**智能计算实验室初试题-项目题**

**请从以下四道题中任选一道作答。**

**题目一：工业缺陷分类**

缺陷检测是工业上非常重要的一个应用，然而，由于缺陷种类多样复杂，传统的算法很难做到对缺陷特征的准确建模，无法实现缺陷类别的精准识别。近年来，深度学习在特征提取和定位上取得了非常好的效果，越来越多的学者和工程人员开始将深度学习算法引入到缺陷检测领域中。

本实验的目标是 开发深度学习算法实现对缺陷图片的分类。提供的数据集中包含9种缺陷类型，每个子文件夹下的图片即为一种缺陷类别的图片，缺陷类型包括裂纹(cr)，划痕(sc)等，图片大小为64x64。

数据集包含有训练集和测试集，可通过以下两种方式下载获取

https://pan.baidu.com/s/1DBKlSWRFT-TMMoJgAYAZ3A

提取码：ehw1

或清华云盘

https://cloud.tsinghua.edu.cn/f/a6224bc7a3134eaba148/



部分训练数据展示

**实验要求**

1. 建立并训练神经网络，对缺陷进行分类。使用的深度学习框架不限，神经网络结构不限。
2. 使用以下两个准则对模型表现进行评估：

评价准则1： 准确率，即为分类正确的样本除以总样本数。请代码中实现该评价函数。

评价准则2： AUC值，请在代码中实现该评价准则，并绘制ROC曲线。

1. 对网络做一些修改并对结果进行分析
   1. 分析数据增强对网络性能的影响
   2. 分析Dropout机制对网络性能的影响

**提交材料：**除了面试规则中要求的提交材料，还需提交最终的模型文件(.pth)。

**题目二：模型迁移**

缺陷检测是工业上非常重要的一个应用，然而，由于缺陷种类多样复杂，传统的算法很难做到对缺陷特征的准确建模，无法实现缺陷类别的精准识别。近年来，深度学习在特征提取和定位上取得了非常好的效果，越来越多的学者和工程人员开始将深度学习算法引入到缺陷检测领域中。

本实验的目标是 开发深度学习算法实现对缺陷图片的分类，并实现不同应用场景之间的模型迁移。

本实验提供两个不同场景的数据集：

**数据集一：**包含9种缺陷类型，每个子文件夹下的图片即为一种缺陷类别的图片，缺陷类型包括裂纹(cr)，划痕(sc)等，图片大小为64x64。

数据集包含有训练集和测试集，可通过以下两种方式下载获取

https://pan.baidu.com/s/1DBKlSWRFT-TMMoJgAYAZ3A

提取码：ehw1

或清华云盘

https://cloud.tsinghua.edu.cn/f/a6224bc7a3134eaba148/



部分训练数据展示

**数据集二：**本部分提供另一个数据集NEU-CLS。该数据集收集了 热轧带钢表面的6类典型缺陷：氧化皮(rolled-in scale， RS)、斑块(patches, Pa)、裂纹(crazing, Cr)、点蚀(pitted surface, PS)、包含(inclusion, In)、划痕(scratches, Sc)。每种类型缺陷有 300 个图像样本，共 1800 张灰度图像，每张图像的 原始分辨率为 200×200 像素。该数据集根据图像名来确定图像类别，如图像名 “Cr\_1.bmp” 表明该图像的缺陷类别属于 裂纹(crazing, Cr)。数据集不进行训练集，测试集的划分，由同学自行划分。数据集可通过以下两种方式下载获取

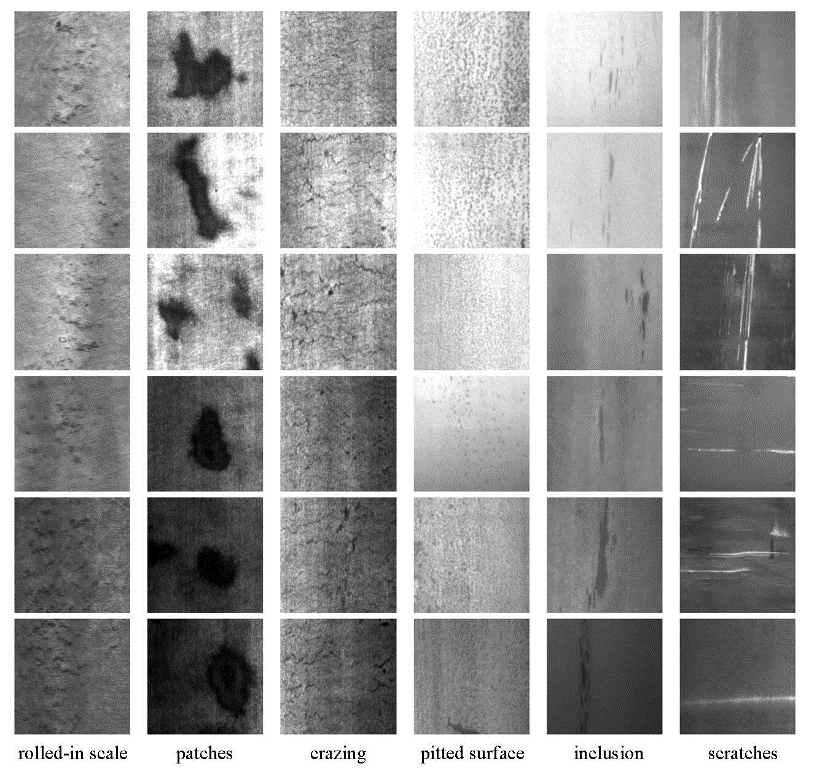
链接：https://pan.baidu.com/s/15gxgYGyM4mwLGbbcTGGeDw

提取码：6fav

或清华云盘

https://cloud.tsinghua.edu.cn/f/3416dac6e2c846178317/

下图展示了六种缺陷的样本图像。



**实验要求**

1. 针对数据集一，建立并训练神经网络，实现对该数据集的缺陷分类，并获得**模型1；**
2. 针对数据集二，建立并训练神经网络，实现对该数据集的缺陷分类，并获得**模型2；**
3. 实现迁移学习，将**模型1**的网络参数迁移到 **实验要求（2）**建立的神经网络中，并对该神经网络进行训练，获得**模型3**。
4. 分析和比较模型2和模型3的表现，探讨迁移学习的作用。

**提交材料：**除了面试规则中要求的提交材料，还需提交最终的模型文件(.pth)。

**题目三：Task on AutoIE2: Sub-Event Identification**

**实验背景：**

Sub-events identification is a very fundamental problem in the field of information extraction, especially in emergency situations (e.g., terrorist attacks). It is challenging for two reasons:

1. data confusing and imbalance. Events usually evolve rapidly and successive sub-events occur. Only a few target sub-events data need to be identifid from the large volume of events related data.
2. low resource. Usually only a limited amount of labelled seed data is given for learning and more annotating datasets are expensive and time consuming.

However, the existing works cannot fully meet the requirements, and thus better few shot learning and data selection models for sub-event identification are crucial.

**实验要求：**

The goal of this task is to build an IE system (Information Extraction system) that can quickly adapt to a new occurring sub-event. Specifically, there are two settings of this task:

1. Given a large number of event-related corpus and a few labelled seed data, the task aims to build an IE system which may identify the target sub-events.
2. Besides the machine learning model designing, annotating data selected from the unlabeled corpus is also allowed, but the size of the labelled data from the unlabeled corpus is fixed. How to select the best data to annotate and supply training dataset is also an important step in this task.

**实验数据：**

All corpus provided are obtained from comments (generally 8 to 120 characters long). The corpus are split into three parts, i.e., unlabeled dataset, seed dataset and testing dataset. The labelled seed dataset(100 samples per event) and unlabeled dataset(100K for 3 events) are released to participants to construct their own training set and developing set, and the testing dataset(around 2k per event) is used for final evaluation.

More details about these three datasets are as follows:

1. Unlabelled dateset: totally 100,000 samples related to the three sub-events.
2. Seed dataset: 100 labeled samples per sub-event.
3. Test dataset : 2000 labeled samples per sub-event.

实验数据可通过以下方式获取

链接：https://pan.baidu.com/s/1a3jdnQ3FMhiCkMG2IpyqUw

提取码：5pnb

或

https://cloud.tsinghua.edu.cn/f/6f6e06480375403aaea8/

**题目四：论文精读**

从以下四篇论文中任选一篇精读

**论文一：**《Global Filter Networks for Image Classification》

论文获取：<https://arxiv.org/pdf/2107.00645.pdf>

**论文二：**《Multimodal Graph-based Transformer Framework for Biomedical

Relation Extraction》

论文获取：https://arxiv.org/pdf/2107.00596.pdf

**论文三：**《Knowledge Distillation for Quality Estimation》

论文获取：https://arxiv.org/pdf/2107.00411.pdf

**论文四**：《Plug-and-Play Image Restoration with Deep Denoiser Prior》

论文获取：https://arxiv.org/pdf/2008.13751.pdf

**要求**：**精读并复现**该论文。在工作汇报中需要对该论文进行详细讲解，包括但不限于论列举该论文想要解决的问题，并总结论文的解决方案；对解决方案进行详细阐述；对论文的工作进行评价并给予改进建议或改进方案。复现的数据集不限。**复现可能比较难，可以把复现的思路写在工作汇报的PPT里，**比如如何搭建网络结构，复现该篇论文主要需要实现哪些功能等。