Deep Learning I

10. 现代循环神经网络

WU Xiaokun 吴晓堃

xkun.wu [at] gmail

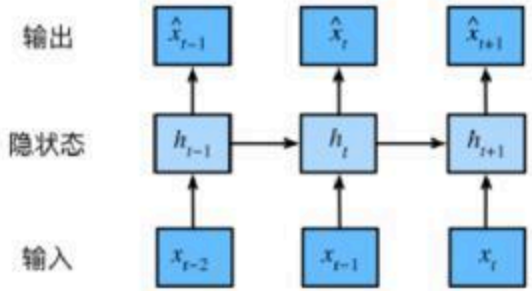
2022/05/04

长短期记忆网络（LSTM）

回顾：隐变量模型的信息传递

当前隐状态$\mathbf{H}_t$整合过去所有的观测信息

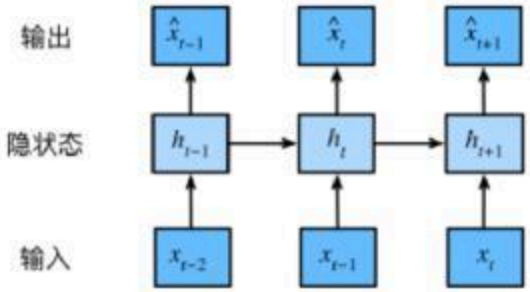
$$\mathbf{H}_t = \sigma(\mathbf{X}_t \mathbf{W}_x^h + \mathbf{H}_{t-1} \mathbf{W}_h^h + \mathbf{b}^h)$$
$$\mathbf{O}_t = \mathbf{H}_t \mathbf{W}_h^o + \mathbf{b}^o$$



回顾：隐变量模型的信息传递

当前隐状态$\mathbf{H}_t$整合过去所有的观测信息

$$\mathbf{H}_t = \sigma(\mathbf{X}_t \mathbf{W}_x^h + \mathbf{H}_{t-1} \mathbf{W}_h^h + \mathbf{b}^h)$$
$$\mathbf{O}_t = \mathbf{H}_t \mathbf{W}_h^o + \mathbf{b}^o$$



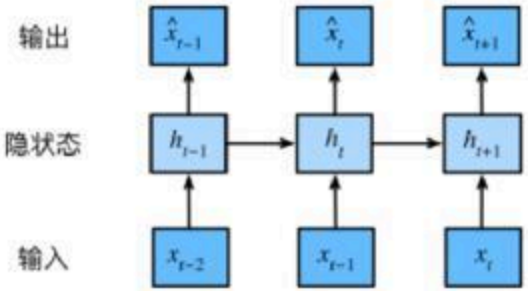
不同时间步使用**相同参数**：$\mathbf{W}_x^h, \mathbf{W}_h^h, \mathbf{W}_h^o, b^h, b^o$

- 每个时间步整合新的观测：长期信息在不断更新中逐渐消失
  - 模型架构、**存储空间固定**：新信息必定挤占旧信息的位置，类比监控磁盘

回顾：隐变量模型的信息传递

当前隐状态$\mathbf{H}_t$整合过去所有的观测信息

$$\mathbf{H}_t = \sigma(\mathbf{X}_t \mathbf{W}_x^h + \mathbf{H}_{t-1} \mathbf{W}_h^h + \mathbf{b}^h)$$
$$\mathbf{O}_t = \mathbf{H}_t \mathbf{W}_h^o + \mathbf{b}^o$$



不同时间步使用**相同参数**：$\mathbf{W}_x^h, \mathbf{W}_h^h, \mathbf{W}_h^o, b^h, b^o$

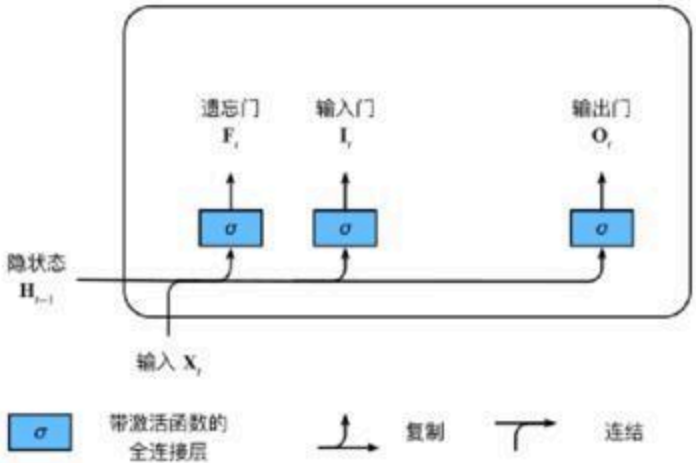
- 每个时间步整合新的观测：长期信息在不断更新中逐渐消失
  - 模型架构、**存储空间固定**：新信息必定挤占旧信息的位置，类比监控磁盘

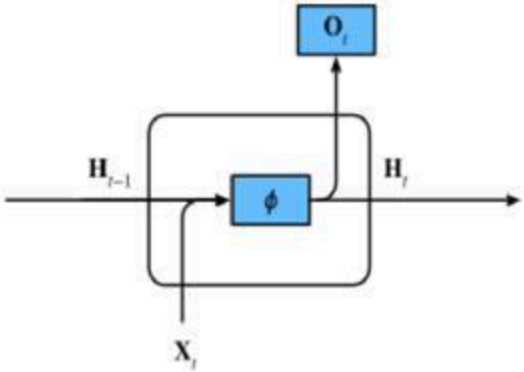
[Hochreiter 1997] LSTM：引入类似逻辑门电路的**记忆元**

- 根据需要（重要程度）选择保存、丢弃信息：解决**信息选取**的问题

LSTM：输入门、忘记门、输出门

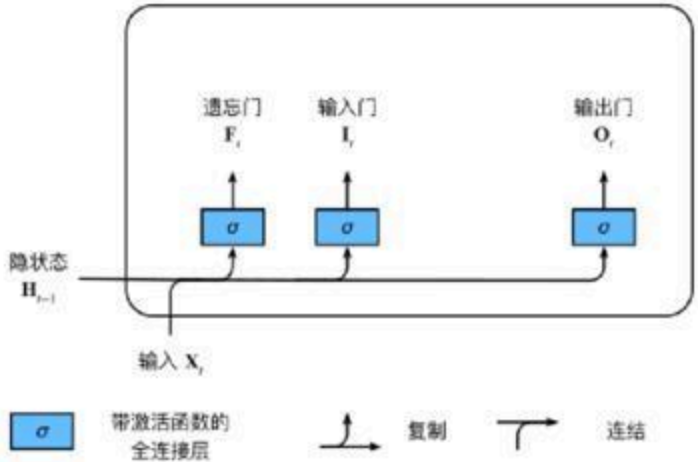
门控：决定信息**整合比例**（类比电流，可想象成水阀）

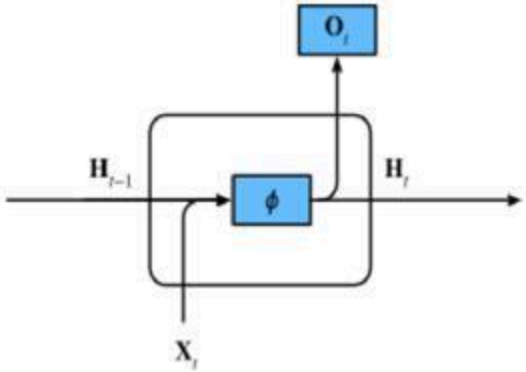




LSTM：输入门、忘记门、输出门

门控：决定信息整合比例（类比电流，可想象成水阀）





$$\mathbf{I}_t = \sigma(\mathbf{X}_t\mathbf{W}_x^i + \mathbf{H}_{t-1}\mathbf{W}_h^i + \mathbf{b}^i)$$
$$\mathbf{F}_t = \sigma(\mathbf{X}_t\mathbf{W}_x^f + \mathbf{H}_{t-1}\mathbf{W}_h^f + \mathbf{b}^f)$$
$$\mathbf{O}_t = \sigma(\mathbf{X}_t\mathbf{W}_x^o + \mathbf{H}_{t-1}\mathbf{W}_h^o + \mathbf{b}^o)$$

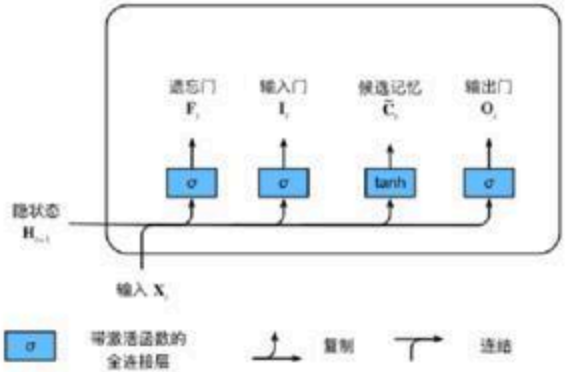
$$\mathbf{H}_t = \sigma(\mathbf{X}_t\mathbf{W}_x^h + \mathbf{H}_{t-1}\mathbf{W}_h^h + \mathbf{b}$$
$$\mathbf{O}_t = \mathbf{H}_t\mathbf{W}_h^o + \mathbf{b}^o$$

LSTM：候选记忆元

候选记忆：来自当前时间步的**编码信息**

$$\tilde{\mathbf{C}}_t = \tanh(\mathbf{X}_t \mathbf{W}_x^c + \mathbf{H}_{t-1} \mathbf{W}_h^c + \mathbf{b}^c$$

- tanh激活：只挤压，不截取

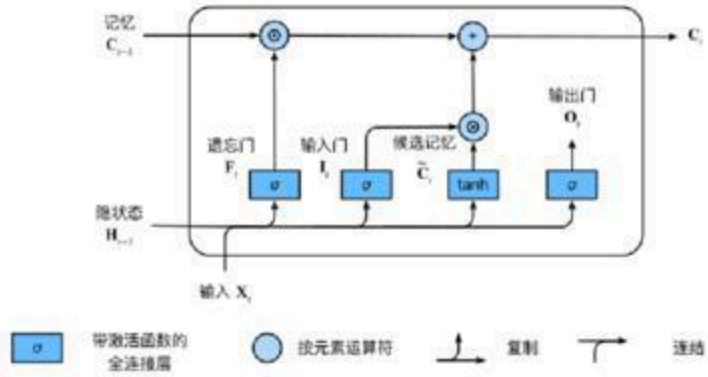


- 对比门控（sigma激活）：从信息本身分析重要程度，即**权重值**

LSTM：记忆元

输入门$\mathbf{I}_t$控制采用多少来自$\tilde{\mathbf{C}}_t$的新数据；遗忘门$\mathbf{F}_t$控制保留多少过去的记忆$\mathbf{C}_{t-1}$

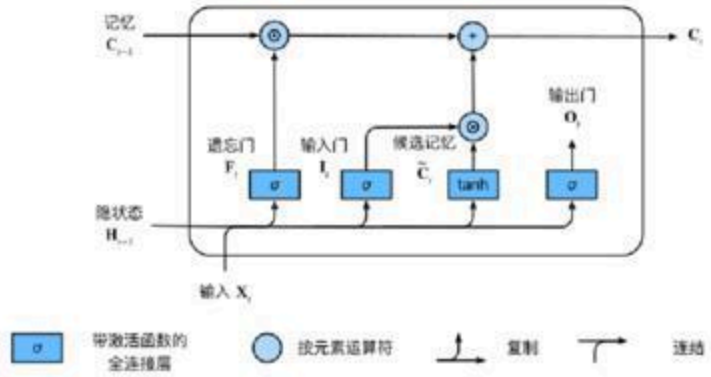
$$\mathbf{C}_t = \mathbf{F}_t \odot \mathbf{C}_{t-1} + \mathbf{I}_t \odot \tilde{\mathbf{C}}_t$$



LSTM：记忆元

输入门$\mathbf{I}_t$控制**采用**多少来自$\tilde{\mathbf{C}}_t$的新数据；遗忘门$\mathbf{F}_t$控制**保留**多少过去的记忆$\mathbf{C}_{t-1}$

$$\mathbf{C}_t = \mathbf{F}_t \odot \mathbf{C}_{t-1} + \mathbf{I}_t \odot \tilde{\mathbf{C}}_t$$



例如$\mathbf{F}_t = 1, \mathbf{I}_t = 0$：**所有记忆都保存**并传递到当前时间步

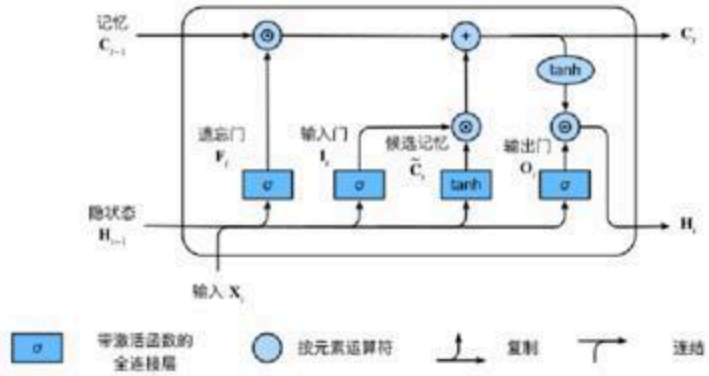
- 缓解梯度消失问题，并捕获**长距离依赖**关系
- 反之，$\mathbf{F}_t = 0, \mathbf{I}_t = 1$：重置记忆，由当前时间步初始化

LSTM：隐状态

输出门$\mathbf{O}_t$控制**传出**多少更新后的记忆$\mathbf{C}_t$

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

- tanh激活：只挤压，不截取

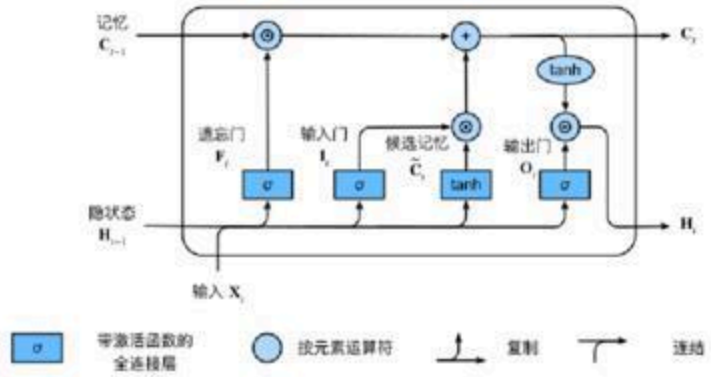


LSTM：隐状态

输出门$\mathbf{O}_t$控制**传出**多少更新后的记忆$\mathbf{C}_t$

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

- tanh激活：只挤压，不截取



例如$\mathbf{O}_t = 1$：**所有记忆都传递**给预测

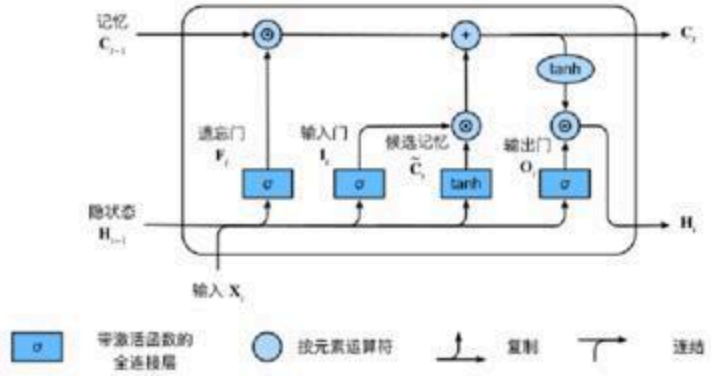
- $\mathbf{O}_t = 0$：记忆传递给下一时间步；隐状态清零

LSTM：隐状态

输出门$\mathbf{O}_t$控制**传出**多少更新后的记忆$\mathbf{C}_t$

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

- tanh激活：只挤压，不截取



例如$\mathbf{O}_t = 1$：**所有记忆都传递**给预测

- $\mathbf{O}_t = 0$：记忆传递给下一时间步；隐状态清零

注意：只有隐状态$\mathbf{H}_t$**传递到输出层**；记忆元$\mathbf{C}_t$不直接参与输出计算

- 思考：似乎记忆元与隐状态的作用有些类似？

实验：LSTM

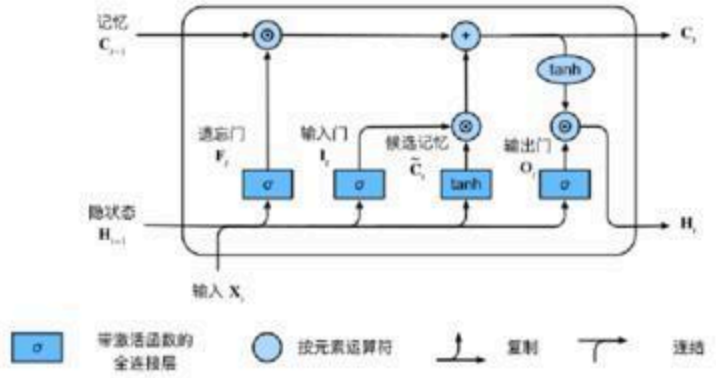
小结：LSTM

- LSTM使用三种门控单元：输入门、遗忘门、输出门
  - 输出小数（百分比）：想象成水阀
- 隐藏层输出：隐状态（传递到输出层）、记忆元（内部传递）
  - 输出挤压到$(-1, 1)$：归一化的信息
- 缓解梯度消失、梯度爆炸；可以捕获长距离依赖关系

门控循环单元（GRU）

LSTM的计算问题

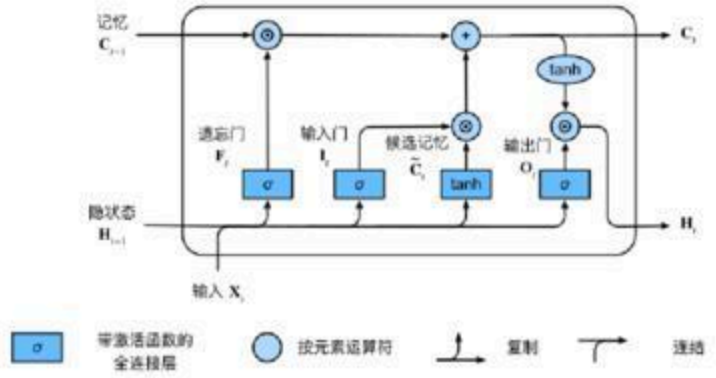
$$\mathbf{I}_t = \sigma(\mathbf{X}_t \mathbf{W}_x^i + \mathbf{H}_{t-1} \mathbf{W}_h^i + \mathbf{b}^i)$$
$$\mathbf{F}_t = \sigma(\mathbf{X}_t \mathbf{W}_x^f + \mathbf{H}_{t-1} \mathbf{W}_h^f + \mathbf{b}^f)$$
$$\mathbf{O}_t = \sigma(\mathbf{X}_t \mathbf{W}_x^o + \mathbf{H}_{t-1} \mathbf{W}_h^o + \mathbf{b}^o)$$
$$\tilde{\mathbf{C}}_t = \tanh(\mathbf{X}_t \mathbf{W}_x^c + \mathbf{H}_{t-1} \mathbf{W}_h^c + \mathbf{b}^c)$$
$$\mathbf{C}_t = \mathbf{F}_t \odot \mathbf{C}_{t-1} + \mathbf{I}_t \odot \tilde{\mathbf{C}}_t$$
$$\mathbf{H}_t = \mathbf{O}_t \odot \tanh(\mathbf{C}_t)$$
$$\mathbf{Y}_t = \mathbf{H}_t \mathbf{W}_h^y + \mathbf{b}^y$$



参数：$\mathbf{W}_x^i, \mathbf{W}_h^i, \mathbf{W}_x^f, \mathbf{W}_h^f, \mathbf{W}_x^o, \mathbf{W}_h^o, \mathbf{W}_x^c, \mathbf{W}_h^c, \mathbf{W}_h^y, b^i, b^f, b^o, b^c, \mathbf{b}^y$

LSTM的计算问题

$$\mathbf{I}_t = \sigma(\mathbf{X}_t \mathbf{W}_x^i + \mathbf{H}_{t-1} \mathbf{W}_h^i + \mathbf{b}^i)$$
$$\mathbf{F}_t = \sigma(\mathbf{X}_t \mathbf{W}_x^f + \mathbf{H}_{t-1} \mathbf{W}_h^f + \mathbf{b}^f)$$
$$\mathbf{O}_t = \sigma(\mathbf{X}_t \mathbf{W}_x^o + \mathbf{H}_{t-1} \mathbf{W}_h^o + \mathbf{b}^o)$$
$$\tilde{\mathbf{C}}_t = \tanh(\mathbf{X}_t \mathbf{W}_x^c + \mathbf{H}_{t-1} \mathbf{W}_h^c + \mathbf{b}^c)$$
$$\mathbf{C}_t = \mathbf{F}_t \odot \mathbf{C}_{t-1} + \mathbf{I}_t \odot \tilde{\mathbf{C}}_t$$
$$\mathbf{H}_t = \mathbf{O}_t \odot \tanh(\mathbf{C}_t)$$
$$\mathbf{Y}_t = \mathbf{H}_t \mathbf{W}_h^y + \mathbf{b}^y$$



参数：$\mathbf{W}_x^i, \mathbf{W}_h^i, \mathbf{W}_x^f, \mathbf{W}_h^f, \mathbf{W}_x^o, \mathbf{W}_h^o, \mathbf{W}_x^c, \mathbf{W}_h^c, \mathbf{W}_h^y, b^i, b^f, b^o, b^c, \mathbf{b}^y$

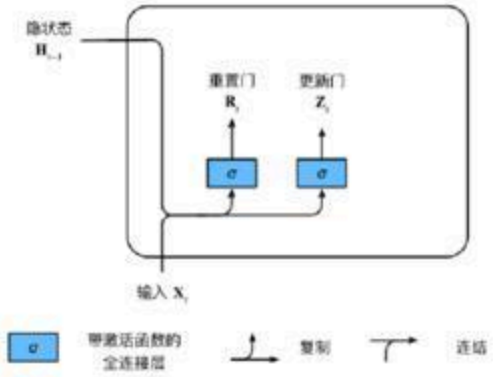
[Cho 2014] GRU：（**简化**的）门控循环网络，且将记忆元**合并**到隐状态

GRU：重置门、更新门

重置门：控制“**可能还想记住**”的候选隐状态数量；更新门：控制**更新比例**

$$\mathbf{R}_t = \sigma(\mathbf{X}_t \mathbf{W}_x^r + \mathbf{H}_{t-1} \mathbf{W}_h^r + \mathbf{b}^r)$$
$$\mathbf{Z}_t = \sigma(\mathbf{X}_t \mathbf{W}_x^z + \mathbf{H}_{t-1} \mathbf{W}_h^z + \mathbf{b}^z)$$

- sigma激活：转换成权重（百分比）值

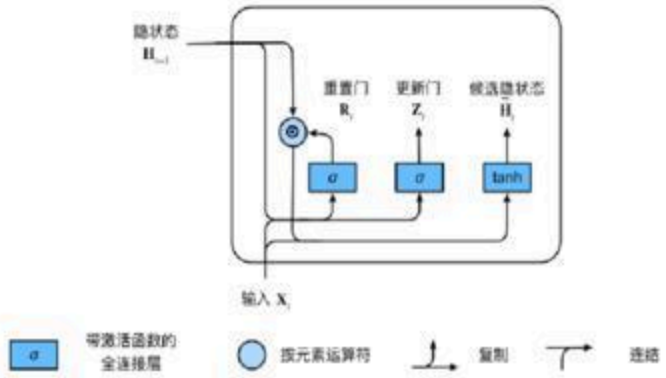


GRU：候选隐状态

$\mathbf{R}_t$门控：取值$(0,1)$，减少以往隐状态的影响

$$\tilde{\mathbf{H}}_t = \tanh(\mathbf{X}_t \mathbf{W}_x^h + (\mathbf{R}_t \odot \mathbf{H}_{t-1})\mathbf{W}_h^h +$$

- tanh激活：只挤压，不截取

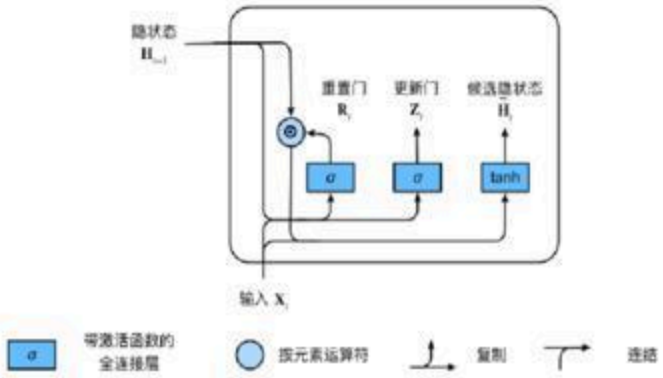


GRU：候选隐状态

$\mathbf{R}_t$门控：取值$(0, 1)$，减少以往隐状态的影响

$$\tilde{\mathbf{H}}_t = \tanh(\mathbf{X}_t \mathbf{W}_x^h + (\mathbf{R}_t \odot \mathbf{H}_{t-1}) \mathbf{W}_h^h +$$

- tanh激活：只挤压，不截取



例如$\mathbf{R}_t = 1$：**所有隐状态都传递**，等价于普通循环网络

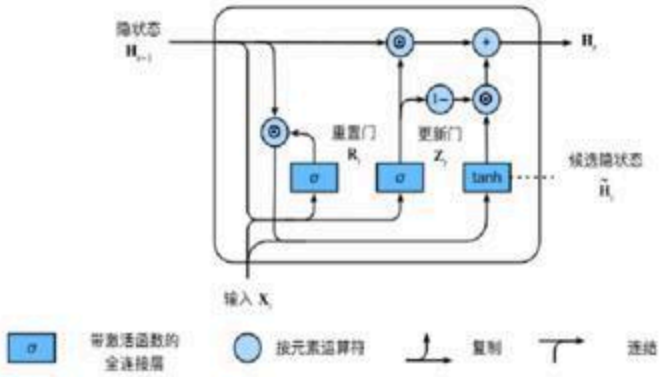
- $\mathbf{R}_t = 0$：所有隐状态重置，等价于多层感知机

GRU：隐状态

$\mathbf{Z}_t$门控：决定新、旧隐状态整合的比例

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

- 记忆元合并到隐状态

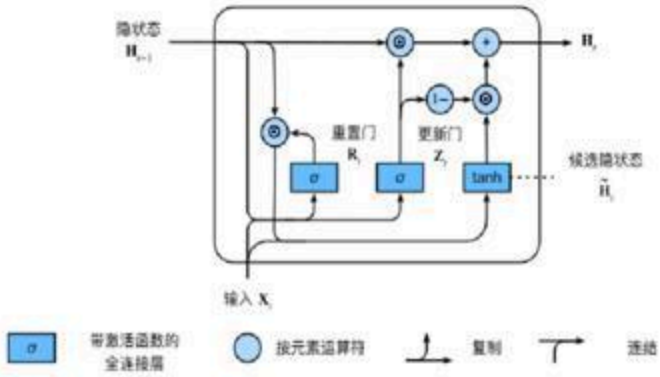


GRU：隐状态

$\mathbf{Z}_t$门控：决定新、旧隐状态整合的比例

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

- 记忆元合并到隐状态

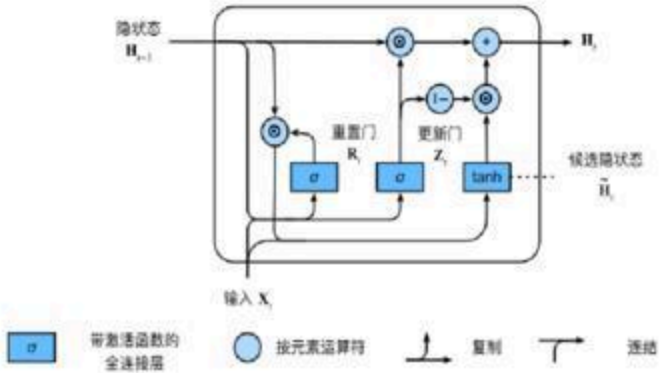


例如$\mathbf{Z}_t = 1$：只保留旧状态，即跳过当前时间步

- 缓解梯度消失问题，并捕获**长距离依赖**关系
- $\mathbf{Z}_t = 0$：只传递候选隐状态

GRU：计算量

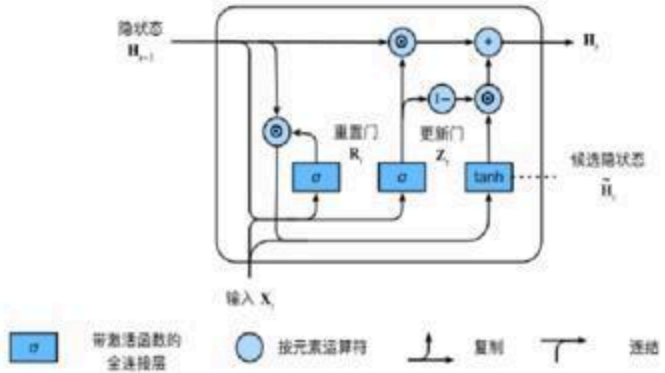
$$\mathbf{R}_t = \sigma(\mathbf{X}_t \mathbf{W}_x^r + \mathbf{H}_{t-1} \mathbf{W}_h^r + \mathbf{b}^r)$$
$$\mathbf{Z}_t = \sigma(\mathbf{X}_t \mathbf{W}_x^z + \mathbf{H}_{t-1} \mathbf{W}_h^z + \mathbf{b}^z)$$
$$\tilde{\mathbf{H}}_t = \tanh(\mathbf{X}_t \mathbf{W}_x^h + (\mathbf{R}_t \odot \mathbf{H}_{t-1}) \mathbf{W}_h^h +$$
$$\mathbf{H}_t = \mathbf{Z}_t \odot \mathbf{H}_{t-1} + (1 - \mathbf{Z}_t) \odot \tilde{\mathbf{H}}_t$$
$$\mathbf{Y}_t = \mathbf{H}_t \mathbf{W}_h^y + \mathbf{b}^y$$



参数：$\mathbf{W}_x^r, \mathbf{W}_h^r, \mathbf{W}_x^z, \mathbf{W}_h^z, \mathbf{W}_x^h, \mathbf{W}_h^h, \mathbf{W}_h^y, b^r, b^z, b^h, \mathbf{b}^y$

GRU：计算量

$$\mathbf{R}_t = \sigma(\mathbf{X}_t \mathbf{W}_x^r + \mathbf{H}_{t-1} \mathbf{W}_h^r + \mathbf{b}^r)$$
$$\mathbf{Z}_t = \sigma(\mathbf{X}_t \mathbf{W}_x^z + \mathbf{H}_{t-1} \mathbf{W}_h^z + \mathbf{b}^z)$$
$$\tilde{\mathbf{H}}_t = \tanh(\mathbf{X}_t \mathbf{W}_x^h + (\mathbf{R}_t \odot \mathbf{H}_{t-1}) \mathbf{W}_h^h +$$
$$\mathbf{H}_t = \mathbf{Z}_t \odot \mathbf{H}_{t-1} + (1 - \mathbf{Z}_t) \odot \tilde{\mathbf{H}}_t$$
$$\mathbf{Y}_t = \mathbf{H}_t \mathbf{W}_h^y + \mathbf{b}^y$$

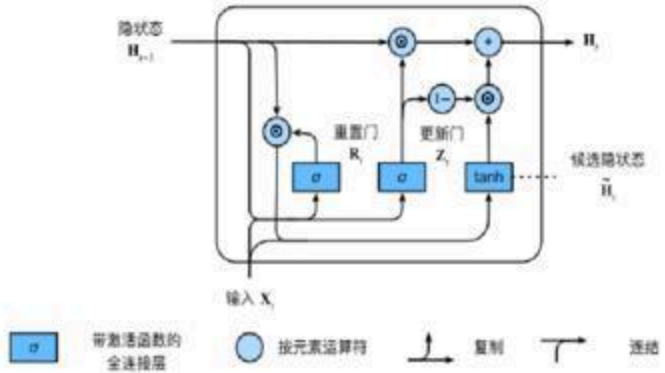


参数：$\mathbf{W}_x^r, \mathbf{W}_h^r, \mathbf{W}_x^z, \mathbf{W}_h^z, \mathbf{W}_x^h, \mathbf{W}_h^h, \mathbf{W}_h^y, b^r, b^z, b^h, \mathbf{b}^y$

LSTM：$\mathbf{W}_x^i, \mathbf{W}_h^i, \mathbf{W}_x^f, \mathbf{W}_h^f, \mathbf{W}_x^o, \mathbf{W}_h^o, \mathbf{W}_x^c, \mathbf{W}_h^c, \mathbf{W}_h^y, b^i, b^f, b^o, b^c, \mathbf{b}^y$

GRU：计算量

$$\mathbf{R}_t = \sigma(\mathbf{X}_t \mathbf{W}_x^r + \mathbf{H}_{t-1} \mathbf{W}_h^r + \mathbf{b}^r)$$
$$\mathbf{Z}_t = \sigma(\mathbf{X}_t \mathbf{W}_x^z + \mathbf{H}_{t-1} \mathbf{W}_h^z + \mathbf{b}^z)$$
$$\tilde{\mathbf{H}}_t = \tanh(\mathbf{X}_t \mathbf{W}_x^h + (\mathbf{R}_t \odot \mathbf{H}_{t-1}) \mathbf{W}_h^h +$$
$$\mathbf{H}_t = \mathbf{Z}_t \odot \mathbf{H}_{t-1} + (1 - \mathbf{Z}_t) \odot \tilde{\mathbf{H}}_t$$
$$\mathbf{Y}_t = \mathbf{H}_t \mathbf{W}_h^y + \mathbf{b}^y$$



参数：$\mathbf{W}_x^r, \mathbf{W}_h^r, \mathbf{W}_x^z, \mathbf{W}_h^z, \mathbf{W}_x^h, \mathbf{W}_h^h, \mathbf{W}_h^y, b^r, b^z, b^h, \mathbf{b}^y$

LSTM：$\mathbf{W}_x^i, \mathbf{W}_h^i, \mathbf{W}_x^f, \mathbf{W}_h^f, \mathbf{W}_x^o, \mathbf{W}_h^o, \mathbf{W}_x^c, \mathbf{W}_h^c, \mathbf{W}_h^y, b^i, b^f, b^o, b^c, \mathbf{b}^y$

参数少，但功能不少：更简单的效能更好

实验：GRU

小结：GRU

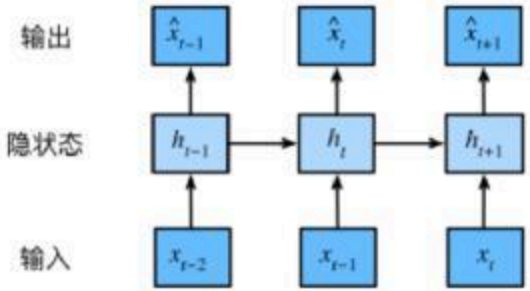
- GRU使用两种门控单元：重置门、更新门
  - 只有一个隐状态：记忆元合并到隐状态
- 缓解梯度消失问题；可以捕获长距离依赖关系

深度循环神经网络

回顾：隐变量模型的表达能力

当前隐状态$\mathbf{H}_t$整合过去所有的观测信息

$$\mathbf{H}_t = \sigma(\mathbf{X}_t \mathbf{W}_x^h + \mathbf{H}_{t-1} \mathbf{W}_h^h + \mathbf{b}^h)$$
$$\mathbf{O}_t = \mathbf{H}_t \mathbf{W}_h^o + \mathbf{b}^o$$

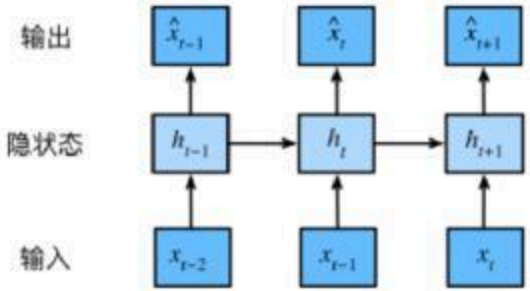


- 隐状态$\mathbf{H}_t$产生**信息瓶颈**：融合之前所有信息
  - 对当前时间步的观测信息$\mathbf{X}_t$**编码、描述不足**

回顾：隐变量模型的表达能力

当前隐状态$\mathbf{H}_t$**整合过去所有**的观测信息

$$\mathbf{H}_t = \sigma(\mathbf{X}_t \mathbf{W}_x^h + \mathbf{H}_{t-1} \mathbf{W}_h^h + \mathbf{b}^h)$$
$$\mathbf{O}_t = \mathbf{H}_t \mathbf{W}_h^o + \mathbf{b}^o$$



- 隐状态$\mathbf{H}_t$产生**信息瓶颈**：融合之前所有信息
  - 对当前时间步的观测信息$\mathbf{X}_t$**编码、描述不足**

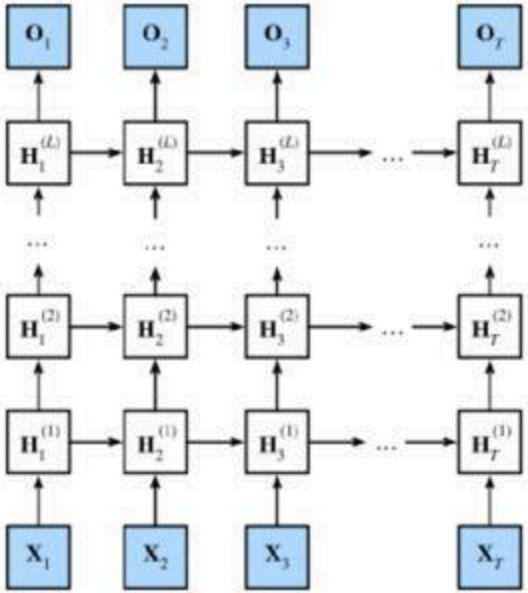
如何获得更多**非线性**、构造更复杂模型？思考感知机的设计（多层变换解决 XOR 问题）

DRNN：架构

深度循环神经网络：隐藏层的**堆叠**

隐状态的**两个信息流向**

- 当前层的下一时间步
  - 横列构成 HMM
- 下一层的当前时间步
  - 纵列构成感知机

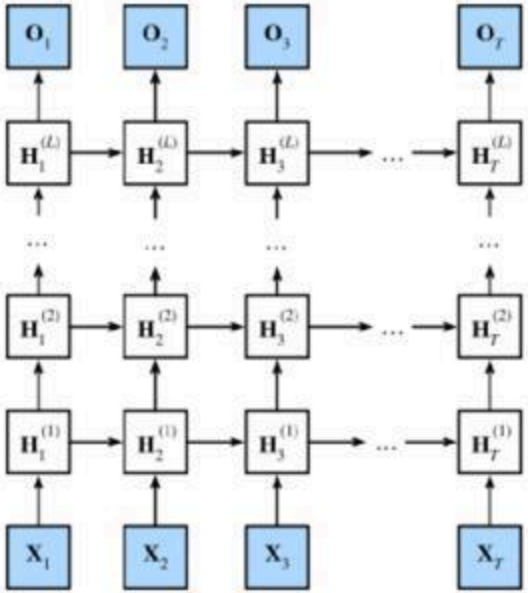


DRNN：计算

深度循环神经网络：隐藏层的**堆叠**

$$\mathbf{H}_t^{(l)} = \sigma(\mathbf{X}_t^{(l-1)}\mathbf{W}_x^{h^{(l)}} + \mathbf{H}_{t-1}^{(l)}\mathbf{W}_h^{h^{(l)}} + \mathbf{b}^{h^{(l)}})$$

$$\mathbf{O}_t = \mathbf{H}_t^{(L)}\mathbf{W}_h^o + \mathbf{b}^o$$



实验：DRNN

小结：DRNN

- DRNN：隐藏层横向、纵向分别传递信息
- 超参数：隐变量数、层数

双向循环神经网络

文本生成、完形填空

文本生成：预测下一个位置

他下课后去了 __ __ 。

文本生成、完形填空

文本生成：预测下一个位置

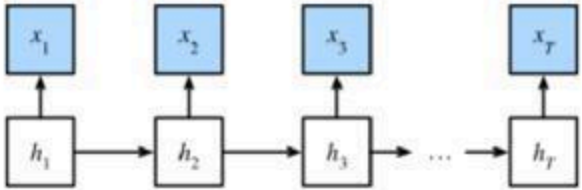
他下课后去了 __ __ 。

完形填空：后文对当前预测**反向传递**提示信息

她把 __ __
她把 __ __ 递给了老师。
她把 __ __ 递给了老师，然后飞走了。

隐Markov模型 HMM

一阶HMM的两个假设

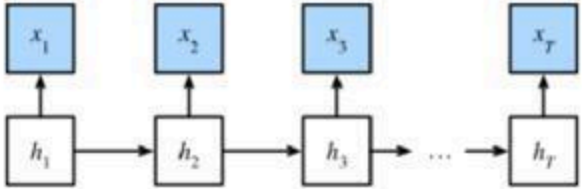


- 一阶Markov假设：状态概率只取决于**前一个**状态
  - 称为**状态转移概率**

$$P(h_t|h_1,..,h_{t-1}) \approx P(h_t|h_{t-1})$$

隐Markov模型 HMM

一阶HMM的两个假设



- 一阶Markov假设：状态概率只取决于**前一个**状态
  - 称为**状态转移概率**

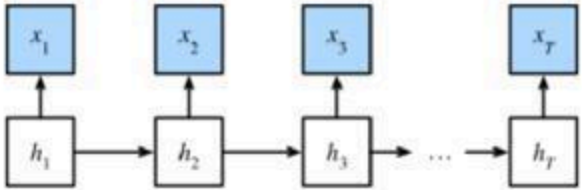
$$P(h_t|h_1,..,h_{t-1}) \approx P(h_t|h_{t-1})$$

- 独立性假设：输出变量的概率只取决于**直接关联**的隐变量
  - 称为**观测似然**，或“**发射概率**”

$$P(x_t|h_1,..,h_T,x_1,..,x_T) \approx P(x_t|h_t)$$

观测序列的概率

文本生成：预测下一个位置；等价于计算整个预测序列的概率



$$
\begin{aligned}
P(x_1, .., x_T) &= \sum_{h_1,..,h_T} P(x_1, .., x_T, h_1, .., h_T) \\
&= \sum_{h_1,..,h_T} \prod_t P(x_t|h_t) P(h_t|h_{t-1})
\end{aligned}
$$

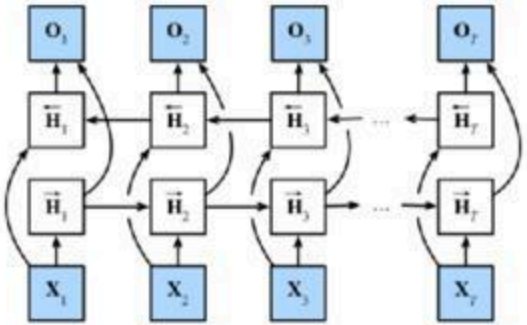
- Viterbi 算法：动态规划

双向RNN模型

完形填空：后文对当前预测**反向传递**提示信息

隐状态的两个信息流向

- 隐藏层：前向、后向分别传递
- 输出：整合两个方向的隐状态

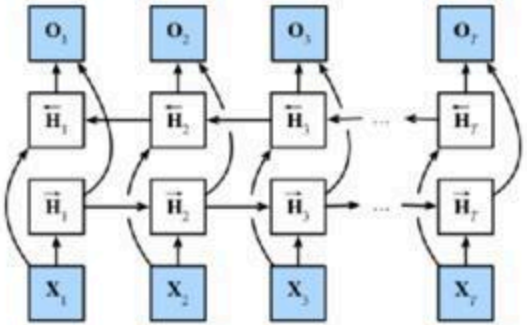


双向RNN模型：计算

完形填空：后文对当前预测**反向传递**提示信息

$$\mathbf{H}_t^{(f)} = \sigma(\mathbf{X}_t \mathbf{W}_x^{h^{(f)}} + \mathbf{H}_{t-1}^{(f)} \mathbf{W}_h^{h^{(f)}} + \mathbf{b}^{h^{(f)}})$$
$$\mathbf{H}_t^{(b)} = \sigma(\mathbf{X}_t \mathbf{W}_x^{h^{(b)}} + \mathbf{H}_{t-1}^{(b)} \mathbf{W}_h^{h^{(b)}} + \mathbf{b}^{h^{(b)}})$$
$$\mathbf{O}_t = \mathbf{H}_t \mathbf{W}_h^o + \mathbf{b}^o$$

- $\mathbf{H}_t$：$\mathbf{H}_t^{(f)}$、$\mathbf{H}_t^{(b)}$的拼接

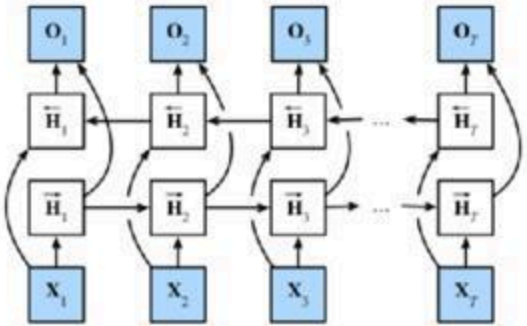


双向RNN模型：应用

不能用于“单边预测”，例如不能用未来信息预测现在

- 预测时无法获得对应数据

计算缓慢：两倍传递链

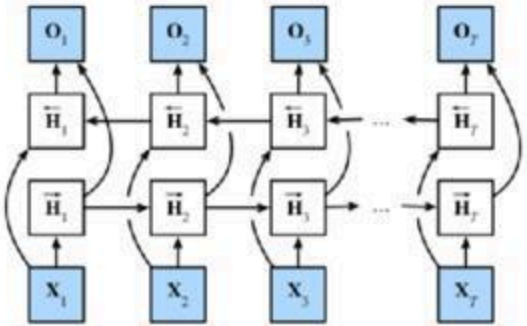


双向RNN模型：应用

不能用于“单边预测”，例如不能用未来信息预测现在

- 预测时无法获得对应数据

计算缓慢：两倍传递链



常用场合：填充缺失、词元标注、机器翻译

实验：BiRNN

小结：BiRNN

- BiRNN：隐藏层前向、后向分别传递时序信息
- 应用：特征提取、完形填空、词元标注、机器翻译
  - 不能用于“单边预测”

Review

本章内容

长短期记忆网络（LSTM）。门控循环单元（GRU）。深度循环神经网络。双向循环神经网络。

重点：LSTM 的设计特点、网络架构；GRU 的设计特点、网络架构；深度循环神经网络的信息流向、网络架构；双向循环神经网络的信息流向、网络架构。

难点：现代循环神经网络的实现。

学习目标

- 理解 LSTM 的设计特点（门控，记忆元）、网络架构
- 理解 GRU 设计特点（简化门控，合并记忆元到隐状态）、网络架构
- 理解深度循环神经网络的信息流向、网络架构。
- 理解双向循环神经网络的信息流向、网络架构。

问题

简述 LSTM 的设计特点、网络架构。

简述 GRU 的设计特点、网络架构。

简述深度循环神经网络的信息流向、网络架构。

简述双向循环神经网络的信息流向、网络架构。