Modelos de Linguagem Neurais Com Redes Recorrentes

"Time will explain"

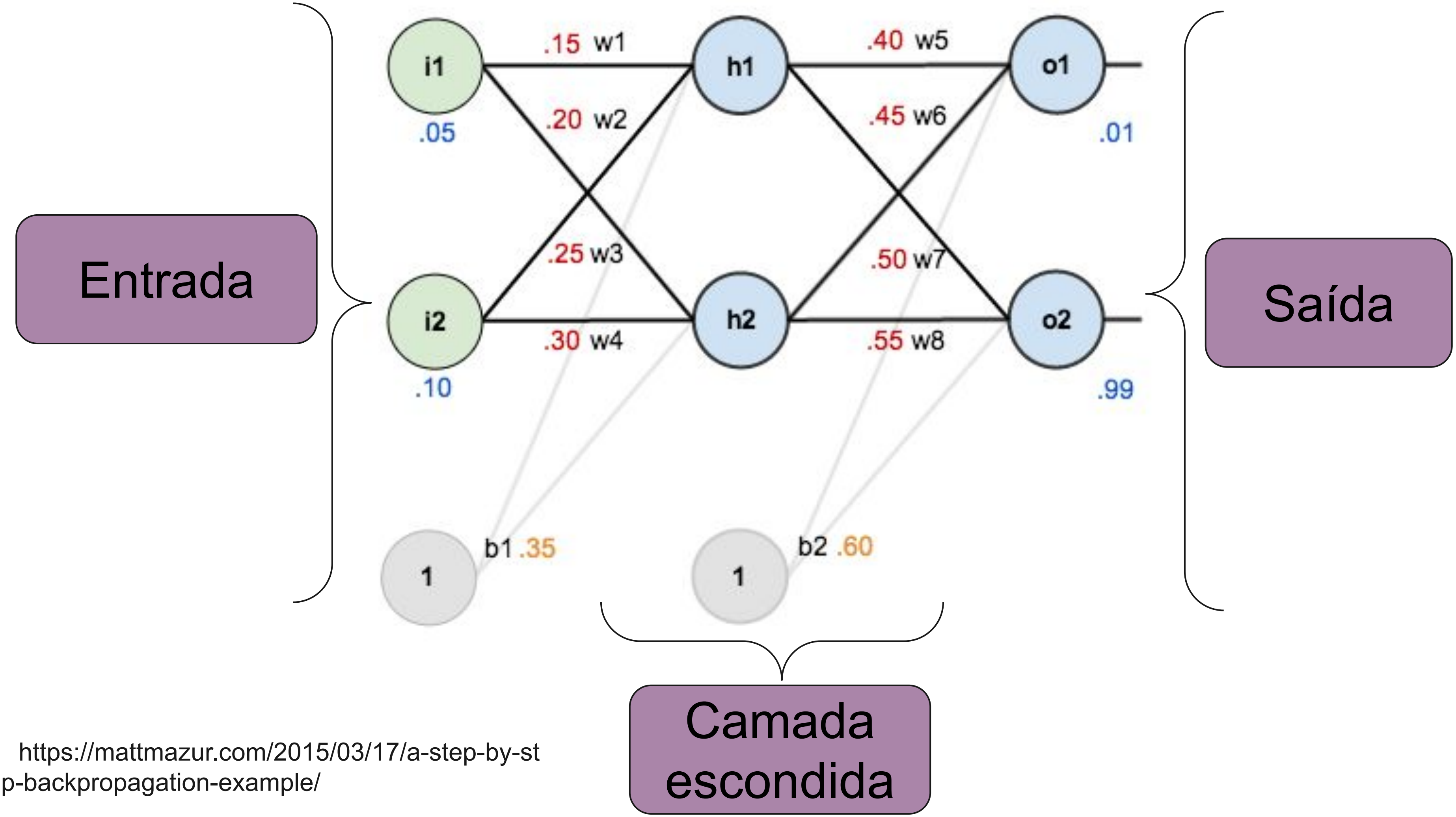
centro tomorrow

POSITIVO TECNOLOGIA

computação UFBA

Profa Aline Paes
alinepaes@ic.uff.br

Rede neural feed-forward



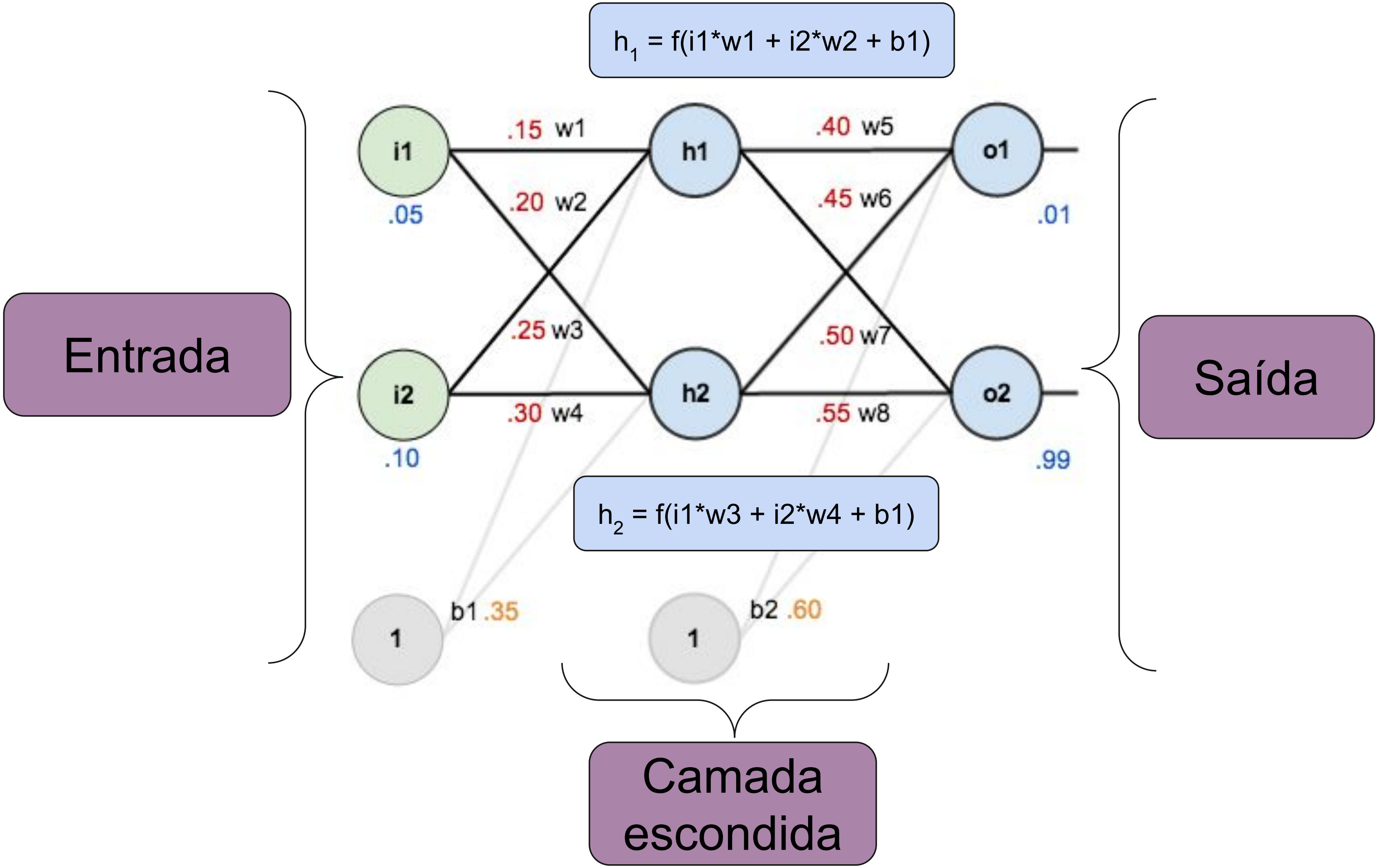
Entrada

Camada escondida

Saída

https://mattmazur.com/2015/03/17/a-step-by-step-backpropagation-example/

Rede neural feed-forward



$h_1 = f(i1*w1 + i2*w2 + b1)$

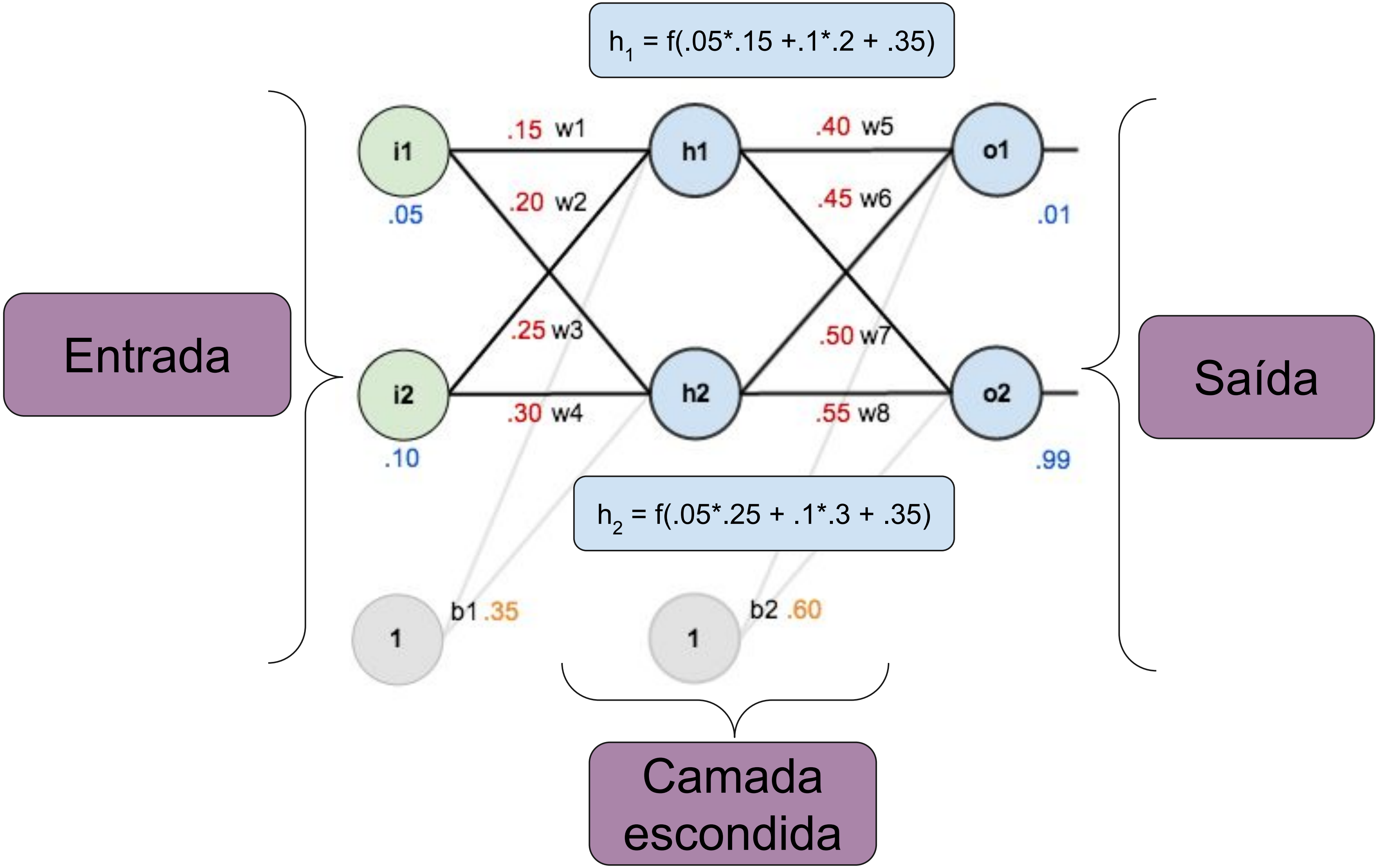
$h_2 = f(i1*w3 + i2*w4 + b1)$

Entrada

Saída

Camada escondida

Rede neural feed-forward



h_1 = f(.05*.15 +.1*.2 + .35)

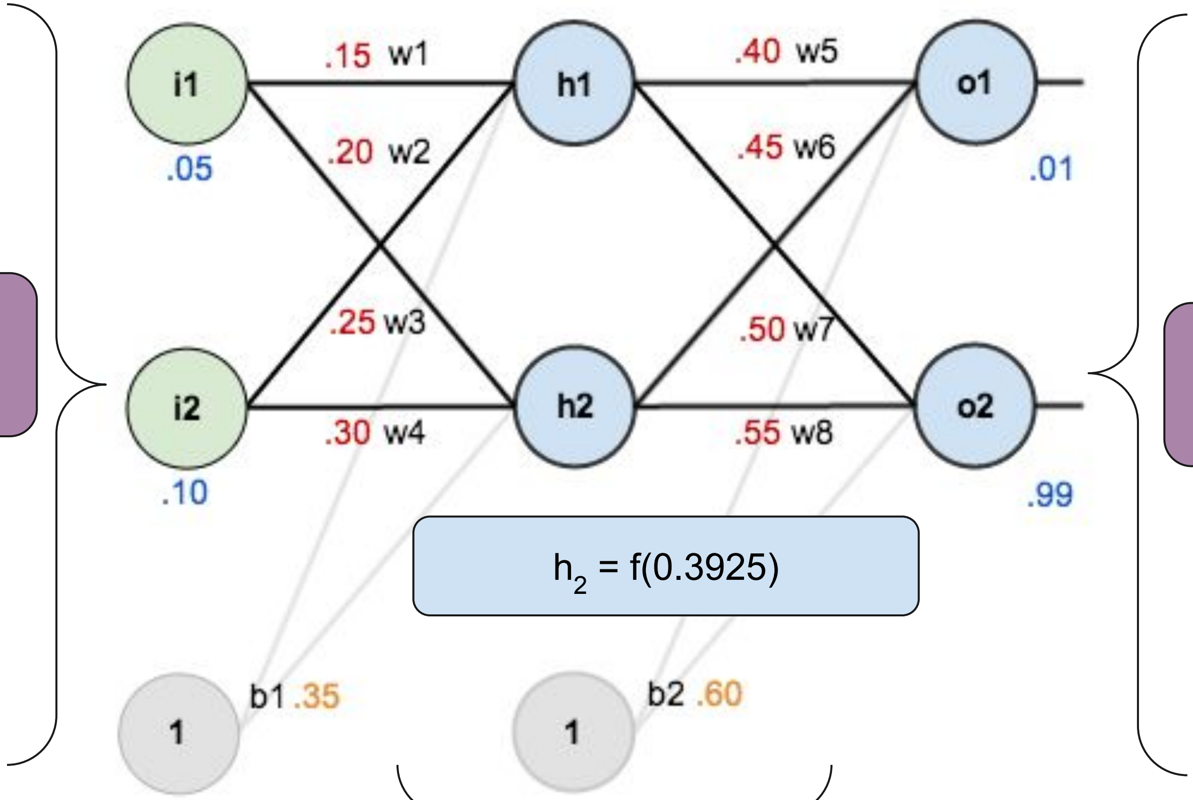
h_2 = f(.05*.25 + .1*.3 + .35)

Entrada

Saída

Camada escondida

Rede neural feed-forward



$h_1 = f(0.3775)$

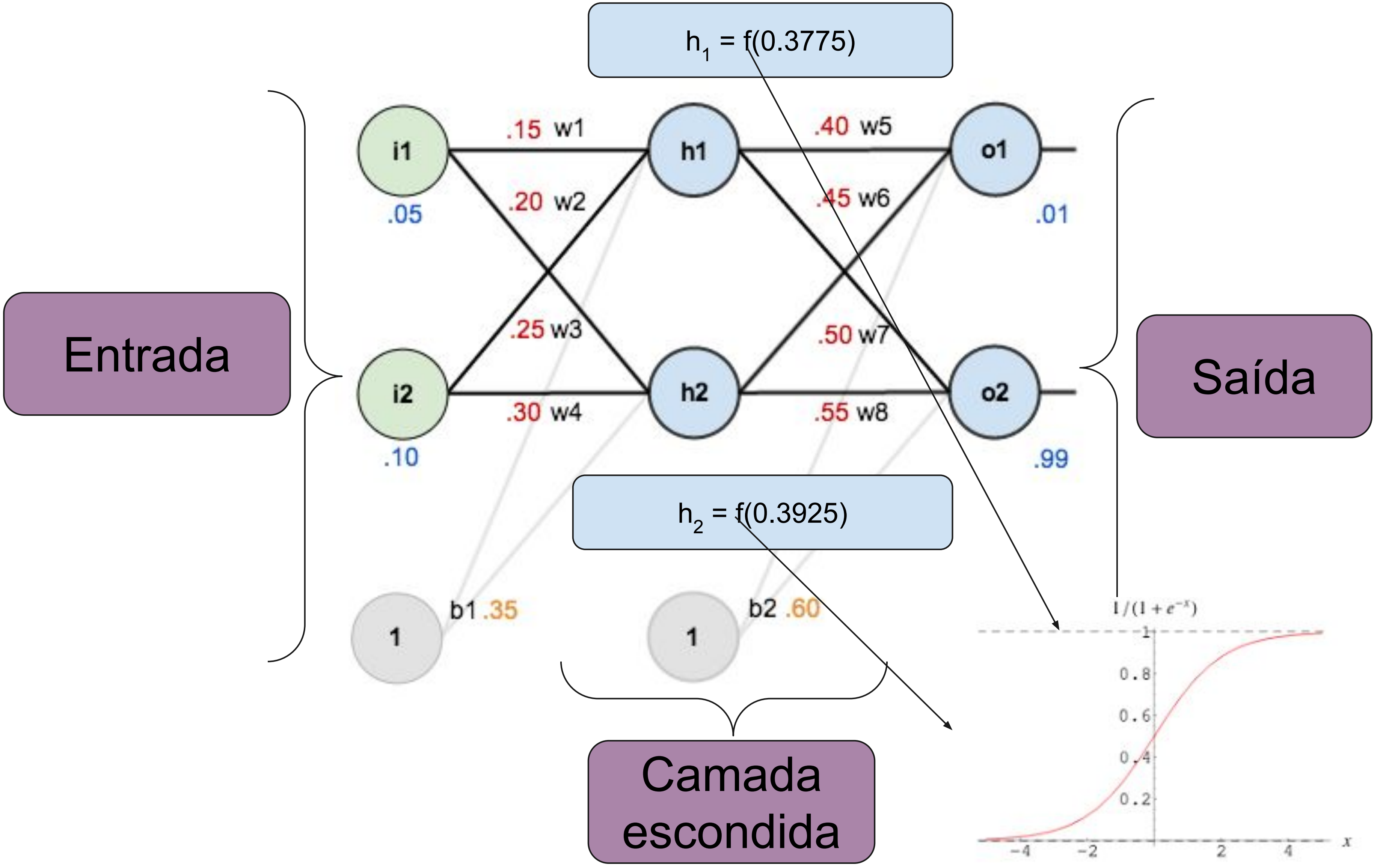
$h_2 = f(0.3925)$

Entrada

Saída

Camada escondida

Rede neural feed-forward



$h_1 = f(0.3775)$

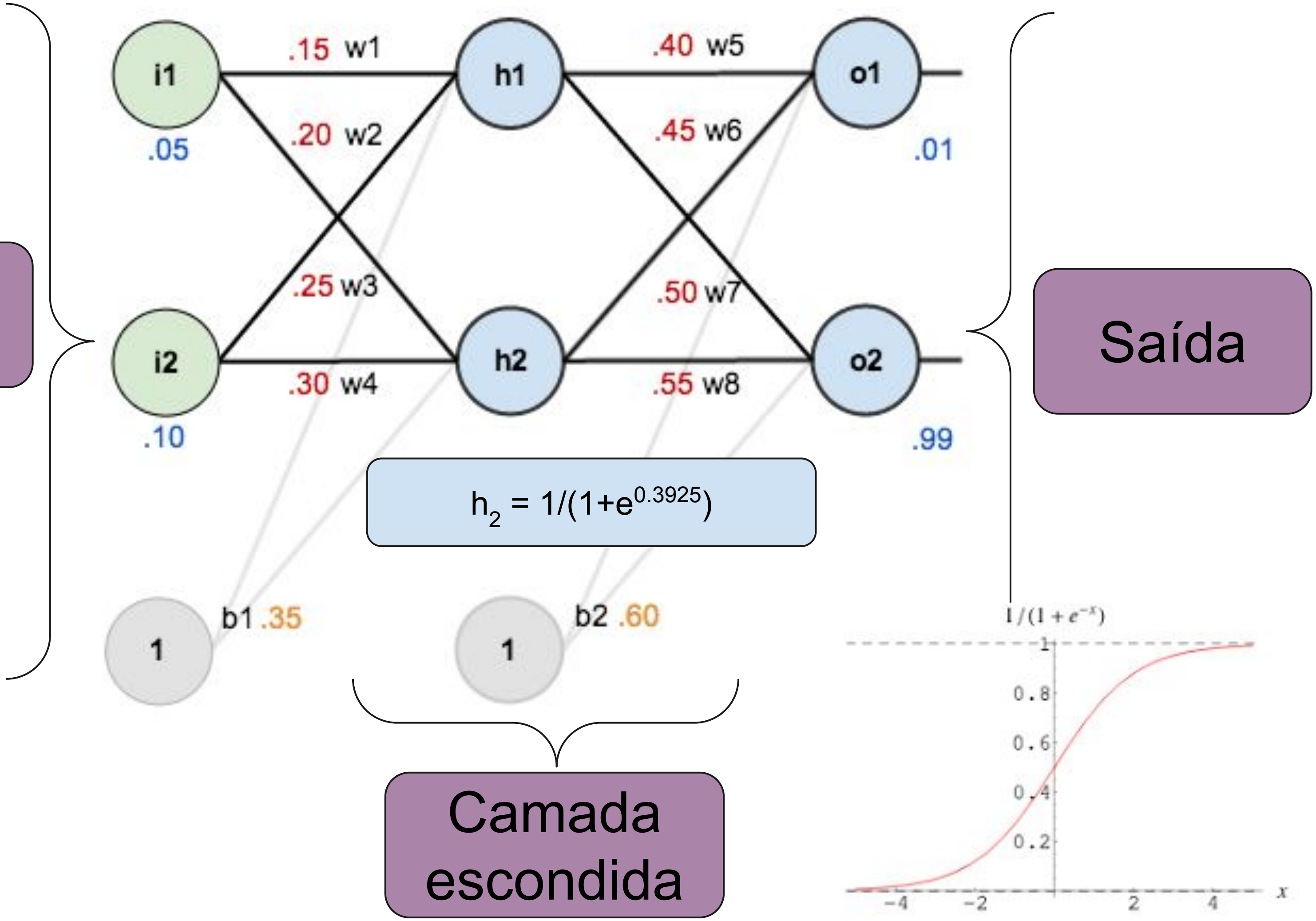
$h_2 = f(0.3925)$

Entrada

Saída

Camada escondida

Rede neural feed-forward



$h_1 = 1/(1+e^{0.3775})$

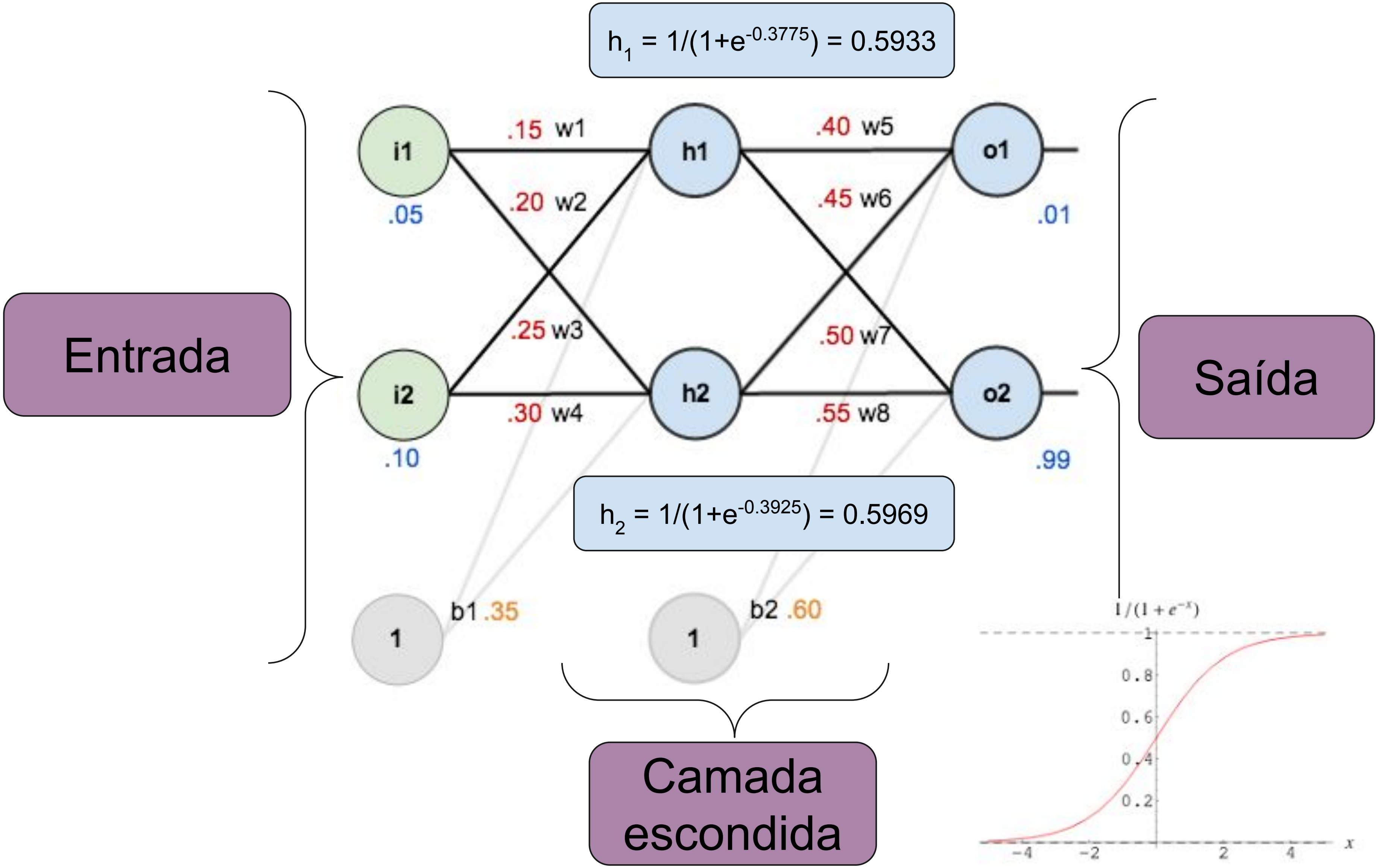
Entrada

Saída

$h_2 = 1/(1+e^{0.3925})$

Camada escondida

Rede neural feed-forward



$h_1 = 1/(1+e^{-0.3775}) = 0.5933$

$h_2 = 1/(1+e^{-0.3925}) = 0.5969$

Entrada

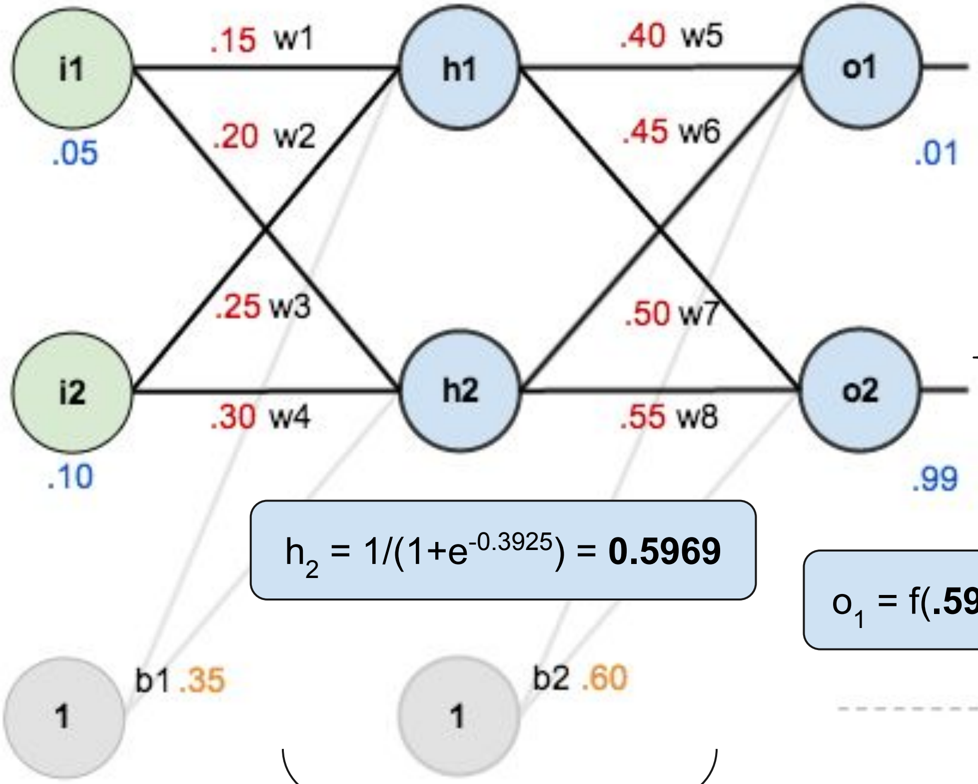
Saída

Camada escondida

Rede neural feed-forward

$h_1 = 1/(1+e^{-0.3775}) = \mathbf{0.5933}$

$o_1 = f(\mathbf{.5933}*.4 + \mathbf{.5969}*.5 + .6)$

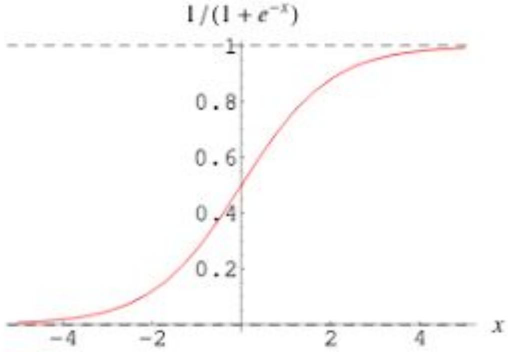


Entrada

Saída

$h_2 = 1/(1+e^{-0.3925}) = \mathbf{0.5969}$

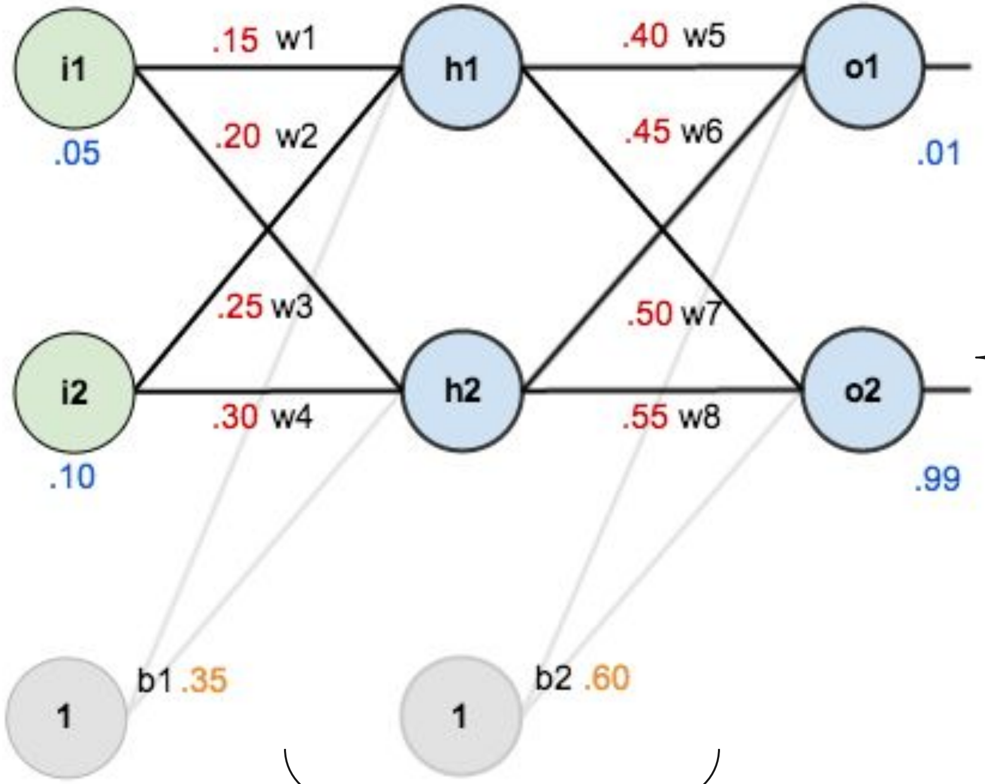
$o_1 = f(\mathbf{.5933}*.45 + \mathbf{.5969}*.55 + .6)$



Camada escondida

Rede neural feed-forward

Entrada: tamanho fixo

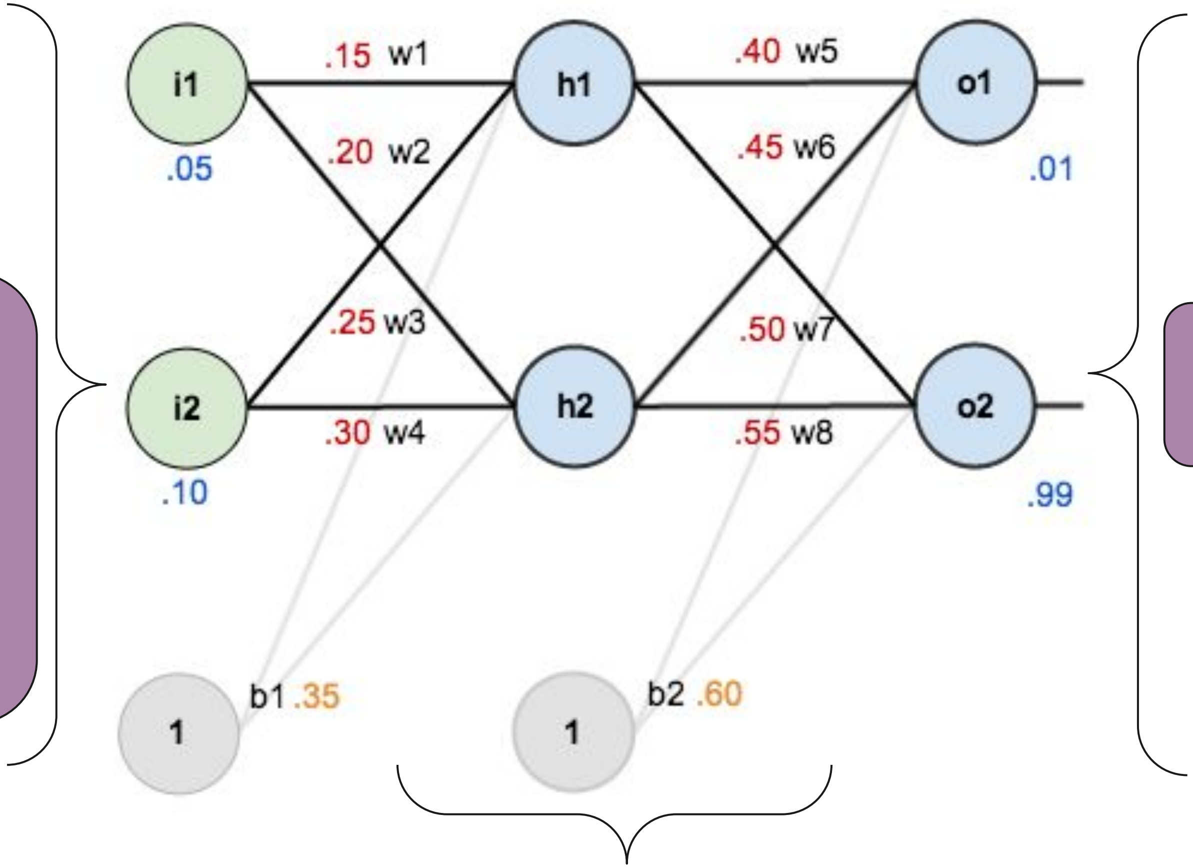


Saída

Camada escondida

https://mattmazur.com/2015/03/17/a-step-by-step-backpropagation-example/

Rede neural feed-forward



Entrada: ordem não é relevante

Saída

Camada escondida

https://mattmazur.com/2015/03/17/a-step-by-step-backpropagation-example/

Sequências em NLP

"De casa eu decidi antes que as coisas mais difíceis ficassem mudar."

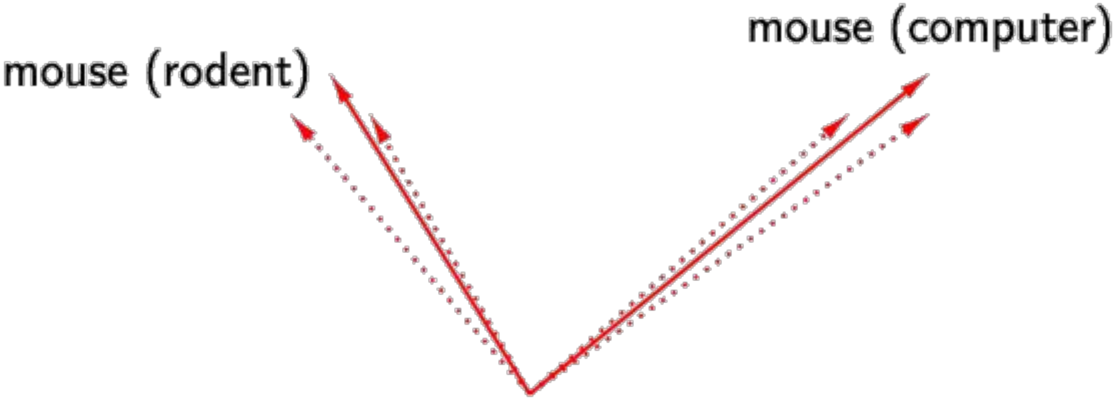
"De eu decidi casa antes que as coisas mais difíceis ficassem mudar todo dia."

"Casa de antes decidi casa eu as que coisas mais mudar ficassem difíceis."

Sequências em NLP

- Diversos problemas precisam considerar as dependências entre termos
    - Co-referência
        - *Apesar das suas obrigações familiares, Wilma consegue se dedicar aos estudos.*
    - Concordância de número e gênero
        - *Lula e FHC foram presidentes do Brasil.*
    - Coesão de textos
        - *A caixa não coube na mala pois ela era muito…*
            - *Grande:*
            - *Pequena:*

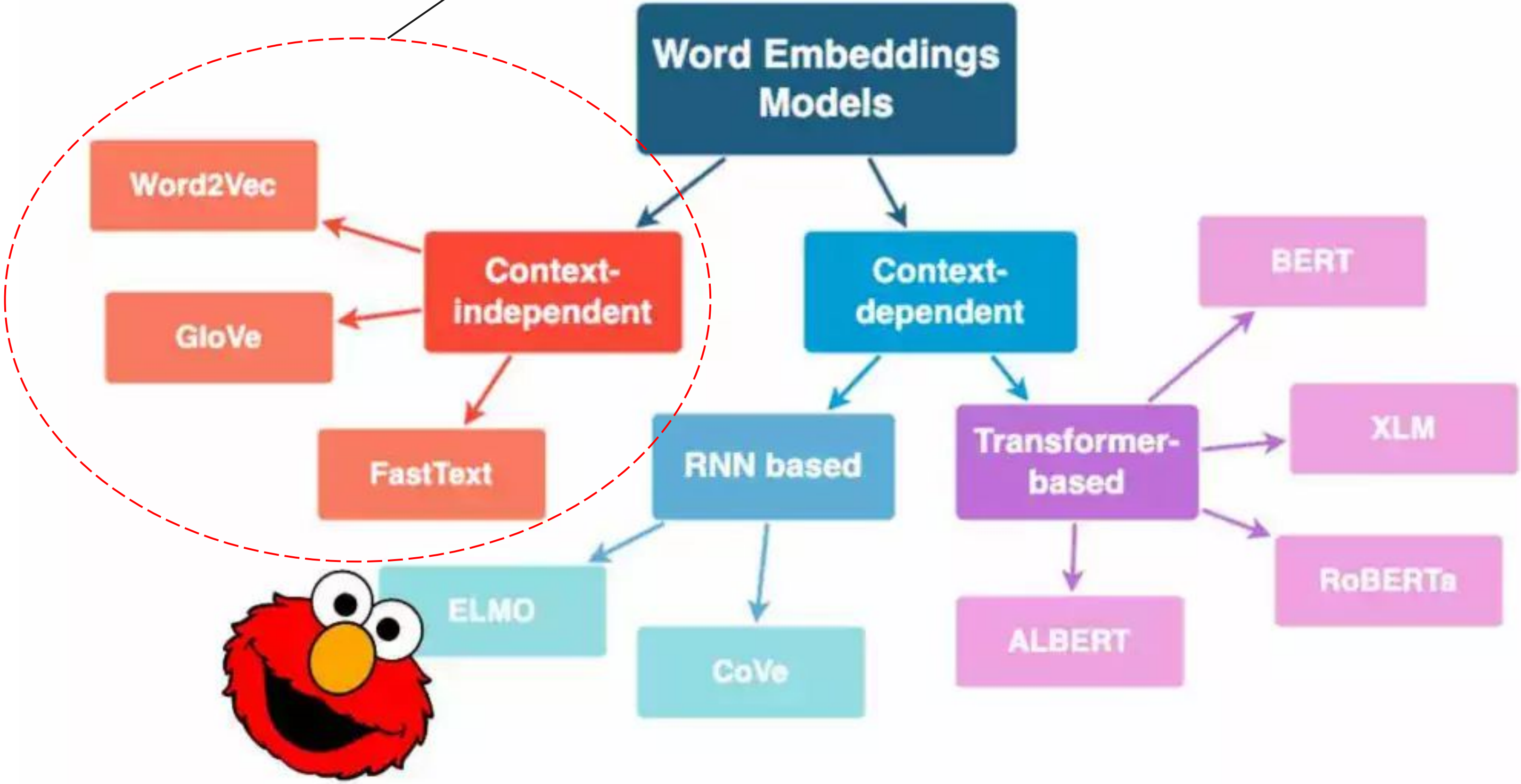
Contexto - de verdade



https://ai.stanford.edu/blog/contextual/

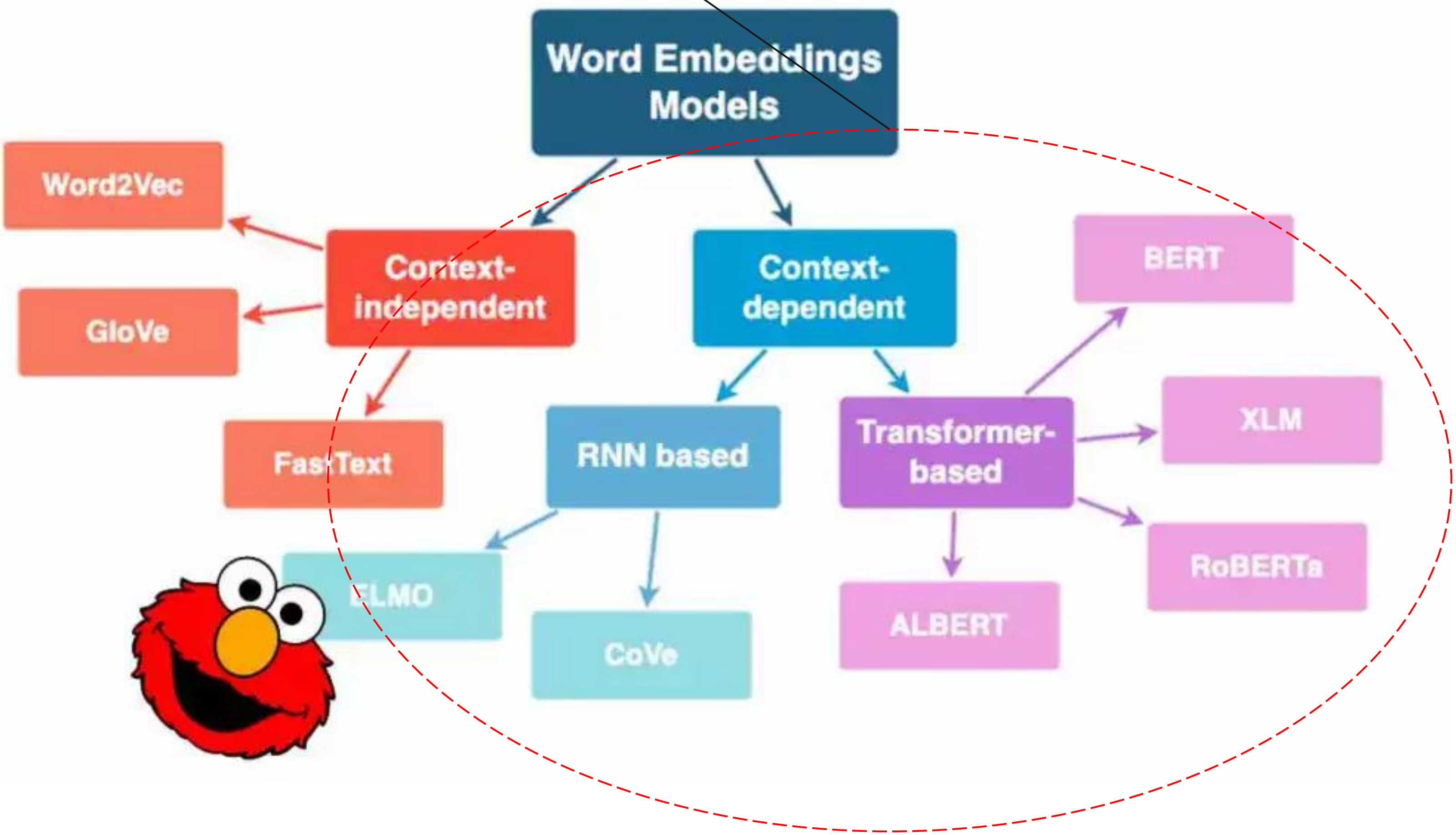
Contexto - de verdade

Embeddings estáticos



https://spotintelligence.com/2023/12/26/embeddings-from-language-models-elmo/

Contexto - "de verdade"



Embeddings contextualizados

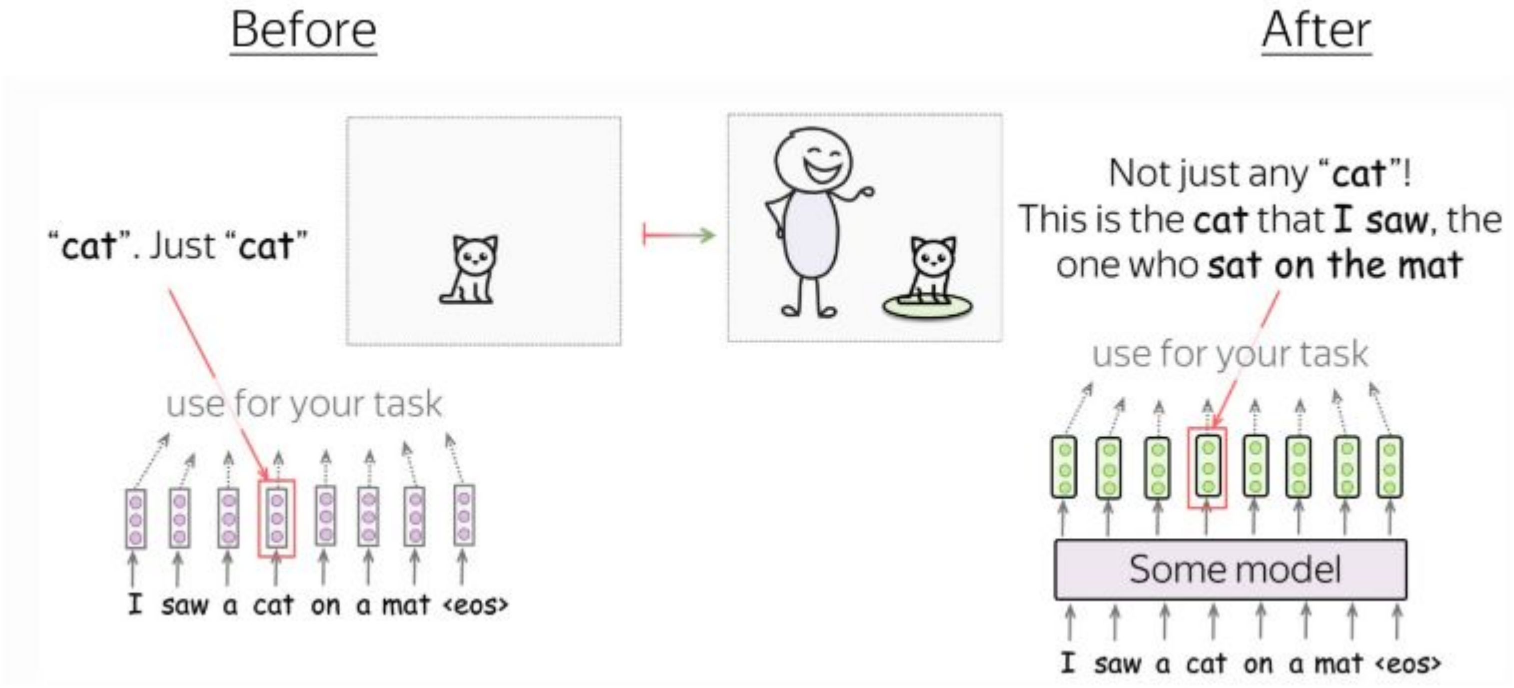
https://spotintelligence.com/2023/12/26/embeddings-from-language-models-elmo/

Embeddings contextualizados : Elmo (Peters et al., 2018)

Contextualized word-embeddings can give words different embeddings based on the meaning they carry in the context of the sentence. Also, RIP Robin Williams

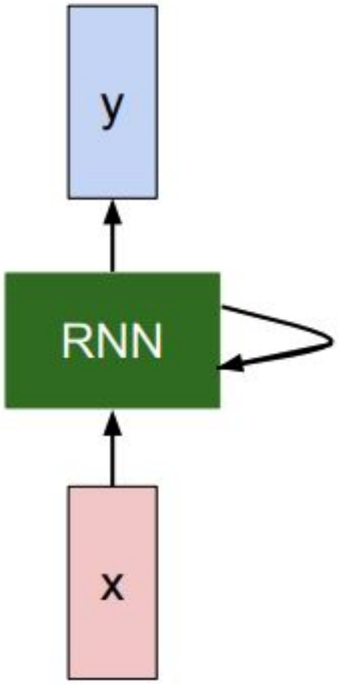
* http://jalammar.github.io/illustrated-bert/

Embeddings contextualizados



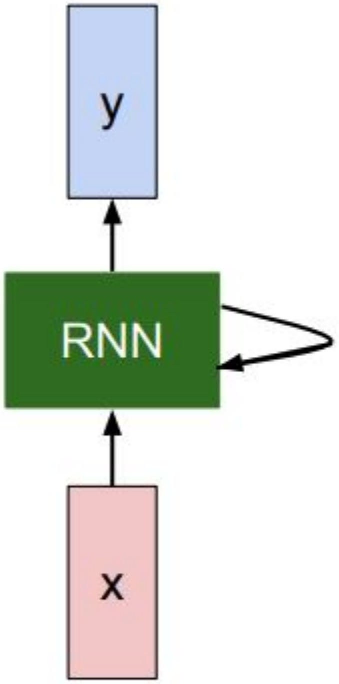
https://lena-voita.github.io/nlp_course.html#preview_transfer

Rede neural recorrente (Elman, 1990)



Estado interno: atualizado conforme uma **sequência** é processada

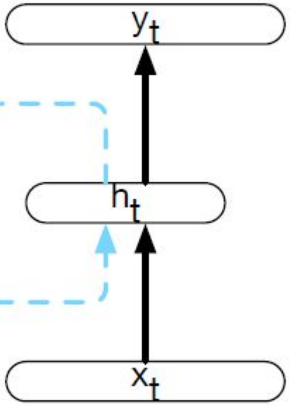
Rede neural recorrente (Elman, 1990)



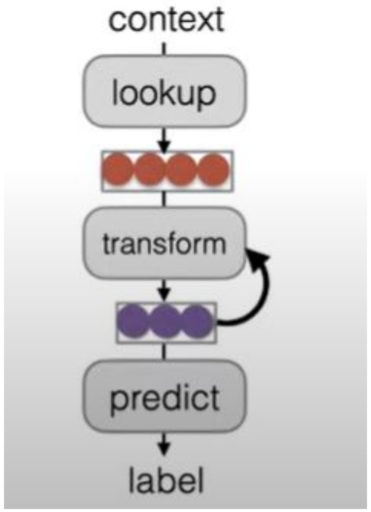
Permitem entradas de tamanho variável

Redes Neurais Recorrentes (Elman, 1990)

- Contém um ciclo em suas conexões
  - O valor de uma unidade é direta ou indiretamente dependente de uma saída anterior
- Simulam memória
- Permitem entradas de tamanho variável

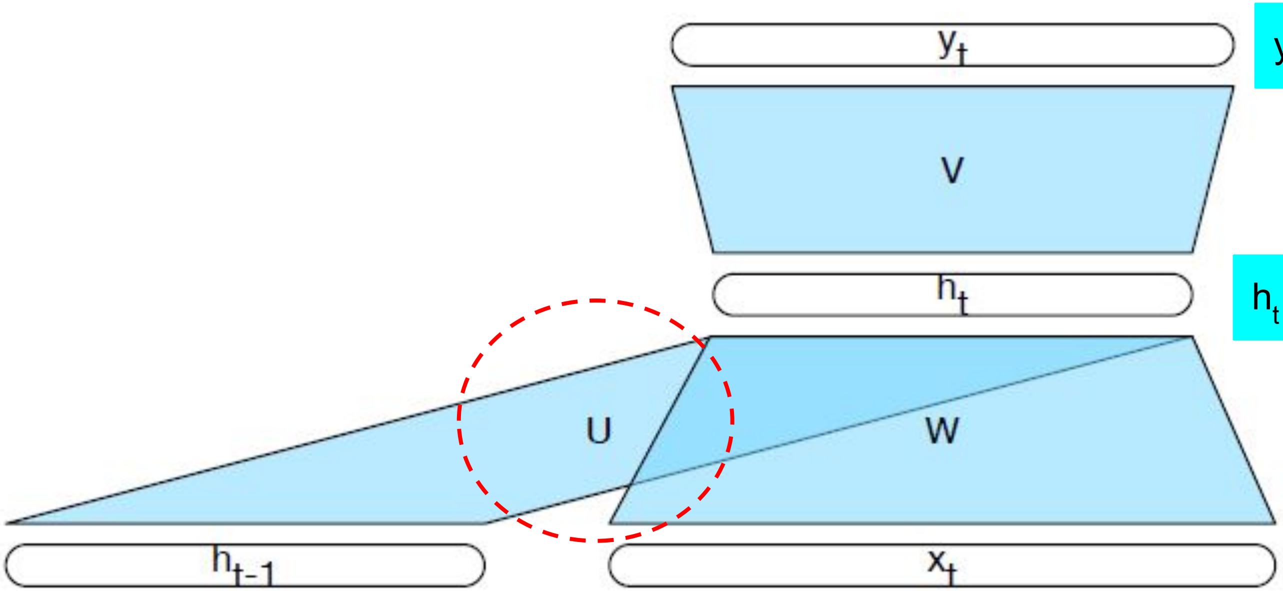


*Speech and Language Processing



*CMU Neural Nets for NLP

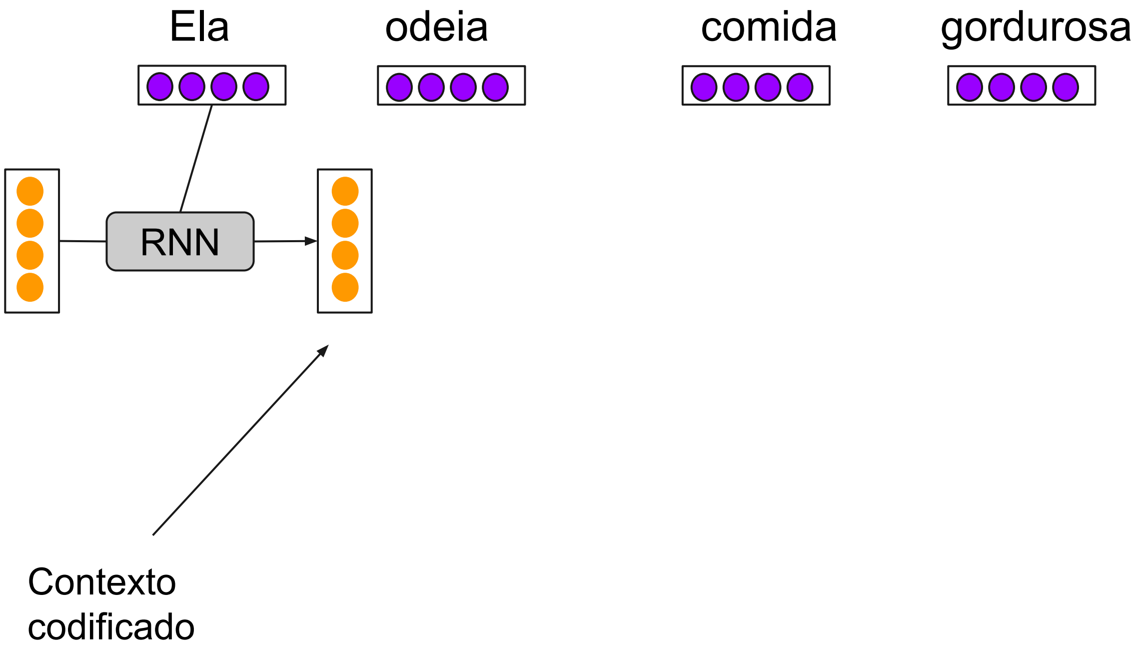
Unfolding



$y_t = f(Vh_t)$

$h_t = g(Uh_{t-1}+Wx_t)$

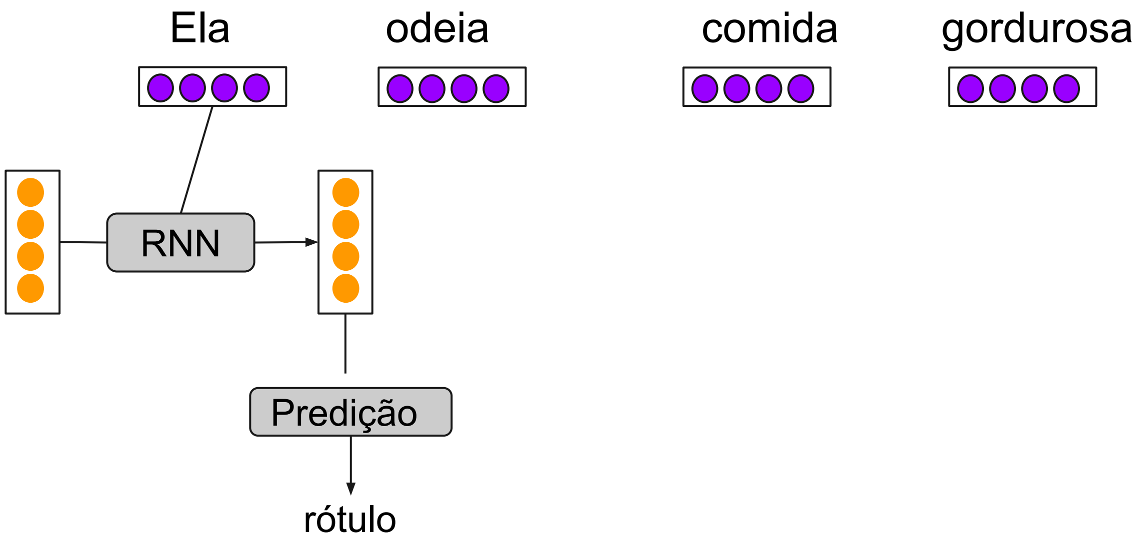
Unfolding



Contexto
codificado

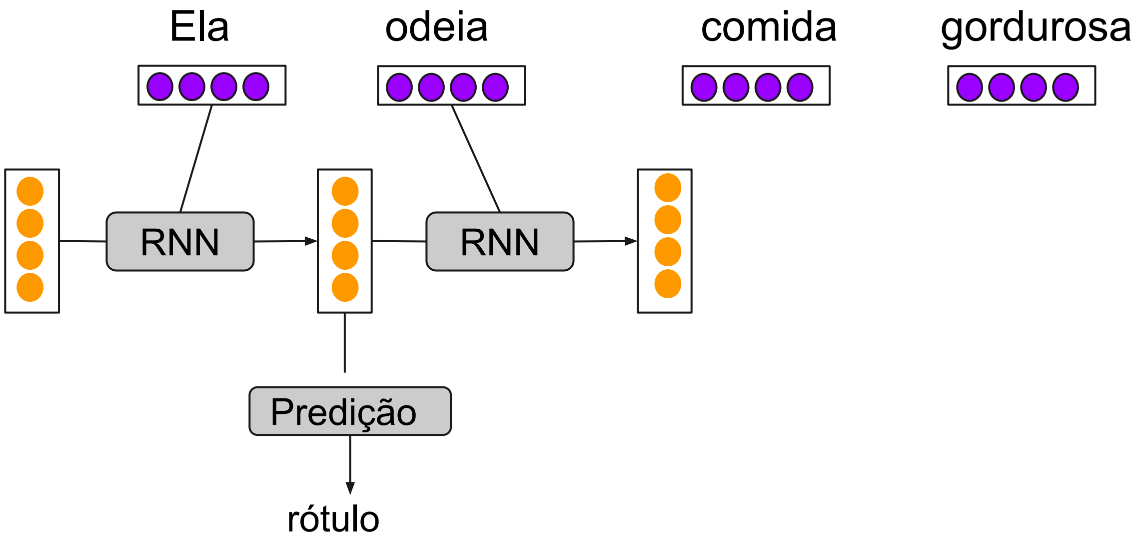
*CMU Neural nets for NLP

Unfolding



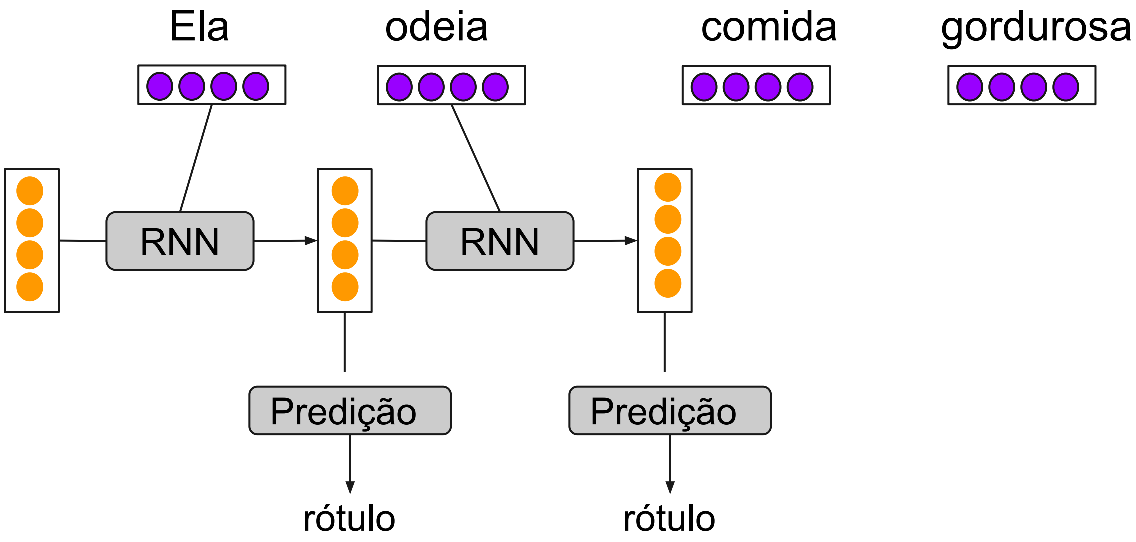
*CMU Neural nets for NLP

Unfolding



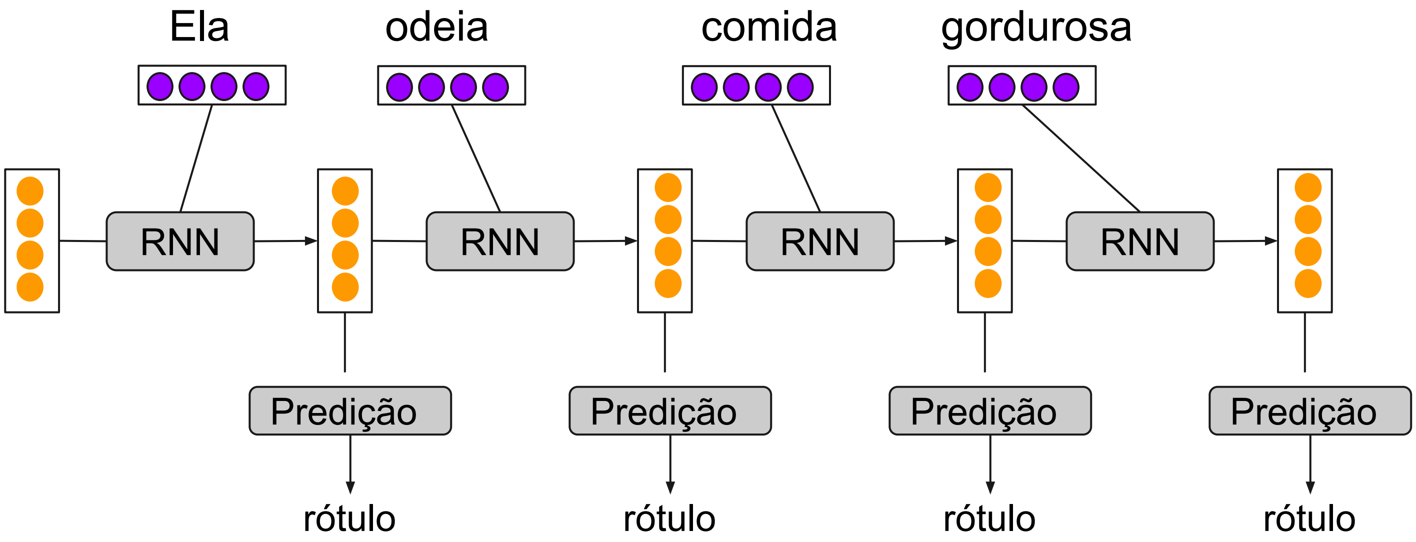
*CMU Neural nets for NLP

Unfolding



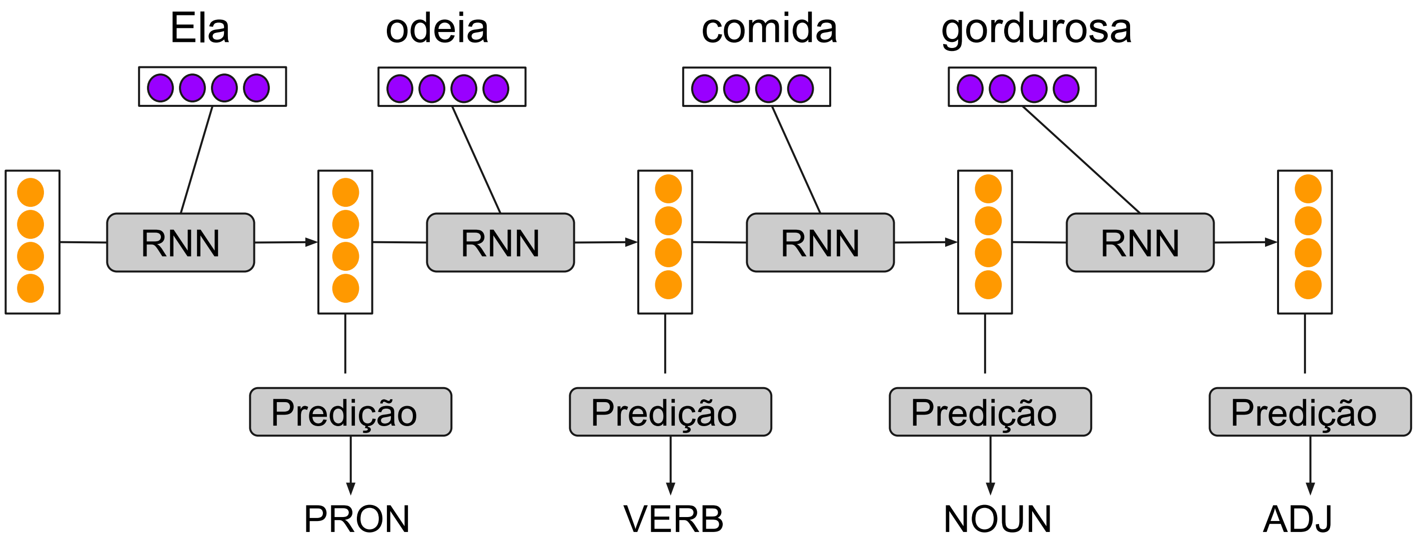
*CMU Neural nets for NLP

Unfolding



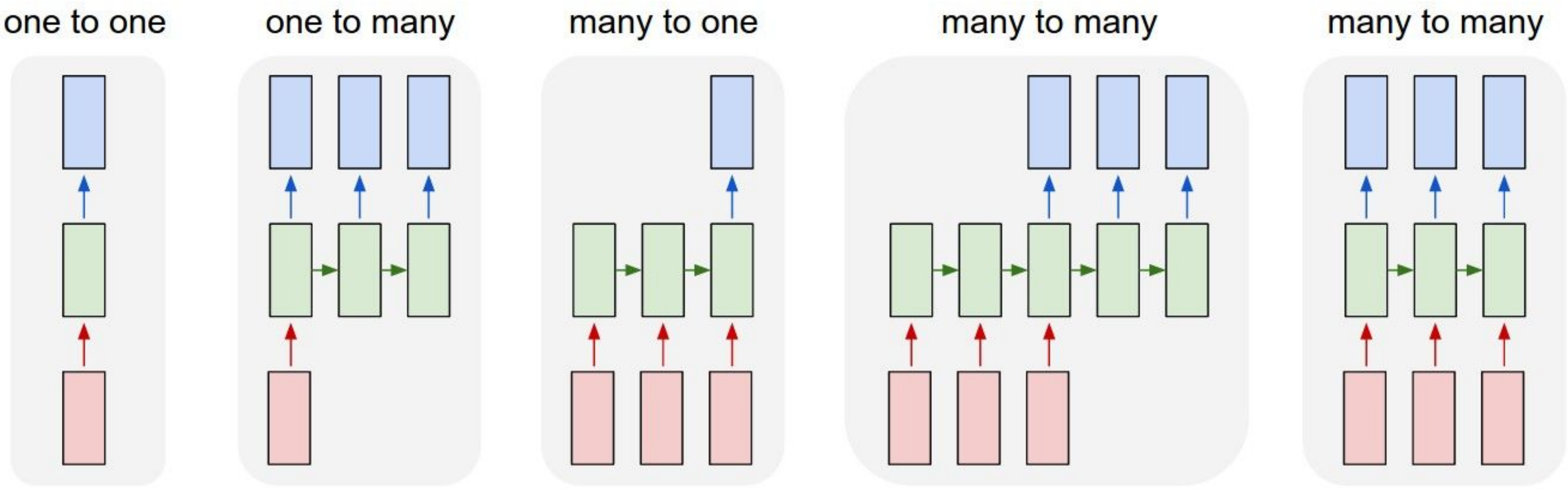
*CMU Neural nets for NLP

Unfolding - Tagging, parsing



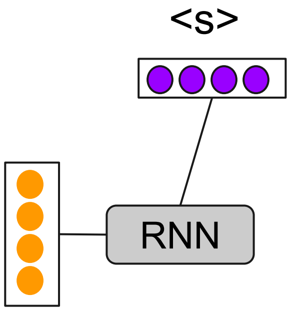
*CMU Neural nets for NLP

Resumo de tipos



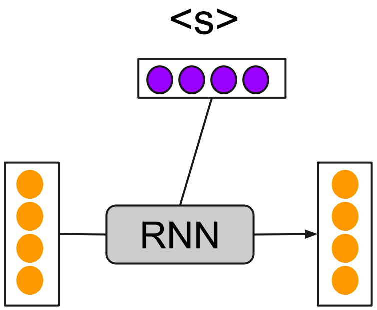
*The Unreasonable Effectiveness of Recurrent Neural Networks, Andrej Karpathy blog, 2015

Inferência - geração de texto



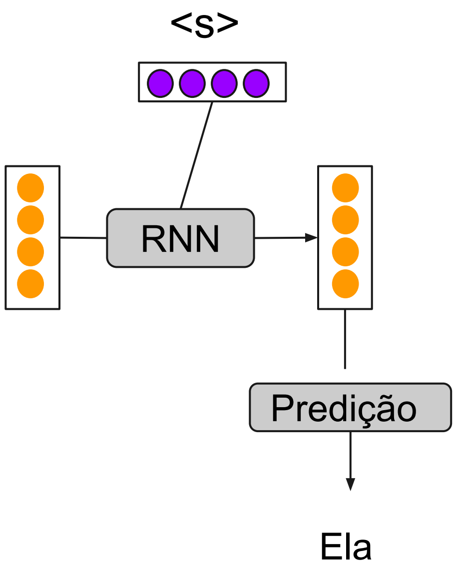
*CMU Neural nets for NLP

Inferência - geração de texto



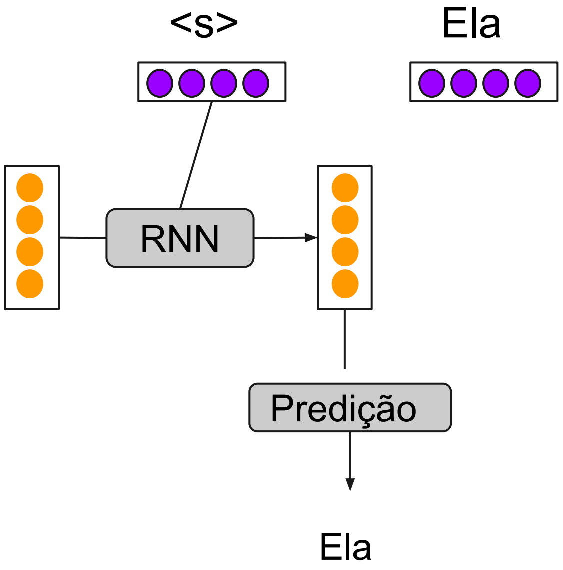
*CMU Neural nets for NLP

Inferência - geração de texto



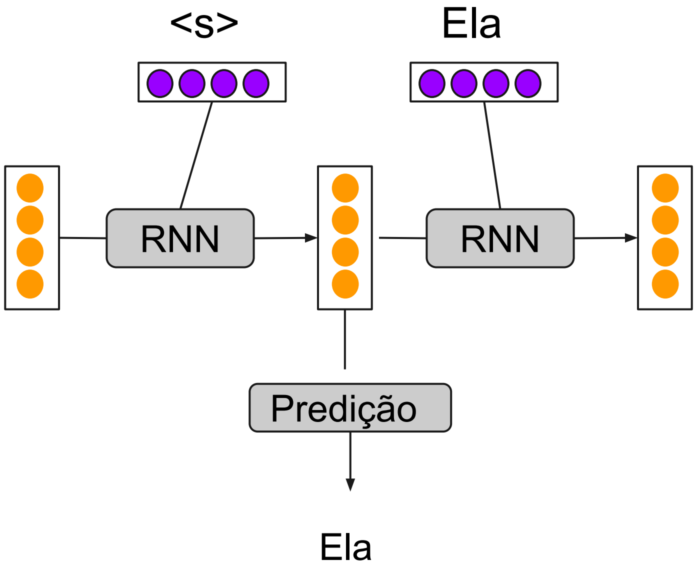
*CMU Neural nets for NLP

Inferência - geração de texto



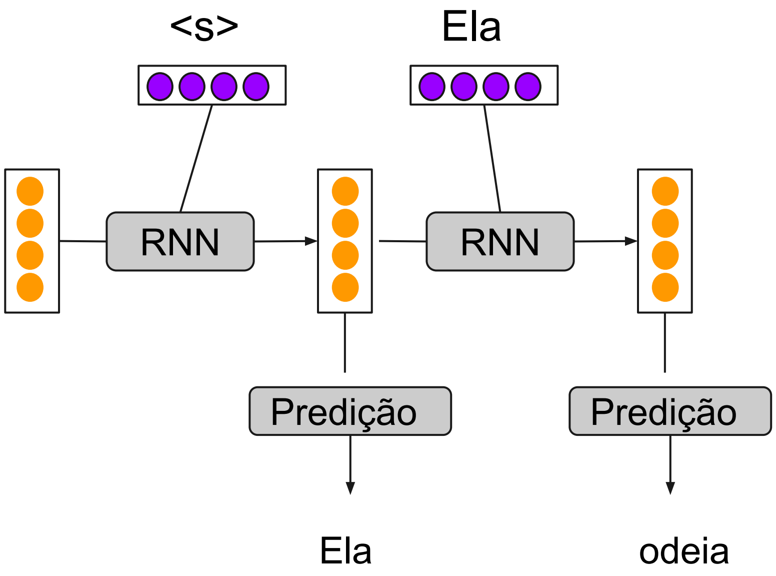
*CMU Neural nets for NLP

Inferência - geração de texto



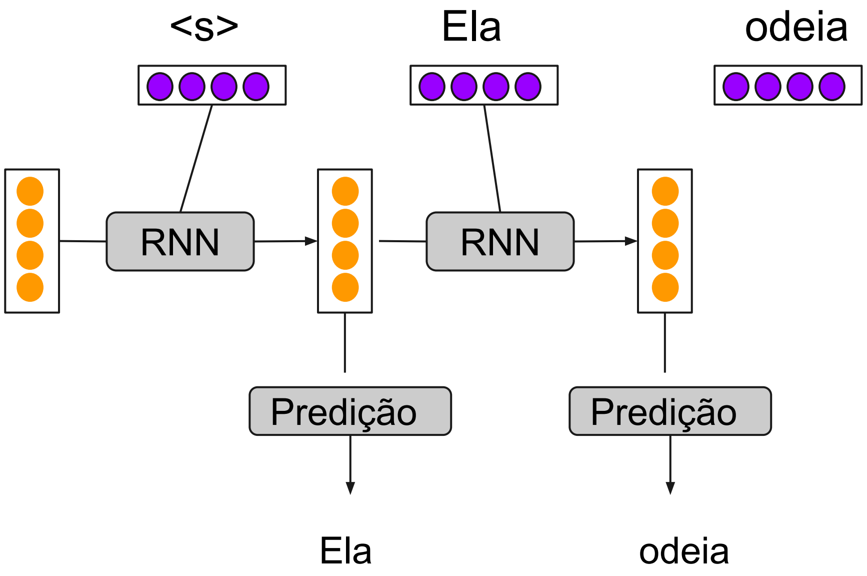
*CMU Neural nets for NLP

Inferência - geração de texto



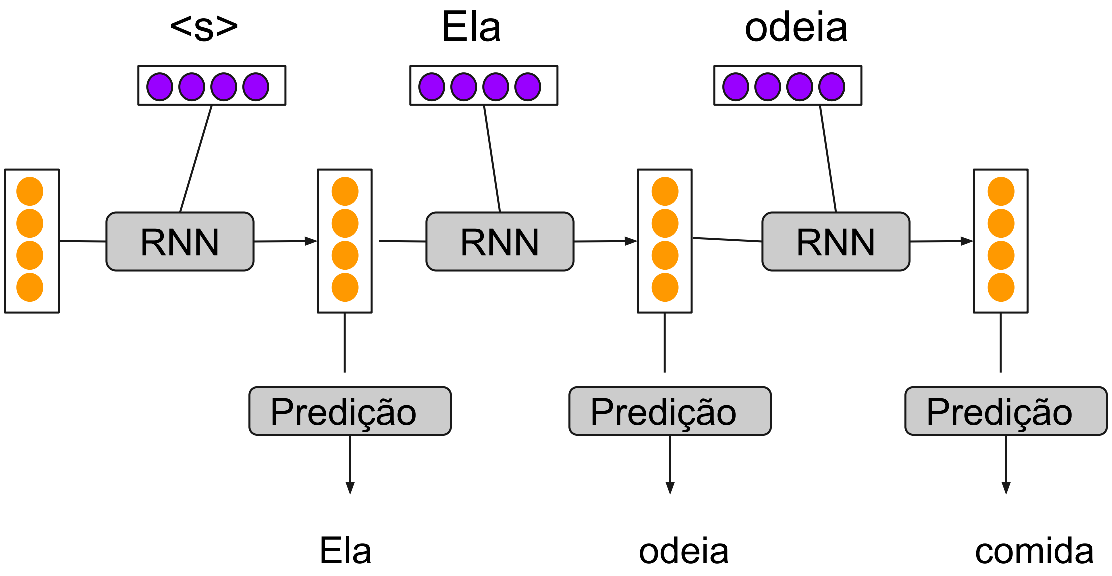
*CMU Neural nets for NLP

Inferência - geração de texto



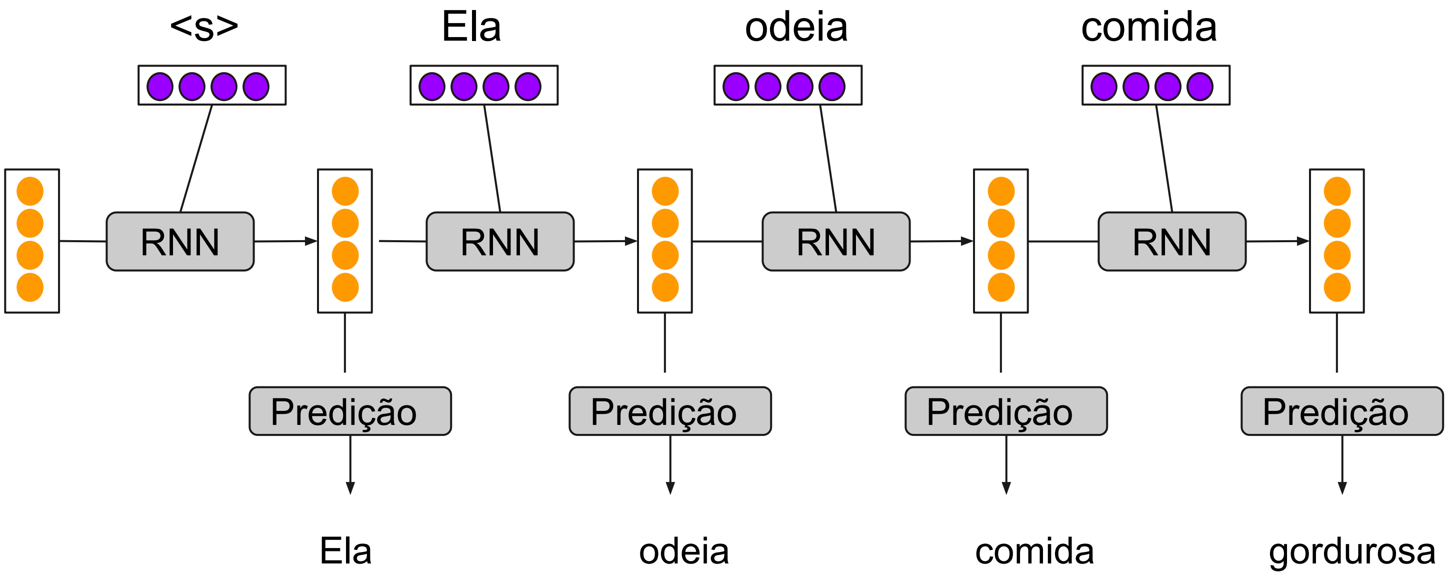
*CMU Neural nets for NLP

Inferência - geração de texto



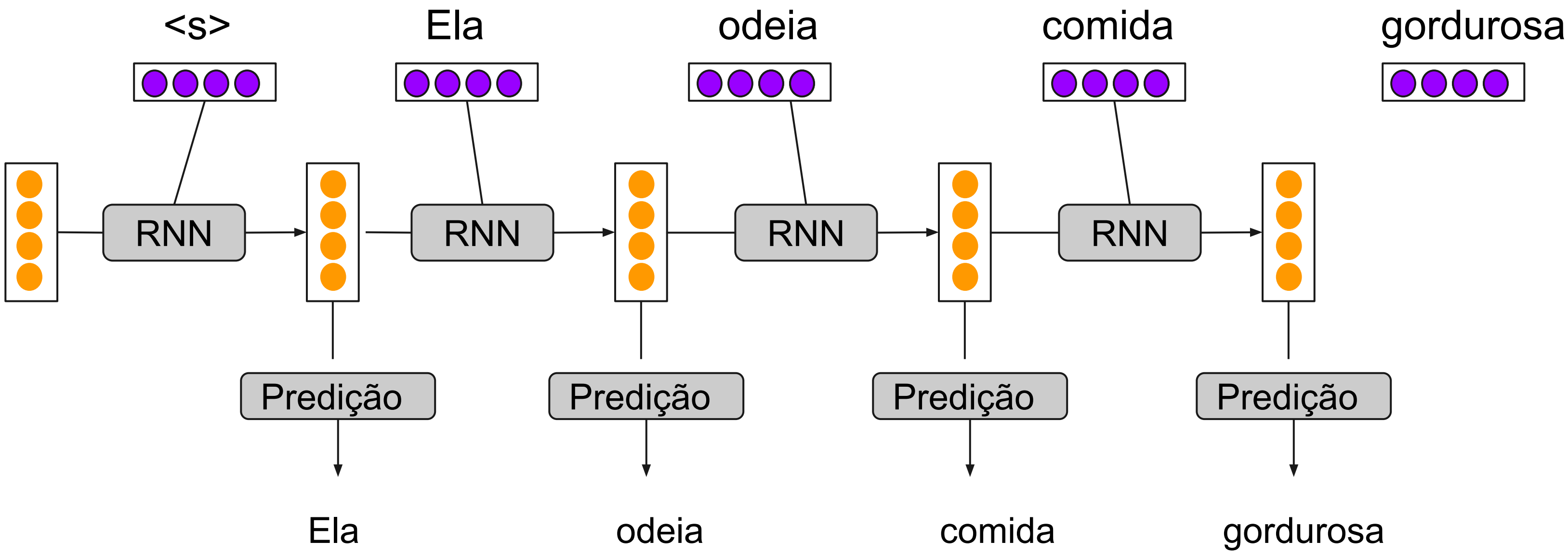
*CMU Neural nets for NLP

Inferência - geração de texto



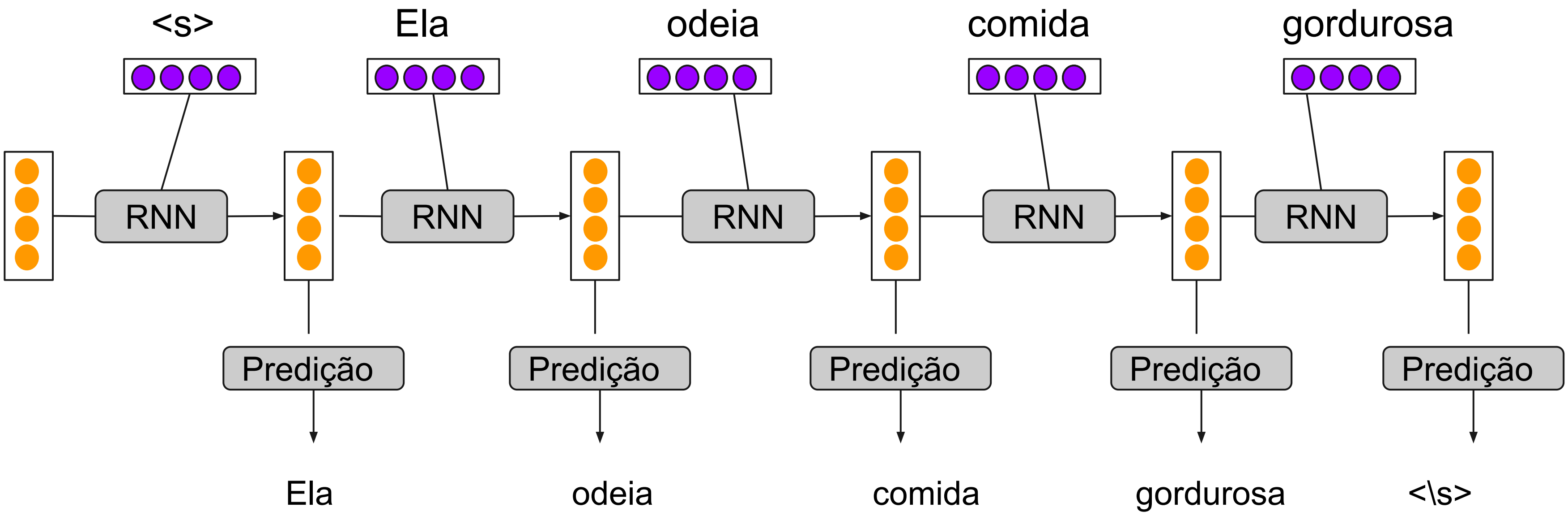
*CMU Neural nets for NLP

Inferência - geração de texto



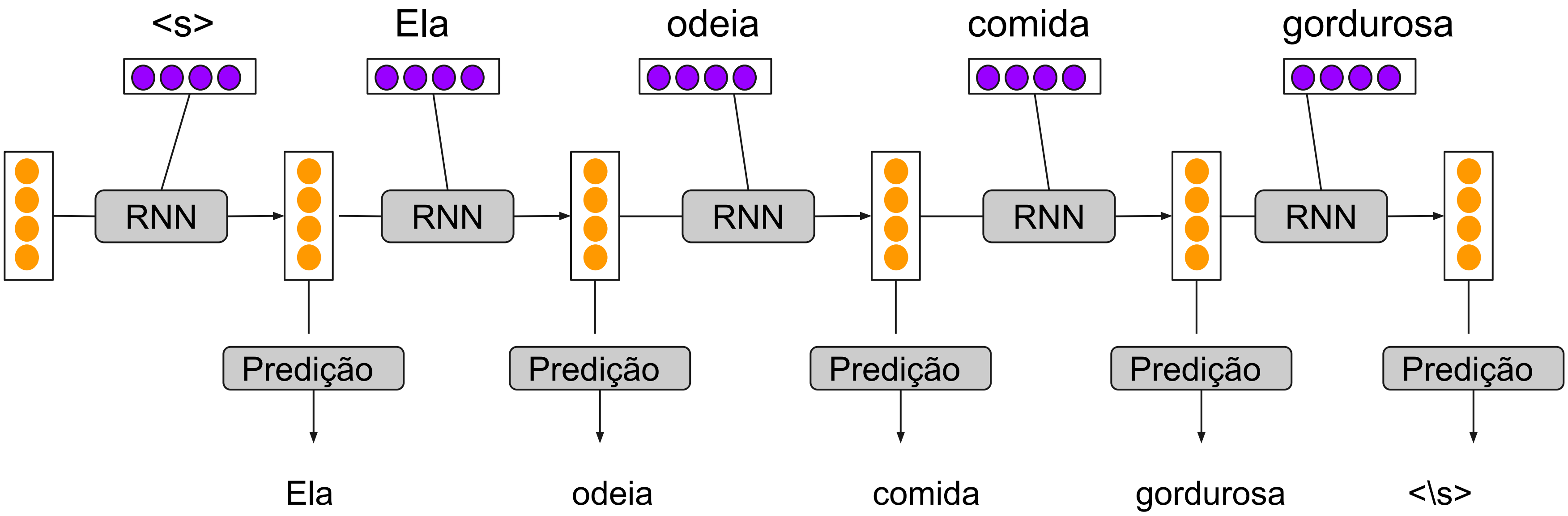
*CMU Neural nets for NLP

Inferência - geração de texto



*CMU Neural nets for NLP

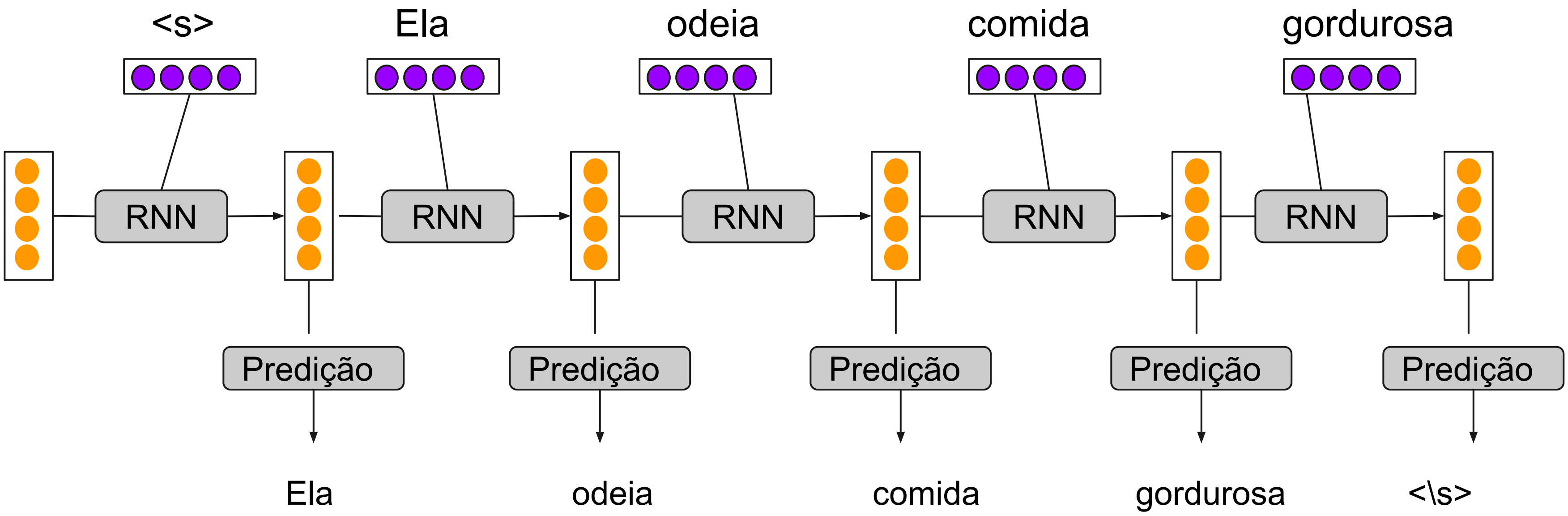
Inferência - geração de texto



Geração autorregressiva

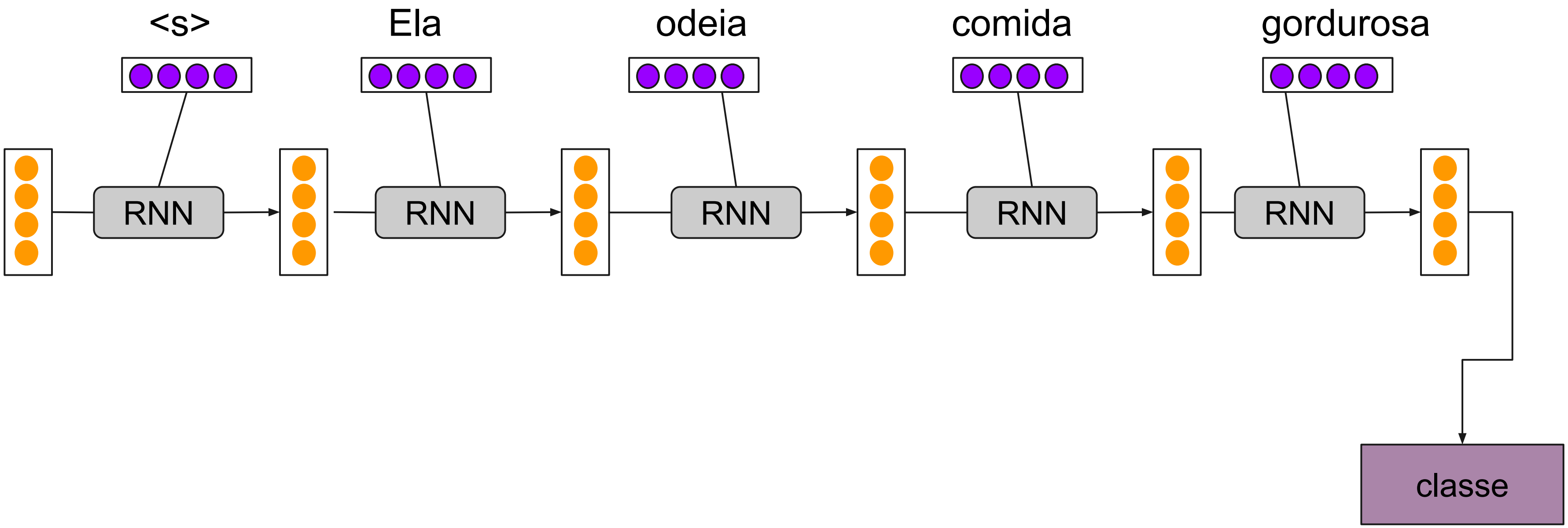
*CMU Neural nets for NLP

Inferência - classificação de texto?



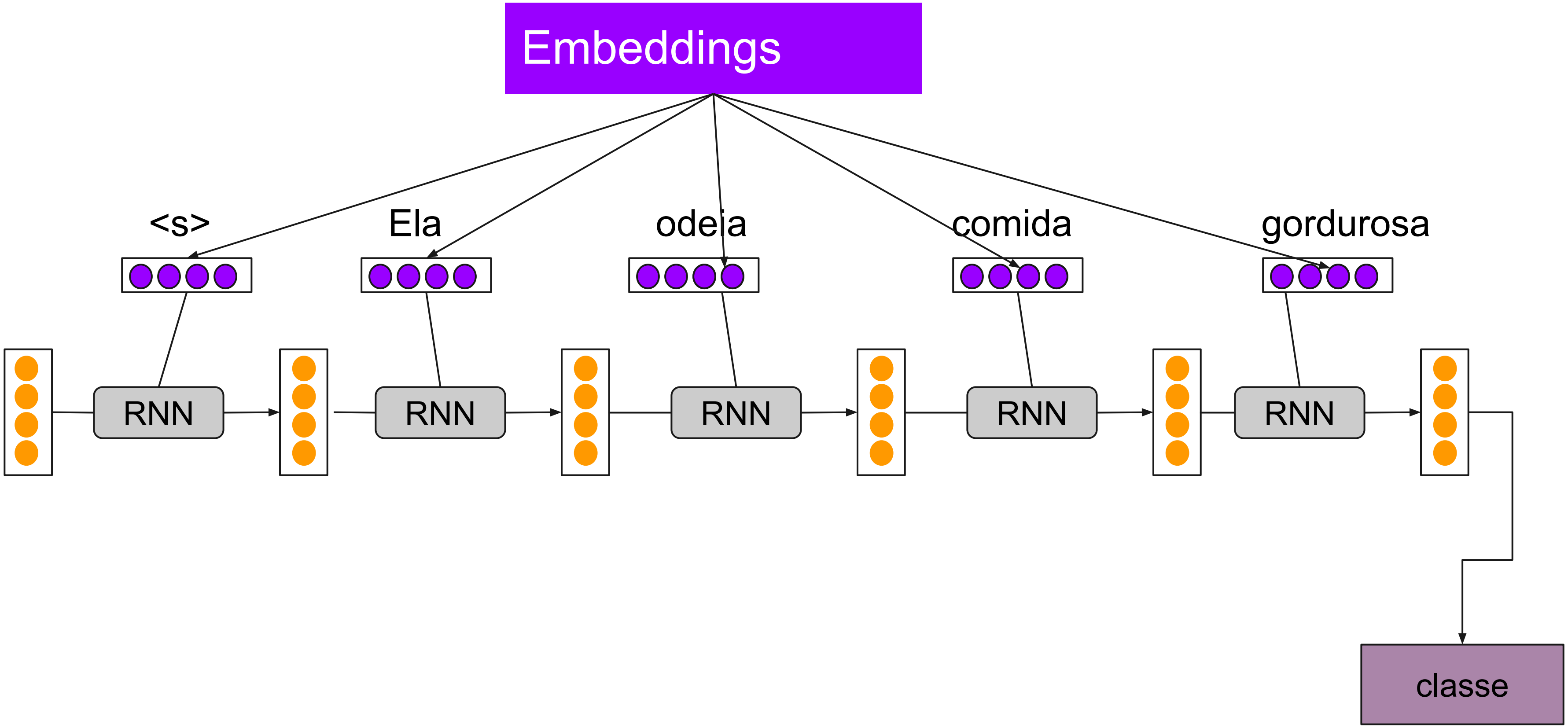
*CMU Neural nets for NLP

Inferência - classificação de texto



*CMU Neural nets for NLP

Inferência - classificação de texto



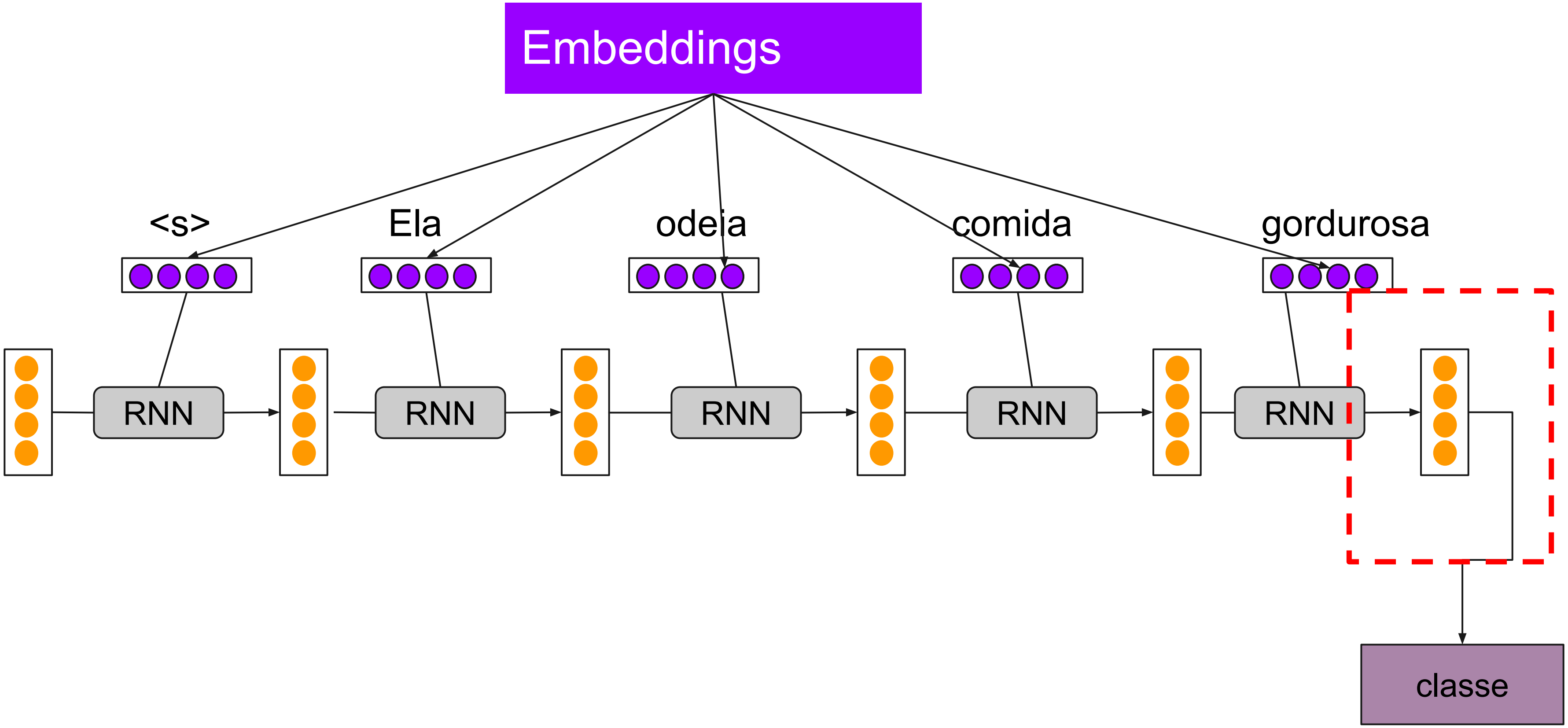
*CMU Neural nets for NLP

Pergunta

Como representar uma sentença com RNN?

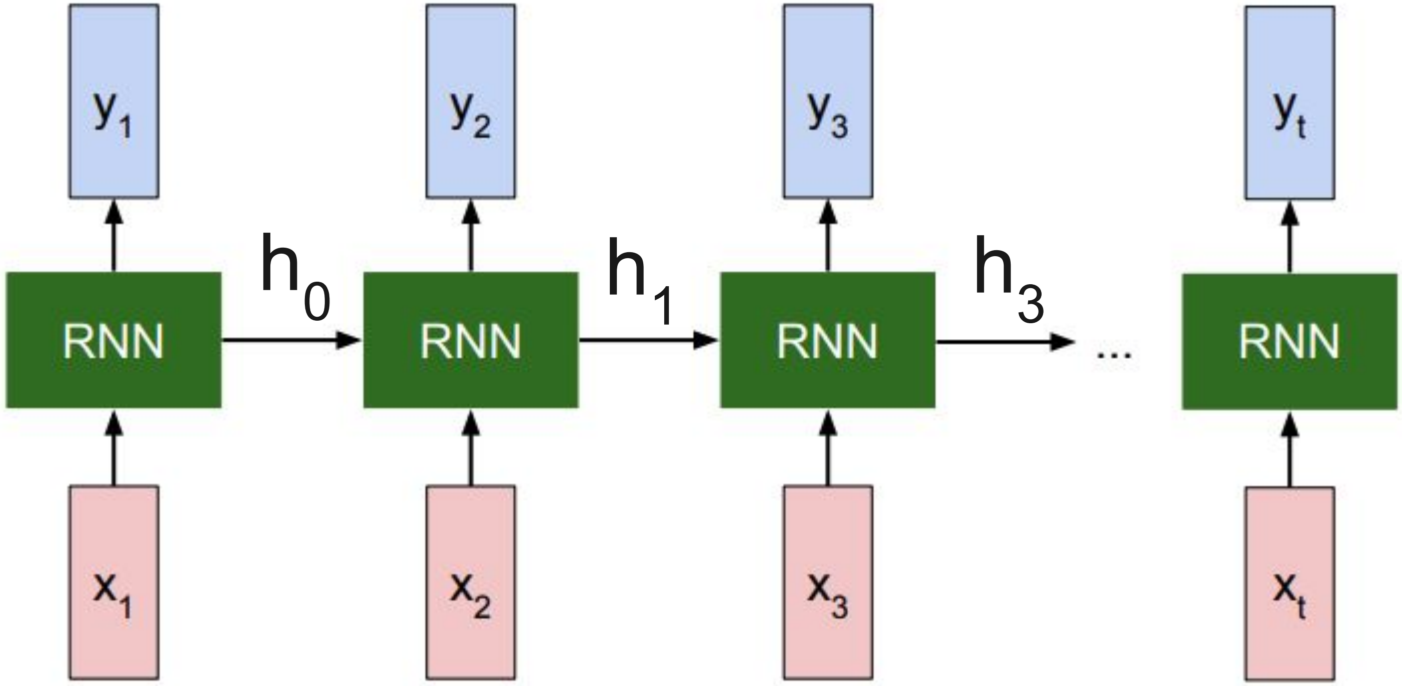
a) Média dos embeddings
b) Saída no instante t
c) Soma dos embeddings
d) Concatenação dos embeddings

Inferência - classificação de texto



*CMU Neural nets for NLP

Rede neural recorrente

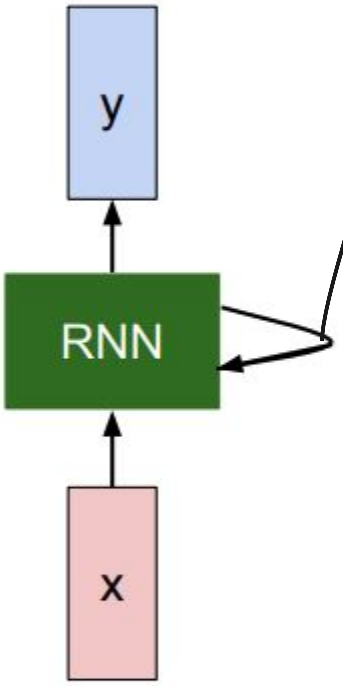


Rede neural recorrente (Elman, 1990)

Fórmula de recorrência

$$\boxed{h_t} = f_{\mathbf{W}}(\boxed{h_{t-1}}, \boxed{x_t})$$

Entrada no instante *t*

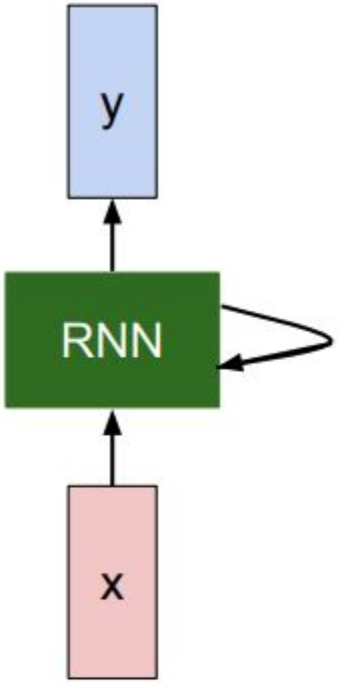


Rede neural recorrente (Elman, 1990)

Fórmula de recorrência

$$\boxed{h_t} = f_{\mathbf{W}}(\boxed{h_{t-1}}, \boxed{x_t})$$

Estado anterior

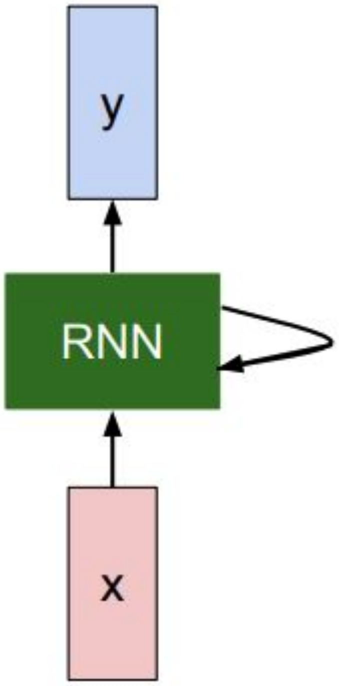


Rede neural recorrente (Elman, 1990)

Fórmula de recorrência

$$h_t = f_{\mathbf{w}}(h_{t-1}, x_t)$$

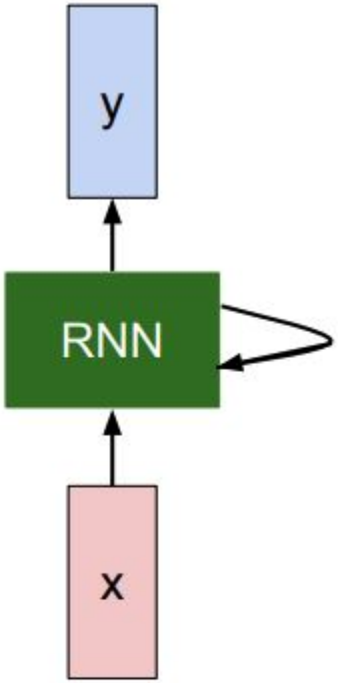
Novo estado



Rede neural recorrente (Elman, 1990)

$$\boxed{h_t} = f_{\mathbf{W}}(\boxed{h_{t-1}}, \boxed{x_t})$$

Função parametrizada

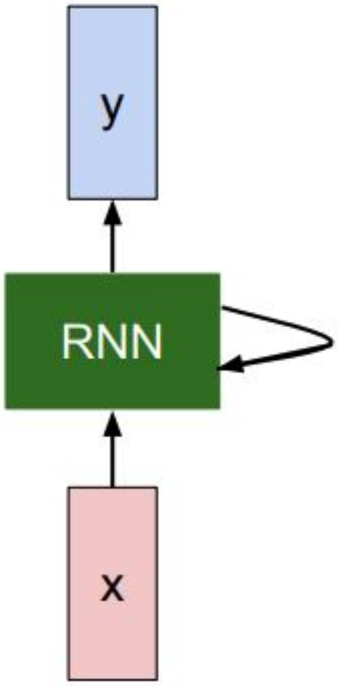


Rede neural recorrente (Elman, 1990)

Fórmula de recorrência

$$\boxed{h_t} = f_{\mathbf{W}}(\boxed{h_{t-1}}, \boxed{x_t})$$

os mesmos parâmetros para todos os estados

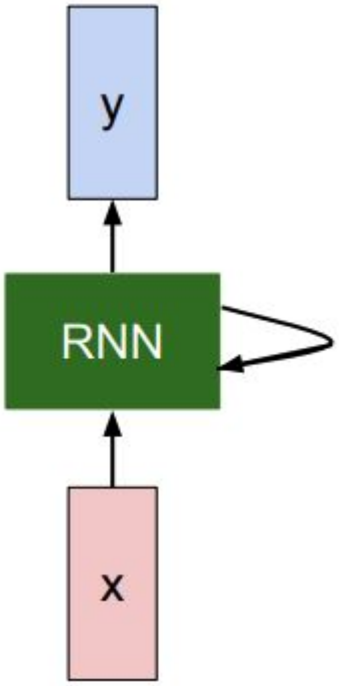


Rede neural recorrente (Elman, 1990)

Fórmula da saída

$$\boxed{y_t} = f_{W'}(\boxed{h_t})$$

Estado novo

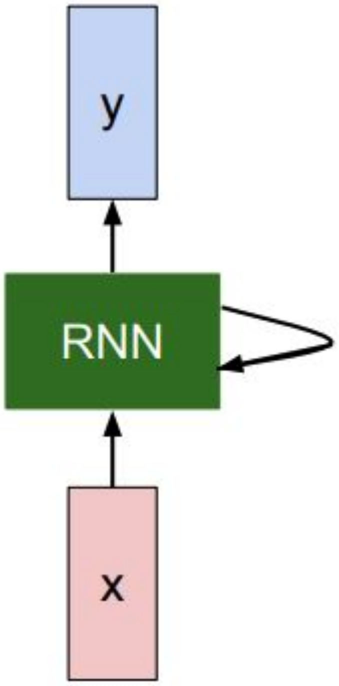


Rede neural recorrente (Elman, 1990)

Fórmula da saída

$$\boxed{y_t} = f_{W'}(\boxed{h_t})$$

Saída

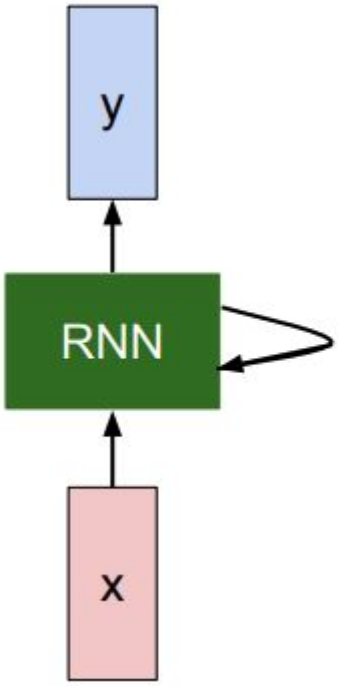


Rede neural recorrente (Elman, 1990)

Fórmula da saída

$$\boxed{y_t} = f_{\mathbf{W'}}(\boxed{h_t})$$

Outra função parametrizada

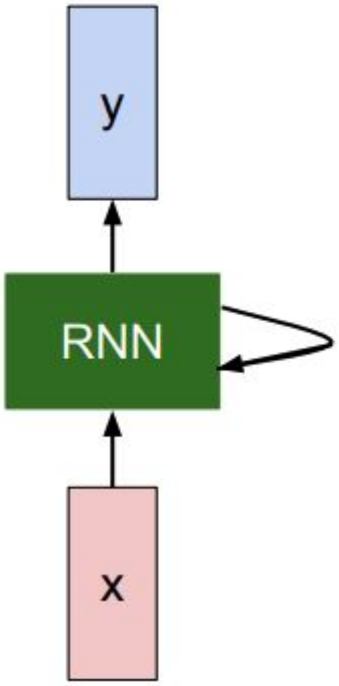


Rede neural recorrente (Elman, 1990)

Fórmula da saída

$$\boxed{y_t} = f_{W'}(\boxed{h_t})$$

COnjunto diferente de parâmetros



Rede neural recorrente (Elman, 1990)

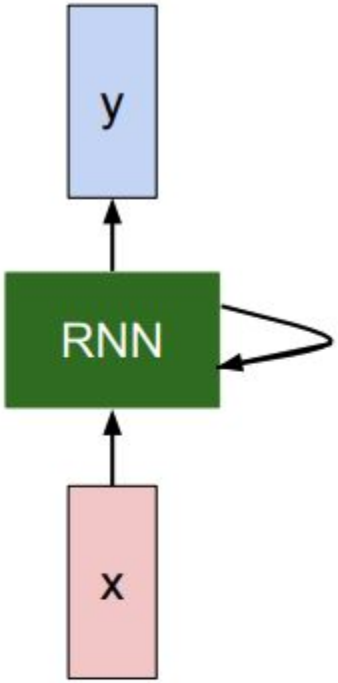
Vanilla RNN

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

$$\downarrow$$

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

$$y_t = W_{hy}h_t$$



cs231:Stanford

Rede neural recorrente (Elman, 1990)

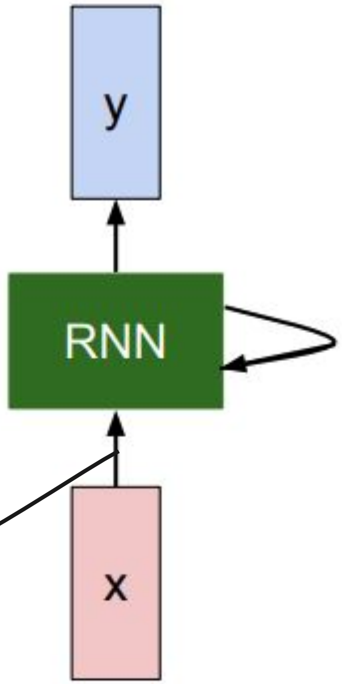
Vanilla RNN

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

$$\downarrow$$

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

$$y_t = W_{hy}h_t$$



cs231:Stanford

Rede neural recorrente (Elman, 1990)

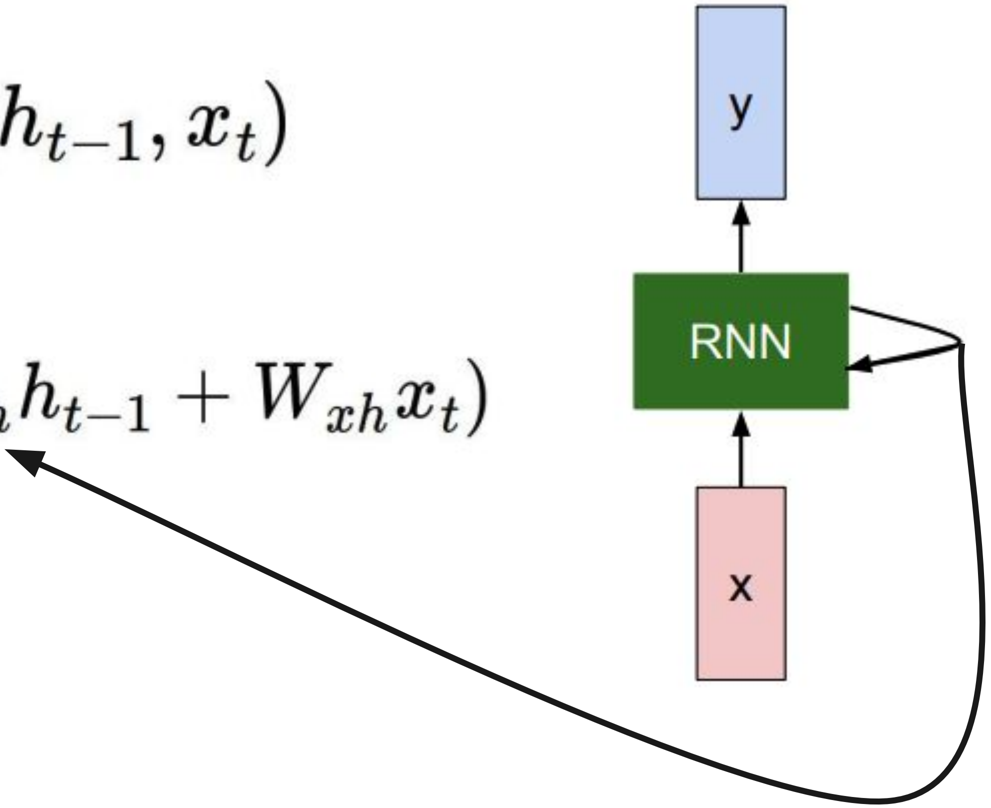
Vanilla RNN

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

$$\downarrow$$

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

$$y_t = W_{hy}h_t$$



https://cs231n.stanford.edu/slides/2024/lecture_7.pdf

Rede neural recorrente (Elman, 1990)

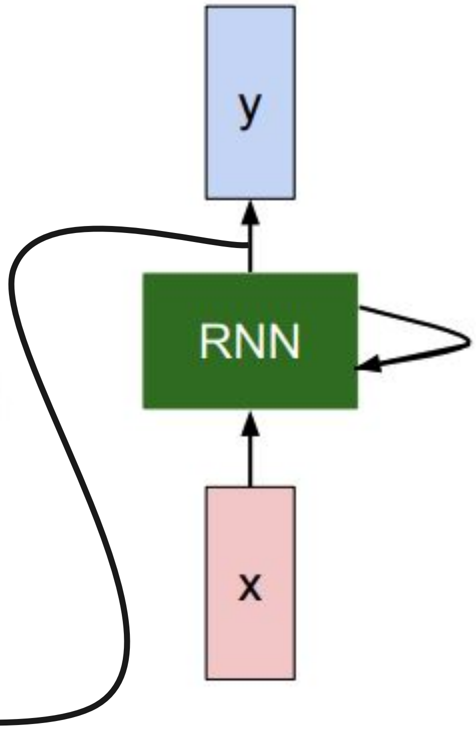
Vanilla RNN

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

$$\Big\downarrow$$

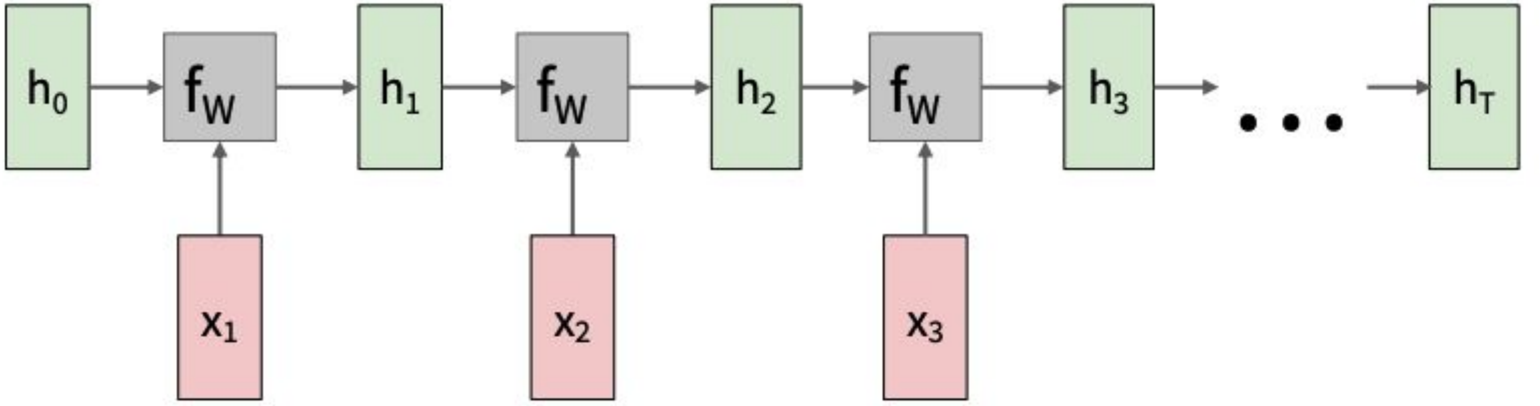
$$h_t = \tanh(W_{hh} h_{t-1} + W_{xh} x_t)$$

$$y_t = W_{hy} h_t$$



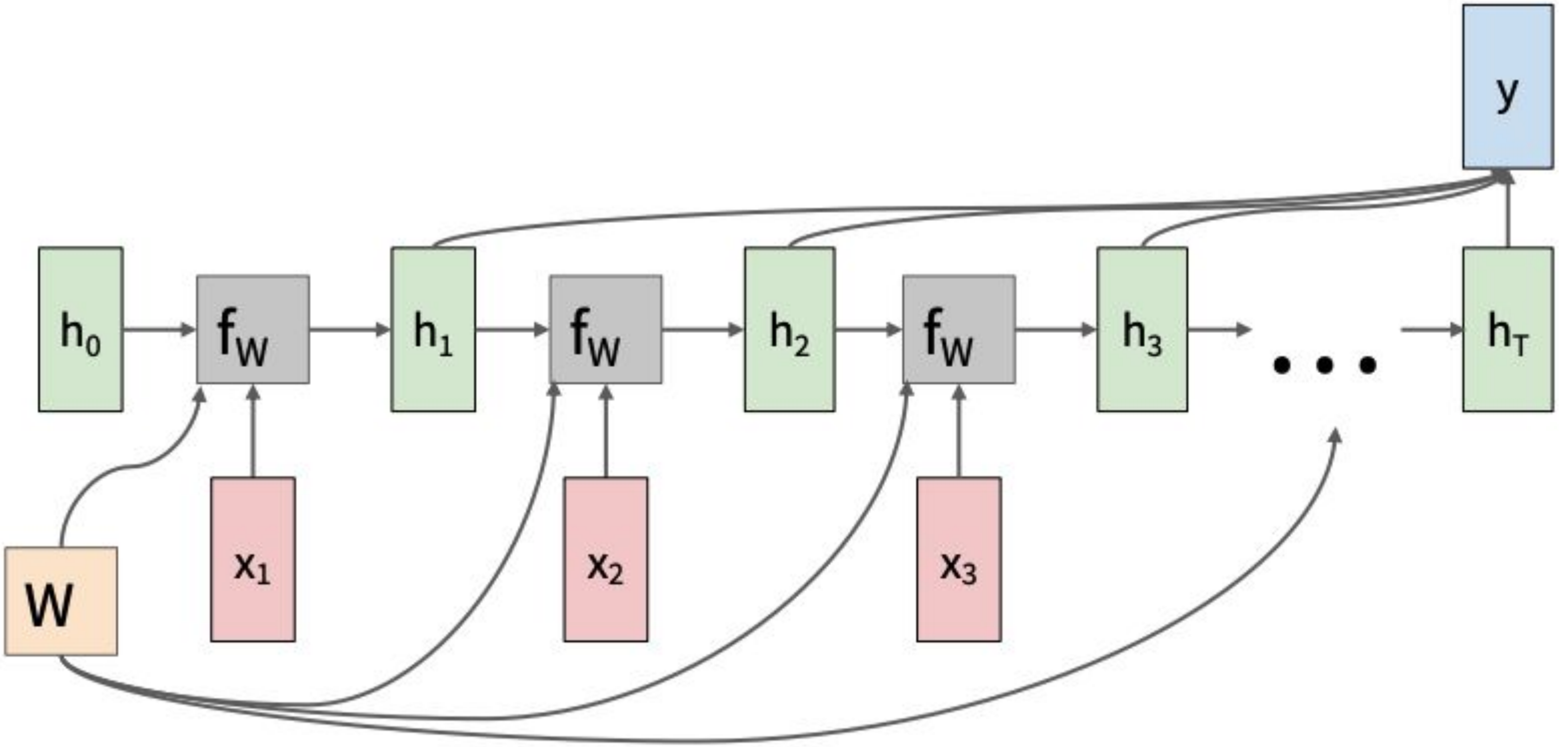
https://cs231n.stanford.edu/slides/2024/lecture_7.pdf

Grafo computational



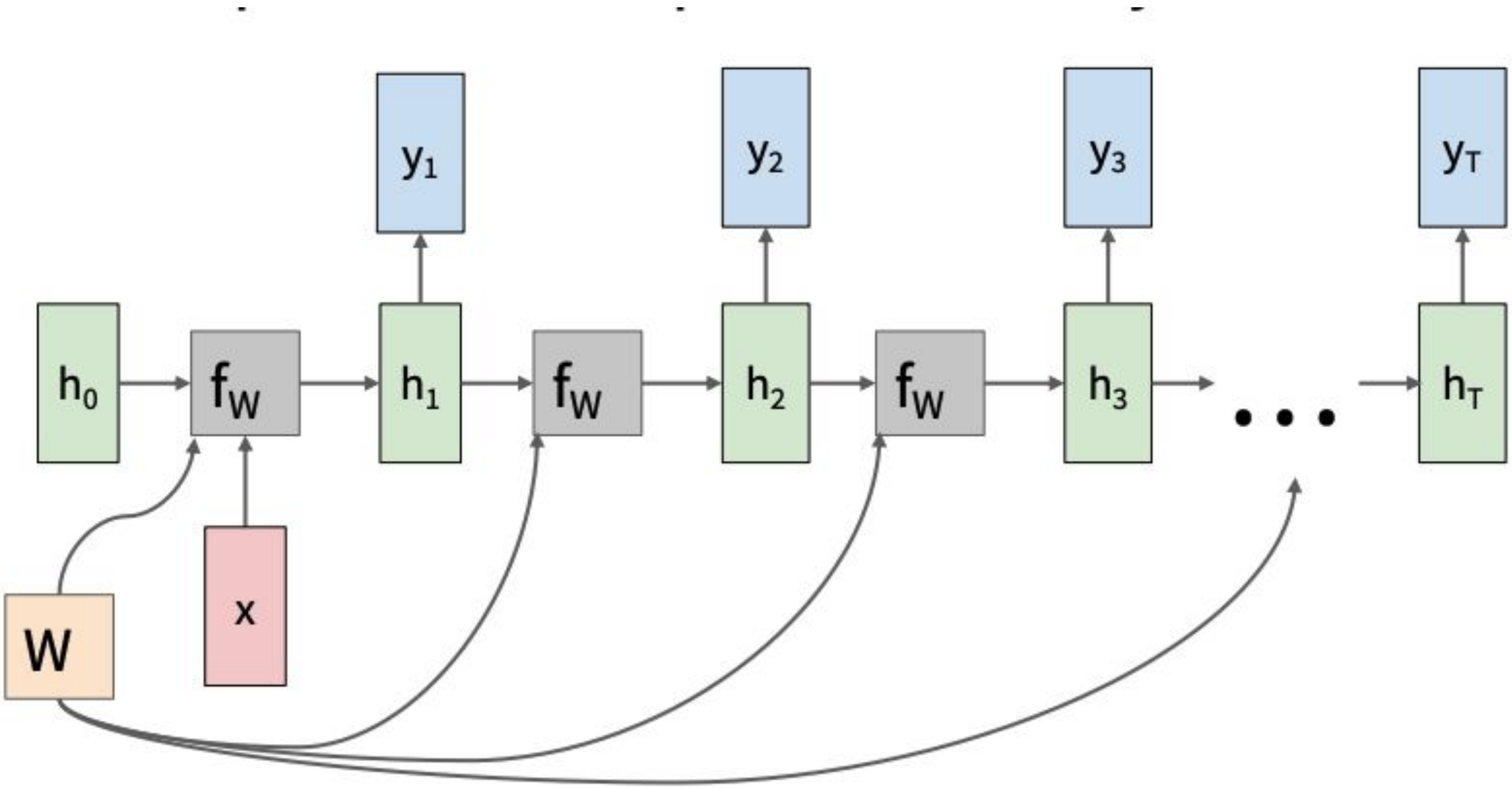
https://cs231n.stanford.edu/slides/2024/lecture_7.pdf

Muitos para um



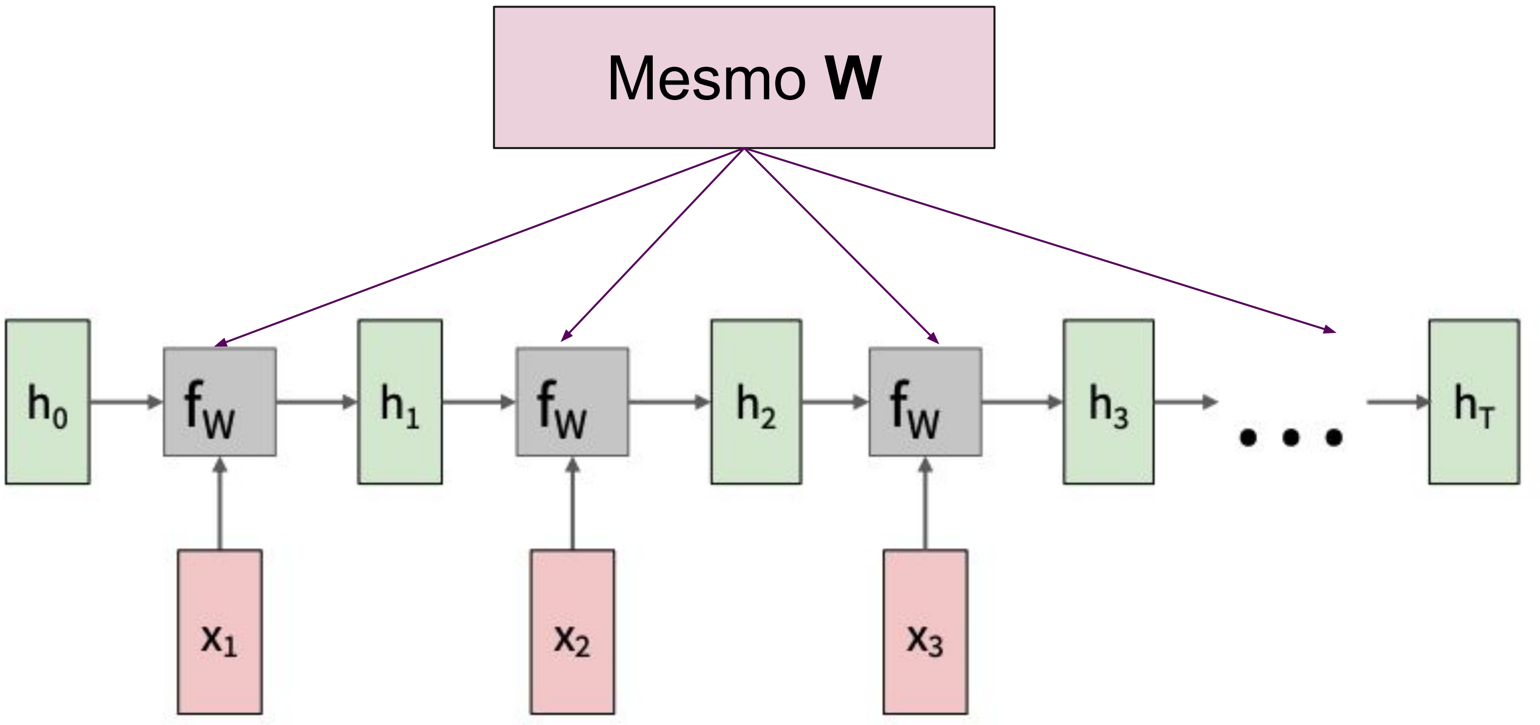
https://cs231n.stanford.edu/slides/2024/lecture_7.pdf

Um para muitos



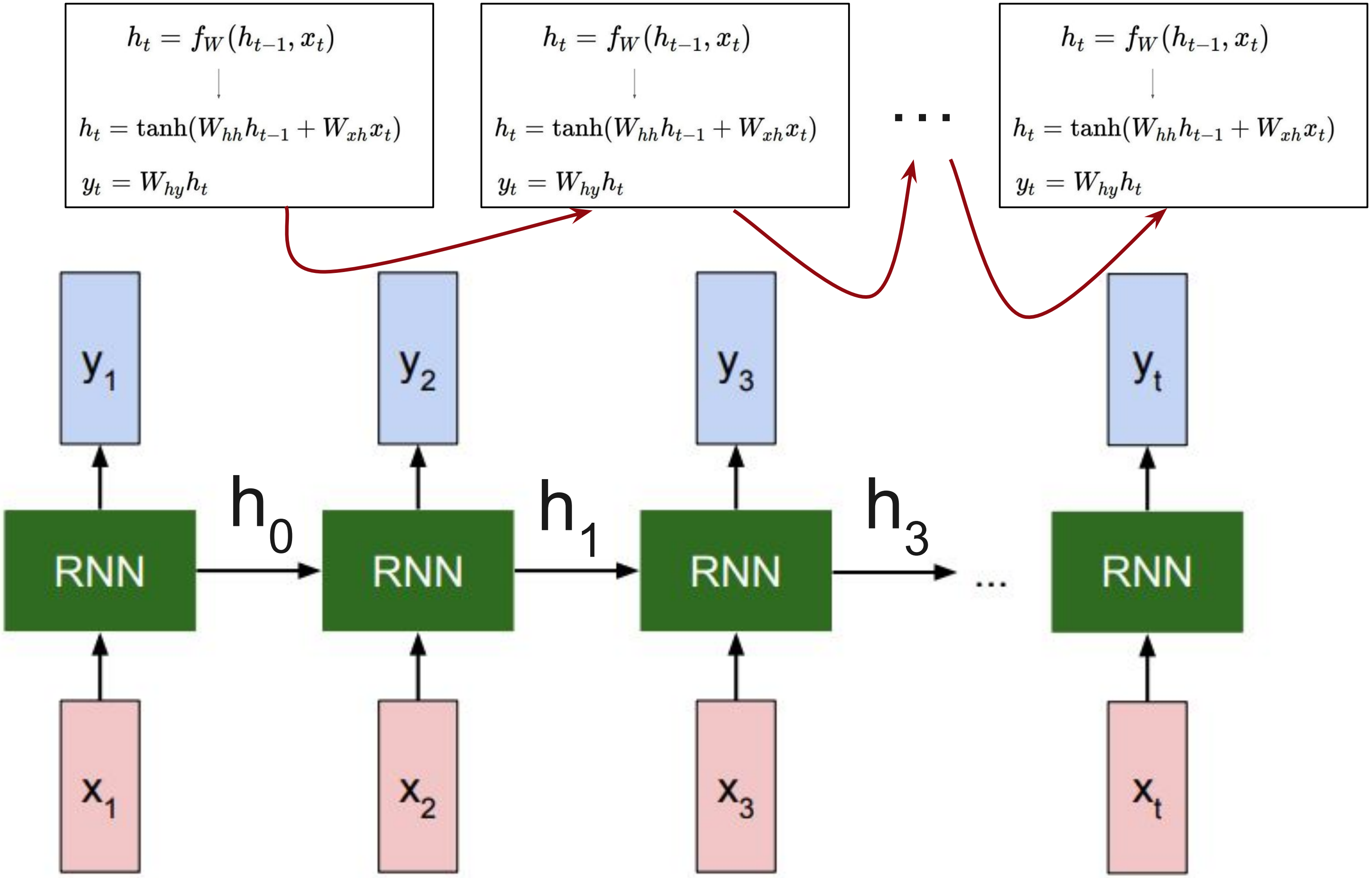
https://cs231n.stanford.edu/slides/2024/lecture_7.pdf

Grafo computational



https://cs231n.stanford.edu/slides/2024/lecture_7.pdf

Rede neural recorrente



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

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

$$y_t = W_{hy}h_t$$

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

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

$$y_t = W_{hy}h_t$$

...

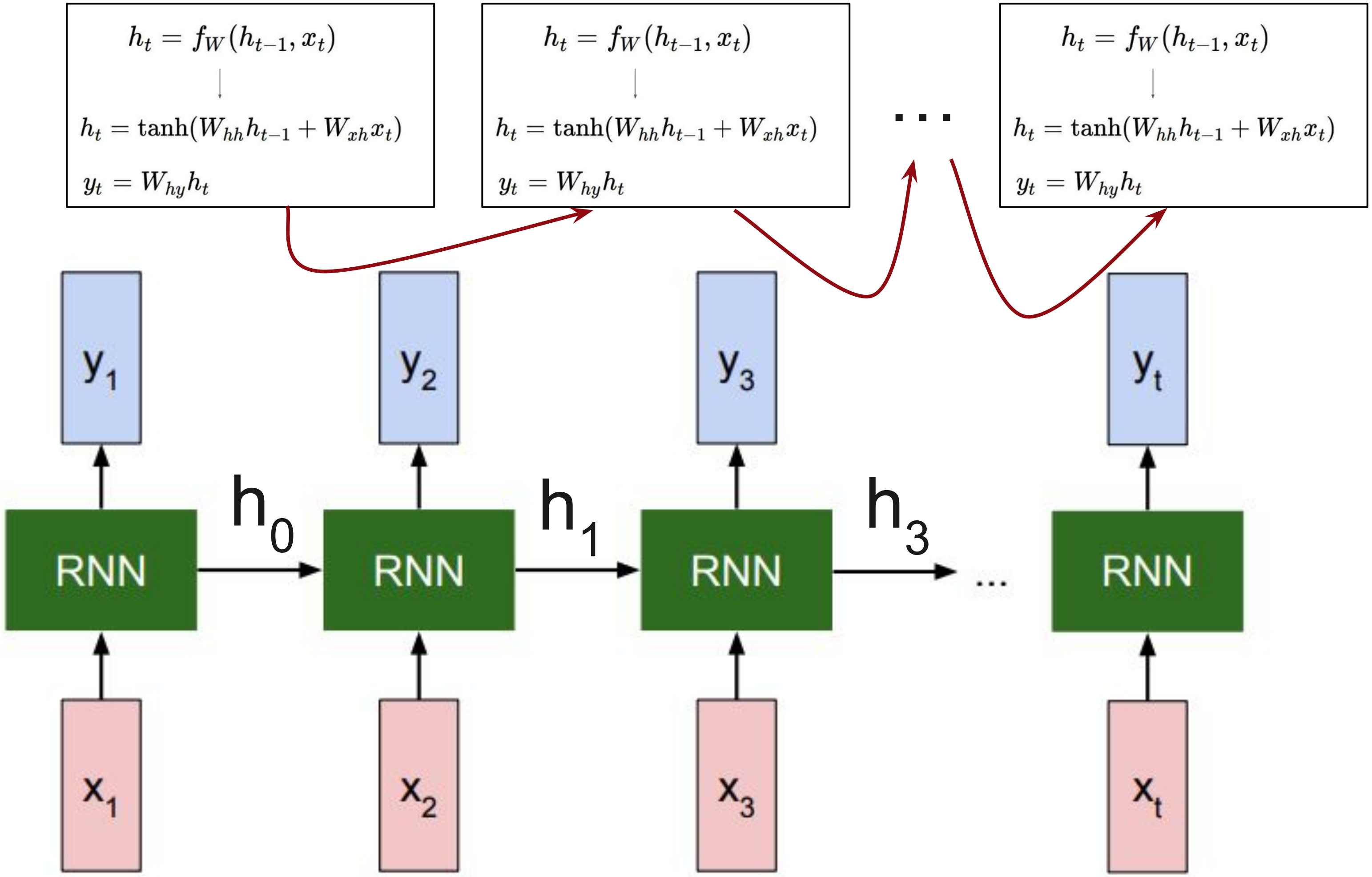
$$h_t = f_W(h_{t-1}, x_t)$$

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

$$y_t = W_{hy}h_t$$

https://cs231n.stanford.edu/slides/2024/lecture_7.pdf

De onde vêm os pesos?



$$h_t = f_W(h_{t-1}, x_t)$$
$$\downarrow$$
$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$$
$$y_t = W_{hy}h_t$$

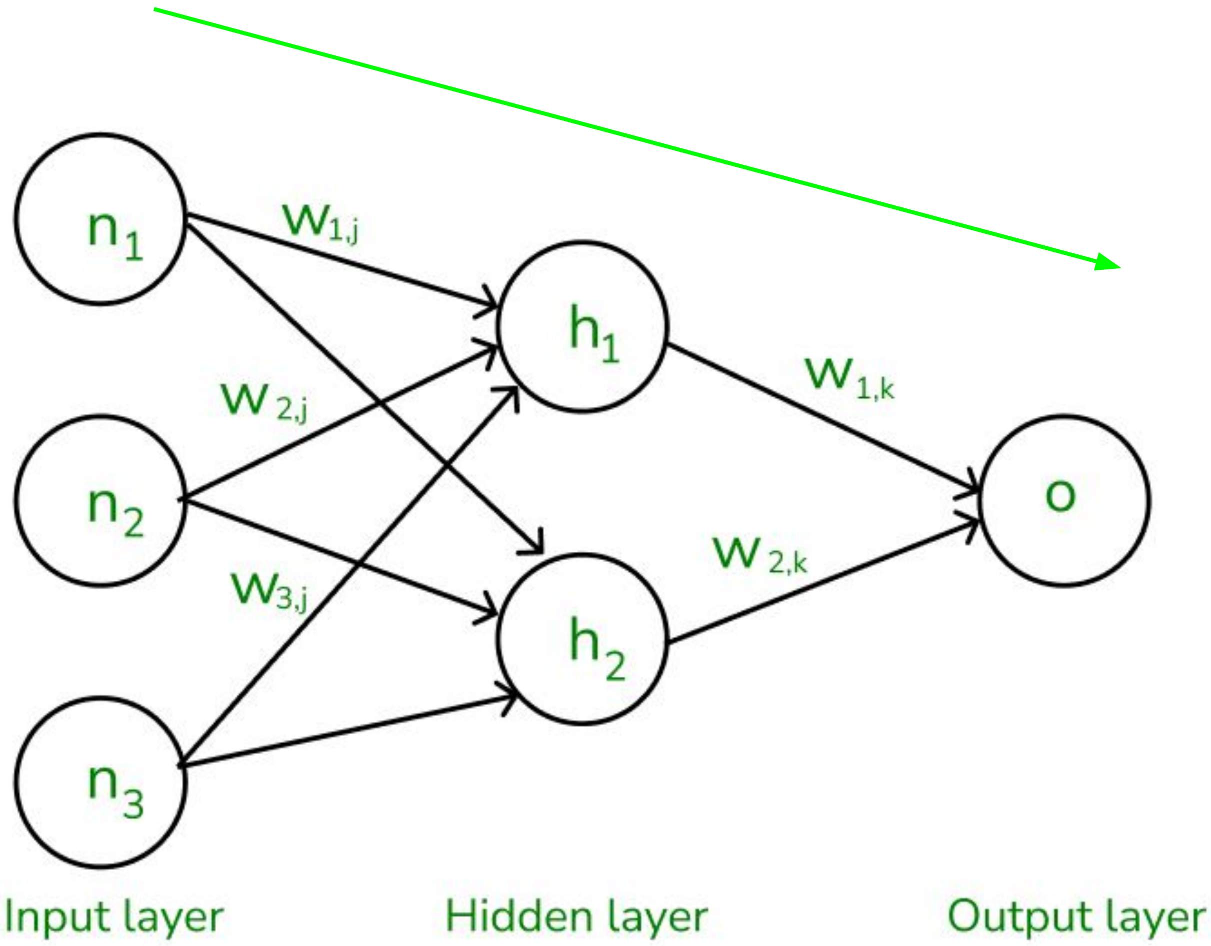
$$h_t = f_W(h_{t-1}, x_t)$$
$$\downarrow$$
$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$$
$$y_t = W_{hy}h_t$$

...

$$h_t = f_W(h_{t-1}, x_t)$$
$$\downarrow$$
$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$$
$$y_t = W_{hy}h_t$$

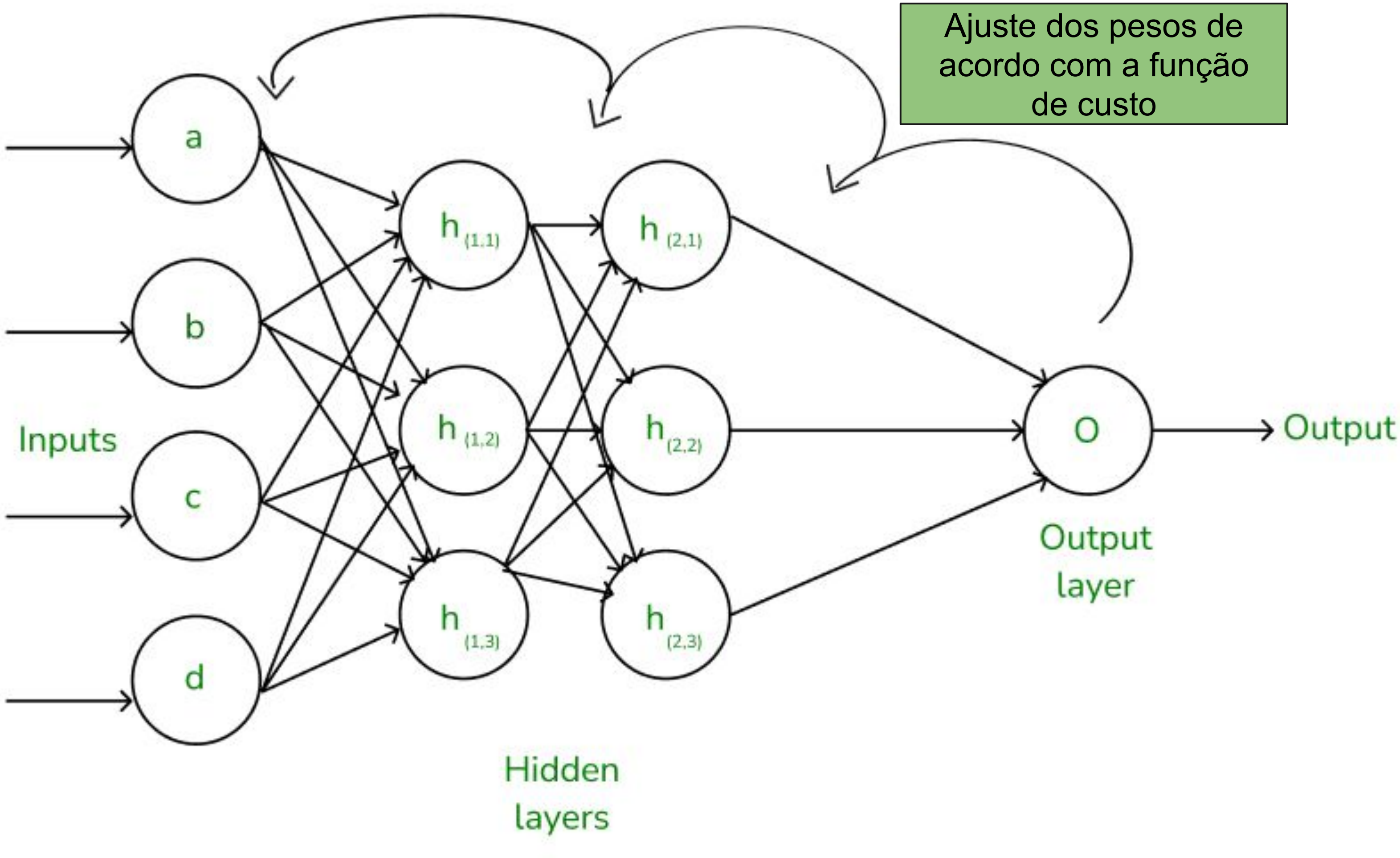
Adaptado de https://cs231n.stanford.edu/slides/2024/lecture_7.pdf

Forward para redes não recorrentes



https://www.geeksforgeeks.org/backpropagation-in-neural-network/

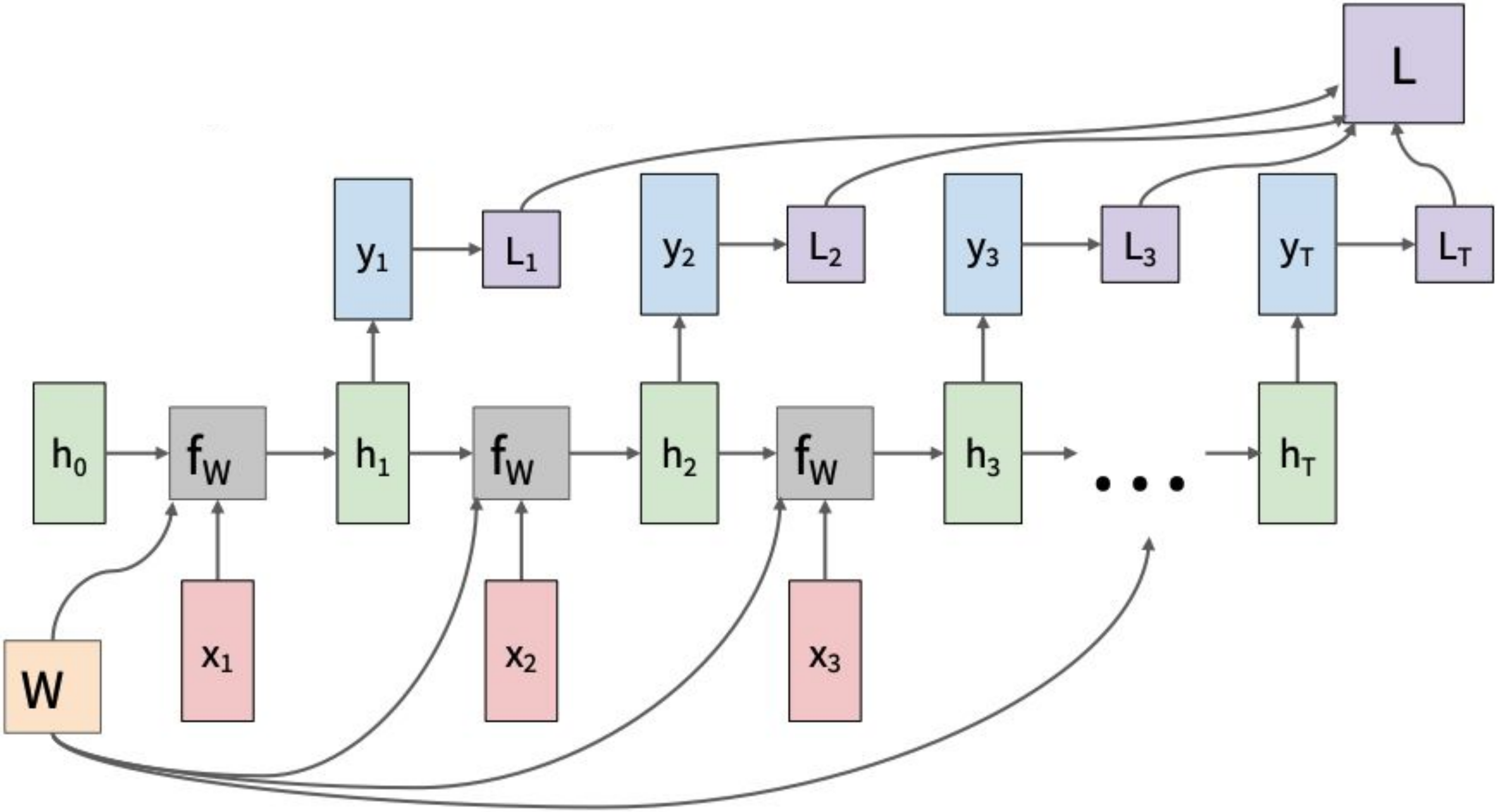
Backpropagation para redes não recorrentes



Ajuste dos pesos de acordo com a função de custo

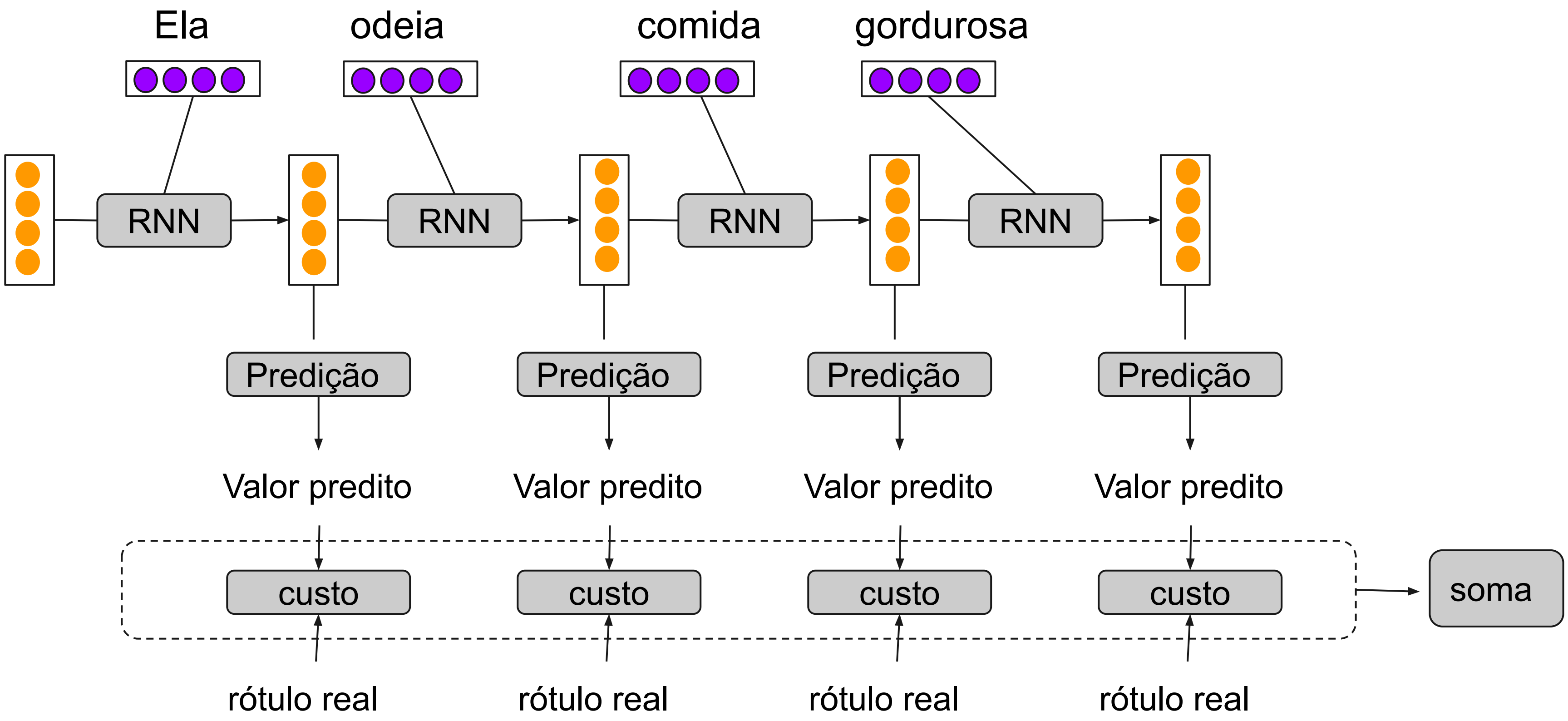
https://www.geeksforgeeks.org/backpropagation-in-neural-network/

Muitos para muitos



https://cs231n.stanford.edu/slides/2024/lecture_7.pdf

Treinamento

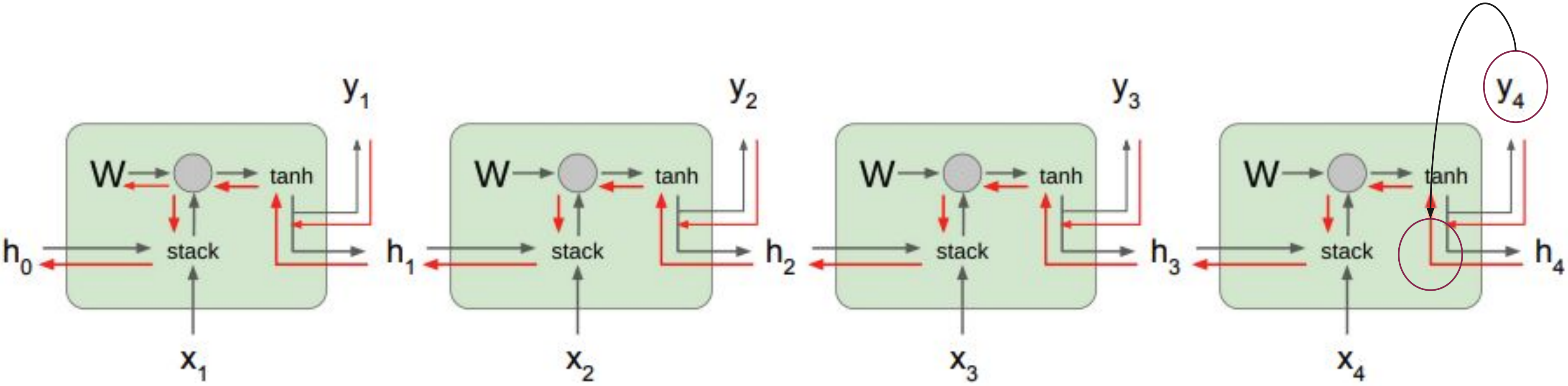


*CMU Neural nets for NLP

Backpropagation through time

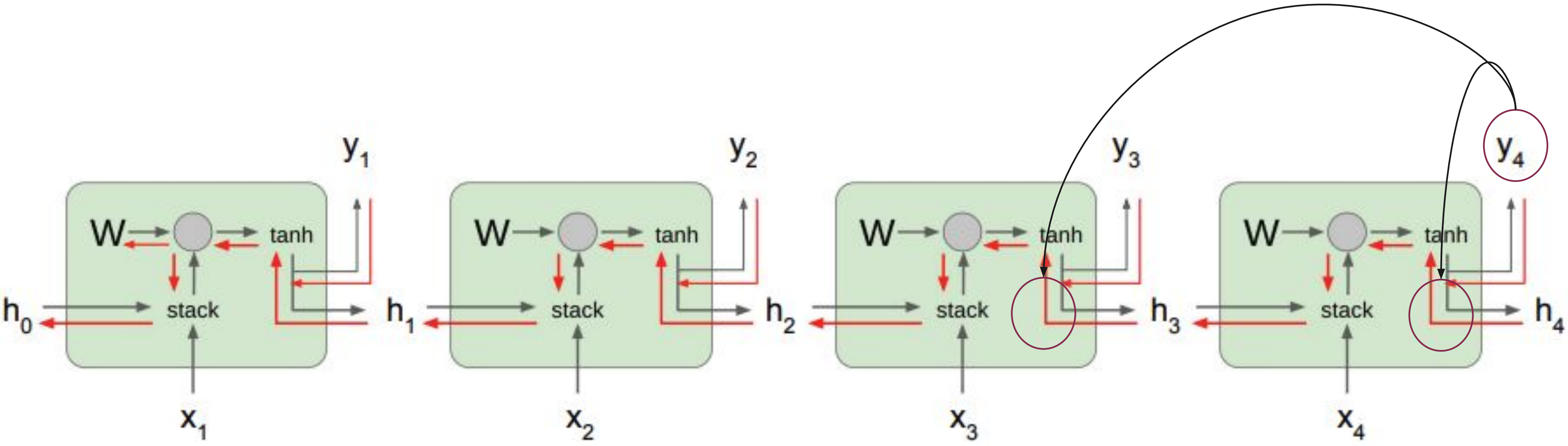
- O estado da camada escondida no instante de tempo *t* contribui para
    - A saída e seu erro associado no tempo *t*
    - A saída e o erro no instante de tempo *t+1*

Treinamento



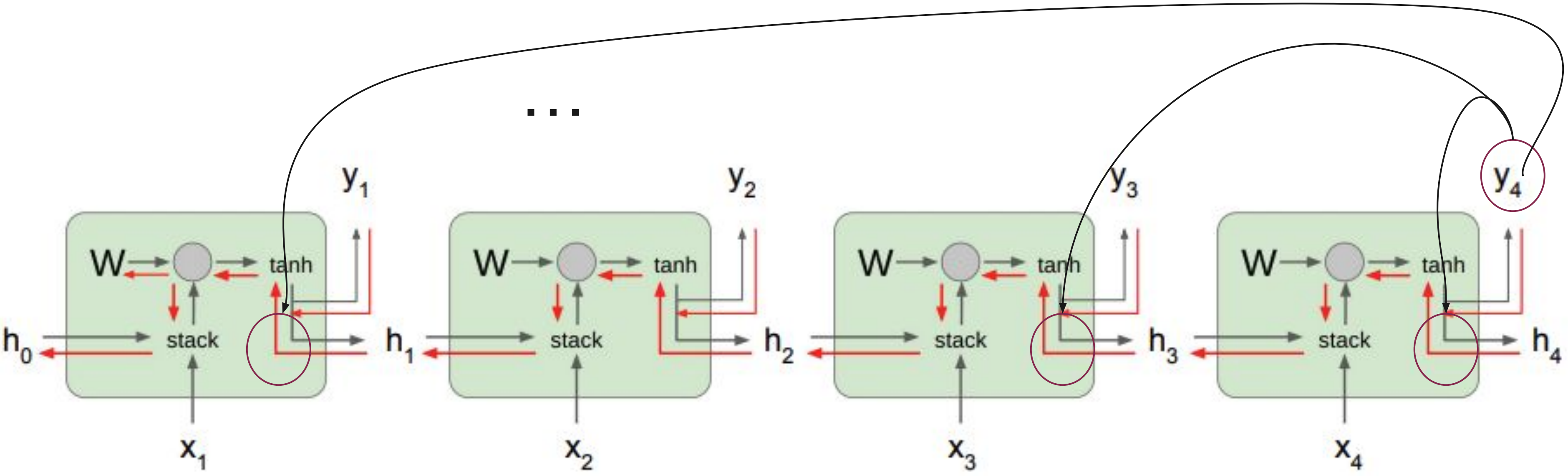
Bengio et al, “Learning long-term dependencies with gradient descent is difficult”, IEEE Transactions on Neural Networks, 1994 Pascanu et al, “On the difficulty of training recurrent neural networks”, ICML 2013 Cs231n. Stanford 2022

Treinamento



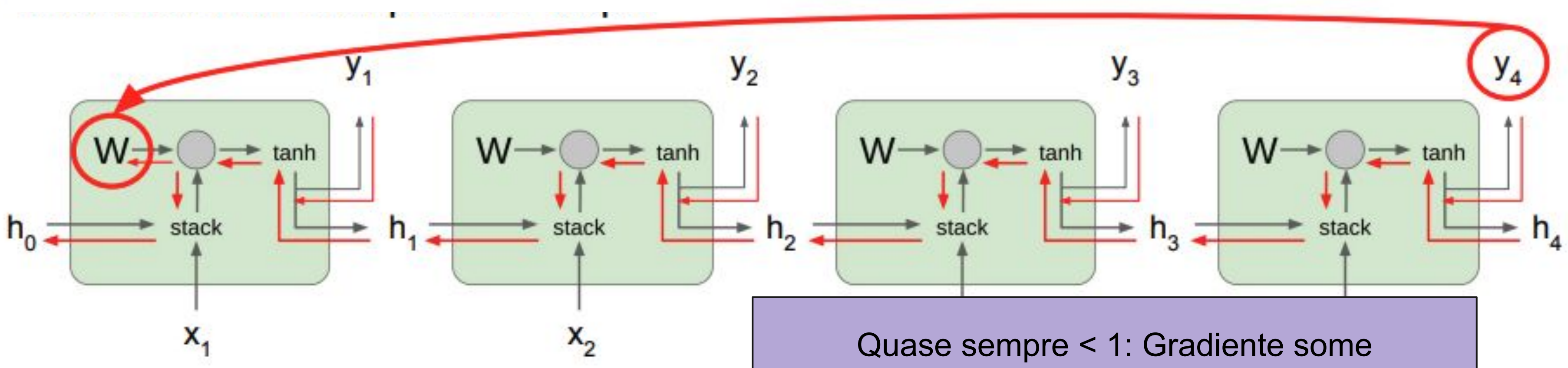
Bengio et al, “Learning long-term dependencies with gradient descent is difficult”, IEEE Transactions on Neural Networks, 1994 Pascanu et al, “On the difficulty of training recurrent neural networks”, ICML 2013
Cs231n. Stanford 2022

Treinamento



Bengio et al, “Learning long-term dependencies with gradient descent is difficult”, IEEE Transactions on Neural Networks, 1994 Pascanu et al, “On the difficulty of training recurrent neural networks”, ICML 2013 Cs231n. Stanford 2022

Treinamento



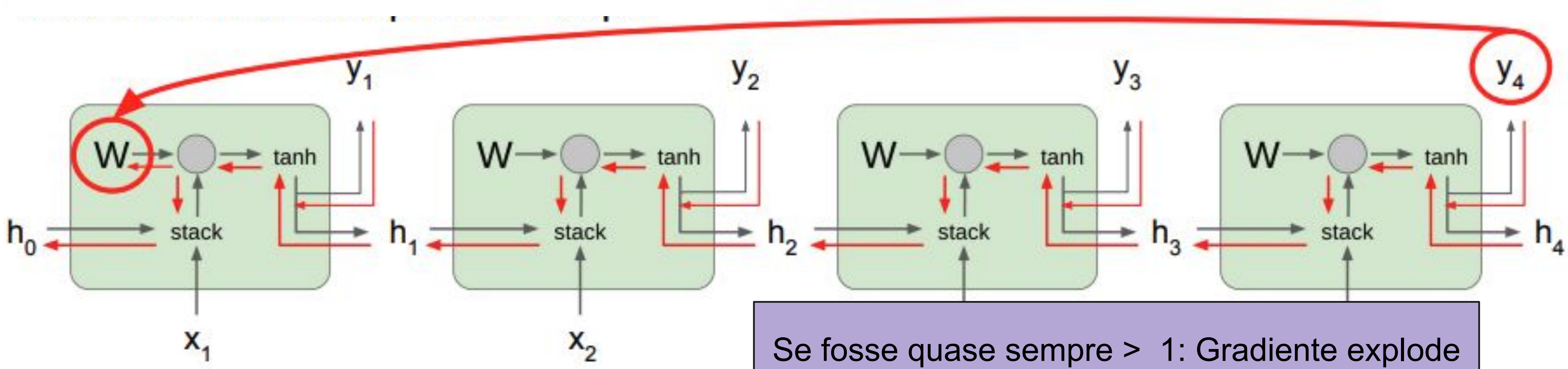
Quase sempre < 1: Gradiente some

$$\frac{\partial L}{\partial W} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W} \qquad \boxed{\frac{\partial h_t}{\partial h_{t-1}} = tanh'(W_{hh}h_{t-1} + W_{xh}x_t)W_{hh}}$$

$$\frac{\partial L_T}{\partial W} = \frac{\partial L_T}{\partial h_T}\frac{\partial h_t}{\partial h_{t-1}} \cdots \frac{\partial h_1}{\partial W} = \frac{\partial L_T}{\partial h_T}\left(\prod_{t=2}^{T}\boxed{\frac{\partial h_t}{\partial h_{t-1}}}\right)\frac{\partial h_1}{\partial W}$$

Bengio et al, “Learning long-term dependencies with gradient descent is difficult”, IEEE Transactions on Neural Networks, 1994 Pascanu et al, “On the difficulty of training recurrent neural networks”, ICML 2013 Cs231n. Stanford 2022

Treinamento



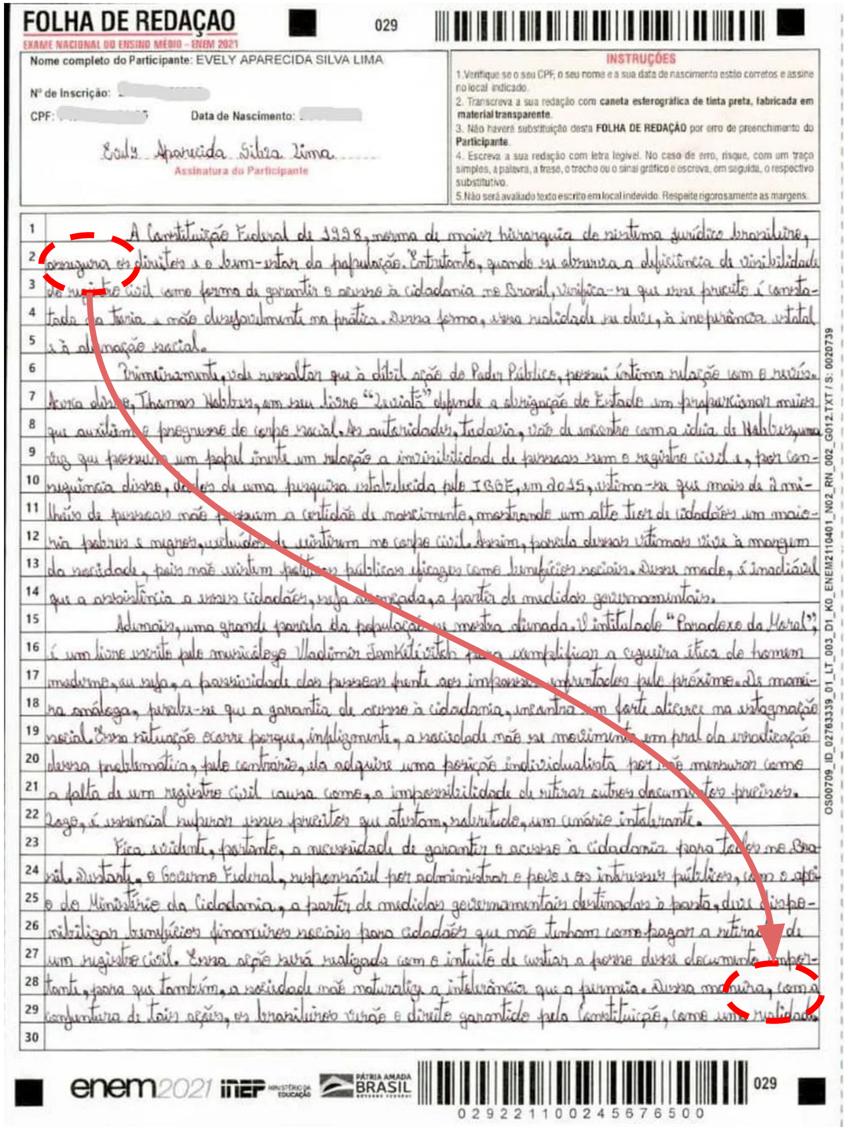
Se fosse quase sempre > 1: Gradiente explode

$$\frac{\partial L}{\partial W} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W} \qquad \boxed{\frac{\partial h_t}{\partial h_{t-1}} = tanh'(W_{hh}h_{t-1} + W_{xh}x_t)W_{hh}}$$

$$\frac{\partial L_T}{\partial W} = \frac{\partial L_T}{\partial h_T}\frac{\partial h_t}{\partial h_{t-1}} \cdots \frac{\partial h_1}{\partial W} = \frac{\partial L_T}{\partial h_T}\left(\prod_{t=2}^{T} \boxed{\frac{\partial h_t}{\partial h_{t-1}}}\right)\frac{\partial h_1}{\partial W}$$

Bengio et al, “Learning long-term dependencies with gradient descent is difficult”, IEEE Transactions on Neural Networks, 1994 Pascanu et al, “On the difficulty of training recurrent neural networks”, ICML 2013 Cs231n. Stanford 2022

Mas e se eu precisar processar uma dependência de longo prazo?



FOLHA DE REDAÇÃO 029

EXAME NACIONAL DO ENSINO MÉDIO – ENEM 2021

Nome completo do Participante: EVELY APARECIDA SILVA LIMA

Nº de Inscrição: [redacted]

CPF: [redacted] Data de Nascimento: [redacted]

Evely Aparecida Silva Lima

Assinatura do Participante

INSTRUÇÕES

1. Verifique se o seu CPF, o seu nome e a sua data de nascimento estão corretos e assine no local indicado.
2. Transcreva a sua redação com **caneta esferográfica de tinta preta, fabricada em material transparente**.
3. Não haverá substituição desta **FOLHA DE REDAÇÃO** por erro de preenchimento do **Participante**.
4. Escreva a sua redação com letra legível. No caso de erro, risque, com um traço simples, a palavra, a frase, o trecho ou o sinal gráfico e escreva, em seguida, o respectivo substitutivo.
5. Não será avaliado texto escrito em local indevido. Respeite rigorosamente as margens.

A Constituição Federal de 1988, norma de maior hierarquia do sistema jurídico brasileiro, assegura os direitos e o bem-estar da população. Entretanto, quando se observa a deficiência de visibilidade do registro civil como forma de garantir o acesso à cidadania no Brasil, verifica-se que esse preceito é constatado na teoria e não devidamente na prática. Dessa forma, essa realidade se deve à ineficiência estatal e à alienação social.

Primeiramente, vale ressaltar que a débil ação do Poder Público possui íntima relação com o revés. Acerca disso, Thomas Hobbes, em seu livro "Leviatã" defende a obrigação do Estado em proporcionar meios que auxiliem o progresso do corpo social. As autoridades, todavia, vão de encontro com a ideia de Hobbes, uma vez que possuem um papel inerte em relação a invisibilidade de pessoas sem o registro civil e, por consequência disso, dados de uma pesquisa estabelecida pelo IBGE, em 2015, estima-se que mais de 3 milhões de pessoas não possuem a certidão de nascimento, mostrando um alto teor de cidadãos, em maioria pobres e negros, excluídos de existirem no corpo civil. Assim, parcela dessas vítimas vive à margem da sociedade, pois não existem políticas públicas eficazes como benefícios sociais. Desse modo, é inadiável que a assistência a esses cidadãos, seja alcançada, a partir de medidas governamentais.

Ademais, uma grande parcela da população se mostra alienada. O intitulado "Paradoxo da Moral", é um livro escrito pelo musicólogo Vladimir Jankélévitch para exemplificar a cegueira ética do homem moderno, ou seja, a passividade das pessoas frente aos impasses enfrentados pelo próximo. De maneira análoga, percebe-se que a garantia de acesso à cidadania, encontra um forte alicerce na estagnação social. Essa situação ocorre porque, infelizmente, a sociedade não se movimenta em prol da erradicação dessa problemática, pelo contrário, ela adquire uma posição individualista por não mensurar como a falta de um registro civil causa como, a impossibilidade de retirar outros documentos precisos. Logo, é essencial superar esses preceitos que atestam, sobretudo, um cenário intolerante.

Fica evidente, portanto, a necessidade de garantir o acesso à cidadania para todos no Brasil. Destarte, o Governo Federal, responsável por administrar o povo e os interesses públicos, com o apoio do Ministério da Cidadania, a partir de medidas governamentais destinadas à pasta, deve disponibilizar benefícios financeiros sociais para cidadãos que não tinham como pagar a retirada de um registro civil. Essa ação será realizada com o intuito de custear a posse desse documento importante, para que também, a sociedade não naturalize a intolerância que a permeia. Dessa maneira, com a conjuntura de tais ações, os brasileiros terão o direito garantido pela Constituição, como uma realidade

OS00709_ID_02763339_01_LT_003_D1_K0_ENEM2110401_N02_RN_002_G012.TXT / S: 00207739

enem2021 INEP MINISTÉRIO DA EDUCAÇÃO PÁTRIA AMADA BRASIL

029

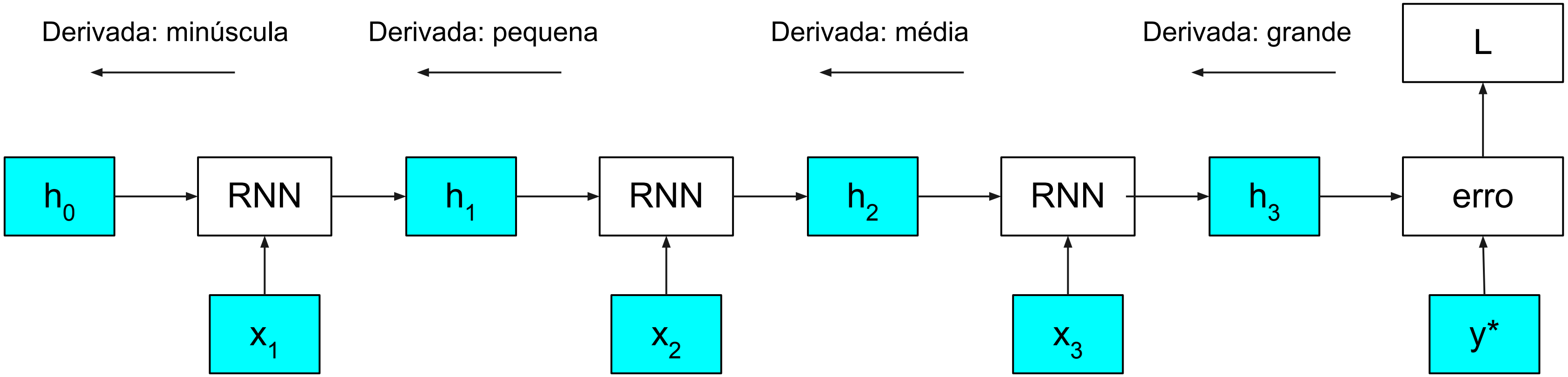
029221100245676500

Memória de longo prazo

- Redes recorrentes têm dificuldade em lidar com dependência de longa distância
    - Camadas intermediárias devem
        - Fornecer informação útil para o instante corrente
        - Atualizar e carregar informação de contexto para decisões futuras

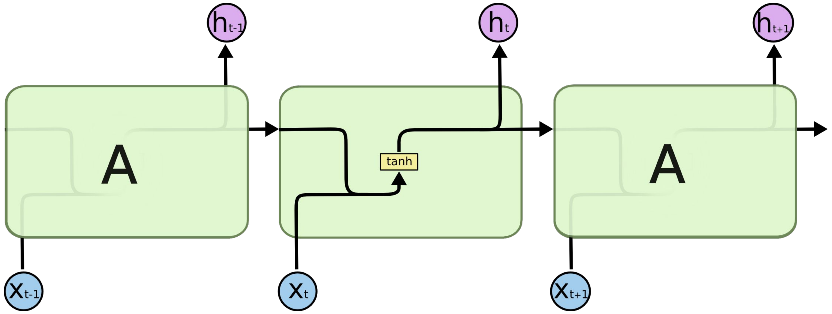
Vanishing gradient

- Redes recorrentes têm dificuldade em lidar com dependência de longa distância
  - *Backpropagação* do sinal do erro através do tempo
    - Camada escondida contribui para a perda do instante de tempo seguinte

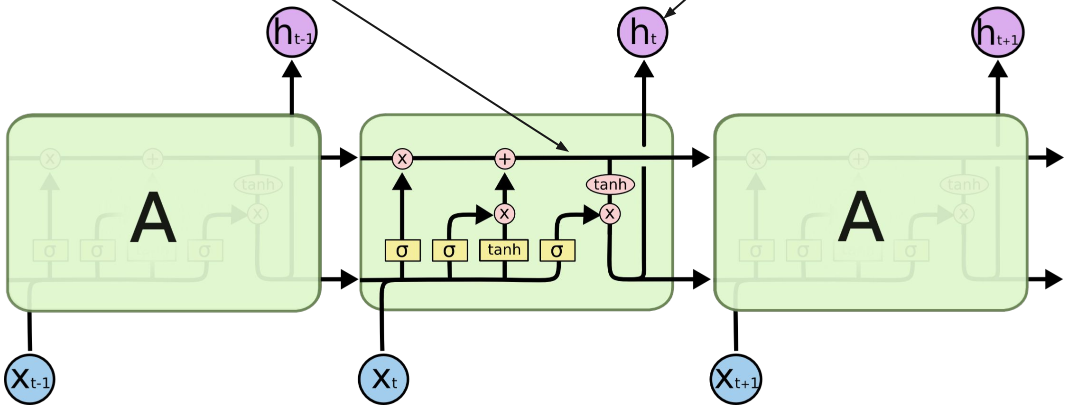


Long Short-Term memory (LSTM) (Hochreiter and Schmidhuber, 1997)

- RNN com uma estrutura de “memória”
- A cada passo, existe um estado escondido e uma célula de estado (vetores)
  - Contexto explícito
- A célula armazena informação de “longo termo”



RNN



Memória de longo prazo

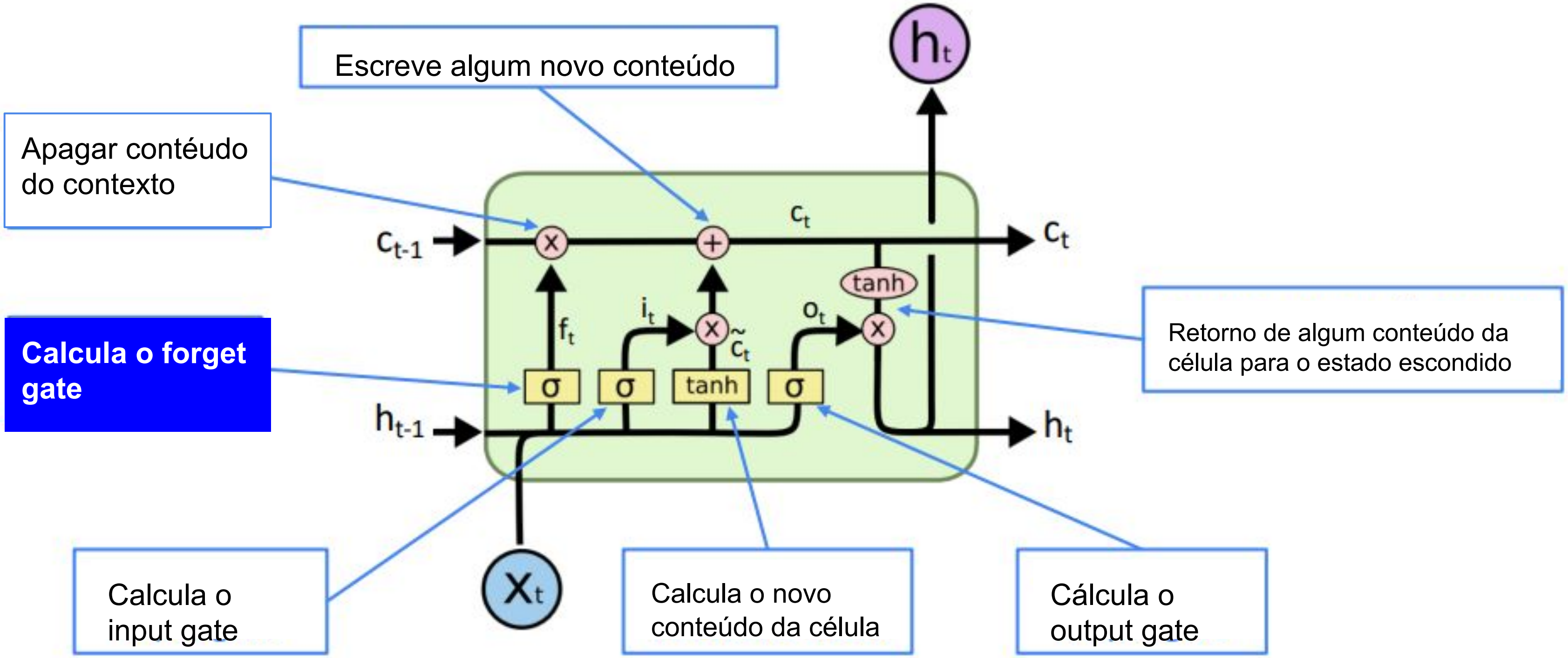
LSTM

Memória de curto prazo

Long Short-Term memory (LSTM) (Hochreiter and Schmidhuber, 1997)

- A rede pode apagar, escrever e ler informação da célula
    - A seleção de qual informação passará por cada operação é controlada por *gates* (vetores)
        - Conexões aditivas
        - Camada feedforward + sigmoid + multiplicação
    - A cada passo, as operações nos gates podem ser: abrir (1), fechar (0) ou algo no meio do caminho
    - Gates são dinâmicos: seu valor é calculado com base no contexto corrente

Long Short-Term Memory (LSTM)



Escreve algum novo conteúdo

Apagar contéudo do contexto

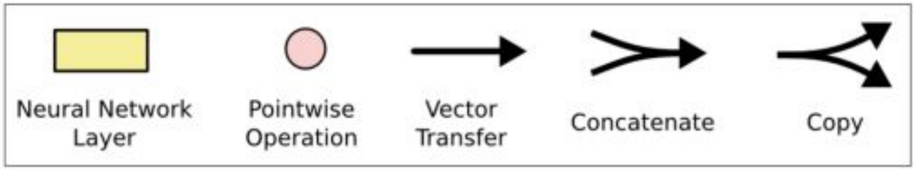
Calcula o forget gate

Retorno de algum conteúdo da célula para o estado escondido

Calcula o input gate

Calcula o novo conteúdo da célula

Cálcula o output gate



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

Long Short-Term Memory

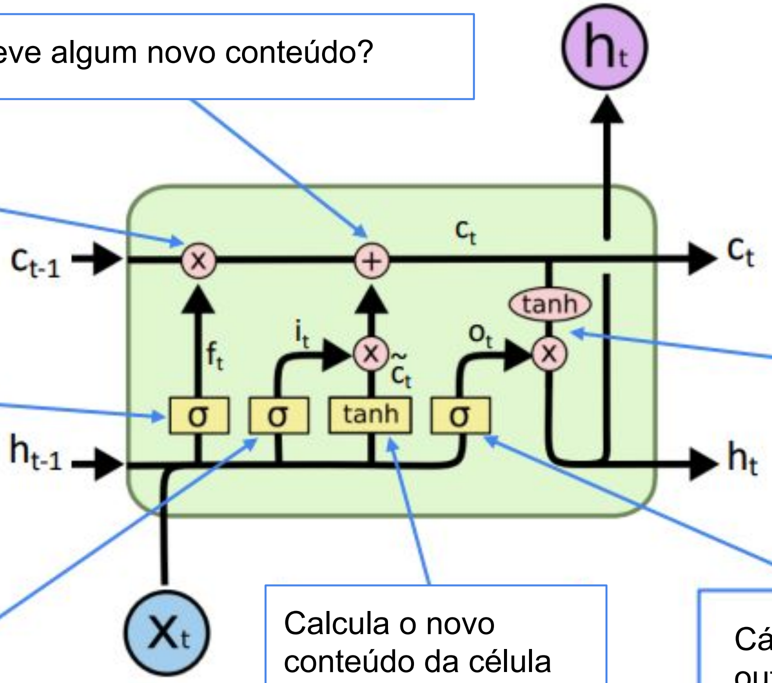
Forget gate: o que eu não devo esquecer (ou o que eu devo lembrar do curto prazo)

Sigmoid: valores entre 0 e 1

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

concatenação

Long Short-Term Memory (LSTM)



Escreve algum novo conteúdo?

Esquece algum conteúdo?

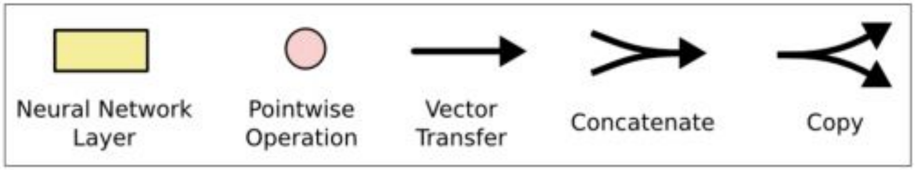
Calcula o forget gate

Calcula o input gate

Calcula o novo conteúdo da célula

Cálcula o output gate

Retorno de algum conteúdo da célula para o estado escondido



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

Long Short-Term Memory

Forget gate: o que eu não devo esquecer

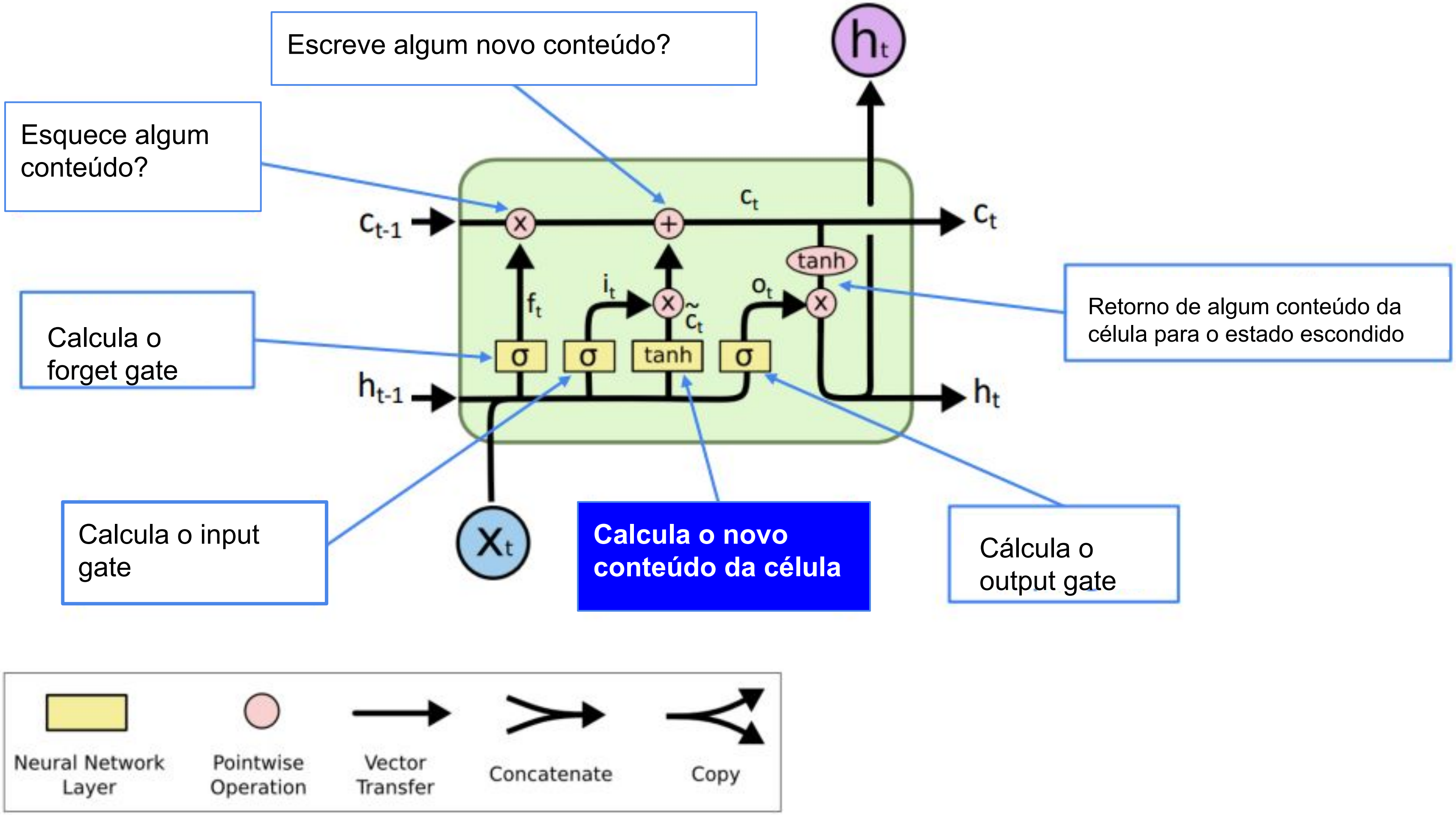
Input gate: Escreve na célula? (escrita)

Sigmoid: valores entre 0 e 1

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

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

Long Short-Term Memory (LSTM)



Escreve algum novo conteúdo?

Esquece algum conteúdo?

Calcula o forget gate

Retorno de algum conteúdo da célula para o estado escondido

Calcula o input gate

Calcula o novo conteúdo da célula

Cálcula o output gate

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

Long Short-Term Memory

Forget gate: o que é armazenado vs o que é esquecido, a partir da célula anterior

Input gate: Escreve na célulA?

Sigmoid: valores entre 0 e 1

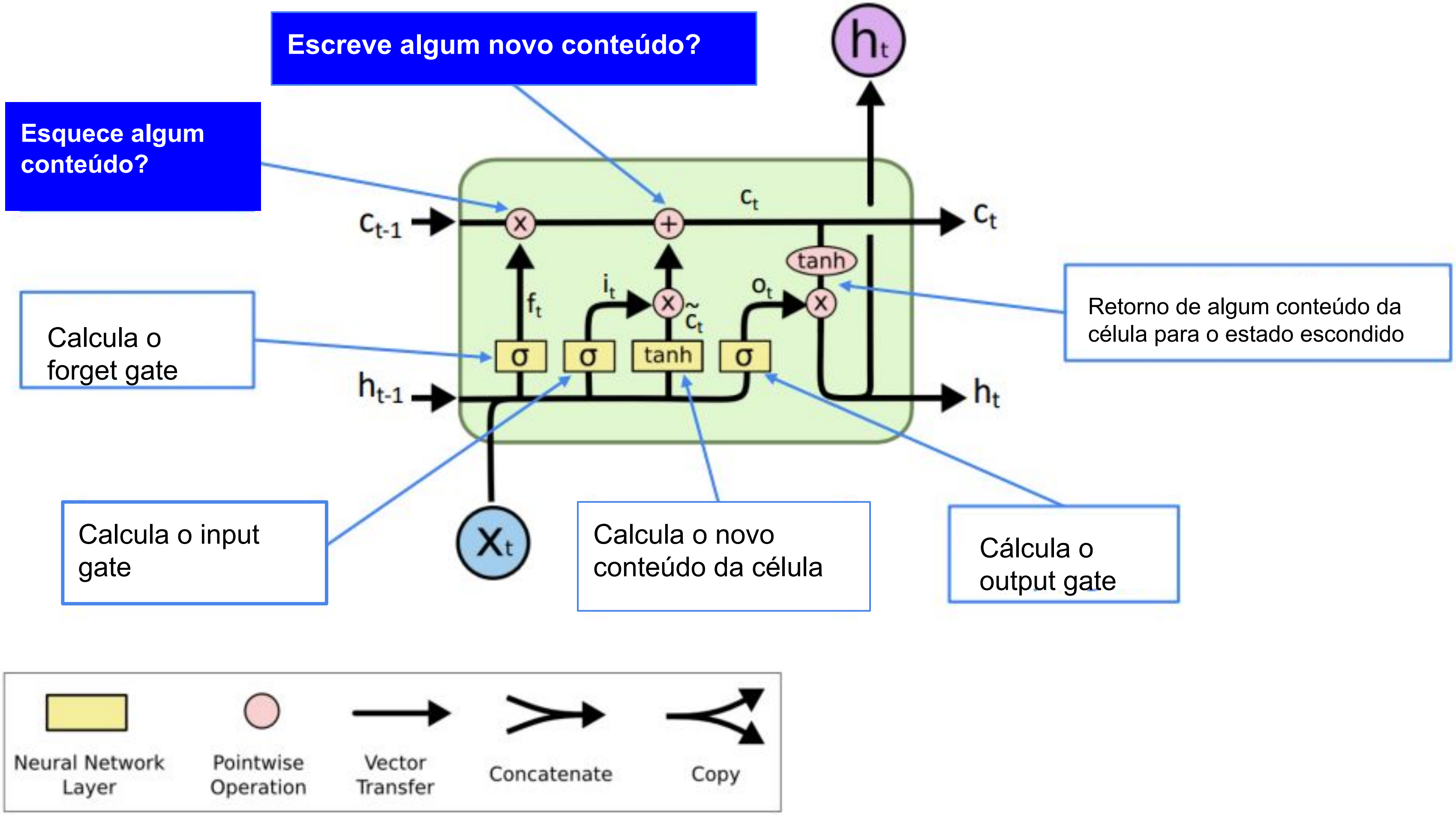
$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \ + \ b_f\right)$$

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

Novo conteúdo a ser escrito na célula

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

Long Short-Term Memory (LSTM)



Escreve algum novo conteúdo?

Esquece algum conteúdo?

Calcula o forget gate

Retorno de algum conteúdo da célula para o estado escondido

Calcula o input gate

Calcula o novo conteúdo da célula

Cálcula o output gate

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

Long Short-Term Memory

Sigmoid: valores entre 0 e 1

Forget gate: o que é armazenado vs o que é esquecido, a partir da célula anterior

Input gate: ESCRITA

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

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

Novo conteúdo a ser escrito na célula

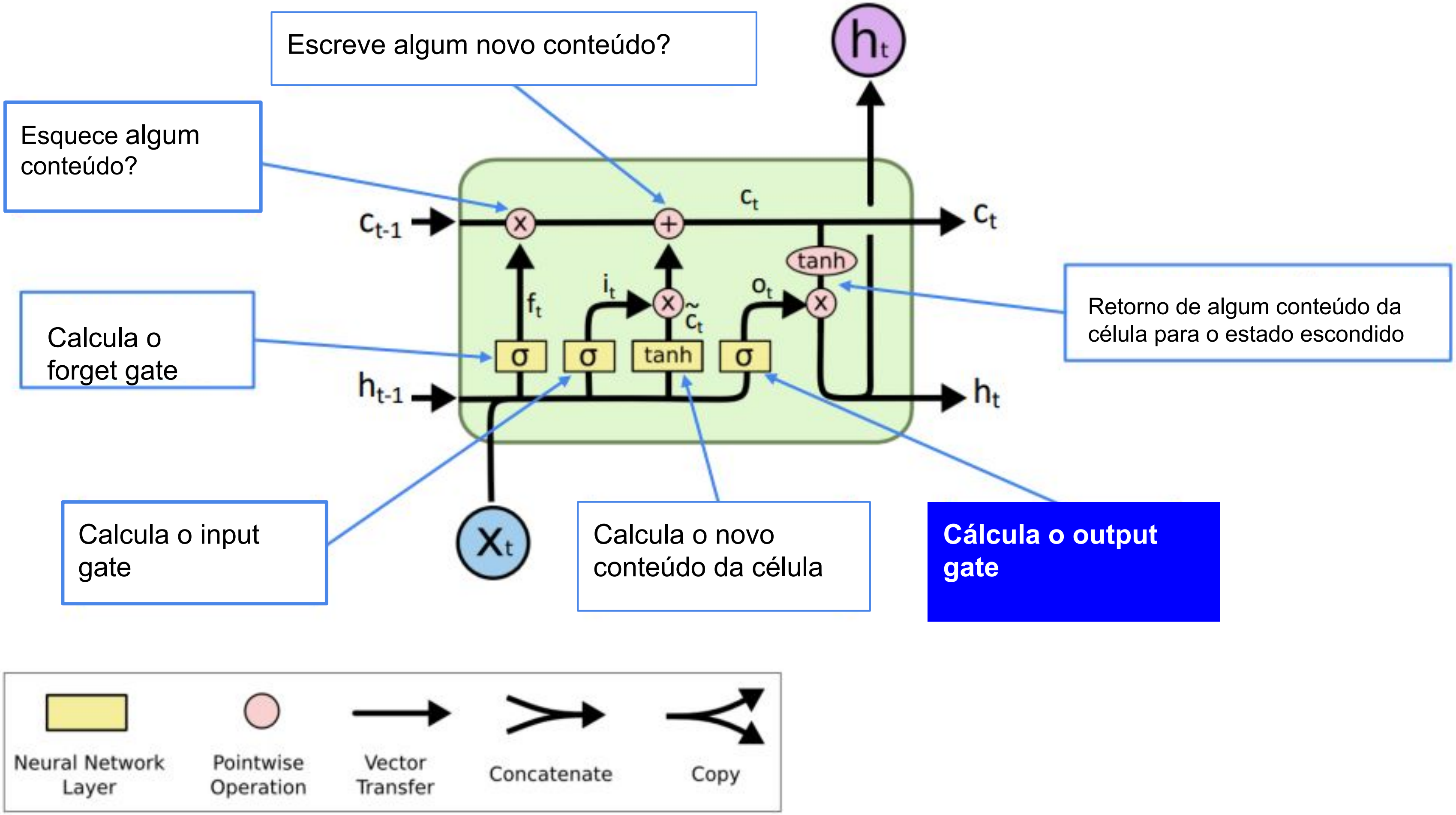
Estado da célula: “esquece” algum conteúdo do estado anterior da célula e escreve algum conteúdo novo

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

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

Produto de elemento a elemento

Long Short-Term Memory (LSTM)



Escreve algum novo conteúdo?

Esquece algum conteúdo?

Calcula o forget gate

Retorno de algum conteúdo da célula para o estado escondido

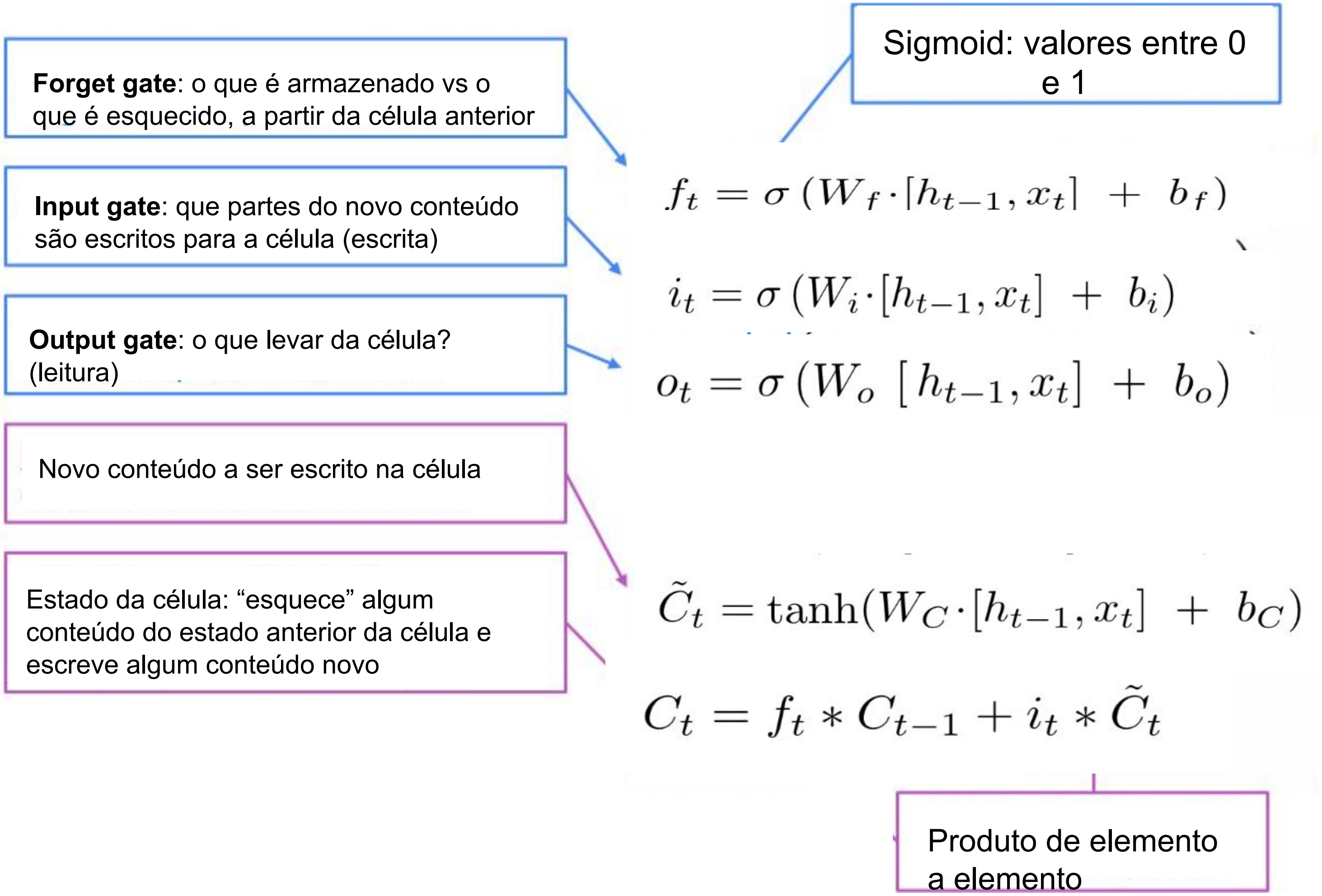
Calcula o input gate

Calcula o novo conteúdo da célula

Cálcula o output gate

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

Long Short-Term Memory



Sigmoid: valores entre 0 e 1

Forget gate: o que é armazenado vs o que é esquecido, a partir da célula anterior

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

Input gate: que partes do novo conteúdo são escritos para a célula (escrita)

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

Output gate: o que levar da célula? (leitura)

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

Novo conteúdo a ser escrito na célula

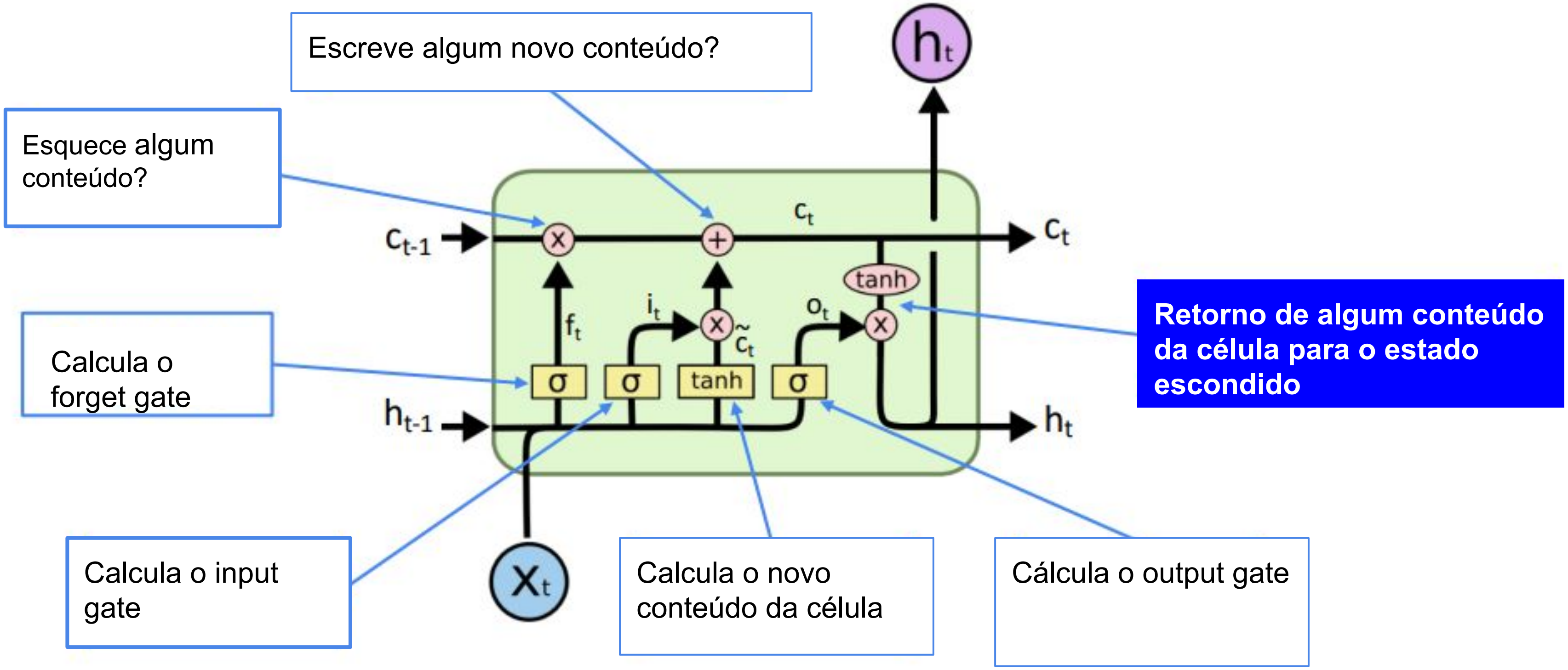
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \;+\; b_C)$$

Estado da célula: “esquece” algum conteúdo do estado anterior da célula e escreve algum conteúdo novo

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

Produto de elemento a elemento

Long Short-Term Memory (LSTM)



Escreve algum novo conteúdo?

Esquece algum conteúdo?

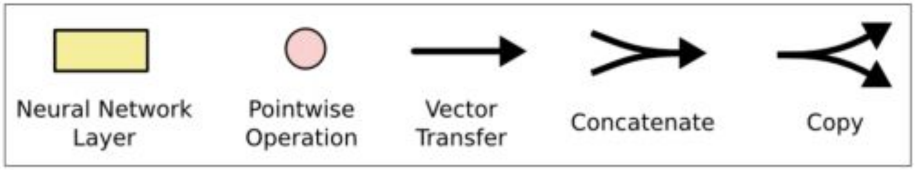
Calcula o forget gate

Retorno de algum conteúdo da célula para o estado escondido

Calcula o input gate

Calcula o novo conteúdo da célula

Cálcula o output gate



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

Long Short-Term Memory

Sigmoid: valores entre 0 e 1

Forget gate: o que é armazenado vs o que é esquecido, a partir da célula anterior

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

Input gate: escrita

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

Output gate: leitura

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

Novo conteúdo a ser escrito na célula

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

Estado da célula: "esquece" algum conteúdo do estado anterior da célula e escreve algum conteúdo novo

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

Estado escondido: lê algum conteúdo da célula

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

Produto de elemento a elemento

Vetores de tamanho n

Gated Recurrent Units (GRUs) (Cho et al., 2014)

- LSTM pode ser bem custosa para treinar
    - 8 matrizes de peso (duas para cada gate)
- GRUs
    - Dispensam o vetor de contexto (célula)
    - Usam apenas dois gates
        - Reset
            - O que é relevante no estado anterior e o que pode ser ignorado?

$$r_t = \sigma(W_r[h_{t-1} ; x_t])$$

$$h'_t = \tanh(U(r_t \odot h_{t-1})+Wx_t)$$

Gated Recurrent Units (GRUs) (Cho et al., 2014)

- LSTM pode ser bem custosa para treinar
    - 8 matrizes de peso (duas para cada gate)
- GRUs
    - Dispensam o vetor de contexto (célula)
    - Usam apenas dois gates
        - Reset

$$r_t = \sigma(W_r[h_{t-1} ; x_t])$$

$$h'_t = \tanh(U(r_t \odot h_{t-1})+Wx_t)$$

        - Update
            - O que de h'_t será usado diretamente no novo estado escondido h_t e o que precisa ser preservado de h_{t-1}

$$z_t = \sigma(U_z h_{t-1} + W_z x_t)$$

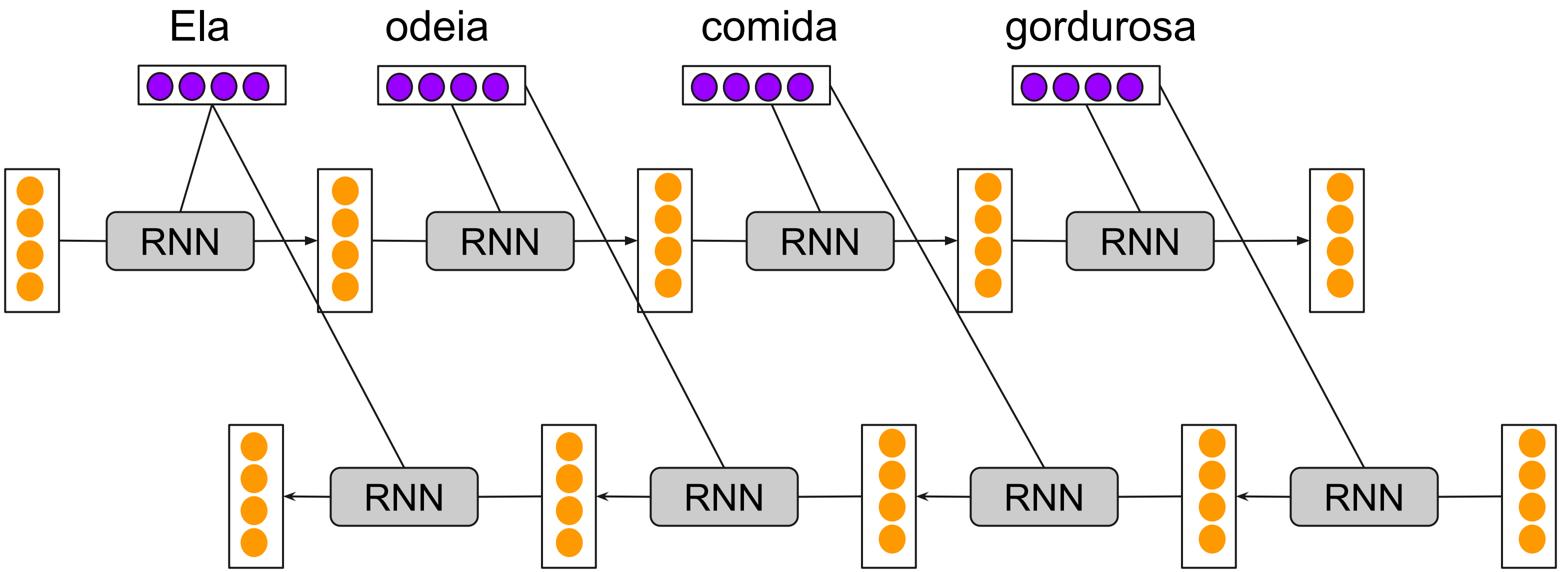
$$h_t = (1-z_t) h_{t-1} + z_t h't$$

Embeddings contextualizados : Elmo (Peters et al., 2018)

Contextualized word-embeddings can give words different embeddings based on the meaning they carry in the context of the sentence. Also, RIP Robin Williams

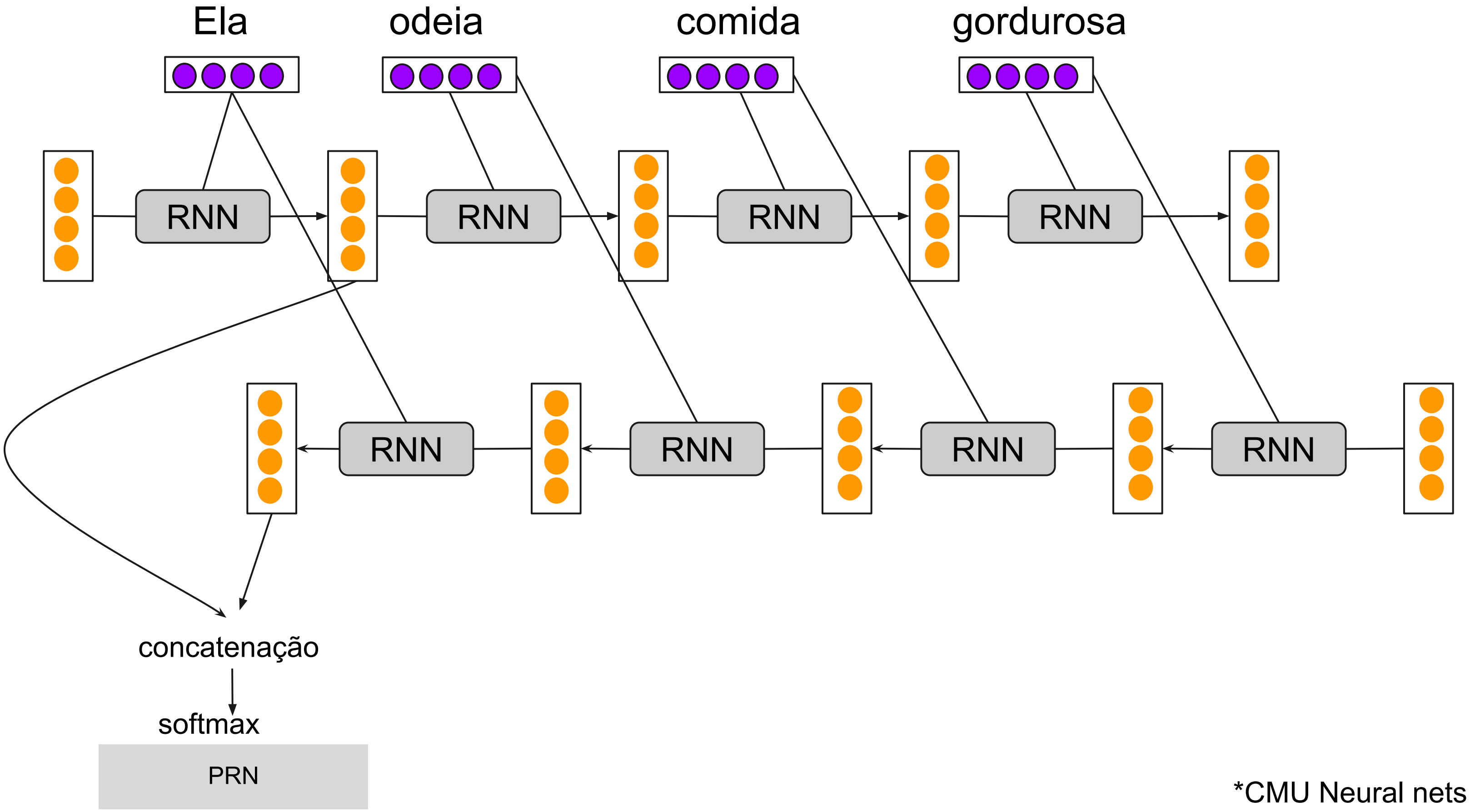
* http://jalammar.github.io/illustrated-bert/

RNN bidirecional



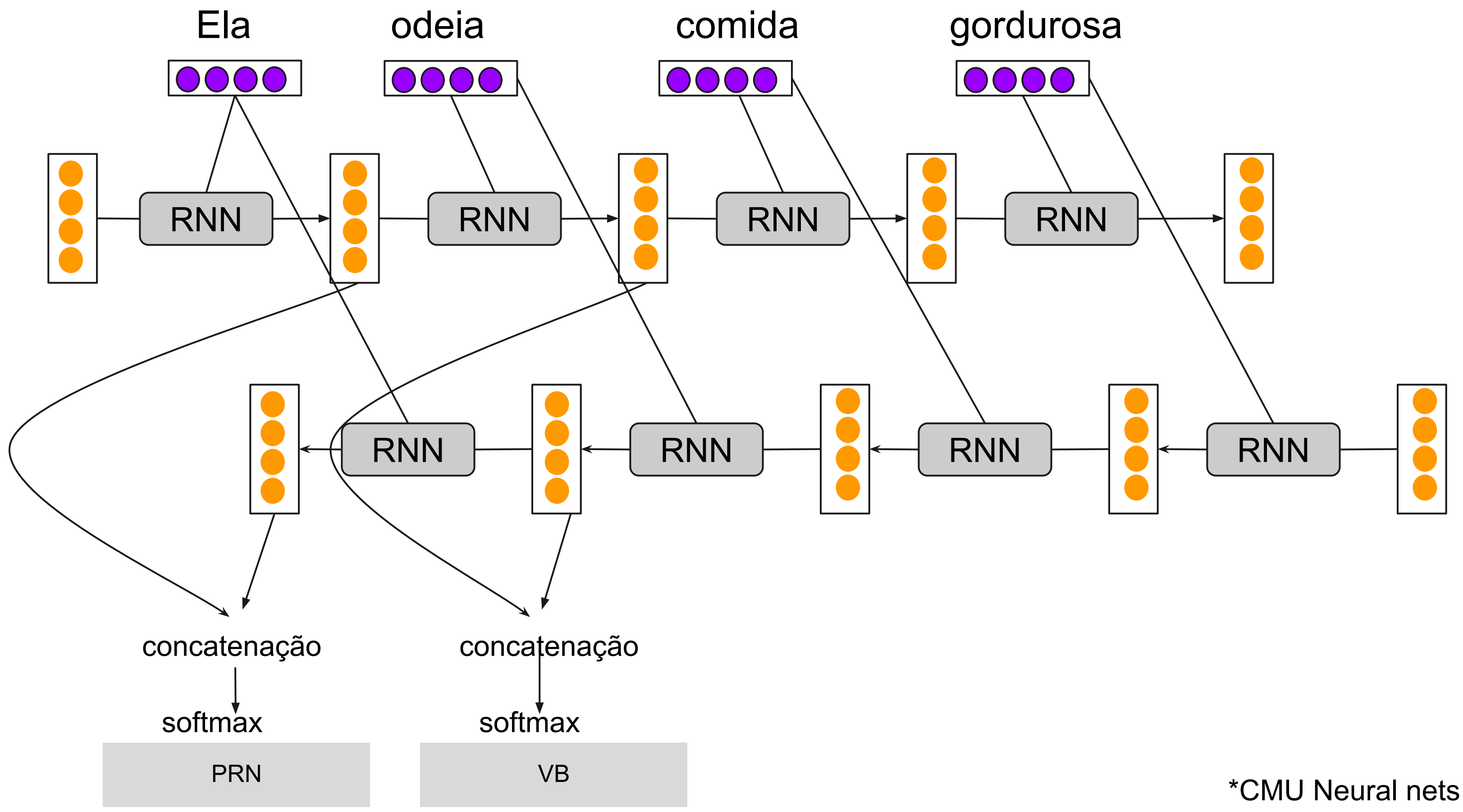
*CMU Neural nets for NLP

RNN bidirecional



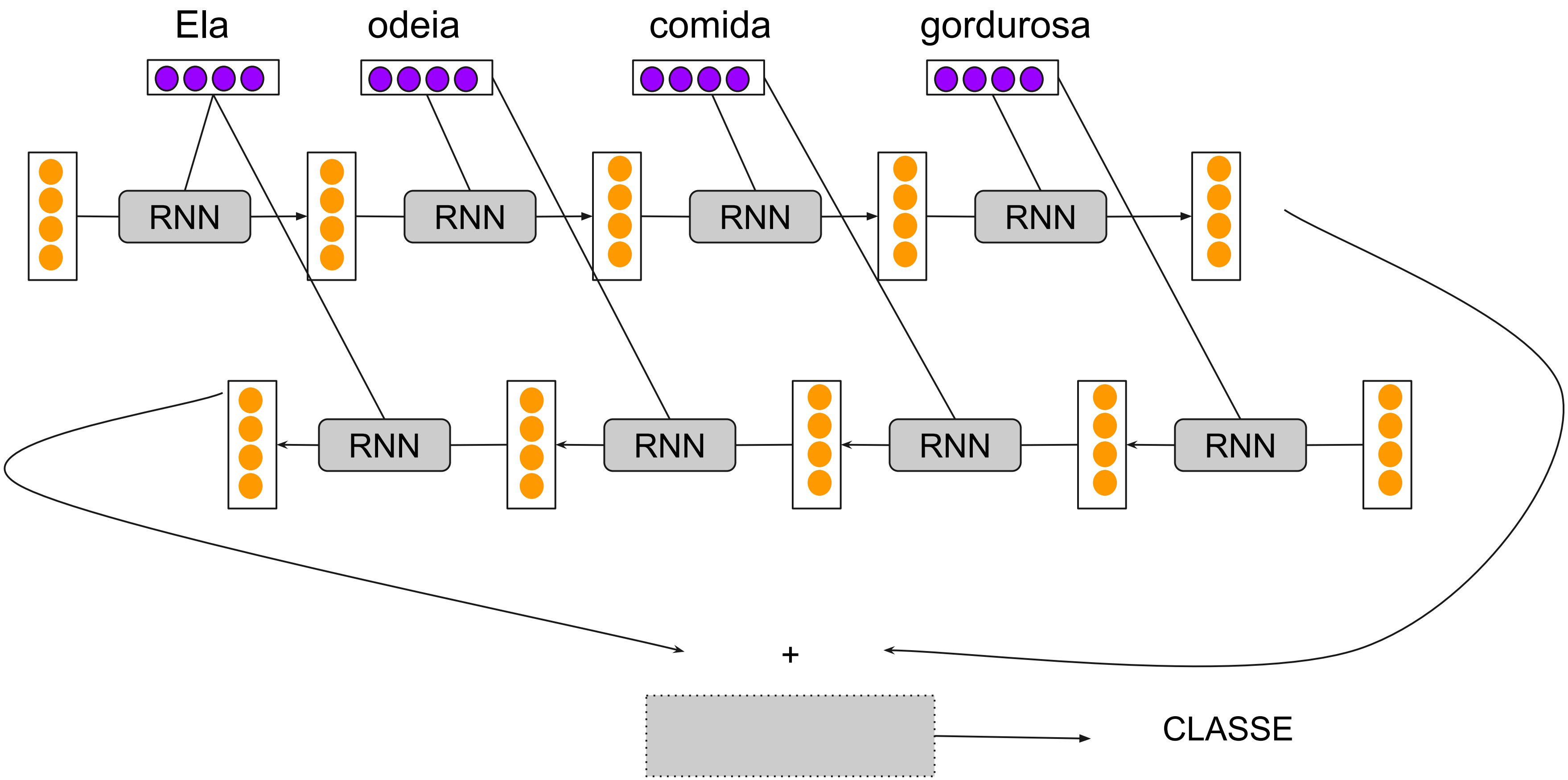
*CMU Neural nets for NLP

RNN bidirecional



*CMU Neural nets for NLP

RNN bidirecional



Embeddings contextualizados : Elmo (Peters et al., 2018)

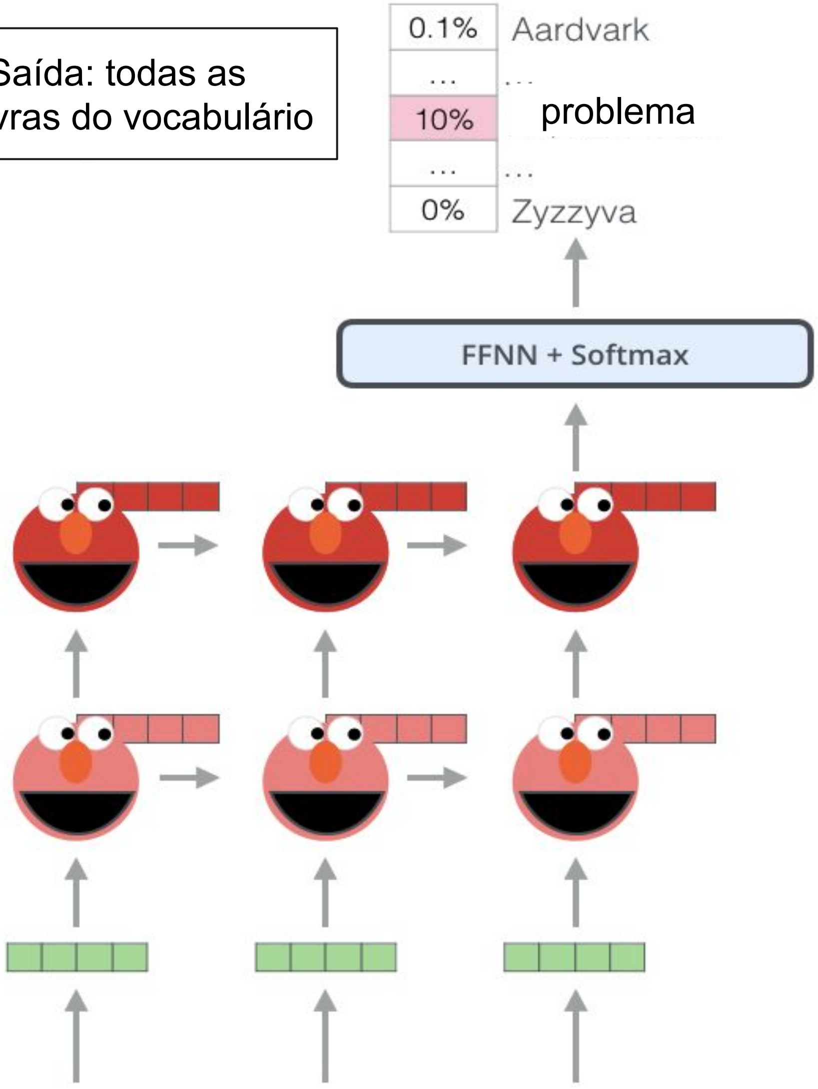
- Modelo de linguagem bidirecional

$$p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k \mid t_1, t_2, \ldots, t_{k-1})$$

$$p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k \mid t_{k+1}, t_{k+2}, \ldots, t_N)$$

- Embedding da palavra: combinação linear das camadas escondidas correspondentes

Treinamento genérico contextualizado



Saída: todas as palavras do vocabulário

Camada de saída

LSTM 2

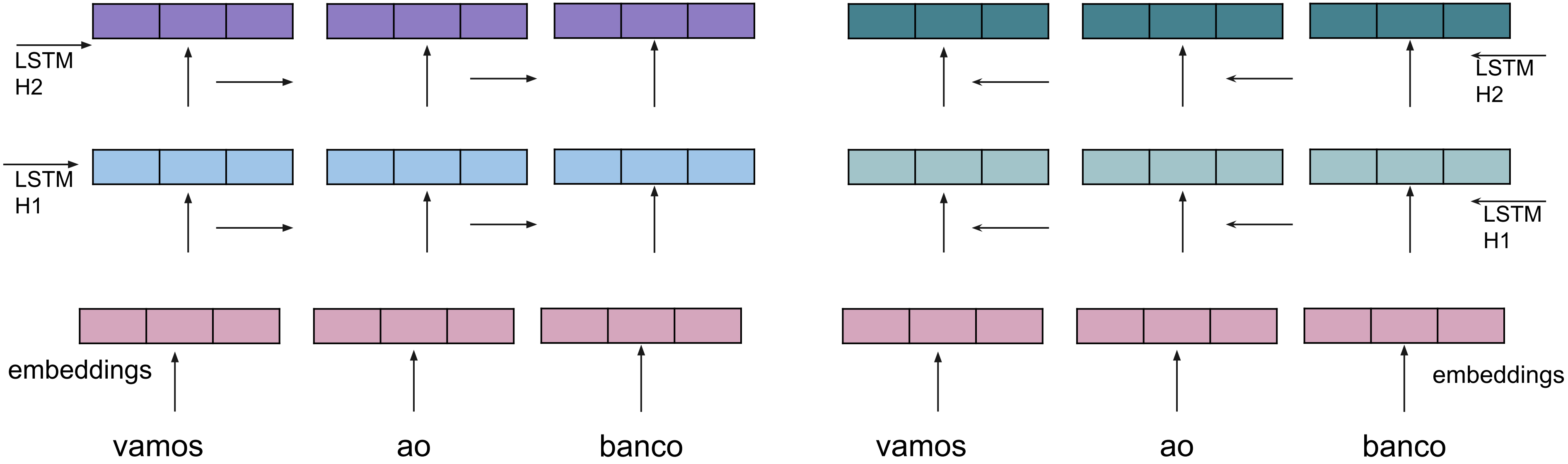
LSTM 1

Embeddings

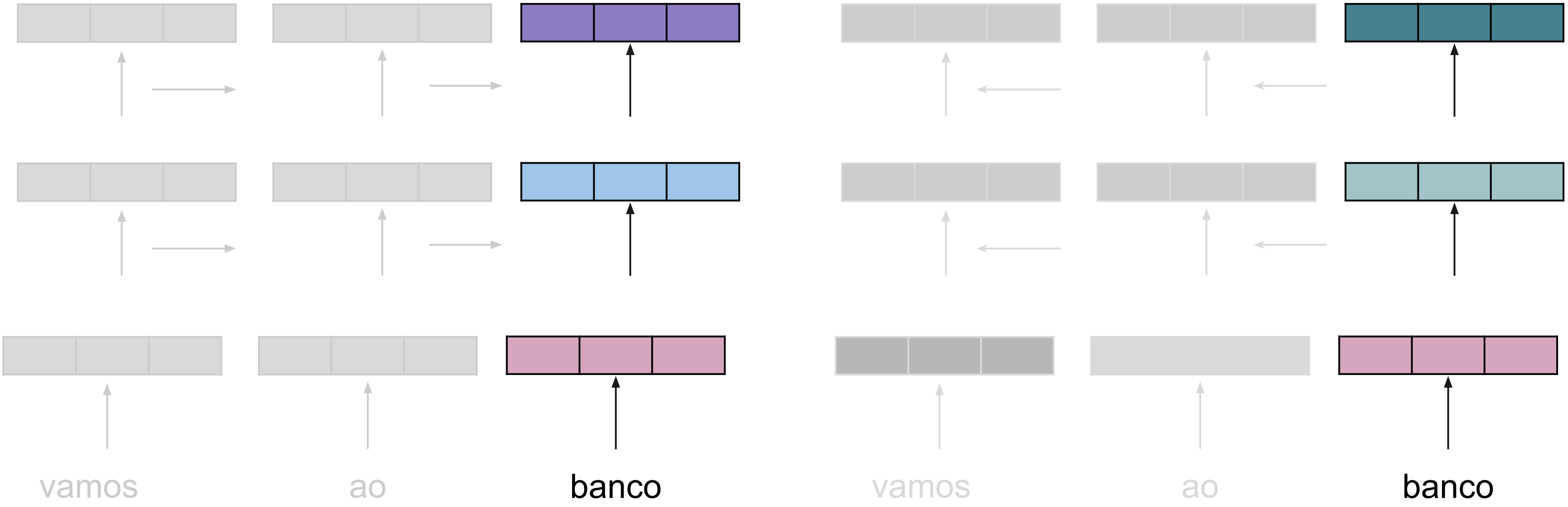
Chocolate resolve qualquer

Adaptado de https://jalammar.github.io/illustrated-bert/

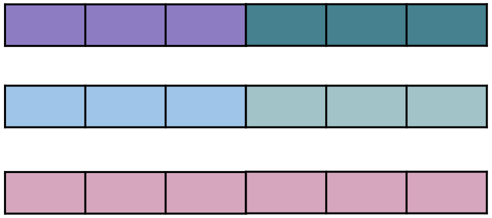
Elmo


LSTM
H2
LSTM
H1
embeddings
vamos
ao
banco
vamos
ao
banco
LSTM
H2
LSTM
H1
embeddings

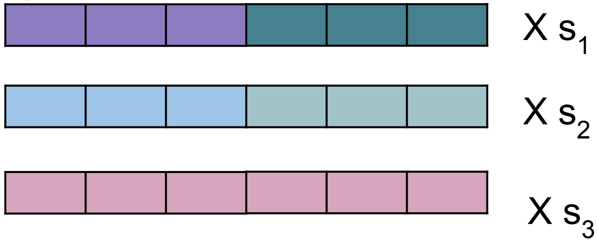
Elmo



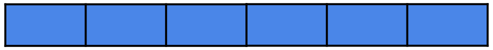
Concatena pesos das camadas escondidas e embeddings



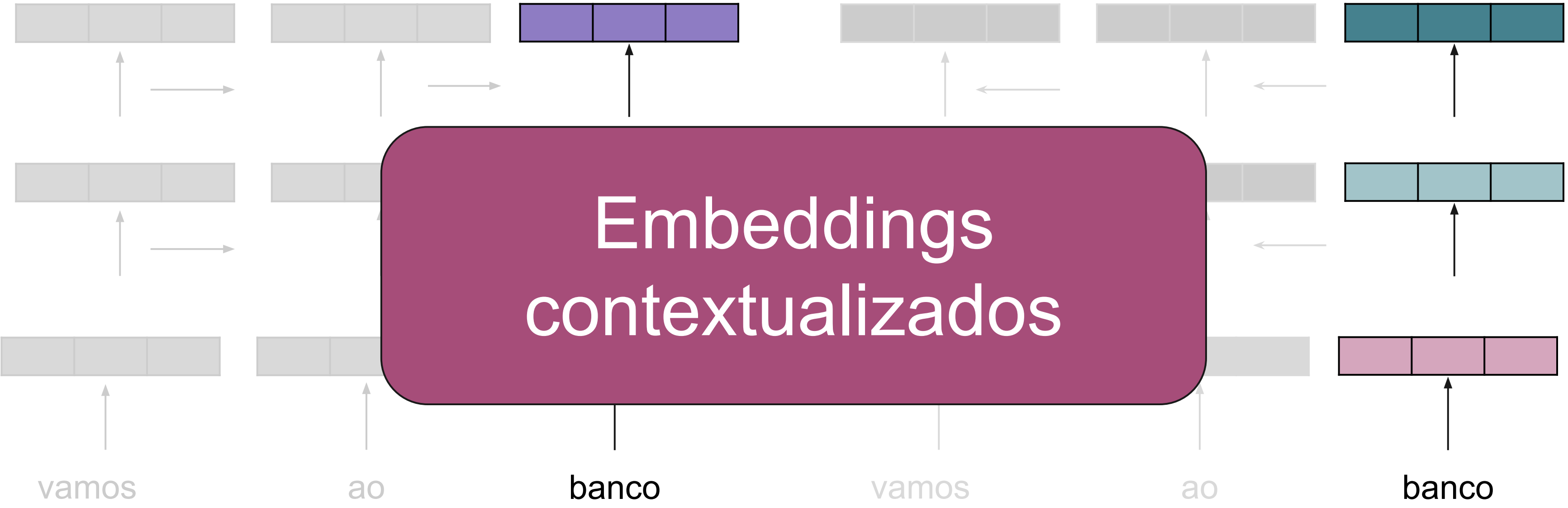
Cada um é multiplicado por um peso, baseado na tarefa



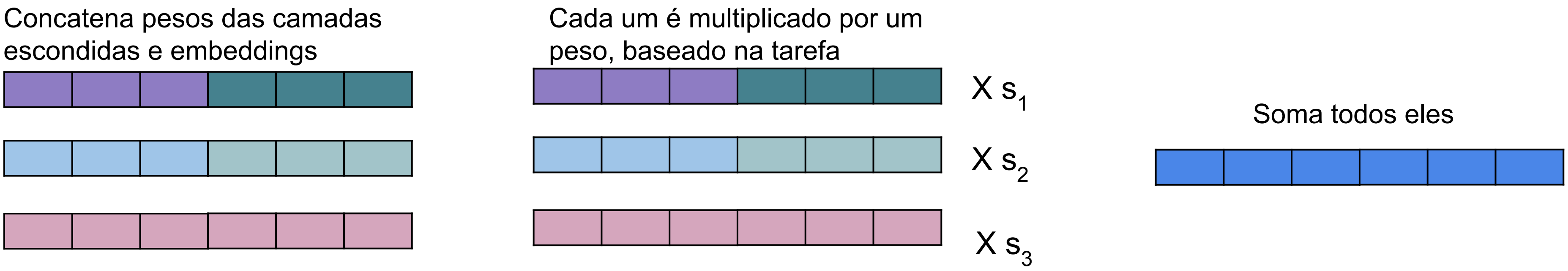
Soma todos eles



Elmo



Embeddings contextualizados

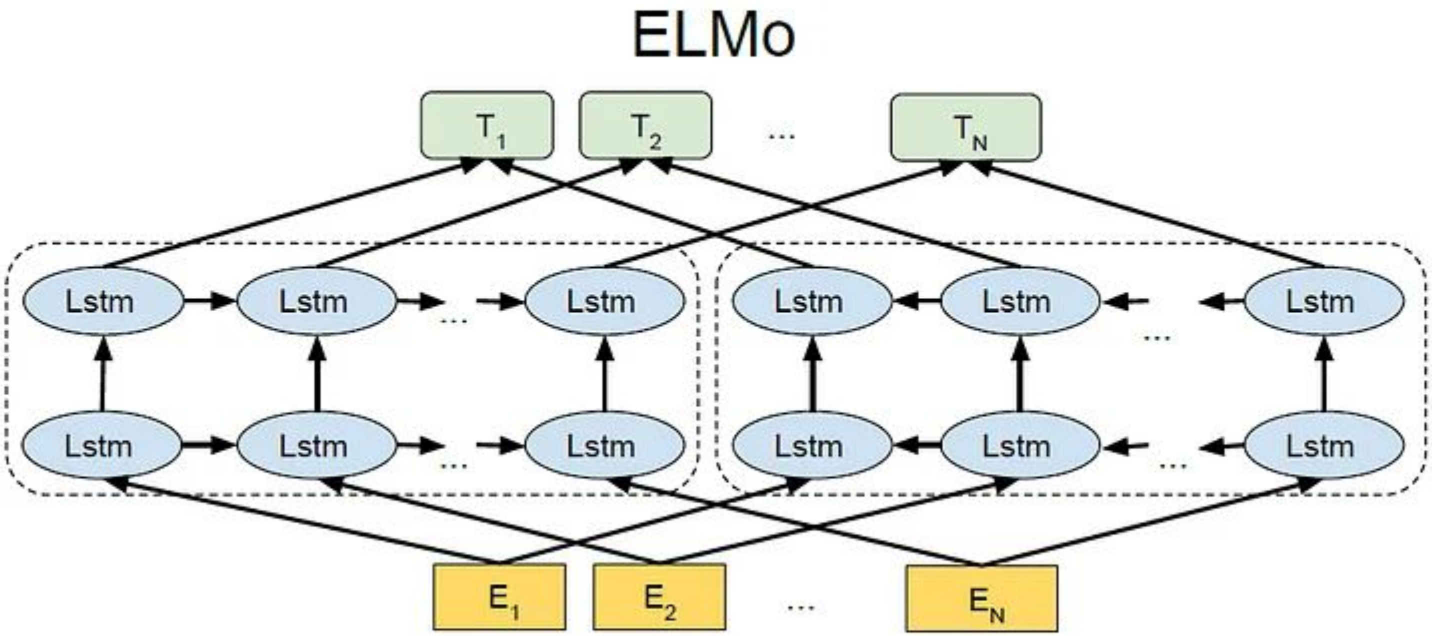


Concatena pesos das camadas escondidas e embeddings

Cada um é multiplicado por um peso, baseado na tarefa

Soma todos eles

ELMo


ELMo
T_1
T_2
...
T_N
Lstm
Lstm
...
Lstm
Lstm
Lstm
...
Lstm
Lstm
Lstm
...
Lstm
Lstm
Lstm
...
Lstm
E_1
E_2
...
E_N