软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(3):1125−1139 [doi: 10.13328/j.cnki.jos.007073]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向多模态模型训练的高效样本检索技术*

唐 秀[1], 伍 赛[1,2], 侯 捷[1], 陈 刚[1,2]

[1](浙江大学 软件学院, 浙江 宁波 315103)

[2](浙江大学 计算机科学与技术学院, 浙江 杭州 310027)

通信作者: 伍赛, E-mail: wusai@zju.edu.cn

摘 要: 深度学习中, 多模态模型的训练通常需要大量高质量不同类型的标注数据, 如图像、文本、音频等. 然而, 获取大规模的多模态标注数据是一项具有挑战性和昂贵的任务. 为了解决这一问题, 主动学习作为一种有效的学习范式被广泛应用, 能够通过有针对性地选择最有信息价值的样本进行标注, 从而降低标注成本并提高模型性能. 现有的主动学习方法往往面临着低效的数据扫描和数据位置调整问题, 当索引需要进行大范围的更新时, 会带来巨大的维护代价. 为解决这些问题, 提出了一种面向多模态模型训练的高效样本检索技术 So-CBI. 该方法通过感知模型训练类间边界点, 精确评估样本对模型的价值; 设计了半有序的高效样本索引, 通过结合数据排序信息和部分有序性, 降低了索引维护代价和时间开销. 在多组多模态数据集上通过与传统主动学习训练方法实验对比, 验证了 So-CBI 方法在主动学习下的训练样本检索问题上的有效性.

关键词: 多模态模型训练; 主动学习; 样本检索

中图法分类号: TP18

中文引用格式: 唐秀, 伍赛, 侯捷, 陈刚. 面向多模态模型训练的高效样本检索技术. 软件学报, 2024, 35(3): 1125−1139. http://www.jos.org.cn/1000-9825/7073.htm

英文引用格式: Tang X, Wu S, Hou J, Chen G. Efficient Sample Retrieval Techniques for Multimodal Model Training. Ruan Jian Xue Bao/Journal of Software, 2024, 35(3): 1125−1139 (in Chinese). http://www.jos.org.cn/1000-9825/7073.htm

Efficient Sample Retrieval Techniques for Multimodal Model Training

TANG Xiu[1], WU Sai[1,2], HOU Jie[1], CHEN Gang[1,2]

[1](College of Software, Zhejiang University, Ningbo 315103, China)

[2](College of Computer Science and Technology and College of Software, Zhejiang University, Hangzhou 310027, China)

Abstract: Training multimodal models in deep learning often requires a large amount of high-quality annotated data from diverse modalities such as images, text, and audio. However, acquiring such data in large quantities can be challenging and costly. Active learning has emerged as a powerful paradigm to address this issue by selectively annotating the most informative samples, thereby reducing annotation costs and improving model performance. However, existing active learning methods encounter limitations in terms of inefficient data scanning and costly maintenance when dealing with large-scale updates. To overcome these challenges, this study proposes a novel approach called So-CBI (semi-ordered class boundary index) that efficiently retrieves samples for multimodal model training. So-CBI incorporates inter-class boundary perception and a semi-ordered indexing structure to minimize maintenance costs and enhance retrieval efficiency. Experimental evaluations on various datasets demonstrate the effectiveness of So-CBI in the context of active learning.

Key words: multimodal model training; active learning; sample retrieval

* 基金项目: 国家重点研发计划(2022YFB3304100)
本文由“面向多模态数据的新型数据库技术”专题特约编辑彭智勇教授、高云君教授、李国良教授、许建秋教授推荐.
收稿时间: 2023-07-17; 修改时间: 2023-09-05; 采用时间: 2023-10-24; jos 在线出版时间: 2023-11-08
CNKI 网络首发时间: 2023-12-26

近年来, 深度神经网络在各个领域取得了令人兴奋的进展, 并成为许多任务的核心技术. 然而, 这些深度神经网络的训练通常需要大量高质量标注数据[1], 以达到最佳性能. 以 ChatGPT 为例, 其完整训练需要数千万级别的语料库. 在多模态数据的背景下, 例如包含图像、文本、音频等多种类型信息的数据, 获取大规模的高质量多模态标注数据是一项具有挑战性的任务[2]. 尽管已经存在一些大型多模态数据集, 但这些数据集的质量和数量仍然无法满足深度神经网络模型训练的需求. 主动学习作为一种能够通过有效地选择和标注最有价值样本的技术, 已经被广泛应用于多模态模型训练中降低标注成本和提高模型性能.

主动学习(active learning)作为一种有效的学习范式, 能够通过有针对性地选择最有信息价值的样本进行标注, 以减少标注成本并提高模型性能[3]. 在多模态数据的背景下, 例如包含图像、文本、音频等多种类型信息的数据, 多模态模型训练中的主动学习方法备受关注. 在主动学习中, 样本选择的关键是能够快速有效地找到对当前模型训练最有效的样本. 近年来, 研究者们提出了许多适用于多模态主动学习的方法和技术, 以改进样本选择的效率和性能.

Xie 等人[4]提出了一种通用且高效的主动学习方法, 该方法利用基于不确定性的样本选择策略来选择最具信息量的样本, 适用于选择在多模态数据中最具信息量的样本. 作者通过研究不同的标注器选择策略和主动学习策略, 并在多个数据集上进行了实验, 证明了该方法的高效性和性能优势. Bengar 等人[5]专注于图像分类任务中的类别平衡问题, 适用于通过避免多模态训练数据类间差异来提升多模态模型性能. 作者针对不平衡数据集提出了一种类别平衡的主动学习方法, 该方法能够有效地选择既能够提高整体模型性能, 又能够保持类别平衡的样本进行标注. 实验结果表明, 该方法在处理不平衡数据集时具有显著的优势. Emam 等人[6]关注的是在大规模图像数据集(如 ImageNet)上进行主动学习的挑战, 适用于对多模态数据进行衡量信息密度来选择更有价值的样本. 作者提出了一种基于扩展的信息密度近似的样本选择算法, 以应对海量图像数据的标注问题. 该算法有效地降低了标注数据的需求, 同时保持了良好的模型性能.

现有的多模态模型研究已经取得了一定的成果, 但在处理大规模数据集时, 传统的样本选择方法往往面临着一些问题: (1) 在传统的多模态模型训练下, 存在很多相似数据的重复训练, 往往难以有效地检索到对模型更有用的样本; (2) 传统的主动学习样本选择方法往往面临着低效的数据扫描和数据位置调整问题; (3) 当多模态数据索引需要进行大范围的更新时, 涉及数据的移动和重新排序, 带来了显著的维护开销.

针对上述问题, 本文提出了一种面向多模态模型训练的高效样本检索技术 So-CBI (semi-ordered class boundary index). 该方法更关注当前多模态模型训练的类边界数据, 通过建立一类特殊的数据索引来支持训练样本的高效检索模型框架图. 首先, 本文提出一种感知模型训练多模态类间边界点的方法, 用于准确评估样本对模型的价值. 通过分析样本之间的相似性和差异性, 本文能够更精确地选择对当前多模态模型训练最有效的样本, 避免了对相似样本的重复训练, 从而提高了训练效率和模型性能. 其次, 通过设计高效的数据索引和检索算法, 能够快速定位和访问对模型训练最有价值的样本, 避免了对整个样本集的遍历和排序操作, 从而显著提高了样本选择的效率和速度. 最后, 为了降低索引的维护代价, 提出了一种基于批更新优化的数据半有序索引的设计. 该索引结合了数据的排序信息和部分有序性, 使得在索引需要进行大范围更新时, 只需对部分数据进行调整, 避免了对整个索引的重建和排序操作, 从而降低了维护代价和时间开销.

通过多组对比实验, 验证了本文所提方法面向多模态模型训练的训练样本检索问题上的有效性. 本文根据选择具有最大信息价值的样本进行标注, 发现仅使用 50%的样本进行主动学习即可达到与其他方法使用全部样本更高的模型准确率, 大幅减少标注样本的数量, 从而降低了标注成本. 其次, 本文提出的数据索引可以有效地存储和检索未训练样本, 能够在更短的时间内获得对当前模型训练最有效的样本, 进一步提高了主动学习的效率. 通过数据批更新和半有序索引的结合应用, 我们能够快速地找到对当前模型训练最有效的样本, 从而加速深度学习模型的训练过程.

本文第 1 节介绍主动学习样本检索的相关方法和研究现状. 第 2 节介绍本文构建的模型训练边界点感知技术. 第 3 节介绍本文构建的基于半有序索引的高效样本检索技术. 第 4 节通过对比实验验证了所提模型的有效性. 最后总结全文.

1 相关工作

1.1 主动学习

主动学习算法是通过优先标注最有价值样本, 从而在有限标注预算下实现最高的模型精度的技术, 其中, 查询策略是影响训练效果的关键因素. 国内外学术界对主动学习进行了广泛的研究. 早期的主动学习方法主要集中在基于不确定度的采样策略上, 其中一种常见的方法是基于不确定度最大化(uncertainty sampling)[7], 它通过评估样本的不确定度来选择最具挑战性的样本进行标注, 例如, 选择使得模型预测概率最接近 0.5 的样本. 此外, 基于边界的采样策略(margin sampling)和基于熵的采样策略(entropy sampling)[8]也是常见的不确定度采样方法. 随着研究的深入, 人们意识到, 仅仅使用不确定度采样方法可能无法充分利用标注样本的信息, 因此提出了一系列基于多样性的主动学习方法. 这些方法试图选择与已标注样本最不相似的样本, 以便提供更多多样化的信息. 多样性采样方法[9]包括最大化最小距离采样(maximum minimization sampling)和最小化最大距离采样(minimum maximization sampling)等.

近年来, 深度学习的兴起为主动学习带来了新的挑战和机遇. 研究者开始探索如何将主动学习与深度学习相结合. 一种常见的方法是: 基于梯度的采样策略(gradient-based sampling), 通过评估样本对模型参数的梯度来选择最具信息量的样本. 此外, 还有一些基于生成对抗网络(GANs)的主动学习方法, 通过利用生成器和判别器之间的竞争来选择最具挑战性的样本. Gal 等人证明了带有 Dropout 结构的神经网络是深度高斯过程(deep Gaussian process)的近似, 并用模型的 Dropout 近似估计模型在样本上的不确定程度. 基于样本多样性和分布的查询策略, 其优先选择最具多样性、最具数据总体分布的样本. Sener 等人将主动学习任务定义为核心集合(core-set)等选择问题, 每次查询所得到的样本, 是最能代表整体样本的核心集合. Gissin 等人[10]将主动学习任务视为一个二分类问题, 其运行过程中保证标注数据符合真实场景的数据分布. Bengar 等人[11]提出了一个考虑类间数据平衡的主动学习通用算法, 通过智能地选择样本来减少标注成本, 并提高模型性能.

1.2 样本检索

样本检索优化方法是指在大规模样本库中高效检索目标样本的技术, 该领域的研究旨在解决在海量数据中快速找到相关样本的问题. 在国内外学术界, 样本检索优化方法得到了广泛的研究. 一种常见的样本检索优化方法是基于特征向量的相似度度量, 在这种方法中, 样本通常被表示为高维特征向量, 例如图像、文本或音频的特征表示, 然后, 利用特定的相似度度量方法计算样本之间的距离或相似度. 传统的相似度度量方法包括欧氏距离、余弦相似度等. 然而, 在大规模数据集上计算样本之间的相似度是非常耗时的. 因此, 研究者们提出了一些高效的相似度度量方法[12], 如局部敏感哈希(LSH)和 k 最近邻(k-NN)等. 这些方法通过构建索引结构或哈希函数, 将样本映射到特定的空间, 以加速相似度计算和样本检索过程.

随着深度学习的兴起, 基于深度特征的样本检索优化方法[13]成为研究热点. 深度学习模型能够学习更加高级的特征表达, 从而提供更准确和语义丰富的样本表示. 在这方面, 一种常见的方法是使用预训练的卷积神经网络(CNN)提取图像特征[14], 然后通过计算特征之间的距离来进行样本检索. 此外, 还有一些工作提出了使用基于注意力机制的深度特征来加强样本表示的方法[14]. 通过引入注意力机制, 模型能够更加关注与目标样本最相关的特征信息, 从而提高检索的准确性.

为了进一步提高样本检索的效率, 一些研究者提出了索引结构优化方法. 索引结构是一种用于组织和管理样本库的数据结构, 可以加速样本检索过程. 常见的索引结构包括球树(ball tree)[15]、kd 树(kd-tree)[16]等. 这些索引结构能够将样本按照一定的规则进行划分和组织, 从而提高检索的效率. 此外, 基于哈希的索引结构[17]如哈希图(hashing graph)和哈希表(hash table)等也被广泛应用于样本检索中. 通过将样本映射到二进制编码或哈希码, 哈希索引可以快速过滤出候选样本减少计算量, 对于解决大规模样本库中的快速检索问题具有重要意义.

2 问题描述与技术框架

2.1 问题描述

给定训练多模态数据集 $D=\{(x_i,y_i)\}_{i=1}^{N}$，其中，$x_i$ 是输入样本，y_i 是对应的标签，以及一个模型 $f(x;\theta)$，其中，θ 是模型参数. 本文的目标是设计一个基于主动学习的高效样本检索技术，解决以下问题.

(1) 重复训练问题：引入一个信息量度量函数 $I(x)$来衡量样本 x 的信息量，以及一个模型贡献度度量函数 $C(x)$来衡量样本 x 对模型的贡献度. 目标是最大化选择样本集合 $D_{selected}\subset D$ 的总信息量和总贡献度，即

$$\max_{D_{selected}}\sum_{x\in D_{selected}}I(x)+\sum_{x\in D_{selected}}C(x) \tag{1}$$

(2) 低效的数据扫描和位置调整问题：对于给定的样本选择方法 $M(D)$，其中，$M(\cdot)$是一个函数，用于选择样本并调整数据位置，本文的目标是最小化样本选择和数据位置调整的时间复杂度. 定义优化目标为

$$\min_{P} Time(M(D)) \tag{2}$$

其中，$Time(M(D))$表示样本选择方法 $M(D)$的时间复杂度. 需要设计一个高效的样本选择算法，以降低时间复杂度，并保持样本选择的准确性.

(3) 数据索引的大范围更新维护开销问题：对于给定的数据索引维护策略 $P(D)$，其中，$P(\cdot)$是一个函数，用于在数据集变化时更新数据索引，目标是最小化数据索引的更新时间复杂度和维护开销. 定义优化目标为

$$\min_{P} Time(P(D))+Space(P(D)) \tag{3}$$

其中，$Time(M(D))$表示数据索引维护策略 $P(D)$的时间复杂度，$Space(P(D))$表示维护过程所需的额外空间复杂度. 需要设计一个高效的数据索引维护策略，以降低时间复杂度和维护开销，并保持数据索引的准确性和一致性.

2.2 高效样本检索技术框架

本文提出了一个面向多模态模型训练的高效样本检索技术框架，该框架主要包括 3 个关键组成部分，分别是基于预训练的样本表征模型、基于边界感知的样本预筛选模型和基于半有序索引的样本检索优化模型，如图 1 所示.

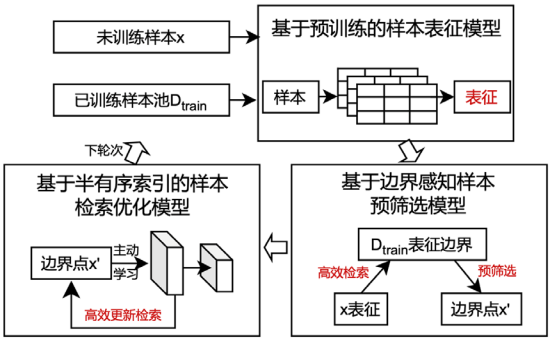


图 1 基于主动学习的高效样本检索技术框架图

(1) 基于预训练的样本表征模型：该模型利用预训练技术，通过对大规模多模态数据集进行离线训练，学习得到高质量的样本表征表示. 通过将样本映射到一个低维向量空间，该模型能够捕捉到样本之间的相似性和关联性，为后续的样本检索提供基础.

(2) 基于边界感知的样本预筛选模型: 在样本表征的基础上, 该模型通过边界感知方法对样本进行预筛选. 通过计算样本与边界之间的距离或相关性度量, 该模型能够快速地识别出具有较高价值的样本, 即与模型边界最为接近的样本. 这种预筛选机制能够减少样本选择的搜索空间, 提高后续样本检索的效率.

(3) 基于半有序索引的样本检索优化模型: 为了进一步提升样本检索的效率, 该模型设计了基于半有序索引的优化方案. 该方案参考了半有序跳表结构, 通过构建高效的索引结构, 实现对样本的快速检索. 该索引结构充分利用样本之间的排序关系, 使得样本的搜索和排序操作变得高效, 从而降低了检索过程的计算和存储开销.

综上所述, 本文提出的高效样本检索技术框架通过基于预训练的样本表征模型、基于边界感知的样本预筛选模型和基于半有序索引的样本检索优化模型的组合, 能够有效地解决传统主动学习方法中的问题, 提升样本选择和多模态数据索引维护的效率, 为主动学习任务提供了一个高效可行的解决方案.

3 基于主动学习的高效样本检索技术

3.1 基于预训练的样本表征模型

为了更好地度量样本之间的相似性和差异性, 本文基于自监督预训练模型来进行学习样本表征, 如图 2 所示. 样本表征模型的设计基于 DINO[18]模型和对自监督学习的理解, 它使用无标签数据来学习有用的特征表示. 通过将输入图像分成多个小块, 并在这些小块之间进行交互来学习特征表示.

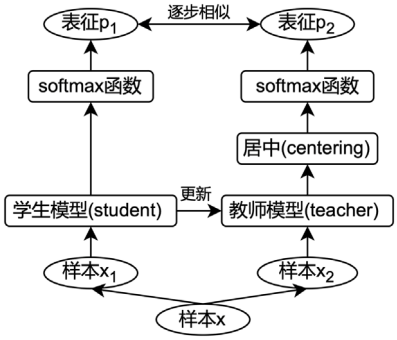


图 2 基于预训练的样本表征模型图

为了简化自监督训练的繁琐过程, 本文使用一种名为“教师-学生(teacher-student)”框架的方法. 其中, 教师(teacher)模型生成目标表示, 而学生(student)模型生成与之相似的表示. 通过最小化这两个表示之间的差异来训练学生网络. 此外, 模型使用了动量编码器(momentum encoder)和多重裁剪训练(multi-crop training)等技巧来提高性能.

DINO 模型中的教师模型使用动量编码器模型来生成目标表示, 它将输入图像编码为归一化的向量:

$$h_t(x)=\frac{f_{\theta_t}(x)}{\| f_{\theta_t}(x)\|_2} \tag{4}$$

其中, f_{θ_t} 是由参数θ_t控制动量编码器函数. 然后, 在每个小块上, 学生模型使用另一个编码器模型 f_{θ_t} 来生成与 $h_t(x)$相似的向量 $h_s(x)$:

$$h_s(x)=\frac{f_{\theta_s}(x)}{\| f_{\theta_s}(x)\|_2} \tag{5}$$

教师模型的参数被固定, 用于生成特征向量. 学生模型使用教师模型生成的特征向量作为输入, 以帮助学生模型更好地学习到教师模型的特征. 学生模型使用交叉熵损失函数最小化学生模型特征 $h_s(x)$和教师模型

特征 $h_t(x)$之间的差异，使用标准的反向传播算法来更新模型的参数:

$$\mathcal{L}=-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}p_t(x_i)_j\log p_s(x_i)_j \tag{6}$$

其中，N 是训练样本的数量，C 是类别数，$p_t(x)$和 $p_s(x)$分别是教师模型和学生模型在输入 x 上的预测概率分布. 模型使用居中(centering)技术的一些技巧来提升模型的性能，通过减去每个小块的目标表示在当前批次中的均值来减少特征表示之间的冗余:

$$\hat{h}_t(x)=h_t(x)-\mu \tag{7}$$

这样，可以使得教师网络输出的特征表示更加具有代表性. 此外，DINO 还使用了指数移动平均(EMA)来平滑教师网络的参数. 具体地，在每个训练步骤中，模型更新教师网络参数θ_t和一个 EMA 变量$\theta_{t,ema}$:

$$\theta_{t,ema}=\alpha\theta_{t,ema}+(1-\alpha)\theta_t \tag{8}$$

其中，α是一个控制 EMA 衰减率的超参数. 基于 DINO 自监督预训练模型的样本表征方法能够学习到有用的特征表示，为后续的多模态数据选择和主动学习任务提供了重要的基础. 通过充分利用无标签数据进行预训练，该方法能够提高样本的表征能力，进一步优化模型的性能和泛化能力.

3.2 基于边界感知的样本预筛选模型

本文提出了一种基于边界感知的样本预筛选模型，用于选择对深度学习模型训练更有价值的样本，如图 3 所示. 该模型利用已训练样本和未训练样本之间的距离，来对未训练样本进行评分判断更有价值的样本进行训练. 同时，模型使用 HNSW (hierarchical navigable small world)[19]算法来加速检索已训练样本在未训练样本池中的 K 个最近邻，并更新其分数.

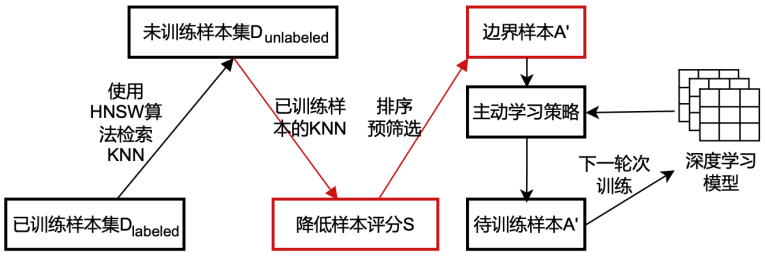


图 3 基于边界感知的样本预筛选模型图

首先，对于每个已标记样本 x，在未标记样本集$\mathcal{D}_{unlabeled}$中检索出其 K 个最近邻样本集$\mathcal{N}(x)$:

$$N(x)=\{knn\_search(f(D_{unlabeled}),f(x),K),x\in D_{unlabeled}\} \tag{9}$$

其中，$f(\cdot)$是第 3.1 节中基于预训练的样本表征模型函数. 同时，我们使用 HNSW 算法进行最近邻检索加速. HNSW 算法是一种高效的最近邻检索算法，通过构建多层的小世界网络来加速近邻搜索过程. 采用 HNSW 算法进行最近邻检索，能够快速地找到未训练样本的最近邻，大大减少了计算复杂度，提高了模型训练的效率. 通过对不同最近邻检索算法实验对比，结合固定召回率下的检索速率与空间占比上的对比分析，综合分析后使用HNSW算法作为本文的最近邻检索算法，具体分析参考后文第4.3节实验方法部分. 然后，我们记录未训练样本 x'被选中为最近邻的频次，每个未训练样本的选中频次初始化为 0: $\{\{S'_x\}_{init}=0,x'\in D_{unlabeled}\}$; 随后，每轮训练时迭代更新:

$$S'_{x_i}=S'_{x_i}+1,\ \{x'_i\in N(x_i),x_i\in D_{labeled}\}_{i=0}^{i=len(D_{labeled})} \tag{10}$$

通过这种方式计算所有未训练样本的分数，我们可以得到未训练样本集$\mathcal{D}_{unlabeled}$的得分表示集合 $D_{score}=\{S'_x,x'\in D_{unlabeled}\}$. 根据得分表示$\mathcal{D}_{score}$集合对未训练样本进行排序，并根据排序预筛选出边界样本，模型训练将更加关注那些能够提供更多信息、更具挑战性的样本，这有助于提高模型的鲁棒性和泛化能力，避免过度拟合和欠拟合的问题.

最后, 选择较小频次的样本作为有价值的样本, 加入下一轮次的模型训练中. 基于样本之间的距离判断, 该模型能够自适应地选择有价值的边界样本. 这意味着模型可以根据当前训练状态和多模态数据分布的特点, 动态地调整样本选择的策略, 以获得更好的训练效果.

综上所述, 基于边界感知的样本预筛选模型利用 HNSW 算法进行最近邻检索, 并通过记录选中为最近邻的频次来判断样本的边界程度, 从而选择对深度学习模型训练更有价值的样本. 该模型在深度学习模型训练中具有高效、边界感知、自适应和高质量样本选择等优点, 为提升模型性能和泛化能力提供了有力支持.

3.3 基于半有序索引的样本检索优化模型

在主动学习过程中, 随着新样本的标注和加入, 未标记样本在索引中的位置会在短时间内进行大范围的更新, 这会导致巨大的维护代价. 在当前算法下, 每个已标记样本在一个单独的线程内, 更新其最近邻 K 个样本的得分. 传统的索引在这种场景下, 不仅在更新时具有较高的维护代价, 同时不同线程往往会多次更新某个特定样本的位置, 这带来了显著的计算和存储开销, 严重影响了系统的效率和可扩展性. 为了解决这个问题, 本文设计了基于半有序索引的样本检索优化模型, 该模型对传统方法进行了两方面的优化, 如图 4 所示.

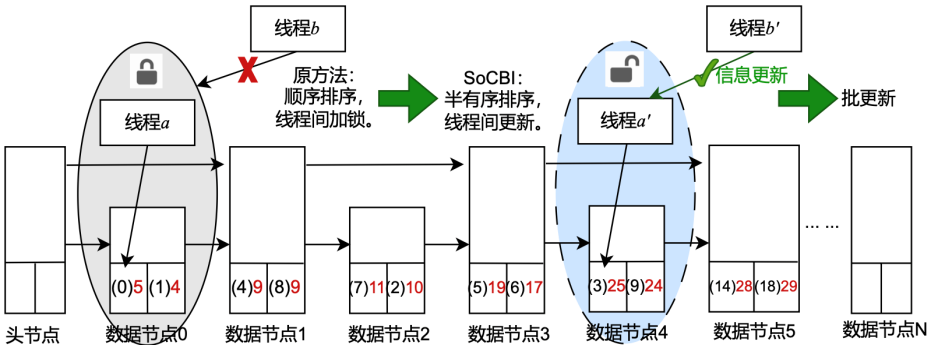


图 4 基于半有序索引的样本检索优化模型图

首先, So-CBI 设计半有序索引对数据进行半有序存储, 来优化数据更新与检索效果. 半有序索引是一种多层链表结构[20], 每个数据节点可以存放多个数据项, 同时, 每个数据节点内的数据是无序存放的, 而数据节点之间是有序的. 半有序索引的主要目标是通过半有序性质减少索引的更新代价, 同时保持高效的检索性能. 然后, So-CBI 利用批更新的思想设计线程间更新机制, 解决线程间更新冲突加锁后导致数据更新变慢的问题. 具体来说, 模型存储用于主动学习的样本更新价值表, 用于记录当前代表各样本价值的分数, 以及更新状态. 其可选的字段有: 无、正在更新. 其中, “无”意味着当前样本可以被任何线程更新, “正在更新”意味着当前已有线程在执行该样本的更新操作.

通过记录样本更新价值表, 记录当前样本 ID 以及样本分数的对应关系. 在该索引框架下, 使用分数作为跳表索引的主键, 从而能够在常数级别的时间复杂度下, 找到得分 Top N 的未标记样本. 此外, 样本价值更新表的每一行还记录了当前样本的状态, 当状态为“无”时, 负责更新该样本的线程, 原子地将状态变为“正在被更新”, 然后执行更新任务; 当状态为“正在被更新”时, 当前线程需要原子的更新样本分数, 然后将更新任务, 传递给正在执行更新的线程. 线程完成其更新任务时, 需要检查当前的样本分数, 与其任务开始时分数的差异, 若存在差异, 则需要继续完成更新操作. 以下是该并发同步协议的完整流程(见算法 1).

综上所述, 基于半有序索引的样本检索优化模型通过引入半有序的索引结构和批更新概念, 解决了在主动学习过程中样本位置大范围更新带来的维护代价大的问题. 该模型利用半有序性质减少了索引的更新代价, 并保持了高效的检索性能. 通过记录样本更新价值表, 记录样本的分数和更新状态, 可以在常数级别的时间复杂度下找到得分 Top N 的未标记样本. 同时, 利用原子操作和任务传递机制, 确保多个线程能够协调进行样本更新操作, 避免重复更新和计算开销.

算法 1. 样本索引更新算法流程.

获得该样本对应的样本更新价值表条目.

(1) 如果当前样本正在被更新, 则原子更新条目中的分数字段.

(2) 如果当前样本没有被更新, 则:

a) 设置当前条目状态为“正在被更新”.

b) 记录当前条目中的分数.

c) 调用索引暴露出的更新接口, 按照记录的分数, 找到样本所要插入的位置.

d) 判断当前样本分数和位置是否匹配: 如果匹配, 则直接插入, 然后更新状态为“无”; 否则, 执行 b 步骤.

4 实验分析

4.1 实验数据

4.1.1 多模态分类数据

本文在 3 个多模态数据集上评估我们所提出的方法对多模态模型训练的效果. 这些多模态数据集涵盖图像、文本、数字和分类输入的数据. 表 1 提供了数据集的来源、统计数据和模态标识的摘要.

表 1 多模态匹配数据集的基本信息

数据集	训练集	测试集	衡量标准	图像	文字	表格
PetFinder [21]	11 994	2 999	Kappa 系数	√	√	√
Hateful Memes [22]	7 134	1 784	Accuracy	√	√	–
Visual Genome [23]	108 077	2 000	Accuracy	√	√	–

PetFinder 数据集是一个用于预测宠物的受欢迎程度的多模态数据集, 包含图像、文字和表格. PetFinder 根据宠物的个人资料的照片预测该宠物的受欢迎程度, 使用可爱指数来排名宠物照片. Hateful Memes 数据集是一个用于仇恨言论检测的多模态数据集, 包含图像和文本. 该数据集包含了 Facebook AI 创建的 10 000 多个新的多模态实例. Visual Genome 数据集相较于前两个数据集来说规模更大, 是斯坦福大学于 2016 年提出的语言与视觉数据集, 标注了大量在图片中的对象和关系以及关于图片的问答对.

4.1.2 图像分类数据

同时, 为了评估我们的方法在其他模型的适用性, 本文在 4 个图像分类基准数据集评估我们所提出的方法对其他模型训练的效果. 表 2 提供了数据集的一些基本信息. 实验中, 初始标记集 DL 包含了 1 000 个训练集样本, 这些数据是从所有类别中均匀随机选择的. 在每个循环中, 均从头开始训练基础模型, 或者在 Tiny ImageNet 数据集中, 从预训练的 ImageNet 模型开始. 实验进行 c 个循环的模型训练, 直到耗尽预算 B. 每个循环的预算为原始数据集的 5%.

表 2 图像分类数据集的基本信息

数据集	训练集	测试集	衡量标准
CIFAR-10 [4]	50k	10k	准确度(Acc)
CIFAR-100 [5]	50k	10k	准确度(Acc)
Tiny ImageNet [5]	100k	10k	准确度(Acc)
SVHN [24]	73k	26k	准确度(Acc)

我们在 CIFAR-10、CIFAR-100、Tiny ImageNet、SVHN 数据集上对本文方法进行评估. CIFAR-10 和 CIFAR-100 数据集有 50k 张图像用于训练, 10k 张用于测试. CIFAR-10 和 CIFAR-100 数据集分别有 10 个和 100 个对象类别, 图像大小为 32×32. Tiny ImageNet 数据集有 90k 张图像用于训练, 10k 张用于测试, 共包含 200 个对象类别, 图像大小为 64×64. SVHN 数据集包含了超过 60 000 张训练图像, 其内容是 Google 街景视图的数码照片. 该数据集的任务是对图像上的数字进行分类.

4.1.3 倾斜版本数据

为了评估本文方法在倾斜数据集的有效性, 实验构造图像分类数据集的倾斜版本进行全面的考量. 为了验证方法在不平衡数据集上的效果, 本文使用数据分类数据集的类别倾斜版本. 同样, 保留这两个数据集中剩余的样本用于初始标记集. 参考倾斜数据集构造工作[4]中所述, 通过随机删除训练样本来创建长尾数据集. 具体而言, 每个类别的样本数量从原来的 n_v 个样本减少到 $n_v \cdot IF$ 个样本, 其中, n_v 是类别 y 中的原始训练样本数量, 不平衡因子(imbalance factor) $IF \in (0,1)$. 为了构建长尾数据集, 我们将 IF 应用于一半的类别, 并使用 $IF \in \{0.1,0.3\}$.

4.2 评价指标及基准模型

为了衡量所选择样本的平衡性, 本文使用 $L1$ 分数通过计算样本分布和均匀分布之间的 $L1$ 距离来衡量. 为了得到一个从 0 到 1 的度量, 对 $L1$ 进行了归一化. 本文通过测试集上的准确率来评估性能, 所有实验的结果都是在 5 次运行中平均得到的. 对于每种方法, 我们绘制了所有运行的平均性能, 并使用垂直条表示标准差. 本文采用两种不同的维度进行度量方法有效性.

(1) 模型准确率增益: 本文通过主动学习方法对模型进行改进后, 相较于基准模型在准确率方面所获得的提升. 准确率是指模型在预测时正确分类的样本比例, 因此, 准确率增益表示改进后的模型在正确分类样本方面相较于基准模型的改进程度.

(2) 样本检索效率: 本文定义每轮次样本检索时间 T, $T=T_{update}+T_{sort}+T_{search}$, 其中: T_{update} 为根据每轮次模型训练状态, 使用主动学习策略后对样本进行更新所需要的时间; T_{sort} 为样本状态更新后对样本进行排序所需要的时间; T_{search} 为样本排序后检索样本来进行下一轮标注训练的时间. 即: 在每轮次更新主动学习样本时, 检索下轮次所需样本所花费的时间, 作为主动学习方法的样本检索效率衡量.

在本文中, 我们将方法在最低置信度算法(least confidence)和 K 中心贪心算法(K center greedy sampling, KCenter)两种信息丰富和代表性方法进行实验.

(1) 最低置信度算法: 选择最不确定的样本进行标注, 以提高模型的性能. 该算法通过计算模型对每个未标记样本的预测概率中最低的值, 来衡量样本的不确定性.

(2) K 中心贪心算法: 一种启发式算法, 该策略的基本思想是选择一组最能够代表整个数据集的样本.

4.3 实验方法

在本文中, 我们对 So-CBI 进行了详细的实验评估, 采用了最低置信度算法和 K 中心贪心算法作为主动学习策略, 以验证本文方法的有效性. 对于多模态数据匹配任务来说, 实验使用 CLIP 模型[25]提取多模态数据集的样本特征, 然后使用 LogisticRegression 进行多模态数据匹配, 验证 So-CBI 在 Hateful Memes 数据集和 PetFinder 数据集的正常与倾斜版本上的性能. 对于图像分类任务来说, 实验使用基于 PyTorch 实现的 ResNet18 模型[26], 验证其在 CIFAR-10、CIFAR-100、Tiny ImageNet 和 SVHN 数据集的正常与倾斜版本上的性能. 在每轮样本筛选后重新训练模型, 基于当前标记样本池进行训练.

针对最近邻检索算法进行固定召回率下的检索速率与空间占比上的对比分析, 对比球树、kd 树、局部敏感哈希、HNSW 算法的性能, 结果见表 3.

表 3 最近邻检索算法在检索速率和空间开销上的分析对比

索引结构	检索速率(请求数/s)		空间开销(MB)	
	Hateful Memes	PetFinder	Hateful Memes	PetFinder
球树	93	102	2 262.06	3 123.38
kd 树	85	96	2 164.31	3 046.91
LSH	327	375	33.20	46.74
HNSW	**4 639**	**5 374**	**41.71**	**56.20**

球树适用于高维空间和不均匀数据, 但构建耗时; kd 树适用于低维静态数据, 但在高维空间性能下降; 局部敏感哈希适用于高维空间且可处理非欧氏距离, 但需要参数调整; HNSW 适用于高维大规模数据, 通过构

建图实现搜索, 性能平衡可调. 针对多模态数据集(Hateful Memes 数据集和 PetFinder 数据集), 我们对这几种索引结果进行检索速率与空间开销的权衡对比分析, 综合考虑下选择检索效率最高且与空间开销最低方法(LSH)相近的 HNSW 算法作为本文的最近邻检索的索引结构.

在我们的实验设置中, 固定所有随机种子, 并将批量大小设置为 256, 初始样本池大小为 1 000, 每次查询的样本数量为 128. 我们使用学习率为 0.001 和默认参数的 Adam 优化器来训练模型. 对于每个数据集, 模型训练 40 个周期. 在训练过程中, 我们应用了通用的数据增强方法, 包括图像随机裁剪、水平翻转以及使用均值和标准差进行归一化. 这些增强方法有助于提高模型的鲁棒性和泛化能力. 通过实验评估, 我们得出了一系列结果和结论. 我们比较了不同主动学习策略的性能, 并分析了它们在不同数据集和倾斜程度下的表现. 我们考察了模型的准确率、所需标记样本数量、每轮次样本检索效率等指标, 以评估 So-CBI 的有效性和优越性. 本文所使用的实验平台为 Intel(R) Xeon(R) Gold 6248R, 主频为 3.00 GHz.

4.4 实验结果与分析

为了评估基于主动学习的高效样本检索方法 So-CBI 的有效性, 我们研究了以下 3 个问题.

(1) So-CBI 是否能够帮助多模态模型训练, 在倾斜数据集上获得更好的模型训练准确率?

(2) So-CBI 是否使多模态模型训练需要更少的训练样本, 从而减少模型所需训练时间?

(3) So-CBI 是否在多模态模型训练中, 比其他主动学习方法拥有更高效的样本检索效率?

4.4.1 模型训练准确率对比

(1) 多模态模型训练

我们将 So-CBI 使用到多模态模型训练中, 准确度结果如图 5 所示.

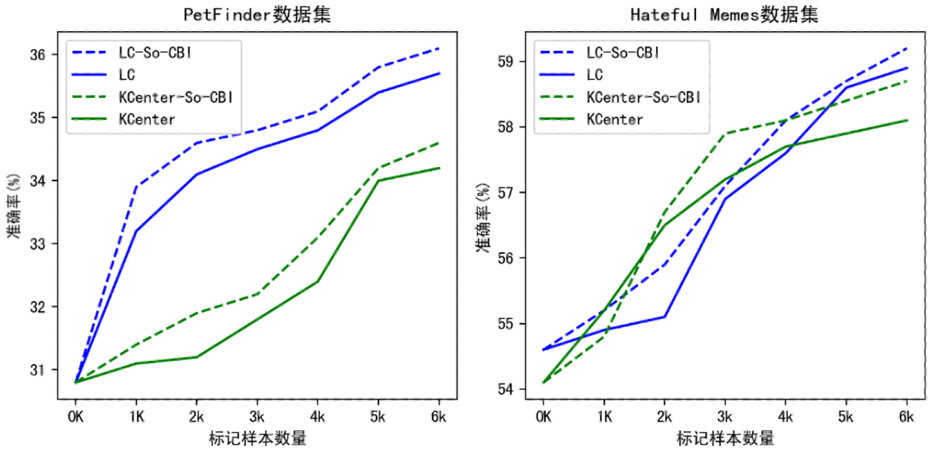


图 5 仅使用 50%数据在多模态数据集上的性能评估

首先, 我们观察 PetFinder 数据集的效果: 当主动学习算法使用最低置信度算法(LC)时, So-CBI 方法在图像分类准确度方面表现出更为优秀的效果, 达到 0.5%的准确度提升; 当主动学习算法使用 K 中心贪心算法(KCenter)主动学习时, So-CBI 方法也表现出比原始方法更高的准确率, 达到 0.9%的准确度提升.

然后, 我们观察 Hateful Memes 数据集的效果: 当主动学习算法使用最低置信度算法(LC)时, So-CBI 方法在图像分类准确度方面表现出更为优秀的效果, 达到 0.7%的准确度提升; 当主动学习算法使用 K 中心贪心算法(KCenter)主动学习时, So-CBI 方法也表现出比原始方法更高的准确率, 达到 0.8%的准确度提升.

值得注意的是: 多模态数据集是极具挑战性的数据集, 在此类数据集上进行主动学习, 鲜有效果提升的工作. 可以看出, So-CBI 在多模态数据上体现出了极高的适用性.

(2) 图像分类模型训练

Tiny ImageNet[27]是深度学习领域的具有挑战性的大规模数据集, 我们使用它来评估我们方法在大规模数据上的可扩展性, 我们展示了最低置信度算法(LC)、K 中心贪心算法(KCenter)主动学习策略, 在应用我们的方法前后的性能对比. 图 6 展示了我们的具体实验效果.

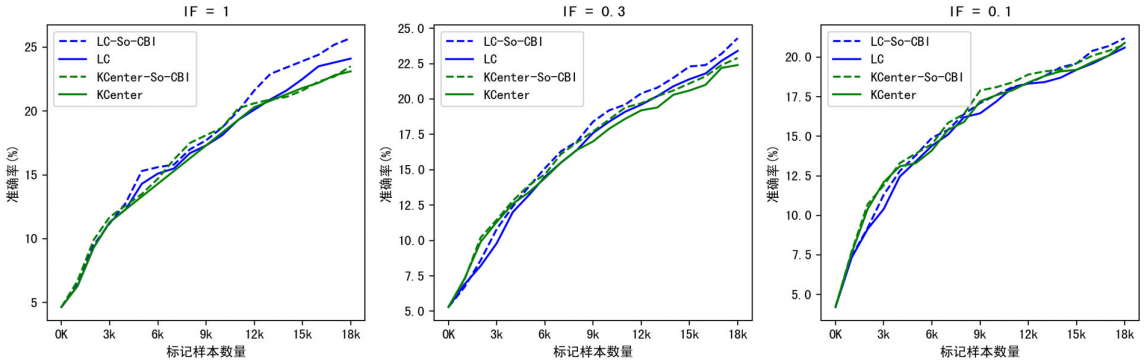


图 6 仅使用 50%数据在不同倾斜程度的 Tiny ImageNet 数据集上的性能评估

首先, 我们观察主动学习算法使用最低置信度算法(LC)时的效果: 在倾斜率为 1 时, 即数据集不变时, So-CBI 方法在图像分类准确度方面表现出更为优秀的效果, 达到 2.3%的准确度提升; 在倾斜率为 0.3 时, So-CBI 方法在图像分类准确度同样优于原始算法, 达到 0.9%的准确度提升; 在倾斜率为 0.1 时, So-CBI 方法在图像分类准确度仍然保持领先水平, 达到 1.2%的准确度提升. 值得注意的是: Tiny ImageNet 是数据规模极大极具挑战性的数据集, 在此数据上, So-CBI 体现出了极高的适用性. 然后, 我们观察使用 K 中心贪心算法(KCenter)的效果. 可以看出, So-CBI 方法仍然保持与原始方法竞争性的效果. 例如: 倾斜率为 1 时, 有 1.2%准确度提升; 倾斜率为 0.3 时, 有 0.8%准确度提升; 倾斜率为 0.1 时, 有 0.7%的准确度提升.

在 SVHN、CIFAR-10、CIFAR-100 数据集上, 我们进行了相同的实验, 训练参数与 Tiny ImageNet 保持一致. 为了更好地表示方法之间的性能差别, 我们采用表格的形式对我们方法的每轮次的准确率增益进行表示. 表 4 展示了在不同倾斜程度的 SVHN 数据集上的性能评估, 表中数值表示 So-CBI 方法的增益效果, 数值后的括号里的数值代表原始算法的精度.

表 4 仅使用 50%数据在不同倾斜程度的 SVHN 数据集上的性能评估

不平衡指数	方法	批次(每批次 128 个数据)				
		10	20	30	40	50
1	LC-So-CBI (%)	0.12(+68.9)	**0.21(+72.3)**	0.11(+74.2)	0.13(+77.1)	0.14(+79.2)
	KCenter-So-CBI (%)	0.04(+73.9)	0.02(+77.2)	**0.09(+79.3)**	**0.09(80.1)**	0.05(+81.7)
0.3	LC-So-CBI (%)	0.07(+69.3)	0.12(+72.9)	**0.13(+73.6)**	0.11(+76.6)	0.12(+77.9)
	KCenter-So-CBI (%)	0.03(+72.7)	0.06(+74.5)	**0.07(+78.8)**	0.05(+79.6)	0.04(+80.9)
0.1	LC-So-CBI (%)	**0.12(+67.3)**	0.06(+70.4)	0.07(+73.6)	0.05(+76.5)	0.04(+77.4)
	KCenter-So-CBI (%)	**0.09(+71.6)**	0.03(+76.8)	0.05(+77.9)	0.03(+79.4)	−0.02(80.7)

由表 4 可以看出, So-CBI 方法在图像分类准确度仍然保持领先水平. 在使用最低置信度算法(LC)时, 倾斜率为 1 时, 最高有 0.21%的准确度提升; 倾斜率为 0.3 时, 有 0.13%的准确度提升; 倾斜率为 0.1 时, 有 0.21%的准确度提升. 在使用 K 中心贪心算法(KCenter)时, 可以看出, So-CBI 方法仍然保持与原始方法竞争性效果.

表 5 展示了在不同倾斜程度的 CIFAR-10 数据集上的性能评估. 由表 5 可以看出, So-CBI 方法在图像分类准确度与原始方法保持领先水平. 在使用最低置信度算法(LC)时, 倾斜率为 1 时, 最高有 0.3%的准确度提升; 倾斜率为 0.3 时, 有 0.06%的准确度提升; 倾斜率为 0.1 时, 有 0.08%的准确度提升. 在使用 K 中心贪心算法(KCenter)的效果, 可以看出, So-CBI 方法仍然保持与原始方法竞争性的效果. 表 6 展示了在不同倾斜程度的 CIFAR-100 数据集上的性能评估.

表 5 仅使用 50%数据在不同倾斜程度的 CIFAR-10 数据集上的性能评估

不平衡指数	方法	批次(每批次 128 个数据)				
		10	20	30	40	50
1	LC-So-CBI (%)	0.05(+56.7)	0.02(+63.5)	0.03(+69.2)	–0.04(+76.6)	**0.3(+77.3)**
	KCenter-So-CBI (%)	0.1(+56.1)	0.06(+63.7)	–0.05(+70.4)	0.08(+75.7)	**0.15(+74.9)**
0.3	LC-So-CBI (%)	0.02(+56.3)	–0.03(+63.4)	0.04(+67.5)	–0.11(+72.3)	**0.06(+74.5)**
	KCenter-So-CBI (%)	0.03(+54.8)	**0.07(+61.5)**	0.05(+66.1)	0.01(+70.2)	0.02(+75.1)
0.1	LC-So-CBI (%)	**0.08(+54.6)**	–0.01(+59.2)	0.07(+64.7)	0.05(+69.3)	0.07(+70.7)
	KCenter-So-CBI (%)	–0.01(+53.1)	0.02(+57.5)	**0.05(+62.1)**	0.02(+65.7)	0.03(+69.6)

由表 6 可以看出, So-CBI 方法在图像分类准确度仍然保持领先水平. 在使用最低置信度算法(LC)时, 倾斜率为 1 时, 最高有 0.04%的准确度提升; 倾斜率为 0.3 时, 有 1.4%的准确度提升; 倾斜率为 0.1 时, 有 0.67%的准确度提升. 在使用 K 中心贪心算法(KCenter)的效果, 可以看出, So-CBI 方法仍然保持与原始方法竞争性的效果.

表 6 仅使用 50%数据在不同倾斜程度的 CIFAR-100 数据集上的性能评估

不平衡指数	方法	批次(每批次 128 个数据)				
		10	20	30	40	50
1	LC-So-CBI (%)	0.02(+18.3)	–0.01(+22.2)	**0.04(+26.1)**	0.03(+29.1)	–0.01(+31.9)
	KCenter-So-CBI (%)	0.04(+19.6)	**0.14(+22.7)**	0.13(+26.8)	0.09(+29.4)	0.11(+31.5)
0.3	LC-So-CBI (%)	0.3(+18.2)	0.62(+21.6)	0.11(+23.2)	0.8(+25.7)	**1.4(+29.9)**
	KCenter-So-CBI (50%)	0.14(+18.7)	0.09(+22.1)	0.13(+24.7)	0.41(+26.9)	**0.64(+29.5)**
0.1	LC-So-CBI (50%)	0.24(+17.9)	0.43(+20.4)	**0.67(+22.1)**	0.53(+25.1)	0.59(+26.4)
	KCenter-So-CBI (50%)	0.19(+18.2)	0.26(+21.9)	0.11(+23.1)	**0.32(+25.6)**	0.25(+27.9)

4.4.2 样本检索效率对比

为了对比我们的索引框架在该更新场景下的性能, 我们使用在多模态数据集和图像分类数据集上得到的样本间 KNN 关系数据, 统计我们的索引技术以及其他方案在该场景下, 实现样本分数更新、样本分数排序、根据分数检索样本所花费的时间. 对比使用的其他方案使用的具体方法为:

- 方案 1: 借助数组存储样本分数, 每次全部重新排序获得 Top N 的样本;
- 方案 2: 借助数组存储样本分数, 使用堆过滤获得 Top N 的样本;
- 方案 3: 借助 B+树存储样本分数, 每次遍历叶子节点获得 Top N 的样本.

(1) 多模态模型训练

首先, 我们将 So-CBI 使用到多模态模型训练中. 为更好地显示对比效果, 数据集选用数据规模更大的 Visual Genome 数据集, 样本检索结果如表 7 所示. 在多模态数据集 Visual Genome 数据集上, 当使用 So-CBI 样本检索算法时表现出更为优秀的效果, 相比于方案 2, 总计达到 1.9 倍的速度提升. 相比于方案 1, 总计达到 1.4 倍的速度提升; 相比于方案 2, 总计达到 1.9 倍的速度提升; 相比于方案 3, 总计达到 1.1 倍的速度提升. 由结果可以看出, So-CBI 方法在模型训练中对样本检索速度有着大幅提高.

表 7 在不同索引方案下的多模态数据集(Visual Genome)上操作花费时间对比

索引方案	更新线程数	样本更新(min)	样本检索(min)	总计(min)
So-CBI	1	12.5	**0.01**	12.5
	5	9.4		9.4
	10	7.4		7.4
	20	**5.7**		**5.7**
方案 1	1	0.04	7.9	7.9
方案 2	1	14.3	0.02	14.3
	5	13.1		13.1
	10	12.3		12.3
	20	10.8		10.8
方案 3	1	16.4	0.01	16.4
	5	11.9		11.9
	10	8.2		8.2
	20	6.4		6.4

(2) 图像分类模型训练

然后, 为了验证我们方法的普适性, 我们将 So-CBI 使用到图像分类模型训练中, 样本检索结果如表 8 所示. 在图想分类数据集 ImageNet 数据集上, 当使用 So-CBI 样本检索算法时表现出更为优秀的效果, 相比于方案 1, 总计达到 1.76 倍速度提升; 相比于方案 2, 总计达到 2.7 倍的速度提升; 相比于方案 3, 总计达到 1.3 倍的速度提升. 当由结果可以看出, So-CBI 方法在模型训练中对样本检索效率提升有着非常重要的作用.

表 8 在不同索引方案下的图像数据集(ImageNet)操作花费时间对比

索引方案	更新线程数	样本更新(min)	样本检索(min)	总计(min)
So-CBI	1	178	**0.01**	178
	5	127		127
	10	109		109
	20	**81**		**81**
方案 1	1	0.03	143	143
方案 2	1	228	0.02	228
	5	224		224
	10	223		223
	20	219		219
方案 3	1	247	0.01	247
	5	171		171
	10	129		129
	20	103		103

根据准确率和样本检索时间的实验结果可以得出: 基于主动学习的高效样本检索方法 So-CBI 可以在保证模型训练准确率提升 2.3%的情况下, 减少模型训练所需要检索的 50%的样本数量, 从而模型的训练时间缩短至仅仅需要一半的时间. 同时, 主动学习在更新样本时, 样本检索效率提升 2.7 倍.

综上所述, 基于主动学习的高效样本检索方法 So-CBI 在多个实验中展现出了显著的有效性. 通过引入预训练的样本表征模型、边界感知的样本预筛选模型以及半有序索引的样本检索优化模型, So-CBI 能够在保持模型准确率的同时, 大幅减少所需的训练样本数量, 从而缩短了模型的训练时间. 实验结果表明: 使用 So-CBI 方法进行主动学习更新样本时, 样本的检索效率也得到了显著提升. 这些结果表明: So-CBI 方法在高效样本检索方面具有较高的实用性和有效性, 能够为样本选择和模型训练过程带来重要的改进.

5 总 结

本文提出了一种基于主动学习的高效样本检索技术, 旨在解决传统样本选择方法在处理大规模数据集时的问题. 通过引入预训练的样本表征模型、基于边界感知的样本预筛选模型和基于半有序索引的样本检索优化模型, 本文的方法(So-CBI)在保持准确率的同时, 显著减少了训练样本的数量, 并提高了样本检索的效率. 实验结果表明: 使用 So-CBI 方法进行样本选择和模型训练, 可以显著缩短训练时间, 降低训练样本的需求量, 并提高样本检索的效率. 与传统方法相比, 本文方法能够准确选择对模型训练更有价值的样本, 避免了重复训练和低效的数据扫描与调整. 此外, 基于半有序索引的样本检索优化模型还降低了数据维护的开销, 提高了系统整体效率. 未来的研究可以进一步优化和扩展这一技术, 以应对更复杂的数据集和任务.

References:

[1] Yin C, Menglin J, Tsung-Yi L, *et al*. Class-balanced loss based on effective number of samples. In: Proc. of the CVPR. 2019. 9268–9277.

[2] Du PF, Li XY, Gao YL. Survey on multimodal visual language representation learning. Ruan Jian Xue Bao/Journal of Software, 2021, 32(2): 327−348 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6125.htm [doi: 10.13328/j.cnki.jos.006125]

[3] Zhao WZ, Ma HF, Li ZQ, Shi ZZ. Efficiently active learning for semi-supervised document clustering. Ruan Jian Xue Bao/Journal of Software, 2012, 23(6): 1486−1499 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4073.htm [doi: 10.3724/SP.J.1001.2012.04073]

[4] Xie Y, Tomizuka M, Zhan W. Towards general and efficient active learning. arXiv:10.48550, 2021.

[5] Bengar J, Weijer J, Fuentes L, *et al*. Class-balanced active learning for image classification. In: Proc. of the IEEE/CVF Winter Conf. on Applications of Computer Vision. 2022. 1536−1545.

[6] Emam Z, Chu H, Chiang P, *et al*. Active learning at the ImageNet scale. arXiv:2111.12880, 2021.

[7] Dan W, Yi S. A new active labeling method for deep learning. In: Proc. of the Int'l Joint Conf. on Neural Networks (IJCNN). 2014. 112−119.

[8] Claude ES. A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review, 2001, 5(1): 3−55.

[9] Jordan TA, Zhang CCh, Akshay K, *et al*. Deep batch active learning by diverse, uncertain gradient lower bounds. arXiv:1906.03671, 2019.

[10] Daniel G, Shai SS. Discriminative active learning. arXiv:1907.06347, 2019.

[11] Bengar JZ, Joost VDW, Fuentes LL, *et al*. Class-balanced active learning for image classification. In: Proc. of the IEEE/CVF Winter Conf. on Applications of Computer Vision. 2021. 1536−1545.

[12] Tang X, Wu S, Chen G, *et al*. A learning to tune framework for LSH. In: Proc. of the Int'l Conf. on Data Engineering (ICDE). 2021. 2201−2206.

[13] Gordo A, Almazan J, Revaud J, *et al*. End-to-end learning of deep visual representations for image retrieval. Int'l Journal of Computer Vision, 2017, 124(2): 237−254.

[14] Girdhar R, Ramanan D. Attentional pooling for action recognition. arXiv:1711.01467, 2017.

[15] Fukunaga K, Narendra PM. A branch and bound algorithm for computing *k*-nearest neighbors. IEEE Trans. on Computers, 1975, 100(7): 750−753.

[16] Muja M, Lowe DG. Fast approximate nearest neighbors with automatic algorithm configuration. In: Proc. of the VISAPP. 2009. 331−340.

[17] Jégou H, Douze M, Schmid C. Product quantization for nearest neighbor search. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2010, 33(1): 117−128.

[18] Zhang H, Li F, Liu S, *et al*. DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection. arXiv:2203.03605, 2022.

[19] Malkov YA, Yashunin DA. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2018, 42(4): 824−836.

[20] Zhang J, Wu S, Tan Z, *et al*. S3: A scalable in-memory skip-list index for key-value store. Proc. of the VLDB Endowment, 2019, 12(12): 2183−2194.

[21] Workman M, Hoffman C. An evaluation of the role the Internet site Petfinder plays in cat adoptions. Journal of Applied Animal Welfare Science, 2015, 18(4): 388−397.

[22] Kiela D, Firooz H, Mohan A, *et al*. The hateful memes challenge: Detecting hate speech in multimodal memes. In: Proc. of the Advances in Neural Information Processing Systems. 2020. 2611−2624.

[23] Krishna R, Zhu Y, Groth O, *et al*. Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int'l Journal of Computer Vision, 2017, 123: 32−73.

[24] Dong C, Loy CC, He KM, *et al*. Learning a deep convolutional network for image super-resolution. In: Proc. of the 13th European Conf. on Computer Vision (ECCV 2014). Springer, 2014. 184−199.

[25] Radford A, Kim JW, Hallacy C, *et al*. Learning transferable visual models from natural language supervision. In: Proc. of the Int'l Conf. on Machine Learning. 2021. 8748−8763.

[26] Odusami M, Maskeliūnas R, Damaševičius R, *et al*. Analysis of features of Alzheimer's disease: Detection of early stage from functional brain changes in magnetic resonance images using a finetuned ResNet18 network. Diagnostics, 2021, 11(6): Article No. 1071.

[27] Le Y, Yang X. Tiny ImageNet Visual Recognition Challenge. CS 231N, 2015, 7(7): Article No. 3.

附中文参考文献:

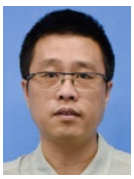
[2] 杜鹏飞, 李小勇, 高雅丽. 多模态视觉语言表征学习研究综述. 软件学报, 2021, 32(2): 327−348. http://www.jos.org.cn/1000-9825/6125.htm [doi: 10.13328/j.cnki.jos.006125]

[3] 赵卫中, 马慧芳, 李志清, 史忠植. 一种结合主动学习的半监督文档聚类算法. 软件学报, 2012, 23(6): 1486−1499. http://www.jos.org.cn/1000-9825/4073.htm [doi: 10.3724/SP.J.1001.2012.04073]

唐秀(1995－), 女, 博士, CCF 专业会员, 主要研究领域为数据库查询优化, 数据库测试, 数据智能.

侯捷(1999－), 男, 硕士生, 主要研究领域为数据库索引优化.

伍赛(1980－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为分布式数据库检索和查询, 大数据分析处理, 基于机器学习的数据库智能化算法.



陈刚(1973－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为数据库系统, 大数据技术, 数据智能计算.