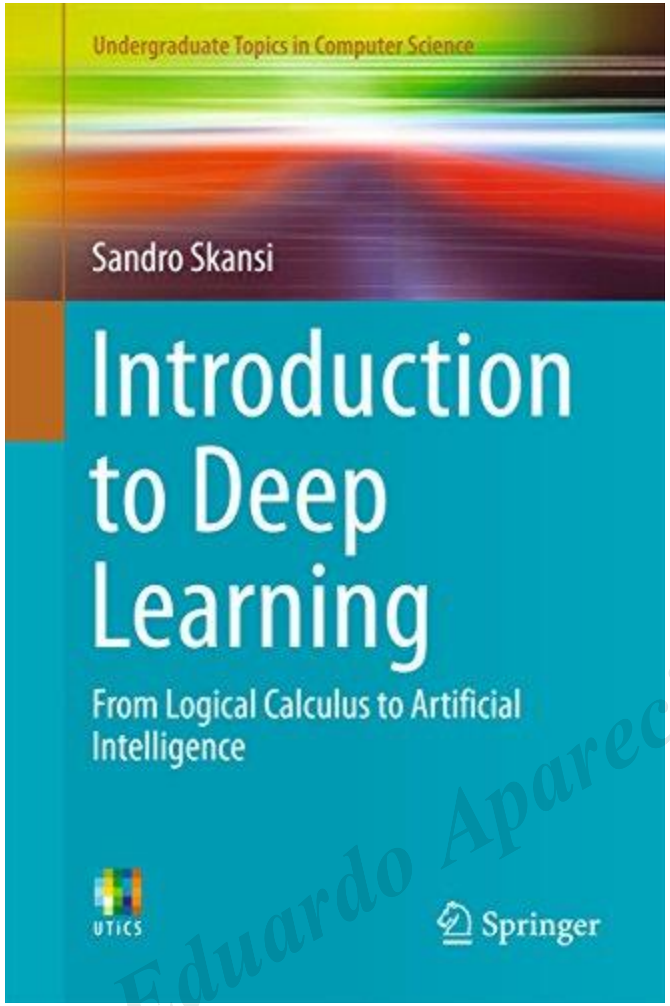
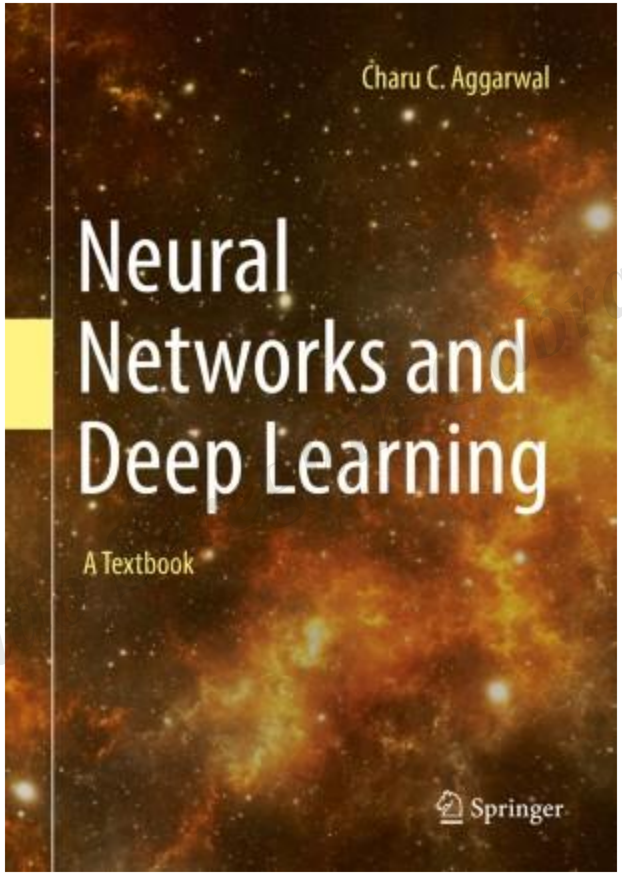
MBA USP ESALQ

Deep Learning

Prof. Dr. Jeronymo Marcondes

Introduction





MBA USP ESALQ

Introduction

- Some important problems:

1. Text data

2. Time Series

3. Watch a movie

RNN

- Example of a time serie.

Introduction

- What is the characteristic of these data? Sequence.

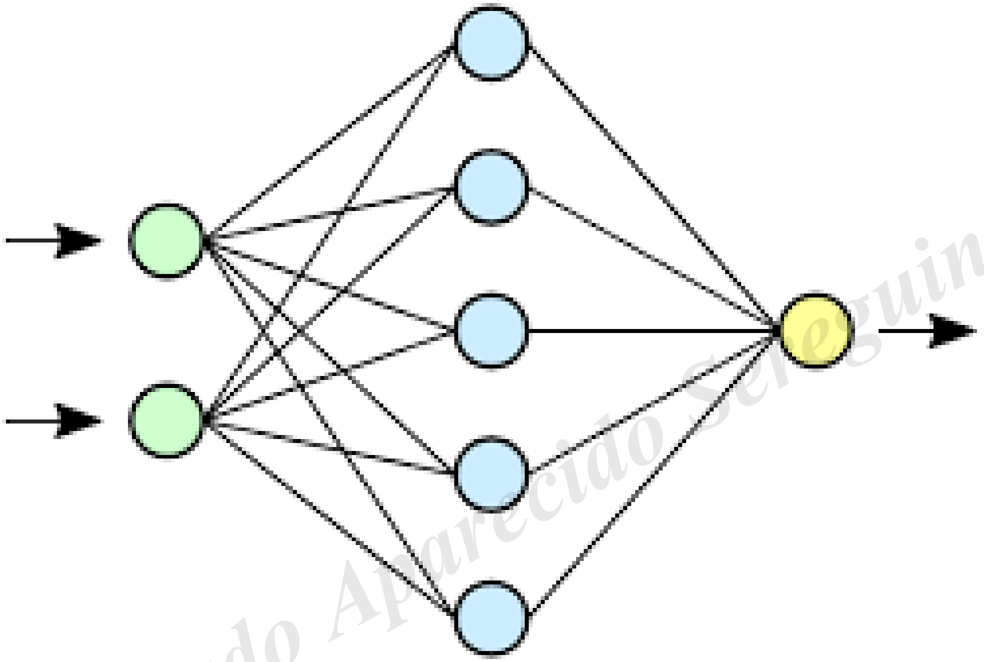
- FNN can be used to reproduce this process, but it is not the best choice.

"The cat chased the mouse"

"The mouse chased the cat"

RNN

- What would happen if we use a FNN?



RNN

(‘$’,‘all’)
(‘$ all’,‘I’)
(‘$ all I’,‘want’)
(‘$ all I want’, ‘for’)
(‘$ all I want for’, ‘Christmas’)
(‘$ all I want for Christmas’, ‘is’)
(‘$ all I want for Christmas is’, ‘you’)
(‘$ all I want for Christmas is you’, ‘&’).

Source: Introduction to Deep Learning from Logical Calculus to Artificial Intelligence

RNN

- RNN builds probability distribution

- Example

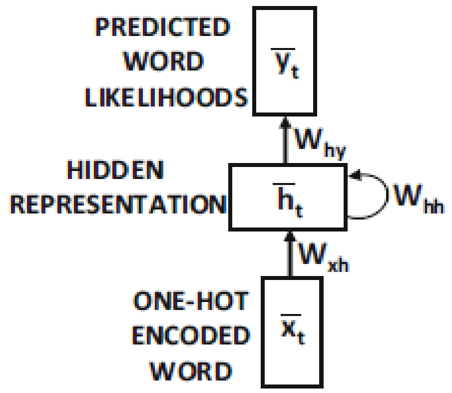
'My name is Cassidy'
'My name is Myron'
'My name is Marcus'
'My name is Marcus'
'My name is Marcus'.

Source: Introduction to Deep Learning from Logical Calculus to Artificial Intelligence

RNN

- Goal: neural network with "memory".

- Recurrent: it performs the same task for all elements and their output depends on the previous calculations.

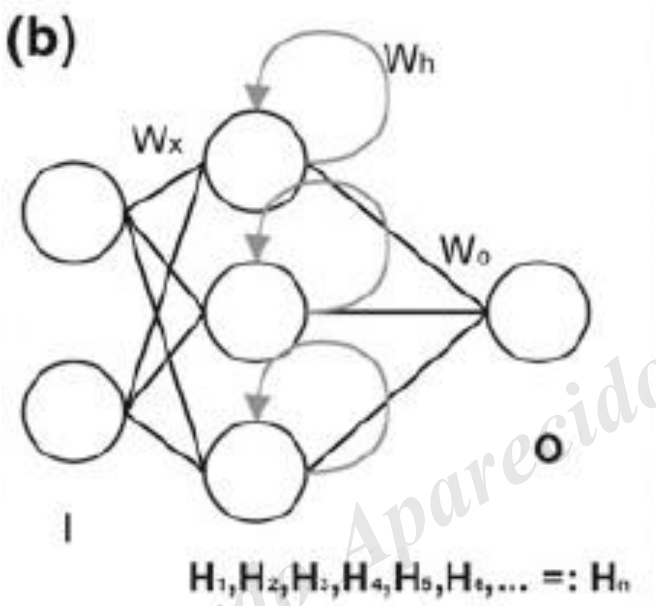


(a) RNN

Source: Neural Networks and Deep Learning

RNN

- We can visualize RNN in another way:



Source: Introduction to Deep Learning from Logical Calculus to Artificial Intelligence

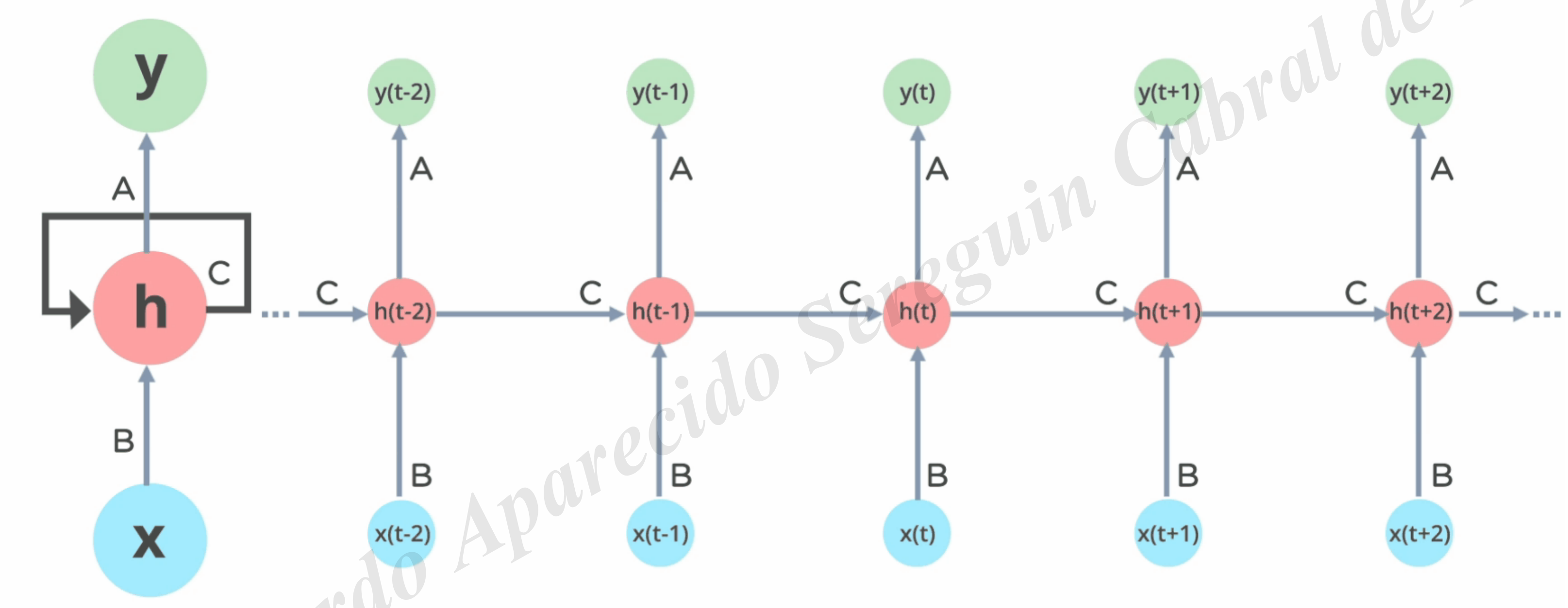
RNN

The input layer 'x' receives the input to the neural network, and processes and transfer it to the intermediate layer.

The intermediate layer 'h' can consist in several hidden layers, each one with their own activation functions, weights and biases.

Source: https://www.simplilearn.com/tutorials/deep-learning-tutorial/rnn

RNN

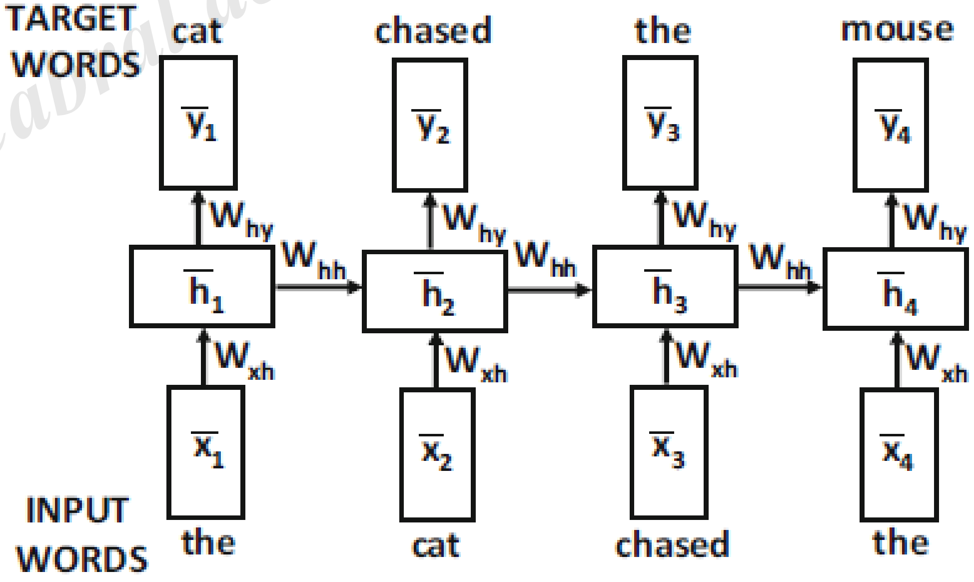


Source: https://www.simplilearn.com/tutorials/deep-learning-tutorial/rnn

RNN

- We can visualize RNN in another way:
- The cat chased the mouse

Source: Neural Networks and Deep Learning



(b) Time-layered representation of (a)

RNN

- This means that we have a network that "keeps" all the past
- Explain economic growth based on the trust level.
- Simple Recurrent Neural Network

$$h(t) = f_h(\mathbf{w}_h^\top h(t-1) + \mathbf{w}_x^\top x(t))$$
$$y(t) = f_o(\mathbf{w}_o^\top h(t)),$$

RNN

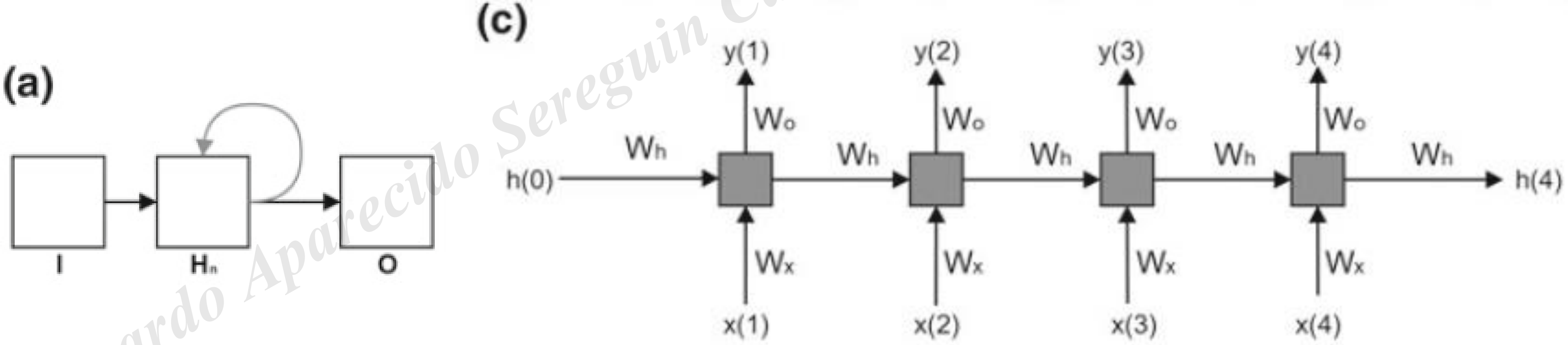
- Elman Network

$$
\begin{aligned}
y(t) &= f(\mathbf{w}_o^\top h(t)) = && (7.1)\\
&= f(\mathbf{w}_o^\top f(\mathbf{w}_h^\top h(t-1) + \mathbf{w}_x^\top x(t))) = && (7.2)\\
&= f(\mathbf{w}_o^\top f(\mathbf{w}_h^\top f(\mathbf{w}_h^\top h(t-2) + \mathbf{w}_x^\top x(t-1)) + \mathbf{w}_x^\top x(t))) = && (7.3)\\
&= f(\mathbf{w}_o^\top f(\mathbf{w}_h^\top f(\mathbf{w}_h^\top f(\mathbf{w}_h^\top h(t-3) + \mathbf{w}_x^\top x(t-2)) + \mathbf{w}_x^\top x(t-1)) + \mathbf{w}_x^\top x(t))). && (7.4)
\end{aligned}
$$

- We have the results that the values of the hidden layer will be multiplied by high weights values raised to greater powers as older the information is.

Backpropagation

- Everything is affected at the same time!
- Backpropagation Through Time



Backpropagation

- How is the calculation of the gradient?

- How much does the error vary for a given weight variation?

- A problem appears: past multiplied by weights raised to greater powers as older it is.

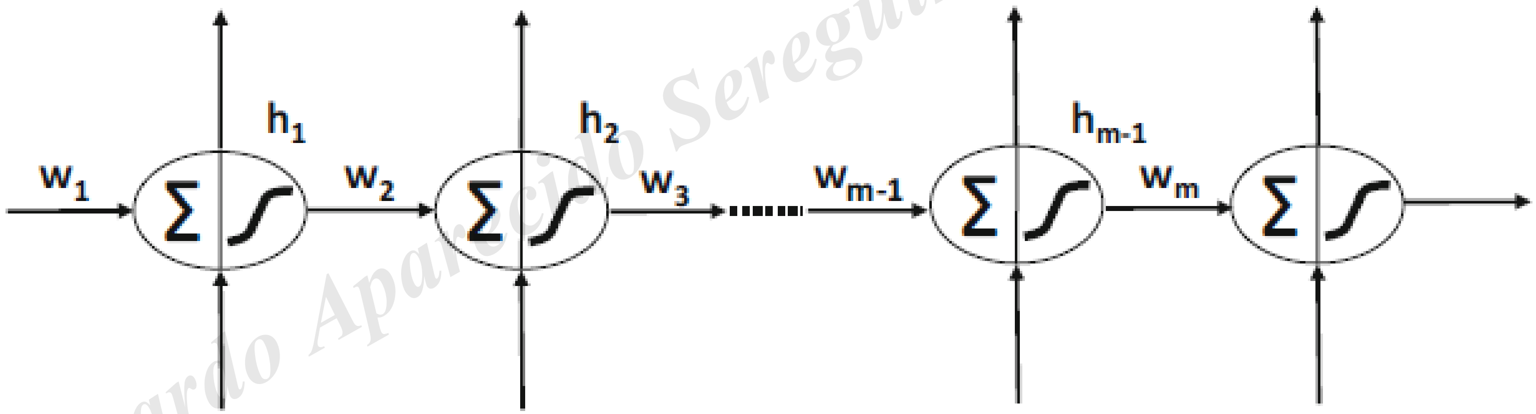
- Intuition: the further in the past, the harder to see the influence, because a lot things happened.

Gradient Vanishing Problem

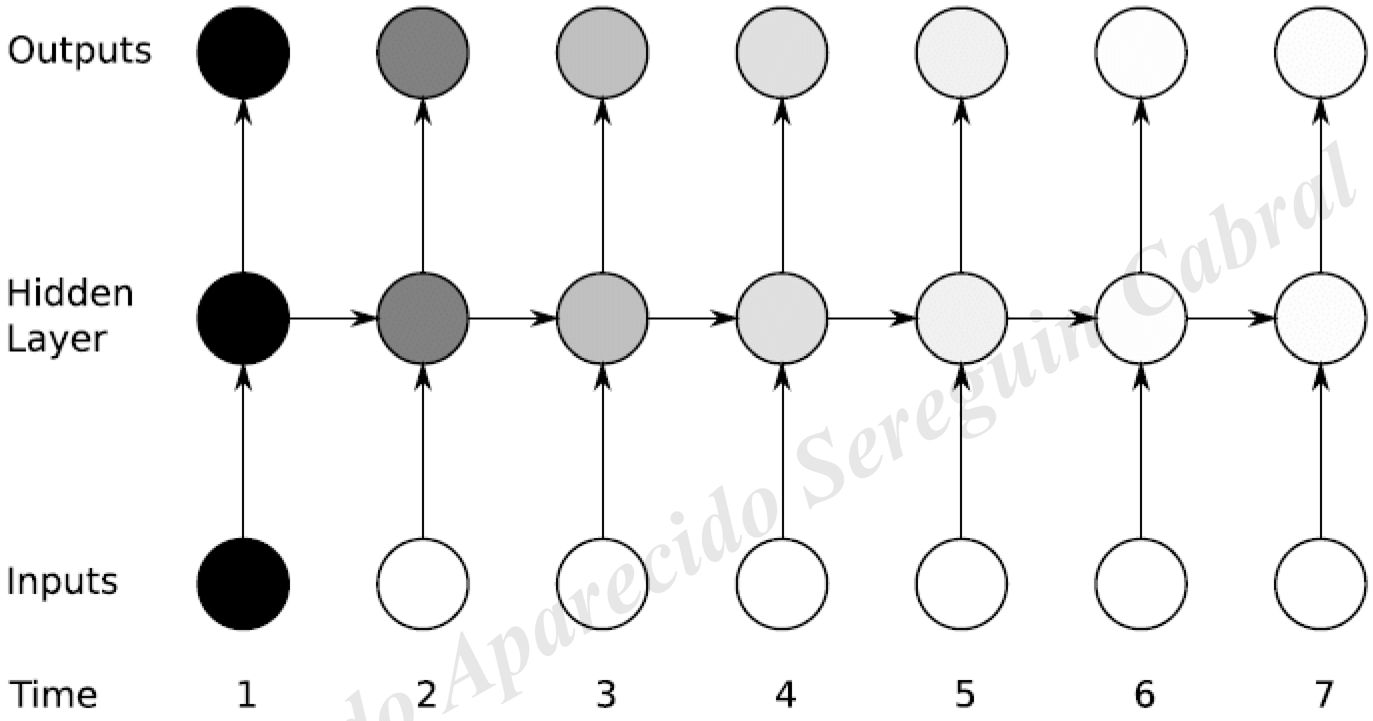
- Intuition
- Multiplication of numbers smaller than 1
- Multiplication of numbers greater than 1

Problems with Gradient

- "Disappearance" (Vanishing" and Gradient Explosion
- w>1 or w<1



Gradient Vanishing Problem



Problems with Gradient

- It occurs in any network - most common in RNN
- Is RNN deeper?
- How to solve?

Truncated Backpropagation

- The truncated backpropagation process consists of stopping the evaluation of weights change until a certain point. The update will not take into account all the past but only until a certain limit of it.

- Computational cost

- Arbitrary Solution

Solve Vanishing Gradient

- Initialization of weight matrix

- Activation Function ReLU::

$$f(x) = \max(0, x)$$

Gradient Clipping

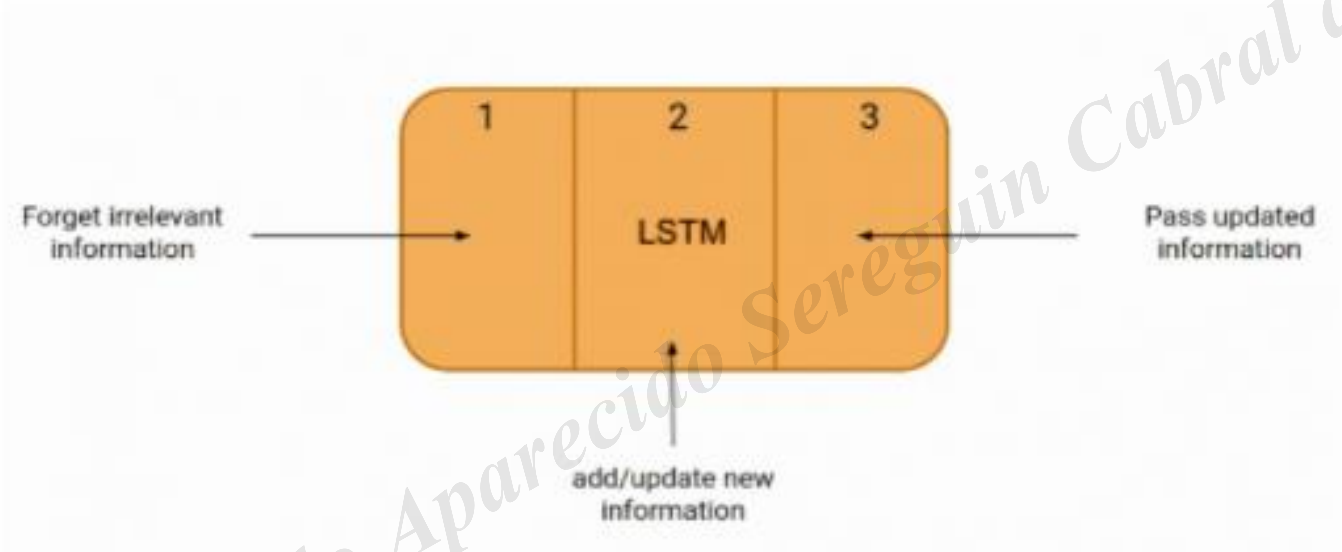
- Possible solution for Dissipation and Explosion

The clipping defines a limit value defined in the gradient, which means that, even if a gradient increases beyond the predefined value during the training, its value will still be limited to the defined limit. Therefore, the direction of the gradient remains unchanged and only the magnitude of the gradient is changed. (deeplearningbook.com).

LSTM

- Long Short-Term Memory
- The same thing as in the RNN – but we have the "cell state"
- Based on "gates"
- Should we maintain or keep an information?

LSTM - intuition



Source: https://www.analyticsvidhya.com/

LSTM

- Some important functions

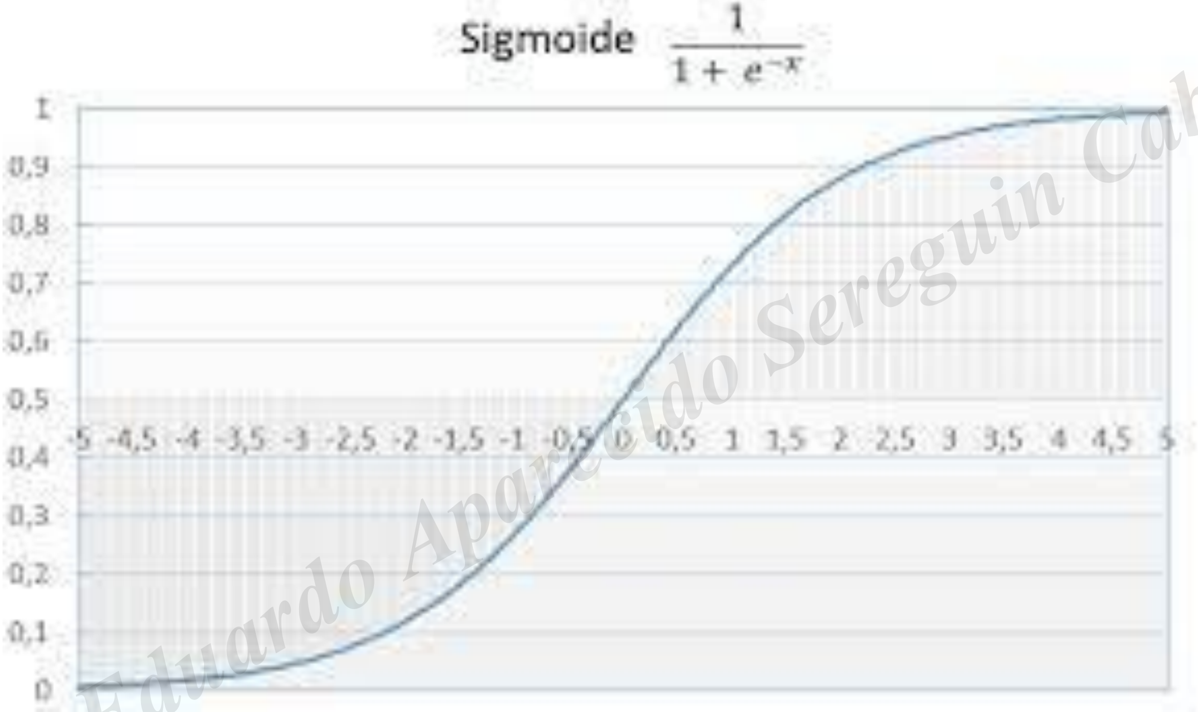
- TANH - hyperbolic tangent

- Result between -1 and 1 => "negative", "neutral" and "positive"

$$\tanh = \frac{senh(t)}{\cosh(t)}$$

LSTM

- Sigmoid
- Result between 0 and 1 => "yes" or "no"



Source: Research Gate

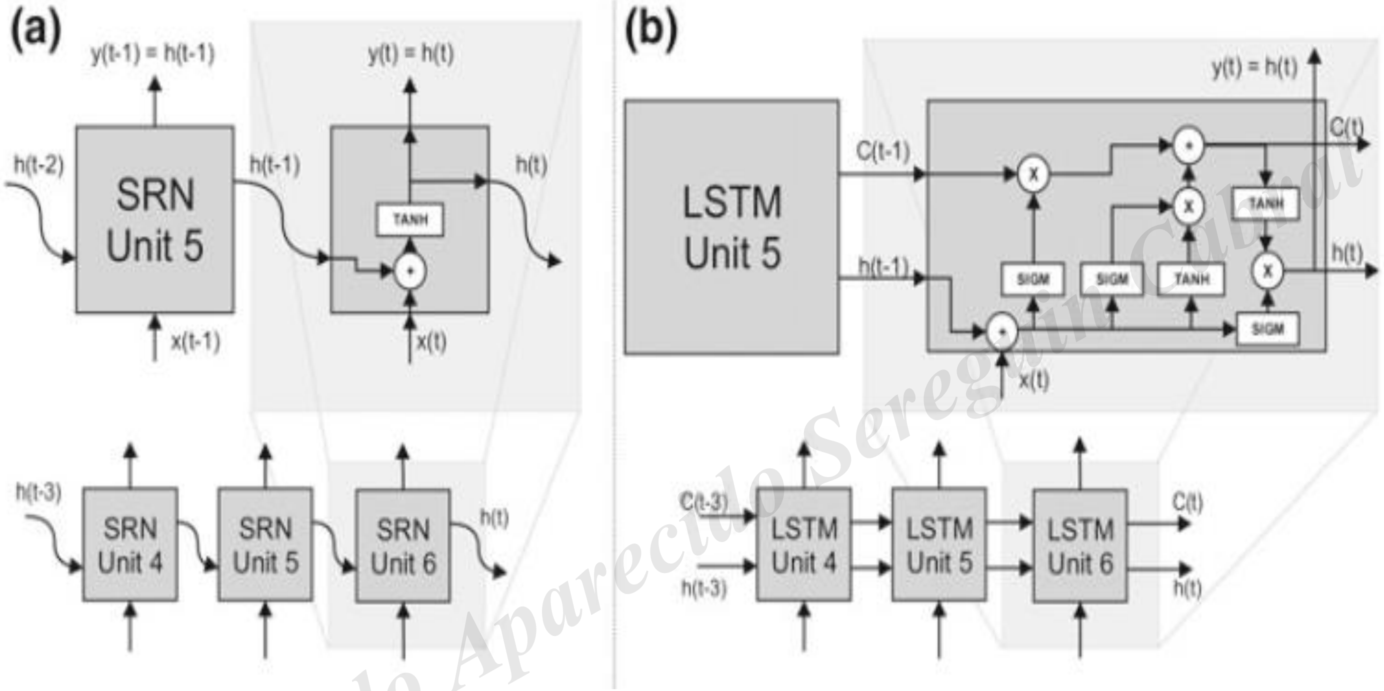
LSTM

- Connection of the input with each hidden - equal
- Cell State - long memory

LSTM

- Forget Gate - How much to remember?
- Input Gate - How much to maintain inputs? What to add to the cell state?
- Output Gate - What will be used from the cell state and the hidden state as a result?

LSTM



MBA USP ESALQ

LSTM

It is possible to observe that given an instant in time t, the LSTM cell have the current moment of network information feeding as inputs, identified as xt, the hidden state ht-1 and the cell state ct-1, both states from the recurrence of the instant of past time t-1. The cell outputs are the cell state ct of the current moment, the hidden state ht and the information output yt. For the case of the cell belonging with the last layer of the network, the ht is understood as the final output yt, for the case of the layer being internal to the network, the ht will serve as ht-1 for the next layer in the network.

LSTM

In addition to the input and output, a LSTM cell is internally composed of combinations between activation function, addition and products. These internal operations of the LSTM cell are called gates, which consists of forget gate, input gate, cell gate, and output gate. In addition to these gates, the LSTM cell has a region responsible for grouping the output of some of these gates to produce the ct, which is one of the outputs of the cell. (OLIVEIRA, E.V., 2020)

LSTM – Forget Gate

LSTM

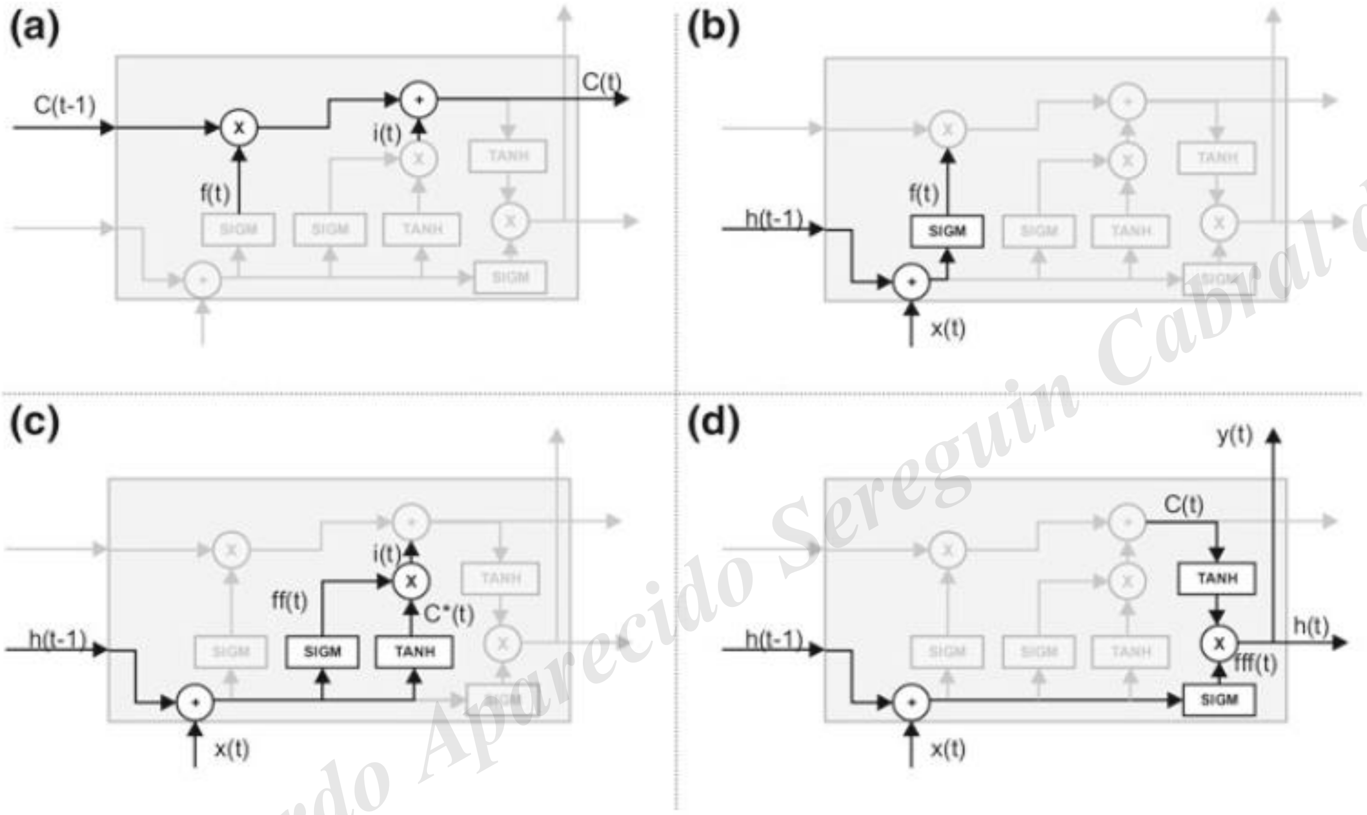


Fig. 7.4 Cell state (**a**), forget gate (**b**), input gate (**c**) and output gate (**d**)

LSTM Animation

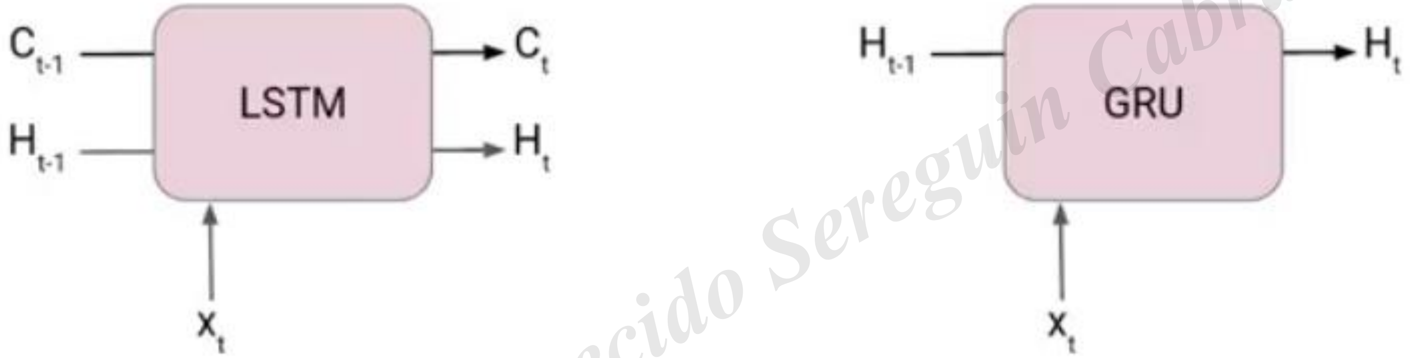
GRU – Gate Recurrent Unit

- It solves the problem of the gradient dissipation
- Based on gates: reinitialization and update
- Only 1 hidden state

GRU

- It retains long dependences
- Reset Gate: How much previous information will we ignore?
- Update Gate: How much previous information will we maintain?

GRU



Source: https://www.analyticsvidhya.com/

Extra: Transformers

- Most of the Natural Language Processing was done with RNN
- Attention is all you need
- Use of RNN => lose information as it is distant from the beginning of a serie
- The context is essential in NLP

Extra: Transformers

- Encoder - decoder

- Encoder - it processes information about input and relationships between them

- Decoder - it does the opposite, it collects all the codes and processes them, using its incorporating contextual information to generate an output sequence.

Extra: Transformers

- Logic: which is faster to find a solution: read a entire book or seek in the index?

- Context Vector - it retains position inside the sequence

- Solution: transfer all hidden states

https://www.linkedin.com/in/jeronymo-marcondes-585a26186