修改建议

> KingXHJ

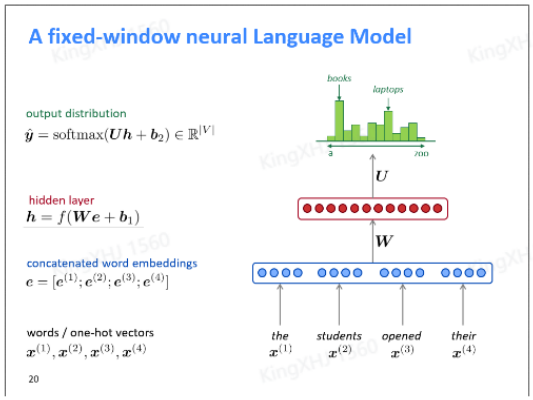
2023.12.05

1. 6.6节
   - 5-gram模型应该由前四个词元组成序列？还是您想表达的意思是，前四个词元+label一块输入呢

现在我们用FFNN去优化一个5-gram模型，即以前4个单词为输入，预测下一个词的输出。假设词嵌入是512维，词典中有3000个Token。

图源CS224N，右图展示5-gram前馈神经语言模型。

1. Input Layer：输入的句子就是前五个词元构成的序列，每一个Token都对应一个稀疏的one-hot向量。
2. Projection Layer：我们把上一层输入每个one-hot向量通过乘一个权重矩阵 W_e，映射到词嵌入 $\vec{e}$，并且把这四个词嵌入给拼起来变成一个2048维的向量。
3. Hidden Layer：输入上一层拼出来的2048维向量，以一个权重矩阵 W 做一个全连接。
4. Output Layer：是一个Softmax层，也是与上一个隐层的输出以矩阵 U 做全连接，输出一个3000维的、每个元素都为正值的、元素之和为1的向量，即下一个词的概率分布函数了。



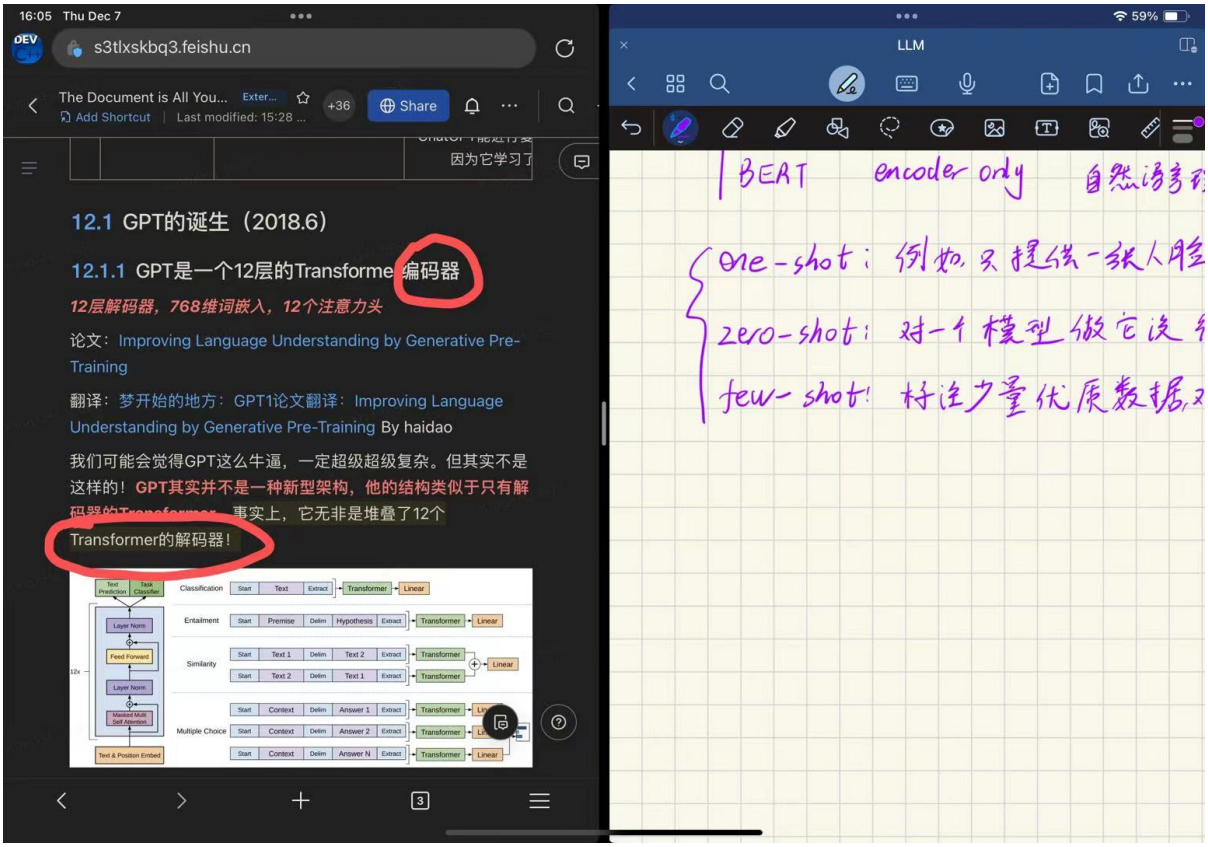
<参数集 $\theta = (E, W, b_1, U, b_2)$ >

但只用FFNN去升级n-gram模型还是太low了，并没有克服一个本质缺点：他的窗口数是固定的，每次预测新词 w_i 都只依赖前n个词。

2023.12.07

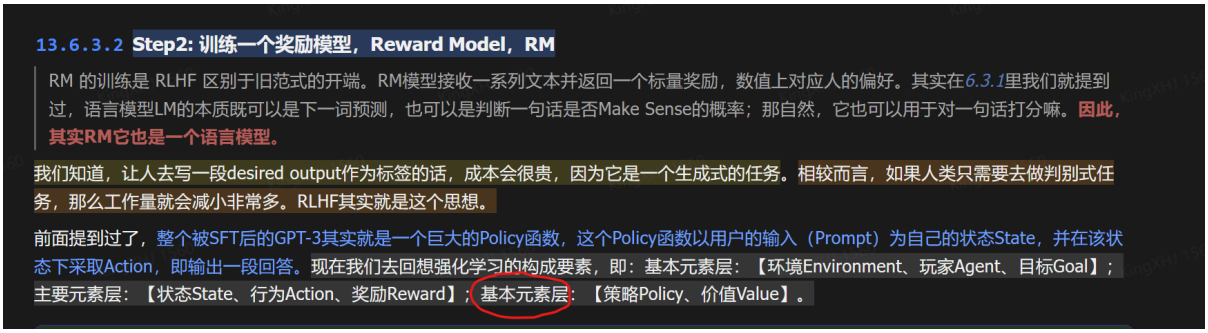
1. 12.1.1节

- 这块是不是有笔误呢？



2. 13.6.3.2节

- 这块用词应该是没对应上滴



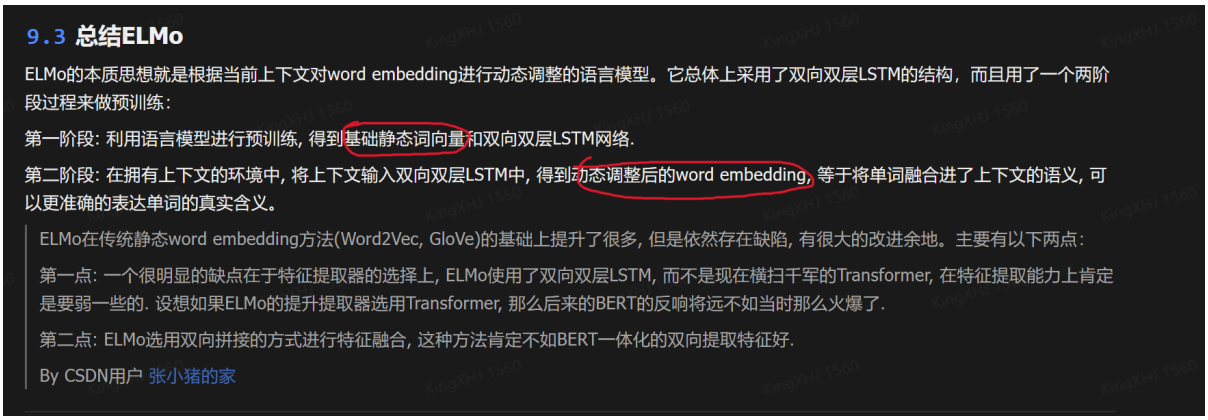
3. 15节

- 这个位置的日期写错了

2023.12.08

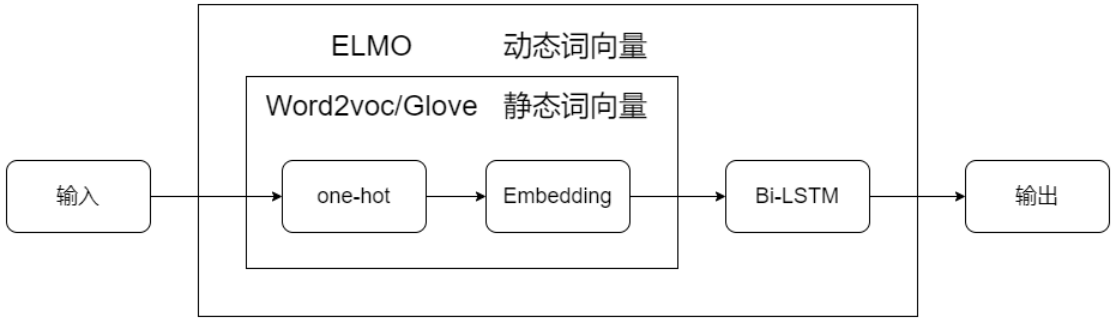
1. 9.3节

- 当我看到这里，我会疑惑一下：什么是基础静态词向量，什么是动态调整后的词向量？我感觉可能作为初学者，突然之间很难去理解ELMO作为一个Embedding预训练模型的意义。



- 我推荐在第9章章首增加一个总结性的语句：

> Word2vec和Glove通过训练后的词向量会直接变成下游任务的输入，词向量不会随着下游任务再改变，称为静态词向量；而ELMO在Word2vec或Glove训练后的静态词向量基础上，增加了Bi-LSTM(双向LSTM)模块，相当于增加了可学习参数，在新的训练中，Bi-LSTM会学习到新的参数，从而能够根据整个模型的输入，对Word2vec或Glove训练后的静态词向量进行"动态"调整，因此，从ELMO模型中输出给下游任务的词向量，是静态词向量经过Bi-LSTM调整过的动态词向量

> 配图：
>
> 
> ELMO 动态词向量
Word2voc/Glove 静态词向量
输入
one-hot
Embedding
Bi-LSTM
输出