Jupyter 笔记本书面评估 2023-24

创建用于蛋白质二级结构预测的全卷积 PyTorch 模型

入门

课程作业设置为 Kaggle 竞赛。Kaggle 是领先的数据科学竞赛网站之一，值得熟悉一下。

我们的 Kaggle 竞赛将是非公开的，只有课程讲师和您的同学**可以**查看，**因此我们将使用格拉斯哥大学的电子邮件地址，以便我在给您打分时识别您的身份。**

因此，您首先需要使用您的 UofG 电子邮件地址，通过 Kaggle 网站上的 Kaggle 注册链接（https://www.kaggle.com/）注册参加本次比赛。

如果您对 Kaggle 不熟悉，那么您可能应该从《泰坦尼克号教程》开始学习：

https://www.kaggle.com/alexisbcook/titanic-tutorial

登录 Kaggle（使用您的 UofG 电子邮件地址）后，您可以使用以下链接参加 "2022-23 年理学硕士深度学习 "竞赛：
https://www.kaggle.com/t/59d729144530466b9513a7528ea8c462

(请勿与本班以外的其他人共享此链接）。

您需要选择一个队名来参加 Kaggle 竞赛。通常 Kaggle 竞赛都是以团队形式进行的，但这次是**个人课业，**所以你的团队中只有你自己！

总体目标

您的总体目标是编写一个**全卷积 PyTorch 模型**，该**模型**可以输入蛋白质序列数据（通常称为蛋白质一级结构，或使用 PSSM Profiles 预测蛋白质二级结构（H = 螺旋，E = 延展片，C = 螺旋符号）。
PDB 数据库包含 20 多万种蛋白质的结构。每个蛋白质都有一个独特的 PDB_ID 代码，如 1A0S（训练数据中的第一个），它就是上图所示的结构（沙门氏菌的蔗糖特异性孔蛋白），用于将蔗糖转移到引起食物中毒的沙门氏菌的细胞膜上。该蛋白质具有三维

结构显示，这种蛋白质的大部分是延伸的β片（平箭头）和线圈（随机线）。

Kaggle 上的 "数据 "选项卡允许您浏览用于训练的可用数据。您应该使用该数据选项卡浏览数据，以便了解数据的情况。您会发现一个 seqs_train.csv 文件，这是一个 CSV 文件，其中给出了每个蛋白质的 PDB_ID（唯一标识符）和序列。您还会发现一个 train.zip 文件，其中包含大量的

<PDB_ID>_train.csv 文件包含该特定蛋白质中每个残基的残基号、氨基酸和 PSSM 曲线。
labels_train.csv文件包含不同训练蛋白质的二级结构标签（以H = Helix、E = Extended Sheet、C = Coil符号表示）。
seqs_test.csv 和 test.zip 中包含您需要预测二级结构的测试序列的类似数据。

此外，您还需要通过 Moodle 网页提交生成这些输出结果的 Jupyter 笔记本。

我应该如何开发我的代码（我从哪里获得 GPU/TPU 能力？）

到目前为止，您主要使用 Google Colab 笔记本进行实验，但这将涉及到传输相当大的数据文件，而且您可能会发现 Google Colab 的 GPU 时间不够用（特别是如果您还在使用它进行其他课程作业时）。

对于本课程作业，我们将使用 Kaggle Notebooks！这不仅能让您熟悉另一个 Jupyter Notebook 系统，还意味着您可以直接访问本次比赛的数据文件，而无需转移文件。此外，由于 Kaggle Notebook 内置了版本系统，因此还能让您保持文件的有序性（重要的是，您提交的笔记本必须与您在 Kaggle 最佳尝试中生成预测时使用的笔记本相同，除非您记录不同提交时使用的笔记本版本）。

如果您进入 "代码标签"，然后选择 "新建笔记本"--这将创建一个竞赛笔记本，您可以直接访问竞赛数据。有关 Kaggle 笔记本系统的更多信息，请参阅 Kaggle 笔记本文档：
https://www.kaggle.com/docs/notebooks。

您使用 GPU 的方式与使用普通 PyTorch 代码的方式相同（您需要以类似于 Google Colab 的方式打开 GPU）。如果您想尝试使用 TPU，以下是一个很好的入门课程：
https://www.kaggle.com/competitions/tpu-getting-started

关于使用 TPU 和 PyTorch 的更具体教程请参见：
https://www.kaggle.com/code/tanlikesmath/the-ultimate-pytorch-tpu-tutorial-jigsaw-xlm-r/notebook

成功的步骤！

这在很大程度上是一个 "顶点 "项目，你将把前 5 周从不同的实验和讲座中理解的大量材料整合在一起。

熟悉 Kaggle 基础架构后，笔记本开发的第一阶段将是为 PDB ID csv 数据和 PSSM 数据编写一个自定义数据加载器，方法与实验室 5 类似。

首先要确保理解第 4 讲--机器学习工作流程中的关键概念。在将数据分成合适的训练数据集和

验证数据集方面，这些概念中的很多都是必不可少的（您应该使用验证数据集来评估您的性能，而不是依靠重新提交到 Kaggle 来评估您的性能--您每天只允许提交 5 次--提交更多将导致过度拟合测试集）。

然后，第一阶段将是为这种特殊数据编写一个类似于实验室 5 的自定义数据加载器。然后可以综合实验 3 中的 ConvNets 材料，但要对其进行修改，以便与新型数据一起使用，并将模型转化为完全卷积网络，以便同时预测蛋白质的多个残基标签。此时，您可能需要加入实验室 4 中用于 Ray Tune 或 Ax 超参数优化的代码。

你必须

1. **用 PyTorch 开发一个模型！(我本不必说这些.....但每年我们都会收到 Keras 和 TensorFlow 模型的提交.....通常都是从 GitHub 上截取的！）。**
2. **您需要为这项任务设计并实现一个全卷积模型，该模型将接收输入张量（作为张量的完整序列或作为张量的完整 PSSM 序列剖面），然后通过该模型生成一个输出二级结构标签的完整张量。请看全卷积模型是如何用于将图像分割成若干标记区域的。您要做的也是类似的事情，只不过是将序列 "分割 "成若干二级结构标签。**
3. **您需要演示使用 Ray Tune 或 Ax 进行一些适当的超参数优化，如实验 4。显然，鉴于 Kaggle GPU/TPU 资源有限，您需要选择合理的超参数优化方法。**

你应该

4. **让笔记本生成损耗和精确度曲线，以评估训练效果并诊断任何问题。**
5. **作为一项延伸挑战，请尝试使用 Captum 了解您的模型用于预测α螺旋、β薄片和线圈区域的特征！**

显然，你还应该向 Kaggle 提交每个模型的预测结果，并确定其中哪个模型可能做得最好（通常是通过创建并提交 submission.csv 文件来实现）。这可以直接从 Kaggle 笔记本的 "输出 "目录中完成。

提交

请将您的方法结果（submission.csv）提交到 Kaggle 网站。我们将对此进行测试，并在排行榜上公布**未见测试集的**准确率。您每天最多可以提交 5 次，以评估什么是最佳方法。最终的私人排行榜（仅在比赛结束后公布）将显示您提交的最佳作品的分数，**该分数将占总分的 50%**（该分数将基于优于特定阈值的分数，而不是准确率分数的直接转换）。
!)

重要提示--同时将您的最终结果 Jupyter Notebook 提交到 Moodle（从 Kaggle Notebooks 系统导出）。

您的 Jupyter 笔记本将在多个方面进行评分，例如显示上述关键部分（使用训练和验证数据、绘制和解释损失曲线、超参数调整、使用 Captum 解释拉伸挑战以及从这些方面讨论您的

两个模型）。您的 Jupyter

作为数据科学实验笔记本，笔记本文件应该有很好的注释--解释你在做什么以及为什么，解释你的结果以及它们意味着什么。**您提交的笔记本需要运行所有单元格，以便全部显示输出结果，从而获得分数！**提交的笔记本将占课程作业分数的另外 50%。

再次强调--您提交的 Jupyter 笔记本应具有可见的所有输出，这样就可以作为数据科学笔记本阅读，*而无需再次运行。*

请使用**团队 Jupyter 书面课件频道**来澄清有关课件的任何信息。