深度对比学习综述

张重生 [1] 陈 杰 [1] 李岐龙 [1] 邓斌权 [1] 王 杰 [1] 陈承功 [1]

摘 要 在深度学习中, 如何利用大量、易获取的无标注数据增强神经网络模型的特征表达能力, 是一个具有重要意义的研究问题, 而对比学习是解决该问题的有效方法之一, 近年来得到了学术界的广泛关注, 涌现出一大批新的研究方法和成果. 本文综合考察对比学习近年的发展和进步, 提出一种新的面向对比学习的归类方法, 该方法将现有对比学习方法归纳为 5 类, 包括: 1) 样本对构造; 2) 图像增广; 3) 网络架构; 4) 损失函数; 5) 应用. 基于提出的归类方法, 对现有对比研究成果进行系统综述, 并评述代表性方法的技术特点和区别, 系统对比分析现有对比学习方法在不同基准数据集上的性能表现. 本文还将梳理对比学习的学术发展史, 并探讨对比学习与自监督学习、度量学习的区别和联系. 最后, 本文将讨论对比学习的现存挑战, 并展望未来发展方向和趋势.

关键词 对比学习, 深度学习, 特征提取, 自监督学习, 度量学习

引用格式 张重生, 陈杰, 李岐龙, 邓斌权, 王杰, 陈承功. 深度对比学习综述. 自动化学报, 2023, **49**(1): 15−39

DOI 10.16383/j.aas.c220421

Deep Contrastive Learning: A Survey

ZHANG Chong-Sheng[1] CHEN Jie[1] LI Qi-Long[1] DENG Bin-Quan[1] WANG Jie[1] CHEN Cheng-Gong[1]

Abstract In deep learning, it has been a crucial research concern on how to make use of the vast amount of unlabeled data to enhance the feature extraction capability of deep neural networks, for which contrastive learning is an effective approach. It has attracted significant research effort in the past few years, and a large number of contrastive learning methods have been proposed. In this paper, we survey recent advances and progress in contrastive learning in a comprehensive way. We first propose a new taxonomy for contrastive learning, in which we divide existing methods into 5 categories, including 1) sample pair construction methods, 2) image augmentation methods, 3) network architecture level methods, 4) loss function level methods, and 5) applications. Based on our proposed taxonomy, we systematically review the methods in each category, and analyze the characteristics and differences of representative methods. Moreover, we report and compare the performance of different contrastive learning methods on the benchmark datasets. We also retrospect the history of contrastive learning and discuss the differences and connections among contrastive learning, self-supervised learning, and metric learning. Finally, we discuss remaining issues and challenges in contrastive learning and outlook its future directions.

Key words Contrastive learning, deep learning, feature extraction, self-supervised learning, metric learning

Citation Zhang Chong-Sheng, Chen Jie, Li Qi-Long, Deng Bin-Quan, Wang Jie, Chen Cheng-Gong. Deep contrastive learning: A survey. *Acta Automatica Sinica*, 2023, **49**(1): 15−39

近年来, 以深度学习为代表的新一代人工智能技术取得了迅猛发展, 并成功应用于计算机视觉、智能语音等多个领域. 然而, 深度学习通常依赖于海量的标注数据进行模型训练, 才能获得较好的性能表现. 当可用的标注数据较少、而无标注数据较多时, 如何提高深度学习的特征表达能力是亟需解决的重要现实需求. 自监督学习[1] 是解决该问题的有效途径之一, 能够利用大量的无标注数据进行自我监督训练, 得到更好的特征提取模型.

早期的对比学习起源于自监督学习, 通过设置实例判别代理任务完成自监督学习的目标. 具体而言, 对比学习首先对同一幅图像进行不同的图像增广, 然后衡量得到的图像对特征之间的相似性, 旨在使同一幅图像增广后的图像对特征之间的相似度增加, 而不同图像特征之间的相似度减小. 随着技术的发展, 对比学习已经扩展到监督和半监督学习中, 以进一步利用标注数据提升模型的特征表达能力.

近年来, 基于深度学习的对比学习技术取得了突飞猛进的发展. 典型的对比学习方法有 SimCLR[2] (Simple framework for contrastive learning of visual representations), MoCo[3] (Momentum con-

收稿日期 2022-05-22 录用日期 2022-08-22
Manuscript received May 22, 2022; accepted August 22, 2022
科技部高端外国专家项目 (G2021026016L) 资助
Supported by the High-end Foreign Expert Project of the Ministry of Science and Technology of China (G2021026016L)
本文责任编委 金连文
Recommended by Associate Editor JIN Lian-Wen
1. 河南大学 河南省大数据分析与处理重点实验室 开封 475001
1. Henan Key Lab of Big Data Analysis and Processing, Henan University, Kaifeng 475001

trast), BYOL[4] (Bootstrap your own latent), SwAV[5] (Swapping assignments between multiple views of the same image), SimSiam[6] (Simple siamese networks) 等算法. 这些技术通常基于类孪生神经网络的网络架构, 但训练过程中所用的图像对为同一幅图像分别增广后得到的图像对 (正样本对), 或不同图像分别增广后构成的图像对 (负样本对). 深度对比学习通过大量的正负样本对间的比对计算, 使得神经网络模型能够对数据自动提取到更好的特征表达. CPC[7] (Contrastive predictive poding) 是深度对比学习的奠基之作, 该算法通过最大化序列数据的预测结果和真实结果之间的相似度 (一致性程度), 优化特征提取网络, 并提出 InfoNCE 损失, 该损失如今已广泛地应用在对比学习研究中. Khosla 等[8] 提出监督对比学习损失 (Supervised contrastive learning loss, SCL loss), 将对比学习的思想扩充到了监督学习中, 旨在利用有标注的数据进一步提升模型的特征表达能力. Chen 等[9] 设计了半监督对比学习算法, 首先对所有数据进行对比学习预训练, 然后使用标注数据将预训练模型的知识通过蒸馏学习的方法迁移到新的模型中.

目前, 很少有系统总结对比学习最新进展的英文综述论文[10−12], 中文综述论文更是极度缺乏. 因此, 学术界迫切需要对深度对比学习的最新文献及进展进行全面系统的总结、归纳和评述, 并分析存在的问题, 预测未来发展趋势. 本文聚焦视觉领域的深度对比学习技术, 系统梳理深度对比学习 2018 年至今的技术演进, 总结该方向代表性的算法和技术. 如图 1 所示, 本文首先将深度对比学习的相关技术归纳为样本对构造方法层、图像增广层、网络架构层、损失函数层及应用层 5 大类型. 然后, 综合归纳现有技术的特点及异同之处, 并分析其性能表现, 指出尚未解决的共性问题及相关挑战, 最后勾勒该领域的未来发展方向与趋势.

本文的主要贡献概括如下:

1) 基于一种新的归类方法, 将现有的深度对比学习工作进行了系统总结;

2) 比较分析了不同对比学习方法的区别和联系, 及其在基准数据集上的性能表现;

3) 讨论了当前对比学习研究存在的挑战, 展望了未来的研究方向.

本文的剩余章节结构安排如下: 第 1 节介绍对比学习的背景知识. 第 2 节引入本文提出的归类方法, 并对每种类型的方法进行详细总结. 第 3 节对现有的对比学习技术进行整体分析, 并比较性能表现. 第 4 节探讨对比学习当前存在的挑战, 及未来发展方向. 最后是全文总结.

1 背景介绍

1.1 对比学习思想

对比学习的思想有两个起源, 一个是同类数据对比的思想[13], 另一个是自监督学习中的实例判别任务[14]. 文献 [13] 最早提出了使用两幅图像进行对比学习的思想, 主要使用孪生神经网络[15] 进行训练, 旨在拉近同类图像的特征之间的距离、推远不同类图像之间的距离, 以获得更好的特征提取模型. 而自监督学习中的实例判别任务, 将同一批次中的每个样本视作一个独立的类, 故类别的数量与该批次的样本数量相同. 通过该设计, 将无监督学习任务转化为分类任务 (实例判别任务, 寻找图像集中与输入图像特征相似度最高的图像). SimCLR[2] 和 MoCo[3] 是最早结合上述两个思想的方法, 它们通过同一幅图像分别增广后的图像对之间的特征比对

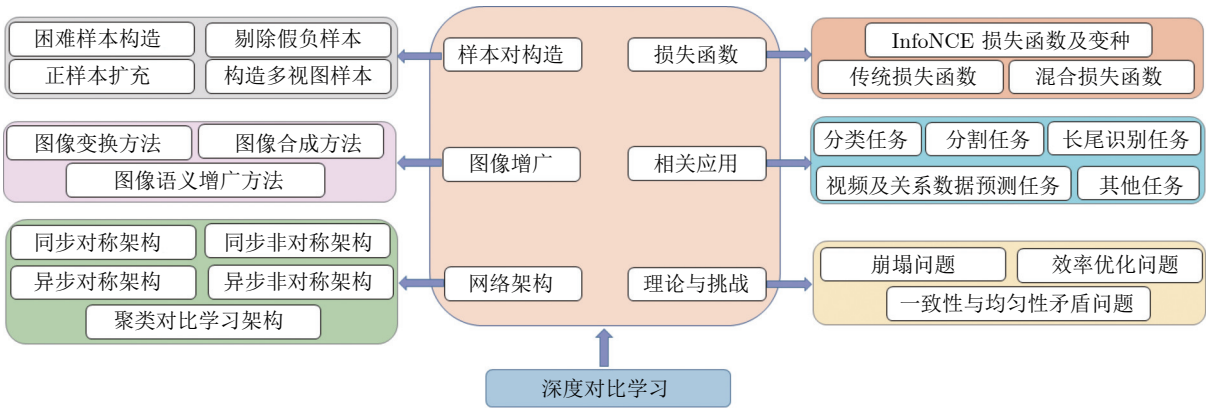


图 1 对比学习方法归类

Fig. 1 Taxonomy of contrastive learning methods

计算, 增强神经网络模型的特征提取能力, 再应用于下游任务中.

1.2　对比学习定义

对比学习的研究目前仍处于不断发展的阶段, 尚无明确的定义. 为了更清晰直观的阐述对比学习问题, 本文尝试给出一个对比学习的公式化定义及常用网络架构, 如式 (1) 及图 2 所示.

$$\max_{f_1, f_2} \sum_{i=1}^{N} \{s(z_1(x_{i1}), z_2(x_{i2})) - \sum_{j=1, j\neq i}^{N} [s(z_1(x_{i1}), z_1(x_{j1})) + s(z_1(x_{i1}), z_2(x_{j2}))]\} \tag{1}$$

在式 (1) 中, x_i 表示数据集 $X = \{x_1, x_2, \cdots, x_n\}$ 中的一个样本, $z_1(x_{i1})$ 是图像 x_i 先经过图像增强方法 T_1, 再经过图 2 中的分支 1 (上部分支) 得到的特征, $z_2(x_{i2})$ 是同一幅图像 x_i 先经过图像增强方法 T_2, 再经过分支 2 (下部分支) 得到的特征 (正样本), $z_1(x_{j1})$, $z_2(x_{j2})$ 是另一幅图像 x_j 经过增广得到的特征 (负样本). s 为两个特征向量之间的相似度度量方法, 默认使用余弦相似度. 该公式的目标是最大化相同图像增广后的两个样本之间的相似度, 同时最小化不同图像经过不同增广后的特征之间的相似度及不同图像经过相同增广后的特征之间的相似度. 图 2 是对比学习常用的网络架构, 采用了本文所归纳的同步对称网络架构, 后续章节将进行详述.

1.3　对比学习与度量学习的关系

对比学习与度量学习[16] 之间有较高的关联性. 其相似之处在于学习目标类似, 二者都在优化特征空间, 使得数据在特征空间中类内距离减小、类间距离增大. 度量学习通常需要基于标注数据, 而对比学习则不要求样本必须有标注, 其目标是使得同一幅图像的不同增广之后的图像对在特征空间中靠近.

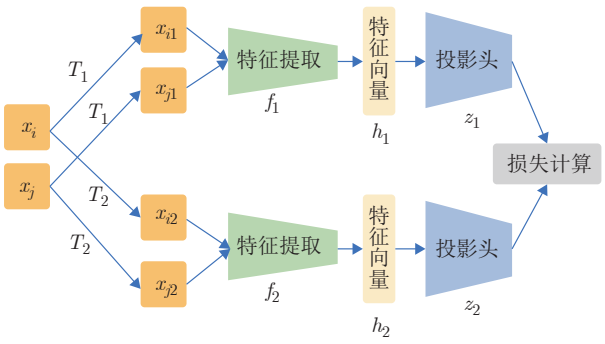


图 2　常用的对比学习网络架构

Fig. 2　Commonly used contrastive learning network architecture

两者的不同之处在于:

1) 训练数据构建方式不同

对一个数据集, 度量学习通过标签信息筛选样本对, 分别得到正样本对和负样本对. 而对比学习的正样本对是通过同一幅图像的两个随机增广得到. 图像增广操作是对比学习的关键技术之一, 多样、有效的图像增广方法能够提高对比学习的训练效果[2].

2) 网络架构不同

对比学习通常需要在特征提取网络之后增加小型的投影头网络, 将特征空间转换到投影空间中进行损失计算. 训练结束后, 将该小型投影头去掉, 仅保留特征提取网络, 用于下游任务. 而在度量学习任务中, 通常只有特征提取网络, 通过设计度量损失函数优化特征提取模型.

3) 适用情况不同

度量学习需要使用有标注的数据, 适用于监督学习; 而对比学习不要求数据有标注, 适用于无标注的数据及有标注或部分标注的数据, 支持监督、无监督和半监督学习.

1.4　对比学习与自监督学习的关系

初始的对比学习方法[2−3] 是自监督学习的一种, 目标是通过无标注数据的自我监督学习, 获得良好的特征表达. 随着技术的发展, 对比学习研究已经拓展到监督学习和半监督学习中 (如第 1.3 节所述).

对比学习与无监督学习的关系. 无监督学习大体可分为生成式学习和对比式学习. 对比学习是实现无监督学习的一种重要途径, 能够学习到更好的特征表示.

对比学习与监督学习的关系. 随着对比学习技术的发展, 在构造对比学习所需的样本对时, 若数据有标注或部分标注, 则属于同一类的图像都可以用于构造正样本, 而不再局限于同一幅图像, 这样可以增加正样本对的多样性, 有利于提取更好的特征表达. 另外, 对比学习也可以作为监督学习的前置, 进行模型预训练, 以进一步提高监督学习的模型性能.

1.5　常用数据集介绍

本节所介绍的数据集总结如表 1 所示:

ImageNet-1K[17] 是对比学习方法最常用的数据集, 该数据集是 ImageNet 数据集的一个子集, 包含 1 000 个类别, 训练集约 128 万幅彩色图像. 该

表 1 对比学习常用数据集总结
Table 1 Summary of common datasets

数据集	任务	类别个数	图像总量
ImageNet-1K[17]	分类	1 000	128 万
Cifar10[18]	分类	10	6 万
Cifar100[18]	分类	100	6 万
Food101[19]	分类	101	10 万
Birdsnap[20]	分类	500	5 万
Sun397[21]	分类	397	11 万
Cars[22]	分类	196	1.6 万
Aircraft[23]	分类	100	1 万
DTD[24]	分类	47	5 640
Pets[25]	分类	37	3 680
Caltech-101[26]	分类	101	9 144
Flowers[27]	分类	102	7 169
VOC[28]	检测&分割	20	1 万
COCO[29]	检测&分割	80	33 万

数据集因其包含的图像种类多、数量多以及样本质量高的特点常常被认为是最有挑战性的数据集之一. 该数据集各类样本数量分布相对均衡.

Cifar10 和 Cifar100[18] 是常见的应用于图像分类的数据集, Cifar10 包含 10 个类别的 6 万幅彩色图像, Cifar100 包含 100 个类别的 6 万幅彩色图像. Cifar10 和 Cifar100 的图像尺寸均为 32×32 像素.

常用于分类任务评估的数据集还有: Food101[19] 包含 101 个类别的食物, 共 10 万幅图像. Birdsnap[20] 包含 500 个类别的鸟的图像, 共 4.7 万幅训练集图像. Sun397[21] 包含 397 个类别的场景图像, 共有近 11 万幅图像. Cars[22] 包含 196 个类别的汽车图像, 共有 1.6 万幅图像. Aircraft[23], 包含 102 个类别的飞行器图像, 共有约 1 万幅图像. DTD[24], 包含 47 个类别的纹理图像, 共有 5 640 幅图像. Pets[25], 包含 37 个类别的宠物图像, 共有 3 680 幅训练集图像. Caltech-101[26], 包含 101 个类别的图像, 共有 9 144 幅图像. Flowers[27], 包含 102 个类别的花卉图像, 共有 7 169 幅图像.

VOC[28] 数据集常用来评估目标检测和分割任务, 总共包含 20 类图像, 共约 1 万幅图像, 包含 13 000 多个物体目标.

COCO[29] 数据集是一个可用于大规模目标检测、分割和关键点检测的数据集, 该数据集包含有 80 个目标类别、91 个物品类别, 约 33 万幅彩色图像组成.

1.6 评价方法

目前, 衡量对比学习模型的效果最常用两种评估方法是线性评估方法和微调 (Finetune) 评估方法. 在进行评估之前, 将预训练好的模型作为下游任务的主干网络, 然后, 根据下游任务和数据集的要求, 在主干网络后加入对应的任务头网络, 最后采用线性或微调方法进行评估.

线性评估方法. 采用该方法评估主干网络时, 需要冻结主干网络参数, 只训练下游任务头网络. 然后采用下游任务数据集中的测试集对模型的能力进行评估.

微调评估方法. 采用该方法评估主干网络时, 采用下游任务的数据训练由主干网络及任务头组成的整体网络, 然后同样采用下游任务数据集的测试集评估模型.

1.7 本文符号定义

本文中所使用的数学符号定义说明如表 2 所示:

表 2 本文所用符号总结
Table 2 Summary of the symbols used in this paper

符号	说明
X	数据集合, 小写为其中的数据
Y	标签集合, 小写为其中的数据
T	图像增广方法
f	特征提取网络
g	投影头
s	相似度度量函数
h	特征向量, $h = f(x)$
z	投影向量, $z = g(h)$
c	聚类中心向量
τ	温度系数

2 对比学习研究现状

在本章内容中, 首先引入所提出的归类方法, 在此基础上归纳总结国内外对比学习研究成果. 值得注意的是, 对比学习亦深受我国学者的关注, 很多优秀的对比学习论文虽然发表在国外学术会议和刊物上, 但作者为国内学者.

2.1 归类方法

本文提出一种新的归类方法, 以对最新的对比学习工作进行系统归纳总结. 如图 1 所示, 按照对比学习的整体流程, 将现有方法划分为样本对构造层、图像增广层、神经网络架构层、损失函数层以及应用层等类型.

图 3 将每种类型的对比学习方法进行细分及可视化表示. 其中, 样本对构造层方法可以细分为困

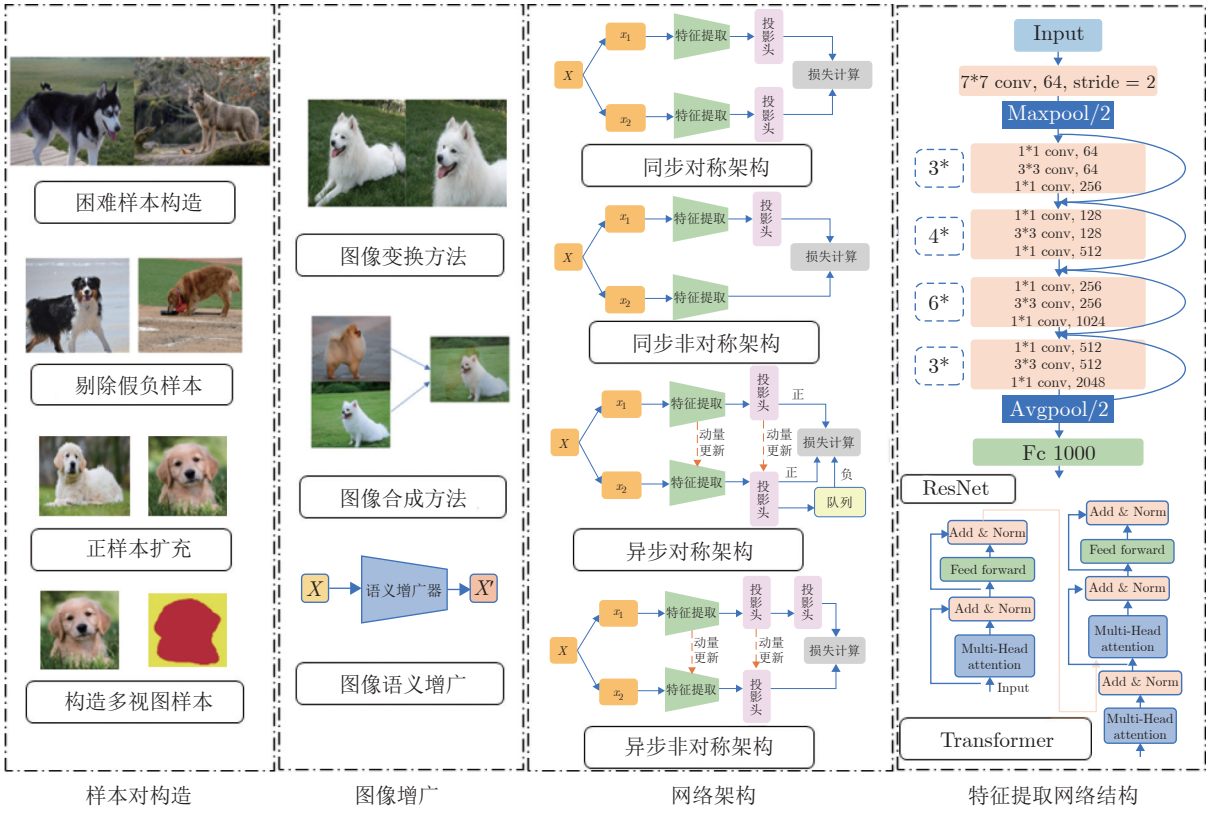


图 3　对比学习的整体流程及各模块的细分类方法

Fig. 3　Overall framework of the contrastive learning process and the sub-category of each module

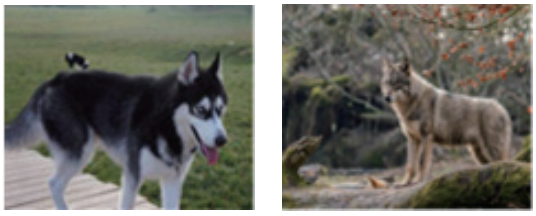
难样本构造、剔除假负样本、正样本扩充及构造多视角样本四种方法. 图像增广层可以细分为图像变换、图像合成及图像语义增广三种方法. 网络架构层方法可以分为同步对称、同步非对称、异步对称、异步非对称对比学习, 及基于聚类的网络架构. 特征提取网络主要使用 ResNet[30] (Residual neural network)、Transformer[31] 等主流神经网络结构, 损失函数层分为基于互信息的损失函数、传统损失函数和混合损失函数. 下面将分别介绍各类型的方法.

2.2　样本处理及样本对构造方法

在对比学习过程中, 样本对的选择指的是对数据集的采样过程. 1) 对无标注数据集, 通常采用随机采样的方法构建一个批次的数据, 因此一个批次的数据可能存在类别分布不均匀的情况, 导致假负样本的出现及困难负样本过少的问题; 2) 对有标注数据集, 通过标签信息采样训练数据, 能够有效提高对比学习效果. 下面将分别介绍各种细分的样本处理及样本对构造方法.

2.2.1　困难样本构造

如图 4 所示, 图 (a) 是狗, 图 (b) 是狼, 在选择样本进行模型训练的时候, 这两种动物在视觉上相似度很高, 同时在经过网络提取特征之后, 他们的特征之间的余弦相似度也比较高, 容易被误认为是相同类的样本, 而这两张图像并非同类, 这种情况称为困难负样本对, 同理, 若两张图片属于同一类, 但特征相似度不高则称为困难正样本对. 多个研究表明, 困难负样本和困难正样本对在对比学习中具有至关重要的作用.



(a) 狗　(b) 狼

(a) Dog　(b) Wolf

图 4　困难负样本对示例

Fig. 4　Example of hard negative pair

Zhu 等[32] 通过对 MoCo 算法训练过程的可视化分析, 发现增加困难样本在同批次中的比例能够提升网络在下游任务中的表现. 通过这一现象, 作者提出了在特征空间上将负样本对图像对应的两个特征向量插值, 正样本对图像对应的两个特征向量

外推, 构造新的、更加困难的样本对, 最终提高了对比学习的效果. 采用类似的思想, Kalantidis 等[33] 提出了一种合成困难负样本的方法, 该方法也作用于特征空间中, 首先通过余弦相似性度量方法将输入样本对应的所有负样本进行降序排序并取前 K 个样本, 然后使用与文献 [32] 相同的方法依次合成 K 个困难负样本. 但是该方法只适用于带队列存储库 (Memory bank) 的对比学习模型. Zhong 等[34] 提出了一种结合图像混合 (Mixup) 技术的困难样本构造方法, 该方法首先利用 Mixup 方法合成新的样本, 然后通过余弦相似性度量方法挑选困难的合成样本, 加入到原始数据中进行对比学习训练, 提高了对比学习的训练效果.

2.2.2 剔除假负样本

如图 5 所示, 在同一批次的训练图像中, 如果输入图像和同一个批次的其他某一张图像是同一类, 但是在计算损失的时候把它们误归为负样本对, 这种情况就称为假负样本对. 在对比学习中, 负样本的质量和数量是制约最后训练效果的关键, 如果在训练的批次中出现假负样本, 会降低最后的网络效果.

图 5 假负样本示例

Fig. 5 Example of false negative pair

为了解决该问题, Huynh 等[35] 提出一种启发式的方法来避免假负样本对训练的影响. 该方法将同一幅图像使用多种增广方法得到的所有图像作为该幅图像对应的支持集, 然后将同一个批次中该图像对应的每个负样本依次与其支持集中的每幅图像进行特征相似性计算, 并将前 K 个最相似的负样本视为假负样本, 认为这些负样本与输入图像具有相同的类标签. 得到假负样本后, 可以直接剔除假负样本, 或将假负样本移入到正样本集合. 通过大量实验, 作者证明了两种处理方式都能提高对比学习的效果. 针对相同的问题, Chuang 等[36] 提出修正 InfoNCE 损失函数中所有负样本相似度的指数幂和, 以降低可能采样到的假负样本对损失值带来的影响.

2.2.3 正样本扩充法

扩充正样本有助于提高对比学习的效果. 这里的扩充不包含图像增广方法, 而是指从可用的数据资源中寻找隐藏的、与输入样本类别相同的图像的方法. 在监督对比学习中, 由于同类样本已知, 一般无需扩充正样本. 而在无监督对比学习中, 若图像数据没有标注信息, 一般无法进行正样本扩充. 但在一些特殊场景下, 如行人重识别、遥感图像处理、视频分析, 可以基于某些假设, 对正样本进行扩充.

Kim 等[37] 将 Mixup 方法应用到了对比学习领域, 提出 MixCo 方法, 该方法将经过增广之后的两幅异类图像进行 Mixup 操作合成新的图像. 然后对所有合成的图像, 计算其 InfoNCE 损失, 由于每幅合成图像对应两个类别, 因此可以认为是扩充了每个样本对应的正样本个数. 最终损失由原始图像上的对比损失和上述合成图像上的对比损失构成.

在行人重识别领域, 王梦琳[38] 提出一种度量不同相机中行人关联性的方法, 将关联性强的样本视作同类样本, 以扩充正样本的数量.

在遥感图像处理中, 基于同一地点的遥感图像中的内容随时间变化较小的特点, Ayush 等[39] 使用同一地点的不同时刻的遥感图像扩充正样本.

在视频分析中, Qian 等[40] 基于邻近的视频帧图像可作为同类样本的假设, 扩充正样本, 用于对比学习. Kumar 等[41] 同样采用这一假设扩充正样本. Han 等[42] 发现对于两个不同的视频片段, 做的动作相同时, 在 RGB 视角下, 提取的特征相似度不高, 但在光流 (Optical flow) 视角下, 提取的特征相似度很高, 反之亦然. 基于以上发现, 作者提出将光流视角下相似度高的视频片段作为 RGB 视角下的正样本, 同样, 可采用相同方法扩充光流视角下的正样本.

在半监督学习领域, Wang 等[43] 提出通过部分标签训练的分类器, 在当前训练批次中利用余弦相似度指标寻找正样本的方法, 扩充正样本数量. Yang 等[44] 采用同样的思路设计类感知模块, 扩充正样本数量.

2.2.4 构造多视图样本

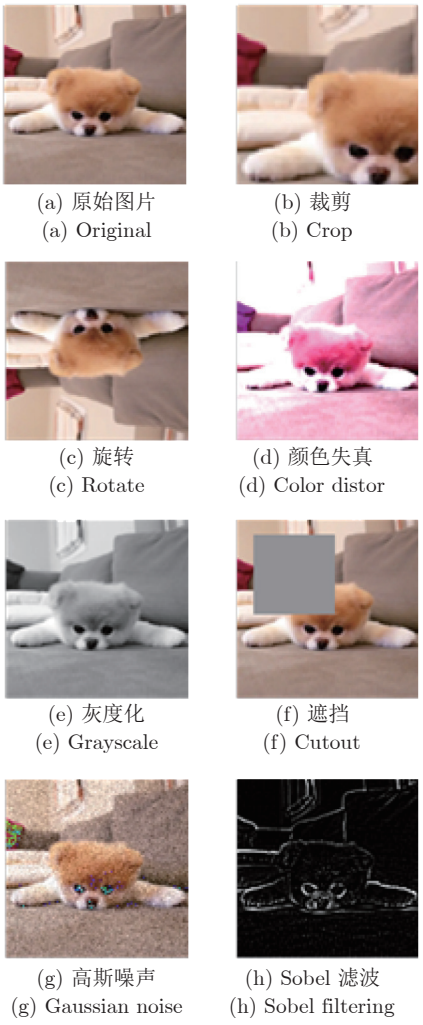
多数对比学习方法所使用的数据只有一个来源 (又称为同一个视图). 针对该问题, 研究人员探索使用多视图数据的方法提升对比学习的效果. Tian 等[45] 提出基于多视图特征的对比学习方法. 该方法将同一幅图像在多个不同视图下的表达分别进行特征提取, 然后进行对比学习, 有利于提升模型的效果. 在视频分析中, Rai 等[46] 对同一幅图像分别提取光流、语义分割、关键点等多视图特征, 然后进行对比学习, 提升了视频特征表达能力.

2.3 图像增广方法

2.3.1 图像变换方法

图像增广是对比学习的天然组成部分, 以产生

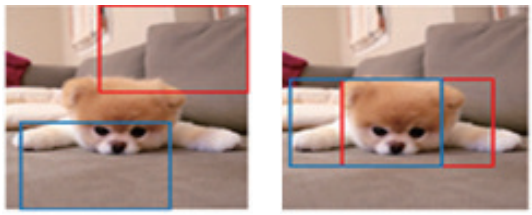
网络所需的输入样本对. 如图 6 所示, 传统的图像变换方法有随机裁剪、颜色失真、灰度化等方法. 需要说明的是, 单一的图像变换方法不利于有效的对比学习. SimCLR 综合对比了一系列图像变换的效果, 发现由随机裁剪和颜色失真组成的变换组合能够获得更好的对比学习效果.



(a) 原始图片 (a) Original
(b) 裁剪 (b) Crop
(c) 旋转 (c) Rotate
(d) 颜色失真 (d) Color distor
(e) 灰度化 (e) Grayscale
(f) 遮挡 (f) Cutout
(g) 高斯噪声 (g) Gaussian noise
(h) Sobel 滤波 (h) Sobel filtering

图 6 常用图像变换方法示例
Fig. 6 Example of common image augmentations

在众多的图像变换方法之中, 裁剪通常是必须的变换方法之一, 但可能存在经过裁剪后的正样本对信息重叠过多或者完全不重叠的情况. 如图 7(a) 所示, 如果裁剪到的正样本对是两个矩形框中的区域, 那么该图像对并不能促进对比学习, 这种情况称为错误正样本 (False positive, FP). 图 7(b) 给出了裁剪的两个图像区域重叠过多的情况, 若选择它们作为正样本对, 也无助于对比学习. Peng 等[47] 针对以上问题, 首先利用热图定位目标区域, 然后采用中心抑制抽样策略, 在目标区域内离物体中心越远的像素点被采样到的概率越高. 得到采样点后, 以采样点为中心, 并使用随机生成的宽高, 进行最终的图像裁剪操作. 在 SwAV 中, 作者提出了一种多裁剪策略 (Multi-crop), 该策略给编码器提供多个不同尺度的裁剪图像作为正样本, 并验证了该策略有助于提高对比学习算法的性能.



(a) (b)

图 7 裁剪操作对正样本对构造的影响示例
Fig. 7 The influence of constructing positive pairs by image crop

2.3.2 图像合成方法

图像合成不同于图像变换, 前者能够生成新的内容. 在视频分析中, Ding 等[48] 提出复制粘贴的图像合成方法, 将所选视频帧的前景区域粘贴到其他视频帧的背景图像中, 得到合成的视频帧, 通过这种简单的合成方法, 使得对比学习模型能够更加关注前景信息, 提取到更加良好的特征表达.

2.3.3 图像语义增广方法

图像语义增广是一种直接对图像中物体的语义进行修改的图像增广方法, 如将图像中的物体的颜色或角度进行改变. 现有的一些语义增广算法[49−51] 已证明其在不同问题和应用中的有效性. 相较于图像变换方法和图像合成方法, 图像语义增广是一种更强的图像增广方法. Tian 等[52] 认为最小化在不同增广下样本间的互信息有利于提高对比学习的效果, 为了最小化训练样本对之间的互信息, 作者引入了对抗训练策略寻找可以最小化训练样本对互信息的图像变换编码器 S_g, 通过与 InfoNCE 联合训练, 获得了更适用于下游任务的模型, 经过训练得到的图像变换编码器 S_g 可认为是一个语义增广器.

2.4 对比学习网络架构设计

根据对比学习网络架构的更新方式是同步或异步, 及该网络架构是对称或非对称, 本文将对比学习所涉及的网络架构划分为同步对称、同步非对称、异步对称和异步非对称 4 种类型. 同步更新指的是对比学习网络的两个分支 (分支 1 和分支 2) 同时进行梯度更新, 异步更新指的是两个分支网络的权值更新方法不同. 对称指的是分支 1 和分支 2 的网

络结构完全相同, 非对称则指相反的情况. 除此之外, 本节还将含有聚类算法的对比学习方法总结为聚类对比学习架构.

2.4.1 同步对称网络架构

典型的同步对称网络架构如图 8 所示, 分支 1 和分支 2 采用相同的网络结构, 并且同时采用梯度更新.

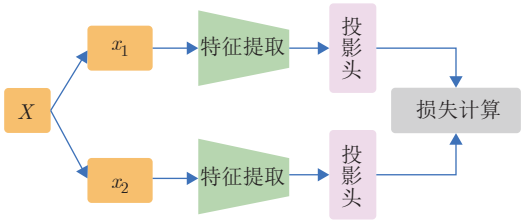


图 8 同步对称网络架构

Fig. 8 The architecture of synchronous symmetrical network

SimCLR[2] 是最早提出采用同步对称网络架构的对比学习工作. 该网络架构采用了结构相同的两个网络分支, 每一个分支都包含特征提取网络和投影头. 特征提取网络可以是任意的网络结构, 例如 ResNet 和 Transformer, 投影头一般使用多层感知机 (Multilayer perceptron, MLP). 如图 8 所示, 使用该网络架构进行对比学习的前向过程为: 首先, 输入一个样本, 经过两种图像增广方法产生成对的训练样本. 然后, 通过特征提取网络和投影头将样本输出到投影空间. 最后, 在投影空间进行损失计算并回传梯度, 更新特征提取网络和投影头的参数. 同步对称网络架构除了可以用于自监督对比学习之外, Khosla 等[8] 将该类网络架构应用到了监督对比学习之中. 值得注意的是, 采用该类网络架构的对比学习算法若想获得良好效果, 往往需要在训练时采用很大的批次进行训练, 比如 1 024 或 2 048.

2.4.2 同步非对称网络架构

典型的同步非对称网络结构如图 9 同步非对称网络所示, 分支 1 和分支 2 采用不同的网络结构, 但同时采用梯度更新. 同步非对称网络架构与同步对称网络架构的相同之处在于两个分支都进行了梯度更新, 不同之处在于前者的两个分支的网络结构不同, 而后者相同.

Van 等[7] 提出一种基于该种网络架构的对比预测编码方法 CPC. 在该方法中, 分支 1 的输入为某一个时间点的数据, 分支 2 的输入为未来某一个时间点的数据, 训练的目标是利用分支 1 的输出预测分支 2 的输出, 如图 9 所示, 然后计算对应的预测损失. 在 CPC 算法之后, 衍生出很多改进的算法[41, 53], 这些方法均采用典型的同步非对称网络架构, 其主要改进之处在于采样方法或损失函数.

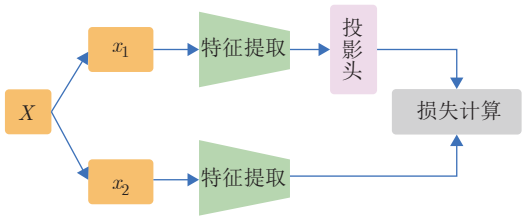


图 9 同步非对称网络架构

Fig. 9 The architecture of synchronous unsymmetrical network

除了如图 9 所示的同步非对称架构之外, 还存在一些其它形式的同步非对称的网络架构, 主要有如下两种形式: 1) 分支 1 与分支 2 均含有投影头, 但投影头结构不相同[54−55]. 2) 分支 1 和分支 2 的特征提取网络数量不相等[56−57].

Nguyen 等[54] 将对比学习方法应用于网络结构搜索任务 (Neural architecture search, NAS), 提出 CSNAS 方法, 该方法构造了一个类似于上述第 1 种形式的同步非对称网络架构. Misra 等[55] 在分支 2 中使用拼图代理任务进行特征学习, 因此在分支 2 的投影头之前还包含拼接操作以及多层感知机的映射, 构成了符合上述第 1 种形式的同步非对称网络架构.

Bae 等[56] 提出自对比学习 (Self contrastive learning, SelfCon) 方法, 该方法构造了一个类似于上述第 2 种形式的同步非对称网络架构, 在该方法中, 将一个特征提取网络和一个投影头组合在一起, 称为一个主干网络块, 分支 1 采用多个主干网络块进行特征映射, 分支 2 从分支 1 中的某个节点引出, 通过一次主干网络块映射得到另一个视图下的特征, 最后进行损失计算. 类似地, Chaitanya 等[57] 提出可用于图像分割的对比学习方法, 在该方法中, 分支 1 和分支 2 共享编码器, 但分支 1 直接在编码器后加入投影头, 分支 2 在编码器后加入一个解码器, 然后再加入投影头.

2.4.3 异步对称网络架构

在异步对称的网络架构中, 分支 1 和分支 2 采用相同的网络结构, 但是两个分支的神经网络权值更新方式不同, 因此称作异步更新, 如图 10 所示.

MoCo 是经典的采用异步对称网络架构的对比学习方法. 其中, 分支 1 中的特征提取网络和投影头采用 SGD 等梯度更新方法更新权值, 而分支 2 的特征提取网络和投影头采用动量更新的方式, 动量更新的公式如式 (2) 所示:

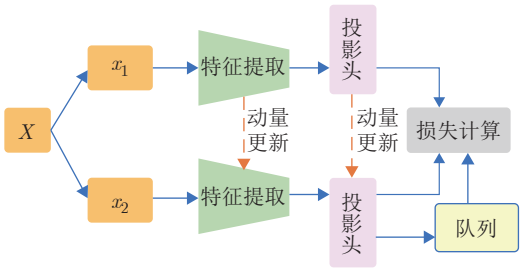


图 10　异步对称网络架构

Fig. 10　The architecture of asynchronous symmetrical network

$$\theta_k^{t+1} \leftarrow m\theta_k^t + (1-m)\theta_q^{t+1} \tag{2}$$

其中, θ_k^t 是分支 2 中动量特征提取网络在时刻 t 的参数值, θ_q^{t+1} 是分支 1 中特征提取网络经过 t 时刻梯度反传训练过后的参数值, m 是 $0-1$ 之间的动量系数. 这种动量更新机制能够有效防止极端样本对参数更新影响过大的问题.

在对比学习中, 负样本的数量对最终的模型效果起决定性的作用. 在原始的对比学习方法中, 负样本的来源是同批次的训练数据, 为了增加负样本的数量, 必须采用大批次的数据来支撑训练, 但这样会造成存储资源过大, 计算消耗巨大的问题. MoCo 设计了队列 (Queue) 结构, 存储一定数量的之前批次用到的样本的特征, 当队列容量满时, 最老样本的特征出队, 最新样本的特征入队. 在 MoCo 方法中, 对比学习所需的负样本将从该队列中抽样产生, 避免了原始对比学习方法为获得较多负样本而构建大的批容量带来的问题. 简言之, 该队列结构减小了对存储资源的要求, 又取消了训练必须采用大批次数据的约束, 提升了计算速度. 后续的基于异步对称网络架构的方法都将该队列结构作为其默认的组成部分.

近年来, 随着 Transformer 技术的发展, 研究者将其引入到计算机视觉领域, 例如 ViT[58]. 在对比学习的研究中. Caron 等[59] 提出了一种名为 DINO (Self-distillation with no labels) 的异步对称对比学习方法, 该方法采用 Transformer 作为骨干网络, 并将自监督对比学习转换为蒸馏学习的任务, 将分支 1 视作学生网络, 将分支 2 视作教师网络. 学生网络采用梯度回传更新参数, 教师网络采用动量更新方式更新参数.

2.4.4　异步非对称网络架构

异步非对称网络架构指的是分支 1 和分支 2 采用不同的网络结构, 同时参数更新方式也不同. 异步非对称网络架构与异步对称网络架构的相同之处在于两个分支的更新方式相同, 不同之处在于前者的两个分支的网络结构不同, 而后者相同.

异步非对称网络架构存在两种形式, 一种是采用 MoCo 作为主干网络, 但是投影头数量不一致, 以 BYOL[4] 方法为典型代表. 另一种是网络架构不对称, 同时采用梯度交叉更新的方法, 以 SimSiam[6] 方法为典型代表.

BYOL 的网络结构如图 11 所示. 其主干网络结构是 MoCo, 分支 2 的网络参数采用动量更新机制. BYOL 是首个只采用正样本进行对比学习的工作, 但由于训练集中不存在负样本, 如果上下网络分支结构完全相同, 训练就有可能出现 “捷径解”的问题 (亦可称为 “网络崩塌”问题), “捷径解”指的是对不同的输入, 输出的特征向量完全相同. 为了解决这一问题, BYOL 在分支 1 中额外增加了一个预测头 (MLP), 构成了非对称网络架构, 并将对比学习的实例判别代理任务替换为比对任务, 即分支 1 最终得到的特征与分支 2 最终得到的特征要尽可能的相似. BYOL 的性能提升归因于以下两点设计: 1) 良好的模型初始化. 2) 在投影头和预测头中加入正则化技术. 如果没有这两种设计, BYOL 仍可能出现 “捷径解”的问题[60].

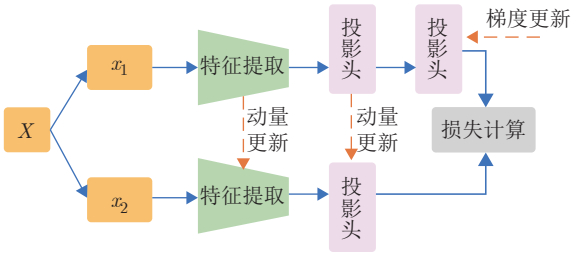


图 11　BYOL 网络架构

Fig. 11　The architecture of BYOL

Chen 等[61] 提出了一种基于视觉 Transformer 的异步非对称网络架构 MoCov3. MoCov3 采了 BYOL 型异步非对称网络架构. 在分支 1 中额外加入了一个预测头, 分支 2 的编码器采用动量更新, 获得了更好的对比学习效果.

另一种代表性的异步非对称架构是 SimSiam. 如图 12 所示, SimSiam 方法也不需要利用负样本训练网络. 为了避免网络出现 “捷径解”, 该方法的两个分支的网络结构采用交叉梯度更新的方式. 具体而言, 训练分支 1 的时候, 投影头放在分支 1 的编码器之后, 计算其损失, 而分支 2 的梯度停止回传. 在训练分支 2 的时候, 将分支 1 的投影头接入到分支 2 的编码器之后, 并计算损失, 而分支 1 的梯度停止回传. 这就是交叉梯度更新.

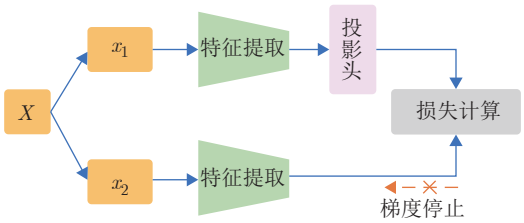


图 12 SimSiam 网络架构

Fig. 12 The architecture of SimSiam

2.4.5 聚类对比学习结构

在上述四种网络架构的基础上, 还可以结合聚类技术进行进一步的优化.

基于实例的对比学习算法在计算损失差别过大, 不利于模型学习到良好的语义特征. 而聚类算法能够在无监督情况下自动学习数据的语义信息. 因此, 一些方法将对比学习与聚类方法结合在起来, 嵌入到上述四种网络架构中.

Caron 等[62] 首次将 K-Means 聚类算法与无监督深度学习结合, 提出深度聚类算法 (Deep cluster), 该算法通过 K-Means 聚类产生伪标签, 然后利用伪标签对模型进行自监督训练. 在该方法的基础上, 作者又提出了一种基于聚类的对比学习算法 SwAV, 该算法采用同步对称网络架构训练特征提取网络. 在分支 1 和分支 2 之间引入聚类中心信息, 并计算相关损失, 帮助网络进行训练. 相关的细节在第 2.5.2 节中给出.

Li 等[63] 提出原型对比学习算法 (Prototypical contrastive learning, PCL), 该方法采用异步对称网络架构, 分支 1 采用梯度更新, 分支 2 采用动量更新. 在 PCL 的前向传播过程中, 当分支 2 对输入样本提取特征之后, 采用 K-Means 算法依据样本特征进行聚类. 然后, 将分支 1 得到的每个样本特征与分支 2 的所有聚类中心计算聚类对比学习损失, 将分支 1 的样本与其所属的聚类中心的向量视作正样本对, 其余为负样本对. 最终的损失函数由聚类对比学习损失和原有 (实例) 对比学习损失构成. 聚类对比学习损失计算的目标是最大化样本实例与其相对应的聚类中心特征之间的相似度, 因此这种设计也能在一定程度上缓解实例对比学习中的假负样本问题. PCL 仅在分支 2 上进行聚类计算, 后续工作尝试在两个分支上都进行聚类, 如 Wang 等[64] 提出的方法, 将分支 1 和分支 2 都进行聚类计算, 然后进行交叉聚类对比学习损失计算.

在现实图像数据集中, 往往存在多层级的语义结构, 例如在"狗"这个分类的数据中, 还存在 “哈士奇”、“金毛”、“雪纳瑞”等不同品种的狗. 现有的很多对比学习方法没有考虑到数据集的这个特点. 针对这一问题, Guo 等[65] 提出了一种分层聚类的对比学习方法 (Hierarchical contrastive selective coding, HCSC), HCSC 的网络架构如图 13 所示, 和 PCL 的神经网络架构相同, 都属于异步对称网络架构. HCSC 通过多层聚类来实现对数据集中存在的分层语义现象的模拟. 具体而言, 在分支 2 的特征提取操作之后, 首先, 通过 K-means 算法计算第一层聚类中心, 随后针对第一层聚类中心再使用 K-means 算法寻找下一层的聚类中心, 循环 N 次, 得到 N 个层次的聚类向量, 完成对 N 层语义的模拟. 同时 HCSC 还提出一种新的负样本采样方法: 在每一层语义结构中寻找与当前输入样本特征的原型相似度较低的样本作为负样本. 这种采样负样本的方法能够有效缓解采样到假负样本的问题.

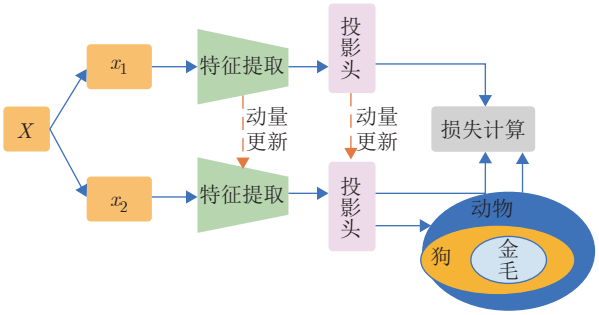


图 13 HCSC 网络架构

Fig. 13 The architecture of HCSC

此外, Li 等[66] 提出了一种聚类中心可自动学习的聚类对比学习算法. 该网络设计了一个共享特征网络的双分支同步对称网络架构, 两个分支分别构建投影头, 即用于实例对比学习的投影头 I, 和用于聚类对比学习的投影头 C. 投影头 I 采用线性激活, 投影头 C 采用 softmax 激活. 将投影头 C 输出的矩阵中的列向量作为聚类中心, 通过 InfoNCE 损失和交叉熵损失联合训练模型, 在训练过程中由于投影头 C 参数在更新, 聚类中心因此也获得自动更新的能力.

Cui 等[67] 提出了一种新的异步非对称聚类网络架构 PaCo, 该方法将聚类思想和 MoCo 结合在一起, 通过构建一个可随着模型一同学习的参数化类别中心 c, 使数据少的类别在训练中的重要性提升, 从而在长尾学习中实现对损失的再平衡.

2.5 损失函数设计

为了实现对比学习的优化目标, 研究者使用了多种针对性的损失函数, 可以分为基于互信息的损失函数 (InfoNCE 类)、传统损失函数和混合损失函数.

值得注意的是, 对比损失[13] 与对比学习损失很容易在字面上产生混淆, 实际上, 对比损失是一种度量学习损失, 只能用于监督学习, 且样本对数据来源与对比学习不同, 因此, 对比损失不一定是对比学习的损失函数.

2.5.1 InfoNCE 损失函数及变种

对于深度学习模型来说, 理想的训练目标是获得一个从样本到特征的映射模型 $p(z|x;\theta)$, 其中 x 是样本, z 是模型输出的特征, θ 为需要学习的参数. 很多自监督学习方法通过交叉熵或均方误差训练网络达到上述目标. 然而, 基于交叉熵或均方误差这样的单峰函数的损失往往效果不好[7]. 因此 Van 等[7] 提出通过最大化互信息的方法训练特征提取网络. 互信息 $I(X;Y)$ 是一个概率论中的概念, 可以衡量两个随机变量 X 和 Y 之间的相关性. 互信息的定义如式 (3) 所示:

$$I(X;Y)=\mathrm{E}_{x,\,y\sim p(x,\,y)}\left[\ln\frac{p(x|y)}{p(x)}\right] \tag{3}$$

由于直接最大化互信息是十分困难的, 因此作者提出了通过 InfoNCE 损失来间接优化互信息的方法. InfoNCE 损失的思想是将最大化互信息转换为真实分布占合成分布的概率密度比的预测问题[68]. 原始 InfoNCE 公式 (4) 所示:

$$L_{\text{infoNCE}}=-\ln\frac{\exp(s(q,\,h^{+}))}{\sum_{x_i\in X}\exp(s(q,\,h_i))} \tag{4}$$

其中, $X=\{x_1,\cdots,x_N\}$ 为采样的 N 个样本, f 为特征提取网络, $q=f(x_q)$ 为查询样本的特征向量, $h^{+}=f(x_i)$ 为 x_q 为对应的正样本的特征向量. 在训练中, 每个查询样本 x_q 只对应 1 个正样本 x_i, 其余都视为从噪声分布里面采样的负样本. s 为相似度度量函数, 这里采用余弦相似度. Van 等[7] 和 Poole 等[69] 证明了优化 InfoNCE 损失等价于优化变量之间的互信息的下界, 并且 $I(x_q,\,x_i)\geq\ln N-L_{\text{infoNCE}}$, 可知如果想获得一个较高的互信息值, 需要大批次的样本参与训练.

SimCLR 首先将 InfoNCE 损失引入到对比学习中来, 由于训练所用的样本对来自于同一幅图像的两次不同增广, 因此 InfoNCE 的优化目标变为最大化同一个样本的两个不同增广图像之间的互信息, SimCLR 中采用的 InfoNCE 如式 (5) 所示:

$$L=-\ln\frac{\exp(s(z_i,\,z_i^{+})/\tau)}{\sum\limits_{j=1}^{2N}1_{[j\neq i]}\exp(s(z_i,\,z_j)/\tau)} \tag{5}$$

其中, $X=\{x_1,\cdots,x_N\}$ 为采样的 N 个样本, f 为特征提取网络, g 为投影头, T 为图像增广函数, $z_i=g(f(T_1(x_i)))$ 为样本经过第 1 种图像增广之后产生的投影向量, $z_i^{+}=g(f(T_2(x_i)))$ 为样本经过第 2 种图像增广之后产生的投影向量, $1_{[j\neq i]}$ 为一个指示器函数, 当 $j\neq i$ 时为 1, 否则为 0. s 为相似度度量函数, 此处使用余弦相似度函数, 即 $s(u,\,v)=u^{\mathrm{T}}v/(||u||||v||)$. 总结而言, 在 SimCLR 的 InfoNCE 公式中, 分子表示同一个样本的两种不同变换得到的正样本对的特征相似度, 分母是不同图像组成的负样本集合中的每个负样本对的相似度的和.

在 InfoNCE 引入到对比学习之后, 许多方法[3, 9, 32−33, 40, 48, 52−54, 57, 70] 对其直接调用并扩展到更多的学习任务中. 表 3 列出了一些对 InfoNCE 改动较大的方法, 包括 ProtoNCE[63], DCL[71], DirectNCE[72], SCL[8], FNCL[35]. 其中, 第一行为原始的 InfoNCE 损失函数, 其他部分为对其改进的损失函数, 包括 ProtoNCE, DCL, DirectNCE, SCL, FNCL 等损失函数.

1) ProtoNCE 损失函数. 与原始 InfoNCE 损失函数相比, ProtoNCE 的分子部分将正样本实例之间的相似度修改为正样本与其所在的聚类中心之间的相似度, 分母部分改为计算样本与其它聚类中心之间的相似度. 具体而言, 当前实例特征与其对应的聚类中心互为正样本对, 与其它聚类中心互为负样本对. 在实际操作中, 通过设置不同的聚类中心个数对一个批次的数据进行 M 次聚类, 计算 M 次 ProtoNCE 损失, 然后求其损失均值, 最后得到的效果会更好.

ProtoNCE 方法是在同一个语义层级上进行聚类, 但无法获得分层的语义结构. Guo 等[65] 针对该问题提出分层聚类的方法, 该方法对每个聚类迭代地进行下一级细分聚类, 使特征网络能够学习到分层的语义结构, 获得更好的效果.

2) DCL 损失函数. Yeh 等[71] 通过对 InfoNCE 的反传梯度进行分析, 发现损失的梯度中存在一个负正耦合系数, 该系数体现了采用 InfoNCE训练网络时大批次样本的重要性, 同时揭示当训练样本中存在简单正样本对和简单负样本时, 会明显降低对比学习训练的效率, 因此作者提出了解耦对比学习损失 (Decoupled contrastive loss, DCL), 解决上述问题. 与 InfoNCE 损失函数相比, DCL 去除 InfoNCE 损失函数中的负正耦合项, 并将损失展开成两部分, 具体如表 3 中的 DCL 损失函数公式所示, 该式展示的是单个样本的 DCL 损失计算公式, 原 InfoNCE 包含的负正耦合系数在该式中已去除. 该式的第一项表示正样本对之间的相似度, 第二项是当前样本与同批次其他样本组成的负样本对之间的相似度之和. 从该式可以看出, DCL 损失的计算相较于 In-

表 3 InfoNCE 损失函数及其变种

Table 3 InfoNCE loss and some varieties based on InfoNCE

损失名	文献	年份	会议/刊物	公式	主要改进
InfoNCE	[7]	2019	arXiv	$-\ln \frac{\exp(s(q,\ h^+))}{\sum_{x_i\in X}\exp(s(q,\ h_i))}$	初始的 InfoNCE 损失. 文献 [3, 9, 30–31, 38] 等均采用 InfoNCE 损失函数.
ProtoNCE	[63]	2021	ICLR	$-\ln \frac{\exp(s(z_i,\ c_i)/\tau)}{\sum_{j=0}^{r}\exp(s(z_i,\ c_j)/\tau)}$	从两个样本增广间的对比变为样本增广与聚类中心的对比. 注: c_i , c_j 为聚类中心. 总损失包括实例间的对比和实例–原型对比损失, 此处只列出实例-原型对比损失.
DCL	[71]	2021	ECCV	$-\frac{s(z_i,\ z_i^+)}{\tau}+\ln\sum_{j=1}^{2N}1_{[j\neq i]}\exp(s(z_i,\ z_j)/\tau)$	去除负正耦合系数后通过化简得到该损失.
DirectNCE	[72]	2022	ICLR	$-\ln \frac{\exp s((\widehat{h}_i',\ \widehat{h}_i'^+)/\tau)}{\sum_j \exp s((\widehat{h}_i',\ \widehat{h}_i'^+)/\tau)}$	$\widehat{h}_i'=\widehat{h}_i'[0:d]$, 即取特征向量的前 d 个维度.
FNCL	[35]	2022	WACV	$-\ln \frac{\exp(s(z_i,\ z_i^+)/\tau)}{\sum_{j=1}^{2N}1_{[j\neq i,\ j\neq F_i]}\exp(s(z_i,\ z_j)/\tau)}$	F_i 为第 i 个样本的假负样本集.
SCL	[8]	2020	NIPS	$-\sum_{i\in I}\frac{1}{p(i)}\sum_{p\in p(i)}\ln\frac{\exp(s(z_i,\ z_p)/\tau)}{\sum_{a\in A(i)}\exp(s(z_i,\ z_a)/\tau)}$	将标签引入对比学习, $P(i)$ 是与第 i 个样本相同类的数据集合.

foNCE 损失更加简单、高效.

3) DirectNCE 损失函数. Jing 等[72] 从解决对比学习中的网络崩塌问题入手, 经过一系列分析, 提出去掉投影头, 然后将特征提取网络输出向量的前 d 个维度单独取出, 计算 InfoNCE 损失. 由于论文中并没有给改动后的损失起名, 因此为了便于对比, 本论文中将其命名为 DirectNCE. DirectNCE 与 InfoNCE 的区别在于所使用的样本特征维度大小, 在 DirectNCE 中, 样本只取前 d 个维度计算损失.

4) FNCL 损失函数. 针对在训练对比学习模型时可能存在的假负样本问题, Huynh 等[35] 首先提出了一种假负样本的检测策略, 然后对 InfoNCE 进行改进, 在损失函数中剔除了假负样本的干扰. 本文将该方法暂定名为 FNCL. 与 InfoNCE 相比, FNCL 方法首先确定当前正在处理的样本对应的假负样本, 然后在分母中计算负样本对之间的相似度时, 去除假负样本的部分.

5) SCL 损失函数. SCL[8] 损失函数面向有监督学习, 对 InfoNCE 损失函数进行改进, 旨在解决深度有监督学习中采用交叉熵损失时神经网络对噪声标签敏感[73] 的问题.

令 $P(i)$ 表示当前批中正在处理的样本对应的同类样本集合 (因为数据为有标注数据), $|P(i)|$ 表示该集合中样本的数量, 与 InfoNCE 损失函数相比, SCL 损失函数的分子计算当前样本与其对应的 $P(i)$ 集合中每个正样本之间的相似度的和, 而分母部分并无变化. 该损失在特定情况下可以等价于三元组损失[74] 和 N-Pair 损失[75]. 在 SCL 的基础上, 研究者还研究了利用对比学习提升长尾学习效果的方法, 此部分工作将在第 2.6.4 节进行详细介绍.

2.5.2 传统损失函数在对比学习中的应用

如何衡量特征空间中不同特征点之间的距离是对比学习中一个很重要的问题. 欧氏距离是衡量特征点之间距离的一个最直观的方法. 通过最小化均方误差损失 (Mean square error, MSE), 可以直接减小同类特征点之间的欧式距离, 实现让同类特征靠近的目的. BYOL 先对两个分支的特征进行 L_2 正则化, 然后采用 MSE 损失计算两个特征之间的距离, 对网络进行优化.

除了欧氏距离外, 余弦相似度也能够衡量特征之间的相似性, 因此直接最大化相同样本的不同增广的特征之间的余弦相似度也能达到对比学习的目标. SimSiam[6] 对其中一个分支的投影向量 z 和另一个分支的特征向量 h 计算余弦相似度, 直接将负的余弦相似度作为损失函数进行训练, 由于网络结构是一个非对称结构, 为了平衡训练, 作者将投影头依次放在两个分支后面进行损失计算, 从而提出对称损失, 其计算公式如式 (6) 所示:

$$L=-\frac{1}{2}\left(\frac{h_1 z_2}{\|h_1\|_2\|z_2\|_2}+\frac{h_2 z_1}{\|h_2\|_2\|z_1\|_2}\right) \tag{6}$$

欧式距离和余弦相似度都是非监督的特征相似度度量方法. 除此之外, 还有一些有监督的相似度计算方法.

在 SwAV 方法中, 首先初始化聚类中心, 然后将聚类中心矩阵分别与分支 1 和分支 2 的特征矩阵相乘, 计算交换预测编码矩阵 Q_1 和 Q_2, 最后, 采用交叉熵损失训练模型, 训练分支 1 时将 Q_2 作为标

签, 训练分支 2 时将 Q_1 作为标签. 聚类中心利用回传梯度进行更新.

此外, Shah 等[76] 将支持向量机与对比学习结合在了一起, 采用改进后的合页损失 (Hinge loss) 优化对比学习网络.

2.5.3 混合损失函数

在某些情况下, 只采用 InfoNCE 损失不能获得良好的效果, 而将多种损失函数结合, 有助于提升对比学习的效果.

有监督混合损失. Wang 等[77] 提出了一种将 SCL 与交叉熵损失结合起来的损失. 该方法采用一个平滑因子, 在训练早期, SCL 占据损失的主导地位, 随着学习过程的进行, 交叉熵损失会逐渐占据主导地位. Li 等[78] 将 ProtoNCE 的思想带入到 SCL 中, 即在 SCL的分子中计算每个样本与其同类样本所形成聚类中心之间的相似度. 在该算法中, 聚类中心直接由同类样本的特征求平均值得到. 最后, 作者将上述损失与交叉熵分类损失联合训练, 获得了较好的遥感图像分类效果.

半监督混合损失. Li 等[79] 设计了一个基于伪标签图结构对比学习方法 CoMatch. CoMatch 方法采用三部分损失训练网络. 对比学习网络中的一个分支, 对于有标签的数据, 采用交叉熵计算损失, 得到分类模型. 在对比学习网络的另外一个分支, 首先利用分类模型对无标签数据进行预测, 产生软标签, 该分支对每个无标注的数据, 进行特征提取, 并利用预测头产生预测结果, 当样本对应的软标签的置信度较高时, 采用交叉熵损失优化网络, 同时, 每个无标签的样本还将采用 InfoNCE 计算损失. Yang 等[44]采用与 CoMatch 相似的损失构建方法, 对有标签的数据采用交叉熵损失训练, 对软标签置信度高的无标签数据采用交叉熵训练, 对其余数据采用 InfoNCE 进行对比学习训练, 在该算法中, 由于包含类感知模块, 可以获得与当前训练样本相同类的样本集合, 因此对比学习损失部分采用实例 InfoNCE 和 SCL 相结合的损失函数. 此外, Wang 等[43] 将交叉熵损失和 SCL 结合在一起, 获得了良好的效果.

无监督混合损失. Park 等[80] 将对比学习损失融合到基于 GAN 的图像风格迁移任务中, 该方法对变换前后相同位置的图像块进行对比学习, 将对比学习损失辅助于 GAN 损失, 训练图像风格迁移模型, 获得了良好的效果.

其他混合损失. Rai 等[46] 对于具有多视图特征的数据, 对同一样本在不同视角下的特征, 分别采用 InfoNCE、MSE 和合页损失计算这些特征之间的相似性, 并将三个相似性度量结果进行混合, 用于网络模型优化. Kim 等[37] 将 Mixup 方法用在了对比学习中, 该方法假定合成后的样本同时属于合成前的两个样本的类别, 然后将合成样本分别与合成前的样本进行对比学习, 得到两个损失, 这两个损失的混合系数与合成图像时所产生的混合系数一致. 基于相似的 InfoNCE 与交叉熵损失结合的思想, Kumar 等[41]解决了无监督视频动作分割问题, Yang 等[81] 解决了文本-图像跨模态特征提取问题, Dong 等[82] 实现对五种模态数据的跨模态特征提取.

2.6 相关应用

对比学习在分类、分割、预测等下游任务中均有重要应用. 本文针对每种下游任务, 按照数据的类型, 介绍相关的应用. 本文将数据的类型概括为静态数据和序列数据, 其中静态数据主要有图像、关系型数据、点云和图结构等类型, 序列数据主要有视频、音频、信号等类型.

2.6.1 分类任务

分类任务是对比学习最常见的下游应用. 在静态数据中, 针对图像分类任务, Hou 等[83] 提出基于对比学习的半监督高光谱图像分类算法, 解决有标注数据不足时高光谱图像分类问题. 该算法分为 2 个阶段对模型进行训练, 第 1 阶段, 对于无标签样本, 利用对比学习方法对模型进行预训练. 第 2 阶段, 利用有标注的样本对模型进行监督学习. 针对小样本遥感场景分类问题, Li 等[78] 将无监督对比学习方法融合到小样本学习的框架中, 提高了模型的特征提取能力. 郭东恩等[84] 将监督对比学习方法引入到遥感图像场景分类任务中, 通过监督对比学习预训练, 提高了遥感图像分类精度. Aberdam 等[85] 将对比学习应用到文本图像识别任务. 由于对文本图像采用随机增广的方法可能会导致文本内容的丢失等问题, 因此, 作者首先设计可对齐的文本图像增广技术, 然后, 基于同步对称网络架构进行对比学习训练, 最终提高了文本图像识别的准确率. 在细粒度分类问题中, Zhang 等[86] 直接采用数据集中包含的分层语义标签, 利用 SCL 损失构建了一个细粒度分类对比学习算法. 对遥感图像数据而言, 同一个地理位置的图像语义信息几乎不随时间的变化而变化. 基于该特点, Ayush 等[39] 设计了一个针对遥感图像的对比学习方法, 该方法将同一地理位置不同时间的两幅图像作为对比学习中的正样本对. 基于 MoCo 架构, 该方法将图像定位的代理任务添加到其中一个分支的特征提取网络之后, 辅助模型训练, 从而提高了下游任务的预测性能.

卢绍帅等[87] 将监督对比学习应用到了文本数

据的情感分类研究中. 在弱监督预训练阶段, 采用三元组损失预训练模型; 随后, 在下游的分类器训练阶段, 采用 SCL 损失和交叉熵损失联合优化网络, 获得更好的分类结果.

在序列数据的分类中, 也涌现出了一些基于对比学习的算法. 李巍华等[88] 将 MoCo 方法迁移到故障信号诊断研究领域中, 首先, 通过对信号进行无监督对比学习预训练, 获得良好的特征提取网络. 然后, 再进行分类网络训练, 解决了信号故障诊断问题.

自监督对比学习的训练通常分为两个独立的阶段, 即特征提取网络训练和分类器训练. 在分类器训练阶段, 有是否冻结特征提取网络参数的两种选择. Wang 等[77] 认为这种两阶段的学习方式会损害特征提取网络和分类器的兼容性, 因此提出一个混合框架进行特征提取和分类器的联合学习. 该方法的对比学习部分采用同步对称网络架构, 并在特征提取网络后面加入一个分类器. 在训练过程中, 通过一个平滑因子来调整两个损失的权重, 使得对比学习在训练开始时起主导作用, 随着训练时间的推移, 分类器学习过程逐渐主导训练.

2.6.2 分割任务

分割任务指的是对图像的语义分割、实例分割, 视频中的动作分割等任务. 图像分割任务关注像素级的分类, 因此在此类任务中, 特征提取网络能否学习到良好的局部特征至关重要. Wang 等[89] 为了更好地学习到图像的局部空间特征, 提出密集对比学习算法 (Dense contrastive learning, DenseCL). 该方法提出全局对比学习框架和局部对比学习框架, 每个框架均采用同步对称网络架构, 两个框架共享同一个特征提取网络. 其中, 局部对比学习框架对卷积得到的特征取消拉平操作, 从而保留特征的空间信息, 使得学习到的特征提取网络更适合于分割任务. 在医学图像分割领域, 由于数据的标注过程非常依赖专家知识, 获取大量的有标注数据代价十分高昂, 因此, 如何利用大量的无标签医学数据训练图像分割模型是一个很重要的研究问题. Chaitanya 等[57] 将对比学习的思想应用到该领域, 提出基于自编码器框架的全局 _ 局部对比学习网络, 在该方法中, 全局对比学习目标是学习图像的全局语义信息, 局部对比学习目标是学习局部特征信息. 全局网络和局部网络共享同一个编码器. 医学图像中有一个 “卷” (Volume) 的概念, 对于全局网络, 正样本对来自于同一幅图像的不同卷. 对于局部网络, 正样本对来自同一幅图像编码后特征的同一个空间位置. 两个分支均采用 InfoNCE 进行损失计算. 康健等[90] 采用监督对比学习方法解决高分辨率 SAR 图像的分割问题, 通过改进的 SCL损失提高同类建筑像素特征之间的相关性, 最终提高模型对建筑物的分割精度. Wang 等[91] 在 Mask R-CNN[92] 框架中加入对比学习模块, 提高了像素级特征的可分辨能力, 获得更好的图像分割结果.

在视频动作分割问题中, Kumar 等[41] 提出基于对比学习的无监督视频动作分割方法. 该方法对 SwAV 算法进行改进, 且不需要图像增广, 利用 “视频数据的相邻帧为同类样本” 这一假设, 将相邻帧的图像作为正样本对进行对比学习, 完成视频动作分割任务.

2.6.3 视频及关系数据预测任务

在视频预测问题中, 研究者使用密集预测编码 (Dense predictive coding, DPC)[53], 预测视频未来帧的信息. Han 等[42] 将 DPC 与存储库思想结合起来, 提出存储增强密集预测编码方法, 该方法将视频的特征保存到存储库模块中, 并设计了存储库寻址机制. 通过该存储库模块的设计, 网络在训练过程中能够考虑更长时间段的特征, 使预测结果更好. 此外, 为了更好地捕捉到视频中的重要信息, Zhang 等[93] 提出对编码视频同时进行帧间以及帧内的对比学习, Han 等[53] 对不同视角的特征进行对比学习.

Bahri 等[94] 将对比学习方法应用到了关系型数据预测任务中. 为了构建训练所需的正样本对, 作者提出了一种面向关系型数据的增广方法. 该方法受启发于关系型数据中同一维度 (属性) 下的信息语义相同的特点, 先在输入样本的对应维度上随机抹除一部分数据, 然后, 从其他样本的相同维度的数据中随机抽取信息填充到当前输入样本被抹除的位置中, 最后, 基于同步对称架构的网络进行对比学习. 训练得到的模型可用于预测关系型数据中丢失区域的信息.

2.6.4 长尾识别任务

粗略地说, 长尾数据是指尾部类众多的不均衡数据, 而长尾学习的主要研究目标是提升尾部类的识别正确率. 最近几年, 研究者们开始尝试将对比学习的思想和技术应用到长尾学习任务中.

文献 [95] 提出 K-正样本对比损失 (K-positive contrastive loss, KCL), 将对比学习与长尾识别任务结合起来. 具体而言, 在长尾学习的特征学习阶段, KCL 使用对比学习方法, 但每个训练样本仅随机选取 K 个同类样本. 而在后续的分类器训练阶段, 仍采用传统的交叉熵损失, 但使用类均衡采样, 以平衡不同类的样本量、提升少数类的分类准确度. 文献 [96] 提出目标监督对比学习 (Targeted supervised contrastive learning, TSC) 方法, 将监督对

比学习用于长尾识别的任务中. 该方法首先在特征空间中设定均匀分布的聚类中心, 然后, 在 KCL 损失的基础上, 增加样本到其聚类中心之间的距离计算的损失项, 以将样本逼近其聚类中心, 使得不同类别之间的分类界限更加清晰.

文献 [97] 提出平衡对比学习损失 (Balanced contrastive learning, BCL), 用于长尾识别. 该方法将类别中心加入到对比学习计算中, 并求批中每个类别的样本的梯度的平均值, 以减少多数类样本在梯度方面的影响.

2.6.5 其他任务

对比学习除了在分类、分割、预测这些任务中有广泛的应用以外, 在多模态学习任务中, 也有重要作用. 多模态学习任务通常包含两大类子问题, 即模态内的特征学习问题和模态间特征对齐问题.

Yang 等[81] 提出了一种视觉 _ 语言跨模态对比学习方法. 该方法提出将对比学习方法应用于模态内特征提取网络的训练, 使各模态的特征提取网络能力更强, 解决了跨模态学习过程中模态内的特征学习问题. Dong 等[82] 将对比学习运用在包含 5 个模态数据的模型训练中. 针对模态内特征学习问题, 该方法采用掩码恢复任务作为模态内模型的代理任务. 然后, 针对不同模态特征之间的对齐问题, 设计模态间对比学习模块, 利用模态间的对齐得分矩阵衡量不同模态间信息的相似度, 进行更好的对比学习. Afham 等[98] 将对比学习算法引入到 3D 点云表示学习中. 该模型采用两个并行的对比学习方法, 其中一个对比学习方法基于 SimCLR 网络结构, 在点云数据内部进行对比学习, 另一个对比学习方法引入二维图像数据, 将点云数据与对应的二维图像数据进行跨模态的对比学习.

Laskin 等[99] 将对比学习结合到强化学习中, 用于提高特征提取网络的能力. 该方法基于 MoCo 架构, 分支 1 的输出, 送入强化学习中; 分支 1 和分支 2 的输出, 组对送入对比学习中.

3 综合对比分析

本节根据前文提出的对比学习归类方法, 对现有的方法进行归纳汇总和整体分析. 表 4 汇总了代表性的对比学习方法及其所属归类 (包括具体样本对构造、图像增广方法、网络架构类型和损失函数). 另外, 本节还对代表性对比学习方法的性能进行了对比分析.

3.1 整体分析

表 4 以时间为序, 依据前文提出的归类方法, 对主要的深度对比学习算法进行归纳汇总, 从样本对构造、图像增广、网络架构和损失函数 4 个层面进行分析. 该表的最后一列主要是为了区分是否为有监督对比学习方法, “无标签”表示数据无标注, 对应无监督对比学习方法, “部分标签”表示半监督对比学习算法, 而 “有标签”表示监督对比学习算法.

图 14 从宏观上统计各细分类所采用方法的占比情况. 其中, 1) 在样本对构造层, 最常采用的为随机采样方法, 即将数据集打乱顺序, 从其中随机抽样一个批次进行训练. 但其它样本对构造方法越来越受到研究者的关注 (尤其是正样本扩充和困难样本构造类方法); 2) 在网络架构层, 最常用的是同步对称网络架构, 但以 MoCo 为代表的异步对称网络架构和同步非对称网络架构也得到了大量关注; 3) 在损失函数层面, 最常用的是 InfoNCE 损失函数及其变种, 占比约 70%, 但随着对比学习下游应用的发展, 混合损失函数不断涌现; 4) 在应用领域层面, 目前对比学习最常用于无监督学习领域, 但也有越来越多的工作将其应用在有监督学习和半监督学习中. 需要说明的是, 在图像增广层, 大部分工作都是直接采用简单的图像变换方法, 而数据合成及图像语义增强方法较少, 故未对其进行可视化展示.

分析不同对比学习方法随时间演进的规律, 发现对比学习呈现以下发展趋势: 从样本对构造层面来看, 对比学习从一开始需要大批次采样负样本, 发展到通过队列采样负样本, 再发展到不需要负样本的对比学习方法, 以减轻对比学习对计算资源的要求, 这亦是其必须解决的关键问题之一.

从网络架构层面来看, 对比学习有以下发展趋势: 最初的对比学习架构主要采用同步对称和异步对称架构, 但随着 BYOL 和 SimSiam 等异步非对称网络发展, 因其只需要正样本和小批次数据就能进行对比学习训练的优点, 这种网络架构得到了越来越多的关注. 同时, 由于聚类方法能够有效地提取数据的语义信息, 因此聚类方法与上述 4 种架构的结合也越来越受到研究者的关注, 并得到了一定的发展. 此外, 所采用的主干网络的结构也逐渐从卷积神经网络向 Transformer 过渡. 从损失函数层面来看, 对比学习呈以下发展趋势: 从初始的 InfoNCE 损失函数发展到各种变种, 再到尝试采用多种传统损失函数进行对比学习, 再到混合多种不同类型的损失函数进行训练, 以提升模型的性能.

从下游任务来看, 对比学习从开始时只关注图像分类, 逐渐向视频分析、遥感图像处理、医学影像分析、文本分析和多模态学习等任务拓展, 且具有进一步的发展空间.

表 4 对比学习方法整体归类分析

Table 4 Analysis of different contrastive learning methods based on our proposed taxonomy

文献	年份	会议/刊物	样本对构造	图像增广	网络架构	损失函数	数据标注
[7]	2019	arXiv	随机采样	图像变换	同步非对称	InfoNCE 类	无
[2]	2020	ICML	随机采样	图像变换	同步对称	InfoNCE 类	无
[3]	2020	CVPR	随机采样	图像变换	异步对称	InfoNCE 类	无
[4]	2020	NIPS	随机采样	图像变换	异步非对称	传统损失	无
[5]	2020	NIPS	随机采样	图像变换	聚类/同步对称	传统损失	无
[8]	2020	NIPS	随机采样	图像变换	同步对称	InfoNCE 类	有
[9]	2020	NIPS	随机采样	图像变换	同步对称	混合损失	部分
[33]	2020	NIPS	困难样本构造	图像变换	异步对称	InfoNCE 类	无
[36]	2020	NIPS	剔除假负样本	图像变换	同步对称	InfoNCE 类	无
[37]	2020	arXiv	正样本扩充	图像变换	异步对称	InfoNCE 类	无
[42]	2020	NIPS	正样本扩充	图像变换	同步对称	InfoNCE 类	无
[45]	2020	ECCV	构造多视图	图像变换	同步对称	InfoNCE 类	无
[52]	2020	NIPS	随机采样	语义增广	同步对称	InfoNCE 类	无
[55]	2020	CVPR	随机采样	图像变换	同步非对称	InfoNCE 类	无
[57]	2020	NIPS	随机采样	图像变换	同步非对称	InfoNCE 类	无
[70]	2020	arXiv	随机采样	图像变换	异步对称	InfoNCE 类	无
[109]	2020	ECCV	随机采样	图像变换	同步非对称	InfoNCE 类	无
[6]	2021	CVPR	随机采样	图像变换	异步非对称	传统损失	无
[32]	2021	ICCV	困难样本构造	图像变换	异步对称	InfoNCE 类	无
[34]	2021	CVPR	困难样本构造	图像变换	同步对称	混合损失	部分
[39]	2021	ICCV	正样本扩充	图像变换	异步对称	混合损失	无
[40]	2021	CVPR	正样本扩充	图像变换	同步对称	InfoNCE 类	无
[46]	2021	CVPRW	构造多视图	图像变换	同步对称	混合损失	无
[61]	2021	ICCV	随机采样	图像变换	异步非对称	InfoNCE 类	无
[63]	2021	ICLR	随机采样	图像变换	聚类/异步对称	InfoNCE 类	无
[64]	2021	CVPR	随机采样	图像变换	聚类架构	InfoNCE 类	无
[66]	2021	AAAI	随机采样	图像变换	聚类/同步对称	InfoNCE 类	无
[77]	2021	CVPR	随机采样	图像变换	同步对称	混合损失	有
[79]	2021	ICCV	随机采样	图像变换	同步对称	混合损失	部分
[83]	2021	TGRS	随机采样	图像变换	同步对称	混合损失	部分
[78]	2021	TGRS	随机采样	图像变换	同步对称	InfoNCE 类	无
[85]	2021	CVPR	随机采样	图像变换	同步对称	InfoNCE 类	无
[35]	2022	WACV	剔除假负样本	图像变换	同步对称	InfoNCE 类	无
[41]	2022	CVPR	正样本扩充	图像变换	同步非对称	混合损失	无
[43]	2022	ICLR	正样本扩充	图像变换	异步对称	混合损失	部分
[44]	2022	CVPR	正样本扩充	图像变换	同步对称	混合损失	部分
[47]	2022	CVPR	随机采样	图像变换	任意架构	InfoNCE 类	无
[48]	2022	CVPR	随机采样	图像合成	异步对称	InfoNCE 类	无
[54]	2022	TAI	随机采样	图像变换	同步非对称	InfoNCE 类	无
[65]	2022	CVPR	剔除假负样本	图像变换	聚类/异步对称	InfoNCE 类	无
[72]	2022	ICLR	随机采样	图像变换	同步对称	InfoNCE 类	无
[76]	2022	AAAI	随机采样	图像变换	同步对称	传统损失	无
[94]	2022	ICLR	随机采样	图像变换	同步对称	InfoNCE 类	无
[98]	2022	CVPR	随机采样	图像变换	同步对称	InfoNCE 类	无
[103]	2022	ICLR	随机采样	图像变换	同步对称	混合损失	无

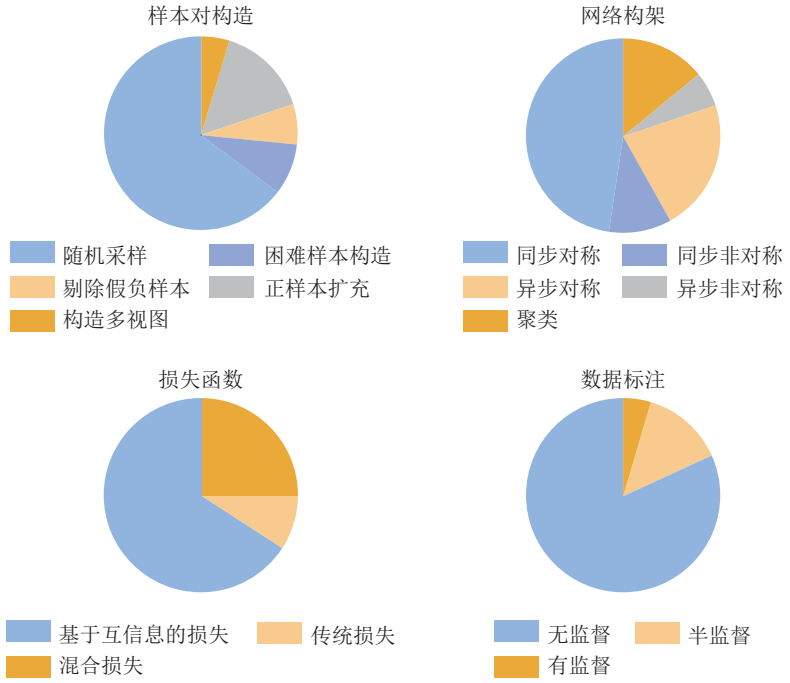


图 14　不同类型的对比学习方法统计展示

Fig. 14　The statistical results of different contrastive learning methods

3.2　不同对比方法的性能分析

本部分内容将不同对比学习算法在常用数据集上的性能表现进行对比和分析.

1) 图像分类. 图像分类是最常见的下游任务, 现有方法通常在 ImageNet 数据集上采用第 1.6 节中给出的线性评估方法进行比较. 表 5 给出了不同算法在 ImageNet 数据集上的分类准确度. 可以观察到, 在 ImageNet 数据集上的分类任务中, 在无监督学习中, SwAV 方法获得了最佳的分类效果, 在有监督学习中, PaCo 获得了最好的效果. 除了 ImageNet 之外, 常用于图像分类评估的数据集还有 Cifar10 和 Cifar100, Food101, Birdsnap, Sun397, Cars, Aircraft, DTD, Pets, Caltech-101, Flowers 等. 为全面评估对比学习方法得到模型的可迁移性能, 在这些数据集上采用第 1.6 节中给出的评估方法进行对比分析, 其中, 主干网络模型均采用 ResNet50, 在 ImageNet 数据集上训练得到. 结果如表 6 所示. 表中 VOC07 采用 mAP (Mean average precision) 指标进行验证, Aircraft、Pets、Caltech 和 Flowers 数据集采用平均准确率 (Mean per class accuracy) 进行验证, 其余数据集采用 Top 1 分类进度进行验证. 半监督学习根据所采用的有标注数据的比例进行分类性能的评估. 如表 7 所示, 1% 指的是训练过程中使用了 1% 的有标注数据, 10% 指的是训练过程中使用了 10% 的有标注数据. 其中, 所有方法采用的主干网络均为 ResNet50. 整体而言, SimCLRv2 取得了最佳的半监督分类效果.

表 5　不同对比学习算法在 ImageNet 数据集上的分类效果

Table 5　The classification results of different contrastive learning methods on ImageNet

文献	主干网络	Top 1 (%)	Top 5 (%)	数据标注
MoCov1[3]	ResNet50	60.6	—	无
CPCv2[100]	ResNet50	63.8	85.3	
	ResNet161	71.5	90.1	
PCL[63]	ResNet50	67.6	—	
SimCLR[2]	ResNet50	69.3	89	
MoCov2[70]	ResNet50	71.1	—	
SimSiam[6]	ResNet50	71.3	—	
BT[101]	ResNet50	73.2	91	
VICReg[103]	ResNet50	73.2	91.1	
HCSC[65]	ResNet50	73.3	—	
MoCov3[61]	ResNet50	73.8	—	
	Transformer	76.5	—	
BYOL[4]	ResNet50	74.3	91.6	
SwAV[5]	ResNet50	**75.3**	—	
DINO[59]	ResNet50	**75.3**		
	Transormer	**77**	—	
TSC[96]	ResNet50	77.1	—	有
SCL[8]	ResNet50	78.7	94.3	
PaCo[67]	ResNet50	79.3	—	

表 6 不同对比学习算法在各数据集上的迁移学习效果

Table 6 The transfer learning results of different contrastive learning methods on each dataset

文献	Food (%)	Cifar10/Cifar100 (%)	Birds (%)	SUN (%)	Cars (%)	Aircraft (%)	VOC (%)	DTD (%)	Pets (%)	Caltech (%)	Flowers (%)
线性评估											
SimCLR[2]	68.4	90.6/71.6	37.4	58.8	50.3	50.3	80.5	74.5	83.6	90.3	91.2
SimCLRv2[9]	73.9	92.4/76	44.7	61	54.9	51.1	81.2	76.5	85	91.2	93.5
BYOL[4]	75.3	91.3/78.4	57.2	62.2	67.8	60.6	82.5	75.5	90.4	94.2	96.1
微调评估											
MMCL[76]	82.4	96.24/82.1	—	—	89.2	85.4	—	73.5	—	87.8	95.2
SimCLR[2]	88.2	97.7/85.9	75.9	63.5	91.3	88.1	84.1	73.2	89.2	92.1	97
SimCLRv2[9]	88.2	97.5/86	74.9	64.6	91.8	87.6	84.1	74.7	89.9	92.3	97.2
BYOL[4]	88.5	97.8/86.1	76.3	63.7	91.6	88.1	85.4	76.2	91.7	93.8	97
FNC[35]	88.3	97.7/86.8	76.3	64.2	92	88.5	84.7	76	90.9	93.6	97.5
SCL[8]	87.2	97.42/84.3	75.2	58	91.7	84.1	85.2	74.6	93.5	91	96

表 7 不同半监督对比学习算法在 ImageNet 上的分类效果

Table 7 The classification results of different semi-supervised contrastive learning methods on ImageNet

文献	Top 1 (%)		Top 5 (%)	
	1%	10%	1%	10%
PIRL[55]	30.7	60.4	57.2	83.8
PCL[63]	—	—	75.3	85.6
SimCLR[2]	48.3	65.6	75.5	87.8
BYOL[4]	53.2	68.8	78.4	89
SwAV[5]	53.9	70.2	78.5	89.9
BT[101]	55	69.7	79.2	89.3
HCSC[65]	55.5	68.7	80.9	88.6
CoMatch[79]	67.1	73.7	87.1	91.4
SimCLRv2[9]	**73.9**	**77.5**	**91.5**	**93.4**

2) 图像检测和分割. 在图像检测和分割任务中, VOC 和 COCO 数据集是较为常用的数据集. 不同对比学习算法在这两个数据集上的分割性能在表 8 中进行了汇总展示.

表 8 中的所有模型均在 ImageNet 数据集上进行预训练, 随后在 VOC 数据集或 COCO 数据集上进行微调. 表中所列出的算法均采用图像变换方法进行图像增广. 网络结构方面, SimCLR 采用同步对称结构, MoCo 和 DenseCL 采用异步对称结构, BYOL 和 SimSiam 采用异步非对称结构, SwAV 采用聚类对比学习结构. 损失函数层面, SimCLR、MoCo 以及 DenseCL 均采用 InfoNCE 损失函数; BYOL、SwAV 及 SimSiam 均采用传统损失函数 (如第 2.5.2 节所述).

由于 DenseCL 在训练过程中加入了局部对比学习结构, 使得学习到的模型对局部信息的提取能力提高, 进而提高了模型的检测和分割能力, 获得了目前最好的图像检测和分割效果.

表 8 不同对比学习算法在图像分割任务上的性能表现

Table 8 The image segmentation results of different contrastive learning methods on VOC and COCO dataset

文献	AP		APm
	VOC (%)	COCO (%)	COCO (%)
BYOL[4]	55.3	37.9	33.2
SwAV[5]	55.4	37.6	33.1
SimCLR[2]	55.5	37.9	33.3
MoCov2[70]	57	39.2	34.3
SimSiam[6]	57	39.2	34.4
DenseCL[89]	**58.7**	**40.3**	**36.4**

4 现存挑战和未来发展方向

4.1 现存挑战

4.1.1 对比学习中的崩塌问题研究

崩塌问题是对比学习研究中经常会遇到的问题, 崩塌包含两种现象, 第 1 种称为完全崩塌, 第 2 种称为维度崩塌. 完全崩塌指的是对于任意输入, 模型会将其输出到同一个特征向量上. 维度崩塌是指特征向量只能占据特征空间的某一个子空间中的现象. 完全崩塌如图 15(a) 所示, 维度崩塌如图 15(b) 所示.

一般来说, 有两种办法观测网络是否发生崩塌: 1) 将训练好的特征提取网络应用到下游任务, 若下游任务表现不好, 则可能出现崩塌. 2) 特征提取网络训练完成后, 首先, 将一个批次的测试数据进行特征提取. 然后, 计算该批次特征矩阵的协方差矩阵. 最后, 对协方差矩阵进行奇异值分解. 如果奇异值矩阵对角线上的值在某一个维度开始发生断层现象, 则网络发生了崩塌[72].

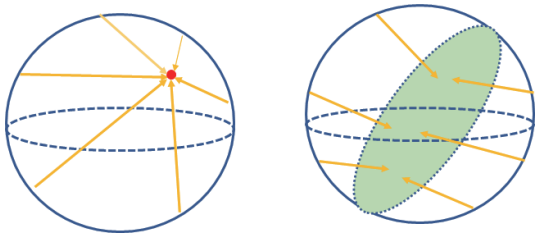


图 15　完全崩塌与维度崩塌示例

Fig. 15　Example of complete collapse and dimensional collapse

对于采用负样本进行对比学习的方法来说, 由于训练过程中网络可以见到大量不同的负样本, 因此在一定程度上可以避免维度崩塌问题. 而对于只采用正样本进行对比学习的算法来说, 则需要一些特殊的设计以避免崩塌[4−6]. SwAV 采用聚类的思想避免了不同的输入输出到同一个特征向量上的问题, 从而防止了完全崩塌的出现. BYOL 采用了非对称的网络设计和动量更新方法避免网络的完全崩塌. SimSiam 通过实验证明了, 通过交叉梯度回传, 能够有效的防止网络完全崩塌. 在避免维度崩塌问题上, 有一些论文专门对其进行研究和讨论, 根据这些论文的讨论内容, 可以将其分为两种思路, 即缓解维度崩塌思路和避免维度崩塌思路.

1) 缓解维度崩塌

Jing 等[72] 探讨了对比学习中出现的维度崩塌问题, 作者发现, 在对比学习中, 强大的图像增广方法和隐式正则化可能是产生维度崩塌的原因. 在该论文中, 作者提出了一个简单有效的解决维度崩塌办法: DirectCLR. DirectCLR 采用与 SimCLR 相似的同步对称网络架构, 但进行了下述两点改进: 1) 抛弃投影头. 2) 在完成特征提取之后, 只在特征向量的前 d 个维度上计算损失. 对于这种方法的一个直观的理解是: 通过训练, 将没有崩塌的维度全部集中到前 d 个维度中去, 因此最后在前 d 个维度上就没有了崩塌问题. 在这种理解下, 超参数 d 的物理意义为没有发生崩塌的子空间维度.

DirectCLR 中的实验证明, 与没有投影头的 SimCLR 相比, DirectCLR 的效果要好一些. 然而当 SimCLR 有一个两层的非线性投影头的时候, 表现会比 DirectCLR 要好很多. 这个现象可能在一定程度上证明了如下观点: 投影头在对比学习方法中承担着寻找无崩塌子空间的任务. 在该观点下, 首先, 特征提取网络将输入样本进行特征提取, 获得特征矩阵. 然后, 投影头将高维特征矩阵映射为低维投影矩阵. 最后, 在低维的投影空间中计算损失. 通过一次投影头的降维变换, 将在高维特征空间中可能出现的崩塌现象转移到没有崩塌的低维投影空间中, 从而保证了损失计算的有效性.

2) 避免维度崩塌

对维度崩塌的直观理解是特征空间中的部分维度失去了信息. 为了避免这种情况的出现, Zbontar 等[101] 等计算分支 1 特征矩阵和分支 2 特征矩阵的协方差矩阵, 通过将协方差矩阵的学习目标设计为同形状的单位矩阵, 来完成以下两个目标: 1) 保证每个维度信息的有效性. 2) 消除不同维度之间信息的相关性. 通过设计上述学习目标, 在训练过程中, 协方差矩阵的主对角线上的元素值会逐渐向 1 靠近, 其它元素的值会逐渐向 0 靠近. 由于协方差矩阵的主对角线元素可以代表该维度信息的信息量, 大于 0 可以保证该维度没有发生崩塌, 非对角线元素代表不同维度之间信息的相关性, 等于 0 即可保证各维度信息独立. 因此通过学习, 在避免维度崩塌出现的同时也消除了维度间的冗余性, 使得训练过程更加稳定有效. Hua 等[102] 深度分析了自监督学习中的特征崩塌问题, 同样提出特征去相关方法避免崩塌.

Bardes 等[103] 进行了更加深入的探索, 提出分别在分支 1 和分支 2 的特征矩阵上采用与 Zbontar 类似的方法去除维度间的冗余性. 与 Zbontar 方法不同的是, Bardes 引入了特征矩阵的方差信息, 当某一个维度的方差信息向 0 靠近时, 可以认为该维度发生了崩塌, 因此作者通过合页损失使得每个维度的方差信息向超参数 γ 靠近, 从而避免了崩塌. 此外, 该方法还引入了不变项约束, 该约束采用 MSE 损失减小同一个样本在分支 1 中特征和在分支 2 中特征之间的距离.

4.2　算法效率优化

在对比学习获得成功的背后, 因其需要大批次样本参与学习的特点, 从而对计算资源和存储资源产生较高要求, 这在一定程度上限制了对比学习的发展. Bao 等[104] 将对比学习过程中使用的负样本数量 N 与分类任务训练过程进行联合分析, 发现增加 N 的值可以使得分类损失的上界和下界之间的截距减小. 这篇论文从理论角度解释了负样本数量与下游任务性能之间的关系, 证明了在无监督预训练阶段使用的负样本数量越大, 对后续的分类任务的训练结果就越稳定. Yeh 等[71] 通过对 InfoNCE损失求解梯度发现对比学习存在负正耦合效应. 此处的负正耦合效应指的是以下两种情况: 1) 当正样本对相似度高 (简单正样本对) 时, 损失对负样本回传的梯度变小. 2) 当查询样本与所有负样本相似度都很

小 (简单负样本) 时, 损失对正样本回传的梯度变小. 以上两种情况说明无论是训练中采用简单正样本对或简单负样本集均会造成训练效率低的问题. 为了解决这个问题, 论文提出解耦对比学习损失函数 DCL. 该损失去除了 InfoNCE 中的负正耦合系数, 从而可以用更小的批次来训练对比学习网络, 同时提高计算效率. 除此以外, 很多经典的对比学习方法如 BYOL, SimSiam 等都对训练批次大小进行了不同的实验. 然而, 如何利用更小的批次获得更好的对比学习效果, 这仍是未来的一个挑战.

4.3 一致性与均匀性矛盾问题

在对比学习中, 由于输出特征经过正则化, 因此所有图像均会被映射到特征空间的单位超球面上. 为了使模型具有良好的泛化能力, 希望在超球面上分布的特征具有以下两点特质: 1) 同类样本特征集合具有一致性; 2) 所有特征分布具有均匀性[62]. 一致性指的是相同类的样本特征应当集合在超球面的同一片区域内, 均匀性指的是所有样本的特征应当在超球面均匀分布, 直观理解如图 16 所示. 避免神经网络崩塌的最好办法就是同时满足一致性和均匀性. 均匀性有助于对比学习学习到可分离的特征, 但是过度追求均匀性, 将会导致一些语义相似的样本的特征一定程度上互相远离. Wang 等[105] 针对这个问题进行详细分析, 认为在对比学习训练中, 一致性与均匀性是一个互相对抗的关系. 该论文通过对 InfoNCE 损失函数中的温度参数 τ 进行分析, 发现温度参数 τ 的取值影响一致性与均匀性的结果, 因此对于采用 InfoNCE 损失进行对比学习的方法来说, 温度参数 τ 的取值对模型最终的表现非常重要.

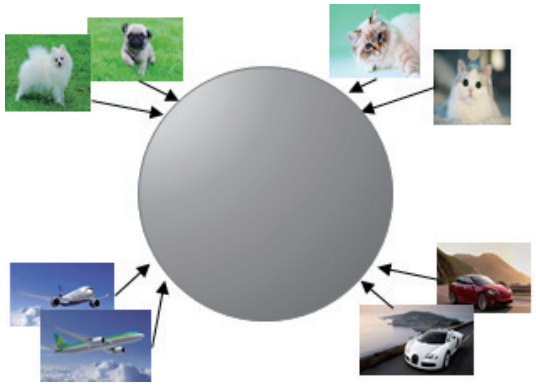


图 16 对比学习中一致性和均匀性的概念

Fig. 16 The concept of uniformity and alignment in contrastive learning

4.4 未来发展方向

本文在对目前的对比学习论文进行归纳和总结后, 认为该研究领域还存在许多可以探索的问题, 同时存在一些可以与其他领域互相借鉴和发展的方向, 具有广泛的研究前景, 以下是对该领域发展的展望:

1) 对比学习中样本对的选择方法仍存在发展空间, 在训练过程中剔除假负样本以及选择合适的正样本对能够有效地提高特征学习网络的学习效果. 因此如何更加合理地剔除假负样本和选择正样本对是一个值得研究的关键问题.

2) 解决对比学习训练过程中的一致性与均匀性矛盾是一个十分重要的问题, 如果该问题得到解决, 能在很大程度上提高特征提取网络在下游任务上的泛化能力.

3) 主动学习是一种通过最少的标注样本获得最好的训练效果的学习技术[106]. 在深度主动学习领域, 网络模型需要首先在一个含有标签的数据集 L_0 上进行预训练, 然后通过查询策略从无标签数据集 U 中筛选最有用的样本给专家进行标注, 最后更新当前训练的有标签数据集 L, 采用 L 的数据继续训练网络. 重复以上过程直到标注预算耗尽或触发停止策略[106]. 对比学习是一种良好的模型预训练方法, 可以自发的通过无标签数据或少量标签数据训练出特征提取模型, 因此可以将对比学习算法引入到主动学习的网络模型预训练过程中, 或作为辅助主动学习挑选待标注样本的方法.

4) 对比学习和无监督域自适应[107−108] 的结合. 在无监督域自适应问题中, 源域数据存在标签, 目标域数据不存在标签, 源域数据和目标域数据分布相近或相同, 且拥有相同的任务[107], 如何将源域数据和目标域数据一同训练, 使得模型能够在目标域上获得良好的效果是无监督域自适应的核心问题. 在无监督域自适应研究中, 源域数据和目标域数据可以通过自监督训练方法联合训练模型, 对比学习就是一种先进的自监督训练算法, 因此如何将对比学习方法与无监督域自适应方法进行有效结合是一个值得研究的问题.

5) 目前对比学习主要的下游应用是分类任务, 如何设计更多的对比学习方法应用到检测、追踪等下游任务中, 也将是未来的发展方向之一.

5 结束语

对比学习是近年的研究热点. 本文系统梳理了对比学习的研究现状, 提出一种将现有方法划分为样本对构造层、图像增广层、网络架构层、损失函数层和应用层的归类方法, 并从自监督对比学习算法入手, 分析和归纳近四年主要的对比学习方法. 而且, 本文还全面对比了不同方法在各种下游任务中

的性能表现, 指出了对比学习现存的挑战, 勾勒了其未来发展方向. 对比学习研究作为一个快速发展的研究领域, 在理论依据、模型设计、损失函数设计及与下游任务结合等方面还有较大的研究空间.

References

1 Jing L L, Tian Y L. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, **43**(11): 4037−4058

2 Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning (ICML). Vitual: ACM, 2020, 1597−1607

3 He K M, Fan H Q, Wu Y X, Xie S N, Girshick R. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020. 9726−9735

4 Grill J B, Strub F, Altché F, Tallec C, Richemond P H, Buchatskaya E, et al. Bootstrap your own latent a new approach to self-supervised learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS). Vancouver, Canada: Curran Associates Inc., 2020. 21271−21284

5 Caron M, Misra I, Mairal J, Goyal P, Bojanowski P, Joulin A. Unsupervised learning of visual features by contrasting cluster assignments. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. 9912−9924

6 Chen X L, He K M. Exploring simple Siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 15745−15753

7 Van Den Oord A, Li Y Z, Vinyals O. Representation learning with contrastive predictive coding [Online], available: https://arxiv.org/abs/1807.03748, January 22, 2019

8 Khosla P, Teterwak P, Wang C, et al. Supervised contrastive learning. In: Proceedings of the Advances in Neural Information Processing Systems, 2020, 33: 18661−18673

9 Chen T, Kornblith S, Swersky K, Norouzi M, Hinton G. Big self-supervised models are strong semi-supervised learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 1865

10 Jaiswal A, Babu A R, Zadeh M Z, Banerjee D, Makedon F. A survey on contrastive self-supervised learning. *Technologies*, 2020, **9**(1): 2

11 Le-Khac P H, Healy G, Smeaton A F. Contrastive representation learning: A framework and review. *IEEE Access*, 2020, **8**: 193907−193934

12 Liu X, Zhang F J, Hou Z Y, Mian L, Wang Z Y, Zhang J, et al. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 2021, **35**(1): 857−576

13 Hadsell R, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). New York, USA: IEEE, 2006. 1735−1742

14 Wu Z R, Xiong Y J, Yu S X, Lin D H. Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: 2018. 3733−3742

15 Bromley J, Guyon I, LeCun Y, Säckinger E, Shah R. Signature verification using a "Siamese" time delay neural network. In: Proceedings of the 6th International Conference on Neural Information Processing Systems. Denver, USA: Morgan Kaufmann Publishers Inc., 1993. 737−744

16 Li D W, Tian Y J. Survey and experimental study on metric learning methods. *Neural Networks*, 2018, **105**: 447−462

17 Jia D, Wei D, Richard S, Li J J L, Kai L, Li F F. ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, USA: IEEE, 2009. 248−255

18 Krizhevsky A. Learning Multiple Layers of Features from Tiny Images [Master thesis], University of Toronto, Canada, 2009

19 Bossard L, Guillaumin M, Van Gool L. Food-101-mining discriminative components with random forests. In: Proceedings of the 13th European Conference on Computer Vision (ECCV). Zurich, Switzerland: Springer, 2014. 446−461

20 Berg T, Liu J X, Lee S W, Alexander M L, Jacobs D W, Belhumeur P N. Birdsnap: Large-scale fine-grained visual categorization of birds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, USA: IEEE, 2014. 2019−2026

21 Xiao J X, Hays J, Ehinger K A, Oliva A, Torralba A. SUN database: Large-scale scene recognition from abbey to zoo. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). San Francisco, USA: IEEE, 2010. 3485−3492

22 Jonathan K, Michael S, Jia D, and Li F F. 3D object representations for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV). Sydney, Australia: IEEE, 2013: 554−561

23 Maji S, Rahtu E, Kannala J, Blaschko M, Vedaldi A. Fine-grained visual classification of aircraft [Online], available: https://arxiv.org/abs/1306.5151, June 6, 2013

24 Cimpoi M, Maji S, Kokkinos I, Mohamed S, Vedaldi A. Describing textures in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, USA: IEEE, 2014. 3606−3613

25 Parkhi O M, Vedaldi A, Zisserman A, Jawahar C V. Cats and dogs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, USA: IEEE, 2012. 3498−3505

26 Li F F, Rob F, Pietro P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop (CVPR-W). Washington, USA: IEEE, 2004: 178−178

27 Nilsback M E, Zisserman A. Automated flower classification over a large number of classes. In: Proceedings of the Sixth Indian Conference on Computer Vision, Graphics & Image Processing (ICVGIP). Bhubaneswar, India: IEEE, 2008. 722−729

28 Everingham M, Van Gool L, Williams C K I, Winn J, Zisserman A. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 2010, **88**(2): 303−338

29 Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft coco: Common objects in context. In: Proceedings of the 13th European Conference on Computer Vision (ECCV). Zurich, Switzerland: Springer, 2014. 740−755

30 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Ve-

gas, USA: IEEE, 2016. 770−778

31 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc., 2017. 6000−6010

32 Zhu R, Zhao B C, Liu J E, Sun Z L, Chen C W. Improving contrastive learning by visualizing feature transformation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 10286−10295

33 Kalantidis Y, Sariyildiz M B, Pion N, Weinzaepfel P, Larlus D. Hard negative mixing for contrastive learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 1829

34 Zhong Z, Fini E, Roy S, Luo Z M, Ricci E, Sebe N. Neighborhood contrastive learning for novel class discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 10862−10870

35 Huynh T, Kornblith S, Walter M R, Maire M, Khademi M. Boosting contrastive self-supervised learning with false negative cancellation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2022. 986−996

36 Chuang C Y, Robinson J, Yen-Chen L, Torralba A, Jegelka S. Debiased contrastive learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. 8765−8775

37 Kim S, Lee G, Bae S, Yun S Y. MixCo: Mix-up contrastive learning for visual representation [Online], available: https://arxiv.org/abs/2010.06300, November 15, 2020

38 Wang Meng-Lin. Contrastive Learning Based Person Re-identification. [Ph. D. dissertation], Zhejiang University, China, 2021 (王梦琳. 基于对比学习的行人重识别[博士学位论文], 浙江大学, 中国, 2021)

39 Ayush K, Uzkent B, Meng C L, Tanmay K, Burke M, Lobell D, et al. Geography-aware self-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 10161−10170

40 Qian R, Meng T J, Gong B Q, Yang M H, Wang H S, Belongie S, et al. Spatiotemporal contrastive video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 6960−6970

41 Kumar S, Haresh S, Ahmed A, Konin A, Zia M Z, Tran Q H. Unsupervised action segmentation by joint representation learning and online clustering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 20142−20153

42 Han T D, Xie W D, Zisserman A. Self-supervised co-training for video representation learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 477

43 Wang H B, Xiao R, Li S, et al. Contrastive Label Disambiguation for Partial Label Learning. In: Proceedings of the 10th International Conference on Learning Representations. Virtual: ICLR, 2022.

44 Yang F, Wu K, Zhang S Y, Jiang G N, Liu Y, Zheng F, et al. Class-aware contrastive semi-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 14401−14410

45 Tian Y L, Krishnan D, Isola P. Contrastive multiview coding. In: Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK: Springer, 2020. 776−794

46 Rai N, Adeli E, Lee K H, Gaidon A, Niebles J C. CoCon: Cooperative-contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Nashville, USA: IEEE, 2021. 3379−3388

47 Peng X Y, Wang K, Zhu Z, Wang M, You Y. Crafting better contrastive views for siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 16010−16019

48 Ding S R, Li M M, Yang T Y, Qian R, Xu H H, Chen Q Y, et al. Motion-aware contrastive video representation learning via foreground-background merging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 9706−9716

49 Li S, Gong K X, Liu C H, Wang Y L, Qiao F, Cheng X J. MetaSAug: Meta semantic augmentation for long-tailed visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 5208−5217

50 Wang Y L, Pan X R, Song S J, Zhang H, Wu C, Huang G. Implicit semantic data augmentation for deep networks. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2019. Article No. 1132

51 Li S, Xie M X, Gong K X, Liu C H, Wang Y L, Li F. Transferable semantic augmentation for domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 11511−11520

52 Tian Y L, Sun C, Poole B, Krishnan D, Schmid C, Isola P. What makes for good views for contrastive learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 573

53 Han T D, Xie W D, Zisserman A. Video representation learning by dense predictive coding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCV). Seoul, Korea: IEEE, 2019. 1483−1492

54 Nguyen N, Chang J M. CSNAS: Contrastive self-supervised learning neural architecture search via sequential model-based optimization. *IEEE Transactions on Artificial Intelligence*, 2022, **3**(4): 609−624

55 Misra I, Van Der Maaten L. Self-supervised learning of pretext-invariant representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 6706−6716

56 Bae S, Kim S, Ko J, Lee G, Noh S, Yun S Y. Self-contrastive learning [Online], available: https://arxiv.org/abs/2106.15499, February 9, 2021

57 Chaitanya K, Erdil E, Karani N, Konukoglu E. Contrastive learning of global and local features for medical image segmentation with limited annotations. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS). Vancouver, Canada: Curran Associates Inc., 2020. Article No. 1052

58 Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of

9th International Conference on Learning Representations. Austria: ICLR, 2021.

59 Caron M, Touvron H, Misra I, Jegou H, Mairal J, Bojanowski P, et al. Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 9630−9640

60 Richemond P H, Grill J B, Altché F, Tallec C, Strub F, Brock A, et al. BYOL works even without batch statistics [Online], available: https://arxiv.org/abs/2010.10241, October 20, 2020

61 Chen X L, Xie S N, He K M. An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 9620−9629

62 Caron M, Bojanowski P, Joulin A, Douze M. Deep clustering for unsupervised learning of visual features. In: Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018. 139−156

63 Li J N, Zhou P, Xiong C, et al. Prototypical contrastive learning of unsupervised representations. In: Proceedings of 9th International Conference on Learning Representations. Austria: ICLR, 2021.

64 Wang X D, Liu Z W, Yu S X. Unsupervised feature learning by cross-level instance-group discrimination. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 12581−12590

65 Guo Y F, Xu M H, Li J W, Ni B B, Zhu X Y, Sun Z B, et al. HCSC: Hierarchical contrastive selective coding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 9696−9705

66 Li Y F, Hu P, Liu Z T, Peng D Z, Zhou J T, Peng X. Contrastive clustering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, **35**(10): 8547−8555

67 Cui J Q, Zhong Z S, Liu S, Yu B, Jia J Y. Parametric contrastive learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 695−704

68 Gutmann M U, Hyvärinen A. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *The Journal of Machine Learning Research*, 2012, **13**: 307−361

69 Poole B, Ozair S, Van Den Oord A, Alemi A A, Tucker G. On variational bounds of mutual information. In: Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR, 2019. 5171−5180

70 Chen X L, Fan H Q, Girshick R, He K M. Improved baselines with momentum contrastive learning [Online], available: https://arxiv.org/abs/2003.04297, March 9, 2020

71 Yeh C H, Hong C Y, Hsu Y C, Liu T L, Chen Y B, LeCun Y. Decoupled contrastive learning. In: Proceedings of the 17th European Conference. Tel Aviv, Israel: ECCV, 2021.

72 Jing L, Vincent P, LeCun Y, et al. Understanding dimensional collapse in contrastive self-supervised learning. In: Proceedings of the 10th International Conference on Learning Representations. Virtual: ICLR, 2022.

73 Zhang Z L, Sabuncu M R. Generalized cross entropy loss for training deep neural networks with noisy labels. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook, USA: Curran Associates Inc., 2018. 8792−8802

74 Weinberger K Q, Saul L K. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, 2009, **10**: 207−244

75 Sohn K. Improved deep metric learning with multi-class n-pair loss objective. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona: Spain, Curran Associates Inc., 2016. 1857−1865

76 Shah A, Sra S, Chellappa R, Cherian A. Max-margin contrastive learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2022, **36**(8): 8220−8230

77 Wang P, Han K, Wei X S, Zhang L, Wang L. Contrastive learning based hybrid networks for long-tailed image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 943−952

78 Li X M, Shi D Q, Diao X L, Xu H. SCL-MLNet: Boosting few-shot remote sensing scene classification via self-supervised contrastive learning. *IEEE Transactions on Geoscience and Remote Sensing*, 2021, **60**: Article No. 5801112

79 Li J N, Xiong C M, Hoi S C H. CoMatch: Semi-supervised learning with contrastive graph regularization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 9455−9464

80 Park T, Efros A A, Zhang R, Zhu J Y. Contrastive learning for unpaired image-to-image translation. In: Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK: Springer, 2020. 319−345

81 Yang J Y, Duan J L, Tran S, Xu Y, Chanda S, Chen L Q, et al. Vision-language pre-training with triple contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 15650−15659

82 Dong X, Zhan X L, Wu Y X, Wei Y C, Kampffmeyer M C, Wei X Y, et al. M5Product: Self-harmonized contrastive learning for e-commercial multi-modal pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 21220−21230

83 Hou S K, Shi H Y, Cao X H, Zhang X H, Jiao L C. Hyperspectral imagery classification based on contrastive learning. *IEEE Transactions on Geoscience and Remote Sensing*, 2021, **60**: Article No. 5521213

84 Guo Dong-En, Xia Ying, Luo Xiao-Bo, Feng Jiang-Fan. Remote sensing image scene classification based on supervised contrastive learning. *Acta Photonica Sinica*, 2021, **50**(7): 0710002
(郭东恩, 夏英, 罗小波, 丰江帆. 基于有监督对比学习的遥感图像场景分类. 光子学报, 2021, **50**(7): 0710002)

85 Aberdam A, Litman R, Tsiper S, Anschel O, Slossberg R, Mazor S, et al. Sequence-to-sequence contrastive learning for text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 15297−15307

86 Zhang S, Xu R, Xiong C M, Ramaiah C. Use all the labels: A hierarchical multi-label contrastive learning framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022. 16639−16648

87 Lu Shao-Shuai, Chen Long, Lu Guang-Yue, Guan Zi-Yu, Xie Fei. Weakly-supervised contrastive learning framework for few-shot sentiment classification tasks. *Journal of Computer Research and Development*, 2022, **59**(9): 2003−2014
(卢绍帅, 陈龙, 卢光跃, 管子玉, 谢飞. 面向小样本情感分类任务的弱监督对比学习框架. 计算机研究与发展, 2022, **59**(9): 2003−2014)

88 Li Wei-Hua, He Chen, Chen Zhu-Yun, Huang Ru-Yi, Jin Gang.

Unsupervised fault diagnosis of gearbox based on symmetrical contrast learning. *Chinese Journal of Scientific Instrument*, 2022, **43**(3): 131−131
(李巍华, 何琛, 陈祝云, 黄如意, 晋刚. 基于对称式对比学习的齿轮箱无监督故障诊断方法. 仪器仪表学报, 2022, **43**(3): 131−131)

89 Wang X L, Zhang R F, Shen C H, Kong T, Li L. Dense contrastive learning for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 3023−3032

90 Kang Jian, Wang Zhi-Rui, Zhu Ruo-Xin, Sun Xian. Supervised contrastive learning regularized high-resolution synthetic aperture radar building footprint generation. *Journal of Radars*, 2022, **11**(1): 157−167
(康健, 王智睿, 祝若鑫, 孙显. 基于监督对比学习正则化的高分辨率SAR图像建筑物提取方法. 雷达学报, 2022, **11**(1): 157−167)

91 Wang X H, Zhao K, Zhang R X, Ding S H, Wang Y, Shen W. ContrastMask: Contrastive Learning to Segment Every Thing. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 11594−11603

92 He K M, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 2980−2988

93 Zhang L, She Q, Shen Z Y, Wang C H. Inter-intra variant dual representations for self-supervised video recognition. In: Proceedings of 32nd British Machine Vision Conference. UK: BMVA Press, 2021.

94 Bahri D, Jiang H, Tay Y, et al. SCARF: Self-supervised contrastive learning using random feature corruption. In: Proceedings of the 10th International Conference on Learning Representations. Virtual: ICLR, 2022.

95 Kang B, Li Y, Xie S, Yuan Z, Feng J. Exploring Balanced Feature Spaces for Representation Learning. In: Proceedings of the 9th International Conference on Learning Representations. Australia: ICLR, 2021.

96 Li T H, Cao P, Yuan Y, Fan L J, Yang Y Z, Feris R, et al. Targeted supervised contrastive learning for long-tailed recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 6908−6918

97 Zhu J G, Wang Z, Chen J J, Chen Y P P, Jiang Y G. Balanced contrastive learning for long-tailed visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 6898−6907

98 Afham M, Dissanayake I, Dissanayake D, Dharmasiri A, Thilakarathna K, Rodrigo R. CrossPoint: Self-supervised cross-modal contrastive learning for 3D point cloud understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 9892−9902

99 Laskin M, Srinivas A, Abbeel P. CURL: Contrastive unsupervised representations for reinforcement learning. In: Proceedings of the 37th International Conference on Machine Learning (ICML). Vitual: ACM, 2020. 5639−5650

100 Hénaff O, Srinivas A, De FauwJ, Razavi A, Doersch C, Ali Eslami S M, et al. Data-efficient image recognition with contrastive predictive coding. In: Proceedings of the 37th International Conference on Machine Learning (ICML). Vitual: ACM, 2020. 4182−4192

101 Zbontar J, Jing L, Misra I, et al. Barlow twins: Self-supervised learning via redundancy reduction. In: Proceedings of the International Conference on Machine Learning (ICML). Vitual: ACM, 2021: 12310−12320

102 Hua T Y, Wang W X, Xue Z H, Ren S C, Wang Y, Zhao H. On feature decorrelation in self-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 9578−9588

103 Bardes A, Ponce J, Lecun Y. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. In: Proceedings of the 10th International Conference on Learning Representations. Virtual: ICLR, 2022.

104 Bao H, Nagano Y, Nozawa K. Sharp learning bounds for contrastive unsupervised representation learning [Online], available: https://arxiv.org/abs/2110.02501v1, October 6, 2021

105 Wang F, Liu H P. Understanding the behaviour of contrastive loss. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 2495−2504

106 Ren P Z, Xiao Y, Chang X J, Huang P Y, Li Z H, Gupta B B, et al. A survey of deep active learning. *ACM Computing Surveys*, 2022, **54**(9): Article No. 180

107 Sun Qi-Yu, Zhao Chao-Qiang, Tang Yang, Qian Feng. A survey on unsupervised domain adaptation in computer vision tasks. *Scientia Sinica Technologica*, 2022, **52**(1): 26−54
(孙琦钰, 赵超强, 唐漾, 钱锋. 基于无监督域自适应的计算机视觉任务研究进展. 中国科学: 技术科学, 2022, **52**(1): 26−54)

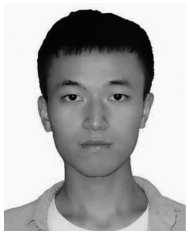
108 Fan Cang-Ning, Liu Peng, Xiao Ting, Zhao Wei, Tang Xiang-Long. A review of deep domain adaptation: General situation and complex situation. *Acta Automatica Sinica*, 2021, **47**(3): 515−548
(范苍宁, 刘鹏, 肖婷, 赵巍, 唐降龙. 深度域适应综述: 一般情况与复杂情况. 自动化学报, 2021, **47**(3): 515−548)

109 Han T D, Xie W D, Zisserman A. Memory-augmented dense predictive coding for video representation learning. In: Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK: Springer, 2020. 312−329

张重生 河南大学计算机与信息工程学院教授. 主要研究方向为长尾学习与不均衡学习, 基于深度学习的汉字识别和古文字计算.
E-mail: cszhang@henu.edu.cn
(**ZHANG Chong-Sheng** Professor at the School of Computer and Information Engineering, Henan University. His research interest covers long-tail learning and imbalanced learning, deep learning based OCR and ancient character computing.)

陈 杰 河南大学计算机与信息工程学院硕士研究生. 主要研究方向为计算机视觉与模式识别.
E-mail: jiechen@henu.edu.cn
(**CHEN Jie** Master student at the School of Computer and Information Engineering, Henan University. Her research interest covers computer vision and pattern recognition.)

李岐龙 河南大学计算机与信息工程学院博士研究生. 主要研究方向为对比学习和文字识别. 本文通信作者.
E-mail: qilonghenu@henu.edu.cn
(**LI Qi-Long** Ph.D. candidate at the School of Computer and Information Engineering, Henan University. His research interest covers contrastive learning and scene text recognition. Corresponding author of this paper.)



邓斌权 河南大学计算机与信息工程学院硕士研究生. 主要研究方向为计算机视觉与模式识别.
E-mail: bqdeng@henu.edu.cn
(**DENG Bin-Quan** Master student at the School of Computer and Information Engineering, Henan University. His research interest covers computer vision and pattern recognition.)

王 杰 河南大学计算机与信息工程学院硕士研究生. 主要研究方向为计算机视觉与模式识别.
E-mail: wangjie@henu.edu.cn
(**WANG Jie** Master student at the School of Computer and Information Engineering, Henan University. His research interest covers computer vision and pattern recognition.)

陈承功 河南大学计算机与信息工程学院硕士研究生. 主要研究方向为计算机视觉与模式识别.
E-mail: cgcheng@henu.edu.cn
(**CHEN Cheng-Gong** Master student at the School of Computer and Information Engineering, Henan University. His research interest covers computer vision and pattern recognition.)