Report of Deep Learning for Natural Langauge Processing\_3

ZY2303109 金子棋

**Abstract**

基于Word2Vec神经语言模型，以金庸小说(共16篇)为内容的中文语料库进行词向量训练，并计算词向量之间的语义距离以验证训练效果。

**Introduction**

Word2Vec是一种用于将单词表示为连续向量空间中的密集向量的神经语言模型，通过学习单词在语料库中的上下文关系来捕捉单词之间的语义相似性。Word2Vec模型有两种主要架构：连续词袋（CBOW）和Skip-gram。在CBOW中，模型根据上下文单词预测目标单词，而在Skip-gram中，模型根据目标单词预测上下文单词。通过大规模文本数据的训练，Word2Vec可以生成高质量的词向量，使得语义相近的单词在向量空间中距离较近，从而实现了词汇的向量化表示。Word2Vec的优点包括高效的训练速度、对稀有词汇的泛化能力和对语义关系的良好捕捉，使其成为自然语言处理领域中不可或缺的工具之一。本文将采用Word2Vec神经语言模型对中文语料库进行词向量训练，计算所得词向量之间的平均语义距离以验证训练效果，并观察神经网络随Epoch变化的情况

**Methodology**

1. 文本预处理

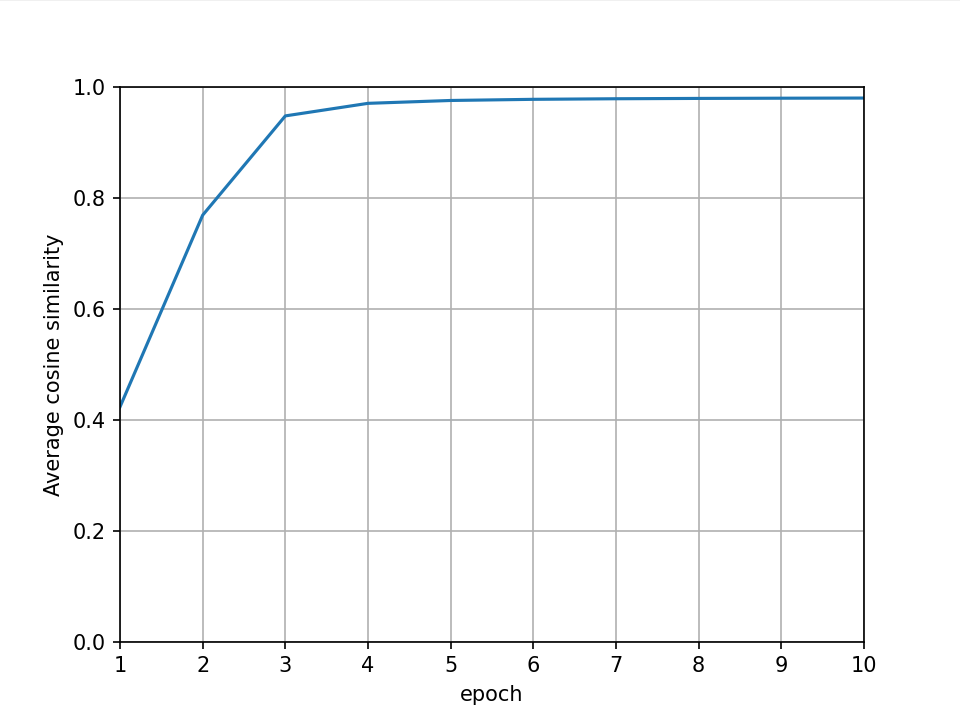
本次使用的中文语料库为金庸小说集，有效正文文本文件共16个，此外还有包含所有文件的zip包、文本文件来源网站url链接以及包含所有章节标题的文本文件inf.txt，直接在文件夹中删去。

对于正文文本，一需要删除开头结尾网站附上的广告信息，二需要去除文本中非中文字符、空格换行等无效信息，三需要去除中文文本中的标点符号，最后仅剩中文文字作为后续处理的信息源。

1. 神经语言模型训练与效果验证

得到预处理后的文本后，对共计16篇文章每篇基于jieba分词模块进行分词，并将其作为一个可迭代的语句序列使用append()方法加入训练集列表。Word2Vec模型基于gensim模块实现，直接将训练集输入即可得到词向量模型。为评估词向量训练效果，根据Word2Vec模型的目的，引入余弦相似度代表词向量之间的语义距离，公式如下：

显然余弦相似度的取值范围为。为验证整体效果，使用Word2Vec.wv.most\_similar()方法找到每个词在模型下语义最接近的词并计算所有词与其语义最接近的词的余弦相似度最后取平均作为准确率。首先在默认设置下进行训练，即取词频≥5次的词语进行训练，词向量维度为100维，窗口大小设为当前词的前2词至后2词(window = 5)。



可以看到词向量平均余弦相似度随着训练不断上升，最终稳定在0.98左右。接下来探究上述提到的不同参数对最终效果的影响，为充分训练下述结果都是epoch20情况下的。首先改变词频数(vector\_size = 100,window = 5)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 词频阈值 | 2 | 3 | 5 | 8 | 10 |
| 平均余弦相似度 | 0.8356 | 0.9160 | 0.9821 | 0.9972 | 0.9964 |

可以看到随着词频阈值的上升平均余弦相似度也在上升，实际上就是样本少了使得神经网络更加容易拟合。然后改变词向量维度(window = 5,min\_count = 3)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 5 | 10 | 20 | 50 | 80 | 100 | 200 | 500 |
| 0.9899 | 0.9654 | 0.9443 | 0.9266 | 0.9193 | 0.9160 | 0.9090 | 0.9027 |

其中第一列是词向量维度(vector\_size)，第二列是平均余弦准确率。可以看到随着词向量维度的增加余弦准确率从0.99逐渐下降至0.90，并在词向量维度达到大于100维时逐渐趋于饱和。可以看出词向量维度越大训练难度越大，神经网络更难拟合，但词向量维度比较小时准确率很高，但增大后也没有无限制下降而是趋于饱和，也说明了词向量维度过小时不太有准确意义。实际上，官方推荐合理的词向量维度在数十到数百之间，也说明了这个问题。最后改变训练窗口大小(vector\_size = 100,min\_count = 3)

|  |  |  |  |
| --- | --- | --- | --- |
| 窗口大小 | 5 | 3 | 1 |
| 平均余弦相似度 | 0.9160 | 0.9157 | 0.9134 |

可以看到窗口大小对于最后的平均余弦相似度影响不大。

**Conclusion**

Word2Vec作为常用的词向量训练模型，在vector\_size = 100,min\_count = 3,window =5的设置下最终平均余弦相似度约为0.91，可见在本次中文语料库上取得了较好的效果。同时对Word2Vec网络的不同超参数进行了调整实验，可见网络拟合程度有一定限制，对于大规模的语料库拟合程度有限，可能可以通过调节神经网络层数等方式进一步提高拟合能力。