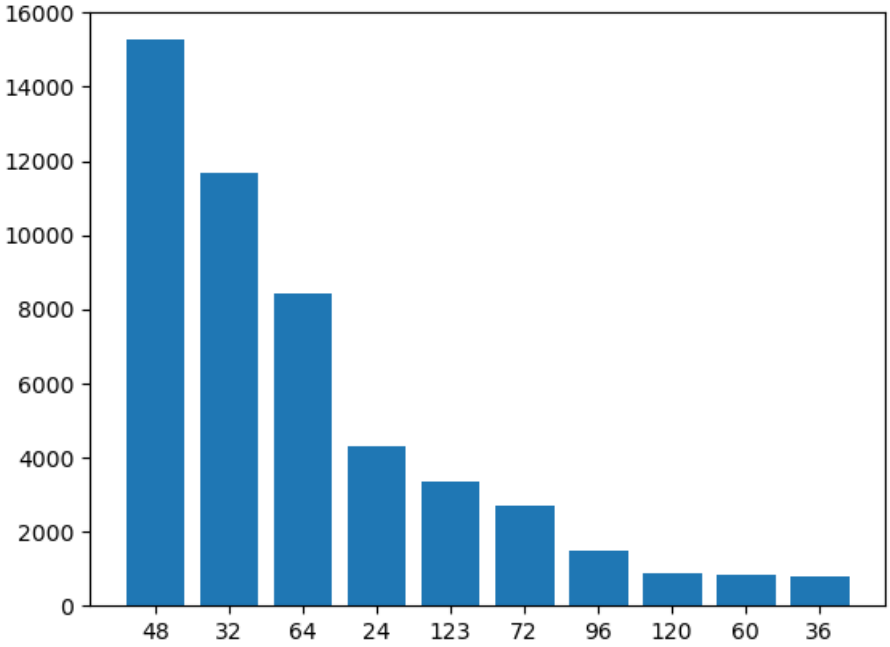
自动写诗实验报告

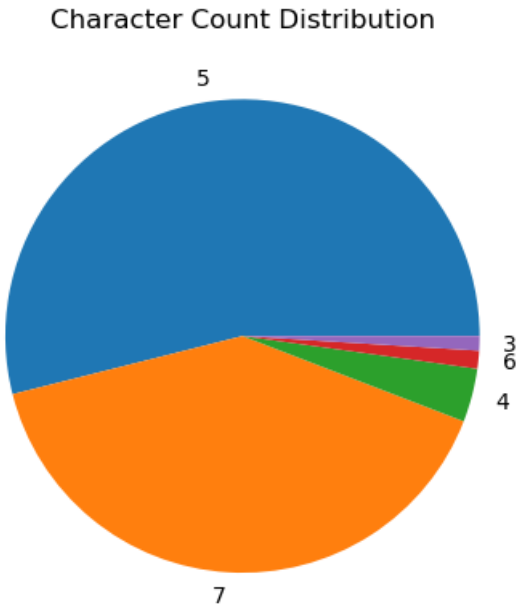
本实验使用 PyTorch 实现了使用长短期记忆网络自动生成诗歌的任务。

数据集

数据集中一共有 57580 首诗，其中大部分诗歌的长度是 48 个字（15267 首）、 32 个字（11661 首）和 64 个字（8406 首）（包含标点）。字数的分布如下：



其中大部分诗歌是五言（30266 首）和七言（22543 首），分布如下：



包含各种标记在内，词典的大小为 8293。

实验环境

软件环境：Python 3.8.13，PyTorch 1.13.1。

硬件环境：Intel(R) Xeon(R) Gold 6136 CPU @ 3.00GHz，Tesla P100-PCIE-16GB

预处理

原数据集的大小为 57580×125，考虑到其中大部分都是填充长度的 `</s>`，如果直接输入模型，学习效率很低。因此对其进行如下预处理：

- 将数据集中所有诗歌按顺序全部拼接在一起，形成一个一维向量；
- 移除向量中用于填充长度的 `</s>`；
- 从第一个字开始，固定每 48 个字合并为一首诗，从而重新构成了 65260 首诗，作为模型的输入。
- 从第二个字开始，固定每 48 个字合并为一首诗，从而重新构成了 65260 首诗，作为模型的输出。

这样，模型就会在训练过程中逐渐学会根据之前的信息预测下一个字的概率。

实验设置

使用长短期记忆网络结构进行生成：

- 嵌入层（Embedding）：用于生成 128 维词向量
- 长短期记忆网络层（LSTM），一共 3 层，隐藏层为 512 维
- 线性层：一共两个隐藏层，分别有 1024 和 2048 个节点，激活函数为 ReLU，输出维度为 8293（词典的大小）

超参数设置：

- 初始学习率：10^{-3}。
- 优化算法：Adam
- 批量大小：256
- 迭代次数：50
- 损失函数：交叉熵损失函数
- Dropout rate：0.1

生成规则

完全依靠模型计算的最大概率生成诗歌会有很多问题，例如不遵守字数格式、重复字过多等，也无法控制其每句诗歌的字数。因此，根据模型的输出，依靠下列规则来逐字生成诗歌：

1. 由用户设定好诗歌长度、每句诗的字数、诗的开头、藏头等限制；
2. 开始编写诗歌，模型的第一个输入是用于开头的 `START`；
3. 模型根据输入，计算下一个字的概率分布；
4. 检查是否需要按照用户给定的开头或者藏头写诗，如果需要，直接选择对应的字作为生成的字，跳到第 8 步；
5. 检查是否需要输出标点符号，如果需要，直接选择逗号或者句号作为生成的字，跳到第 8 步；
6. 检查是否已经编写完成，如果是，打印完整诗歌，结束；
7. 从模型计算的最大概率的 K 个字中，随机选取一个字，作为生成的字；
8. 将生成的字作为模型的下一个输入；
9. 重复执行 3-8 步。

代码说明

文件 `main.py` 用于训练和测试模型。其中核心的函数和类：

- `RNN`：实现了一个长短期记忆网络。通过设定参数还可以更改为 GRU 网络。
- `train_epoch`：使用训练集对模型参数进行一轮完整更新，并计算样本平均损失。
- `train`：通过调用 `train_epoch` 函数，对模型参数进行 `EPOCH` 轮更新，并保存测试集表现最好的模型参数，绘制训练中的损失曲线。

- `generate`：通过模型生成一首诗，结合了一些生成诗歌的规则（如前所述）。
- `PoemDataset`：继承 PyTorch 提供的 Dataset 类，对诗歌进行预处理（如前所述）。
- `get_loader`：加载诗歌数据集。
- `test_poem`：生成一系列诗歌，测试实验效果。
- `main`：函数入口。

此外还实现了一些辅助功能：

- `args`：命令行参数
- `log`：日志记录
- `seed_evetything`：固定随机数种子，确保实验可复现

由于模型权重较大（100MB），因此没有随作业提交。

运行代码

安装依赖：

```
pip install torch matplotlib tqdm
```

训练模型：（参数均为可选）

```
python main.py --batch_size 256 --epoch 50 --learning_rate 0.001 --dropout 0.1
```

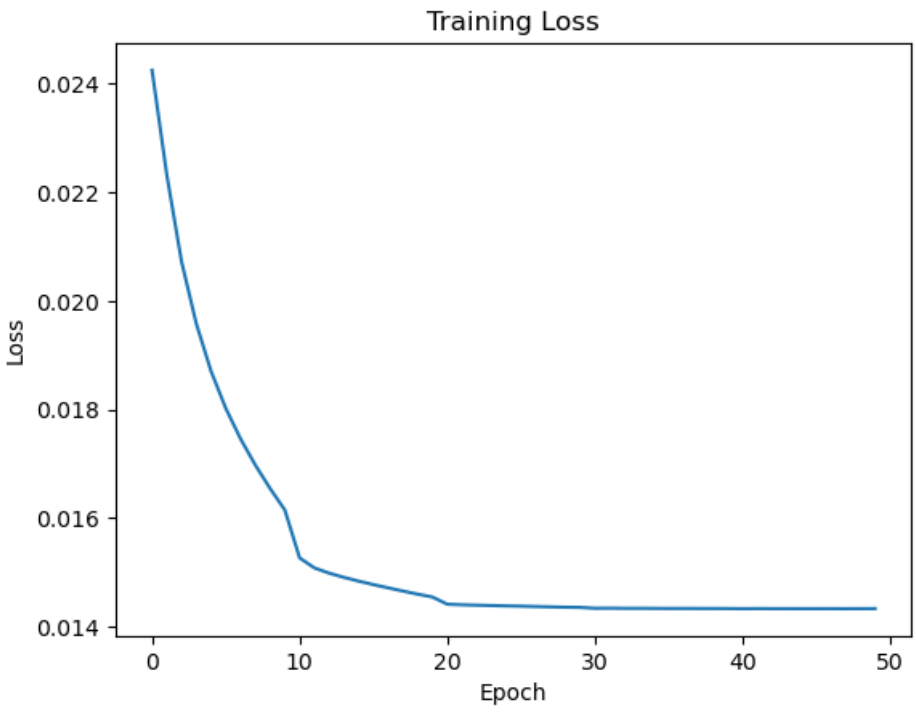
测试模型：（`model_path` 为模型参数文件路径）

```
python main.py --test true --model_path models.pth
```

将依次测试自由生成、给定开头生成和藏头诗生成的效果，如实验结果一节所述。

实验结果

训练损失如下：（最佳损失为第 48 个 epoch）



以下效果展示中，从左到右分别是 K 值为 $1, 3, 5, 10$ 的效果对比。K 越大，生成诗歌时的随机性就越大。

首先测试自由写诗的效果（没有给定开头）：

一片蛩云屏，	天下兵戈尽，	三月三五月，	天生汉祖后，
千寻万象间。	天骄事事多。	花落最高峰。	山色四明中。
风吹金谷雨，	一身同汉主，	人事何须问，	一带松花树，
风动玉山云。	一日换燕支。	人间是几年。	分分玉指林。
玉佩千年盛，	城阙千官拥，	风光连鬓髮，	山河分远岸，
金丹一日新。	天书五道开。	春意起春晖。	城树映高原。
何当一枝桂，	还将汉飞去，	何况蘼闲事，	别有千金乐，
不是别离情。	不见虏尘飞。	不应长是情。	长谣万万春。

一片蛩云屏不定，	一片长松一片冰，	三月风高夜半明，	不是安仁事更长，
一条红豔影参差。	不能骑马上金堤。	玉关云起一千年。	未曾论拙似知贫。
不知何处堪惆怅，	不如今古偏言去，	只应得见长城里，	如能济世无人法，
不见春风不见人。	不是长沙不是人。	犹得当来得路来。	却被闲时与子钱。
香气满时光照灼，	万里烟波千古在，	不用倚花兼不整，	花木自能多兴在，
麝脐初熟夜移缸。	一身风水一乖空。	不妨花叶为谁栽。	竹竿空媿未成行。
不知何事偏相忆，	谁能为我同归去，	闲来若与东归约，	闲中不似君家计，
不是春风不肯开。	不得无家别钓船。	犹自归来似等丹。	为向青冥老钓矶。

给定一个字“好”作为开头：

好是人间事，	好住江城路，	好住江南春，	好是天仙相，
无心亦不知。	门中见此身。	江亭月已暄。	其常不易逢。
不知人事少，	白云归路僻，	不因春景晚，	人非真地下，
不得有情知。	沧岛旅僧归。	犹喜落年多。	地远与人深。
月照池塘月，	远树千峰色，	鸟向庭花落，	草色无风雨，
风吹竹树风。	寒江一月风。	香多杨菜香。	云深有地苔。
何须问知己，	何当重相望，	谁能为君去，	莫忘名画里，
不是有诗人。	应是别家愁。	相见不胜倾。	无用向心边。

给定“好雨知”作为开头：

好雨知何处，	好雨知春晚，	好雨知春尽，	好雨知君子，
高楼望不穷。	新晴入夏晴。	新秋觉夏余。	新花爱晚妆。
云随天仗转，	风光连雨歇，	不知春色好，	未胜芳月好，
云逐蚌胎回。	风物带花飞。	只有客心惊。	更忆旧交期。
日月临天阙，	露重红樱浅，	远水人归寺，	水暗风吹影，
星河入御楼。	风吹绿草长。	空山鸟过林。	楼明柳动花。
还同汉明主，	不知何岁月，	谁怜北山路，	不应成楚老，
应此咏尧年。	来夜是天台。	何日得依过。	明豔正堪携。

给定“好雨知时节”作为开头：

好雨知时节，好雨知时节，好雨知时节，好雨知时节，
高风入夜阑。高风满故枝。孤灯坐夜凉。当寒独掩津。
风吹寒食雨，春风一夜起，清风不相待，雨晴余月夜，
风起一声蝉。吹落不知春。不是别君何。林暗暮烟归。
月下风来好，春色犹相似，万古千家静，日出空山晓，
风前酒上迟。年年独不同。孤城百尺来。声悲落叶中。
明朝不相见，春风不可折，不须为酒伴，此君心更远，
犹自忆佳期。还有泪霑裳。相伴不能开。何惜此芳心。

给定“雪”作为开头：

雪岭风烟里，雪岭秋来尽，雪中花下无，雪后秋光冷，
山川道路长。天河暮不还。花下柳如霜。霜消暮露凋。
山川连晋塞，风烟生楚泽，何况春风早，夜蚌生夜色，
山水接秦川。云雨入隋宫。今年别几多。空树动啼风。
树色连天阔，水色浮云外，春深春欲老，旅恨缘愁易，
潮声入塞深。风烟隔水西。花里晚还归。悲酸恨别忧。
何当一攀桂，谁知此时意，日晚谁相识，空令相见分，
千里共氛氲。不及望夫人。愁人独自知。犹可叹无穷。

给定“白”作为开头：

白日照长川，白日东西来，白首三千岁，白帝高台万，
青山入汉宫。黄榆郁如削。相看又几年。金陵第一过。
云开万里树，君不闻我游，一生同病老，为君如陇日，
树隔九江风。君不见尔为。万恨是秋愁。应在汉宫朝。
别后谁相忆，昔时有子不，夜半胡沙暗，日日星芒满，
离家又几年。今人不得知。春惊碛口悲。河边道吏归。
何时见明月，一朝十九载，何时重此别，还将百司帐，
更上望乡台。四体一相随。应为故乡情。长戟在狼沙。

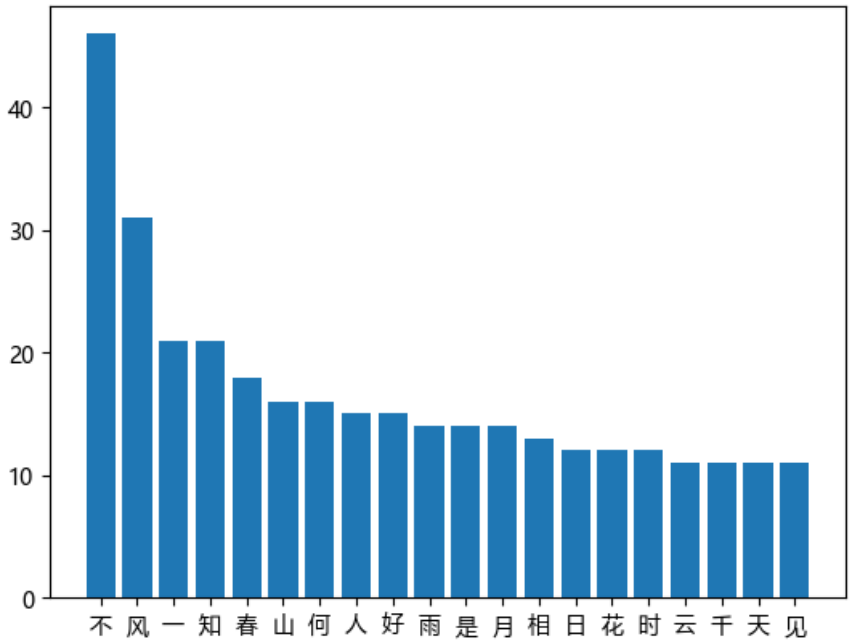
给定“深度学习”的藏头诗：

深山无定所，深山不见山，深宫漏滴时，深入天山山，
度岭有归期。度日望山水。度烛舞腰光。度临嶓梁曲。
学道无人识，学道有余情，学巧风轻动，学领千余树，
习居无事时。习然有遗策。习声月色长。习雪十二绿。

实验分析

在上面实验效果的展示中，一些问题是很明显的：

- 部分字的频率过高，尤其是在同一首诗中重复出现多次。对以上诗中字进行统计的结果如下，“不”和“风”的出现频率过高了。



- 模型学到了一种奇怪的模式：每句诗的第三四句开头的字，有很大概率是重复的。例如：

山川连晋塞， 山水接秦川。	云随天仗转， 云逐蚌胎回。	风吹金谷雨， 风动玉山云。	风光连雨歇， 风物带花飞。

这在 $K = 1$ 的结果中非常明显，而 $K > 1$ 的结果中也有若干出现。

- 模型似乎学到了一些押韵的模式。但也有可能是巧合。

好雨知时节， 高风入夜阑。	白日照长川， 青山入汉宫。	深宫漏滴时， 度烛舞腰光。	深入天山山， 度临嶓梁曲。
风吹寒食雨， 风起一声蝉。	云开万里树， 树隔九江风。	学巧风轻动， 习声月色长。	学领千余树， 习雪十二绿。