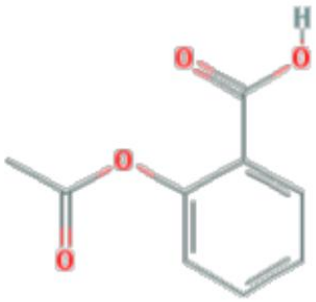
Project Specification

几乎所有人在其一生中都通过不同的来源（包括食物，家用清洁产品和药品）接触到不同的化学物质。但是，在某些情况下，这些化学物质可能有毒并影响人体健康。实际上，尽管有希望在动物模型中进行临床前研究，但仍有超过 30％的药物在人类临床试验中失败，因为它们被确定为有毒的。考虑到现实世界中用于评估药物的临床试验非常耗时，因此，如果可以开发出一种计算性药物分子毒性评估方法来快速测试某些化学物质是否有可能破坏人体高度关注的人体过程，则是理想的选择健康。

近年来，深度神经网络已成为机器学习中的热门研究课题。与其他方法相比，深度学习已显示出其在处理大量数据和实现更好性能方面的优势。在这个单独的项目中，将为您提供带有 SMILES 表达的药物分子数据集（将在后面进行解释）以及指示一个药物分子是否有毒的二进制标签。您将要开发一个深度神经网络，该网络可以从提供的数据中学习有用的模式，并根据使用 TensorFlow 或 PyTorch 软件包获得的知识预测新分子列表的毒性。

SMILES Expression

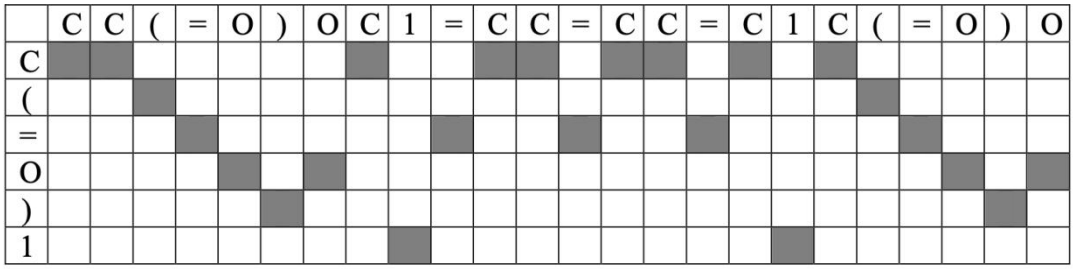
简化分子线性输入规范（SMILES）是使用一维 ASCII 字符串表示分子结构的线性表示。（https://www.jianshu.com/p/8c915de5ad4d）例如，阿司匹林是日常生活中常用的药物，其 2D 结构为



它的 SMILES 是：

CC(=O)OC1=CC=CC=C1C(=O)O

SMILES 的一种热门格式是 2D {0,1}矩阵，其中每一列代表当前分子的 SMILES 表示法中的一个符号，每一行是出现在数据集的 SMILES 词典中的一个 ASCII 字符。 2D 矩阵的大小是数据集的 SMILES 字典的大小*最长的分子 SMILES 的长度，这意味着在短分子 SMILES 之后填充零。对于 SMILES 表示法，第 i 行第 j 列表示该 SMILES 的第 j 个符号是词典中的第 i 个字符。阿司匹林的一个热门例子是：



Dataset

提供的数据集是关于一些小分子的毒性的。我们为您提供两个文件夹，一个是“ train”文件夹下的训练数据（8169 个样本），另一个是“ validation”文件夹下的验证数据（272 个样本）。每个文件夹中有三个文件：

文件	类型	描述
names_smiles.txt	String	一个 txt 文件，每行包含一个药物分子的名称及其 SMILES 表达，以逗号（，）分隔
names_labels.txt	String	一个 txt 文件，每行包含一个药物分子的名称和毒性标签，其中 0 表示无毒，1 表示有毒，以逗号（，）分隔
names_onehots.npy	Numeric	一个可以由 numpy 包加载的 npy 文件，存储两个 ndarray；一个是分子的名称，另一个是药物分子的 SMILES 表达的一种热门表示

源数据是 names_smiles.txt 和 names_labels.txt 文件。您的神经网络应该从 (SMILES, 标签）记录中学习，并最终能够从 SMILES 预测标签。 names_onehots.npy 文件是从 names_smiles.txt 派生的，该文件存储药物分子的 SMILES 表达的一键表示。提供 names_onehots.npy 可以减轻数据预处理的负担，并专注于神经网络的构建。只要这些数据文件是唯一的训练数据和验证数据，您就不必受限于它们的使用方式。您可以根据需要构建卷积神经网络或其他类型的神经网络。

“ ../train”和“ ../validation”中的分子彼此不重叠。我们用来标记模型的数据位于名为“ test”（610 个样本）的文件夹中，您无权访问其标签，但这些样本位于../test/names_smiles.txt 和../test/names_onehots 中。 npy 为您生成自己的概率预测，应将其存储在../output_student_id.txt 中。这两个文件的格式：names_smiles.txt 和 names_onehots.npy 与用于训练和验证的文件相同，但是分子是新的。

Assignment Requirement

Data

a）训练
b）验证
c）测试（只能访问样本）
您可以在“ train”文件夹下的数据上训练模型并验证其性能，在“validation”文件夹下的数据上，也可以同时使用两者训练模型。

Feature

您可以使用 SMILES 的热点作为分子的特征，也可以直接使用随心所欲的符号。

Model

您可以在此分配中使用任何深度神经网络架构来实现良好的性能，例如卷积神经网络（CNN），递归神经网络（RNN），长短期记忆（LSTM），图卷积网络（GCN）等。

Prediction Task

您将根据分子的结构预测分子的毒性。模型的输出为[0，1]，表示当前药物有毒的可能性

Output and Marking

您的模型将在数据集“ test”下的数据集上进行测试。您的输出文件应遵循../test/output_sample.txt 中提供的格式和顺序。第一列应该是已经给出的药物分子名称，第二列应该是当前药物有毒可能性的预测。

您提交的文件夹将被提取并放置在“ test”文件夹旁边，这意味着您必须使用提交的 test.py 文件中的相对路径“ ../test/”来访问标记数据。在您的 test.py 文件中，您需要首先还原模型参数，然后在标记数据（“ ../test/”）上测试模型，然后应将预测结果按照以下格式输出到名为 output_student_id.txt 的文件中的 output_sample.txt 文件, 应与 train.py 和 test.py 位于同一目录中。

您对这项作业的最终分数将取决于您的预测和真实标签中的 AUC（大区下限）。

Deep Learning Library

您可以选择以下三个版本之一

a）Numpy，Pandas 和 TensorFlow-gpu 1.15.0。

b）NumPy，Pandas 和 TensorFlow-cpu 1.15.0。

c）NumPy，Pandas 和 PyTorch 1.1.0。

请不要使用其他库。否则，您将获得此作业的零分。

Programming Language

此作业支持的唯一语言是 Python 3.6 和 Python 3.7。

Submission Requirement

您必须同时在 Canvas 和 Kaggle inClass 竞赛中提交。 加入 URL 分享 Kaggle 比赛。

1）提交清单（到 Canvas）

a）培训的源文件。 将其命名为 train.py

b）用于恢复网络模型的源文件。 将其命名为 test.py。

c）您对“ ../test”下的数据集的预测。 将其命名为 output_ student_id.txt，并严格遵循 output_sample.txt 中提供的格式。 您上载到 Canvas，上载到 Kaggle 和提交的模型的运行输出的 output_ student_id.txt 应该相同，否则您违反了荣誉码。

d）TensorFlow 和 PyTorch 在“ ../weights”中生成文件，这些文件存储您的训练模型。 使用经过训练的模型，可以执行 test.py 并输出结果。

e）其他有助于程序运行的文件，例如预处理文件，格式转换文件。

f）报告。

Important Points

为了使该项目公平，有意义，您还必须注意其他几点：

1）运行 test.py 的时间限制为 60 秒

2）整个 zip 文件的大小应小于 200 MB

3）抄袭行为将受到严厉处罚（零分加报告部门）

4）您上传到 Canvas 的 output_ student_id.txt，上传到 Kaggle 并运行提交的模型的输出应该相同，否则您违反了荣誉守则。