讲解LSTM源码

one-hot

- 有多少个状态就有多少个比特，可以用来编码词向量，有N个单词，词向量就是1*N维
- 使用one-hot编码，将离散特征的取值扩展到了欧式空间，离散特征的某个取值就对应欧式空间的某个点。
- 独热编码的问题
  - 无法利用单词之间的关系

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

Bi-LSTM

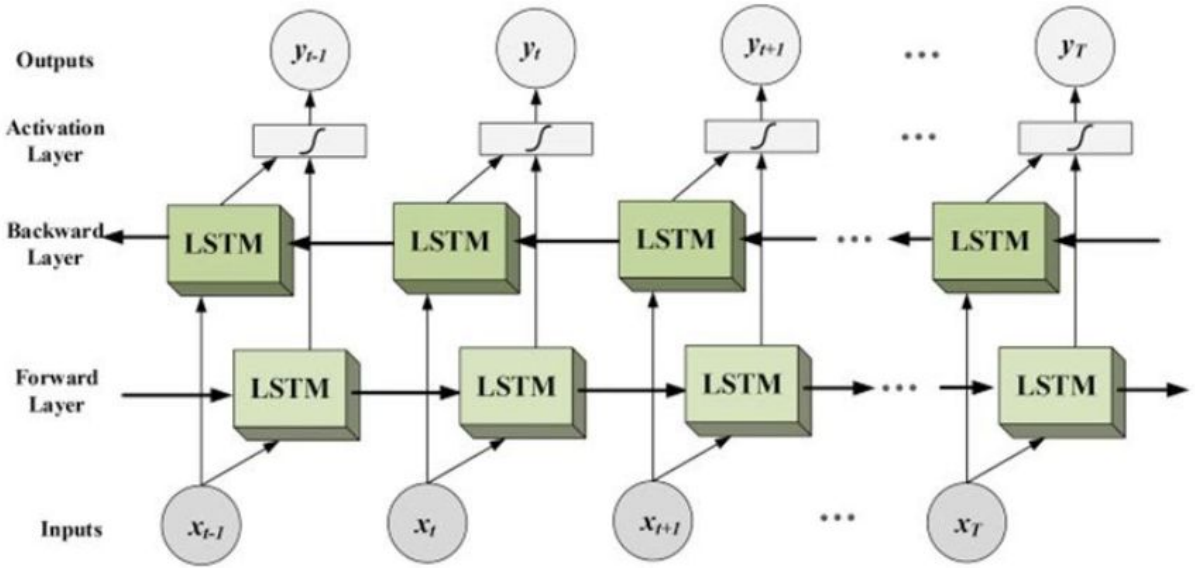
如图就是双向处理的lstm

正向，反向输出都一起放入activation layer得到结果

$$h_t = f(w_1 x_t + w_2 h_{t-1})$$

$$h'_t = f(w_3 x_t + w_5 h'_{t+1})$$

$$o_t = g(w_4 h_t + w_6 h'_t)$$



隐喻和动词的关系

- 「结构隐喻」，以一个概念去构建另一个概念。例如「争论是战争」。
- 「方位隐喻」隐喻的本体与想表达的对象（客体）有相同的特点。例如：up表达高兴，down表达伤心。
- 「本体隐喻」，将经验视作物质实体，包括「指称」、「量化」、「拟人」等。例如「**通货膨胀**降低了我们的生活水平」、「完成这本书要有**许多**耐心」、「心碎」（将心灵隐喻为易碎品）、「生活**欺骗**了我」。

代码从哪里来

- kaggle代码
- kaggle不仅有比赛，还类似百度的aistudio，有用户共享的代码与数据

导入依赖，导入数据

- nltk
- `sklearn`
- `tensorflow`
-

文本处理

- 词干提取
- 词性还原
- 删除停用词
- 分词
- 句子序列长度标准化
- 词嵌入

训练参数

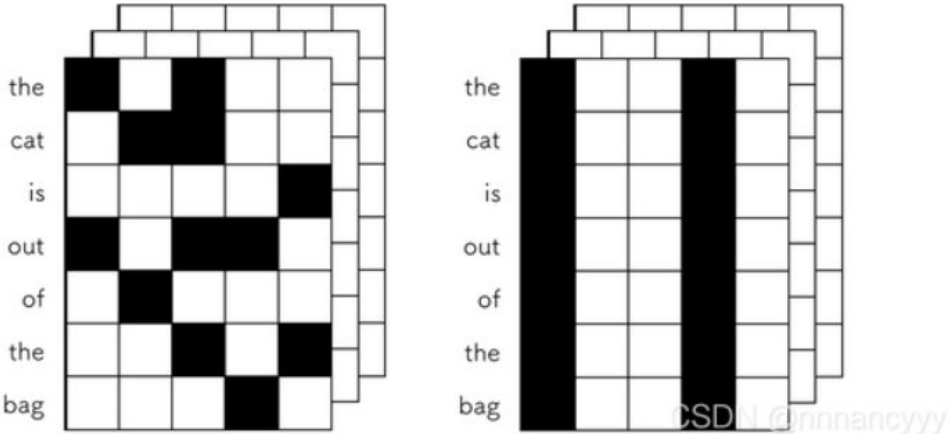
- 各个参数解释

名词	定义
Epoch	使用训练集的全部数据对模型进行一次完整训练，被称之为“**一代训练**”
Batch	使用训练集中的一小部分样本对模型权重进行一次反向传播的参数更新，这一小部分样本被称为 “**一批数据**”
Iteration	使用一个 Batch 数据对模型进行一次参数更新的过程，被称之为 “**一次训练**”

Modle training（依靠API）

- `SpatialDropout1D`

  随机丢弃某几个维度的数据
- 作用
  - 减少模型训练量
  - 每次随机丢弃一部分信息，逼着模型关注所以信息，而不是侧重一部分节点（防止过拟合）



模型评估

precision准确率 recall召回率（查全）

f1-score精确率和召回率的调和平均数

support各分类样本的数量或测试样本的总数量

macro avg (宏平均值)：所有标签结果的平均值。

weighted avg（加权平均值）：所有标签结果的加权平均值。

	precision	recall	f1-score	support
Negative	0.80	0.75	0.77	160542
Positive	0.76	0.81	0.79	159458
accuracy			0.78	320000
macro avg	0.78	0.78	0.78	320000
weighted avg	0.78	0.78	0.78	320000

总结

- 文本处理，训练模型相关操作基本都被封装起来了
- 许多模型 比如 lstm模型内部细节都没有表现出来
- 可以直接使用训练好的wordembedding