**Report of Deep Learning for Natural Langauge Processing**

Xintong Zhang

zhangxintong0810@icloud.com

**Abstract**

This assignment utilizes the given corpus of Jin Yong's novels to complete a text generation model based on LSTM. The text generation model is implemented through the LSTM neural network method. A known paragraph from Jin Yong's novels is input as a prompt to generate a new paragraph and conduct quantitative and qualitative analyses.

**Introduction**

**LSTM Model**

Long Short-Term Memory (LSTM) is a commonly used Recurrent Neural Network (RNN) architecture for processing sequential data, especially those with long-term dependencies.

LSTM solves the problems of vanishing and exploding gradients in traditional RNNS by introducing a mechanism called "gate", enabling it to effectively capture and utilize long-term dependencies. The following are the main components of LSTM and their working principles:

1. Input Gate: Controls whether new input information is added to the cell state. It performs a weighted sum of the input and the hidden state of the previous moment, and then generates a value between 0 and 1 through a sigmoid function, representing the importance of each input element.

2. Forget Gate: Controls which information from the cell state at the previous moment is retained. It performs a weighted sum of the hidden states of the input and the previous moment, and then generates a value between 0 and 1 through a sigmoid function, representing the retention degree of each cell state element.

3. Cell State Update: Based on the results of the input gate and the forget gate, calculate the new candidate cell states. First, use the input gate to determine which information will be added to the cell state. Then, a forgetting gate is used to determine which information in the cell state at the previous moment should be forgotten. Finally, combine the two to obtain a new cell state.

4. Output Gate: Controls the output of the current moment based on the input and the hidden state of the previous moment. It performs a weighted sum of the input and the hidden state of the previous moment, and then generates a value between 0 and 1 through a sigmoid function, representing the contribution degree of each cell state element to the output. Meanwhile, the current cell state is processed through a tanh function to obtain a value between -1 and 1, representing the output at the current moment.

The unit structure of LSTM is shown in the following figure:

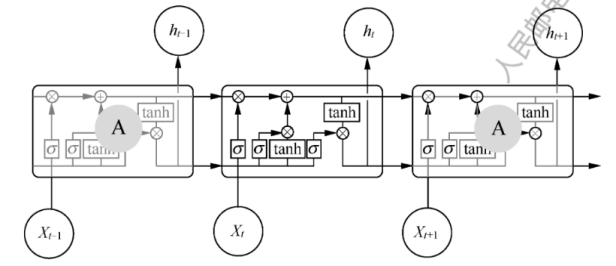


Figure1 LSTM model structure

Through the above steps, LSTM can effectively handle sequential data and retain and utilize long-term dependencies during the learning process. Its main advantage is that it has no limit on the time steps of input and output, and can capture dependencies over longer distances. This enables LSTM to be widely applied in sequence data processing tasks such as language modeling, machine translation, and speech recognition.

**Methodology**

1. **Import the required libraries**

Import the relevant libraries required for the code, among which jieba is used for word segmentation and tqdm is used for printing the progress bar.

****

1. **Define the functions for data reading and preprocessing**

Define the function for reading the article and performing preprocessing. The preprocessing operation here cannot remove all punctuation marks as in the previous preprocessing. Because for the text generation model, punctuation marks also have their own significance in Chinese text. They can represent pauses - such as periods and commas, and can also represent emotions - such as question marks for questions and exclamation marks for surprises. It can also represent what is said - for example, double quotation marks and colons. So it cannot be deleted at will.

By observing the text of "The Legend of the Condor Heroes", it can be found that there are many symbols that should not be present in Chinese text. These symbols should be deleted during the preprocessing stage, including carriage returns, TAB characters, Spaces, and some garbled characters, such as.

Meanwhile, in "The Legend of the Condor Heroes", in some places, double and single quotation marks are represented in the form of "" and" ", which need to be replaced with the double and single quotation marks in the normal Chinese text.

At the same time, some Chinese texts that do not belong to the content of "The Legend of the Condor Heroes" should also be deleted



1. **Read the data and build the mapping dictionary**

Because the corpus is quite large, if all 16 novels were used for training, the training process would be too long. Therefore, I only chose the story "The Legend of the Condor Heroes" as the corpus.

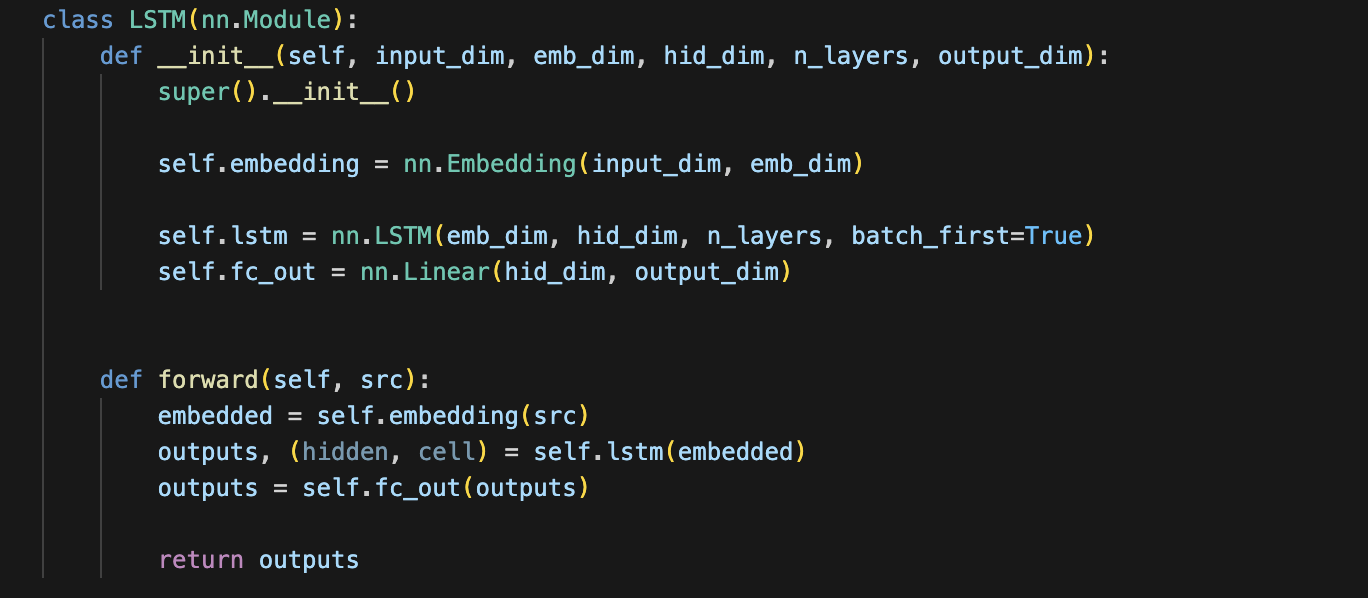
Meanwhile, since the LSTM network does not accept Chinese characters as input, it is necessary to construct a mapping dictionary from Chinese characters to indexes and convert the list of Chinese character words into an index list for the network to train.

Since the output index of the network still needs to be converted into Chinese characters during the testing stage to view the text generation effect of the model, a mapping dictionary from the index to Chinese characters also needs to be constructed.



1. **Define the LSTM model**

Give the basic parameter structure definitions of LSTM.。



1. **Set hyperparameters**

The definition of the embedding layer in the LSTM model requires two parameters. One is the total number of words in the corpus dictionary, vocab\_size, and the other is the feature dimension embed\_size for expanding each word in the corpus.

The definition of the LSTM layer in the LSTM model requires three parameters. One is the feature dimension embed\_size that the embedding layer expands for each word in the corpus, and the other is the number of neurons in each hidden layer at each time step of the LSTM model hidden\_size. The other one is the number of hidden layers num\_layers for each time step of the LSTM model.

In the final fully connected layer of the LSTM model, it is necessary to convert the output of LSTM into the probability of each word in the corpus dictionary. Therefore, the output dimension remains the total number of words in the corpus dictionary, vocab\_size.

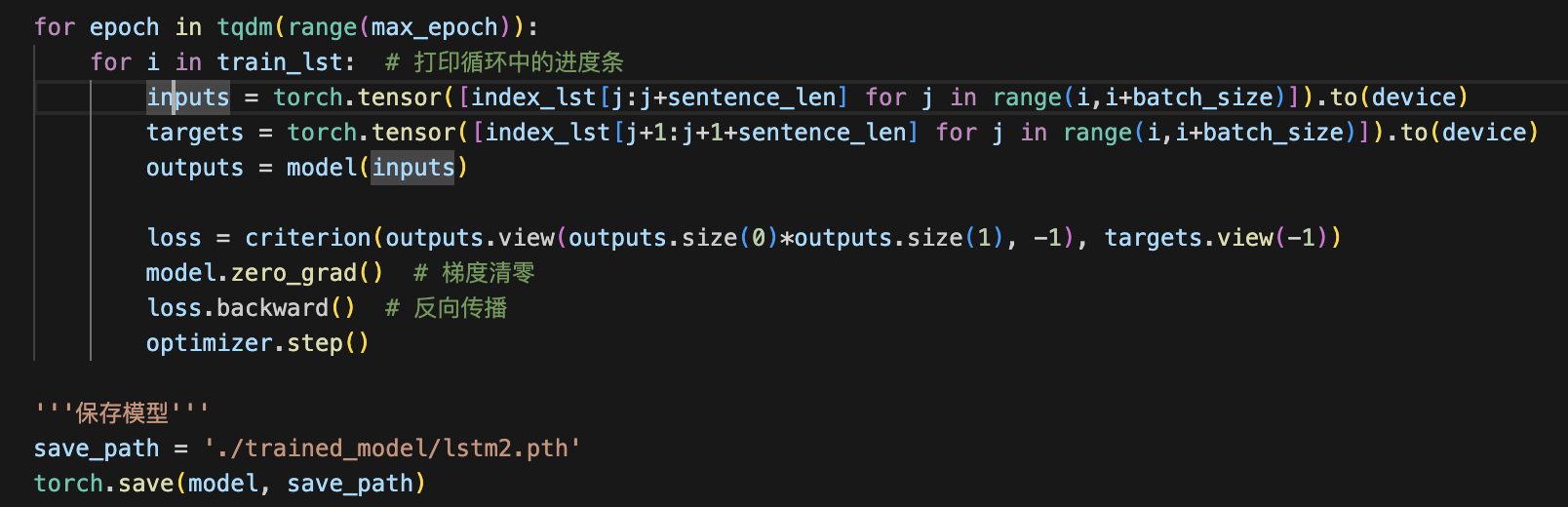
For the training process, it is necessary to set the maximum training period, batch size, learning rate, the sentence length accepted by the LSTM model, the range of the training set, whether to use GPU training, the loss function and the optimizer.



1. **Model Training**

When training an LSTM model, one point needs to be noted: the target output language sequence of the model should be a sequence of words one position after the input sequence, that is, each input sequence moves one position to the right as the target sequence. Only in this way can it be ensured that the model can generate words in sequence based on the input after training.

Meanwhile, after the training is completed, the model needs to be saved for testing.

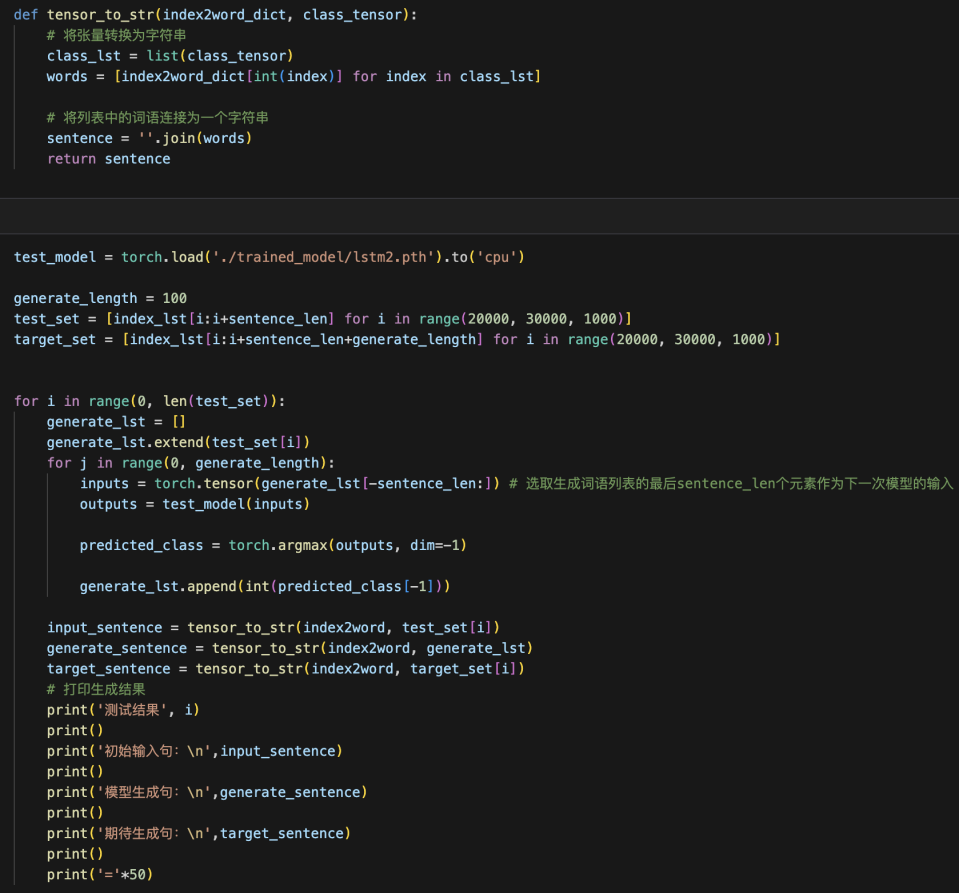


1. **Model testing**

From the non-training set part of "The Legend of the Condor Heroes", find 10 sentences of sentence\_len length as the initial input of the test model, and let the model generate respectively according to these 10 sentences.

During the testing phase, set the generated sequence length "generate\_length" to indicate how many words the model is allowed to generate during the testing phase.

Let the model generate a new sequence based on the initial input sequence, and take the last word of the sequence output by the model and store it in generate\_lst. Then let the model continue to take the last sentence\_len words of generate\_lst as input to obtain the last word of the new output sequence. Continue to store it in generate\_lst, and so on until the length of the generated sequence reaches generate\_length.



**Experimental Studies**

1. **Test Results**

测试结果 0

初始输入句： ，但见那乘马奔到大街转弯角处，忽然站住。完颜洪烈又是一奇，心想马匹

模型生成句： ，但见那乘马奔到大街转弯角处，忽然站住。完颜洪烈又是一奇，心想马匹，就留给两个还没出世，忽然转念：“别鬼使神差的，偏偏有人这时过来撞见。”鼓起勇气，过去拉那尸首，想拉入草丛之中藏起，再去叫丈夫。不料她伸手一拉，那尸首又呻吟了一下，声音甚是微弱。她才知此人未死。定睛看时，见他背后肩头中了一枝狼牙利箭，深入肉里，箭枝上染满了血污。天空雪花兀自不断飘下，那人全身已罩上了薄薄一层白雪，

期待生成句： ，但见那乘马奔到大街转弯角处，忽然站住。完颜洪烈又是一奇，心想马匹疾驰，必须逐渐放慢脚步方能停止，此马竟能在急行之际斗然收步，实是前所未睹，就算是武功高明之人，也未必能在发力狂奔之时如此神定气闲的蓦地站定。只见那矮胖子飞身下马，钻入一家店内。完颜洪烈快步走将过去，只见店中直立着一块大木牌，写着“太白遗风”四字，却是一家酒楼，再抬头看时，楼头一块极大的金字招牌，写着“醉仙楼”三个大字

测试结果 1

初始输入句： 。那人腰里插了一柄砍柴用的短斧，斧刃上有几个缺口。两人刚

模型生成句： 。那人腰里插了一柄砍柴用的短斧，斧刃上有几个缺口。两人刚笑道：“说不定！”这时了五年龙廷，那人全身已罩上了薄薄一层白雪，只须过得半夜，便冻也冻死了。她自幼便心地仁慈，只要见到受了伤的麻雀、田鸡、甚至虫豸蚂蚁之类，必定带回家来妥为喂养，直到伤愈，再放回田野，若是医治不好，就会整天不乐，这脾气大了仍旧不改，以致屋子里养满了诸般虫蚁、小禽小兽。她父亲是个屡试不第的村

期待生成句： 。那人腰里插了一柄砍柴用的短斧，斧刃上有几个缺口。两人刚坐定，楼下脚步声响，上来两人。那渔女叫道：“五哥、六哥，你们一齐来啦。”前面一人身材魁梧，少说也有二百五六十斤，围着一条长围裙，全身油腻，敞开衣襟，露出毛茸茸的胸膛，袖子卷得高高的，手臂上全是寸许长的黑毛，腰间皮带上插着柄尺来长的尖刀，瞧模样是个杀猪宰羊的屠夫。后面那人五短身材，头戴小毡帽，白净面皮，手里提了一杆秤，

测试结果 2

模型生成句： 查觉，当即带同亲随，由临安府的捕快衙役领路，亲自追拿刺客。追她，只听在断墙残瓦的破败之地，直到人挟持，当下捡出丈夫的止血散金创药，拿了小刀碎布，在灶上提了半壶热酒，又奔的孩子都是男儿的村学究，按着她性子给她取个名字，叫作惜弱。红梅村包家老公鸡老母鸡特多，原来包惜弱饲养鸡雏之后，决不肯宰杀一只，父母要吃，只有到市上另买，是以家里每只小鸡都是得享天

期待生成句： 道：“你在营里当官，不去欺负别人，人家已谢天谢地啦，又有谁敢欺负你啦？”段天德满脸惭容，说道：“侄儿不争气，给一个恶道赶得东奔西逃，无路可走。求伯父看在我过世的爹爹面上，救侄儿一命。”枯木听他说得可怜，问道：“那道人追你干什么？”段天德知道越是将自己说得不堪，越是易于取信，当下连称：“侄儿该死，该死。前日侄儿和几个朋友，到清冷桥西的瓦子去玩耍……”枯木鼻中哼了一声，脸色

其余7个生成结果见测试代码中。

1. Result analysis

By observing several test results, it can be found that the generated sentences are logical, relatively coherent, understandable to people, and the use of punctuation marks is also relatively correct. At the same time, the characters and language styles in the generated sentences also conform to the content of "The Legend of the Condor Heroes". By comparing the generated sentences with the real original sentences, it can be found that there is still a certain gap in content comparison.

(2) Areas that can be improved

Due to the limitations of computing power and time, the training set I chose only included 10,000 sequences of length 20 from "The Legend of the Condor Heroes". Under the condition of sufficient time and computing power, the scale of the dataset can be further increased, and thus the statement generation ability of the trained model will be better.

Meanwhile, in terms of the network structure, in order to save computing power and training time, the LSTM model network I set up is relatively simple. The number of embedded features for each word is only 30, the number of neurons in each hidden layer of LSTM is only 512, and the number of hidden layers in each time step is only 2. All these can be further increased under the condition of sufficient time and computing power

**Conclusions**

This assignment implemented the text generation task of Jin Yong's novel "The Legend of the Condor Heroes" based on the LSTM model. By constructing a mapping dictionary from Chinese characters to indexes, the Chinese text is transformed into sequence data that can be processed by LSTM, and a network structure including an embedding layer, an LSTM layer and a fully connected layer is designed.

The experimental results show that the text generated by the model performs well in terms of logic, coherence and the use of punctuation, and is capable of imitating the narrative style and character language characteristics of the original work. For example, the character dialogues and scene descriptions in the generated text are close to the style of the original text, and the use of punctuation marks (such as periods, commas, and quotation marks) conforms to Chinese grammar rules. However, compared with the original text, the generated content still has gaps in terms of detail richness and plot rationality. For example, some generated paragraphs have problems such as semantic jumps or logical incoherence.

**Problems and Future Improvements:**

1. Data scale: The current training set only contains 10,000 sequences of length 20. Due to the limitations of computing power and time, the model has not fully learned the complex patterns of the corpus. In the future, it can be expanded to all 16 novels, and the sequence length and quantity can be increased to enhance the model's ability to capture long-distance dependencies and diverse themes.

2. Model structure: The embedding dimension (30), the number of neurons in the hidden layer (512), and the number of layers (2) of the existing model are relatively simple. By increasing the dimension of embedded features (such as 128 or 256), expanding the scale of LSTM hidden layers (such as 1024 neurons) and depth (such as 4 layers), or introducing an attention mechanism, the model's understanding of the context and generation quality can be enhanced.

In conclusion, LSTM has shown potential in the task of Chinese text generation, but the generation effect needs to be optimized through larger-scale data and more complex architectures. This method can be further applied to the automatic creation or assisted writing scenarios of other literary works

**References**

[1] Zenchang Qin (2025)，LLMs and Vision-Language Problems.Vol. 3: 23: pp. 1-12.