浙江大学电子信息类工程硕士专业必修课

人工智能算法与系统 （课程号：2142001）

算法实践报告

学号：22060319　　姓名：刘文顺

一. 算法实践拟解决问题

根据给出的一小段文字，判断其文章风格属于哪位作者，其中包含鲁迅、莫言、钱钟书、王小波、张爱玲五位作者。

二. 描述所实现算法内容

1、相关算法

jieba库将文字切分为词汇，按词汇顺序构建词汇表，在神经网络最前面插入Embedding层构建词向量，Long Short-term Memory（LSTM）结构分析词向量，前馈神经网络分析并识别文风作者。

jieba库是做分词和词分析非常成熟的一个库，能够大大提高数据处理的准确性。Embedding层生成的随机词向量，能够直接将整数映射为一个向量，因此词汇表课直接使用整数编码，减小模型的复杂度，且能够让词向量也在模型的训练中学习训练。LSTM是NLP中表现非常好的结构，它能够接受一个序列的词向量，并且在训练中存在着记忆性，能够用来分析词汇在其前后文中的信息。且LSTM能够在长序列训练过程中RNN所带来的梯度消失和梯度爆炸问题。因此用LSTM分析并获取文段的信息。获取其段落信息之后，使用简单的前馈神经网络，分析其文风信息。

2、算法概貌

首先读入数据，将每段读入的数据和其作者打包成一个元组（文段，作者），只用jieba库将文段进行分词。将词汇进行编码，这里使用的是整数编码，即用一个整数index代表一个词汇。建立神经网络，第一层为词向量的Embedding层，第二层为LSTM结构层，后面两层为前馈神经网络。在神经网络中加入Embeding层，能够将整数映射称为一个向量，在模型的学习的同时能够让向量也得到了训练。在经过Embedding层后，获得的是一个文段词汇的词向量。将文段的时序词向量传入LTSM结构，取得文段信息（向量）。将文段信息向量输入简单的前馈神经网络，使用交叉熵损失函数和Adam梯度下降优化算法，对模型进行训练。流程图见图2.1。

图2.1 算法流程图

3、模块描述

文章处理。将文件夹中的文章按行读入程序，每行打包成一个元祖（文段，作者），将5个作者映射为0-4整数型。使用jieba库，将每个文段切分为词汇数组。同时用整数编码的方式，生成词汇表。即一个整型数字代表一个词汇。本次训练的数据集中大概有8000多个词汇。

算法模型。使用embedding层，将整数计算生成32维词向量。将存在时序的词向量按顺序输入到LSTM层中，LSTM层输入维度为32维，输出维度为16维，即16个LSTM单元，层数为1层。经过LSTM层分析文段信息，获取到16维向量。最后将16维向量输入到全连接层，最后一层输出5个节点。见图2.2。

图2.2 算法模型详情

4、算法实验

数据构成。初始数据有5个txt文件构成，每个文件中存放着作者的文章。将文件内容以行形式读入程序，打包为元组（文段，作者），使用jieba库将文段进行分词，生成词汇表。

算法优化训练方法。损失函数使用交叉熵损失函数。梯度下降优化算法Adam梯度下降优化算法。

算法结果呈现和分析。

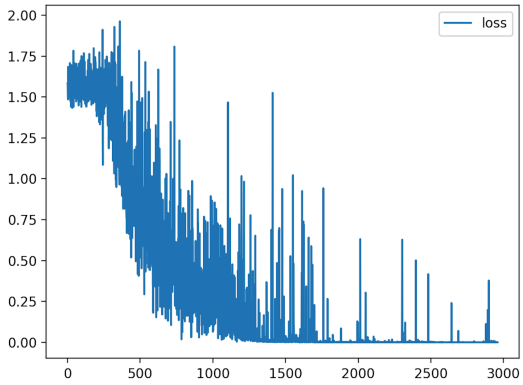

图2.3 损失函数值

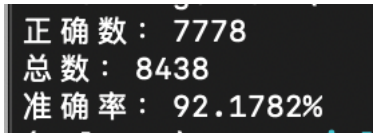

图2.4 测试结果

三. 算法所涉及内容的未来趋势和挑战分析

LSTM结构是特殊的RNN，它能够解决有效的分析时序数据并解决RNN中梯度消失和梯度爆炸的问题。这很适合分析和时序相关的信息。embedding层能够从简单的词汇表生成词向量，能够NLP中发挥很大的作用，但是现实世界中词汇量很大，输入维度就会非常大，本实验中的词汇量也是有限的，因此模型的应用也是有限制的。要使用适用性更为广泛的模型，词向量层和LSTM层的挑战会更大。

提交形式：PDF

提交日期：2020/11/24