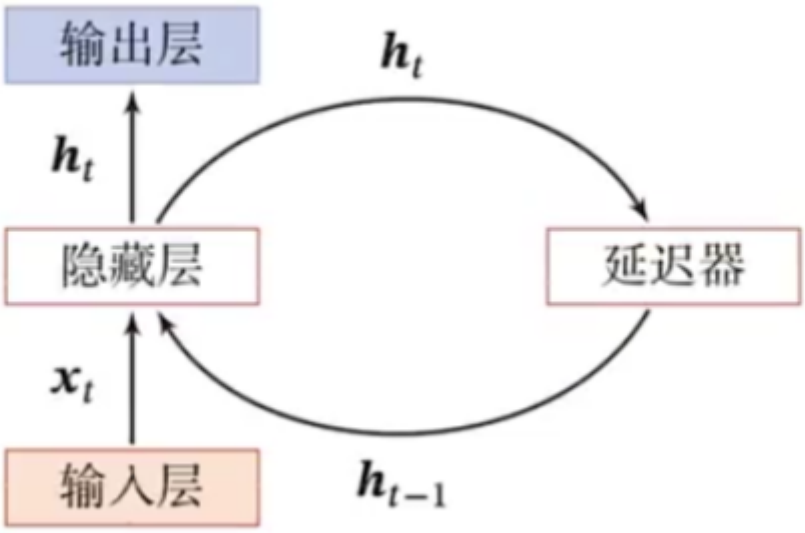
循环神经网络

> 循环神经网络具有记忆性、参数共享(状态转移矩阵U,状态输入矩阵W,偏置项b)、图灵完备

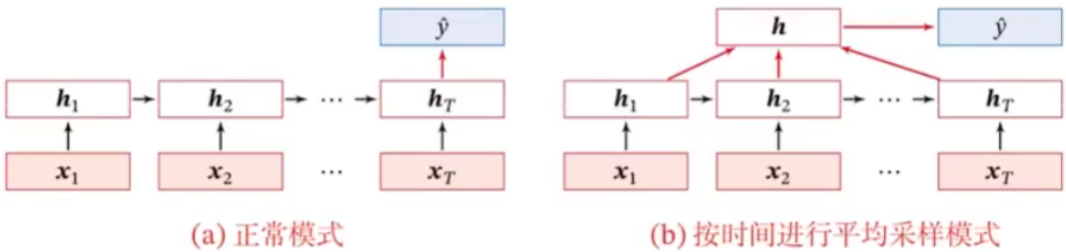
循环神经网络通过使用带自反馈的神经元,能够处理任意长度的时序数据.

$$h_t = f(h_{t-1}, x_t) \tag{1}$$



序列到类别

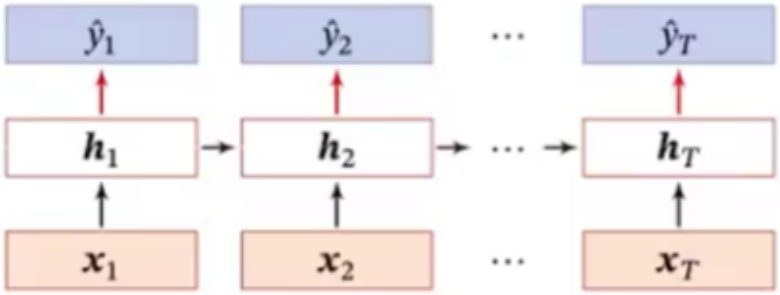
指输入是个序列,输出是个类别



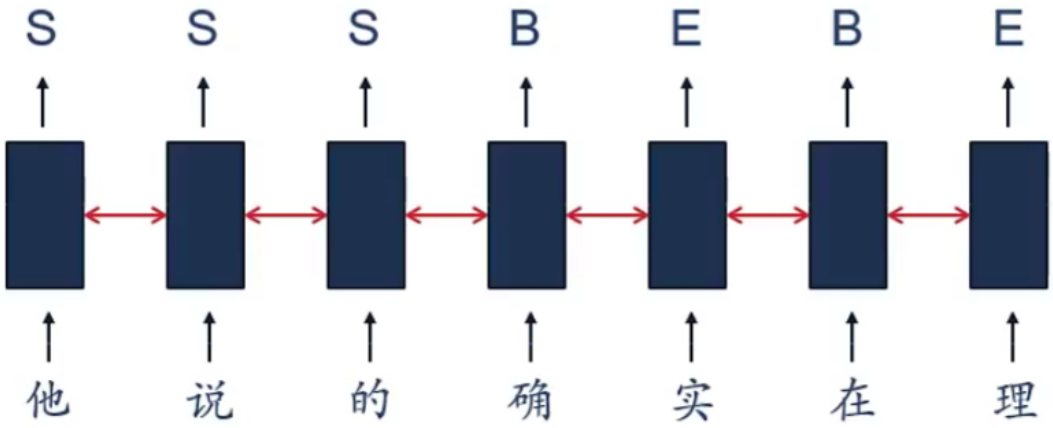
Case1:情感分类

同步的序列到序列

输入的是一个序列,输出的也是一个序列,并且是一一对应的.



Case1: 中文分词

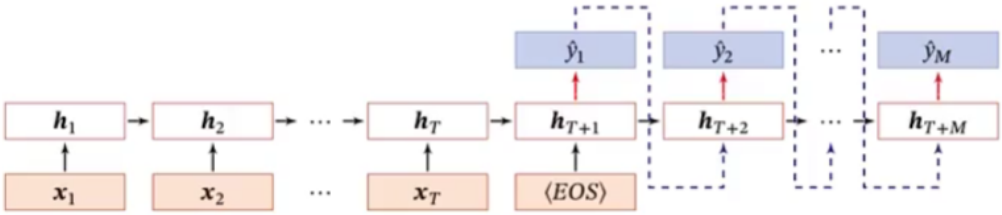


图中下面为输入(需要进行word2vec操作),图的上面为类别,其中S表示单个字为词,B表示词的开始,E表示词的结束.

Case2:信息抽取(就是从文字中提取重要信息,即实体识别)

Case3:语音识别

异步的序列到序列



Case:机器翻译

参数学习

以同步的序列到序列为例

- 给定一个训练样本(x, y),其中
  - 长度为T的输入序列 $x = x_1, x_2, \cdots, x_T$
  - 长度为T的输出序列 $y = y_1, y_2, \cdots, y_T$

- 时刻t的瞬间损失函数为

$$\mathcal{L}_t = \mathcal{L}(y_t, g(h_t)) \tag{2}$$

- 总损失为

$$\mathcal{L} = \sum_{i=1}^{T} \mathcal{L}_t \tag{3}$$

计算梯度

把原来的式子写成两个式子,方便对照前馈神经网络

$$\begin{aligned} z_t &= Uh_{t-1} + Wx_t + b \\ h_t &= f(z_t) \end{aligned} \tag{4}$$

梯度:

$$\frac{\partial \mathcal{L}}{} = \sum^{T} \frac{\partial \mathcal{L}_t}{} \tag{5}$$

$$\frac{\partial \mathcal{L}}{\partial U} = \sum_{t=1} \frac{\partial \mathcal{L}_t}{\partial U} \tag{5}$$

其中, 这里的U表示的是总体,如果对应一个位置的话

$$\frac{\partial \mathcal{L}_t}{\partial U} = \sum_{k=1}^{t} \frac{\partial \mathcal{L}_t}{\partial z_k} h_{k-1}^T \tag{6}$$

对于中间的$\frac{\partial \mathcal{L}_t}{\partial z_k}$有:

$$\begin{aligned} 令\delta_{t,k} &= \frac{\partial \mathcal{L}_t}{\partial z_k} \\ &= \frac{\partial h_k}{\partial z_k} \frac{\partial z_{k+1}}{\partial h_k} \frac{\partial \mathcal{L}_t}{\partial z_{k+1}} \\ &= diag(f'(z_k)) U^T \delta_{t,k+1} \end{aligned} \tag{7}$$

上式就是BPTT(BackPropagation Through Time)算法.

> Note: 长程依赖问题
>
> 由于梯度爆炸或梯度消失问题,实际上只能学到短周期的依赖关系

由上面的式子可以得到:

$$\delta_{t,k} = \prod_{\tau=k}^{t-1} (diag(f'(z_\tau)) U^T) \delta_{t,t} \tag{8}$$

于是有:

$$\delta_{t,k} \cong \gamma^{t-k} \delta_{t,t} = \begin{cases} \infty & if \quad \gamma \leq 1 \\ 0 & if \quad \gamma \geq 1 \end{cases} \tag{9}$$

即出现长程依赖问题.

如何解决长程依赖问题?

循环神经网络在时间维度上非常深!

- 梯度消失或梯度爆炸

如何改进?

- 梯度爆炸问题
    - 权重衰减
    - 梯度截断

- 梯度消失问题
    - 改进模型

门控机制

控制信息的累计速度,包括有选择地加入新的信息,并有选择地遗忘之前累计的信息.

- 门控循环单元(Gate Recurrent Unit,GRU)
- 长短期记忆网络(Long Short-Term Memory,LSTM)

GRU

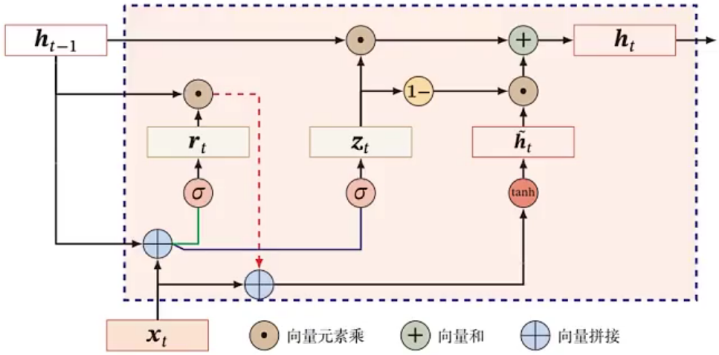
$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot g(x, h_{t-1}; \theta) \tag{10}$$

其中$z_t \in (0,1)^d$

$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$称为更新门,因为logistics函数限制在(0,1)

$$g(x, h_{t-1}; \theta) = \tanh(W_h x_t + U_h r_t(h_{t-1}) + b_h) \tag{11}$$

其中r_t称为重置门,当$z_t = 0$且$r_t = 0$,则输出与前一项无任何关系.



更新门

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

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$

重置门

$$\tilde{h}_t = \tanh\left(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h\right)$$

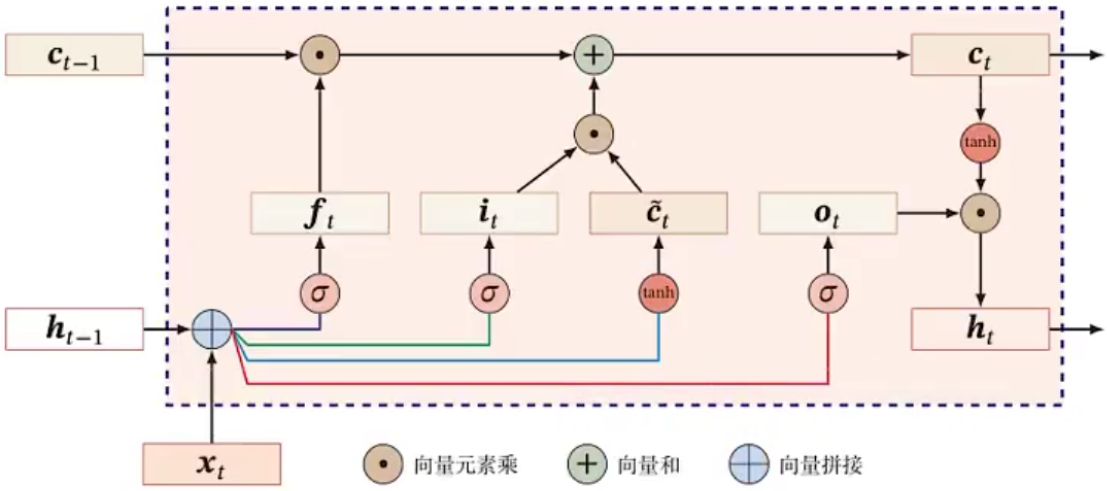
$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot g(x_t, h_{t-1}; \theta)$$

LSTM

$$\begin{aligned} \widetilde{c}_t &= \tanh(W_c x_t + U_c h_{t-1} + b_c) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \widetilde{c}_t \\ h_t &= o_t \odot \tanh(c_t) \end{aligned} \tag{12}$$

其中i_t叫做记忆门,f_t叫做遗忘门,o_t叫做输出门

$$\begin{aligned} i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\ f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\ o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \end{aligned} \tag{13}$$



LSTM的各种变体形式

没有遗忘门

$$c_t = c_{t-1} + i_t \odot \widetilde{c}_t \tag{14}$$

耦合输入门和遗忘门

$$\begin{aligned} f_t &= 1 - i_t \\ c_t &= (1 - i_t) \odot c_{t-1} + i_t \odot \widetilde{c}_t \end{aligned} \tag{15}$$

peephole连接

$$\begin{aligned} i_t &= \sigma(W_i x_i + U_i h_{t-1} + V_i c_{t-1} + b_i) \\ f_t &= \sigma(W_f x_i + U_f h_{t-1} + V_f c_{t-1} + b_f) \\ o_t &= \sigma(W_o x_t + U_o h_{t-1} + V_o c_t + b_o) \end{aligned} \tag{16}$$

深层循环神经网络

1. 堆叠循环神经网络(就是num_layers不为1)
2. 双向循环神经网络

小结

优点	缺点
引入(短期)记忆	长程依赖问题
图灵完备	记忆容量问题
	并行能力