4.2.3 其他分类

除了上文所述的数据分类，根据应用场景的不同，还衍生出了其他分类，具体如表 4-2 所示。

表4-2 数据的其他分类

应用场景	分类
描述事物的角度	可分为状态类数据、事件类数据和混合类数据
数据处理的角度	可分为原始数据和衍生数据
数据粒度	可分为明细数据和汇总数据
更新方式	可分为批量数据和实时数据

4.3 构建高质量数据集

下面以图像类任务为例，解释如何构建高质量的数据集，以获取更好的训练结果。

4.3.1 获取足够的数据量

如果数据集过小，训练模型将因为没有足够多的数据，而难以找到其中的特征，若在此基础上训练数据，发生过拟合的概率较大。因此，足够的数据对于模型训练来说是必需的。然而，目前还无法做到明确某一确定算法所需的数据量，因为每个训练项目所涉及的元素是独特且多样的，涉及模型的复杂度、训练方法、标签需求、误差容忍度等，这些复杂多变的因素会导致开发者无法准确预估所需数据集的大小。而使用 10 倍法则能够帮助开发者快速估算、分析项目的训练数据需求量。

10 倍法则是常见的经验法则，指的是模型通常需要超出其自由度 10 倍的数据量。自由度可以简单地理解为数据集中的列，是影响模型输出的参数，也是数据点的属性。10 倍法则的目标是抵消这些组合参数给模型输入带来的变化。该法则对于复杂模型的作用微乎其微，但使用它能够迅速估算数据集的容量，以保证项目持续运行。

获取更多数据的方法有两个，即收集更多的数据和增强数据。

1. 收集更多的数据

一般来讲，数据来源主要分为两大类：内部来源和外部来源。其中，内部来源的数据指的是企业内部真实场景下的数据，如工业设计图纸、年度销售数据等。内部数据的收集往往有赖于人工，其人力、时间成本较大。在数据相似度允许的情况下，可通过外部来源获取更多数据。外部来源的数据包括外部购买的数据集、网络爬取的数据、免费获取的开源数据等。

可以利用网络爬虫爬取一些需要的数据，再将数据存储为表格的形式。网络爬虫又称网络蜘蛛、网络机器人等，网络爬虫是用计算机语言编写的程序或脚本，可爬取互联网上网站服务器的内容，用于自动从网络上获取信息或数据。

另外，也可以使用一些开源的数据集。以下 3 个开源数据集很常用，这 3 个数据集对于深度学习网络的发展，通用的分类、分割和检测任务的评测有其他数据集所不可比拟的作用。

（1）MNIST

MNIST 是一个用于分类任务的手写数字的开源数据集，数据集来自 250 个不同人员的手写

数字图像，这些人员中，50% 是高中学生，50% 是人口普查相关部门的工作人员。数据为 0 ～ 9 的手写数字图像，图像大小是 28px × 28px，训练集包含 60000 个图像，测试集包含 10000 个图像。图 4-4 所示为 MNIST 数据集示例。

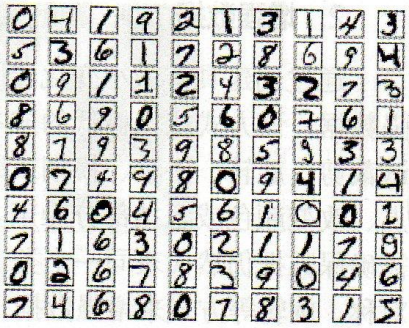


图 4-4　MNIST 数据集示例

（2）CIFAR-10 和 CIFAR-100

MNIST 数据集是入门级的数据集，其缺陷在应用场景下显而易见，所以，将 MNIST 数据集用于评估越来越深的神经网络不太合适，此时需要 CIFAR-10 数据集，该数据集是更加多样的真实的彩色数据集，MNIST 与 CIFAR-10 数据集各自的特征如表 4-3 所示。

表4-3　MNIST与CIFAR-10数据集各自的特征

数据集	图像特征	数量特征	统计特征
MNIST 数据集的特征	灰度图像	类别少	非真实数据，无局部统计特性
CIFAR-10 数据集的特征	真实图像	单个主体目标	可辨识、可整理

CIFAR-10 是一个只用于分类的数据集，该数据集共有 60000 个彩色图像，图像大小是 32px × 32px，共 10 个类，每类 6000 个图像。其中 50000 个图像组成训练集，每一类包括 5000 个图像。另外 10000 个图像组成测试集，每一类包括 1000 个图像。图 4-5 所示为 CIFAR-10 数据集示例。

可以看出，CIFAR-10 类似于扩充的 MNIST 的彩色增强版，在数据集大小方面，CIFAR-10 略小于 MNIST。

CIFAR-100 数据集则包含 100 个小类，每个小类包含 600 个图像，其中有 500 个训练图像和 100 个测试图像。与 CIFAR-10 不同的是，100 个小类被分为 20 个大类，而每一个大类，又可以细分为子类，所以每个图像带有 1 个小类的“fine”标签和 1 个大类的“coarse”标签。

（3）PASCAL

PASCAL 是一个用于模式分析和统计建模的数据集，包括图像分类、物体检测、图像分割等任务，它是由 PASCAL VOC 挑战赛衍生出的数据集。PASCAL VOC 挑战赛是世界级的计算机视觉挑战赛，很多优秀的计算机视觉模型，如分类、定位、检测、分割、动作识别等，都是基于 PASCAL VOC 挑战赛及其数据集推出的。该数据集中用于分类和检测的训练数据为 11540 个图像，包含 27450 个已被标注的感兴趣的标注对象；用于分割的训练数据为 2913 个图像，包含 6929 个分割对象。图 4-6 所示为 PASCAL 数据集示例。



图 4-5　CIFAR-10 数据集示例



图 4-6　PASCAL 数据集示例

2. 增强数据

通过创建具有微小变化的同一图像的多个副本可以实现增强数据，该方式能以极低的成本生成大量额外的图像。可以尝试裁剪、旋转、平移、模糊或缩放图像，也可以为图像添加噪点、改变颜色或阻挡部分噪声。但是，在上述所有情况下，均需确保数据仍然代表同一个类。如图 4-7 所示，可以通过裁剪、对称、旋转等方式来增强数据，同时，这些图像仍然代表“狗”这个类别。这个方式相对强大，因为堆叠这些效果会使得数据集的数据量呈指数级增长。



图 4-7　增强数据的方式

4.3.2　采集高质量的数据

数据量十分重要，但是若不能保证数据的质量，数据量再多对模型训练也无益。因此，在数据准备过程中，需要确保数据的质量，以下从 3 个方面介绍采集高质量数据的注意事项。

1. 采集与真实世界匹配的数据

当数据集是由与真实世界不匹配的数据组成时，如图像来自完全不同渠道，可能会使模型提取出的数据存在于现实世界中无法使用的特征。因此，需要考虑技术应用的长期性，尝试使用相同的技术查找或构建数据集。图 4-8 所示为公仔玩具等的图像，与现实世界中的“狗”并不匹配，因此需要去除。

2. 规范数据

如果数据没有特定格式，或者其尺寸或者相应数值不在特定范围内，则模型可能无法处理这些数据。为此，可以规范数据，使每个数据都在相同的取值范围内。如图 4-9 所示，可以通过裁剪或拉伸操作，规范图像的长横宽比和像素值，使规范化后的数据具有与其他数据相同的格式。

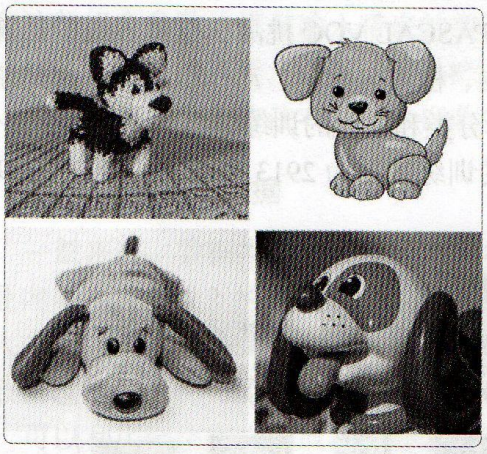


图 4-8　需去除的与真实世界不匹配的图像数据

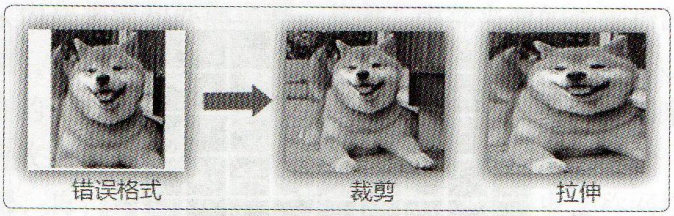


图 4-9　规范数据示例

3. 去除糟糕的数据

低质量的数据会导致低质量的模型训练效果。如果数据集中的数据与所需要的数据相差太大，则这些数据可能会引起混乱，会对模型训练效果产生干扰。因此，需要去除糟糕的数据，如图 4-10 所示，图中的 3 个图像均代表“狗”，但模型可能无法使用这些图像数据。

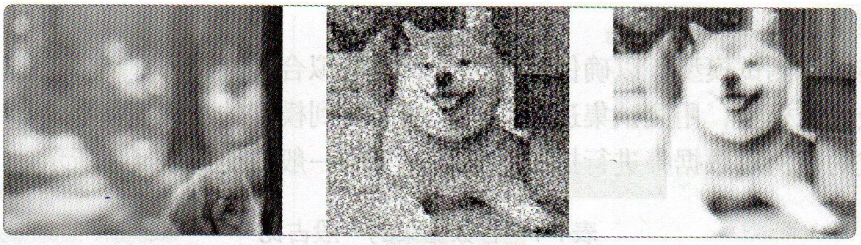


图 4-10　去除糟糕的数据示例

4.3.3　创建高质量的分类

数据采集完成后，需要对数据进行分类，以方便后续进行数据标注，创建高质量的分类的具体方法如下。

1. 选择正确的粒度级别

首先要想好应如何设计分类，每个类为希望模型识别出的一种结果。例如，要识别水果，则可以将“apple”“pear”等设计为一类；如果是判断合规性，则可以将“qualified”“unqualified”设计为两类，或者将“qualified”“unqualified1”“unqualified2”等设计为多类。

2. 创建平衡的分类

如果每类数据的数量不是大致相同的，则该模型可能会出现较低级的错误，即该模型存在偏差，因为类分布是偏态的。因此，需要创建平衡的分类，可以从过度表示的类中删除一些数据，或向表示不足的类中增加数据。如图 4-11 所示，可以补充狗的图像，减少兔子的图像，使得数据集不同的分类更平衡。

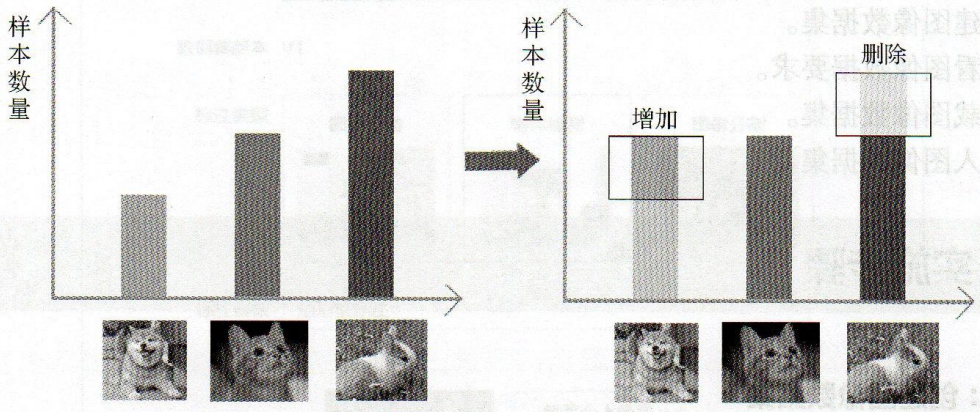


图 4-11　平衡分类的方式

3. 验证数据标签

此外，需要浏览所创建的数据集，并验证每个数据的标签，因为在数据集中使用反例会对学习过程产生不利影响，所以需要花一段时间对每个数据的标签进行验证，以确保标签正确。

4.3.4　拆分数据集

清洗、扩充和正确标记数据集后，需要将数据集拆分为 3 个，包含训练集、验证集、测试集，分别用于模型训练的不同阶段。

（1）训练集用于训练模型。

（2）验证集用于测试模型，以确保模型没有发生过拟合。

（3）选择最优的模型，用测试集进行测试，从而得到模型的准确度。

可按照相应的比例对数据集进行拆分，各数据集的一般占比如表 4-4 所示。

表4-4　各数据集的一般占比

拆分类型	训练集	验证集	测试集
拆分比例	60%	20%	20%

项目实施

4.4　实施思路

基于项目描述与知识准备内容的学习，读者应该已经了解了数据的定义和分类，以及如何构建高质量的数据集。现在回归 EasyData 平台，介绍在该平台上导入安全帽佩戴检测项目所需的数据集。

安全帽是各行各业安全生产工作者必不可少的安全用具，正确佩戴安全帽不仅可以有效防止安全事故的发生和减轻各种事故的伤害，而且可以保障工作者的生命安全。为了防止出现未戴安全帽导致的安全事故，安全帽佩戴检测成了监督工作者佩戴安全帽的利器。在本教材项目 8 中，将开展识别工作者是否佩戴安全帽的项目实施。此处介绍通过以下步骤导入安全帽佩戴检测项目所需的数据集。

（1）创建图像数据集。

（2）查看图像数据要求。

（3）下载图像数据集。

（4）导入图像数据集。

4.5　实施步骤

步骤 1：创建图像数据集

首先，可以通过以下步骤来创建图像数据集，用于存放安全帽佩戴检测项目所需的数据。

（1）登录人工智能交互式在线实训及算法校验系统，进入本项目的实验环境。单击“控制台”中“AI 平台实验”的百度 EasyData 的“启动”按钮，进入 EasyData 平台。单击“立即使用”按钮进入登录界面，输入账号和密码。

（2）进入 EasyData 数据服务控制台后，单击“创建数据集”按钮，如图 4-12 所示，进入信息填写界面。

（3）按照提示输入个人信息，如图 4-13 所示。在“项目归属”一栏选择“个人”选项，“所属行业”可选择“教育培训”选项，在“联系方式”一栏输入自己的手机号码，以方便了解数据处理的进度。信息填写完后，单击“下一步”按钮，进入数据集信息填写界面。