《自动化测试 2020》AI 测试大作业

一．项目背景

由于深度学习技术在开发范式上与软件程序的不同，深度学习技术无法由软件开发人员通过形式化业务逻辑进行系统功能的实现。结合深度学习等 AI 技术数据驱动的特性，数据及其质量将对 AI 模型的存在强关联。出了数据质量外，深度学习测试中还存在着数据获取成本高昂的问题。

针对以上提及的问题，本作业可分为两个阶段：一是测试数据生成，即通过应用扩增、变异等方法，生成更多维持语义的测试数据；二是数据质量评估，即对一个（组）测试数据集本身发现模型问题的能力以及提升模型健壮性的能力进行评估。

二. 项目要求

1. 整体要求

a) 本项目由两阶段组成，测试数据生成，及测试结果评估（其中一个阶段要包含自己的方法，另一个阶段可以使用已有的方法）

b) 为使得作业有统一的评判依据，项目代码需使用 Python 进行编写，且 Python 版本 >= 3.6。

c）本项目将提供 4 个数据集，并在作业方向确定后，学生需选取其中的两个（MNIST, Fashion MNIST，CIFAR-10 三选一，和 CIFAR-100）数据集。因为这四个数据集均是公开数据集，且集成在 keras dataset 中，推荐各位同学直接在项目中使用 keras.dataset 进行数据集的加载。

d）为了方便同学进行测试生成数据的效果评估，本项目将为每个数据集训练提供一个数据集对应的训练完的待测深度学习模型(hdf5 格式)。具体模型文件的下载链接会在课后给出。

2. 测试数据生成

本阶段的评分标准为：

i. 本阶段需要基于所选择的数据集，实施对应的数据生成方法

ii. 采用扰动方法的合理性说明

3. 测试数据质量评估

本阶段的评分标准为：

i. 生成的测试数据与原始数据的差异

ii. 测试数据集中使得待测模型出错的测试数据的数量和测试所发现的问题类型

iii. 测试数据对于模型鲁棒性的提升效果

三. 提交要求

1. 本项大作业分数分配：报告展示 20%，测试数据生成 40%，测试数据评估 40%。原则上，对于所选择的数据集数量不做限制，但选择的数据集的数量（需不少于 2 个）并不会对成绩产生影响。

2. 项目报告，包括以下内容：

a) 测试数据生成方法介绍：包括但不局限于方法实现原理及采用该方法的原因，相关的参数设置，参考文献等；

b) 测试数据质量度量方法介绍：包括但不局限于度量指标以及选用该指标的原因…

c) 项目代码实现的流程图介绍：用于辅助说明代码中的关键流程和步骤

3. 自己生成的测试数据集

4. 提交一份代码运行的 demo 视频，要求能够清晰展示没阶段任务的输入及输出。

5. 所有提交产物提交 Github。

四．未尽事宜发邮件咨询

通常助教会尽快（24 小时内）给出回复，为确保能够快速响应，请同时添加两位主角邮箱到收件人列表。

刘佳玮：eudemoniajw@gmail.com

章许帆：zhangxufan@smail.nju.edu.cn