

Text Generation with LSTM and Transformer based on Jin Yong Novels

Hao Lu

lh2002@buaa.edu.cn

Abstract

This paper explores Chinese text generation using LSTM and Transformer models trained on the Jin Yong novel corpus. We tokenize the text at the character level and investigate the effects of different training epochs, Transformer layer depths, and sampling temperatures on model performance. Our results show that the LSTM model performs relatively well given the limited corpus, while the Transformer struggles to generate coherent text, likely due to its higher parameter requirements. We also propose the idea of fine-tuning a large pre-trained language model (Qwen3-1.7B) with LoRA, though limited hardware prevented full experimentation.

Introduction

In this experiment, we conducted a Chinese text generation task based on the Jin Yong novel corpus, using both LSTM and Transformer models for training and comparison. We further explored the effects of different numbers of training epochs, varying Transformer layer depths, and sampling temperatures on the generation performance.

Methodology

Data Preprocess

We directly use Chinese characters as tokens to build the vocabulary. After merging all Jin Yong novels into a single long text, we apply `set()` to extract the unique characters and then construct the `index2char` and `char2index` dictionaries accordingly.

Sentences separated by newline characters are converted into sequences by shifting a fixed-length window over them, and these sequences are used to construct the training corpus.

```
[LSTM_e1] Temperature=0.7
张无忌只道：“你还不是。”当下将那马长凳上击落。金轮法王左右右手
碰了两柄长剑，便即住口，将六剑奇快，不动片刻，跃出窗外。这一直不
是他右手臂弯，一人背上，自己一数合、钻研、慕容博、变化、黑白子、
青竹、武当、海所有的一路，想是害怕，他说得到了极为奇怪。杨过失了
眼睛，不由得心中一动，道：「你死啦！」那人身影，往往一片片面相
撞。杨过一人，一齐举起，转身跃出，叫道：「你去练你手臂，
你.....我在我站起身来，这才

[LSTM_e1] Temperature=1.0
张无忌所说的，动手 他学先行分往双围，闪不勤力支撑猛踢的面皮，于盘
夜间已近刀部，奏我如此容动之毒倒伤的，但知郑、晓帖、写四、各友这
样一横的工匠倒有宝贝？各人虽然斗出停船，便即撤回家里，也已窥痛，
还是他用心削作。他们站起身来，只听得李秋水站起身来，纵身的盾马驰
向后翼，两人相距一起，再以身躯也跃得高下，自身飘飘、要用力护得
东、一阵。和陈家洛将如银相交，包来一伸不加手。杨过在康亲王身上
的人争头，心道：“这人副

[LSTM_e1] Temperature=1.3
张无忌从树下磕头。只个楼梯清透轻难的进庙白墙身子，伤势挺弱，只听
刘正人仰小道旁淑女皮太髯斥骂笑声号定。陈智泣道：“浚活为隐升官
文”，八月初译，武人百气身形功俘。庭了十余二三，还是收带旗帜应
兼，大雨行人想到，封这小姐相列习弱向单小必女同不投及方便盖笔，会
做饮老，司徒两东交不了暹罗雷的珍珠搏部上卷土炉，另也必从兵为每大
一百年、安坚兼营博俱俘哭，然后调教不为；其後举起中务，可就皇去虑
继续搬过暂罢。孟平道华不解
```

Figure 1: LSTM 1 epoch

Params

We set the LSTM model to have one layer, with both the embedding and hidden dimensions set to 128. For the Transformer model, the embedding dimension is also 128. We experimented with different numbers of layers: 2, 4, and 6, with the number of heads in the multi-head attention layers set to 4. During the sampling process, we only used the temperature parameter and explored values of 0.7, 1.0, and 1.3.

Experimental Studies

LSTM

We trained the LSTM model for 1, 2, 5, and 10 epochs respectively. The results are as follows. It can be observed that the output is better when the temperature is set to 1.0.

LSTM Visualization

Transformer

Due to the limitations of the corpus, the generation quality of the Transformer model is very poor, so we only present one example. This finding is similar to the observations made with CNNs and ViTs: the Transformer model appears to require a much larger corpus, likely due to its higher parameter count.

[LSTM_e2] Temperature=0.7
张无忌、尼摩星、黄蓉、郭靖、杨逍、范遥、耶律齐、师娘等在每日之中，日后再也去见他们知道。值得一番所见，说道：「你想这二人，你既有好多，我也不能出家人日夜里便此出手。」周伯通道：「你不用瞧你，你.....你.....」周伯通喜道：「是！」杨过一惊，急忙跳出数寸，向左冷禅，喝道：「你养了这麽屎，他便有何门派的？」苗人凤道：「我从来没法 见到。」袁千尺心想：「这几句话说得是甚麽几名弟子，不知铁轮’一个字辈的

[LSTM_e2] Temperature=1.0
张无忌命也要母亲，自己破口尚花了，天武修军的大英雄颇感忧患，但上前时不放良心，见他这般“化为国子”的天意中独散并体，跟随郭靖向牛肉搏，听再见文书之文书房门、六七千里北京，官上乱成两亩七星，神史之中众也颇为失敬。韦小宝和周芷若并非失意。此刻竟悬在寺中庭毯。方不痴情意无微，只说到虎目之中，向来只有余气。徐达、常氏华山二大应不全然，说道：“洋兵难以改认为之，反而提气之士，孤苦思索，吃全干观赏乱闹王，一叙是甚么

[LSTM_e2] Temperature=1.3
张无忌，期又为甚出来瞧清，人非喜欢，景了太干踢一较按怎么噤泥朵？丁春秋本如成功，点一个人，也忙深入宫所殉愿，却行截彩。侯通海侧身闪避不避，登时失了哭齐声 意的玉簪拭一沉，总还练隐传把年纪珍小。却听上了张北岸，走到庭院，韦小宝已似蜷从床前抱住，垫了半晌，说道：“我爹爹何赏之前人混进最天滋的脾滋肉？”桃谷六仙从腰间手林向崔秋山料想摆明的白失落地，供竟忙上涌。那老脸无拢大破的椅子露了多心形。眉前虽愧，青将惊

Figure 2: LSTM 2 epoch

[LSTM_e5] Temperature=0.7
张无忌道：“多谢，我不是他妻子。”郭靖道：“你再见你，你.....你.....你.....一切！”忽听得山洞后一阵风声飒然，已瞧着杨过，眼见他抓住他手腕，便击毙手足，这时凌厉异常，但他武功未及，但掌法奇妙，便觉得心贯通，当下自己剑法这两招「毒蛇针上显是金针，不必逼迫，徒弟为师，但因此事说过，只是蒙古 长兵刃，和周伯通四肢无比，无法决胜之人，眼见 她一路拚命，但觉他手臂酸软，嘴角边有一股长剑，跃出一步，於是以一招，见那

[LSTM_e5] Temperature=1.0
张无忌双臂劳烟，原来倒在人手中之多，但他竟会从势渐渐开开。那店伴当时还顾二哥，只是一个小女子来，我姑爷又不喜欢，在我们红花会之后，大师哥不可收拾了他，胡说八道呢。他身边恒山派变化成拳的技艺。”任我行道：“我可也有谁能怪她相助强我，除非非辟邪师者。”当下作别。虽有这般惊张无忌，暗想此人对自己霍青桐所用心中好问，只听对方号称各擅兵武功较强，当身柱香深厚，变化罢了，只是见到了福康安。段誉渐渐躲到近西，更无意间

[LSTM_e5] Temperature=1.3
张无忌挺托倒重，自己闪电闪避。骆驼奔到山石两招数光。小姑娘见来已久，火圈两下两项球均般的伤势“刹，还男装居盘堂的陌生人所带之人，第三日便曾见过一恶人。初时拚拳把他书生已用来势共留射他扶气。巫师太师便该欲疗伤凌于晨曦照于腹上投变化解，遍无疾德一吸，他虽然喝甚极，当即开指。十多归见乐的少女取出，这阵势对他不知是明王而伟坐后。玄生忽然吓了一跳，嚷道：“这厮的总镖头出手品有不纯人，假照身悬尤厉不打得提，更钻研碎

Figure 3: LSTM 5 epoch

```
[LSTM_e10] Temperature=0.7
张无忌左手一扯，这一拳打在他的肩头，说道：“这位是谁？你就算是你，哪里还有什么不好？”黄蓉道：“不错，我倘若真是不知如何？我.....我.....我.....”桃根仙道：“是我的，那也不会再说了，这真是不是？”桃实仙道：“有甚么？”令狐冲道：“不错，要我带一人。”众人齐道：“不过他们不要下毒手，那是晚辈便是。”青青道：“我就是了，只怕你老人家一起去，我去给小师妹跟你说。”他自知段公子已经死了，那两个儿子分心，知道是对方，

[LSTM_e10] Temperature=1.0
张无忌大声喝道：“这小贼也有甚麽江南七怪？他岂能不那行网？不能再算误会，来到青城四个，七十二总，小路啦。”周芷若又道：“这位是哥萨果子、大段文武，东剑西方掌门的来罢，这位小小第三七种妙极，就算已经究普日月天而复，咱们求救你。”孟健雄道：“佩服！那我不要杀他来报仇，料想不赢。虽然佛教主有谁能好，实是武功之人之前不放在上，大家义门下子，贵派之个所为，原因所振海御用，说三祖非常大哥误认。”赵煦皱眉不语，抚传令

[LSTM_e10] Temperature=1.3
张无忌道：“且得意即挡起，不与就同向封万夫齐下眺祠，悉江渡山道人化去了。后来捣夷麻斗，后着崖边。”穆一旁肋消一展，骂倒金猴，拉着她腰，床上紫气直流，桑鸣其中数时的彩头没右髀身上石块；胡斐祖夫传进丰悄生敬，苦亲神‘岳爷扬州魔外寺馆中物’。空智告知应声，何巧时身过恶物，甚至大事自然。两名宫女昨晚竟会典之教了没术；有的投入湖广接地，觉得中了一晚众程野伐，六十天之具之于高林中。未毕攻狮萧峰之中情景，要收出租布长
```

Figure 4: LSTM 10 epoch

Finally, we attempted to perform LoRA fine-tuning on the Jin Yong novels using the pre-trained Qwen3-1.7B model. However, due to hardware limitations and extremely slow training speed, we only discuss the idea here without further experimentation.

Conclusions

This experiment systematically compared the performance of LSTM and Transformer models on the task of Chinese character-level text generation using Jin Yong novels. The LSTM model demonstrated strong performance even with limited data, producing coherent and contextually meaningful text when trained over sufficient epochs. In contrast, the Transformer model failed to generate plausible results, which we attribute to its higher data requirements and complexity. This outcome aligns with similar findings in vision tasks comparing CNNs and ViTs. Although large-scale pre-trained models like Qwen3-1.7B offer promising alternatives, they demand substantial computational resources. Overall, LSTM remains a viable choice for character-level generation in low-resource scenarios.

[illegible]

Figure 5: Transformer 4 layer, 10 epoch