计算 LSTM 的梯度

16337341 朱志儒

设 LSTM 的损失函数为 L(t)，有

$$\mathrm{L(t)} = \sum_{t=1}^{T} L^{(t)}$$

对于两个隐藏状态$\mathrm{h}^{(t)}$和$\mathrm{C}^{(t)}$有

$$\delta_h^{(t)} = \frac{\partial L}{\partial h^{(t)}} = \left(\frac{\partial O^{(\tau)}}{\partial \mathrm{h}^{(\tau)}}\right)^T \frac{\partial L^{(\tau)}}{\partial O^{(\tau)}} = V^T\left(\widehat{y^{(\tau)}} - y^{(\tau)}\right)$$

$$\delta_C^{(t)} = \frac{\partial L}{\partial \mathrm{C}^{(t)}} = \left(\frac{\partial h^{(\tau)}}{\partial \mathrm{C}^{(\tau)}}\right)^T \frac{\partial L^{(\tau)}}{\partial h^{(\tau)}} = \delta_h^{(\tau)} \odot o^{(\tau)} \odot \left(1 - tanh^2\left(C^{(\tau)}\right)\right)$$

接着由$\delta_C^{(t+1)}, \delta_h^{(t+1)}$反向推导$\delta_C^{(t)}, \delta_h^{(t)}$。

$\delta_h^{(t)}$的梯度由本层 t 时刻的输出梯度误差和大于 t 时刻的误差两部分决定，即：

$$\delta_h^{(t)} = \frac{\partial L}{\partial \mathrm{h}^{(t)}} = \frac{\partial l(t)}{\partial \mathrm{h}^{(t)}} + \left(\frac{\partial h^{(t+1)}}{\partial \mathrm{h}^{(t)}}\right)^T \frac{\partial L(t+1)}{\partial \mathrm{h}^{(t+1)}} = V^T\left(\widehat{y^{(t)}} - y^{(t)}\right) + \left(\frac{\partial h^{(t+1)}}{\partial \mathrm{h}^{(t)}}\right)^T \delta_h^{(t+1)}$$

$$\begin{aligned}\frac{\partial h^{(t+1)}}{\partial h^{(t)}} = {} & W_o^T\left[o^{(t+1)} \odot \left(1 - o^{(t+1)}\right) \odot tanh\left(C^{(t+1)}\right)\right] \\ & + W_f^T\left[\Delta C \odot f^{(t+1)} \odot \left(1 - f^{(t+1)}\right) \odot C^{(t)}\right] + W_a^T\{\Delta C \odot i^{(t+1)} \\ & \odot \left[1 - \left(a^{(t+1)}\right)^2\right]\} + W_i^T\left[\Delta C \odot a^{(t+1)} \backslash\mathrm{odot}\, i^{(t+1)} \odot \left(1 - i^{(t+1)}\right)\right]\end{aligned}$$

而$\delta_C^{(t)}$的反向梯度误差由前一层$\delta_C^{(t+1)}$的梯度误差和本层的从$\mathrm{h}^{(t)}$传回来的梯度误差两部分组成，即：

$$\begin{aligned}\delta_C^{(t)} &= \left(\frac{\partial C^{(t+1)}}{\partial \mathrm{C}^{(t)}}\right)^T \frac{\partial L}{\partial C^{(t+1)}} + \left(\frac{\partial h^{(t)}}{\partial C^{(t)}}\right)^T \frac{\partial L}{\partial h^{(t)}} \\ &= \left(\frac{\partial C^{(t+1)}}{\partial \mathrm{C}^{(t)}}\right)^T \delta_C^{(t+1)} + \delta_h^{(t)} \odot o^{(t)} \odot \left(1 - tanh^2\left(C^{(t)}\right)\right) \\ &= \delta_C^{(t+1)} \odot f^{(t+1)} + \delta_h^{(t)} \odot o^{(t)} \odot \left(1 - tanh^2\left(C^{(t)}\right)\right)\end{aligned}$$

有了$\delta_C^{(t)}$和$\delta_h^{(t)}$，计算$\mathrm{W_f}$的梯度计算过程：

$$\frac{\partial L}{\partial W_f} = \sum_{t=1}^{\tau}\left[\delta_C^{(t)} \odot \mathrm{C}^{(t-1)} \odot f^{(t)} \odot \left(1 - f^{(t)}\right)\right]\left(h^{(t-1)}\right)^T$$