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

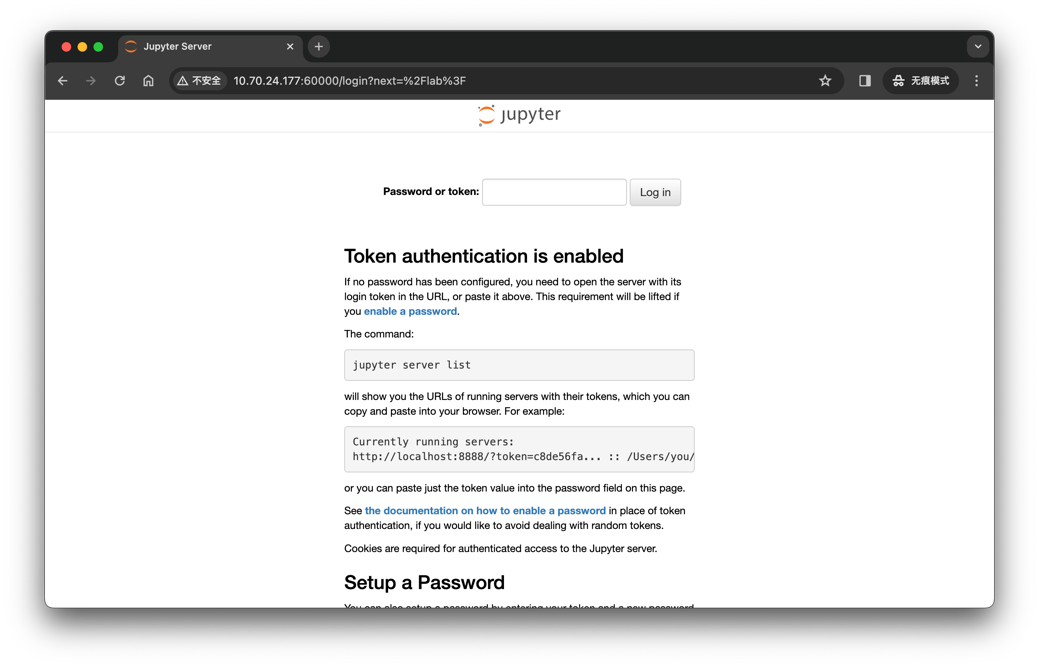
Transformer模型是一种通用的深度学习架构，已经在各种自然语言处理和其他领域的任务中取得了巨大成功。其中Decoder-only的Transformer模型（如GPT系列）在许多领域中发挥着重要作用。这种模型主要集中于生成任务。

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。

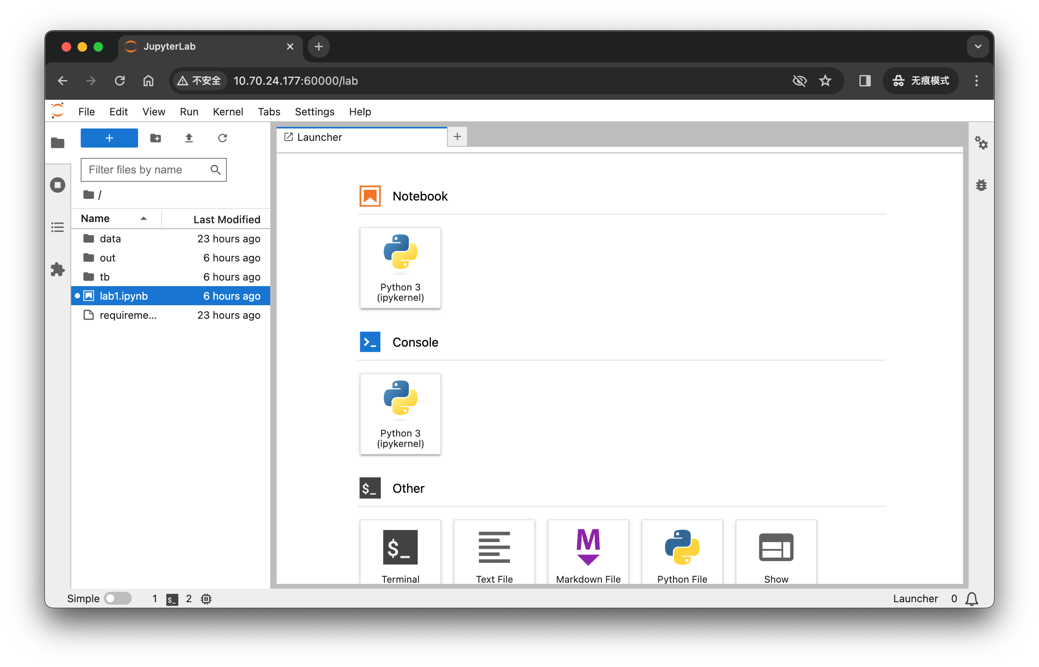
## 二、实验环境

本实验为每组同学准备了1张NVIDIA 2080TI，并提供Jupyter Lab服务。需要连接机房WiFi登录到Jupyter：

1. 连接WiFi：AI-Lab-704，密码：HITAILab
2. 浏览器登录<http://10.70.24.177:<PORT>>，其中每组同学有独立的端口号，具体分配见附件。连接到Jupyter服务的示例如下：



1. 连接Jupyter后，使用密码（初始密码见附件）进行登录，进入服务器。成功进入服务器的示例如下：



1. 在服务器中可以正常运行Python代码或Jupyter Notebook

上述为浏览器登录服务器的方法。使用VS Code或Pycharm连接远程Jupyter Kernel的方法见附录。

注意：由于安全原因，只能用机房WiFi连接NVIDIA 2080TI服务器。请同学们尽量在实验课内调试代码和训练。

## 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、设计逻辑、遇到的困难和解决方式等，提供程序运行效果的截图。（5分）

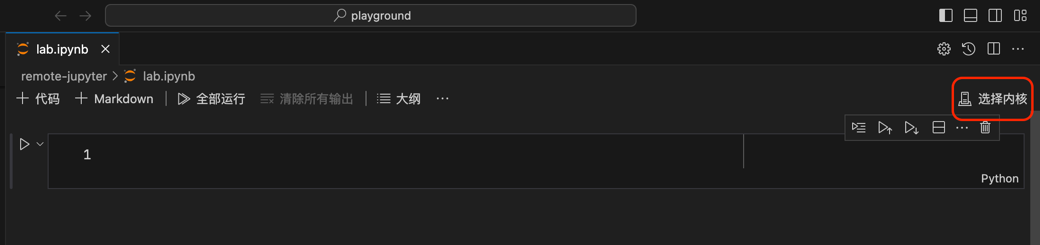
## 四、提交方式

请将代码（仅ipynb文件）、实验报告打包提交。

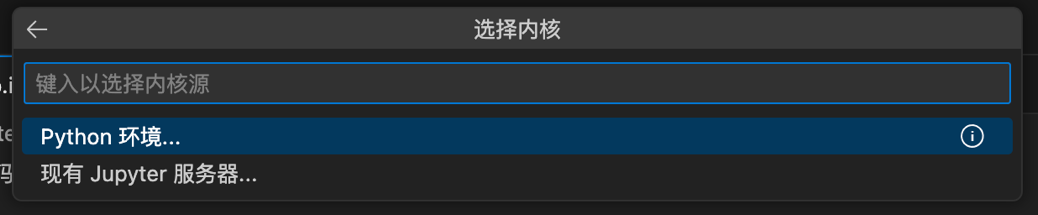
附录

## VS Code连接远程Jupyter Kernel

在本地新建一个Jupyter Notebook，右侧会提示“选择内核”：



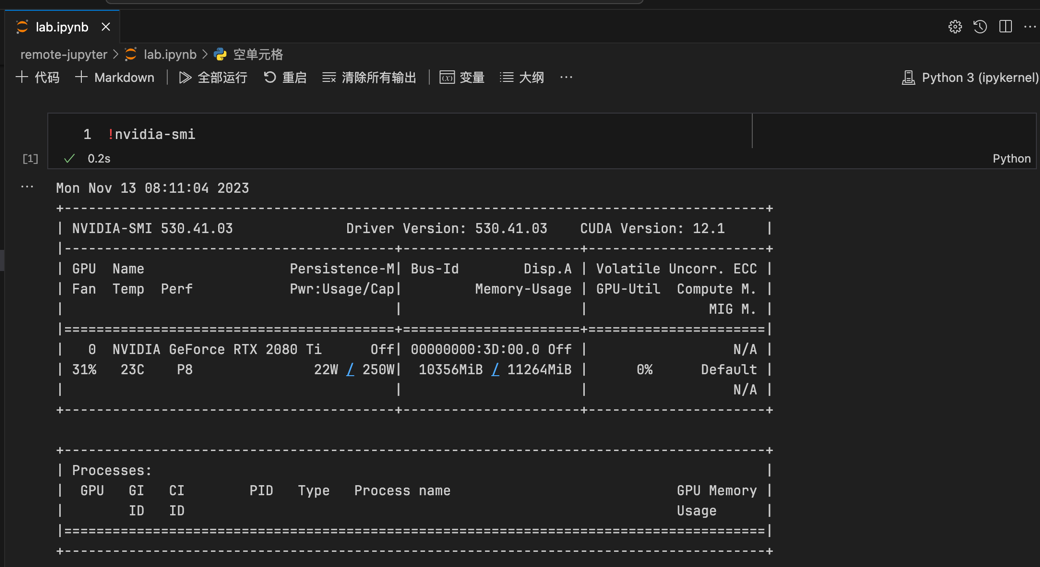
选择“现有Jupyter服务器”：



输入服务器ip和端口，同时加上token字段（密码），格式如下：

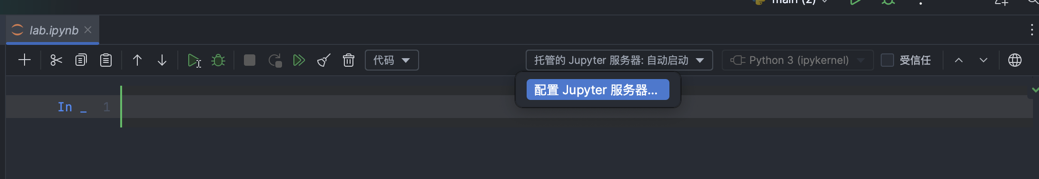


成功连接后，当前Notebook的Python解释器就是远程的解释器，同时目录和显卡等也是远程环境：

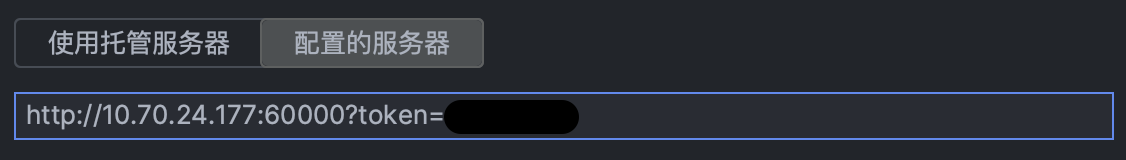


## Pycharm连接远程Jupyter Kernel

在本地新建一个Jupyter Notebook，上方会提示“配置Jupyter服务器”：



选择“配置的服务器”，输入服务器ip和端口，同时加上token字段（密码），格式如下：



成功连接后，当前Notebook的Python解释器就是远程的解释器，同时目录和显卡等也是远程环境：

