Deep Learning através das abstrações

Recurrent Neural Nets(RNN)

O que são?

- Uma arquitetura diferente de redes neurais
- Mais importantemente, redes neurais que usam camadas recorrentes
  - LSTM
  - GRU
- Criadas para reconhecer padrões em sequências de dados
  - Textos
  - Aúdios
  - Mercado financeiro
  - Vídeos
  - etc

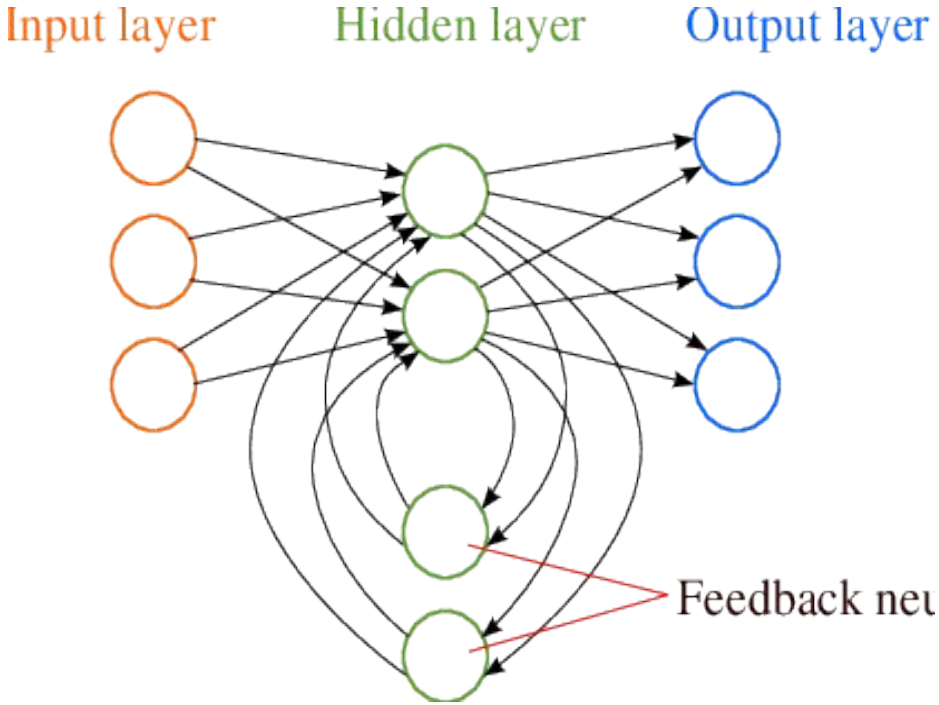
Um pouco mais sobre o funcionamento

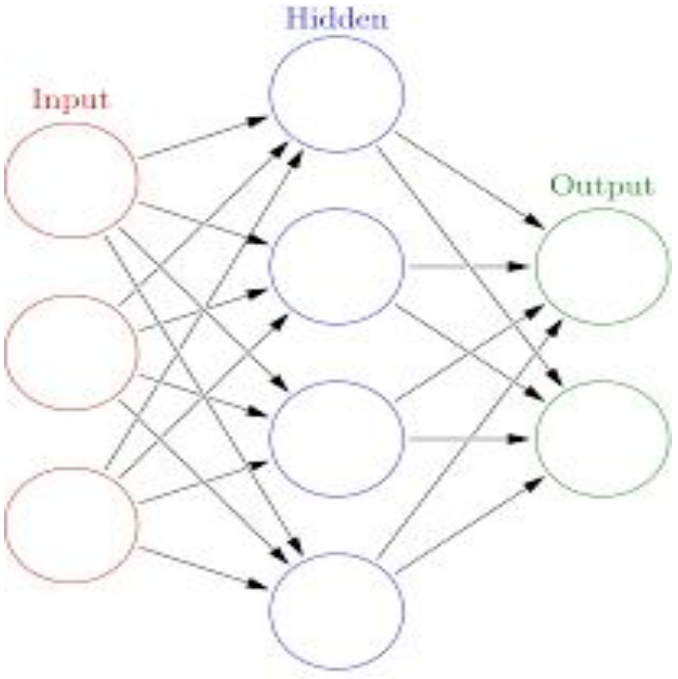
- Usam neurônios com retroalimentação
  - Ou seja, o resultado do input i - 1, faz parte do input i
  - Podemos ver isso como sendo uma rede com memória
  - Isso nos permite ter uma noção de tempo ou de dados sequenciais
- Dizemos que é um modelo profundo no tempo e não no espaço
  - Matematicamente temos

$$\mathbf{h}_t = \phi\left(W\mathbf{x}_t + U\mathbf{h}_{t-1}\right).$$

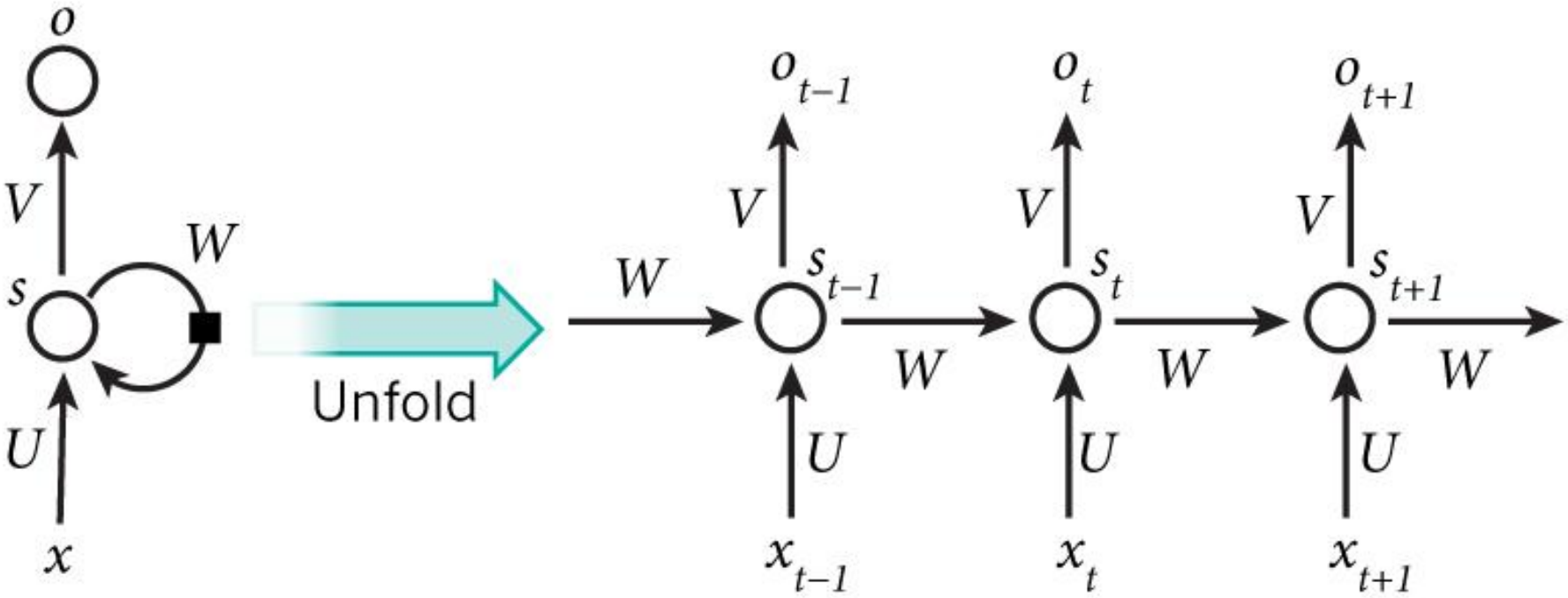
  - Gera problemas, uma vez que uma rede que é treinada para usar 100 unidades de tempo anteriores para prever a 101 é equivalente e uma rede de 100 camadas (muito profunda)

Comparando com feedforward





Unfolding

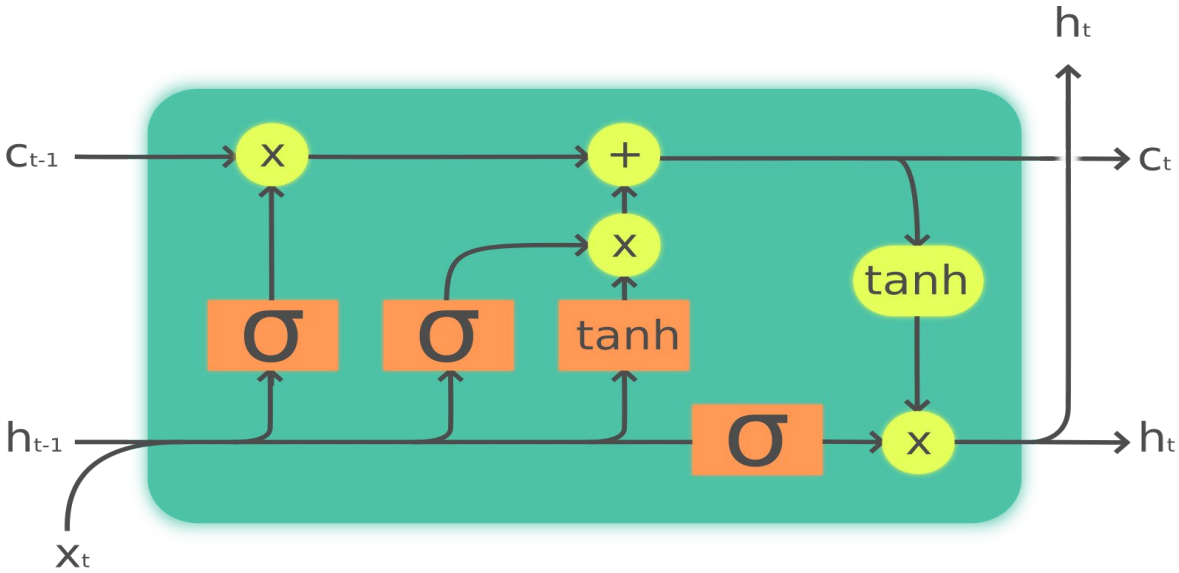


Treinando essas redes

- Backpropagation Through Time
  - Na prática usamos uma versão aproximada (Ilya Sutskever)
  - As bibliotecas já implementam isso de boas
- Vanishing and Exploding Gradients
  - Uso de otimizadores melhores (comumente o RMSProp)
  - Métodos bons de inicialização
  - Usar LSTM
  - No caso de Exploding, usar gradient clipping

Long Short-Term Memory Units (LSTMs)

- Proposta pelo Schmidhuber e o Hochreiter
- Solucionam o problema do vanishing gradient
- Usam memory gates para tentar manter “lembrar” o que é importante



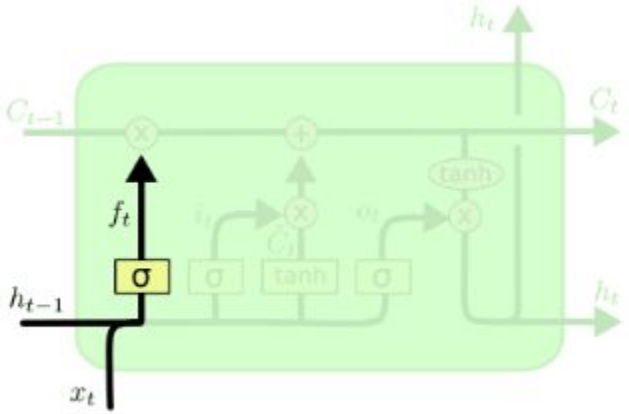
Legend:

Layer

Pointwize op

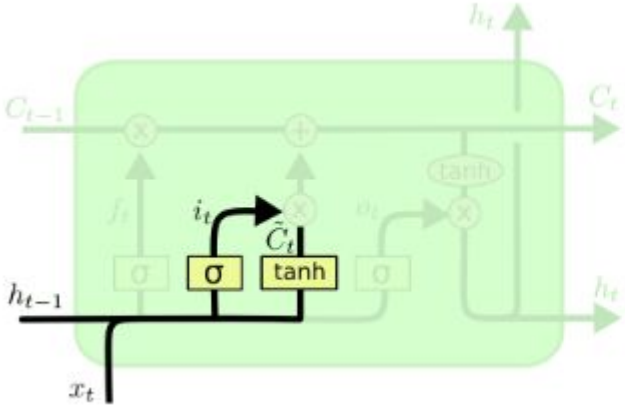
Copy

Passo 1: O que jogar fora



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

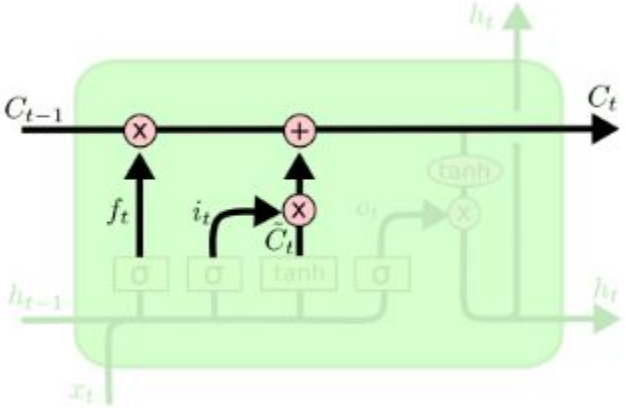
Parte 2: O que adicionar



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

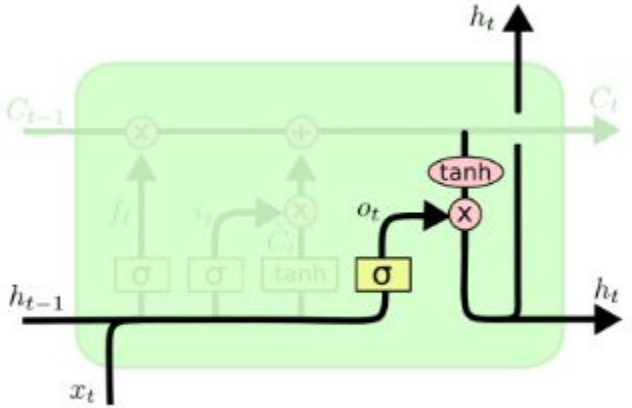
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

Parte 3: Atualiza o Cell State



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

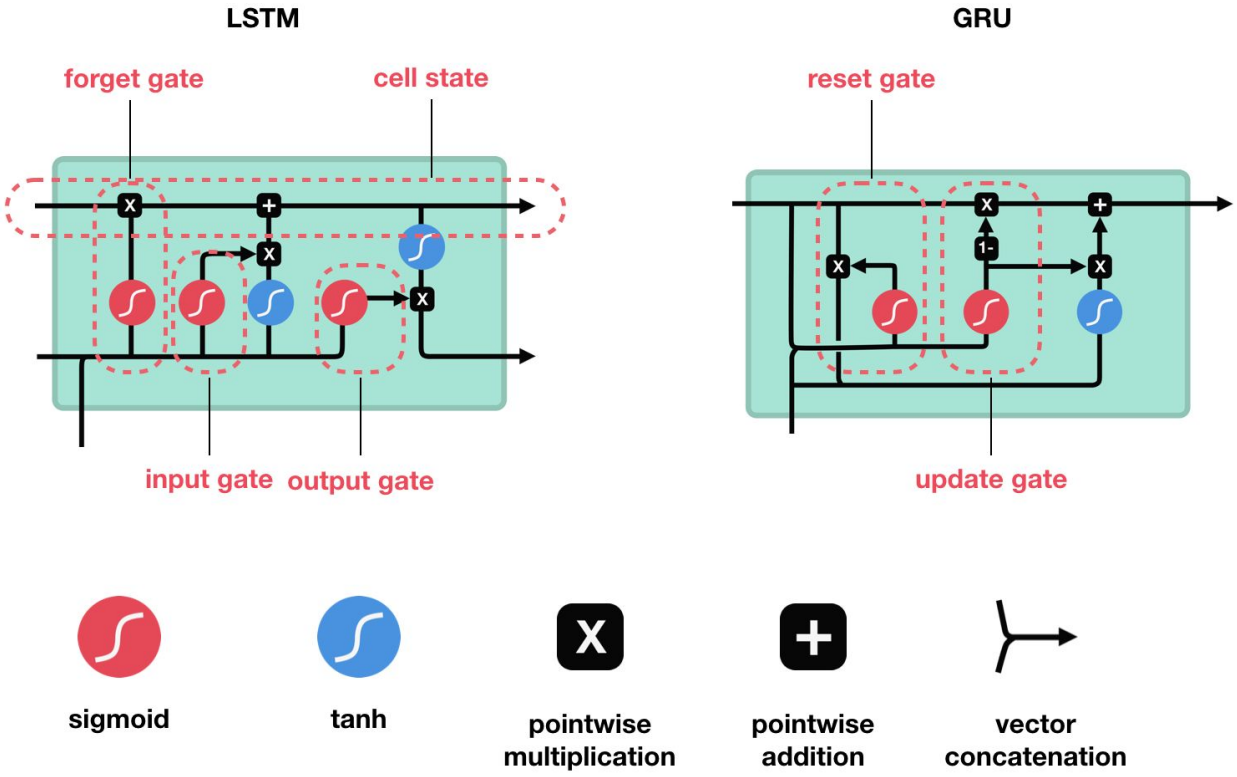
Parte 4: Output da camada



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

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

Gated Recurrent Unit (GRU)



Resources

- Artigo da Skymind
- Blog do Colah
- Blog do Karpathy
- Illustrated Guide do RNNs
- Tese do Ilya