

Redes Neurais para Processamento Temporal

Germano Vasconcelos

Centro de Informática

Universidade Federal de Pernambuco

- Muitas tarefas requerem aprendizagem de sequência de eventos

Reconhecimento de Sequências: produzir saída específica quando uma sequência de entrada específica é apresentada

Exemplo: reconhecimento e processamento de voz,

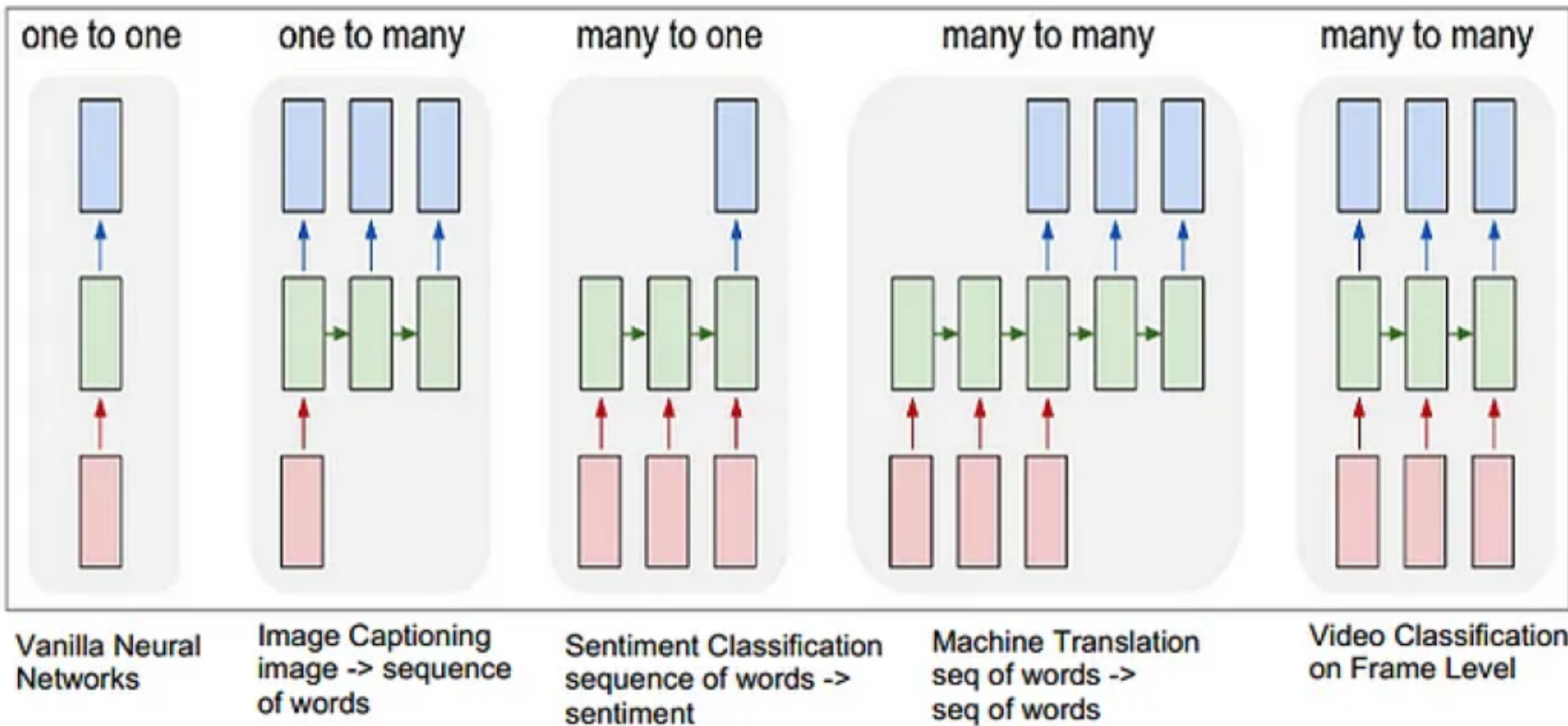
Reprodução de sequências: gerar o restante de uma sequência quando a rede recebe parte da sequência.

Exemplo: previsão de séries temporais (ações, índices, previsão do tempo e fluxos)

Associação temporal: produzir sequência específica de saída em resposta a uma sequência específica de entrada.

Exemplo: geração de fala, geração de textos

Aplicações Atuais de Redes Recorrentes



Redes Recorrentes

- ◆ Possuem 1 (um) ou mais loops de feedback que permitem a representação de informação temporal
- ◆ Representações que criam memórias de curto prazo

Motivação para Redes Recorrentes



➤ Redes Feedforward

- Informação flue apenas no sentido para frente
- Um padrão de entrada produz Uma saída
- Sem noção de tempo ou memória de estados anteriores

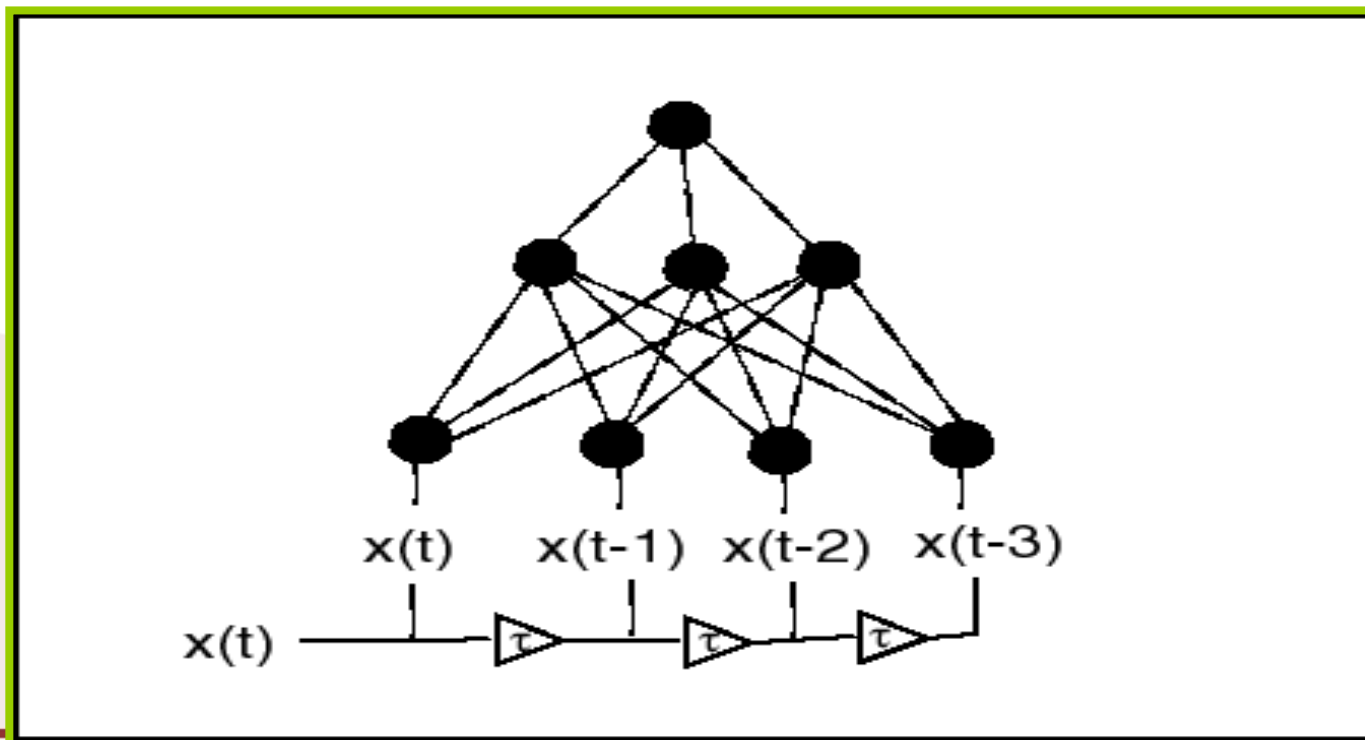
➤ Recorrência

- Nós se conectam para trás com outros nós e consigo mesmos
 - Fluxo de informação é multi-direcional
 - Sentido de tempo e memória de estados anteriores
-
- Sistemas biológicos possuem recorrência (estruturas feedforward também existem)



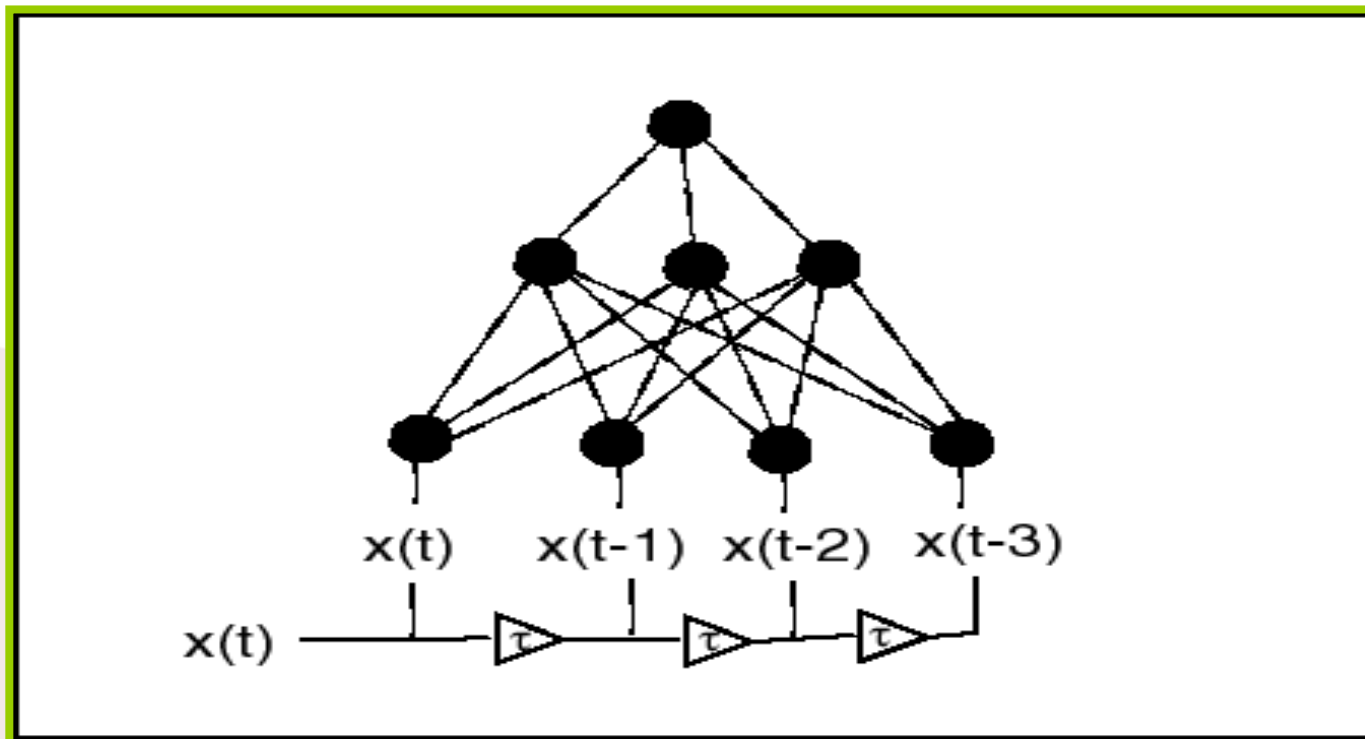
Como Representar Tempo em Redes Feedforward?

- Apresentando a informação temporal relevante ao mesmo tempo!
- Permite uso de backpropagation convencional (e outros)



Como Representar Tempo em Redes Feedforward?

- Potencial desvantagem: memória limitada ao tamanho da “linha de retardo” (número de retardos)
- Se um número grande de entradas for necessário pode tornar treinamento lento e necessidade de muitos exemplos



Previsão de Séries Temporais



- Séries Temporais = sequência de valores medidos em momentos sucessivos no tempo

- Exemplos:
 - Evolução de variações cambiais
 - Evolução do preço de ações
 - Monitoramento de sinais biológicos
 - Previsão meteorológica
 - Previsão de demanda de produtos e serviços (*bike*)

- Objetivo: prever o valor futuro em um ou vários passos



Exemplo: Previsão (*Forecasting*)

- Dado um conjunto de n valores de uma variável ($y(t_1), y(t_2), \dots, y(t_n)$) em uma sequência de tempo $t_1, t_2, \dots, t_n$,
- Prever o valor $y(t_{n+1})$ num futuro t_{n+1}

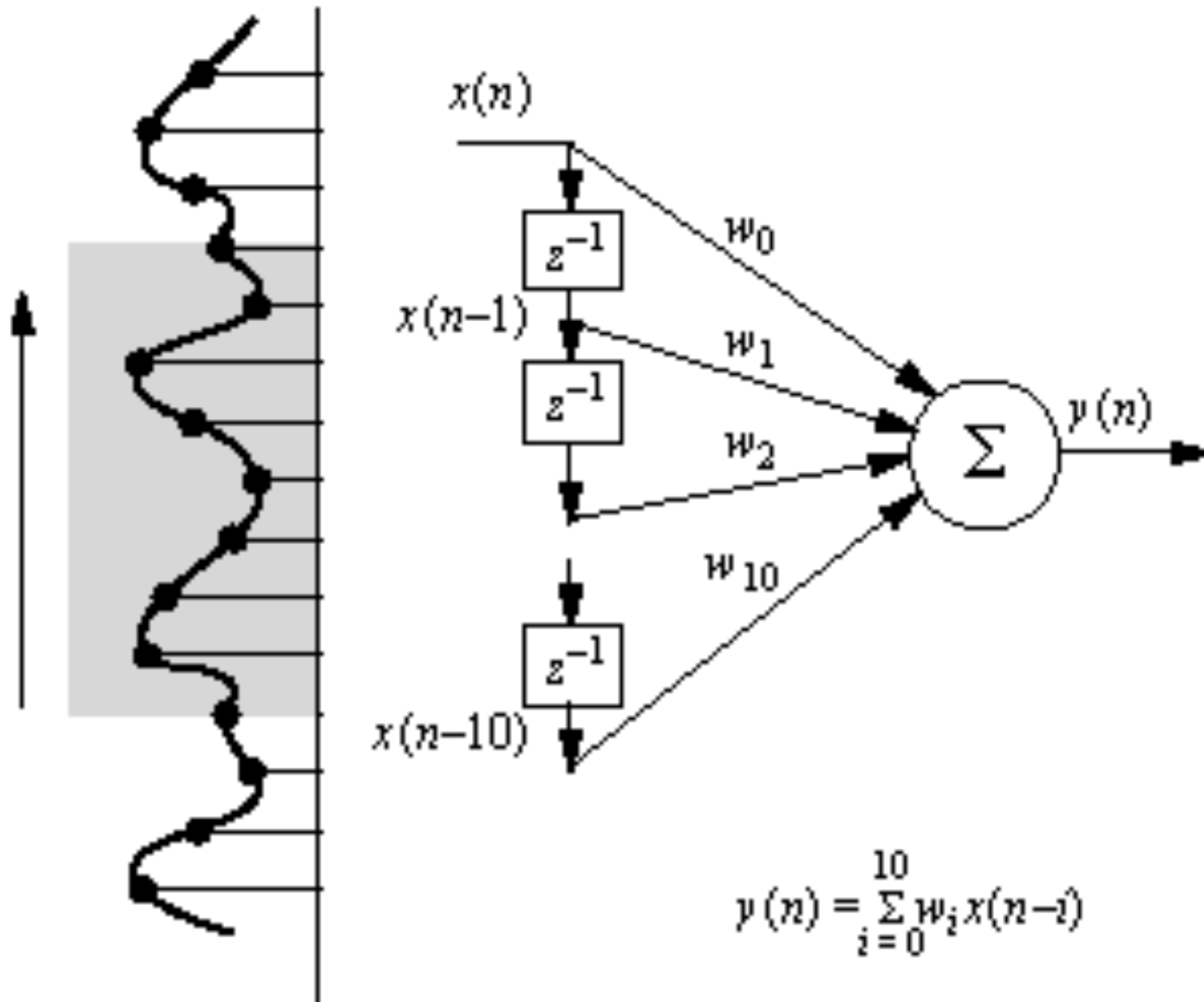
Previsão – Questões Relevantes



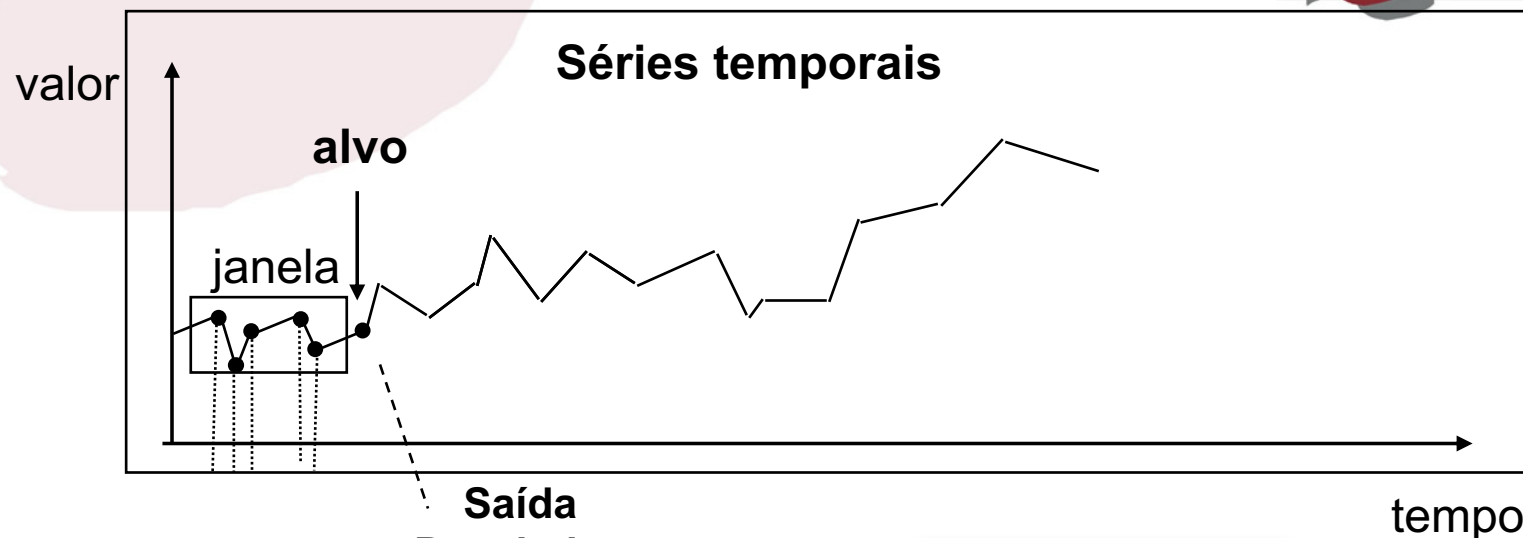
- Definição da janela de entrada
- Definição do horizonte de previsão
- Definição de outras variáveis explicativas



Previsão com uma Rede MLP



Previsão de Séries Temporais



Entradas
da rede =
 n valores
passados

Ex: 5 valores
passados

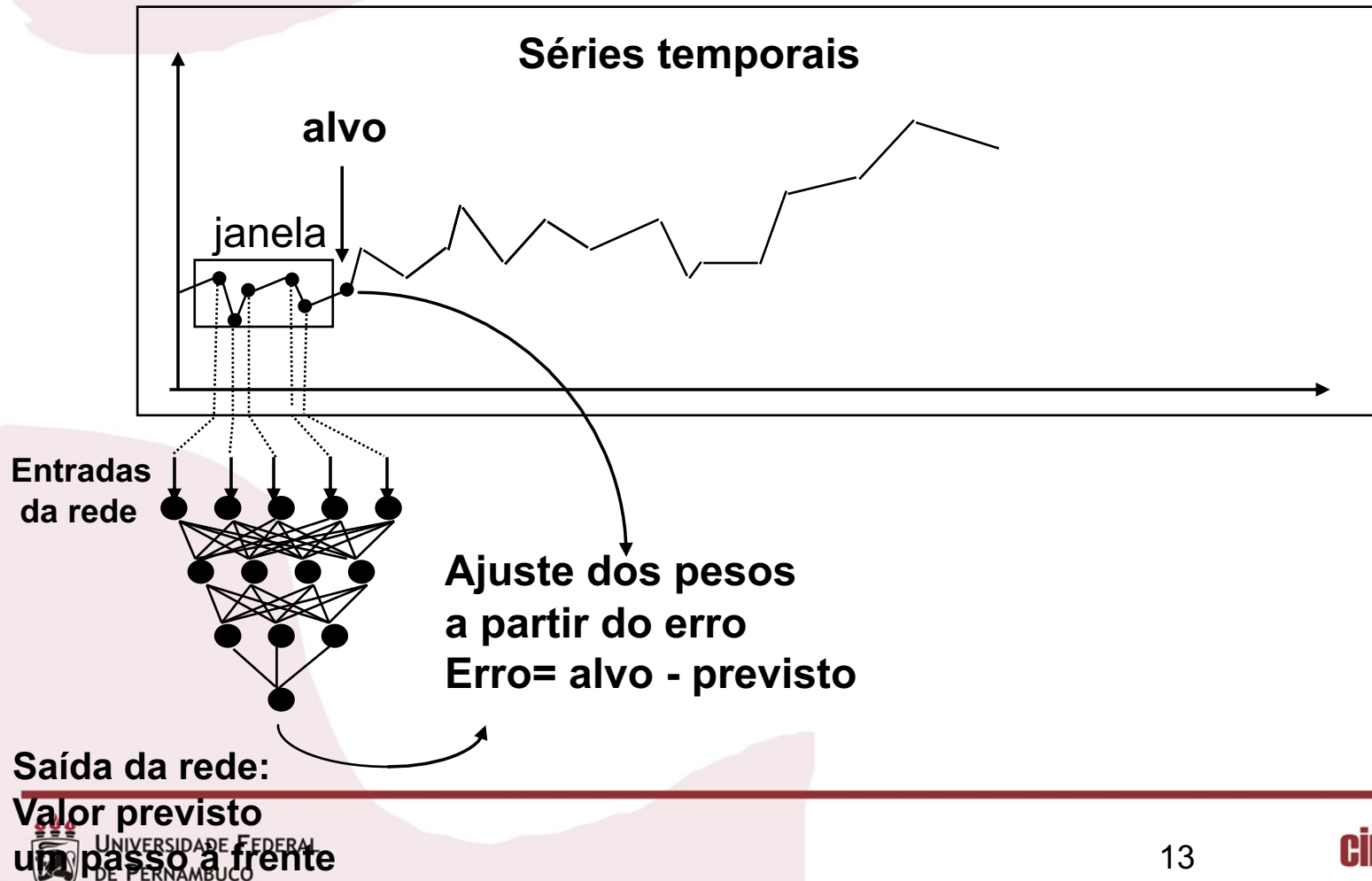
Saída
Desejada =
valor da série
 k passos à
frente

**Definição da
janela de entrada**

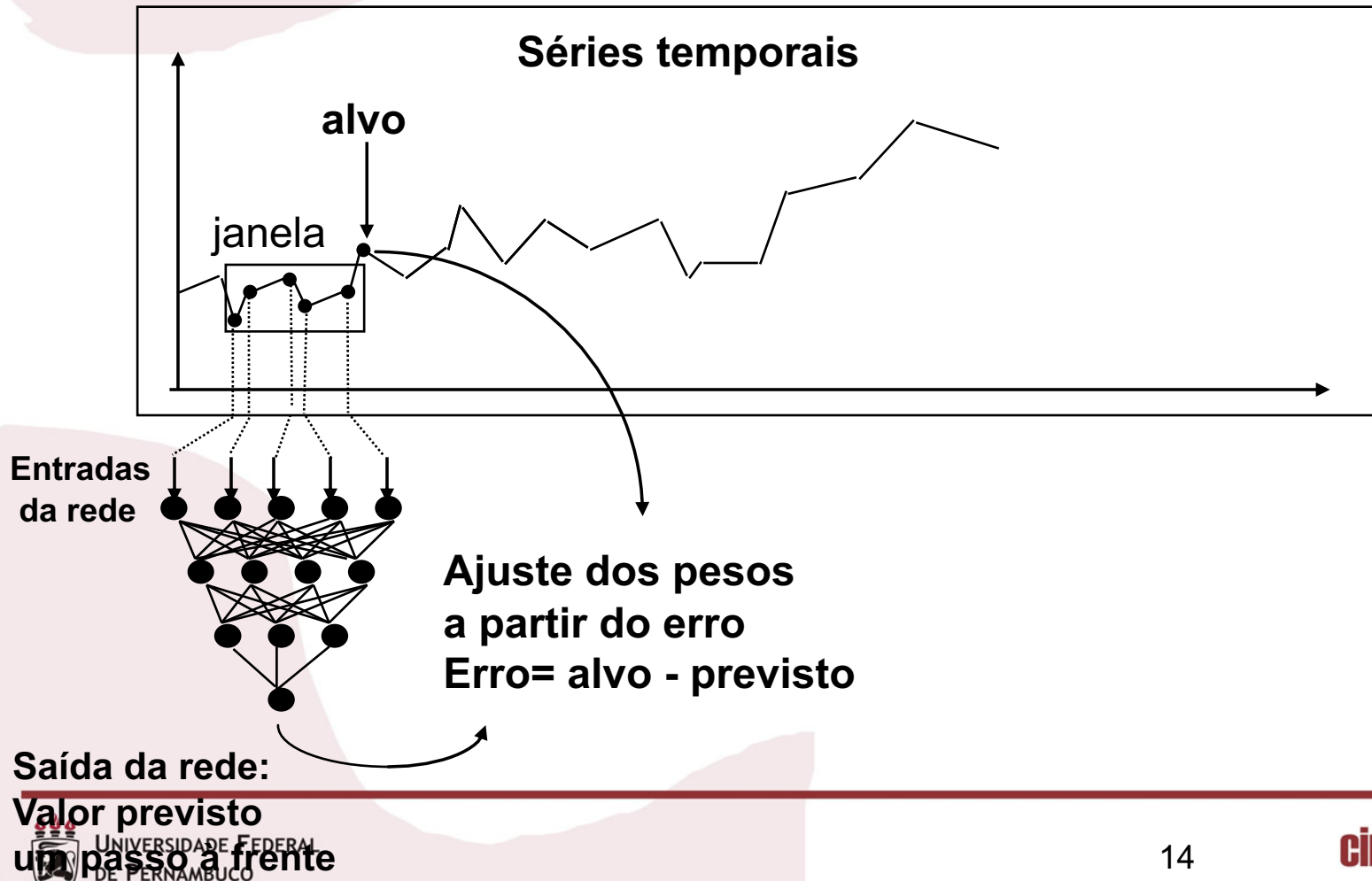
**Definição da
janela de saída**

Ex: valor um
passo à frente

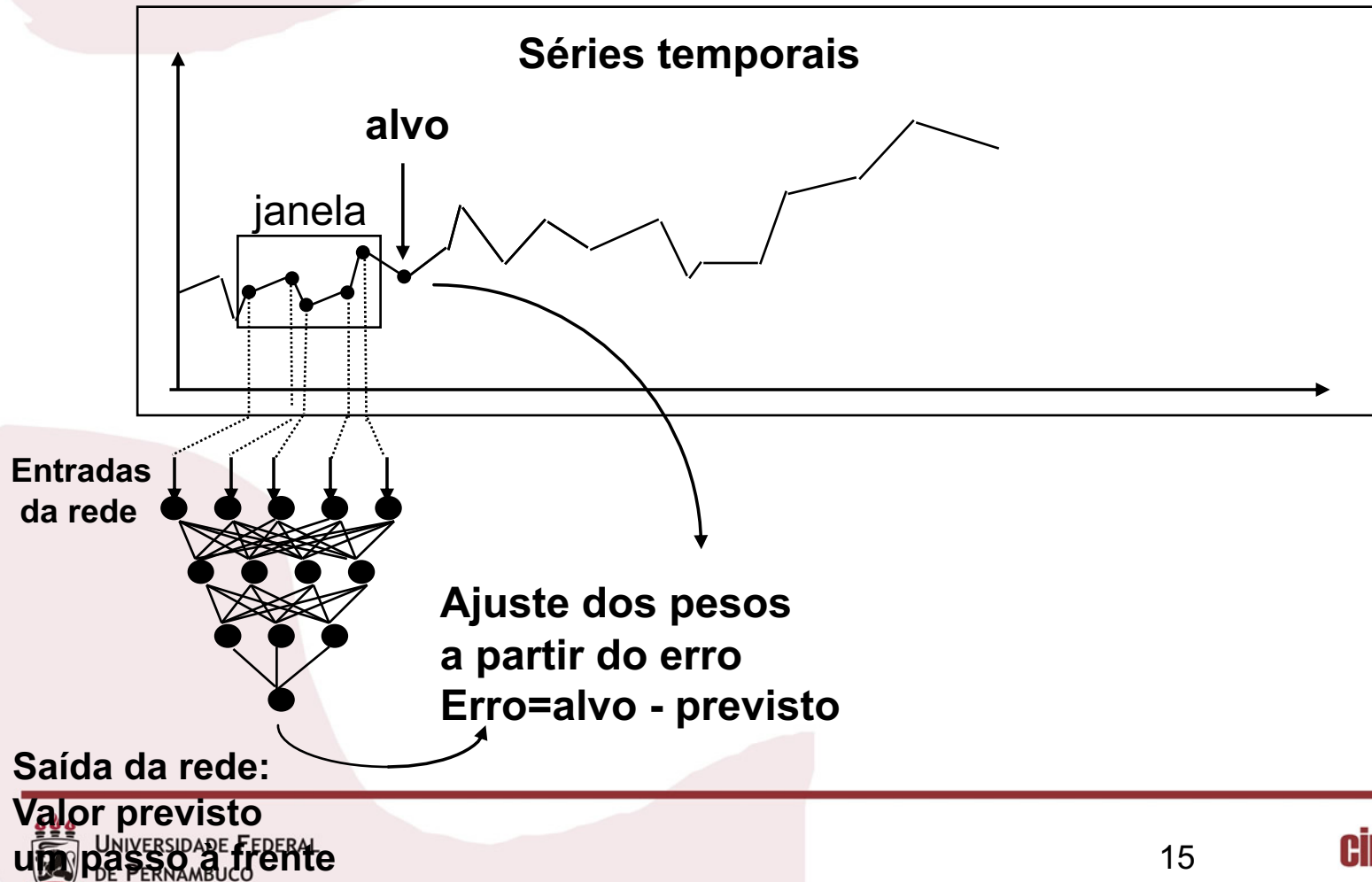
Exemplo: previsão utilizando apenas a série histórica como entrada



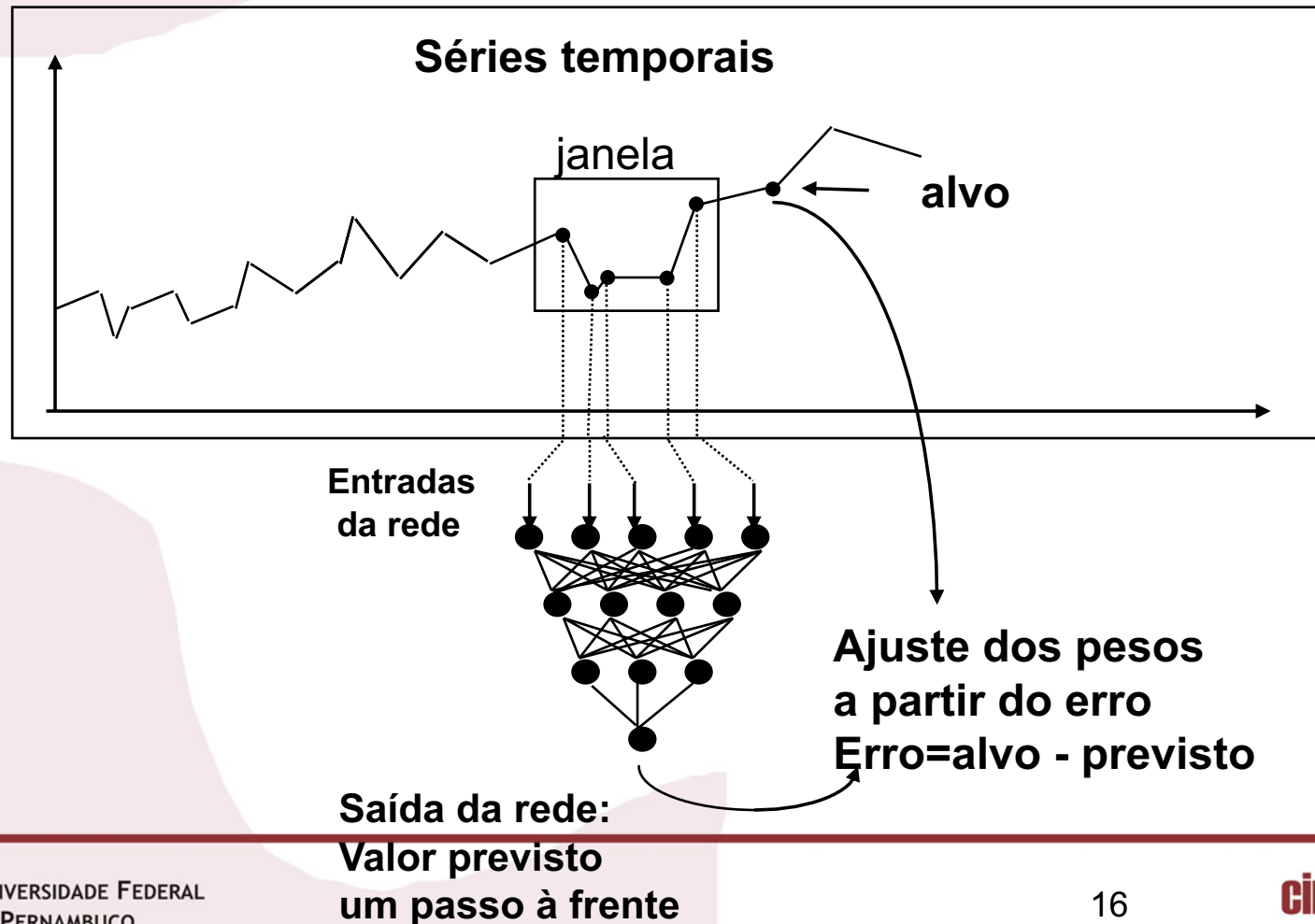
Exemplo: previsão utilizando apenas a série histórica como entrada



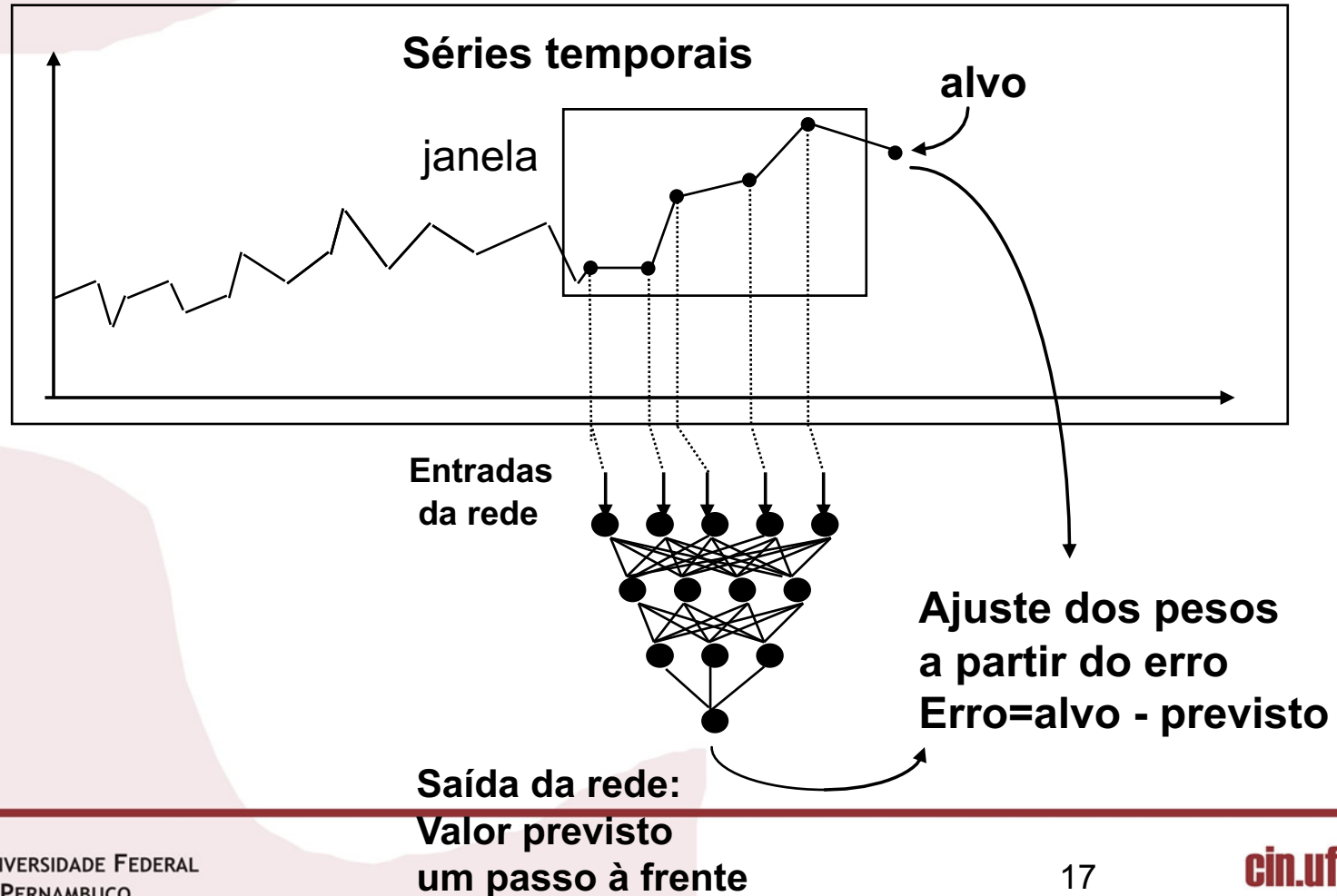
Exemplo: previsão utilizando apenas a série histórica como entrada



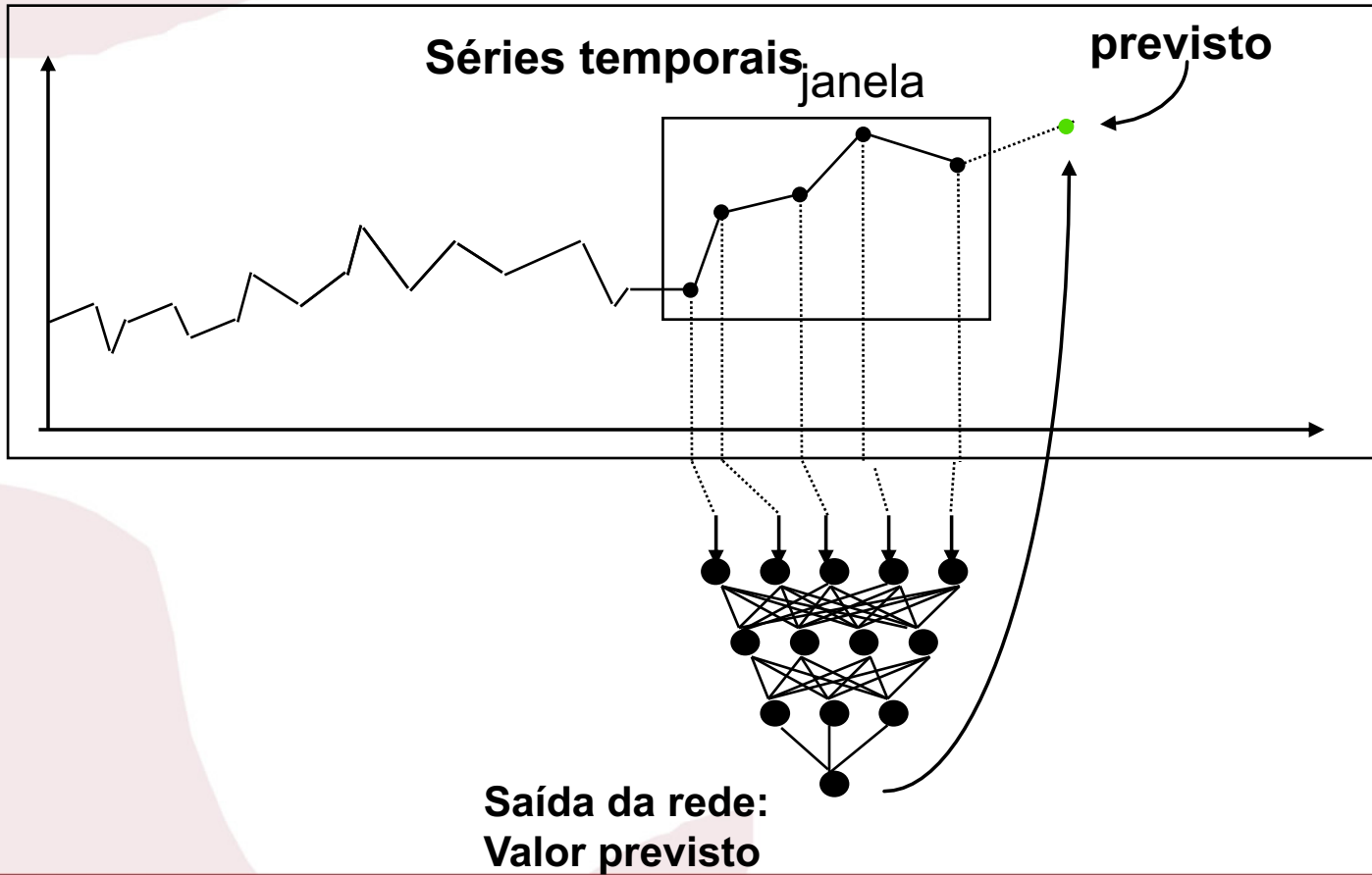
Exemplo: previsão utilizando apenas a série histórica como entrada



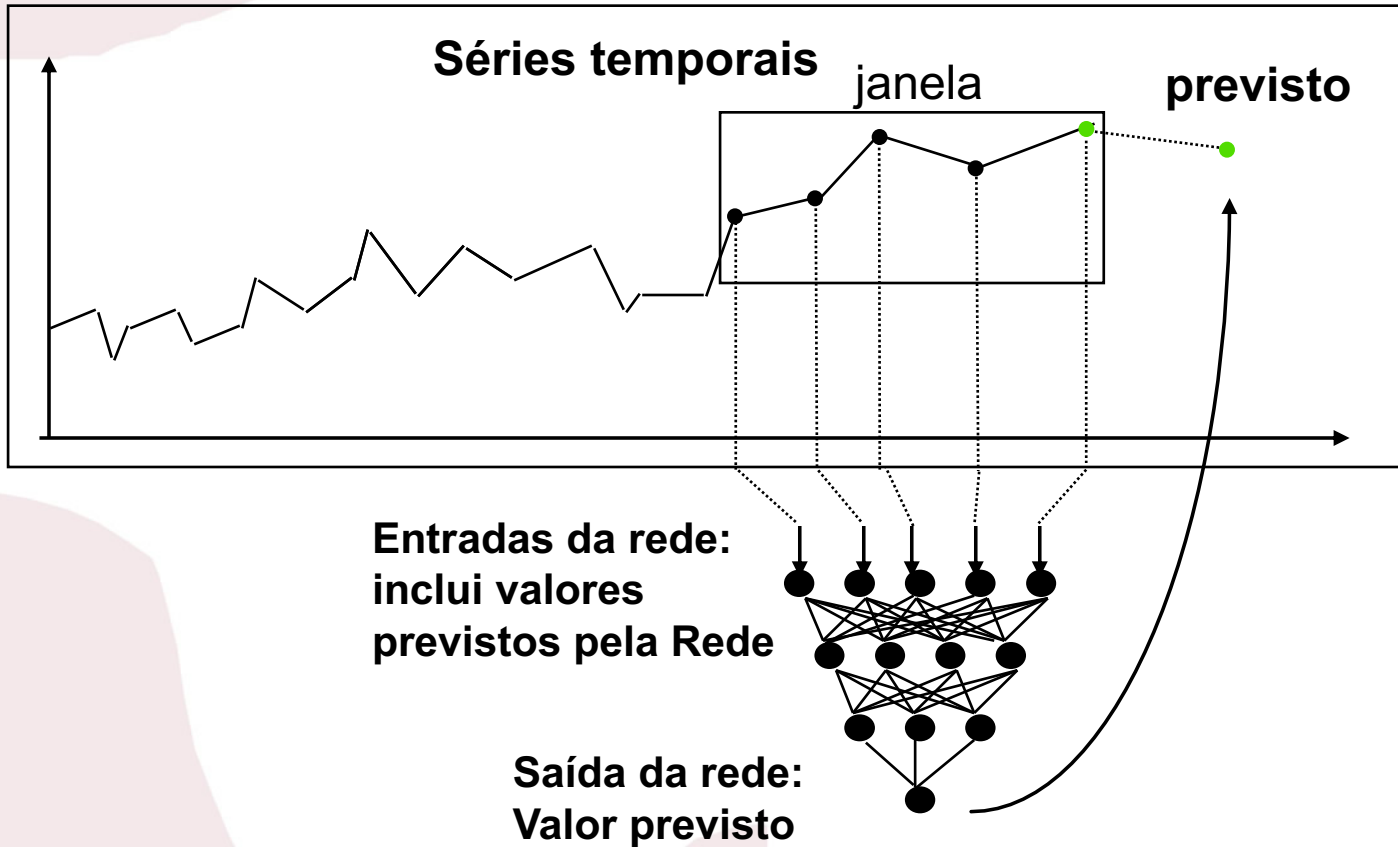
Exemplo: previsão utilizando apenas a série histórica como entrada



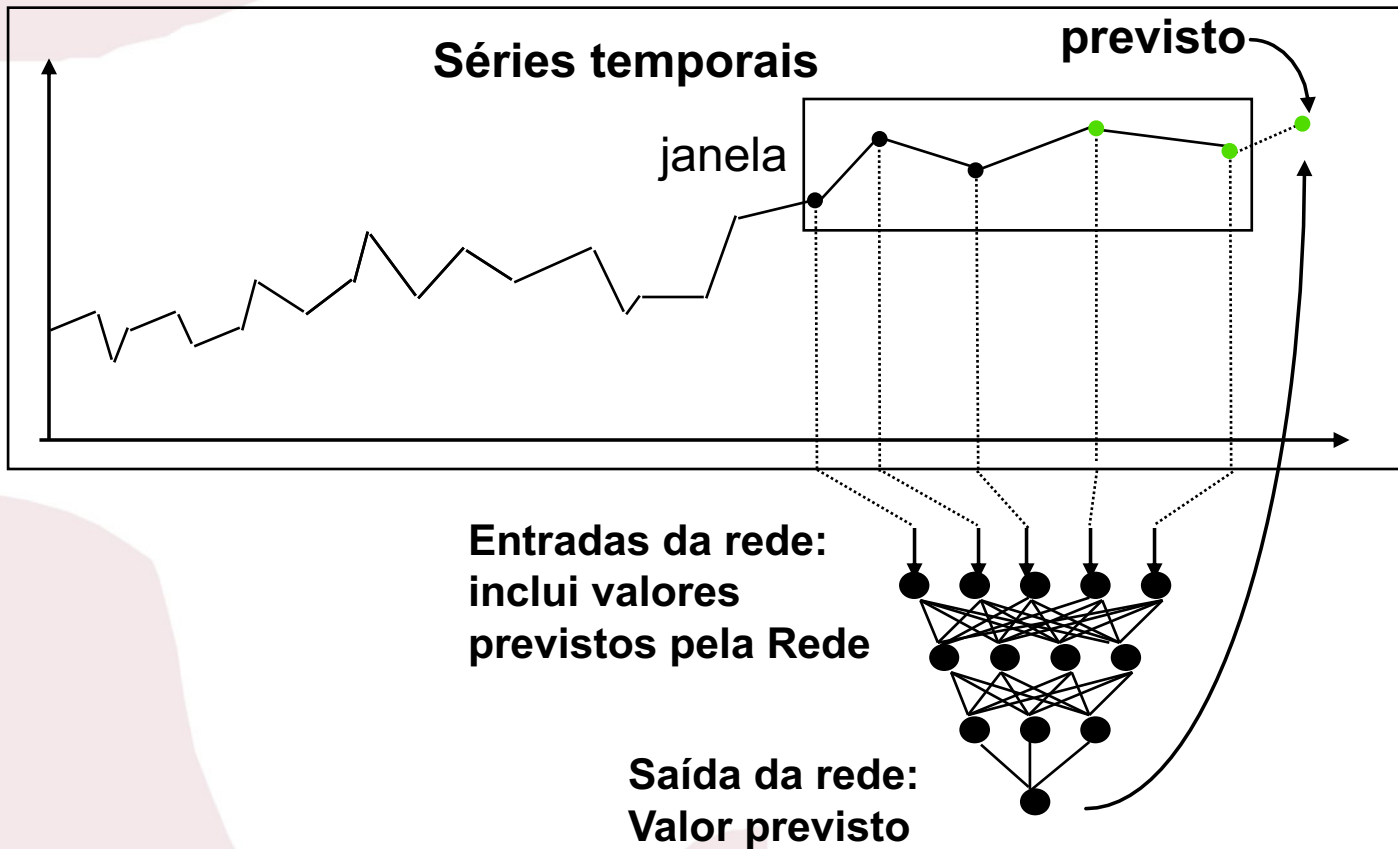
Exemplo: previsão utilizando apenas a série histórica como entrada



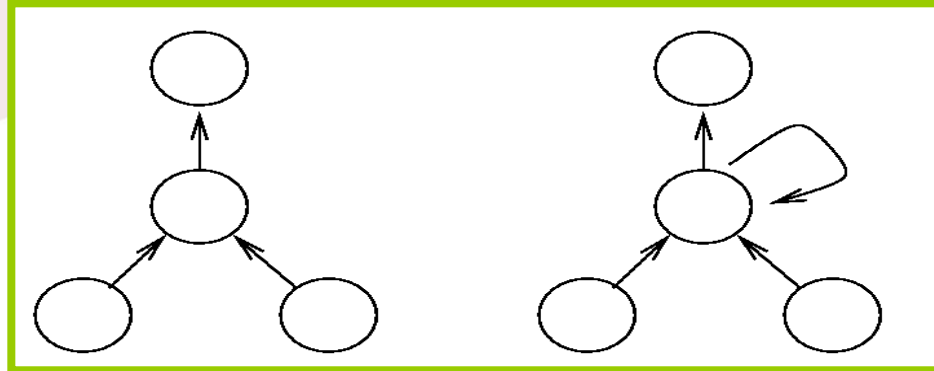
Exemplo: previsão utilizando apenas a série histórica como entrada



Exemplo: previsão utilizando apenas a série histórica como entrada



Recorrente vs. Feedforward



- Entrada para a Rede Recorrente contem termo que reflete o estado da rede (ativação da unidade escondida) antes da apresentação do padrão atual
- Quando padrões seguintes são apresentados, os estados das unidades escondidas e de saídas são uma função de tudo o que a rede viu até então
- *Comportamento da rede é baseada na sua história, então apresentação e representação precisa ser pensada temporalmente*

 **Centro
de Informática**
U · F · P · F

```
graph BT; INPUT[INPUT] --> HIDDEN[HIDDEN]; HIDDEN --> OUTPUT[OUTPUT]; CONTEXT[CONTEXT] --> HIDDEN; HIDDEN -.-> CONTEXT;
```

Unidades de contexto no tempo t
guardam uma cópia das ativações
unidades escondidas no tempo $t-1$

output: years ago boy and girl ..

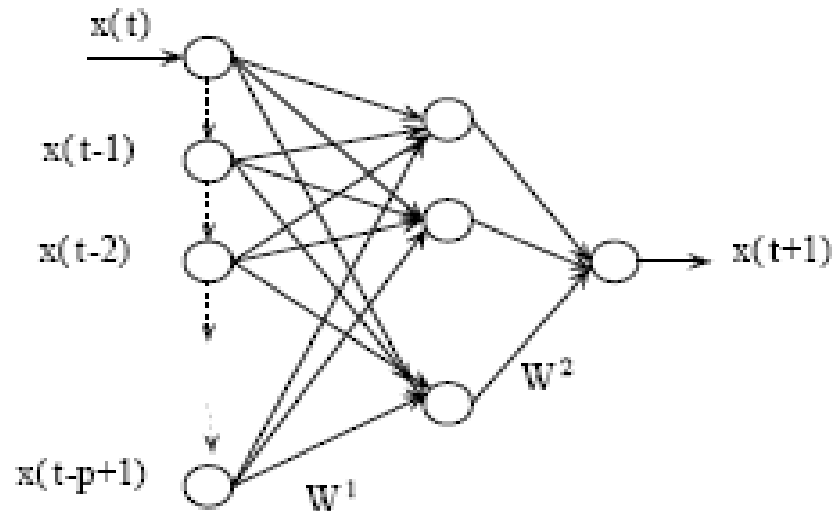
Mémoria Curta em Redes Recorrentes Simples (SRN)



- Unidades de contexto tem o papel de lembrar o estado interno anterior
- Unidades escondidas tem a tarefa de mapear ambas, a entrada externa corrente e o estado interno anterior a uma resposta desejada

Redes Feedforward para Processamento Temporal

Arquitetura:



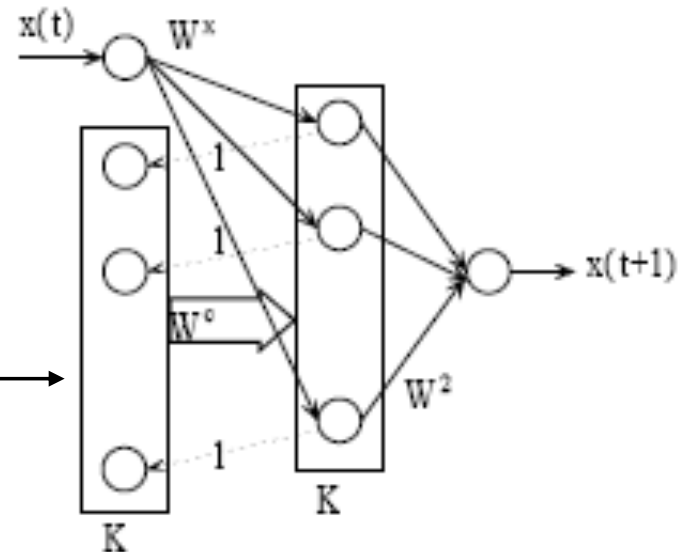
Funcionamento:

$$y = \sum_{k=1}^K w_k^2 f\left(\sum_{j=0}^{p-1} w_{kj}^1 x(t-j)\right)$$

Rede de Elman

Arquitetura:

Unidades de Contexto



Funcionamento

$$y = \sum_{k=1}^K w_k^2 h_k(t)$$

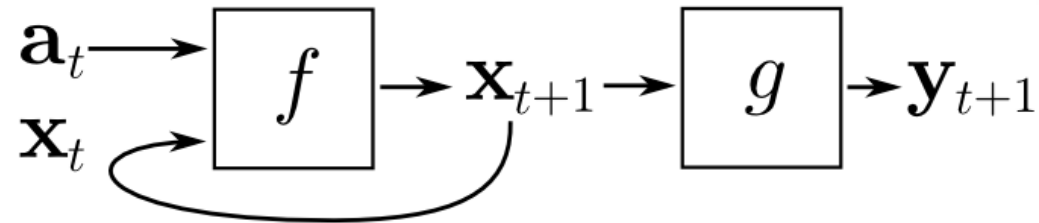
$$h_k(t) = f \left(w_k^x x(t) + \sum_{j=1}^K w_{kj}^c h_j(t-1) \right) \quad h_j(0) = 0.$$

unidades de contexto
contém cópias das
saídas das unidades
escondidas em
instantes anteriores

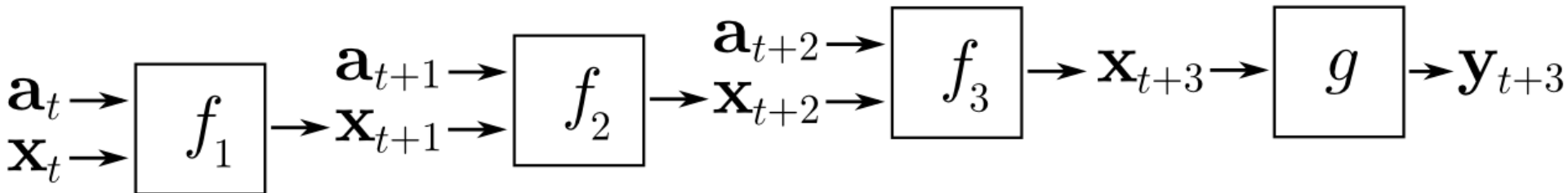
Backpropagation Through Time (BPTT)

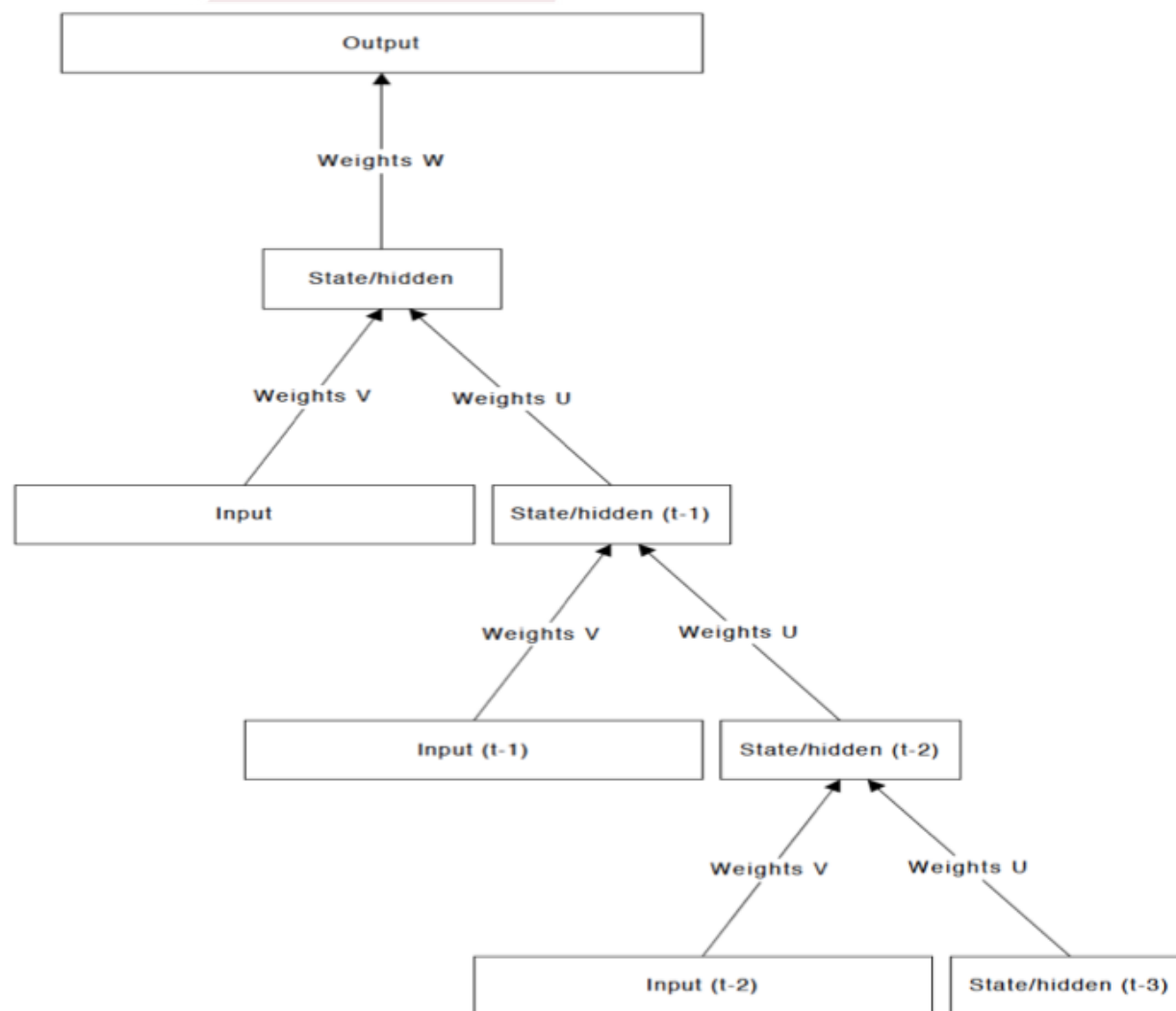
- Generalização da idéia é copiar ativações de entrada e escondidas por vários passos no tempo
- Quanto mais camadas de contexto (*copy layers*), mais história é incluída de forma explícita na computação do gradiente do erro
- Abordagem ficou conhecida como *Back Propagation Through Time (BPTT)*
- BPTT entendido como uma “aproximação” de computação do gradiente que leva em consideração não apenas as entradas mais recentes como todas vistas até o momento

Backpropagation Through Time (BPTT)



↓ unfold through time ↓



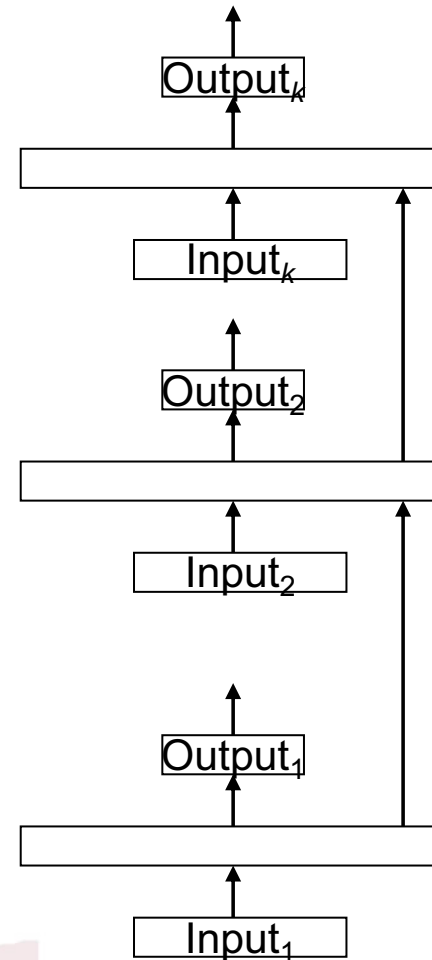
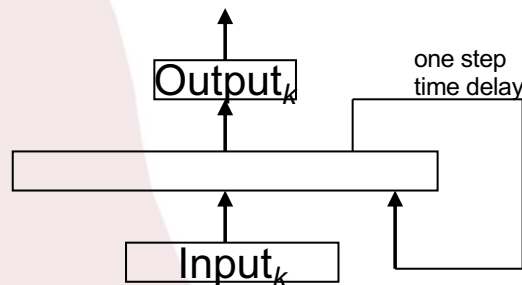


Note $k=1$
is just standard
BP with no
feedback

Figure 5: The effect of unfolding a network for BPTT ($\tau = 3$).

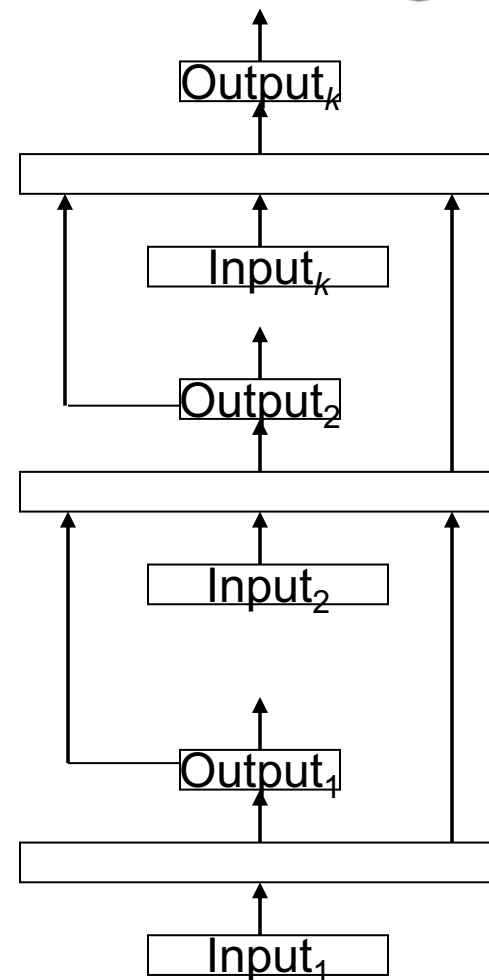
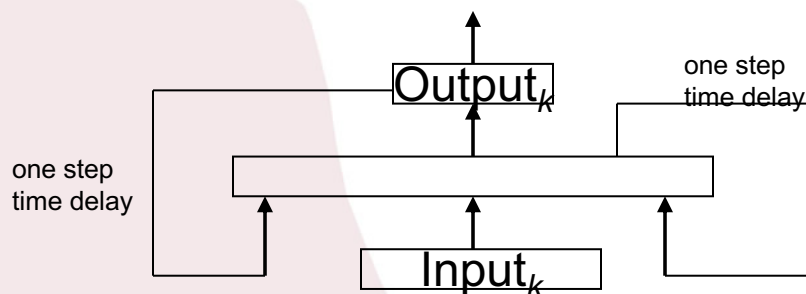
BPTT - Unfolding in Time ($k=3$) with output connections

Weights at each layer are maintained as exact copies



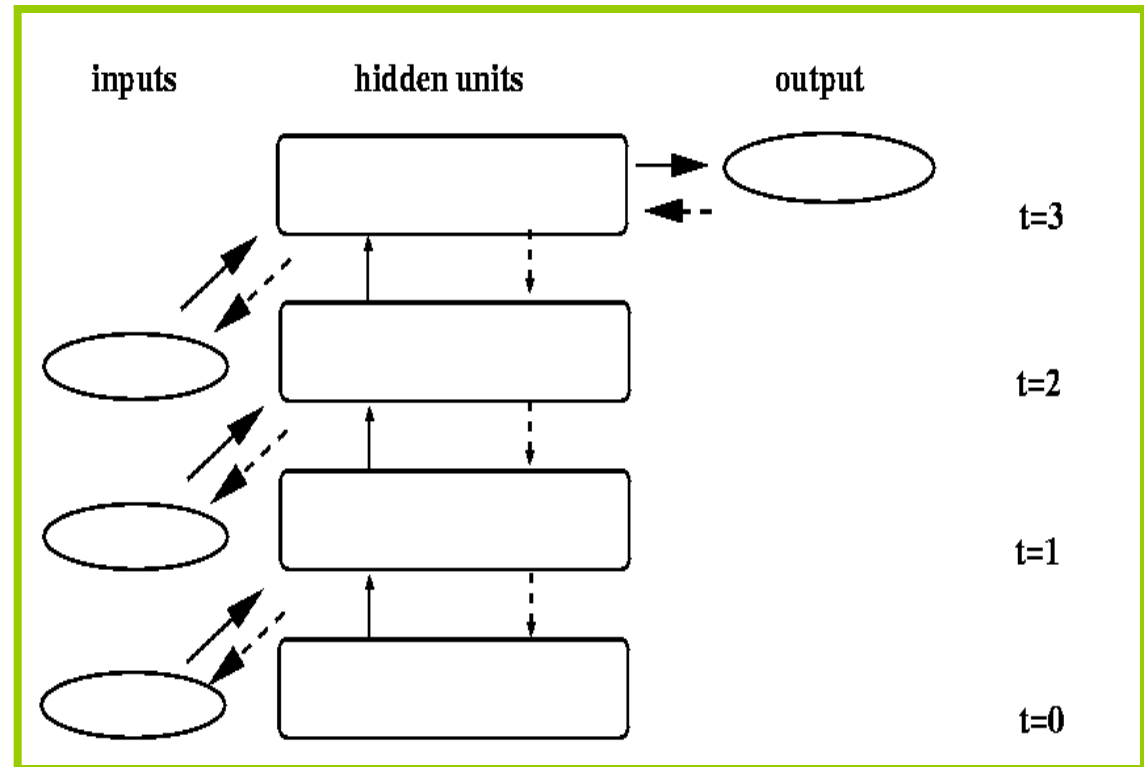
BPTT - Unfolding in Time ($k=3$) with output connections

Weights at each layer are maintained as exact copies



Backpropagation Through Time (BPTT)

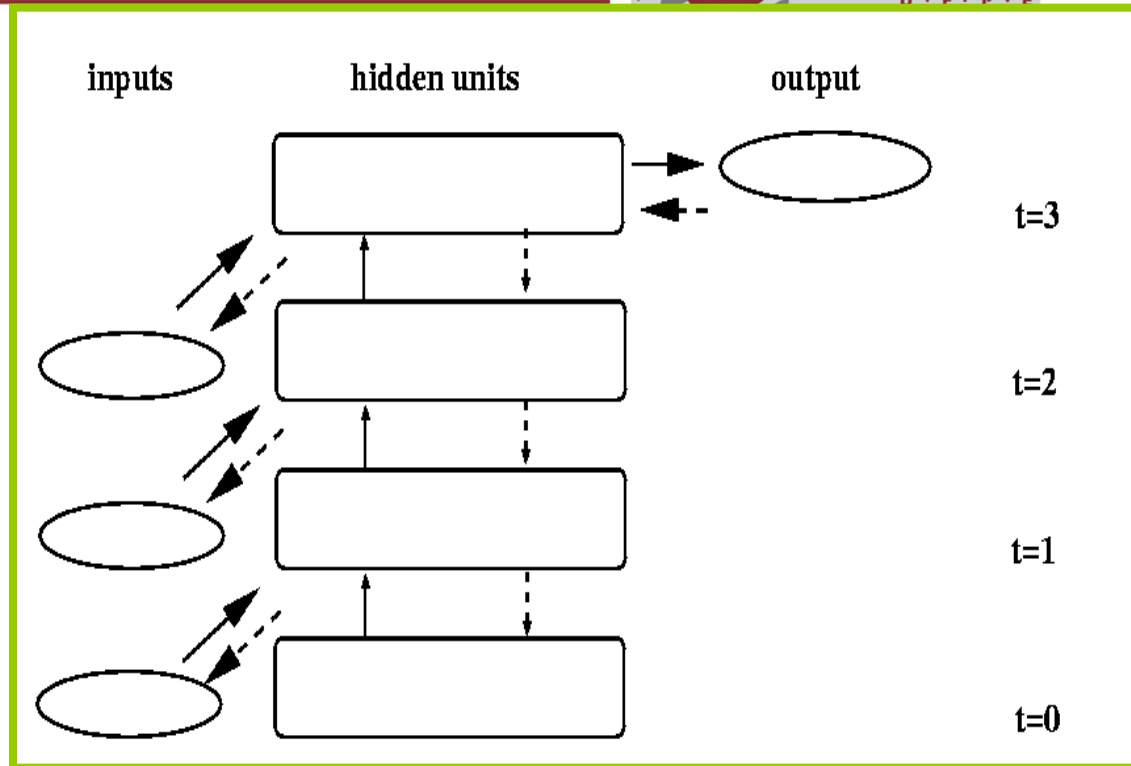
- ▶ Ativações das unidades escondidas e de entrada dos últimos 3 passos são armazenadas
- ▶ Linhas sólidas mostram que as ativações das unidades são definidas das entradas e camada escondida no passo anterior



- ▶ Passo para trás é executado para determinar valores separados de delta (erro) para cada unidade e passo de tempo separadamente

Backpropagation Through Time

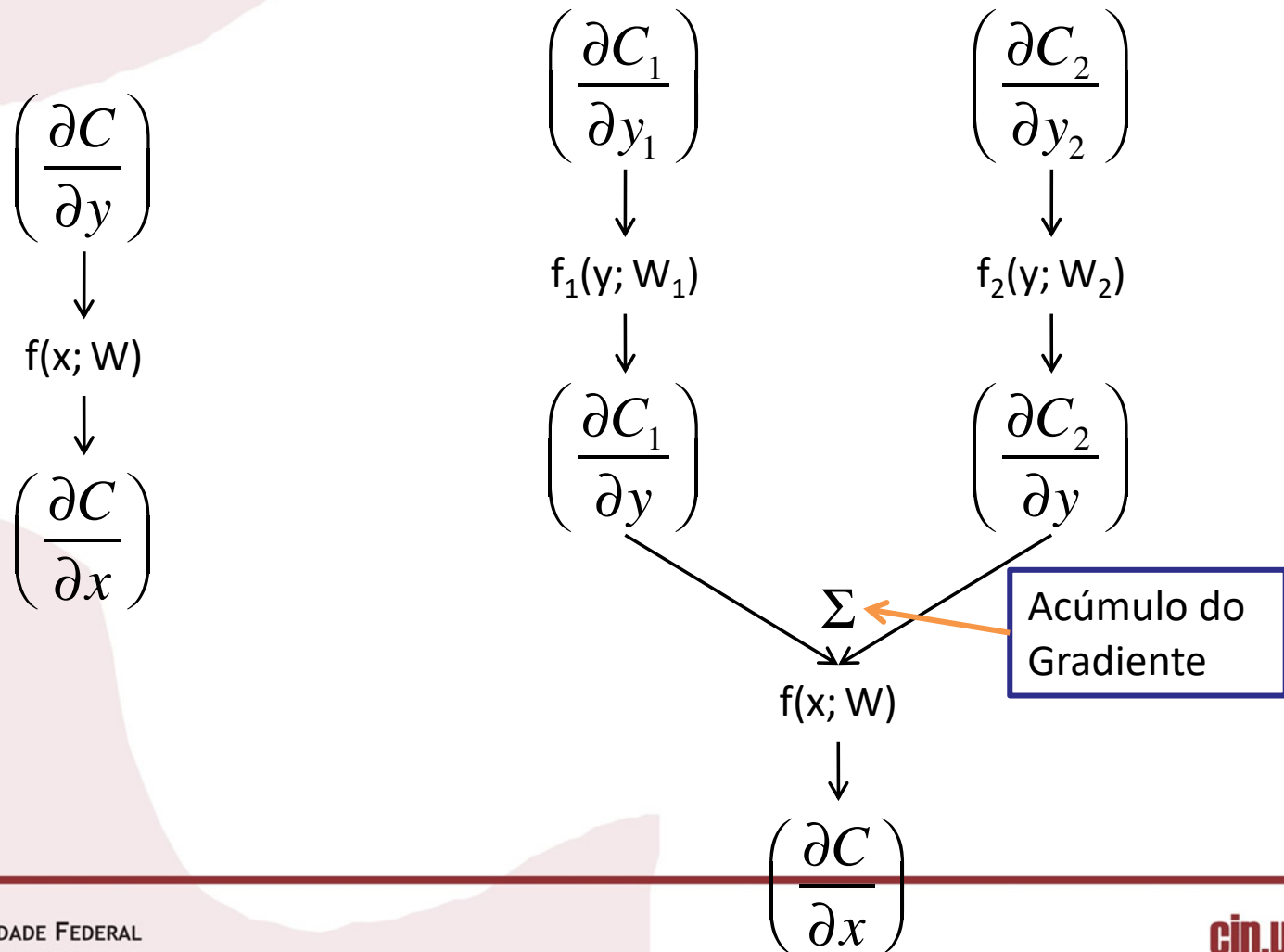
- Como cada camada anterior é uma cópia da camada um nível acima, introduzimos uma nova restrição: pesos em cada nível são idênticos
- Estamos “espalhando” o tempo em múltiplas camadas



A derivada parcial do erro com respeito a $w_{i,j}$ é a soma das derivadas parciais calculada para a cópia de $w_{i,j}$ entre cada 2 camadas

$$E_{total}(t_0, t_n) = \frac{1}{2} \sum_{t=t_0}^{t_n} \sum_{j \in A} e_j^2(t)$$

Backpropagation Through Time



Backpropagation Through Time



- A rede “desdobrada (no passo forward) é tratada como uma “grande” rede feedforward
- Recebe a série temporal inteira como entrada (ou pelo menos do tamanho da janela considerada)
- Atualização dos pesos é computada para cada cópia na rede desdobrada e somada (ou tirada a média)
- E aplicada em todos os pesos da RNN (pesos iguais)

- Rede de Elman e BPTT (generalização) buscam aproximar a computação do gradiente baseada nas entradas passadas
 - Mantendo algoritmo BP padrão
 - BPTT foi aplicado com sucesso em um número de aplicações (exemplo: modelagem de sinais de eletrocardiogramas – ECG)
 - Tarefa principal, nesse caso, é produzir uma resposta específica para uma sequência de entradas

- Principal desvantagem é que requer uma quantidade alta de recursos
 - Armazenamento – da história inteira
 - Processamento – cálculo do gradiente para todas as camadas
 - Dados – muitos exemplos para melhores representações

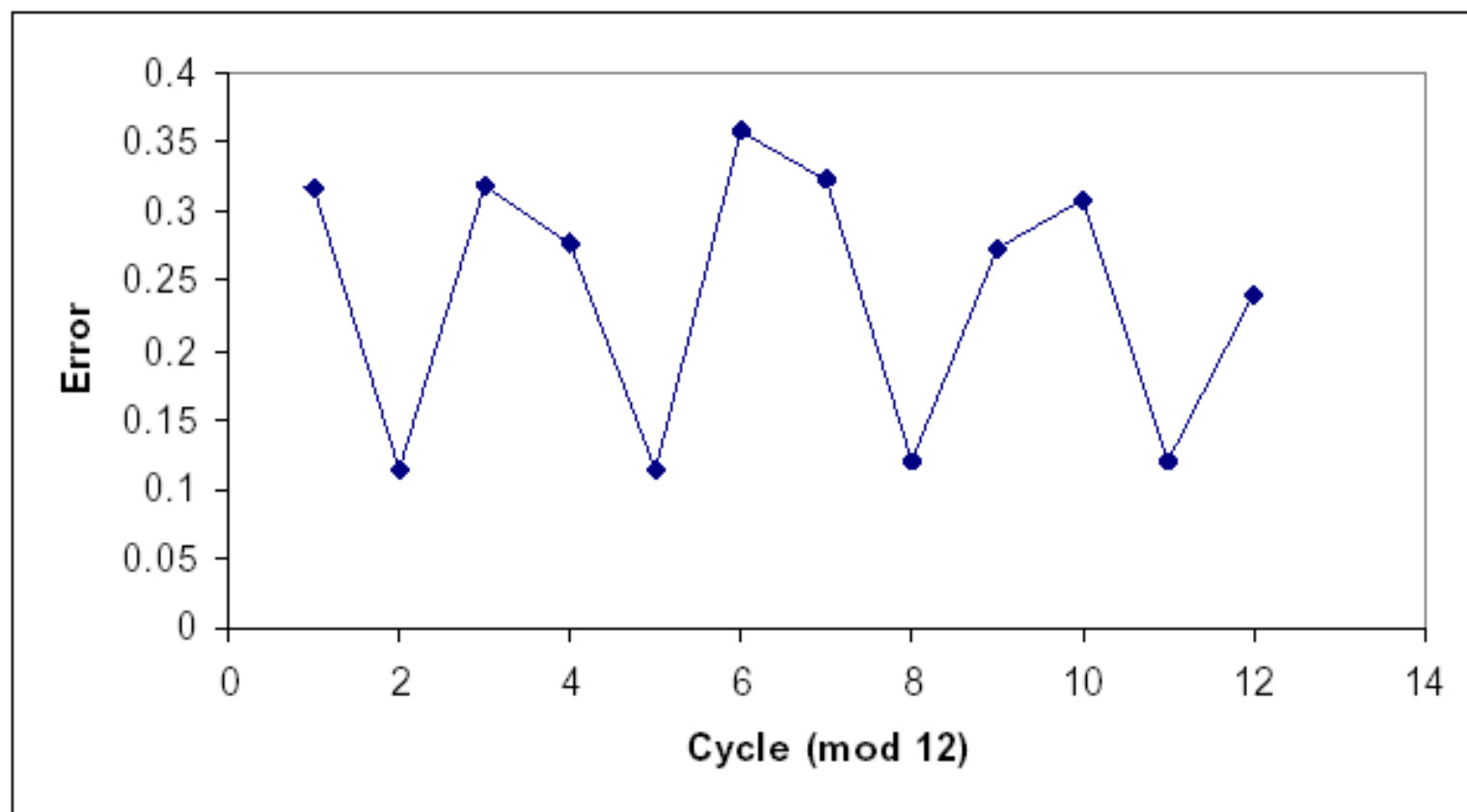
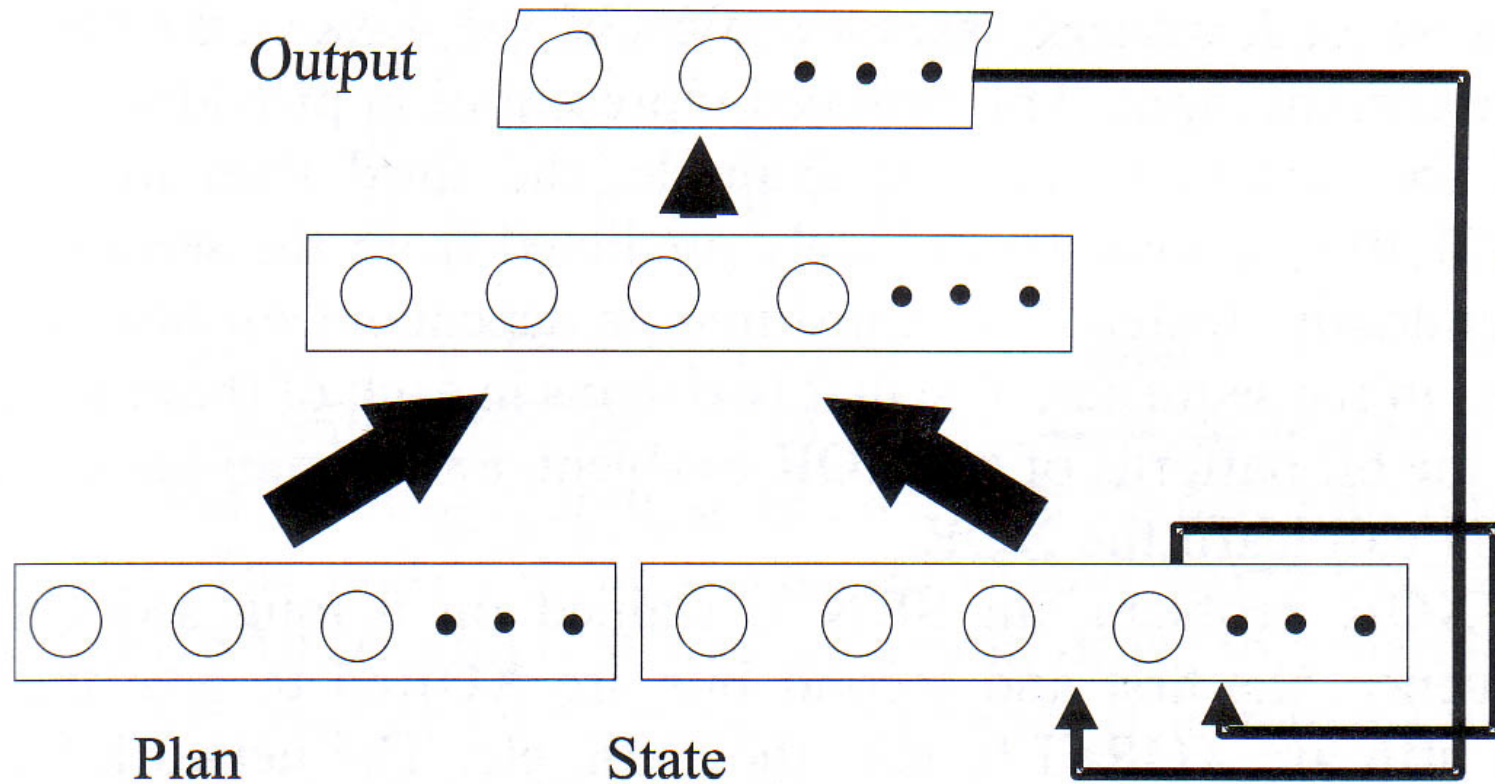


Figure 3. Graph of root mean squared error over 12 consecutive inputs in sequential XOR task. Data points are averaged over 1200 trials.

Rede de Jordan

Feedback das Unidades de Saída



Rede de Jordan

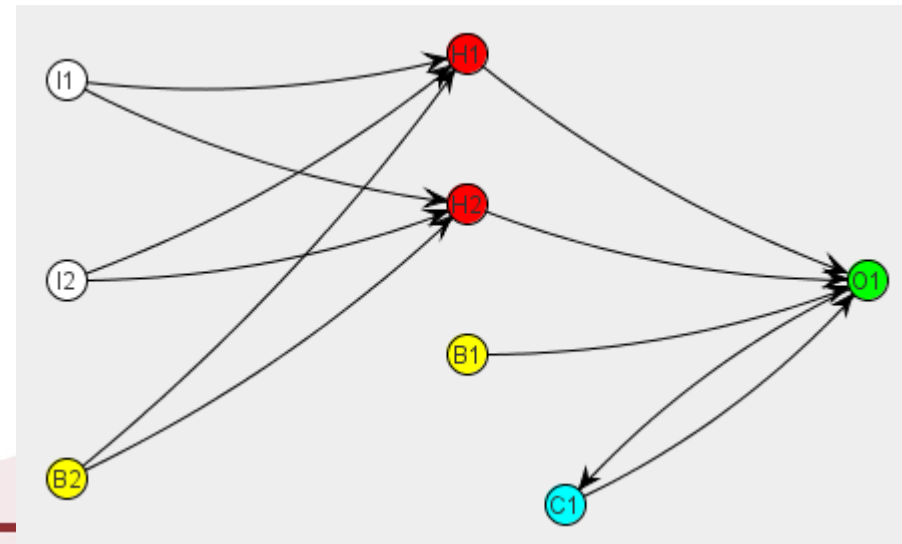
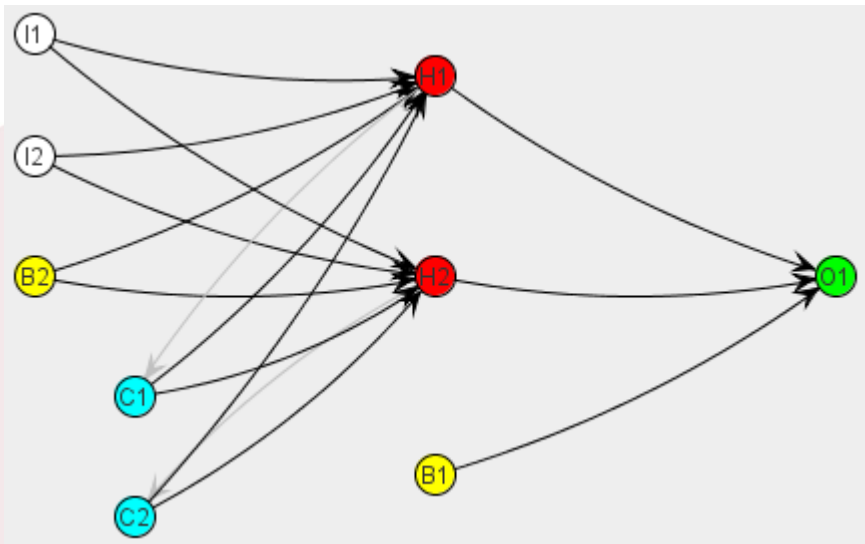


- Conexões da camada de saída para a camada de entrada e conexões das entradas para elas mesmas
- Aplicada em tarefas dependentes de sequências de estados sucessivos (padrões temporais)
- Pode ser treinada com backpropagation
- Rede possui uma forma de guardar memórias curtas (short-term memory)
- Semelhante à Rede de Elman, para treinamento com backpropagation



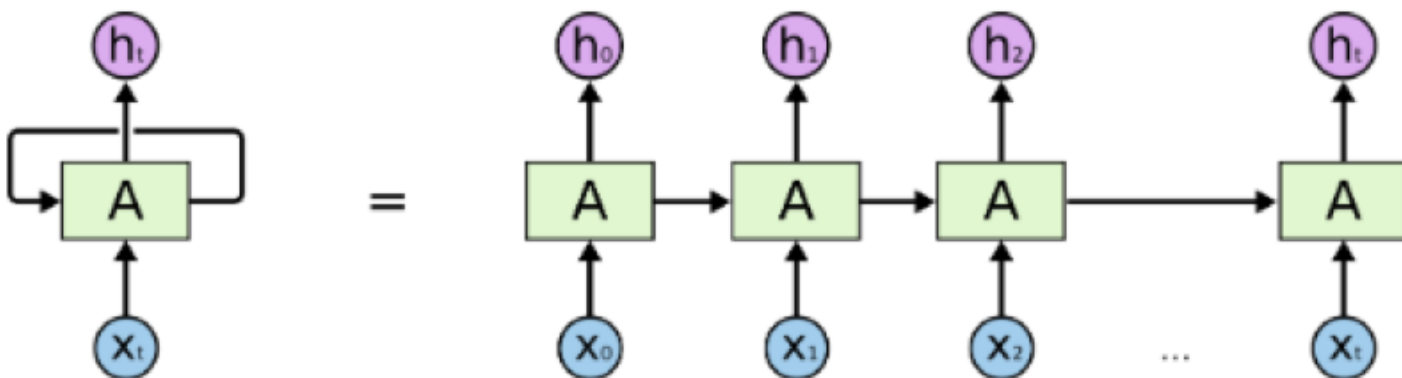
Diferenças

- Elman – guardam memória das unidades escondidas
- Jordan – guardam memória das unidades de saída
- Podem produzir desempenhos diferentes em aplicações diferentes



Long Short Term Memory Networks (LSTM Networks)

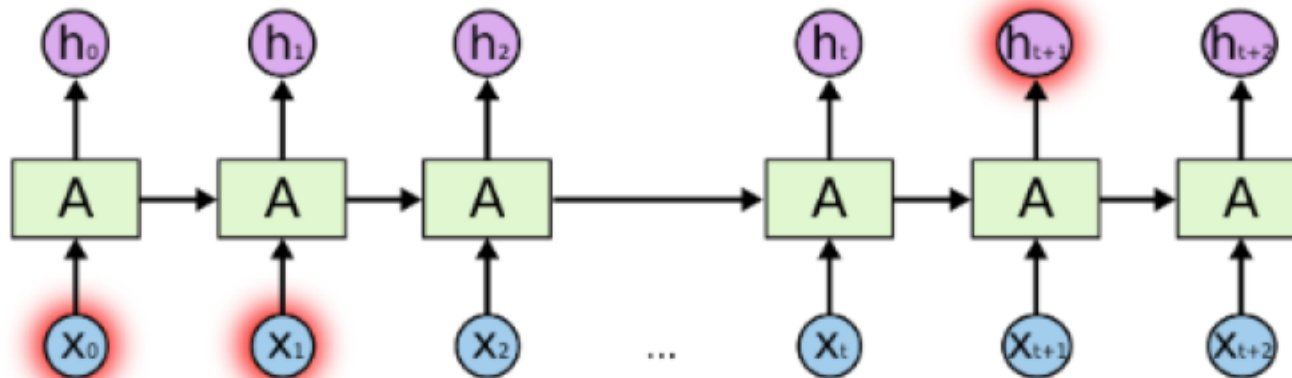
Rede Recorrente (RNN) Clássica



An unrolled recurrent neural network.

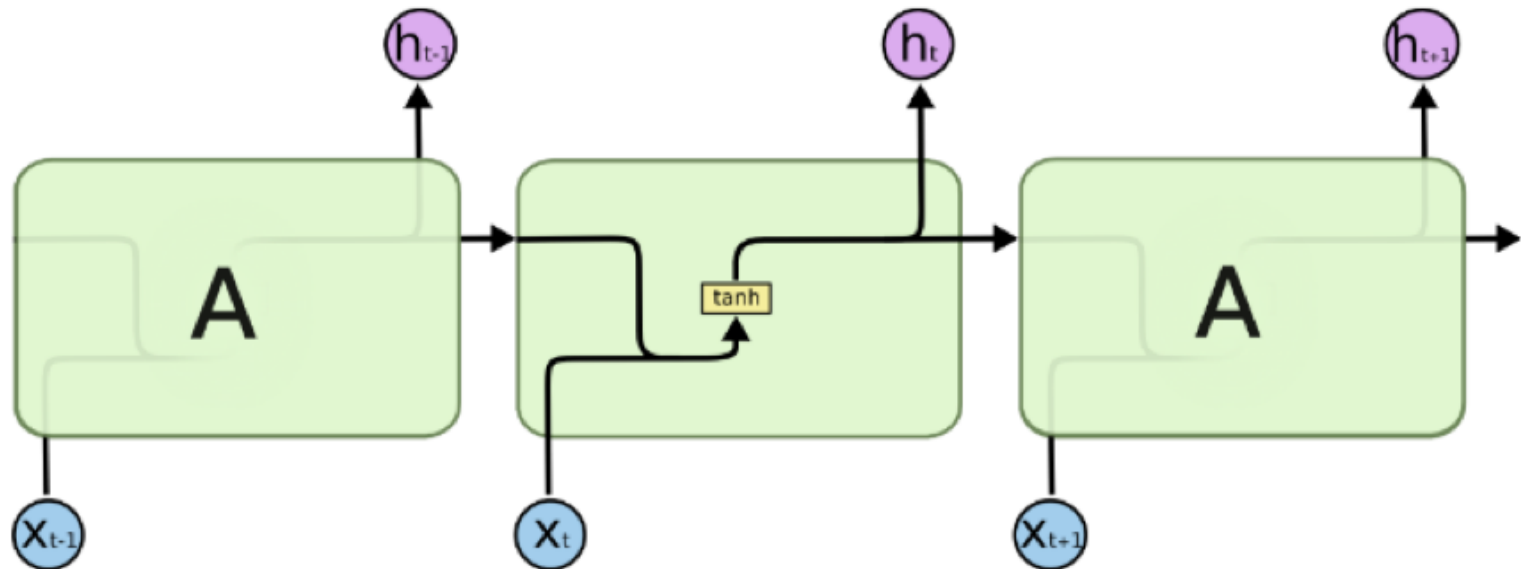
Problemas com Redes Recorrentes (RNN) Clássicas

Dependências temporais longas!!



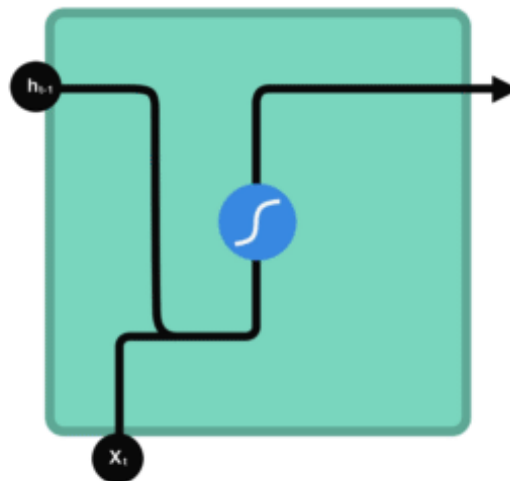
Na teoria, OK!! Na prática, “*the vanishing gradient problem...*”

Na RNN Clássica



Uma rede neural simples em cada módulo

Função da Unidade



Tanh function



new hidden state



previous hidden state

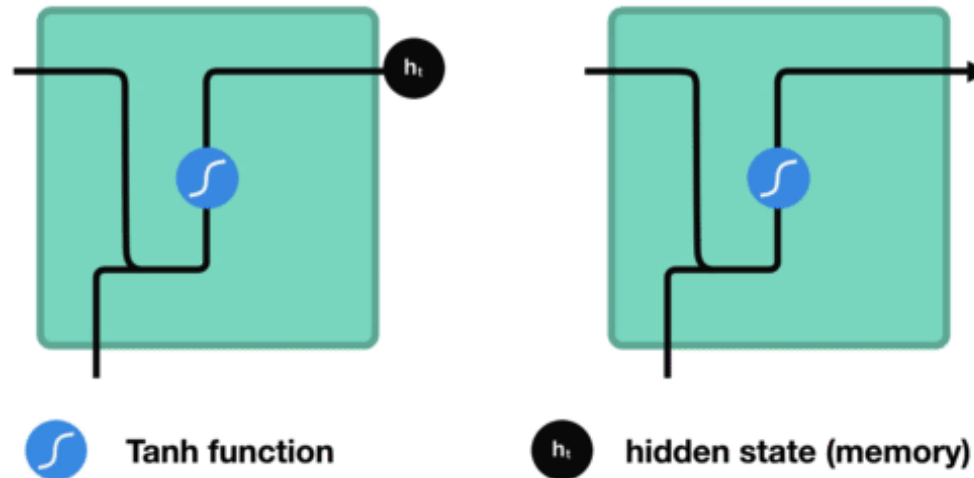


input



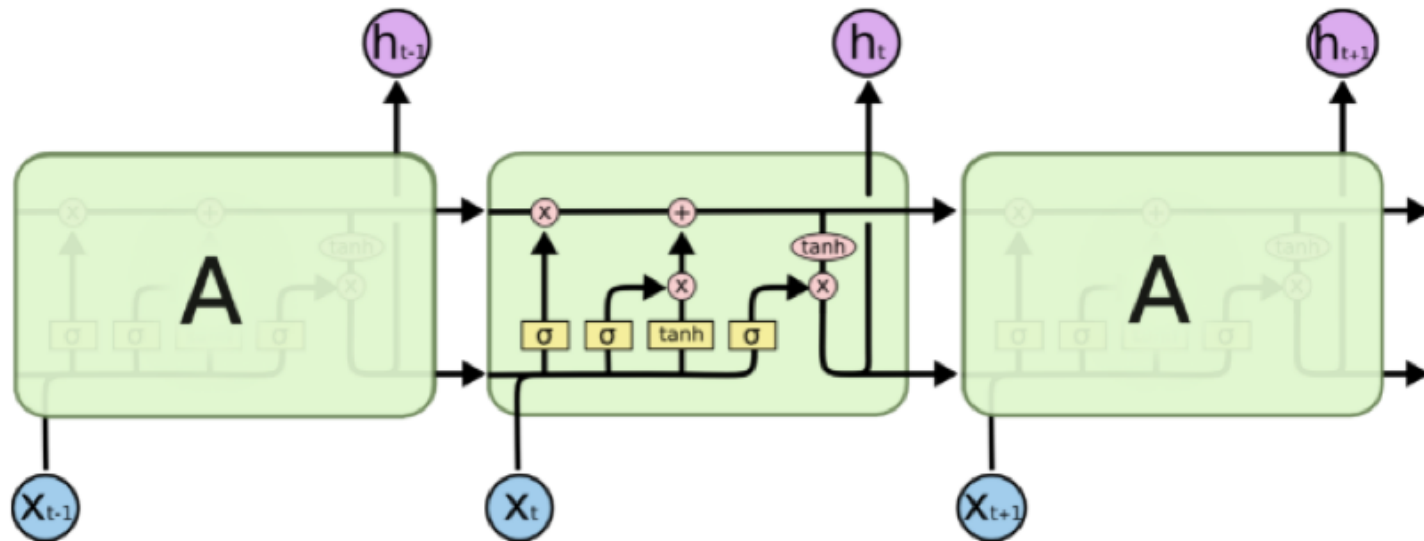
concatenation

Recorrência - Feedback

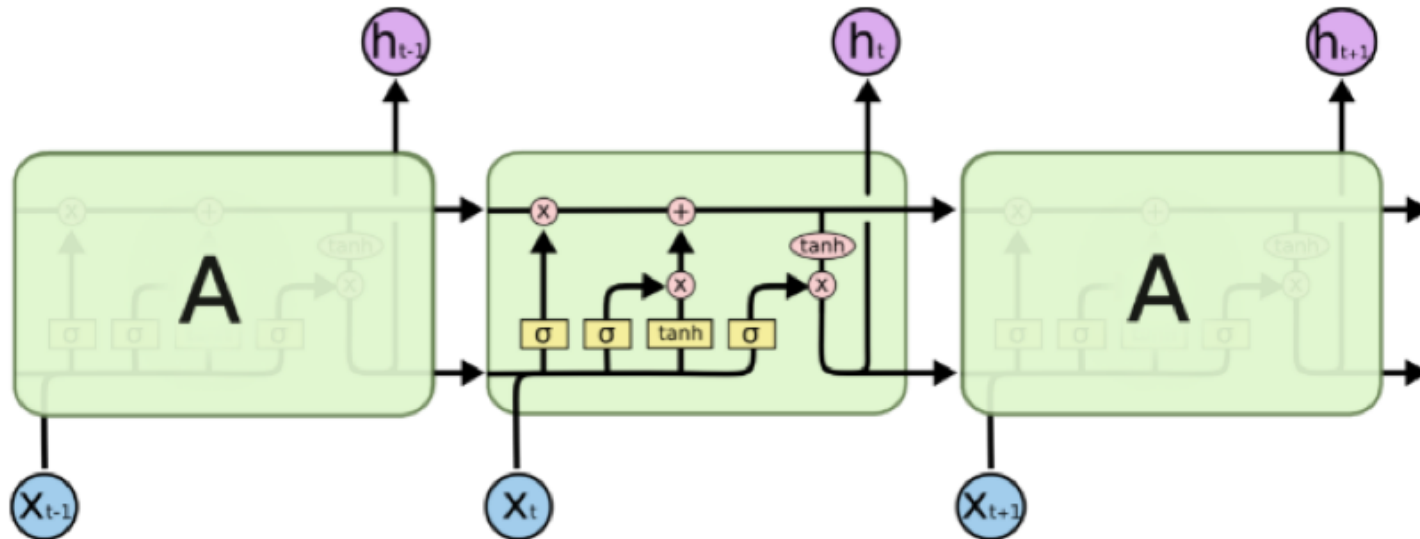


Na LSTM

Estrutura com 4 redes neurais simples em cada módulo

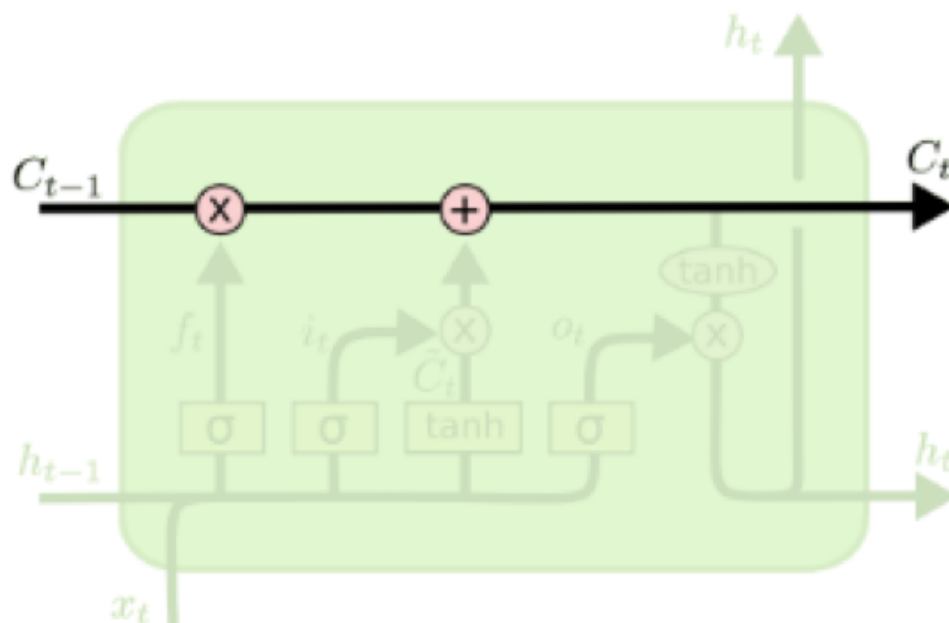


Na LSTM



Conceito de célula de memória que carrega a informação

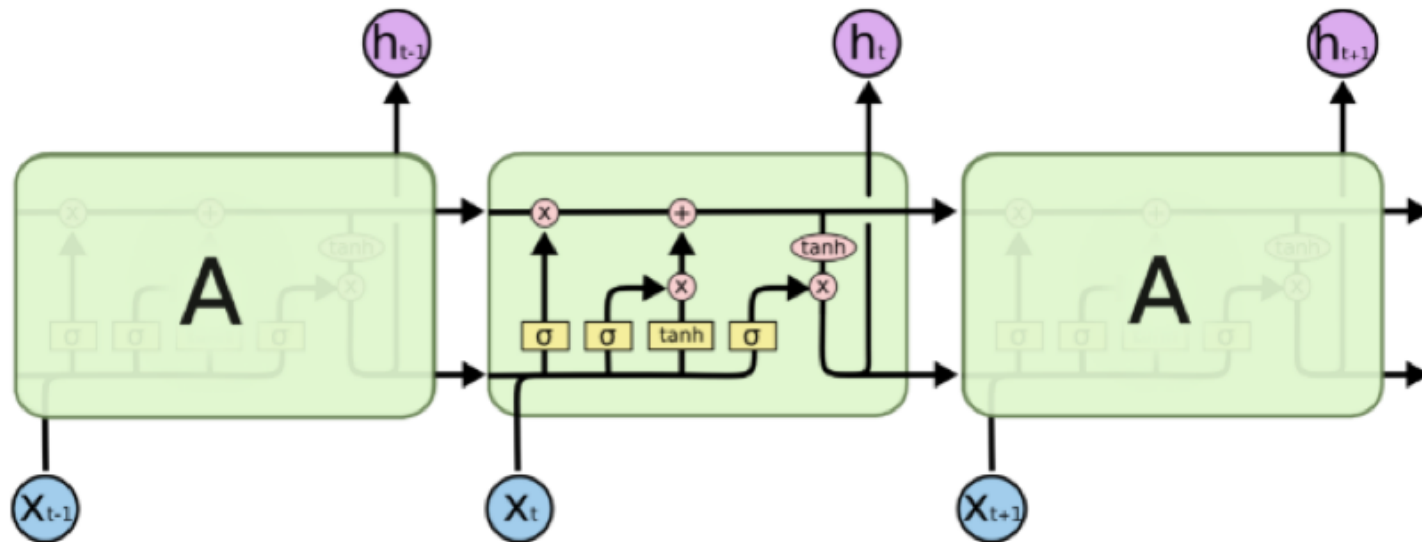
Essa é Ideia Central da LSTM



Conceito de célula de memória (cell state) que carrega a informação

Na LSTM

Esteira transportadora (*conveyor belt*)



Memória é esquecida, alterada ou mantida através de filtros!

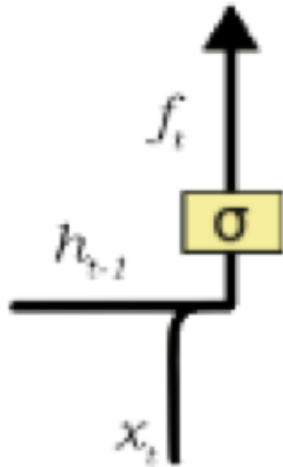
**Centro
de Informática**
U.F.P.E.



UNIVERSIDADE FEDERAL
DE PERNAMBUCO

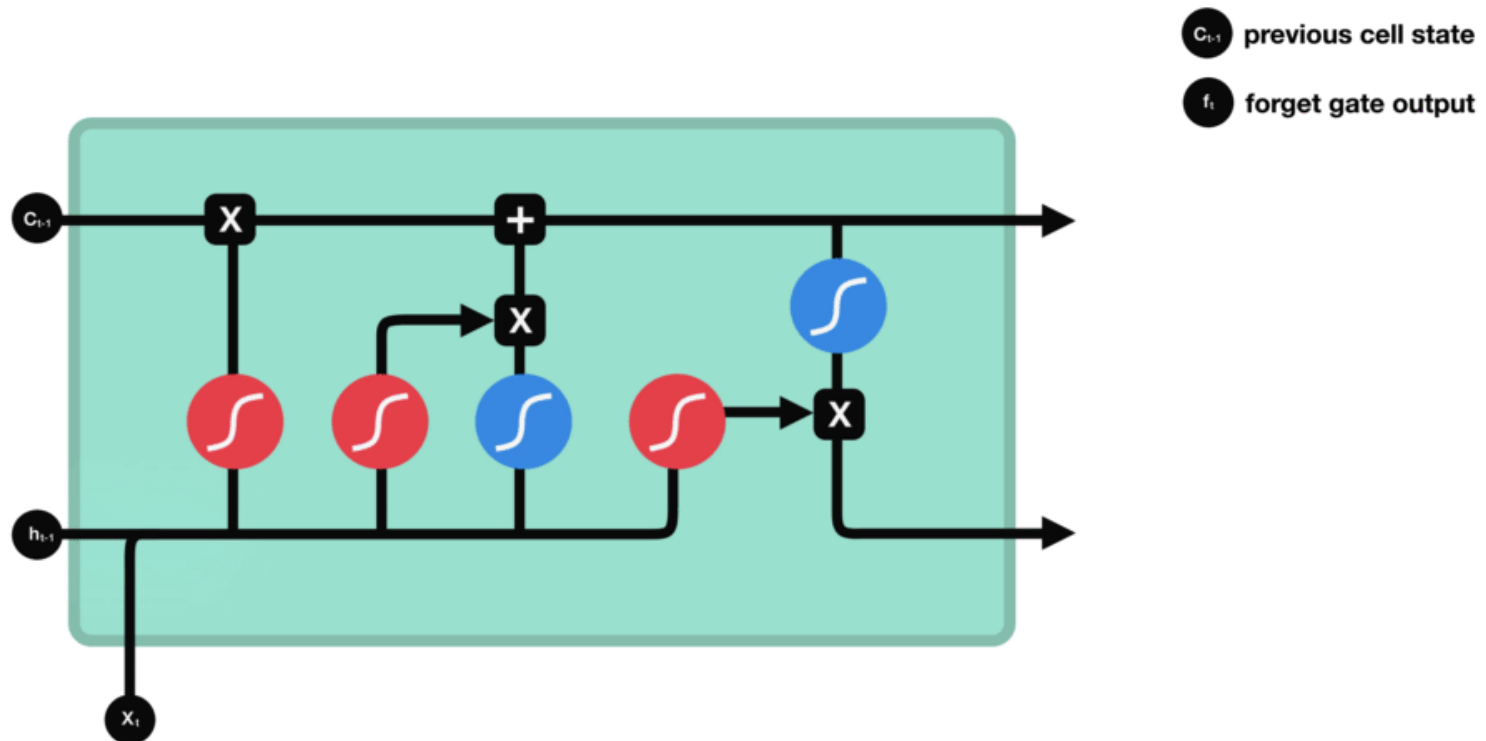
The Forget Gate (Filtro do Esquecimento)

Bob is a nice person. Dan on the other hand is evil.



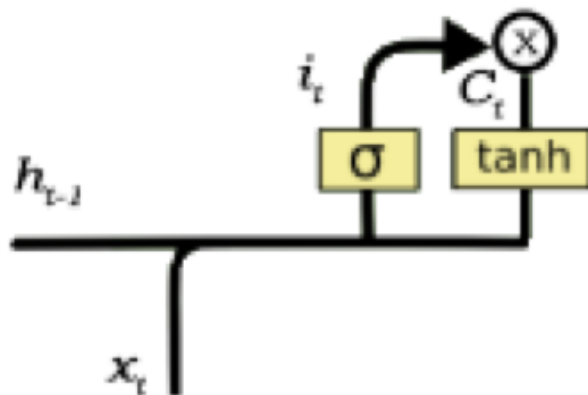
$$f_t = \sigma (W_f \cdot [h_{t-1}, x_t] + b_f)$$

The Forget Gate (Filtro do Esquecimento)



The Input Gate (Filtro da Entrada)

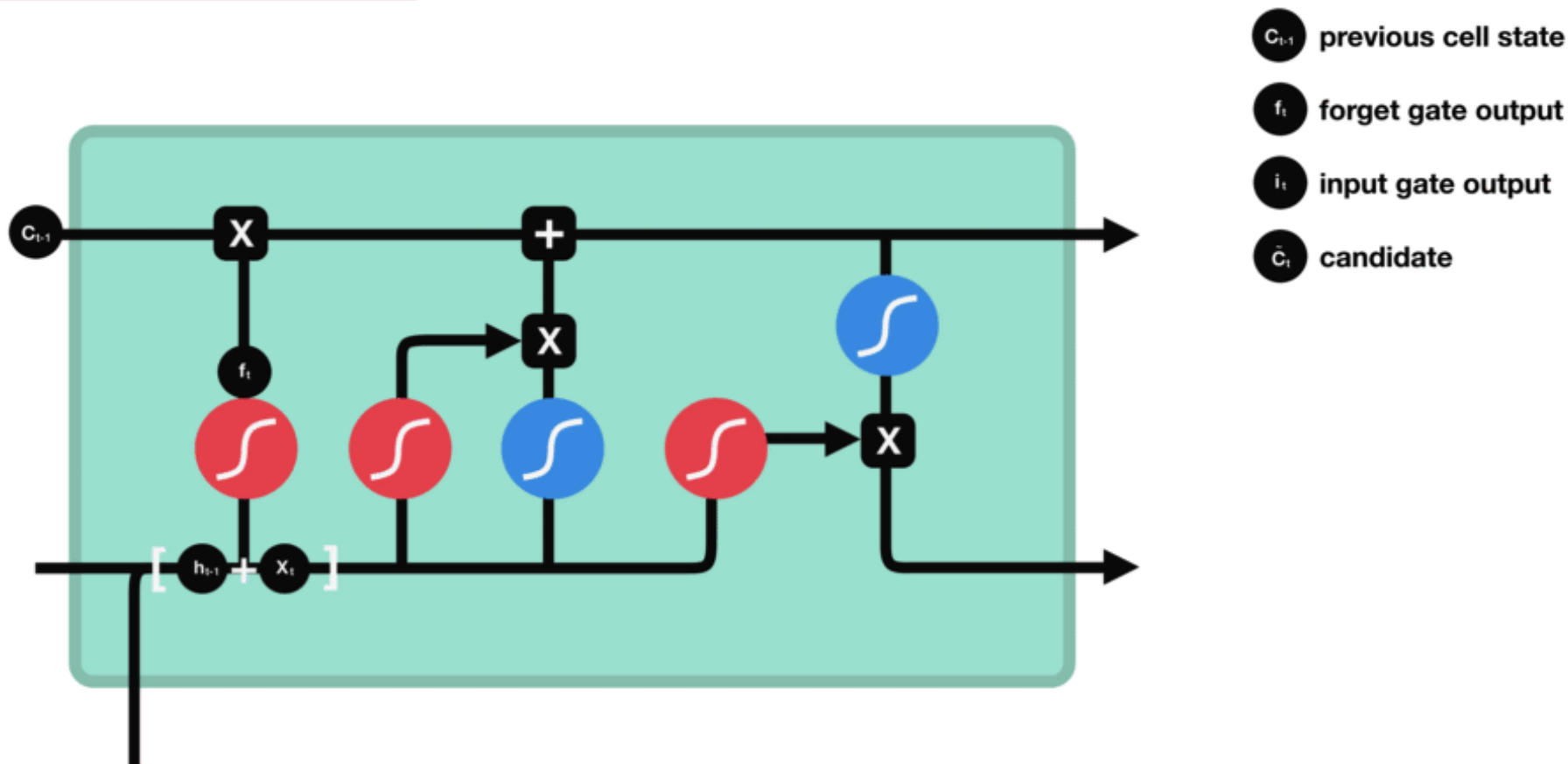
Bob knows swimming. He told me over the phone that he had served the navy for 4 long years.



$$i_t = \sigma (W_i \cdot [h_{t-1}, x_t] + b_i)$$

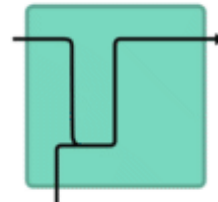
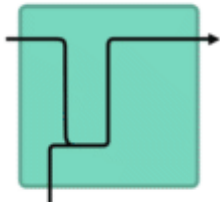
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

The Input Gate (Filtro da Entrada)

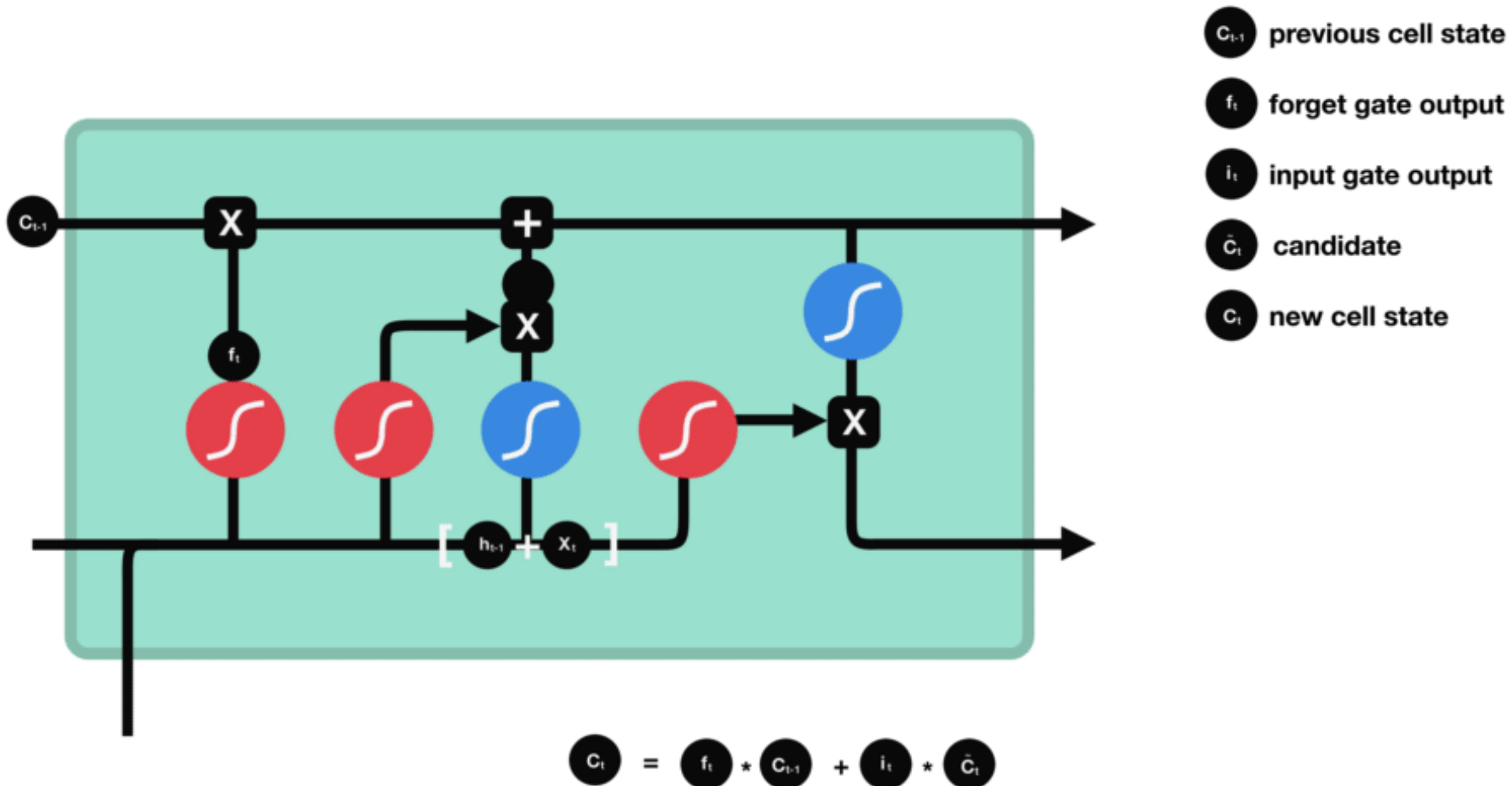


Papel da Tahn – Regularizar os Valores

5
0.01
-0.5

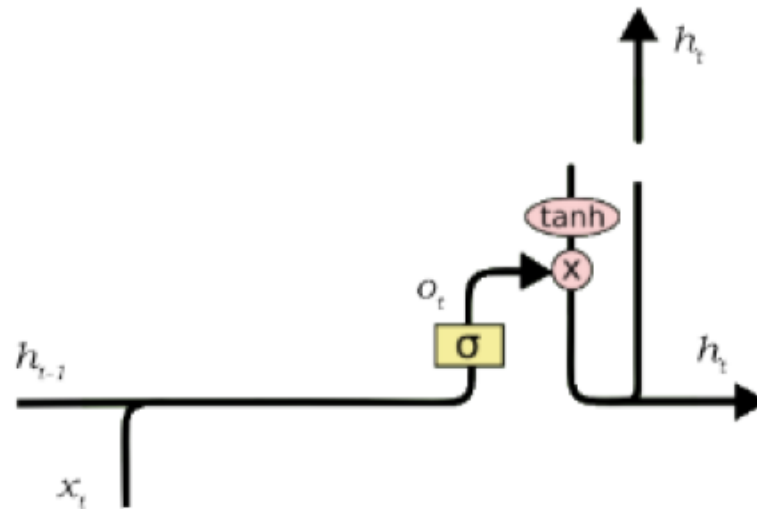


Atualização da Memória (Cell State)

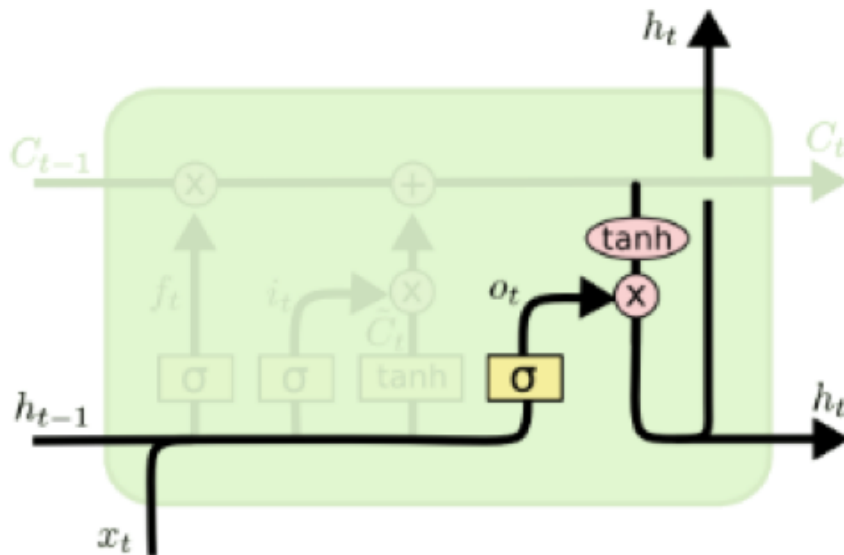


The Output Gate (Filtro da Saída)

Bob fought single handedly with the enemy and died for his country. For his contributions brave ____.



Saida do Módulo

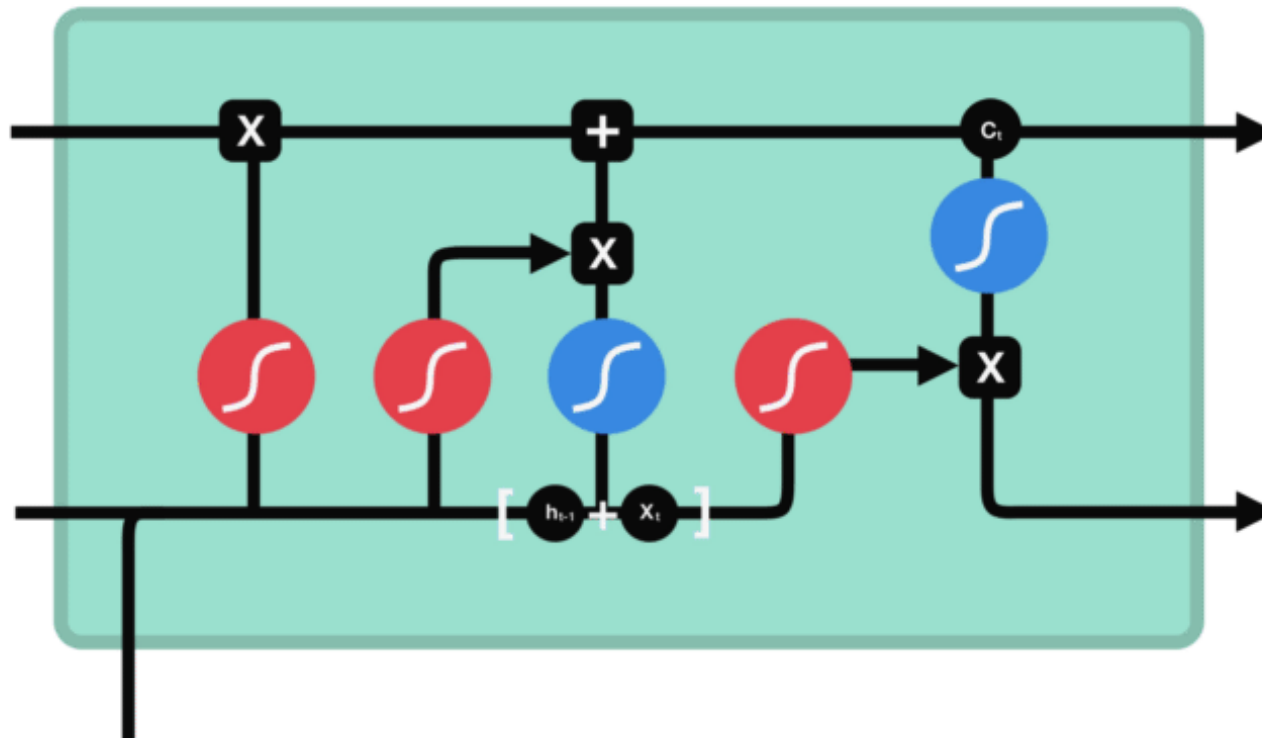


$$o_t = \sigma(W_o [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh(C_t)$$

<https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21>

Output Gate (Saída do Módulo)



- c_{t-1} previous cell state
- f_t forget gate output
- i_t input gate output
- $\tilde{c}_t$ candidate
- c_t new cell state
- o_t output gate output
- h_t hidden state

Como Codificar os Dados?

- Opção mais simples: one-hot encoding

Hello! Welcome to Educative. Happy learning.

One Hot Encoding

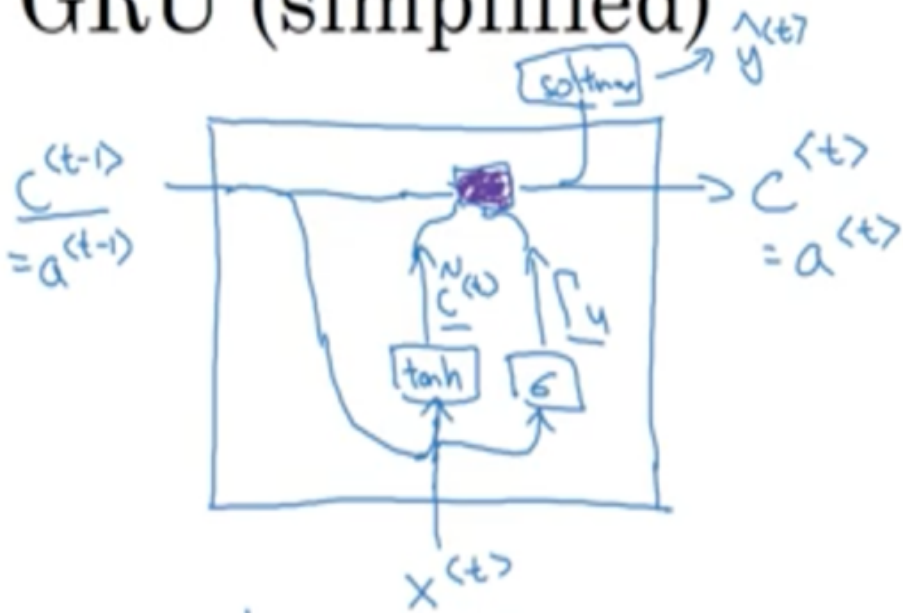
Hello	1	0	0	0	0	0
to	0	1	0	0	0	0
Educative	0	0	1	0	0	0
learning	0	0	0	1	0	0
Happy	0	0	0	0	1	0
Welcome	0	0	0	0	0	1

Centro
de Informática
U · F · P · E



-

GRU (simplified)



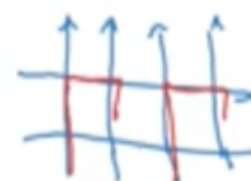
C = memory cell

$$\rightarrow \boxed{C^{(t)}} = \underline{a}^{(t)}$$

$$\rightarrow \boxed{\tilde{C}^{(t)}} = \tanh(W_c [C^{(t-1)}, x^{(t)}] + b_c)$$

$$\rightarrow \boxed{\Gamma_u} = \sigma(W_u [C^{(t-1)}, x^{(t)}] + b_u)$$

$$\boxed{C^{(t)}} = \underbrace{\Gamma_u}_{\text{"update"}=1} * \tilde{C}^{(t)} + (1 - \Gamma_u) * \boxed{C^{(t-1)}}$$



element-wise
Gate

$$\Gamma_u = 0.000001$$

$\Gamma_u = 1$ $\Gamma_u = 0$ $\Gamma_u = 0$ $\Gamma_u = 0$...
 $C^{(t)} = 1$
 The cat, which already ate ..., was full.

[Cho et al., 2014. On the properties of neural machine translation: Encoder-decoder approaches]

[Chung et al., 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling]

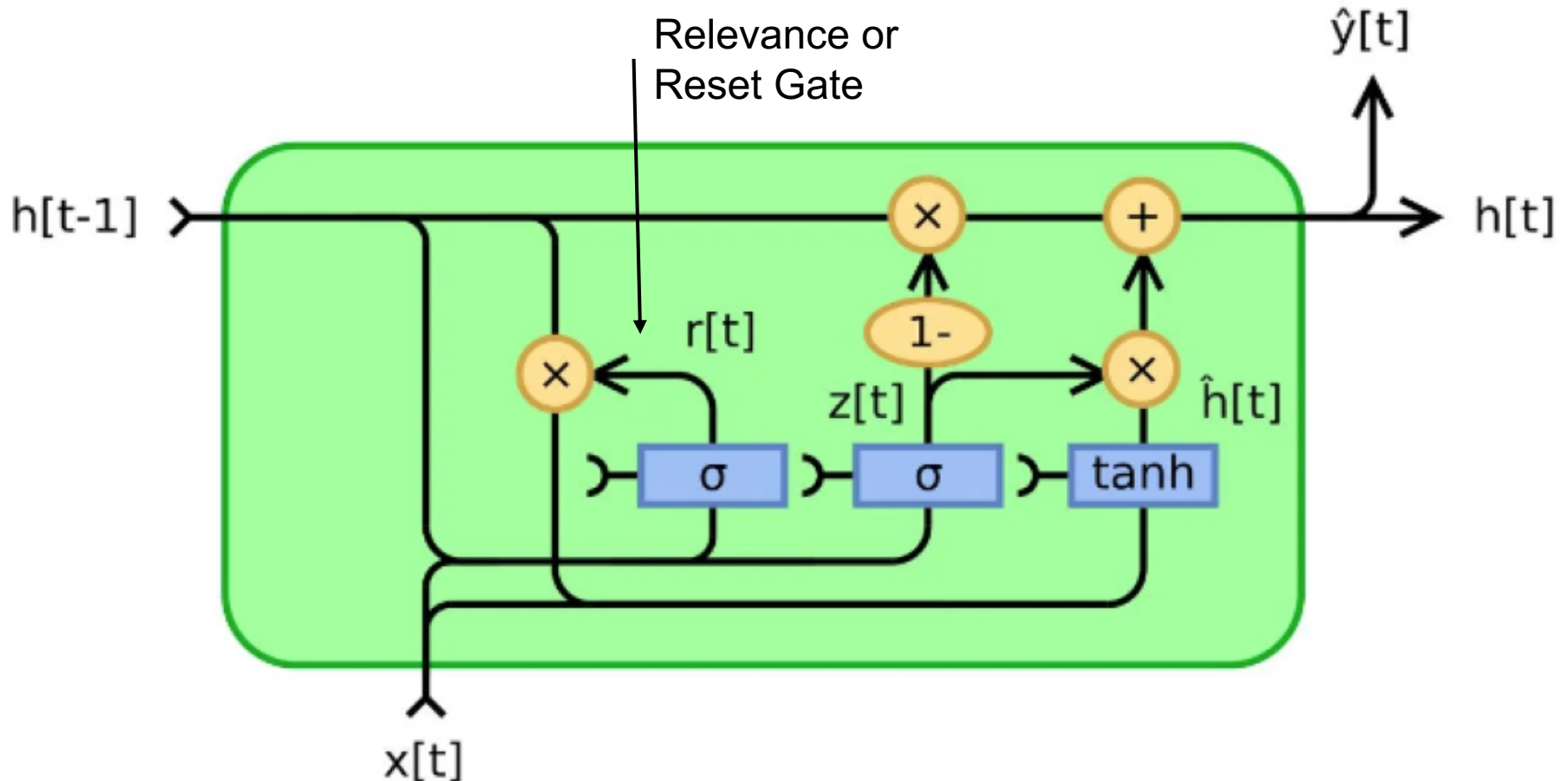
Andrew Ng



UNIVERSIDADE FEDERAL
DE PERNAMBUCO

cin.ufpe.br

GRU (Gated Recurrent Unit)



<https://em360tech.com/tech-article/gated-recurrent-unit-gru>

Full GRU

$$\tilde{c}^{<t>} = \tanh(W_c[\Gamma_r * c^{<t-1>}, x^{<t>}] + b_c)$$

$$\Gamma_u = \sigma(W_u[c^{<t-1>}, x^{<t>}] + b_u)$$

$$\Gamma_r = \sigma(W_r[c^{<t-1>}, x^{<t>}] + b_r)$$

$$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + (1 - \Gamma_u) * c^{<t-1>}$$

The cat, which ate already, was full.

Links de Materiais



<https://scikiq.com/blog/how-generative-adversarial-network-gan-is-transforming-data-analytics/> (exemplos de uso e mal uso)

<http://colah.github.io/posts/2015-08-Understanding-LSTMs/>

<https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21>

<https://www.educative.io/answers/one-hot-encoding-of-text-data-in-natural-language-processing>

<https://www.youtube.com/watch?v=pYRIOGTPRPU> – Andrew NG

<https://medium.com/turing-talks/word-embedding-fazendo-o-computador-entender-o-significado-das-palavras-92fe22745057>

65

<https://nzlul.medium.com/the-classification-of-text-messages-using-lstm-bi-lstm-and-gru-f79b207f90ad>

https://wandb.ai/wandb_fc/wb-tutorials/reports/Tutorial-Text-Generation-with-LSTMs-and-GRUs--Vmlldzo0NTIxMjY1

https://slazebni.cs.illinois.edu/spring17/lec11_gan.pdf

