doi:10.3969/j.issn.1672-1292.2022.02.010

长尾识别研究进展

张　明[1,2],翟俊海[1,2],许　垒[1,2],高光远[1,2]

(1.河北大学数学与信息科学学院,河北 保定 071002)
(2.河北大学河北省机器学习与计算智能重点实验室,河北 保定 071002)

[摘要]　长尾识别是目前深度学习领域最热门的研究方向之一,长尾识别的工作重点是解决长尾分布数据的计算机视觉识别任务. 长尾分布的显著特征为 2-8 分布,即 20%的类占据 80%的样本. 将少数几个类占据了大部分数据的类称之为头部类;而大多数类占据了很少部分数据的类称之为尾部类. 首先,列举解决长尾识别问题的各种方法. 然后,将其划分为重采样、重加权、迁移学习、解耦特征学习和分类器学习以及其它方法进行阐述. 最后,阐述对相关方法的理解.

[关键词]　深度学习,长尾识别,计算机视觉,研究方法,神经网络
[中图分类号]TP181　[文献标志码]A　[文章编号]1672-1292(2022)02-0063-10

Research Advance in Long-tailed Recognition

Zhang Ming[1,2],Zhai Junhai[1,2],Xu Lei[1,2],Gao Guangyuan[1,2]

(1.School of Mathematics and Information Science,Hebei University,Baoding 071002,China)
(2.Hebei Key Laboratory of Machine Learning and Computational Intelligence,Hebei University,Baoding 071002,China)

Abstract:Long tail recognition is one of the most popular research directions in the field of deep learning. The focus of long tail recognition is to solve the computer vision recognition task of long-tail distributed data. The prominent feature of the long-tail distribution is the 2-8 distribution,that is,20% of the classes account for 80% of the sample. We call a class with a few classes that make up most of the data a header class. Classes where most classes occupy a small portion of the data are called tail classes. Firstly, various methods are introduced to solve the problem of long tail recognition. Then, they are divided into resampling,re-weighting,transfer learning,decoupling feature learning,classifier learning and other methods. Finally, our understanding of the related methods are introduced.

Key words:deep learning,long-tailed recognition,computer vision,research method,neural network

近年来,深度学习在图像分类[1]、目标检测[2]、人脸识别[3]等计算机视觉任务中取得了巨大的成绩. 然而,当深度学习模型遇到长尾分布数据时,往往会表现不佳. 这是因为长尾分布数据集的头部少数类占据了大多数数据,而尾部多数类却占据了很少的一部分数据. 深度学习模型在处理长尾分布数据集时会偏向头部类,但对其期望是可以在所有类上都表现良好,而不是偏向头部类. 长尾识别的本质便是这两者之间的不匹配问题[4].

由于尾部类也是整个长尾分布数据集的重要组成部分,所以仅仅为了获得均匀的分布而将其去除是不可取的. 但传统的机器学习方法和具有先进性能的深度模型对长尾分布建模是具有挑战性的.

将长尾学习困难的本质分为以下 3 个方面:一是长尾分布数据集中尾部类样本太少,整个数据集的不平衡程度太高. 二是深度模型的损失由头部类主导,使分离的超平面严重偏离尾部类. 三是尾部类数据太少,类内多样性太低.

我们并以这些困难作为解决方案的线索将现有的解决方案分为 5 类:一是重采样,其策略是试图构造出平衡的数据集. 二是重加权,其策略是在损失函数中给小的类别分配大的权重. 三是迁移学习,分别对

收稿日期:2021-08-31.
基金项目:河北省科技计划重点研发项目(19210310D)、河北省自然科学基金项目(F2021201020).
通讯作者:翟俊海,博士,教授,研究方向:机器学习、云计算与大数据处理、深度学习. E-mail:mczjh@126.com

头部类和尾部类建模,将学到的头部类的信息、表示、知识迁移给少数类使用. 四是解耦特征学习和分类器学习,由于类别重新平衡方法(重采样和重加权)都会损害深度网络学习到的深度特征的表示能力,该方法把不平衡学习分为两个阶段,在特征学习阶段正常采样,在分类器学习阶段平衡采样,以此来克服类别重新平衡方法的缺点. 五是其它方法,是一些长尾学习在其它领域的相关研究工作.

本文针对已经提出解决长尾识别的方法进行综述,并对这些方法按照重采样、重加权、迁移学习、解耦特征学习和分类器学习方法以及其它方法进行分类. 对每一种方法的创新点、优缺点进行论述,这为从事相关研究的人员提供了一些有价值的内容.

1 相关工作

长尾识别问题涉及数据不平衡问题和少样本问题. 下文分 5 类介绍,包括重采样、重加权、迁移学习、解耦特征学习和分类器学习以及其它方法.

1.1 重采样

数据重采样可以分为两种类型:一种是对拥有小部分数据的类别进行过采样[5-7]. 由于随机过采样这种重复采样方法往往会导致严重的过拟合,所以现在主流的过采样方法是数据合成[8-10]. 即通过某种方式人工合成一些少数类样本,从而达到类别平衡的目的. 经典方法 Smote[11] 的思路就是选取任意的少数类样本,用 K 近邻[12] 选取其相似样本,通过对样本线性插值得到新样本. 生成对抗网络[13] 在最近几年中取得了显著的进步. 然而,增加具有高多样性的样本仍然具有挑战性.

另一种是对于拥有大部分数据的类别进行欠采样[14-15]. 随机欠采样是从多数类样本中随机选取一些样本去除掉. 这种方法的缺点是被去除的样本可能包含着一些重要信息,这样会导致学习出来的模型效果不好. 而 EasyEnsemble 和 BalanceCascade[16] 便是采用集成学习的机制来处理传统随机欠采样中的信息丢失问题.

1.2 重加权

经典的重加权方法[17-20] 通常在目标函数上对尾部类的训练数据施加较大惩罚,借此克服不平衡问题. 一般而言,损失函数中的惩罚权重与类别对应样本数成反比. 但最近,Cui 等提出了一种基于有效样本数来重新加权损失的方法,替代了之前根据样本数目比例确定惩罚权重的做法[21].

除了上述在类别上调整权重的方法外,还有一些研究侧重于在样本上重新加权损失[22-24]. 这些方法通过提高难学样本的权重和降低易学样本的权重,来进行重新加权损失.

目前,重加权方法还包括对已有损失函数的改进[25] 和提出新型的损失函数[26]. 其中,对分类器做正则化[27] 和修改损失函数来为不同的类施加不同的惩罚的方法都有局限性. 分类器权重正则化与优化器的选择有关,修改损失函数大多不能满足损失函数在降到最低时,错误率也要达到最小. 因此,Menon 等[28] 提出了一种基于 logit 的训练方法来克服重加权方法带来的不稳定影响.

1.3 迁移学习

迁移学习[29-32] 通常分别对头部类和尾部类建模,将学到的头部类的信息、表示、知识迁移给少数类使用. Wang 等[33] 构造了一个元网络,学习如何逐步将元知识从头部类迁移到尾部类. Mostafa 等[34] 提出了一种长尾信息网络模型,使用元数据驱动的方法在网络发布的数据集之间建立语义信息网络,改善了跨学科的信息可重用性.

1.4 解耦特征学习和分类器学习

解耦特征学习和分类器学习方法是把不平衡学习分为两个阶段,在特征学习阶段正常采样,在分类器学习阶段平衡采样,这样可以带来更好的学习效果,这也是我目前认为最优的长尾识别方法.

Zhou 等[35] 提出了一种具有特定累积学习策略的双边分支网络,彻底改善了长尾识别的性能. Kang 等[36] 提出了一个解耦的训练模式,首先联合特征学习和分类器学习,然后通过使用类别平衡采样方法重新训练分类器来获得平衡分类器,这样有助于模型更好的学习尾部类特征.

1.5 其它方法

此外,还有一些其它类型的方法[37-38] 来解决长尾识别问题. Ma 等[39] 提出了一种新的聚合-分散训练方式,从而实现了对每个类别的准确分类. Tong 等[40] 提出了一种低复杂度的大规模多标签学习算法,其

目标是通过自适应地修剪尾部标签来促进模型的快速预测.

以上都是用来解决长尾识别问题的,有些方法虽然不是专门为解决长尾问题而设计的,但是也可以部分解决. 下面对其中的一些方法进行详细的论述,并对每一类方法的创新点、优缺点进行讨论.

2　主要方法

2.1　数据重采样

2.1.1　Smote

过采样中的数据合成方法是利用已有样本生成更多样本,其中最常见的一种方法为 Smote[11],它利用少数类样本在特征空间的相似性来生成新样本. 对于少数类样本 $x_i \in S_{\min}$,从它属于少数类的 K 近邻中随机选取一个样本点 $\hat{x}_i$,生成一个新的少数类样本 x_{new},

$$x_{\text{new}} = x_i + (\hat{x} - x_i) * \delta. \tag{1}$$

式中,$\delta \in [0,1]$,是一个随机数.

Smote 为每个少数类合成相同数量的新样本,这样就存在一些潜在的问题:一方面是增加了类别之间重叠的可能性,另一方面是生成一些没有提供有益信息的样本. 为了解决这个问题,Han 等[8] 提出了 Borderline-Smote 方法.

Borderline-Smote 为每个少数类样本计算 K 近邻,但只为其中 K 近邻中有一半以上多数类样本的少数类样本生成新样本. 直观地讲,只为那些周围大部分是多数类样本的少数类样本生成新样本,因为这些样本往往是边界样本. 在确定了为哪些少数类样本生成新样本后再利用 Smote 生成新样本.

2.1.2　NearMiss

NearMiss 本质上是一种原型选择方法,即从多数类样本中选取最具代表性的样本用于训练,主要是为了缓解随机欠采样中的信息丢失问题. NearMiss 采用一些启发式的规则来选择样本,根据规则的不同可分为 3 类:

(1) NearMiss-1:选择到最近的 K 个少数类样本平均距离最近的多数类样本.

(2) NearMiss-2:选择到最远的 K 个少数类样本平均距离最近的多数类样本.

(3) NearMiss-3:对于每个少数类样本选择 K 个最近的多数类样本,目的是保证每个少数类样本都被多数类样本包围.

总的来说,重采样就是在已有数据不平衡的情况下,人为地让模型学习时接触到的训练数据是类别平衡的. 不过由于尾部类的少量数据往往被反复学习,缺少足够多的样本差异,鲁棒性较差. 而头部拥有丰富多样性的大量数据又往往得不到充分学习. 因此,重采样在非平衡程度相对较小的长尾分布数据集上往往可以取得比较好的效果. 但是如果在极端失衡的长尾分布下,这种方法的效果就微乎其微了,所以重采样也并非是个真正完美的解决方案.

2.2　重加权

2.2.1　基于有效样本数的重加权方法

在直觉上,我们认为数据越多越好,但是由于数据之间存在信息重叠,随着样本数量的增加,模型可以从数据中提取的有益信息逐渐减少. 基于此观察,Cui 等[21] 提出了一种考虑数据重叠来帮助量化有效样本数的方法,它与每个类别的有效样本数成反比来重新加权损失.

给定一个类,将其特征空间中所有可能的数据集合表示为 S. 我们假设 S 的体积为 N,$N \geqslant 1$. 将每个样本表示为 S 的一个子集,其单位体积为 1,并且可能与其他样本重叠. 本文定义样本的有效数量是样本的期望体积. 将样本的有效数量(期望体积)表示为 E_n,其中 n 是样本数量.

$$E_n = (1-\beta^n)/(1-\beta). \tag{2}$$

式中,$\beta \in [0,1]$,$\beta = (N-1)/N$ 是一个超参数.

当 N 较大时,有效样本数与样本数 n 趋近相同. 这时唯一的原型数 N 较大,因此不存在数据重叠,每个样本都是唯一的. 在另一个极端,如果 $N=1$,这意味着只有一个原型,所以这个类中的所有数据都可以通过数据扩充、转换等由这个原型表示.

为了解决不平衡数据的训练问题,文中还提出了一种类别平衡损失. 引入一个与有效样本数成反比的加权因子,该因子与 i 类样本的有效样本数量成反比:$\alpha_i = 1/E_{n_i}$. 类别平衡损失为:

$$\mathcal{L}(p,y) = \frac{1}{E_{n_y}}\mathcal{L}(p,y) = \frac{1-\beta}{1-\beta_{n_y}}\mathcal{L}(p,y). \tag{3}$$

式中,n_y 是真实类别 y 中的样本数. $\beta=0$ 对应没有重新加权损失,而 $\beta\to1$ 对应通过相反的类别频率重新加权损失.

文中在大规模数据集上进行了对比实验,证实了类别平衡损失比 Softmax 损失、Sigmoid 损失、Focal 损失[23]等有了显著的性能提升. 但是该方法还做不到给每个类别都找到超参数. 下面通过在数据分布中纳入合理的假设或设计,用基于自适应的方法来扩展当前方法的框架.

2.2.2　均衡损失

当某个类别的样本用于训练时,模型对其它类别的预测参数就会受到副作用梯度的影响,使得对其它类别的预测概率变低. 由于稀有类别的样本几乎不会出现,因此在网络参数更新期间,起副作用的梯度会严重影响这些类别的预测参数.

为了解决这个问题,Tan 等[25]提出了一种均衡损失函数,它忽略了尾部类的梯度. 在网络参数更新过程中,保护了模型对尾部类的学习能力.

在原始的 Softmax 交叉熵损失函数中引入权重项 w,以减少负样本对尾部类的影响. 均衡损失可以表示为

$$\mathcal{L}_{SEQL} = -\sum_{j=1}^{C} y_j \lg(\tilde{p}_j). \tag{4}$$

式中,

$$\tilde{p}_j = \frac{\mathrm{e}^{z_j}}{\sum_{k=1}^{C} w\mathrm{e}^{z_k}}. \tag{5}$$

并且权重项 w 为

$$w = 1-\beta T_\lambda(f_k)(1-y_k). \tag{6}$$

式中,β 是一个随机变量,用于维护负样本的梯度. f_k 是类别 k 在数据集中出现的频率. $T_\lambda(x)$ 是阈值函数,当 $x<\lambda$ 时输出 1,否则输出 0. 对于数量频率低于阈值的稀有类别,忽略负样本的副作用梯度. 文中利用 λ 将尾部类别与所有其他类别区分开来.

通过大规模数据集进行了对比实验,证实了均衡损失比 Focal 损失[23]、Class-aware 采样[6]、Repeat Factor 采样[41]和类别平衡损失[21]有了显著的性能提升. 这种方法简单而有效,在目标检测和图像分类方面效果显著.

2.3.3　距离损失

Zhang 等[42]在研究了长尾数据对神经网络训练的影响时,提出了一种新的距离损失函数,以便在训练过程中有效地利用尾部数据.

距离损失与中心损失[43]很类似,都是增大了类别间的距离,减小了类别内的距离. 但是距离损失还可以减小尾部部分数据对识别率的影响.

距离损失可以表示为

$$\mathcal{L}_R = \alpha\mathcal{L}_{R_{\text{intra}}} + \beta\mathcal{L}_{R_{\text{inter}}}. \tag{7}$$

$\mathcal{L}_{R_{\text{intra}}}$ 表示类别内损失

$$\mathcal{L}_{R_{\text{intra}}} = \sum_{i\subseteq I}\mathcal{L}^i_{R_{\text{intra}}} = \sum_{i\subseteq I}\frac{k}{\sum_{j=1}^{k}\frac{1}{D_j}}. \tag{8}$$

文中利用调和平均值来解决类别内的距离,其中 k 的值,按照经验设置为 2.

$\mathcal{L}_{R_{\text{inter}}}$ 表示类别间的损失

$$\mathcal{L}_{R_{\text{inter}}} = \max(M - D_{\text{Center}}, 0). \tag{9}$$

最终损失函数可以表示为

$$\mathcal{L}=\mathcal{L}_S+\lambda\mathcal{L}_R. \tag{10}$$

式中,$\mathcal{L}_S$ 是 Softmax 交叉熵损失函数. 为了在提高模型分类能力,联合采用距离损失和 Softmax 损失.

在 LFW 和 YTF 数据集上进行了对比实验,证实了距离损失比 Softmax 损失、中心损失[43]、DeepFace[44]等有了显著的性能提升,为解决长尾识别问题提供了新方法.

通过对损失函数的改进来解决长尾识别问题. 大部分实现简单,往往只需修改下损失函数,就可以取得非常具有竞争力的结果. 但重加权方法也是具有局限性的,因为其以牺牲深度网络的特征学习模块的性能为代价,显著提升了深度网络的分类器学习模块的性能,所以重加权也并非是个真正完美的解决方案.

2.3　迁移学习

2.3.1　开放长尾识别

2019 年,UC Berkeley 的研究人员们深入研究了视觉识别问题的背景和设定,重新定义了开放长尾识别问题,通过融合不平衡分类、少样本学习和开放集识别三方面,大幅度提升长尾数据识别的表现[45].

长尾识别不仅要处理封闭世界中的不平衡分类和少样本学习,而且还要处理开放集识别. 基于此观察,Liu 等[45]提出了一种集成的长尾识别模型,来处理开放集识别.

如图 1 所示,提出的长尾识别模型主要由两个模块组成:动态元嵌入和调节注意力. 前者在头部类和尾部类之间建立联系并传递知识,而后者则在头部类和尾部类之间保持区别.

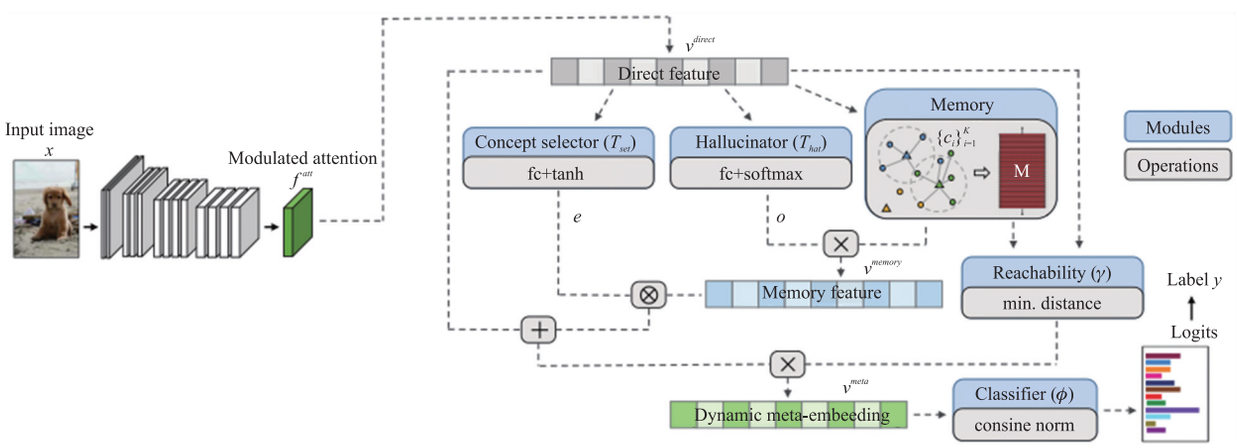

图 1　长尾识别模型

Fig. 1　Long-tailed recognition model

由于长尾分布的头部类数据足够丰富,因此特征提取器一般可以很好地提取出头部类的特征. 而尾部类严重缺乏数据,导致特征提取器无法很好地提取出这些类别的特征.

为了解决尾部类问题,文中引入了一个记忆模块,里面存储着每个类的原型. 通过类似中心损失的方式考虑了类别内与类别间信息构造出来的. 当用特征提取器提取特征时,会从记忆模块中借鉴一些有用的信息来扩充当前的特征表示. 另外,提出了用概念选择器来控制融合特征的数量和种类. 由于头部类别已经具有丰富的数据,只需小部分特征用于针对它们的融合. 而对于尾部类别来说,数据较少所以记忆特征中对它们有着较大的提升作用.

通过实验发现,对特征图添加空间注意力可以进一步增强特征的判别能力. 对于不同类别的图像,具有判别能力的信息往往分布在图片的不同位置上,如果能自适应地给出一个注意力更关注这些位置,或许会使最终学到的特征更适合分类这样. 所以为了让模型能具备这种自适应的提取判别特征的效果,文中提出了调节注意力,它由图像上的自注意力和空间注意力构成.

将这一模型应用到 ImageNet-LT、Places-LT 数据集上,比 Focal 损失[23]、距离损失[42]、Plain Model[46]等有了显著的性能提升.

基于开放长尾识别问题的视觉任务更适合于数据的自然分布,能够更准确贴切地描述真实状况,将会为目标检测、分割和强化学习带来新的提升.

2.3.2　为尾部类样本构造特征云

2020 年,Liu 等提出了一种为尾部类样本构造特征云的方法,在训练过程中来扩展尾部类的分布,在特征空间中将头部类的类内分布转移到尾部类[47]. 目标是使尾部类在训练中实现与头部类相似的类内角度变化.

首先,计算头部类特征与其对应类中心之间的夹角分布. 然后平均所有头部类的角方差,就得到了头部类的总体方差. 然后,考虑将头部类的方差传递给每个尾部类. 文中提出在每个尾部类实例的深层特征空间中增加一定的扰动. 随着特征向量的增加,一个特定的特征向量变成一组散布在其周围的可能特征,这被称为特征云.

特征云的本质是一个概率模型. 既然尾部类自身由于样本数量稀少,导致类内多样性不足,那么就要向头部类学习,将头部类的类内多样性迁移到尾部类用特征云来填充尾部类的特征空间.

每个具有相应特征云的实例都会有一个相对较大的分布范围,使得尾部类与头部类具有相似的角度分布. 因此,它减轻了学习到的特征空间的失真,并改善了对长尾数据的深度表示学习.

大规模进行了对比实验,证实了文中方法比 Advesarial[48]、AACN[49]等有了显著的性能提升,为解决长尾识别问题提供了新的思路.

2.3.3　IEM

通常用一个原型表示一个类别,但是长尾数据有更高的类内方差,导致学习起来更困难. 因此,Zhu 等[50]引入膨胀片段记忆(inflated episodic memory,IEM)来存储每个类别最具判别性的特征. 同时,IEM 各个类的参数是独立更新的,不会受到多数类的影响.

IEM 遵循键值存储格式. IEM 中的每个键存储器都对应一个值存储器. IEM 通过查找与键存储器中相关关键字之间的相似性度量来生成权重. 当进行读取操作时,可以获得检索到的预测值,使用均方差损失(mean square error,MSE)来评估检索到的预测值与真实标签之间的距离.

提出一种新的区域自注意力机制(regional self-attention,RSA),来从特征图中提取局部特征. 在训练过程中,在分类前的最后一个卷积块插入 RSA. 然后聚合区域统计信息,在每个尺度上为区域生成一个特征. 为所有类别提供了更强大的区分功能,从而提高了识别能力.

模型的整体损失为分类的交叉熵损失和 IEM 的 MSE 损失.

$$\mathcal{L}=CE(y,y')+MSE(READ(M_y),1)+MSE(READ(M_{\hat{y}}),0). \tag{11}$$

式中,$READ$ 是 IEM 的读取操作.

大规模进行了对比实验,证实了 IEM 比 Focal 损失[23]、距离损失[42]、OLTR 等有了显著的性能提升[45]. 通过引入 IEM 来存储每个类别最具判别性的特征,加快了尾部类的学习速度,同时也加强了模型的识别能力,但模型的泛化性能不高.

对于迁移学习方法而言,从头部类中学习通用知识,然后迁移到尾部类中的这种理念和实际效果都非常好. 但是这类方法实现起来比较困难,因为往往还需要设计额外比较复杂的模块. 不过目前的偏好也并非绝对,也许未来可以设计出简单有效的迁移模型.

2.4　解耦特征学习和分类器学习

2.4.1　双边分支网络(bilateral-branch network,BBN)

2020 年,Zhou 等[35]首次揭示了重采样和重加权这类类别重新平衡方法,其奏效的原因在于它们显著提升了深度网络的分类器学习的性能. 但是,也损害了所学习的深层特征的表示能力.

基于这一观察,如图 2 所示,Zhou 等提出了一个 BBN,以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.

BBN 主要由 3 部分组成:常规学习分支、再平衡分支、累计学习策略. 常规学习分支和再平衡分支分别用于特征学习和分类器学习. 这两个分支使用了同样的残差网络结构,除最后一个残差模块,两个分支的网络参数是共享的. 为这两个分支分别配备均匀采样器和逆向采样器,得到两个样本(x_c,y_c)和(x_r,y_t)作为输入数据,其中前者用于常规学习分支,后者用于再平衡分支. 然后,将这两个样本送入各自对应的分支后,通过残差网络和全局平均池化(gobal overage pooling,GAP)得到特征向量$\boldsymbol{f}_c$和$\boldsymbol{f}_r$. 最后,通过使用

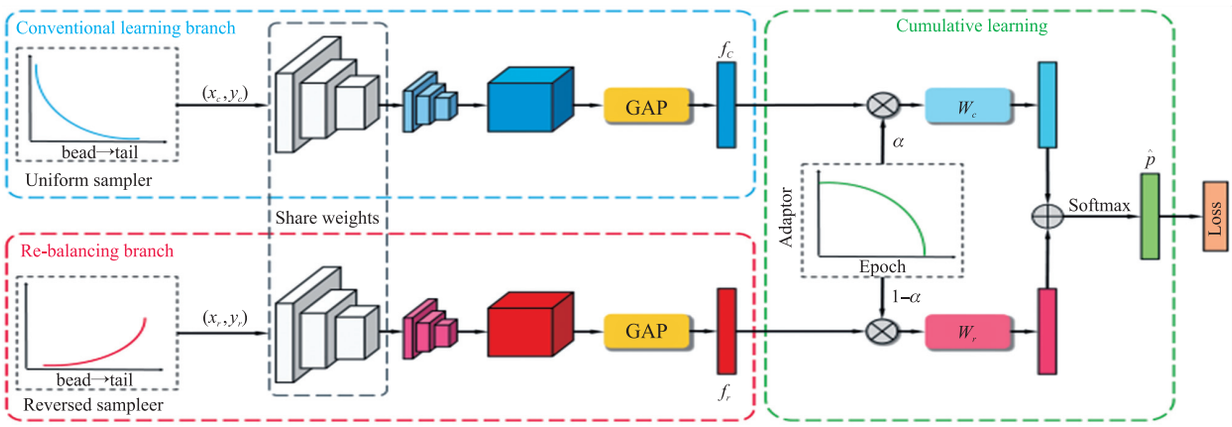

图 2　双边分支网络的框架

Fig. 2　The framework of a bilateral branch network

一个自适应权衡参数 α 来控制 $\boldsymbol{f}_c$ 和 $\boldsymbol{f}_r$ 的权重,将加权特征向量 $\alpha\boldsymbol{f}_c$ 和 $(1-\alpha)\boldsymbol{f}_r$ 分别发送到分类器 W_c 和 W_r,再通过逐元素累加的方式将其输出整合到一起.

经实验验证,BBN 在多个长尾分布的标准数据集(iNaturalist2017/2018、CIFAR-10-LT 和 CIFAR-100-LT)上均取得了目前最佳的视觉识别性能,彻底改善了长尾任务的识别性能.

2.4.2　一种辅助学习方法

Zhang 等[51]提出了一种简单有效的辅助学习方法. 其核心思想与 BBN 相同,也是将深度模型的分类器和特征提取两部分进行解耦,然后对每个部分采用不同的训练策略.

为了避免训练过程被头部类主导,采用类别平衡采样方法(class balanced sampling,CBS)对整个网络进行训练. 为了避免过拟合的风险,对于特征提取部分,提出了一个辅助训练方法,在常规随机采样(routine random sampling,RRS)方法下训练一个分类器.

深度神经网络可以分解为特征提取器 $\varphi(\cdot)$ 和分类器 $h(\cdot)$. 分类器 $h(\cdot)$ 使用 CBS 方案进行训练,在原分类器 $h(\cdot)$ 之外构造了另一个分类器 $h_a(\cdot)$. $h(\cdot)$ 和 $h_a(\cdot)$ 都附加到相同的特征提取器 $\varphi(\cdot)$ 上,进行联合训练.

训练 $h_a(\cdot)$ 时不使用 CBS,而是使用 RRS. 这样分类器只受 CBS 训练的影响,而特征提取器是从 CBS 和 RRS 方案中学习的. CBS 损失的头部类信息可以通过 $h_a(\cdot)$ 来恢复. 因此,特征提取器可以充分利用整个数据集的信息,从而避免了过拟合问题.

通过实验证明,此方法比 Focal 损失[23]、距离损失[42]和 Memory Net[50]具有更好的性能.

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. 解耦特征学习和分类器学习这类方法,将深度模型这两个模块进行解耦,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. 该方法简单有效,易于实现,是目前最优的长尾识别解决方法.

2.5　其它方法

长尾的方法,一般以牺牲头部类的性能为代价,提升了尾部类的性能. Wang 等[52]设计了一种能同时提高头部类和尾部类性能的模型融合方法. 首先,训练多个 classifiers experts,这些 experts 共享一部分模块,再对 classifiers experts 进行差异性地训练. 然后,设计了 expert assignment 模块来处理难学样本. 对难学样本特殊处理这一做法并不新鲜,比如 Focal Loss[23]. 但文中设计的模块可以动态调整,而不是像 Focal Loss 依赖先验的统计进行加权.

因为设计了多个 classifier experts,而一些简单样本显然不需要这么多的 experts,所以就根据样本难易程度进行动态分配. 简单来说就是通过一个路由模块,去动态地决定哪些 classifier expert 应该参与分类,这样可以更高效地对难学和易学样本,进行不同程度的处理. 之后对多个 classifier experts 的输出取几何均值.

该方法在包括 CIFAR100-LT、ImageNet-LT 和 iNaturist 的数据集上测试,表现明显优于 Focal 损失[23]、OLTR[45]等方法. 但是此方法跟长尾分布似乎没有太大关系,对于多个模型的融合很明显会提升分类性能. 而且模型融合似乎对尾部类性能的提升更明显.

3 结论

本文对解决长尾识别问题的若干方法进行了介绍和讨论,一些传统的非平衡学习方法也可以用来解决长尾识别问题. 例如,重采样和重加权方法,可以直接影响深度网络中分类器权重的更新,从而促进了分类器的学习. 但是,它们也在一定程度上损害所学习深层特征的表示能力. 另外,还介绍了一些迁移学习、解耦特征学习和分类器学习的方法. 迁移学习方法通过从头部类中学习通用知识,然后迁移到尾部类中的思想取得了很好的效果. 但是它们往往还需要设计额外比较复杂的模块,并且还存在领域偏移问题. 解耦特征学习和分类器学习方法其核心思想是将深度模型的特征学习和分类器学习两部分进行解耦,然后对每个部分采用不同的训练策略. 这类方法简单有效,易于实现,是我目前认为最优的长尾识别方法. 关于长尾识别的研究仍然有一些值得进一步深入研究的问题,例如,面向大数据环境的长尾可视识别研究、面向开放环境的长尾识别研究等,这些都值得我们进一步探索.

[参考文献](References)

[1] KRIZHEVSKY A,SUTSKEVER I,HINTON G E. Imagenet classification with deep convolutional neural networks[C]// Conference and Workshop on Neural Information Processing Systems. California,USA,2012:1097-1105.

[2] GIRSHICK R,DONAHUE J,DARRELL T,et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//IEEE Conference on Computer Vision and Pattern Recognition. Columbus,OH,USA,2014:580-587.

[3] MASI I,WU Y,HASSNER T,et al. Deep face recognition:A survey[J/OL]. http://arXiv.org/abs/1804.06655v8.

[4] JAMAL M A,BROWN M,YANG M H,et al. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective[C]//IEEE Conference on Computer Vision and Pattern Recognition. Seattle,USA,2020:7607-7616.

[5] JAPKOWICZ N,STEPHEN S. The class imbalance problem:a systematic study[J]. Intelligent Data Analysis,2002,6(5):429-449.

[6] SHEN LI,LIN Z C,HUANG Q M. Relay backpropagation for effective learning of deep convolutional neural networks[C]// European Conference on Computer Vision. Amsterdam,Netherlands:Springer,2016:467-482.

[7] HE H,GARCIA E A. Learning from imbalanced data[J]. IEEE Transactions on Knowledge and Data Engineering,2009,21(9):1263-1284.

[8] HAN H,WANG W Y,MAO B H. Borderline-smote:a new over-sampling method in imbalanced data sets learning[J]. Lecture Notes in Computer Science,2005:878-887.

[9] GAO H,SHOU Z,ZAREIAN A,et al. Low-shot learning via covariance-preserving adversarial augmentation networks[J]. Neural Information Processing Systems,2018,31:975-985.

[10] MACIEJEWSKI T,STEFANOWSKI J. Local neighbourhood extension of smote for mining imbalanced data[C]//IEEE International Conference on Data Mining. Paris,France,2011:104-111.

[11] CHAWLA N V,BOWYER K W,HALL L O,et al. Smote:synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research,2002:321-357.

[12] COVER T,HART P. Nearest neighbor pattern classification[J]. IEEE Transactions on Information Theory,1967,13(1):21-27.

[13] GOODFELLOW I J,POUGET A J,MIRZA M,et al. Generative adversarial networks[J]. Advances in Neural Information Processing Systems,2014,3:2672-2680.

[14] DRUMMOND C,HOLTE R C. C4.5,Class imbalance,and cost sensitivity:Why under-sampling beats over-sampling[C]// Workshop on Learning from Imbalanced Datasets II. Washington,DC,USA,2003:1-8.

[15] BUDA M,MAKI A,MAZUROWSKI M A. A systematic study of the class imbalance problem in convolutional neural networks[J]. Neural Networks,2018,106:249-259.

[16] LIU X Y,WU J,ZHOU Z H. Exploratory undersampling for class-imbalance learning[J]. IEEE Transactions on Systems,Man,and Cybernetics,Part B(Cybernetics),2008,39(2):539-550.

[17] TING K M. A comparative study of cost-sensitive boosting algorithms[C]//International Conference on Machine Learning.

Ithaca,New York,USA,2000:983-990.

[18] ZADROZNY B,LANGFORD J,ABE N. Cost-sensitive learning by cost-proportionate example weighting[C]//Third IEEE International Conference on Data Mining. Melbourne,FL,USA,2003:435.

[19] MIKOLOV T,SUTSKEVER I,CHEN K,et al. Distributed representations of words and phrases and their compositionality[J]. Advances in Neural Information Processing Systems,2013:3111-3119.

[20] HUANG C,LI Y N,TANG X O,et al. Learning deep representation for imbalanced classification[C]//IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas,USA,2016:5375-5384.

[21] CUI Y,JIA M L,LIN T Y,et al. Class-balanced loss based on effective number of samples[C]//IEEE Conference on Computer Vision and Pattern Recognition. Los Angeles,USA,2019:9268-9277.

[22] LI B,LIU Y,WANG X. Gradient harmonized single-stage detector[C]//AAAI conference on artificial intelligence. Honolulu,Hawaii,USA,2019,33(1):8577-8584.

[23] LIN T Y,GOYAL P,GIRSHICK R,et al. Focal loss for dense object detection[C]//International Conference on Computer Vision. Venice,Italy,2017:2980-2988.

[24] DONG Q,GONG S G,ZHU X T,et al. Class rectification hard mining for imbalanced deep learning[C]//IEEE International Conference on Computer Vision. Venice,USA,2017:1869-1878.

[25] TAN J R,WANG C B,LI B Y,et al. Equalization loss for long-tailed object recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. Seattle,USA,2020:11659-11668.

[26] CAO K D,WEI C L,GAIDON A,et al. Learning imbalanced datasets with label-distribution-aware margin loss[C]//Neural Information Processing Systems. Vancouver,Canada,2019:1-18.

[27] ZHOU Y C,HU Q H,WANG Y,et al. Deep super-class learning for long-tail distributed image classification[J]. Pattern Recognition,2018,80:118-128.

[28] MENON A K,JAYASUMANA S,RAWAT A S,et al. Long-tail learning via logit adjustment[J/OL]. http://arXiv.org/abs/2007.07314.

[29] MAHAJAN D,GIRSHICK R,RAMANATHAN V,et al. Exploring the limits of weakly supervised pretraining[C]//European Conference on Computer Vision. Munich,Germany,2018:181-196.

[30] YIN X,YU X,SOHN K,et al. Feature transfer learning for face recognition with under-represented data[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach,Los Angeles,USA,2019:5704-5713.

[31] PAN S J,YANG Q. A survey on transfer learning[J]. IEEE transactions on knowledge and data engineering,2010,22(10):1345-1359.

[32] ZAMIR A R,SAX A,SHEN W. Taskonomy:disentangling task transfer learning[C]//IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City,USA,2018.

[33] WANG Y X,DEVA R,MARTIAL H,et al. Learning to model the tail[C]//Conference and Workshop on Neural Information Processing Systems. California,USA,2017:7032-7042.

[34] MOSTAFA M E,PRAVEEN K,LUIGI M. Identification and characterization of information-networks in long-tail data collections[J]. Environmental Modelling & Software,2017:100-111.

[35] ZHOU B Y,CUI Q,WEI X S,et al. BBN:Bilateral-branch network with cumulative learning for long-tailed visual recognition[C]//Computer Vision and Pattern Recognition. Seattle,USA,2020:9716-9724.

[36] KANG B Y,XIE S,ROHRBACH M,et al. Decoupling representation and classifier for long-tailed recognition[C]//International Conference on Learning Representations. Montreal,USA,2020.

[37] ZHU X X,ANGUELOV D,RAMANAN D,et al. Capturing long-tail distributions of object subcategories[C]//IEEE Conference on Computer Vision and Pattern Recognition. Columbus,USA,2014:915-922.

[38] SINHA S,EBRAHIMI S,DARRELL T,et al. Variational adversarial active learning[C]//IEEE International Conference on Computer Vision. Seoul,Korean,2019:5972-5981.

[39] MA Y H,KAN M N,SHAN S G,et al. Learning deep face representation with long-tail data:anaggregate-and-disperse approach[J]. Pattern Recognition Letters,2020,133:48-54.

[40] TONG W,LI Y F. Does tail label help for large-scale multi-label learning[J]. IEEE Transactions on Neural Networks and Learning Systems,2020,31(7):2315-2324.

[41] GUPTA A,DOLLAR P,GIRSHICK R. LVIS:A dataset for large vocabulary instance segmentation[C]//IEEE Conference on Computer Vision and Pattern Recognition.Los Angeles,USA,2019.

[42] ZHANG X,FANG Z Y,WEN Y D,et al. Range loss for deep face recognition with long-tailed training data[C]//IEEE International Conference on Computer Vision. Venice,USA,2017:5419-5428.
[43] WEN Y,ZHANG K,LI Z,et al. A discriminative feature learning approach for deep face recognition[C]//European Conference on Computer Vision. Amsterdam,Netherlands,2016:499-515.
[44] TAIGMAN Y,YANG M,RANZATO M,et al. Deepface:closing the gapto human-level performance in face verification[C]//IEEE Conference on Computer Vision and Pattern Recognition. Columbus,USA,2014:1701-1708.
[45] LIU Z W,MIAO Z Q,ZHAN X H,et al. Large-scale long-tailed recognition in an open world[C]//IEEE Conference on Computer Vision and Pattern Recognition. Los Angeles,USA,2019:2532-2541.
[46] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas,USA,2016.
[47] LIU J L,SUN Y F,HAN C C,et al. Deep representation learning on long-tailed data:a learnable embedding augmentation perspective[C]//IEEE Conference on Computer Vision and Pattern Recognition. Seattle,USA,2020:2967-2976.
[48] HUANG H,LI D,ZHANG Z,et al. Adversarially occluded samples for person re-identification[C]//IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City,USA,2018:5098-5107.
[49] XU J,ZHAO R,ZHU F,et al. Attention-aware compositional network for person reidentification[C]//IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City,USA,2018:2119-2128.
[50] ZHU L C,YANG Y. Inflated episodic memory with region self-attention for long-tailed visual recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. Seattle,USA,2020:4343-4352.
[51] ZHANG J J,LIU L Q,WANG P,ET AL. To balance or not to balance:a simple-yet-effective approach for learning with long-tailed distributions[C]//IEEE Conference on Computer Vision and Pattern Recognition. Seattle,USA,2020.
[52] WANG X D,LIAN L,MIAO Z,et al. Long-tailed recognition by routing diverse distribution-aware experts[J/OL]. http://arXiv.org/abs/2010.01809.

[责任编辑:陈　庆]

(上接第 62 页)

[24] ARUMUGAM M S,RAO M V C. On the performance of the particle swarm optimization algorithm with various inertia weight variants for computing optimal control of a class of hybrid systems[J]. Discrete Dynamics in Nature and Society,2006(3):79295.
[25] NOMAN N,IBA H. Accelerating differential evolution using an adaptive local search[J]. IEEE Transactions on Evolutionary Computation,2008,12(1):107-125.
[26] GARCÍA S,MOLINA D,LOZANO M,et al. A study on the use of non-parametric tests for analyzing the evolutionary algorithms' behaviour:a case study on the CEC'2005 special session on real parameter optimization[J]. Journal of Heuristics,2009,15(6):617-644.
[27] GONG W Y,CAI Z H. Differential evolution with ranking-based mutation operators[J]. IEEE Transactions on Cybernetics,2013,43(6):2066-2081.
[28] WANG L J,ZHONG Y W. Cuckoo search algorithm with chaotic maps[J]. Mathematical Problems in Engineering,2015(1):715635.

[责任编辑:严海琳]