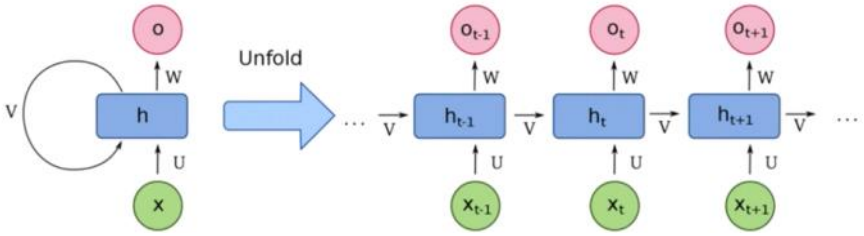
Recurrent Neural Network

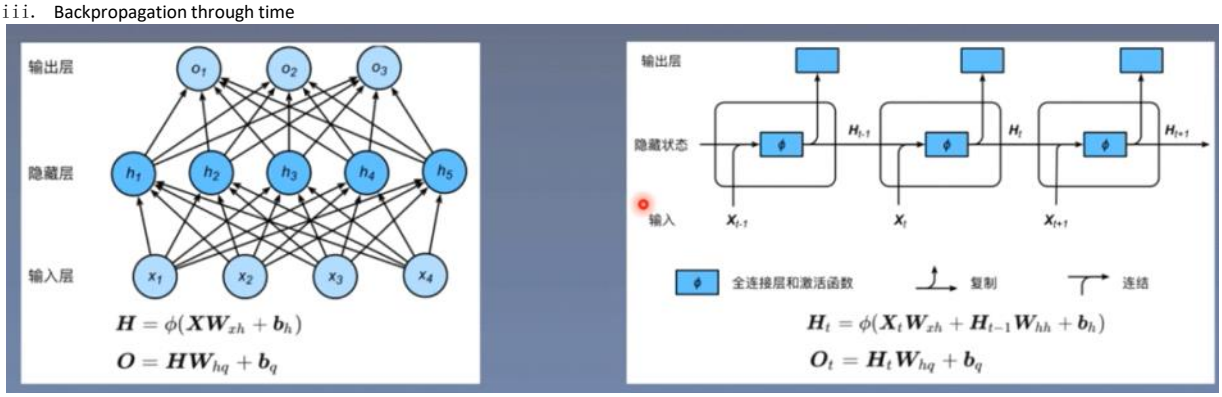
2023年8月17日 15:36

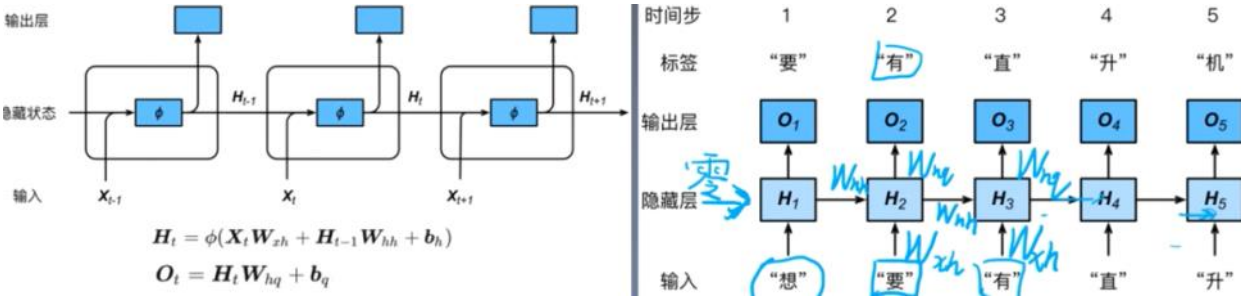
1. Sequence Data:序列数据是常见的数据类型，前后数据通常具有关联性。
   1) Speech recognition
   2) Music generation
   3) Sentiment classification: movie rate
   4) DNA sequence analysis
   5) Machine translation
   6) Video activity recognition
   7) Name entity recognition
2. Language model(NLP) :NLP中常把文本看为离散时间序列，一段长度为T的文本的词依次为$w_1, w_2, \ldots, w_T$其中w_t（$1 \le t \le T$)是时间步（Time Step）t的输出或标签$P(w_1, w_2, \ldots, w_T) = \prod_{t=1}^{T} P(w_t | w_1, \ldots, w_{t-1})$

   缺点：时间步t的词需要考虑t-1步的词，其计算量随t呈指数增长
3. Recurrent Neural Network
   - RNN是针对序列数据而生的神经网络结构，核心在于循环使用网络层参数，避免时间步增大带来的参数激增，并引入hidden state用于记录历史信息，有效的处理数据的前后关联性。



   - Hidden States:用于记录历史信息，有效处理数据的前后关联性，激活函数采用tanh， 将输出值域控制在(-1,1)，防止数值呈指数级变化。
   - 特点：
     i. 循环神经网络的隐藏状态可以捕捉截至当前时间步的序列的历史信息
     ii. 循环神经网络模型参数的数量不随时间步的增加而增长
     iii. Backpropagation through time





   - $\frac{\partial L}{\partial W_{qh}} = \sum_{t=1}^{T} prod\left(\frac{\partial L}{\partial o_t}, \frac{\partial o_t}{\partial W_{qh}}\right) = \sum_{t=1}^{T} \frac{\partial L}{\partial o_t} h_t^T$ T=3

$$\frac{\partial L}{\partial h_T} = prod\left(\frac{\partial L}{\partial o_T}, \frac{\partial o_T}{\partial h_T}\right) = W_{qh}^T \frac{\partial L}{\partial o_T}$$

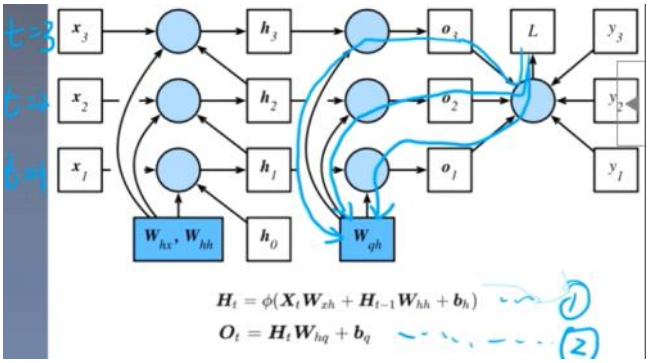
$$\frac{\partial L}{\partial h_t} = prod\left(\frac{\partial L}{\partial h_{t+1}}, \frac{\partial h_{t+1}}{\partial h_t}\right) + prod\left(\frac{\partial L}{\partial o_t}, \frac{\partial o_t}{\partial h_t}\right) = W_{hh}^T \frac{\partial L}{\partial h_{t+1}} + W_{qh}^T \frac{\partial L}{\partial o_t}$$

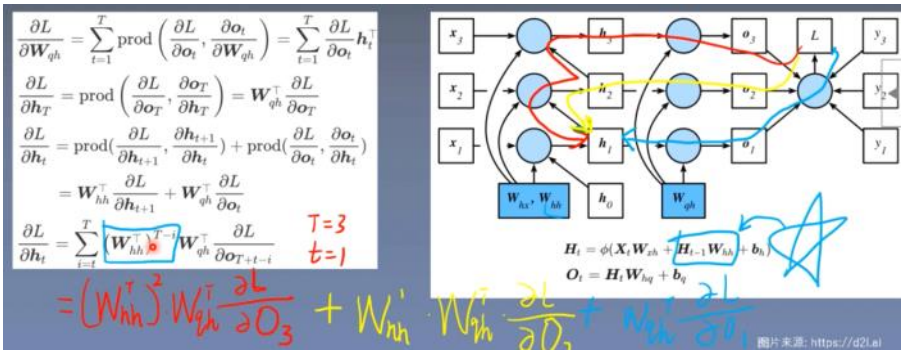
$$\frac{\partial L}{\partial h_t} = \sum_{i=t}^{T} (W_{hh}^T)^{T-i} W_{qh}^T \frac{\partial L}{\partial o_{T+t-i}}$$

$$\frac{\partial L}{\partial W_{hx}} = \sum_{t=1}^{T} prod\left(\frac{\partial L}{\partial h_t}, \frac{\partial h_t}{\partial W_{hx}}\right) = \sum_{t=1}^{T} \frac{\partial L}{\partial h_t} x_t^T$$

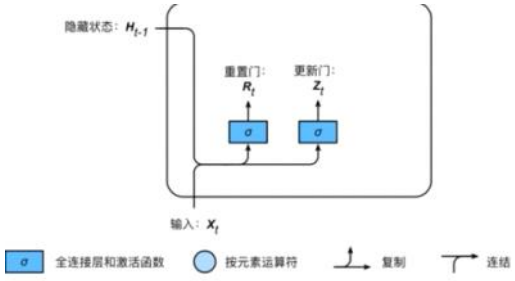
$$\frac{\partial L}{\partial W_{hh}} = \sum_{t=1}^{T} prod\left(\frac{\partial L}{\partial h_t}, \frac{\partial h_t}{\partial W_{hh}}\right) = \sum_{t=1}^{T} \frac{\partial L}{\partial h_t} h_{t-1}^T$$

缺点：梯度随时间t呈指数变化，易引发梯度消失或梯度爆炸。 $W_{hh} < 1 \to 0$ $W_{hh} > 1 \to \infty$



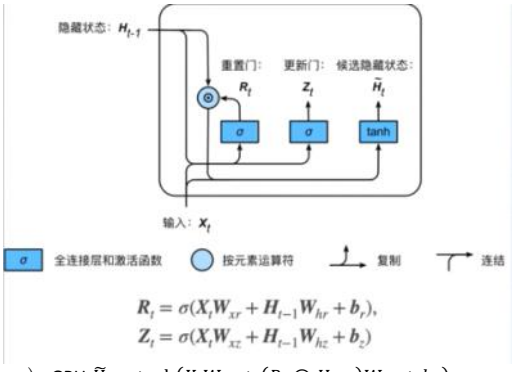


4. Gated recurrent unit(门控循环单元 GRU)：缓解RNN梯度消失带来的问题，引入门概念，来控制信息流动，使得模型更好的记住长远时期的信息，并缓解梯度消失
   - 重置门：哪些信息需要遗忘，用于候选隐藏状态计算过程中用于控制上一时间步隐藏状态中要遗忘哪一些信息
   - 更新门：哪些信息需要注意，用于更新当前时间步隐藏状态
   - 激活函数为：sigmoid，值域（0，1）,0表示遗忘，1表示保留



$$R_t = \sigma(X_t W_{xr} + H_{t-1} W_{hr} + b_r),$$
$$Z_t = \sigma(X_t W_{xz} + H_{t-1} W_{hz} + b_z)$$

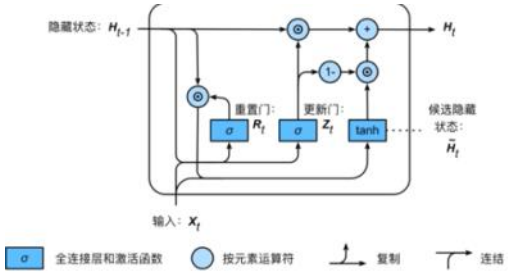
1) 候选隐藏状态：输入与上一时间步隐藏状态共同计算得到候选隐藏状态，用于隐藏状态计算。通过重置门，对上一时间步隐藏状态进行选择性遗忘，对历史信息更好的选择。



$$R_t = \sigma(X_t W_{xr} + H_{t-1} W_{hr} + b_r),$$
$$Z_t = \sigma(X_t W_{xz} + H_{t-1} W_{hz} + b_z)$$

   a) GRU:$\tilde{H}_t = \tanh\left(X_t W_{xh} + \left(R_t \odot H_{t-1}\right) W_{hh} + b_h\right)$
   b) RNN：$H_t = \phi\left(X_t W_{xh} + H_{t-1} W_{hh} + b_h\right)$

2) 隐藏状态：由候选隐藏状态及上一时间步隐藏状态组合的得来

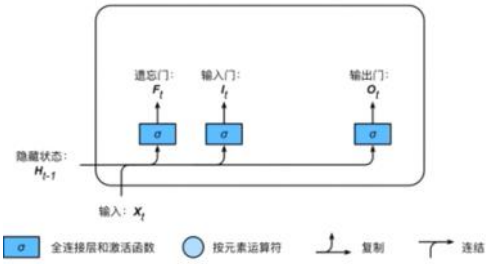


$$R_t = \sigma(X_t W_{xr} + H_{t-1} W_{hr} + b_r),$$
$$Z_t = \sigma(X_t W_{xz} + H_{t-1} W_{hz} + b_z)$$

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

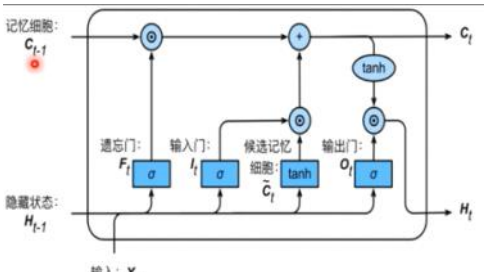
- GRU特点：若更新门从第一个时间步到t-1时间过程中，一直保持1，信息可有效传递到当前时间步。

5. Long short-term memory(LSTM):引入三个门和记忆细胞，控制信息传递



$$I_t = \sigma(X_t W_{xi} + H_{t-1} W_{hi} + b_i),$$
$$F_t = \sigma(X_t W_{xf} + H_{t-1} W_{hf} + b_f),$$
$$O_t = \sigma(X_t W_{xo} + H_{t-1} W_{ho} + b_o),$$

1) 遗忘门：哪些信息需要被遗忘
2) 输入门：哪些信息需要流入当前记忆细胞
3) 输出门：哪些记忆信息流入隐藏状态
4) 记忆细胞：特殊的隐藏状态，记忆历史信息
   i. 激活函数可变 $\tilde{C}_t = \tanh(X_t W_{xc} + H_{t-1} W_{hc} + b_c)$
   ii. 记忆细胞由候选记忆细胞以及上一时间步记忆细胞组合得来：$C_t = F_t \odot C_{t-1} + I_t \odot \tilde{C}_t$
   由输出门控制记忆细胞信息流入隐藏状态：$H_t = O_t \odot \tanh(C_t)$



$$I_t = \sigma(X_t W_{xi} + H_{t-1} W_{hi} + b_i),$$
$$F_t = \sigma(X_t W_{xf} + H_{t-1} W_{hf} + b_f),$$
$$O_t = \sigma(X_t W_{xo} + H_{t-1} W_{ho} + b_o),$$