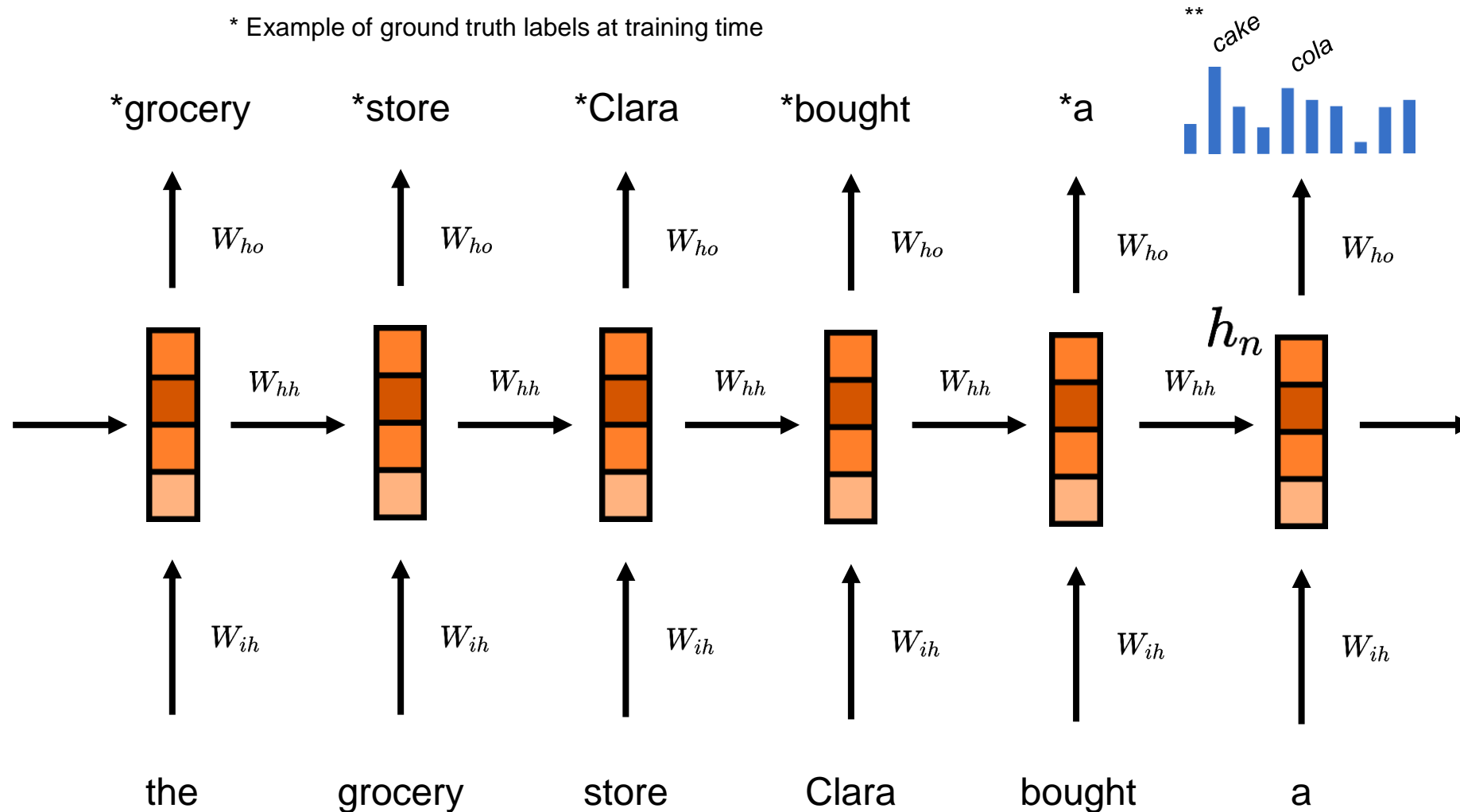
Back at the grocery store, Clara bought a _____

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

RNNs for Language Modeling

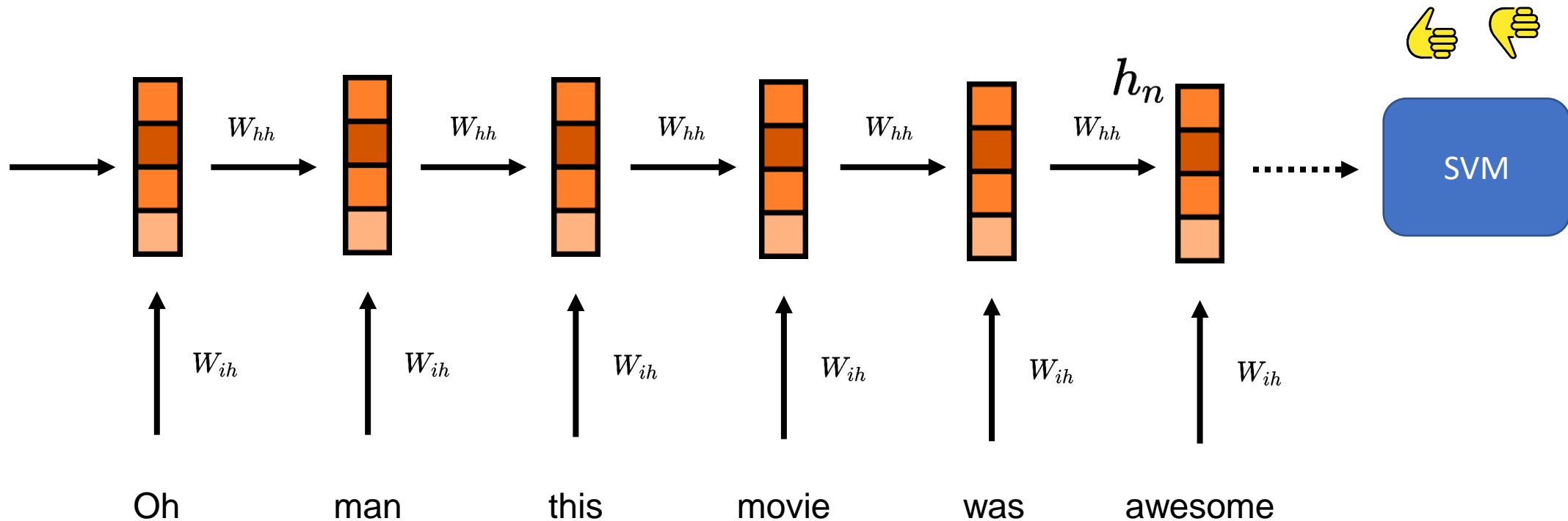
** Example of PDF at inference time

* Example of ground truth labels at training time



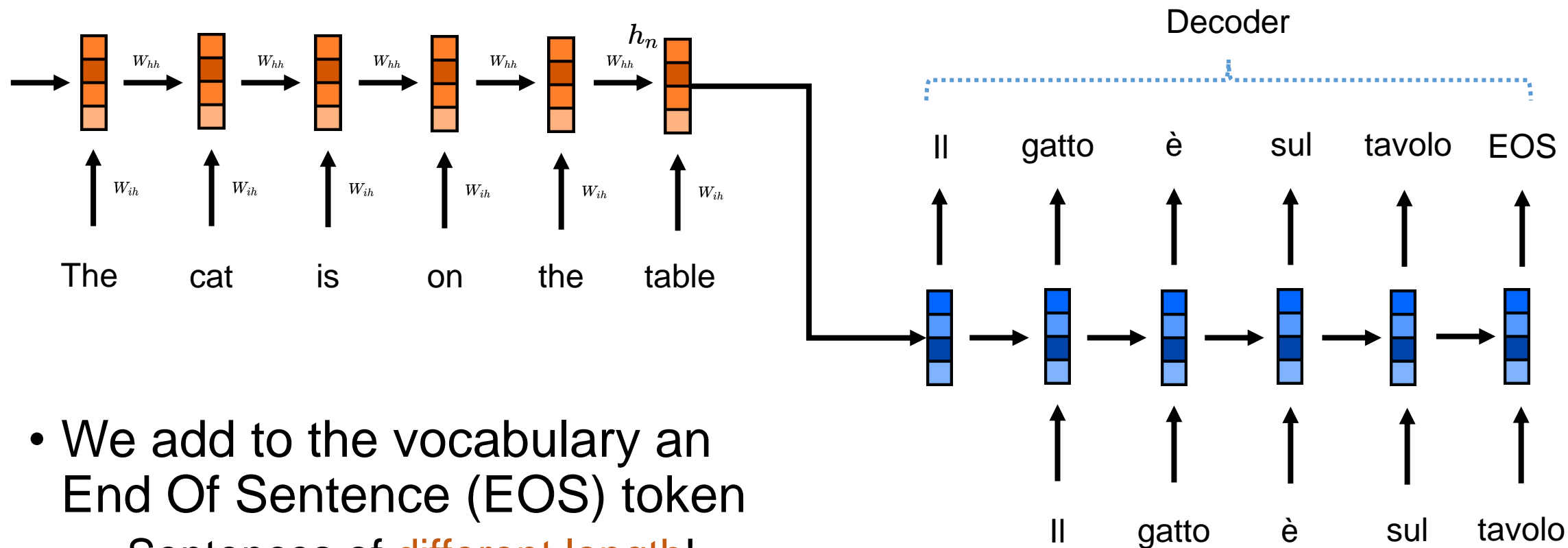
RNNs for Sentiment Analysis

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



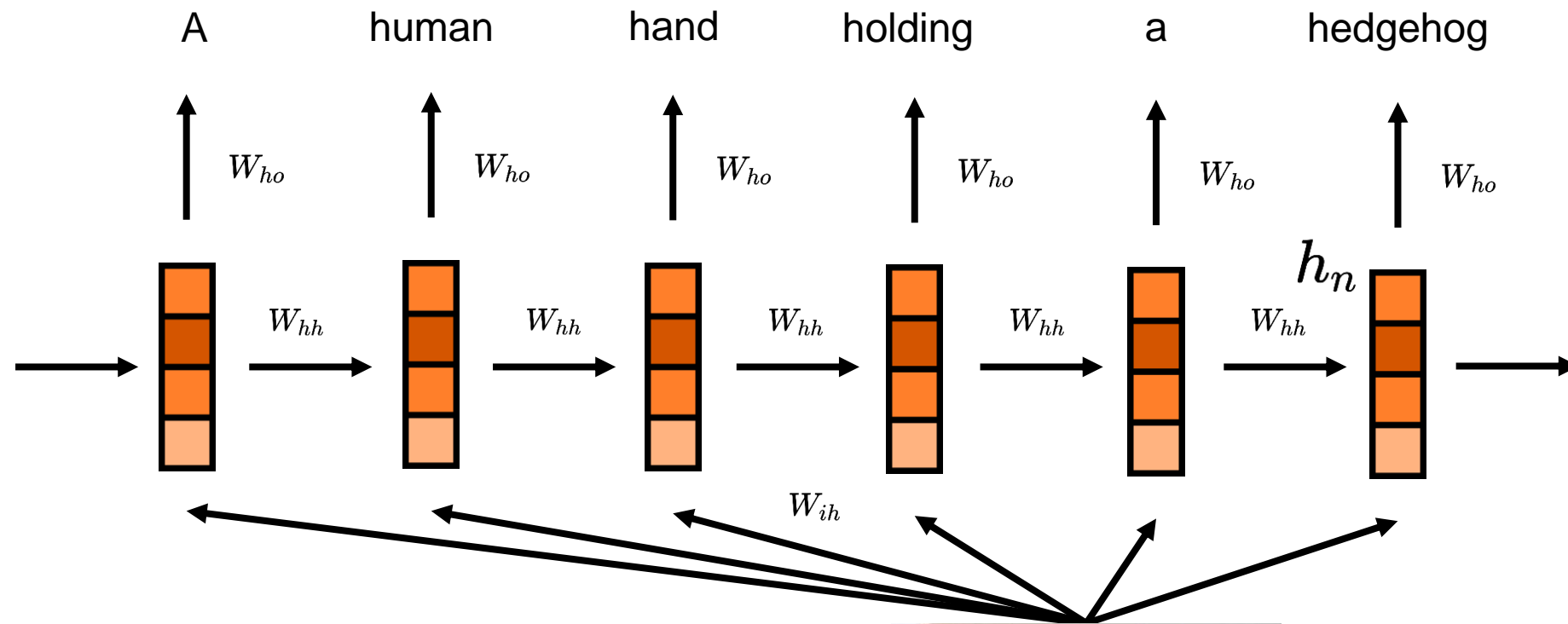
Actually, you can use *all* the hidden states (e.g. concatenating them).

RNNs for Neural Machine Translation



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



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.”

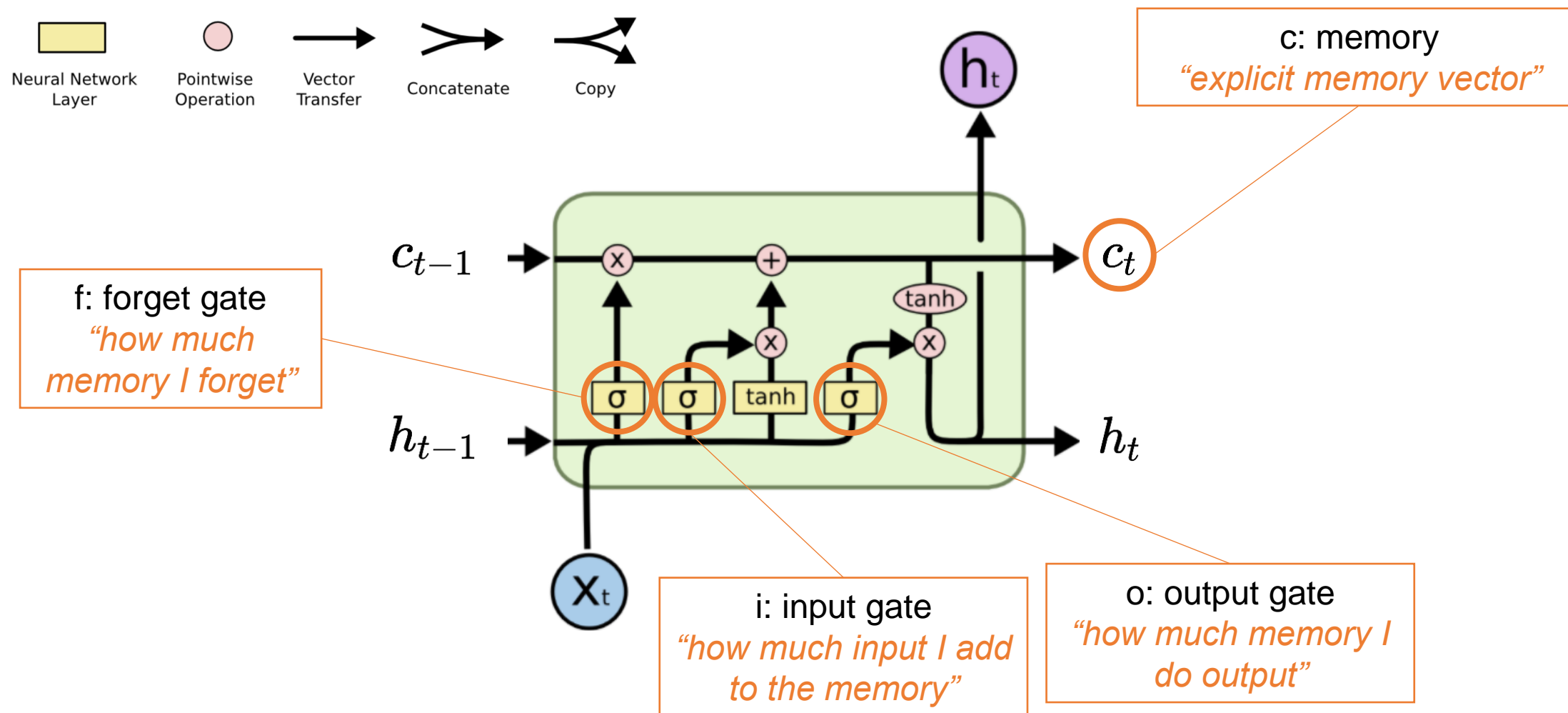
He wanted to strangle me, considering the beautiful boy I’d become wearing his boxers.”

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)
- The 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



A large crowd of people is gathered at an outdoor event at night. The scene is illuminated by warm string lights hanging from a wooden structure above. The crowd is diverse, with people of various ages and ethnicities. In the foreground, a man wearing a red baseball cap and a green and white striped shirt is visible. The word "Attention" is overlaid in large white text in the center of the image.

Attention

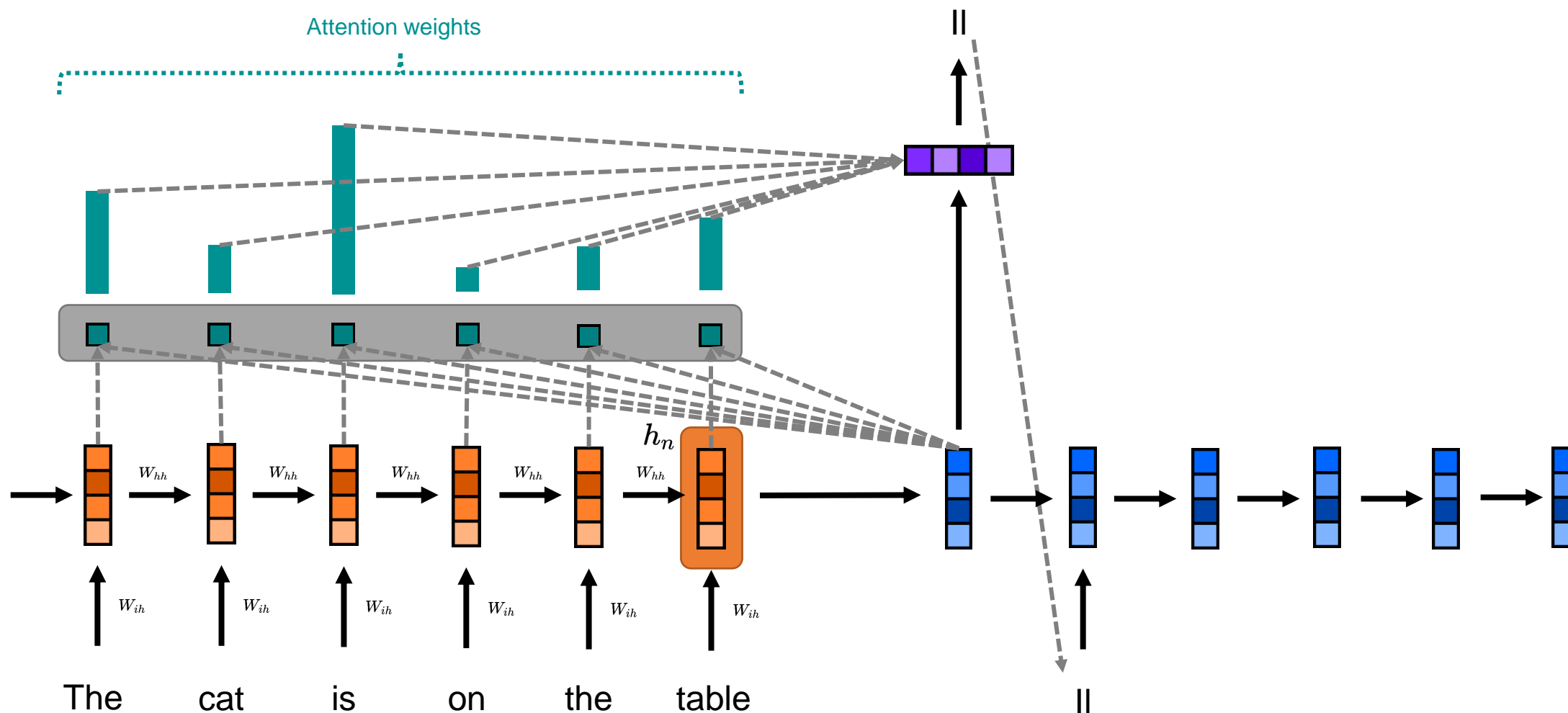
Attention [~2014-2016]

- 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)

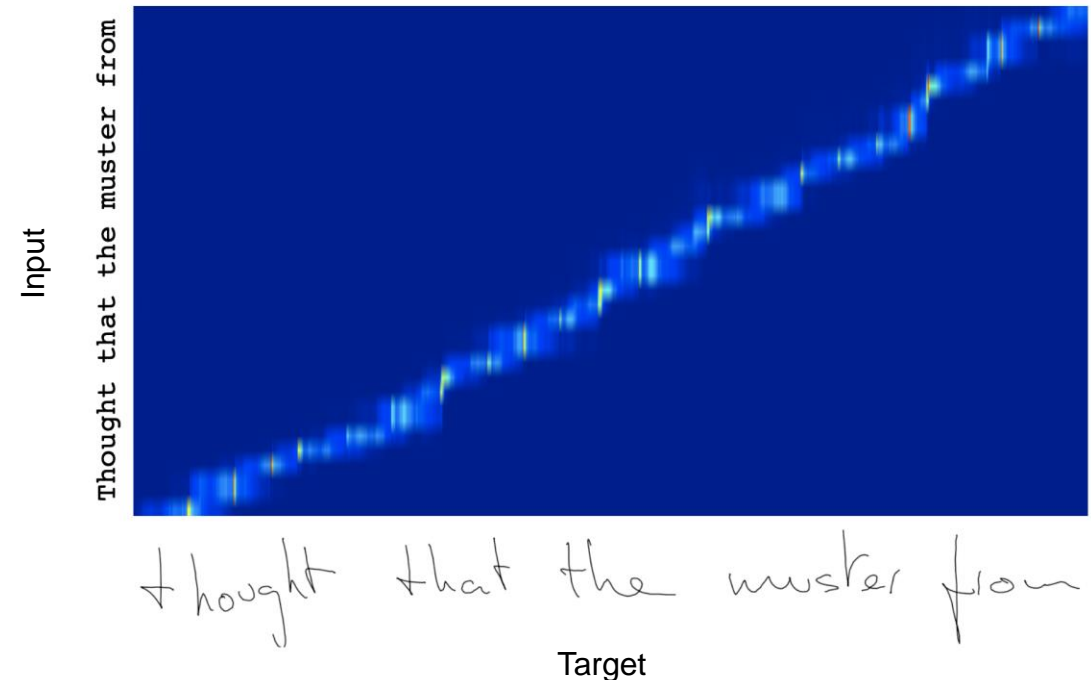


Generating sequences with RNNs


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

more of national temperament
more of national temperament
more of national temperament
more of national temperament
more of national temperament
more of national temperament

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



Attention [2016-today]

- **Attention Is All You need.** Waswani et al.
- The paper introduces **the Transformer** 
 - No more recurrent units
 - Learns sequence meaning using **attention only**
- Building block of modern language models
 - BERT, XLNet, GPT-*, T5, Megatron, ...

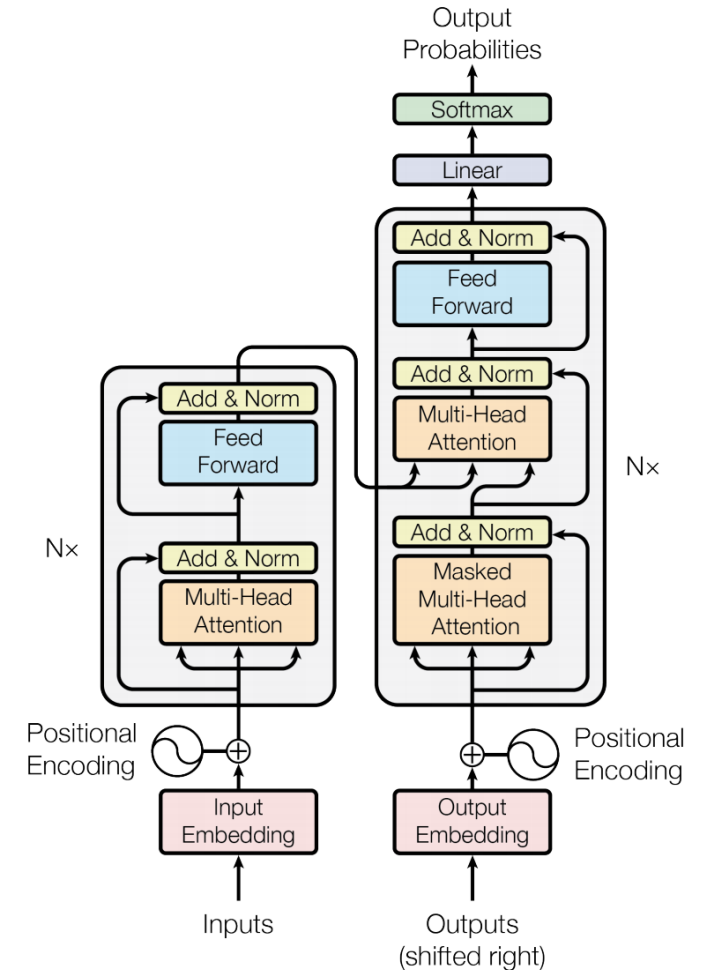


Figure 1: The Transformer - model architecture.

There's more

- Attention as **explanation**
- Transformers in new domains
 - Google's Vision Transformer (ViT) for Vision
 - Meta's XLS-R model for Speech
- Efficient Transformers (attention scales $O(n^2)$)
 - Google's FNet
 - Deed Mind's Perceiver

Thank for your attention!

 giuseppe.attanasio@polito.it

 gattanasio.cc

 @peppeatta