

INF721

2023/2



Aprendizado em Redes Neurais Profundas

A16: Redes Neurais Recorrentes

Logística

Avisos

- ▶ Entrega da PF: Proposta de Problema nesta quarta-feira (18/10)!

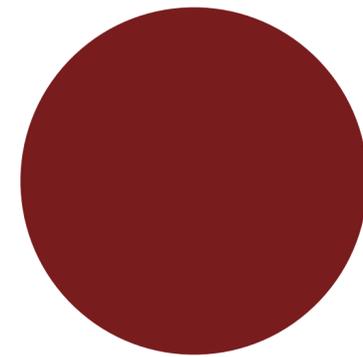
Última aula

- ▶ Estudo de casos de CNNs
- ▶ CNNs clássicas (LeNet-5, AlexNet, VGG-16)
- ▶ ResNet
- ▶ Inception Network

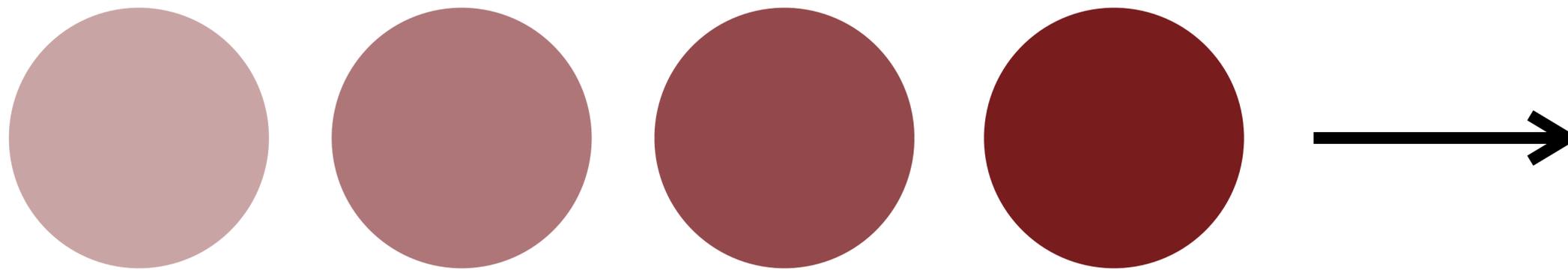
Plano de Aula

- ▶ Problemas sequenciais
- ▶ Modelos de linguagem
- ▶ Codificação vetorial de palavras e vocabulários
- ▶ Diagrama, formalização e treinamento de RNNs
- ▶ Geração de sequências

Para qual direção essa bola vai se mover?



Para qual direção essa bola vai se mover?



Que letra vem depois de T no alfabeto?



Que letra vem depois de T no alfabeto?

R S T _

Que letra vem depois de T no alfabeto?

R S T U

Que letra vem depois de T no alfabeto?

R S T U

Redes Neurais Recorrentes são utilizadas para classificação, regressão ou geração de **dados sequenciais!**

Problemas Sequenciais

Entrada

Saída

Reconhecimento de voz



"Olá, seja bem vindo."

Análise de sentimento

"O produto veio com defeito."

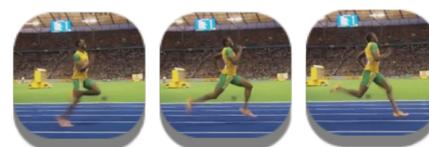


Tradução automática

"The book is on the table."

"O livro está em cima da mesa."

Reconhecimento de atividade em vídeos



Correndo

Geração de músicas

<vazio>



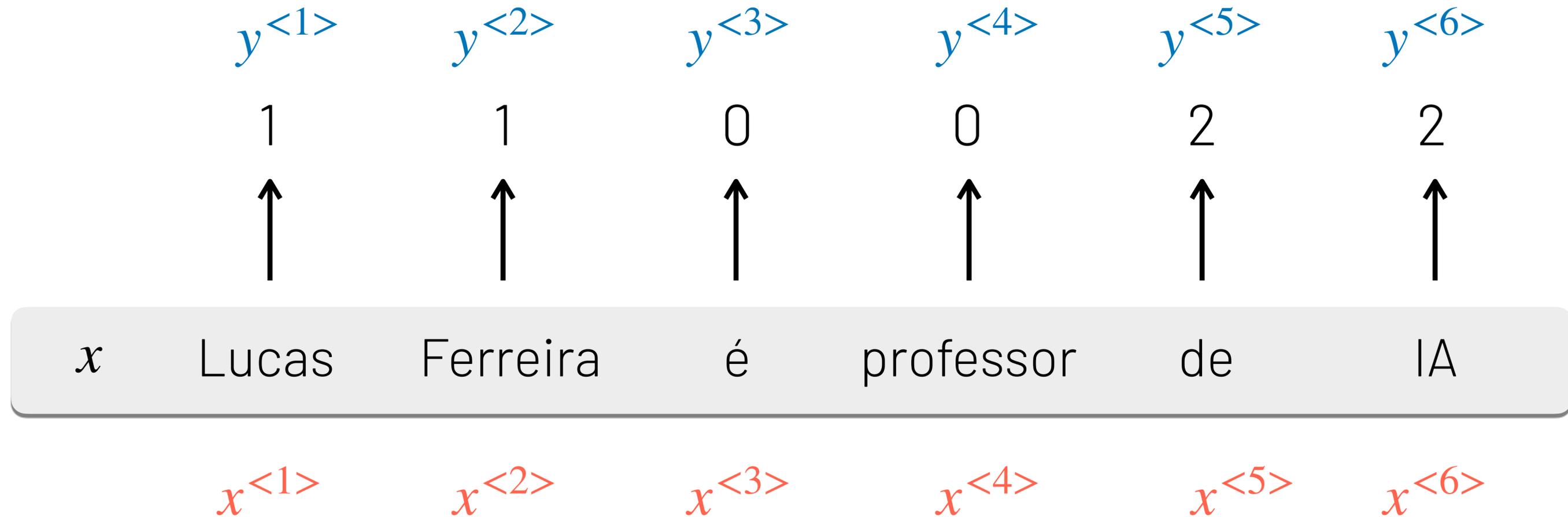
Reconhecimento de Entidade Nomeada

"Lucas Ferreira é professor de IA."

"Lucas Ferreira é professor de IA."

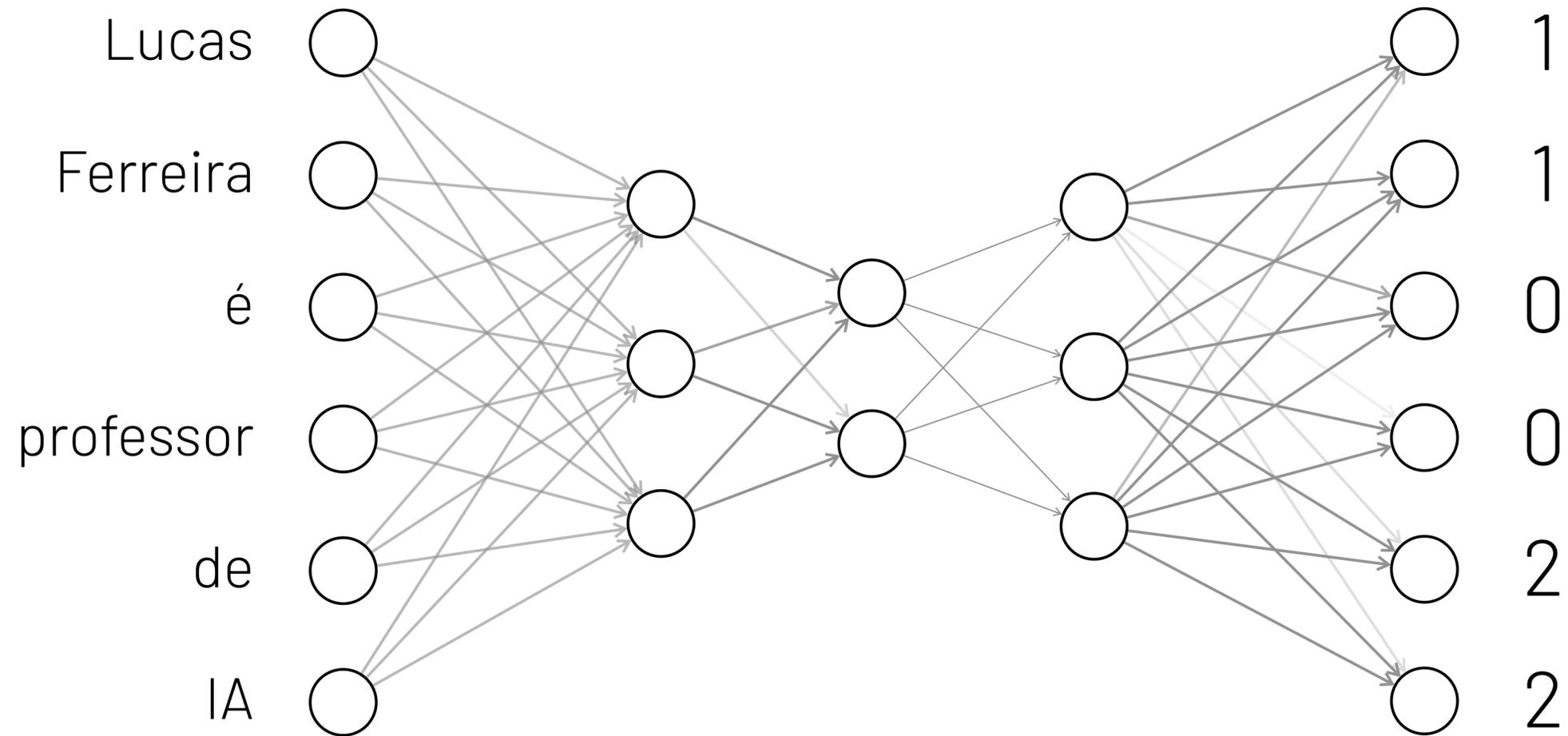
Exemplo de entrada e saída

Reconhecimento de entidade nomeada (REN): professor (1) e disciplina (2)

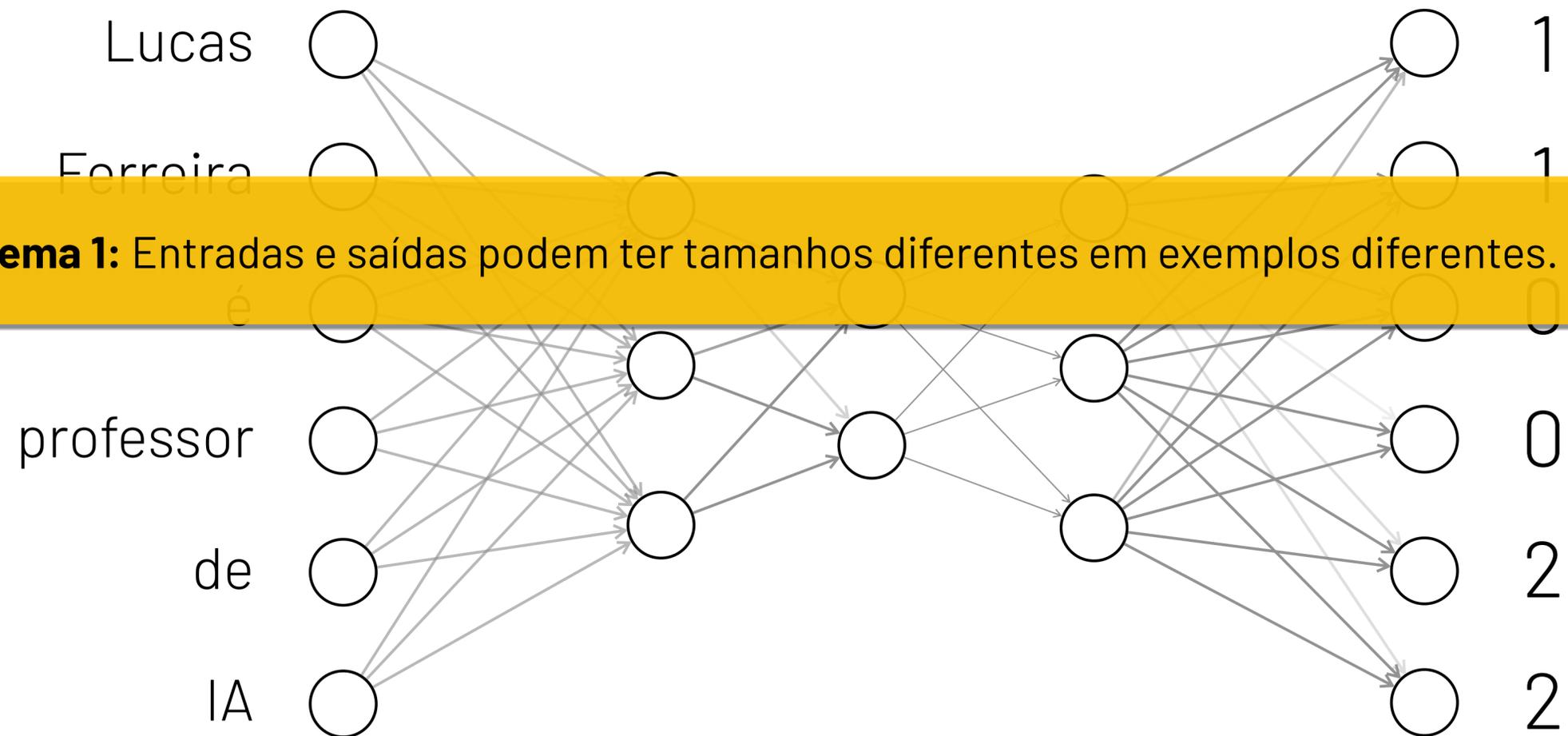


Em problemas sequencias, cada elemento da entrada $x^{<t>}$ pode ter uma saída associada $y^{<t>}$

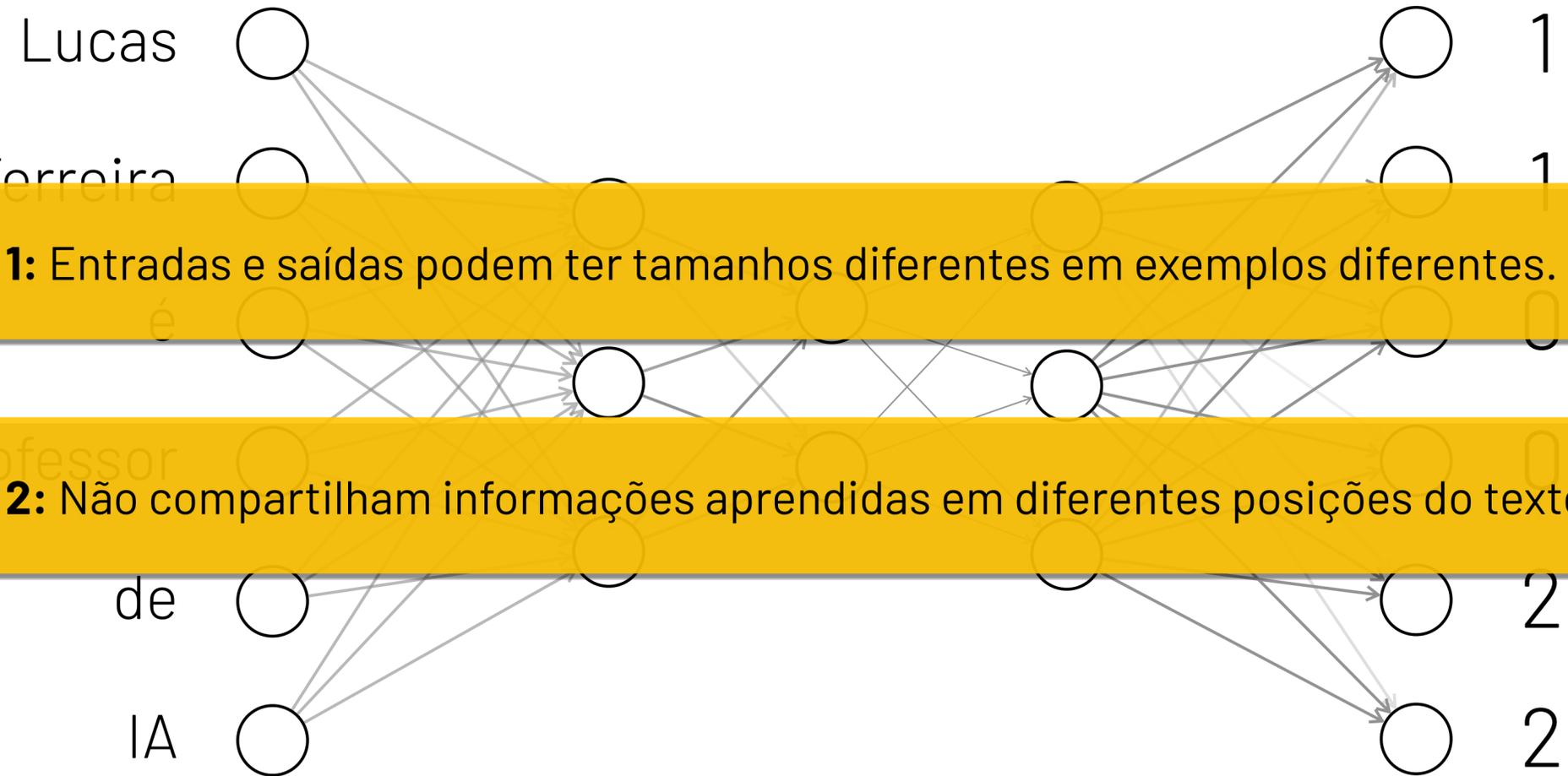
Porque não MLPs para processamento de sequências?



Porque não MLPs para processamento de sequências?



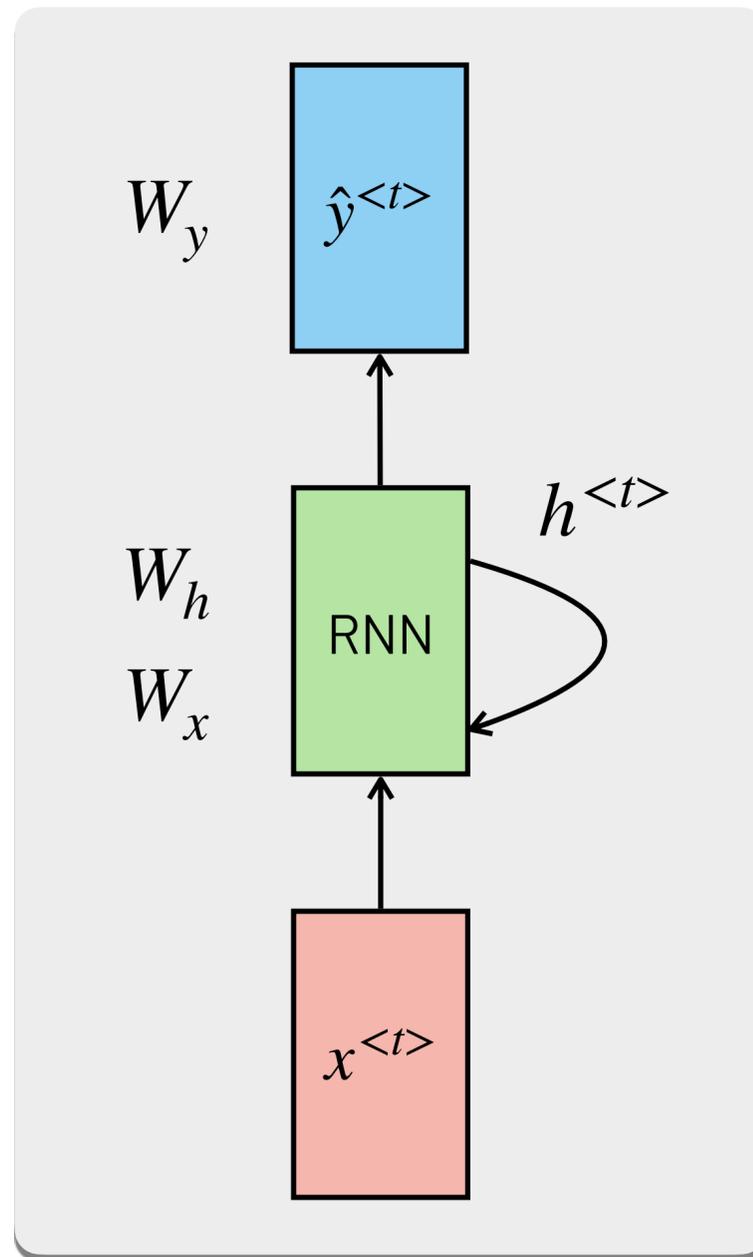
Porque não MLPs para processamento de sequências?



Problema 1: Entradas e saídas podem ter tamanhos diferentes em exemplos diferentes.

Problema 2: Não compartilham informações aprendidas em diferentes posições do texto.

Rede Neural Recorrente (RNN)



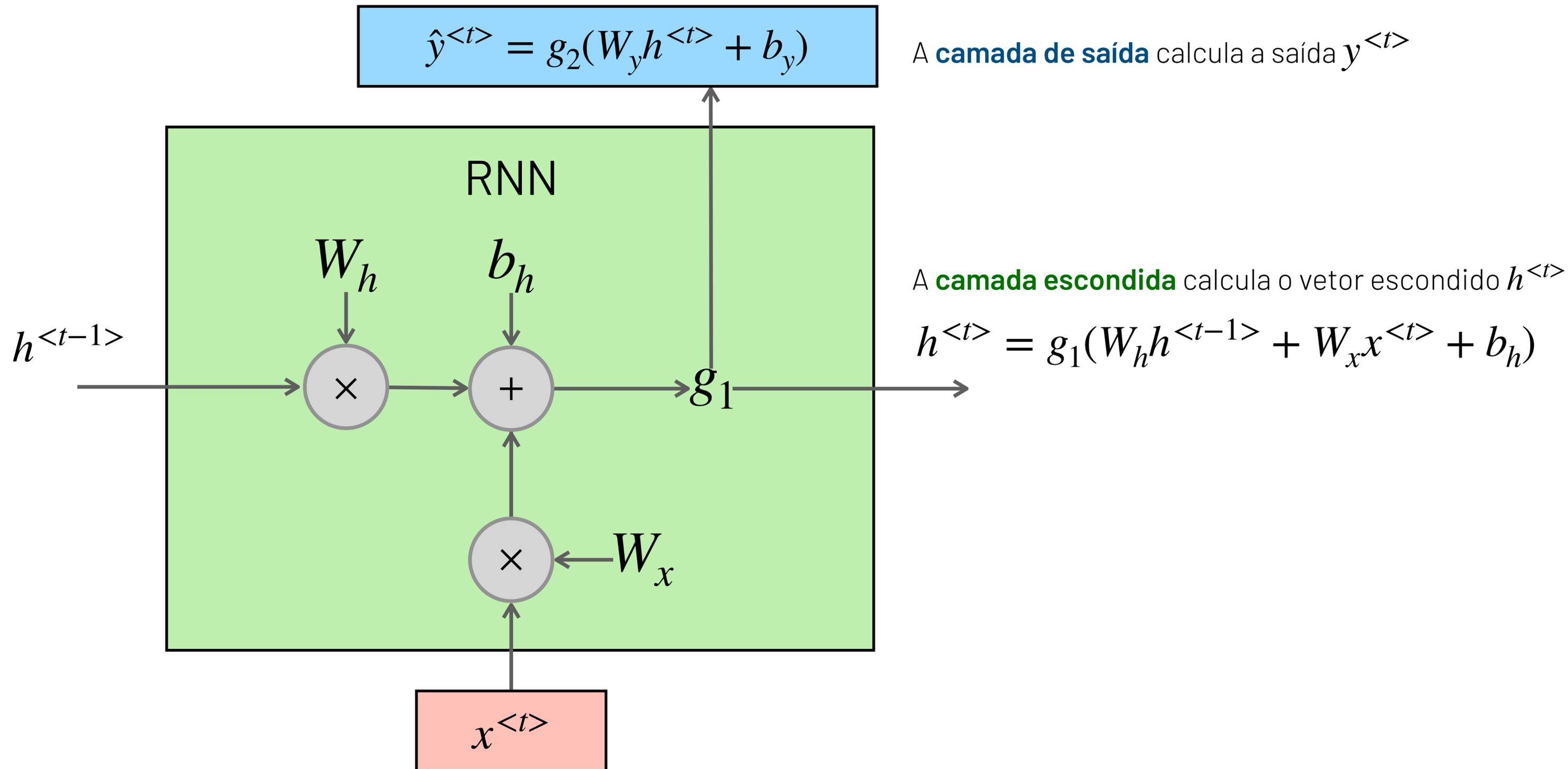
A RNN processa cada elemento da **entrada** $x^{<t>}$ de uma vez, mantendo um estado (vetor) $h^{<t>}$ que é atualizado a cada intervalo de tempo para gerar uma **saída** $y^{<t>}$

$$h^{<t>} = g_1(W_h h^{<t-1>} + W_x x^{<t>} + b_h)$$

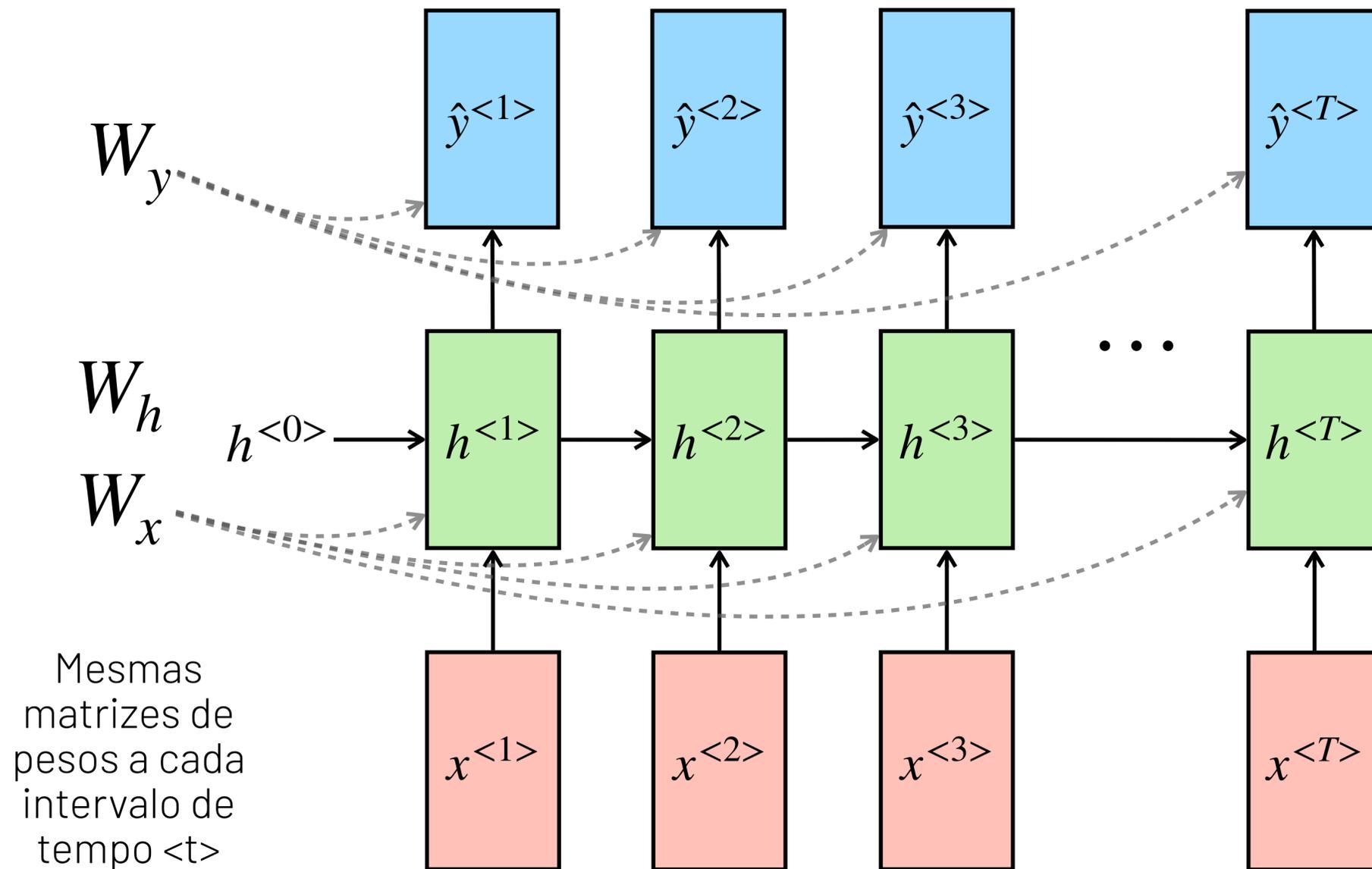
$$\hat{y}^{<t>} = g_2(W_y h^{<t>} + b_y)$$

- ▶ g_1 : função de ativação da camada escondida (tanh/relu)
- ▶ g_2 : função de ativação da camada de saída (sigmoid/softmax)

Rede Neural Recorrente (RNN)



Rede Neural Recorrente (RNN)



$$h^{<1>} = g_1(W_h h^{<0>} + W_x x^{<1>} + b_h)$$

$$\hat{y}^{<1>} = g_2(W_y h^{<1>} + b_y)$$

$$h^{<2>} = g_1(W_h h^{<1>} + W_x x^{<2>} + b_h)$$

$$\hat{y}^{<2>} = g_2(W_y h^{<2>} + b_y)$$

$$h^{<3>} = g_1(W_h h^{<2>} + W_x x^{<3>} + b_h)$$

$$\hat{y}^{<3>} = g_2(W_y h^{<3>} + b_y)$$

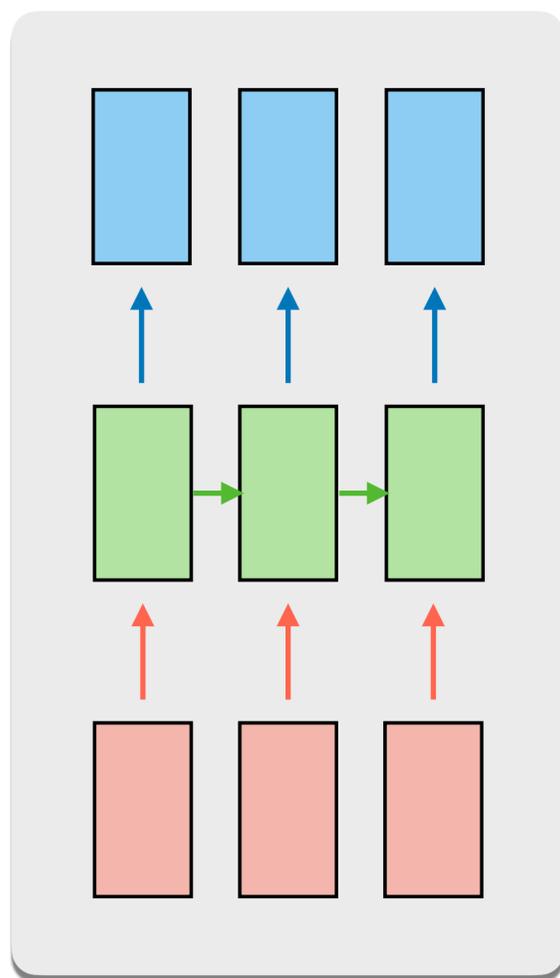
⋮

$$h^{<T>} = g_1(W_h h^{<T-1>} + W_x x^{<T>} + b_h)$$

$$\hat{y}^{<T>} = g_2(W_y h^{<T>} + b_y)$$

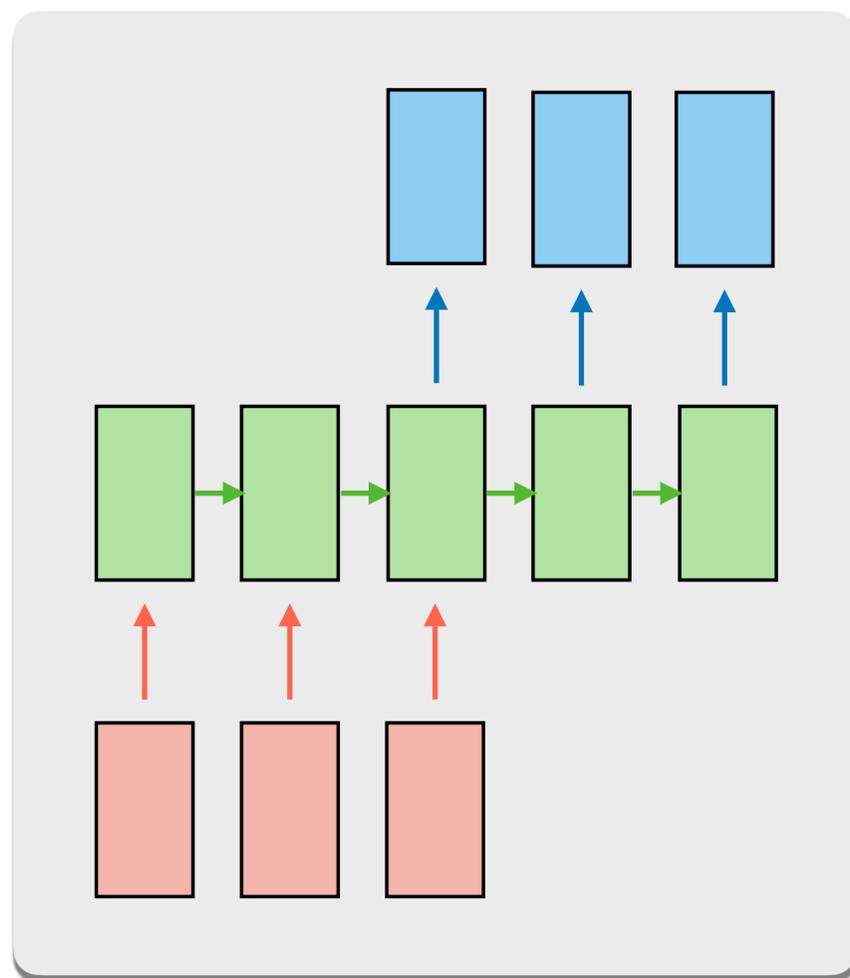
Tipos de RNNs

muitas para muitas



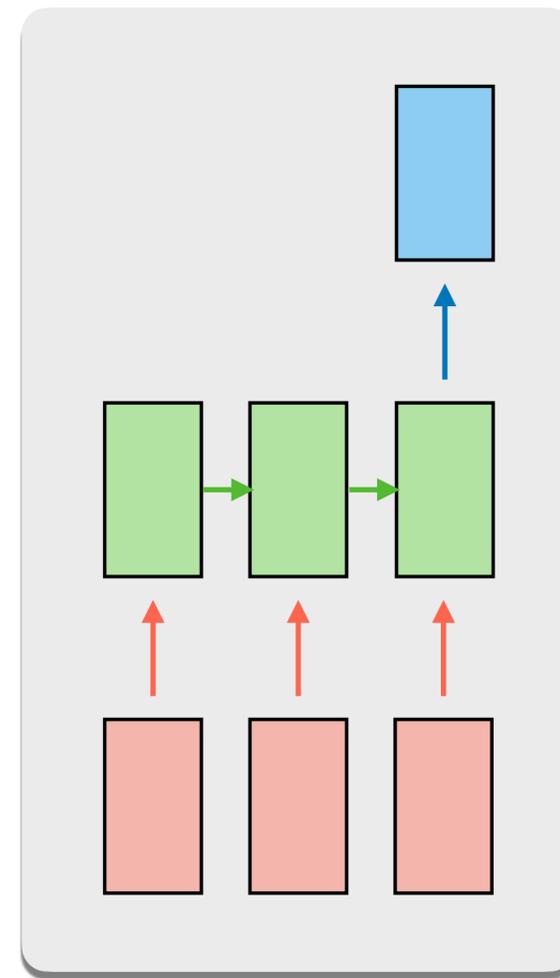
Exemplo
REN

muitas para muitas
(sequência para sequência)



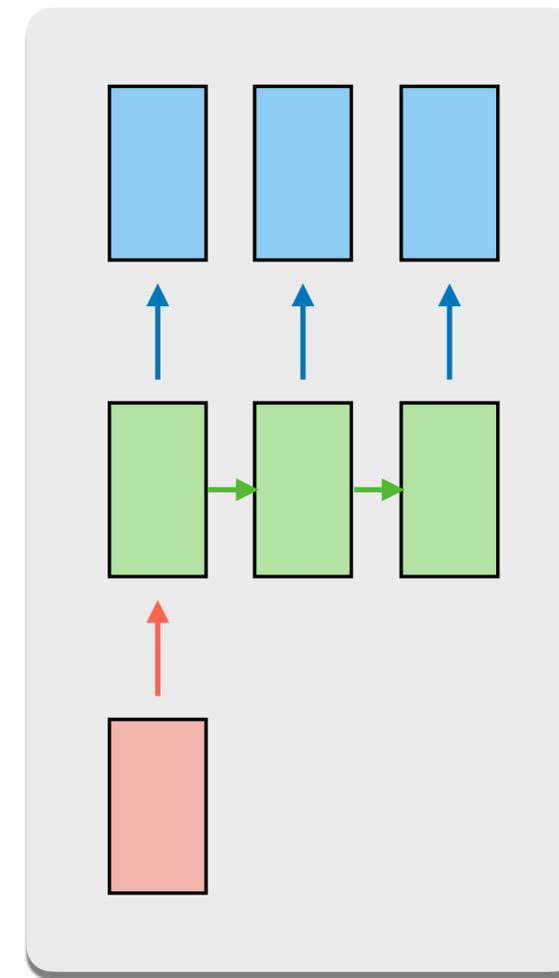
Exemplo
Tradução Automática

muitas para uma



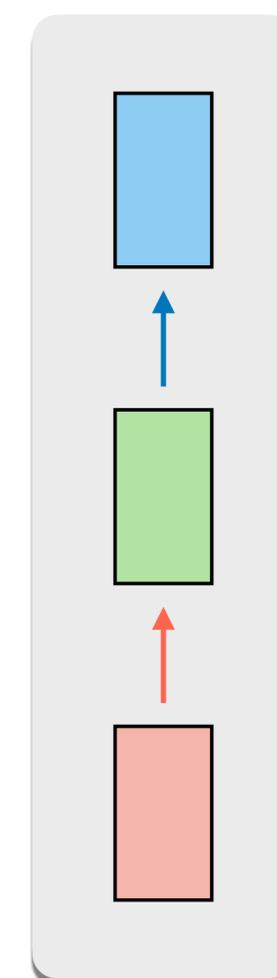
Exemplo
Análise de Sentimento

uma para muitas



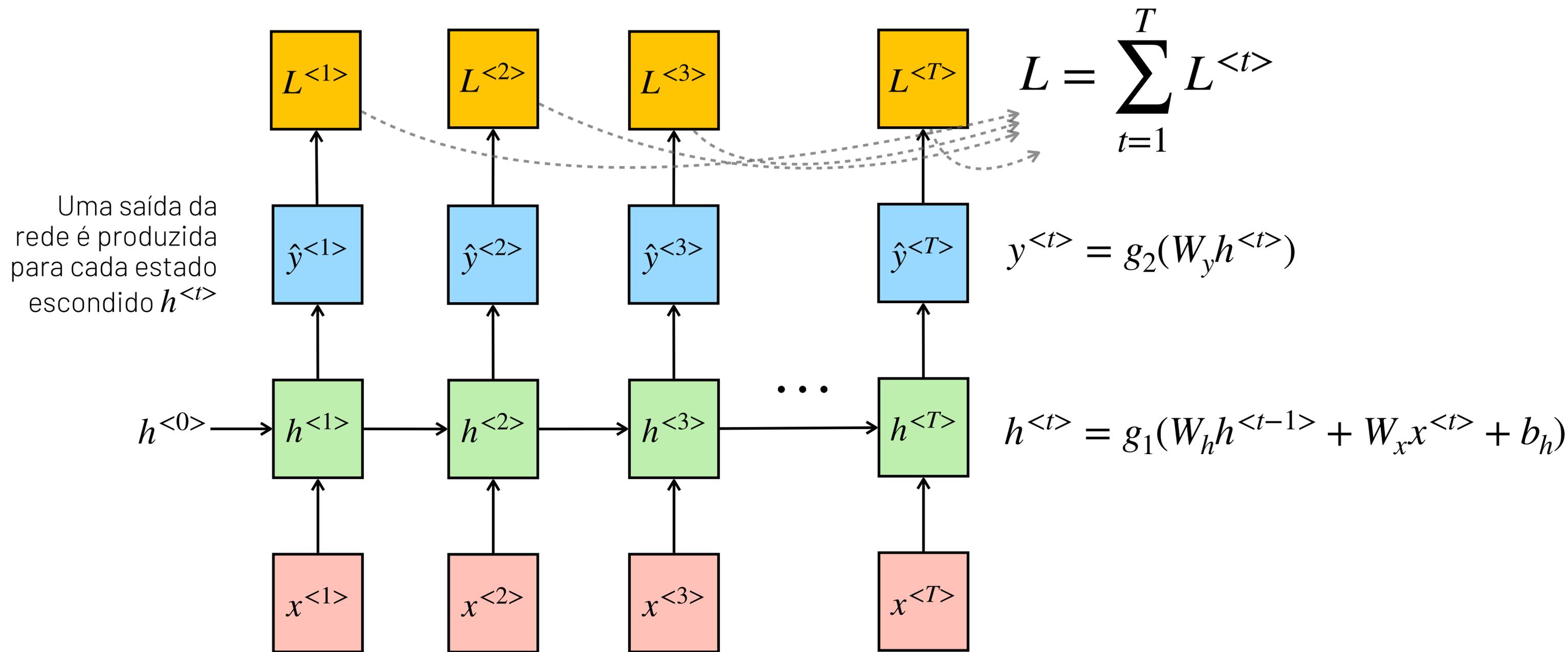
Exemplo
Legenda de imagens

uma para uma

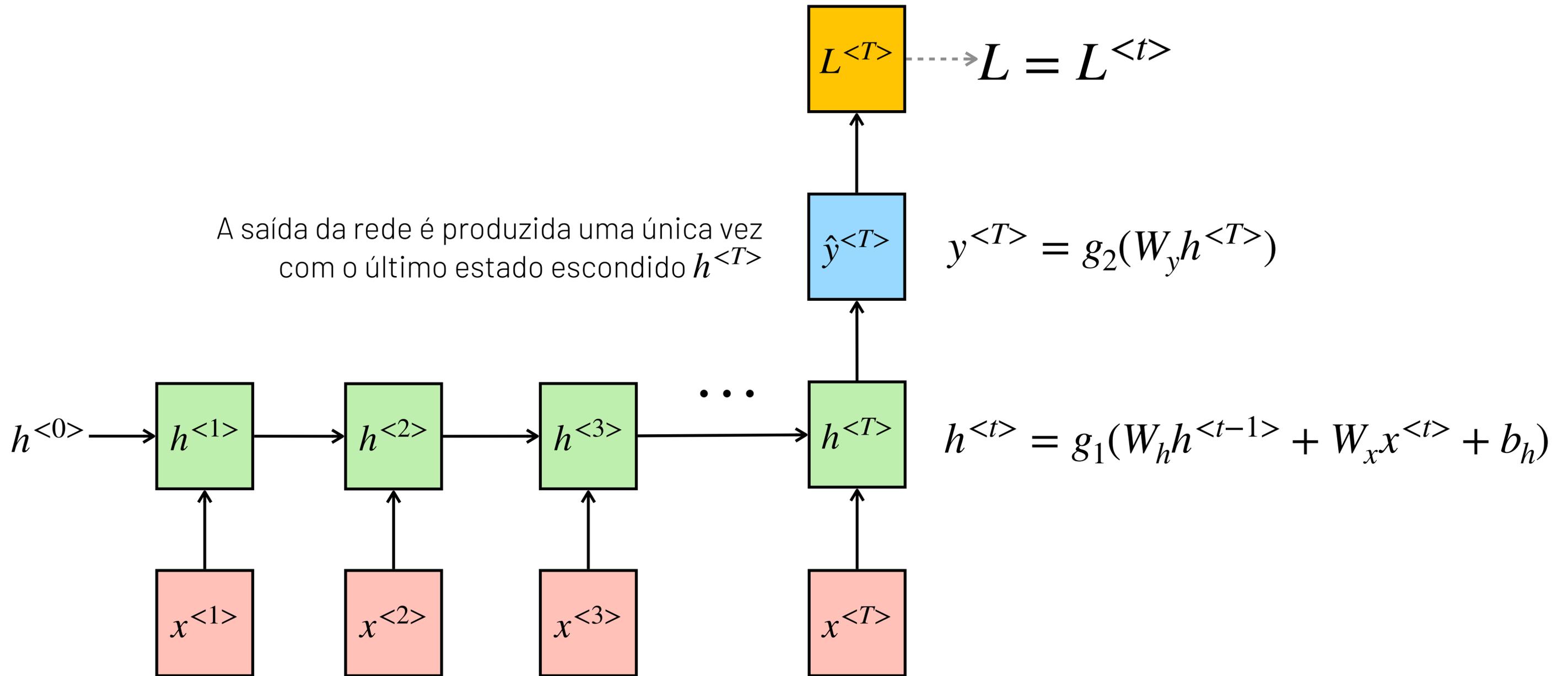


Exemplo
MLP

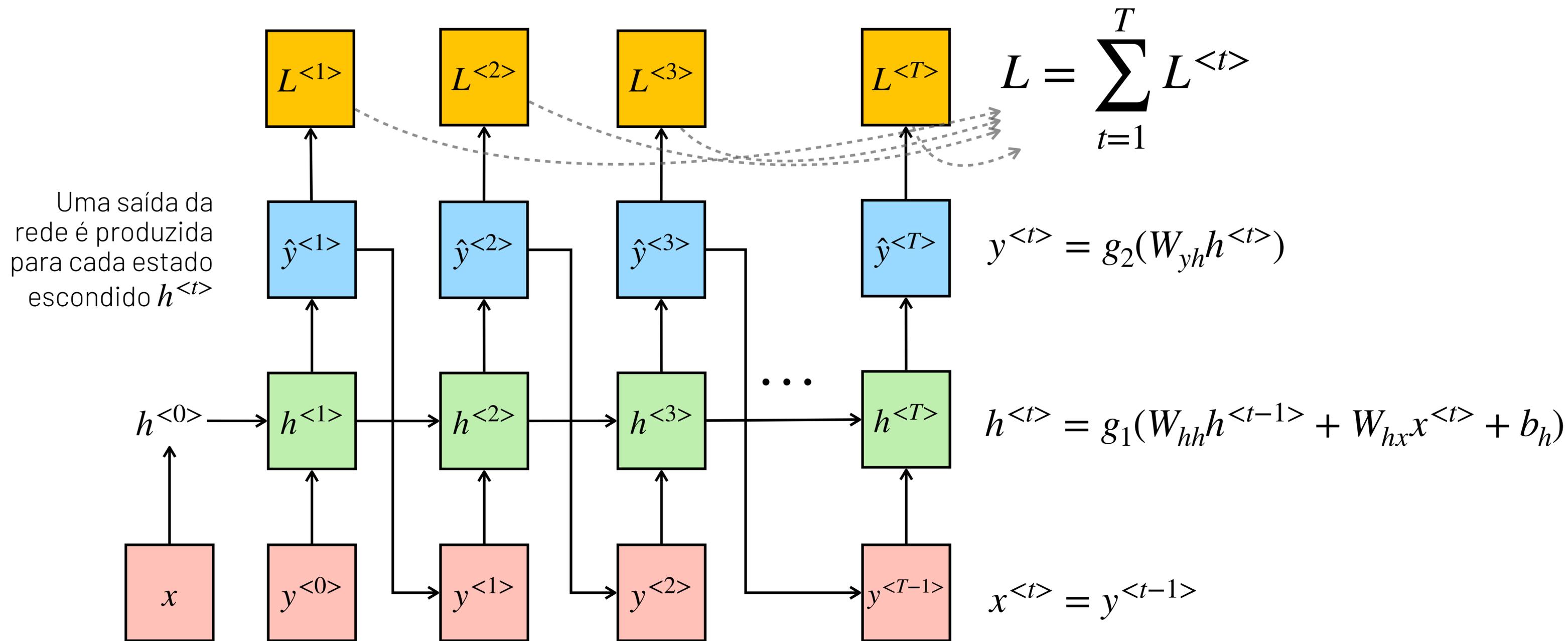
Muitas para Muitas



Muitas para Uma



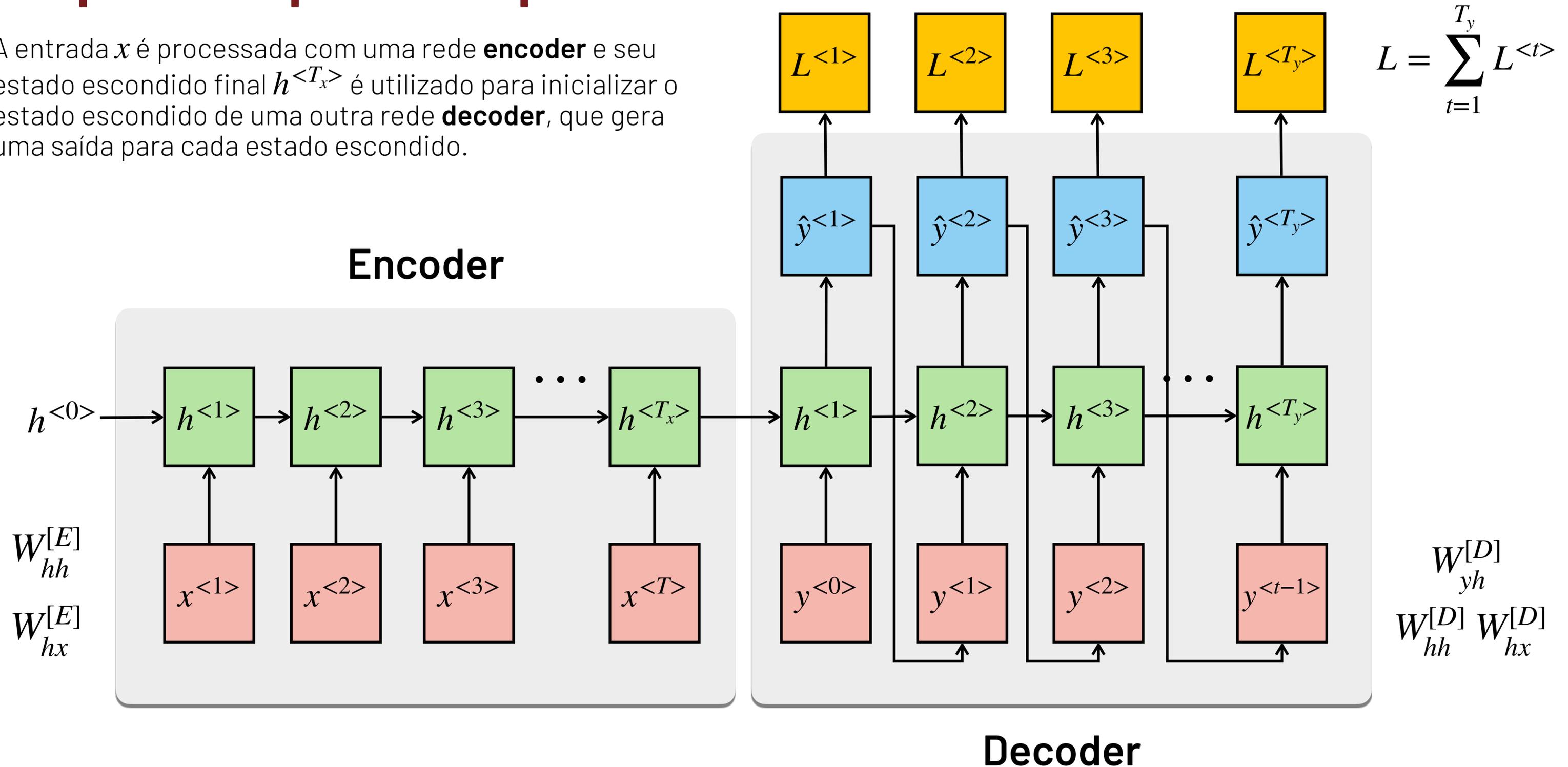
Uma para Muitas



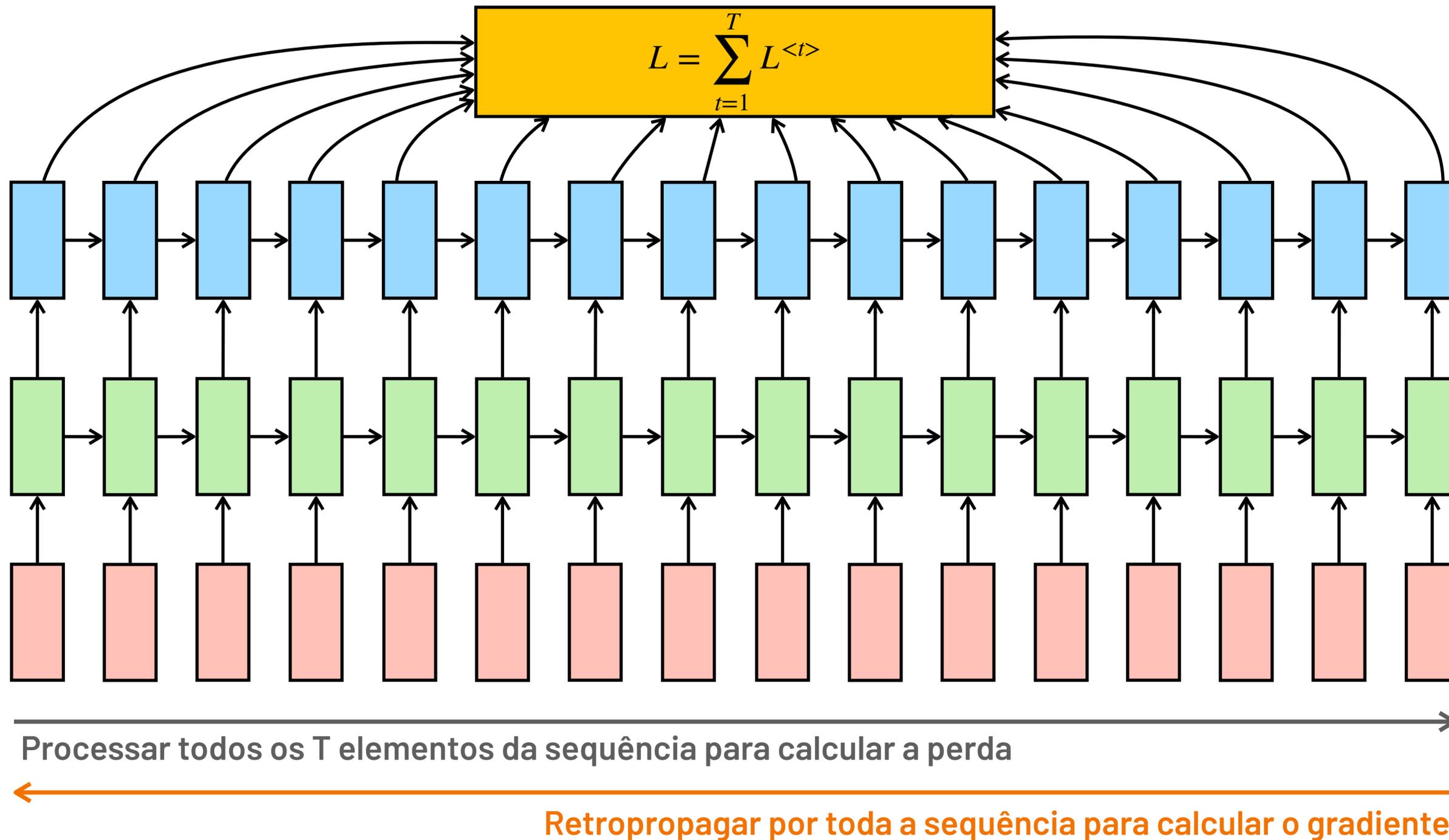
Utilizar x para inicializar $h^{<0>}$

Sequência para Sequência

A entrada x é processada com uma rede **encoder** e seu estado escondido final $h^{<T_x>}$ é utilizado para inicializar o estado escondido de uma outra rede **decoder**, que gera uma saída para cada estado escondido.

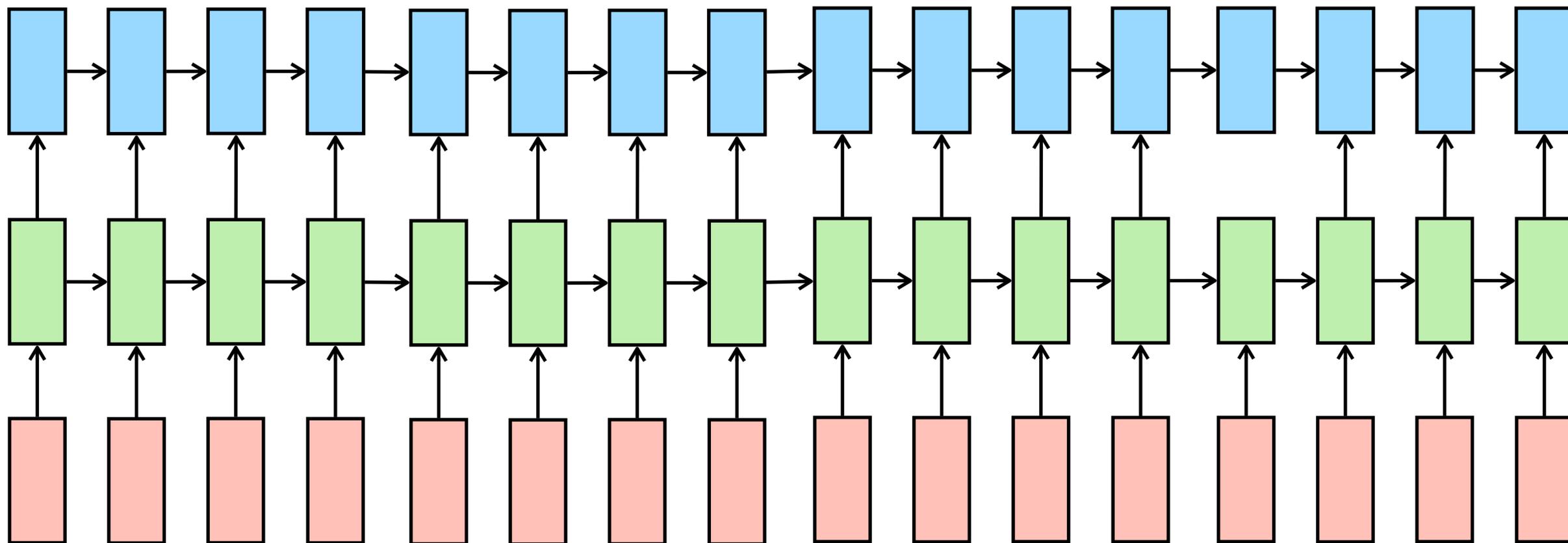


Retropropagação no Tempo



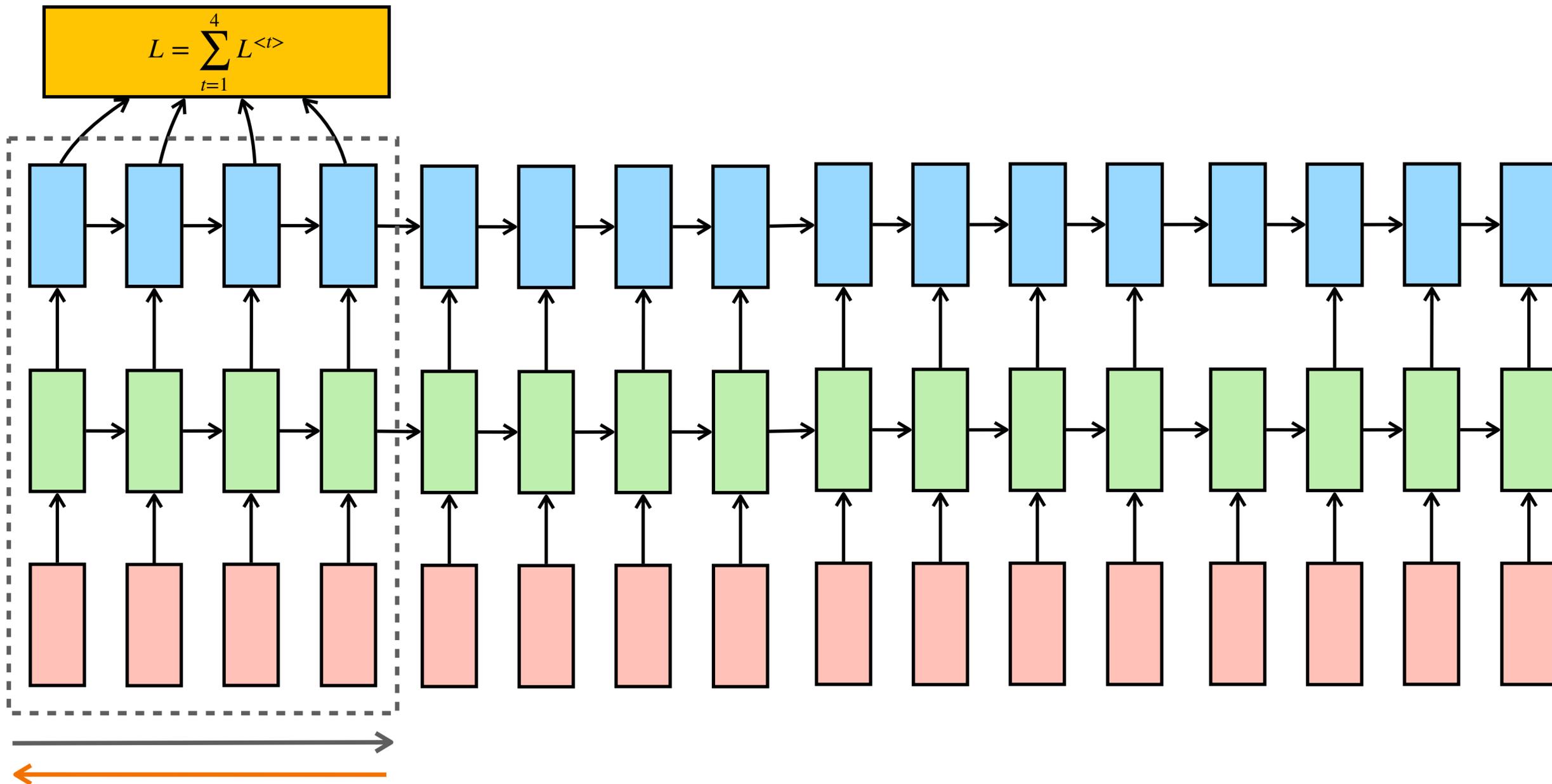
Retropropagação Truncada no Tempo

Se o tamanho da sequência a ser processada é muito grande ou infinita (e.g., séries temporais), executar a **propagação** e a **retropopagação** em partes de tamanho j (e.g., 4)



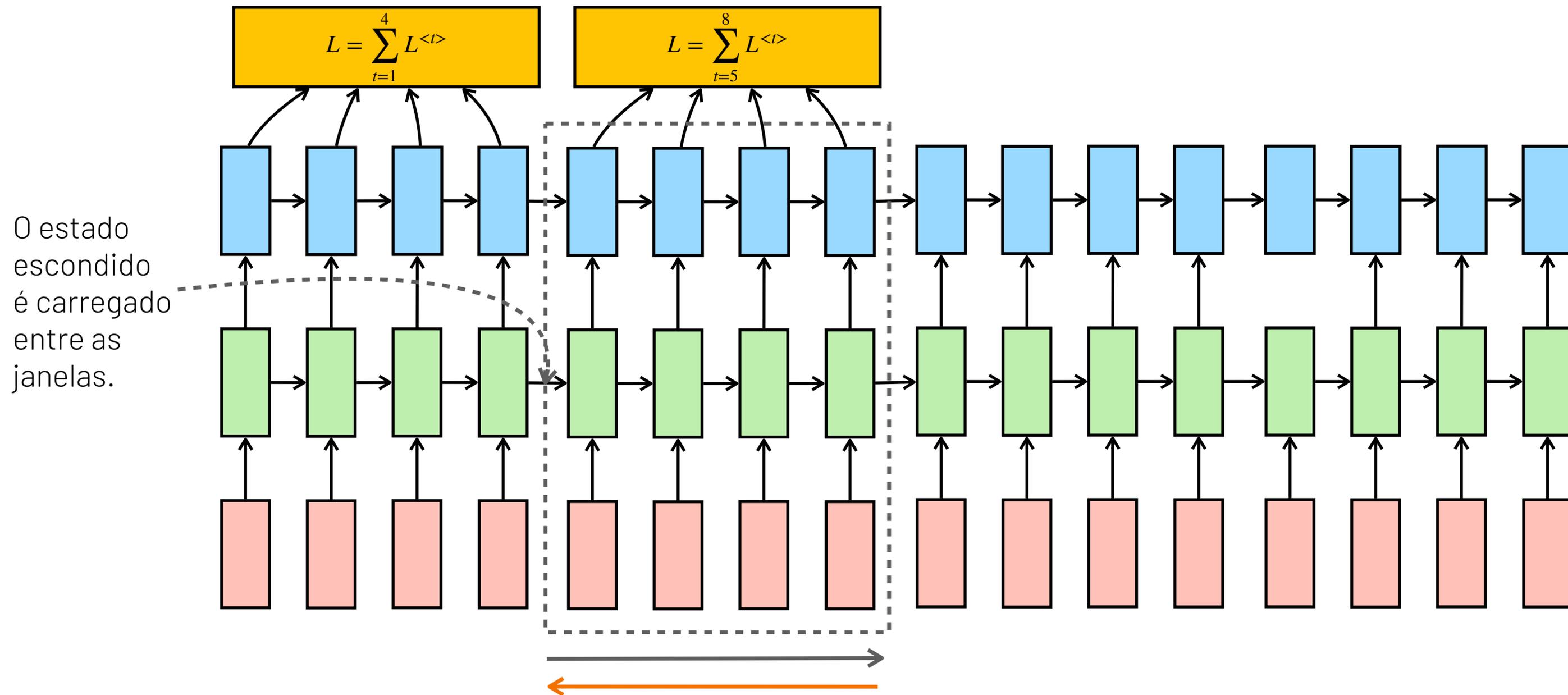
Retropropagação Truncada no Tempo

Se o tamanho da sequência a ser processada é muito grande ou infinita (e.g., séries temporais), executar a **propagação** e a **retropopagação** em partes de tamanho j (e.g., 4)



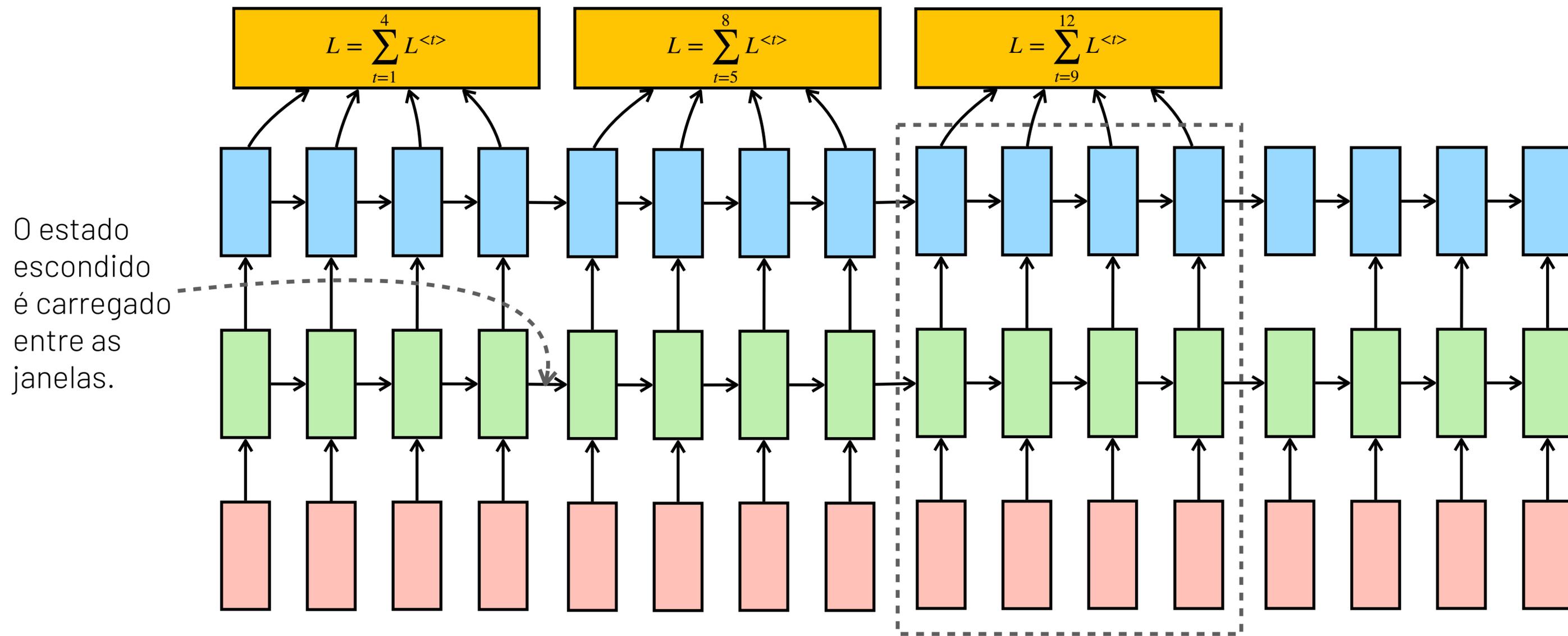
Retropropagação Truncada no Tempo

Se o tamanho da sequência a ser processada é muito grande ou infinita (e.g., séries temporais), executar a **propagação** e a **retropopagação** em partes de tamanho j (e.g., 4)



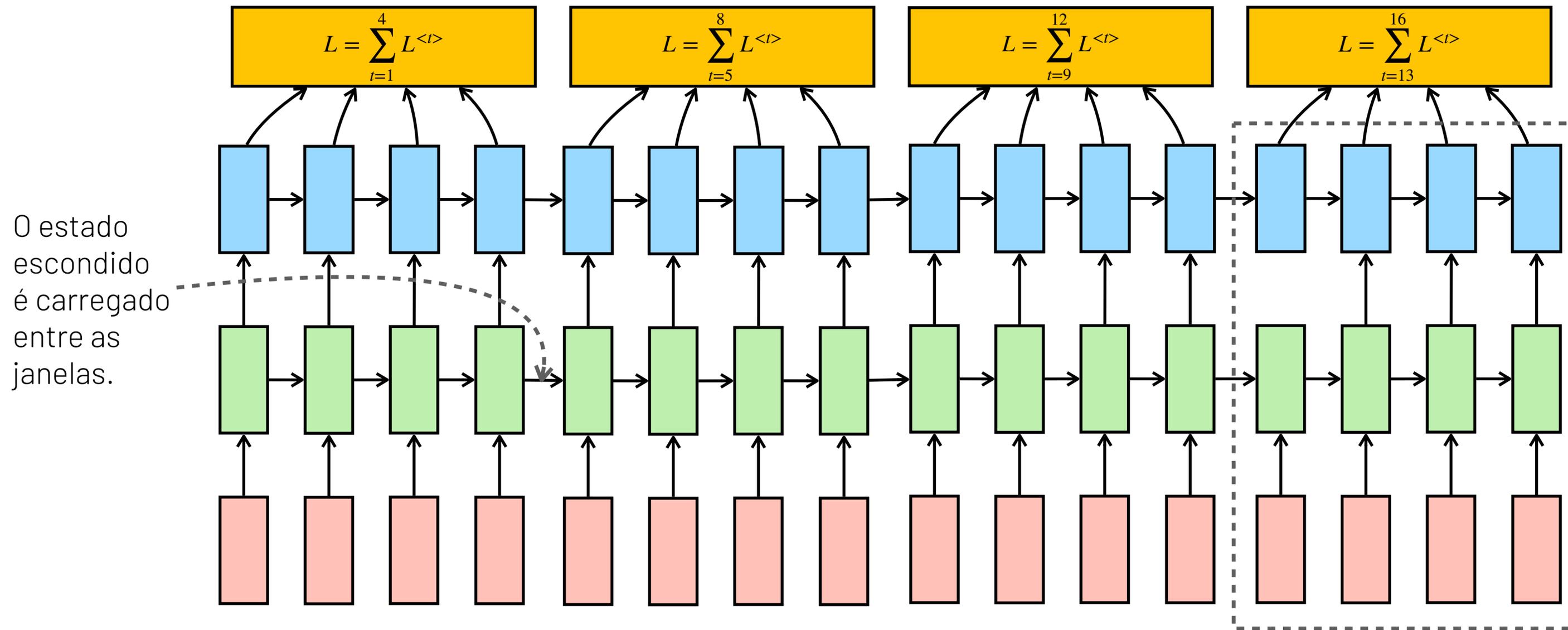
Retropropagação Truncada no Tempo

Se o tamanho da sequência a ser processada é muito grande ou infinita (e.g., séries temporais), executar a **propagação** e a **retropopagação** em partes de tamanho j (e.g., 4)



Retropropagação Truncada no Tempo

Se o tamanho da sequência a ser processada é muito grande ou infinita (e.g., séries temporais), executar a **propagação** e a **retropopagação** em partes de tamanho j (e.g., 4)





Explique em poucas palavras como você funciona.



Eu sou um modelo de linguagem desenvolvido pela OpenAI chamado GPT-3.5. Funciono processando texto ou linguagem escrita para gerar respostas ou textos coerentes e relevantes com base nas informações e no contexto fornecido. Isso é feito por meio do aprendizado de máquina e da análise de vastas quantidades de texto durante o treinamento. Posso ser usado para uma variedade de tarefas, como responder a perguntas, gerar texto, traduzir idiomas e muito mais. Minha capacidade de resposta depende das informações disponíveis até a minha data de corte em setembro de 2021.



Modelo de Linguagem

Essa aula está muito

Modelo de Linguagem

Prever a próxima palavra (ou caractere) a partir de um contexto textual.

Problema fundamental em Processamento de Linguagem Natural (PLN).

Essa aula está muito

contexto x

$$P(x_t | x_{t-1}, x_{t-2}, \dots, x_1)$$

0.31 legal
0.28 interessante
...
0.05 casa
0.01 cachorro

Modelo de Linguagem

<https://pt.wikipedia.org/wiki/>

Universidade Federal de Viçosa

A Universidade Federal de Viçosa (UFV) é uma universidade pública brasileira, com sua sede localizada na cidade de Viçosa, no estado de Minas Gerais, possuindo campus também nas cidades de Rio Paranaíba e Florestal.

Conjunto de dados

Coletar um volume gigante de texto (e.g., wikipedia) e criar exemplos (x, y) usando uma janela deslizante (e.g., tamanho $J=8$)

$x^{(1)}$	A	Universidade	Federal	de	Viçosa	(UFV)	é
$y^{(1)}$	Universidade	Federal	de	Viçosa	(UFV)	é	uma
$x^{(2)}$	universidade	pública	brasileira	,	com	sua	sede
$y^{(2)}$	pública	brasileira	,	com	sua	sede	localizada
$x^{(3)}$	na	cidade	de	Viçosa	,	no	estado
$y^{(3)}$	cidade	de	Viçosa	,	no	estado	de
$x^{(4)}$	Minas	Gerais	,	possuindo	campus	também	nas
$y^{(4)}$	Gerais	,	possuindo	campus	também	nas	Cidades
$x^{(5)}$	de	Rio	Paranaíba	e	Florestal	.	<PAD>
$y^{(5)}$	Rio	Paranaíba	e	Florestal	.	<PAD>	<PAD>

Representação Vetorial de Palavras

x Lucas Ferreira é professor de IA

$x^{<1>}$

$x^{<2>}$

$x^{<3>}$

$x^{<4>}$

$x^{<5>}$

$x^{<6>}$

1	a
2	à
3	Aarão
4	aba
...	...
261.796	zurro
261.797	zurros
261.798	zurrou

0
0
⋮
0
1
0
⋮
0

0
0
⋮
0
1
0
⋮
0

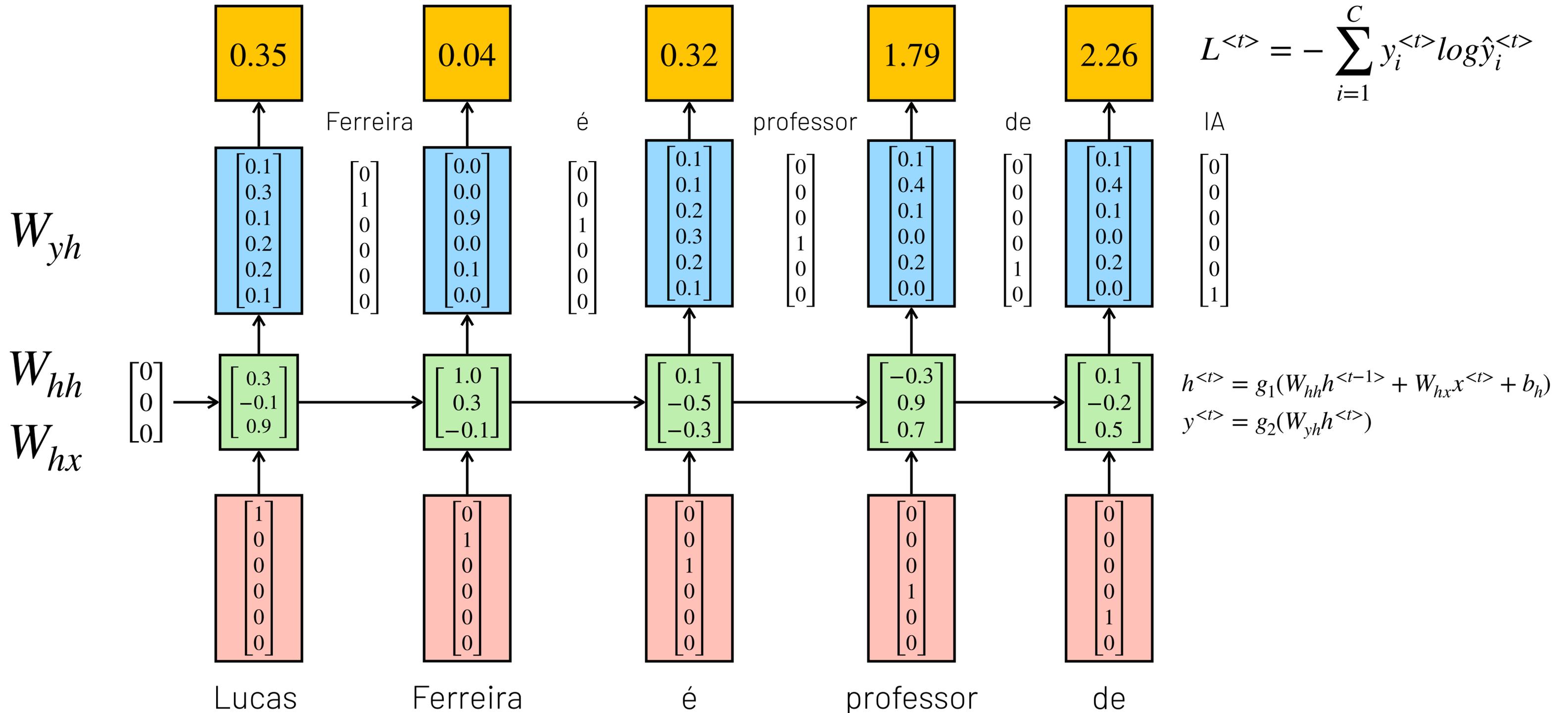
0
0
⋮
0
1
0
⋮
0

261.798

Codificação **one-hot**

Vocabulário

Modelo de Linguagem



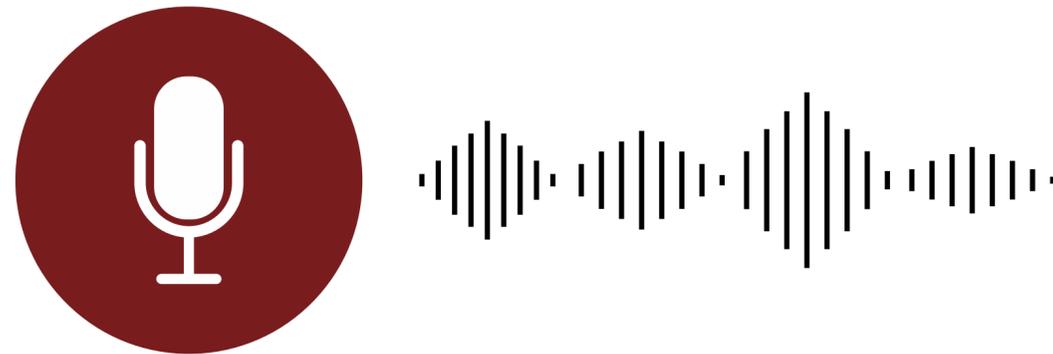
Aplicações de modelos de linguagem

Sugerir próximas palavras ao escrever mensagens de texto.



Aplicações de modelos de linguagem

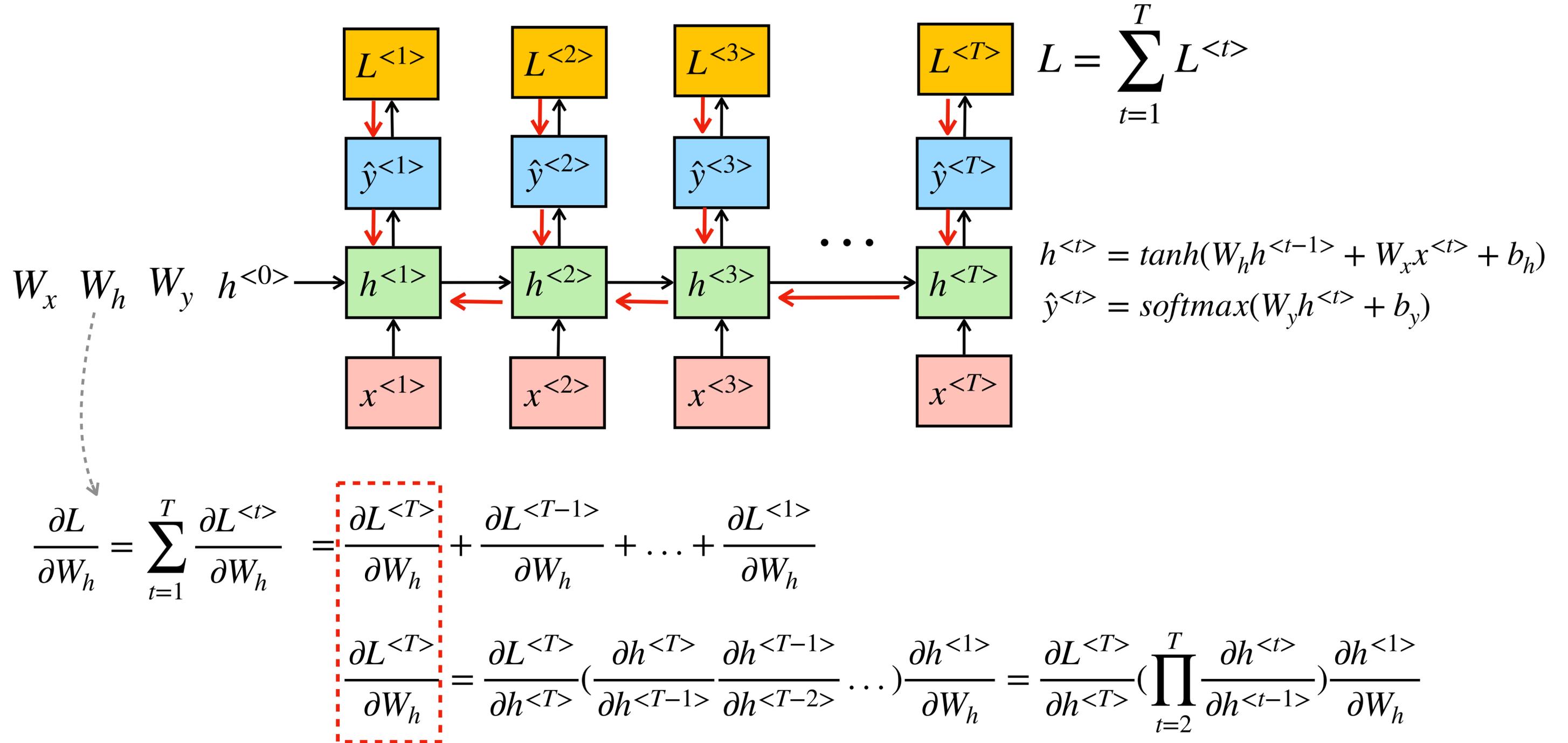
Calcular a probabilidade de uma sequência de palavras.



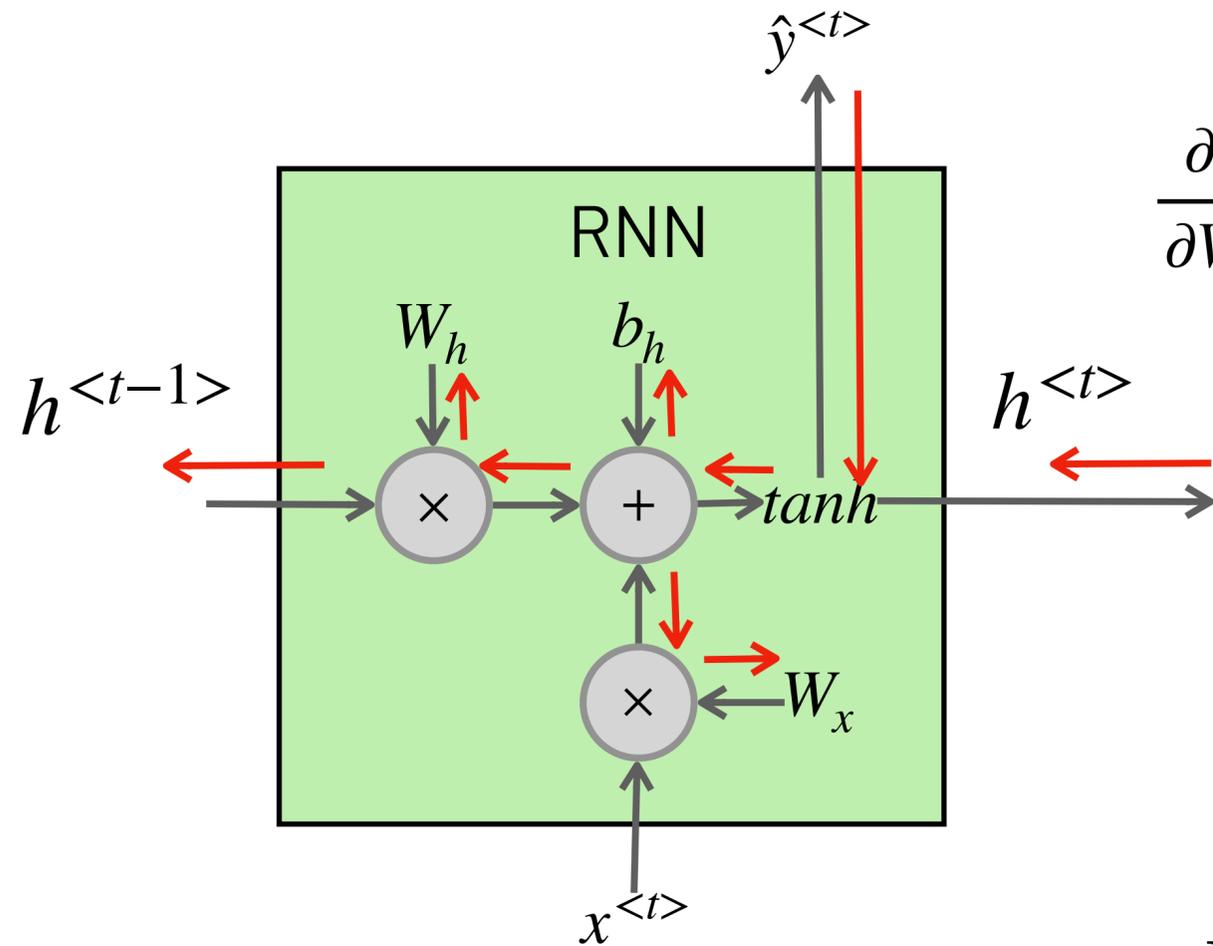
$P(\text{"A palavra rubrica tem **acento**?"}) = 0.23$

$P(\text{"A palavra rubrica tem **assento**?"}) = 0.1$

Explosão e Desaparecimento de Gradientes



Explosão e Desaparecimento de Gradientes



$$\frac{\partial L}{\partial W_h} = \frac{\partial L^{<T>}}{\partial W_h} + \frac{\partial L^{<T-1>}}{\partial W_h} + \dots + \frac{\partial L^{<1>}}{\partial W_h}$$

$$\frac{\partial L^{<T>}}{\partial W_h} = \frac{\partial L^{<T>}}{\partial h^{<T>}} \left(\prod_{t=2}^T \frac{\partial h^{<t>}}{\partial h^{<t-1>}} \right) \frac{\partial h^{<1>}}{\partial W_h}$$

$$\frac{\partial h^{<t>}}{\partial h^{<t-1>}} = \tanh'(W_h h^{<t-1>} + W_x x^{<t>}) W_h$$

$\prod_{t=2}^T \frac{\partial h^{<t>}}{\partial h^{<t-1>}}$ produz uma série de multiplicações de W_h por ela mesma:

- ▶ Se os pesos de $W_h > 1 \rightarrow$ explosão de gradientes
- ▶ Se os pesos de $W_h < 1 \rightarrow$ desaparecimento de gradientes

Próxima aula

A17: GRU e LSTM

Arquiteturas recorrentes avançadas para processamento de sequências longas.