哈尔滨工业大学

**自然语言处理**

**实验指导书**

**(2024年春季学期)**

实验一 从零开始实现基于Transformer的语言模型

## 一、实验内容

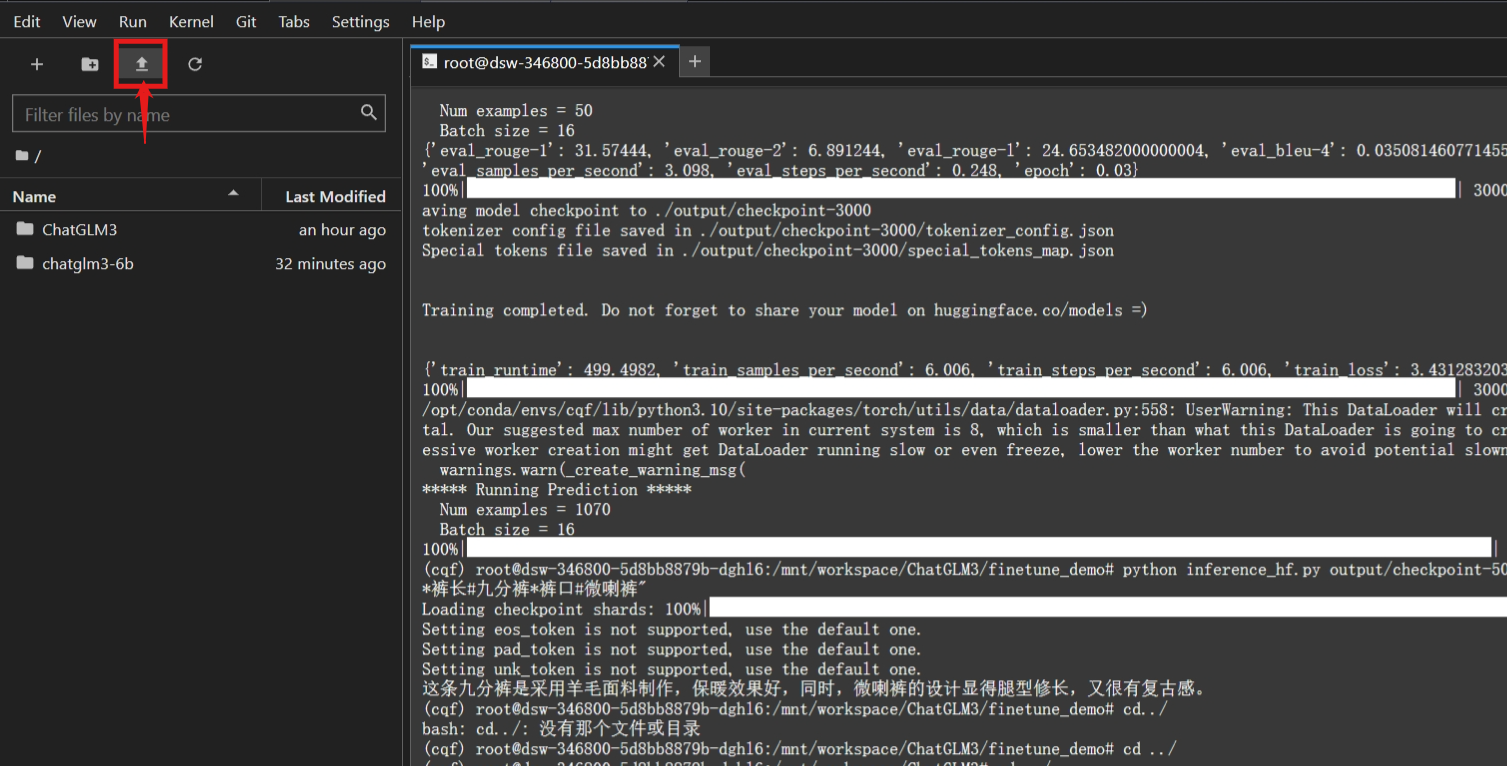
Transformer模型是一种通用的深度学习架构，已经在各种自然语言处理和其他领域的任务中取得了巨大成功。其中Decoder-only的Transformer模型（如GPT系列）在许多领域中发挥着重要作用。这种模型主要集中于生成任务。

在本实验中，请你基于实验提供的代码框架，实现一个Decoder-only的Transformer模型：GPT，并使用SkyPile中文预训练数据集的一个子集进行模型训练。

## 二、实验环境

本实验的实验环境为阿里云服务器提供的NVIDIA V100 ，16GB显存，每名同学请按照实验环境使用手册自行准备。建议在配置环境依赖时使用conda创建虚拟环境。Conda的虚拟环境（也称为Conda环境）是一种强大的工具，它可以帮助你管理和隔离项目的依赖关系，可以使你的开发工作更加高效和可靠，特别是当你需要处理复杂的依赖关系或者需要在多个项目之间切换时。

把students.zip文件上传到阿里云，再自行解压。



解压完进入students文件，**安装依赖后**就可以打开lab1.ipynb进行实验了。

## 三、实验要求

实验总分为20分，包括以下内容：

1. 基于提供的代码框架，实现GPT模型
   1. 实现Token Embedding（1分）
   2. 实现Positional Encoding（1分）
   3. 实现简易Attention Mask（4分）
   4. 实现多头自注意力（2分）
   5. 实现FFN层（2分）
   6. 实现Decoder Block（2分）
   7. 实现完整GPT（2分）
2. 使用SkyPile数据集和提供的训练框架，训练GPT模型，观察训练日志。（0.5分）
3. 加载训练后的模型，测试其生成能力。（0.5分）
4. （可选）增大模型规模、调整超参数，使模型拥有更好的生成能力
5. 撰写实验报告，简要描述设计思路、遇到的困难和解决方式等，提供程序运行效果的截图。（5分）

## 四、提交方式

请将代码（仅ipynb文件）、实验报告打包提交。