

一、Review 词向量的改进点:

本方案的改进主要借鉴 VistaNet 模型,该模型针对词向量的降维问题,采用自注意力机制训练学习到相应的权重,然后在词维度加权相加得到句维度,从而实现降维。这个方法需要解决两个问题:

1)降维方向:原方法是在词维度方向降维的,比如给定一个数据,该数据维度为 $[m,n,k]$,其中 m 是 batch 大小, n 是单词个数,而 k 是词向量维度,在 bert 上, k 是 768,然后原方法则是将 $[m,n,k]$ 降维为 $[m,k]$ 。所以我打算在原程序的基础上,对部分程序修改,让其 $[m,n,k]$ 最后降维为 $[m,n]$ 。

2)自注意力机制的训练:原程序需要经过训练学习到合适的权重进行降维,我对此想将这部分编写一个继承 nn.Module 的类函数,其中这个权重设置为可学习的超参数。其权重可以通过最后的预测 label 和真实 label 之间的损失反向传播学习到。

我目前正在对 DA_HGNH 和 evolveGCN 这两个程序修改,目前除了词向量降维的这一类函数以外,DA_HGNN 的其他要改的程序已经修改完毕(比如更改超边构建的函数等),但是将 DA_HGNN 应用到 evolveGCN 这部分程序我还没想好怎么写。

二、数据集改进

之前的数据集我是以 1000 行(因为原始数据集的天不是连续,用 1000 个天不太严谨,容易令人误会)为一个时间戳来划分的,不过这个导致每个时间戳的数据集过于大,也导致时间戳个数比较少,不好训练,所以我直接改为每隔 100 行为一个时间戳,这样最后得到 3000 个时间戳的训练集数据。

还有,在 review 一词的词向量上,我没有将其展开为一维,仍然保持 $[100,70,768]$ 这样的三维矩阵,另起一个字典,该数据集的具体格式如图 1 所示。

```
Output exceeds the size limit. Open the full output data in a text editor
{'__header__': b'MATLAB 5.0 MAT-file Platform: nt, Created on: Fri Jun 17 09:16:42 2022',
 '__version__': '1.0',
 '__globals__': [],
 'rur': <100x100 sparse matrix of type '<class 'numpy.int32'>'
      with 968 stored elements in Compressed Sparse Column format>,
 'rpr': <100x100 sparse matrix of type '<class 'numpy.int32'>'
      with 222 stored elements in Compressed Sparse Column format>,
 'features': <100x2 sparse matrix of type '<class 'numpy.float64'>'
            with 194 stored elements in Compressed Sparse Column format>,
 'text': array([[-0.16176604, -0.10799506, -0.28463998, ..., -0.05414923,
                0.6224663 , 0.456627  ],
               [ 0.43537915, -0.9114257 , -0.02543993, ..., -0.21906473,
                0.7397299 , 0.19465987],
               [ 0.09081059, -1.5069907 , 0.06811708, ..., -0.23250109,
                0.11629332, 0.01919228],
               ...,
               [ 0.05886769, -0.66331977, 0.05971935, ..., -0.09282777,
                0.5895977 , 0.09413048],
               [-0.3042255 , -1.2610446 , -0.09593259, ..., 0.73638296,
                0.8037771 , -0.35535902],
```

图 1 具体格式