自动写诗

杨安东

目录

摘要

自然语言处理是教会机器如何去处理或者读懂人类语言的系统，是目前较为热门的研究方向。让神经网络自己生成诗句具有很大的应用潜力，例如为杂志，新闻稿生成素材等。本实验的目的是在唐诗数据集上训练一个 RNN 网络，通过给定一个字或词或一句话，由网络自动书写之后的内容。网络内容应当与给定的字、词或句子具有相似的风格，同时要满足唐诗的一些格式规定。最好可以写出具有连贯的主旨与意境的完整律诗。

1 问题描述

在预处理后的唐诗数据集上进行训练，实现一个可以写出较为工整同时具有一定意境的诗句。

2 解决方案

在查阅资料与课程网站的指导后，整体方案分为三部分：数据获取，网络训练，生成诗句。接下来将对各部分进行详细的介绍，本方案基于 Pytorch 编写。

2.1 数据获取

由于数据只有一首一首诗，并没有分好的数据，也没有标签。因此需要自己编写数据集读取过程。pytorch 数据读取主要使用 torch.utils.data.DataLoader，其需要 Dataset 类型作为参数。torch.utils.data.Dataset 是抽象类，可以通过继承 Dataset 类并重写 ___len___ 与 ___getitem___ 方法实现自定义的数据读取。

在实验中发现，数据集中有过多的空格，这对训练会有较大的影响，如果网络直接拟合空格，会出现诗句中有很多空的的情况，也会导致初期收敛很快的假象。

同时在实验中发现，如果按照一般唐诗的取训练单个样本长度为 40 个字符，理论上可以覆盖一首诗的长度，但是实际运行中发现基本上只写了一句就停了。因此在训练集制作时需要把每个样本的字符长度变大，本文使用 96 个字符。因此本实验的数据集结构为：从原始数据集中每 96 个字符作为一个训练样本，其标签为依次向后推移一位获得的 96 个字符。

2.2 网络结构设计

在查阅资料后了解到可以用于写诗的运行对硬件的要求不高的 RNN 网络有：CharRNN 是一个简单但是有效的 RNN 网络，其可以从海量文本中学习文字的组合规律，并能够自动生成相对应的文本。例如作者用莎士比亚的剧集训练 CharRNN，最后得到一个能够模仿莎士比亚写剧的程序。它的作者 Andrej Karpathy 现任特斯拉 AI 主管。

本文方法在 CharRNN 基础上进行了改进，为了使诗可以很快写的很长，添加了两层全联接层，同时为了训练更快调整了长短时网络的参数，只使用一个长短时间网络。

最终网络结构如图 1所示，激活函数全部使用 tanh 函数。

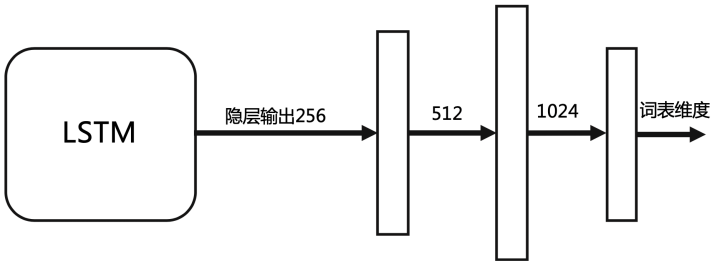

图1 本文方法网络结构

2.3 损失函数设计

损失函数没有很多选择，使用较为常用的交叉熵损失函数。

2.4 超参设计

lr 使用了 pytorch 提供的 lr_schedule 函数实现动态变化的 lr。其可以根据 epoch 动态调节学习率，在本文介绍的方法中，其定义如下：lr_scheduler = torch.optim.lr_scheduler.StepLR(optimizer, step_size=10, gamma=0.8)

3 实验分析

3.1 数据集介绍

实验提供预处理过的数据集，含有 57580 首唐诗，每首诗限定在 125 词，不足 125 词的以 </s> 填充。数据集以 npz 文件形式保存，包含三个部分：

1. data: 诗词数据，将诗词中的字转化为其在字典中的序号表示。
2. ix2word: 序号到字的映射。
3. word2ix: 字到序号的映射。

本文介绍方法通过实现 torch.utils.data.Dataset 抽象类来实现数据读取。

3.2 实验结果与分析

给予初始诗句“雨余芳草净沙尘”，在训练不同阶段获得的结果如下：

1. epoch0: 雨余芳草净沙尘。。出。。出。出。。出。。出。出。。出。。出。出。。出。。出。出。。出。。出。出。出。
2. epoch1: 雨余芳草净沙尘，一片云根一片心。一叶一竿红叶下，一枝红杏一枝梅。
3. epoch2: 雨余芳草净沙尘，花落花时满地春。莫道相逢相见尽，不知何处是非关。
4. epoch3: 雨余芳草净沙尘，风雨萧条古木深。不是人间无事事，不知何处是人家。
5. epoch4: 雨余芳草净沙尘，风雨萧萧落叶稀。不是故乡归未得，一年长在故乡中。
6. epoch5: 雨余芳草净沙尘，春色新妆满眼明。
7. epoch6: 雨余芳草净沙尘，独自闲吟不可闻。
8. epoch7: 雨余芳草净沙尘，独向江南旧隐沦。

9. epoch8: 雨余芳草净沙尘，花落花枝不是春。
10. epoch9: 雨余芳草净沙尘，风送春风满眼前。今日相看犹未得，不知何处更相亲。
11. epoch10: 雨余芳草净沙尘，风起江南草木春。惆怅故人千里别，故人千载在烟霞。
12. epoch11: 雨余芳草净沙尘，山色青山白日斜。不见故人无限意，不知春色不堪愁。

可见在训练初期，写出的诗句并没有什么规律，字数，标点都不一定正确。但随着训练的进行，写出的诗句越来越符合规范，并且渐渐出现了风格，例如 epoch4 训练结束后的句子已经出现了类似思乡的主旨，使用的意向如落叶，风雨都可以表现漂泊的艰难。