UNIVERSIDADE CATÓLICA PORTUGUESA

BRAGA

Deep Learning

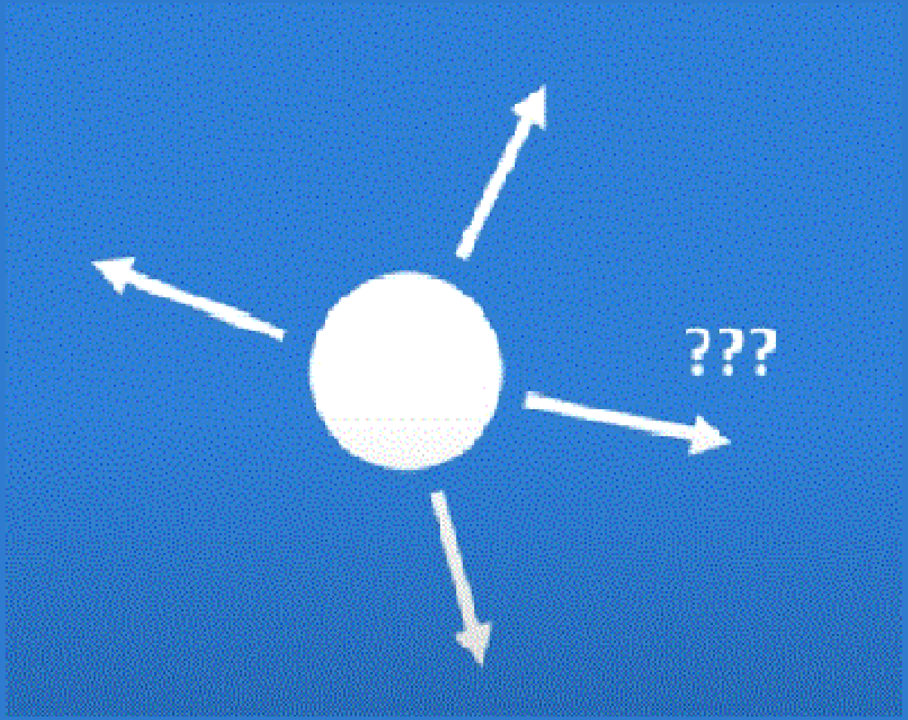
Session 14

Introduction to Recurrent Neural Networks (RNNs)

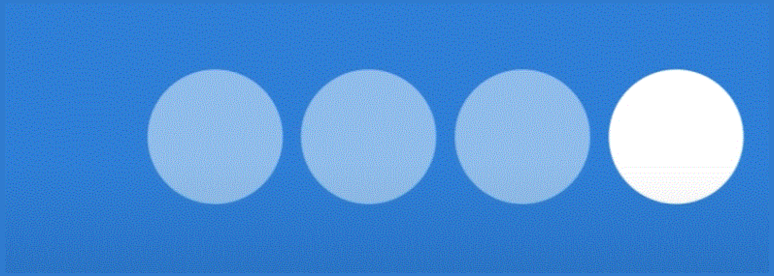
Applied Data Science

2024/2025

Given an image of a ball can you predict where it will go next?

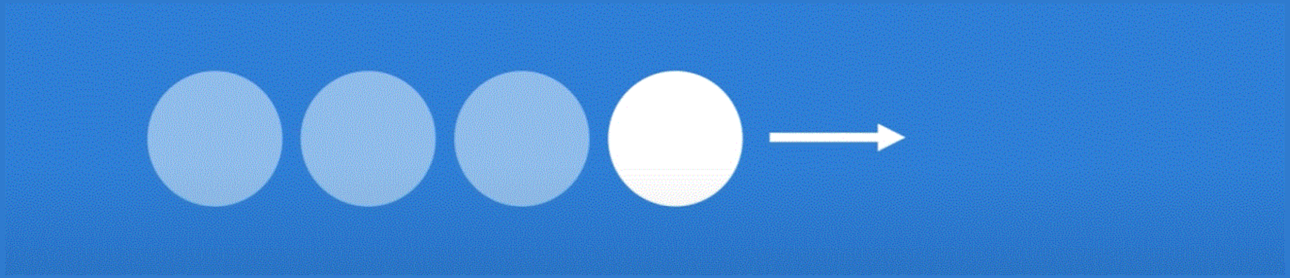


Given an image of a ball can you predict where it will go next?



Given an image of a ball can you predict where it will go next?

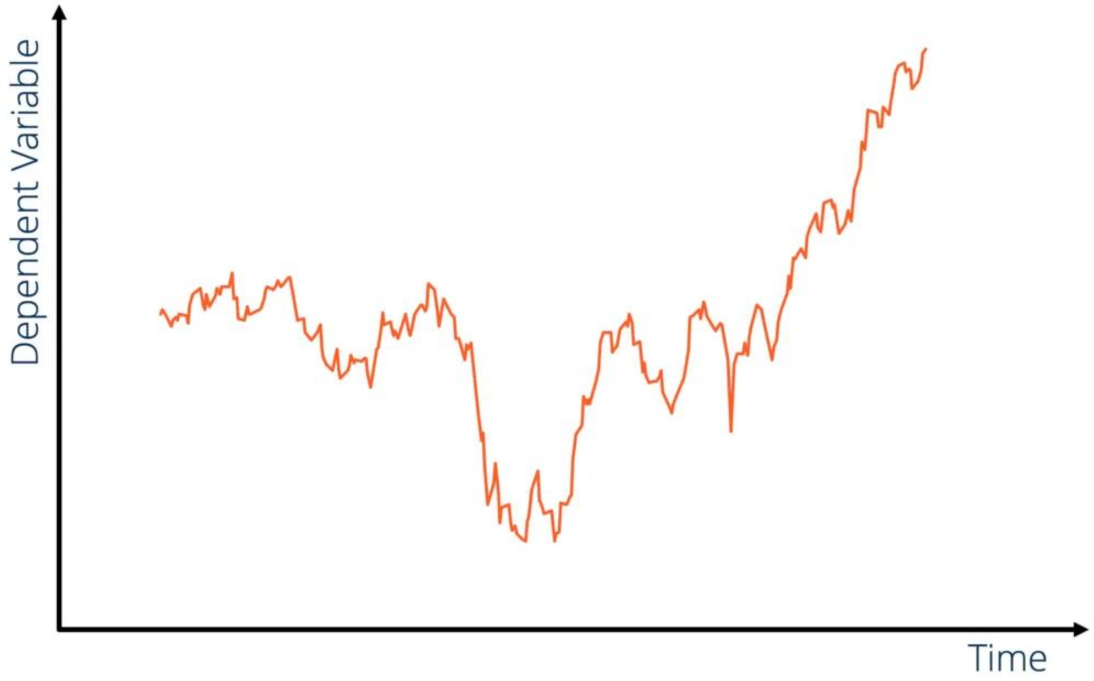
UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

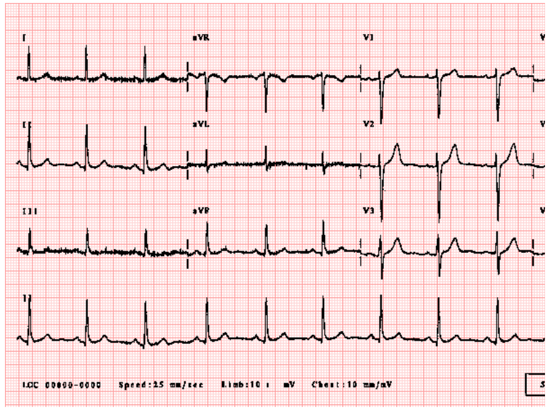


Sequential Data

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- Elements in a sequence occur in a certain order
- Elements depend on each other





Tuesday, Sep 20, YYYY

NEWS TODAY

Latest news and bulletin updates

Issue #10

Mirjam Nilsson

The scoop of the day

The latest updates

Video provides a powerful way to help you prove your point. When you click Online Video, you can paste in the embed code for the video you want to add. You can also type a keyword to search online for the video that best fits your document.

To make your document look professionally produced, Word provides header, footer, cover page, and text box designs that complement each other. For example, you can add a matching cover page, header, and sidebar.

Click Insert and then choose the elements you want from the different galleries.

Themes and styles also help keep your document coordinated. When you click Design and choose a new Theme, the pictures, charts, and SmartArt graphics change to match your new theme. When you apply styles, your headings change to match the new theme.

Save time in Word with new buttons that show up where you need them. To change the way a picture fits in your document, click it and a button for layout options appears next to it.

When you work on a table, click where you want to add a row or a column, and then click the plus sign.

Picture Caption: To make your document look professionally produced, Word provides header, footer, cover page, and text box designs that complement each other.

Mirjam Nilsson

The scoop of the day

The latest updates to get you through the day

Video provides a powerful way to help you prove your point. When you click Online Video, you can paste in the embed code for the video you want to add. You can also type a keyword to search online for the video that best fits your document.

To make your document look professionally produced, Word provides header, footer, cover page, and text box designs that complement each other. For example, you can add a matching cover page, header, and sidebar.

Save time in Word with new buttons that show up where you need them. To change the way a picture fits in your document, click it and a button for layout options appears next to it. When you work on a table, click where you want to add a row or a column, and then click the plus sign.

Themes and styles also help keep your document coordinated. When you click Design and choose a new Theme, the pictures, charts, and SmartArt graphics change to match your new theme. When you apply styles, your headings change to match the new theme.

Save time in Word with new buttons that show up where you need them. To change the way a picture fits in your document, click it and a button for layout options appears next to it. When you work on a table, click where you want to add a row or a column, and then click the plus sign.

Reading is easier, too, in the new Reading view. You can collapse parts of the document and focus on the text you want. If you need to stop reading before you reach the end, Word remembers where you left off - even on another device.

The scoop of the day

The latest updates

Mirjam Nilsson

Reading is easier, too, in the new Reading view. You can collapse parts of the document and focus on the text you want.
Page XX

The scoop of the day

The latest updates

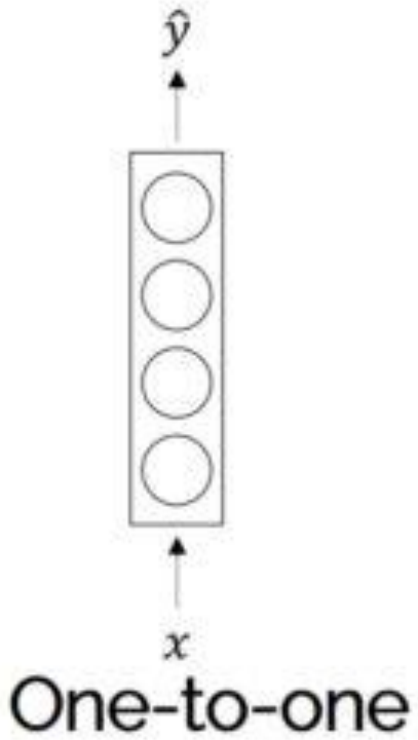
Mirjam Nilsson

To make your document look professionally produced, Word provides header, footer, cover page, and text box designs that complement each other.
Page XX

Page 1

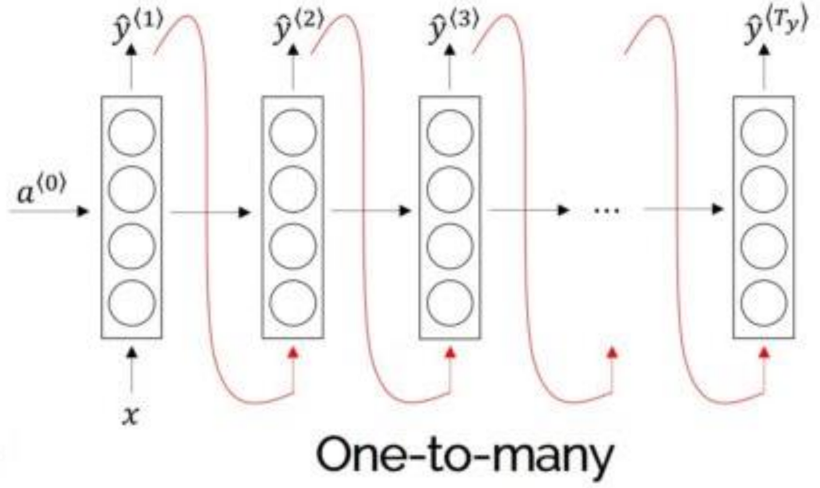
Sequence Applications: One-to-One

- **Description:** Processes a **single input** to produce a **single output**. Typically used in traditional feedforward networks, **not truly recurrent**.



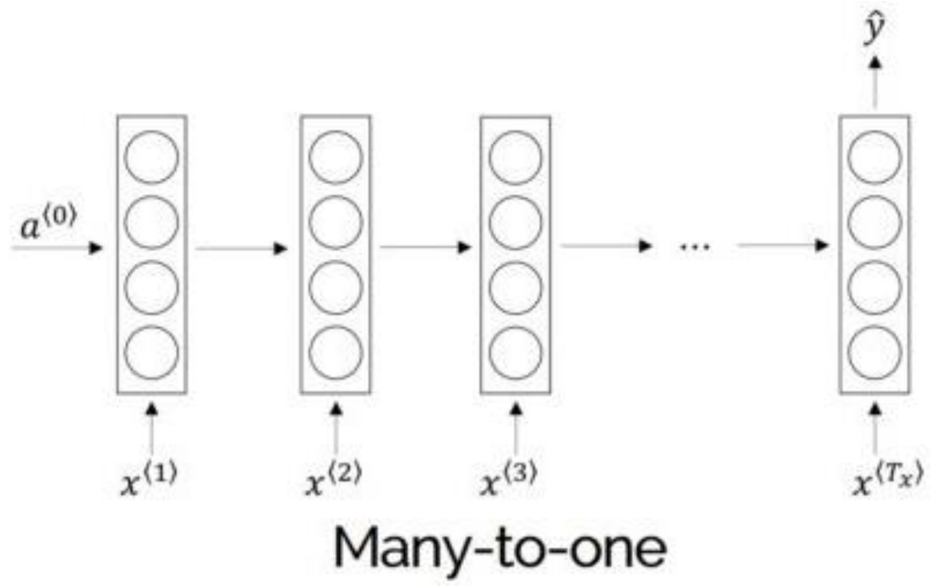
Sequence Applications: One-to-Many

- **Description:** Takes **one input** and generates a **sequence of outputs**, often used for generating text or music.



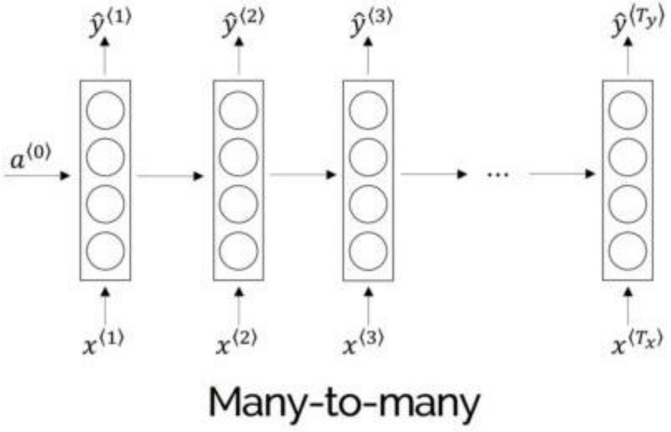
Sequence Applications: Many-to-One

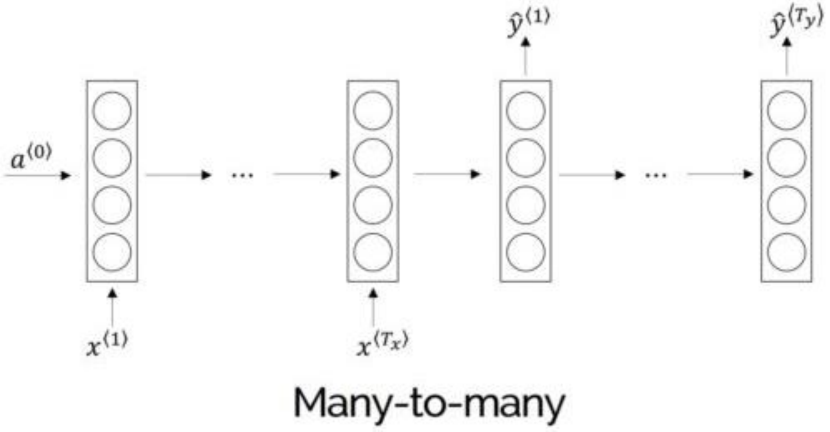
- **Description:** Takes a **sequence of inputs** and produces a **single output**, ideal for analyzing sequences.



Sequence Applications: Many-to-Many

- **Description:** Processes a **sequence of inputs** to generate a **sequence of outputs**, useful for tasks where both input and output are sequential.



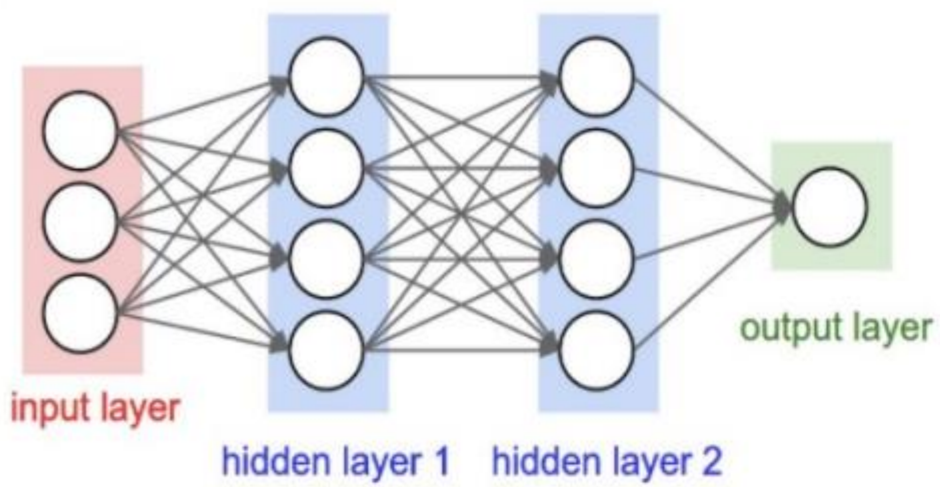


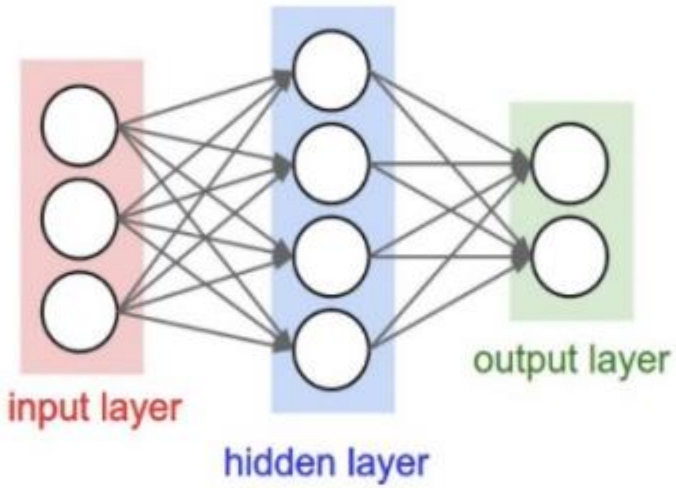
Sequence Applications: Many-to-Many

- **Description:** Processes a **sequence of inputs** to generate a **sequence of outputs**, useful for tasks where both input and output are sequential.

Recall: Feedforward Neural Networks

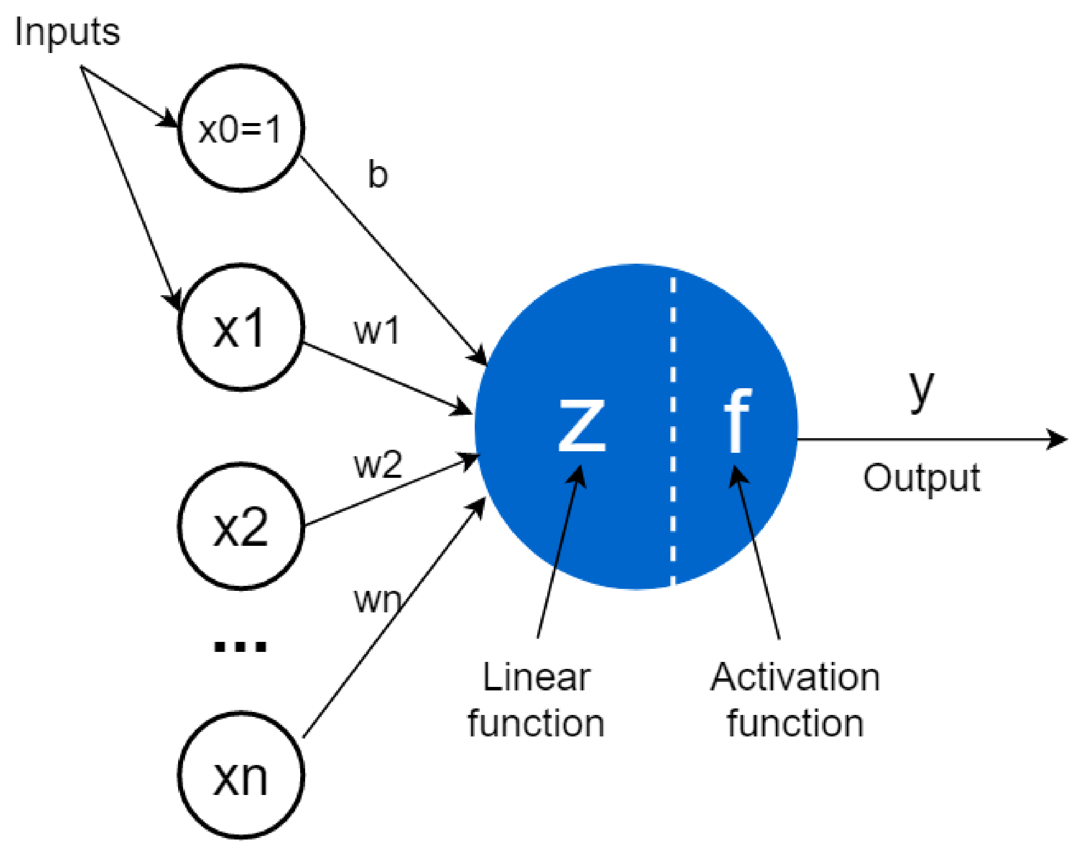
- Problems:
  - Many model parameters;
  - Input / output sizes are fixed;
  - No memory of past since weights are learned independently.





- Each layer serves as input to the next layer with no loops

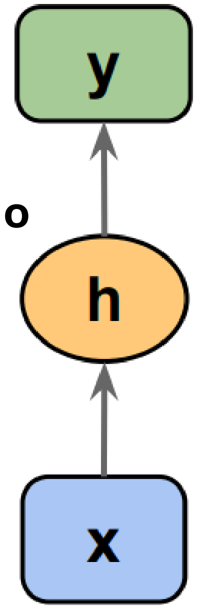
Recall: The Perceptron

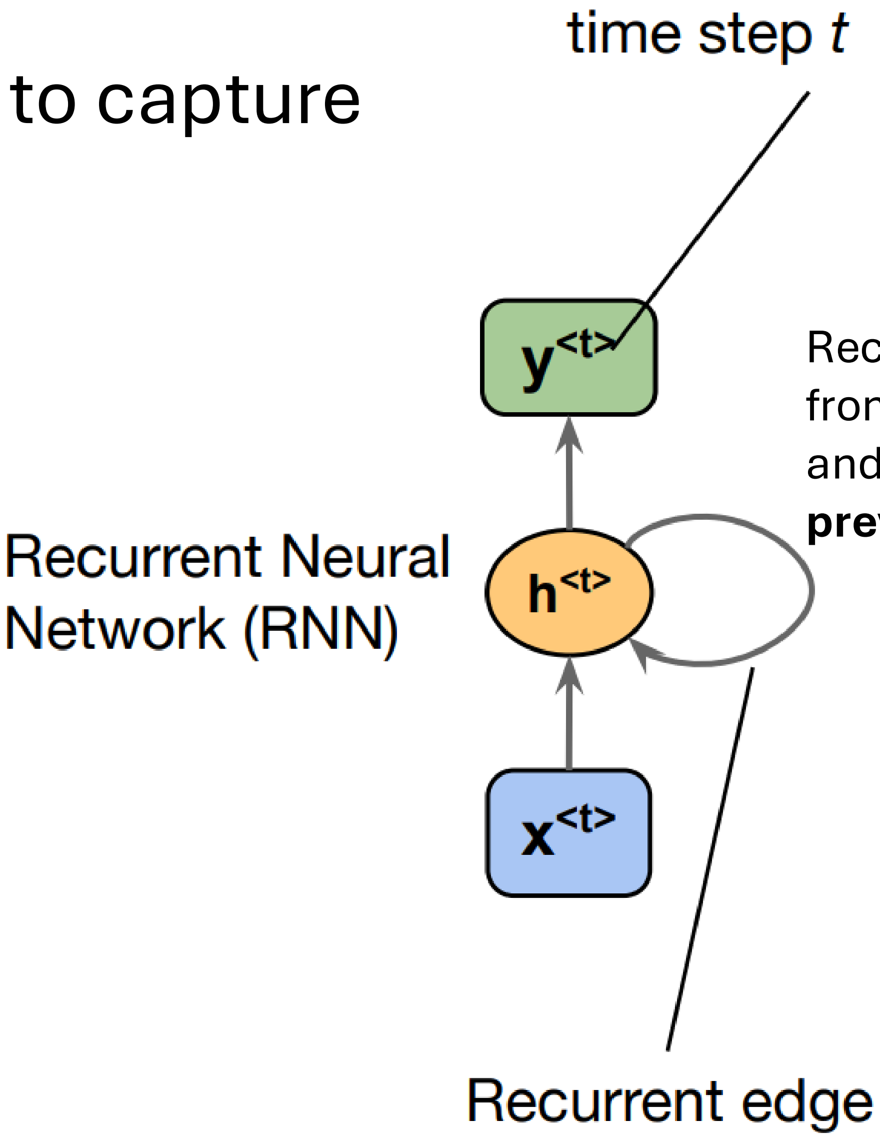


Recurrent Neural Networks (RNNs)

- **Main idea:** use hidden state to capture **information about the past**

Receives input from the previous layer with **no loops**.





Receives input from the previous layer and from **the output of the previous time step**

Image source: Sebastian Raschka, Vahid Mirjalili. *Python Machine Learning. 3rd Edition.* Packt, 2019

Recurrent Neural Networks (RNNs)

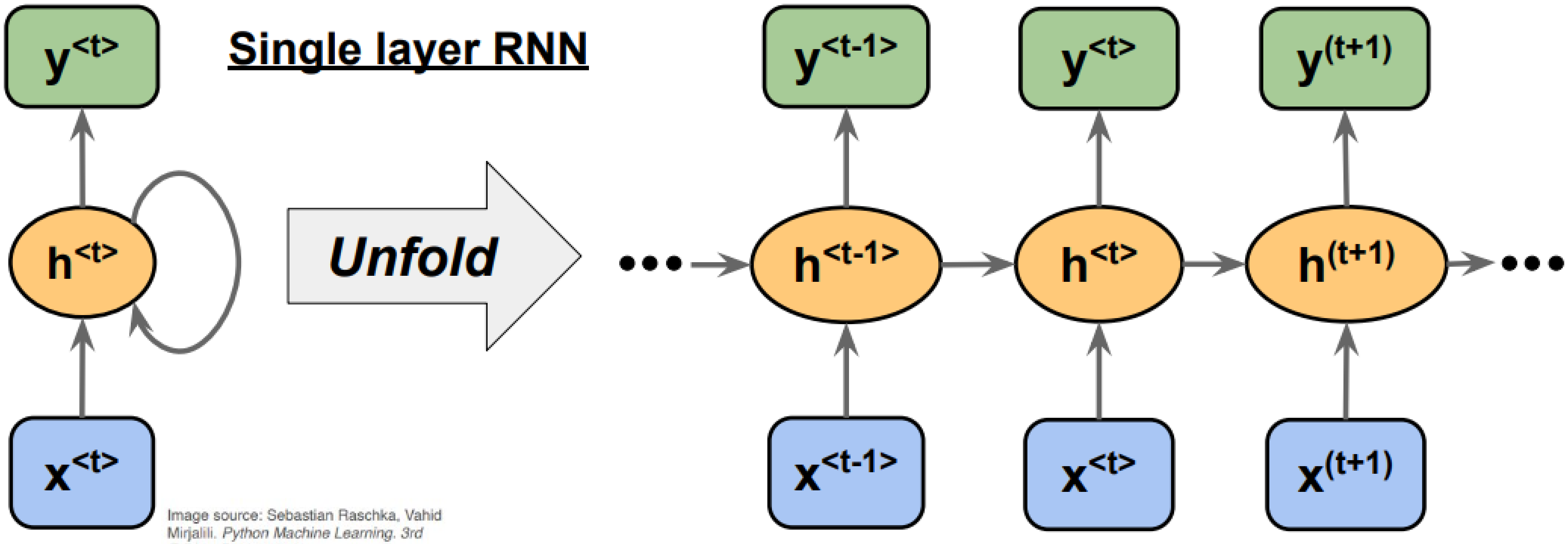


Image source: Sebastian Raschka, Vahid Mirjalili. *Python Machine Learning. 3rd Edition.* Packt, 2019

Recurrent Neural Networks (RNNs)

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

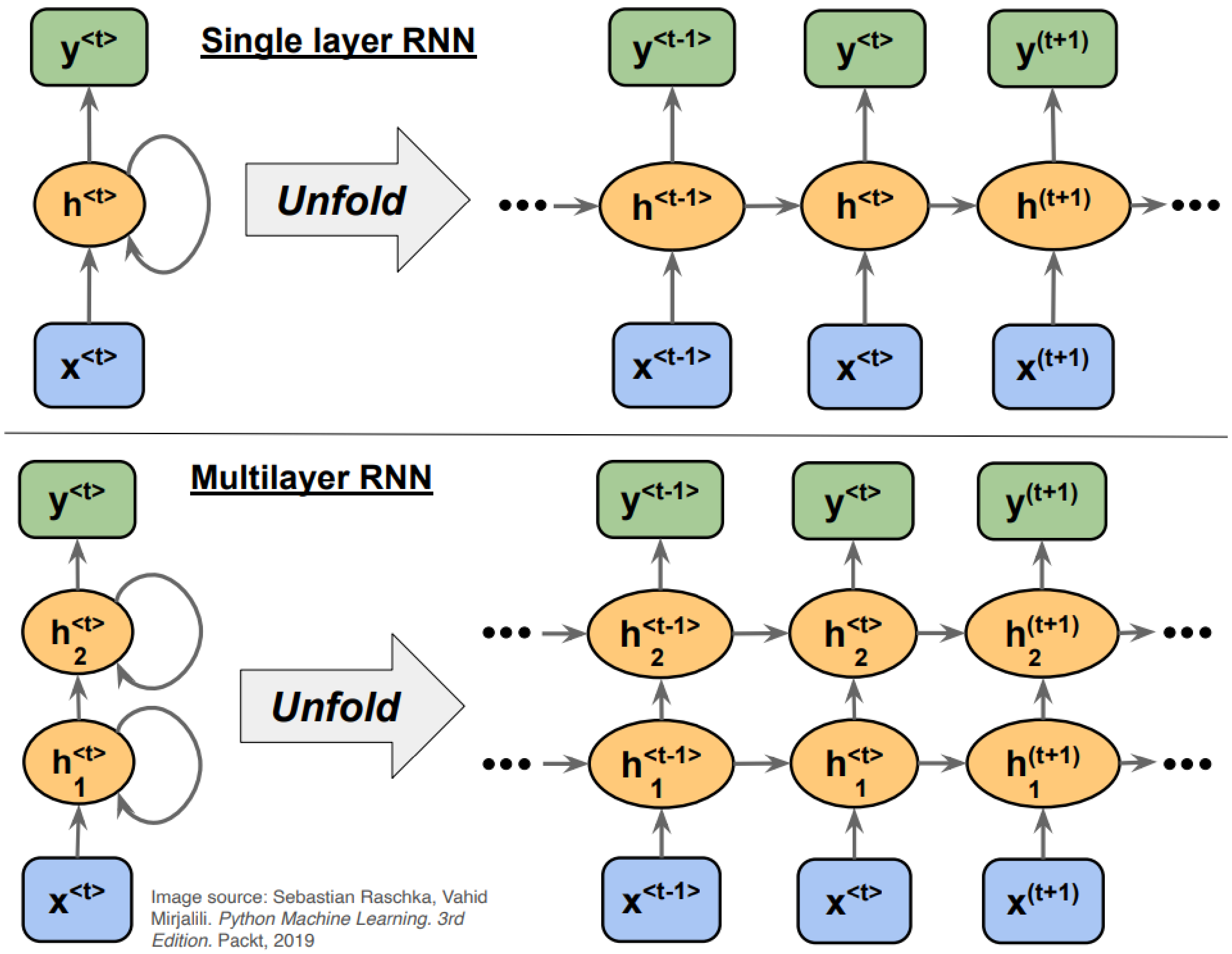


Image source: Sebastian Raschka, Vahid Mirjalili. *Python Machine Learning. 3rd Edition.* Packt, 2019

Recurrent Neural Networks (RNNs)

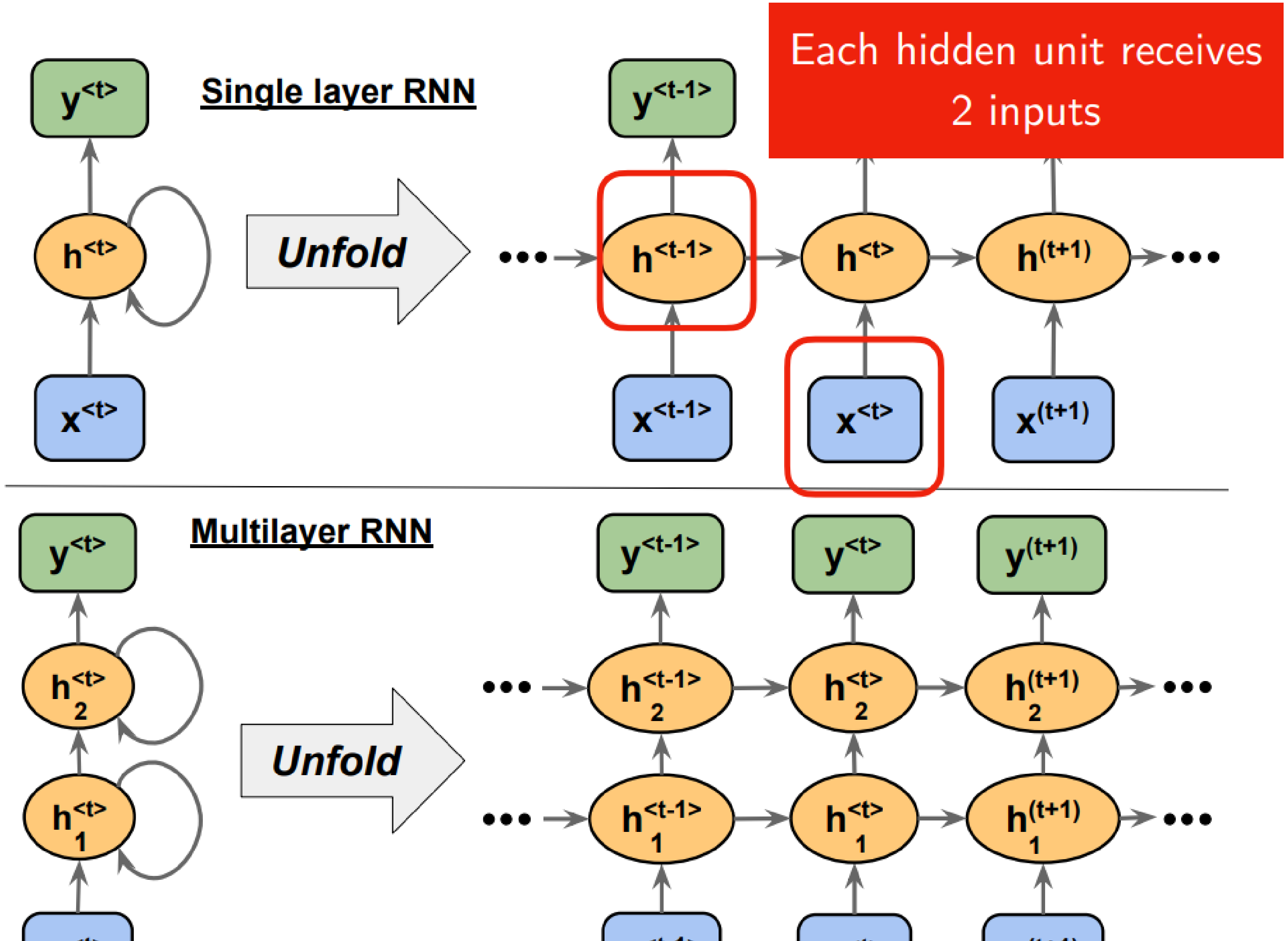


Image source: Sebastian Raschka, Vahid Mirjalili. *Python Machine Learning. 3rd Edition.* Packt, 2019

$$\underbrace{\hat{y}_t}_{\text{output}} = f(\underbrace{x_t}_{\text{input}}, \underbrace{h_{t-1}}_{\text{past memory}})$$

Backpropagation Through Time

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- Weight matrices in a single-hidden layer RNN

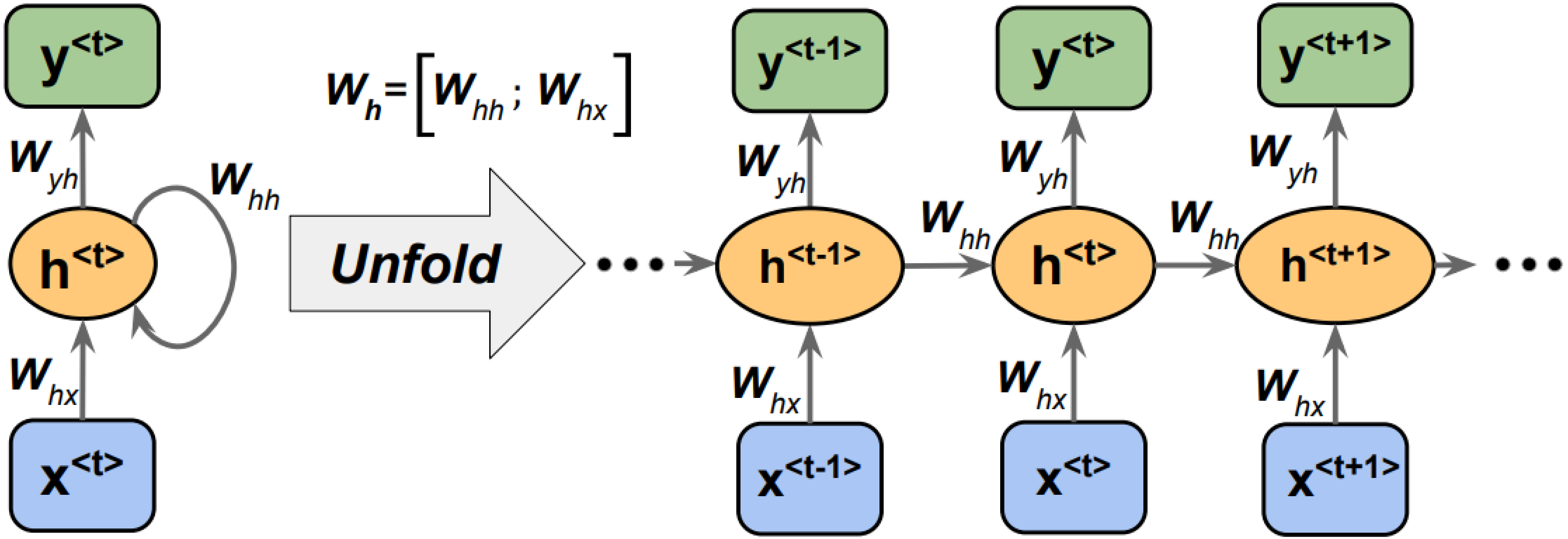


Image source: Sebastian Raschka, Vahid Mirjalili. *Python Machine Learning. 3rd Edition*. Packt, 2019

Backpropagation Through Time

- Weight matrices in a single-hidden layer RNN

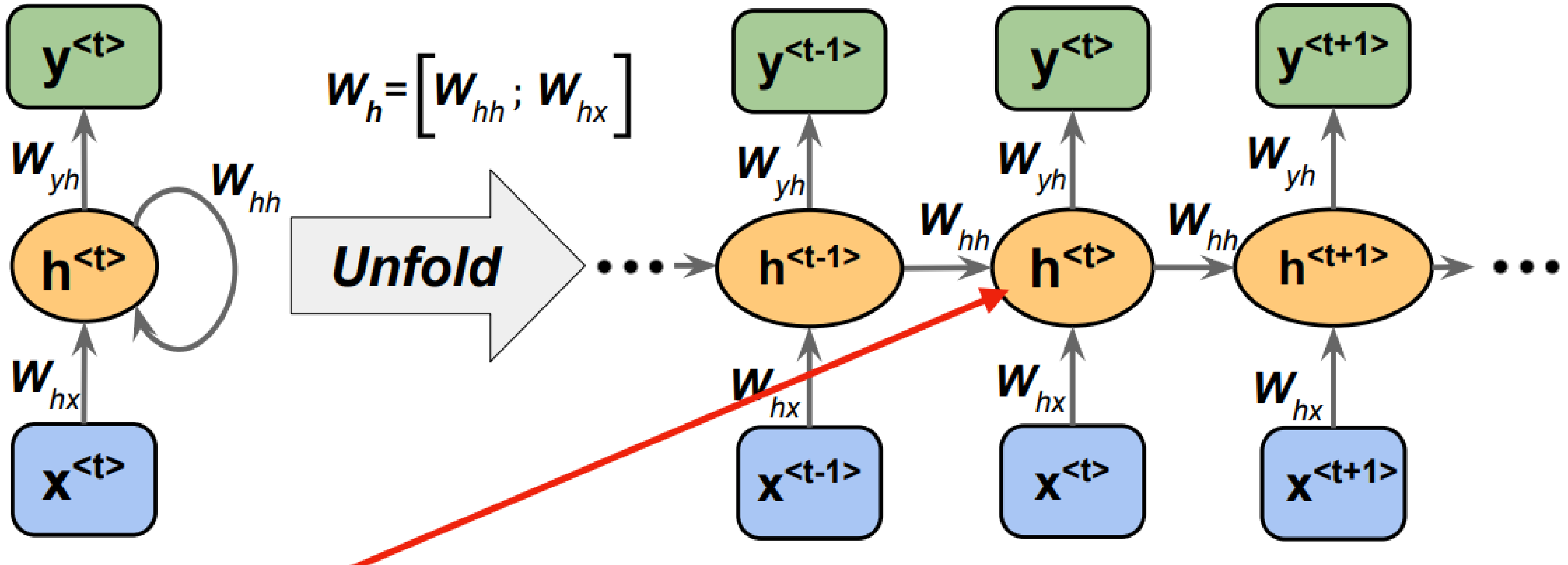


Image source: Sebastian Raschka, Vahid Mirjalili. *Python Machine Learning. 3rd Edition.* Packt, 2019

Net input:

$$\mathbf{z}_h^{\langle t \rangle} = \mathbf{W}_{hx}\mathbf{x}^{\langle t \rangle} + \mathbf{W}_{hh}\mathbf{h}^{\langle t-1 \rangle} + \mathbf{b}_h$$

Activation:

$$\mathbf{h}^{\langle t \rangle} = \sigma_h\left(\mathbf{z}_h^{\langle t \rangle}\right)$$

Backpropagation Through Time

- Weight matrices in a single-hidden layer RNN

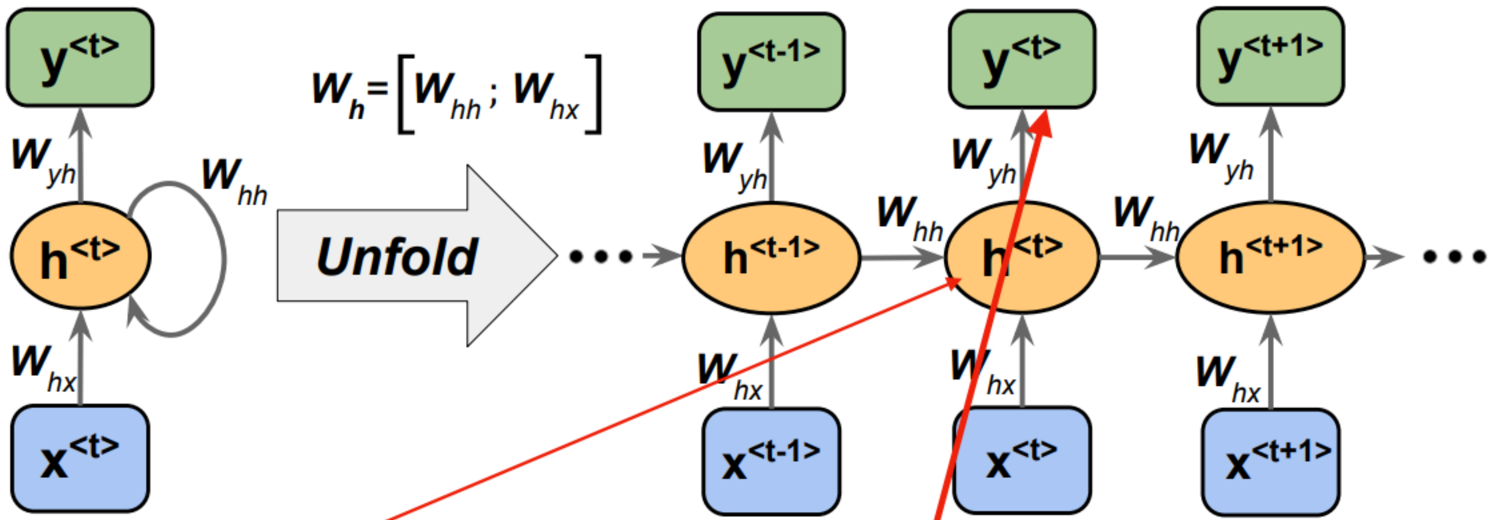


Image source: Sebastian Raschka, Vahid Mirjalili. *Python Machine Learning. 3rd Edition.* Packt, 2019

Net input:

$$\mathbf{z}_h^{\langle t \rangle} = \mathbf{W}_{hx}\mathbf{x}^{\langle t \rangle} + \mathbf{W}_{hh}\mathbf{h}^{\langle t-1 \rangle} + \mathbf{b}_h$$

Activation:

$$\mathbf{h}^{\langle t \rangle} = \sigma_h\left(\mathbf{z}_h^{\langle t \rangle}\right)$$

Net input:

$$\mathbf{z}_y^{\langle t \rangle} = \mathbf{W}_{yh}\mathbf{h}^{\langle t \rangle} + \mathbf{b}_y$$

Output:

$$\mathbf{y}^{\langle t \rangle} = \sigma_y\left(\mathbf{z}_y^{\langle t \rangle}\right)$$

Backpropagation Through Time

- The overall loss can be computed as the sum over all time steps

$$L = \sum_{t=1}^{T} L^{\langle t \rangle}$$

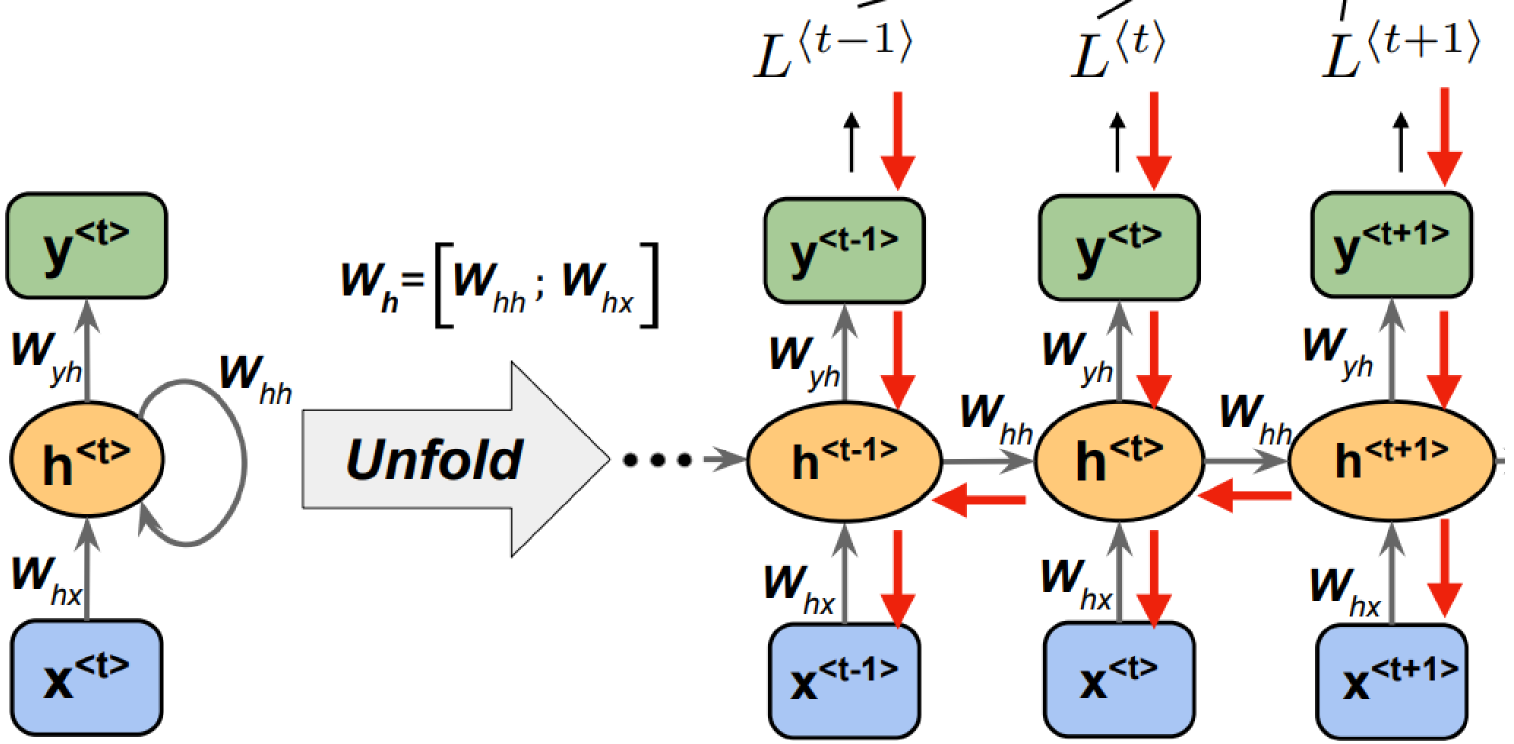
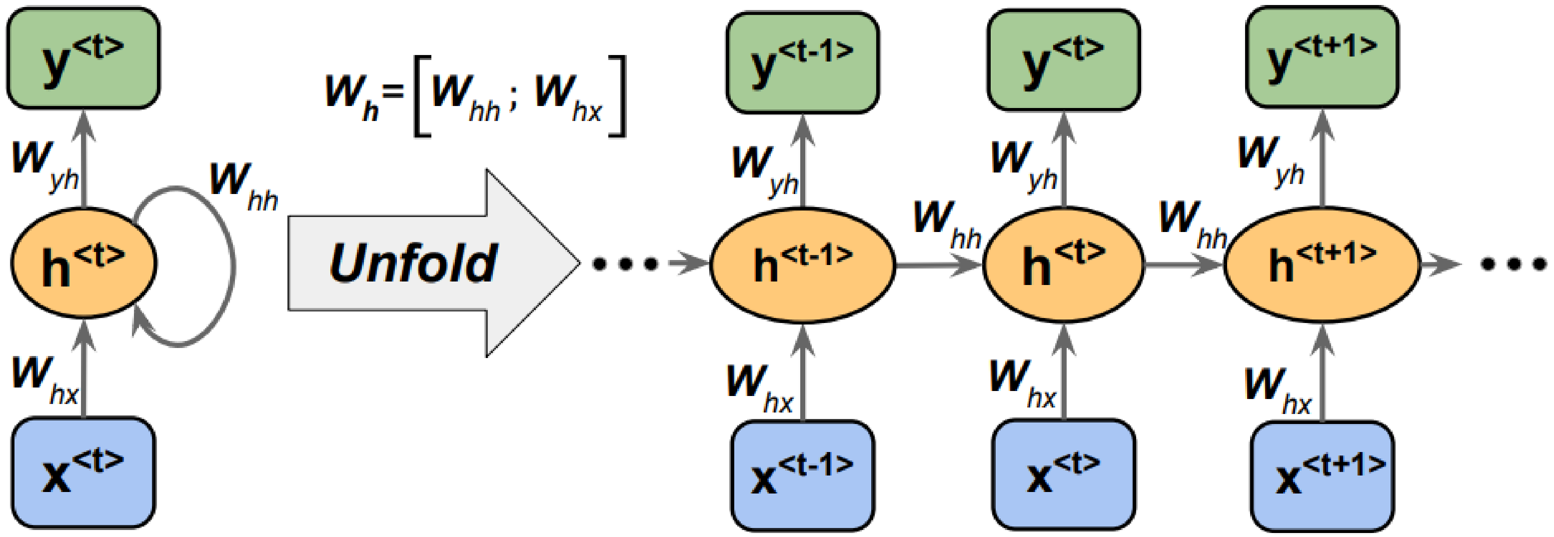


Image source: Sebastian Raschka, Vahid Mirjalili. *Python Machine Learning. 3rd Edition.* Packt, 2019

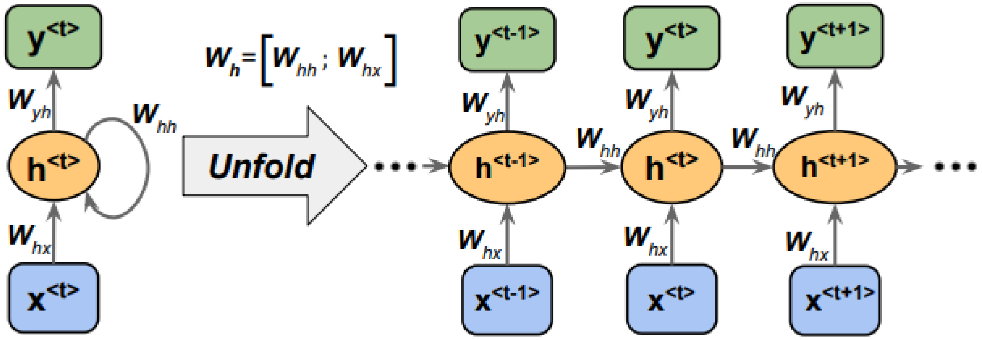
Backpropagation Through Time



$$L = \sum_{t=1}^{T} L^{(t)}$$

$$\frac{\partial L^{(t)}}{\partial \mathbf{W}_{hh}} = \frac{\partial L^{(t)}}{\partial y^{(t)}} \cdot \frac{\partial y^{(t)}}{\partial \mathbf{h}^{(t)}} \cdot \left(\sum_{k=1}^{t} \frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{h}^{(k)}} \cdot \frac{\partial \mathbf{h}^{(k)}}{\partial \mathbf{W}_{hh}} \right)$$

Backpropagation Through Time

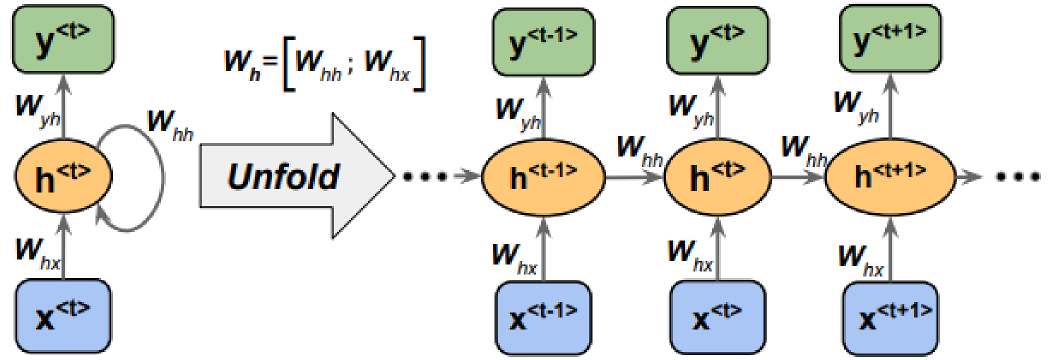


$$\frac{\partial L^{(t)}}{\partial \mathbf{W}_{hh}} = \frac{\partial L^{(t)}}{\partial y^{(t)}} \cdot \frac{\partial y^{(t)}}{\partial \mathbf{h}^{(t)}} \cdot \left(\sum_{k=1}^{t} \boxed{\frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{h}^{(k)}}} \cdot \frac{\partial \mathbf{h}^{(k)}}{\partial \mathbf{W}_{hh}} \right)$$

computed as a multiplication of adjacent time steps:

$$\frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{h}^{(k)}} = \prod_{i=k+1}^{t} \frac{\partial \mathbf{h}^{(i)}}{\partial \mathbf{h}^{(i-1)}}$$

Backpropagation Through Time



Werbos, Paul J. "[Backpropagation through time: what it does and how to do it.]" *Proceedings of the IEEE* 78, no. 10 (1990): 1550-1560.

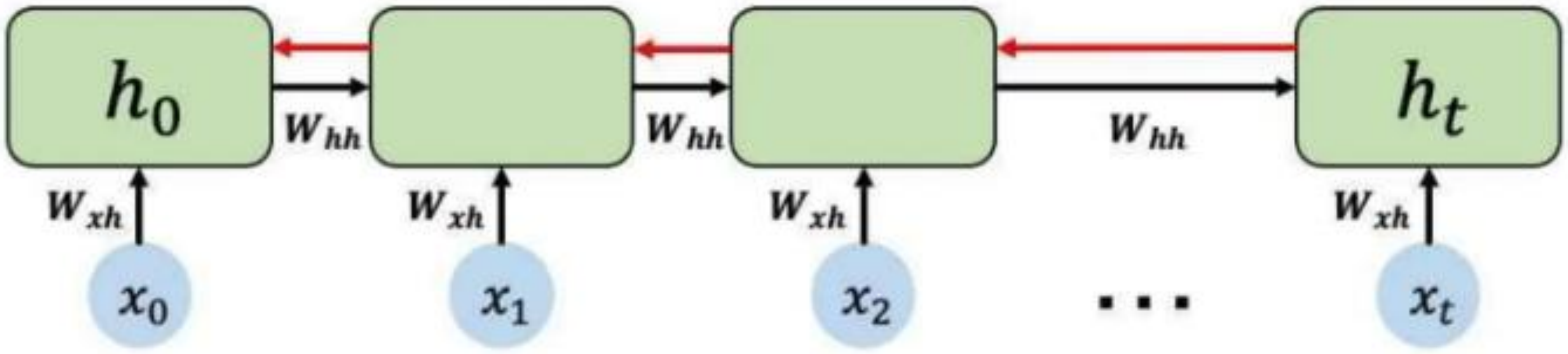
$$L = \sum_{t=1}^{T} L^{(t)} \qquad \frac{\partial L^{(t)}}{\partial \mathbf{W}_{hh}} = \frac{\partial L^{(t)}}{\partial y^{(t)}} \cdot \frac{\partial y^{(t)}}{\partial \mathbf{h}^{(t)}} \cdot \left(\sum_{k=1}^{t} \boxed{\frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{h}^{(k)}}} \cdot \frac{\partial \mathbf{h}^{(k)}}{\partial \mathbf{W}_{hh}} \right)$$

computed as a multiplication of adjacent time steps:

$$\frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{h}^{(k)}} = \prod_{i=k+1}^{t} \frac{\partial \mathbf{h}^{(i)}}{\partial \mathbf{h}^{(i-1)}}$$

This is very problematic:
Vanishing/Exploding gradient problem!

Standard RNN Gradient Flow: Exploding Gradients

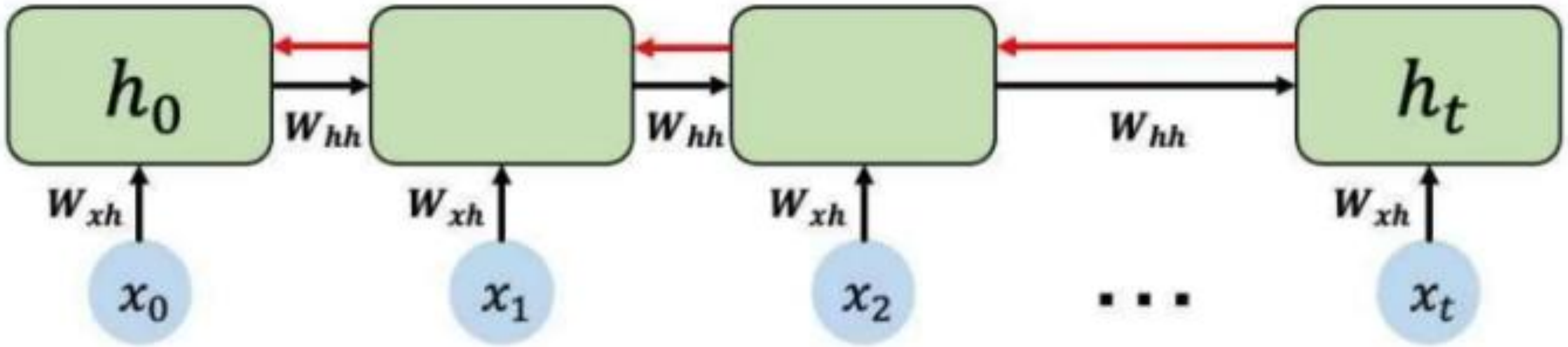


Computing the gradient wrt h_0 involves **many factors of W_{hh} + repeated gradient computation!**

Many values > 1:
exploding gradients

Gradient clipping to scale big gradients

Standard RNN Gradient Flow: Vanishing Gradients



Computing the gradient wrt h_0 involves **many factors of W_{hh} + repeated gradient computation!**

Many values > 1:
exploding gradients

Gradient clipping to scale big gradients

Many values < 1:
vanishing gradients

1. Activation function
2. Weight initialization
3. Network architecture

The Problem of Long-Term Dependencies

Why are vanishing gradients a problem?

Multiply many **small numbers** together

↓

Errors due to further back time steps have smaller and smaller gradients

↓

Bias parameters to capture short-term dependencies

The Problem of Long-Term Dependencies

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

"The clouds are in the ___"

Why are vanishing gradients a problem?

Multiply many **small numbers** together

↓

Errors due to further back time steps have smaller and smaller gradients

↓

Bias parameters to capture short-term dependencies

The Problem of Long-Term Dependencies

Why are vanishing gradients a problem?

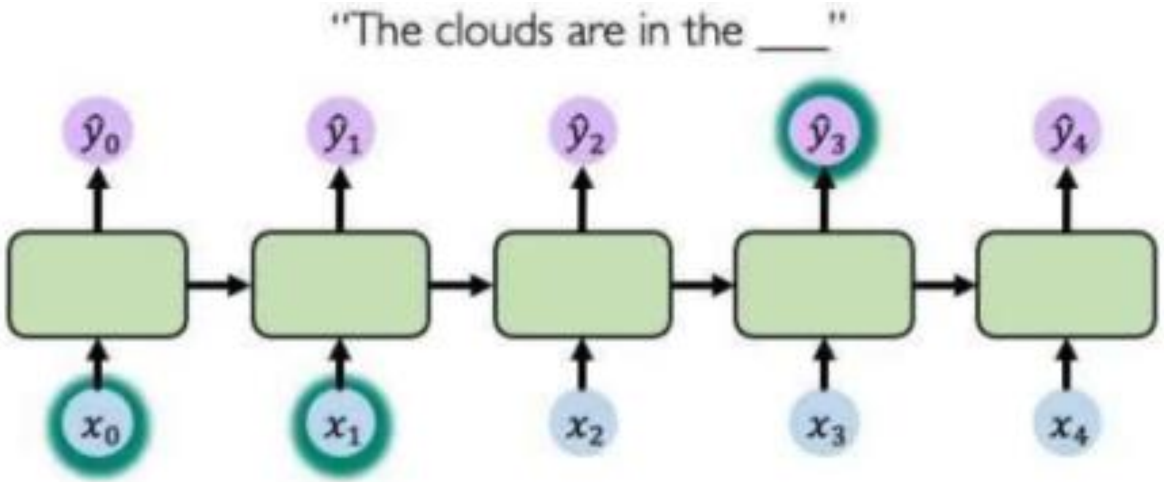
Multiply many **small numbers** together

↓

Errors due to further back time steps have smaller and smaller gradients

↓

Bias parameters to capture short-term dependencies



The Problem of Long-Term Dependencies

Why are vanishing gradients a problem?

Multiply many **small numbers** together

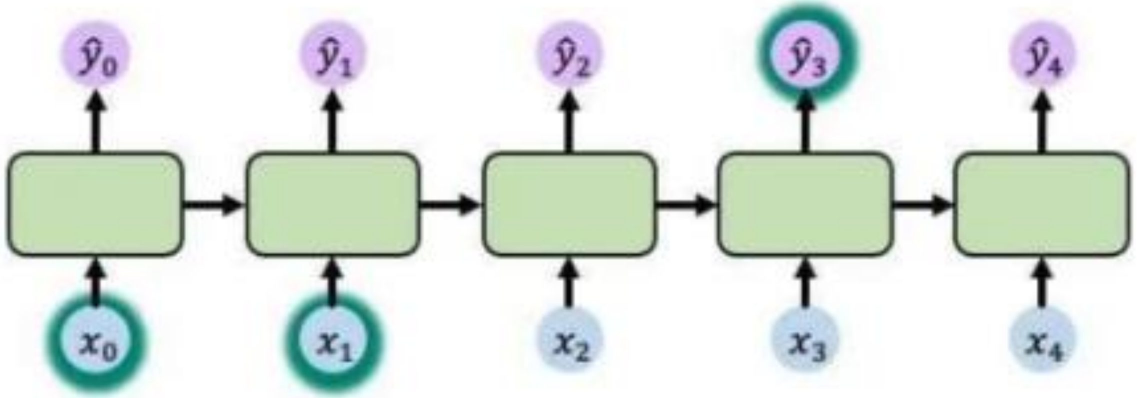
↓

Errors due to further back time steps have smaller and smaller gradients

↓

Bias parameters to capture short-term dependencies

"The clouds are in the ___"



"I grew up in France, … and I speak fluent___ "

The Problem of Long-Term Dependencies

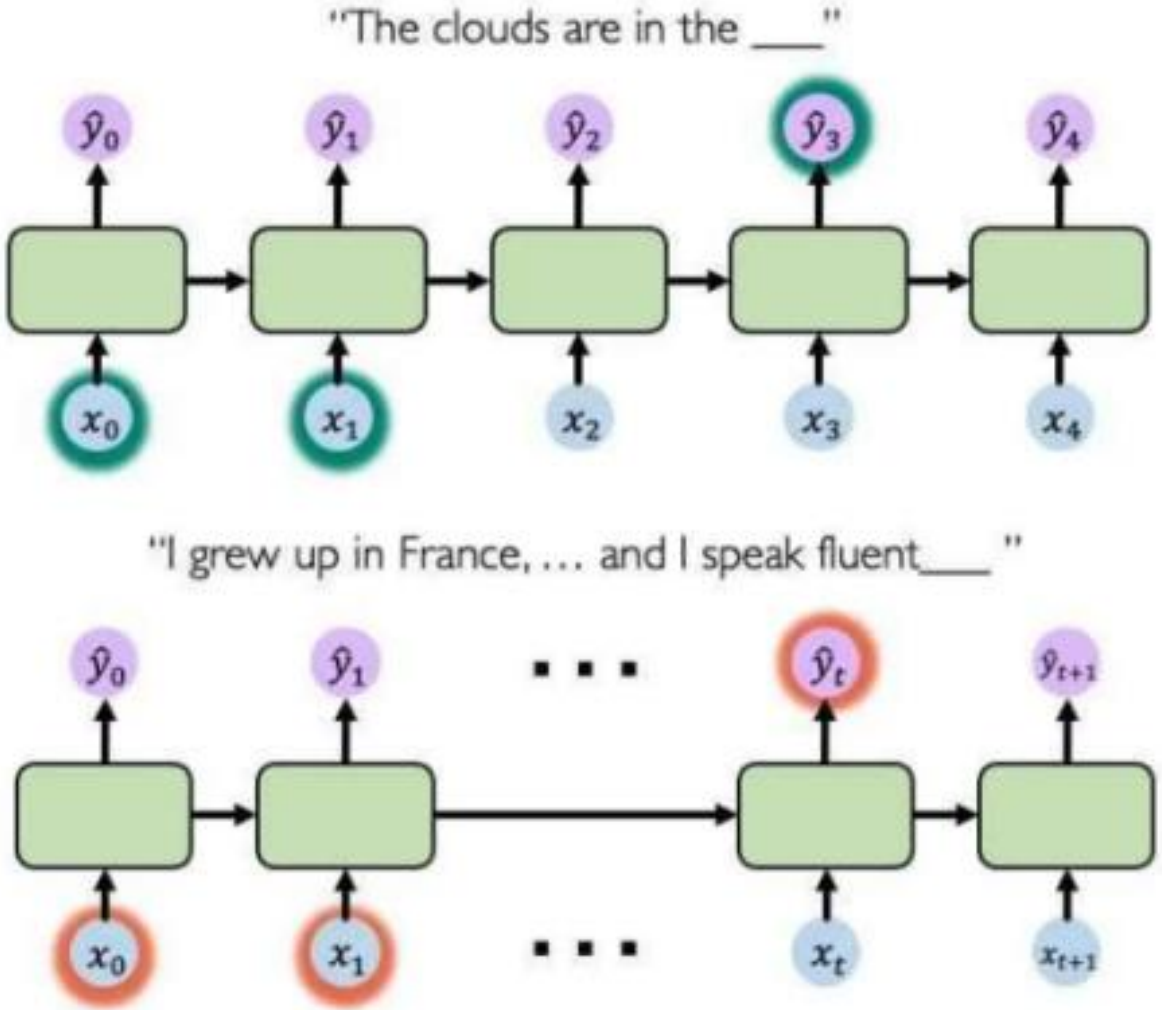
UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

Why are vanishing gradients a problem?

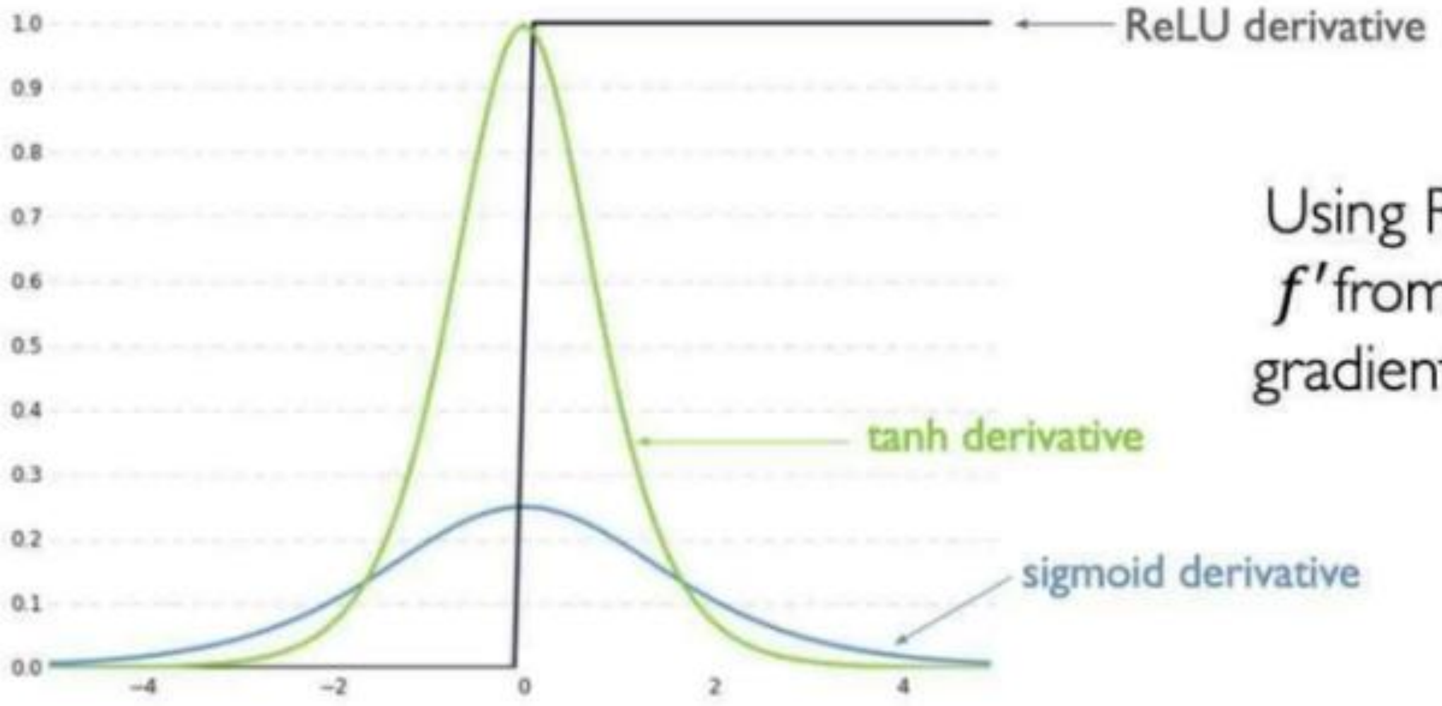
Multiply many **small numbers** together

Errors due to further back time steps have smaller and smaller gradients

Bias parameters to capture short-term dependencies



Solutions to the Vanishing/Exploding Gradients



Using ReLU prevents f' from shrinking the gradients when $x > 0$

Solutions to the Vanishing/Exploding Gradients

Initialize **weights** to identity matrix

Initialize **biases** to zero

$$I_n = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

This helps prevent the weights from shrinking to zero.

Solutions to the Vanishing/Exploding Gradients

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

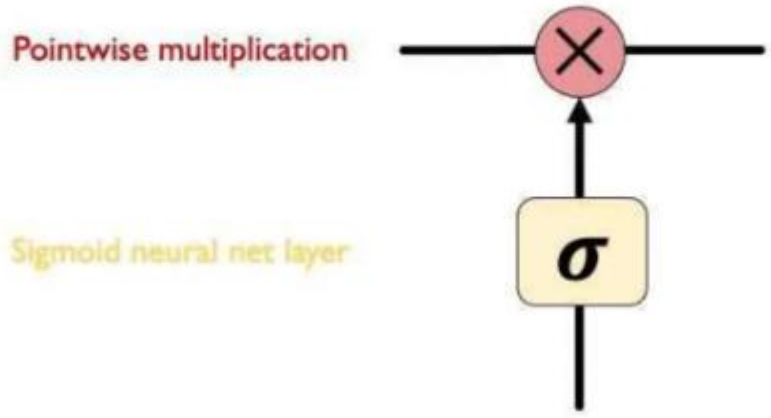
- **Gradient Clipping:** Set a maximum threshold for gradients to prevent them from becoming excessively large, effectively addressing the exploding gradient problem.

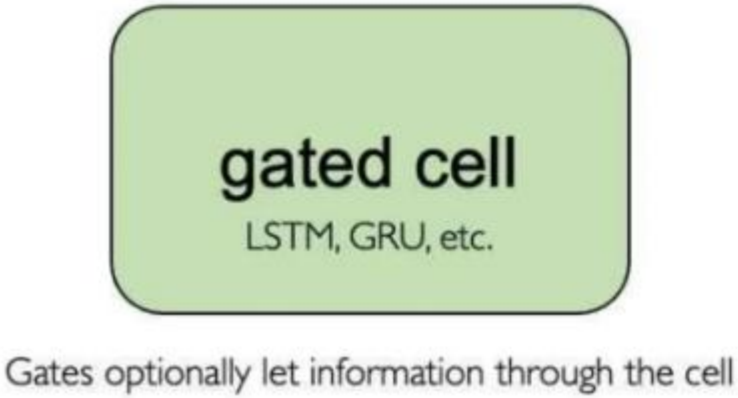
- **Truncated Backpropagation Through Time (TBPTT):** Limits the number of time steps for which gradients are backpropagated after each forward pass. For example, with a sequence of 100 steps, TBPTT may only backpropagate through the most recent 20 steps.

- **Gated Cells:** use gates to selectively add or remove information within each recurrent unit

Gated Cells

- **Idea:** use gates to selectively add or remove information within each recurrent unit

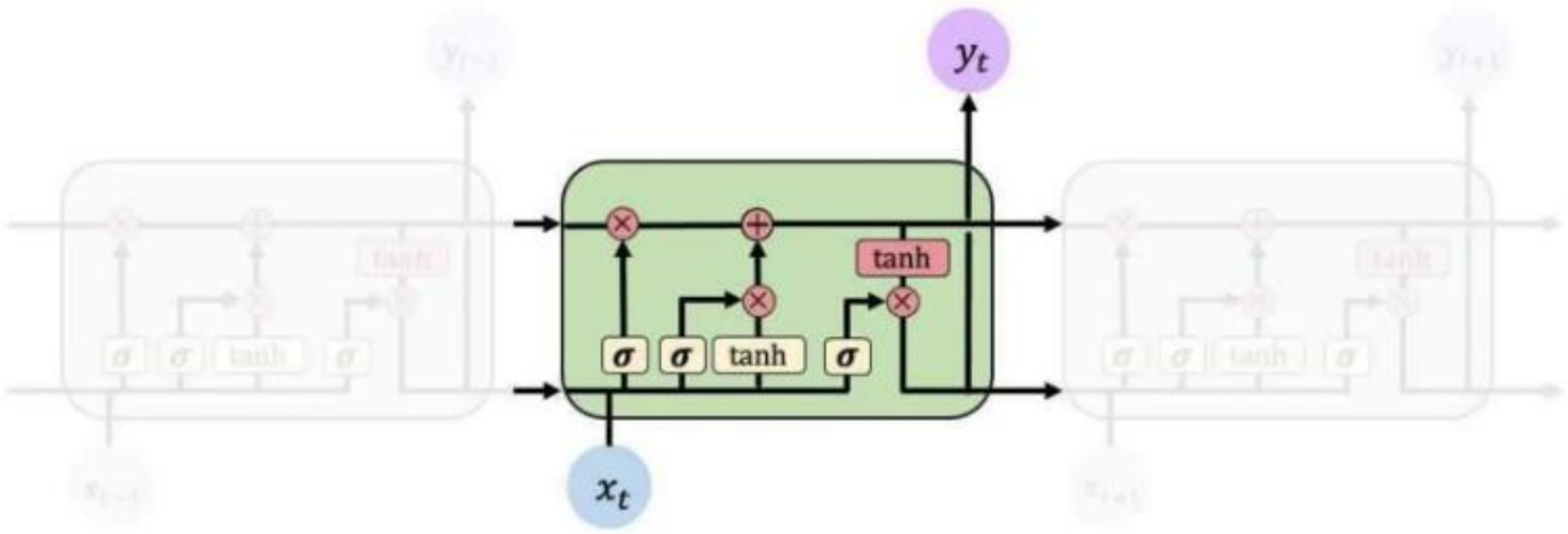




- **Long Short Term Memory** (LSTM) networks rely on gated cells to tracj information throughout many time steps.

Long-Short Term Memory (LSTM)

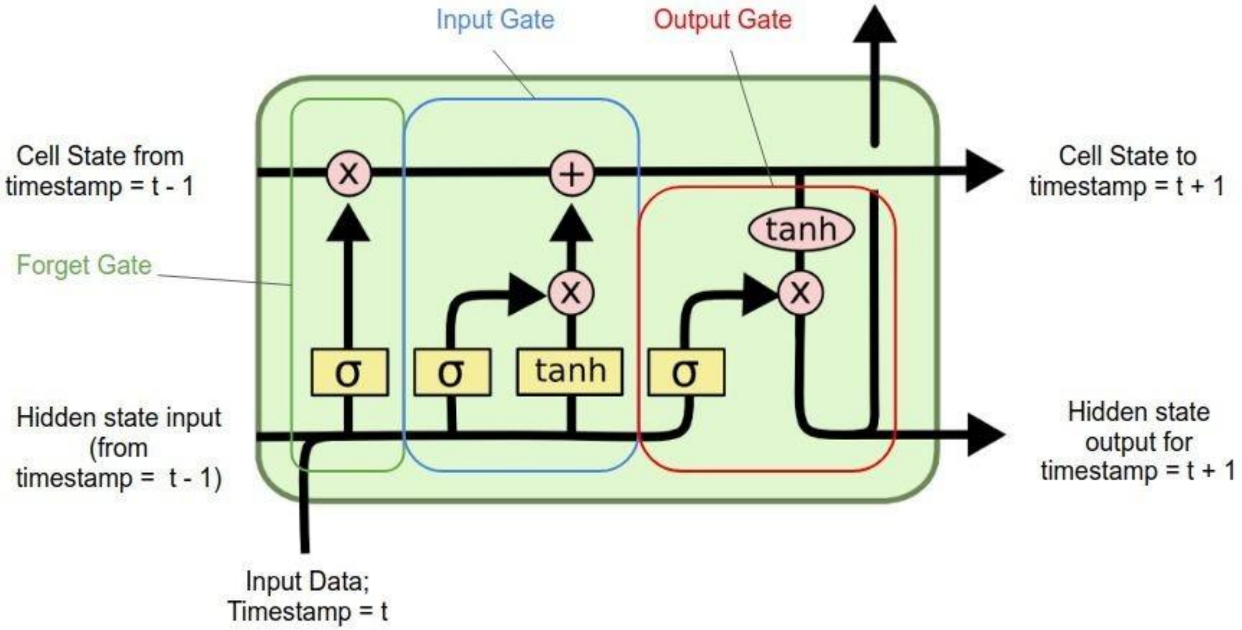
- Gated LSTM cells control information flow:
  1) Forget 2) Store 3)update 4) Output



- LSTM cells are able to track information throughout many timesteps

Long-Short Term Memory (LSTM)

- Gated LSTM cells control information flow:
  1) Forget 2) Store 3)update 4) Output



- LSTM cells are able to track information throughout many timesteps

LSTMs Key Concepts

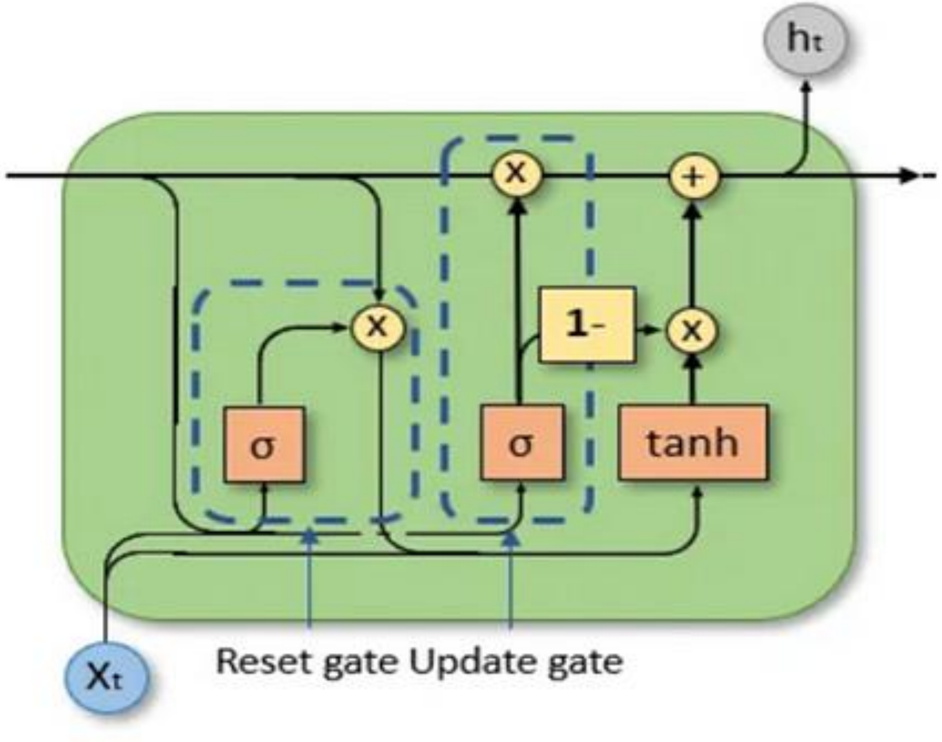
- Mantain a **cell state**

- Use **gates** to control the **flow of information**
  - **Forget** gate gets rid of irrelevant information
  - **Store** relevant information from current input
  - Selectively **update** the cell state
  - **Output** gate returns a filtered version of the cell state

- Backpropagation through time with partially **uninterrupted gradient flow**

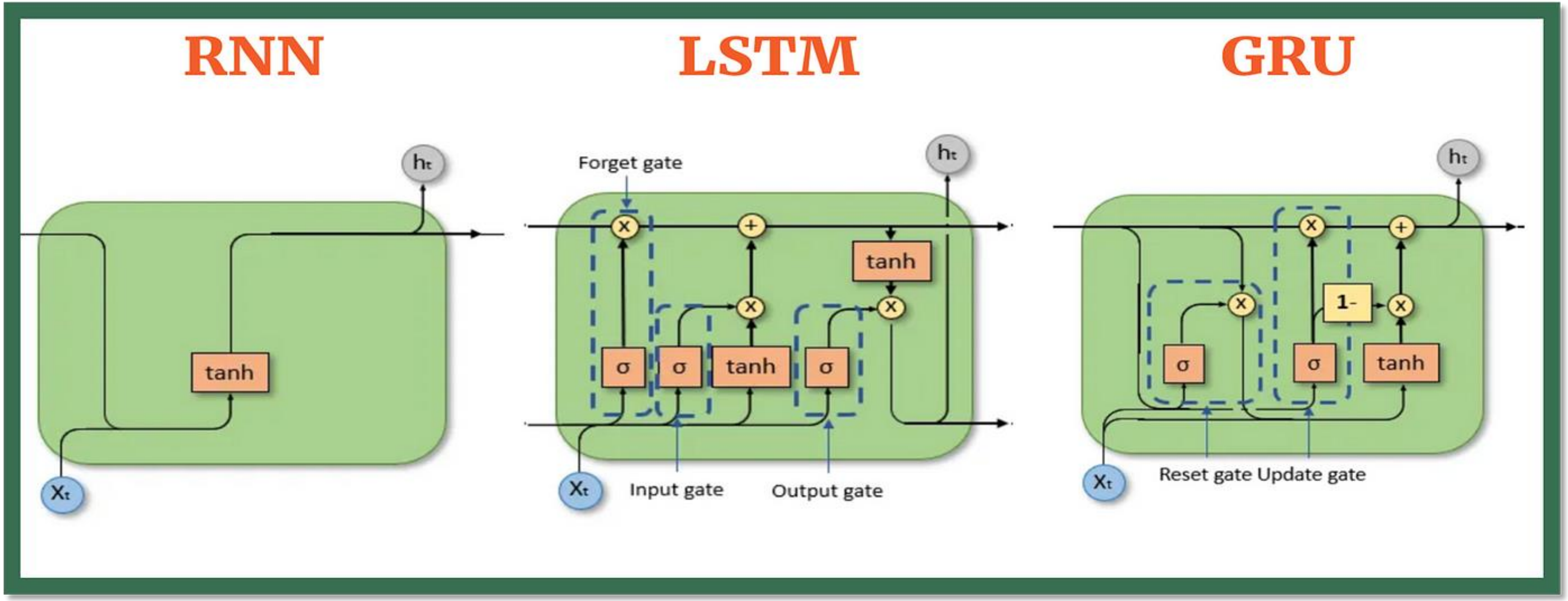
Gated Recurrent Unit (GRU)

- Simplifies LSTM by merging: (1) cell and hidden states and (2) forget and input gates

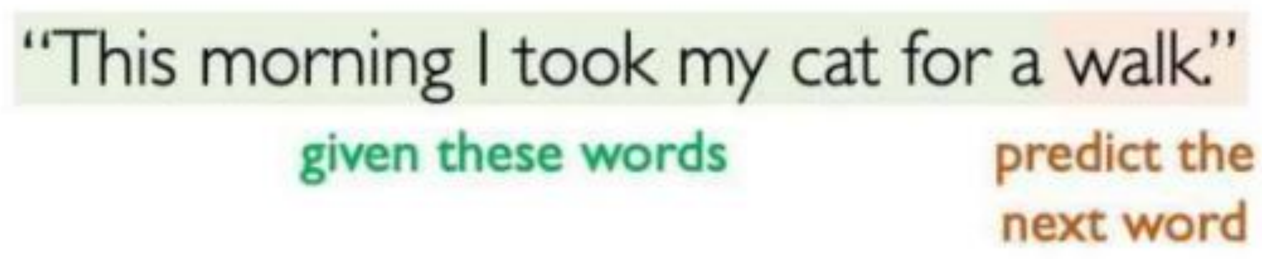


Simple RNN vs LSTM vs GRU

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA



A Sequence Modeling Problem: Predict the Next Word



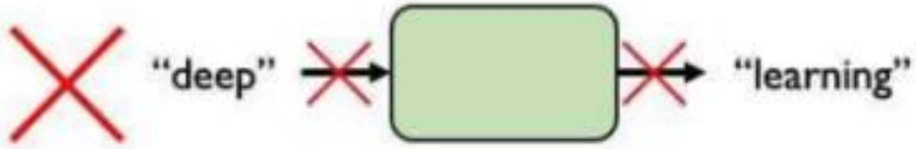
A Sequence Modeling Problem: Predict the Next Word

"This morning I took my cat for a walk."

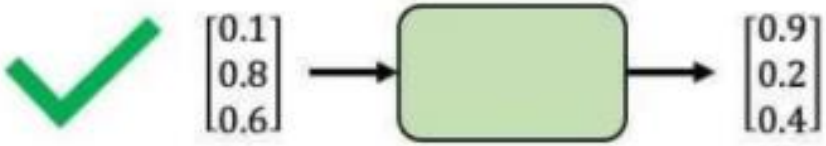
given these words

predict the next word

Representing Language to a Neural Network

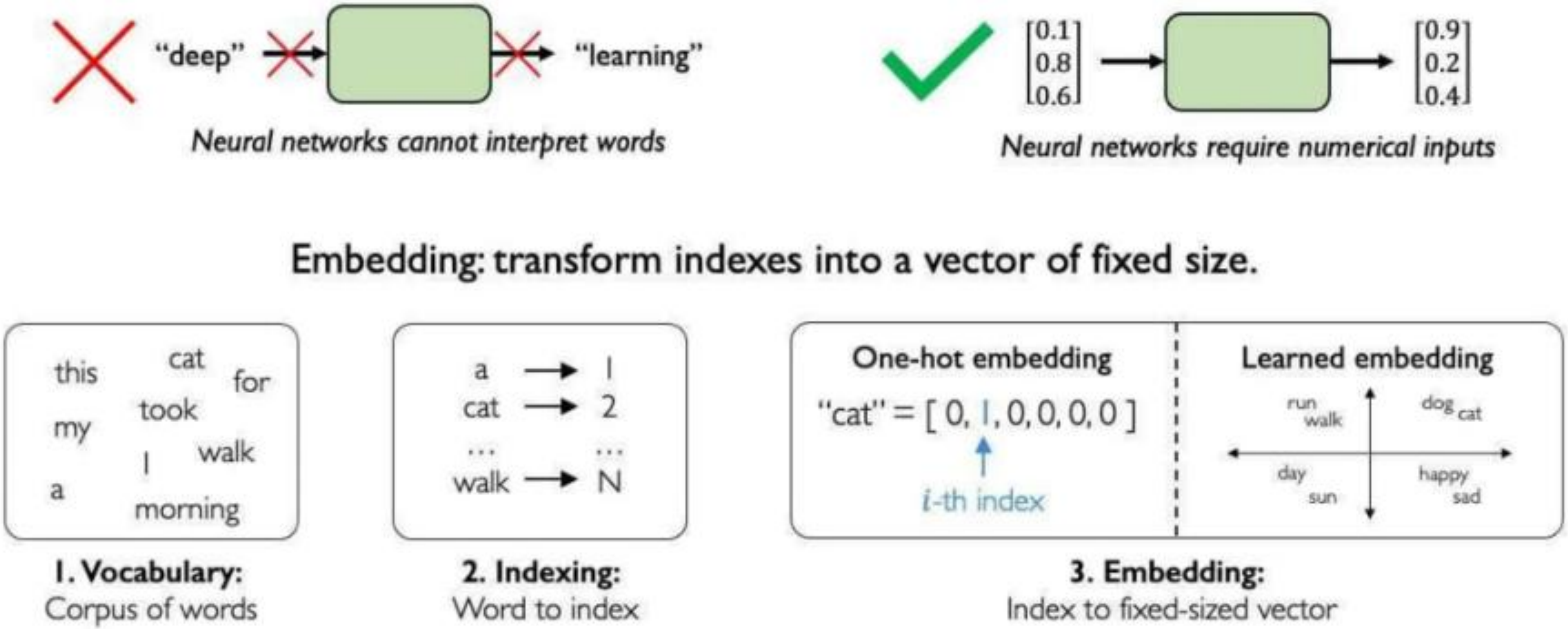


Neural networks cannot interpret words



Neural networks require numerical inputs

A Sequence Modeling Problem: Predict the Next Word



Handle Variable Sequence Lenghts

The food was great

vs.

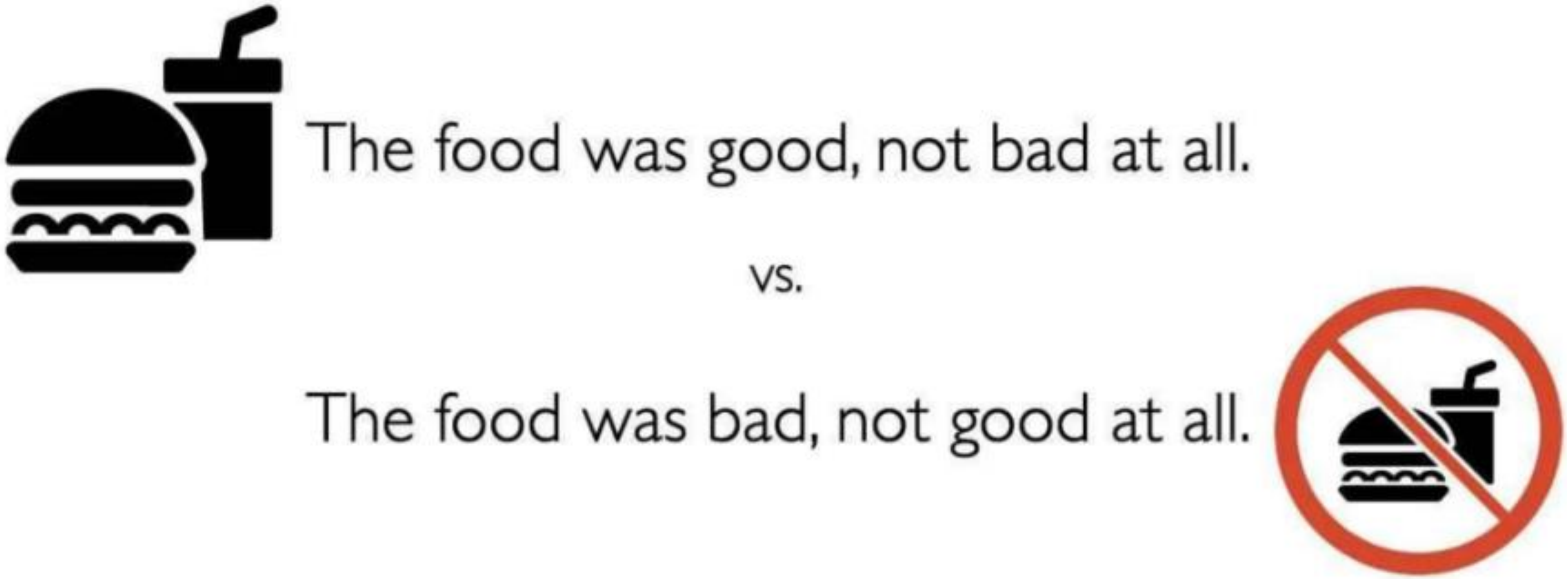
We visited a restaurant for lunch

vs.

We were hungry but cleaned the house before eating

Capture Differences in Sequence Order

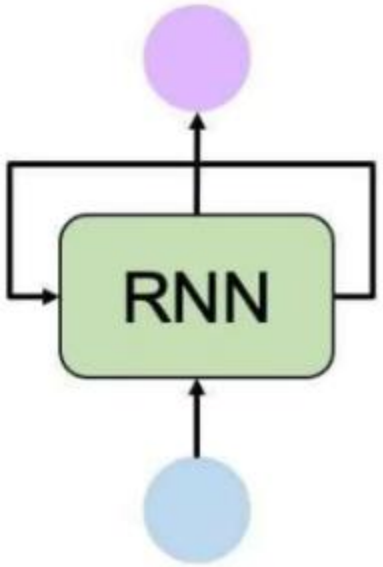
UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA



Sequence Modeling: Design Criteria

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

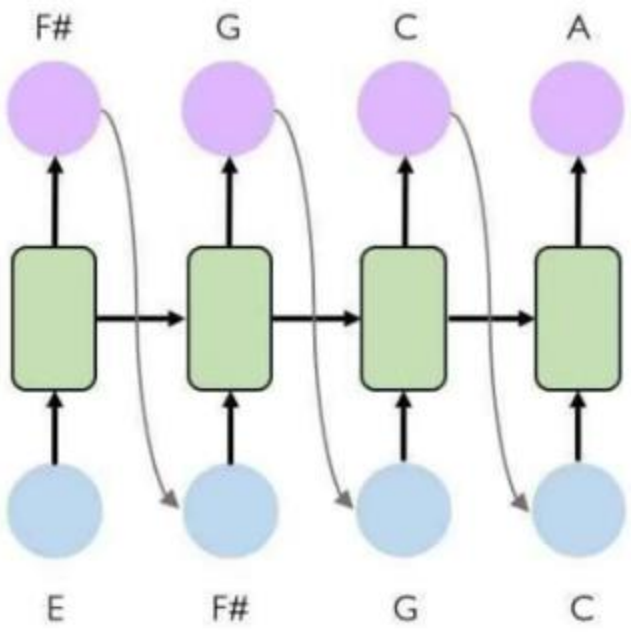
- To model sequences, we need:
  1. Handle variable-lenght sequences
  2. Track long-term depencencies
  3. Mantain information about order
  4. Share parameters across the sequence



- RNNs meet these sequence modeling design criteria

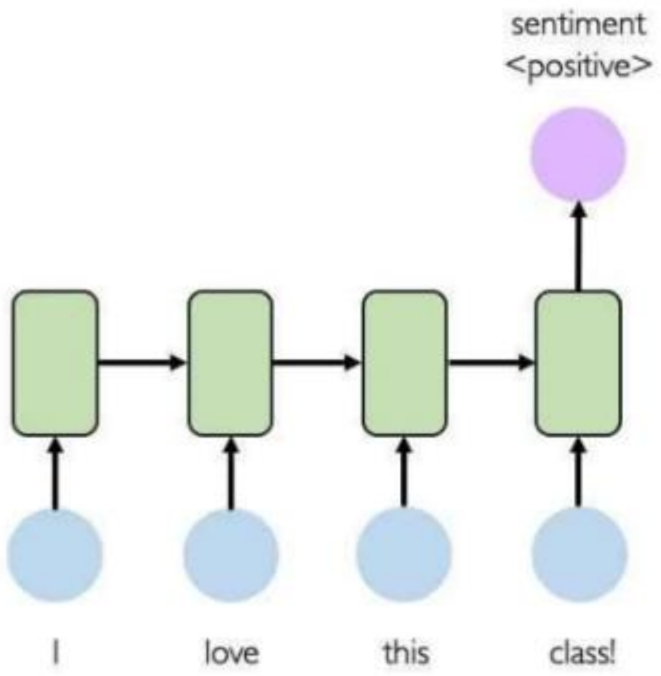
RNNs Applications and Limitations

- Example Task: Music Generation



RNNs Applications and Limitations

- Example Task: Sentiment Classification



Input: sequence of words

Output: probability of having positive sentiment

RNNs Applications and Limitations

- Limitations of RNNs:
  - Encoding bottleneck
  - Slow, no parallelization
  - Not long memory