针对长尾数据分布的深度视觉识别

文 / 魏秀参

摘　要　本文介绍了目前国内外关于长尾数据分布下深度视觉识别的研究进展，主要从常用数据集及应用、经典机器学习解决方案和深度学习解决方案三个维度进行梳理和分析，并针对长尾数据分布的深度视觉识别的未来方向进行了探讨。

关键词　长尾数据分布；深度学习；机器学习；视觉识别；计算机视觉

0 前言

在机器学习及其在视觉识别的应用中，我们处理的标准数据通常都有一个基本假设，即该数据集各类别对应的样本数量是近似服从均匀分布的，即类别平衡。但现实生活中的数据往往呈现较极端的不平衡现象，如日常生活经常看到云朵和狗等物体，却鲜见概念车甚至传说中的“外星生物”，这样的自然规律使得真实数据的分布通常呈现出“长尾”分布的形态，如图 1 所示。可以看到常见（但少量）的物体类别在视觉识别的图像中出现的频次占主导地位，而罕见（却大量）的物体类别出现的频次占比微乎其微。在机器学习和视觉识别的实际应用过程中，长尾分布在某种程度上可以说是比正态分布更加广泛

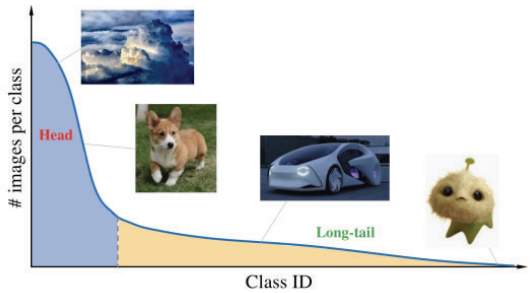

图 1　长尾数据分布示意图

存在的一种自然分布，现实中主要表现在少量个体做出大量贡献（少量类别的样本数占据大量样本比例），人们经常提到的“二八定律”（Pareto 法则）就是长尾分布的形象概括。

长尾分布数据的极度不平衡，给机器学习和视觉识别带来了巨大挑战。类别的极度不平衡导致模型学习非常容易被“头部”类别主导而产生过拟合；同时模型对于“尾部”数据的建模能力极其有限，从而在模型测试阶段表现出对长尾数据（尤其“尾部”数据）预测精度不理想的缺陷。特别是在借助深度学习模型进行的视觉识别应用中，尾部数据的数量缺失还使得深度模型的训练难以充分进行，导致特征学习很难达到理想程度，进而影响整个深度模型的泛化表现。此外，深度模型基于 batch 的训练特性带来的模型“遗忘”问题，在长尾数据分布情况下尤为突出，愈加影响了特征学习的整体质量。

近年来，针对长尾数据分布的深度视觉识别逐渐成为机器学习、计算机视觉和模式识别领域的热门研究课题，在诸多视觉感知任务，如细粒度图像识别、人脸识别、安防监控、车辆识别、商品识别等均有广泛应用。本文主要以长尾数据

分布下的深度视觉识别为主题，重点探讨其常用数据集及应用、经典机器学习解决方案和深度学习解决方案，下面分别从这三方面介绍长尾数据分布下深度视觉识别的研究进展。

1 常用数据集及应用

长尾分布下的视觉识别领域最为著名和常用的数据集为 iNaturalist 系列，其中 iNaturalist 2017 和 iNaturalist 2018 最为令人熟知。iNaturalist 系列数据集是美国加州理工、康奈尔大学和 Google 等机构联合构建的，以植物、鸟类、昆虫和菌类等 13 个自然生物大类下属的上千种物种细分类类别组成的细粒度级别图像数据集（fine-grained dataset），图像量多达近百万张。以 iNaturalist 2017 为例，该数据集共计 5 089 类细粒度物体，其中样本数最多的头部类别含 2 101 张样例图像，样本数最少的尾部类别仅有 4 张样本（见图 2），其数据分布呈现显著的长尾分布状态。而 iNaturalist 2018 则多达 8 142 类细粒度类别，样本最多的头部类别样本数多达 2 917 张，最少者仅有一张图像，呈现出更为极端的长尾现象。这两个著名的标准数据集，一方面验证了长尾分布的现实意义；另一方面其数据复杂性和显著的长尾分布特性，使得它成为长尾分布视觉识别研究中的标准测试“演武场”。此外，围绕 iNaturalist，相关组织者基本每年都在 CVPR 上组织全球视觉识别挑战赛，值得一提的是，我们的团队获得了 2019 届 iNaturalist 旗舰赛事的世界冠军。

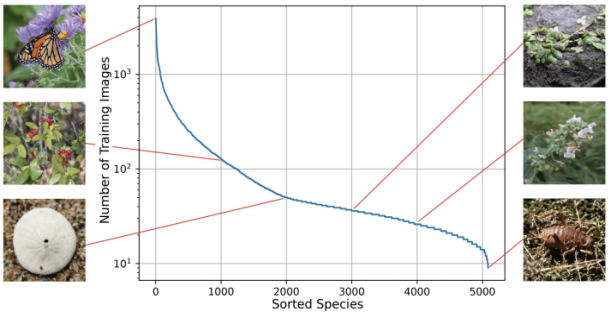

图 2 iNaturalist 2017 数据集示例

除天然的 iNaturalist 外，在人脸识别、通用物体识别和场景分类等应用中均有对应的长尾分布形态数据集。例如，针对人脸识别长尾分布问题构造的 MS1M-LT（2 万余类别），针对通用物体识别长尾分布问题构造的 ImageNet-LT（1 000 类）、CIFAR-10-LT（10 类）、CIFAR-100-LT（100 类），以及针对场景分类长尾分布问题构造的 Places-LT（365 类）等。各数据集的详细对照信息如表 1 所示。

表 1 数据集的详细对照信息

数据集	类别数 / 类	最大“头部”类别图像数 / 张	最小“尾部”类别图像数 / 张	物体类别属性
iNaturalist 2017	5 089	2 101	4	细粒度物体
iNaturalist 2018	8 142	2 917	1	细粒度物体
MS1M–LT	21 000	取决于构造方式	取决于构造方式	人脸
ImageNet–LT	1 000	1 280	5	通用物体
CIFAR–10–LT	10	取决于构造方式	取决于构造方式	通用物体
CIFAR–100–LT	100	取决于构造方式	取决于构造方式	通用物体
Places–LT	365	4 980	5	场景类别

2 经典机器学习解决方案

经典统计机器学习在处理长尾分布带来的挑战时，往往借助一些处理传统类别不平衡问题，以及处理代价敏感学习问题的技术手段和解决方

案。现有技术大体上有三类做法，第一类重采样法，即通过采样方式缓解长尾分布带来的样本极度不平衡；第二类重权重法，即通过改变学习权重来调整不同样本数类别的学习比重；第三类后处理法，即在模型学习后调整分类器参数的做法。

2.1 重采样法

重采样法是对训练集中不同类别训练样本数目直接进行调整，进而保证各类别样本数目平衡的一类方法，主要有“欠采样”和“过采样”两种。“欠采样”法，顾名思义，即去除一些样本较多的头部类别的样例，使得所有类别样本数目基本一致，然后在平衡后的数据上再进行学习；而“过采样”则会复制一些样本较少的尾部类别的样例，从而达到各类别样本数目一致的状态，之后进行学习。

2.2 重权重法

重权重法除应用在长尾数据分布学习任务外，还常应用于代价敏感学习，实际操作时通常在目标函数（或损失函数）上针对尾部类别的训练数据施加较大惩罚，借此克服类别不平衡带来的问题。一般而言，损失函数中的惩罚因子大小与类别对应样本数成反比，即样本数越多的类，其惩罚因子越小；样本数越少的类，其惩罚因子越大。

近期，Cui 等在传统重权重法基础上提出了一种基于“有效样本数”的重权重方法，替代了之前根据样本数目比例确定惩罚权重的做法，在诸多长尾分布数据集上取得了较好的精度。接着，Cao 等也提出了一种基于 margin 的重权重法，一方面表明不同样本数的类别应对应不同 margin；同时提出对于尾部类别须引导学习器得到较大 margin，方能在长尾分布数据上取得满意性能。

2.3 后处理法

后处理法一般用于基于深度神经网络模型训练的分类器，通过调整分类器向量的模长，达到缓解长尾分布数据对网络训练的影响。Kang 等在研究中指出，对于经过反向传播算法训练收敛后得到的深度模型而言，对于第 i 个类别的分类器 ω_i，其模长 $||\omega_i||$ 与该类别样本数基本呈现正相关，即样本多的类别其分类器模长也相对较大，从而长尾数据训练得到的分类器模长的分布也同样呈现接近长尾分布的形态。具体进行分类器模长正则化操作时，可对每类分类器作模长约束，即

$$\omega_i' = \frac{\omega_i}{||\omega_i||^{\tau}}$$

其中 τ 为 (0,1) 区间的超参数，视不同数据集而定。

3 深度学习解决方案

众所周知，深度学习是处理视觉识别应用的利器。长尾数据分布除影响深度模型分类器学习的同时，其极端的不平衡特性还给特征表示学习带来了巨大负面影响。针对长尾数据分布的深度学习解决方案主要分为三类，第一类是二阶段训练法，即通过两个阶段的训练，先后兼顾特征学习和分类器学习，从而克服长尾分布带来的类别极度不平衡问题；第二类是新型损失函数，即构造新式损失函数缓解类别不平衡；第三类是特征学习和分类器学习解耦，即将学习目标不同的二者解耦，各司其职，互无影响，进而协同起来提升模型预测精度。

3.1 二阶段训练法

深度学习应用中较常用的一种技巧是 fine-tuning，针对长尾数据分布的二阶段训练法便源于此。具体而言，二阶段训练法将基于长尾数据分布的模型训练过程分为两个阶段：第一个阶段供给深度神经网络的训练数据仍服从原始长尾分布，从而确保特征表示学习的效果；而第二阶段为缓解长尾分布带来的极度不平衡，此时会使用重采样或重权重法构造类别平衡的训练数据，同时配合较小的学习率进行二阶段 fine-tuning。

3.2 新型损失函数

该类方法主要聚焦在如何设计新型损失函数来指导深度网络学习，比较经典的代表性算法为 Range loss 和 Focal loss。2017 年，Zhang 等首先用切分实验的结果解释了长尾分布带来的性能损失，并受此启发提出 Range loss 来增加类间距离同时减小类内距离，在此基础上该损失函数还可避免模型训练被头部数据主导，且会惩罚由尾部数据（因样本不足）带来的类内松散问题。

另一代表性方法 Focal loss 提出之初是为了解决一阶段的通用物体检测模型，在物体检测任务中带来的类别不平衡问题；随后研究者发现，Focal loss 在处理长尾分布数据时也有较好表现。Focal loss 的设计思想与重权重法一致，本着尽量减小头部数据主导作用的想法，该损失函数在传统的深度学习交叉熵损失函数前添加一个权重项，进而调节不同样本数目类别的学习权重，即

$$\mathrm{FL}(p_t)=-(1-p_t)^{\gamma}\log(p_t)$$

其中，p_t 为类别预测置信度；γ 为大于或等于 0 的超参数，用以控制网络对该预测的惩罚程度。

3.3 特征学习和分类器学习解耦

在深度学习中，特征学习和分类器学习通常被耦合在一起进行端到端的模型训练。但在长尾分布数据的极度不平衡因素影响下，特征学习和分类器学习的效果均会受到不同程度干扰。在我们发表于 CVPR 2020 的工作中，首次揭示了重采样和重权重法这类类别重平衡的方法，其奏效之原因实际在于显著提升了深度网络的分类器学习模块的性能；于此同时还出于意料的发现，这类重平衡方法由于刻意改变样本数目（重采样法）或刻意扭曲数据分布（重权重法），它们在一定程度上会损害深度网络学习到的深度特征的表示能力。基于该发现提出了一个双分支神经网络结构用来同时兼顾特征学习和分类器学习，将深度模型的这两个重要模块进行解耦，从而保证两个模块互不影响，共同达到优异的收敛状态，协同促进深度网络在长尾数据分布上的泛化性能。

如图 3 所示，我们设计了两个网络分支分别负责通用特征学习和处理长尾分布带来的类别不平衡，之后在最后的特征层面（f_c、f_r），以及预

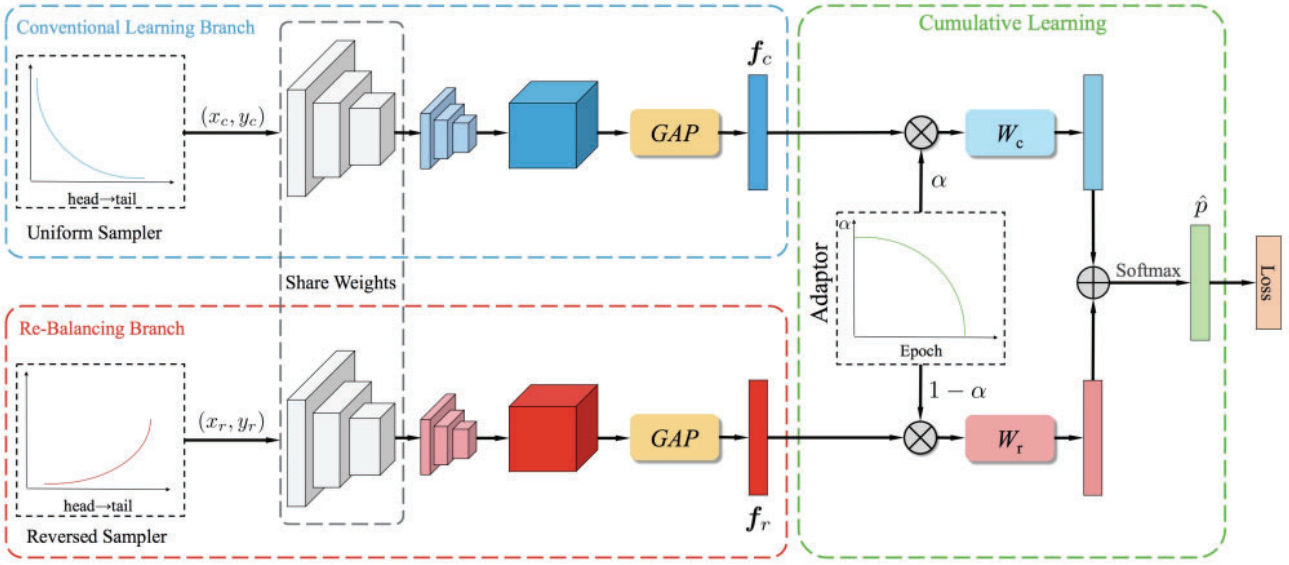

图 3　双分支神经网络

测结果层面（$W_c^T f_c$ 、$W_r^T f_r$）均施加了一个随训练轮数平方递减的调和因子 α。经实验验证，我们的方法在多个长尾分布的标准数据集（iNaturalist 2017/2018、CIFAR-10-LT 和 CIFAR-100-LT） 上均取得了目前最佳的视觉识别性能。

4 结束语

长尾数据分布在日常生活的诸多应用场景广泛出现，但目前针对长尾数据分布，特别是深度学习方向的研究工作还处于起步阶段，未来还有很大的研究和发展空间。现有的针对长尾数据分布的深度视觉识别的研究，主要集中在比较直接的损失函数设计，以及传统机器学习技术（如类别不平衡和代价敏感方法）的应用上，最近一段时间将特征学习和分类器学习解耦的思路逐渐崭露头角变成主流，相信不久的将来，结合更加深入分析深度神经网络本质特性的解决长尾数据分布问题的网络结构和解决方案会被陆续提出。当然，除了视觉识别任务之外，如何处理视觉检测等任务中的长尾数据分布问题也是值得进一步深入研究的课题。

作者介绍

魏秀参

博士，旷视南京研究院院长，南京大学学生创业导师。主要研究领域为计算机视觉和机器学习。

（上接第 25 页）

集中在全监督模式上，通过使用深度神经网络结合人工先验的方式，对合成数据中雨纹（滴）的分布进行拟合，训练出一个端到端的去雨网络。这些方法一方面无法利用真实图像数据进行训练，另一方面也无法很好地泛化到真实去雨任务中。无监督和半监督的单图像深度学习去雨方法可以有效弥补这些不足，但是由于可用的先验信息少，因而研究难度更大，关于此类研究工作目前还处于起步阶段，相关方法还比较少，未来还有很大的研究空间。此外，单图像雨域和非雨域间的迁移学习、图像领域和视频领域去雨的迁移学习等将是未来图像去雨研究值得关注的一些方向。

（参考文献略）

作者介绍

张召

合肥工业大学黄山学者特聘教授，CAAI 机器学习专委会委员和模式识别专委会委员等。主要研究方向为机器学习、深度表示学习等。

韦炎炎

合肥工业大学博士研究生。主要研究方向为机器学习与计算机视觉基础理论研究及其在图像复原中的应用。