清华大学电子工程系

媒体与认知

2023-2024 学年春季学期

作业 4

2024 年 5 月 22 日

理论部分

1 单选题（15 分）

1.1 给定 HMM 的模型参数 $\lambda = \{\pi, A, B\}$，隐含状态总数为 3，观测序列长度为 5。若已经计算得到所有前向变量 $\alpha_t(i)$ 和后向变量 $\beta_t(i)$。在给定观测序列 O 的条件下，$t = 3$ 时刻处于状态 S_2 的概率 $\gamma_3(2) = P(q_3 = S_2|O, \lambda)$ 的计算方式为：

(A) $\alpha_3(2) + \beta_3(2)$

(B) $\alpha_3(2)\beta_3(2)$

(C) $(\alpha_3(2) + \beta_3(2)) \ / \ \sum_{j=1}^{N}(\alpha_3(j) + \beta_3(j))$

(D) $\alpha_3(2)\beta_3(2) \ / \ \sum_{j=1}^{N}\alpha_3(j)\beta_3(j)$

1.2 给定 HMM 的模型参数 $\lambda = \{\pi, A, B\}$，若已经计算得到所有前向变量 $\alpha_t(i)$ 和后向变量 $\beta_t(i)$。在给定观测序列 O 的条件下，t 时刻处于状态 S_i 且 $t+1$ 时刻处于状态 S_j 的概率 $P(q_t = S_i, q_{t+1} = S_j|O, \lambda)$ 的计算方式为：

(A) $\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j) \ / \ P(O|\lambda)$

(B) $\alpha_t(i)a_{ij}\beta_{t+1}(j) \ / \ P(O|\lambda)$

(C) $\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)$

(D) $\alpha_t(i)a_{ij}\beta_{t+1}(j)$

1.3 本题对传统循环神经网络 (recurrent neural network, RNN) 训练中容易出现的梯度消失和梯度爆炸的问题进行一个简单的讨论:

一个 RNN 沿时间展开的示意图如图 1 所示。

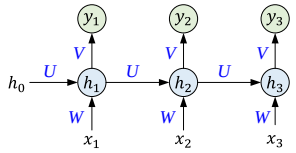

图 1: RNN 沿时间展开的示意图

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 x 的 RNN，$t = 0$ 时刻隐含状态取值为 $\boldsymbol{h}_0$:

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

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:

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

在训练阶段的误差沿时间的反向传播过程 (back-propagation through time, BPTT) 中，通过计算可以得到:

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

在序列长度较长，即 t 较大时，下列说法中正确的是（单选）:

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

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

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

(D) 当 U 的特征值 $\lambda_i < 1$ 时容易造成梯度爆炸问题

1.4 下列有关 Transformer 的说法错误的是：

(A) 通过将位置编码添加到输入序列的嵌入表示中，Transformer 可以在计算注意力权重时考虑到序列的顺序，从而正确地捕捉序列中元素之间的依赖关系

(B) Transformer 编码器的主要作用是编码输入序列中的信息，以供后续的解码步骤使用

(C) Transformer 中的自注意力机制的计算复杂度是线性的，因此在处理长序列时非常高效

(D) 在训练阶段，Transformer 解码器中引入的注意力系数掩码用于防止解码器看到未来的位置，以便并行计算

1.5 在采用缩放点积形式评分函数的注意力机制中，设 query 为 $Q = (1, 0, 1, 1)$，key 为 $K = \begin{pmatrix} 0,0,0,2 \\ 2,0,1,0 \\ 2,1,2,1 \end{pmatrix}$。则注意力系数 $A = \mathbf{softmax}(QK^T/\sqrt{d})$ 为（保留两位小数）：

(A) $(0.04, 0.12, 0.84)$

(B) $(0.14, 0.23, 0.63)$

(C) $(2.00, 3.00, 5.00)$

(D) $(1.00, 1.50, 2.50)$

2 计算题（15 分）

2.1 隐含马尔可夫模型

暑假中，小 E 每天进行一项体育活动，包括跑步（R）、游泳（S）和打球（B），所选择的体育活动受某种潜在因素（如心情）的影响。小 E 每天把进行体育活动的照片发至微信朋友圈，我们可以根据观测信息推测该潜在因素的状态。

假设该潜在因素分为 S_1 和 S_2 两种状态。在 S_1 时，小 E 选择三种体育活动的概率分别为 0.6，0.2，0.2；在 S_2 时，小 E 选择三种体育活动的概率分别为 0.1，0.6，0.3。

该潜在因素的变化也有一定规律，若某天处于 S_1 的状态，第二天处于 S_1 和 S_2 的状态的概率分别为 0.5，0.5；若某天处于 S_2 的状态，第二天处于 S_1 和 S_2 的状态的概率分别为 0.6，0.4。

暑假第一天处于 S_1 和 S_2 的状态的概率均为 0.5。

(1) 采用隐含马尔可夫模型 (HMM) 对小 E 暑假体育活动安排进行建模，请写出 HMM 对应的参数 $\lambda = \{\pi, A, B\}$。

(2) 假设暑假第 1、2、3 天小 E 所进行的体育活动依次为跑步（R）、打球（B）和游泳（S），请计算出现该观测序列的概率。

(3) 在 (2) 的条件下。请利用 Viterbi 算法推测暑假第 1、2、3 天最可能的隐含状态序列。

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。详细说明请参阅第四次习题课课件。

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 库进行可视化。若未安装，请执行

pip install bertviz

进行安装。

3 代码补全（30 分）

在本任务中，完成基于 GPT 的文本生成任务程序。

hw4 文件夹中包含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）	可以选择其他合适的文本
\vis	用于可视化的文本	可以添加其他合适的文本
\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	数据读取与处理	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、Transformer 解码器以及 GPT 模型的前向计算代码，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: 请按照提示完成 model.py 中多头自注意力机制的计算。

序号	行号	说明
Step 1.1	Line 63-65	多头自注意力中的 q,k,v 计算
Step 1.2	Line 69-77	对 q,k,v 进行变形
Step 1.3	Line 82-103	分步完成多头自注意力的计算
Step 1.4	Line 107	对多头自注意力计算结果进行变形
Step 1.5	Line 110	输出结果的计算

TODO 2: 请按照提示完成 model.py 中 Transformer 解码器的计算。

序号	行号	说明
Step 2.1	Line 150-157	多头注意力机制的计算
Step 2.2	Line 161-167	前馈神经网络 FFN 的计算

TODO 3: 请按照提示完成 model.py 中 GPT 的前向计算。

序号	行号	说明
Step 3.1	Line 233	字符位置编号的生成
Step 3.2	Line 236-237	得到词嵌入向量和位置编码
Step 3.3	Line 241-247	初始化 Transformer 网络的输入
Step 3.4	Line 251-256	Transformer 计算并保存 Transformer 层输出的注意力系数
Step 3.5	Line 261-262	预测结果的计算

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（本次作业中每个字符作为一个单词）作为词汇表。**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。手动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 1000，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。

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，在本次作业中，由于模型比较简单，Dropout 概率暂且设置为 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“+++”）。
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。

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

2) 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）

请同学们将代码和作业报告打包为一个文件（例如 *.zip）提交到网络学堂 (请勿将文本数据（“data” 目录）和保存的模型文件（“workdirs” 目录）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。如果同学们使用大模型来完成作业，请在作业中说明，并

指出使用大模型过程中出现的问题。推荐同学们使用随作业发布的 LaTex 模板 HW4-template.zip 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 32，Dropout 概率为 0
train.py	iters	训练轮数，默认为 1000
	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