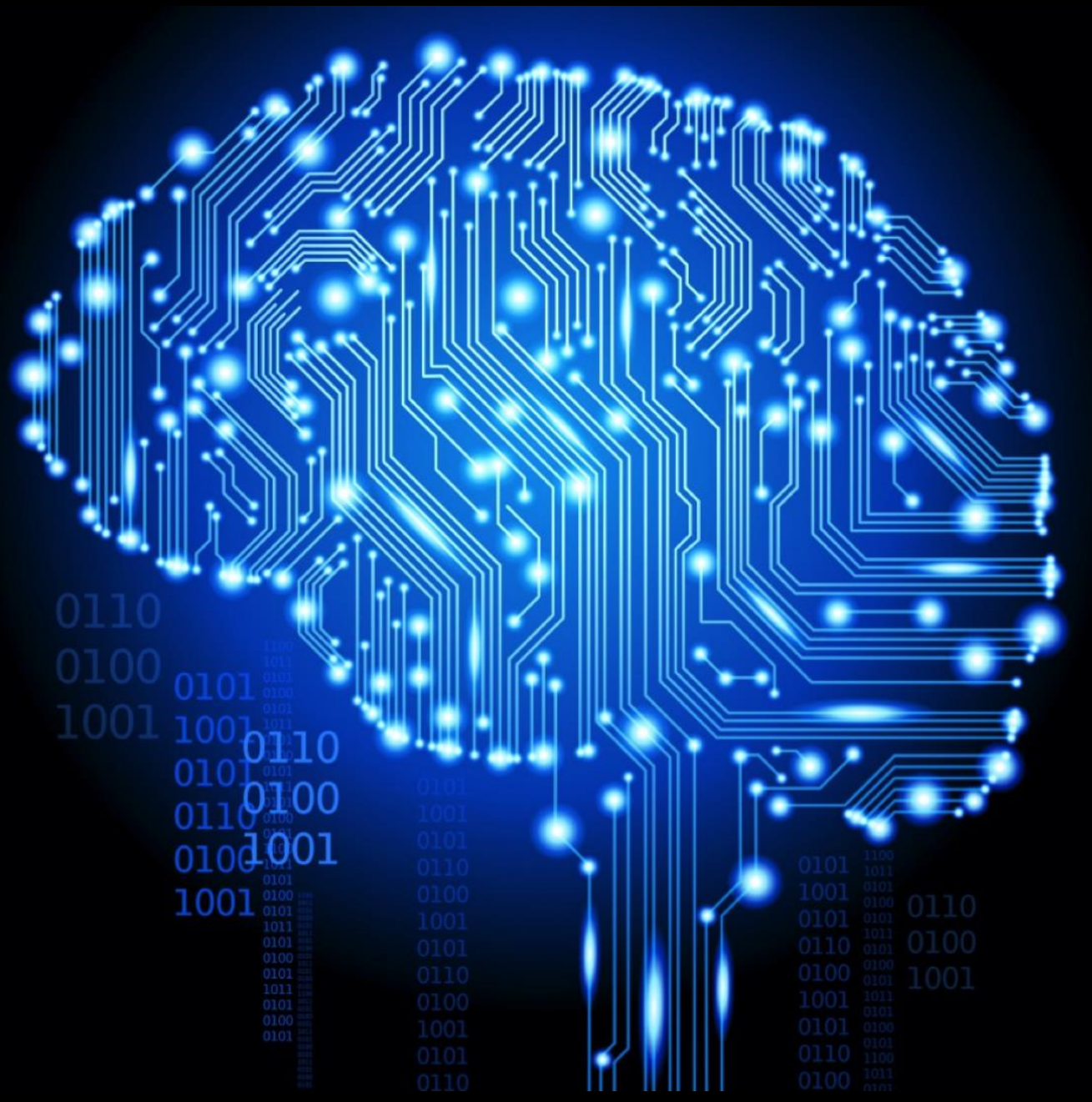
Redes Neurais Recorrentes

RAFAEL TEIXEIRA SOUSA

www.deeplearningbrasil.com.br



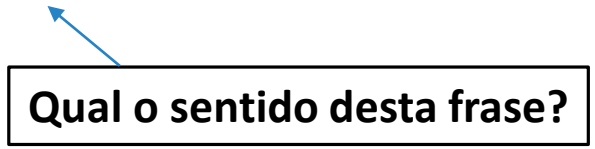
Recorrentes? Por quê?

Quais são as **deficiência** das ANNs (MLP) e das CNNs?

Recorrentes? Por quê?

Quais são as deficiência das ANNs (MLP) e das CNNs?

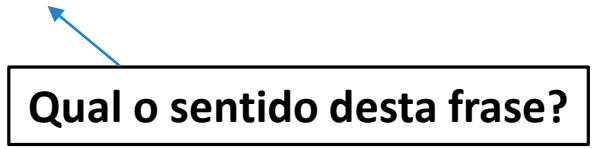
Estamos sem ar...



Recorrentes? Por quê?

Quais são as deficiência das ANNs (MLP) e das CNNs?

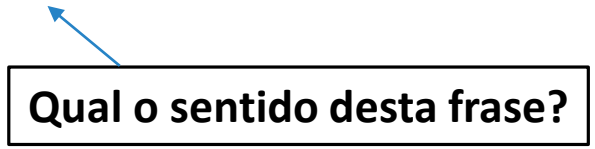
Estamos sem ar condicionado



Recorrentes? Por quê?

Quais são as deficiência das ANNs (MLP) e das CNNs?

Estamos sem ar condicionado



“O filme tenta emocionar e envolver o espectador com atuações carismáticas de bons atores, mas falha devido ao péssimo roteiro.”

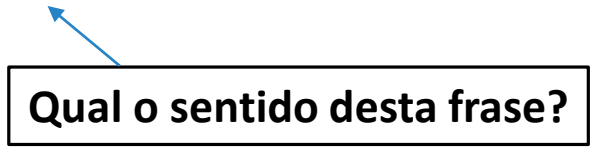
Positivo?

Negativo?

Recorrentes? Por quê?

Quais são as deficiência das ANNs (MLP) e das CNNs?

Estamos sem ar condicionado



“O filme tenta emocionar e envolver o espectador com atuações carismáticas de bons atores, mas falha devido ao péssimo roteiro.”

“Devido ao péssimo roteiro o filme falha em tentar emocionar e envolver o espectador, mesmo contando com atuações carismáticas de bons atores.”

Positivo?

Negativo?

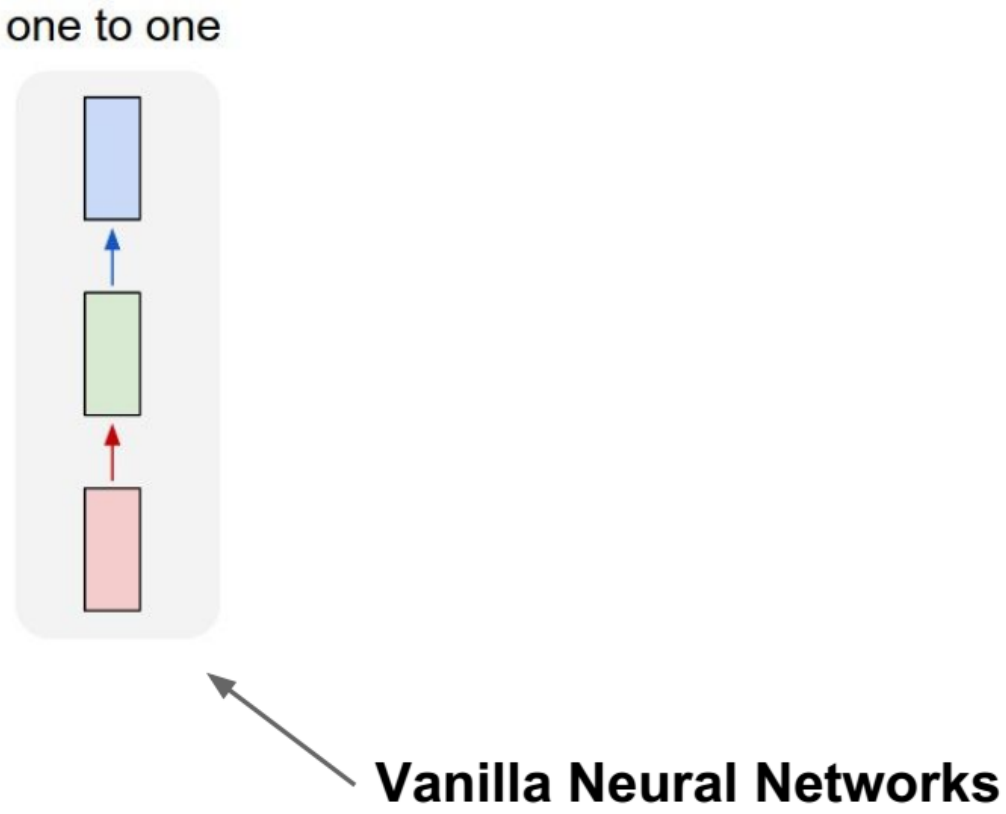
Caso Tesla

Caso Tesla

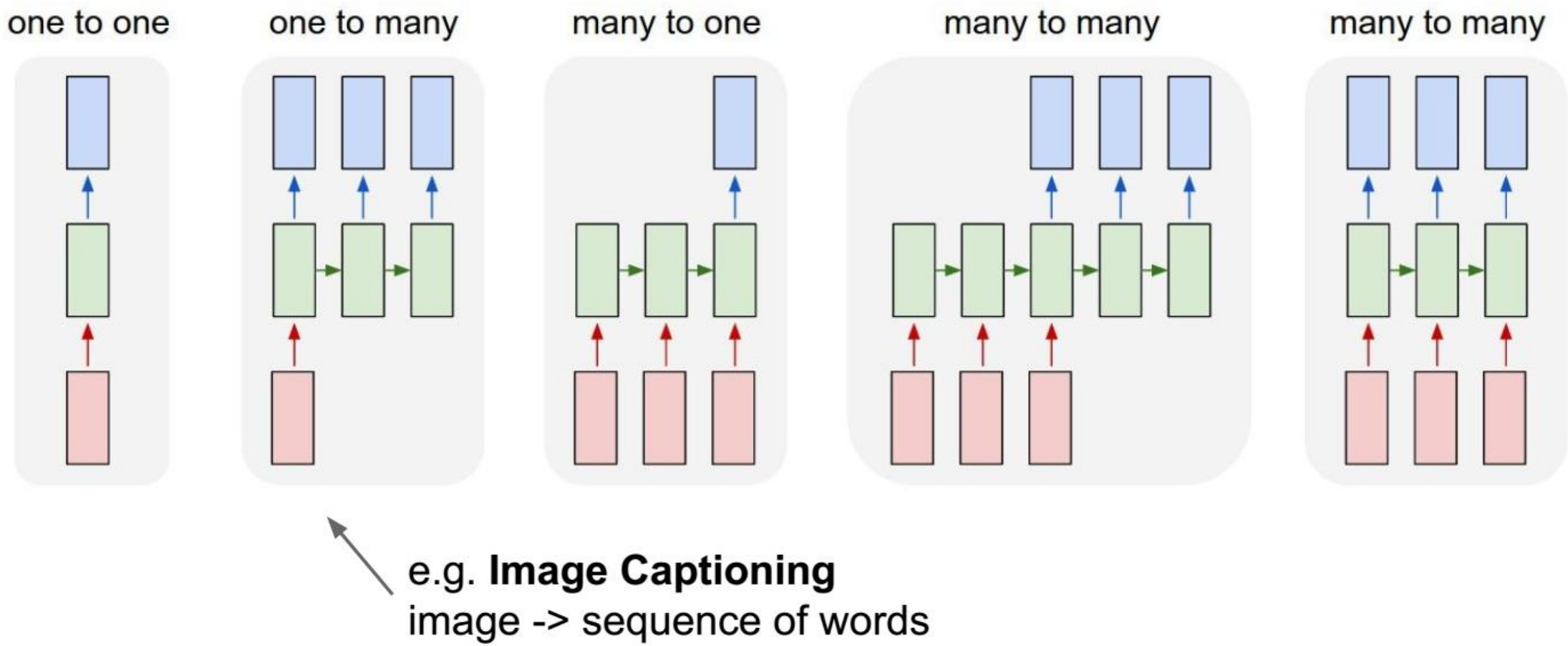
É possível prever o acidente usando apenas uma imagem?


Tempo

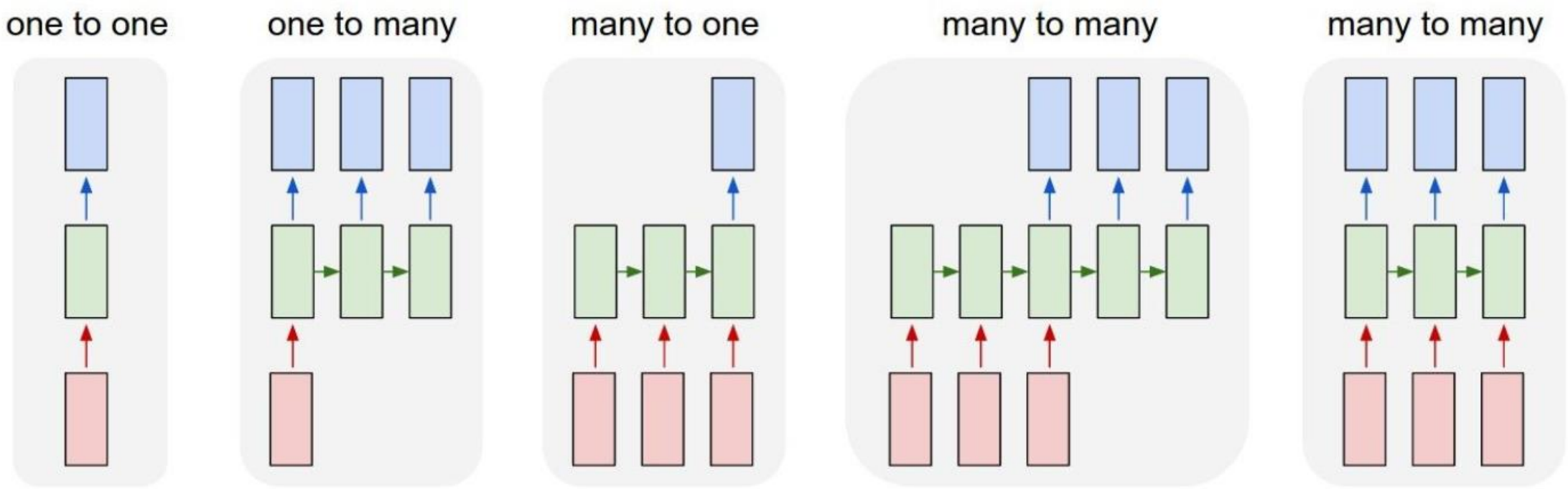
Redes Neurais Recorrentes



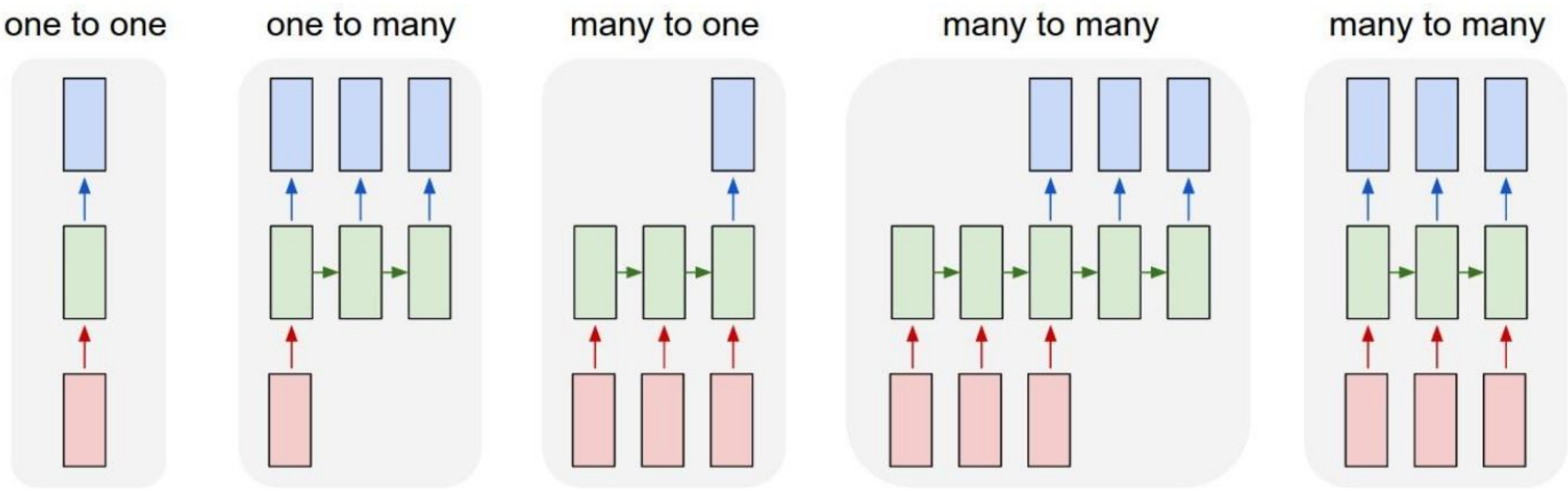
Redes Neurais Recorrentes



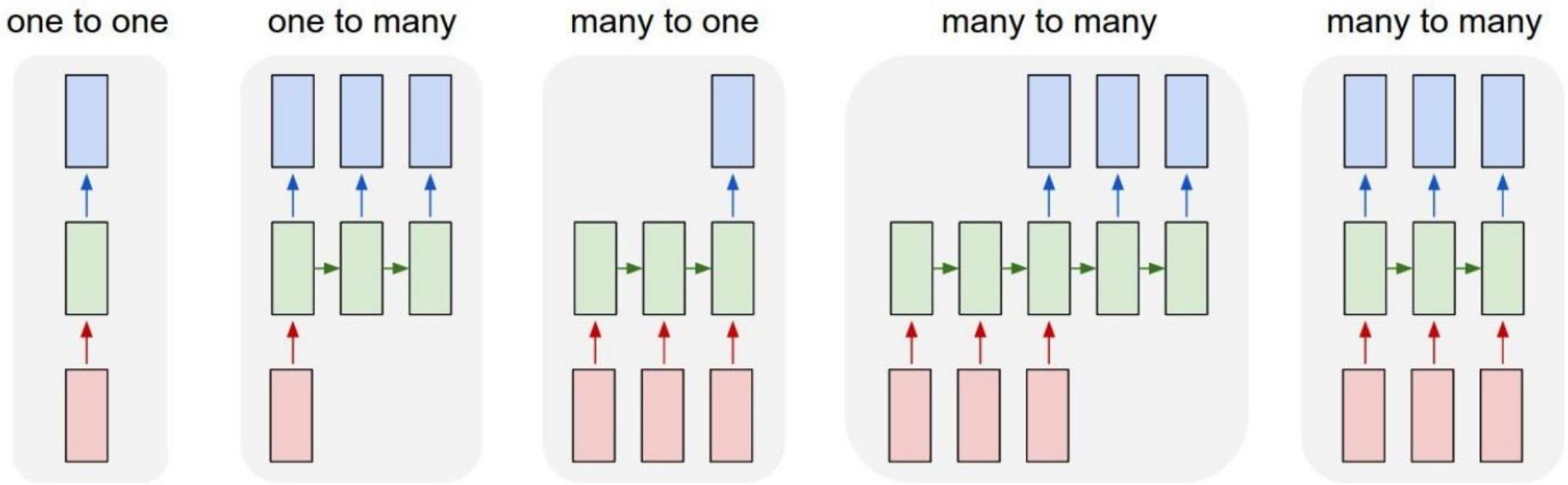
Redes Neurais Recorrentes



Redes Neurais Recorrentes



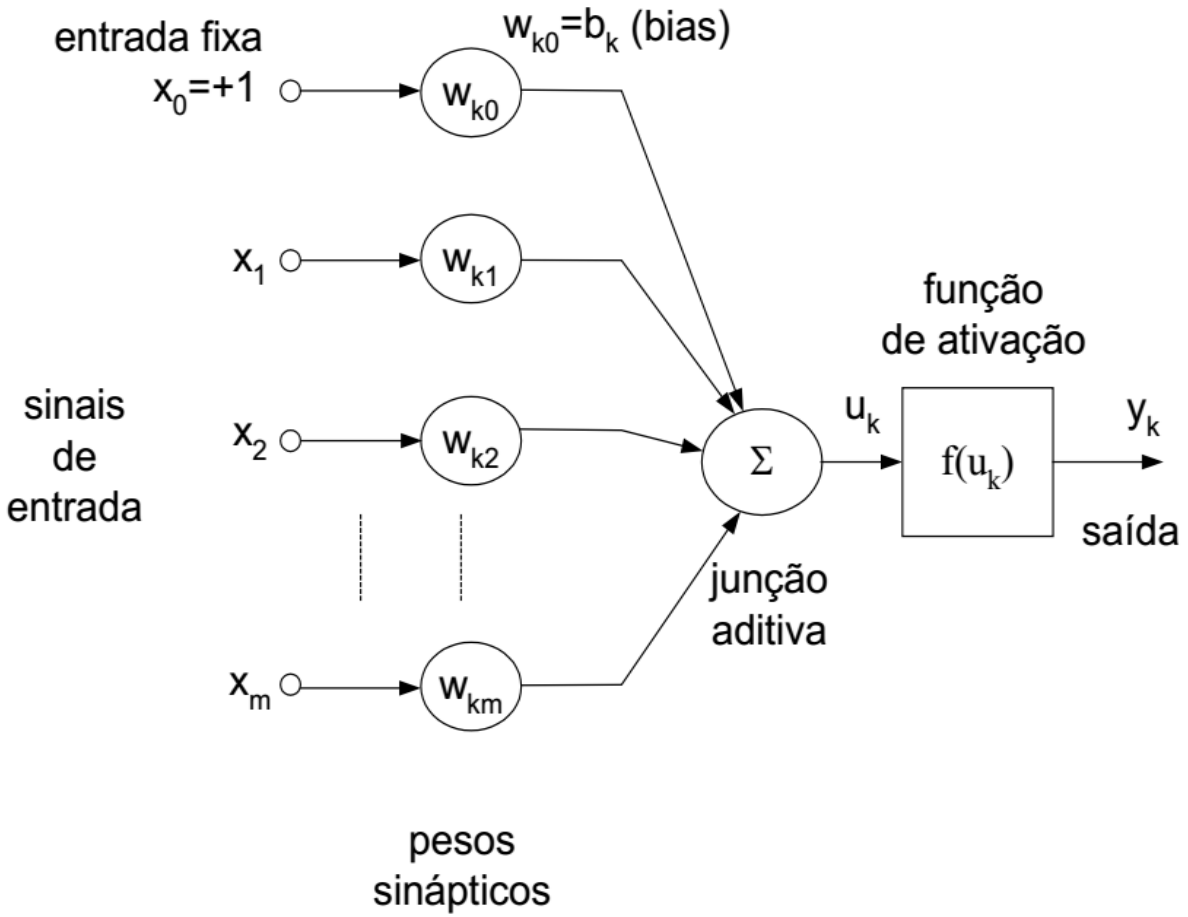
Redes Neurais Recorrentes



Redes Neurais Recorrentes

1. RNN clássica
2. LSTM
3. GRU

Redes Neural



Redes Neural Recorrente (RNN)

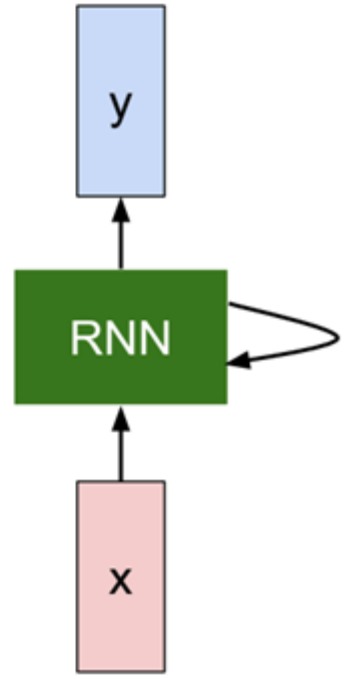
$$h_t = f_W(h_{t-1}, x_t)$$

Input no tempo t

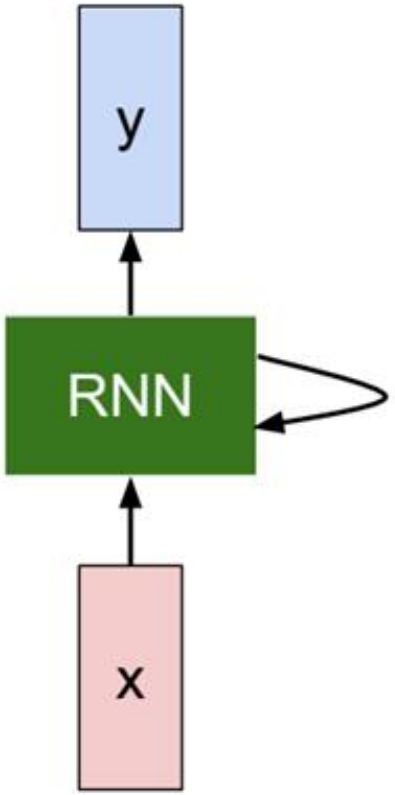
Novo Estado

Estado Anterior

Alguma Função
com Parâmetros W



RNN

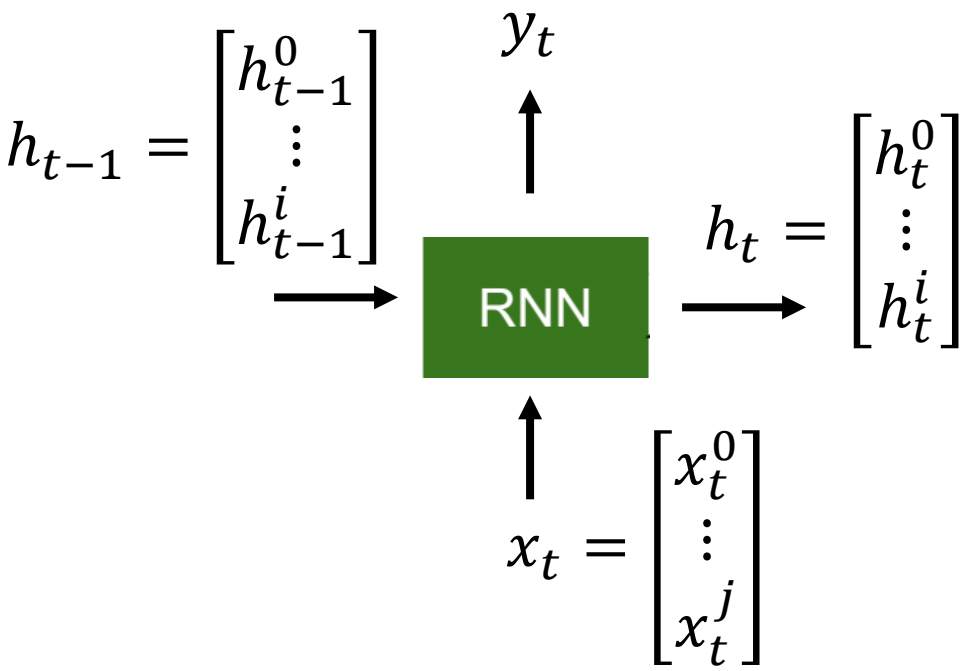


$$h_t = f_W(h_{t-1}, x_t)$$

$$h_t = \tanh(W_{hh}h_{t-1} + W_{hx}x_t)$$

$$y_t = W_{hy}h_t$$

RNN



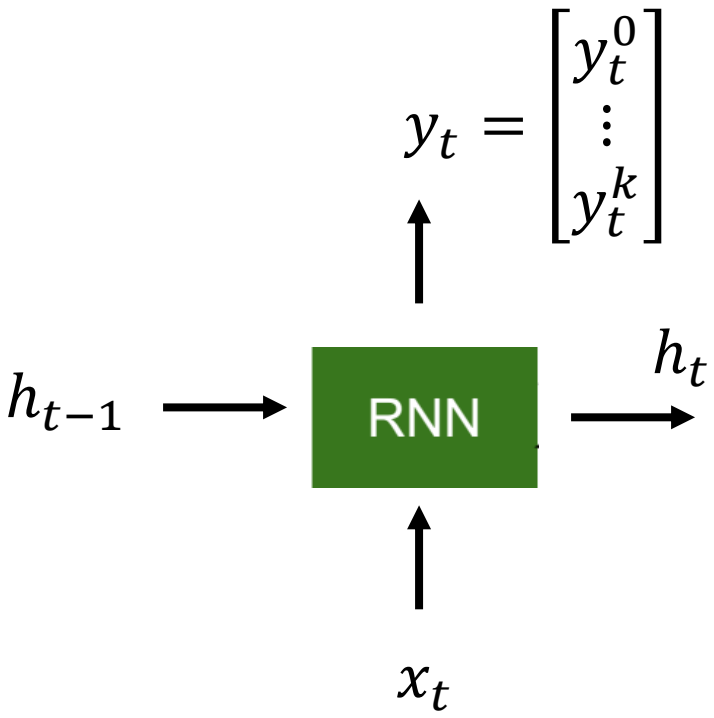
$$h_t = \tanh(W_{hh} h_{t-1} + W_{hx} x_t)$$

$$W_{hh} = \begin{bmatrix} w_{00} & \cdots & w_{0i} \\ \vdots & \ddots & \vdots \\ w_{i0} & \cdots & w_{ii} \end{bmatrix} = w_{ij} \in \mathbb{R}^{i \times i}$$

$$W_{hx} = \begin{bmatrix} w_{00} & \cdots & w_{0j} \\ \vdots & \ddots & \vdots \\ w_{i0} & \cdots & w_{ij} \end{bmatrix} = w_{ij} \in \mathbb{R}^{i \times j}$$

$$h_t = \tanh([W_{hh} \quad W_{hx}] \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix})$$

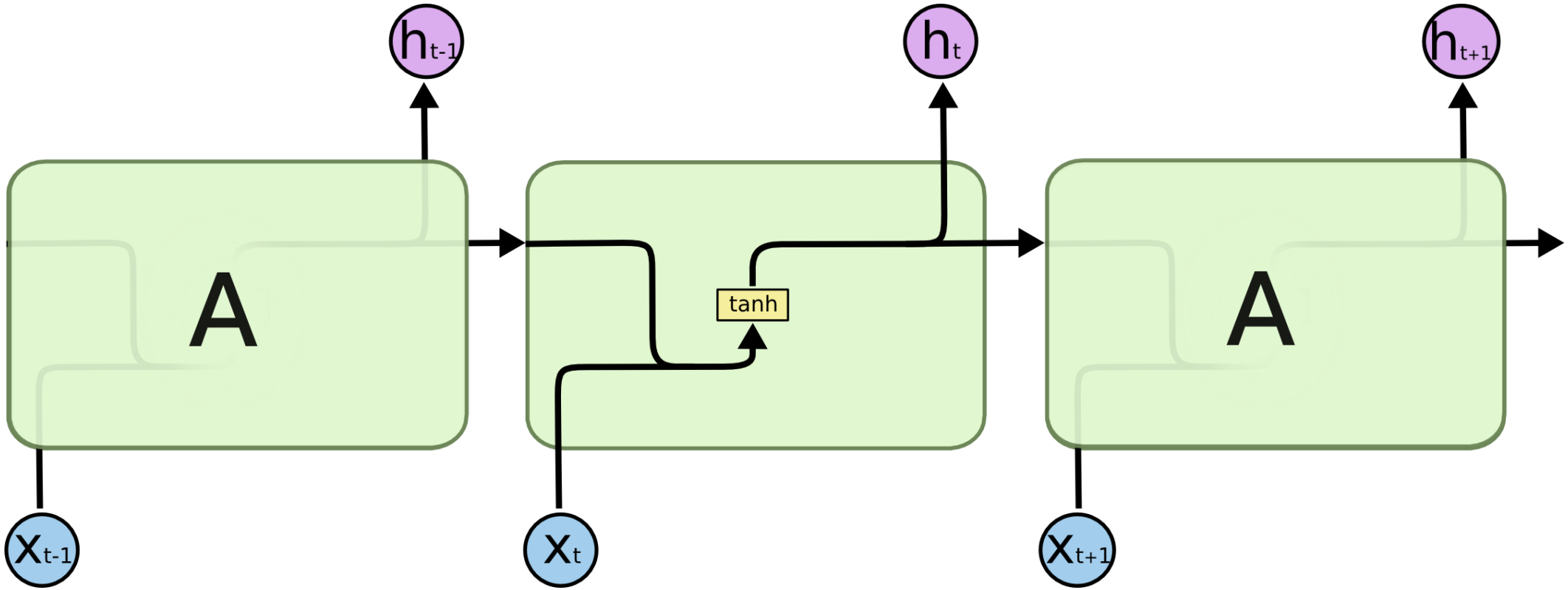
RNN



$$y_t = W_{hy} h_t$$

$$W_{hy} = \begin{bmatrix} w_{00} & \cdots & w_{0k} \\ \vdots & \ddots & \vdots \\ w_{i0} & \cdots & w_{ik} \end{bmatrix} = w_{vi} \in \mathbb{R}^{i \times k}$$

RNN



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

RNN em NLP

Se aplicarmos na saída de uma RNN um **softmax** com todas as possíveis palavras de um idioma:

$$y_t = softmax(W_{hy} h_t)$$

Teoricamente:

$$\hat{P}(x_{t+1} | x_t, \dots, x_0) = y_t$$

#SQN

Problemas das RNNs tradicionais:

- Como treinar?

Treino - RNN

Backpropagation

$$\frac{\partial E}{\partial W} = \sum_{t=1}^{T} \frac{\partial E_t}{\partial W}$$

$$\frac{\partial E_t}{\partial W} = \sum_{k=1}^{t} \frac{\partial E_t}{\partial y_t} \frac{\partial y_t}{\partial h_t} \frac{\partial h_t}{\partial h_k} \frac{\partial h_k}{\partial W}$$

Treino - RNN

Backpropagation

$$\frac{\partial E}{\partial W} = \sum_{t=1}^{T} \frac{\partial E_t}{\partial W}$$

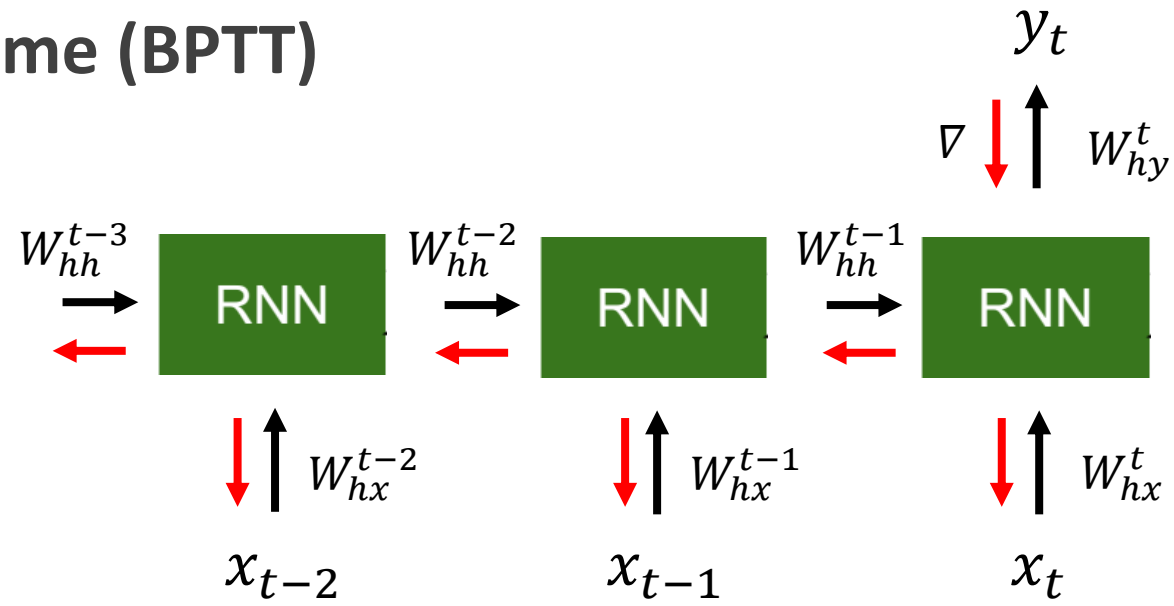
$$\frac{\partial E_t}{\partial W} = \sum_{k=1}^{t} \frac{\partial E_t}{\partial y_t} \frac{\partial y_t}{\partial h_t} \boxed{\frac{\partial h_t}{\partial h_k}} \frac{\partial h_k}{\partial W}$$

Treino - RNN

Backpropagation Through Time (BPTT)

$$\frac{\partial E}{\partial W} = \sum_{t=1}^{T} \frac{\partial E_t}{\partial W}$$

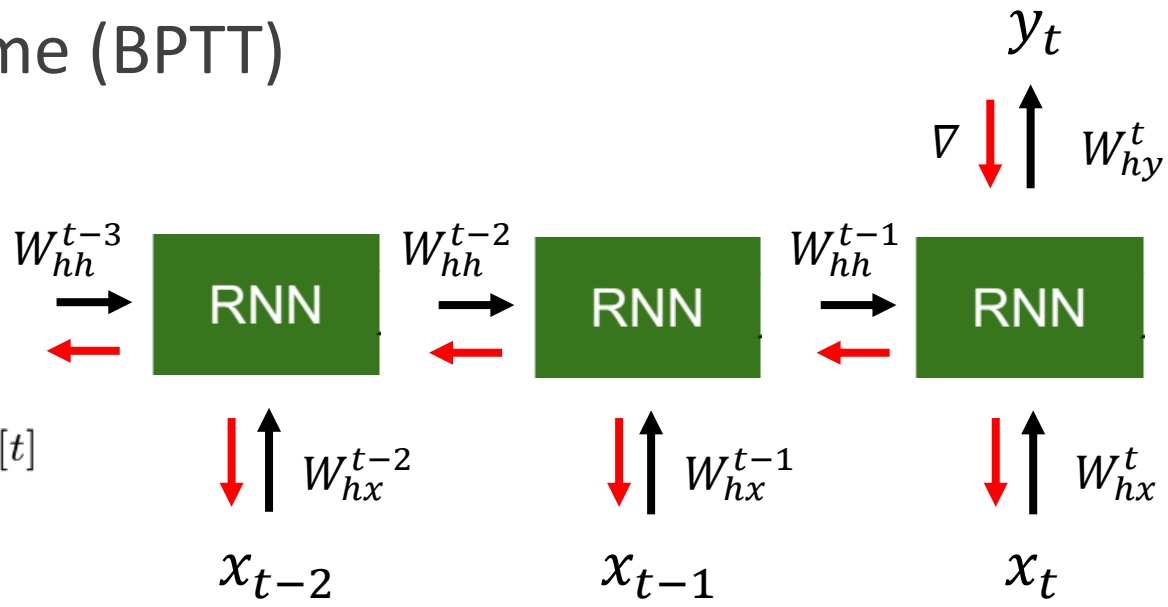
$$\frac{\partial E_t}{\partial W} = \sum_{k=1}^{t} \frac{\partial E_t}{\partial y_t} \frac{\partial y_t}{\partial h_t} \boxed{\frac{\partial h_t}{\partial h_k}} \frac{\partial h_k}{\partial W}$$



Treino - RNN

Backpropagation Through Time (BPTT)

$$\frac{\partial E_t}{\partial W} = \sum_{k=1}^{t} \frac{\partial E_t}{\partial y_t} \frac{\partial y_t}{\partial h_t} \boxed{\frac{\partial h_t}{\partial h_k}} \frac{\partial h_k}{\partial W}$$



Lembrando que: $h_t = W f(h_{t-1}) + W^{(hx)} x_{[t]}$

$$\frac{\partial h_t}{\partial h_k} = \prod_{j=k+1}^{t} \frac{\partial h_j}{\partial h_{j-1}}$$

Cada parcial é uma matriz **Jacobiana**

Treino - RNN

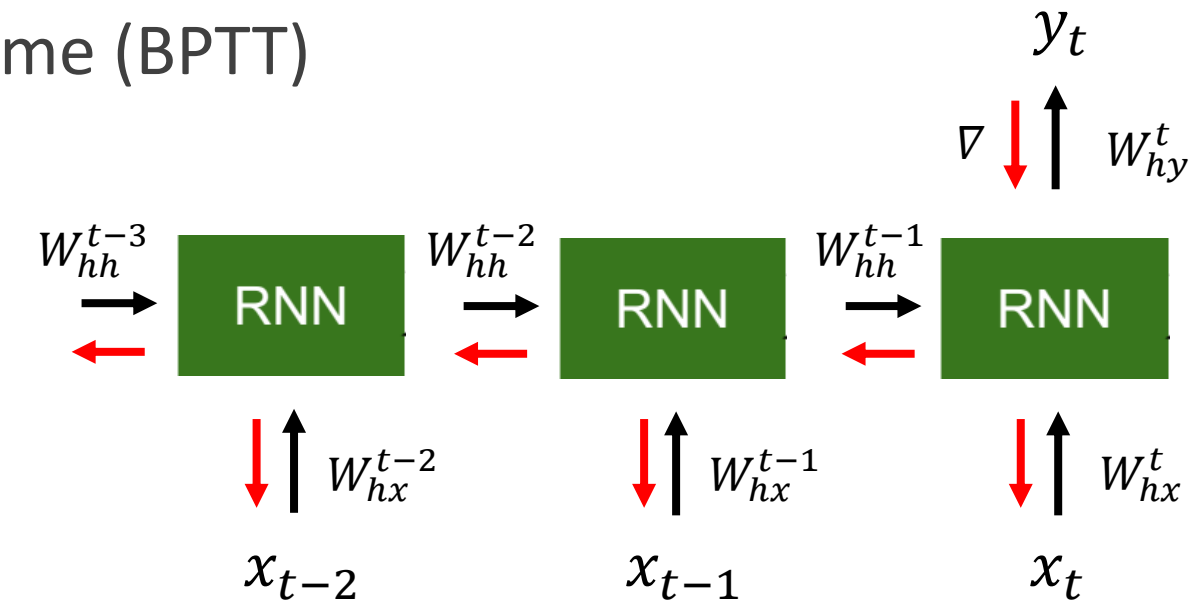
Backpropagation Through Time (BPTT)

Se o gradiente > 1:

Exploding gradients

Se o gradiente < 1:

Vanishing gradients

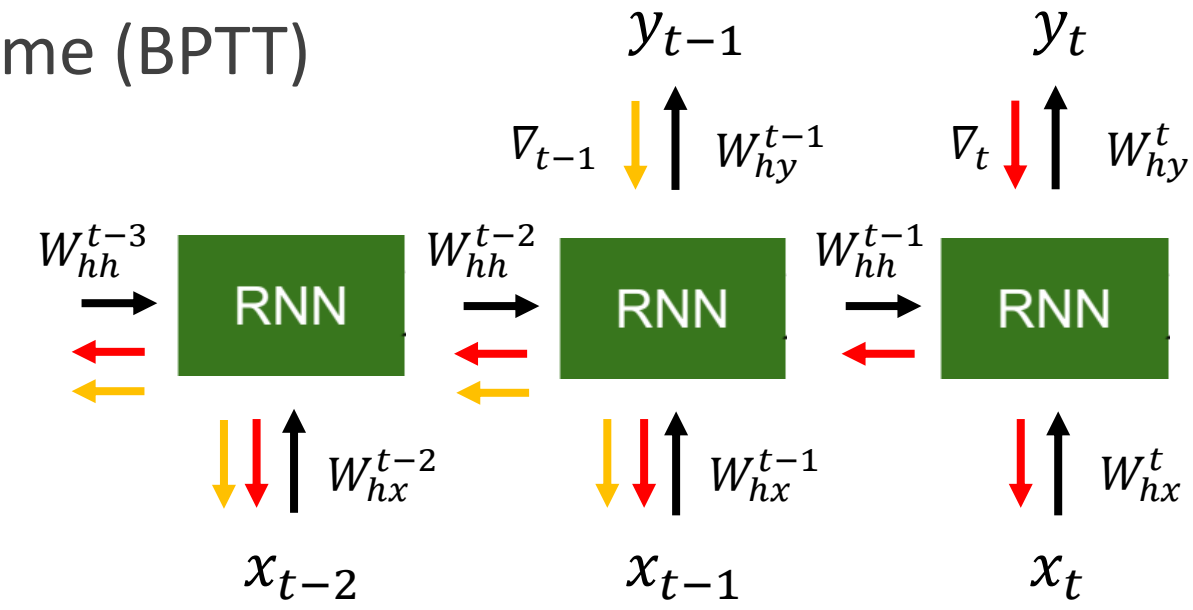


Treino - RNN

Backpropagation Through Time (BPTT)

Ainda tem onde piorar

Se tivermos **múltiplas saídas**,
Teremos **múltiplos gradientes**



RNN

Problemas:

- **Treino ruidoso** devido ao Vanishing/Exploding gradient
- Dificuldade em lidar com **dependências de longo prazo**
  - João entrou na sala. José também. Já é tarde e ambos estão atrasados. João disse oi para ____
- Dificuldade em lidar com **ruído**

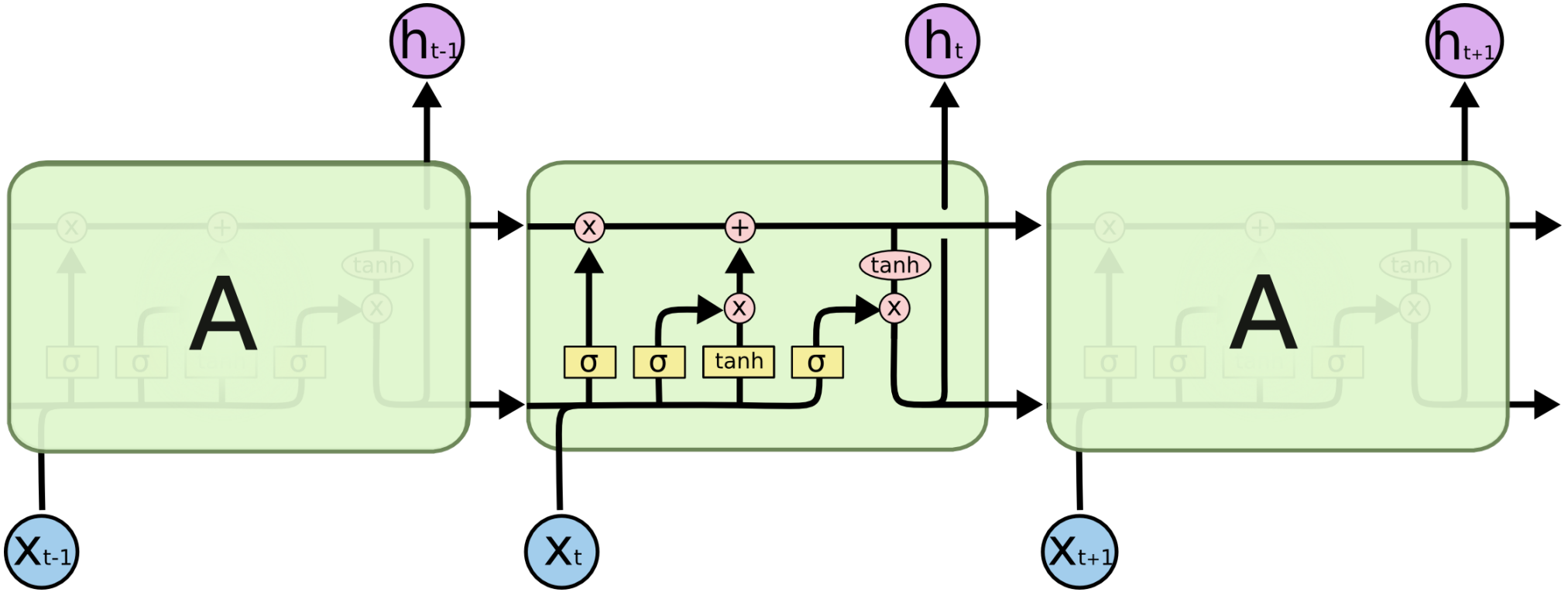
RNN

Problemas:

- **Treino ruidoso** devido ao Vanishing/Exploding gradient
- Dificuldade em lidar com **dependências de longo prazo**
  - João entrou na sala. José também. Já é tarde e ambos estão atrasados. João disse oi para ____
- Dificuldade em lidar com **ruído**
- Teoricamente funciona, mas na prática é difícil treinar para problemas complexos

LSTM

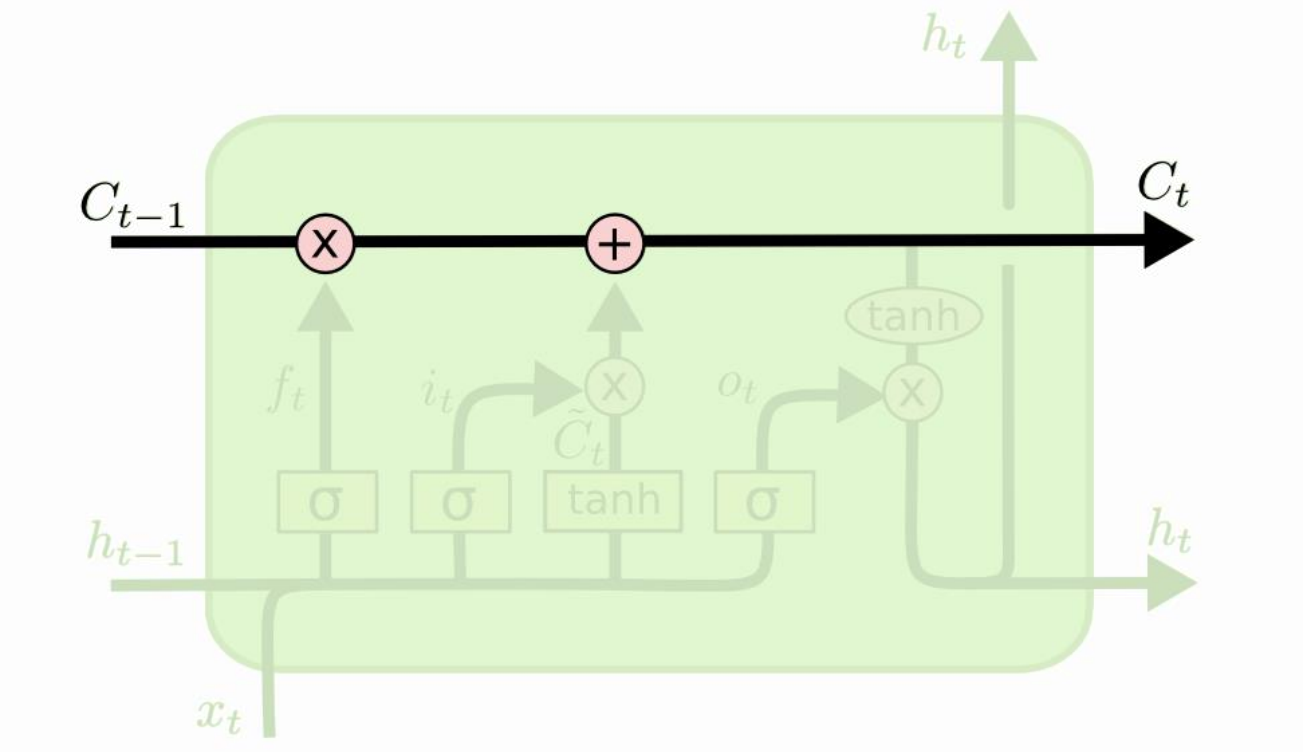
Long Short Term Memory [Hochreiter et al., 1997]



Cell state

Resolve o problema do **Vanishing/Exploding** gradiente

É mais do que uma simples conexão direta, pois as informações propagadas são ponderadas pelas entradas a cada tempo

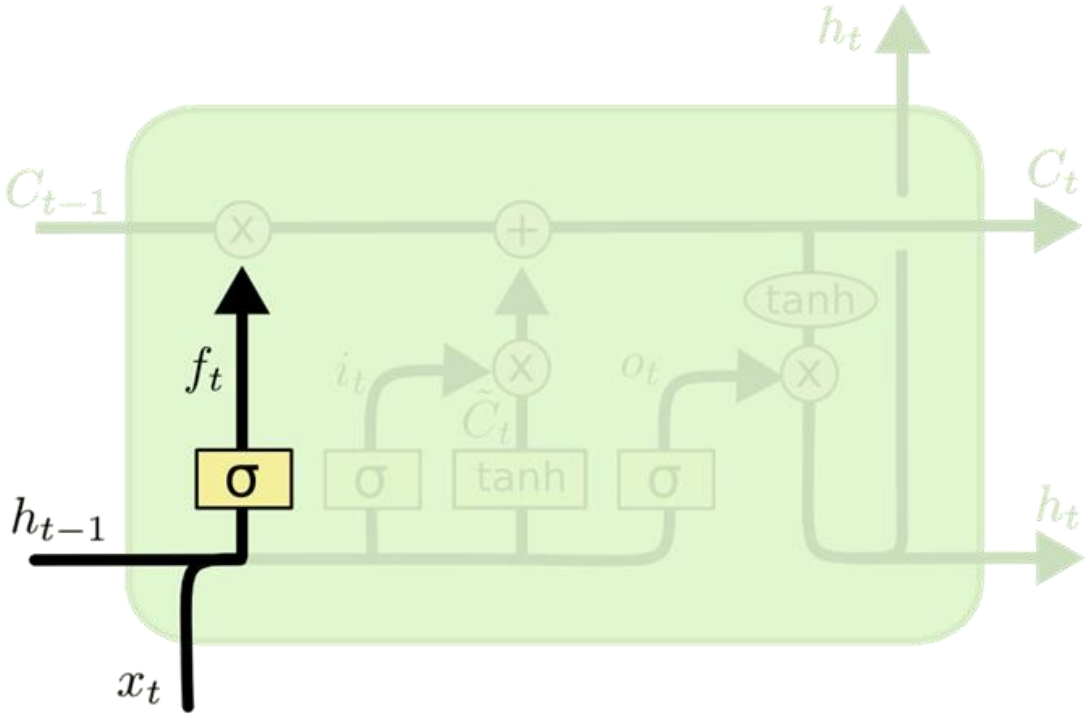


Forget Gate

Decide qual informação que vem do **estado anterior** vai ser jogada fora

1 : “**Mantenha** isso completamente”
0 : “**Esqueça** isso completamente”

Exemplo:
Ao ler um novo substantivo a rede pode **esquecer** o gênero do substantivo recebido anteriormente



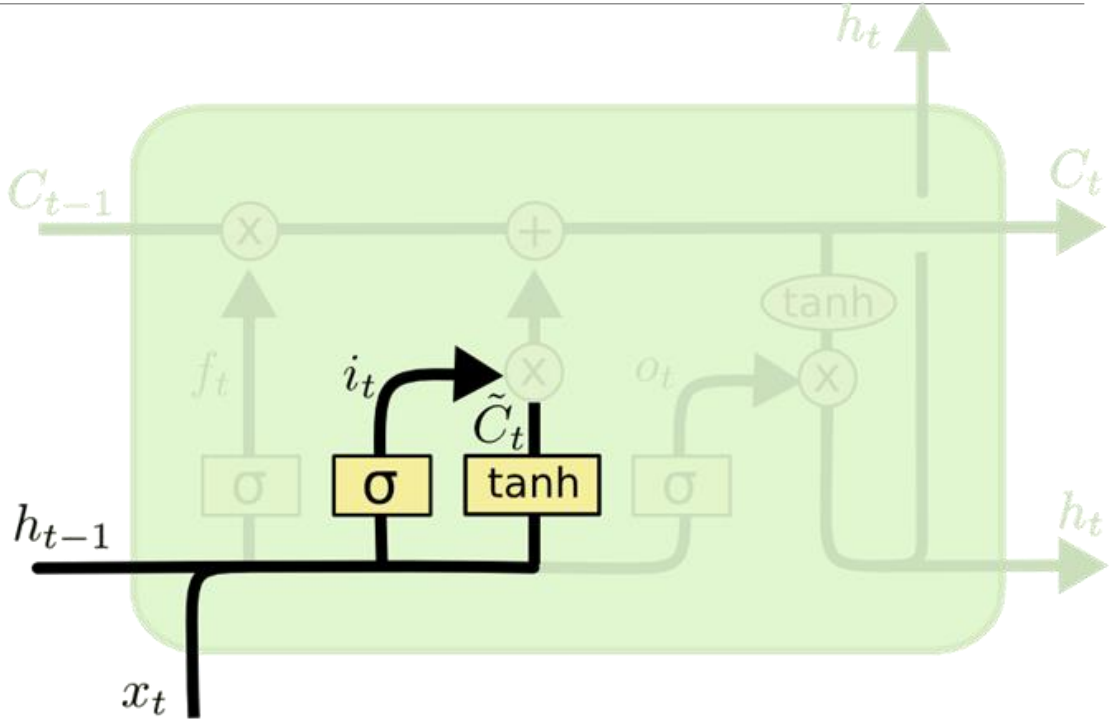
$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

Input Gate

Decide qual informação que vem da **entrada** vai ser inserida
É combinado com o **candidato** a novo C

1 : “**Insira** isso completamente”
0 : “Deixe de lado”

Exemplo:
Ao ler um novo substantivo a rede pode **memorizar** o gênero



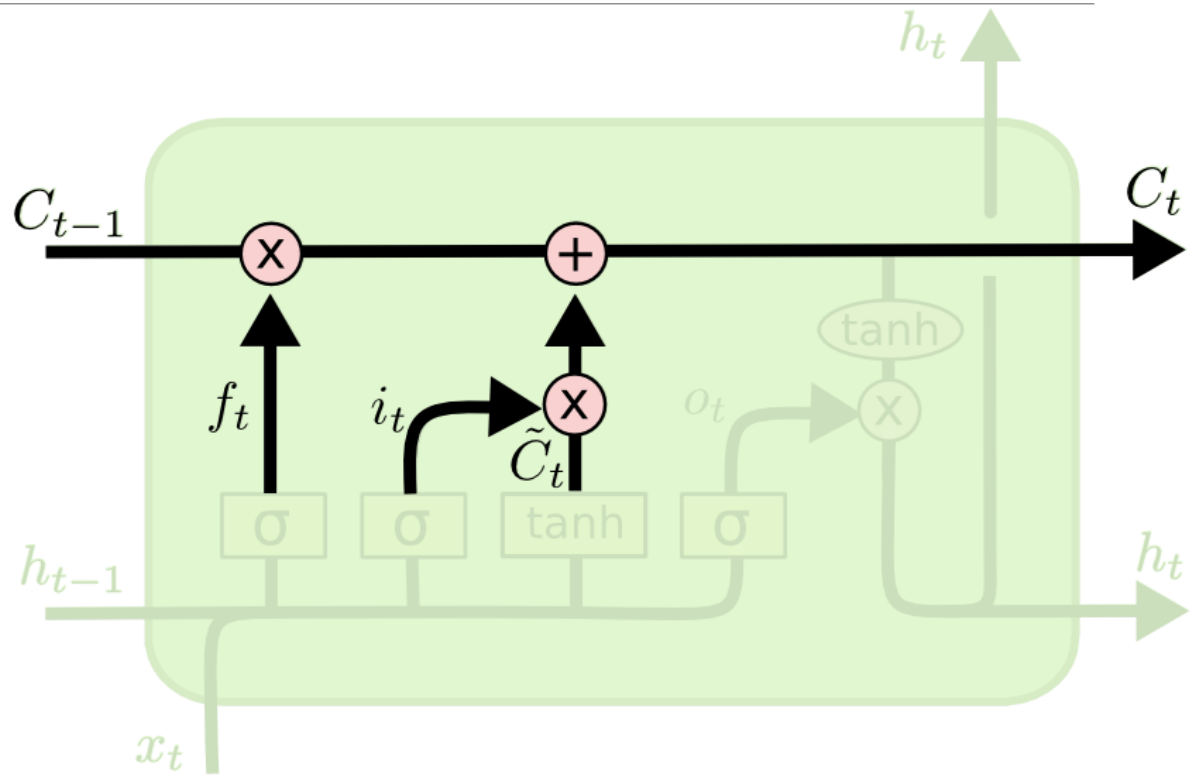
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

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

Cell Update

Multiplica-se o estado anterior por f_t, esquecendo algumas informações. Então adiciona-se $i_t * \tilde{C}_t$, que é o novo candidato **ponderado por quanto queremos lembrar**

Exemplo:
Repassa ao estado celular o novo gênero encontrado



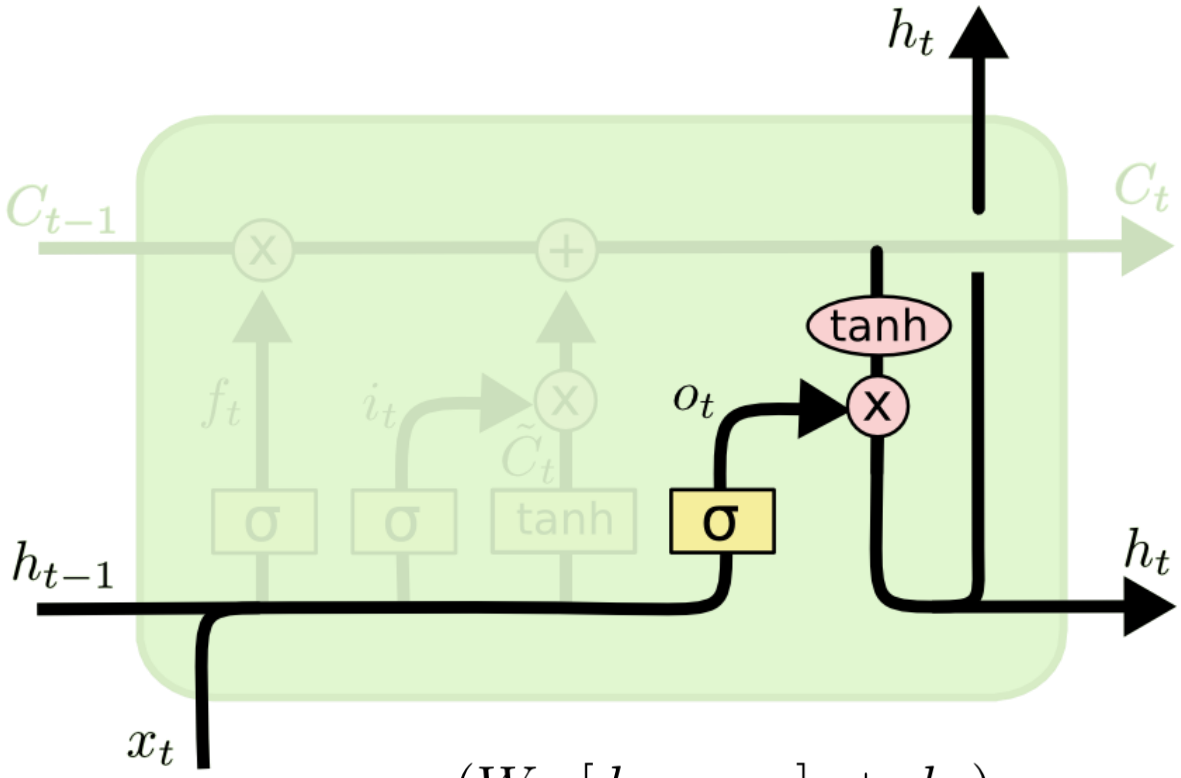
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Output Gate

Primeiro determina-se quais partes do **cell state** será enviado para a saída.

Então usa-se uma tanh para gerar saídas que serão **multiplicadas** pelo o_t

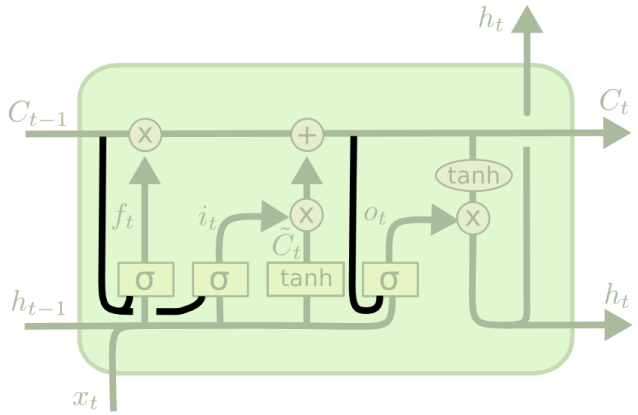
Exemplo:
Prediz que a próxima palavra irá seguir o novo gênero



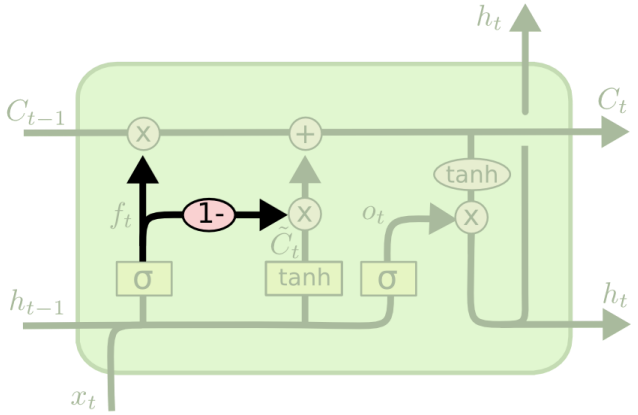
$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$

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

Variações



$$f_t = \sigma\left(W_f \cdot [\boldsymbol{C_{t-1}}, h_{t-1}, x_t] \; + \; b_f\right)$$
$$i_t = \sigma\left(W_i \cdot [\boldsymbol{C_{t-1}}, h_{t-1}, x_t] \; + \; b_i\right)$$
$$o_t = \sigma\left(W_o \cdot [\boldsymbol{C_t}, h_{t-1}, x_t] \; + \; b_o\right)$$



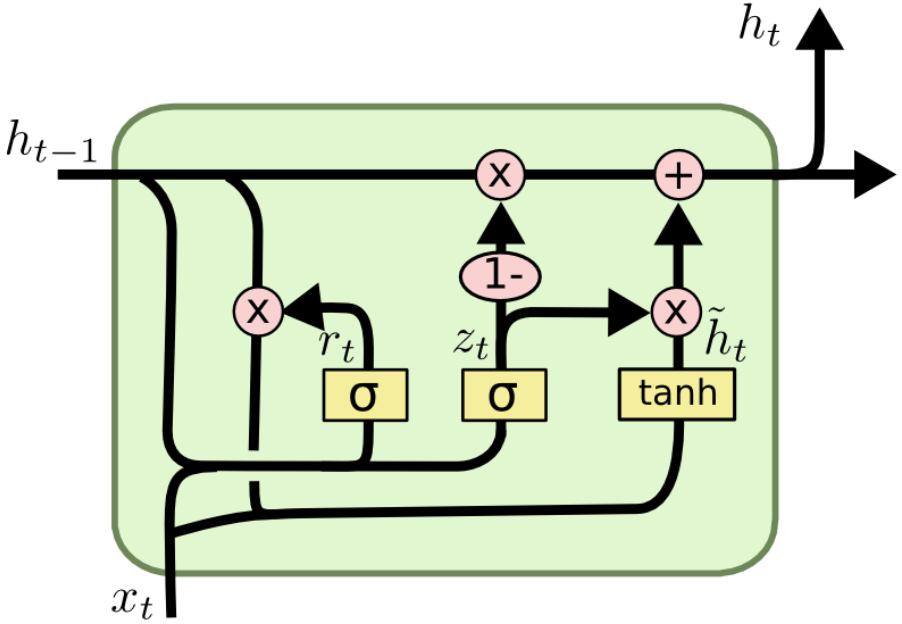
$$C_t = f_t * C_{t-1} + (\mathbf{1} - \boldsymbol{f_t}) * \tilde{C}_t$$

GRU

Gated Recurrent Unit [Cho et al., 2014]

Combina o *input* e *forget gate* no *update gate*

Não possui cell state



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

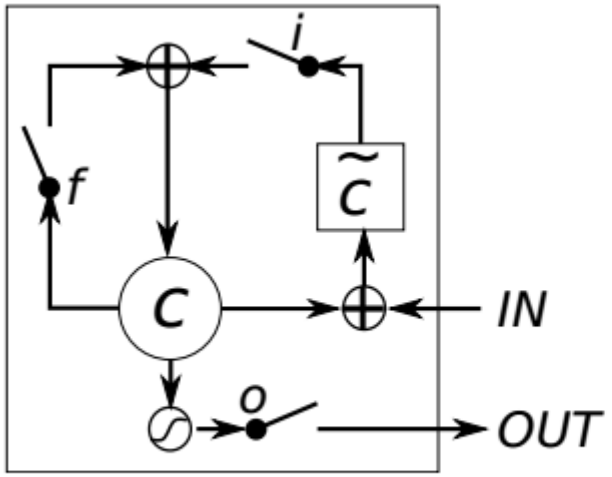
LSTM vs GRU

LSTM
- Mais parâmetros
- Maior custo
- Treino mais “complicado”
- Maior capacidade de combinar informações de formas diferentes

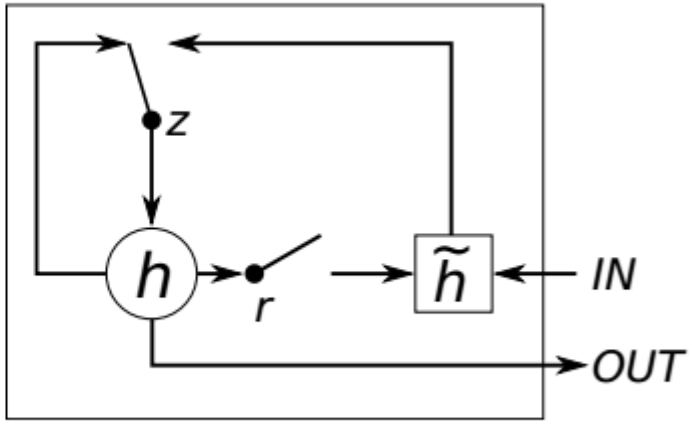
GRU
- Menos parâmetros
- Treino mais fácil
- Apresenta desempenho semelhante a LSTM em várias tarefas

Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling

Junyoung Chung **Caglar Gulcehre** **KyungHyun Cho** **Yoshua Bengio**
Université de Montréal
Université de Montréal
CIFAR Senior Fellow



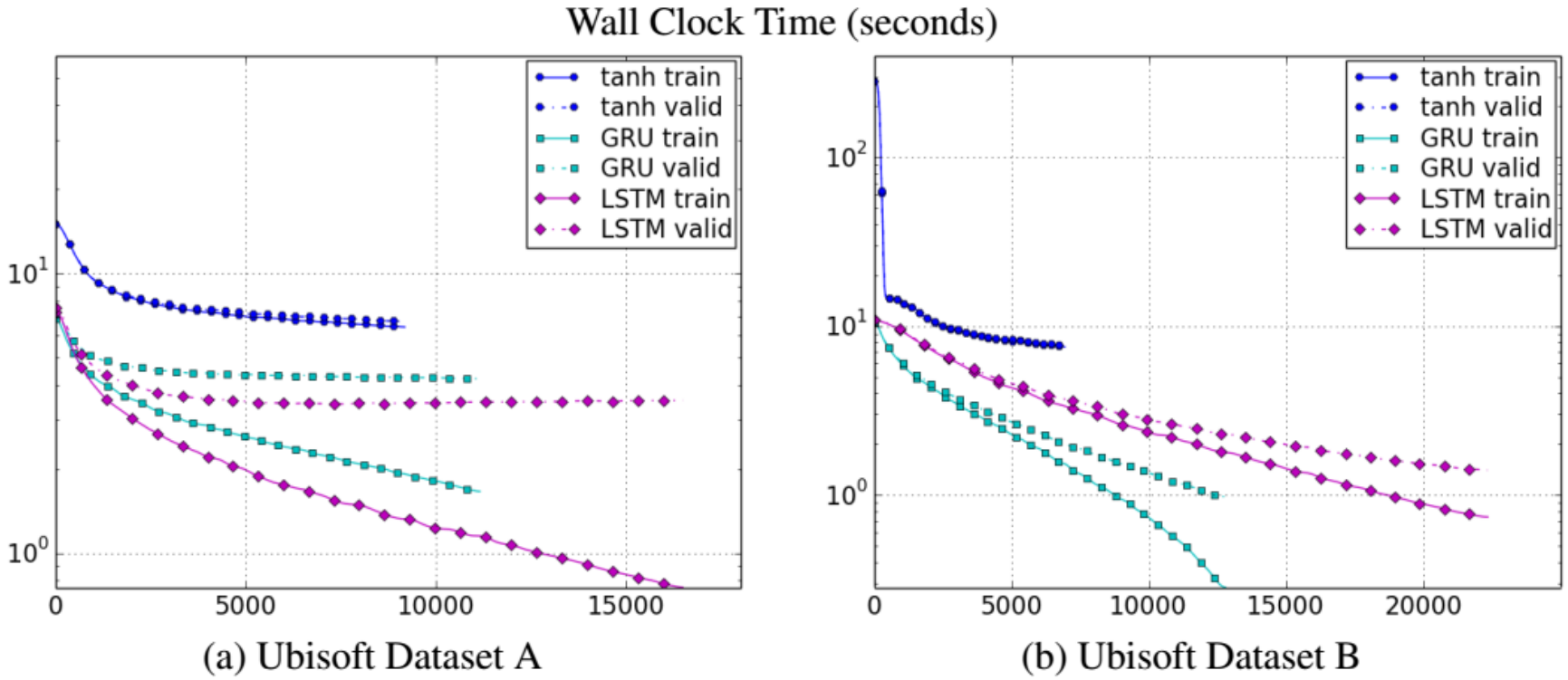
(a) Long Short-Term Memory



(b) Gated Recurrent Unit

LSTM vs GRU

Em problemas de áudio



(a) Ubisoft Dataset A

(b) Ubisoft Dataset B

RNN vs GRU vs LSTM

Dataset:
- Ondas senoidais com diferentes frequências
- Entrada:
  - 100 recortes da função
- Saída:
  - 100 recortes seguintes

Teste 1:
- Ondas perfeitas

Teste 2:
- Ondas corrompidas no conjunto de treino

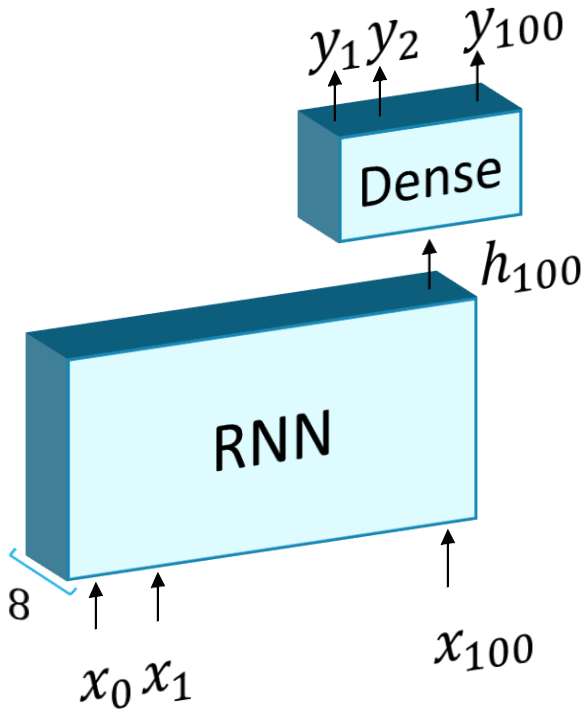
200 épocas

Estado h com 8 neurônios

Camada dense com 100 neurônios de saída

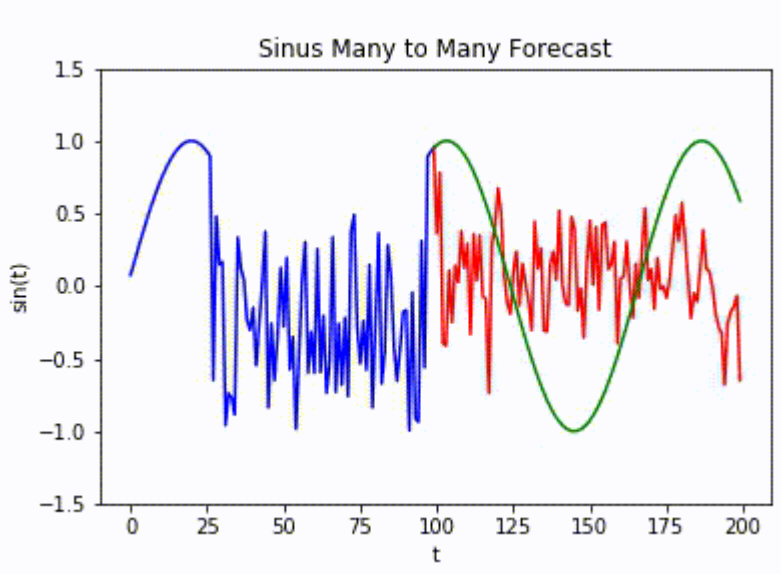
Nº de Parâmetros:
- Camada Dense: 900
- RNN: 80
- LSTM: 320
- GRU: 240

Baseline:
- MLP 100 x 100
- 20.200 parâmetros

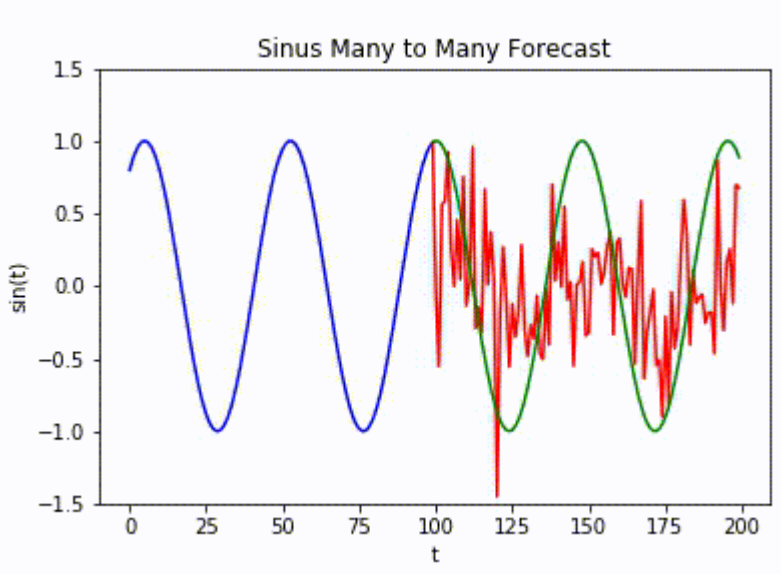


MLP

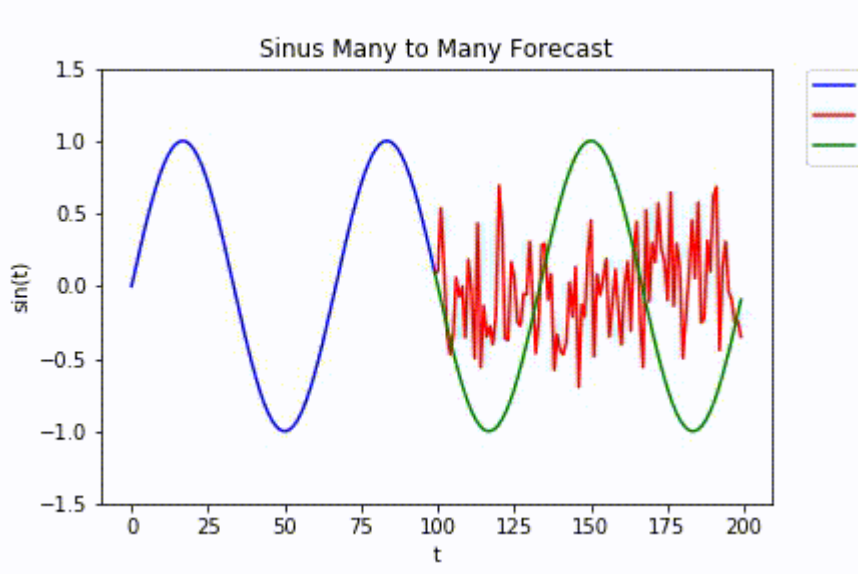
Treino



Teste

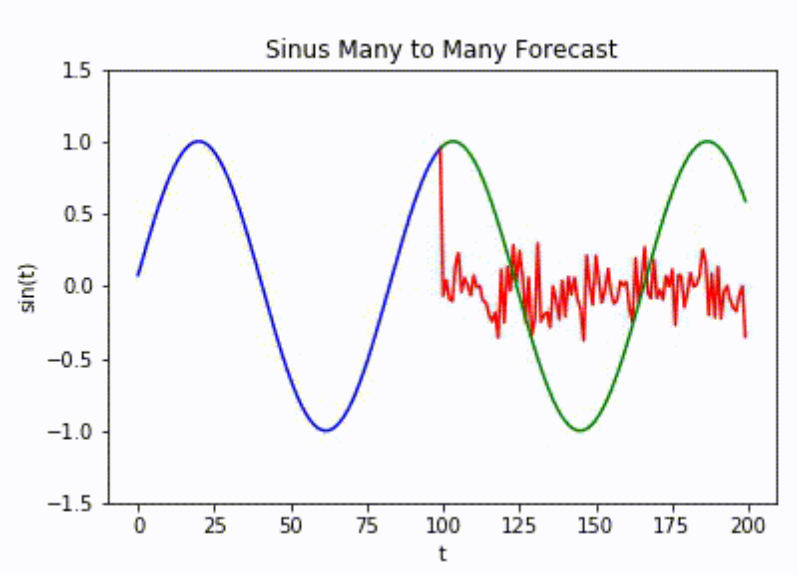


Teste2

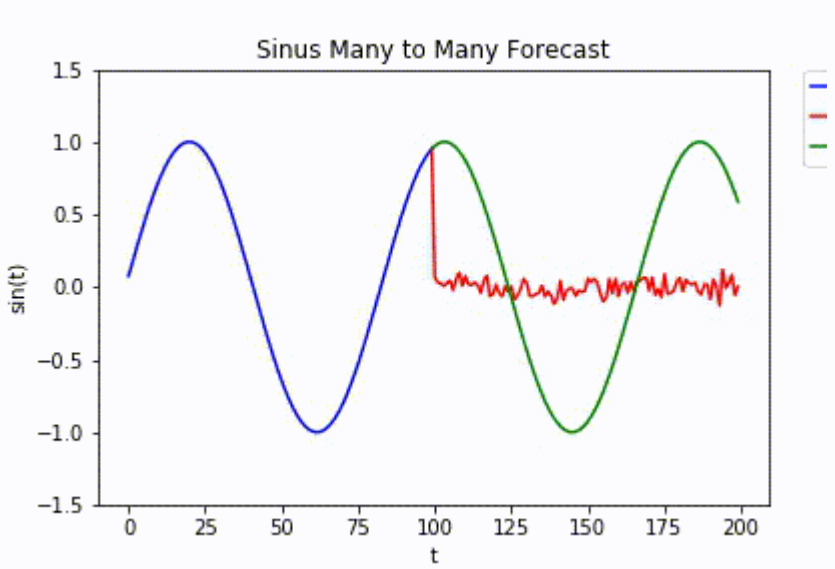


Treino

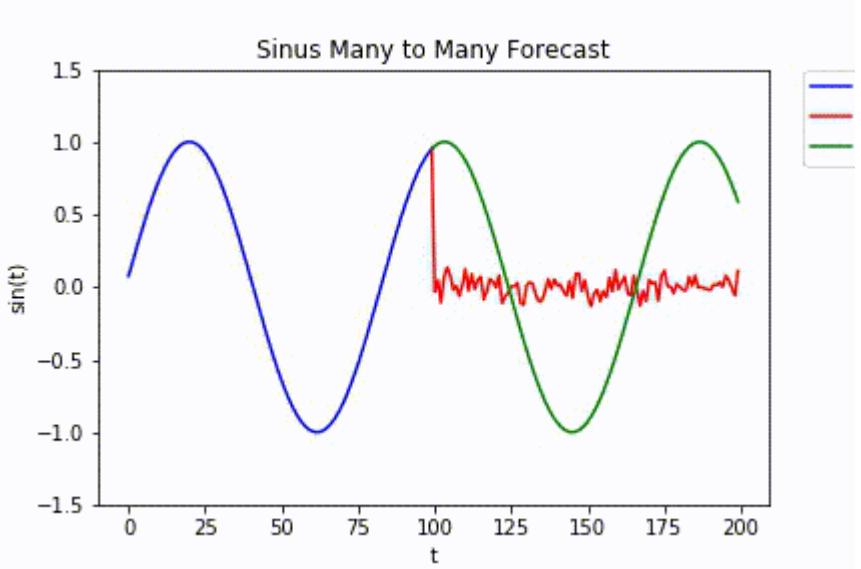
RNN



LSTM

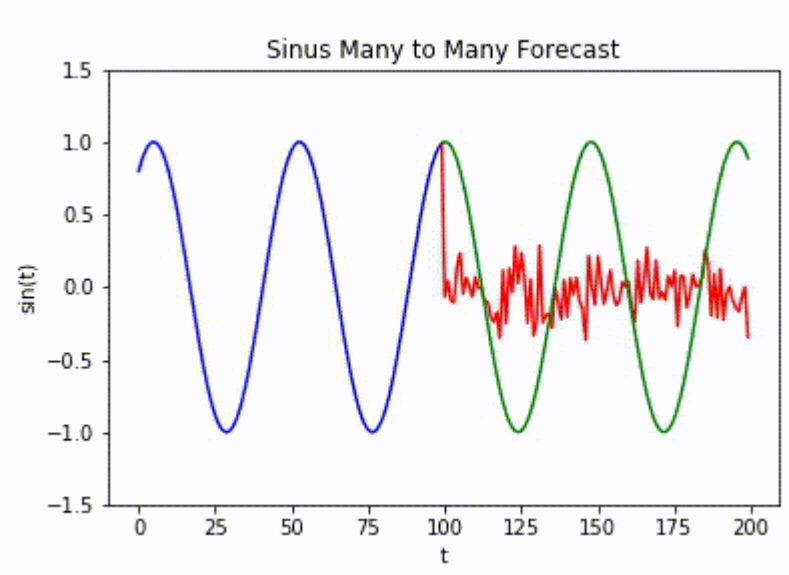


GRU

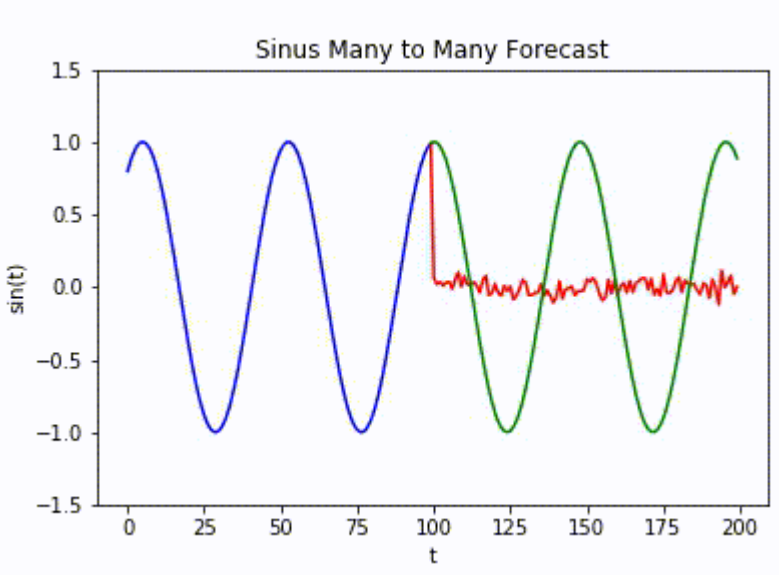


Teste

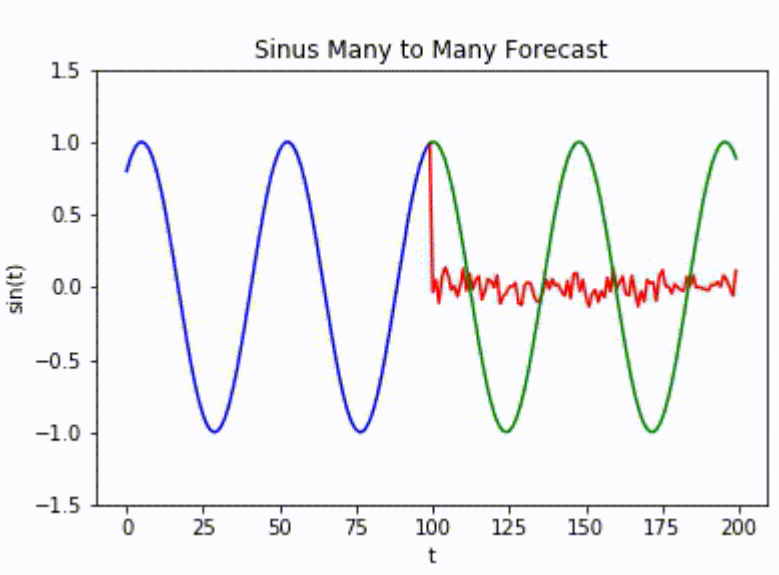
RNN



LSTM

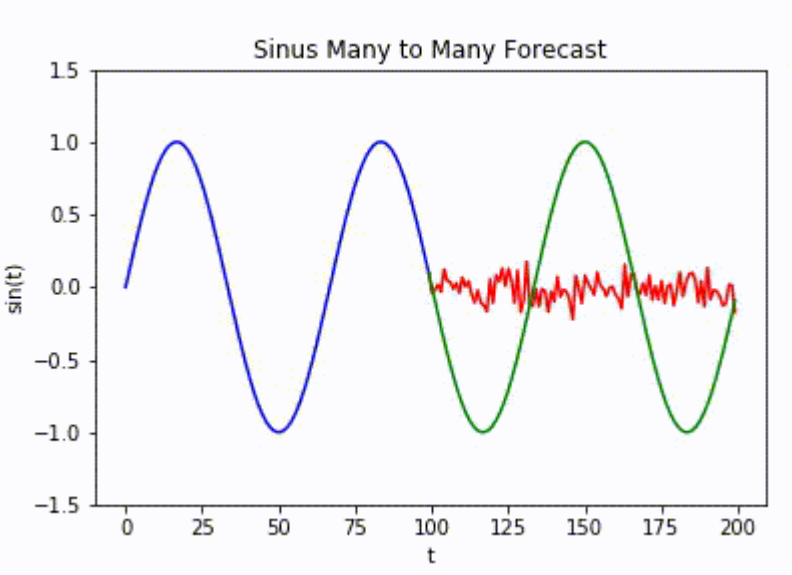


GRU

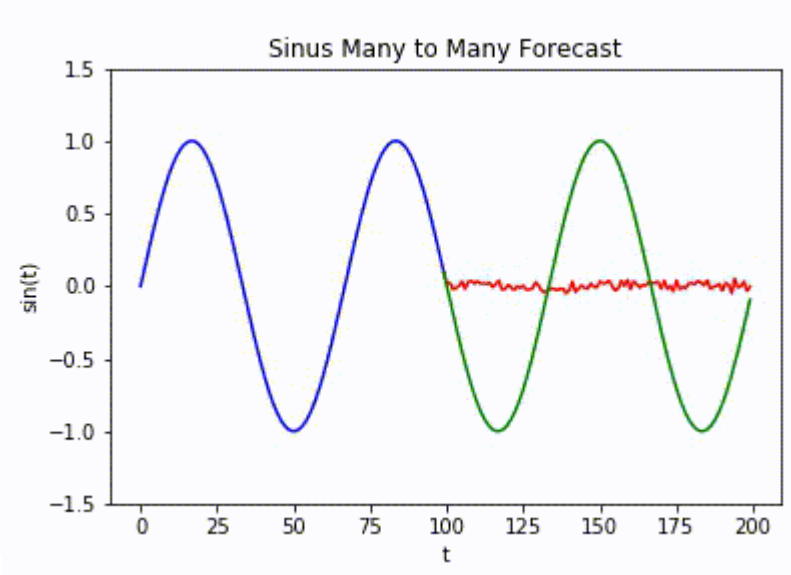


Teste2

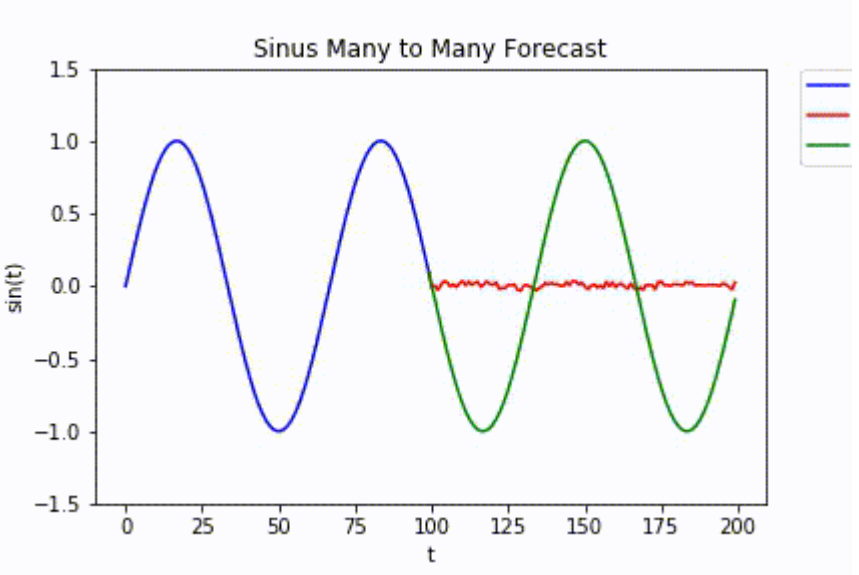
RNN



LSTM

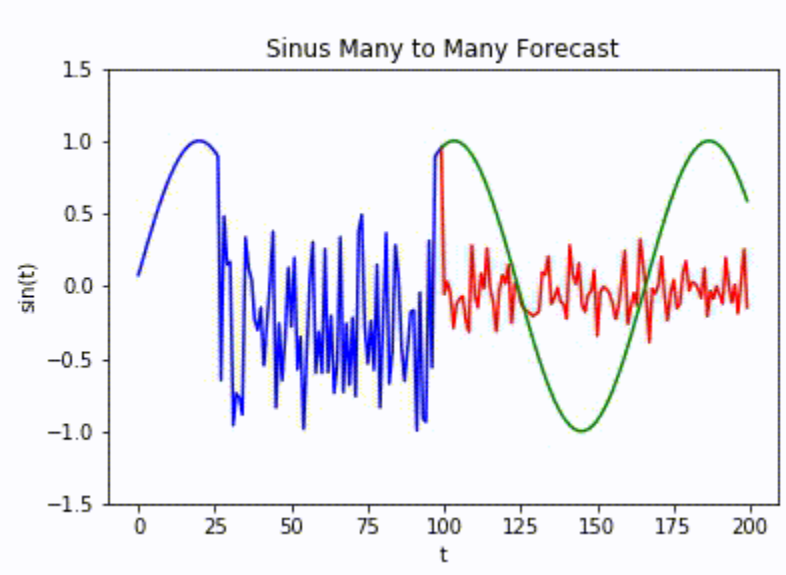


GRU

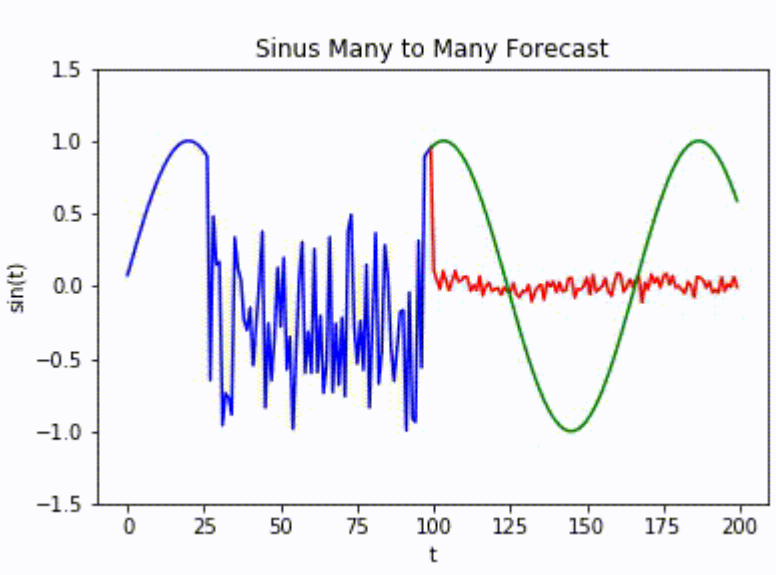


Treino com ruído

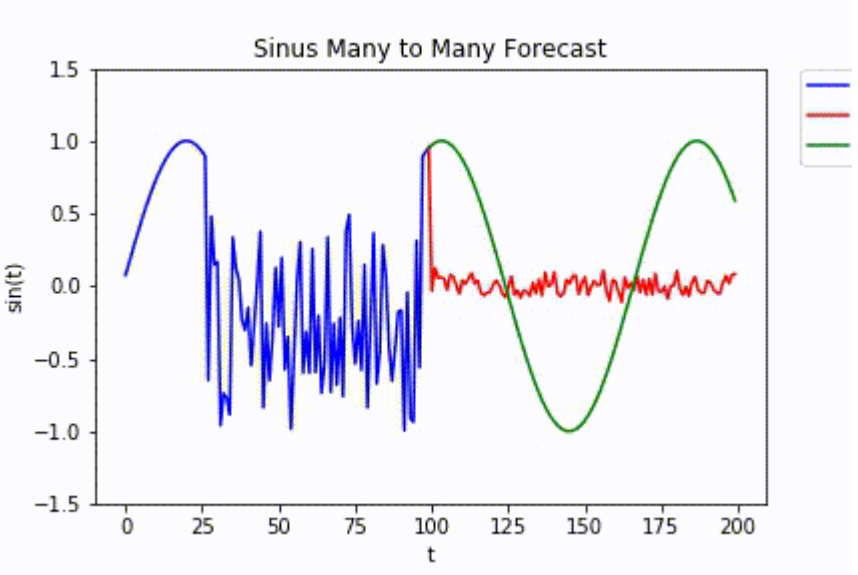
RNN



LSTM

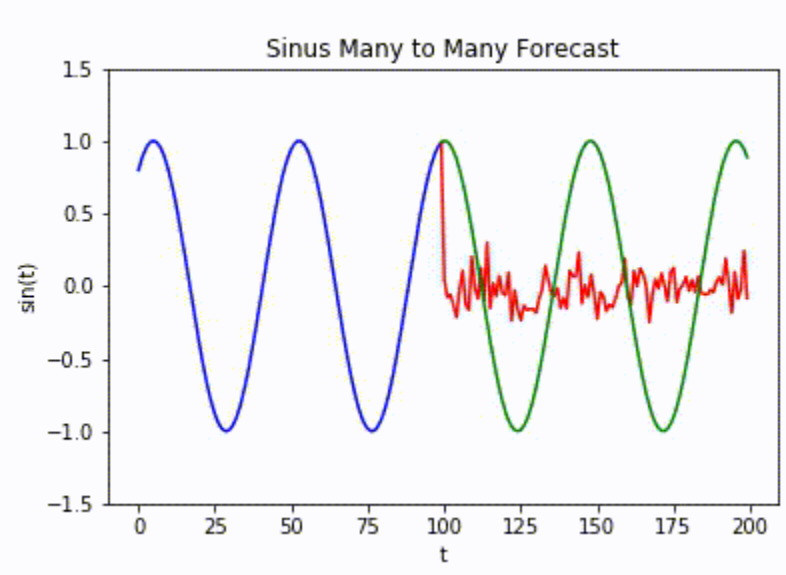


GRU

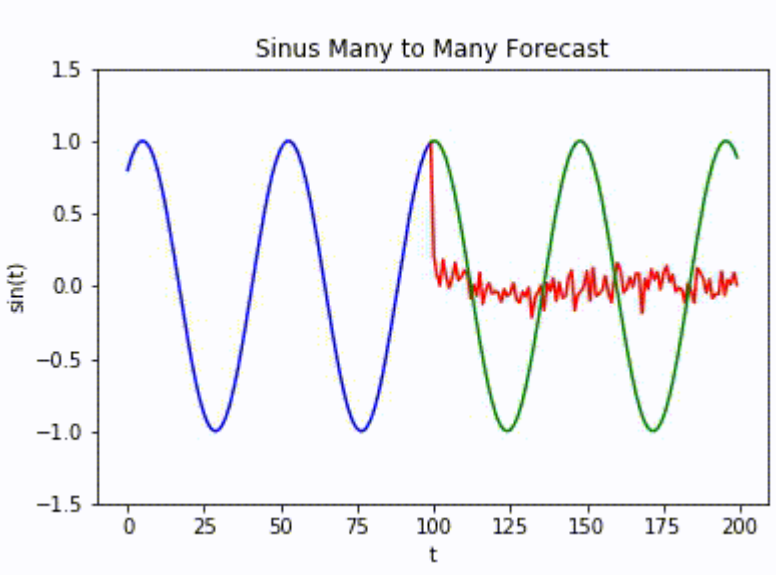


Teste com ruído

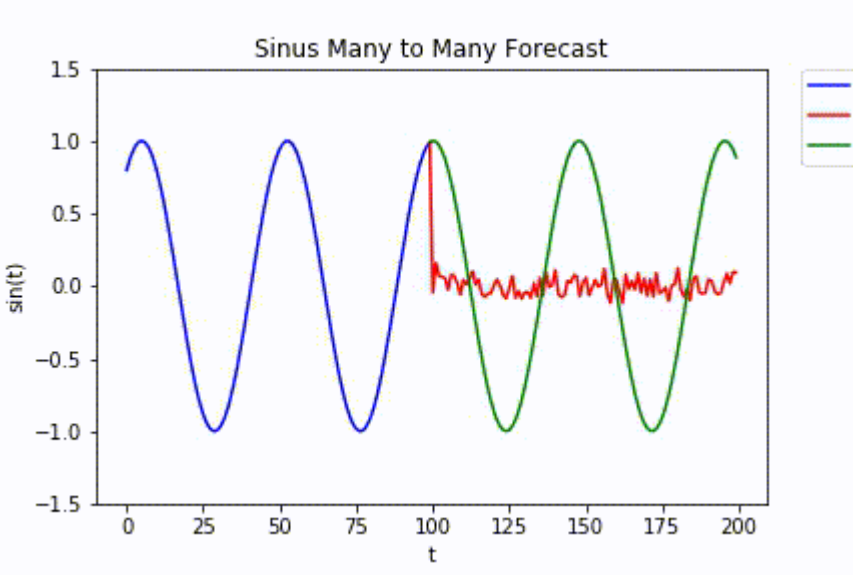
RNN



LSTM

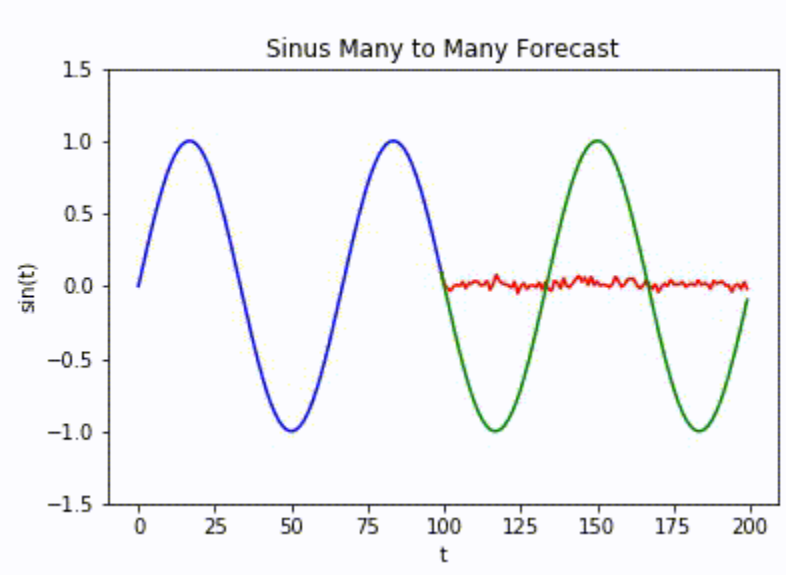


GRU

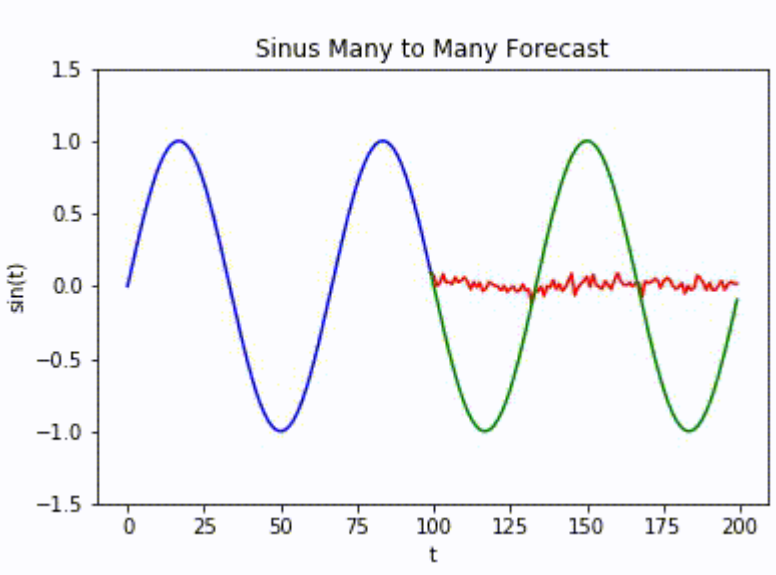


Teste2 com ruído

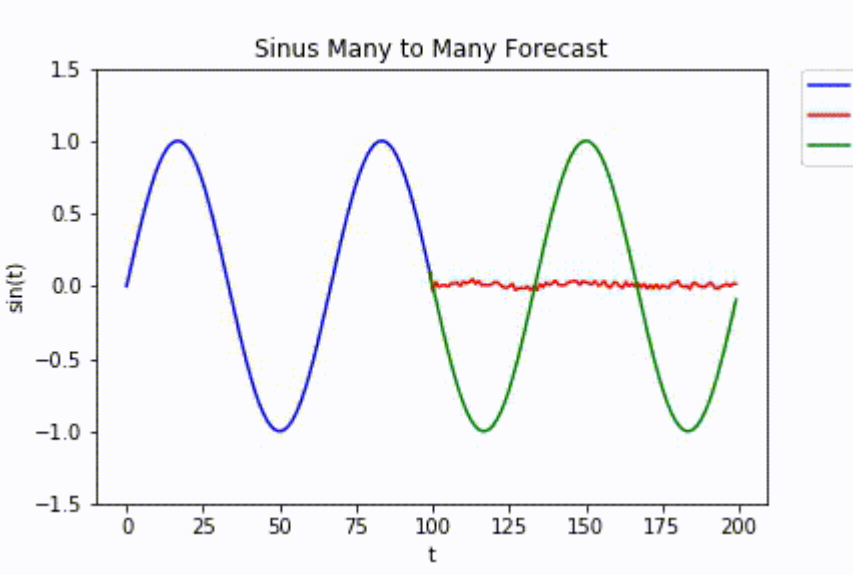
RNN



LSTM

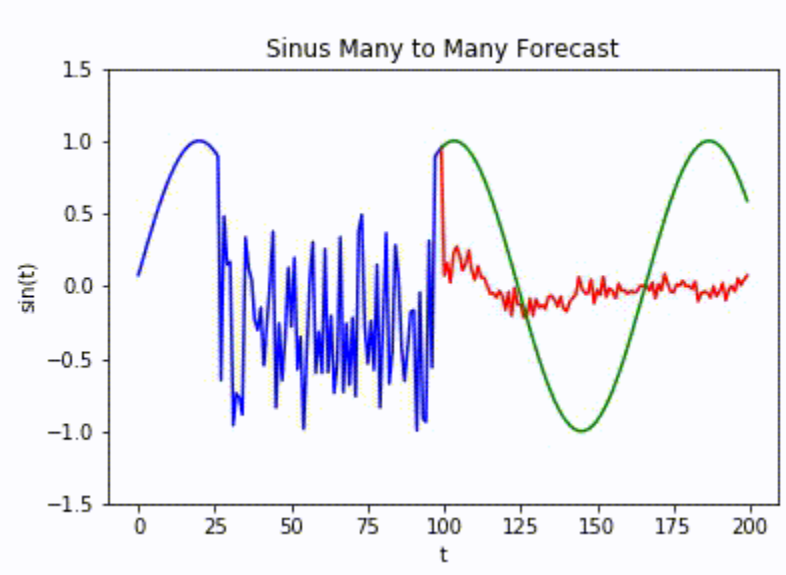


GRU

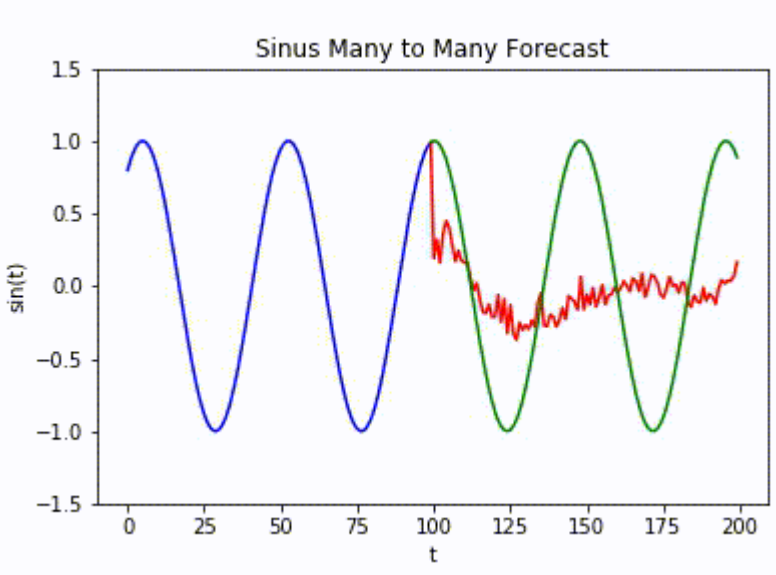


GRU $h = 64$

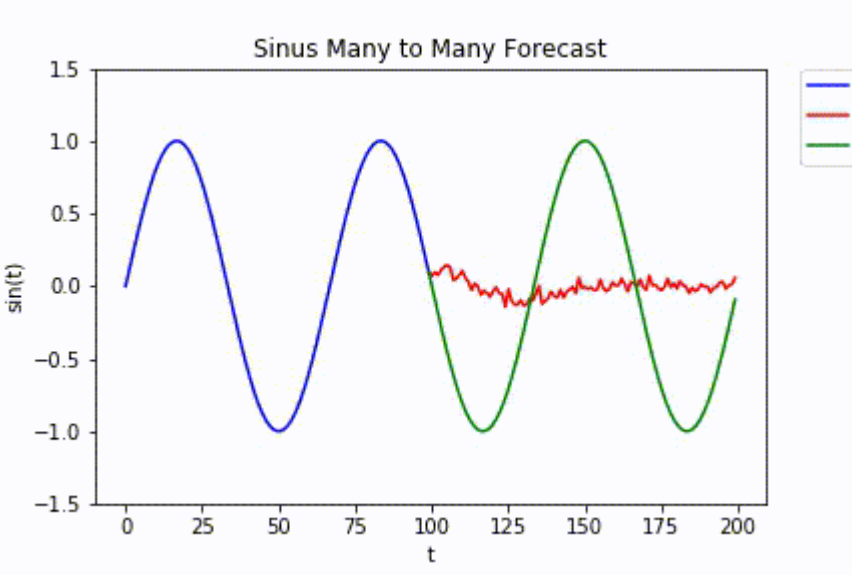
RNN



LSTM



GRU



Implementação

Keras: RNN, LSTM e GRU

```
model = Sequential()
model.add(Embedding(max_features, output_dim=256))
model.add(LSTM(128))
model.add(Dropout(0.5))
model.add(Dense(1, activation='sigmoid'))
```

Recurrent Shop:

https://github.com/farizrahman4u/recurrentshop

Dicas de treino

Tente diferentes otimizadores

Execute o treino muitas vezes

Cuidado com o tamanho do h.

- Crescimento quadrático de parâmetros.
- RNN: $h(x+1)+h^2$
- LSTM: $4(h(x+1)+h^2)$
- GRU: $3(h(x+1)+h^2)$

Cuidado com o número de tempos t

Arquiteturas Recorrentes

Many-to-many

- Seq2seq
- Seq2seq com atenção
- Trabalhos recentes

Seq2seq – Cho, 2014

Formula a ideia de **Encoder-Decoder** recorrente para modelar relação entre sequências.

Inicialmente proposto para tradução

Artigo também introduz a **GRU**

Pode ser usado em:

- Tradução
- Predição
- Q&A (Question & Anwser)
- Resumo de texto
- Image captioning

Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation

Kyunghyun Cho
Bart van Merriënboer **Caglar Gulcehre** **Dzmitry Bahdanau**
Université de Montréal Jacobs University, Germany
firstname.lastname@umontreal.ca d.bahdanau@jacobs-university.de

Fethi Bougares **Holger Schwenk** **Yoshua Bengio**
Université du Maine, France Université de Montréal, CIFAR Senior Fellow
firstname.lastname@lium.univ-lemans.fr find.me@on.the.web

Seq2seq

Na mesma época houve outro artigo propondo uma ideia semelhante

Usa **LSTM**

Propõe a inversão das sequências de saída

Sequence to Sequence Learning with Neural Networks

Ilya Sutskever
Google
ilyasu@google.com

Oriol Vinyals
Google
vinyals@google.com

Quoc V. Le
Google
qvl@google.com

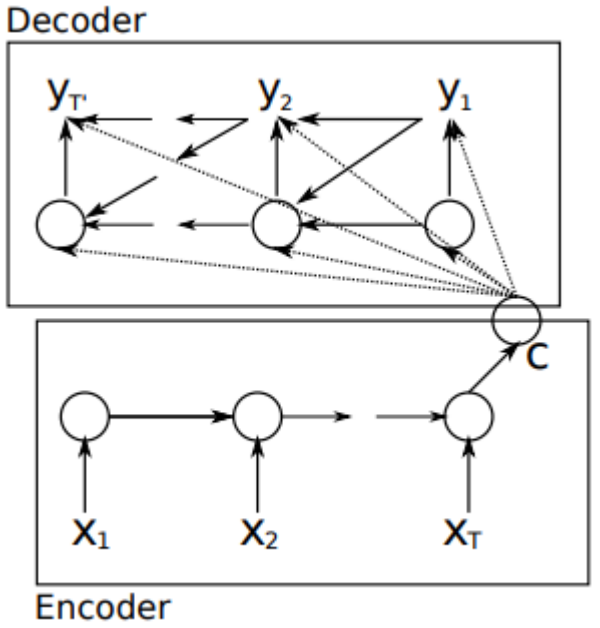
Seq2seq

Parte da ideia de que o Encoder irá **comprimir** a sequência de entrada em seu estado h que se torna um condicional c

O Decoder, que é outra rede, é treinada para construir a sequência de saída condicionado a c

- O estado s_t do Decoder é computado por:

$$s_t = f(h_{t-1}, y_{t-1}, c)$$



Ambas as sequências podem ser de tamanho variável

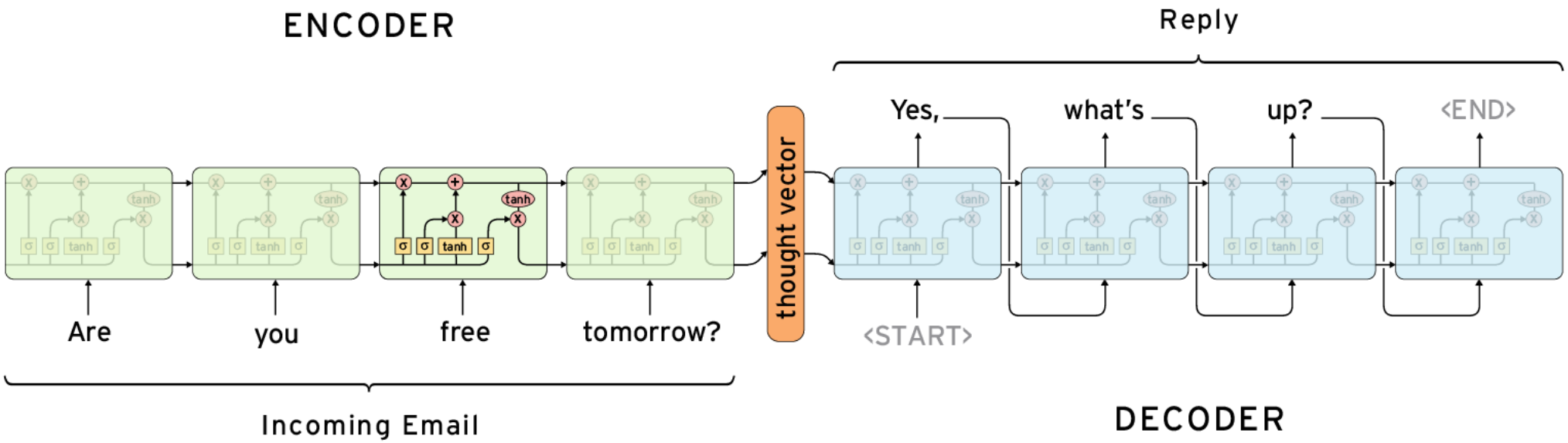
Podemos ter várias camadas

Podemos usar embeddings

Seq2seq

Teacher Forcing

- Treina o decoder com a saída esperada em cada tempo



http://suriyadeepan.github.io/2016-12-31-practical-seq2seq/

Seq2seq

Problemas

- Dificuldade com sequências longas
- Fortemente dependente do tamanho do estado do encoder

Seq2seq com atenção

Propõe um **mecanismo de atenção** para transmitir os estados do Encoder

Permite que o Decoder "**preste atenção**" em partes mais importantes da entrada

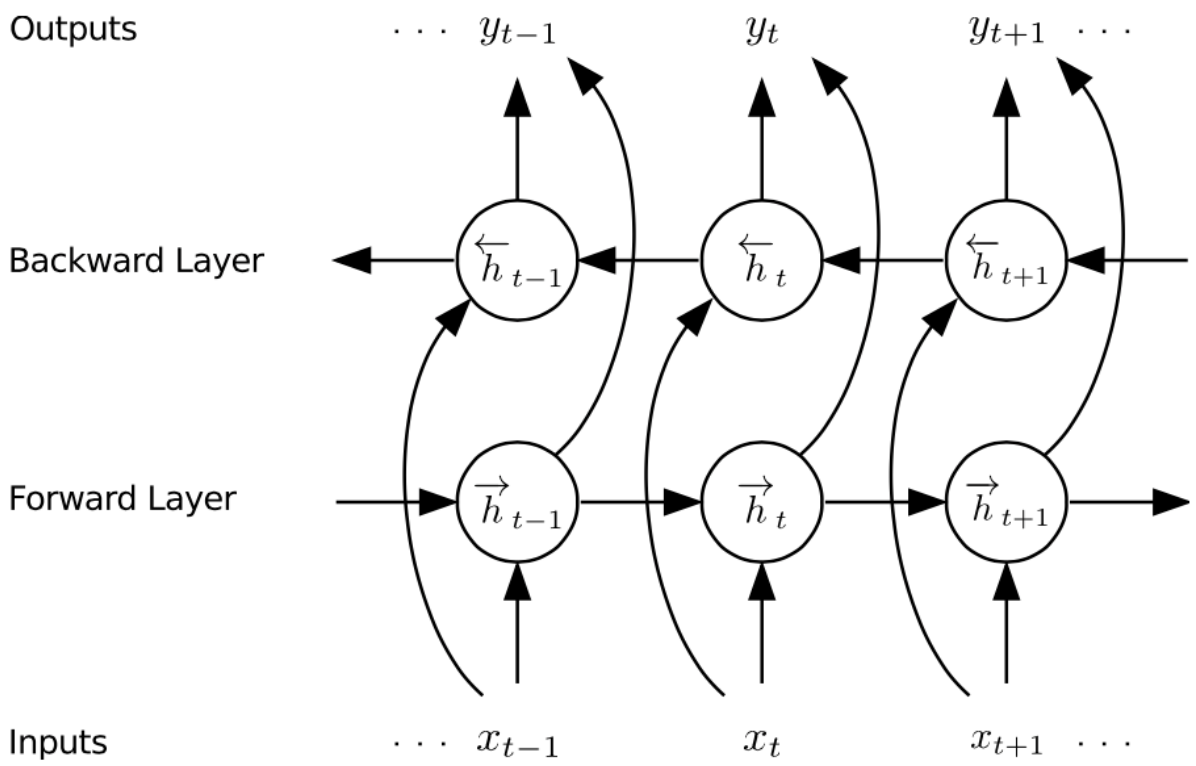
Estado da arte para mapear sequências

NEURAL MACHINE TRANSLATION
BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

Dzmitry Bahdanau
Jacobs University Bremen, Germany

KyungHyun Cho **Yoshua Bengio***
Université de Montréal

Bidirectional LSTM - BiLSTM



Seq2seq com atenção

Usa a **média ponderada** dos estados do Encoder

$s_t = f(s_{t-1}, y_{t-1}, c_t)$

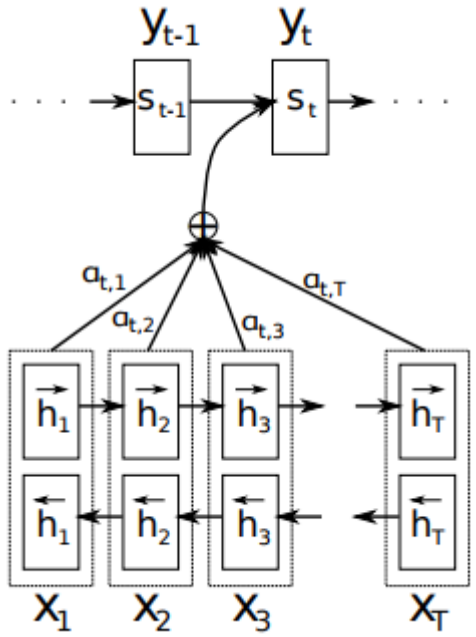
$c_i = \sum_{j=1}^{T} \alpha_{ij} h_j$

$\alpha_{ij} = softmax(e_{ij})$

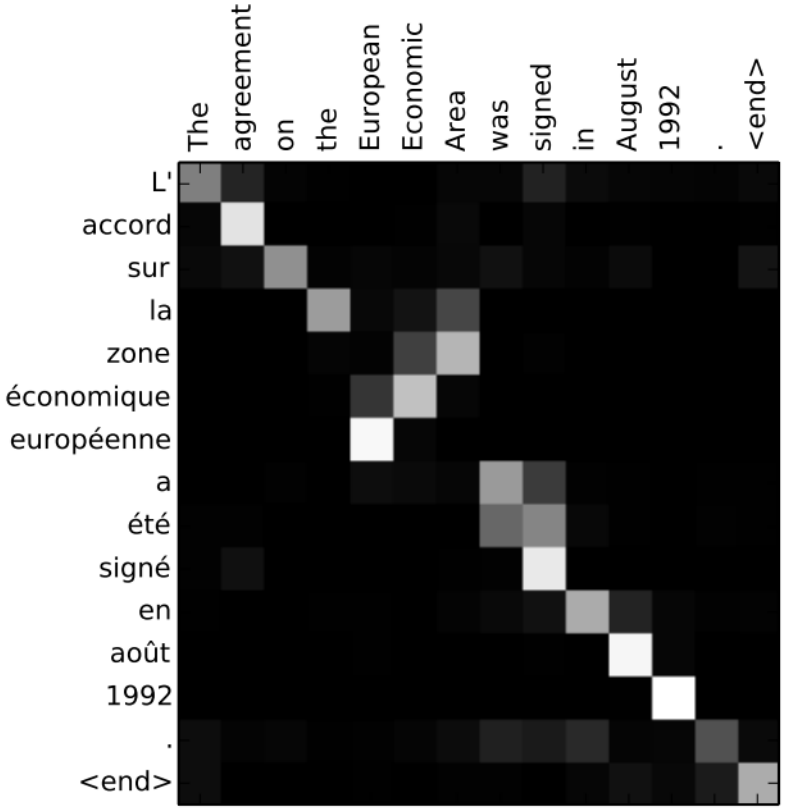
$e_{ij} = f_{att}(s_{i-1}, h_j)$ – modelo de alinhamento = MLP

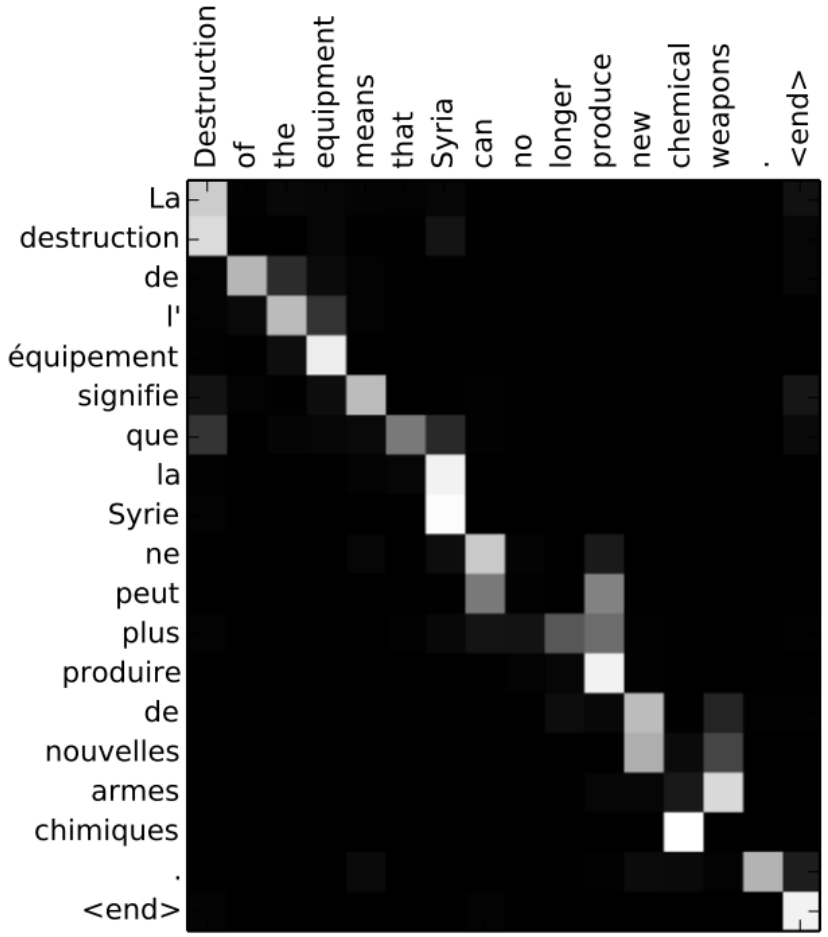
A energia e reflete a importância da anotação do estado h_j com respeito ao estado anterior do Decoder s_{i-1}em decidir o próximo estado s_i e a próxima saída y_i

Assim o Decoder decide em qual parte do input deve prestar atenção



Visualização do alinhamento





Seq2seq – Evolução GNMT - 2016

Conexões residuais

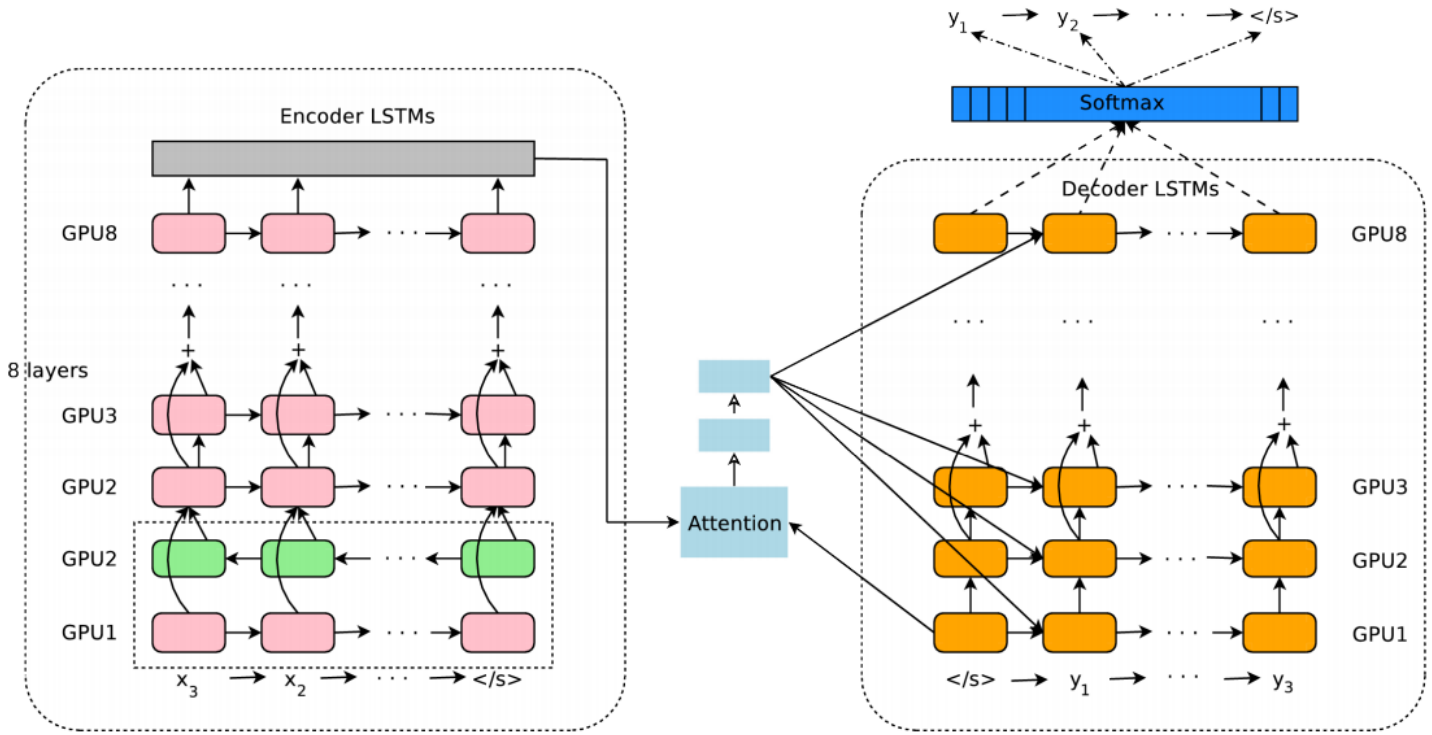
Quebras as palavras em “subpalavras”

Várias otimizações

Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi
yonghui,schuster,zhifengc,qvl,mnorouzi@google.com

Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean



Seq2seq – Evolução GNMT - 2017

Mesmo modelo

Treino realizado com múltiplos idiomas

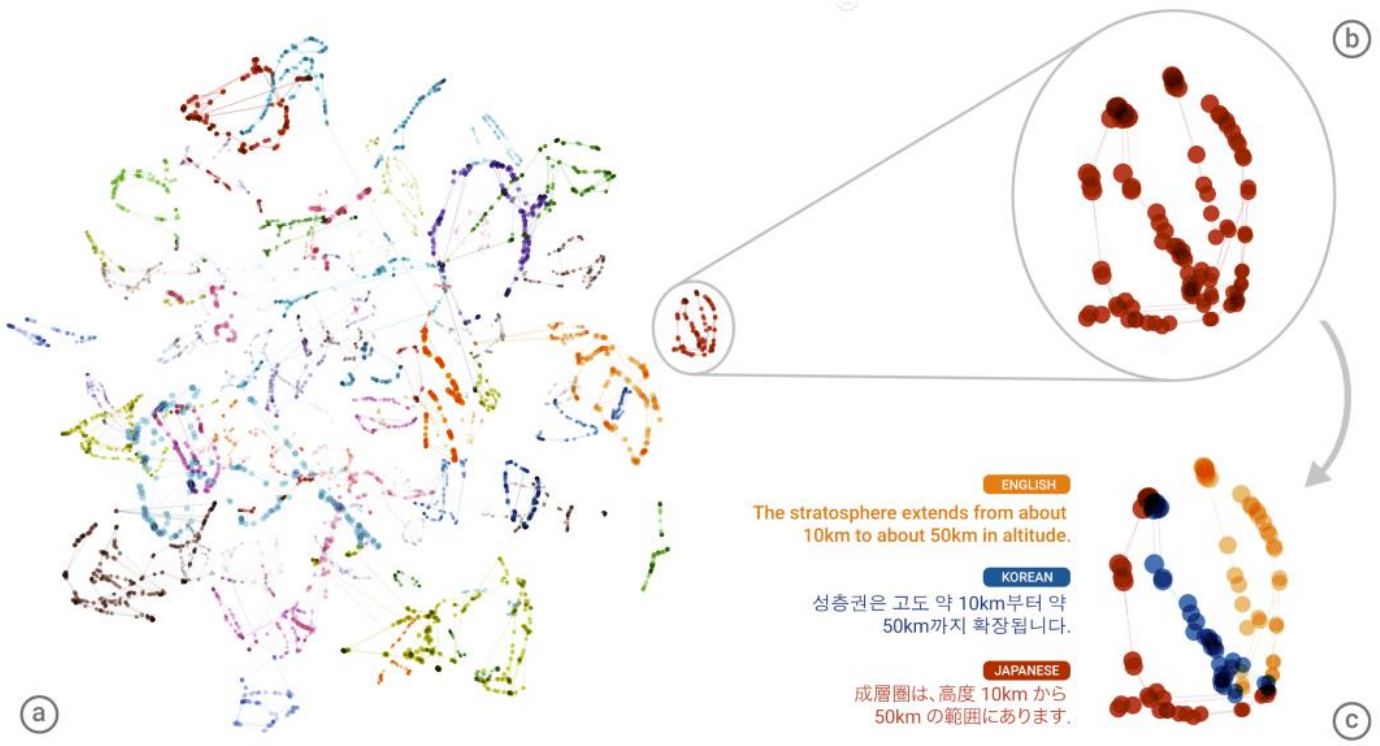
Tonkens indicam qual deve ser o idioma de saída

Sugere evidências de uma “Interlingua”

Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat
melvinp,schuster,qvl,krikun,yonghui,zhifengc,nsthorat@google.com

Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, Jeffrey Dean



Memory Networks - 2015

Propõe a ideia de armazenar os inputs para uso posterior

Tentativa de inserir memória em problemas de Q&A

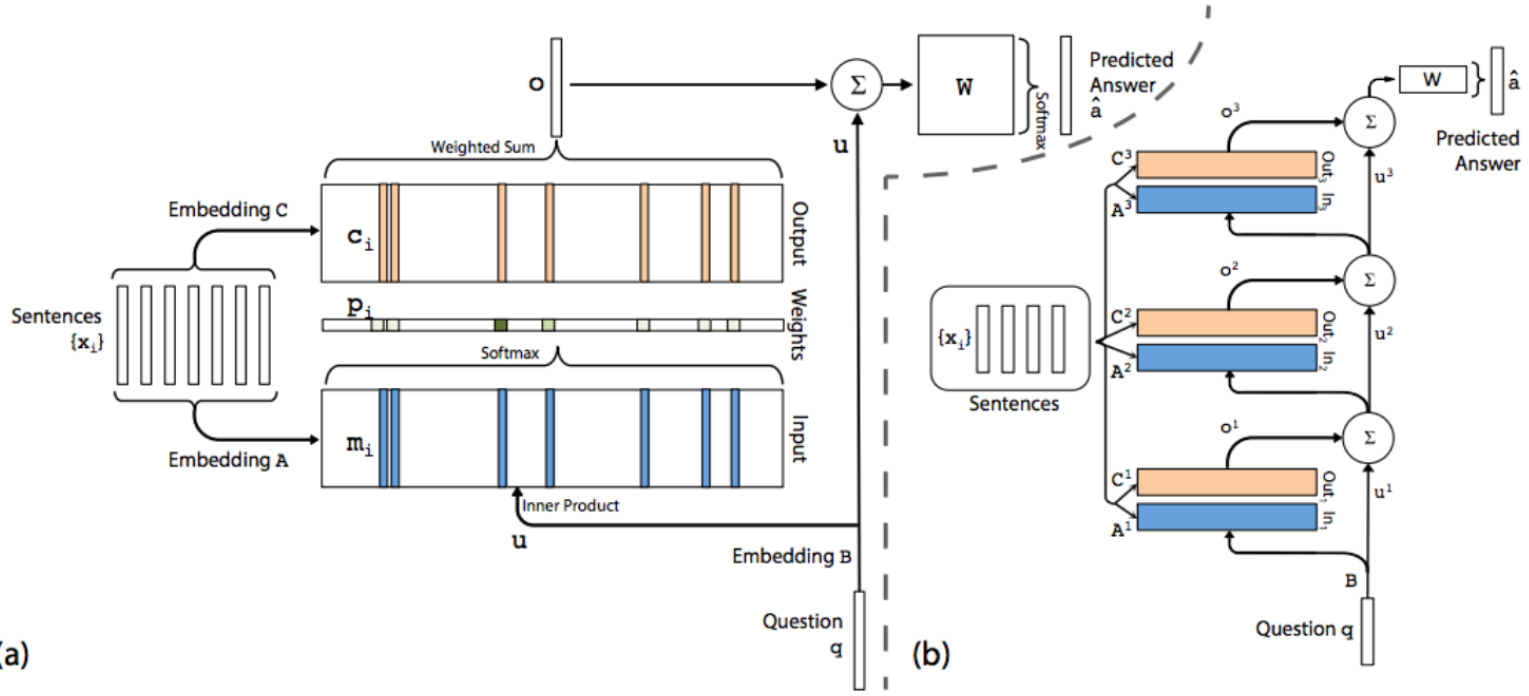
MEMORY NETWORKS

Jason Weston, Sumit Chopra & Antoine Bordes
Facebook AI Research
770 Broadway
New York, USA
{jase,spchopra,abordes}@fb.com

End-To-End Memory Networks

Sainbayar Sukhbaatar
Dept. of Computer Science
Courant Institute, New York University
sainbar@cs.nyu.edu

Arthur Szlam **Jason Weston** **Rob Fergus**
Facebook AI Research
New York
{aszlam,jase,robfergus}@fb.com

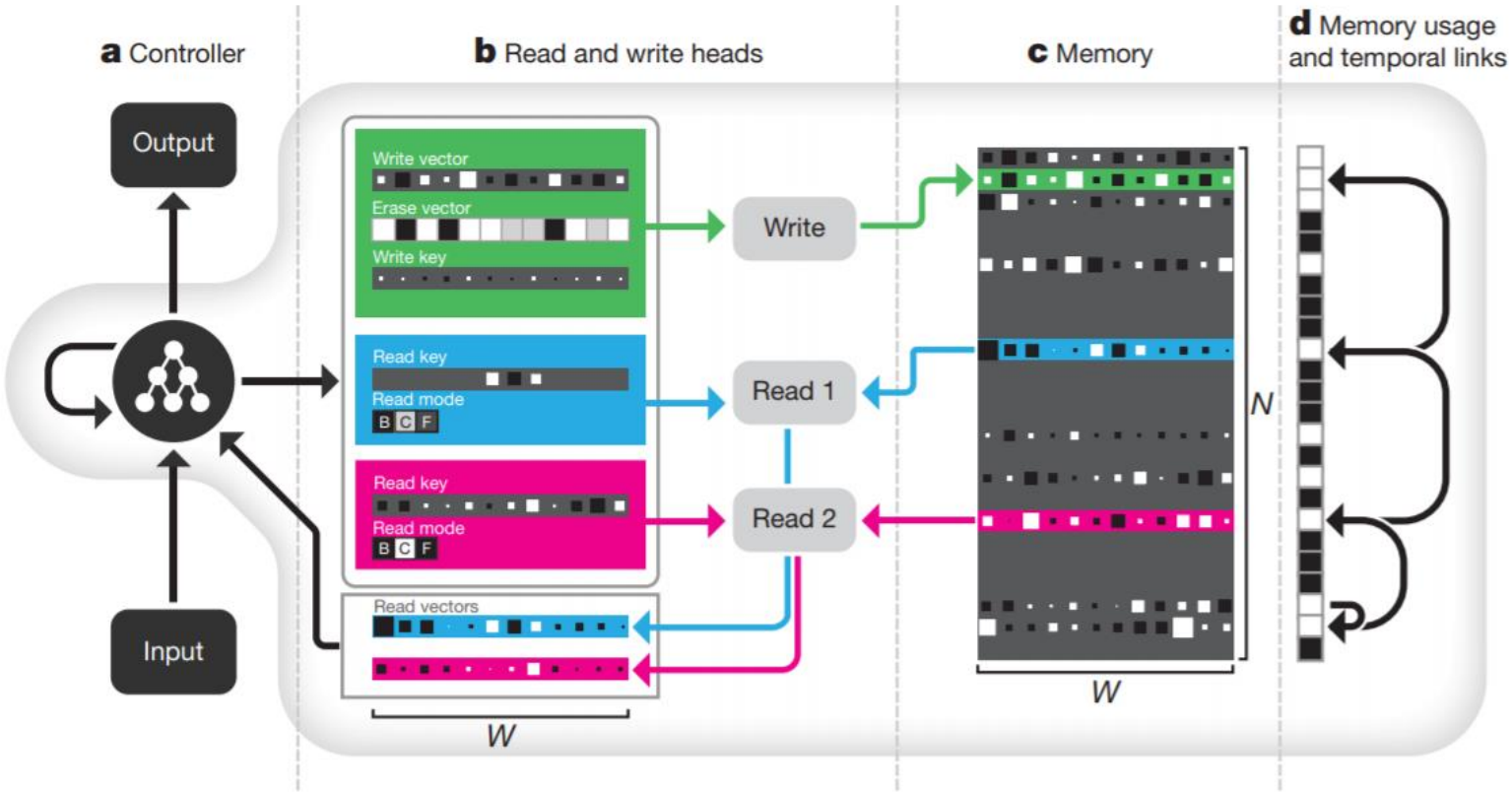


Differentiable Neural Computer [Graves et al, 2016]

Permite acesso a memória como uma RAM

Atenção seletiva

- Endereça a atenção na matriz de memória
- Pondera seus pesos levando em consideração o **conteúdo**, **posição na memória** e **ordem temporal**



E como vou treinar meu chatbot?

Seq2seq atualmente é muito usado, mas não oferece resultados ótimos

Meus 10 centavos sobre o assunto:

- As redes recorrentes atuais modelam as propriedades dos idiomas, mas não capturam o contexto em que se aplicam
- Chatbot para domínios específicos podem ser implementados com seq2seq
- Já para domínios abertos ainda é um problema em aberto
  - Redes com memoria são um possível caminho

Exemplo

Chatbot programado para vender notebooks Dell

Usuário: Quero um notebook novo?

Chatbot: Prefere de qual o tamanho?

Usuário: Quero um de 15?

Chatbot:

Exemplo

Chatbot programado para vender notebooks Dell

Usuário: Quero um notebook novo?

Chatbot: Prefere de qual o tamanho?

Usuário: Quero um de 15?

Chatbot:

Exemplo

Chatbot programado para atender usuários do Buscapé

Usuário: Quero um notebook novo?

Chatbot: Prefere de qual o tamanho?

Usuário: Quero um de 15?

Chatbot:

Exemplo

Chatbot programado para atender usuários do Buscapé

Usuário: Quero um notebook novo?

Chatbot: Prefere de qual o tamanho?

Usuário: Quero um de 15?

Chatbot:

Electrolux Turbo Economia LTD15 Superior 15 Kg Branco

8/10 (Baseado em 114 avaliações)

Veja o histórico de preços

Detalhes do produto

Estudo de caso

Predição de complicações em pessoas diabéticas

30/10/2012	VITAMINA "D" 25 HIDROXI
30/10/2012	CREATINO FOSFOQUINASE TOTAL (CK)
30/10/2012	VITAMINA "D" 25 HIDROXI
30/10/2012	CREATINO FOSFOQUINASE TOTAL (CK)
01/11/2012	CONSULTA EM CONSULTORIO
01/11/2012	CONSULTA EM CONSULTORIO
07/11/2012	CONSULTA EM CONSULTORIO
07/11/2012	CONSULTA EM CONSULTORIO
12/11/2012	ECG CONVENCIONAL
12/11/2012	ECG CONVENCIONAL
13/11/2012	CONSULTA EM CONSULTORIO
13/11/2012	CONSULTA EM CONSULTORIO
13/11/2012	ECG CONVENCIONAL
13/11/2012	ECG CONVENCIONAL
26/11/2012	CONSULTA EM CONSULTORIO
26/11/2012	CONSULTA EM CONSULTORIO
30/11/2012	DOPPLER COLORIDO DE VASOS CERVICAIS ARTERIAIS BILATERAL
30/11/2012	ECODOPPLERCARDIOGRAMA TRANSTORACICO
30/11/2012	ECODOPPLERCARDIOGRAMA TRANSTORACICO
30/11/2012	DOPPLER COLORIDO DE VASOS CERVICAIS ARTERIAIS BILATERAL

Não terá Complicações

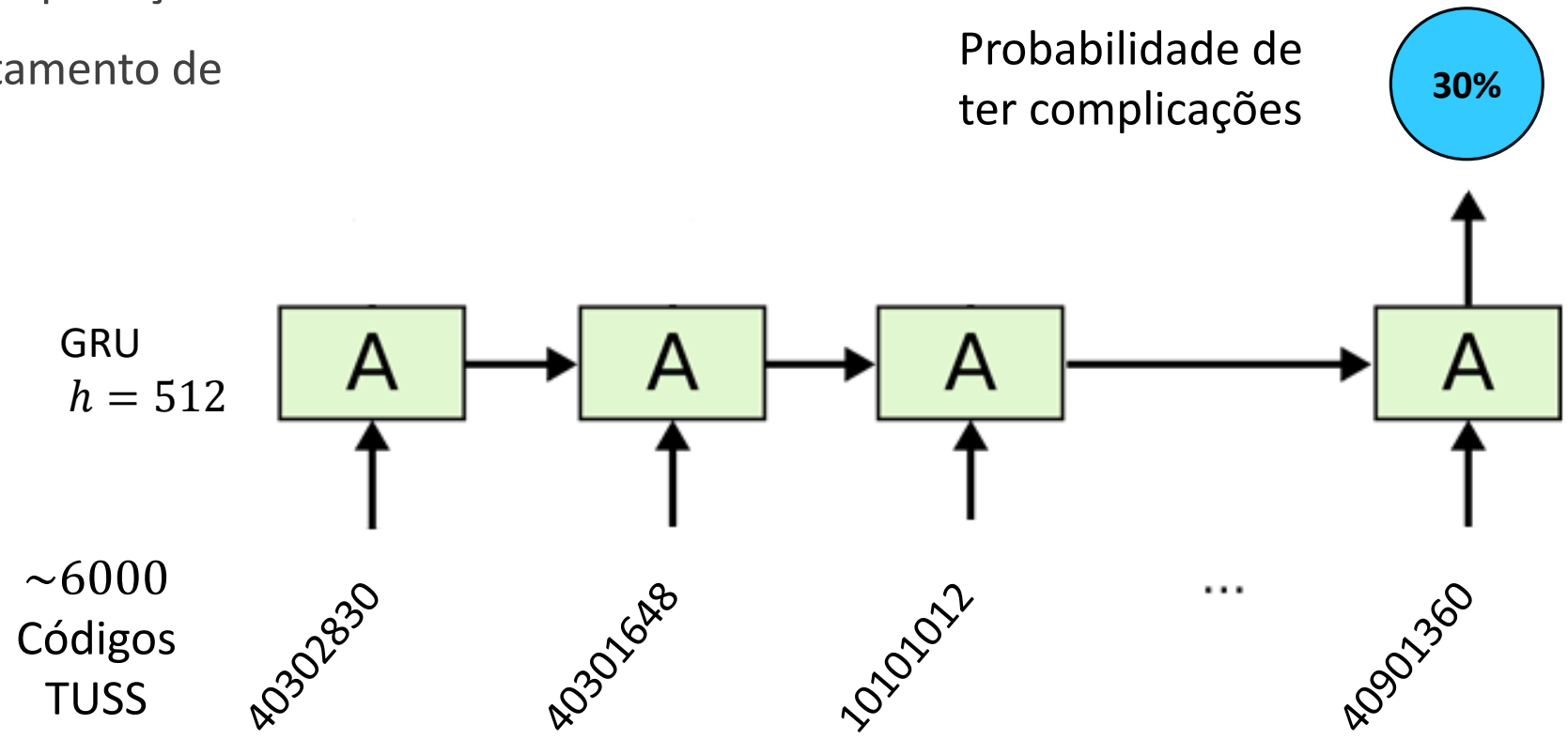
Possível Complicação

Estudo de caso

10.580.481 de parâmetros

Média de 75% de acerto nas predições

A rede modelou o comportamento de um beneficiário diabético



The End

Muito Obrigado