# GEX1082 - Tópicos Especiais em Computação XXXIII
# Deep Learning

## Redes Neurais Recorrentes


1100/1101 - CIÊNCIA DA COMPUTAÇÃO

Prof. Dr. Giancarlo D. Salton

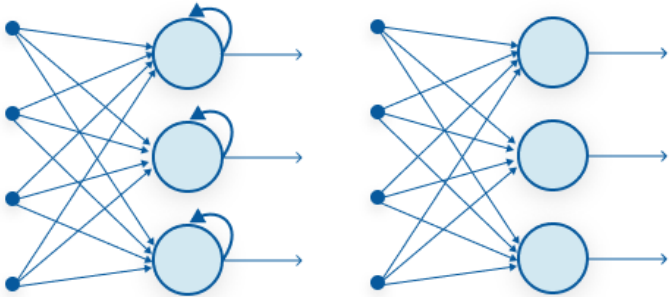
***Backpropagation through time***

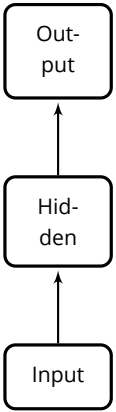
***Long short-term memory***

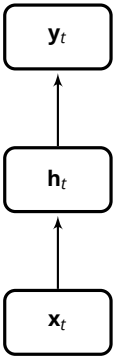
**Arquitetura Encoder-Decoder**

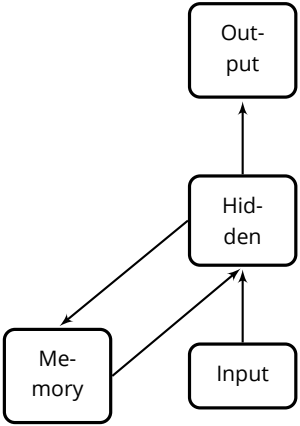
**Mecanismo de Atenção (*Attention Mechanism*)**

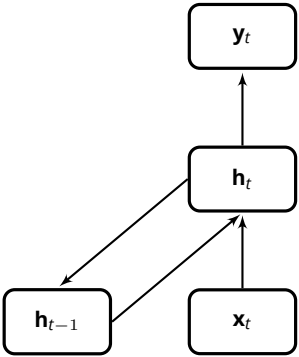
## RNNs vs. *feedforward*




Out-
put
Hid-
den
Input

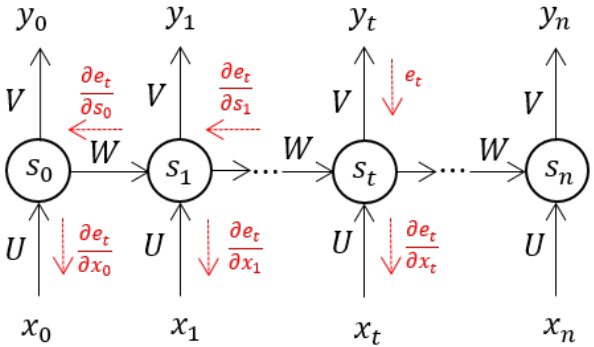

$\mathbf{y}_t$
$\mathbf{h}_t$
$\mathbf{x}_t$


Out-
put
Hid-
den
Me-
mory
Input

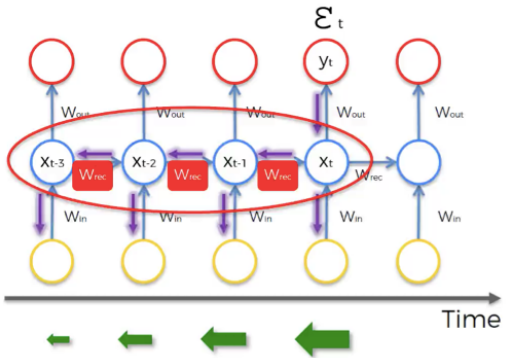

$\mathbf{y}_t$
$\mathbf{h}_t$
$\mathbf{h}_{t-1}$
$\mathbf{x}_t$

# *Backpropagation through time*

## Backpropagation through time



## Backpropagation through time




$$\frac{\partial \mathcal{E}}{\partial \theta} = \sum_{1 \le t \le T} \frac{\partial \mathcal{E}_t}{\partial \theta} \quad (3)$$

$$\frac{\partial \mathcal{E}_t}{\partial \theta} = \sum_{1 \le k \le t} \left( \frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t} \frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} \frac{\partial^+ \mathbf{x}_k}{\partial \theta} \right) \quad (4)$$

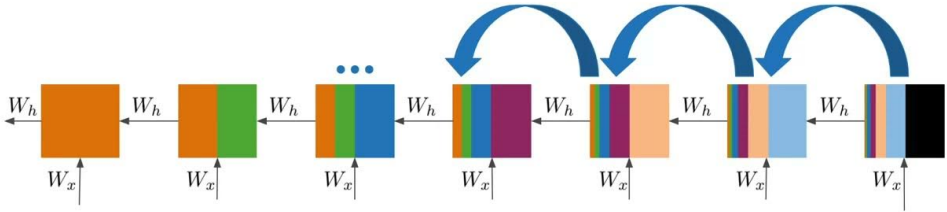
$$\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} = \prod_{t \ge i > k} \frac{\partial \mathbf{x}_i}{\partial \mathbf{x}_{i-1}} = \prod_{t \ge i > k} \mathbf{W}_{rec}^T diag(\sigma'(\mathbf{x}_{i-1})) \quad (5)$$

$W_{rec}$ ~ small ⇨ Vanishing

$W_{rec}$ ~ large ⇨ Exploding

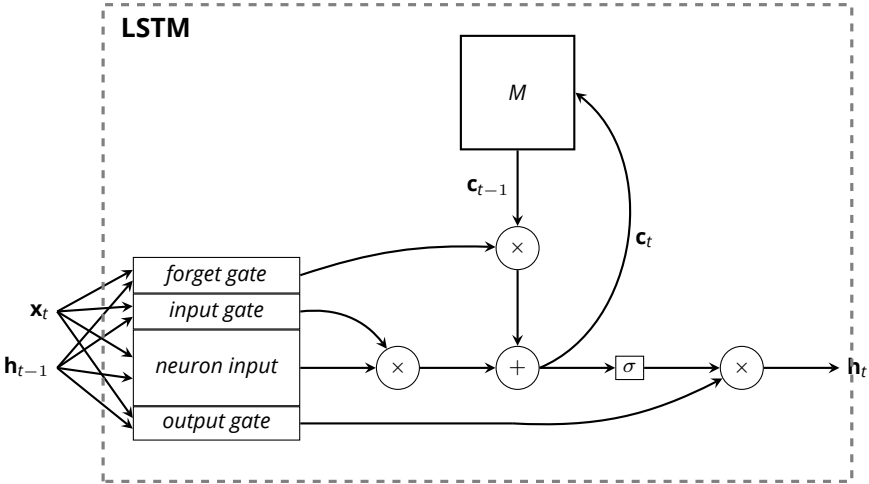
*Formula Source: Razvan Pascanu et al. (2013)*

# Backpropagation through time



# *Long short-term memory*

## *Long short-term memory* (LSTM)

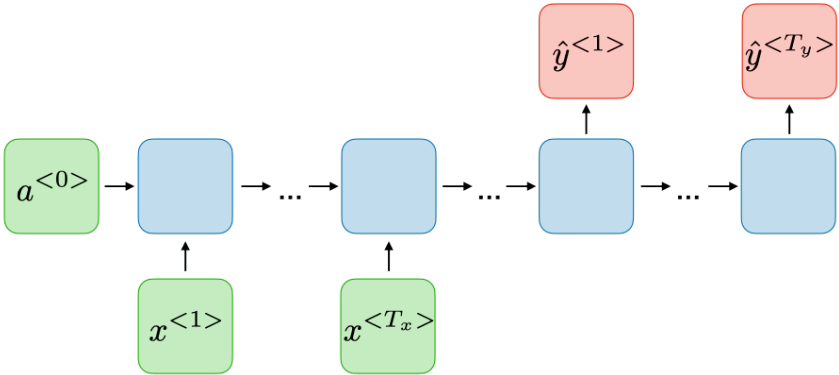


## LSTM equations

$$
\begin{aligned}
\mathbf{c_t} &= tanh(\mathbf{W}\mathbf{x}_t + \mathbf{W}\mathbf{h}_{(t-1)} + \mathbf{b}) \\
\mathbf{i}_t &= \sigma(\mathbf{W}_{ii}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{(t-1)} + \mathbf{b}_i) \\
\mathbf{f}_t &= \sigma(\mathbf{W}_{if}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{(t-1)} + \mathbf{b}_f) \\
\mathbf{c}_t &= \mathbf{f}_t \times \mathbf{c}_{(t-1)} + \mathbf{i}_t \times \mathbf{c_t} \\
\mathbf{o}_t &= \sigma(\mathbf{W}_{io}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{(t-1)} + \mathbf{b}_o) \\
\mathbf{h}_t &= \mathbf{o}_t \times \tanh(\mathbf{c}_t)
\end{aligned}
$$

# Arquitetura Encoder-Decoder

## Encoder/Decoder (M-para-M com tamanhos de sequência diferentes)

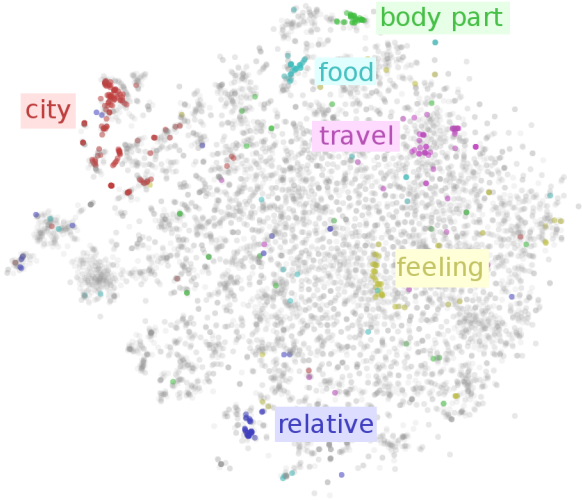


## Word embeddings

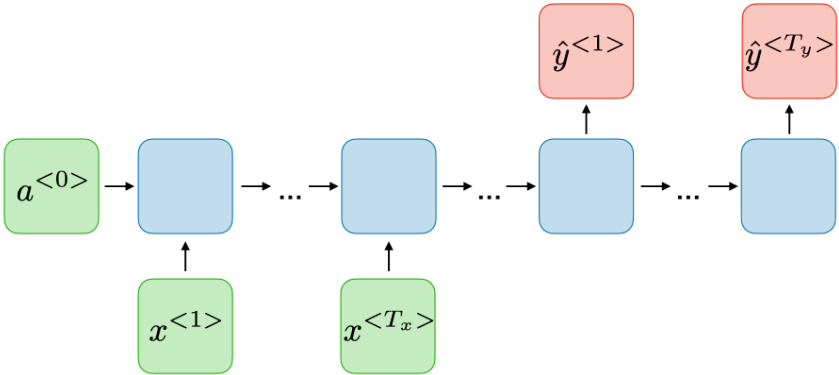
| | d1 | d2 | d3 | d4 | d5 | d6 | d7 |
|---|---|---|---|---|---|---|---|
| *dog* → | 0.6 | 0.9 | 0.1 | 0.4 | −0.7 | −0.3 | −0.2 |
| *puppy* → | 0.5 | 0.8 | −0.1 | 0.2 | −0.6 | −0.5 | −0.1 |
| *cat* → | 0.7 | −0.1 | 0.4 | 0.3 | −0.4 | −0.1 | −0.3 |
| *houses* → | −0.8 | −0.4 | −0.5 | 0.1 | −0.9 | 0.3 | 0.8 |

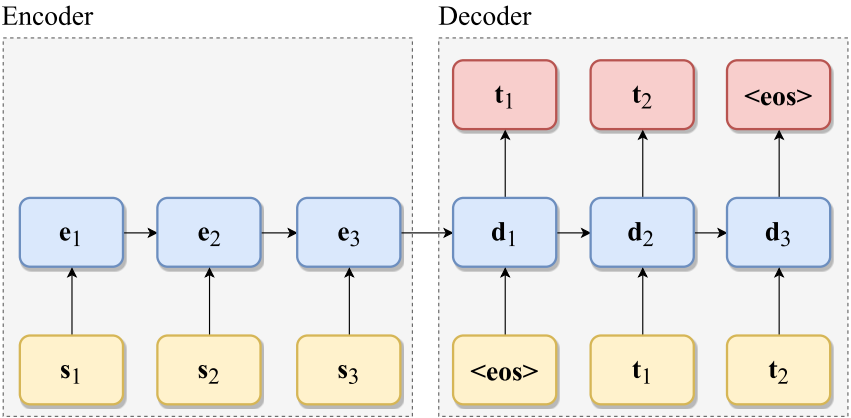
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *man* → | 0.6 | −0.2 | 0.8 | 0.9 | −0.1 | −0.9 | −0.7 |
| *woman* → | 0.7 | 0.3 | 0.9 | −0.7 | 0.1 | −0.5 | −0.4 |
| *king* → | 0.5 | −0.4 | 0.7 | 0.8 | 0.9 | −0.7 | −0.6 |
| *queen* → | 0.8 | −0.1 | 0.8 | −0.9 | 0.8 | −0.5 | −0.9 |

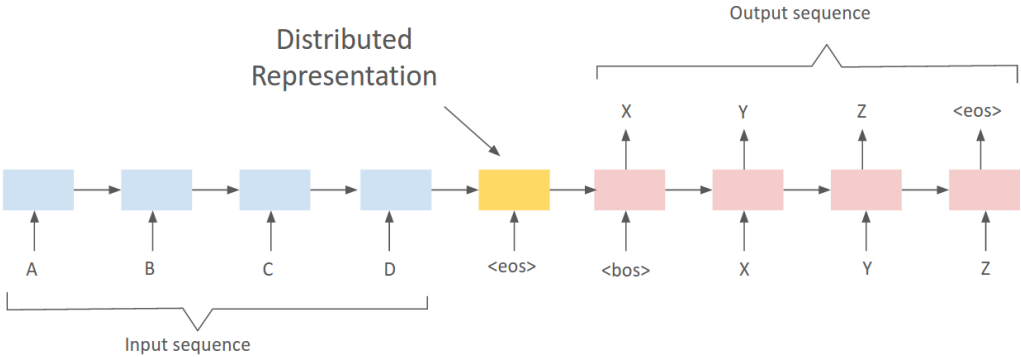
## Word embeddings

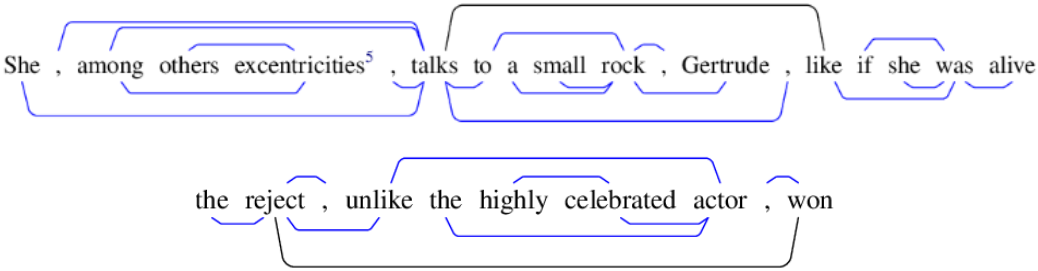


## Encoder/Decoder (M-para-M com tamanhos de sequência diferentes)

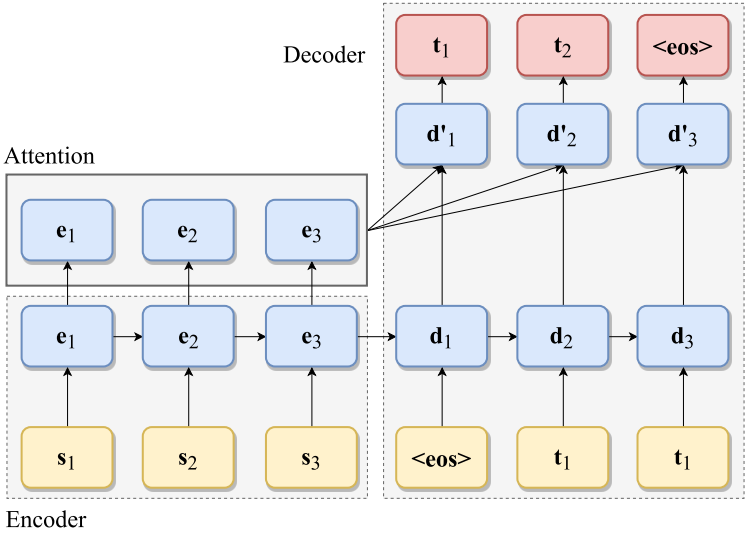


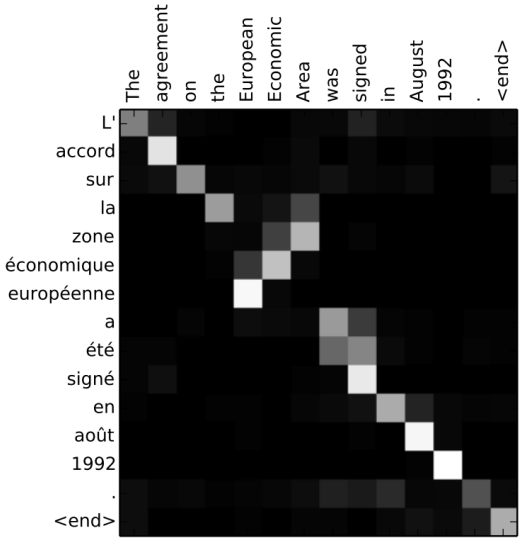

Encoder
Decoder
$e_1$
$e_2$
$e_3$
$s_1$
$s_2$
$s_3$
$d_1$
$d_2$
$d_3$
<eos>
$t_1$
$t_2$
$t_1$
$t_2$
<eos>


Output sequence
Distributed Representation
X
Y
Z
<eos>
A
B
C
D
<eos>
<bos>
X
Y
Z
Input sequence


She , among others excentricities[5] , talks to a small rock , Gertrude , like if she was alive
the reject , unlike the highly celebrated actor , won

# Mecanismo de Atenção (*Attention Mechanism*)


Decoder
Attention
Encoder
t1
t2
<eos>
d'1
d'2
d'3
e1
e2
e3
e1
e2
e3
d1
d2
d3
s1
s2
s3
<eos>
t1
t1


The
agreement
on
the
European
Economic
Area
was
signed
in
August
1992
.
<end>
L'
accord
sur
la
zone
économique
européenne
a
été
signé
en
août
1992
.
<end>

***Backpropagation through time***

***Long short-term memory***

**Arquitetura Encoder-Decoder**

**Mecanismo de Atenção (*Attention Mechanism*)**