# 第三章 数据增广与增强

## 3.1 总论

垃圾分类作为典型的大规模多尺度图像分类问题，其数据有较高的复杂性。

首先，少数种类的垃圾数据量较小。数据量小就要求我们对少量已有的数据进行精细地调整，使图片的信噪比达到最优，另外我们可以使用不同的图像增广方法，获得不同特征的图像以扩充数据集。

其次，对于有效信息离散分布的图像，我们使用可以使用弱监督模型对垃圾区域进行目标检测，使用目标检测的结果生成新图像，以达到扩充数据集的目的。

最后，对于大部分数据，存在明显的多分辨率、多尺度、多粒度、特征复杂的特点。

于是，针对部分分类数据量小，大部分分类数据特征复杂的特点，兹设计如下的数据增广与增强解决方案。

## 3.2解决方案概述

解决方案分为两个子方案：数据增广框架、数据增强框架。数据增广用于扩充数据集、数据增强用于加强有效特征。并且，该解决方案有增强分类特征、优化不平衡数据集、推理时间短的特点。

需要指明的是，在传统上Data Augment可以被译作数据增广或数据增强，这是一个翻译问题。在本文中为了方便表达，使用数据增广一词描述时间成本较高的数据预处理操作，反之使用数据增强一词描述时间成本低的操作。

其中，数据增广的主要目的是扩充数据集的数据量，解决分类不平衡问题；数据增强的目的是增强数据特征，解决图像的光照差异大等问题，使用最简单、最根本的方法有效增强模型的鲁棒性。

### 3.2.1数据增广框架

扩充数据集的经典方法是对图像进行随机裁剪、翻转、色彩变换等操作，将生成的新图片作为新样本，以达到扩充数据集的目的。

扩充数据集的挑战性方案是基于对抗神经网络的样本生成，通过对抗神经网络训练的方式得到优秀的样本解码器，在解码器中输入与垃圾分类图像相似性高的特征，完成样本生成任务。

经典方法的特点是高效、易实现，GAN方案的特点使训练难度大，成本较高，但样本生成的效果较好。

在本方案中，我们经过综合考量，提出新的样本生成方法。

这种方法基于发展日渐成熟的目标检测模型和传统的图像处理技术，具体流程如下：

* 使用少量bounding box数据与mask数据训练目标检测模型，获得弱检测模型1
* 使用弱检测模型1对新获得大量图像进行目标检测，获得推理结果
* 使用推理得到的bounding box及mask目标结果，在目标结果区域进行适应性色彩修正、适应性噪声补充等操作，并添加到其它背景图片中

### 3.2.2数据增强框架

### 3.2.3解决方案特点

#### 特点一：增强分类特征

使用AutoML算法，对限定条件下的数据增强或增广方法进行参数搜索，获得有效的数据增强手段，使得样本特征明显、鲁棒性高。

#### 特点二：优化不平衡数据集

通过样本生成方法优化不平衡数据集。传统样本生成基于GAN的方法，生成器训练难度大、训练成本高。本数据增强解决方案根据垃圾分类问题特点，设计有针对性的解决方案。方案概述如下：

通过弱目标检测模型提取不平衡数据分类的mask或bounding box，再将提取结果与其它图像做拼接，拼接过程中对mask或bounding box做适应性色彩修正、适应性噪声补充、翻转、旋转、移位、缩放、裁剪等操作。通过这种方式，对不平衡分类数据进行扩充，扩充的数据特征自然，适用于模型训练。

#### 特点三：推理时间短

本解决方案基于高效的数据增强第三方库Imgaug，推理时间短，落地方法多。使用于有着海量数据的垃圾分类问题。

## 3.3相关工作

### 谷歌ImageNet搜索得到的Auto-augment策略

### 第三方库Imgaug

参考文献：

Shorten, C., Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. *J Big Data* **6,**60 (2019). <https://doi.org/10.1186/s40537-019-0197-0>

Cubuk E D, Zoph B, Mane D, et al. Autoaugment: Learning augmentation strategies from data[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 113-123.