对比式自监督视觉表征学习研究综述

刘相均

（广东技术师范大学，广东 广州 510665）

摘 要 随着深度学习和大规模数据集的发展，自监督学习（SSL）已成为计算机视觉领域中备受关注的研究方向之一。对比学习（CL）是目前最为流行的自监督学习算法之一，CL方法通过最大化相似样本之间的距离和最小化不同样本之间的距离来进行视觉表征学习。本文从视觉对比学习算法的发展历程、算法原理以及不同CL方法的差异展开阐述，以期为基于对比学习的相关应用及研究提供参考。

关键词 深度学习；自监督学习；计算机视觉；对比学习

中图分类号：TP3 **文献标识码**：A **文章编号**：2097-3365(2023)10-0010-03

近些年，自监督学习（Self-supervised Learning，SSL）在图像分类、图像分割、目标检测等计算机视觉任务中表现出了优异的表征学习能力，广受研究者们的关注。而对比学习（Contrastive Learning，CL）作为一种极具代表性的SSL方法，在自然图像领域取得了进一步成功。与全监督学习不同，CL方法不需要标注数据，其本身可以利用大量的无标注数据来学习图像的特征表示。CL的核心思想是通过数据增强构造样本的多样性，利用损失函数在投影的嵌入空间中构造距离近的相似（正）样本对，同时构建距离相对远的不同（负）样本对，从而学习到不同数据样本之间的相对关系表达。本文将从CL的发展历程，CL的基本原理以及不同CL算法的比较三个方面进行阐述，然后进行总结，旨在帮助后续的研究者们能够快速地对CL算法有一个大致了解。

1 对比学习的发展历程

最早的CL方法可以追溯到基于Siamese网络和三元组网络的方法，这些方法主要应用于验证和检索任务。而随着深度学习的发展，特别是基于深度神经网络的无监督预训练方法的兴起，在计算机视觉领域中，对比学习逐渐受到广泛关注，在2018年Wu等人[1]提出了InstDisc（Instance Discrimination）模型，采用了Memory bank来存储编码器计算得到的表征向量，由此开启了基于正负样本对的对比学习研究思路。随后，一系列基于对比学习的预训练方法相继提出，如MoCo[2]、SimCLR[3]、SwAV[4]等。这些方法都通过对数据样本之间的关系进行建模来实现特征学习，并在多个视觉任务上取得了优异的效果。随着Transformer在视觉领域的热门，研究者尝试利用Transformer的自注意力提取方式进行对比学习的研究，在2021年SwAV的作者提出了DINO[5]模型。随后在2022年Peng等人[6]考虑了随机采样这一增强操作对于视图质量的影响，提出了训练预热Grad-CAM定位ROI区域，然后在定位区域内进行中心压制采样的方法，该方法为CL算法的增强视图提供了更加丰富的图像对比信息，使得CL算法更具鲁棒性。

2 对比学习原理

对比学习的学习模式可以抽象为通过编码器－解码器的架构将图像Embedding排列到嵌入空间中，通过在嵌入空间中的关系来判断图像间的相似性。其中最优特征嵌入是通过实例级判别来学习的，如图1所示，该判别试图最大限度地将训练样本的特征分散在单位球面上。假设输入一个批次的图像，首先使用数据增强策略得到同批次大小的增强视图队列，使用主干CNN网络将两部分批次图像编码为特征向量，然后通过投影层计算使得向量均匀分布，最后将其投影到高维空间并通过相似度损失函数进行归一化聚集。图1中示例的三组向量表示组内近似，而三组向量相互之间是非近似状态。

从得到的嵌入表征向量损失计算观察，对比学习旨在通过噪声对比估计（Noise Contrastive Estimation,NCE）$\mathcal{L}_{NCE}=\mathbb{E}_{x,x^+,x^-}\left[-log\frac{e^{f(x)^Tf(x^+)}}{e^{f(x)^Tf(x^+)}+e^{f(x)^Tf(x^-)}}\right]$进行学习比较，其中$x$、$x^+$以及$x^-$表示输入，$x^+$与$x$为正样本对，$x^-$与$x$为负样本对，$f$表示编码器。由于在实际训练过程中可能涉及很多不相似对，而衍生出了InfoNCE损失函数，其具体公式的表达形式为：

$$\mathcal{L}_{Info}=\mathbb{E}_{x,x^+,x^k}\left[-log\frac{e^{f(x)^Tf(x^+)}}{e^{f(x)^Tf(x^+)}+\sum_{k=1}^{K}e^{f(x)^Tf(x^-)}}\right]$$

通过以上的损失函数，CL模型不断地将编码器参数更新，使得CL算法所构建的相似正样本对更加接近，

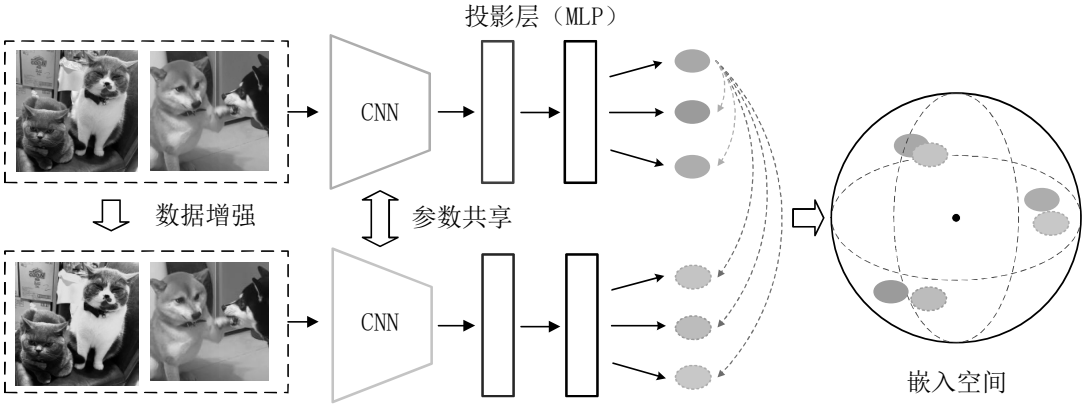

图 1 CL 算法示意图

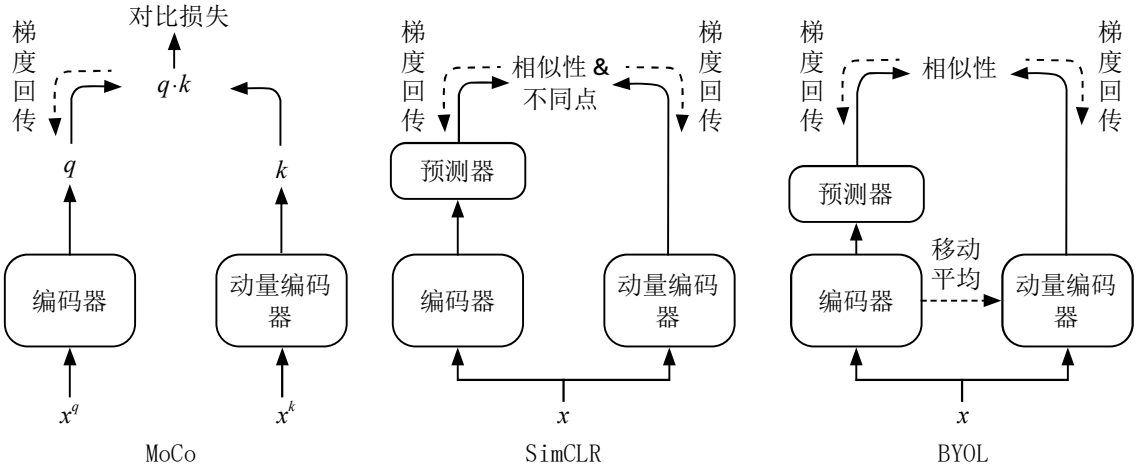

图 2 暹罗网络体系对比算法结构演化

而不同的负样本对更加远离，从而实现无监督条件下对图像的表征学习。

3 不同对比学习算法的分析

目前已经发展出了不少对比学习算法，不同对比学习算法各有优劣，在具体使用时需要根据任务和数据集的特点进行选择和调整。在这我们分析极具代表性的三种暹罗网络结构的模型 MoCo、SimCLR 和 BYOL[7]（如图 2）。MoCo 提出了使用基于动量对比的表征一致性来实现实例区分，利用记忆体（memory bank）形式来进行编码存储。尽管 MoCo 取得了很好的效果，但是其采样方式使得区分正样本对过于简单，需要进一步探索更高效的正样本对采样策略来提升模型的性能。SimCLR 提出使用多种数据增强方法增加输入图像的多样复杂性，不同于 MoCo 模型，SimCLR 直接使用当前批次的记忆体内负样本，但这导致了训练需要使用较大批次的样本容量。此外，SimCLR 还提供了一些技术思路，包括在编码器中添加非线性映射、使用更深的 backbone 网络等。

BYOL 吸收了 MoCo 和 SimCLR 的特点并改进，取得了比二者更好的性能。BYOL 只采用了正样本对，使用随机初始化的网络作为目标编码器，然后将其逐渐替换为经过训练的查询编码器来迭代训练。BYOL 还采用了回归范式的损失计算，使用均方差来度量预测值和目标值之间的差异。通过这些改进，BYOL 取得了很好的表现。BYOL 的设计清晰简洁、方法独特，为对比学习领域的研究提供了一种有效的方法。

$$\mathcal{L}_{\theta}^{\mathrm{BYOL}} \triangleq \left\| \overline{q_{\theta}}(z_{\theta}) - \overline{z\xi}' \right\|_2^2 = 2 - 2 \cdot \frac{\langle q_{\theta}(z_{\theta}), z\xi' \rangle}{\|q_{\theta}(z_{\theta})\|_2 \cdot \|z'\xi\|_2} \quad \#(2-5)$$

一般来说，暹罗结构形式的 CL 网络适用于小规模数据集和特定领域，而 BYOL 和 DINO 等方法则更适合大规模无监督学习和跨领域表示学习。我们在以下列举了一些 CL 算法，见表 1。其中主要比较了不同的 CL 算法的核心思想的转变和随着新技术的改进。

表 1 部分极具代表性的 CL 算法比较

出处 - 年份	模型	创新点
CVPR 2018	InstDisc	提出实例判别任务和使用 Memory bank 进行表征向量的存取。
CVPR 2019	MoCo	把 CL 方法归纳总结为字典查询任务，提出使用队列数据类型保存以及动量更新策略缓慢更新 Memory bank 中的嵌入张量。
ICML 2020	SimCLR	极简的 CL 方法（采用更丰富的数据增强策略、MLP 的非线性投影头、更大 batch size 进行训练）。
CVPR 2020	SwAV	基于聚类的对比学习（采用交换性质的数据增强策略、使用细粒度聚类提高表征学习分布的稳健性、引入使用虚拟样本的损失函数）。
NeurIPS 2020	BYOL	采用 Online 网络和 Target 网络相互预测，舍弃负样本对的构建，只采用正样本对进行损失计算，并采用预测目标动态更新的方法来学习。
CVPR 2020	SimSiam	总结了先前暹罗网络形式的 CL 网络以化繁为简；缩小了编码器和解码器之间的差距，并基于多个视角进行训练，并且去除了负样本对。
CVPR 2021	DINO	利用 Transformer 架构的自注意力机制来学习图像的全局与局部特征。同时引入困难样本挖掘机制，使模型学会自动捕获少样本的信息，同时抑制大量容易分类的样本信息学习。
CVPR 2022	Constrastive Crop	提出了一种适用于 CL 算法的数据增强策略，即中心压制采样方法：可以使得同一图片的不同增强视图暴露更多丰富的图像元素进行对比学习。

4 总结

本文对目前对比学习的发展、原理以及一些 CL 算法的设计创新做出了简要概括，尽管由于不同的体系结构和实现，我们很难详细比较这些方法的性能，但根据这些 CL 算法的设计思路可以总结出对比学习的主要发展的趋势。当前对比学习面临的挑战如下：

1. 数据噪声：CL 方法都要求使用大量的训练数据，但这些数据往往会存在一定的噪声和错误，这会使模型在学习和推断时产生不良影响。如何消除数据噪声和错误，提高数据的质量和可靠性，是对比学习领域需要解决的问题。

2. 训练效率：对比学习方法往往需要进行大量的重复计算和参数更新，对计算资源和存储能力要求很高，这会导致训练和资源成本的增加。如何提高对比学习的训练效率，是未来需要面对的挑战。

3. 多模态学习：大多数对比学习方法只是在单个模态上进行学习，如图像或文本等。但在实际应用中，我们需要对不同模态的数据进行分析。如何将对比学习扩展到不同模态数据之间进行学习，也是未来的一个重要方向。

4. 泛化性能：对比学习方法往往是在特定数据集和任务上进行训练和评估的，但这些方法是否具有比较好的泛化性能，即是否可以在不同数据集和任务上进行推广，仍然需要进一步的探索。

总的来说，对比学习需要解决的挑战包括数据噪声、训练效率、多模态学习和泛化性能等，在未来通过开展更多的研究和实践，对比学习有望在更广泛的场景中发挥重要的作用。

参考文献：

[1] Wu Z,Xiong Y,Yu S X,et al.Unsupervised feature learning via non-parametric instance discrimination[C]//Proceedings of the IEEE conference on computer vision and pattern recognition,2018.

[2] He K,Fan H,Wu Y,et al.Momentum contrast for unsupervised visual representation learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,2020.

[3] Chen T,Kornblith S,Norouzi M,et al.A simple framework for contrastive learning of visual representations[C]//International conference on machine learning. PMLR,2020.

[4] Caron M,Misra I,Mairal J,et al.Unsupervised learning of visual features by contrasting cluster assignments[J].Advances in neural information processing systems,2020(33):9912-9924.

[5] Caron M,Touvron H,Misra I,et al.Emerging properties in self-supervised vision transformers[C]//Proceedings of the IEEE/CVF international conference on computer vision,2021.

[6] Peng X,Wang K,Zhu Z,et al.Crafting better contrastive views for siamese representation learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2022.

[7] Grill J B,Strub F,Altché F,et al.Bootstrap your own latent-a new approach to self-supervised learning[J].Advances in neural information processing systems,2020(33):21271-21284.