LDA主题模型对于文本的建模与分类

刘峻池
2252298@qq.com

摘要

本实验以作家金庸的16本中文小说作为语料库，分别对每本小说进行了数据预处理、字词分割和字词标号，实验中建立了LSTM长短期记忆模型，通过选取合适的模型参数并输入字词对应的编号序列对模型进行训练，对模型生成的语句结果进行了展示并对模型性能和改进方向进行了总结与分析。

一、介绍

LSTM（Long Short-Term Memory）是一种循环神经网络（RNN）的变体，它在处理序列数据方面具有出色的性能。LSTM模型通过使用特殊的记忆单元结构来解决传统RNN中的长期依赖问题。

在传统的RNN中，由于梯度消失或梯度爆炸的问题，网络难以有效地捕捉到长期依赖关系。这意味着在处理长序列数据时，传统RNN的性能受到限制。LSTM通过引入称为"门"的机制来解决这个问题。

LSTM的核心思想是使用三个门：输入门（input gate）、遗忘门（forget gate）和输出门（output gate）。这些门通过学习来控制信息的流动，从而有效地处理长期依赖关系。

输入门决定了当前时刻输入的重要性，遗忘门决定了是否忘记之前的状态，而输出门则决定了当前时刻输出的内容。这种门的机制使得LSTM能够选择性地存储和读取信息，有效地传递重要的上下文。

LSTM在各种序列数据的任务中表现出色，并被广泛应用于自然语言处理（NLP）、语音识别、机器翻译、时间序列预测等领域。例如，对于情感分析任务，LSTM可以有效地建模文本的上下文信息；对于语音识别任务，LSTM可以处理长语音序列并捕捉到语音中的时序信息。

二、方法原理

1、LSTM模型理论

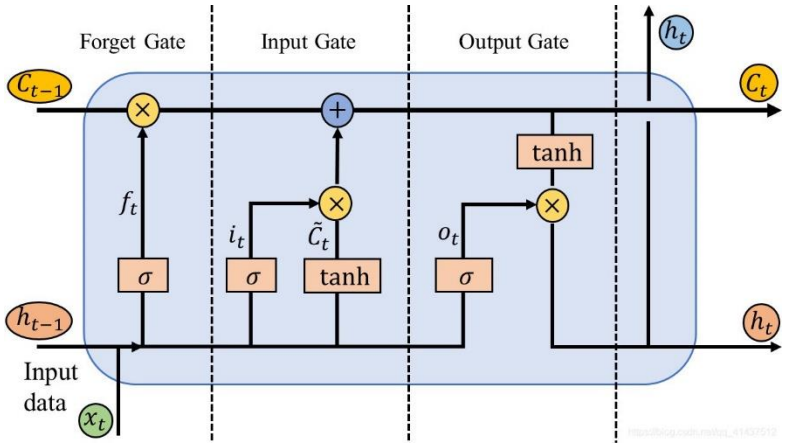

图1 LSTM模型[1]

LSTM（Long Short-Term Memory）是一种循环神经网络（RNN）的变体，通过引入特殊的记忆单元结构来解决传统RNN中的长期依赖问题。下面是LSTM的详细算法原理：

1．LSTM单元结构：

LSTM单元由三个关键组件组成：记忆细胞（cell state）、输入门（input gate）、遗忘门（forget gate）和输出门（output gate）。记忆细胞用于存储和传递信息，输入门控制新信息的输入，遗忘门控制旧信息的遗忘，输出门控制输出的内容。

2．记忆细胞（Cell State）：

记忆细胞是LSTM的核心部分，负责存储和传递信息。记忆细胞的状态在整个序列中保持不变，可以传递给下一个时间步骤。记忆细胞的更新通过输入门、遗忘门和记忆细胞更新公式来控制。

3．输入门（Input Gate）：

输入门决定了当前时刻输入的重要性。它通过使用一个sigmoid激活函数来生成一个0到1之间的值，表示输入的重要程度。输入门的输出与一个tanh激活函数的值相乘，用于确定新的记忆细胞候选值。

4．遗忘门（Forget Gate）：

遗忘门决定了是否忘记之前的状态。它通过使用一个sigmoid激活函数来生成一个0到1之间的值，表示忘记的程度。遗忘门的输出与之前的记忆细胞状态相乘，用于决定要忘记的信息

5．记忆细胞更新：

记忆细胞更新包括两个步骤。首先，通过将输入门的输出和记忆细胞候选值相加，更新记忆细胞的候选值。然后，通过将遗忘门的输出和上一步的记忆细胞状态相加，得到最终的记忆细胞状态。

6．输出门（Output Gate）：

输出门决定了当前时刻输出的内容。它通过使用一个sigmoid激活函数来生成一个0到1之间的值，表示输出的程度。输出门的输出与经过tanh激活函数处理的记忆细胞状态相乘，用于生成LSTM单元的最终输出。

7．循环连接：

LSTM通过循环连接的方式，将上一个时间步骤的记忆细胞状态传递给当前时间步骤，并

接收来自当前时间步骤的输入。

通过以上的算法原理，LSTM能够有效地捕捉到序列数据中的长期依赖关系。在训练过程中，LSTM通过反向传播算法进行参数更新，使得模型能够学习到适合任务的权重参数。在应用阶段，LSTM可以根据输入序列进行预测或分类等任务。

2、LSTM算法原理

Memory Cell 接受两个输入，即上一时刻的输出值和本时刻的输入值，由这两个参数先进入遗忘门，得到决定要舍弃的信息（即权重较小的信息）后，再进入输入门，得到决定要更新的信息（即与上一Cell相比权重较大的信息）以及当前时刻的Cell状态（候选向量，可理解为中间变量，存储当前 Cell State 信息），最后由这两个门（遗忘门，输入门）的输出值进行组合与当前时刻Cell状态进行叠加，到分别的长时和短时信息，最后进行存储操作及对下一个神经元的输入，下图介绍了LSTM在网络中是如何工作的[1]。

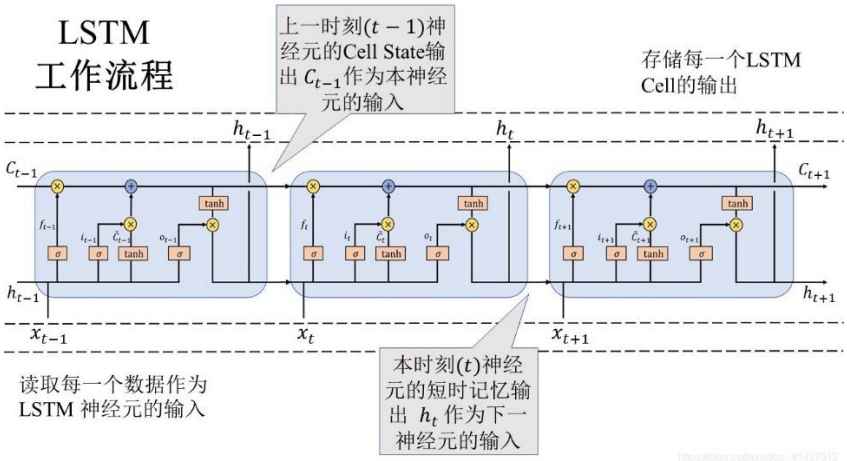

图2 LSTM模型工作原理

其中输入门、遗忘门、输出门的模型公式如下：

1. 遗忘门

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

2. 输入门

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

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_c)$$

以及t时刻的Cell状态(长时)方程：

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

3. 输出门

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

$$h_t = o_t \cdot \tanh(C_t)$$

其中f是门的激活函数，h是Cell输出的激活函数，$\sigma(z)$和$\tanh(z)$分别为Sigmoid激活函数和tanh激活函数。

$$\sigma(z)=\frac{1}{1+e^{-z}}=\frac{1+\tanh(\frac{z}{2})}{2}$$

$$\tanh(z)=\frac{e^{z}-e^{-z}}{e^{z}+e^{-z}}$$

3、LSTM模型优势

LSTM的主要优势包括：

1、处理长期依赖：LSTM通过门机制允许网络选择性地存储和读取信息，有效地解决了传统RNN中的梯度消失和梯度爆炸问题，能够处理长序列数据并捕捉到长期依赖关系。

2、上下文记忆：LSTM通过记忆单元的设计可以有效地存储和保持之前的信息，能够捕捉到更多的上下文信息，有利于理解和处理序列数据中的语义和语境。

3、灵活性：LSTM可以适应不同长度的输入序列，并且可以处理变长序列数据。这种灵活性使得LSTM在处理自然语言文本、语音信号等实际应用中非常有用。

4、并行计算：LSTM中的门机制使得网络的计算可以并行进行，提高了模型的训练和推断效率，适用于大规模数据和计算资源的应用场景。

总之，LSTM作为一种强大的序列模型，在处理长序列数据和捕捉长期依赖关系方面具有显著的优势，被广泛应用于多个领域，为解决序列数据任务提供了有效的解决方案。

三、实验结果与分析

本实验使用Python进行编程实现，主要的实验内容如下：

1、 数据预处理

以作家金庸的16本中文小说作为语料库，分别对每本小说进行了数据预处理，即删除了停词、英文大小写字母、广告等冗余内容，并在初步运行程序后根据实验结果更新了停词内容，降低非法字符或无意义字符对于算法本身的影响

```
with open(path, 'r', encoding="ANSI") as f:
    for line in f.readlines():
        data = f.read()
        data = data.replace(
            '本书来自 www.cr173.com 免费 txt 小说下载站\n 更多更新免费电子书请关注www.cr173.com', '')
        data = re.sub(u'[a-zA-Z0-9’!"#$%&\'()*+,-./:; ;<=>?@. . ?★、…【】 《》？“”‘’！ [\\]^_`{|}~「」『』 〈〉]+', '', data)
        data = data.replace(' ', '')
        data = data.replace('\n', '')
        data = data.replace('\t', '')
        line = line.replace('\u3000', '')
        for item in stopwords:
            data = data.replace(item, '')
```

2、 语句分割与字词标号

使用jieba对于导入的语料库进行分词，并对于每一个不同的字或词进行标号，同时将字或词及其标号存储在一个字典里，后续将用标号代替语料库中出现的字或词进行训练。

3、 建立LSTM模型

通过python中的第三方库torch建立LSTM模型，将语料库以“词”为基本单元输入到模型中进行训练，模型的参数如下：字词特征数embed_size、神经元数量hidden_size、隐藏层数量num_layers、语句序列长度seq_length、训练次数

num_epochs、学习率learning_rate。本实验中模型的参数设定如下：embed_size=16、hidden_size=32、num_layers=5、num_epochs=200、batch_size=100、seq_length=30、learning_rate = 0.01。

4、 实验结果与分析

从语料库中选取部分片段的开头，输入进LSTM模型中生成相应的语句，实验结果如下。

输入语句：

龚光杰大踏步过来，伸剑指向段誉胸口

期望输出语句：

龚光杰大踏步过来，伸剑指向段誉胸口，喝道：“你到底是真的不会，还是装傻？”段誉见剑尖离胸不过数寸，只须轻轻一送，便刺入了心脏，脸上却丝毫不露惊慌之色，说道：“我自然是真的不会，装傻有什么好装？”

模型输出语句：

龚光杰大踏步过来，伸剑指向段誉胸口道理，使异，也说他引诱不成。心下说想要早过，当时即难过，说道：“姑娘不解，你做慕容公子是那坏话，我汪帮主带回家属下脸面余地。也段延庆给我却须送我骂了褚兄弟。此处好大，是半点她成为耳大一个人，此刻流下泪来便向……老命什么，你别生气。”

经对比，训练结果的语句风格与语料库中的相似，但其中的逻辑性仍不足，其原因可能有以下三个：

1. 由于语料库内容较多，逐次训练耗时很长，为了能有效的得出测试结果在实验中降低了训练次数，可能导致模型欠拟合。
2. 模型参数设定不够准确，由于实验中所改变的参数与实验结果并没有直接联系，故在参数设定时具有较大的随机性，后续应再对参数进行单独训练或者采用更优的模型组合。
3. 语料库中的语言风格与句式较为复杂，中文的一词多义以及词性的变化在模型训练的过程中容易使模型混淆多个词语之间的关系，同时文章的中文类文言文句式在语义上省略了较多的内容，使得模型在对语句进行复原时的难度更大。

四、结论

本实验以作家金庸的16本中文小说作为语料库，分别对每本小说进行了数据预处理、字词分割和字词标号，实验中建立了LSTM长短期记忆模型，通过选取合适的模型参数并输入字词对应的编号序列对模型进行训练，对模型生成的语句结果进行了展示并对模型性能和改进方向进行了总结与分析。

参考资料

[1] LSTM公式详解&推导 https://blog.csdn.net/qq_41437512/article/details/113541031