Report of Deep Learning for Natural Language Processing

Shixiang Li BY2203043

lishixiang@buaa.edu.cn

Abstract

This study investigates the differences in the generative capabilities of language models based on the Jin Yong novel corpus. Two different model architectures, LSTM and Transformer, are selected for comparison. For LSTM, a full-scale training approach is adopted, while for the Transformer, full-scale fine-tuning is performed on a pre-trained model. By comparing the content generation results of the two models, it is observed that current pre-trained language models exhibit strong text generation abilities, achieving impressive results with minimal fine-tuning.

Introduction

In recent years, with the continuous evolution of deep learning technologies, text generation techniques have made breakthrough progress in various fields such as news writing, dialogue systems, automatic summarization, and literary creation. Particularly in the domain of literary creation, generating text with specific styles and literary forms using machine learning models has become a challenging and promising research direction in the field of Natural Language Generation (NLG).

This experiment focuses on two mainstream text generation models: LSTM and Transformer. LSTM, with its advantage in time-series modeling, was widely used in early text generation tasks. On the other hand, the Transformer model, with its excellent modeling capabilities and advantages in parallel computation, has nearly become the standard model for natural language processing tasks in recent years. By systematically comparing the performance of these two models in generating wuxia novels, this study aims to analyze their strengths and weaknesses in style learning, text coherence, and generation quality, providing both theoretical insights and practical references for future related research.

LSTM

Long Short-Term Memory (LSTM) [1] is a variant of Recurrent Neural Networks (RNNs) proposed by Hochreiter and Schmidhuber in 1997, designed to address the gradient vanishing and exploding problems that traditional RNNs encounter when learning long sequences. By introducing a gating mechanism, LSTM effectively captures long-range dependencies, making it widely applicable in tasks such as language modeling and sequence generation.

Due to its gated structure, LSTM is particularly effective at capturing dependencies within long sequences, making it suitable for generative tasks that require contextual memory. However, the sequential nature of its computations (i.e., each time step depends on the previous one) results in slower training and inference speeds, and it struggles to fully leverage the parallel computing power of modern hardware. Moreover, when dealing with extremely long texts, LSTM may still face challenges in modeling global information, leading to issues with long-range logical coherence in the generated text.

Transformer

The Transformer model, introduced by Vaswani et al. in 2017 in the paper Attention is All You Need [2], quickly became a mainstream architecture in the field of natural language processing. Unlike traditional structures based on recurrent (RNN) or convolutional (CNN) networks, the Transformer relies entirely on the self-attention mechanism, which enables it to efficiently model relationships between any two points in a sequence, supports parallel processing, and offers exceptional scalability.

Fundamentally, the Transformer model frees itself from the constraints of sequential computation, allowing for global modeling of the entire input sequence. This gives it a natural advantage in capturing long-range dependencies and improving the coherence of generated text. Additionally, due to its high degree of parallelism, the Transformer excels when working with large datasets and models with extensive parameters. However, the Transformer typically requires more computational resources and training data, and when the sample size is limited, it is prone to overfitting. Moreover, its large parameter count and high memory requirements can pose challenges in resource-constrained environments.

Methodology

M1: LSTM Model Architecture

LSTM (Long Short-Term Memory) is an improved variant of Recurrent Neural Networks (RNNs) that effectively captures dependency information in long sequences. The LSTM unit controls the flow of information by introducing three gates: the Forget Gate, the Input Gate, and the Output Gate, which help mitigate the vanishing gradient problem.

The corresponding formulas for the LSTM unit are as follows:

Forget Gate:

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \tag{1.1}$$

Input Gate:

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \tag{1.2}$$

$$\tilde{C}_t = \tanh\left(W_C \cdot \left[h_{t-1}, x_t\right] + b_C\right) \tag{1.3}$$

Update Memory Cell:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{1.4}$$

Output Gate:

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \tag{1.5}$$

$$h_t = o_t * \tanh\left(C_t\right) \tag{1.6}$$

Where σ represents the Sigmoid activation function, $*$ denotes element-wise multiplication, and W and b represent the weights and biases, respectively.

The key characteristic of LSTM is its ability to effectively capture both short-term and long-term dependencies, which is why it has found widespread application in traditional text generation tasks.

M2: Transformer Architecture

The Transformer model, proposed by Vaswani et al. in 2017, is entirely based on the attention mechanism, eliminating the need for recurrent or convolutional structures. This greatly enhances the parallelism of training and the ability to model long-range dependencies.

The core mechanism of the Transformer is self-attention (Self-Attention). Its computation process is as follows:

The input is represented as a sequence of word vectors $X \in \mathbb{R}^{n \times d}$, which undergoes linear transformations to obtain the query (Q), key (K), and value (V):

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V \tag{1.7}$$

Self-attention calculation formula:

$$Attention(Q, K, V) = Concat(head_1, \ldots, head_h)W^O \tag{1.8}$$

The head of each subspace：

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{1.9}$$

In addition, to incorporate positional information, Transformer also adds Positional Encoding：

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{1.10}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{1.11}$$

Transformer leverages the above mechanisms to achieve powerful sequence modeling capabilities, particularly excelling at handling long-text data.

Experimental Studies

E1：Data Processing

Corpus source: Multiple wuxia novels by Mr. Jin Yong (such as *The Legend of the Condor Heroes*, *The Return of the Condor Heroes*, *The Smiling, Proud Wanderer*, etc.).

Processing steps: Remove annotations and special characters; split into fixed-length input sequences (length = 100 characters); construct both character-level and word-level versions (this paper focuses on the character-level version).

E2：Training Setup

LSTM: The LSTM model uses a full training approach.

- 2 layers of LSTM

- 256 hidden layer dimensions

- Adam optimizer

- Learning rate of 1e-3

- 10 epochs

Transformer: The Transformer model utilizes a pre-trained model with full fine-tuning.

- Network architecture: GPT2 (gpt2-chinese-cluecorpussmall)

- 5 epochs

E3：Main Results

Table 1 Example of LSTM-generated results

No	Input	Target output (original text)	Model Output
1	在黄沙莽莽的回疆大漠之上	尘沙飞起两丈来高，两骑马一前一後的急驰而来。前面是匹高腿长身的白马，马上骑著个少妇，怀中搂著个七八岁的小姑娘。後面是匹枣红马，马背上伏著的是个高瘦的汉子。	在黄沙莽莽的回疆大漠之上，突然纵起，双足跟著。苏普见到这绿色，有人同直系敌人的，但是爹爹妈妈逃来回疆，曾听他们说过『高昌迷
2	大明成祖皇帝永乐六年八月，	西南海外渤泥国国王麻那惹加那乃，率同妃子、弟、妹、世子及陪臣来朝，进贡龙脑、鹤顶、玳瑁、犀角、金银宝器等诸般物事。	但十年来念念不忘汉人强盗的深仇大恨，一听「汉人强盗」四字，登时清醒了三分，一跃而起，叫道：「汉人强盗在那里？」苏普向陈达海一

3	一个嘶哑的嗓子低沉地叫着	叫声中充满着怨毒和愤怒，语声从牙齿缝中迸出来，似是千年万年、永恒的咒诅，每一个字音上涂着血和仇恨。	说到这里，声音已是发颤。李文秀道：「咱们族人大队人马就快到来，可得告诉他们，别让兄弟们闯进宫去，一个个的死於非命。」苏鲁克道：「对！只要是在迷宫之外，

Table 2 Example of GPT2 generation results before fine-tuning

No	Input	Target output (original text)	Model Output
1	在黄沙莽莽的回疆大漠之上	尘沙飞起两丈来高，两骑马一前一後的急驰而来。前面是匹高腿长身的白马，马上骑著个少妇，怀中搂著个七八岁的小姑娘。後面是匹枣红马，马背上伏著的是个高瘦的汉子。	一切都那么美好的作品很经典。里面有个片段是关于红军的很多个地方都是描写当时红军是那么渺小但是在经过几百年后所有的变化都让我们看到了那个伟大的中国军事才华。对于这样一部作品我们在了解其它人的历史时却不知如何去理解当中的不同。我们在了解这样一个民族所有的历史的过程中
2	大明成祖皇帝永乐六年八月，	西南海外渤泥国国王麻那惹加那乃，率同妃子、弟、妹、世子及陪臣来朝，进贡龙脑、鹤顶、玳瑁、犀角、金银宝器等诸般物事。	隆庆帝（康熙）和顺天府巡抚李自成被杀，这是顺天府首犯。在其死后的五年半时间内，因安德烈王太后的病逝，顺天府城被毁，北京城也受到极大的破坏。其时

			明朝大臣、藩王、藩王、使持府尹、大将军、总督大臣等均有人丧失官职、丧失执政权利、死亡或者财产灭绝等不同程度的死亡。明朝初年，中国
3	一个嘶哑的嗓子低沉地叫着	叫声中充满着怨毒和愤怒，语声从牙齿缝中迸出来，似是千年万年、永恒的咒诅，每一个字音上涂着血和仇恨。	但不至于呛人。就是声音太响了。。还有就是耳机不能换。我只换个耳机。耳机就不能换了。不过送人的。个耳机还不错。

Table 3 Example of GPT2 generation results after fine-tuning:

No	Input	Target output (original text)	Model Output
1	在黄沙莽莽的回疆大漠之上	尘沙飞起两丈来高，两骑马一前一後的急驰而来。前面是匹高腿长身的白马，马上骑著个少妇，怀中搂著个七八岁的小姑娘。後面是匹枣红马，马背上伏著的是个高瘦的汉子。	能想起的只剩下这几日。此时，天下只剩黄沙，不会流下一滩水，哪里来的这样一口好水？说起黄沙莽莽，黄沙还有几日就要退了。那些人已下马，只见在沙漠里走了好几个人，当然没有走。沙漠里面还有三个敌人，当日晚上，三，又有一个人过来亦蒙，你他？
2	大明成祖皇帝永乐六年八月，	西南海外渤泥国国王麻那惹加那乃，率同妃子、弟、妹、世子及陪臣来朝，进贡龙脑、鹤顶、玳	三人来到御道中，来得有两人，却又都是当时的人物和兵，又有四人居住在一起，当下不由

		瑁、犀角、金银宝器等诸般物事。	得吃了一惊。洪七公又说：我也不知到底哪里做错了，可到底到底为甚么？洪七公怒声道：我跟你爷爷道个好，这可是你皇上的生日礼物。这里有你爸，请收下。洪七公回答说：还想到爷爷的话，我们这里可不能回头。那是，我
3	一个嘶哑的嗓子低沉地叫着	叫声中充满着怨毒和愤怒，语声从牙齿缝中迸出来，似是千年万年、永恒的咒诅，每一个字音上涂着血和仇恨。	拿著长刀，拍的一拳，打量丁同的尸体，又不忍则。如轻子突然大叫，道：「咱们去问计老人动寻地的还是奇怪。他们从这是高狼皮的是」，突然喜欢：「他不知道。师父，你知道麽？」瓦耳拉齐道：

The experimental results show that LSTM, after full training, can generate a certain novel style, but its performance in capturing long-sequence information is poor. The pre-trained GPT model is capable of generating semantically coherent content even before fine-tuning, though it remains relatively generic. After fine-tuning, it can generate text that aligns with the desired style.

Conclusion

This study conducted text generation experiments using both LSTM and Transformer models based on the corpus of Jin Yong's wuxia novels, with a systematic comparison from multiple perspectives. The experimental results indicate that LSTM has certain advantages in local text coherence, but it is significantly inferior to Transformer in modeling long-range dependencies and creativity, and it suffers from longer training times and greater difficulty in training. The

Transformer not only generates text with a more natural language style but also captures long-sequence relationships effectively. Moreover, from the perspective of model complexity and training efficiency, Transformer has a clear advantage in large-scale data training due to its ability to perform parallel computation, suggesting a promising application for large language models (such as the GPT series) in the field of literary creation. Future work could consider incorporating larger pre-trained models (such as Chinese GPT or Wenxin Yiyan) and combining techniques like small-sample fine-tuning (e.g., LoRA) to further enhance the generation quality.

References

[1] Graves A, Graves A. Long short-term memory[J]. Supervised sequence labelling with recurrent neural networks, 2012: 37-45.

[2] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.