清华大学电子工程系

媒体与认知 课堂 2

2022-2023 学年春季学期

作业 4

2023 年 5 月 17 日

本次作业内容包含序列建模任务中常用的隐含马尔可夫模型、循环神经网络和 Transformer。具体任务分为理论部分、编程部分以及作业报告。其中理论部分包含第 1，2 题，所有同学均需完成，答案附在作业报告中；编程部分包含第 3、4 题，采用 Transformer 实现 GPT（Generative Pre-training），并完成文本生成任务。第 5 题为撰写作业报告。已确认自选课题的同学需完成第 6 题。

1. 单选题（15 分）
2. 计算题（15 分）
3. 完成基于 GPT 的文本生成任务的程序代码（30 分）
4. 训练/预测/可视化（30 分）
5. 撰写作业报告（10 分）
6. 汇报自选课题进度（70 分）*

理论部分

1 单选题（15 分）

1.1 给定 HMM 的模型参数 λ，隐含状态总数为 N。设给定观测序列 O 的条件下，在第 t 时刻处于状态 S_i 的概率为 $\gamma_t(i) = P(q_t = S_i|O, \lambda)$。若已经计算得到所有前向变量 $\alpha_t(i)$ 和后向变量 $\beta_t(i)$，则下列计算 $\gamma_t(i)$ 的方法中正确的是哪项？

(A) $\gamma_t(i) = \alpha_t(i)\beta_t(i)$

(B) $\gamma_t(i) = \alpha_t(i)\beta_t(i) \ / \ \sum_{j=1}^{N} \alpha_t(j)\beta_t(j)$

(C) $\gamma_t(i) = \alpha_t(i) + \beta_t(i)$

(D) $\gamma_t(i) = (\alpha_t(i) + \beta_t(i)) \ / \ \sum_{j=1}^{N}(\alpha_t(j) + \beta_t(j))$

1.2 对于一个参数为 $\lambda = \{\pi, A, B\}$ 的 HMM，隐含状态总数为 5，且观测序列 O 长度为 10。若已经计算得到所有前向变量 $\alpha_t(i)$ 和后向变量 $\beta_t(i)$，则下列关于 $P(q_5 = S_2, q_6 = S_4, O|\lambda)$ 的计算方式中正确的是哪项？

(A) $\alpha_5(2)a_{24}$

(B) $b_4(O_6)\beta_6(4)$

(C) $\alpha_5(2)a_{24}b_4(O_6)$

(D) $\alpha_5(2)a_{24}b_4(O_6)\beta_6(4)$

1.3 考虑下图所示的 RNN，其运算过程为

$$\boldsymbol{h}_t = \phi\left(W\boldsymbol{x}_t + U\boldsymbol{h}_{t-1}\right)$$
$$\boldsymbol{y}_t = \mathrm{softmax}\left(V\boldsymbol{h}_t\right)$$

输入特征序列 $X = \{\boldsymbol{x}_t\}_{t=1}^{3}$ 包含三个时刻的数据。每个时刻的输入特征向量 $\boldsymbol{x}_t \in \mathbb{R}^2$，隐含状态 $\boldsymbol{h}_t \in \mathbb{R}^4$，输出特征 $\boldsymbol{y}_t \in \mathbb{R}^3$。则该 RNN 的网络参数量是多少？

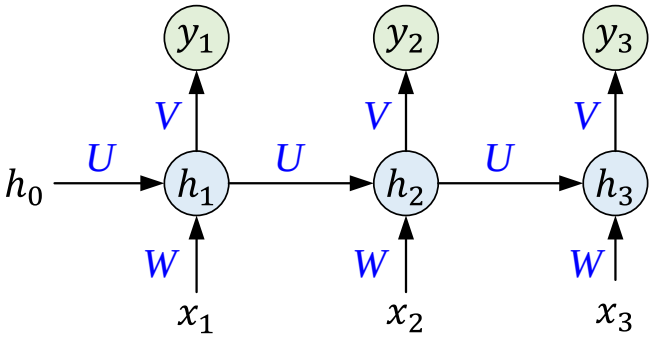

图 1: 沿时间展开后的 RNN 示意图

(A) 12

(B) 36

(C) 72

(D) 108

1.4 在本题中我们对传统 RNN 难以学习长距离相关信息的问题进行一个简单的讨论:

对 RNN 的计算过程进行简化，考虑一个暂不采用激活函数以及输入 $\boldsymbol{h}_0$ 的 RNN:

$$\boldsymbol{h}_t = U\boldsymbol{h}_{t-1} = U\left(U\boldsymbol{h}_{t-2}\right) = ... = U^t\boldsymbol{h}_0$$

其中 U^t 为 t 个 U 矩阵连乘。若矩阵 U 存在如下特征值分解:

$$U = Q\Lambda Q^\top$$

其中 Q 为单位正交矩阵（每一列为模长为 1 的特征向量），$Q^\top$ 为 Q 的转置，Λ 为特征值对角矩阵，则上述的 RNN 计算过程可表示为:

$$\boldsymbol{h}_t = Q\Lambda^t Q^\top \boldsymbol{h}_0$$

通过计算可以得到:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial h_0} &= \frac{\partial \mathcal{L}}{\partial h_t}\frac{\partial h_t}{\partial h_{t-1}}\cdots\frac{\partial h_1}{\partial h_0} \\ &= U^t\frac{\partial \mathcal{L}}{\partial h_t} = Q\Lambda^t Q^\top\frac{\partial \mathcal{L}}{\partial h_t}\end{aligned}$$

下列说法中正确的是:

(A) 当 Q 的特征值 $\lambda_i < 1$ 时容易造成梯度消失问题

(B) 当 Q 的特征值 $\lambda_i = 1$ 时容易造成梯度消失问题

(C) 当 Q 的特征值 $\lambda_i > 1$ 时容易造成梯度消失问题

1.5 在注意力机制中，设 query 为 $Q \in \mathbf{R}^{N\times C}$，key 为 $K \in \mathbf{R}^{M\times C}$，其中 N, M 分别为 query 和 key 的序列长度，C 为特征通道数。则注意力系数 A 和注意力机制的输出 O 的大小为:

(A) $A \in \mathbf{R}^{N\times N}, O \in \mathbf{R}^{N\times C}$

(B) $A \in \mathbf{R}^{M\times N}, O \in \mathbf{R}^{M\times C}$

(C) $A \in \mathbf{R}^{N\times M}, O \in \mathbf{R}^{N\times C}$

(D) $A \in \mathbf{R}^{M\times M}, O \in \mathbf{R}^{M\times C}$

2 计算题（15 分）

2.1 隐含马尔可夫模型的解码

某手机专卖店今年元旦新开业，每月上旬进货时，由专卖店经理决策，采用三种进货方案中的一种：高档手机 (H)，中档手机 (M)，低档手机 (L)。

当月市场行情假设分为畅销 (S_1) 和滞销 (S_2) 两种。畅销时，三种进货方案的概率分别为 0.4, 0.4, 0.2；滞销时，三种进货方案的概率分别为 0.2, 0.3, 0.5。

某月份市场行情为畅销，下一个月份为畅销和滞销的概率分别为 0.6 和 0.4；某月份市场行情为滞销，下一个月份为畅销和滞销的概率分别为 0.5 和 0.5。

开业第一个月市场行情为畅销和滞销的可能性均为 0.5。

(1) 如果我们采用隐含马尔可夫模型 (HMM) 对该专卖店进货环节建模，请写出 HMM 对应的参数 $\lambda = \{\pi, A, B\}$。

(2) 在第一季度中，采购业务员执行的进货方案为 “高档手机，中档手机，低档手机”，即观测序列为 H, M, L。请利用 Viterbi 算法推测前三个月的市场行情。

编程部分

编程部分包括第 3、4 题。在本次任务中，我们将基于 Transformer 构建 GPT（Generative Pre-training），使用给定的中文文本数据集（全宋词）训练语言模型，然后利用训练好的模型生成文本。我们还将探索残差连接和位置编码在模型中的作用。最后，我们将对模型中的注意力系数进行可视化。详细说明请参阅习题课课件。

本次作业需求 python 版本至少为 3.8，若 python 版本过低，请及时更新。本次作业将用到 bertviz 库。若未安装，请执行

pip install bertviz

进行安装。

3 完成基于 GPT 的文本生成任务代码程序（30 分）

通过解压 hw4.zip 可以得到本次编程作业的程序清单如下：

文件或目录	说明	注意事项
\data	存放本次作业所用数据集	
\quansongci	中文文本数据集（全宋词）	可以自行选择其他合适的文本
\vis	用于可视化的文本	可以自行添加用于可视化的文本
\workdirs	存放训练好的模型	请勿修改
prepare.py	数据预处理	已完成代码
dataset.py	数据读取	已完成代码
model.py	网络结构定义	**需要完成代码**
train.py	训练程序	已完成代码
sample.py	文本生成程序	已完成代码
attnvis.ipynb	可视化 jupyter 文件	已完成代码

本次编程作业需要同学完成 model.py 中的多头注意力机制的计算以及 GPT 模型的前向计算，每处需要完成的地方都有代码提示和步骤提示，需要完成的代码清单如下：

TODO 1: 完成多头自注意力机制的计算。

序号	行号	说明
Step 1	Line 62	请按照提示完成多头自注意力中的 q,k,v 计算
Step 2	Line 67	请按照提示对 q,k,v 进行变形
Step 3	Line 71	请按照提示分步完成多头自注意力的计算
Step 4	Line 89	请按照提示对多头自注意力结果进行变形
Step 5	Line 93	请按照提示完成输出结果的计算

假设输入特征为 $X \in \mathbf{R}^{B\times L\times C}$，其中 B 为批次大小 (batch_size)，L 为每个样本序列的长度 (seq_len), $C = h \times d$，h 为多头注意力头数 (num_heads), d 为单头注意力所用的特征维度；注意力掩码为 M，其计算步骤如下:

Step 1: 计算 $q, k, v \in \mathbf{R}^{B\times L\times C}$

$$q = \text{Concat}(q_1, ..., q_h) = \text{Concat}(XW_1^Q, ..., XW_h^Q) = XW^Q$$

$$k = \text{Concat}(k_1, ..., k_h) = \text{Concat}(XW_1^K, ..., XW_h^K) = XW^K$$

$$v = \text{Concat}(v_1, ..., v_h) = \text{Concat}(XW_1^V, ..., XW_h^V) = XW^V$$

Step 2: 对 q, k, v 进行变形，变形后尺寸为 $B \times L \times h \times d$

$q = reshape(q), k = reshape(k), v = reshape(v)$

Step 3: 计算每个注意力头的输出 $head_i \in \mathbf{R}^{B \times L \times d}$

$head_i = \text{Attention}(q_i, k_i, v_i) = \text{Softmax}(q_i k_i^T / \sqrt{d} + M) v_i, i = 1, ..., h$

Step 4-5: 得到输出结果，$head_i$ 尺寸为 $B \times L \times C$

$\text{MultiHead}(q, k, v) = \text{Concat}(head_1, ..., head_h) W^o$

TODO2: 完成 GPT 的前向计算。

序号	行号	说明
Step 1	Line 192	请按照提示完成单词位置编号的生成
Step 2	Line 195	请按照提示得到词嵌入向量和位置编码
Step 3	Line 198	请按照提示初始化 Transformer 网络的输入
Step 4	Line 201	请按照提示生成自注意力掩码
Step 5	Line 210	请按照提示完成 Transformer 计算并保存 Transformer 层输出的注意力系数
Step 6	Line 214	请按照提示完成预测结果的计算

假设输入的文本序列为 $U \in \mathbf{R}^{B \times L}$，其中 B 为批次大小 (batch_size)，L 每个样本序列的长度 (seq_len)，U 中的每个元素为对应单词在词汇表中对应的编号；C 表示特征通道数，n 表示 Transformer 层数，S 表示词汇表中的单词总数。GPT 的前向计算步骤如下：

Step 1: 生成位置编号 P，长度为 L

$P = \{0, 1, ..., L-1\}$

Step 2: 得到词嵌入向量 $E_U \in \mathbf{R}^{B \times L \times C}$ 和位置编码 $E_P \in \mathbf{R}^{L \times C}$

$E_U = \text{TokenEmbed}(U), E_P = \text{PosEmbed}(P)$

Step 3: 初始化 Transformer 网络的输入，使用 if 语句判断是否使用位置编码

$h_0 = E_U$ if no_pos else $E_U + E_P$

Step 4: 生成自注意力掩码

$M = triu(\mathbf{1}_{L \times L}), \mathbf{1}_{L \times L}$为全 1 方阵

Step 5: Transformer 网络

$h_t = \text{TransformerLayer}(h_{t-1}, M), t = 1, ..., n$

Step 6: 得到下一个单词的预测概率，尺寸为 $B \times L \times S$

$(p(\hat{u}_2), ..., p(\hat{u}_{L+1})) = \text{Softmax}(\text{Linear}(h_n))$

4 训练/文本生成/可视化（30 分）

4.1 模型的训练与测试

本次作业要求同学们在中文文本数据集（全宋词）上进行训练，然后利用训练好的模型进行文本生成，在作业报告中记录训练过程中训练集和验证集上 loss 和困惑度 perplexity 的变化，并从生成的文本中选择质量较好和较差的文本进行展示。

在模型训练之前，需要统计数据集中的单词作为词汇表，并将单词转化为其在词汇表中对应的编号。**数据预处理的命令如下**：

```
python prepare.py --data_root data/quansongci
```

训练模型的命令示例如下。通过--ckpt_path 参数可以指定指定模型的保存目录，便于后续测试和可视化：
1）使用中文文本数据集（全宋词）训练：

```
python train.py --ckpt_path workdirs/quansongci
```

训练过程和验证集上 loss 以及困惑度的变化会在训练完成后以图片形式显示，并自动保存在 ckpt_path 文件夹中。关掉图片显示窗口后，程序方能退出。

默认数据集存放路径为 data/quansongci，如果使用其他数据集训练，可以通过--data_root 参数指定数据集目录；默认训练的迭代次数为 2000，可以通过--iters 参数指定训练的迭代次数；默认训练的批次大小为 16，可以通过--batchsize 参数指定训练的批次大小；默认训练时每个样本的序列长度为 64，可以通过--max_seq_len 参数指定训练时每个样本的序列长度。

本次作业实现的 GPT 模型包含 4 层 Transformer Layer，特征通道数为 128，多头自注意力中的注意力头数为 4，每个注意力头的特征通道数为 128/4=32，由于模型比较简单因此 Dropout 设置为 0。

使用训练好的模型进行文本生成的示例如下。通过--ckpt_path 参数指定待加载模型的保存目录；通过--start 参数指定初始文本（默认为 “\n”）。
1) 默认配置下生成样本：

```
python sample.py --ckpt_path workdirs/quansongci
```

2) 指定初始文本生成样本：

```
python sample.py -ckpt_path workdirs/quansongci --start +++ 清平乐
```

默认数据集存放目录为 data/quansongci，如果使用其他数据集训练，可以通过--data_root 参数可以指定数据集目录；默认配置下生成 10 次样本，可以通过--num_samples 参数指定文本生成次数；每次生成 500 个单词, 可以通过--max_new_tokens 参数指定每次生成文本的单词数。

4.2 探究位置编码和残差连接在模型中的作用

在本节中，选择全宋词作为数据集，探究位置编码和残差连接对模型的影响。在默认参数下，位置编码和残差连接都是启用的。请同学们分别关闭位置编码和残差连接训练模型，比较不同参数设置下训练过程中训练集和验证集上 loss 和困惑度 perplexity 的变化以及文本生成的质量。
1) 关闭位置编码的训练与文本生成命令：

python train.py --ckpt_path workdirs/quansongci_no_pos --no_pos

python sample.py --ckpt_path workdirs/quansongci_no_pos

1) 关闭残差连接的训练与测试命令：

python train.py --ckpt_path workdirs/quansongci_no_res --no_res

python sample.py --ckpt_path workdirs/quansongci_no_res

4.3 可视化

本次作业的可视化使用 jupyter 文件 attnvis.ipynb。在 attnvis.ipynb 文件中可以修改数据集目录 (data_root)、模型的保存目录 (ckpt_path) 和用于可视化的文本文件路径 (vis_text_path)，执行可视化程序即可显示文本中单词之间的注意力系数。同学们可以自行选择可视化的层数，将鼠标置于某一单词之上即可显示该单词与文本中所有单词的注意力系数，在报告中试分析注意力系数与文本之间的关系。

5 撰写作业报告（10 分）

将 hw4 目录和作业报告打包为一个文件（例如 *.zip）提交到网络学堂，文本数据（“data” 目录）、保存的模型文件（“workdirs” 目录）不必打包在内。作业报告中包括选择题答案，计算题的解题步骤及答案、任务 3、4 运行结果及分析，本次作业遇到的问题及解决方法，对本次作业的意见及

建议。推荐同学们使用随作业发布的 LaTex 模板 HW4-template 完成作业报告。

6 自选课题进度汇报（70 分）*

请已确认自选课题的同学，完成简短的自选课题工作进度汇报，例如，文献阅读、或者研究方案设计、或者原型系统搭建及实验结果等内容。

关于作业迟交的说明：由于平时作业计入总评成绩，希望同学们能按时提交作业。若有特殊原因不能按时提交，请在提交截止时间之前给本次作业责任助教发 Email 说明情况并给出预计提交作业的时间。对于未能提前说明原因的迟交作业，将酌情扣分。

本次作业责任助教为姚刚 (Email: yg19@mails.tsinghua.edu.cn)。

附录

程序利用 argparse 库进行参数设置，可以查看 train.py 和 sample.py 中可以调节的参数, 可参数说明如下表所示。

	参数	说明
共有参数	data_root	存放数据集的路径
	ckpt_path	存放训练模型的路径
	device	程序运行使用的设备，cpu 或 cuda
	model_name	使用的模型名称，默认为 “mygpt“：Transformer 层数为 4，特征通道数为 128，多头自注意力中的注意力头数为 4，每个注意力头的特征通道数为 128/4=32，Dropout 为 0
train.py	max_seq_len	输入文本序列长度，默认为 64
	iters	训练轮数，默认为 2000
	batchsize	训练批次大小，默认为 16
	no_res	控制是否使用残差连接
	no_pos	控制是否使用位置编码
sample.py	start	指定文本生成的初始文本
	num_samples	生成文本的样本数，默认为 10
	max_new_tokens	每次生成文本的长度，默认为 500