Natural Language Processing

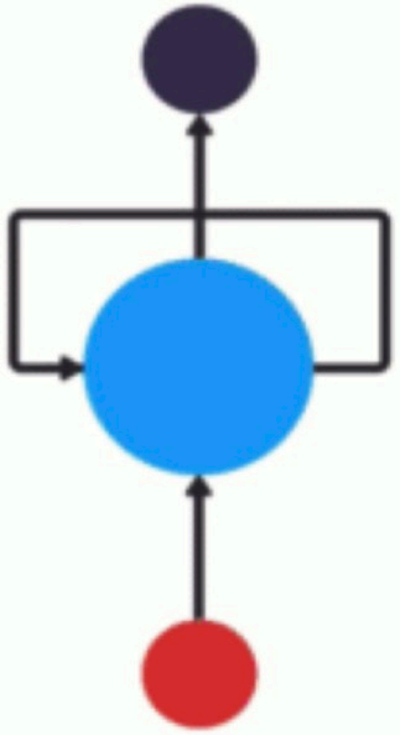
Session-4

RNN, LSTM, GRU

Working Logic of RNN, LSTM and GRU

NLP with Deep Learning

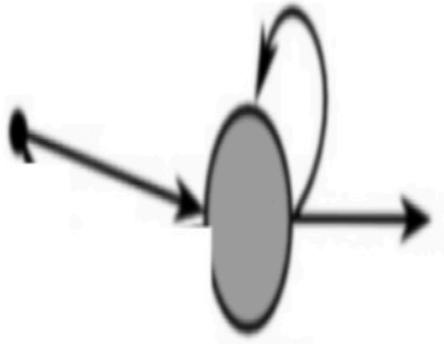
Recurrent Neural Networks (RNN)



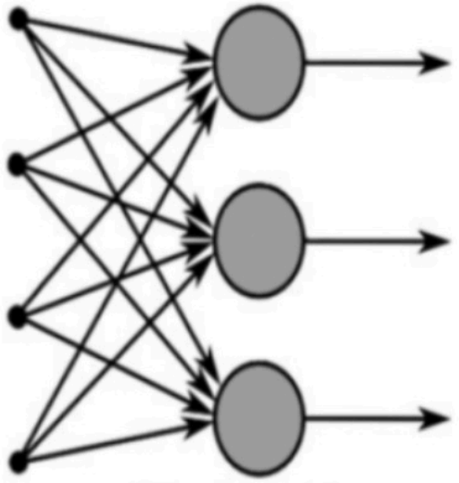
- RNN's are neural networks that are good at modeling sequence data.
- Each layer is dependent on the output from the previous layer.
- RNN's have memory.

NLP with Deep Learning

RNN vs FEED-FORWARD NEURAL NETWORKS



Recurrent Neural Network



Feed-Forward Neural Network

In a RNN the ***information cycles through a loop***. When it makes a decision, it **considers the current input** and also what it has learned from the ***inputs*** it received ***previously***.

A usual RNN has a ***short-term memory***. In combination with a **LSTM** or **GRU** they also have a *long-term memory*

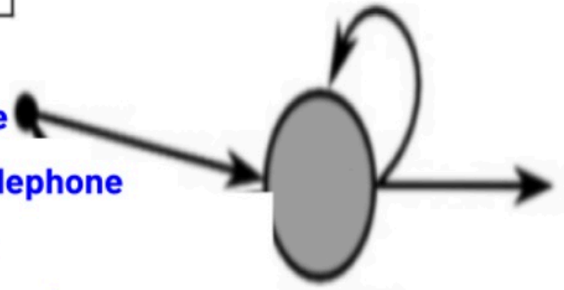
NLP with Deep Learning

RNN vs FEED-FORWARD NEURAL NETWORKS

text
The telephone is amazing
the best TV in my all life

text
the telephone is amazing

1.the
2.telephone
3.is
4.amazing



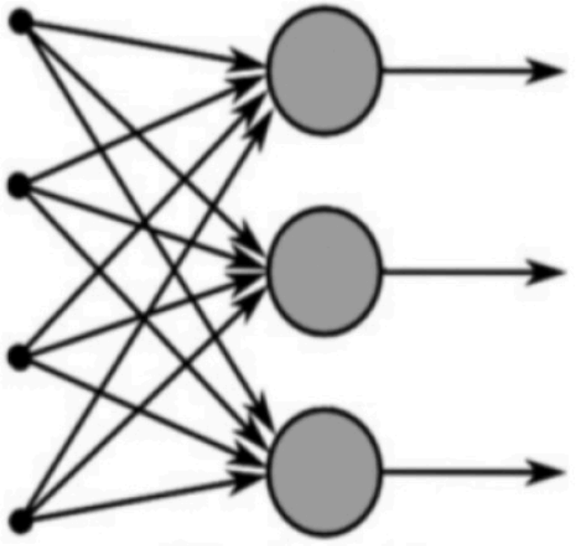
Recurrent Neural Network

km	age	hp
1000	1	130
15000	2	160
20000	3	250

km
1000

age
1

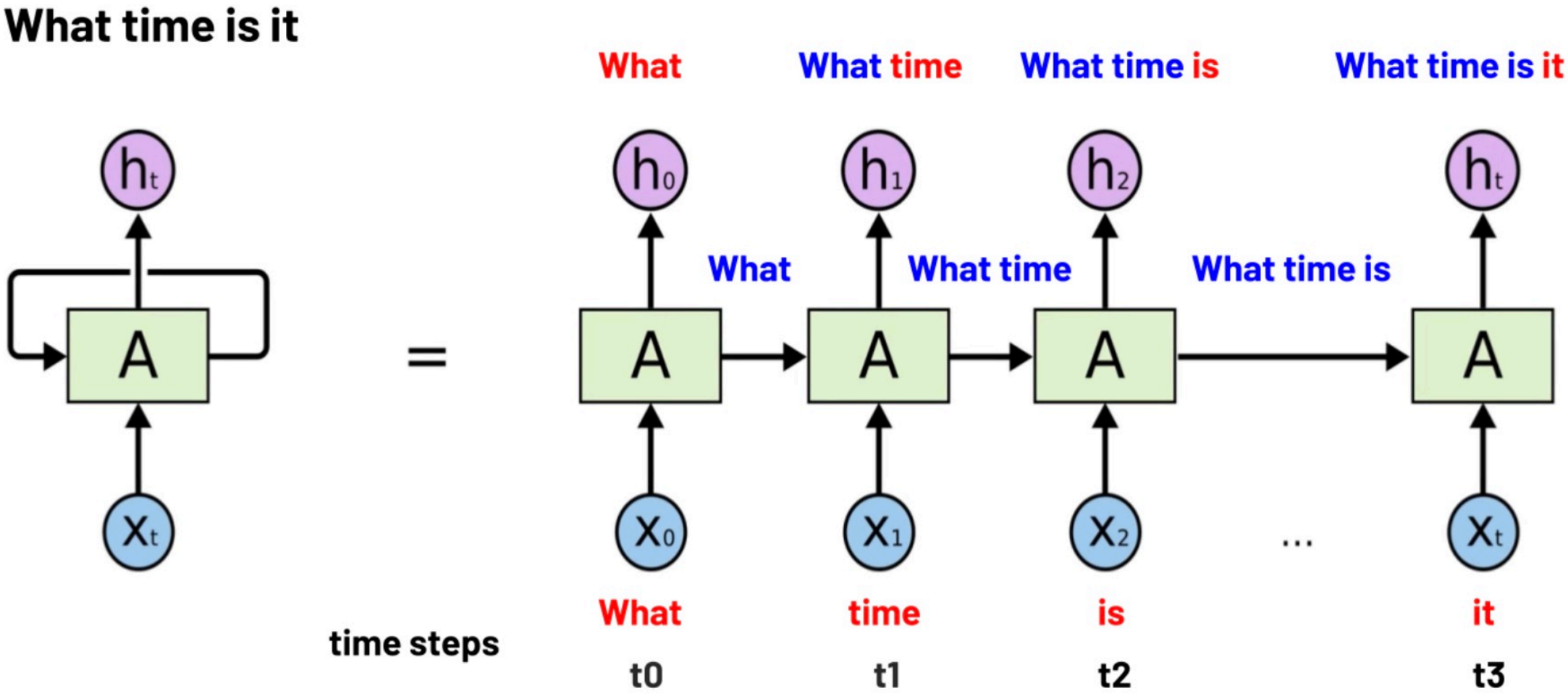
hp
130



Feed-Forward Neural Network

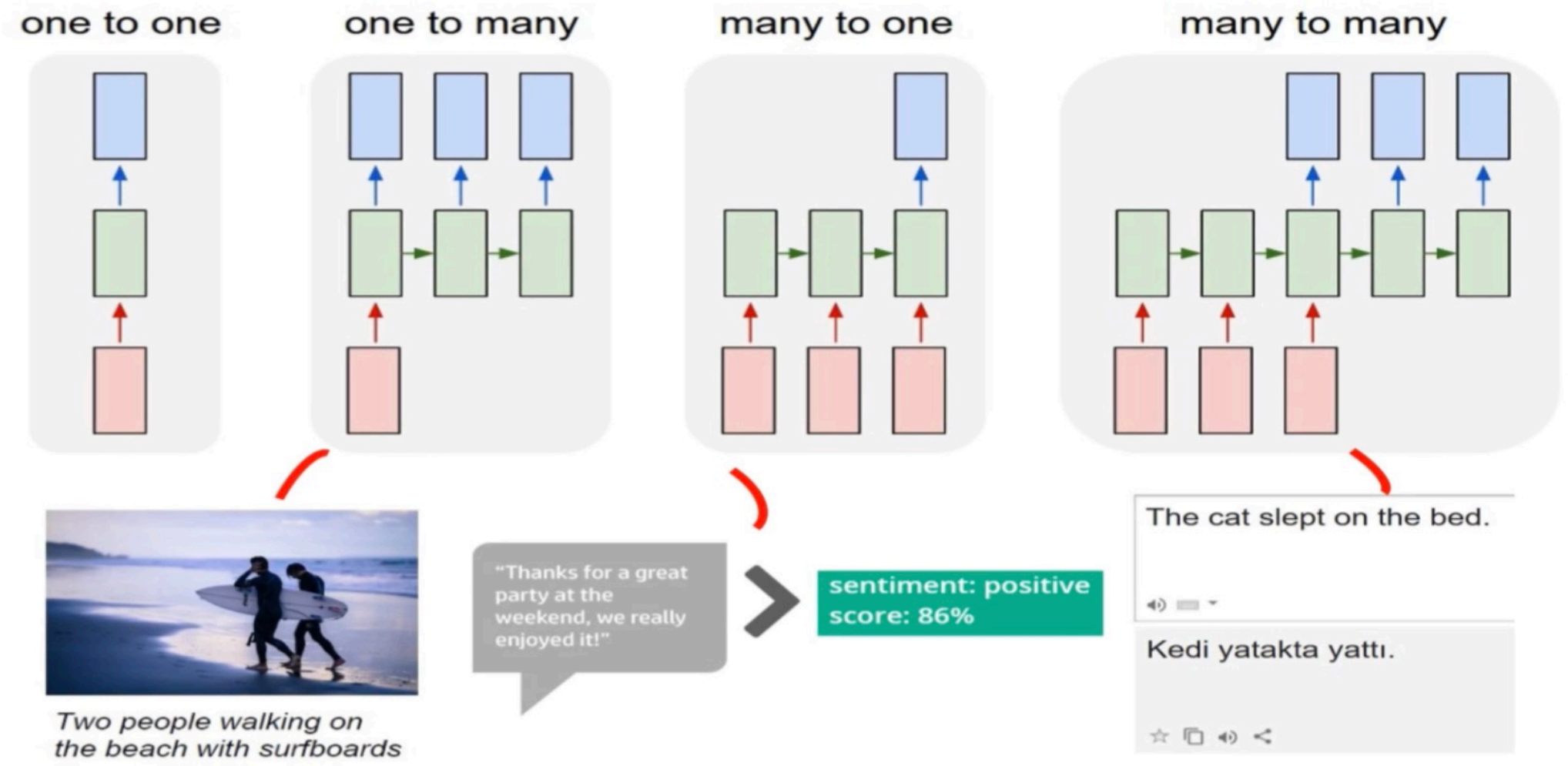
NLP with Deep Learning

Recurrent Neural Networks (RNN)



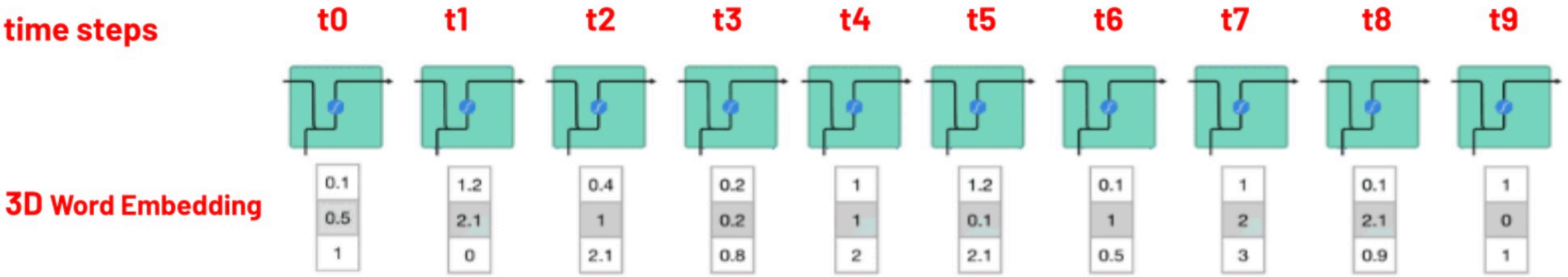
NLP with Deep Learning

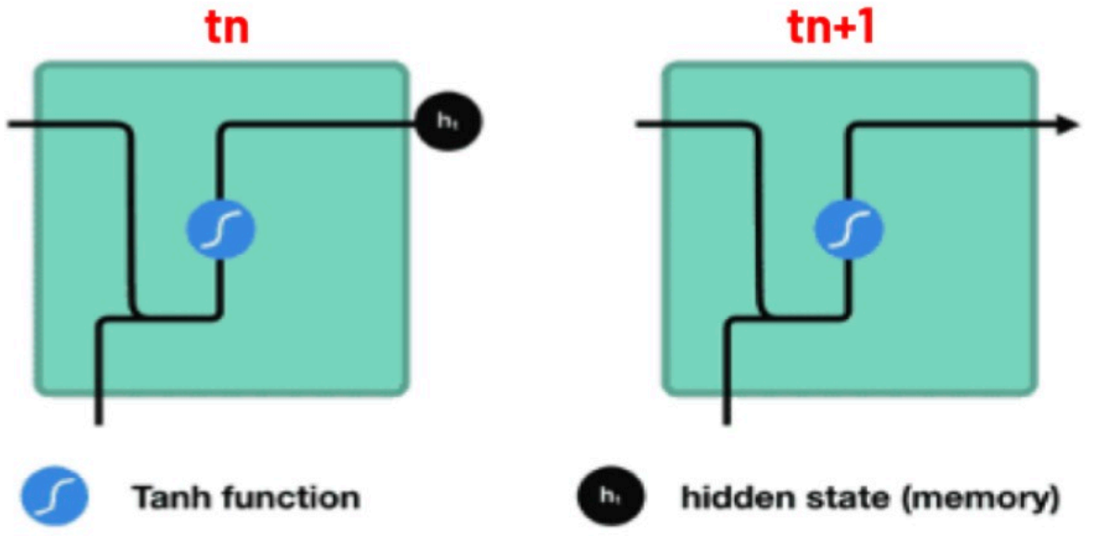
Recurrent Neural Networks (RNN)



NLP with Deep Learning

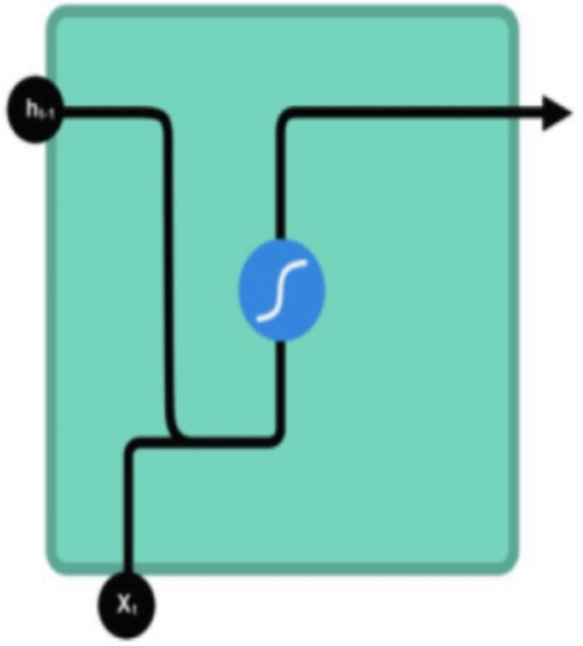
Recurrent Neural Networks (RNN)





NLP with Deep Learning

Recurrent Neural Networks (RNN)



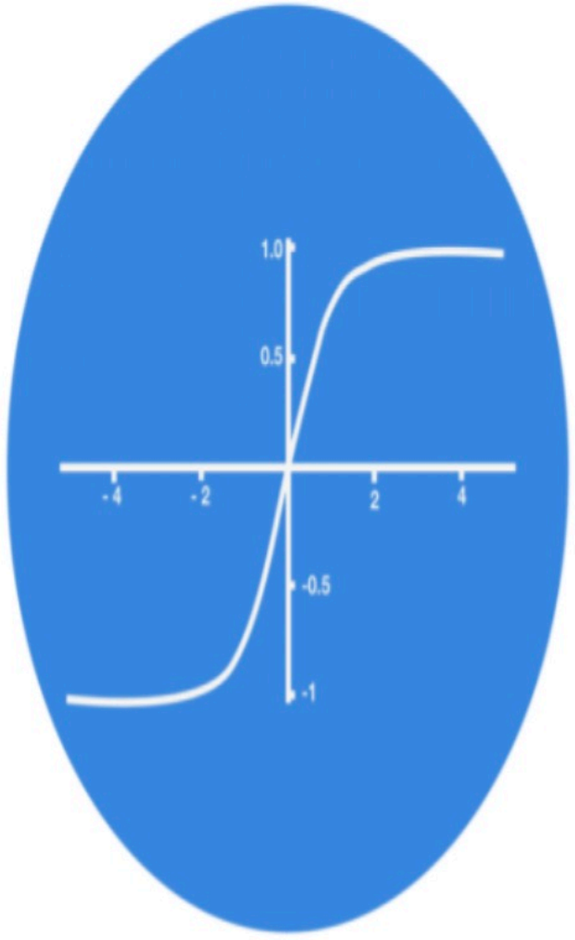
NLP with Deep Learning

Recurrent Neural Networks (RNN) - Tanh Activation

5

0.1

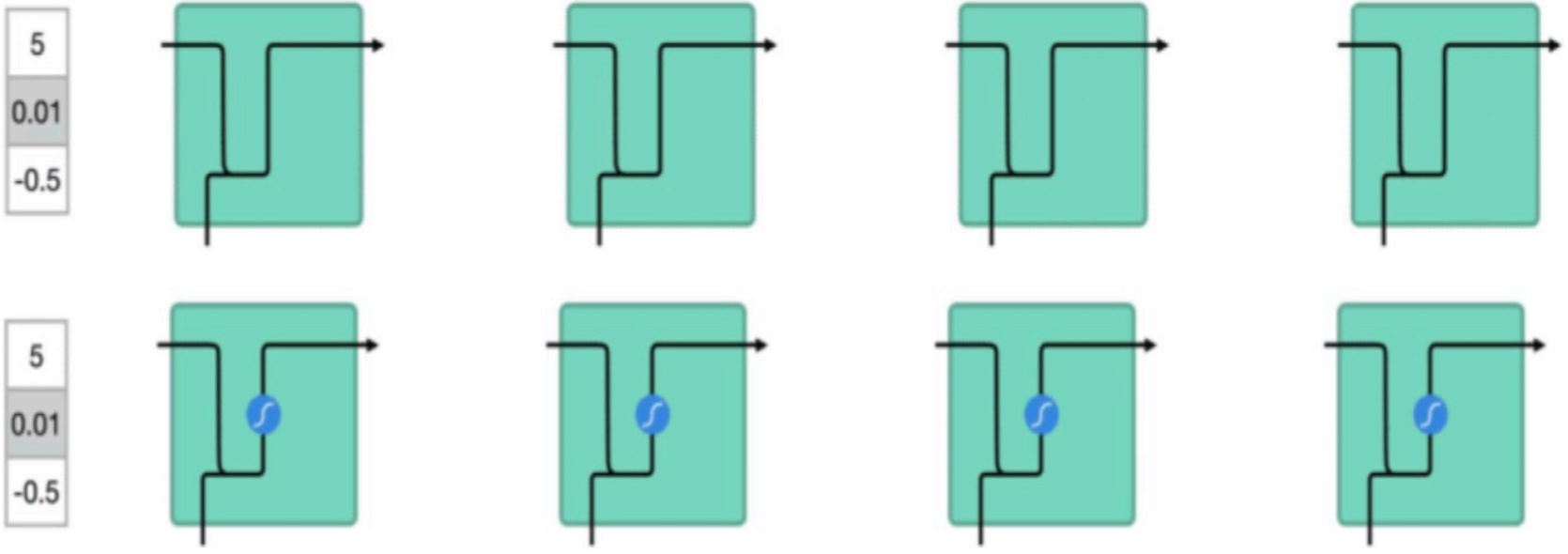
-0.5



The tanh activation is used to help regulate the values flowing through the network. The tanh function squishes values to always be between -1 and 1.

NLP with Deep Learning

Recurrent Neural Networks (RNN)



If Tanh is not used as activation function in RNN. The vector values become too large so the model becomes unstable and training stops.

NLP with Deep Learning

RNN - Sequential Memory

The sequential memory is a mechanism that makes it easier for your brain to recognize sequence patterns.

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

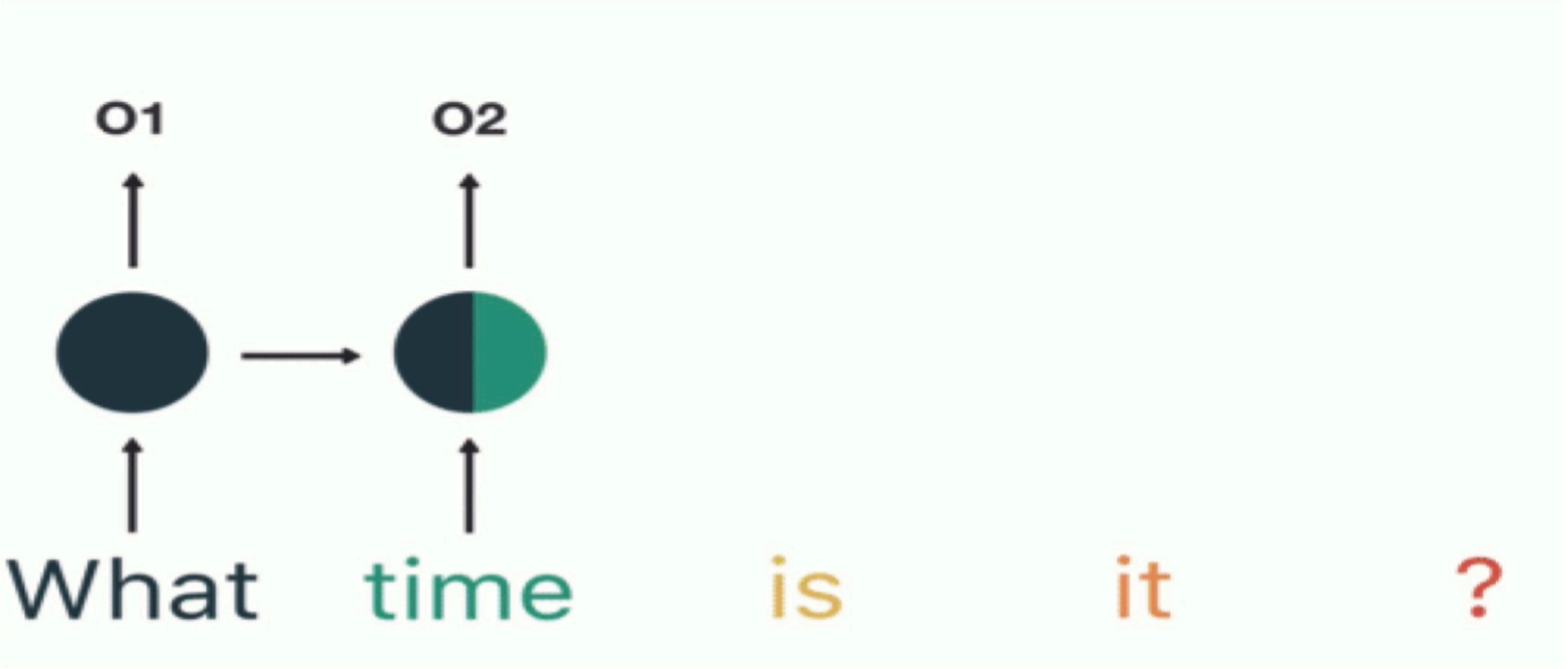
Z Y X W V U T S R Q P O N M L K J I H G F E D C B A

I met my best friend yesterday.

Best yesterday I friend my met.

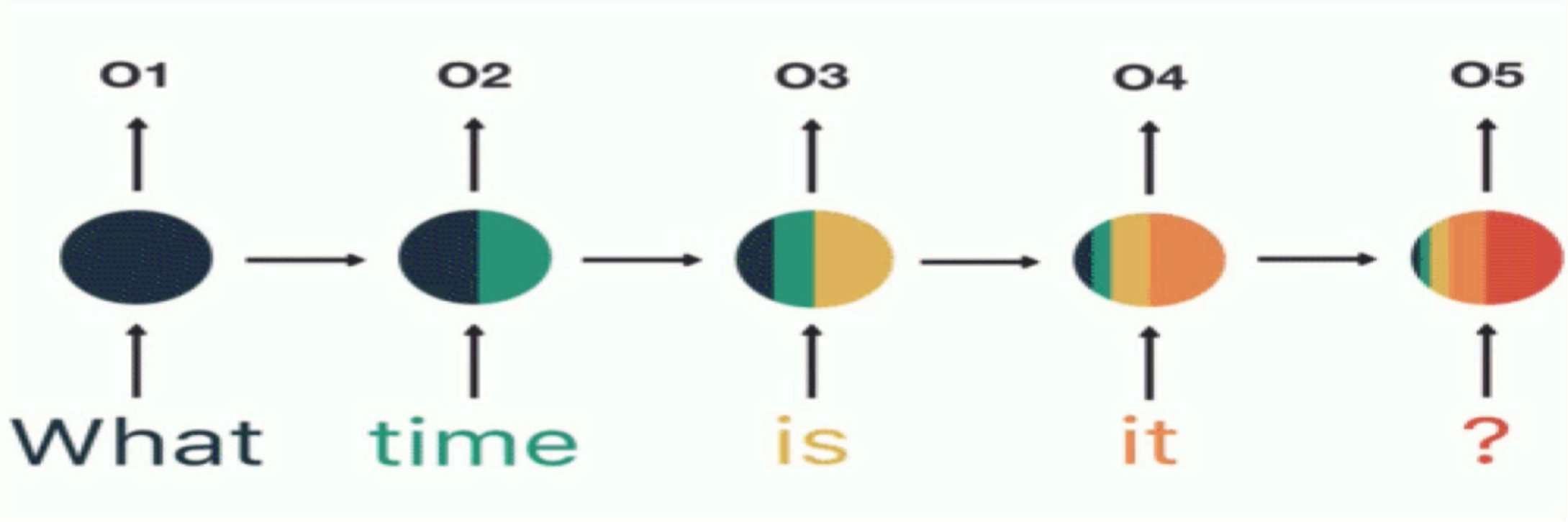
NLP with Deep Learning

RNN - Sequential Memory



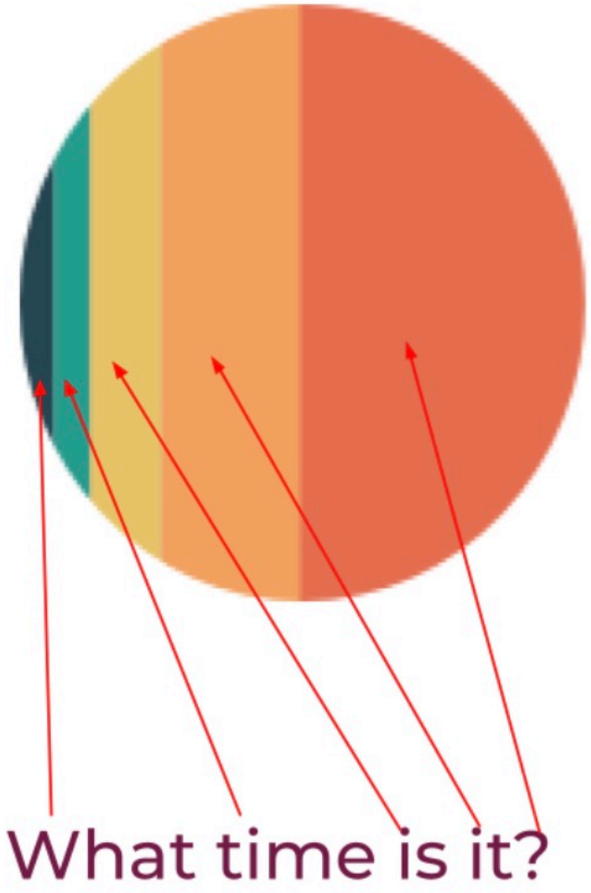
NLP with Deep Learning

RNN - Sequential Memory



NLP with Deep Learning

RNN - Short-Term Memory, Vanishing Gradient



1- Short-term memory is caused by the infamous vanishing gradient problem.

2- As the RNN processes more steps, it has troubles retaining information from previous steps.

3- As you can see, the information from the word “what” and “time” is almost non-existent at the final time step

NLP with Deep Learning

RNN - Vanishing Gradient

updated new weight =

old weight – learning rate * gradient

1.epoch 1.01 = 1.01001 – 0.00001

.

.

10.epoch 1.00898 = 1.009 – 0.00001

.

.

.

200.epoch 1.008789 = 1.00880 – 0.000001

1- During back propagation, recurrent neural networks suffer from the vanishing gradient problem.

2- Gradients are values used to update a neural networks weights.

3- The vanishing gradient problem is when the gradient shrinks as it back propagates through time.

4- If a gradient value becomes extremely small, it doesn't contribute too much

NLP with Deep Learning

RNN - Vanishing Gradient

What time is it?

5- So in recurrent neural networks, layers that get a small gradient update stops learning.

6- Those are usually the earlier layers.

7- So because these layers don't learn, RNN's can forget what it seen in longer sequences, thus having a **short-term memory**.

8-RNNs work very well in short sentences and very poorly in long sentences

NLP with Deep Learning

RNN - Exploding Gradient

updated new weight = old weight – learning rate * gradient

1.epoch -8.98999 = 1.01001 – 10

.

.

10.epoch -19.569.676 = -9.000.899– 10.568.777

.

.

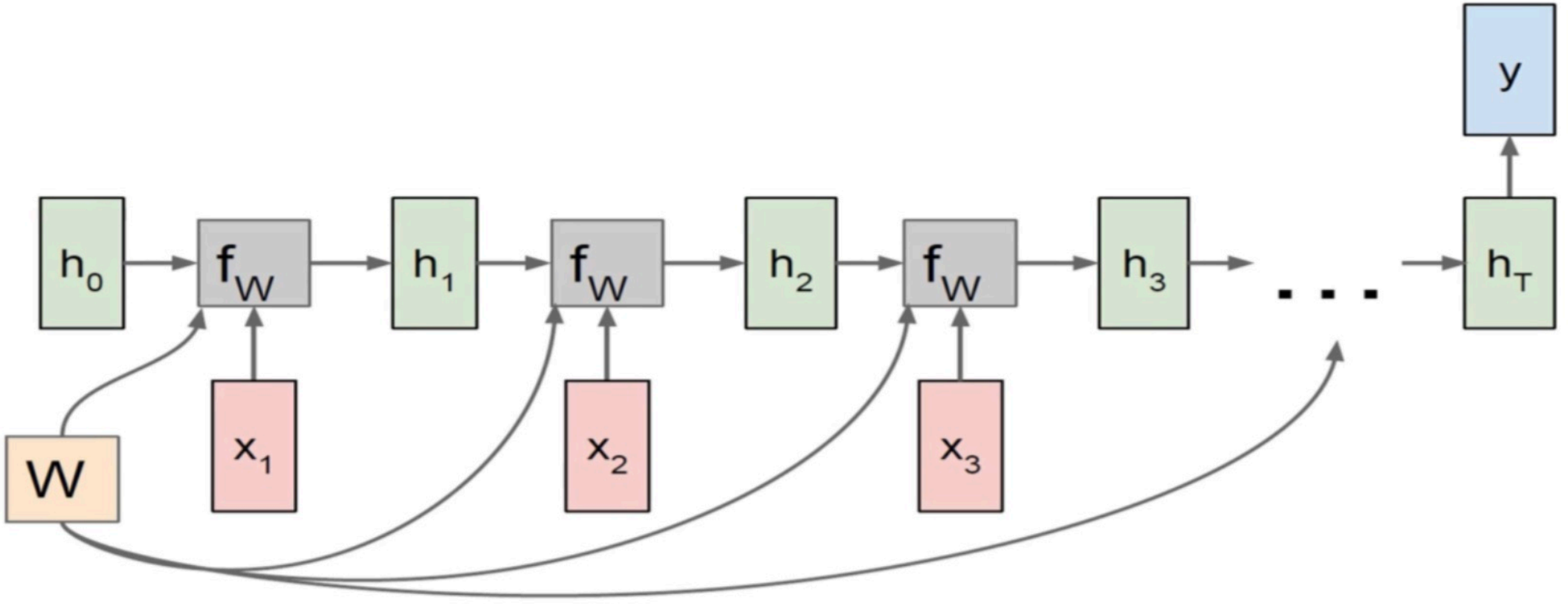
200.epoch NAN

Exploding Gradient is a phenomenon that occurs during the gradient descent optimization in deep learning models.

This is result in the calculation of a **very large gradient** and **prevent the completion of the training process**

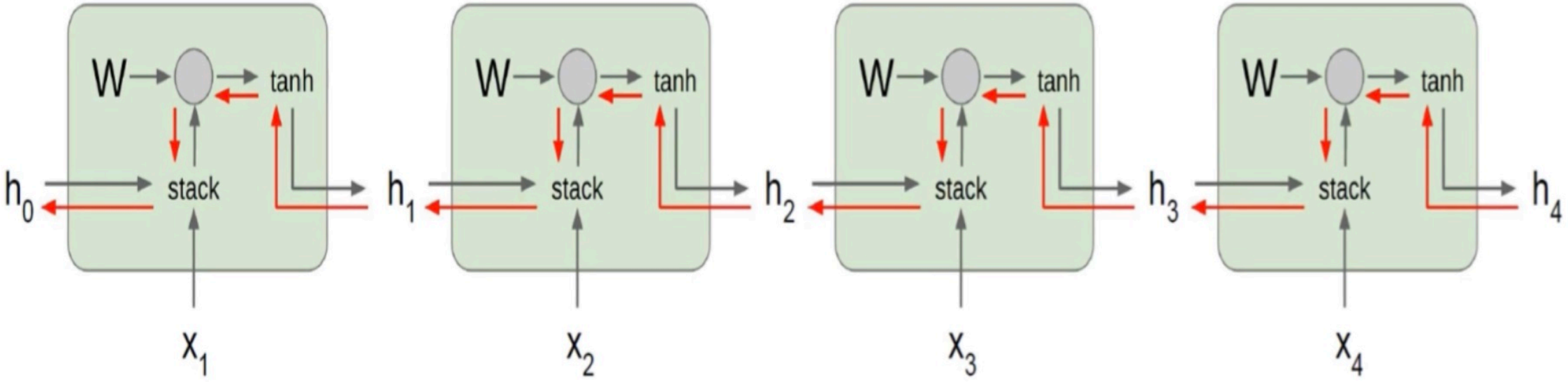
NLP with Deep Learning

RNN - Vanishing & Exploding Gradient



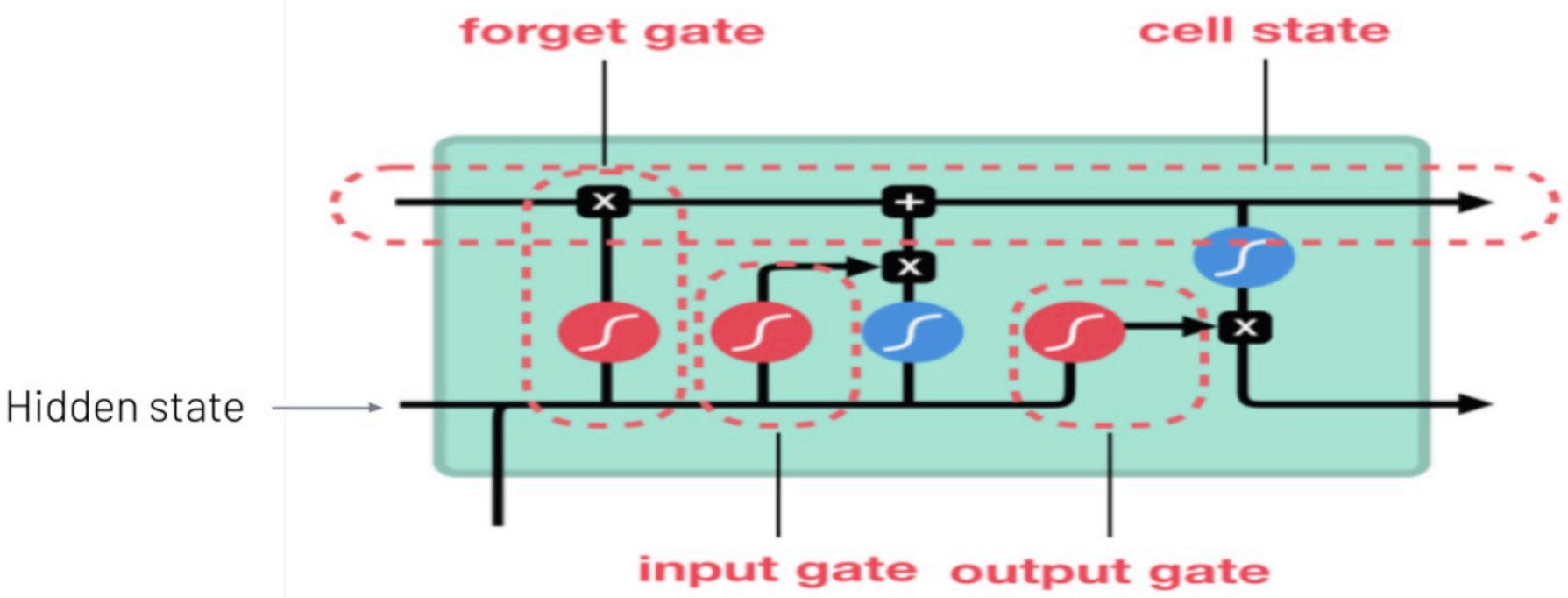
NLP with Deep Learning

RNN - Vanishing & Exploding Gradient



NLP with Deep Learning

Long Short-Term Memory (LSTM)

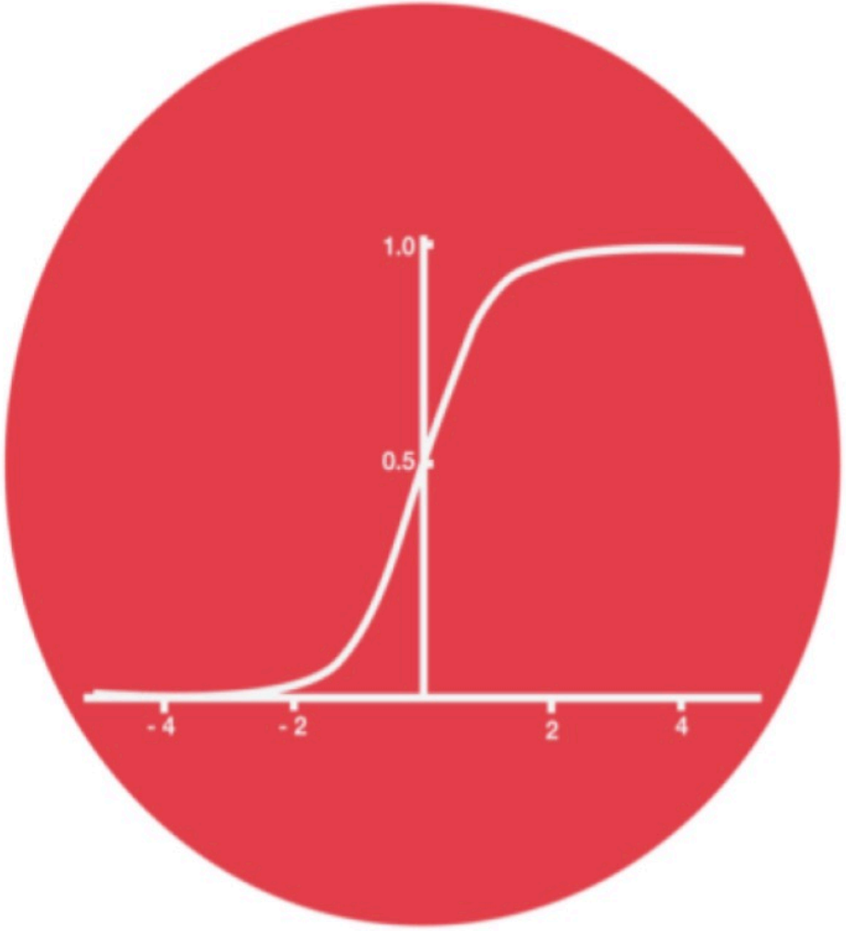


It is the enhanced variant of RNNs. It was started to be used in 1995-1997 and was developed to solve the vanishing gradient problem of RNNs. LSTM has 3 gates (forget, input and output)

NLP with Deep Learning

Long Short-Term Memory (LSTM)-Sigmoid Act.

5
0.1
-0.5



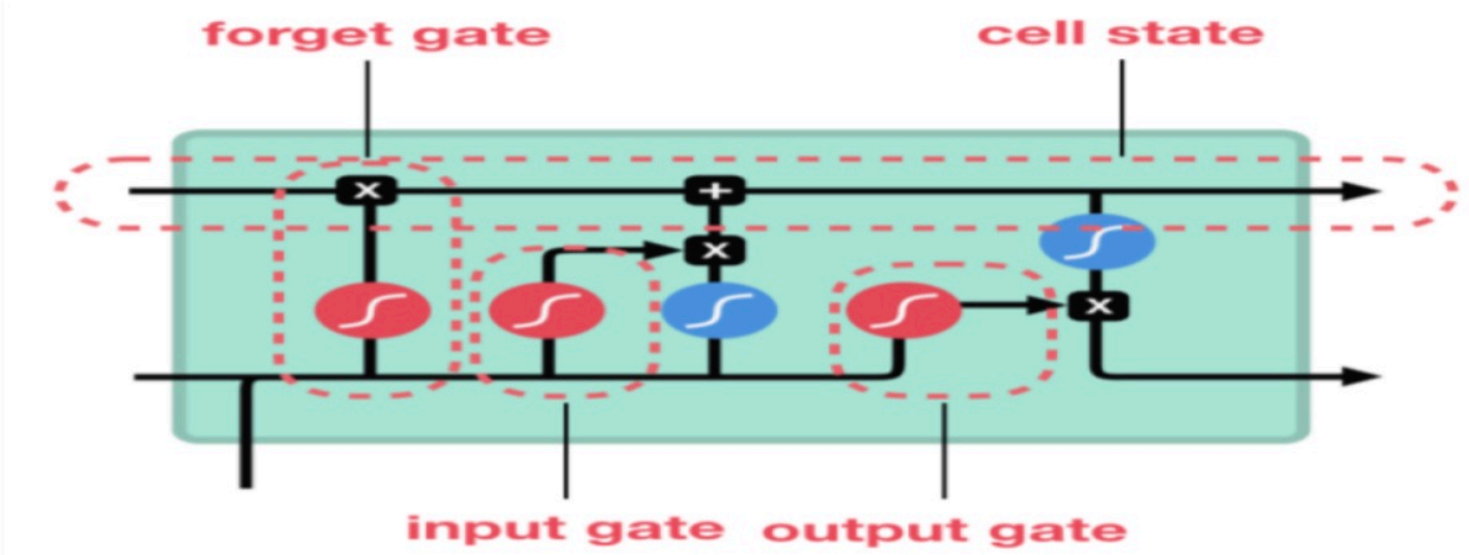
NLP with Deep Learning

Long Short-Term Memory (LSTM)-Sigmoid Act.

- Gates contains sigmoid activations.
- A sigmoid activation is similar to the tanh activation. Instead of squishing values between -1 and 1, it squishes values between 0 and 1.
- That is helpful to update or forget data because any number getting multiplied by 0 is 0, causing values to disappears or be “forgotten.”
- Any number multiplied by 1 is the same value therefore that value stay’s the same or is “kept.”

NLP with Deep Learning

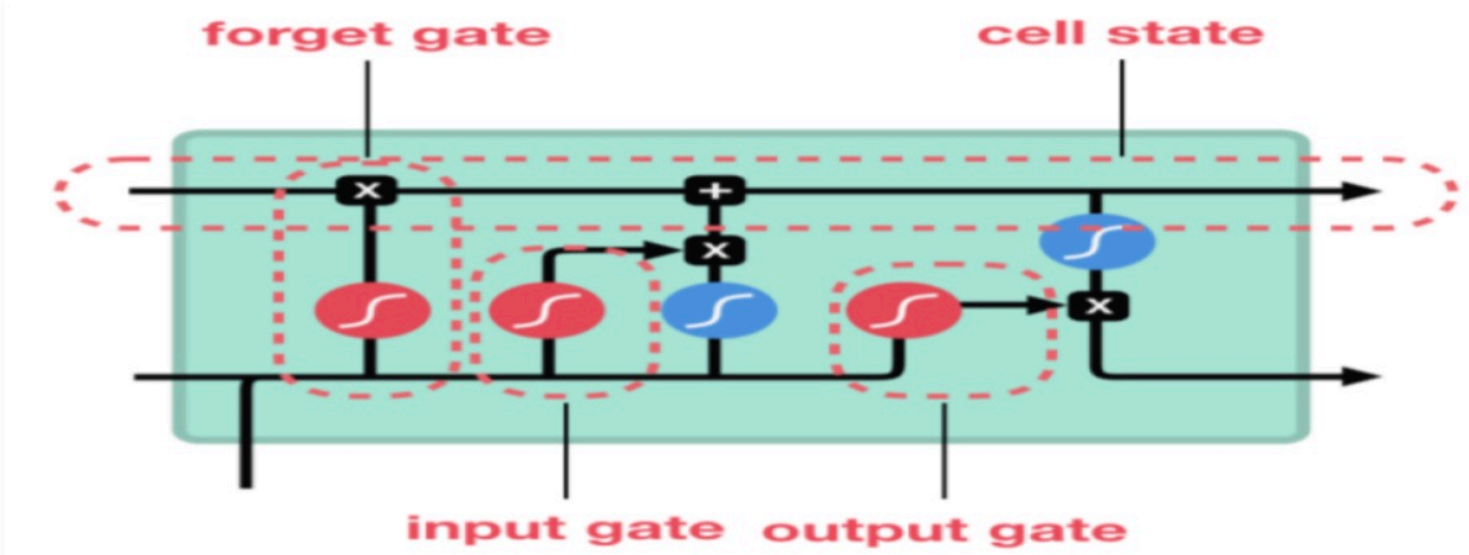
Long Short-Term Memory (LSTM)



Input gate: In the first stage, the current input on the cell state is passed through into the sigmoid function and then into the tanh function. Since the results from the sigmoid function will be between 0 and 1, it decides whether the information from the tanh function will be added to the cell state or not.

NLP with Deep Learning

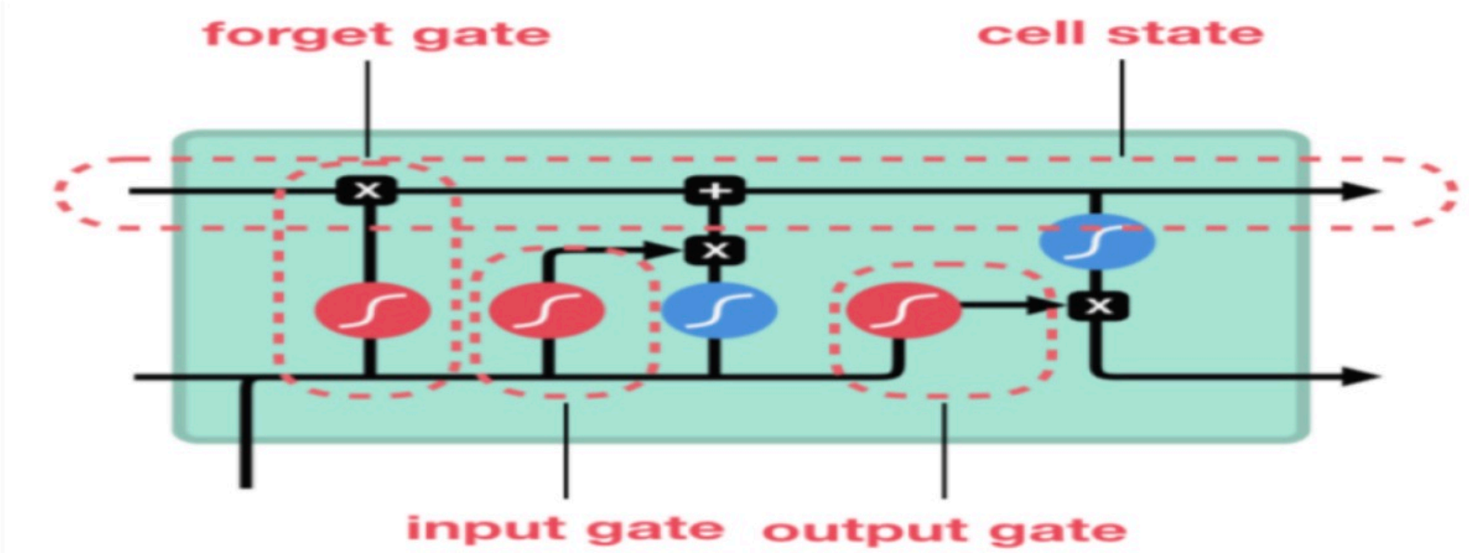
Long Short-Term Memory (LSTM)



Forget Gate : This gate decides what information should be thrown away or kept. Information from the previous hidden state and information from the current input is passed through the sigmoid function. Values come out between 0 and 1. The closer to 0 means to forget, and the closer to 1 means to keep.

NLP with Deep Learning

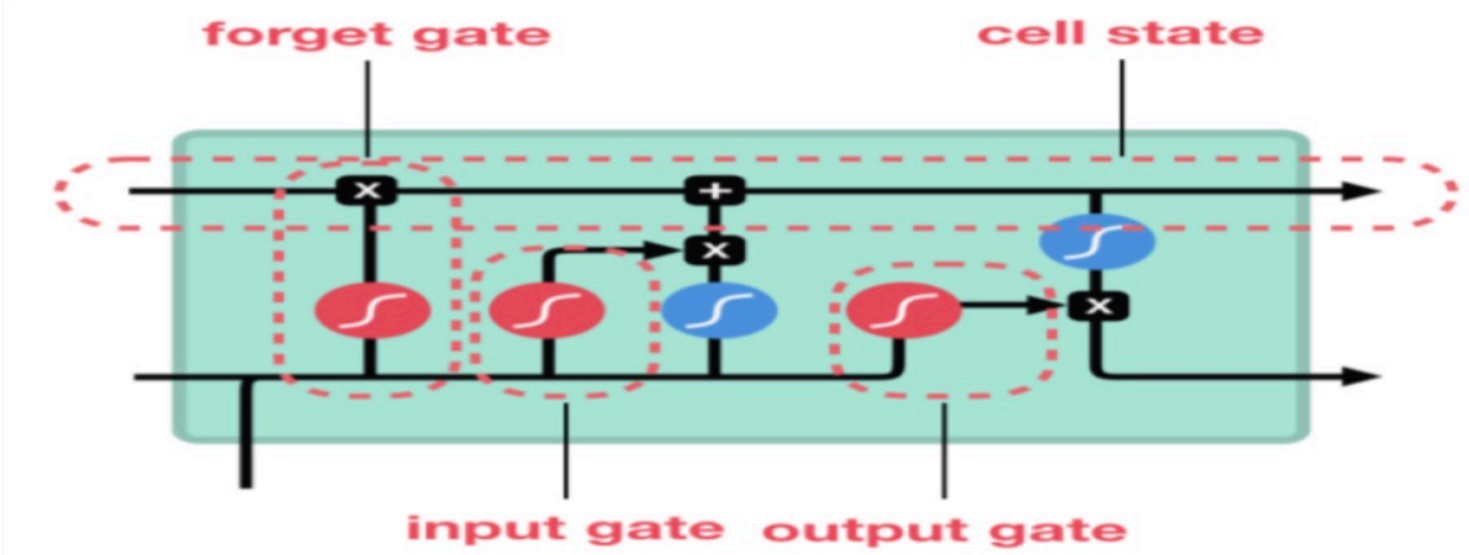
Long Short-Term Memory (LSTM)



Cell state: The cell state carries relevant information throughout the sequence. Thus, previous information isn't forgotten and the model can make more accurate predictions. Therefore, you can think of it as the memory of the network.

NLP with Deep Learning

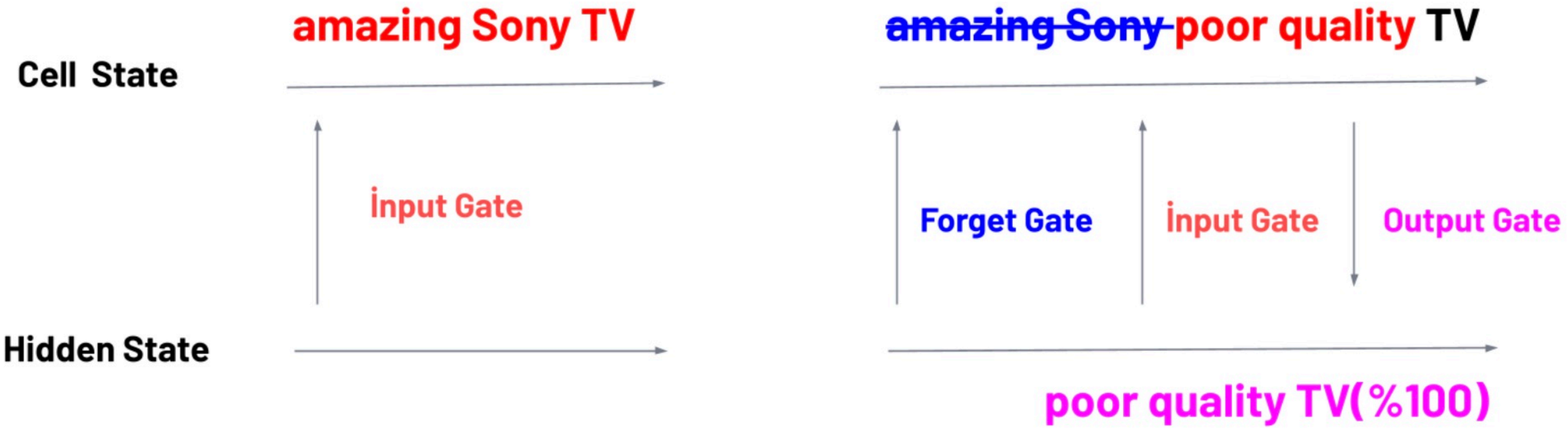
Long Short-Term Memory (LSTM)



Output gate: It is used to determine the hidden state to be transferred to the next step. Since the hidden state contains information about previous inputs (words, letters, etc.), it is used for **predictions** that the model will make.

NLP with Deep Learning

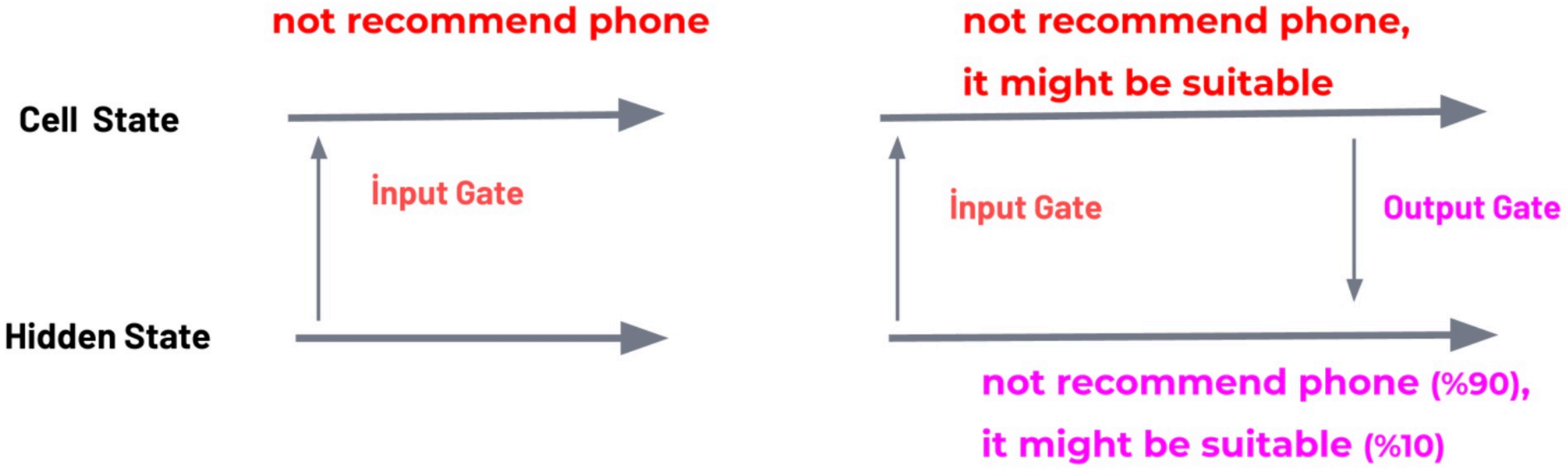
Long Short-Term Memory (LSTM)



My first purchase from Best Buy was an **amazing Sony TV**. However, over time, I realized that it was a **poor quality TV**

NLP with Deep Learning

Long Short-Term Memory (LSTM)



I definitely do **not recommend** this **phone**. But **it might be suitable** for children.

NLP with Deep Learning

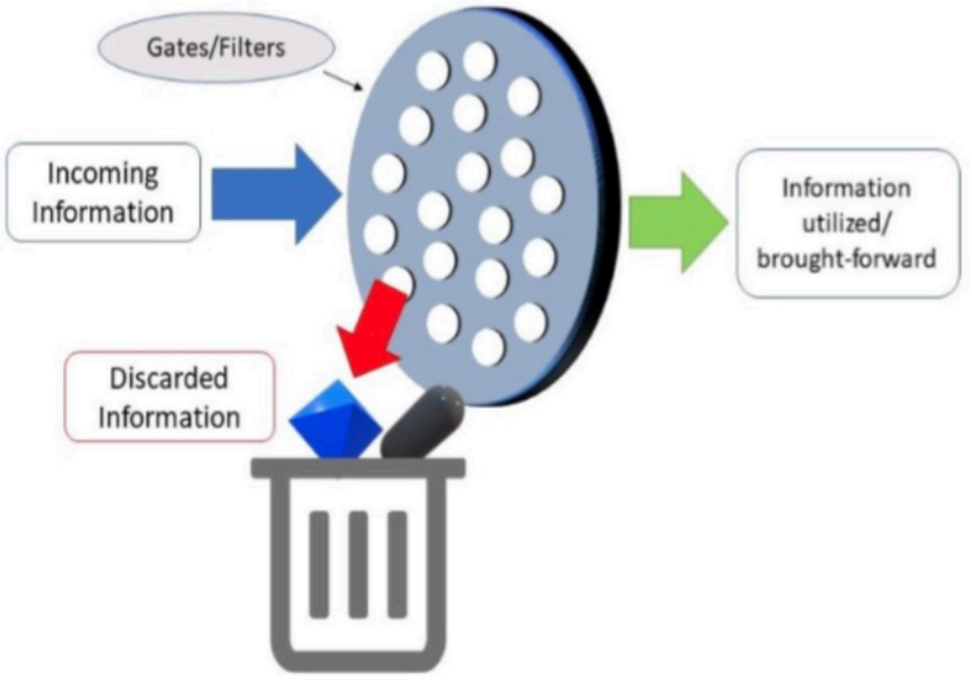
Long Short-Term Memory (LSTM)

To summarize briefly;

The forget gate decides whether to continue transferring the information from the previous steps to the cell state.

The input gate decides whether or not to add to the cell state a piece of new information.

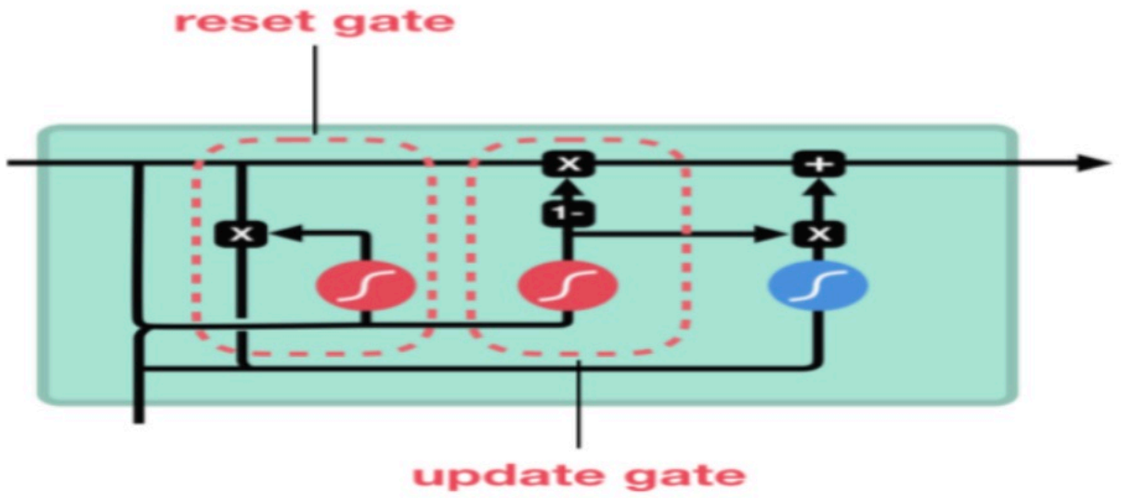
The output gate determines the next hidden state.



LSTM Gates can be seen as filters

NLP with Deep Learning

Gated Recurrent Units (GRU)

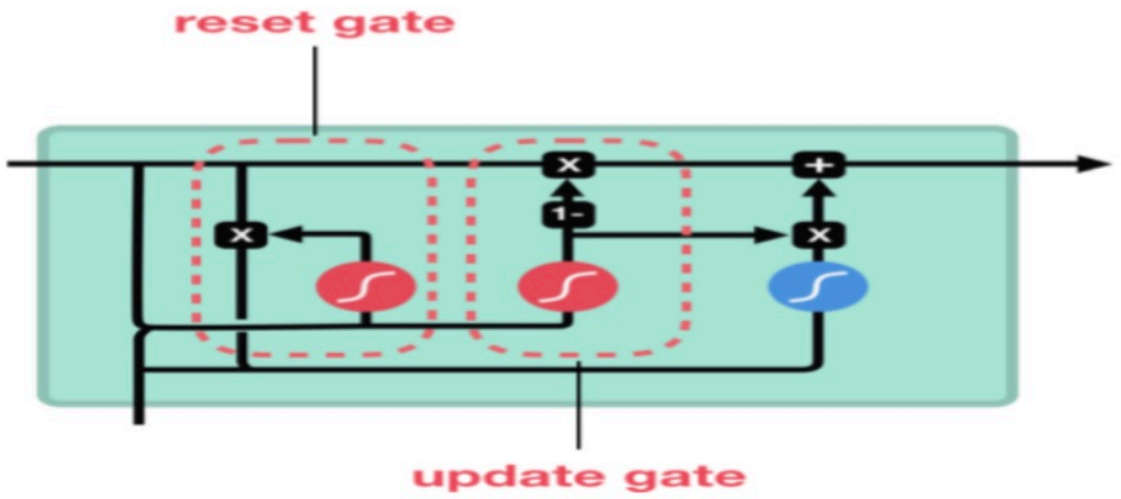


It is the enhanced variant of RNNs. It was started to be used in 2014 and was developed to solve the vanishing gradient problem of RNNs.

The GRU is the newer generation of Recurrent Neural networks and is pretty similar to an LSTM. GRU has 2 gates(reset and update)

NLP with Deep Learning

Gated Recurrent Units (GRU)

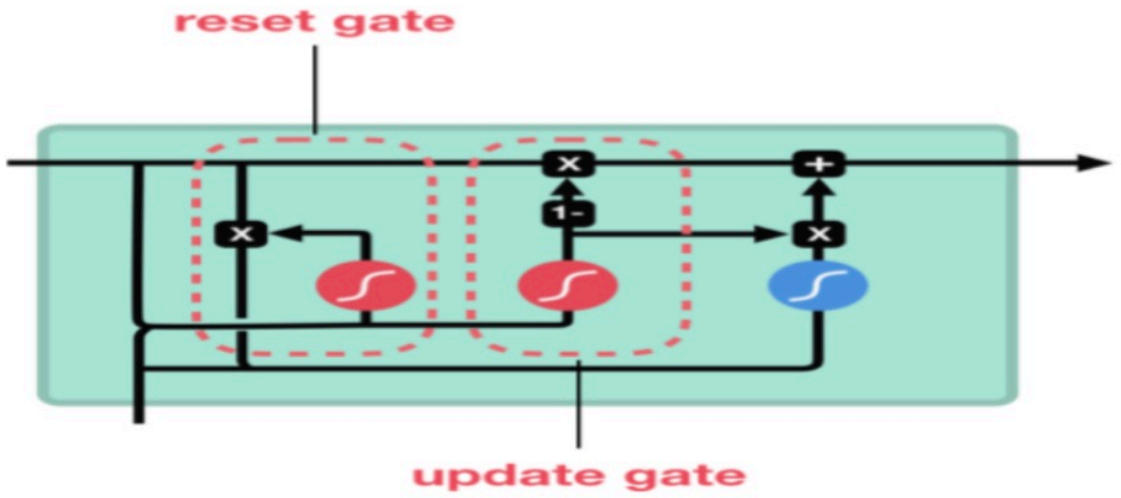


Also, GRUs don't have cell state. Instead, the hidden state and the cell state are combined.

Update gate: It decides what information to discard and what information to include, such as the forget and input gates in LSTM.

NLP with Deep Learning

Gated Recurrent Units (GRU)



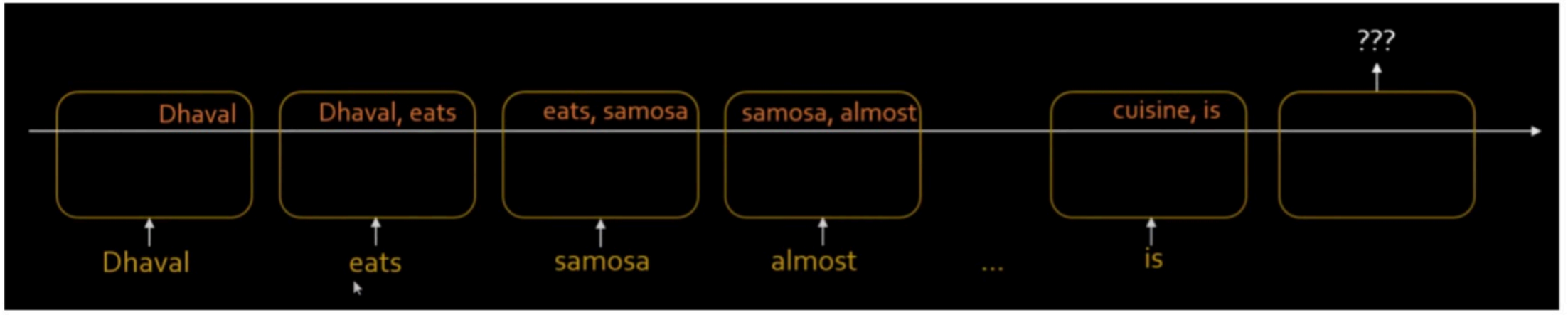
Reset gate: It is the gate that decides how much of the information from other steps should be forgotten.

Although **GRU is faster than LSTM**, **both LSTM, and GRU give very good results**. It's up to you to try both and decide.

NLP with Deep Learning

Working Logic of RNN

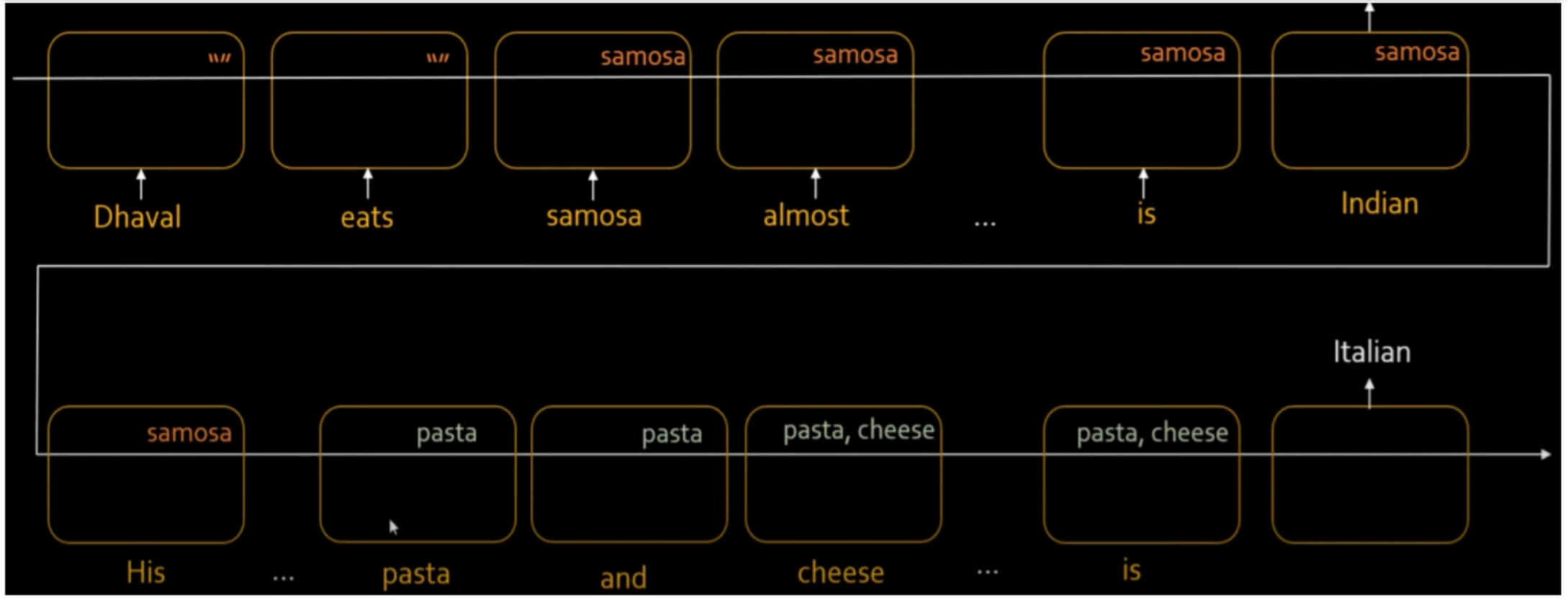
Dhaval eats **samosa** almost everyday, it shouldn't be hard to guess that his favorite cuisine is **Indian**



NLP with Deep Learning

Working Logic of LSTM and GRU

Dhaval eats **samosa** almost everyday, it shouldn't be hard to guess that his favorite cuisine is **Indian.** His brother Bhavin however is a lover of **pastas** and **cheese** that means Bhavin's favorite cuisine is **Italian**

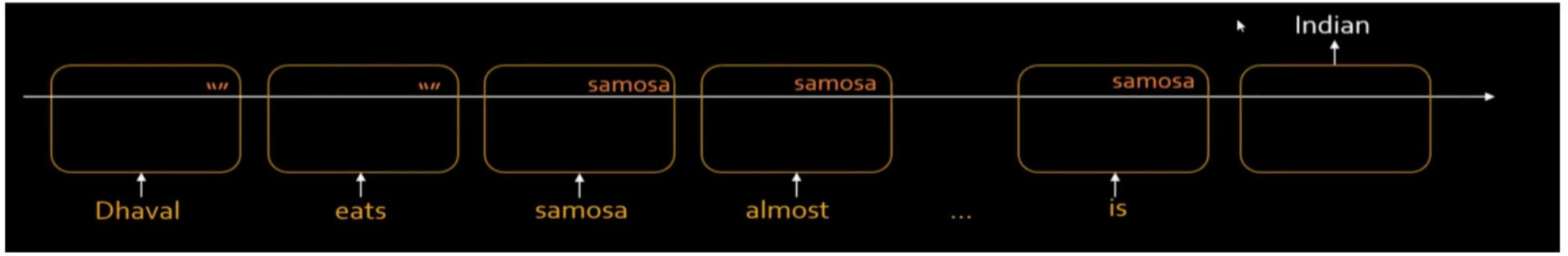


NLP with Deep Learning

Working Logic of LSTM and GRU

Dhaval eats **samosa** almost everyday, it shouldn't be hard to guess that his favorite cuisine is **Indian**

Cell State



Hidden State