基于 LSTM 的文本生成模型

陆旭军 ZY2203320
18376486@buaa.edu.cn

摘要

LSTM 常常用于解决长期依赖问题，有效地捕捉和利用输入序列中的长期依赖关系。本文使用 LSTM 模型进行文本生成，通过构建字典和将文本转换为索引序列，将文本数据转化为模型可接受的形式。使用 LSTM 模型进行训练，通过优化器和损失函数迭代更新模型参数，使模型能够预测下一个单词。训练过程中采用梯度裁剪和 Adam 优化器来防止梯度爆炸和加速训练。在生成文本阶段，代码提供了两种方式。一种是随机选择一个单词作为输入，然后使用模型生成指定长度的文本。另一种是通过自定义输入文本，将其转换为索引序列，并利用模型生成与输入相关的文本。LSTM 模型通过记忆和遗忘机制，可以在生成文本时保持一定的上下文和语义连贯性，这种方法在自然语言处理和文本生成任务中有广泛的应用，如机器翻译、对话系统等。

引言

1.LSTM 总结框架[1]

LSTM（长短期记忆）是一种循环神经网络（RNN）的变体，专门用于处理序列数据的建模和预测。相比于传统的 RNN 结构，LSTM 引入了一种特殊的记忆单元，可以更好地解决长期依赖问题。LSTM 在许多自然语言处理（NLP）和时间序列建模任务中取得了很大的成功。

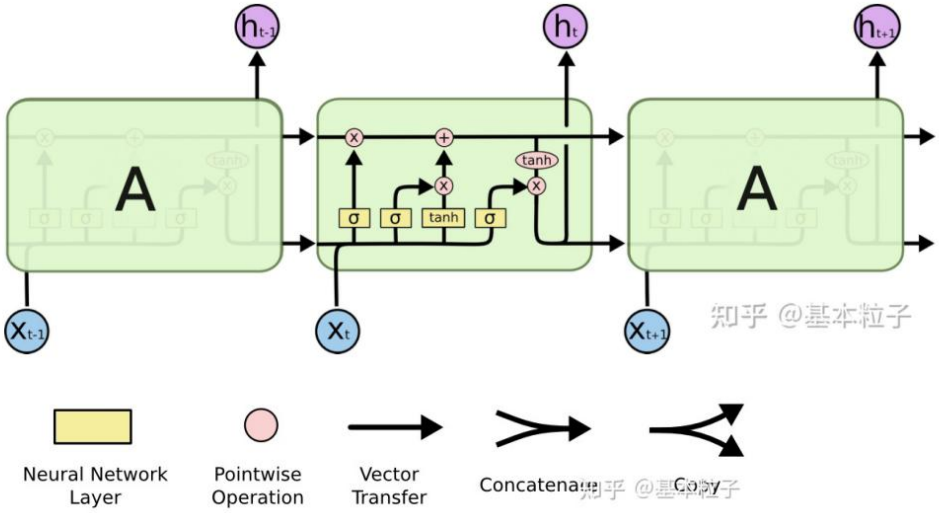


图 LSTM 总体框架及符号意义

Neuial Network layer:一层神经网络，也就是 w^T x + b 的操作。区别在于使用的激活函数不同，σ 表示的是 softmax 函数，他是将数据压缩到[0，1]范围内，如下图所示；tanh 表示的是双曲正切激活函数，他把数据归一化到[-1，1]之间。

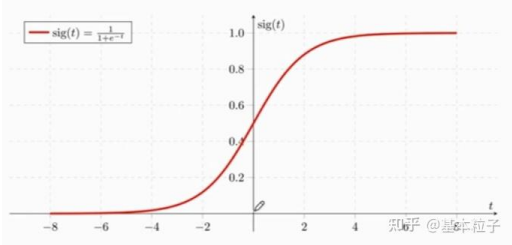


图 sigmoid 函数

Pointwise Operation：这个是两个矩阵按位操作，如果是 X 号表示，这两个维数相同的矩阵，每个位置相同的元素相乘放到新矩阵的该位置上。

Vector Transfer：矩阵传递

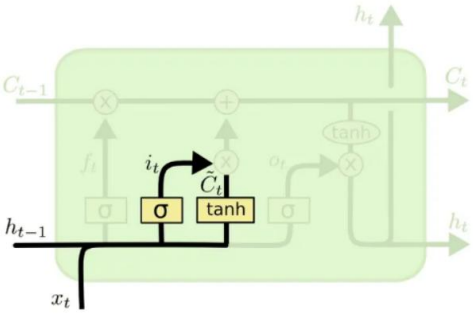
Concatenate：矩阵连接，两个矩阵不做任何计算，只是连接在一起，比如原来 A10 维，B5 维，连接之后 15 维，就像贪食蛇一样。

Copy：一个矩阵变成两个一模一样的。

2.LSTM 关键组件[1]

LSTM 的核心思想是通过使用称为"门"的结构来控制信息的流动和遗忘，从而解决了传统 RNN 面临的梯度消失和梯度爆炸的问题。一个标准的 LSTM 单元由以下几个关键组件组成：

输入门（Input Gate）：决定是否将输入信息存储到记忆单元中。它由一个 Sigmoid 激活函数和一个点积操作组成，用于控制输入的权重。

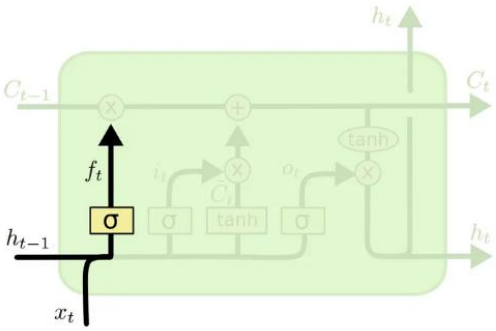


$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$

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

图 输入门及其公式

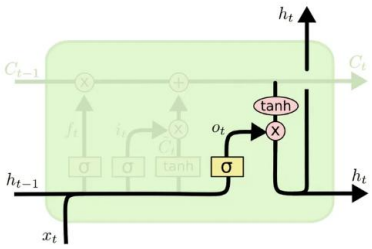
遗忘门（Forget Gate）：决定是否从记忆单元中删除特定的信息。它由一个 Sigmoid 激活函数和一个点积操作组成，用于控制遗忘的权重。



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

图 遗忘门及其公式

输出门（Output Gate）：决定从记忆单元中输出的信息。它由一个 Sigmoid 激活函数和一个点积操作组成，用于控制输出的权重。



$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

图 输出门层及其公式

记忆单元（Cell State）：用于存储和传递信息的主要部分。它通过输入门、遗忘门和输出门的组合来更新和控制信息的流动。

隐藏状态（Hidden State）：是 LSTM 的输出。它是基于当前输入和前一个隐藏状态计算得到的。

在 LSTM 中，每个时间步骤都会接收一个输入和一个隐藏状态作为输入，并输出一个新的隐藏状态和一个输出。这样，LSTM 可以在时间上处理序列数据，并记住长期的上下文信息。

LSTM 的训练过程通常使用反向传播算法和梯度下降来最小化损失函数。通过将 LSTM 应用于大量的训练数据，它可以学习到序列数据的模式和规律，并用于生成文本、语言建模、机器翻译、情感分析等各种 NLP 任务。

实验过程

代码是基于 PyTorch 实现的使用 LSTM（长短期记忆）架构的语言模型，用于生成文本。以下是详细实验过程说明：

1.定义辅助类：

Dictionary 类：用于构建词典，包括初始化字典和添加单词的方法。初始化一个空字典和两个映射字典，用于单词到索引和索引到单词的转换；add_word 方法：将单词添加到字典中，并建立单词到索引的映射关系。

Corpus 类：用于处理语料库，包括获取文件列表、处理文本行、处理文件和获取数据集的方法。初始化一个字典对象和文件路径列表；get_file 方法：获取指定目录下的所有 txt 文件路径列表；process_line 方法：处理文本行，去除空格和制表符；process_file 方法：读取文件并处理文件内容；get_data 方法：获取数据集并构建字典。

2.定义模型类：

LSTMmodel 类：定义了一个 LSTM 模型，包括嵌入层、LSTM 层和线性层，并实现了前向传播方法。初始化一个嵌入层、LSTM 层和线性层；forward 方法：前向传播过程，将输入经过嵌入层、LSTM 层和线性层，返回输出和隐藏状态。

3.主函数部分：

设置批量大小、设备使用 GPU 或 CPU、语料库对象、数据集、词汇表大小。

设置嵌入维度、隐藏层大小、隐藏层数、训练轮数、序列长度和学习率。

创建 LSTM 模型，并定义损失函数和优化器。

进行训练循环，包括迭代数据集、前向传播、计算损失、反向传播和梯度更新。

保存模型到指定路径。

使用生成文本函数生成指定长度的文本。

将生成的文本写入文件。

实验结果

1.输出结果

白衣尼哀哭了良久，站起身来，抱住树干，突然全身颤抖，昏晕了过去，身子慢慢软垂下来。韦小宝吃了一惊，急忙扶住，叫道：“师太，师太，快醒来。”康亲王笑道：“咱们今日庆贺韦大人高升，按理他该坐首席才是。不过他是本宅主人，只好坐主位了。”韦小宝奇道：“什么本宅主人？”康亲人王笑道：“这所宅子，是韦大人的子爵府。做哥哥的跟你预备的。车夫、厨子、仆役、婢女，全都有了。匆匆忙忙的，只怕很不周全，兄弟见缺了什么，只管吩咐，命人到我家里来搬便是。”韦小宝惊喜交集，自己帮了康亲王这个大忙，不费分文本钱，不担丝毫风险，虽然明知他定有酬谢，却万想不到竟会送这样一件重礼，一时说不出话来，只道：“这……这个……那怎么可以？”康亲王捏了捏他手，说道：“咱哥儿俩是过命的交情，哪还分什么彼此？来来来，大伙儿喝酒。哪一位不喝醉的，今日不能放他回去。”这一席酒喝得尽欢而散。韦小宝贵为子爵，大家又早知他那太监是奉旨假扮的，便不能再回宫住宿国。这一晚睡在富丽华贵的卧室之中，放眼不是金器银器，就是绫罗绸缎，忽想：“他奶奶的，我如在这子爵府开座妓院，十间丽春院也比下去了。”宝依样葫芦的说来，果然也引得茅十八放怀大笑。韦小宝继续说道：“沐王爷摆开阵仗，黄黎洲等都吃了一惊，均想：“连这人都要知道了，只怕又是一场大这句话一入耳，韦小宝喜得便想跳了起来，就可惜手足被绑，难以跳跃。又听得阿珂的声音说道：“他……他没穿衣服，不能救啊！”韦小宝大怒，心中大骂：“死丫头，我不穿衣服，为什么不能救，难道定要穿了衣服，才能救么？你不救老公，就是谋杀亲夫。自己做小寡妇，好开心么？”只听九难道：“你闭着眼睛，去割断他手脚的绳索，不就成了？”阿珂道：“不成啊。我闭着眼睛，瞧不见，倘若……倘若碰到他身子，那怎么办？师父，还是你去救他罢。”九难怒道

2.结果分析

生成的结果描述了白衣尼哀哭、韦小宝受康亲王帮助获得子爵府、康亲王与韦小宝举杯庆贺、韦小宝惊喜得到子爵府的赠礼等情节。生成的文本描述了康亲王为韦小宝准备的子爵府，形容其富丽华贵。子爵府内装饰豪华，摆满了金器银器，绫罗绸缎等珍贵物品。文本中表达了韦小宝对于康亲王的帮助和赠礼表示感激，并对子爵府的惊喜表现出开心的情绪。康亲王的友情和慷慨也得到了韦小宝的深深感激。整个文本中还融入了一些幽默和轻松的描写和对话，增添了愉快的氛围。

文本结果展示了一个喜庆的场景，突出了康亲王对韦小宝的友情和慷慨，以及韦小宝对于子爵府的惊喜和开心。生成的结果具有一定的情节连贯性和可读性，为读者呈现了一个活泼有趣的场景。

不过由于是限定词数，因此生成的语段戛然而止，由于时间关系，来不及优化，未来可以继续完善代码。

总结

LSTM 通过引入门结构和记忆单元，有效地解决了传统 RNN 中的梯度消失和梯度爆炸问题，并成为处理序列数据的强大工具。通过 LSTM 生成文本的过程涉及了语料库的准备、模型的训练和文本的生成。LSTM 模型的记忆单元和门控机制使其能够处理长序列数据，并生成具有上下文连贯性的文本。然而，生成的文本结果仍然受到训练数据和模型参数的影响，

需要进行适当的调优和评估，以获得更好的文本生成效果。结合生成的场景化的、情绪化的文本来看，LSTM 取得了显著的成果。

参考文献

[1] https://zhuanlan.zhihu.com/p/518848475