

RNN中反向传播的梯度问题

原创 睡星 代码搜罗屋 2020-08-23

RNN反向传播中梯度问题

梯度计算公式

$$\begin{aligned}
 h_t &= \tanh(W_I x_t + W_R h_{t-1}) \\
 y_t &= W_O h_t \\
 \frac{\partial E_t}{\partial W_R} &= \sum_{i=0}^t \frac{\partial E_t}{\partial y_t} \frac{\partial y_t}{\partial h_t} \frac{\partial h_t}{\partial h_i} \frac{\partial h_i}{\partial W_R} \\
 \frac{\partial h_t}{\partial h_i} &= \frac{\partial h_t}{\partial h_{t-1}} \frac{\partial h_{t-1}}{\partial h_{t-2}} \cdots \frac{\partial h_{i+1}}{\partial h_i} = \prod_{k=i}^{t-1} \frac{\partial h_{k+1}}{\partial h_k} \\
 \frac{\partial h_{k+1}}{\partial h_k} &= \text{diag}(f'(W_I x_i + W_R h_{i-1})) W_R \\
 \frac{\partial h_k}{\partial h_1} &= \prod_i^k \text{diag}(f'(W_I x_i + W_R h_{i-1})) W_R
 \end{aligned}$$

梯度弥散和梯度爆炸

$$\begin{aligned}
 1.01^{365} &= 37.8 \\
 0.99^{365} &= 0.03
 \end{aligned}$$

如果 W_{hh} 中的元素略大于1梯度接近无穷大，如果其中的元素略小于1，最终梯度会接近0
梯度爆炸表现在本来loss是很小的，突然loss变得很大

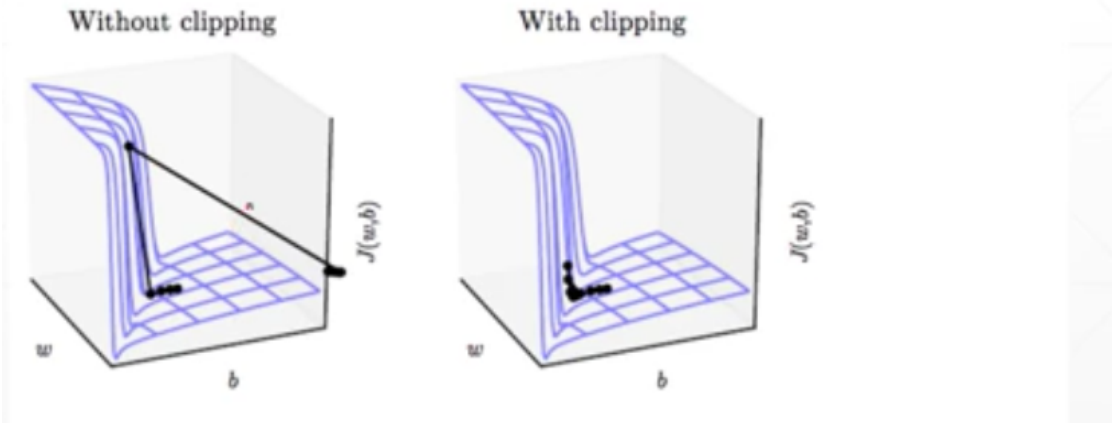
Algorithm 1 Pseudo-code for norm clipping

$$\hat{g} \leftarrow \frac{\partial \mathcal{E}}{\partial \theta}$$

if $\|\hat{g}\| \geq threshold$ then

$$\hat{g} \leftarrow \frac{threshold}{\|\hat{g}\|} \hat{g}$$

end if



梯度爆炸解决方法

设置梯度门限例如，如果梯度向量大于15，让梯度除以它的模，保留方向，进行小数值试探

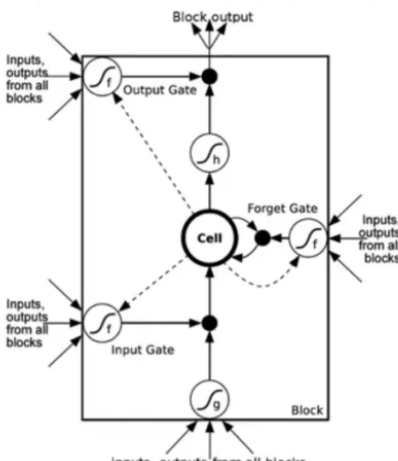
梯度弥散解决方法

一个网络太深的时候前面几层可以的得到更新，后面的层得不到更新，模型训练很多次还是效果不好。LSTM，深度残差网络

RNN记不住长期的信息，short term memory；改进就是LSTM

在Cell传递过程中和CNN中深度残差网络相似，可以跳出层次的限制可以减少梯度弥散

- 当输入门为0，遗忘门为1，即保留上一时间戳所有记忆单元，不加入任何新信息
- 当输入门为1，遗忘门为1，即保留上一时间戳所有记忆单元，加入当前所有新信息
- 当输入门为0，遗忘门为0，清空所有信息
- 当输入门为1，遗忘门为0，清空上一时间戳所有记忆单元，加入当前所有新信息



input gate	forget gate	behavior
0	1	remember the previous value
1	1	add to the previous value
0	0	erase the value
1	0	overwrite the value

LSTM的梯度公式

How to solve Gradient Vanishing?

$$\frac{\partial C_t}{\partial C_{t-1}} = \frac{\partial C_t}{\partial f_t} \frac{\partial f_t}{\partial h_{t-1}} \frac{\partial h_{t-1}}{\partial C_{t-1}} + \frac{\partial C_t}{\partial i_t} \frac{\partial i_t}{\partial h_{t-1}} \frac{\partial h_{t-1}}{\partial C_{t-1}} + \frac{\partial C_t}{\partial \tilde{C}_t} \frac{\partial \tilde{C}_t}{\partial h_{t-1}} \frac{\partial h_{t-1}}{\partial C_{t-1}} + \frac{\partial C_t}{\partial C_{t-1}}$$
$$\frac{\partial C_t}{\partial C_{t-1}} = C_{t-1} \sigma'(\cdot) W_f * o_{t-1} \tanh'(C_{t-1}) + \tilde{C}_t \sigma'(\cdot) W_i * o_{t-1} \tanh'(C_{t-1}) + i_t \tanh'(\cdot) W_C * o_{t-1} \tanh'(C_{t-1}) + f_t$$

<https://weberna.github.io/blog/2017/11/15/LSTM-Vanishing-Gradients.html>
http://www.cs.toronto.edu/~rgrosse/courses/csc321_2017/readings/L15%20Exploding%20and%20Vanishing%20Gradients.pdf

$$\frac{\partial h_{k+1}}{\partial h_k} = \text{diag}(f'(W_I x_i + W_R h_{i-1})) W_R$$
$$\frac{\partial h_k}{\partial h_1} = \prod_i^k \text{diag}(f'(W_I x_i + W_R h_{i-1})) W_R$$

因为有四项累加的存在，很少会出现几个权重同时爆炸，或者消失的情况

喜欢此内容的人还喜欢

高层次人才PPT介绍：关于技术细节，“听不懂”还是“不想听”？

科奖中心

山东涉疫奶枣流入3省，多地检出阳性！这个传播渠道要警惕

中国反邪教