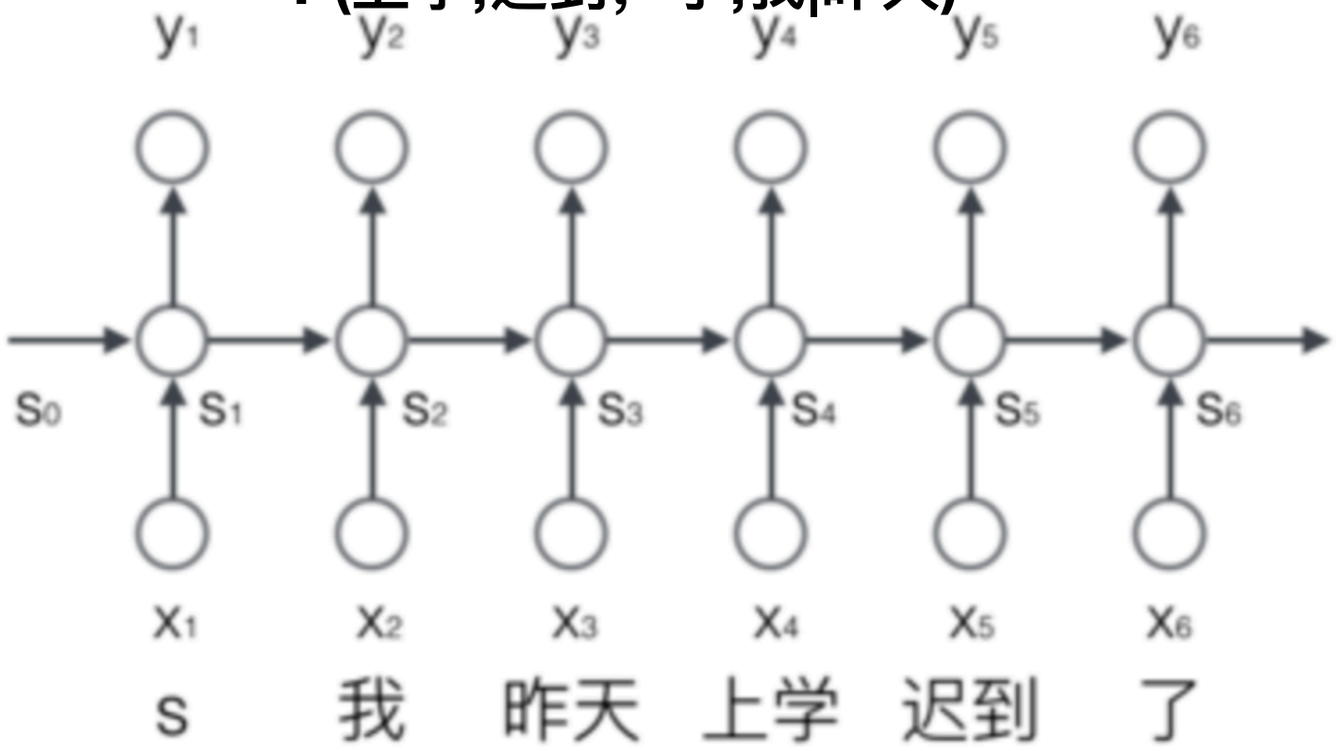
P(昨天,上学,迟到，了|我)

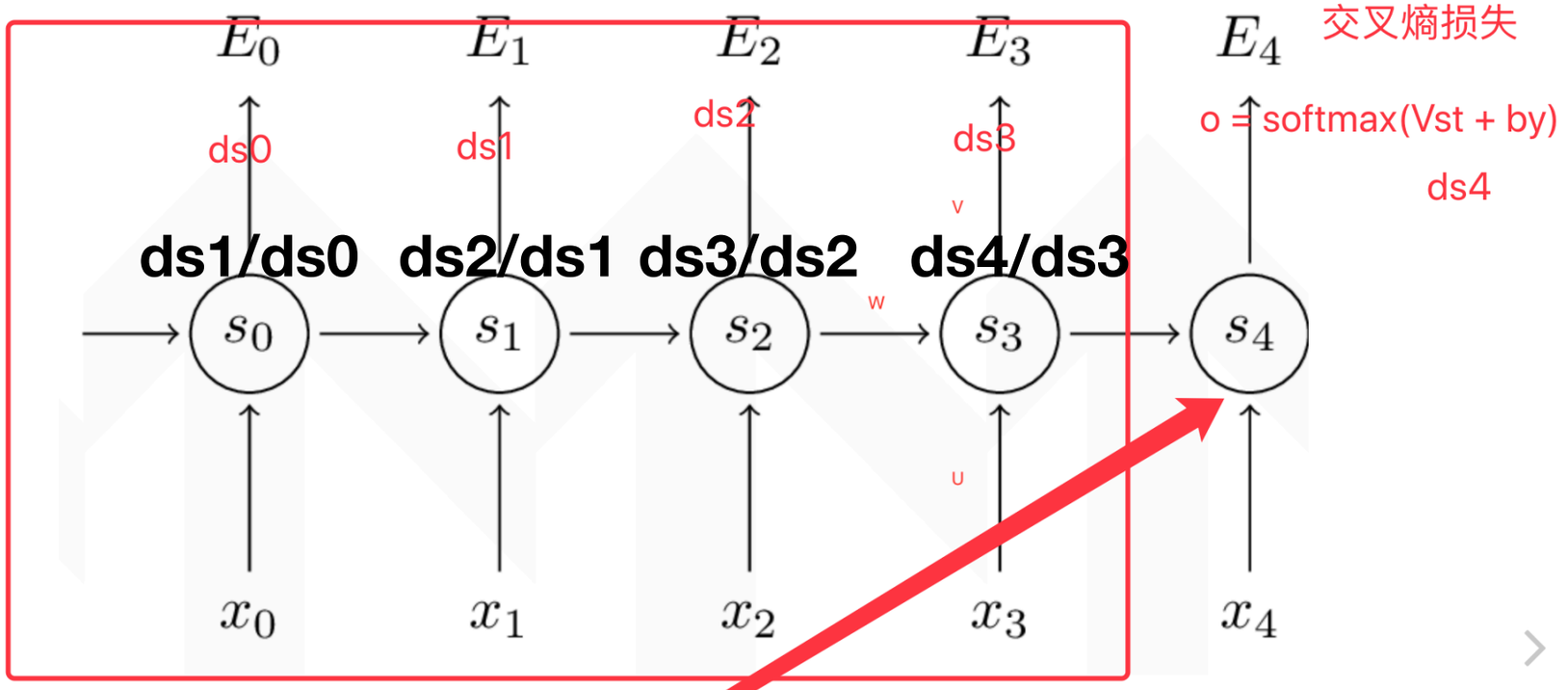
我 昨天 上学 迟到 了 e

30000所有词

P(上学,迟到，了,我|昨天)

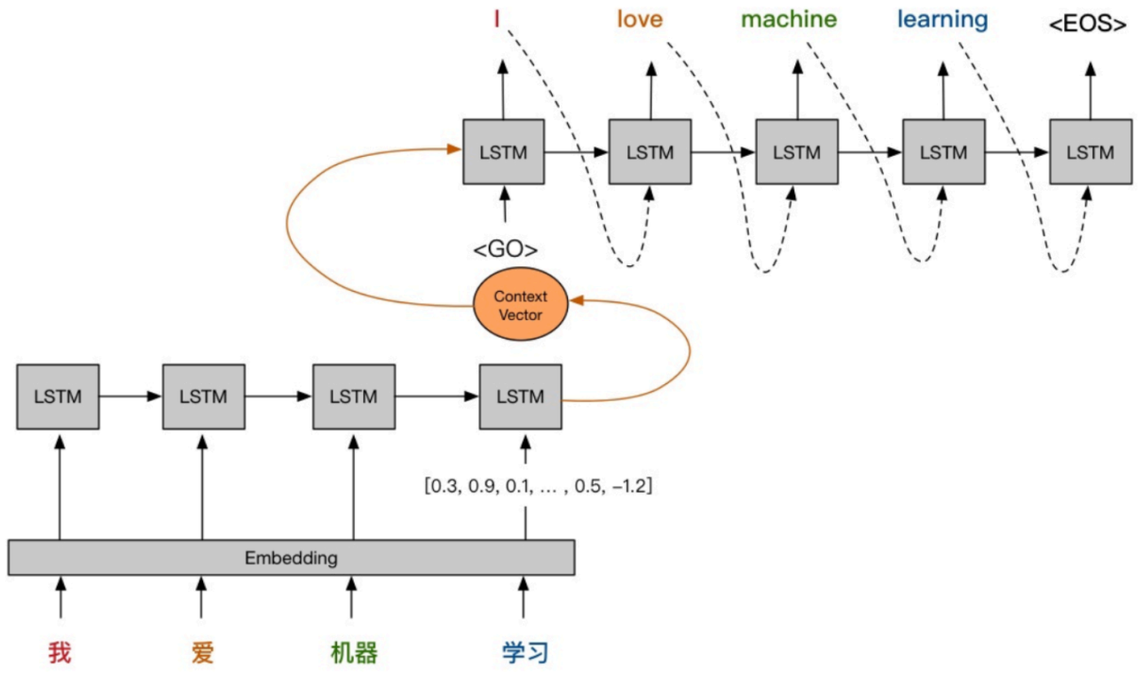


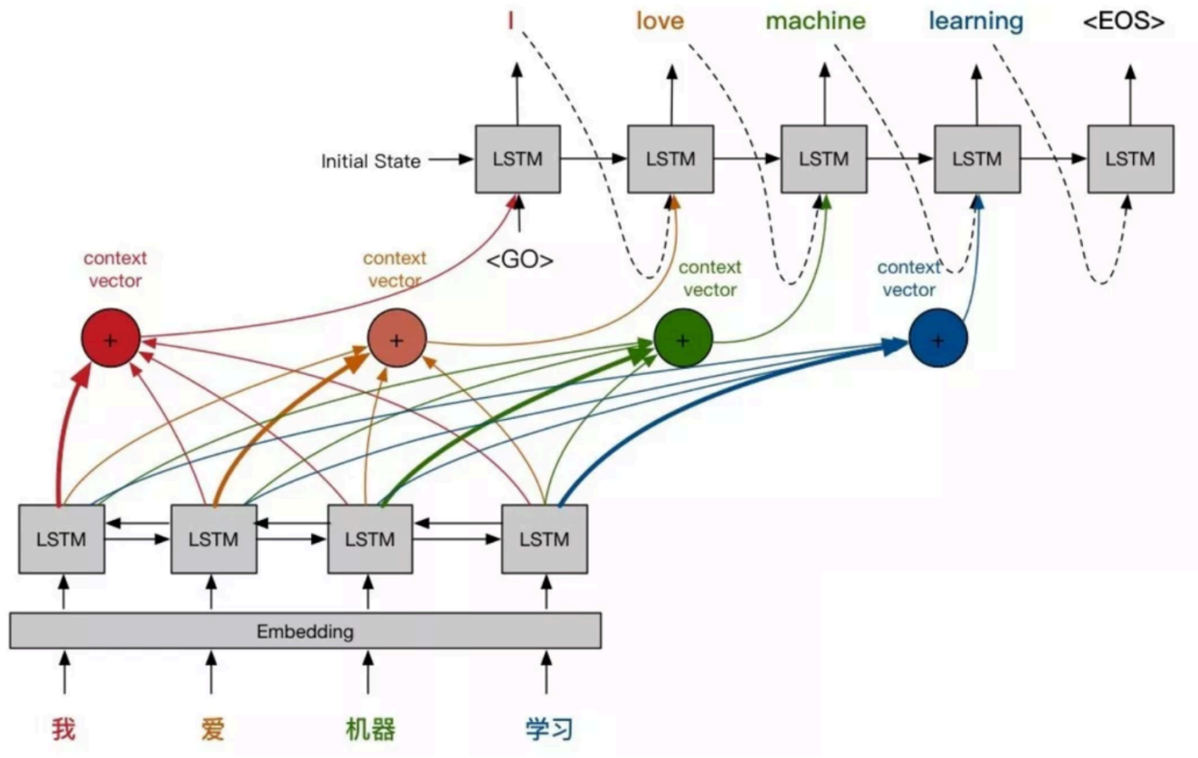
P(30000个词的概率)

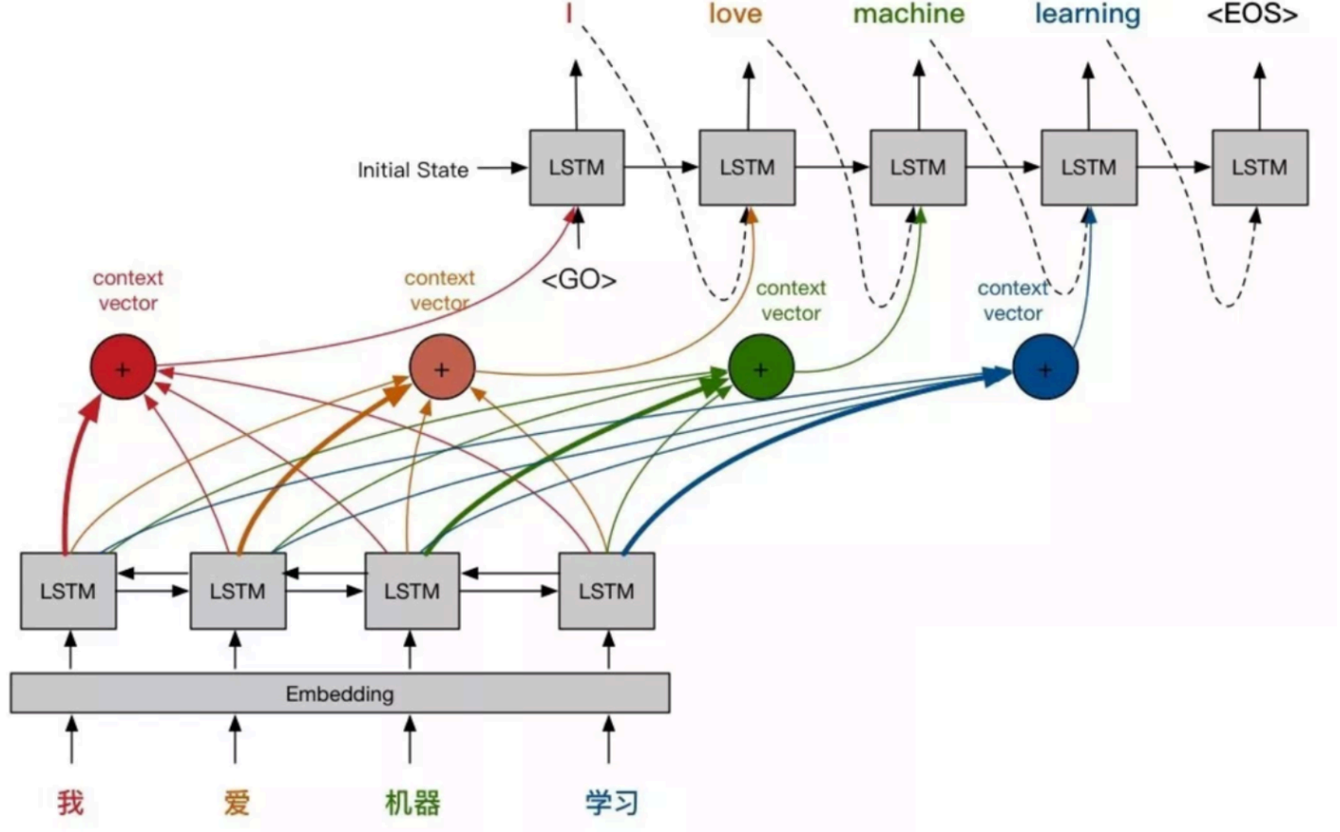


我们的目标是计算误差关于**参数U、V和W以及两个偏置bx,by的梯度**，然后使用梯度下降法学习出好的参数。由于这三组参数是共享的，我们需要将一个训练实例在每时刻的梯度相加。

- 1、**要求：每个时间的梯度都计算出来t=0,t=1,t=2,t=3,t=4，然后加起来的梯度，为每次W更新的梯度值。**
- 2、求不同参数的导数步骤：
  - 最后一个cell:
    - 计算最后一个时刻交叉熵损失对于s_t的梯度，记忆交叉熵损失对于s^t,V,by的导数
    - 按照图中顺序计算
  - 最后一个前面的cell:
    - 第一步：求出当前层损失对于当前隐层状态输出值 s^t 的梯度 + 上一层相对于s^t 的损失
    - 第二步：计算tanh激活函数的导数
    - 第三步：计算$Ux_t + Ws_{t-1} + b_a$的对于不同参数的导数


I
love
machine
learning
<EOS>
LSTM
LSTM
LSTM
LSTM
LSTM
<GO>
Context Vector
LSTM
LSTM
LSTM
LSTM
[0.3, 0.9, 0.1, … , 0.5, −1.2]
Embedding
我
爱
机器
学习


I
love
machine
learning
<EOS>
Initial State
LSTM
LSTM
LSTM
LSTM
LSTM
<GO>
context vector
context vector
context vector
context vector
+
+
+
+
LSTM
LSTM
LSTM
LSTM
Embedding
我
爱
机器
学习



注意上述的几个细节，颜色的连接深浅不一样，假设Encoder的时刻记为t,而Decoder的时刻记为$t^{'}$ 。

- 1、$c_{t'} = \sum_{t=1}^{T} \alpha_{t't} h_t$
  - $\alpha_{t't}$为参数，在网络中训练得到
  - 理解：蓝色的解码器中的cell举例子
  - $\alpha_{41} h_1 + \alpha_{42} h_2 + \alpha_{43} h_3 + \alpha_{44} h_4 = c_4$
- 2、$\alpha_{t't}$的N个权重系数由来?
  - 权重系数通过softmax计算：$\alpha_{t't} = \frac{\exp(e_{t't})}{\sum_{k=1}^{T} \exp(e_{t'k})}, \quad t = 1, \ldots, T$
  - $e_{t't} = g(s_{t'-1}, h_t) = v^{\top} \tanh(W_s s + W_h h)$
    - $e_{t't}$**是由t时刻的编码器隐层状态输出和解码器**$t^{'} - 1$**时刻的隐层状态输出计算出来的**
    - s**为解码器隐层状态输出，**h**为编码器隐层状态输出**
    - v, W_s, W_h**都是网络学习的参数**