RNN

\- Recurrent Neural Network

概要

- 序列模型、语言模型
- 循环神经网络
    - 实现 RNN 语言模型
    - 时间反向传播
- 门控循环单元（GRU）
- 长短期记忆网络（LSTM）

序列模型

- 相关的随机变量

$$(x_1, \ldots x_T) \sim p(x)$$

- 条件概率展开

$$p(x) = p(x_1) \cdot p(x_2|x_1) \cdot p(x_3|x_1, x_2) \cdot \ldots p(x_T|x_1, \ldots x_{T-1})$$

  – 也可以按照反向...

$$p(x) = p(x_T) \cdot p(x_{T-1}|x_T) \cdot p(x_{T-2}|x_{T-1}, x_T) \cdot \ldots p(x_1|x_2, \ldots x_T)$$

序列模型

$$p(x) = p(x_1) \cdot p(x_2|x_1) \cdot p(x_3|x_1, x_2) \cdot \dots p(x_T|x_1, \dots x_{T-1})$$

马尔可夫（Markov）假设

$$p(x) = p(x_1) \cdot p(x_2|x_1) \cdot p(x_3|x_1, x_2) \cdot \ldots p(x_T|x_{T-\tau}, \ldots x_{T-1})$$

假设只依赖固定前几步

语言模型

$$p(w_1, w_2, \dots, w_T) = \prod_{t=1}^{T} p(w_t | w_1, \dots, w_{t-1})$$

$$p(\text{Statistics, is, fun, .})$$
$$= \quad p(\text{Statistics})p(\text{is}|\text{Statistics})p(\text{fun}|\text{Statistics, is})p(.\,|\text{Statistics, is, fun})$$

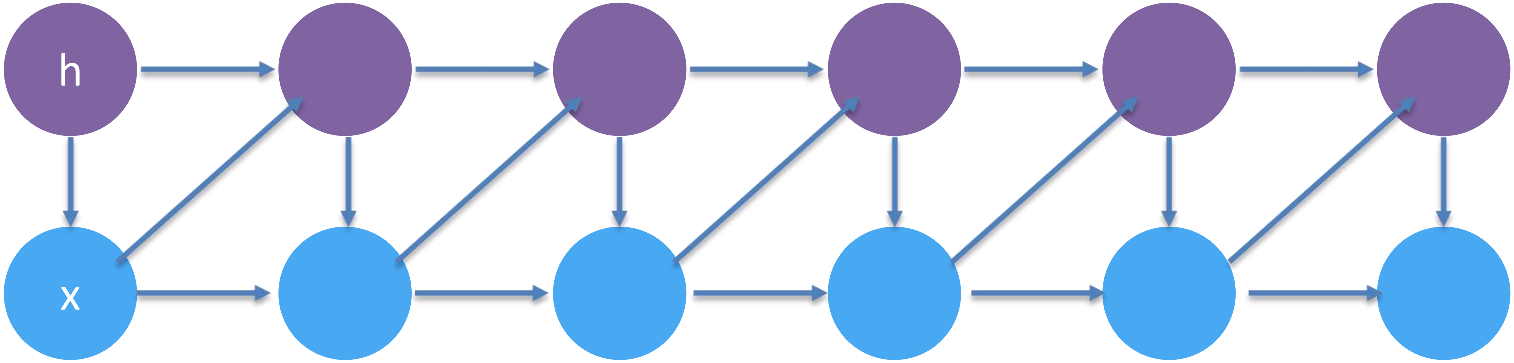
- 估计 $\hat{p}(\text{is}|\text{Statistics}) = \dfrac{n(\text{Statisticsis})}{n(\text{Statistics})}$

隐变量模型

- 隐含状态总结了有关过去所有的相关信息

$$h_t = f(x_1, \dots x_{t-1}) = f(h_{t-1}, x_{t-1})$$

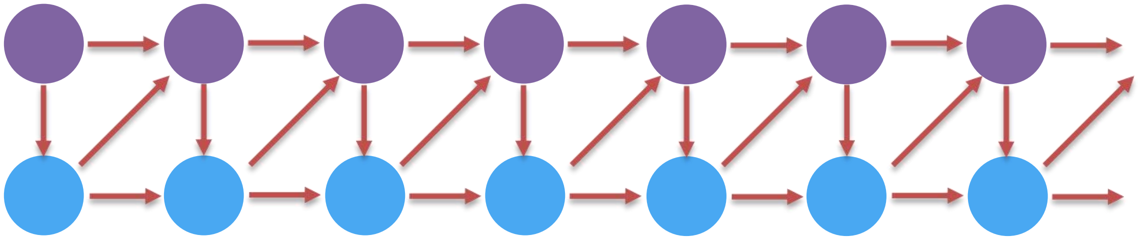
$$p(h_t | h_{t-1}, x_{t-1}) , \quad p(x_t | h_t, x_{t-1})$$



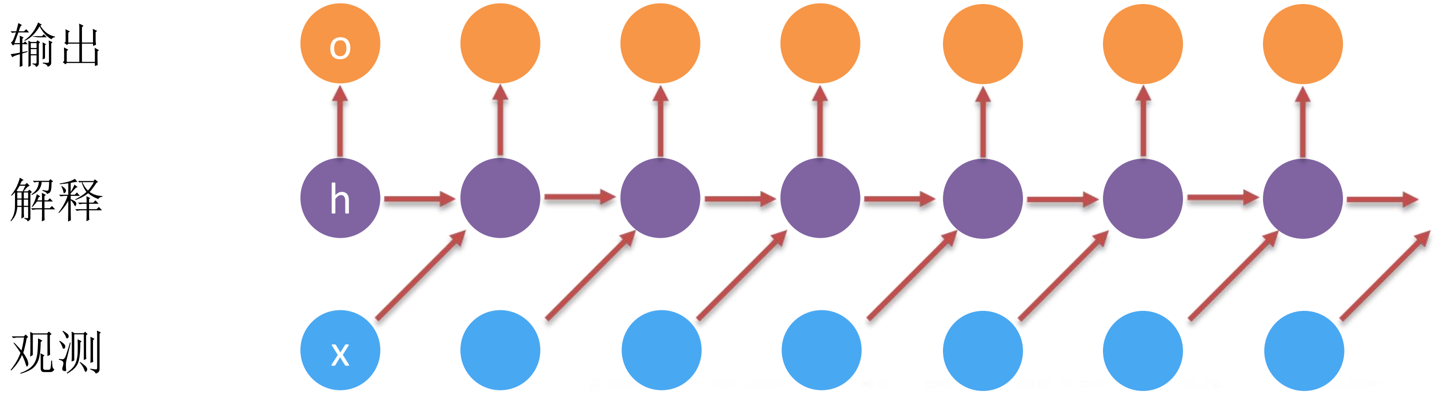
循环神经网络

解释

行动



循环神经网络



- 隐含状态更新

$$\mathbf{h}_t = \phi(\mathbf{W}_{hh}\mathbf{h}_{t-1} + \mathbf{W}_{hx}\mathbf{x}_{t-1} + \mathbf{b}_h)$$

- 观察更新

$$\mathbf{o}_t = \phi(\mathbf{W}_{ho}\mathbf{h}_t + \mathbf{b}_o)$$

独热编码

How to represent each word as a vector?

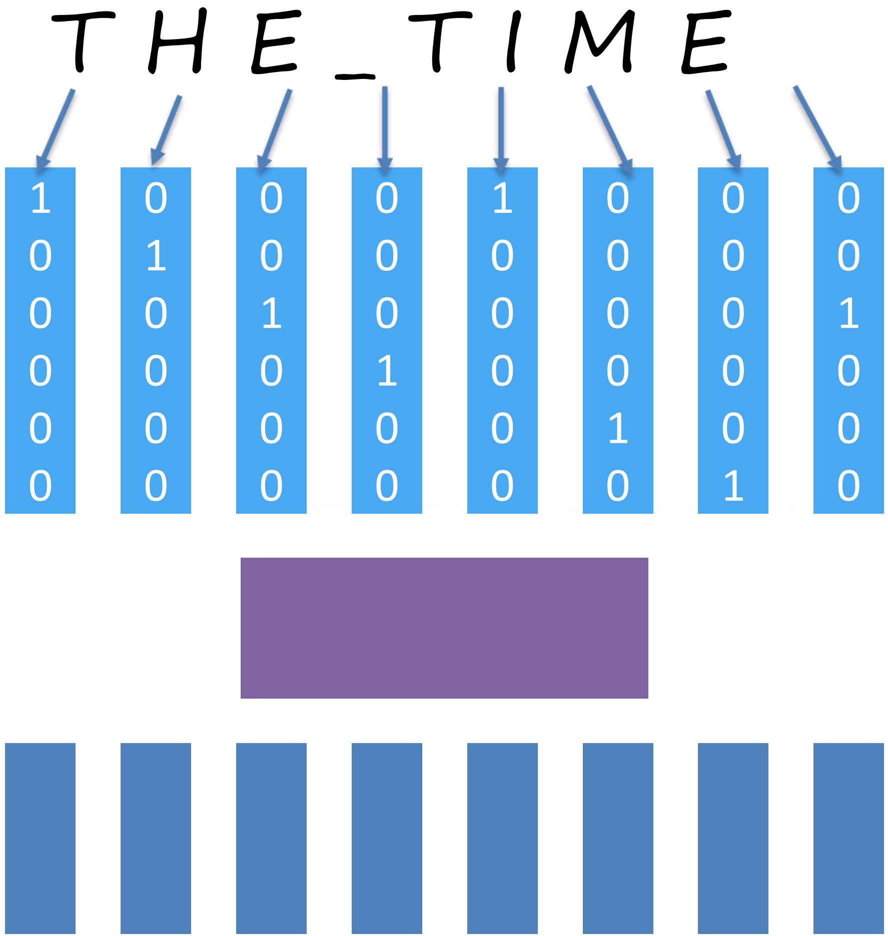
1-of-N Encoding lexicon = {apple, bag, cat, dog, elephant}

The vector is lexicon size.

Each dimension corresponds to a word in the lexicon

The dimension for the word is 1, and others are 0

apple	= [1 0 0 0 0]
bag	= [0 1 0 0 0]
cat	= [0 0 1 0 0]
dog	= [0 0 0 1 0]
elephant	= [0 0 0 0 1]



简洁向量 **v**

嵌入矩阵 **W**

嵌入向量 **v'**

具有隐含状态机制的RNN

- 输入向量序列 $\mathbf{x}_1, \ldots, \mathbf{x}_T$
- 隐含状态向量序列
  - $\mathbf{h}_1, \ldots, \mathbf{h}_T$
  - $\mathbf{h}_t = f(\mathbf{h}_{t-1}, \mathbf{x}_t)$
- 输出向量
  - 序列 $\mathbf{o}_1, \ldots, \mathbf{o}_T$ ； $\mathbf{o}_t = g(\mathbf{h}_t)$
  - 读取序列以生成隐含状态，然后开始生成输出
  - 输出向量通常用作下一个隐含状态的输入

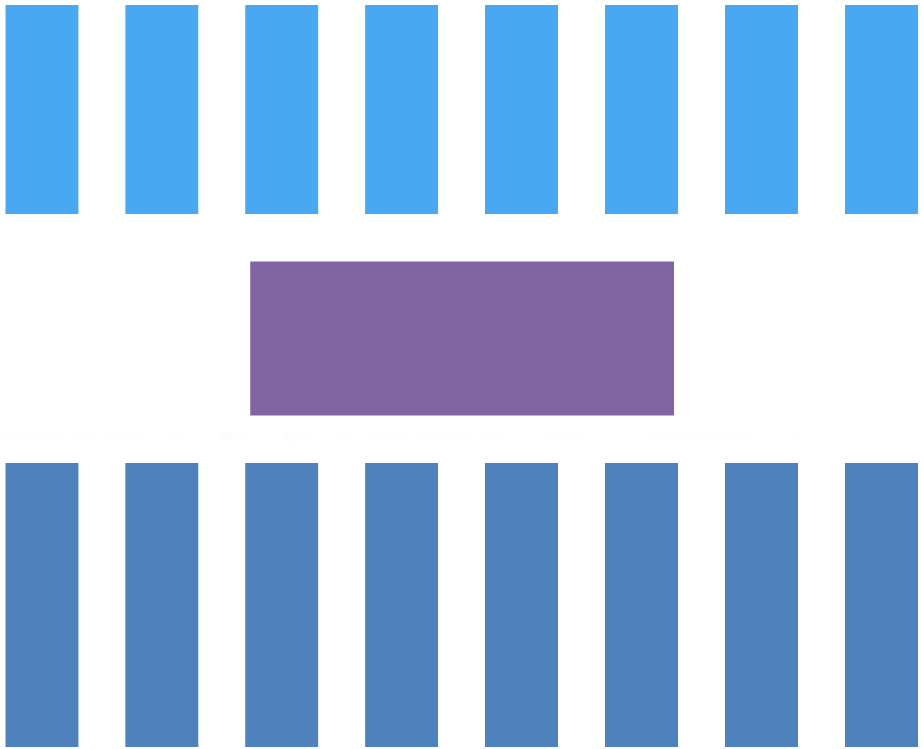
输出编码

输出向量 $\mathbf{o}$

解码矩阵 $\mathbf{W}'$

$$p(y|\mathbf{o}) \propto \exp\left(\mathbf{v}_y^\top \mathbf{o}\right) = \exp(\mathbf{o}[y])$$

独热解码



梯度（时间反向传播）

- 反向传播的长链依赖关系
  - 需要在内存中保留很多中间值
  - 蝴蝶效应
  - 梯度消失或发散（稍后会详细介绍）
- 裁剪梯度以防止发散

$$\mathbf{g} \leftarrow \min\left(1, \frac{\theta}{\|\mathbf{g}\|}\right)\mathbf{g}$$

- 重新缩放到最大尺寸为 θ 的梯度

困惑度

- 通常使用对数似然来测量准确度
- 这使得不同长度的输出无法比较

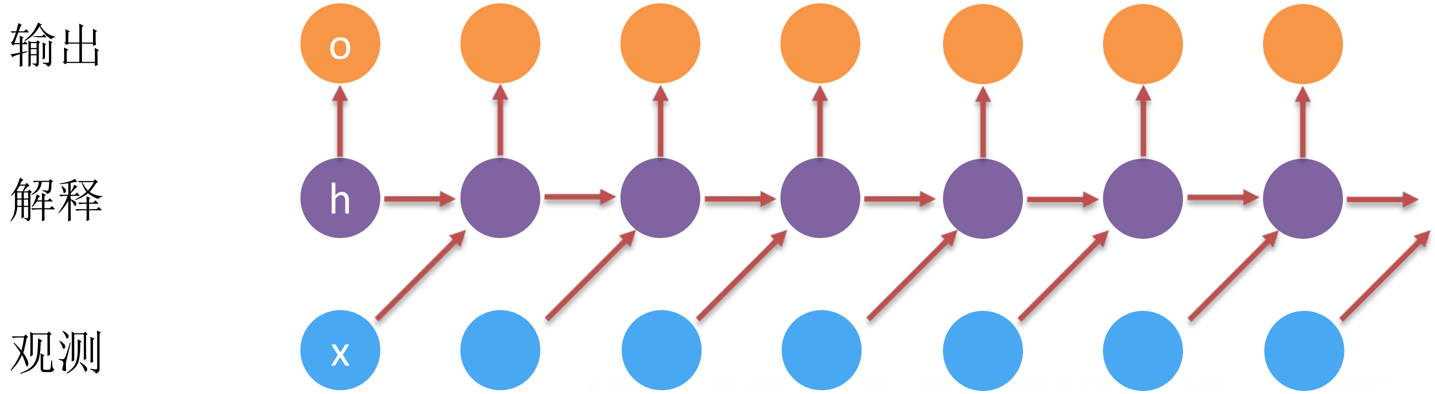
  （例如，一个坏模型的较短输出的效果可能比一个优秀模型的较长输出的性能具有更好的对数似然）
- 将对数似然标准化为序列长度

$$-\sum_{t=1}^{T} \log p(y_t|\text{model}) \qquad vs. \qquad \pi := -\frac{1}{T}\sum_{t=1}^{T} \log p(y_t|\text{model})$$

- 困惑度是指数版本 $\exp(\pi)$

(平均有效选择的数量)

循环神经网络



- 隐含状态更新

$$h_t = f(h_{t-1}, x_{t-1}, w)$$

- 观察更新

$$o_t = g(h_t, w)$$

目标函数

- RNN 生成的输出需要与目标标签进行比较

$$L(x, y, w) = \sum_{t=1}^{T} l(y_t, o_t)$$

- 梯度

$$\partial_w L = \sum_{t=1}^{T} \partial_w l(y_t, o_t)$$

$$= \sum_{t=1}^{T} \partial_{o_t} l(y_t, o_t) \left[\partial_w g(h_t, w) + \partial_{h_t} g(h_t, w) \partial_w h_t \right]$$

$\partial_w h_t$ 隐含状态梯度

- 目标函数

$$\partial_w L = \sum_{t=1}^{T} \partial_w l(y_t, o_t) = \sum_{t=1}^{T} \partial_{o_t} l(y_t, o_t) \left[\partial_w g(h_t, w) + \partial_{h_t} g(h_t, w) \partial_w h_t \right]$$

- 梯度递归

$$\partial_w h_t = \partial_w f(x_t, h_{t-1}, w) + \partial_h f(x_t, h_{t-1}, w) \partial_w h_{t-1}$$

$$= \sum_{i=t}^{1} \left[\prod_{j=t}^{i} \partial_h f(x_j, h_{j-1}, w) \right] \partial_w f(x_i, h_{i-1}, w)$$

$\partial_w h_t$ 隐含状态梯度

- 梯度递归

$$\partial_w h_t = \sum_{i=t}^{1} \left[\prod_{j=t}^{i} \partial_h f(x_j, h_{j-1}, w) \right] \partial_w f(x_i, h_{i-1}, w)$$

太多项

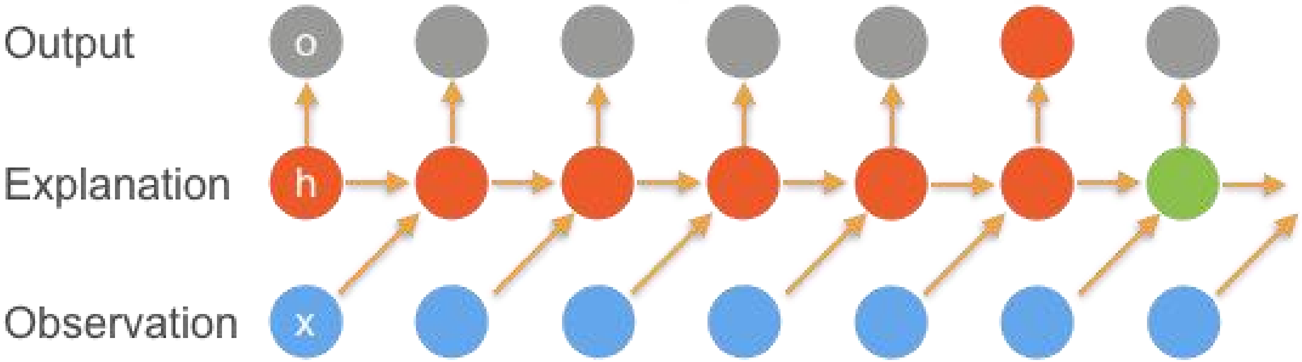
不稳定
(易发散)

开销大

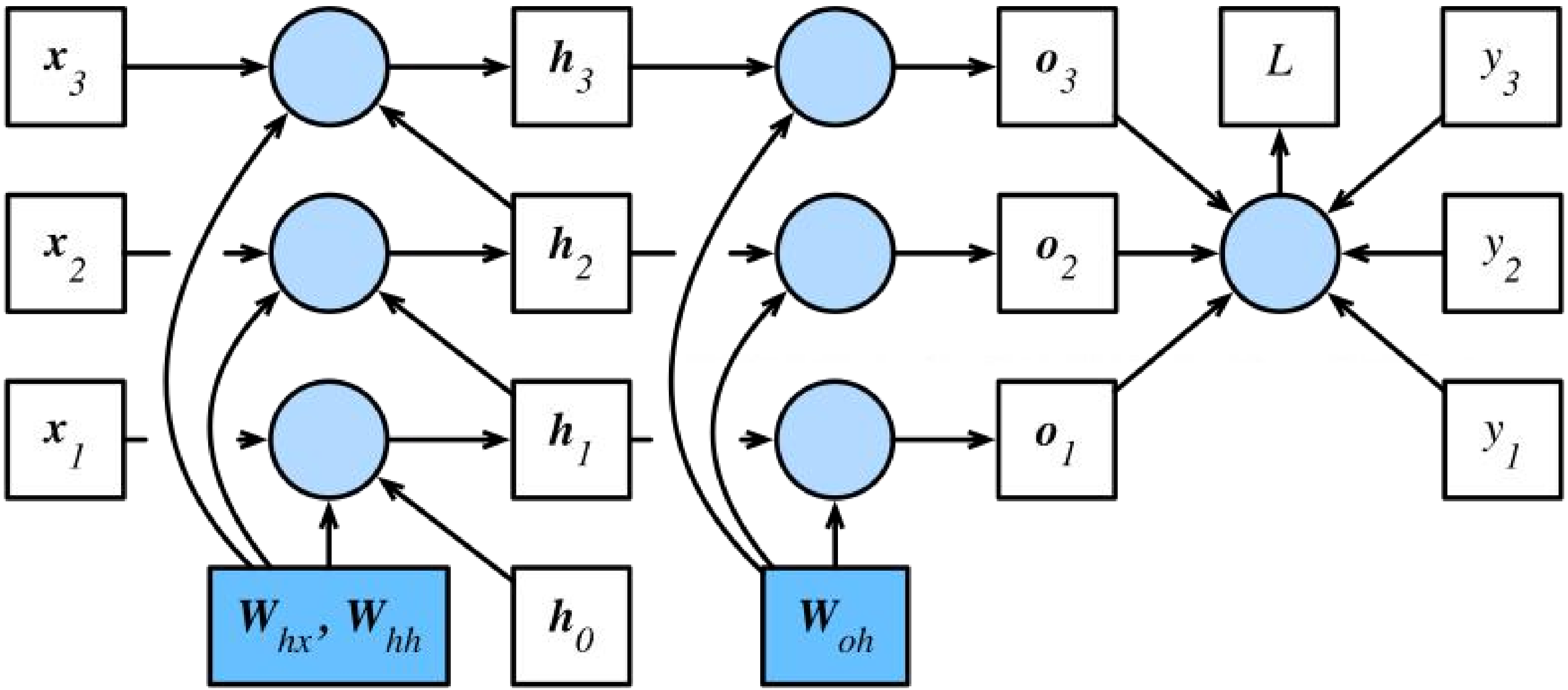
$\partial_w h_t$ 隐含状态梯度

- 梯度递归

$$\partial_w h_t = \sum_{i=t}^{1} \left[\prod_{j=t}^{i} \partial_h f(x_j, h_{j-1}, w) \right] \partial_w f(x_i, h_{i-1}, w)$$



计算图



示例

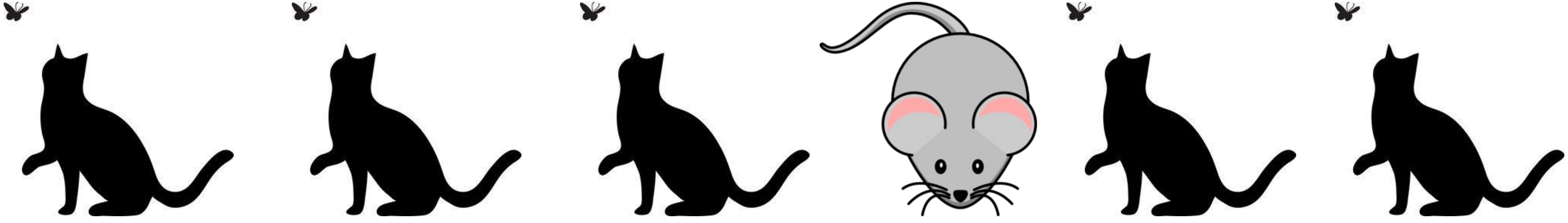
- 线性 RNN $\quad \mathbf{h}_t = \mathbf{W}_{hx}\mathbf{x}_t + \mathbf{W}_{hh}\mathbf{h}_{t-1}$ and $\mathbf{o}_t = \mathbf{W}_{oh}\mathbf{h}_t$
- 输出梯度 $\quad \partial_{\mathbf{W}_{oh}} L = \sum_{t=1}^{T} \operatorname{prod}\left(\partial_{\mathbf{o}_t} l(\mathbf{o}_t, y_t), \mathbf{h}_t\right)$
- 更新梯度

$$\partial_{\mathbf{W}_{hh}} L = \sum_{t=1}^{T} \operatorname{prod}\left(\partial_{\mathbf{o}_t} l(\mathbf{o}_t, y_t), \mathbf{W}_{oh}, \partial_{\mathbf{W}_{hh}} \mathbf{h}_t\right)$$

$$\partial_{\mathbf{W}_{hx}} L = \sum_{t=1}^{T} \operatorname{prod}\left(\partial_{\mathbf{o}_t} l(\mathbf{o}_t, y_t), \mathbf{W}_{oh}, \partial_{\mathbf{W}_{hx}} \mathbf{h}_t\right)$$

门控循环单元（GRU）

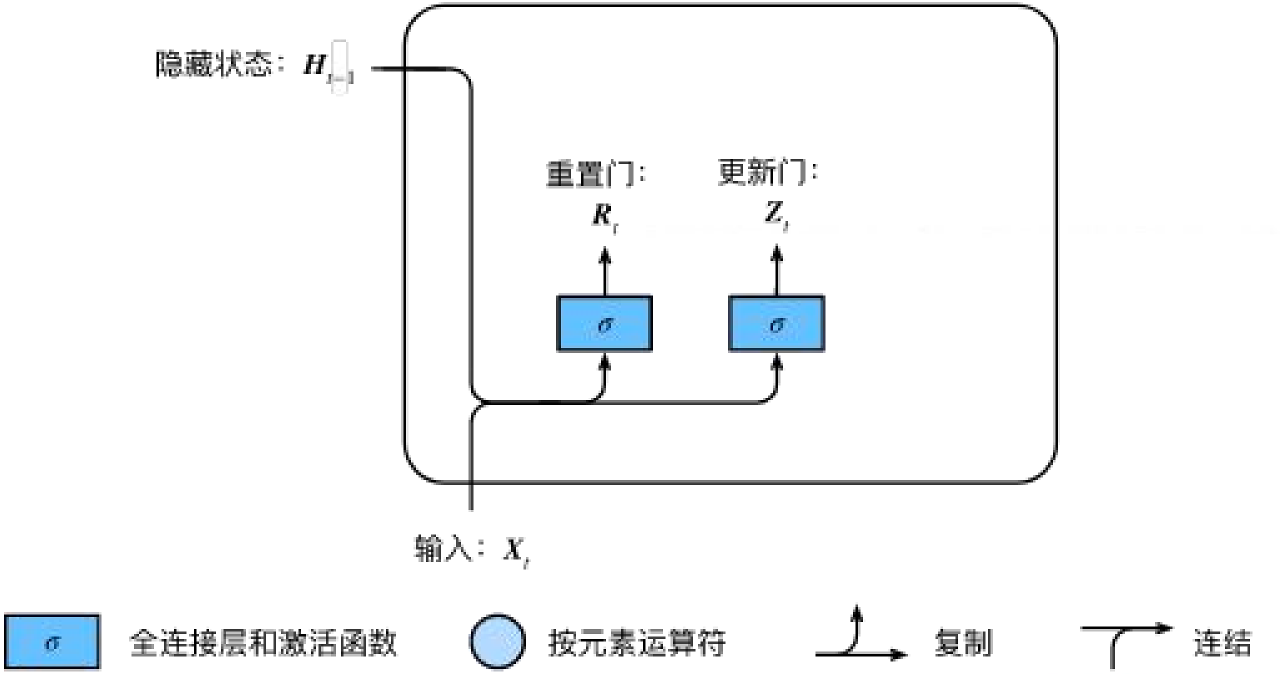
- 并非所有元素都具有同等意义



- 只记住相关的元素
  - 需要注意的机制（更新门）
  - 需要忘记的机制（重置门）

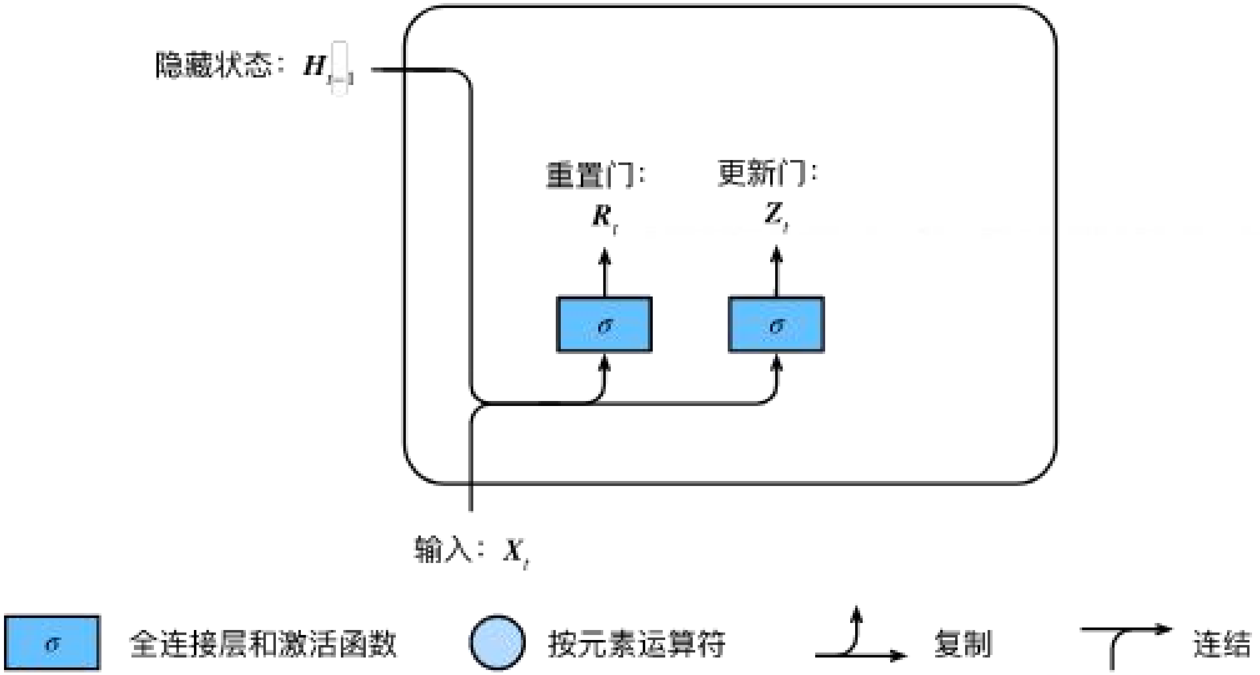
门控循环单元

$$\boldsymbol{R}_t = \sigma(\boldsymbol{X}_t \boldsymbol{W}_{xr} + \boldsymbol{H}_{t-1} \boldsymbol{W}_{hr} + \boldsymbol{b}_r),$$
$$\boldsymbol{Z}_t = \sigma(\boldsymbol{X}_t \boldsymbol{W}_{xz} + \boldsymbol{H}_{t-1} \boldsymbol{W}_{hz} + \boldsymbol{b}_z)$$



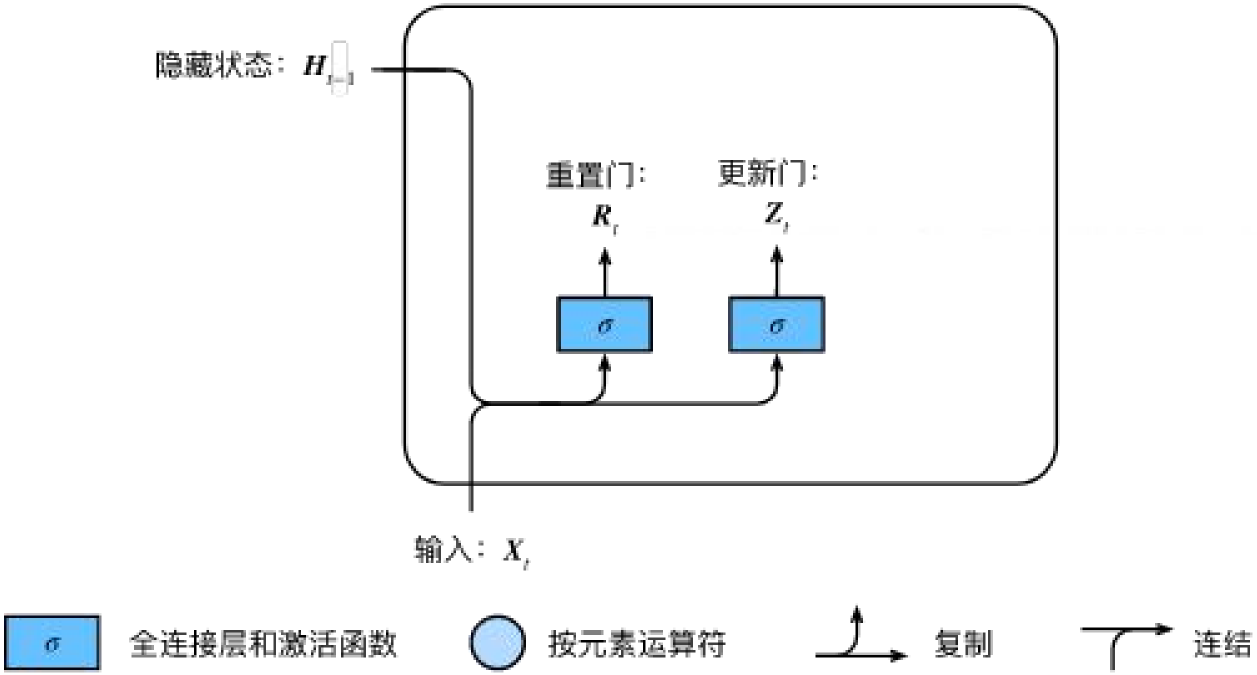
候选隐含状态

$$\tilde{\boldsymbol{H}}_t = tanh(\boldsymbol{X}_t\boldsymbol{W}_{xh} + (\boldsymbol{R}_t \odot \boldsymbol{H}_{t-1})\boldsymbol{W}_{hh} + \boldsymbol{b}_h)$$

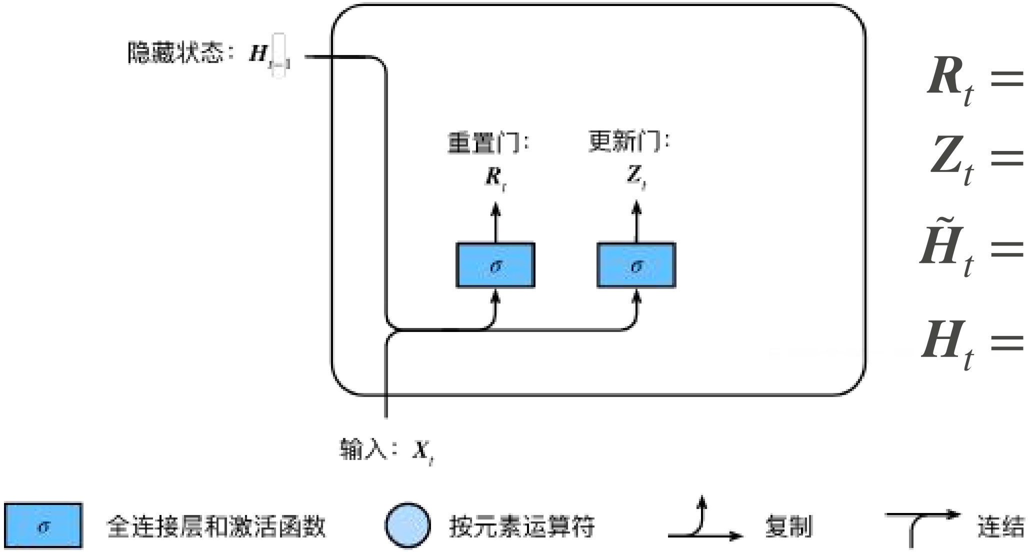


隐含状态

$$H_t = Z_t \odot H_{t-1} + (1 - Z_t) \odot \tilde{H}_t$$



门控循环单元（GRU）总结



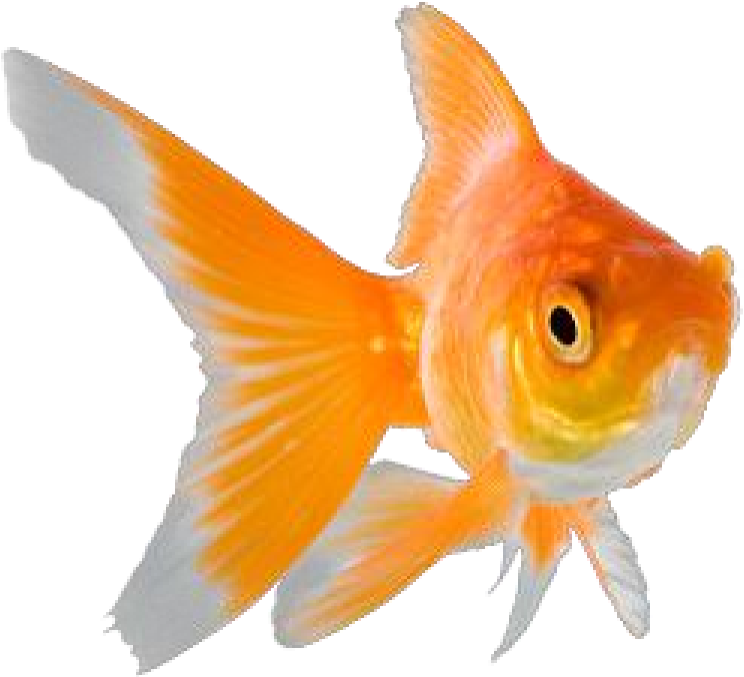
$$\boldsymbol{R}_t = \sigma(\boldsymbol{X}_t\boldsymbol{W}_{xr} + \boldsymbol{H}_{t-1}\boldsymbol{W}_{hr} + \boldsymbol{b}_r),$$

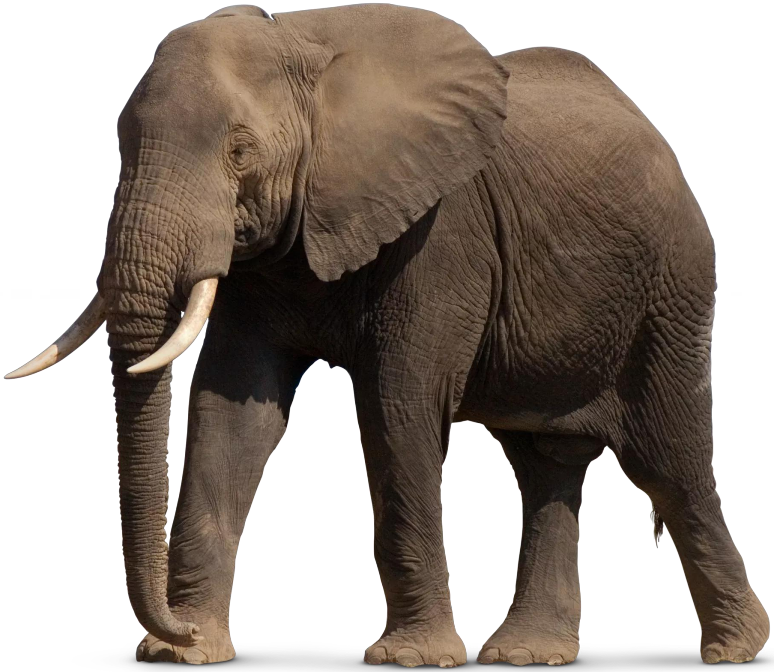
$$\boldsymbol{Z}_t = \sigma(\boldsymbol{X}_t\boldsymbol{W}_{xz} + \boldsymbol{H}_{t-1}\boldsymbol{W}_{hz} + \boldsymbol{b}_z)$$

$$\tilde{\boldsymbol{H}}_t = \tanh(\boldsymbol{X}_t\boldsymbol{W}_{xh} + \left(\boldsymbol{R}_t \odot \boldsymbol{H}_{t-1}\right)\boldsymbol{W}_{hh} + \boldsymbol{b}_h)$$

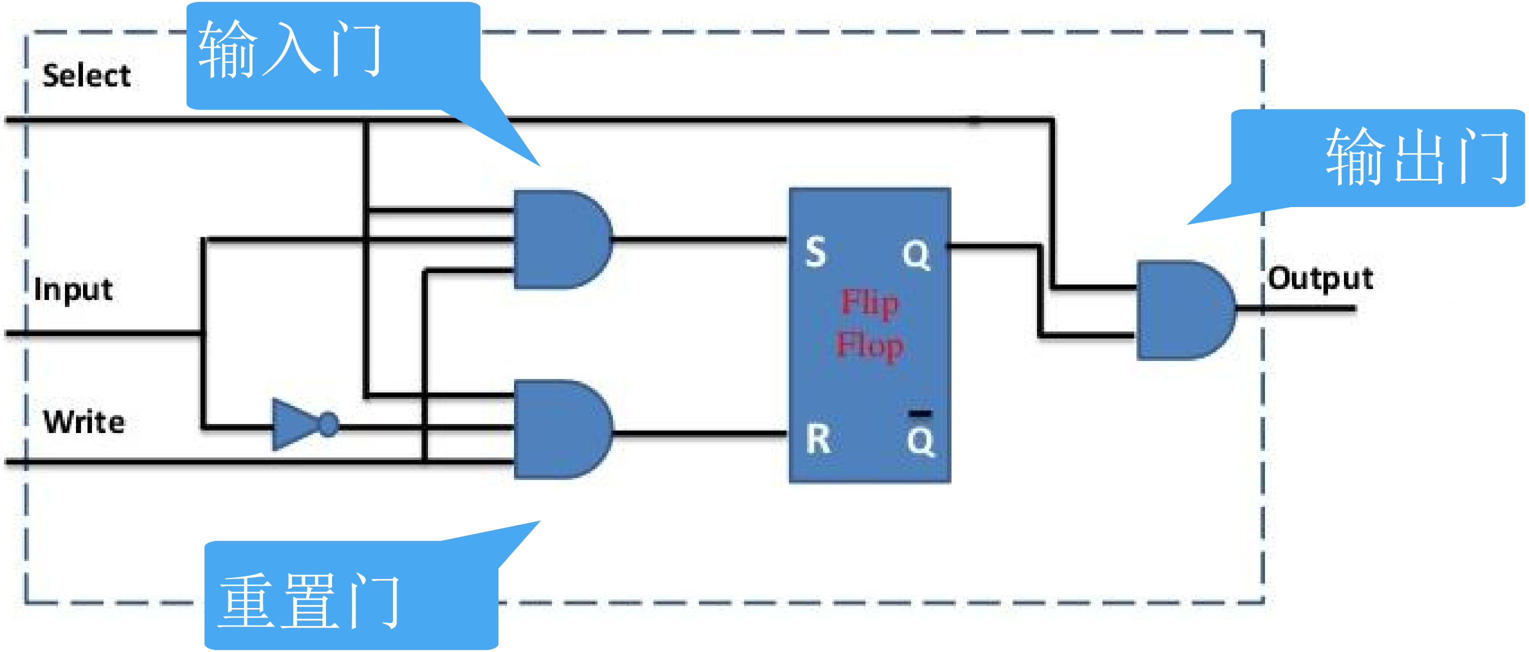
$$\boldsymbol{H}_t = \boldsymbol{Z}_t \odot \boldsymbol{H}_{t-1} + (1 - \boldsymbol{Z}_t) \odot \tilde{\boldsymbol{H}}_t$$

长短期记忆（LSTM）





电路联想



长短期记忆（LSTM）

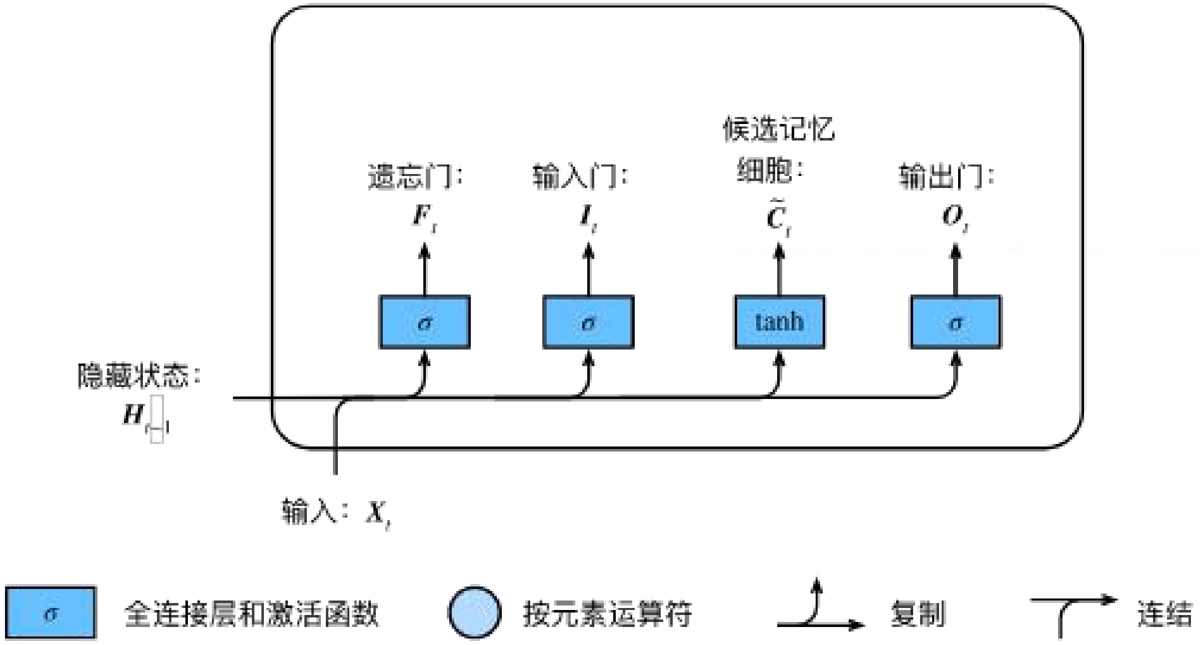
- 遗忘门
  - 将每个值尽可能收缩为零
- 输入门
  - 决定是否应忽略输入数据
- 输出门
  - 决定隐含状态是否用于 LSTM 生成的输出

输入门、遗忘门和输出门

$$I_t = \sigma(X_t W_{xi} + H_{t-1} W_{hi} + b_i)$$

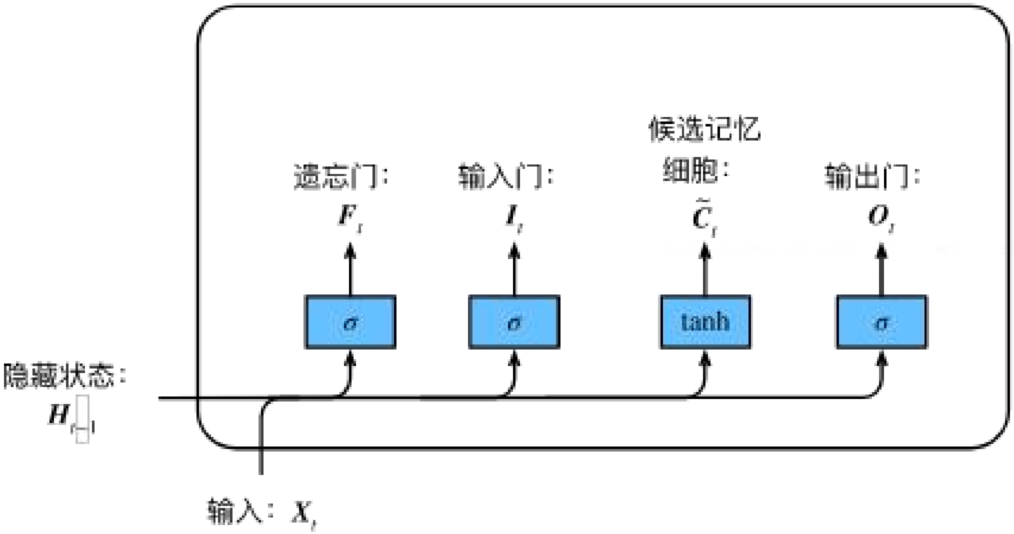
$$F_t = \sigma(X_t W_{xf} + H_{t-1} W_{hf} + b_f)$$

$$O_t = \sigma(X_t W_{xo} + H_{t-1} W_{ho} + b_o)$$



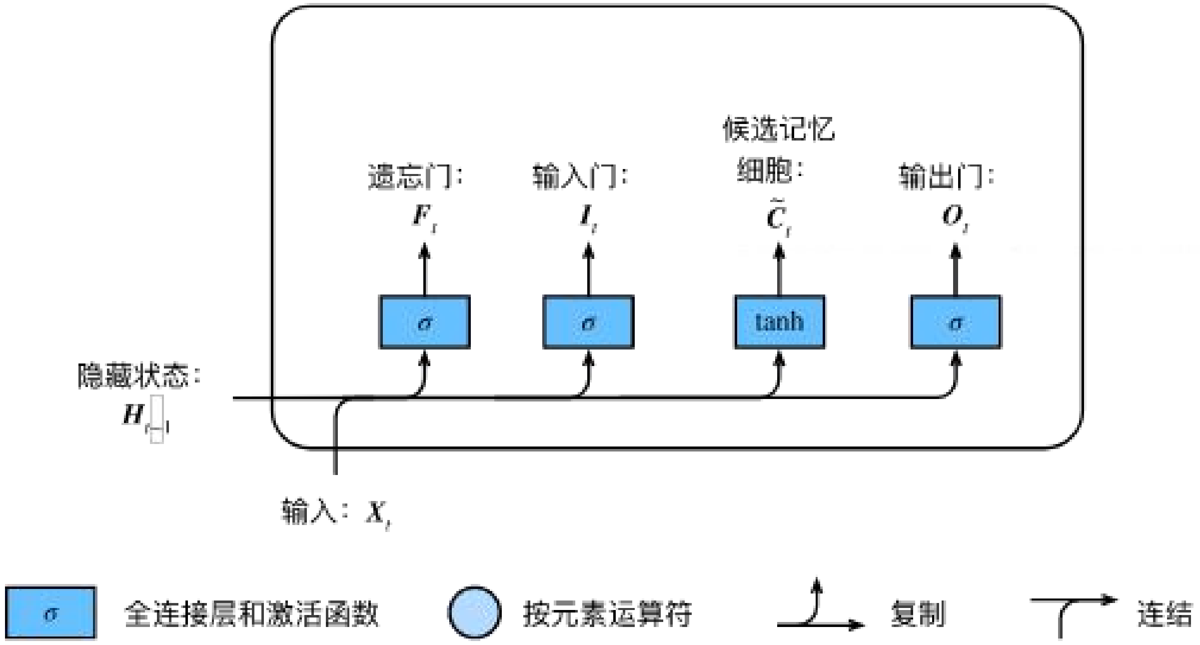
候选记忆细胞

$$\tilde{C}_t = tanh(X_t W_{xc} + H_{t-1} W_{hc} + b_c)$$



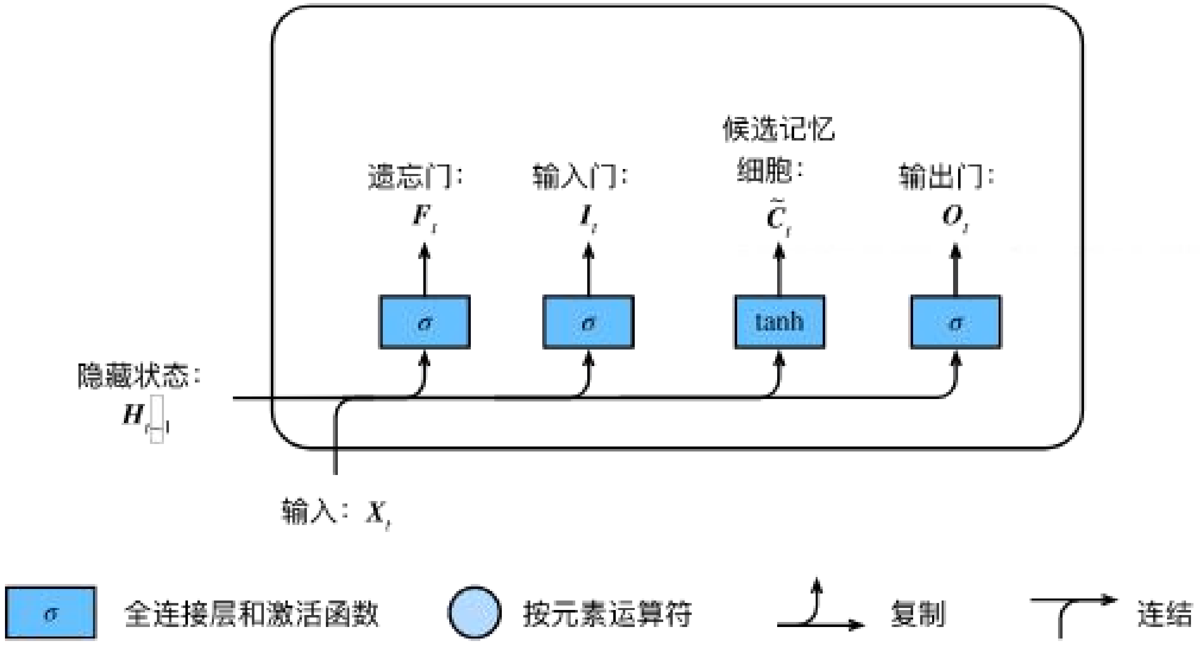
记忆细胞

$$C_t = F_t \odot C_{t-1} + I_t \odot \tilde{C}_t$$

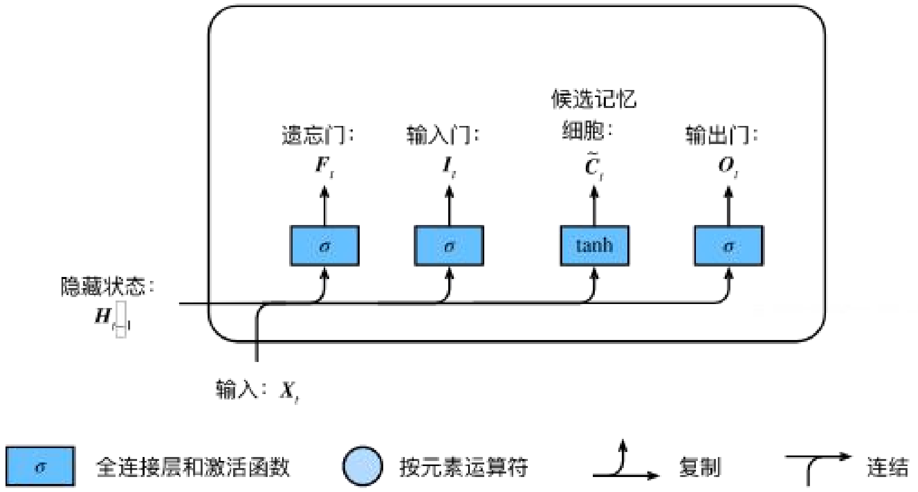


隐含状态

$$\boldsymbol{H}_t = \boldsymbol{O}_t \odot tanh(\boldsymbol{C}_t)$$



长短期记忆（LSTM）总结



$$I_t = \sigma(X_t W_{xi} + H_{t-1} W_{hi} + b_i)$$

$$F_t = \sigma(X_t W_{xf} + H_{t-1} W_{hf} + b_f)$$

$$O_t = \sigma(X_t W_{xo} + H_{t-1} W_{ho} + b_o)$$

$$\tilde{C}_t = \tanh(X_t W_{xc} + H_{t-1} W_{hc} + b_c)$$

$$C_t = F_t \odot C_{t-1} + I_t \odot \tilde{C}_t$$

$$H_t = O_t \odot \tanh(C_t)$$

总结

- 循环神经网络
  - 实现 RNN 语言模型
  - 时间反向传播
- 门控循环单元（GRU）
- 长短期记忆（LSTM）