词向量训练

1. 总体概况

由于面向NLP最主要的深度学习模型就是循环神经网络，因此本次实验考虑使用RNNLM词向量。使用python3+TensorFlow的形式进行项目构建。

1. 思路及实现流程

要构建词向量模型简单来说可以理解为，将语料库里的单词映射到一个新的向量空间，这个新的向量空间应该具有相对低维度的特点。课上介绍的RNNLM模型词向量就是将整个语料库转化成one-hot向量，之后通过look-up表的形式将原始向量嵌入到一个更低维度的向量空间。

本次实验和课上讲的RNNLM模型略有不同，标准的RNNLM的输入是one-hot向量，此处为了增加词向量可解释性，将输入改成了语料库中单词出现频率的顺序排列，即语料库中出现次数最多的被编码成0，以此类推（类似于词袋模型）。

我们可以通过Keras提供的Embedding层来实现词向量的构建。但是我们无法直接获得所有词的词向量，直接从Embedding层获取的是初始化随机的数值，需要确定一个训练任务，通过训练得到有意义的词向量。

这里考虑到可解释性和易操作性，准备构建最简单的任务：将每个句子的除最后一个单词的所有单词作为输入，模型输出预测的最后一个单词，实际的最后一个单词作为标签。选用LSTM网络构建模型，由此我们可以设计实现一个NLP的基础应用。完成模型训练之后，将模型的第一层取出，这一层就是输出词向量的嵌入层。在第一层输出的时候进行记录我们就可以获得整个的词向量空间。

1. 模型概况

使用LSTM网络构建模型；对于中英文都采用出现频率前5000个单词构建语料库；句子最大长度设置为25；嵌入层输出50维的嵌入向量；使用交叉熵损失函数；优化器使用自动地Adam。由于模型的精度与最后的实验目标关系不大，所以这里直接训练20轮结束。模型参数情况如下：

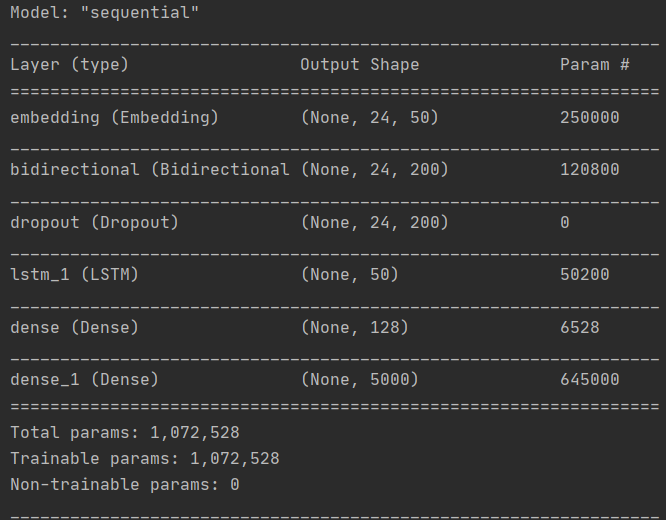


图 1模型的summary

1. 结果分析

通过训练好的模型，获得嵌入层的输出即为词向量空间，将中文和英文的任务分别存放在了对应的csv文件中。其中由于语料库设置的大小为5000，所以词向量只有5001个，剩下的单词对应的词向量全都是用<OOV>的词向量。