大实验：Domain Generalization

实验目的

1. **了解域泛化相关的知识以及相关方法**
   领域泛化涉及到许多机器学习的关键概念，包括迁移学习、特征工程、模型泛化等。通过本次大作业，综合课程中学到的各种机器学习概念和技术，增强对于整个机器学习流程的理解，并尝试一些新的方法和策略以提高模型的泛化能力。
2. **构建分类器对 PACS 数据集进行分类，并在 kaggle 进行准确率排名**
   通过动手编程，加深对于在课程中学到的机器学习概念和技术的理解，包括模型训练、特征工程、评估等。并在代码实践中总结关于调整模型结构，损失，超参数的方法论，提升以后在科研上解决困难的能力

实验背景

域泛化 (Domain Generalization, DG)是近几年较为热门的一个研究方向，旨在让模型从具有不同数据分布(不同域)的数据上学习到一个泛化能力更强的模型，让模型能够提取出域不变的表征，使得模型能够在未知域的数据上依然具有较好的效果。
域泛化的方法和思路主要包含以下几种：

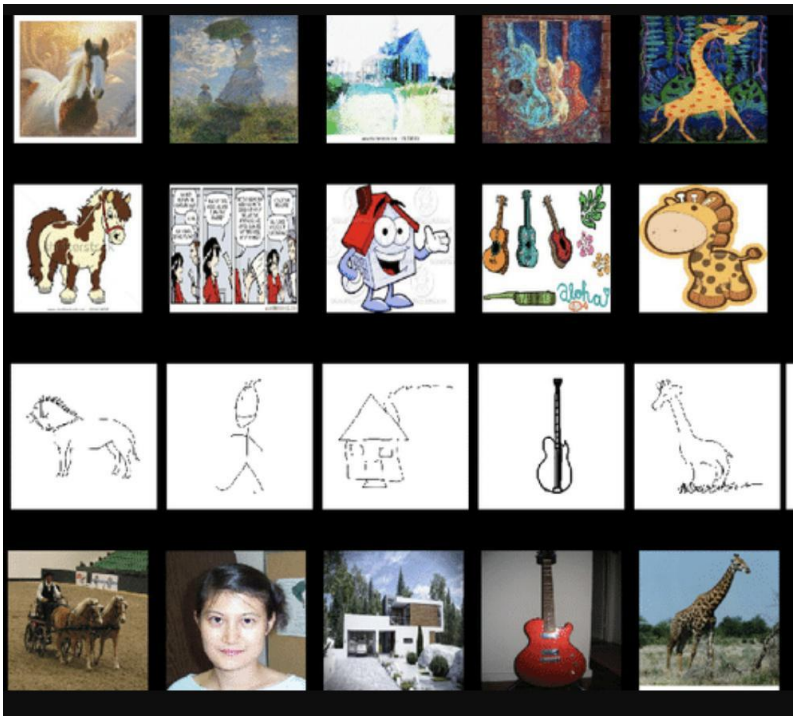
- 基于域对齐（Domain Alignment）的方法：这类方法试图通过显式或隐式地对齐源域数据的分布或特征，来减少不同域之间的差异，从而提高模型在目标域上的泛化能力。
- 基于元学习（Meta-Learning）的方法：这类方法利用元学习的框架，将源域数据划分为训练集和验证集，然后通过优化模型在验证集上的性能，来学习一个适应性强的初始化参数或更新策略，从而使模型能够快速适应新的目标域。
- 基于数据增强（Data Augmentation）的方法：这类方法通过对源域数据进行各种变换或合成，来扩充数据集或增加数据多样性，从而提高模型对目标域数据的鲁棒性。
- 基于集成学习（Ensemble Learning）的方法：这类方法通过结合多个子模型或子空间的输出，来提高模型在目标域上的性能，从而利用多样性或不确定性来增强模型的泛化能力。
- 其他类型的方法：除了上述四类主流的方法外，域泛化还包含一些其他类型的方法，例如基于因果推断（Causal Inference）、基于对抗生成网络（Generative Adversarial Network）、基于正则化（Regularization）等。

域泛化的问题具有重要的理论意义和实际应用价值，例如在计算机视觉、语音识别、自然语言处理、医学图像和强化学习等领域都有相关的研究。

关于该领域更深入得到的研究可以阅读：
[2103.02503] Domain Generalization: A Survey (arxiv.org)

实验数据

PACS is an image dataset for domain generalization. It consists of four domains, namely Photo (1,670 images), Art Painting (2,048 images), Cartoon (2,344 images) and Sketch (3,929 images). Each domain contains seven categories.



数据下载链接:

https://drive.google.com/file/d/13eZSqjGYR3cisCETFuhKmzHuBc_gtgsq/view?usp=sharing

实验内容

1. 熟悉领域泛化要解决的问题，并了解相关理论
复现在 github 上发布的 baseline，对 DG 问题要解决的问题和具体问题设定的数据划分有细致的了解
代码链接: nathanielyvo/Big-data-Machine-learning-Last-experiment (github.com)
2. 构建你自己的分类器用于对 PACS 数据集分类
针对 DG 问题构建你自己的模型，loss，训练框架等，利用 PACS 数据集对你的模型进行训练和验证，并保留你最好的模型用于后续测试，其中训练集的源域为 cartoon,art_paint,photo,用于测试的目标域为 sketch，具体数据的划分将会在后续发布
3. 在 kaggle 上提交你在 test 数据集上的分类结果
Kaggle 的比赛链接: Big data&machine learning DG Competition | Kaggle

作业提交时间

01/07 (周日) 之前
在 kaggle 上提交最终分类的 csv 文件，并在网络学堂提交源码以及实验报告
本次大实验对于使用的库没用限制
提交内容：报告（pdf）和代码（zip），并在 kaggle 上提交你在测试数据的分类结果，请确保你的代码清晰可读、可复现、无bug、无特殊环境依赖，无法复现的代码会极大影响你的得分。

评分依据

实验结果占大实验总分的 80%，实验结果分数评定将由在 kaggle 上的准确率排名进行打分
实验报告占大实验总分的 20%

本次实验是课程的附加分,总分为 10 分
最终课程得分计算方式为:

$$Min(\text{大实验分数} + \text{课程基本分数}，100)$$

代码语言：不限