From Recurrent Models to the advent of Attention

A Recap

Giuseppe Attanasio

Papers We Love Milano. February 15, 2023

Hey there!

- Postdoc @ MilaNLP, Bocconi, Milano
- Studying Transformers to
  - Improve Hate Speech Detection
  - Interpret their decision process
  - Bridge vision and language worlds

gattanasio.cc
@peppeatta

ChatGPT

BERT

Transformer

Attention

Recurrent Neural Network

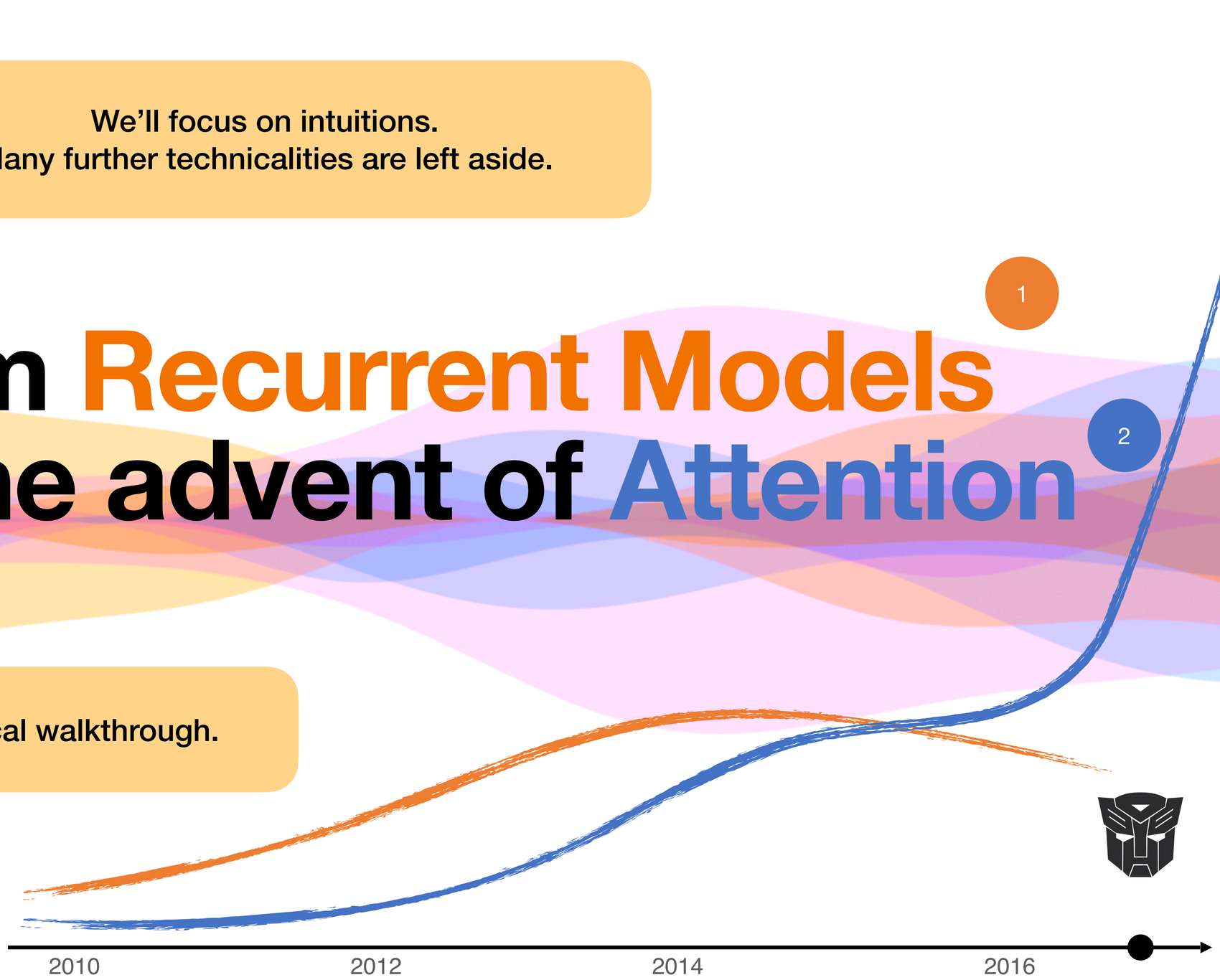
We'll focus on intuitions.
Many further technicalities are left aside.

From Recurrent Models to the advent of Attention

1

2

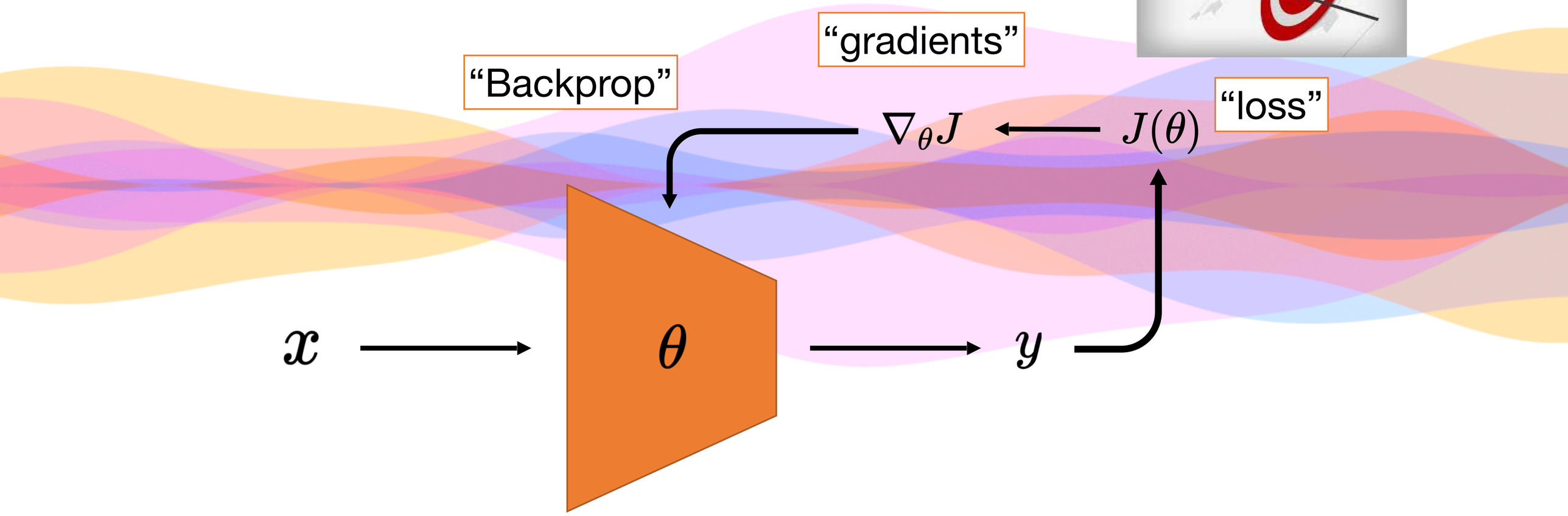
An NLP historical walkthrough.



Attention Is All You Need.
Waswani et al.

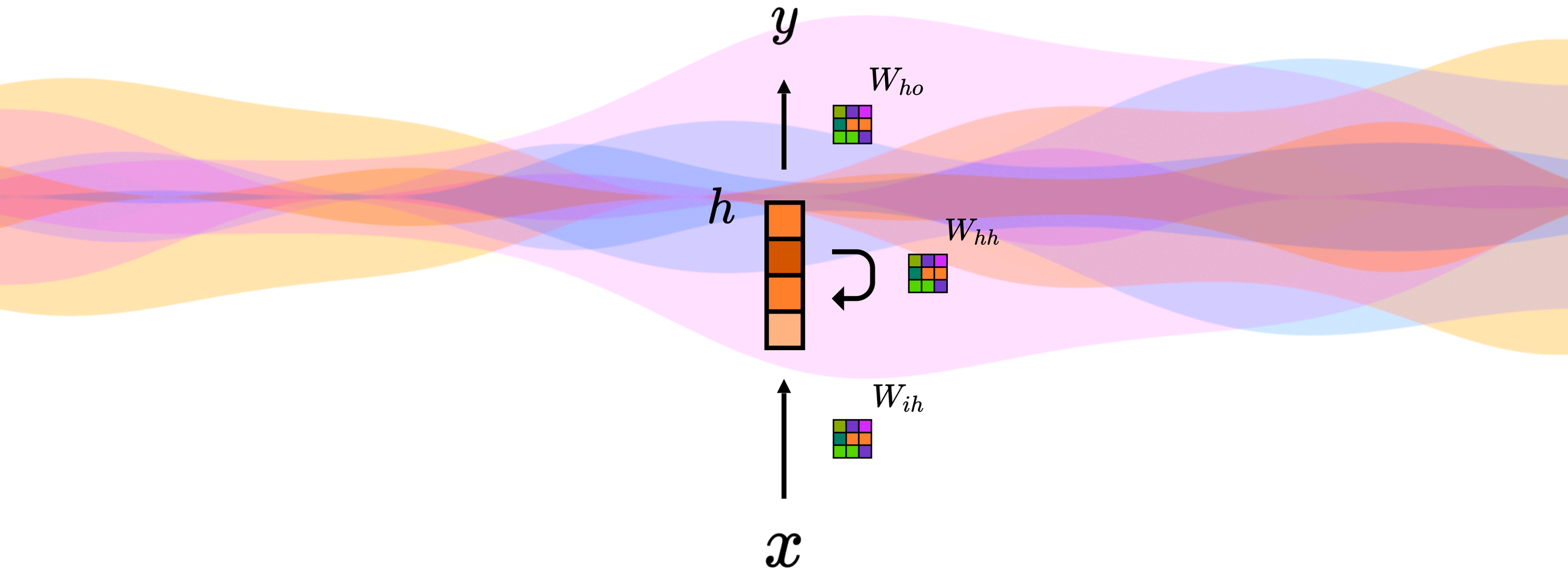
Neural networks

A primer

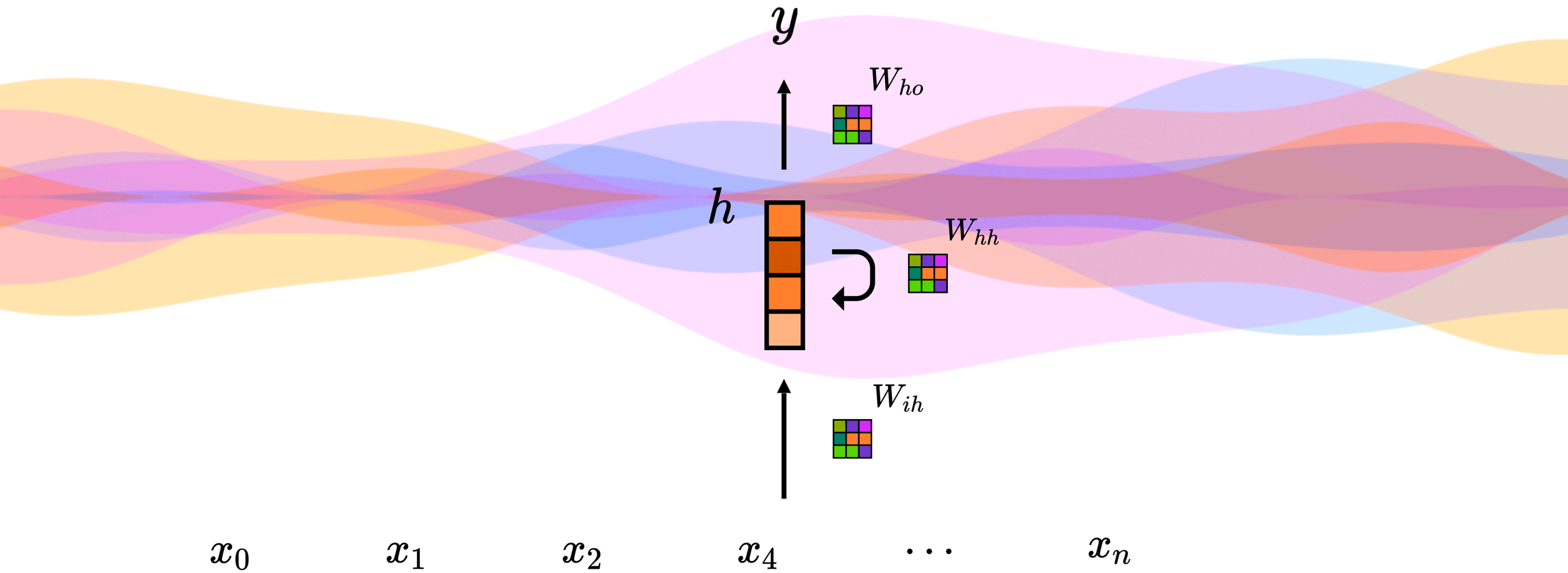


Rumelhart, D.E., Hinton, G.E. and Williams, R.J., 1986. Learning representations by back-propagating errors. Nature

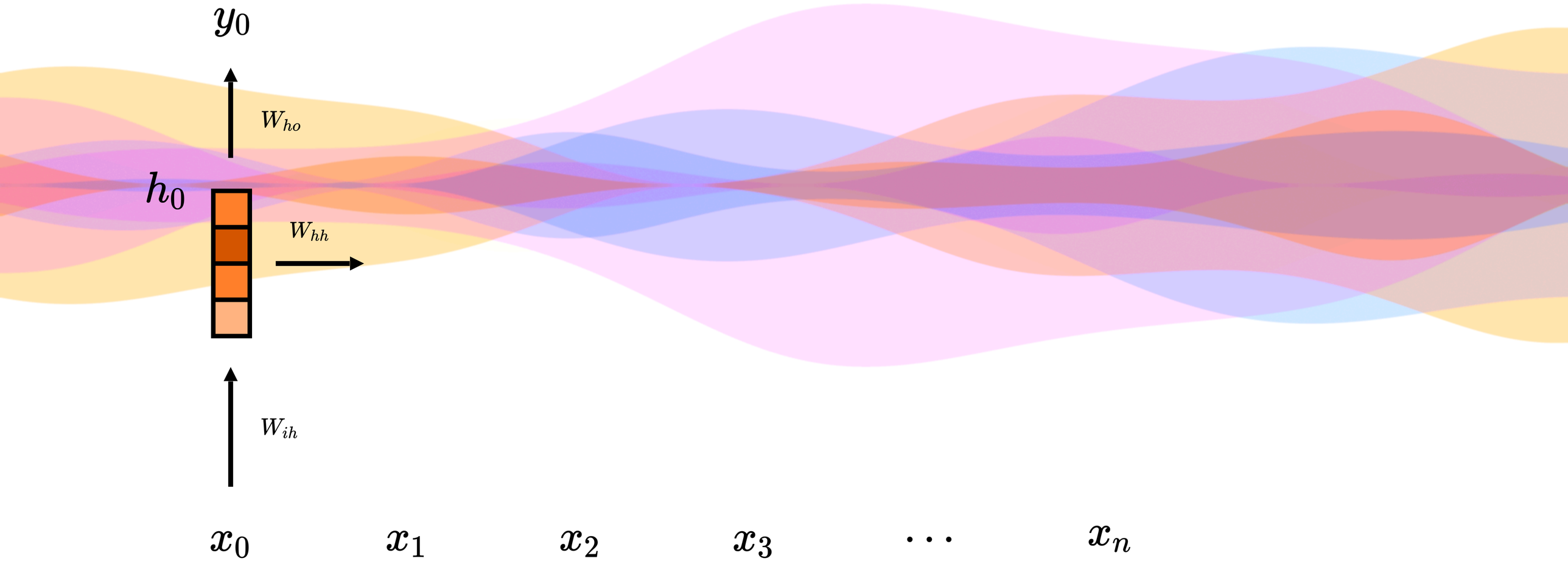
Recurrent Neural Networks



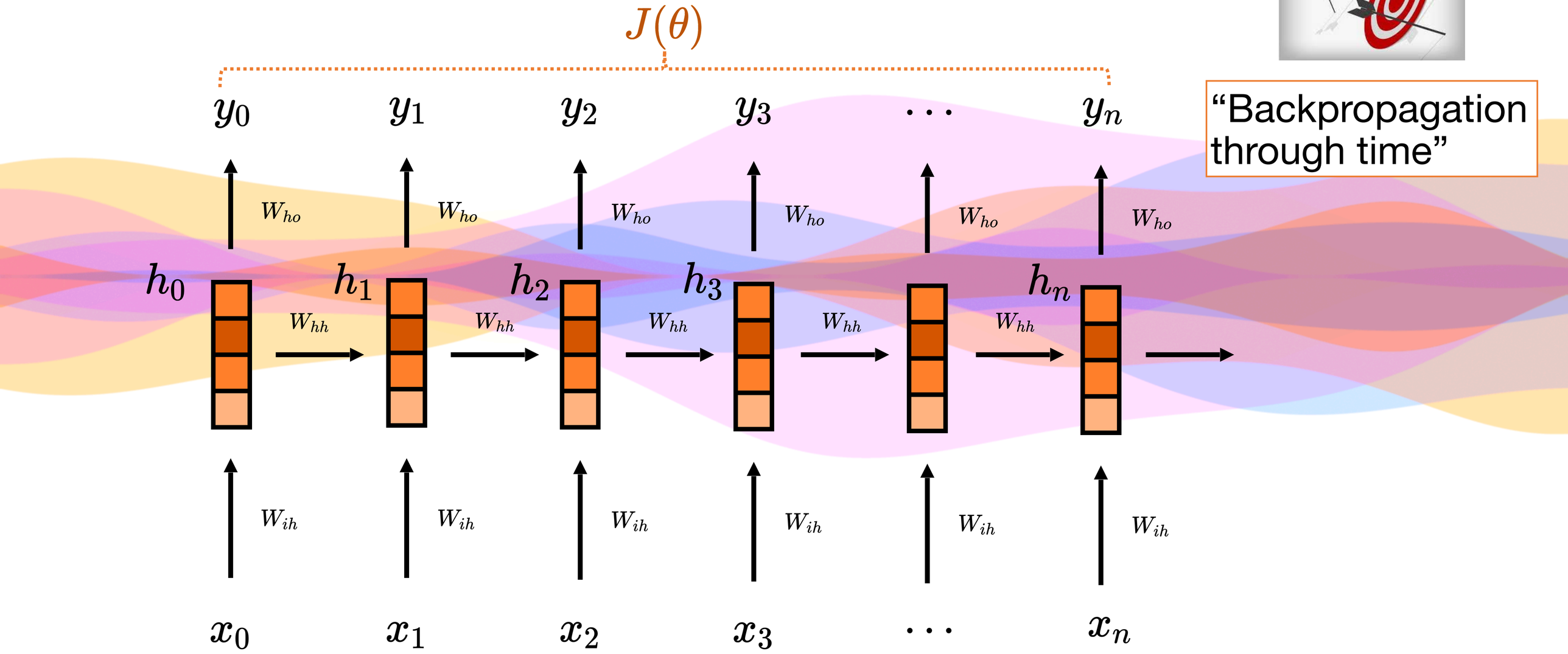
Recurrent Neural Networks



Recurrent Neural Networks

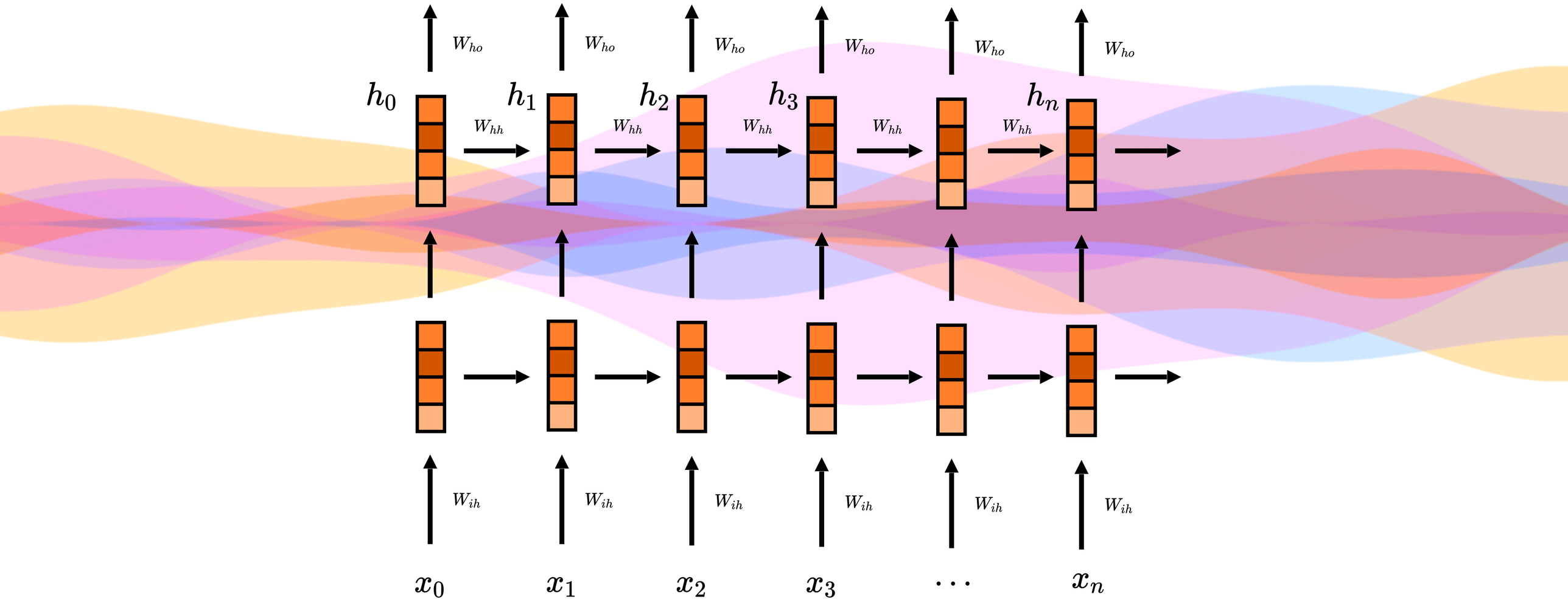


Recurrent Neural Networks



"Backpropagation through time"

Stacked layers in RNNs



Right-to-left units in RNNs

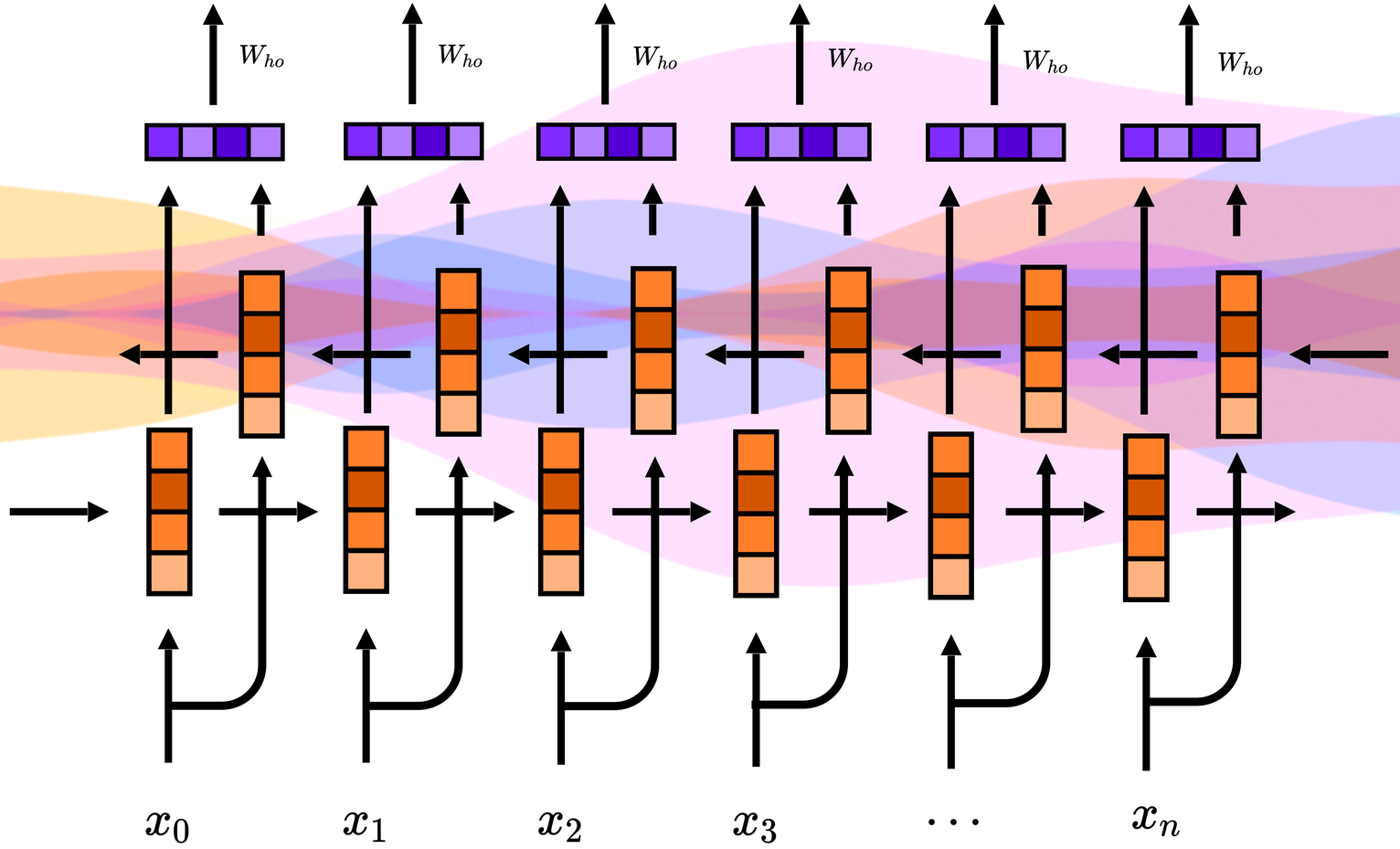
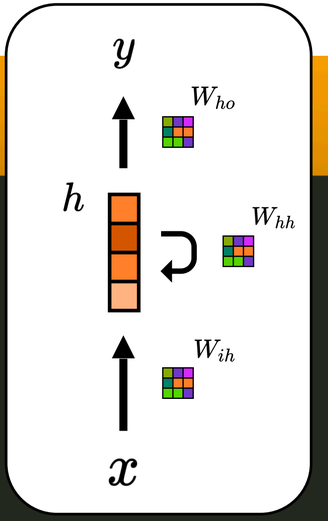


Photo by Mikhail Vasilyev on Unsplash

Not really




```python
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

- **Weights are shared across time**
  - the number of parameters is low (3 matrices in Vanilla RNN)
  - all inputs get equal treatment
- 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 == **no parallelization through “time”**
- 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

Using RNNs

Language modeling

- Modeling language entails predicting what's the most likely item (generally speaking, a *token*) given a context.

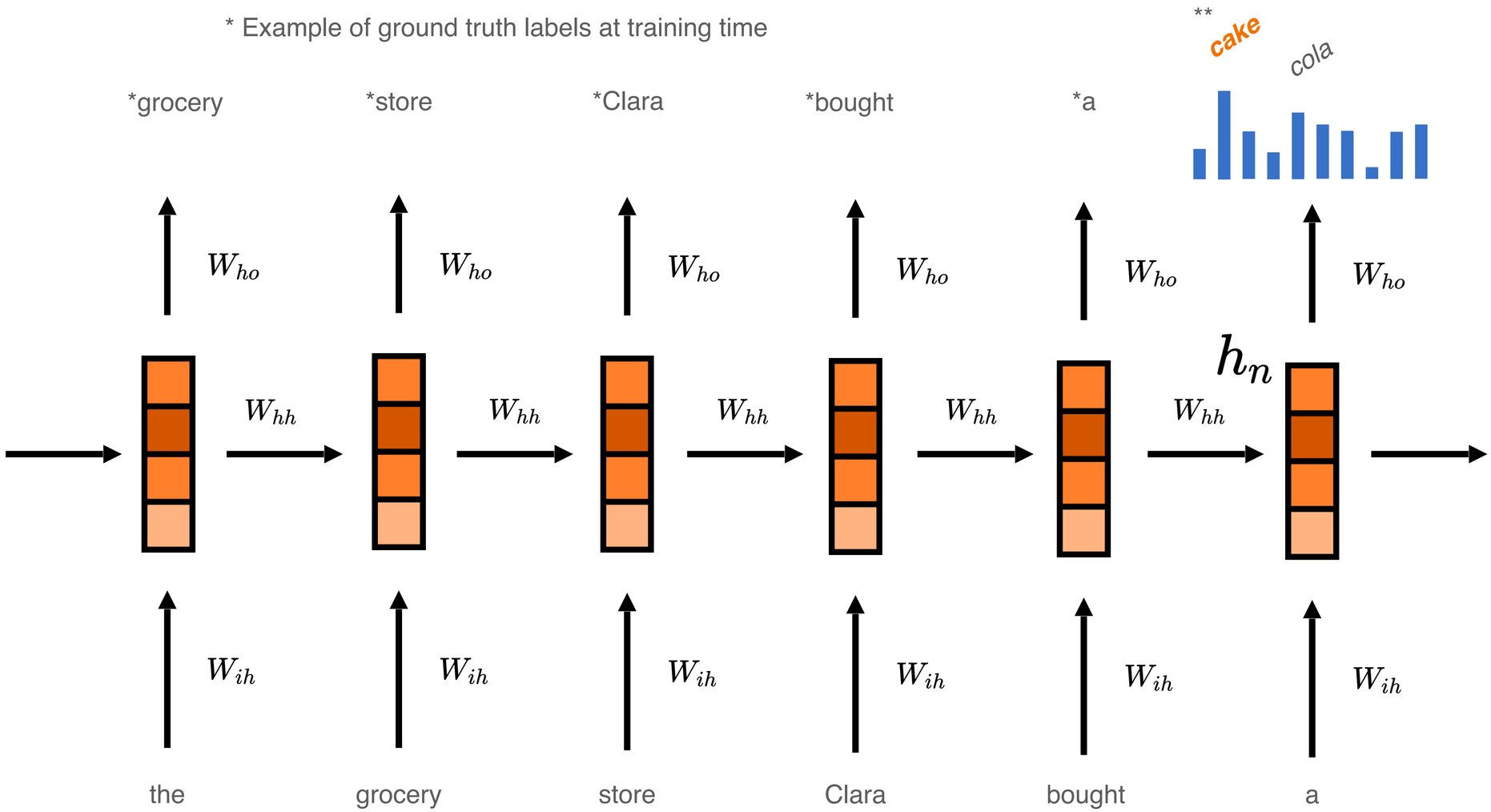
Back at the grocery store, Clara bought a _____

Grocery stuff should be more likely to follow
we are modeling a probability

RNNs for Language Modeling

** Example of PDF at inference time

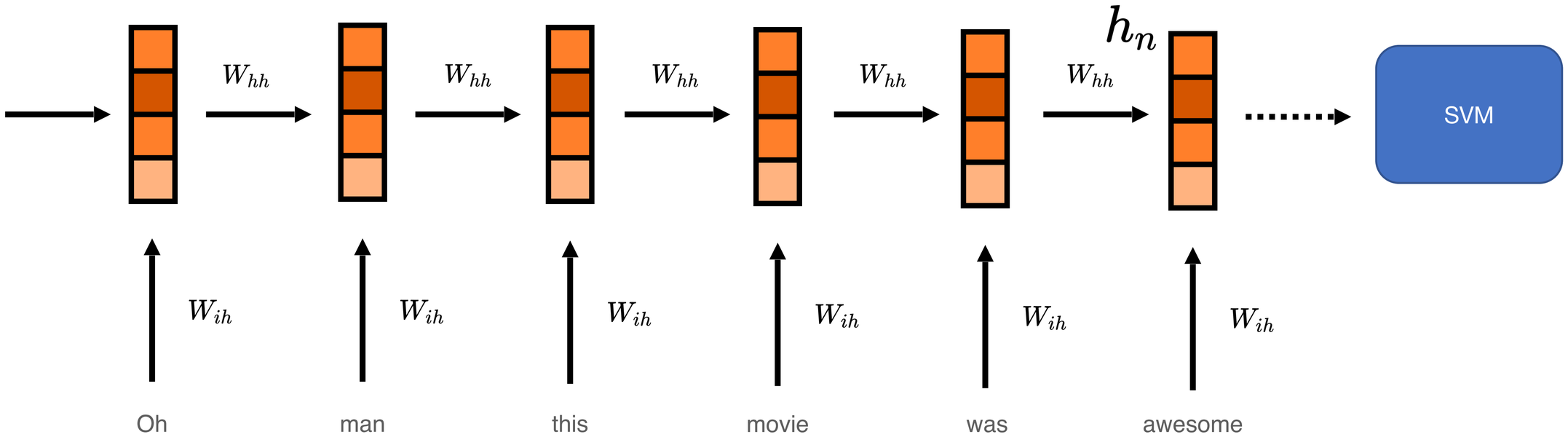
* Example of ground truth labels at training time



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

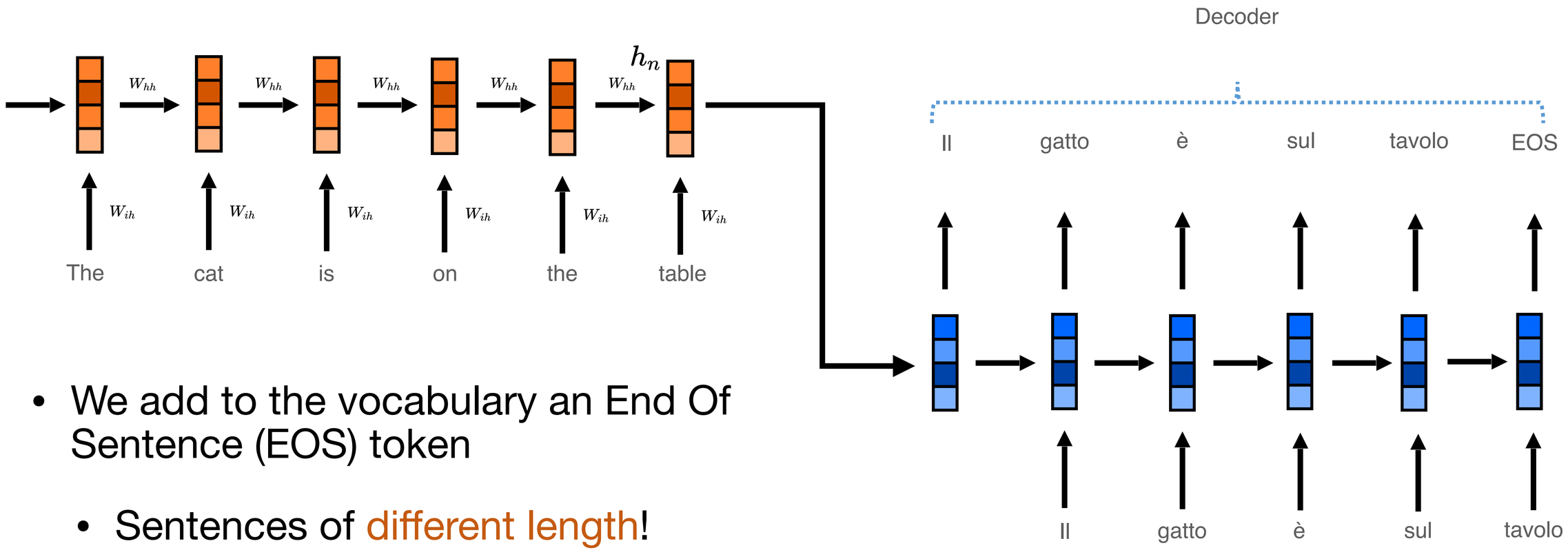
RNNs for Sentiment Analysis

We can use the network as an “encoder” for further downstream tasks.



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

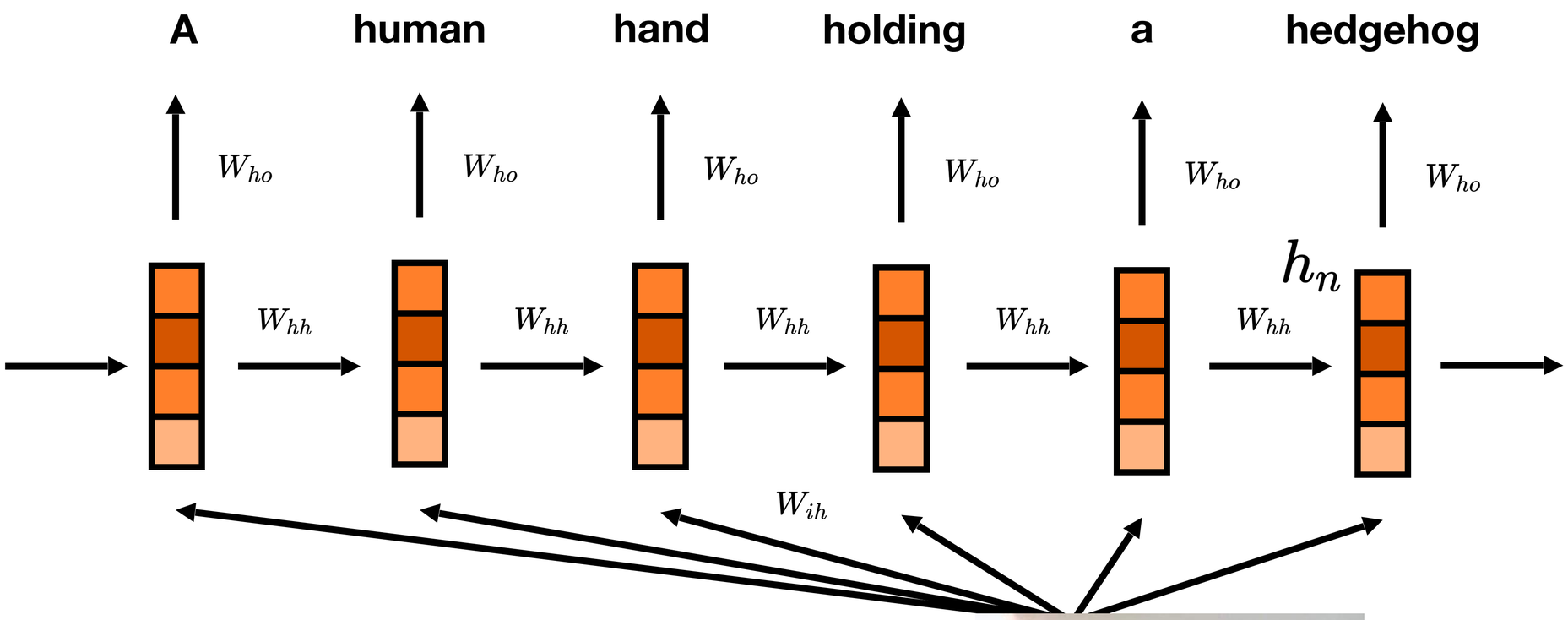
RNNs for Neural Machine Translation



- 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



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

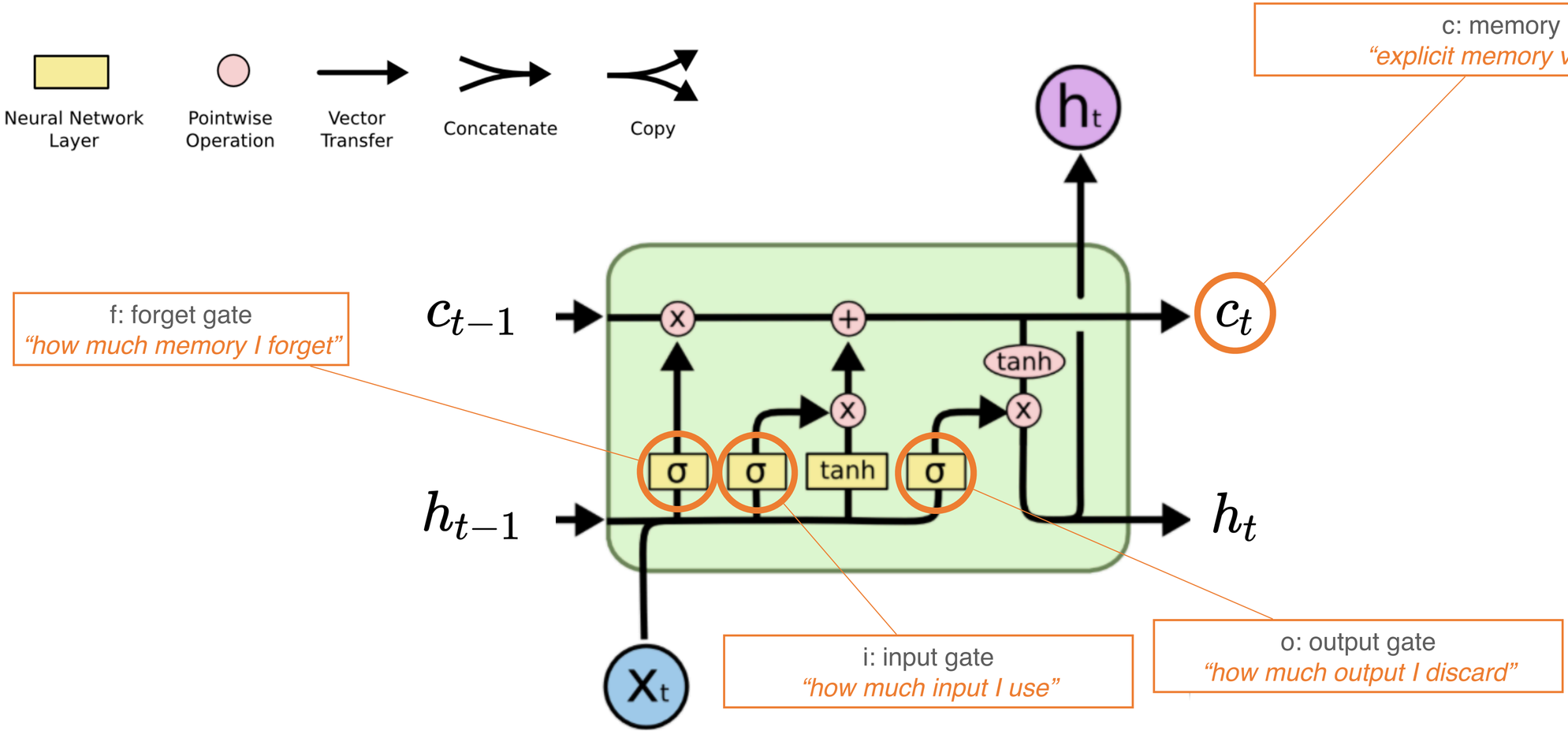
Dealing with long-range dependencies

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., when to forget, what to remember)
- 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/

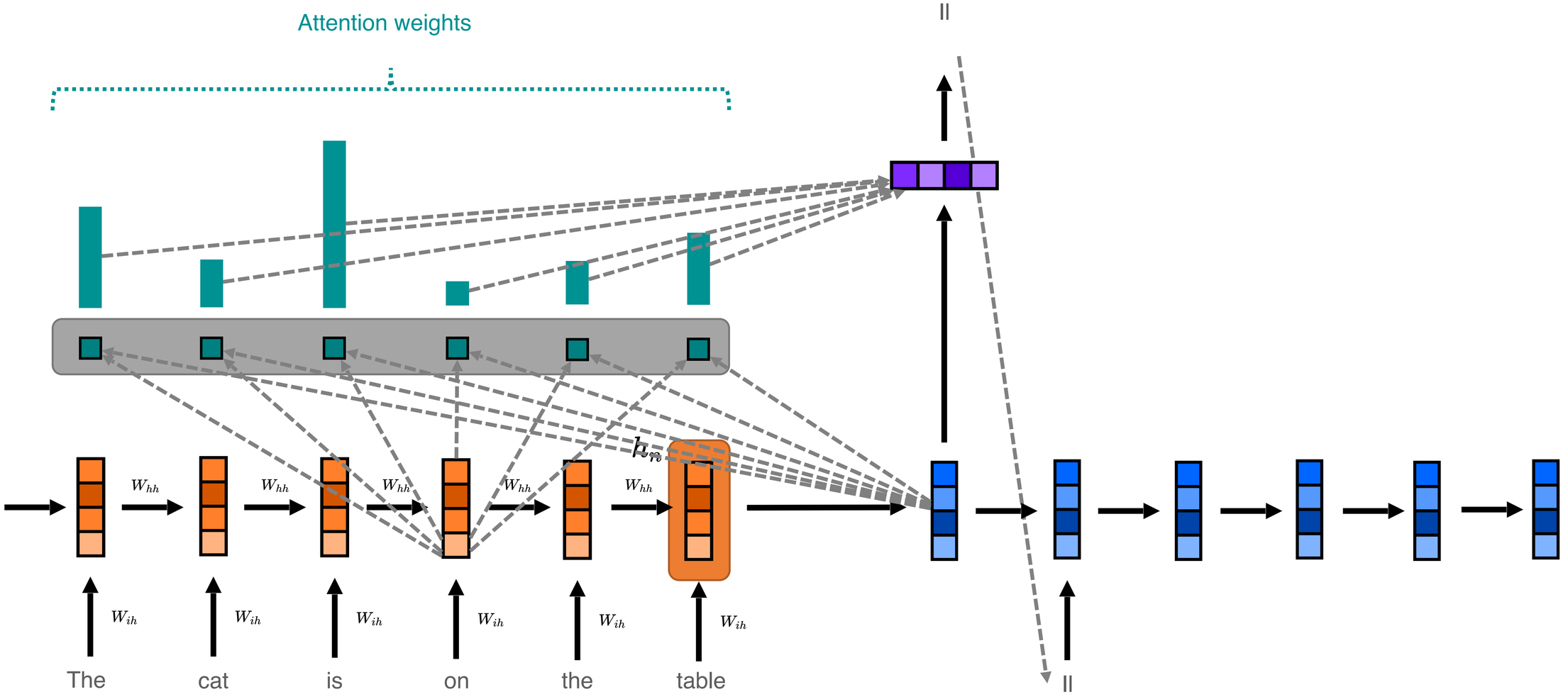
Photo by Samantha Gades on Unsplash

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
  - Direct connection to let information (and gradients) flow
  - Foundational in Transformers

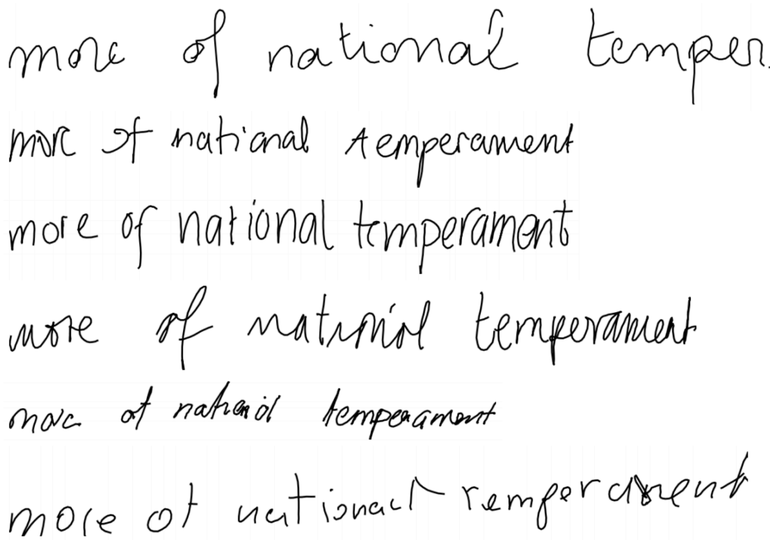
RNNs for Neural Machine Translation (2)



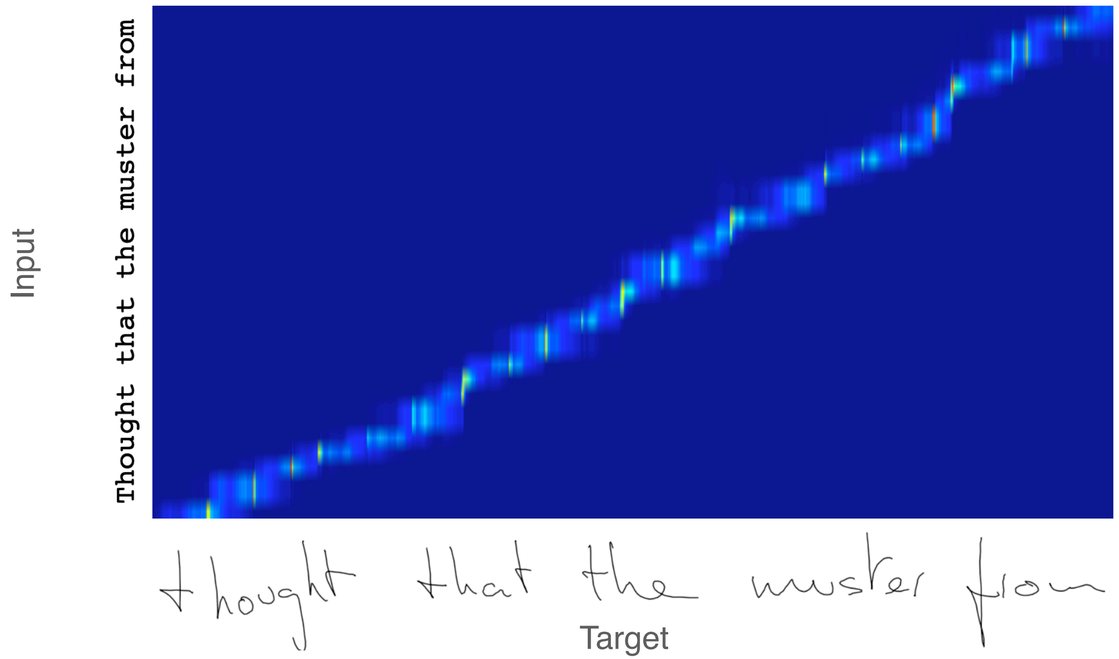
Bahdanau, D., Cho, K. and Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate.

Generating sequences with RNNs

- Architecture: encoder-decoder 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)

Thanks

- Attention Is All You need. Waswani et al.

- Introducing the Transformer
  - No more recurrent units
  - Fundamentally, a machine translation paper
  - Language modeling using attention only

- Building block of modern ~~language~~ models
  - BERT, GPT-*, ViT, Wav2Vec, …

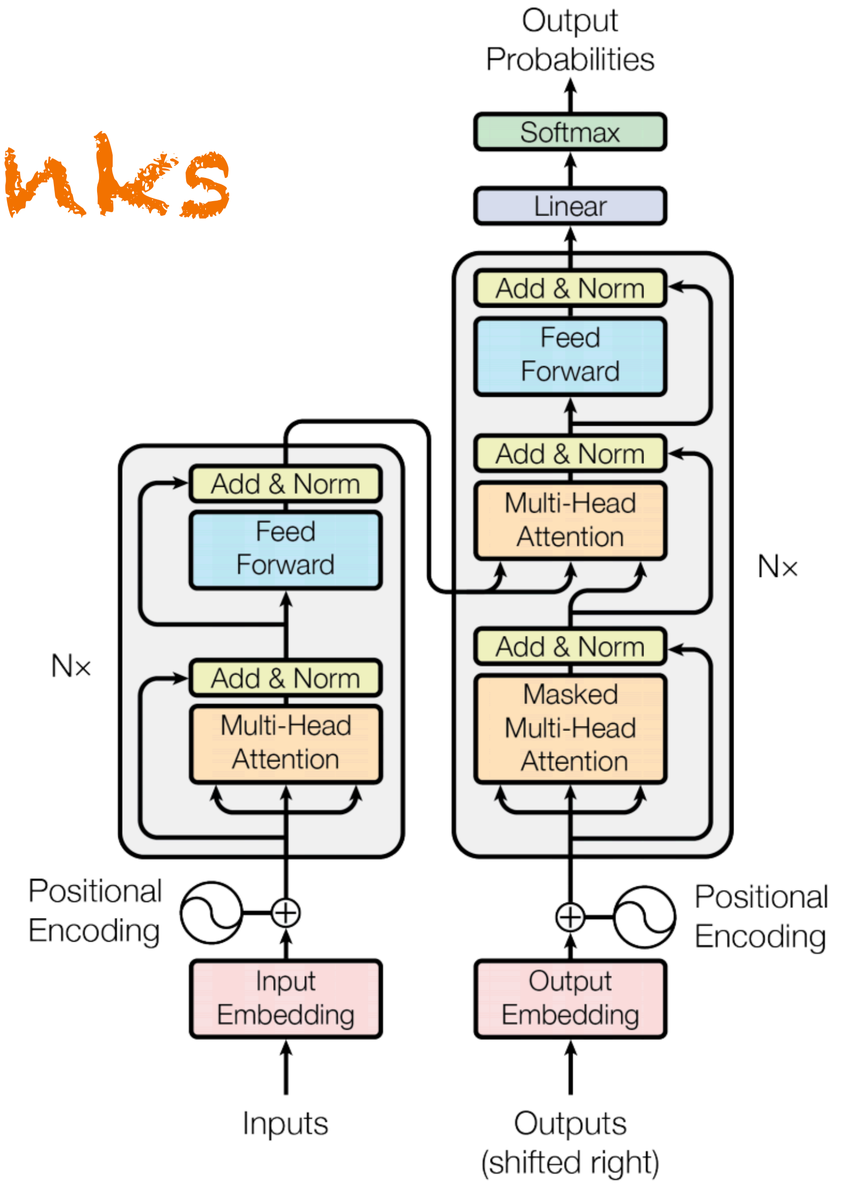


Figure 1: The Transformer - model architecture.

Vaswani et al, 2017. Attention is All You Need.