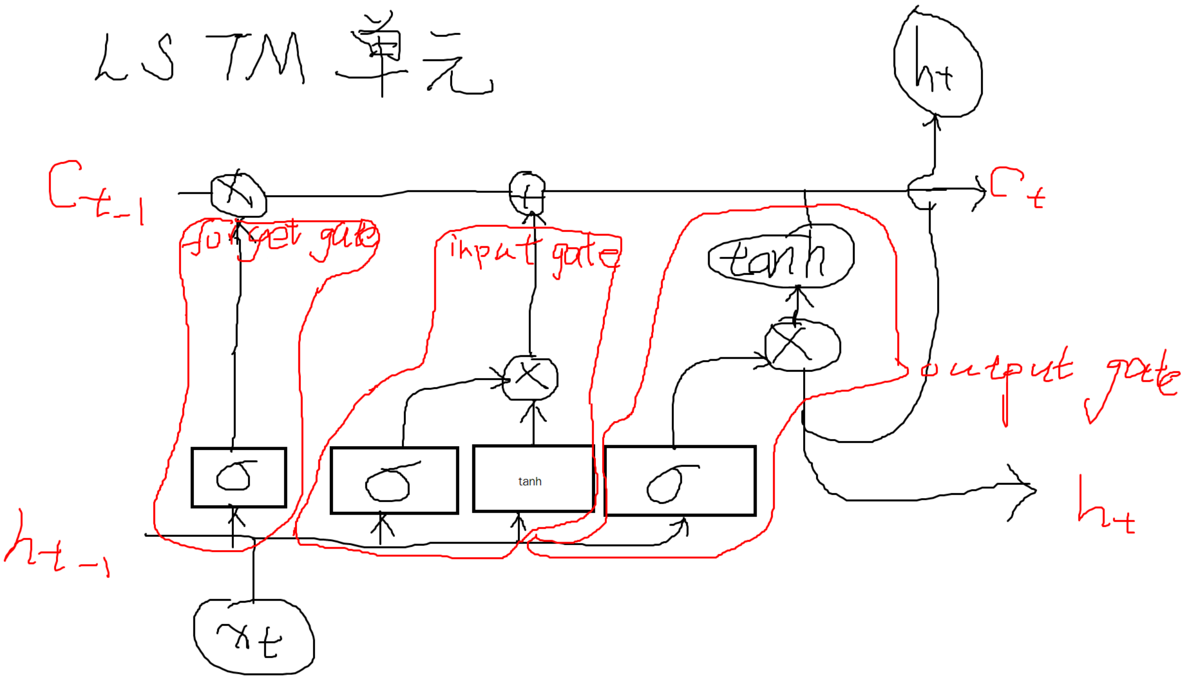
LSTM 单元草稿&笔记

LSTM 单元的结构：



一个 LSTM 单元包括 3 个主要部分：遗忘门(forget gate)、输入门(input gate)和输出门(output gate)。

遗忘门：$f_t = \sigma(w_f h_{t-1} + u_f x_t + b_f)$

输入门：$i_t = \sigma(w_i h_{t-1} + u_i x_t + b_i) \times \tanh(w_a h_{t-1} + u_a x_t + b_a)$

细胞状态：$C_t = C_{t-1} \times f_t + i_t$

输出门：$h_t = \tanh(C_t) \times \sigma(w_o h_{t-1} + u_o x_t + b_o)$

x 是输入。所有的 w，u 和 b 都是待定系数。f，i 和 C 是中间量。最后需要的是输出 h。输入 x 是一个序列时，利用上面 4 条公式反复计算可以得到输出序列 h。初始状态下 C 和 h 可以设置成 0。

注意：

1，输出序列 h 和 x 长度一致，根据实际需要可以舍弃大部分 h 只取最后面的少数 h 作为预测结果。

2，LSTM 的输出 h 的值域是(-1,1)。

当输入序列 x 和输出序列 h 已知时，利用反向传播算法可以计算出待定系数的梯度，然后利用梯度下降法逼近所有待定系数的理想值。

确定所有系数后，需要根据已有序列 x 求出对应序列 h 的时候就可以利用公式算出来。

LSTM 反向传播的计算

LSTM 的算法是一种递归运算。因此在求导的时候需要注意，对确定的任意 ht，需要考虑的不仅是当前参数的影响，还要考虑上一步运算产生的 C(t-1)和 h(t-1)的影响。所以它的导数包含 3 个部分。

当 t=1 的时候，用 MSE 单独计算这个误差（O 是期望值）：

$E=(O_1-h_1)^2$，输出对误差的偏导数，$\frac{\partial E}{\partial h_1}=2(O_1-h_1)$。

由于 C0 和 h0 不存在，一般被认为等于 0，故不需要考虑它们对 h1 的影响。
因此 t=1 时：

$$\frac{\partial E}{\partial p}=\frac{\partial E}{\partial h_1}\cdot\frac{\partial h_1}{\partial p}$$

p 指代任意一个系数参数。
计算 t=2 时的偏导数，由于 h2 受 C1 和 h1 的影响，所以需要用链式求导把之前的影响加进去：

$$\frac{\partial E}{\partial p}=\frac{\partial E}{\partial h_2}\cdot\frac{\partial h_2}{\partial p}$$

将第二个因子展开

$$\frac{\partial h_2}{\partial p}=\frac{\partial h_2}{\partial p}+\frac{\partial h_2}{\partial h_1}\cdot\frac{\partial h_1}{\partial p}+\frac{\partial h_2}{\partial C_1}\cdot\frac{\partial C_1}{\partial p}$$

这里没有重复计算，等号右边的 $\frac{\partial h_2}{\partial p}$ 是 p 直接对 h2 产生的影响，右边两项分别是通过 h1 和 C1 对 h2 产生的影响，这三个加起来才是总的影响。将公式代回去得到：

$$\frac{\partial E}{\partial p}=\frac{\partial E}{\partial h_2}\cdot\left(\frac{\partial h_2}{\partial p}+\frac{\partial h_2}{\partial h_1}\cdot\frac{\partial h_1}{\partial p}+\frac{\partial h_2}{\partial C_1}\cdot\frac{\partial C_1}{\partial p}\right)$$

这里，p 对 h2 直接影响外，还通过上一次运算产生的 h1 和 C1 对这次的计算产生影响。
考虑 t=3 的情况：

$$\frac{\partial E}{\partial p}=\frac{\partial E}{\partial h_3}\cdot\frac{\partial h_3}{\partial p}$$

展开需要拆分的因子：

$$\frac{\partial h_3}{\partial p}=\frac{\partial h_3}{\partial p}+\frac{\partial h_3}{\partial h_2}\cdot\frac{\partial h_2}{\partial p}+\frac{\partial h_3}{\partial C_2}\cdot\frac{\partial C_2}{\partial p}$$

这里的 $\frac{\partial h_2}{\partial p}$ 和 $\frac{\partial C_2}{\partial p}$ 继续展开：

$$\frac{\partial h_2}{\partial p}=\frac{\partial h_2}{\partial p}+\frac{\partial h_2}{\partial h_1}\cdot\frac{\partial h_1}{\partial p}+\frac{\partial h_2}{\partial C_1}\cdot\frac{\partial C_1}{\partial p}$$

$$\frac{\partial C_2}{\partial p}=\frac{\partial C_2}{\partial p}+\frac{\partial C_2}{\partial C_1}\cdot\frac{\partial C_1}{\partial p}+\frac{\partial C_2}{\partial h_1}\cdot\frac{\partial h_1}{\partial p}$$

t=3 时这个拆分合起来太长了，这里不合并了。
想办法归纳一下。由 LSTM 的算法易直接求得的因子：

$\frac{\partial C_t}{\partial h_{t-1}}$，$\frac{\partial C_t}{\partial C_{t-1}}$，$\frac{\partial C_t}{\partial p}$，$\frac{\partial h_t}{\partial p}$，$\frac{\partial h_t}{\partial C_{t-1}}$，$\frac{\partial h_t}{\partial h_{t-1}}$，$t\in Z^+$。

我考虑 p 的变动会对输出产生的影响，x 是自变量不考虑。避免混淆，这里用大写 H 表示考虑以前若干步运算的 p 对当前的影响，小写 h 表示不考虑迭代关系的具体的一步。

然后可以总结出下面的公式。a 表示 p 考虑迭代影响时，对 C 的影响。

$$\frac{\partial H_t}{\partial p}=\frac{\partial h_t}{\partial p}+\frac{\partial h_t}{\partial C_{t-1}}\alpha_{t-1}+\frac{\partial h_t}{\partial h_{t-1}}\cdot\frac{\partial H_{t-1}}{\partial p}，\quad \alpha_t=\frac{\partial C_t}{\partial p}+\frac{\partial C_t}{\partial h_{t-1}}\cdot\frac{\partial H_{t-1}}{\partial p}+\frac{\partial C_t}{\partial C_{t-1}}\alpha_{t-1}$$

t=1：

$$\frac{\partial H_1}{\partial p}=\frac{\partial h_1}{\partial p}，\quad \alpha_1=\frac{\partial C_1}{\partial p}$$

END