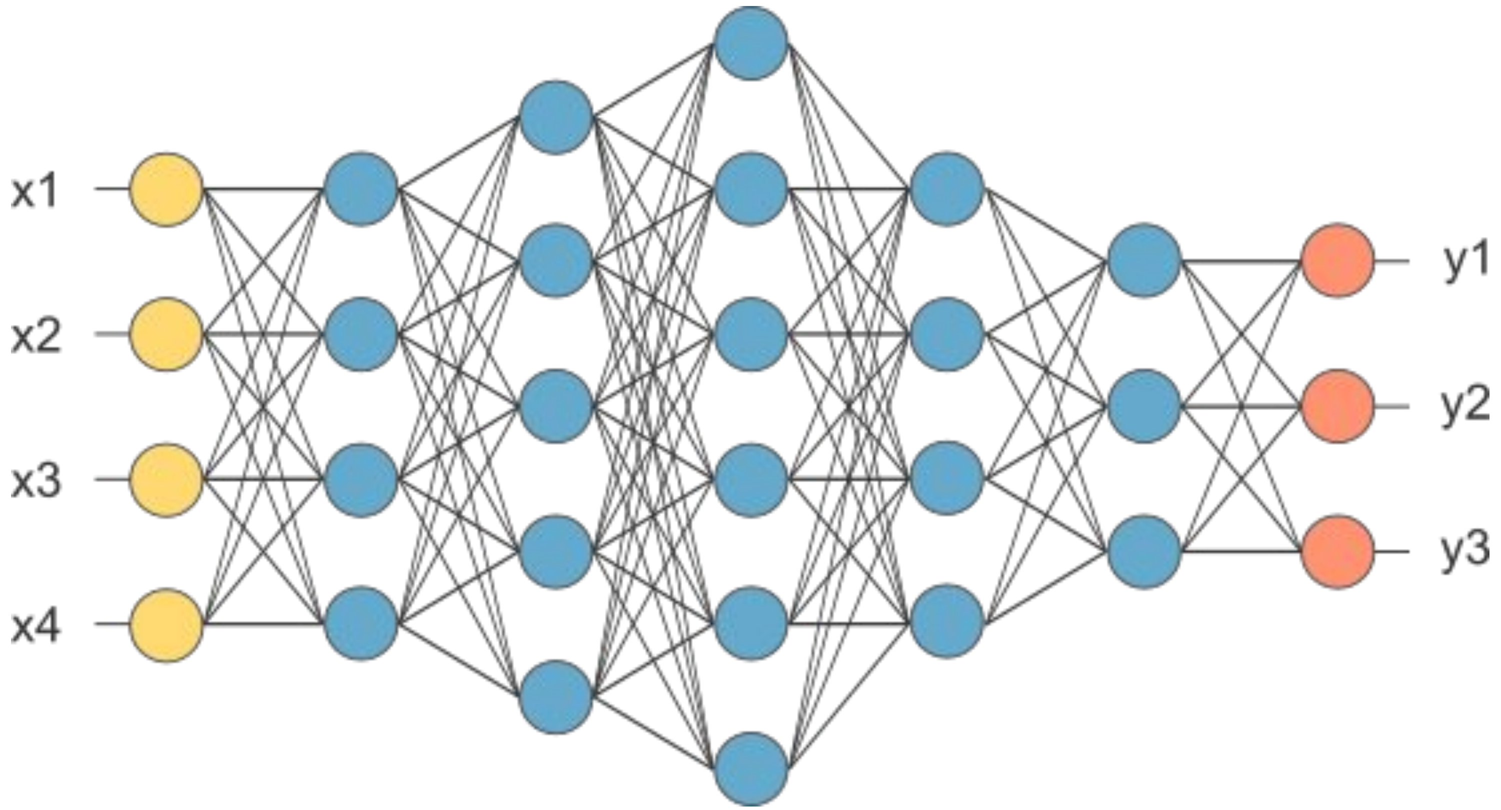
Attention and Transformers

November 1, 2021

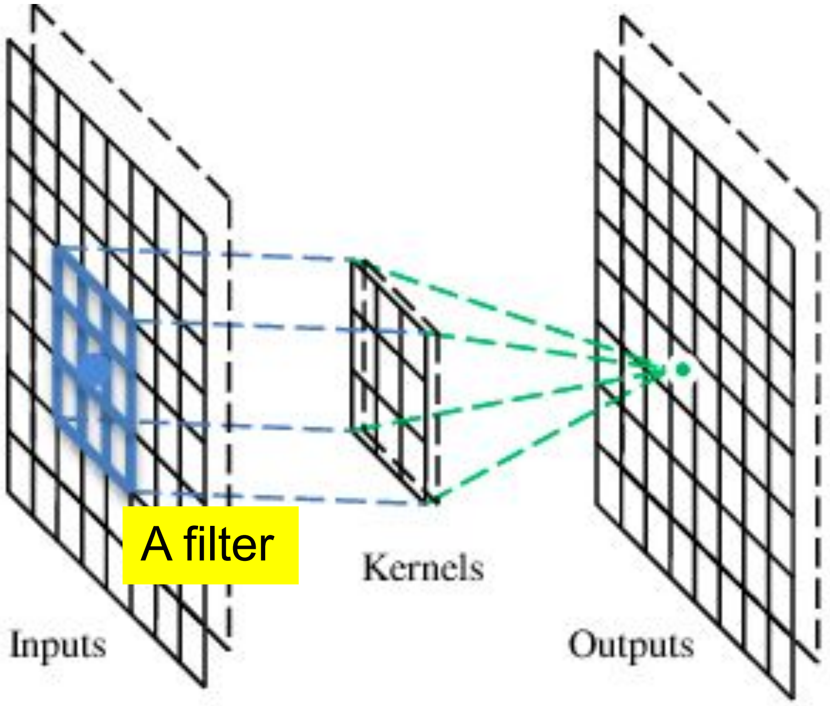
1. Fully connected network, feedforward network



To learn the weights on the edges

2. CNN

A CNN is a neural network with some convolutional layers (and some other layers). A convolutional layer has a number of filters that do convolutional operation.



Convolutional layer

1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0
1	0	0	0	1	0
0	1	0	0	1	0
0	0	1	0	1	0

Input

These are the network parameters to be learned.

1	-1	-1
-1	1	-1
-1	-1	1

Filter 1

-1	1	-1
-1	1	-1
-1	1	-1

Filter 2

⋮ ⋮

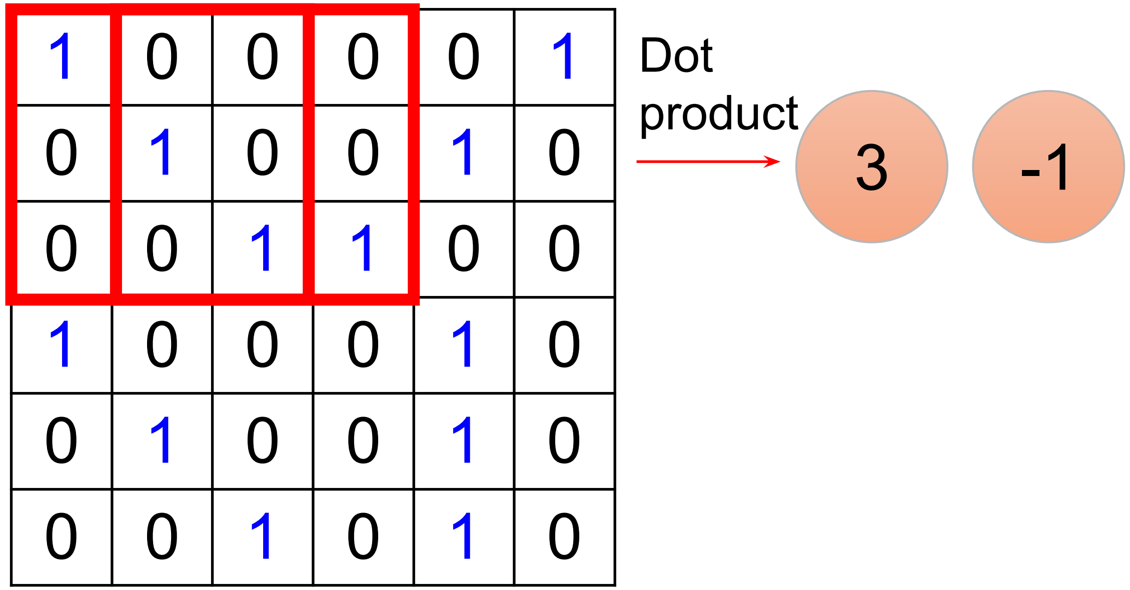
Each filter detects a small pattern (3 x 3).

Convolution Operation

1	-1	-1
-1	1	-1
-1	-1	1

Filter 1

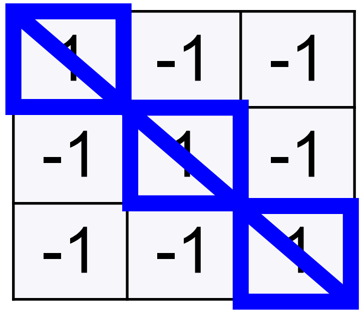
stride=1



1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0
1	0	0	0	1	0
0	1	0	0	1	0
0	0	1	0	1	0

Input

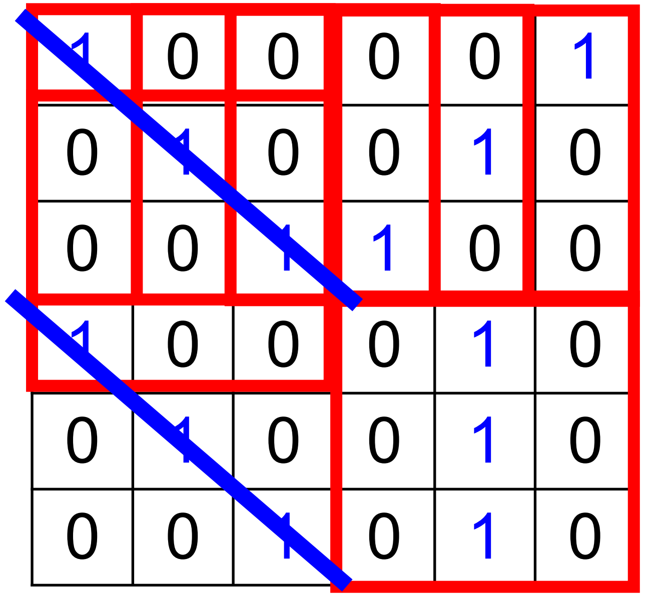
Convolution



1	-1	-1
-1	1	-1
-1	-1	1

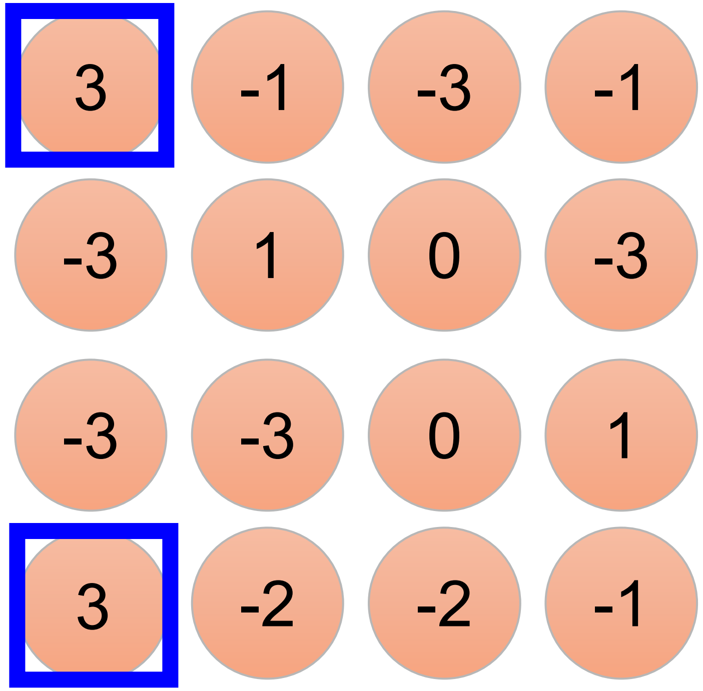
Filter 1

stride=1



1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0
1	0	0	0	1	0
0	1	0	0	1	0
0	0	1	0	1	0

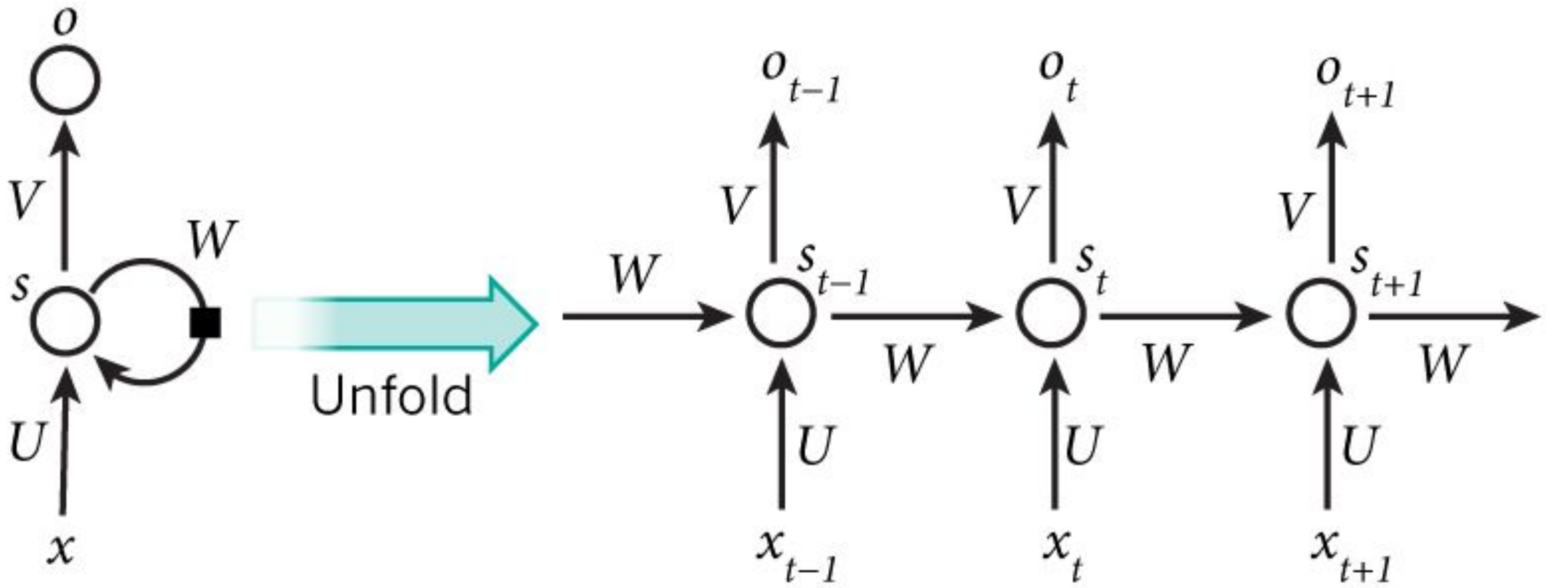
Input



3	-1	-3	-1
-3	1	0	-3
-3	-3	0	1
3	-2	-2	-1

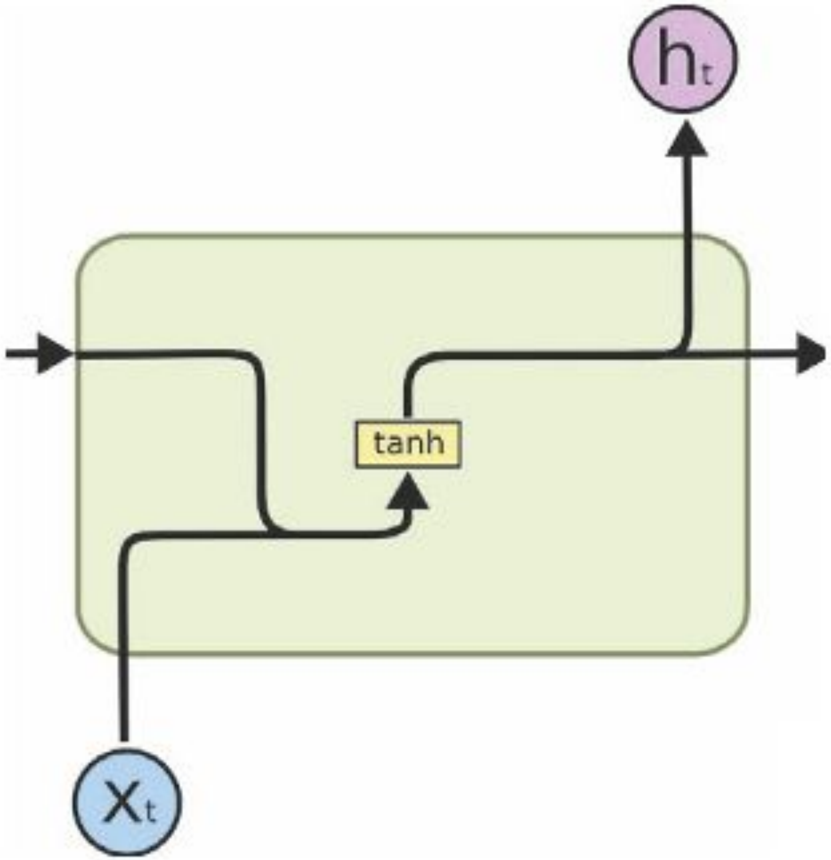
https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

3. RNN

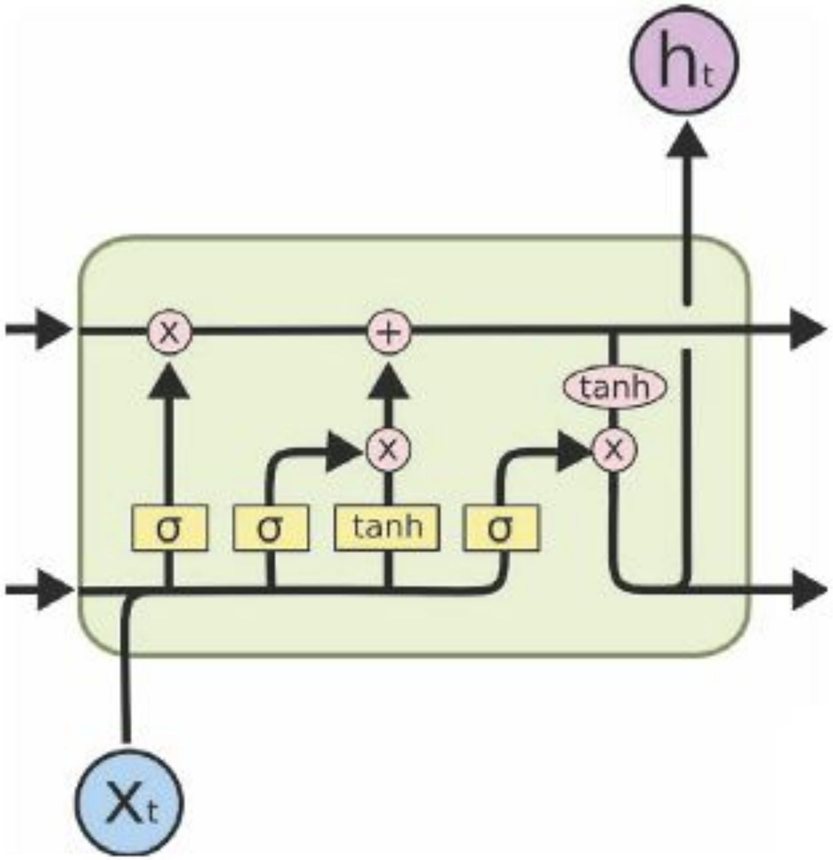


Parameters to be learned:
U, V, W

Simple RNN vs LSTM



(a) RNN

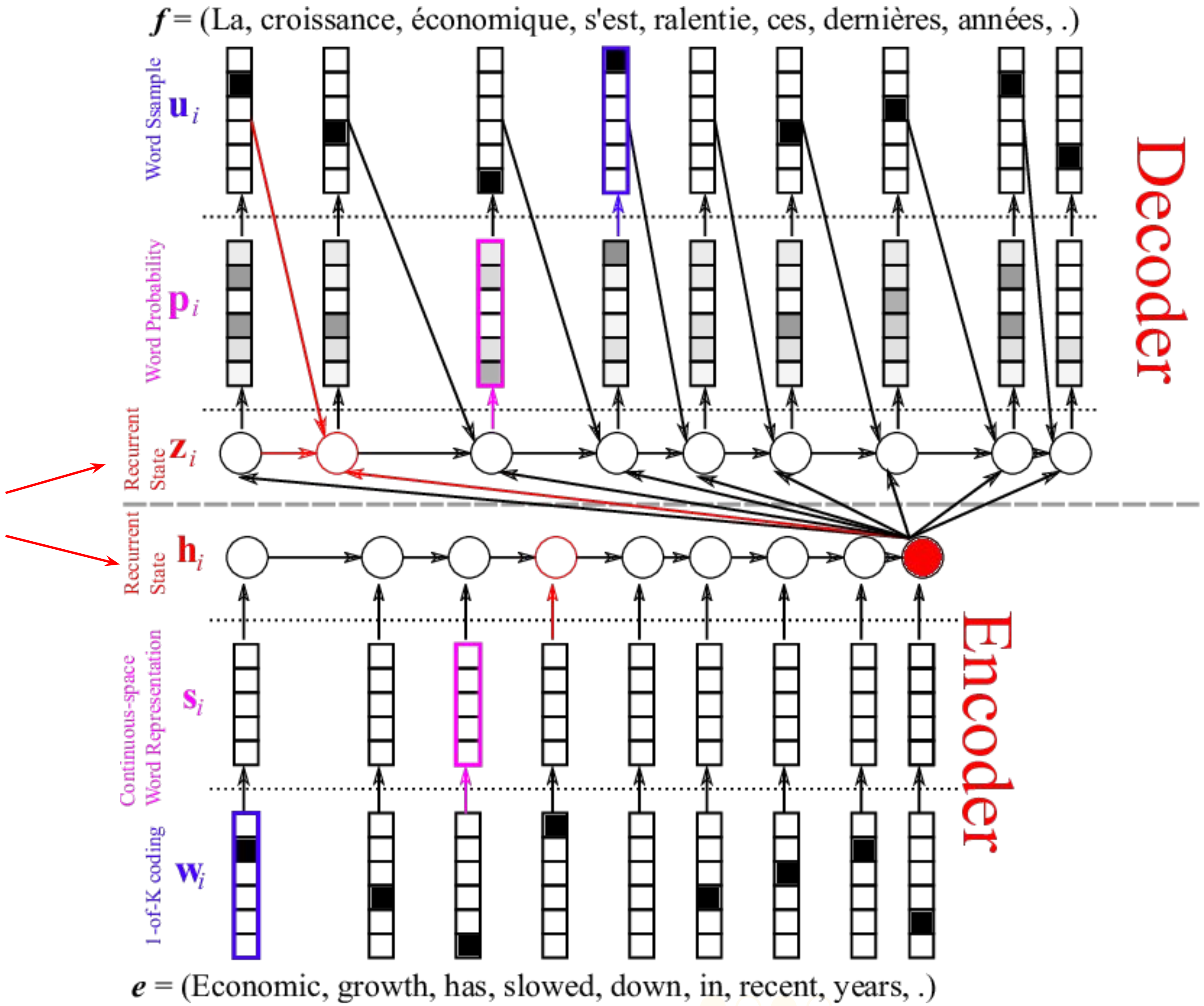


(b) LSTM

Encoder Decoder

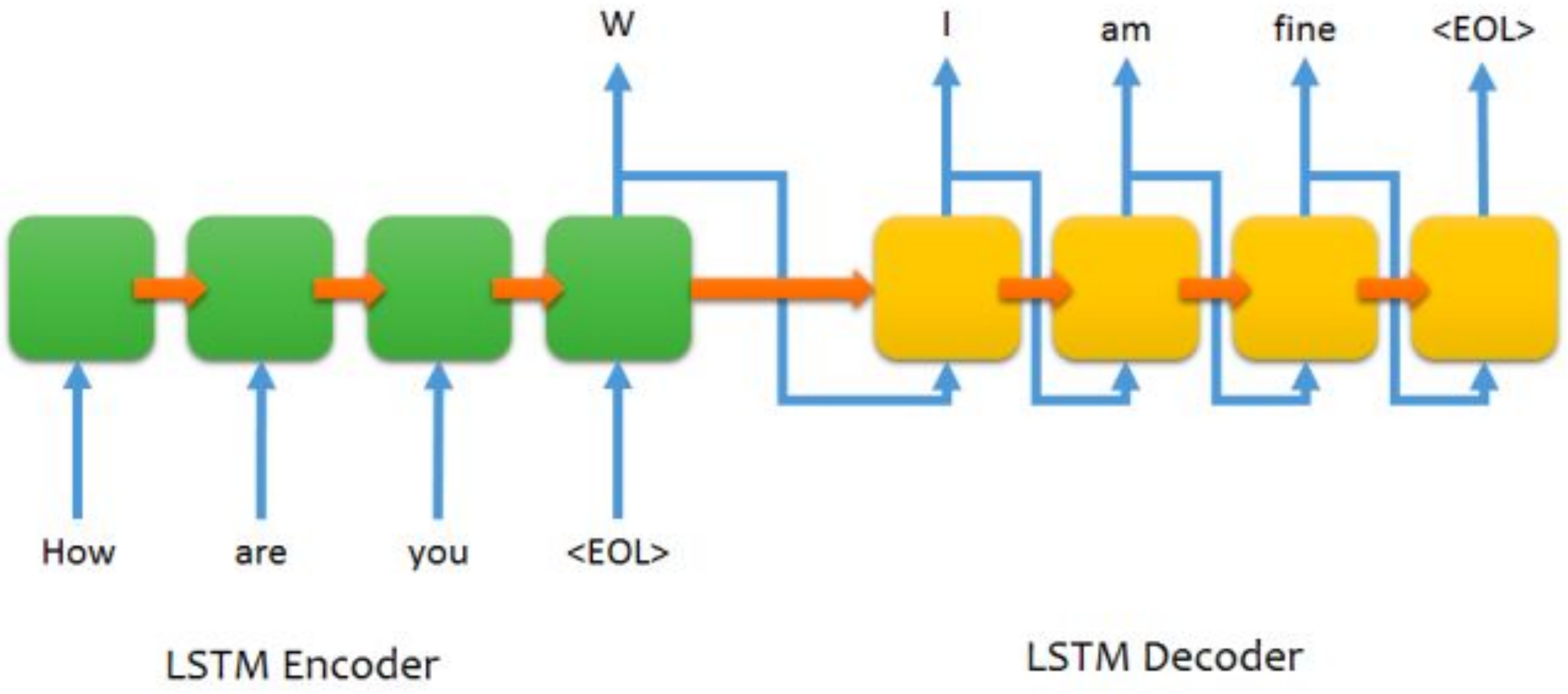
- Sequence to Sequence model transforms a given sequence of elements to another sequence.
- LSTM is one such model.
- Seq2Seq consists of an Encoder and a Decoder
    - Encoder: take input sequence and map it to an n-dimensional vector.
    - Decoder: take the output from an encoder and convert it to an output sequence.

Encoder-Decoder machine translation

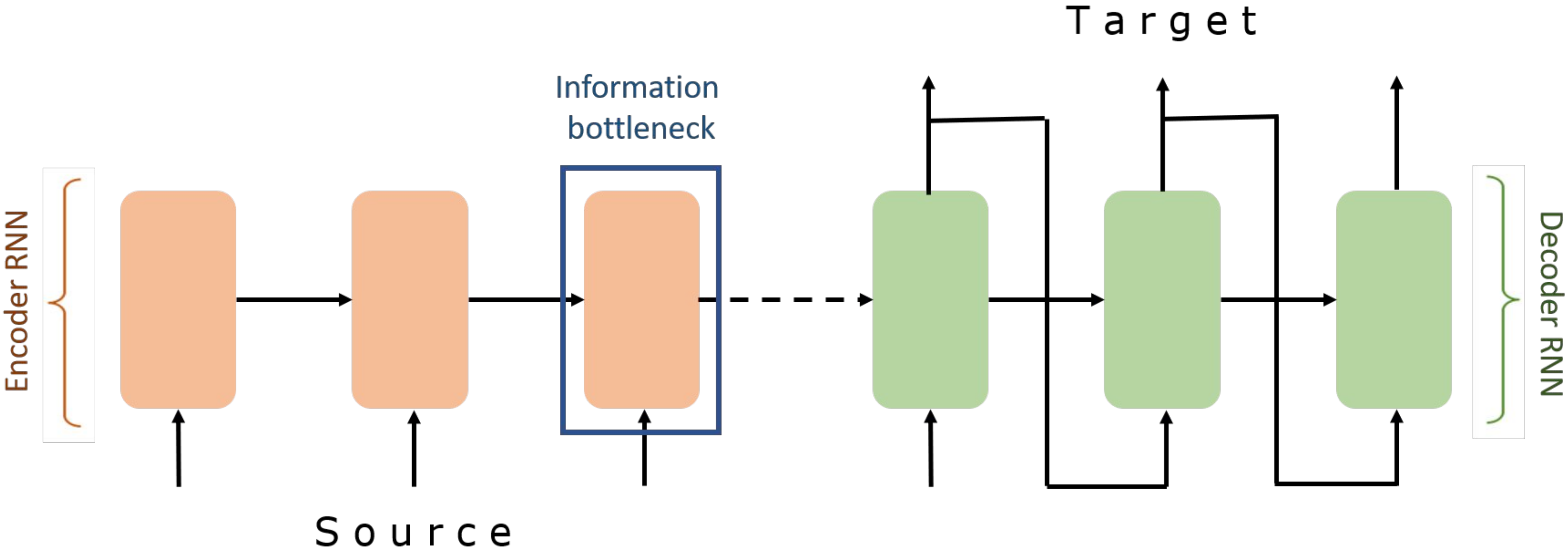

f = (La, croissance, économique, s'est, ralentie, ces, dernières, années, .)
Word Ssample
u_i
Word Probability
p_i
Recurrent State
z_i
Decoder
Recurrent State
h_i
Continuous-space Word Representation
s_i
1-of-K coding
w_i
Encoder
e = (Economic, growth, has, slowed, down, in, recent, years, .)

LSTM

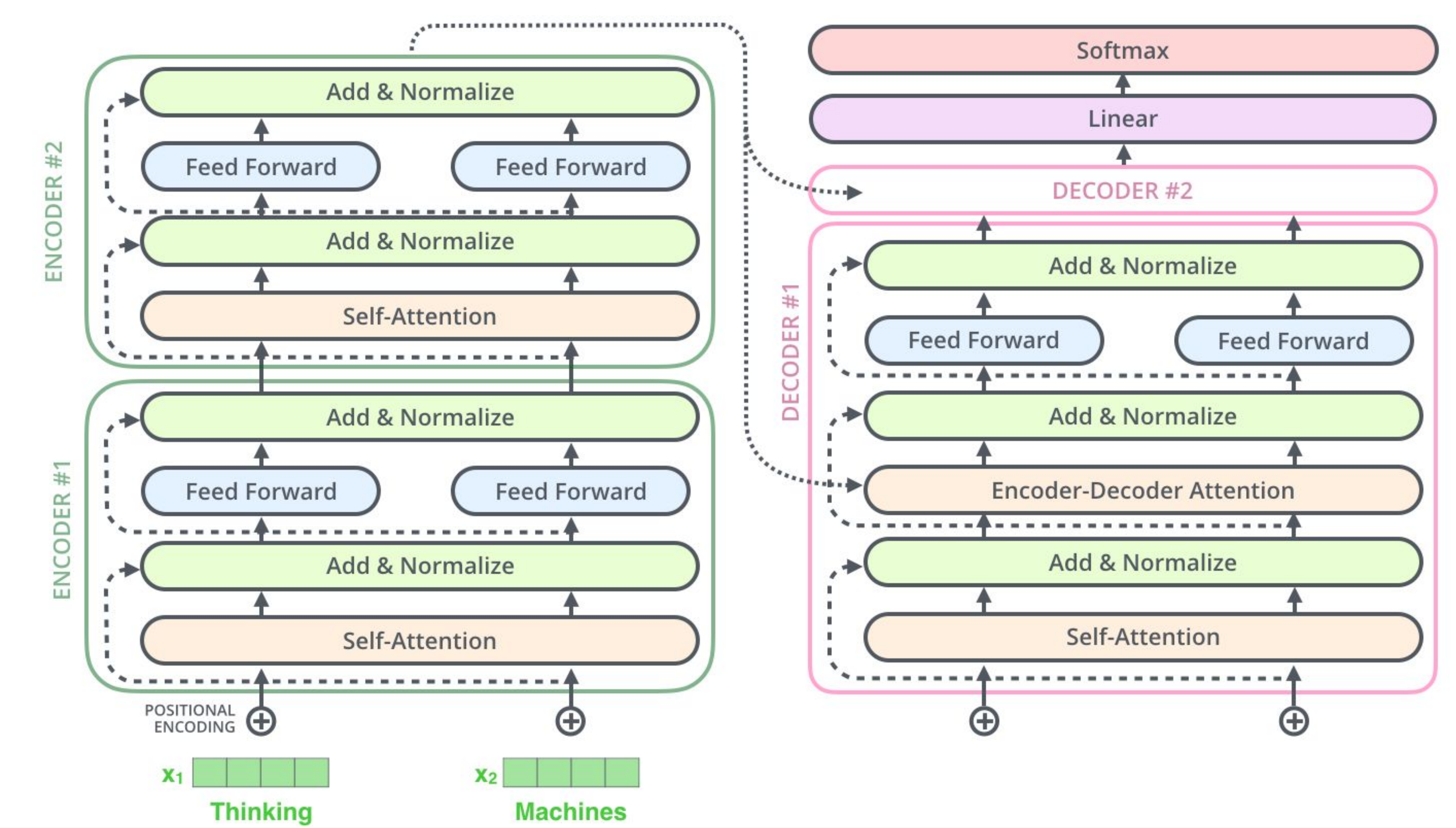
Encoder-Decoder LSTM structure for chatting (for non-intelligent beings)



Avoiding Information bottleneck

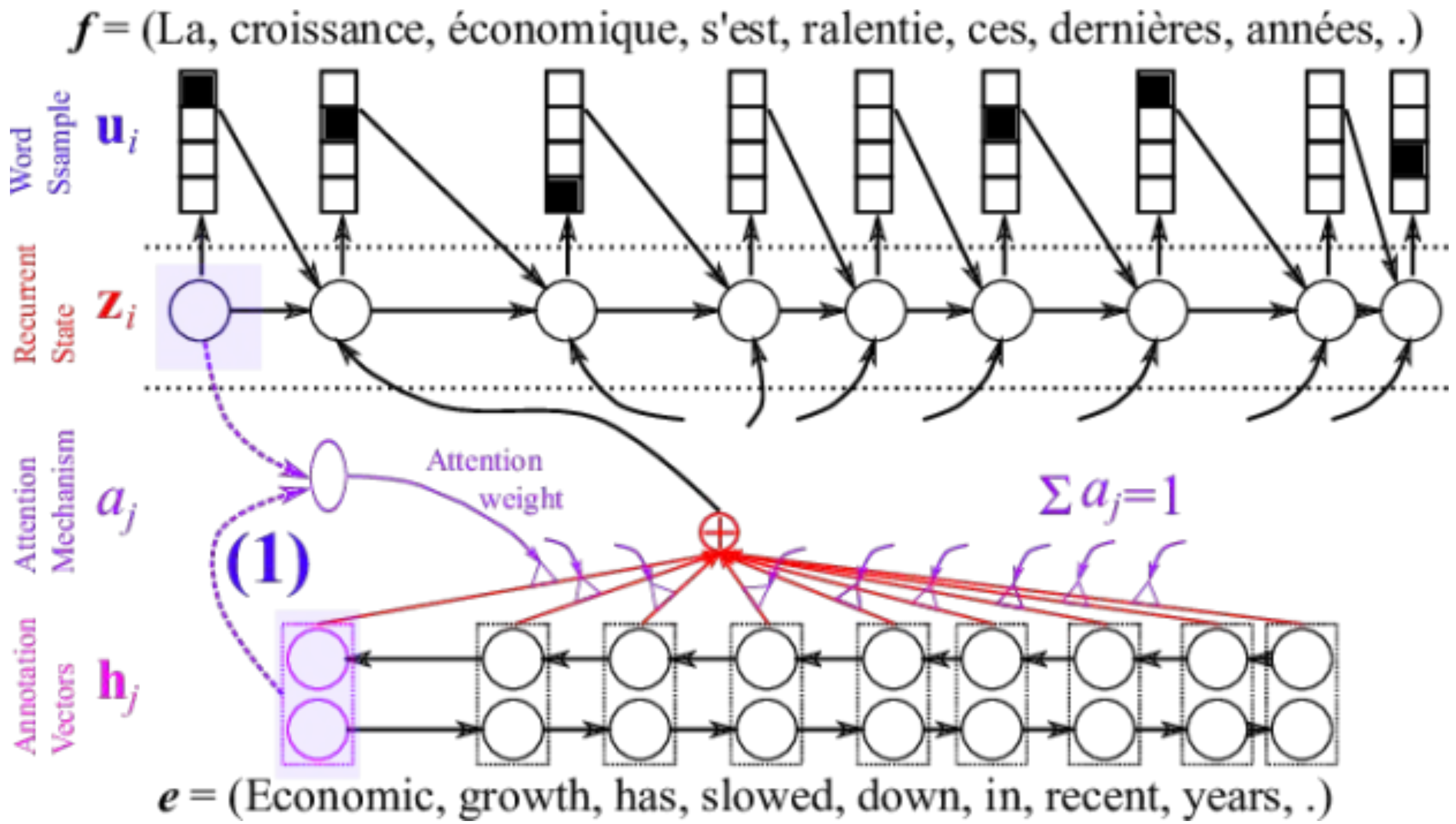


Transformer



Attention

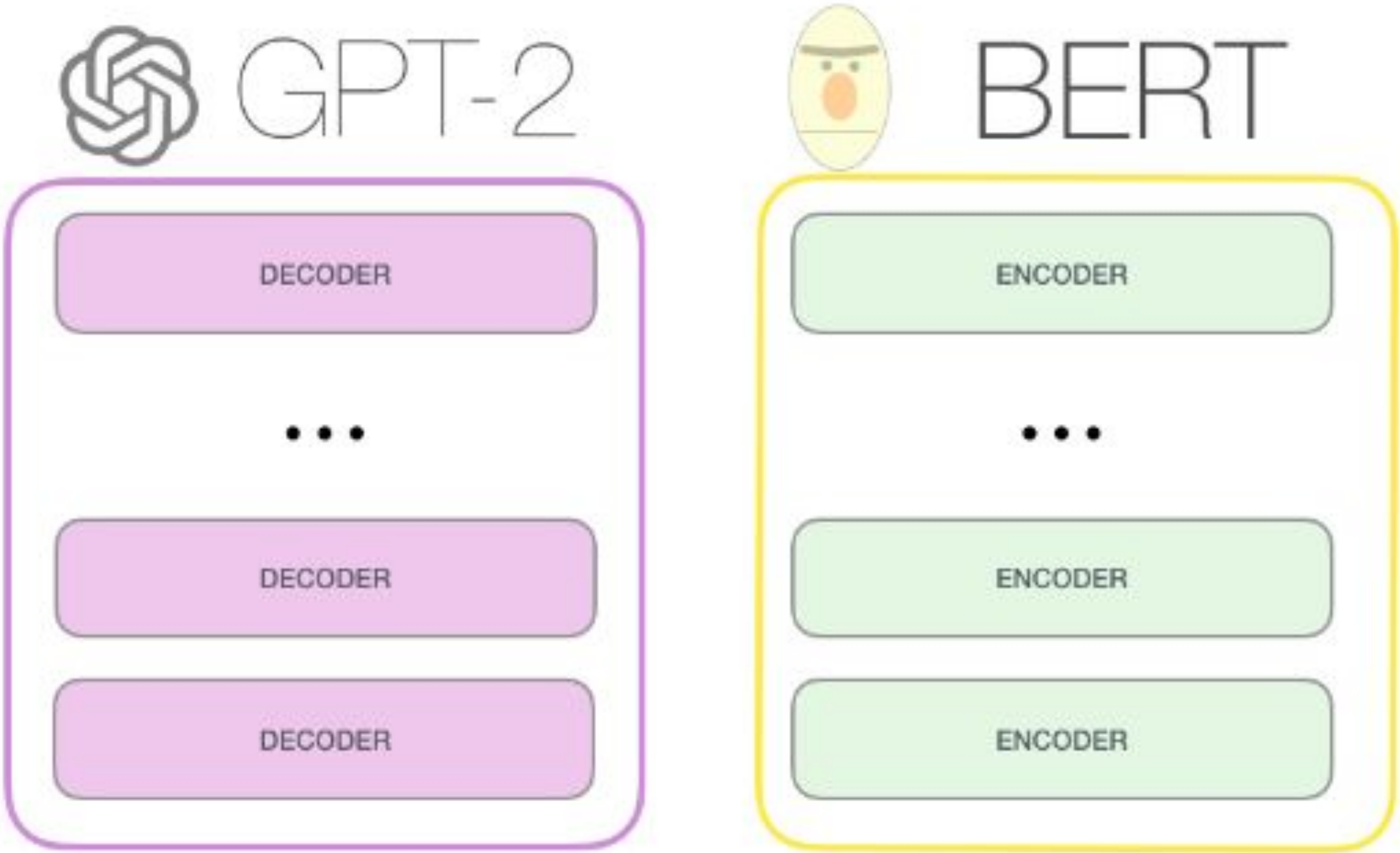
Given the input sequence, the **attention** decides which other parts of the sequence are important.



Transformers, GPT-2, and BERT

1. A transformer uses Encoder stack to model input, and uses Decoder stack to model output (using input information from encoder side).
2. But if we do not have input, we just want to model the "next word", we can get rid of the Encoder side of a transformer and output "next word" one by one. This gives us **GPT**.
3. If we are only interested in training a language model for the input for some other tasks, then we do not need the Decoder of the transformer, that gives us **BERT**.

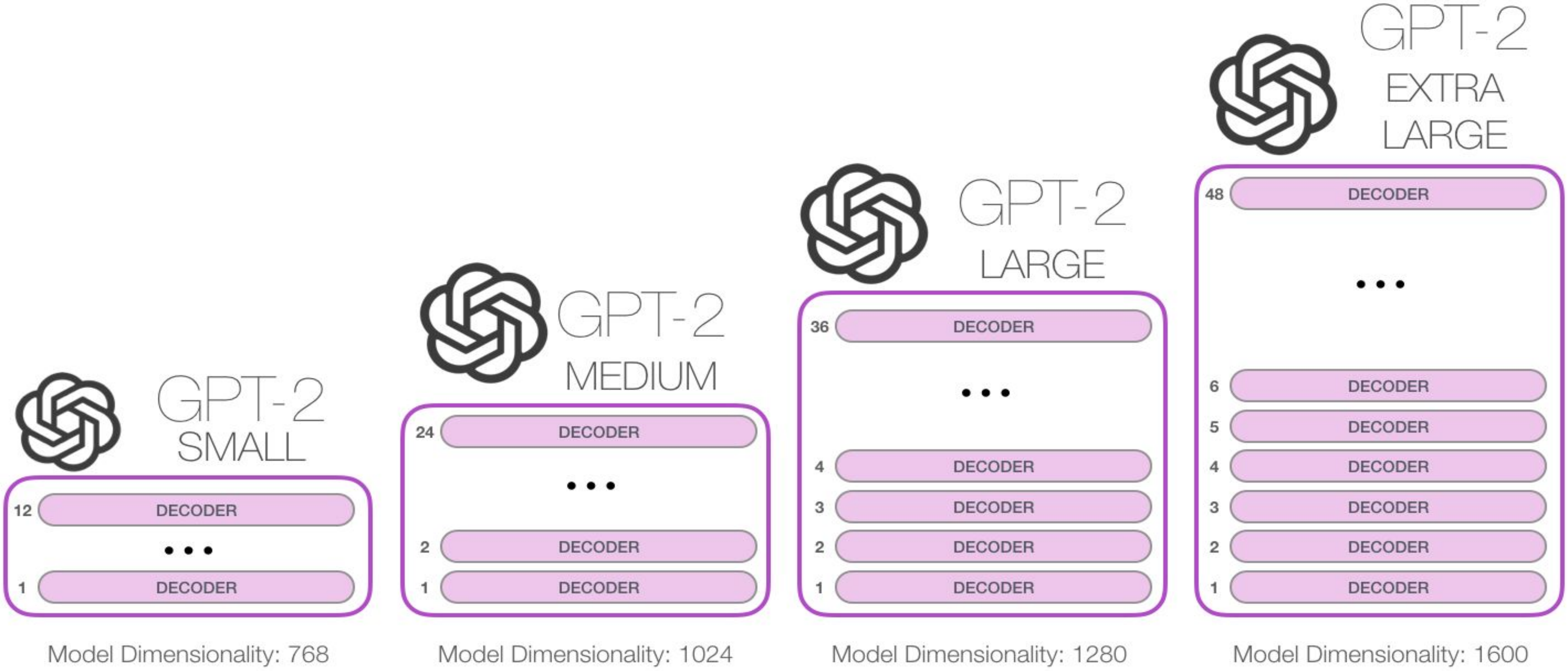
GPT-2, BERT


GPT-2
BERT
DECODER
DECODER
DECODER
ENCODER
ENCODER
ENCODER

GPT-2, BERT

GPT released June 2018

GPT-2 released Nov. 2019 with 1.5B parameters



117M parameters

345M

762M

1542M