LSTM自动写诗

1 概述：

任务：使用LSTM完成自动写诗任务

数据集：[chinese-poetry](https://github.com/chinese-poetry/chinese-poetry) 数据集（https://github.com/chinese-poetry/chinese-poetry）

数据集介绍：[chinese-poetry](https://github.com/chinese-poetry/chinese-poetry)是一个中华古诗词数据库，其中收集了唐诗、宋词等几十万首。

解决方案：将[chinese-poetry](https://github.com/chinese-poetry/chinese-poetry)中的唐诗部分提取出来，筛选出其中的五言诗与7言诗。以7言诗为例，使用Word2Vec对7言诗训练词向量。取得相应的词表并将词向量放入LSTM模型进行训练。

2 解决方案

2.1 训练相关库

pytorch, numpy, json, gensim

2.2 数据集提取

数据集的文件是json格式，这里要将所有古诗提取出并判断其是否为5言古诗或7言古诗，将不符合的古诗去除。



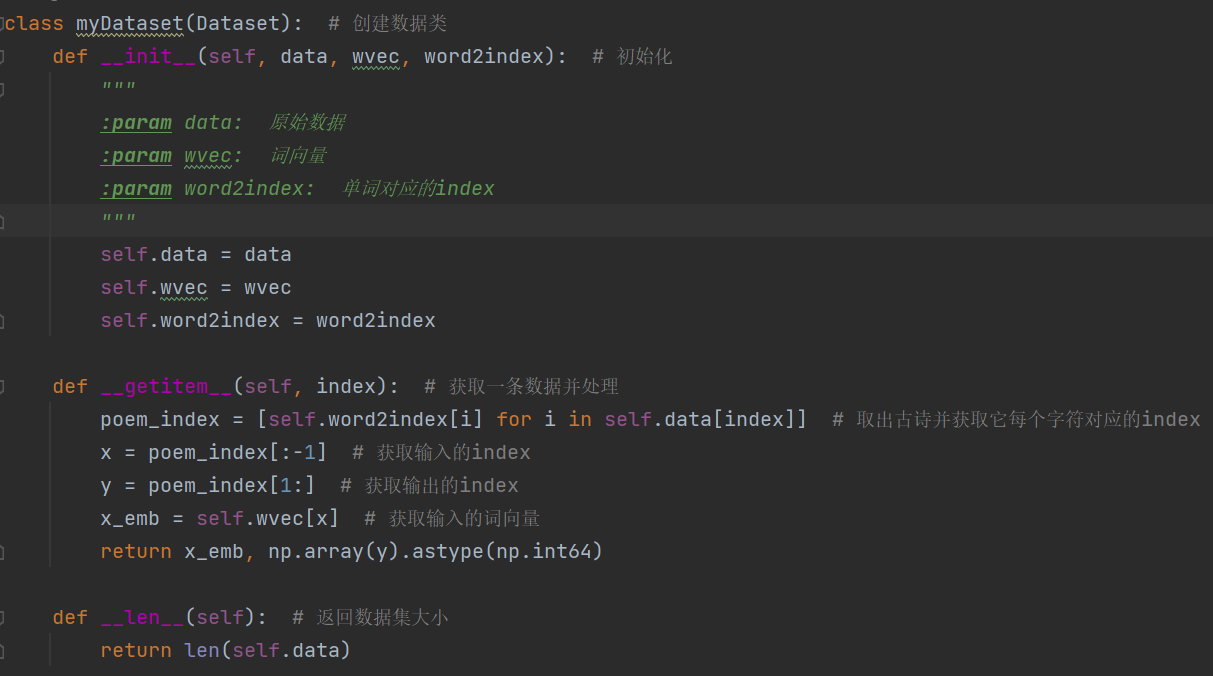
2.3 使用word2vec训练词向量

2.2中已经提取了古诗，这里训练古诗的词向量模型。词向量的维度设置成100，字符最小出现次数1，即表示所有的字符都进行统计，线程数设为6。



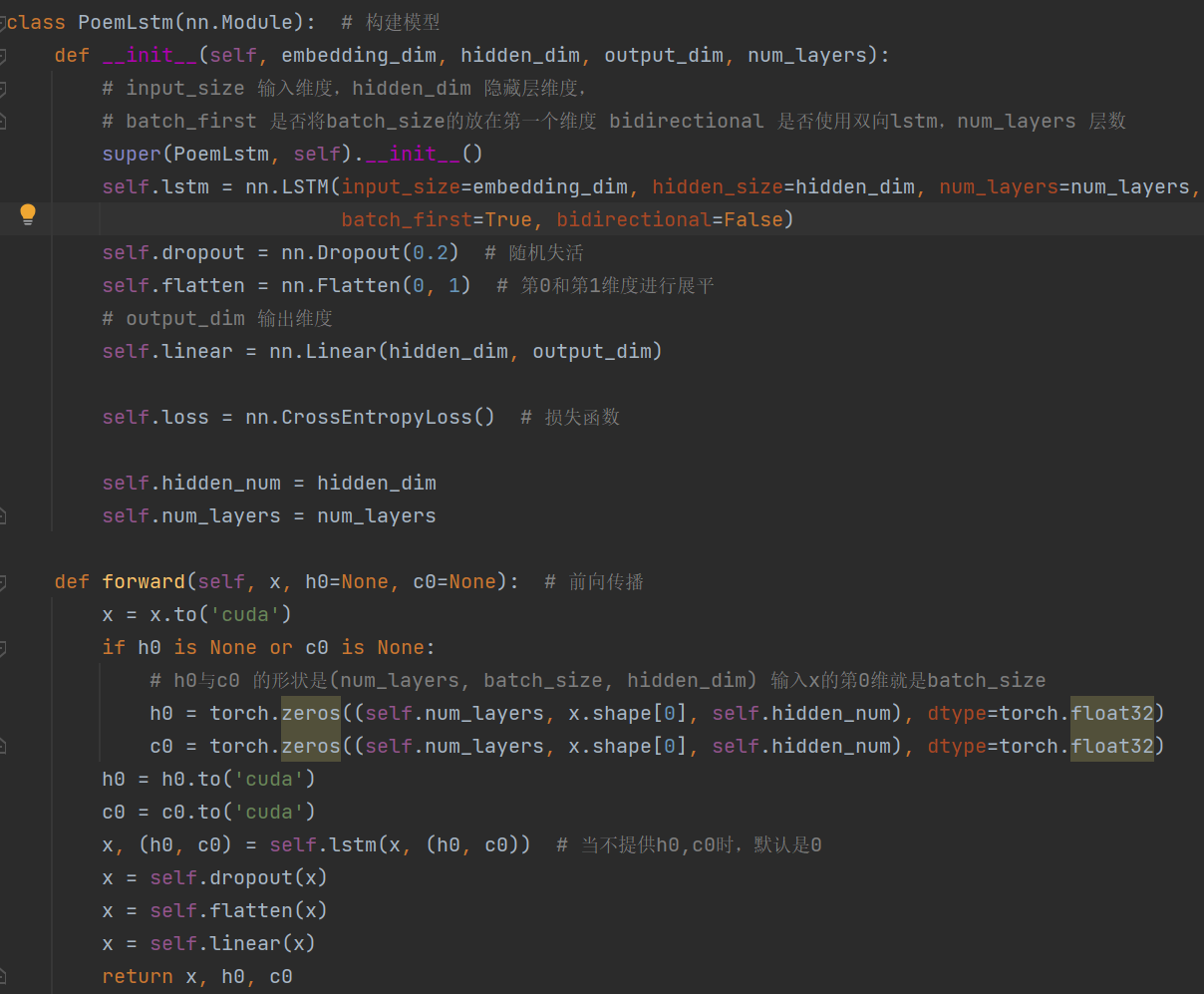
2.4 数据集类

数据集类，获取源数据与目标数据，并将源数据转换为词向量返回出来。



2.5 LSTM网络

建立LSTM网络，h0与c0为LSTM网络的参数，如果未给出具体值，则将其置位全0的矩阵。



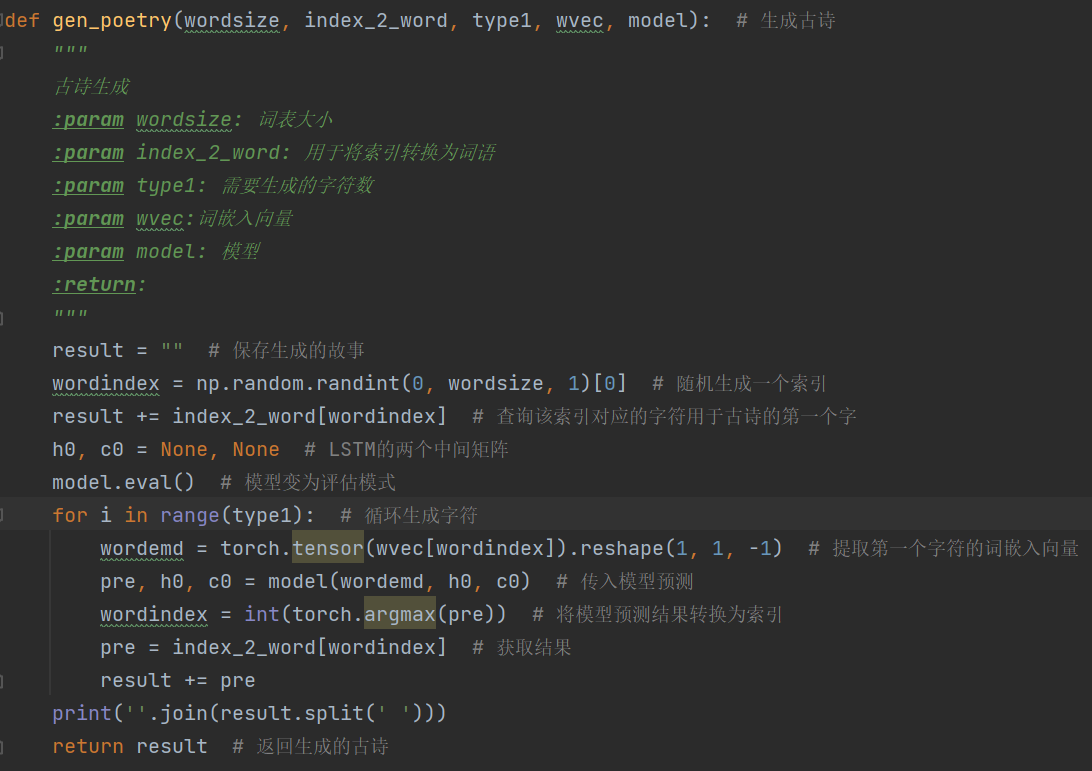
2.6 古诗生成函数

训练好模型后，这里使用训练好的模型进行古诗生成。

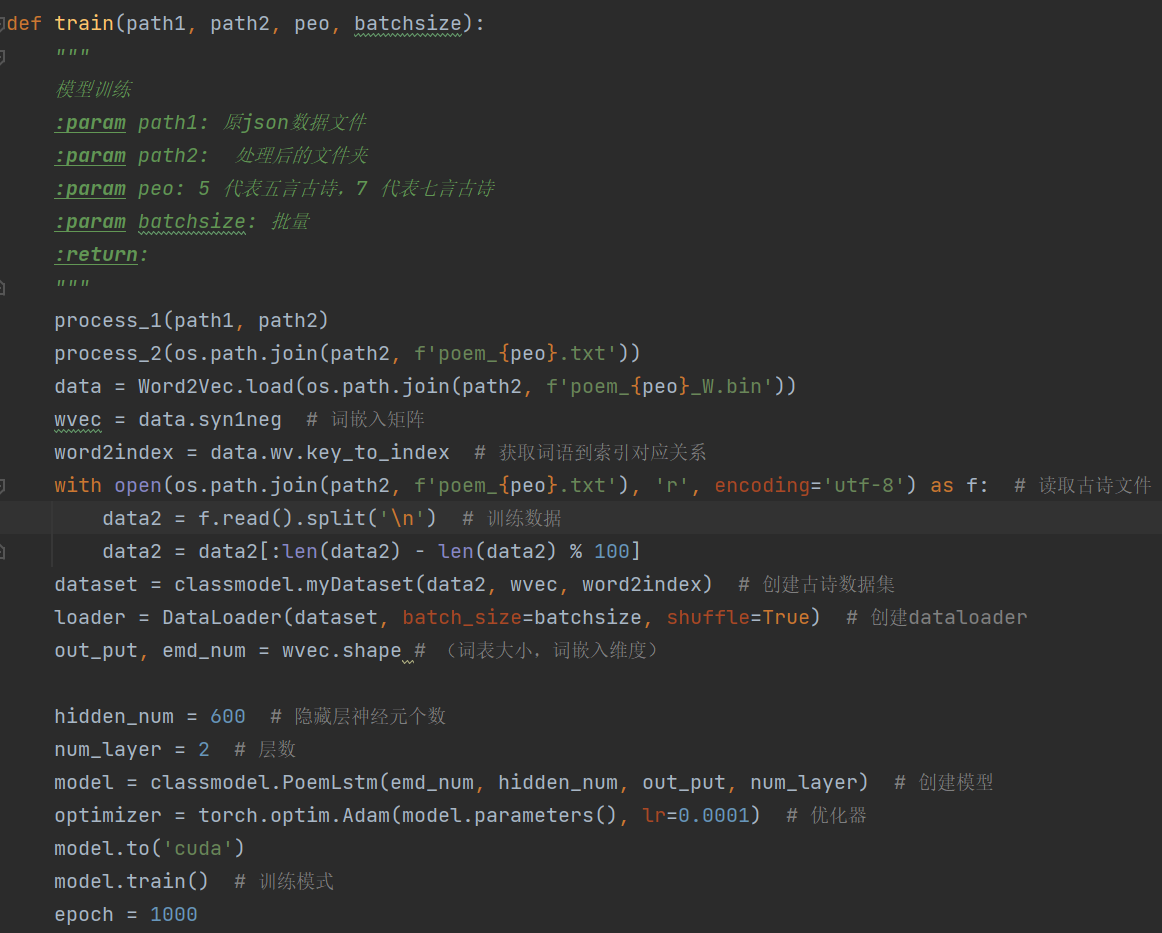
首先随机生成一个值，作为古诗的第一个字。

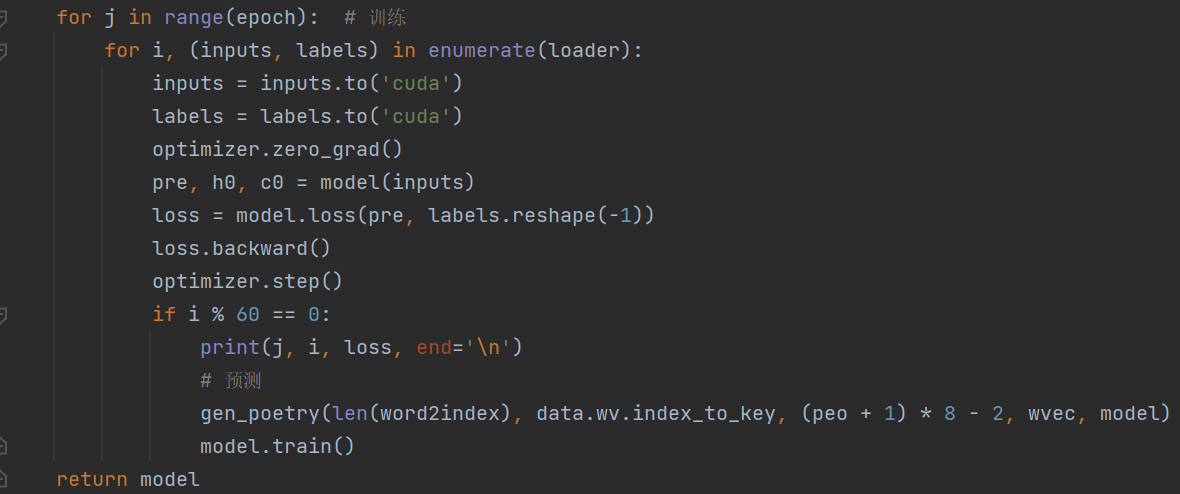
将这个值转换为词向量并传入模型进行预测，并获得预测值，h0,c0。

将获得预测值，h0,c0作为新一轮的输入放入模型进行预测。



2.7 训练函数





2.8 整体网络结果介绍

LSTM模型：层数 2

输入维度 100

输出维度 词表长度

隐藏层神经元个数 600

LSTM层与线性层之间随机失活比例 20%

损失函数：交叉熵

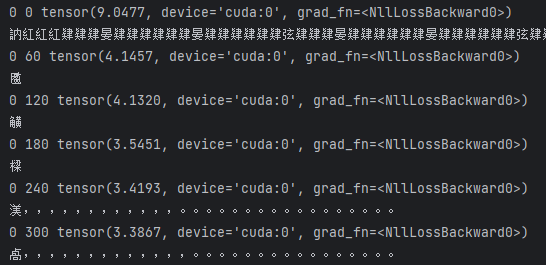
优化器：Adam

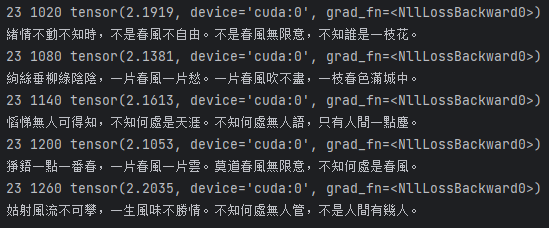
学习率：0.0003

训练轮次：24

3 结果

随着模型的不断优化，能够实现自动写诗。





Loss变化如下：

这里模型并未完全收敛，不过训练比较费时间，于是就停止了训练。

