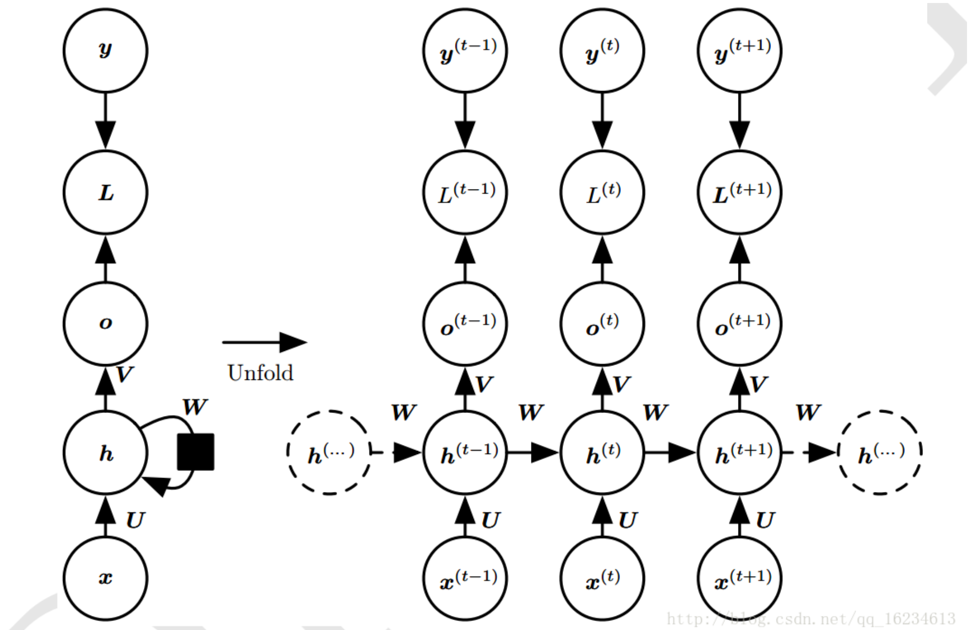
RNN-LSTM-GRU 模型

1. RNN 模型

循环神经网络又叫做 RNN(Recurrent Neural Networks)，主要用来处理连续序列的样本。下图是一个标准的 RNN 网络：



在 RNN 网络中，首先介绍下各个符号的含义：t 表示时间序列；$\boldsymbol{x}$ 代表输入，是一个 n 行一列的向量；$\boldsymbol{U}$ 代表 $\boldsymbol{x}$ 的权重，是一个 m 行 n 列的矩阵；$\boldsymbol{h}$ 代表隐藏状态，是一个 m 行一列的向量；$\boldsymbol{W}$ 和 $\boldsymbol{V}$ 代表权重，分别是一个 m 行 m 列的矩阵和一个 k 行 m 列的矩阵；$\boldsymbol{o}$ 是一个 k 行一列的向量；$\hat{\boldsymbol{y}}^t$ 和 $\boldsymbol{y}^t$ 代表 t 时刻预测输出和真实输出，都是一个 k 行一列的向量；$\boldsymbol{b}$ 和 $\boldsymbol{c}$ 分别是 m 行一列和 k 行一列的向量；L^t 代表 t 时刻的损失值（标量），是一个交叉熵损失函数。

有了上面模型符号的介绍，下面就来介绍 RNN 的前向传播过程。对于任意的隐藏状态 $\boldsymbol{h}^t$ 是由 $\boldsymbol{x}^t$ 和 $\boldsymbol{h}^{t-1}$ 决定，即：

$$\boldsymbol{h}^t = f\left(\boldsymbol{z}^t\right) = f\left(\boldsymbol{U}\boldsymbol{x}^t + \boldsymbol{W}\boldsymbol{h}^{t-1} + \boldsymbol{b}\right)$$

上式中，f 代表 $tanh$ 函数。$\boldsymbol{o}^t$ 是 $\boldsymbol{h}^t$ 由决定，即：

$$\boldsymbol{o}^t = \boldsymbol{V}\boldsymbol{h}^t + \boldsymbol{c}$$

因此，预测输出 $\hat{\boldsymbol{y}}^t$ 为：

$$\hat{\boldsymbol{y}}^t = f\left(\boldsymbol{o}^t\right)$$

上式中，f 代表 $softmax$ 函数。由于 L^t 代表 t 时刻的损失函数，因此最终的损失函数为：

$$L = \sum_{t=1}^{\tau} L^t$$

上式中，τ 代表最后一个时刻。

在 RNN 模型中，模型参数 $\boldsymbol{U}$、$\boldsymbol{V}$、$\boldsymbol{W}$、$\boldsymbol{b}$ 和 $\boldsymbol{c}$ 是共享的，因此可以根据 RNN 的反向传

播算法 BPTT(back-propagation through time)计算各个模型参数的值。首先，计算 L 对 $\boldsymbol{V}$ 和 $\boldsymbol{c}$ 的偏导，即：

$$\frac{\partial L}{\partial \boldsymbol{c}}=\sum_{t=1}^{\tau}\frac{\partial L^t}{\partial \boldsymbol{c}}=\sum_{t=1}^{\tau}\frac{\partial L^t}{\partial \boldsymbol{o}^t}\frac{\partial \boldsymbol{o}^t}{\partial \boldsymbol{c}}=\sum_{t=1}^{\tau}\hat{\boldsymbol{y}}^t-\boldsymbol{y}^t$$
$$\frac{\partial L}{\partial \boldsymbol{V}}=\sum_{t=1}^{\tau}\frac{\partial L^t}{\partial \boldsymbol{V}}=\sum_{t=1}^{\tau}\frac{\partial L^t}{\partial \boldsymbol{o}^t}\frac{\partial \boldsymbol{o}^t}{\partial \boldsymbol{V}}=\sum_{t=1}^{\tau}\left(\hat{\boldsymbol{y}}^t-\boldsymbol{y}^t\right)\left(\boldsymbol{h}^t\right)^T$$

然后计算 $\boldsymbol{W}$、$\boldsymbol{U}$ 和 $\boldsymbol{b}$ 的梯度，定义序列索引 t 位置的隐藏状态的梯度为：

$$\boldsymbol{\delta}^t=\frac{\partial L}{\partial \boldsymbol{h}^t}$$

上式中，$\boldsymbol{\delta}^t$ 是一个 m 行一列的向量。当 t 不等于 τ 时，

$$\begin{aligned}\boldsymbol{\delta}^t&=\frac{\partial L}{\partial \boldsymbol{o}^t}\frac{\partial \boldsymbol{o}^t}{\partial \boldsymbol{h}^t}+\frac{\partial L}{\partial \boldsymbol{h}^{t+1}}\frac{\partial \boldsymbol{h}^{t+1}}{\partial \boldsymbol{h}^t}\\&=\boldsymbol{V}^T\left(\hat{\boldsymbol{y}}^t-\boldsymbol{y}^t\right)+\frac{\partial\left(\boldsymbol{\delta}^{t+1}\right)^T\boldsymbol{h}^{t+1}}{\partial \boldsymbol{h}^t}\\&=\boldsymbol{V}^T\left(\hat{\boldsymbol{y}}^t-\boldsymbol{y}^t\right)+\boldsymbol{W}^T\left(\boldsymbol{\delta}^{t+1}\odot f'\left(\boldsymbol{z}^t\right)\right)\\&=\boldsymbol{V}^T\left(\hat{\boldsymbol{y}}^t-\boldsymbol{y}^t\right)+\boldsymbol{W}^T diag\left(\delta_1^{t+1},\delta_2^{t+1},\ldots,\delta_m^{t+1}\right)f'\left(\boldsymbol{z}^t\right)\\&=\boldsymbol{V}^T\left(\hat{\boldsymbol{y}}^t-\boldsymbol{y}^t\right)+\boldsymbol{W}^T diag\left(1-\left(h_1^{t+1}\right)^2,1-\left(h_2^{t+1}\right)^2,\ldots,1-\left(h_m^{t+1}\right)^2\right)\boldsymbol{\delta}^{t+1}\end{aligned}$$

上式中，$diag$(*)表示对角矩阵，δ_i^{t+1} (i=1,2,..,m)和 h_i^{t+1} (i=1,2,..,m)分别是 $\boldsymbol{\delta}^{t+1}$ 和 $\boldsymbol{h}^{t+1}$ 中的每个元素。当 t 等于 τ 时，

$$\boldsymbol{\delta}^\tau=\frac{\partial L}{\partial \boldsymbol{o}^\tau}\frac{\partial \boldsymbol{o}^\tau}{\partial \boldsymbol{h}^\tau}=\boldsymbol{V}^T\left(\hat{\boldsymbol{y}}^\tau-\boldsymbol{y}^\tau\right)$$

因此，$\boldsymbol{W}$、$\boldsymbol{U}$ 和 $\boldsymbol{b}$ 的梯度计算表达式为：

$$\frac{\partial L}{\partial \boldsymbol{W}}=\sum_{t=1}^{\tau}\frac{\partial L}{\partial \boldsymbol{h}^t}\frac{\partial \boldsymbol{h}^t}{\partial \boldsymbol{W}}=\sum_{t=1}^{\tau}diag\left(1-\left(h_1^{t+1}\right)^2,1-\left(h_2^{t+1}\right)^2,\ldots,1-\left(h_m^{t+1}\right)^2\right)\boldsymbol{\delta}^{t+1}\left(\boldsymbol{h}^{t-1}\right)^T$$
$$\frac{\partial L}{\partial \boldsymbol{b}}=\sum_{t=1}^{\tau}\frac{\partial L}{\partial \boldsymbol{h}^t}\frac{\partial \boldsymbol{h}^t}{\partial \boldsymbol{b}}=\sum_{t=1}^{\tau}diag\left(1-\left(h_1^{t+1}\right)^2,1-\left(h_2^{t+1}\right)^2,\ldots,1-\left(h_m^{t+1}\right)^2\right)\boldsymbol{\delta}^{t+1}$$
$$\frac{\partial L}{\partial \boldsymbol{U}}=\sum_{t=1}^{\tau}\frac{\partial L}{\partial \boldsymbol{h}^t}\frac{\partial \boldsymbol{h}^t}{\partial \boldsymbol{U}}=\sum_{t=1}^{\tau}diag\left(1-\left(h_1^{t+1}\right)^2,1-\left(h_2^{t+1}\right)^2,\ldots,1-\left(h_m^{t+1}\right)^2\right)\boldsymbol{\delta}^{t+1}\left(\boldsymbol{x}^t\right)^T$$

最后，根据如下公式更新 RNN 模型中的参数：

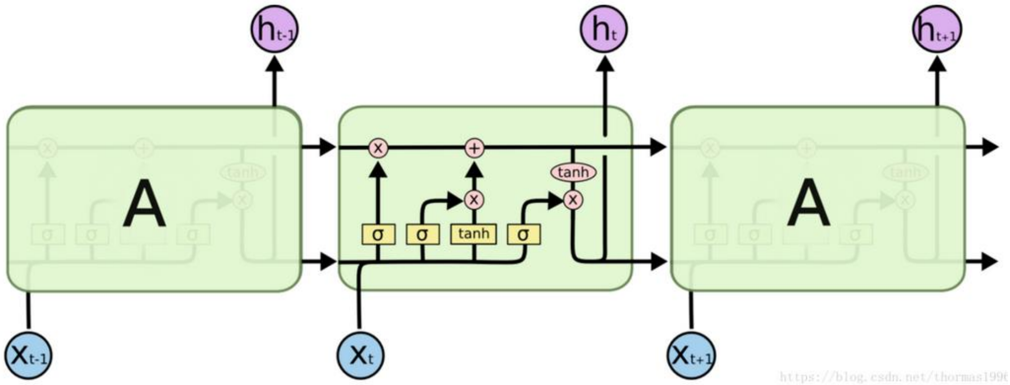
$$\theta_{update}=\theta-\alpha\frac{\partial L}{\partial \theta}$$

上式中，θ 代表的是模型参数 $\boldsymbol{U}$、$\boldsymbol{V}$、$\boldsymbol{W}$、$\boldsymbol{b}$ 和 $\boldsymbol{c}$，α 代表学习率，θ_{update} 代表更新完以后的参数。

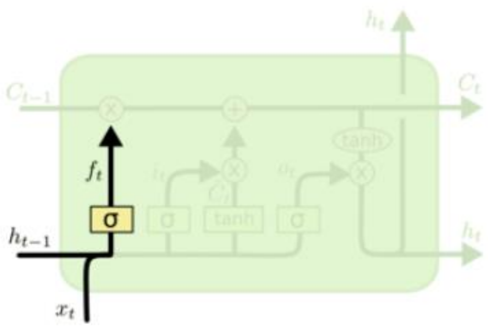
但是在 RNN 模型中会出现梯度爆炸和梯度消失的问题。出现梯度消失，是因为更新模型参数 $\boldsymbol{W}$、$\boldsymbol{U}$ 和 $\boldsymbol{b}$ 时，都与之前的状态有关，而 *tanh* 函数的导数在 0 到 1 之间，经过多次迭代相乘，极容易出现梯度消失。出现梯度爆炸，是因为更新模型参数 $\boldsymbol{W}$、$\boldsymbol{U}$ 和 $\boldsymbol{b}$ 时，也都与之前的状态有关，当 *tanh* 函数的导数不是特别小时同时模型参数 $\boldsymbol{W}$ 又很大，经过多次迭代相乘，就会出现梯度爆炸，这也是梯度爆炸出现概率比较小的原因。

2. LSTM 模型

LSTM(Long Short-Term Memory)又叫长短期记忆，是 RNN 模型的一种改进，主要为了解决 RNN 模型的梯度消失问题。下面是 LSTM 的结构图：



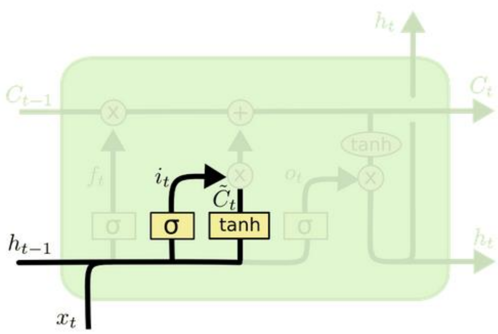
上图中，σ 表示的是 sigmoid 函数。下面深度分析 LSTM 模型，首先是遗忘门，其结构如下图：



上图中用数学表达式为：

$$\boldsymbol{f}^t = \sigma\left(\boldsymbol{W}_f \boldsymbol{h}^{t-1} + \boldsymbol{U}_f \boldsymbol{x}^t + \boldsymbol{b}_f\right)$$

上式中，$\boldsymbol{f}^t$ 是 m 行 1 列，$\boldsymbol{W}_f$ 是 m 行 m 列，$\boldsymbol{h}^{t-1}$ 是 m 行 1 列，$\boldsymbol{U}_f$ 是 m 行 n 列，$\boldsymbol{x}^t$ 是 n 行 1 列，$\boldsymbol{b}_f$ 是 m 行 1 列。其次是输入门，其结构如下图：



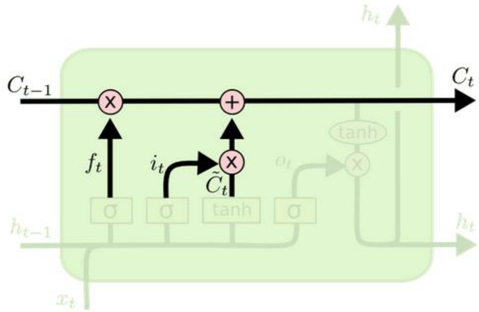
上图中用数学表达式为：

$$\boldsymbol{i}^t = \sigma\left(\boldsymbol{W}_i \boldsymbol{h}^{t-1} + \boldsymbol{U}_i \boldsymbol{x}^t + \boldsymbol{b}_i\right)$$

$$\boldsymbol{a}^t = tanh\left(\boldsymbol{W}_a \boldsymbol{h}^{t-1} + \boldsymbol{U}_a \boldsymbol{x}^t + \boldsymbol{b}_a\right)$$

上式中，$\boldsymbol{i}^t$ 是 m 行 1 列，$\boldsymbol{W}_i$ 是 m 行 m 列，$\boldsymbol{U}_i$ 是 m 行 n 列，$\boldsymbol{b}_i$ 是 m 行 1 列，$\tilde{\boldsymbol{C}}_t = \boldsymbol{a}^t$，$\boldsymbol{a}^t$ 是

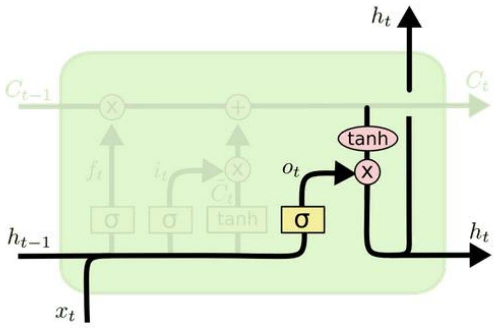
m 行 1 列，$\boldsymbol{W}_a$ 是 m 行 m 列，$\boldsymbol{U}_a$ 是 m 行 n 列，$\boldsymbol{b}_a$ 是 m 行 1 列。其次是状态更新，其结构如下图：



上图中用数学表达式为：

$$\boldsymbol{C}^t = \boldsymbol{C}^{t-1} \odot \boldsymbol{f}^t + \boldsymbol{i}^t \odot \boldsymbol{a}^t$$

最后是输出门，其结构如下图：



上图中用数学表达式为：

$$\boldsymbol{o}^t = \sigma\left(\boldsymbol{W}_o \boldsymbol{h}^{t-1} + \boldsymbol{U}_o \boldsymbol{x}^t + \boldsymbol{b}_o\right)$$
$$\boldsymbol{h}^t = \boldsymbol{o}^t \odot tanh\left(\boldsymbol{C}^t\right)$$

上式中，$\boldsymbol{o}^t$ 是 m 行 1 列，$\boldsymbol{W}_o$ 是 m 行 m 列，$\boldsymbol{U}_o$ 是 m 行 n 列，$\boldsymbol{b}_o$ 是 m 行 1 列。因此，LSTM 模型的前向传播为：

(1).遗忘门输出：

$$\boldsymbol{f}^t = \sigma\left(\boldsymbol{W}_f \boldsymbol{h}^{t-1} + \boldsymbol{U}_f \boldsymbol{x}^t + \boldsymbol{b}_f\right)$$

(2).输入门输出：

$$\boldsymbol{i}^t = \sigma\left(\boldsymbol{W}_i \boldsymbol{h}^{t-1} + \boldsymbol{U}_i \boldsymbol{x}^t + \boldsymbol{b}_i\right)$$
$$\boldsymbol{a}^t = tanh\left(\boldsymbol{W}_a \boldsymbol{h}^{t-1} + \boldsymbol{U}_a \boldsymbol{x}^t + \boldsymbol{b}_a\right)$$

(3).状态更新：

$$\boldsymbol{C}^t = \boldsymbol{C}^{t-1} \odot \boldsymbol{f}^t + \boldsymbol{i}^t \odot \boldsymbol{a}^t$$

(4).输出门输出：

$$\boldsymbol{o}^t = \sigma\left(\boldsymbol{W}_o \boldsymbol{h}^{t-1} + \boldsymbol{U}_o \boldsymbol{x}^t + \boldsymbol{b}_o\right)$$
$$\boldsymbol{h}^t = \boldsymbol{o}^t \odot tanh\left(\boldsymbol{C}^t\right)$$

(5).预测输出：

$$\boldsymbol{d}^t = \boldsymbol{V}\boldsymbol{h}^t + \boldsymbol{c}$$

$$\hat{\boldsymbol{y}}^t = softmax\left(\boldsymbol{d}^t\right)$$

上式中，$\boldsymbol{b}^t$ 是 k 行 1 列，相当于 RNN 模型中的 $\boldsymbol{o}^t$，$\hat{\boldsymbol{y}}^t$ 是 k 行 1 列，$\boldsymbol{V}$ 是 k 行 m 列，$\boldsymbol{c}$ 是 k 行 1 列。通过上面的描述，LSTM 算法中更新的参数为：$\boldsymbol{W}_f$、$\boldsymbol{U}_f$、$\boldsymbol{b}_f$、$\boldsymbol{W}_i$、$\boldsymbol{U}_i$、$\boldsymbol{b}_i$、$\boldsymbol{W}_a$、$\boldsymbol{U}_a$、$\boldsymbol{b}_a$、$\boldsymbol{W}_o$、$\boldsymbol{U}_o$、$\boldsymbol{b}_o$、$\boldsymbol{V}$ 和 $\boldsymbol{c}$，因此通过 BPTT 更新这些参数，首先，计算 L 对 $\boldsymbol{V}$ 和 $\boldsymbol{c}$ 的偏导，即：

$$\frac{\partial L}{\partial \boldsymbol{c}} = \sum_{t=1}^{\tau} \frac{\partial L^t}{\partial \boldsymbol{c}} = \sum_{t=1}^{\tau} \frac{\partial L^t}{\partial \boldsymbol{d}^t} \frac{\partial \boldsymbol{d}^t}{\partial \boldsymbol{c}} = \sum_{t=1}^{\tau} \hat{\boldsymbol{y}}^t - \boldsymbol{y}^t$$

$$\frac{\partial L}{\partial \boldsymbol{V}} = \sum_{t=1}^{\tau} \frac{\partial L^t}{\partial \boldsymbol{V}} = \sum_{t=1}^{\tau} \frac{\partial L^t}{\partial \boldsymbol{d}^t} \frac{\partial \boldsymbol{d}^t}{\partial \boldsymbol{V}} = \sum_{t=1}^{\tau} \left(\hat{\boldsymbol{y}}^t - \boldsymbol{y}^t\right)\left(\boldsymbol{h}^t\right)^T$$

然后定义两个变量，如下：

$$\boldsymbol{\delta}_h^t = \frac{\partial L}{\partial \boldsymbol{h}^t}$$

$$\boldsymbol{\delta}_C^t = \frac{\partial L}{\partial \boldsymbol{C}^t}$$

当 t 等于 τ (最后一个时刻)，则：

$$\boldsymbol{\delta}_h^\tau = \frac{\partial L}{\partial \boldsymbol{d}^t} \frac{\partial \boldsymbol{d}^t}{\partial \boldsymbol{h}^t} = \boldsymbol{V}^T \left(\hat{\boldsymbol{y}}^\tau - \boldsymbol{y}^\tau\right)$$

$$\boldsymbol{\delta}_C^\tau = \frac{\partial L}{\partial \boldsymbol{h}^\tau} \frac{\partial \boldsymbol{h}^\tau}{\partial \boldsymbol{C}^t} = \boldsymbol{\delta}_h^\tau \odot \boldsymbol{o}^\tau \odot \left(1 - tanh\left(\boldsymbol{C}^\tau\right) \odot tanh\left(\boldsymbol{C}^\tau\right)\right)$$

当 t 不等于 τ，则：

$$\boldsymbol{\delta}_h^t = \frac{\partial L}{\partial \boldsymbol{d}^t} \frac{\partial \boldsymbol{d}^t}{\partial \boldsymbol{h}^t} = \boldsymbol{V}^T \left(\hat{\boldsymbol{y}}^t - \boldsymbol{y}^t\right)$$

$$\boldsymbol{\delta}_C^t = \frac{\partial L}{\partial \boldsymbol{C}^{t+1}} \frac{\partial \boldsymbol{C}^{t+1}}{\partial \boldsymbol{C}^t} + \frac{\partial L}{\partial \boldsymbol{h}^t} \frac{\partial \boldsymbol{h}^t}{\partial \boldsymbol{C}^t} = \boldsymbol{\delta}_C^{t+1} \odot \boldsymbol{f}^{t+1} + \boldsymbol{\delta}_h^t \odot \boldsymbol{o}^t \odot \left(1 - tanh\left(\boldsymbol{C}^t\right) \odot tanh\left(\boldsymbol{C}^t\right)\right)$$

因此，其他的参数更新的为：

$$\frac{\partial L}{\partial \boldsymbol{W}_f} = \sum_{t=1}^{\tau} \frac{\partial L}{\partial \boldsymbol{C}^t} \frac{\partial \boldsymbol{C}^t}{\partial \boldsymbol{f}^t} \frac{\partial \boldsymbol{f}^t}{\partial \boldsymbol{W}_f} = \sum_{t=1}^{\tau} \boldsymbol{\delta}_C^t \odot \boldsymbol{C}^{t-1} \odot \boldsymbol{f}^t \odot \left(1 - \boldsymbol{f}^t\right)\left(\boldsymbol{h}^{t-1}\right)^T$$

$$\frac{\partial L}{\partial \boldsymbol{U}_f} = \sum_{t=1}^{\tau} \frac{\partial L}{\partial \boldsymbol{C}^t} \frac{\partial \boldsymbol{C}^t}{\partial \boldsymbol{f}^t} \frac{\partial \boldsymbol{f}^t}{\partial \boldsymbol{U}_f} = \sum_{t=1}^{\tau} \boldsymbol{\delta}_C^t \odot \boldsymbol{C}^{t-1} \odot \boldsymbol{f}^t \odot \left(1 - \boldsymbol{f}^t\right)\left(\boldsymbol{x}^t\right)^T$$

$$\frac{\partial L}{\partial \boldsymbol{b}_f} = \sum_{t=1}^{\tau} \frac{\partial L}{\partial \boldsymbol{C}^t} \frac{\partial \boldsymbol{C}^t}{\partial \boldsymbol{f}^t} \frac{\partial \boldsymbol{f}^t}{\partial \boldsymbol{b}_f} = \sum_{t=1}^{\tau} \boldsymbol{\delta}_C^t \odot \boldsymbol{C}^{t-1} \odot \boldsymbol{f}^t \odot \left(1 - \boldsymbol{f}^t\right)$$

$$\frac{\partial L}{\partial \boldsymbol{W}_i} = \sum_{t=1}^{\tau} \frac{\partial L}{\partial \boldsymbol{C}^t} \frac{\partial \boldsymbol{C}^t}{\partial \boldsymbol{i}^t} \frac{\partial \boldsymbol{i}^t}{\partial \boldsymbol{W}_i} = \sum_{t=1}^{\tau} \boldsymbol{\delta}_C^t \odot \boldsymbol{a}^t \odot \boldsymbol{i}^t \odot \left(1 - \boldsymbol{i}^t\right)\left(\boldsymbol{h}^{t-1}\right)^T$$

$$\frac{\partial L}{\partial \boldsymbol{U}_i}=\sum_{t=1}^{\tau}\frac{\partial L}{\partial \boldsymbol{C}^t}\frac{\partial \boldsymbol{C}^t}{\partial \boldsymbol{i}^t}\frac{\partial \boldsymbol{i}^t}{\partial \boldsymbol{U}_i}=\sum_{t=1}^{\tau}\boldsymbol{\delta}_C^t\odot\boldsymbol{a}^t\odot\boldsymbol{i}^t\odot\left(1-\boldsymbol{i}^t\right)\left(\boldsymbol{x}^t\right)^T$$

$$\frac{\partial L}{\partial \boldsymbol{b}_i}=\sum_{t=1}^{\tau}\frac{\partial L}{\partial \boldsymbol{C}^t}\frac{\partial \boldsymbol{C}^t}{\partial \boldsymbol{i}^t}\frac{\partial \boldsymbol{i}^t}{\partial \boldsymbol{b}_i}=\sum_{t=1}^{\tau}\boldsymbol{\delta}_C^t\odot\boldsymbol{a}^t\odot\boldsymbol{i}^t\odot\left(1-\boldsymbol{i}^t\right)$$

$$\frac{\partial L}{\partial \boldsymbol{W}_a}=\sum_{t=1}^{\tau}\frac{\partial L}{\partial \boldsymbol{C}^t}\frac{\partial \boldsymbol{C}^t}{\partial \boldsymbol{a}^t}\frac{\partial \boldsymbol{a}^t}{\partial \boldsymbol{W}_a}=\sum_{t=1}^{\tau}\boldsymbol{\delta}_C^t\odot\boldsymbol{i}^t\odot\left(1-\boldsymbol{a}^t\odot\boldsymbol{a}^t\right)\left(\boldsymbol{h}^{t-1}\right)^T$$

$$\frac{\partial L}{\partial \boldsymbol{U}_a}=\sum_{t=1}^{\tau}\frac{\partial L}{\partial \boldsymbol{C}^t}\frac{\partial \boldsymbol{C}^t}{\partial \boldsymbol{a}^t}\frac{\partial \boldsymbol{a}^t}{\partial \boldsymbol{U}_a}=\sum_{t=1}^{\tau}\boldsymbol{\delta}_C^t\odot\boldsymbol{i}^t\odot\left(1-\boldsymbol{a}^t\odot\boldsymbol{a}^t\right)\left(\boldsymbol{x}^t\right)^T$$

$$\frac{\partial L}{\partial \boldsymbol{b}_a}=\sum_{t=1}^{\tau}\frac{\partial L}{\partial \boldsymbol{C}^t}\frac{\partial \boldsymbol{C}^t}{\partial \boldsymbol{a}^t}\frac{\partial \boldsymbol{a}^t}{\partial \boldsymbol{b}_a}=\sum_{t=1}^{\tau}\boldsymbol{\delta}_C^t\odot\boldsymbol{i}^t\odot\left(1-\boldsymbol{a}^t\odot\boldsymbol{a}^t\right)$$

$$\frac{\partial L}{\partial \boldsymbol{W}_o}=\sum_{t=1}^{\tau}\frac{\partial L}{\partial \boldsymbol{h}^t}\frac{\partial \boldsymbol{h}^t}{\partial \boldsymbol{o}^t}\frac{\partial \boldsymbol{o}^t}{\partial \boldsymbol{W}_o}=\sum_{t=1}^{\tau}\boldsymbol{\delta}_h^t\odot tanh\left(\boldsymbol{C}^t\right)\odot\left(1-\boldsymbol{o}^t\odot\boldsymbol{o}^t\right)\left(\boldsymbol{h}^{t-1}\right)^T$$

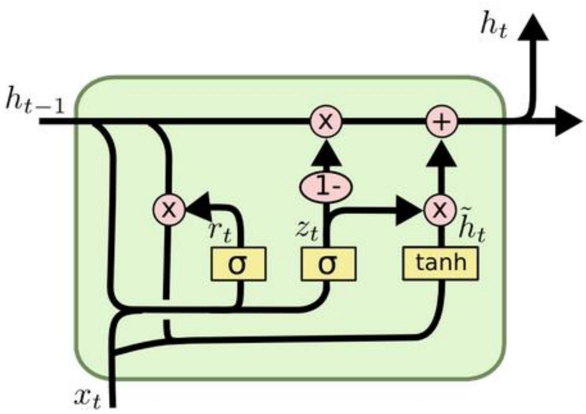
$$\frac{\partial L}{\partial \boldsymbol{U}_o}=\sum_{t=1}^{\tau}\frac{\partial L}{\partial \boldsymbol{h}^t}\frac{\partial \boldsymbol{h}^t}{\partial \boldsymbol{o}^t}\frac{\partial \boldsymbol{o}^t}{\partial \boldsymbol{U}_o}=\sum_{t=1}^{\tau}\boldsymbol{\delta}_h^t\odot tanh\left(\boldsymbol{C}^t\right)\odot\left(1-\boldsymbol{o}^t\odot\boldsymbol{o}^t\right)\left(\boldsymbol{x}^t\right)^T$$

$$\frac{\partial L}{\partial \boldsymbol{b}_o}=\sum_{t=1}^{\tau}\frac{\partial L}{\partial \boldsymbol{h}^t}\frac{\partial \boldsymbol{h}^t}{\partial \boldsymbol{o}^t}\frac{\partial \boldsymbol{o}^t}{\partial \boldsymbol{b}_o}=\sum_{t=1}^{\tau}\boldsymbol{\delta}_h^t\odot tanh\left(\boldsymbol{C}^t\right)\odot\left(1-\boldsymbol{o}^t\odot\boldsymbol{o}^t\right)$$

最后，根据 RNN 模型更新参数的方式更新参数。LSTM 模型之所以可以减小(不是解决)RNN 的梯度消失问题，是因为 LSTM 模型引入了门控开关(模型里面的 sigmoid 函数起着门控开关的作用)和 Hadamard 积，这可以防止对模型参数的偏导进行矩阵连乘，从而可以减小梯度消失。

3. GRU 模型

GRU(Gated Recurrent Unit)又叫门控循环单元，是 LSTM 模型的一种变体，同样是为了解决 RNN 模型的梯度消失问题。下面是 GRU 的结构图：



上图中，σ 表示的是 sigmoid 函数。因此 GRU 模型的前向传播为：

$$
\begin{aligned}
&\boldsymbol{z}^t=\sigma\left(\boldsymbol{W}_z\boldsymbol{h}^{t-1}+\boldsymbol{U}_z\boldsymbol{x}^t+\boldsymbol{b}_z\right)\\
&\boldsymbol{r}^t=\sigma\left(\boldsymbol{W}_r\boldsymbol{h}^{t-1}+\boldsymbol{U}_r\boldsymbol{x}^t+\boldsymbol{b}_r\right)\\
&\boldsymbol{s}^t=tanh\left(\boldsymbol{W}_s\left(\boldsymbol{h}^{t-1}\odot\boldsymbol{r}^t\right)+\boldsymbol{U}_s\boldsymbol{x}^t+\boldsymbol{b}_s\right)\\
&\boldsymbol{h}^t=\left(1-\boldsymbol{z}^t\right)\odot\boldsymbol{h}^{t-1}+\boldsymbol{z}^t\odot\boldsymbol{s}^t\\
&\boldsymbol{o}^t=\boldsymbol{V}\boldsymbol{h}^t+\boldsymbol{c}\\
&\hat{\boldsymbol{y}}^t=softmax\left(\boldsymbol{o}^t\right)
\end{aligned}
$$

上式中，$\boldsymbol{W}_z$是 m 行 m 列，$\boldsymbol{h}^{t\text{-}1}$是 m 行 1 列，$\boldsymbol{U}_z$是 m 行 n 列，$\boldsymbol{x}_t$是 n 行 1 列，$\boldsymbol{b}_z$是 m 行 1 列，$\boldsymbol{z}^t$是 m 行 1 列，$\boldsymbol{W}_r$是 m 行 m 列，$\boldsymbol{U}_r$是 m 行 n 列，$\boldsymbol{b}_r$是 m 行 1 列，$\boldsymbol{r}^t$是 m 行 1 列，$\boldsymbol{W}_s$是 m 行 m 列，$\boldsymbol{U}_s$是 m 行 n 列，$\boldsymbol{b}_s$是 m 行 1 列，$\boldsymbol{s}^t$等于$\tilde{\boldsymbol{h}}^t$，$\boldsymbol{s}^t$是 m 行 1 列，$\hat{\boldsymbol{y}}^t$ 是 k 行 1 列，$\boldsymbol{V}$是 k 行 m 列，$\boldsymbol{c}$ 是 k 行 1 列。因此通过 BPTT 更新这些参数，首先，计算 L 对 $\boldsymbol{V}$ 和 $\boldsymbol{c}$ 的偏导，即：

$$
\begin{aligned}
&\frac{\partial L}{\partial \boldsymbol{c}}=\sum_{t=1}^{\tau}\frac{\partial L^t}{\partial \boldsymbol{c}}=\sum_{t=1}^{\tau}\frac{\partial L^t}{\partial \boldsymbol{o}^t}\frac{\partial \boldsymbol{o}^t}{\partial \boldsymbol{c}}=\sum_{t=1}^{\tau}\hat{\boldsymbol{y}}^t-\boldsymbol{y}^t\\
&\frac{\partial L}{\partial \boldsymbol{V}}=\sum_{t=1}^{\tau}\frac{\partial L^t}{\partial \boldsymbol{V}}=\sum_{t=1}^{\tau}\frac{\partial L^t}{\partial \boldsymbol{o}^t}\frac{\partial \boldsymbol{o}^t}{\partial \boldsymbol{V}}=\sum_{t=1}^{\tau}\left(\hat{\boldsymbol{y}}^t-\boldsymbol{y}^t\right)\left(\boldsymbol{h}^t\right)^T
\end{aligned}
$$

然后计算 $\boldsymbol{W}$、$\boldsymbol{U}$ 和 $\boldsymbol{b}$ 的梯度，定义序列索引 t 位置的隐藏状态的梯度为：

$$
\boldsymbol{\delta}^t=\frac{\partial L}{\partial \boldsymbol{h}^t}
$$

上式中，$\boldsymbol{\delta}^t$ 是一个 m 行 1 列的向量。当 t 不等于 τ (最后一个时刻)时，

$$
\begin{aligned}
\boldsymbol{\delta}^t&=\frac{\partial L}{\partial \boldsymbol{o}^t}\frac{\partial \boldsymbol{o}^t}{\partial \boldsymbol{h}^t}+\frac{\partial L}{\partial \boldsymbol{h}^{t+1}}\frac{\partial \boldsymbol{h}^{t+1}}{\partial \boldsymbol{h}^t}\\
&=\boldsymbol{V}^T\left(\hat{\boldsymbol{y}}^t-\boldsymbol{y}^t\right)+\boldsymbol{\delta}^{t+1}\left(1-\boldsymbol{z}^t\right)
\end{aligned}
$$

当 t 等于 τ 时，

$$
\boldsymbol{\delta}^\tau=\frac{\partial L}{\partial \boldsymbol{o}^\tau}\frac{\partial \boldsymbol{o}^\tau}{\partial \boldsymbol{h}^\tau}=\boldsymbol{V}^T\left(\hat{\boldsymbol{y}}^\tau-\boldsymbol{y}^\tau\right)
$$

因此，$\boldsymbol{W}_z$、$\boldsymbol{U}_z$、$\boldsymbol{b}_z$、$\boldsymbol{W}_r$、$\boldsymbol{U}_r$、$\boldsymbol{b}_r$、$\boldsymbol{W}_s$、$\boldsymbol{U}_s$和 $\boldsymbol{b}_s$ 的更新方式为：

$$
\frac{\partial L}{\partial \boldsymbol{W}_z}=\sum_{t=1}^{\tau}\frac{\partial L}{\partial \boldsymbol{h}^t}\frac{\partial \boldsymbol{h}^t}{\partial \boldsymbol{z}^t}\frac{\partial \boldsymbol{z}^t}{\partial \boldsymbol{W}_z}=\sum_{t=1}^{\tau}\boldsymbol{\delta}^t\odot\left(\boldsymbol{s}^t-\boldsymbol{h}^{t-1}\right)\odot\left(\boldsymbol{h}^{t-1}\right)^T
$$

$$
\frac{\partial L}{\partial \boldsymbol{U}_z}=\sum_{t=1}^{\tau}\frac{\partial L}{\partial \boldsymbol{h}^t}\frac{\partial \boldsymbol{h}^t}{\partial \boldsymbol{z}^t}\frac{\partial \boldsymbol{z}^t}{\partial \boldsymbol{U}_z}=\sum_{t=1}^{\tau}\boldsymbol{\delta}^t\odot\left(\boldsymbol{s}^t-\boldsymbol{h}^{t-1}\right)\odot\left(\boldsymbol{x}^t\right)^T
$$

$$
\frac{\partial L}{\partial \boldsymbol{b}_z}=\sum_{t=1}^{\tau}\frac{\partial L}{\partial \boldsymbol{h}^t}\frac{\partial \boldsymbol{h}^t}{\partial \boldsymbol{z}^t}\frac{\partial \boldsymbol{z}^t}{\partial \boldsymbol{b}_z}=\sum_{t=1}^{\tau}\boldsymbol{\delta}^t\odot\left(\boldsymbol{s}^t-\boldsymbol{h}^{t-1}\right)
$$

$$\frac{\partial L}{\partial \boldsymbol{W}_r}=\sum_{t=1}^{\tau}\frac{\partial L}{\partial \boldsymbol{h}^t}\frac{\partial \boldsymbol{h}^t}{\partial \boldsymbol{s}^t}\frac{\partial \boldsymbol{s}^t}{\partial \boldsymbol{r}^t}\frac{\partial \boldsymbol{r}^t}{\partial \boldsymbol{W}_r}=\sum_{t=1}^{\tau}\boldsymbol{\delta}^t\odot\boldsymbol{z}^t\frac{\partial \boldsymbol{s}^t}{\partial \boldsymbol{r}^t}\frac{\partial \boldsymbol{r}^t}{\partial \boldsymbol{W}_r}$$

$$=\sum_{t=1}^{\tau}\frac{\partial\left(\boldsymbol{\delta}^t\odot\boldsymbol{z}^t\right)^T\boldsymbol{s}^t}{\partial \boldsymbol{r}^t}\frac{\partial \boldsymbol{r}^t}{\partial \boldsymbol{W}_r}=\sum_{t=1}^{\tau}\left(\boldsymbol{W}_s\right)^T\left(\boldsymbol{\delta}^t\odot\boldsymbol{z}^t\odot\left(1-\boldsymbol{s}^t\odot\boldsymbol{s}^t\right)\right)\odot\boldsymbol{h}^{t-1}\frac{\partial \boldsymbol{r}^t}{\partial \boldsymbol{W}_r}$$

$$=\sum_{t=1}^{\tau}\left(\boldsymbol{W}_s\right)^T\left(\boldsymbol{\delta}^t\odot\boldsymbol{z}^t\odot\left(1-\boldsymbol{s}^t\odot\boldsymbol{s}^t\right)\right)\odot\boldsymbol{h}^{t-1}\odot\boldsymbol{r}^t\odot\left(1-\boldsymbol{r}^t\right)\left(\boldsymbol{h}^{t-1}\right)^T$$

$$\frac{\partial L}{\partial \boldsymbol{U}_r}=\sum_{t=1}^{\tau}\frac{\partial L}{\partial \boldsymbol{h}^t}\frac{\partial \boldsymbol{h}^t}{\partial \boldsymbol{s}^t}\frac{\partial \boldsymbol{s}^t}{\partial \boldsymbol{r}^t}\frac{\partial \boldsymbol{r}^t}{\partial \boldsymbol{U}_r}=\sum_{t=1}^{\tau}\boldsymbol{\delta}^t\odot\boldsymbol{z}^t\frac{\partial \boldsymbol{s}^t}{\partial \boldsymbol{r}^t}\frac{\partial \boldsymbol{r}^t}{\partial \boldsymbol{U}_r}$$

$$=\sum_{t=1}^{\tau}\frac{\partial\left(\boldsymbol{\delta}^t\odot\boldsymbol{z}^t\right)^T\boldsymbol{s}^t}{\partial \boldsymbol{r}^t}\frac{\partial \boldsymbol{r}^t}{\partial \boldsymbol{U}_r}=\sum_{t=1}^{\tau}\left(\boldsymbol{W}_s\right)^T\left(\boldsymbol{\delta}^t\odot\boldsymbol{z}^t\odot\left(1-\boldsymbol{s}^t\odot\boldsymbol{s}^t\right)\right)\odot\boldsymbol{h}^{t-1}\frac{\partial \boldsymbol{r}^t}{\partial \boldsymbol{U}_r}$$

$$=\sum_{t=1}^{\tau}\left(\boldsymbol{W}_s\right)^T\left(\boldsymbol{\delta}^t\odot\boldsymbol{z}^t\odot\left(1-\boldsymbol{s}^t\odot\boldsymbol{s}^t\right)\right)\odot\boldsymbol{h}^{t-1}\odot\boldsymbol{r}^t\odot\left(1-\boldsymbol{r}^t\right)\left(\boldsymbol{x}^t\right)^T$$

$$\frac{\partial L}{\partial \boldsymbol{b}_r}=\sum_{t=1}^{\tau}\frac{\partial L}{\partial \boldsymbol{h}^t}\frac{\partial \boldsymbol{h}^t}{\partial \boldsymbol{s}^t}\frac{\partial \boldsymbol{s}^t}{\partial \boldsymbol{r}^t}\frac{\partial \boldsymbol{r}^t}{\partial \boldsymbol{b}_r}=\sum_{t=1}^{\tau}\boldsymbol{\delta}^t\odot\boldsymbol{z}^t\frac{\partial \boldsymbol{s}^t}{\partial \boldsymbol{r}^t}\frac{\partial \boldsymbol{r}^t}{\partial \boldsymbol{b}_r}$$

$$=\sum_{t=1}^{\tau}\frac{\partial\left(\boldsymbol{\delta}^t\odot\boldsymbol{z}^t\right)^T\boldsymbol{s}^t}{\partial \boldsymbol{r}^t}\frac{\partial \boldsymbol{r}^t}{\partial \boldsymbol{b}_r}=\sum_{t=1}^{\tau}\left(\boldsymbol{W}_s\right)^T\left(\boldsymbol{\delta}^t\odot\boldsymbol{z}^t\odot\left(1-\boldsymbol{s}^t\odot\boldsymbol{s}^t\right)\right)\odot\boldsymbol{h}^{t-1}\frac{\partial \boldsymbol{r}^t}{\partial \boldsymbol{b}_r}$$

$$=\sum_{t=1}^{\tau}\left(\boldsymbol{W}_s\right)^T\left(\boldsymbol{\delta}^t\odot\boldsymbol{z}^t\odot\left(1-\boldsymbol{s}^t\odot\boldsymbol{s}^t\right)\right)\odot\boldsymbol{h}^{t-1}\odot\boldsymbol{r}^t\odot\left(1-\boldsymbol{r}^t\right)$$

$$\frac{\partial L}{\partial \boldsymbol{W}_s}=\sum_{t=1}^{\tau}\frac{\partial L}{\partial \boldsymbol{h}^t}\frac{\partial \boldsymbol{h}^t}{\partial \boldsymbol{s}^t}\frac{\partial \boldsymbol{s}^t}{\partial \boldsymbol{W}_s}=\sum_{t=1}^{\tau}\boldsymbol{\delta}^t\odot\boldsymbol{z}^t\odot\left(1-\boldsymbol{s}^t\odot\boldsymbol{s}^t\right)\left(\boldsymbol{h}^{t-1}\odot\boldsymbol{r}^t\right)^T$$

$$\frac{\partial L}{\partial \boldsymbol{U}_s}=\sum_{t=1}^{\tau}\frac{\partial L}{\partial \boldsymbol{h}^t}\frac{\partial \boldsymbol{h}^t}{\partial \boldsymbol{s}^t}\frac{\partial \boldsymbol{s}^t}{\partial \boldsymbol{U}_s}=\sum_{t=1}^{\tau}\boldsymbol{\delta}^t\odot\boldsymbol{z}^t\odot\left(1-\boldsymbol{s}^t\odot\boldsymbol{s}^t\right)\left(\boldsymbol{x}^t\right)^T$$

$$\frac{\partial L}{\partial \boldsymbol{b}_s}=\sum_{t=1}^{\tau}\frac{\partial L}{\partial \boldsymbol{h}^t}\frac{\partial \boldsymbol{h}^t}{\partial \boldsymbol{s}^t}\frac{\partial \boldsymbol{s}^t}{\partial \boldsymbol{b}_s}=\sum_{t=1}^{\tau}\boldsymbol{\delta}^t\odot\boldsymbol{z}^t\odot\left(1-\boldsymbol{s}^t\odot\boldsymbol{s}^t\right)$$

最后，根据 RNN 模型更新参数的方式更新参数。GRU 模型之所以可以减小(不是解决)RNN 的梯度消失问题，和 LSTM 减小 RNN 梯度消失的原因一样，也是因为 GRU 模型引入了门控开关(模型里面的 sigmoid 函数起着门控开关的作用)和 Hadamard 积，这可以防止对模型参数的偏导进行矩阵连乘，从而可以减小梯度消失。

注：

1. 本文用 $\odot$ 表示 Hadamard 积，对于两个维度相同的向量 $\boldsymbol{a}=\left[a_1,a_2,\ldots,a_m\right]^T$ 和 $\boldsymbol{b}=\left[b_1,b_2,\ldots,b_m\right]^T$，则 $\boldsymbol{a}\odot\boldsymbol{b}=\left[a_1b_1,a_2b_2,\ldots,a_mb_m\right]^T$；

2.假设两个维度相同的向量 $\boldsymbol{y}=\left[y_1,y_2,\ldots,y_m\right]^T$ 和 $\boldsymbol{x}=\left[x_1,x_2,\ldots,x_m\right]^T$ 满足如下公式：

$$\boldsymbol{y} = f\left(\boldsymbol{x}\right) \Leftrightarrow \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} f\left(x_1\right) \\ f\left(x_2\right) \\ \vdots \\ f\left(x_m\right) \end{bmatrix}$$

则：

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \begin{bmatrix} f'\left(x_1\right) \\ f'\left(x_2\right) \\ \vdots \\ f'\left(x_m\right) \end{bmatrix}$$

3.假设两个维度相同的向量 $\boldsymbol{y} = \left[y_1, y_2, \ldots, y_m\right]^T$ 和 $\boldsymbol{x} = \left[x_1, x_2, \ldots, x_n\right]^T$ 满足如下公式：

$$\boldsymbol{y} = \boldsymbol{A}_{m\times n}\boldsymbol{x}$$

则：

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \boldsymbol{A}_{m\times n}^T$$
$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{A}_{m\times n}} = \boldsymbol{x}^T$$

4. 如果 LSTM 与 GRU 减小梯度消失原因看不懂，可以参考以下这篇文章：
https://zhuanlan.zhihu.com/p/28297161