韵文散文文本分类任务

任务介绍

任务简介

本实验的主要任务为对古文中的韵文和散文进行分类。

本实验使用多种不同的中文预训练词向量将古文文本转换为词向量，然后使用不同的分类器进行二分类任务

本次实验基于Bert-Chinese-Text-Classfication-Pytorch库进行扩展

任务流程

- 收集数据

- 数据预处理
- 模型训练
- 结果分析

数据处理

收集数据

本次实验的语料包含韵文语料和散文语料

- 韵文语料来源于古诗文网，https://github.com/Werneror/Poetry。
  - 包含历朝历代的诗词曲赋等，共 853385条
  - 从各朝代语料中按照8 - 1 - 1的比例随机抽取训练集8万条，验证集1万条，测试集1万条
- 散文语料来源于 https://disk.pku.edu.cn443/link/94FA2F534E12B25FDD7B1F7F4B8F37F3
  - 包括传记、四书、正史等散文，共220余万条
  - 从中抽取10万条，按照8 - 1 - 1的比例随机抽取训练集8万条，验证集1万条，测试集1万条

数据预处理

数据预处理的代码在 preprocess.py中

预处理过程

- 去除多余的html内容、分隔符和空格
- 去除逗号和句号外所有的特殊符号，让散文和韵文保持统一
- 基于opencc库将所有的繁体字转化为简体字

实验方法

预训练模型选择

所有的预训练模型都使用Base版本，预训练语言模型的地址：
https://github.com/ymcui/Chinese-BERT-wwm

预训练模型	模型特点
BERT	使用google基于中文维基内容的预训练模型，以字为单位训练
ERNIE	在Bert的基础上将MASK分为token、word和phase级
BERT-wwm	Bert的升级版本，引入全词MASK
ROBERTA-wwm	Bert的基础上训练序列更长，引入动态调整Masking

分类器模型

分类器	特点
MLP	Embedding层后直接使用全连接层进行分类
RNN	LSTM代替全连接层进行预测，保留全局记忆
CNN	用三层卷积核的卷积结果堆叠接全连接层 代替基本的全连接层，捕捉更多局部信息

实验模型

基于上面的四种预训练模型和三种分类器模型，共计生成了12中不同的分类模型，其中的baseline为BERT + MLP模型

模型的实现位于 model 文件夹中

实验过程

实验数据准备

基于上文的实验数据提取，在保持测试集不变的情况下，将训练集和验证集等比例缩小进行进一步的实验

对应的四种训练集的大小为2000，4000，10000，40000

生成训练集的代码位于csv_data文件夹中

实验过程

在上文所述的12个模型上在四个训练集上使用3个随机数种子进行3次实验取平均值

实验的脚本见 experiment.py，具体的实验过程见 run.py

实验结果

以准确率Accuracy为指标得到实验的结果如下：

分类器模型	预训练模型	**2000**条训练集	**4000**条训练集	**10000**条训练集	**40000**条训练集
MLP	BERT	50.89%	66.89%	98.81%	99.68%
	ERNIE	52.25%	99.10%	99.53%	99.76%
	WWM	51.33%	68.41%	99.73%	99.84%
	ROBERTA	58.41%	98.17%	99.77%	99.91%
RNN	BERT	62.84%	68.7%	98.84%	98.94%
	ERNIE	65.96%	99.13%	99.4%	99.76%
	WWM	66.90%	69.82%	99.78%	99.62%
	ROBERTA	69.82%	99.57%	99.76%	99.84%
CNN	BERT	45.31%	61.28%	99.56%	99.7%
	ERNIE	50.04%	98.04%	99.61%	99.77%
	WWM	46.98%	63.98%	98.84%	99.12%
	ROBERTA	52.48%	97.08%	99.75%	98.91%

实验结论

根据上述实验结果得到以下基本结论：

- 整体上，在数据量较大的场景（训练集数量 >= 10000）下每个分类器 + 预训练模型的组合在简单的二分类任务上都能取得比较好的效果，因为训练的epoch较少，模型过拟合的概率不大。
- 模型之间的差别主要体现在数据量较小的场景。

- 分类器模型相同的情况下对于预训练模型进行比较可以得出，在准确性上BERT是最差的。ROBERTA是最好的，ERNIE的效果也非常的不错，仅仅在4000条训练集的情况下就取得了很好的效果（可能过拟合）。原因可能因为，ROBERTA和ERINIE均是在BERT的基础上进行了知识层面的改进，更好的提取了语义知识信息
- 预训练模型相同的情况下对于分类器模型进行比较。首先是CNN主要提取的是局部的特征信息，而RNN主要是对于整体特征信息的把握。整体看来，RNN > MLP > CNN，词向量整体代表了文本的语义知识信息，所以对于整体信息的提取更有利于进行二分类。

实验中遇到的问题

- 开源框架使用的问题
  - 解决方法：读README和代码
- 在数据量较小的情况下，根据混淆矩阵，出现了模型将99%以上的测试文本分为某一类的情况
  - 解决方法：训练的时候引入随机种子，在训练集生成的时候保证散文和韵文数量基本一致的情况下进行shuffle。模型的参数的初始化，numpy向量的初始化都使用固定的随机种子。使用不同的随机种子进行多次实验，综合实验结果作为整体的结果

整体结论

- 在数据量较大的情况下，可以使用最基本的baseline完成任务
- 在数据量较小的情况下，可以使用效果较好的预训练模型，根据任务侧重于文本的局部特征还是整体特征，在Embedding层之后基于CNN或者RNN进行合适的扩展