

doi: 10.3969/j.issn.2095-0411.2025.01.005

BatchOOD: 基于能量的批处理式多标记分布外检测

程一飞^{1,2}, 彭欣¹, 程玉胜^{1,2}, 陈启东¹

(1. 安庆师范大学 计算机与信息学院, 安徽 安庆 246011; 2. 安徽省高校智能感知与计算重点实验室 (安庆师范大学), 安徽 安庆 246011)

摘要: 分布外 (Out-of-Distribution, OOD) 检测对于深度模型在开放环境中安全可靠地应用至关重要。现有方法通常利用深度网络提取分布内 (In-Distribution, ID) 表征, 却忽略了对小批量样本间关系的学习, 并且缺乏针对更符合现实设置的多标记 OOD 检测的研究。基于此, 文章提出一种在批处理级上检测 OOD 样本的深度模型 BatchOOD。首先, 利用主干网络提取单个样本的初始特征; 随后, 引入 BatchFormer 模块从批量维度上探索样本间的依赖关系; 最后, 应用基于能量的多标记 OOD 检测器判别 ID 样本和 OOD 样本。在 MS-COCO, PASCAL-VOC 和 NUS-WIDE 3 个多标记数据集上的实验结果表明建模小批量间样本依赖关系更有利于模型学习到精确的 ID 表征, 从而提高 ID-OOD 可分离性。与 JointEnergy 方法相比, 文章所提出的模型在 FPR95 指标上分别实现了 6.42%, 5.72% 和 6.57% 的性能提升。

关键词: 多标记学习; 分布外检测; BatchFormer; 样本关系; 能量函数

中图分类号: TP 391

文献标志码: A

文章编号: 2095-0411(2025)01-0037-11

BatchOOD: energy-based batch-level multi-label out-of-distribution detection

CHENG Yifei^{1,2}, PENG Xin¹, CHENG Yusheng^{1,2}, CHEN Qidong¹

(1. School of Computer and Information, Anqing Normal University, Anqing 246011, China; 2. The University Key Laboratory of Intelligent Perception and Computing of Anhui Province, Anqing Normal University, Anqing 246011, China)

Abstract: Out-of-distribution (OOD) detection is crucial for ensuring the safe and reliable deployment of deep models in open environments. Existing approaches typically rely on deep networks to extract in-distribution (ID) representations. However, these methods overlook the significance of understanding the sample relationships from each mini-batch and lack research on multi-label OOD that bet-

收稿日期: 2024-05-27。

基金项目: 安徽省自然科学基金面上资助项目(2108085MF216)。

作者简介: 程一飞(1976—), 男, 安徽怀宁人, 硕士, 副教授。通信联系人: 程玉胜(1969—), E-mail: chengyusheng@163.com

引用本文: 程一飞, 彭欣, 程玉胜, 等. BatchOOD: 基于能量的批处理式多标记分布外检测[J]. 常州大学学报(自然科学版), 2025, 37(1): 37-47.

ter aligned with real-world settings. Based on this, a deep model named BatchOOD for batch-level OOD detection was proposed in this paper. Firstly, initial features of individual samples were extracted utilizing the backbone. Secondly, the BatchFormer module was introduced to explore dependencies among samples from the batch dimension. Finally, an energy-based multi-label OOD detector to discriminate between ID and OOD was applied. Experimental results on three multi-label datasets, namely MS-COCO, PASCAL-VOC, and NUS-WIDE, demonstrate that modeling inter-sample dependencies within each batch enhances the learning of accurate ID representations, leading to improved ID-OOD separability. Compared with the JointEnergy method, our proposed model achieves performance improvements of 6.42%, 5.72%, and 6.57% in terms of the FPR95 metric, respectively.

Key words: multi-label learning; out-of-distribution detection; BatchFormer; sample relationships; energy function

在开放的环境中, 各种各样的数据可以很自然地与训练数据呈现不同的分布。分布外 (Out-of-Distribution, OOD) 样本的出现对封闭世界的基本假设提出了挑战, 并且研究表明深度神经网络对该类样本往往会产生过度自信的预测^[1], 尤其是在自动驾驶^[2]、医疗诊断^[3]、欺诈检测^[4]等安全关键型应用中, 不可靠地预测可能导致严重的损失或事故, 危及人类的财产和生命安全。分布外检测的目的是在推理阶段检测来自训练集分布之外的样本, 即与分布内 (In-Distribution, ID) 标签集没有交集的样本。该领域在增强模型鲁棒性、提高系统安全性和可靠性等方面具有较高的应用价值, 因此受到研究界的广泛关注^[5-9]。但已有研究集中于探索多类分类设置下的分布外 (OOD) 问题, 其中每个样本只携带一个标记, 对于多标记 OOD 问题的研究仍处于起步阶段。在现实场景中, 标记空间规模通常较大, 往往一个样本被分配多个标记。因此, 相比于多类 OOD 样本, 多标记 OOD 样本的研究更有助于推动 OOD 检测的实际应用。

已有的 OOD 检测技术可以大致分为 3 种策略。第一种策略是基于神经网络分类器的 OOD 检测方法, 即设计新的分类网络模型提取 ID-OOD 的判别性特征, 然后利用模型输出设计 OOD 评分函数^[10]。基于度量学习的 OOD 检测方法是第二种策略, 该策略通过计算样本之间的距离和相似度检测 OOD 样本^[11]。最后一种策略是基于生成模型的 OOD 检测方法, 主要利用生

成模型拟合 ID 分布并在该分布下评估测试样本的似然值^[12]。然而现有方法大都通过重构输入和损失函数朴素地学习样本之间的关系, 并没有挖掘深度网络的内部结构使模型具备建模样本关系的能力, 更没有从训练批次的维度进行数据的协同学习。受 BatchFormer^[13]启发, 文章尝试从批次 (Batch) 的角度执行注意力操作, 促使模型学习每个批量间样本的关联性。

在多标记 OOD 样本中, 标记语义空间庞大且复杂, 一般的卷积网络只能对图像的局部信息进行建模而不能捕获全局上下文信息^[14-15]。众所周知 Transformer^[16]由于其核心的自注意力机制在许多视觉任务中展现出了强大的建模长距离依赖关系的能力, 这对于网络全面理解图像语义非常有益。为了更好的建模样本级的相关性, 在小批量的维度上应用 Transformer, 探索批量级数据间的相互作用, 以学习鲁棒的数据表征。因此, 文章提出的方法是基于神经网络分类器的 OOD 检测算法, 旨在优化神经网络结构, 促使网络学习到更有利于区分 ID 样本和 OOD 样本的判别性表示。

另外, 能量函数在理论上与输入的概率密度保持一致, 不易受到过度置信问题的影响, 并且可以从任何神经分类器中导出, 无需重新训练即可灵活运用^[17]。因此在文中建议使用能量函数^[18-19]作为 OOD 样本评估函数。具有较高能量分数的输入意味着其密度较低, 在基于能量的 OOD 检测器中就会被归类为 OOD 样本。

综上,为了在多标记设置下提升模型的 ID-OOD 分离性能,文章提出了一种在小批量级别上检测 OOD 样本的深度模型 BatchOOD。首先,利用主干网络提取样本的初始特征;然后,引入 BatchFormer 模块探索批量样本间的依赖关系从而学习到更具判别力和更丰富的样本表征;最后,设计一个基于能量的多标记 OOD 检测器捕获跨标签之间的联合信息。在 MS-COCO, PASCAL-VOC 和 NUS-WIDE 多标记数据集上的实验结果表明 BatchOOD 能够更好地获取 ID 样本的表征,从而实现利用模型的输出达到更优的评分结果。主要贡献有:①为学习小批量样本间的上下文表示构建了一种端到端的深度模型 BatchOOD,旨在批处理中传递消息并应用核心的自注意力机制,探索更具判别性表征的样本关系。②实验结果表明利用小批量样本间的相关性进一步塑造了能量差距,实现了更优的 ID-OOD 分离性。

1 相关工作

在真实开放环境中,数据往往不满足独立同分布的假设,即存在大量数据与训练数据呈现不同的分布。分布外(OOD)检测任务旨在检测出与训练集分布不同的样本,并将其预测为 OOD 样本。执行 OOD 检测可以确保模型在面对未见过的或不相关的数据时做出可靠的预测。但由于分布外空间是无穷无尽的,因此想要模型在所有数据上都具有很好的泛化性能十分困难,尤其在面对复杂的标记空间时,模型的性能更会急剧下降,于是对多标记 OOD 样本的研究显得尤为重要。文章旨在多标记设置下提升模型的 ID-OOD 检测性能,以增强模型在多标记空间中的鲁棒性和稳定性。

1.1 多类分布外检测

NGUYEN 等^[1]首次指出深度神经网络在分布外样本上存在过度自信的问题,这一现象可能导致机器学习系统性能的急剧下降。以自动驾驶系统为例,该系统可能会将未知的异常场景错误地预测为正常场景,从而造成事故的发生。而

OOD 检测希望系统检测出异常情况并采取预防措施。为了克服以上问题,在 OOD 检测领域涌现出了许多有意义的研究方法。YANG 等^[20]和郭凌云等^[21]的综述中对已有的 OOD 检测方法进行了深入分析,一般来说,这些方法可以分为判别式和生成式两大类。

1.1.1 判别式分布外检测

对分布外检测的研究可以追溯至一个公认的基线(Maximum softmax probabilities, MSP),即 HENDRYCKS 等^[22]通过对输出 logit 进行 softmax 函数计算,利用得到的预测概率区分 ID 样本和 OOD 样本。先驱工作 ODIN 方法^[23]通过使用温度缩放和在输入中添加扰动对基线进行改进。WANG 等^[24]认为仅依赖单一输出的判别方法不足以准确区分 OOD 数据,提出一种结合类别无关的特征空间得分和 ID 类别相关的 logit 得分方法 Vim,从多个维度上判断 OOD 数据。与文章工作相近的 Oodformer^[25]将 Transformer 作为主要的特征提取器探索全局图像的上下文信息以区分 ID 样本和 OOD 样本之间的非局部对象性,但这种方法将 Transformer 作为主干网络存在忽略图像的局部信息局限。因此,在文中工作仅利用 Transformer 辅助建模长距离依赖关系。除此之外,研究者也探索了基于距离的检测方法,为了尽可能使 OOD 数据远离质心或原型所在的分布内位置以判别其类别,如马氏距离^[11]、IsoMax 损失^[26]和非参数最近邻距离^[27]。这些方法旨在克服多类分类任务中检测分布外数据的挑战,但这种设置在现实场景中并不常见。因此,文章致力于研究更符合真实场景的多标记 OOD 检测方法。

1.1.2 生成式分布外检测

鉴于神经网络仅依赖输入数据难以准确近似真实样本的分布,有研究^[28-30]采用生成模型学习已知和未知样本之间的决策边界,从而应对这一挑战。生成模型可以基于密度估计来学习数据的潜在分布,DINH 等^[31]和 VAN DEN OORD 等^[32]利用密度估计区分 ID 样本和 OOD 样本的分布。文中工作虽然是基于判别式分类模型,而在统计物理学领域已证明了能量函数与数据密度

函数之间的等价性^[17], 所以使用能量得分使得文章的方法也可以从密度估计的角度进行解释。更重要的是, 在设计基于生成模型时存在两个主要挑战: 一是生成与真实数据相似的 OOD 样本比较困难; 二是模型的训练和优化也具有挑战性^[33]。相比之下, 文章的研究规避了这两个挑战, 并且不需要借用辅助 OOD 样本进行训练以及使用标准随机梯度下降法 (SGD) 即可优化。

1.2 多标记分布外检测

HENDRYCKS 等^[34]认为与现实应用更为贴近的是大规模的多类和多标记的 OOD 设置, 并基于此提出了 MaxLogit 方法, 为未来的研究建立了新基线。随后 WANG 等^[19]提出的 JointEnergy 方法则是第一个专门针对多标记 OOD 问题的研究, 该方法通过聚合每个标记上的能量分数, 在多个标记中捕获联合不确定性来估计 OOD 样本的指标得分。WANG 等^[35]认为之前的研究忽略了多标记的独特属性, 提出稀疏标签共现评分 (SLCS) 作为 OOD 检测器, 即利用标

记稀疏性和共现性进行检测。然而, 上述方法只在每个批量内独立地提取数据特征, 均没有考虑到批量与批量间样本存在的关系, 无法学习到完整数据集的表征。基于此, 文中提出使用 BatchFormer 探索小批量样本间的相互联系并使用核心的自注意力机制对长距离交互进行建模, 以学习更具判别性的整体图像表征, 从而更利于区分 ID 样本和 OOD 样本。

2 BatchOOD 算法建模

BatchOOD 的框架示意图如图 1 所示。包括数据的输入、主干网络、样本感知模块和分布外检测 4 个部分。首先, 已知多标记数据集和未知标记的 OOD 样本作为模型的输入; 其次, 利用主干网络提取单个样本的初始特征; 然后, 设计样本感知模块从批维的角度学习样本关系; 最后, 将学习到样本表征送入分类器获得样本 logit, 基于能量的多标记 OOD 检测器根据 logit 计算样本 OOD 得分从而区分 ID 样本和 OOD 样本。

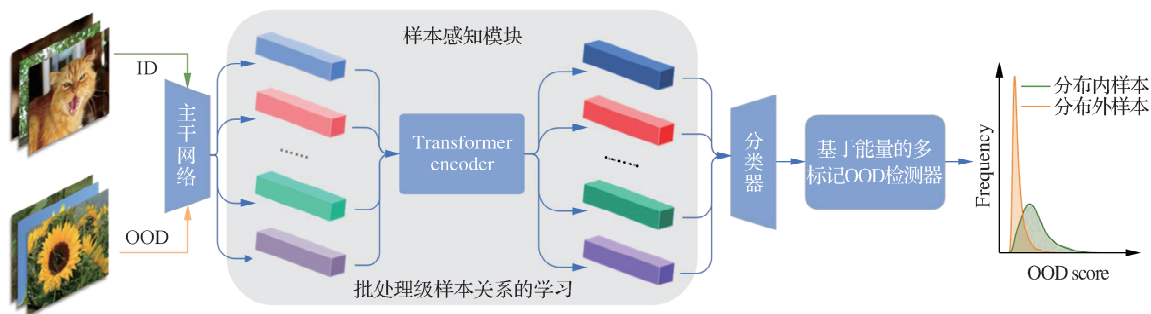


图 1 BatchOOD 框架示意图

Fig.1 Illustration of BatchOOD framework

2.1 OOD 检测问题的定义

给定训练数据集 $D = \{(x_i, y)\}_{i=1}^N$, $x_i \in \mathbb{R}^{N \times d}$, 为对应的第 i 个样本; N 为数据集内包含的样本个数; d 为样本的特征维度; $y = \{y_1, y_2, \dots, y_k\}$ 是 \mathbf{Y}_{ID} 的子集, 代表着标记空间, k 为该数据集所包含的标签个数, 而 $\mathbf{Y}_{ID} \subseteq \mathbb{R}^{N \times k}$ 是已知的分布内标记集, 二元标记 $y_k = 1$ 表示当前标记与样本 x 相关, $y_k = 0$ 表示不相关, 每个样本都携带一个标记集 y 。OOD 检测定义分布内

与分布外标记集不存在任何的交集, 即 $\mathbf{Y}_{OOD} \cap \mathbf{Y}_{ID} = \emptyset$, 其中 $\mathbf{Y}_{OOD} \subseteq \mathbb{R}^{N \times k}$ 代表分布外标记集。通常 OOD 分布在训练时是未知的, OOD 检测旨在训练一个 OOD 检测器, 在推理阶段能够识别和拒绝预测 OOD 样本。即定义 OOD 检测器 $D_\epsilon(x): x \rightarrow \mathbb{R}$ 为一个映射函数, 该函数可以为每个样本映射对应的 OOD 得分, 以代表该样本属于 OOD 样本的可能性。

简而言之在推理阶段, 多标记 OOD 检测可以形式化为一个二分类问题, 其目标就是设计一

个多标记 OOD 检测器,以判断输入 x 是 ID 样本还是 OOD 样本,见式 (1)

$$D_{\epsilon}(x) = \begin{cases} 0, & \text{if } x \in Y_{ID} \\ 1, & \text{if } x \in Y_{OOD} \end{cases} \quad (1)$$

2.2 批处理级样本关系的学习

开放环境中不同样本之间的关系复杂且多样,而现有方法都是通过简单直接的方式去探索样本之间的关联性,即直接在神经网络的输入和输出中建模样本相关性,而没有在训练过程中利用网络本身结构自动学习样本之间的关系。并且神经网络只能独立地探索每个批量内的样本关系,对批量与批量间样本关系建模存在局限。受 BatchFormer 启发,尝试从批次处理的角度出发,利用具有强大关系建模能力的 Transformer^[16]对样本关系进行学习。值得一提的是从批处理级角度出发检测 OOD 样本在该领域中是从未被考虑的。

文章提出一种新型的端到端的深度模型 BatchOOD 以学习小批量样本间的相关性,从而获取更具判别性且丰富的 ID 表征。具体而言,如图 1 所示,首先使用深度密集型连接卷积网络 DenseNet^[36]作为主干网络去学习单个样本的特征,即在每个批次中样本之间不存在交互;将得到的独立样本的特征图建模为序列数据送入样本感知模块,利用该模块核心的自注意力机制建模样本之间的交互;最后利用样本感知模块学习到的样本表征作为最终分类器的输入。

图 1 中样本感知模块的核心是 Transformer 编码部分。利用 Transformer 对样本之间的交互进行建模,实际就是应用自注意力机制有效地捕捉输入序列中不同位置之间的长距离依赖关系,使得模型能够同时捕捉到序列中的局部和全局样本依赖关系。自注意力机制公式为

$$\text{Attention}(\mathbf{q}, \mathbf{k}, \mathbf{v}) = \text{softmax}\left(\frac{\mathbf{q}\mathbf{k}^T}{\sqrt{d'}}\right)\mathbf{v} \quad (2)$$

式中 \mathbf{q} , \mathbf{k} 和 \mathbf{v} 分别对应一个输入样本进入样本感知模块后,通过线性映射变换成维度 (d) 相同的 3 组矩阵向量,即分别代表查询 (query)、

键 (key) 和值 (value)。Transformer 的核心是由多头自注意力机制和多层感知机模块组成。对于一个给定 m 头的多头注意力模块,会进一步将 \mathbf{q} , \mathbf{k} 和 \mathbf{v} 分为 m 组,每一个经过多头注意力操作的向量其维度都会变换为 $d' = d/m$ 。多头注意力的计算过程为:

$$\text{MultiHead}(\mathbf{q}, \mathbf{k}, \mathbf{v}) =$$

$$\text{Concat}(\text{head}_1, \dots, \text{head}_m) \mathbf{w}_0 \quad (3)$$

$$\text{head}_i = \text{Attention}(\mathbf{q}\mathbf{w}_i^q, \mathbf{k}\mathbf{w}_i^k, \mathbf{v}\mathbf{w}_i^v), i \in [1, m] \quad (4)$$

式中 \mathbf{w}_0 , \mathbf{w}_i^q , \mathbf{w}_i^k 和 \mathbf{w}_i^v 为线性映射参数矩阵, $\mathbf{w}_0 \in \mathbb{R}^{d \times d}$, $\mathbf{w}_i^q \in \mathbb{R}^{d \times d}$, $\mathbf{w}_i^k \in \mathbb{R}^{d \times d}$, $\mathbf{w}_i^v \in \mathbb{R}^{d \times d}$ 。

多头注意力已经成功应用于通道和空间维度的关系建模^[37-38],以此为基础,文章将其拓展到探索批处理维度的样本关系上。与传统的 Transformer 不同,文中工作首先会重塑输入,即将每个批处理级中的所有图像视为一个序列,其中每个图像被看作序列中的一个节点,使得模型可以在小批量的样本上进行工作。通过这一操作,样本感知模块中的自注意力机制形成了不同样本之间的交叉注意。经过上述特征计算,该模块充分考虑了序列中每个样本与其他样本的关联性,从而实现了针对不同样本之间关系的建模。

2.3 基于能量的多标记分布外检测器

OOD 检测是一个二分类问题,通常依赖于样本评估方法区分分布内和分布外的数据。LIU 等^[17]指出 softmax 函数可能导致模型对 OOD 数据产生较高的置信度预测,这一局限性促使能量评分方法的提出。LECUN 等^[18]提出的基于能量模型 (EBM) 通过构建一个能量函数,将输入空间中的样本映射到非概率能量值上,作为概率估计的替代方法。这种做法避免了模型中归一化的相关问题,同时由于能量分数在理论上与输入的概率密度保持一致且表现良好^[19],在文中也应用能量评分作为评估 OOD 得分的方法。

能量函数 $E(x): \mathbb{R}^D \rightarrow \mathbb{R}$ 将一个输入 x 映射成其对应的能量得分,一个能量得分的集合可以通过玻尔兹曼分布转化成概率密度

$$p(x) = \frac{\exp(-E(x))}{Z} \quad (5)$$

式中 Z 为关于 x 的标准化常数, 也被称为配分函数。对式 (5) 两边取对数, 得到

$$\log p(x) = -E(x) - \log Z \quad (6)$$

由于 Z 在所有输入 x 上是一个常数, 在训练和推理过程中不对检测结果造成任何影响, 所以可以忽略式 (6) 的最后一项 $\log Z$, 由式 (6) 得到最终简化版能量函数表达式为

$$E(x) = -\log p(x) \quad (7)$$

分析式 (7) 发现能量函数与对数似然函数实际上是线性对齐的, 这正好符合了 OOD 检测的理论需求^[19]。

而对于一个在 ID 数据上训练的分类器 $f_{y_i}(z) = f(z_i; \theta) \cdot w_{cls}^i$, 将深度特征 z_i 映射为第 i 类输出 logit, w_{cls}^i 表示第 i 类的权重向量, $f(z_i; \theta)$ 是倒数第二层的特征向量。将 logit 作为二元逻辑分类器的输入, 于是预测得出的分类概率为

$$p(y_i = 1 | z) = \frac{\exp(f_{y_i}(z))}{1 + \exp(f_{y_i}(z))} \quad (8)$$

遵循^[19]设置, 对于每一个类 y_i , 可定义标签级能量 $E_{y_i}(z)$ 为

$$E_{y_i}(z) = -\log(1 + \exp(f_{y_i}(z))) \quad (9)$$

标签级的能量只能获得单个标签信息, 不能捕获不同标签之间的相关性。因此聚合所有标签上的能量得分以考虑跨标签之间的关联性。由式 (10) 得到的能量值为负数, 为了与 OOD 决策惯例保持一致, 即较大得分的输入表示 ID 样本, 反之为 OOD 样本, 对聚合机制进行取负操作以确保结果为正。于是所有标签上的能量得分形式化定义为

$$E_{all}(z) = \sum_{i=1}^N -E_{y_i}(z) \quad (10)$$

至此设计了一个基于能量的多标记 OOD 检测器。该检测器将集成后的能量作为 OOD 得分, 集成了所有标签的能量分数, 减少了单一类别输出值对整体判断的影响, 提高了鲁棒性。更加适合于多标签 OOD 检测。能量分数用于衡量输入属于 ID 数据的概率, 分数越高表示 ID 样本的置信度越高。多标记 OOD 检测器定义为

$$D_{\epsilon}(x; \epsilon) = \begin{cases} \text{InD, if } E_{all}(z) > \epsilon \\ \text{OoD, if } E_{all}(z) \leq \epsilon \end{cases} \quad (11)$$

式中: InD 表示 ID 样本; OoD 表示 OOD 样本。具有更高得分的输入被视为 ID 样本, 反之被视为 OOD 样本。 ϵ 表示能够使真阳性率 (TPR) 值大于 95% 的决策阈值, 也是判断输入是否为 OOD 样本的阈值。

3 实验与分析

遵循共同的基准^[18], 在 3 个公开可用的多标记图像数据集上训练 BatchOOD 模型, 并在通用计算机视觉数据集 ImageNet 上进行验证。在 3 个指标上将 BatchOOD 与 7 个先进算法进行对比, 以提供一个全面的评估。此外, 为了突出实验结果的差异, 采用雷达图对结果进行分析。为了进一步证明 BatchOOD 的有效性, 进行了消融实验从而确定关键组件对整体性能的贡献。最后对决策阈值 ϵ 进行讨论, 以观察 OOD 检测性能的变化。通过分析, 对所提出模型进行了全面的分析和验证。

3.1 数据集和评价指标

3.1.1 分布内数据集

1) PASCAL-VOC2012^[39]: 包含 20 个标记和 22 531 张图片, 其中训练集含有 5 717 张图片, 验证集含有 5 823 张图片, 测试集含有 10 991 张图片, 每张图片包含不等数量的类别标记。

2) MS-COCO^[40]: 包含 80 个标记, 共有 164 062 张图片, 其中 82 783 张图片为训练集, 40 504 张图片为验证集, 40 775 张图片为测试集, 每张图片包含不等数量的类别标记。

3) NUS-WIDE^[41]: 包含 81 个概念标记和 269 648 张图片。考虑到此数据集中存在无效图片, 所以遵循文献 [42] 设置选择其子集, 即使用 119 986 张图片作为训练集, 80 283 张图片为测试集, 每张图片包含不等数量的类别标记。

3.1.2 分布外数据集

遵循文献^[43]设置从 ImageNet-22k 中选取一些作为 OOD 数据集以评估 BatchOOD 的性能。并且由于文章使用的深度网络是在 ImageNet-1k 上预训练的, 所以从 ImageNet-22k 中选取的类

别不与 ImageNet-1k 重叠。详细地说,选取包含 bamboo, bat, cotton, croissant, cherry tree, dolphin, deer, giant clam, Japanese cherry blossoms, leech, octopus, rhino, raccoon, redwood, rice, sunflower, stick cinnamon, sugar cane, turmeric, venus flytrap 这 20 个类别的数据作为 PASCAL-VOC2012 和 MS-COCO 的 OOD 测试集。由于 NUS-WIDE 包含了 animal, plants 和 flowers 等高级概念标签,选取不同的 20 个类别: asterism, battery, cave, cylinder, delta, fabric, filament, fire bell, hornet nest, kazoo, lichen, naval equipment, newspaper, paperclip, pythium, satellite, thumb, X-ray tube, yeast, zither 来构建 NUS-WIDE 对应的测试集。

3.1.3 评价指标

为了评估性能,采用 OOD 检测任务中 3 个通用的标准指标^[18]进行评估,即① FPR95 代表当 ID 样本真阳性率达到 95% 时 OOD 样本的假阳性率 (False Positive Rate, FPR); ② AUROC 为受试者工作特征 (ROC) 曲线下的面积; ③ AUPR 为精度-召回 (PR) 曲线下的面积。其中越低的 FPR95 和越高的 AUROC 和 AUPR 代表实验性能越好。

3.2 实验结果和分析

3.2.1 实验结果

在 MS-COCO, PASCAL-VOC2012 和 NUS-WIDE 3 个多标记图像数据集上,将 Batch-

hOOD 方法与另外 7 个先进的 OOD 检测方法进行比较,结果见表 1。从表 1 中对比数据可知, BatchOOD 在 3 个评价指标和数据集上均取得最优结果。在这 7 个对比算法中,大多基线方法,如 MaxLogit^[34], MSP^[22], ODIN^[23] 和 Mahalanobis^[11], 是基于所有标签的最大值来检测 OOD 样本。局部离群因子 (Local Outlier Factor, LOF)^[44] 使用 k 近邻 (k-nearest neighbors, KNN) 估计数据密度,其中 OOD 样本的密度相对较低其邻居样本的密度。隔离森林 (Isolation Forest)^[45] 是一种基于树的异常检测方法,依据从根节点到终止节点的路径长度来检测异常。其中 JointEnergy^[19] 是唯一针对多标记设置下 OOD 检测的研究,其余 6 种方法都是经典的单标记 OOD 检测方法在多标记场景下的拓展应用。

为了确保实验的客观性,所有对比算法的实验均采用相同的预训练网络。由表 1 可知, BatchOOD 方法在 3 个数据集 3 个评价指标上表现出色,优于另外 7 种 OOD 检测方法。这表明在小批量维度上应用 Transformer 探索批量级数据间的相关性,相比于现有方法有利于 OOD 检测性能的提升。实验结果也间接说明了单标记 OOD 检测器在解决多标记分类问题上的局限性,同时也凸显了多标记 OOD 检测任务的挑战性。综上, BatchOOD 方法从批维的角度学习样本关系更有利于建模样本级的相关性,从而学习到更具判别性的 ID-OOD 表征,进而利于模型区分 ID 样本和 OOD 样本。

表 1 多标记 OOD 检测方法的性能对比
Table 1 Performance comparison of multi-label OOD detection methods

Method	MS-COCO			PASCAL-VOC2012			NUS-WIDE		
	FPR95	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95	AUROC	AUPR
MaxLogit ^[34]	43.53	89.11	93.74	45.06	89.22	83.14	56.46	83.58	94.32
MSP ^[22]	79.90	73.70	85.37	74.05	79.32	72.54	88.50	60.81	87.00
ODIN ^[23]	45.04	89.32	94.40	38.57	86.53	79.10	50.84	83.32	95.15
Mahalanobis ^[11]	46.86	88.59	93.85	41.74	88.65	81.12	62.67	84.02	95.25
LOF ^[44]	80.44	73.95	86.01	86.34	69.21	58.93	85.21	67.75	89.61
Isolation Forest ^[45]	94.39	49.04	66.87	93.22	50.67	35.78	95.69	53.12	83.32
JointEnergy ^[19]	33.48	92.70	96.25	41.01	91.10	86.33	48.98	88.30	96.40
BatchOOD	27.76	93.70	96.74	34.59	92.35	87.40	42.41	90.07	96.94

3.2.2 实验分析

为了验证文中所提方法的竞争性, 采用雷达图^[46]进行对比实验分析, 如图 2 所示。针对不同数据集和评价指标, 比较了 BatchOOD 与其他 7 种方法在性能和相对关系上的差异。在这些评价指标中, FPR95 指标值越低表明性能越好, 而其他指标则数值越高表示性能越好。在雷达图

上的表现就是 FPR95 指标越接近中心, 表示性能越好; 而 AUROC 和 AUPR 指标越接近外围则表明性能越好。由图 2 可知, BatchOOD 在评价指标上表现出了最优性能, 这论证了在批处理级维度上探索样本关系可以学习到更具有判别性的 ID 表征以区分 ID 样本和 OOD 样本, 从而利于 OOD 检测性能的提升。

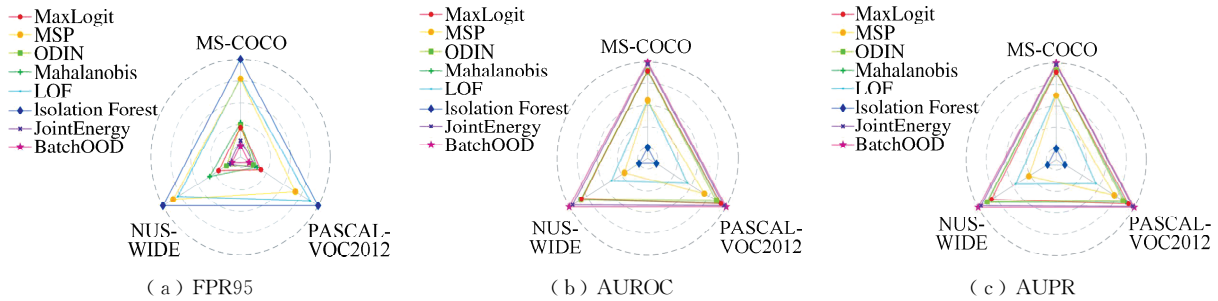


图 2 在不同数据集和评价指标上的对比分析

Fig.2 Comparative analyses on different datasets and evaluation indicators

3.3 消融实验

为了评估 BatchOOD 方法中核心组件的有效性, 对模型进行分解。生成了所提方法的 2 个退化版本: BatchOOD-b 去除 BatchFormer 模板, 仅利用其核心的 Transformer 架构建模整体样本之间的关联性, 从而没有在每个批次上独立地学习批量间样本关系, 而是全局地学习样本表征; BatchOOD-bf 去除了 BatchFormer 组件, 从而只依靠 DenseNet 网络本身的建模能力。

表 2 为文章所提出的方法和其他两种退化方法的消融实验对比结果。由表 2 可知在大多数情况下 BatchOOD 性能优于退化算法, 这表明在批处理角度上使用 Transformer 探索样本关系有利于 OOD 检测。同时也发现 BatchOOD-b 在 AUROC 和 AUPR 两个指标上有时会优于 BatchOOD, 充分说明了 Transformer 强大的表征建模能力。综上, BatchOOD 对于多标记 OOD 检测来说是一种有效方法。

表 2 样本感知模块上的消融实验

Table 2 Ablation experiments on sample-aware modules

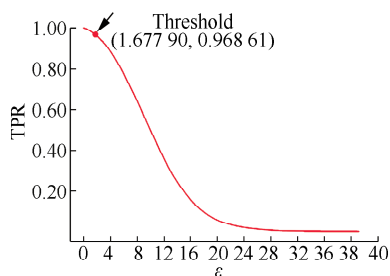
Method	MS-COCO			PASCAL-VOC2012			NUS-WIDE		
	FPR95	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95	AUROC	AUPR
BatchOOD	27.76	93.70	96.74	34.59	92.35	87.40	42.41	90.07	96.94
BatchOOD-b	29.38	93.88	97.03	35.15	93.20	89.78	47.01	89.14	96.69
BatchOOD-bf	33.48	92.70	96.25	41.01	91.10	86.33	48.98	88.30	96.40

3.4 阈值 ϵ 的讨论

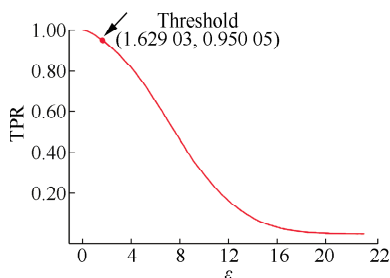
在多标记 OOD 检测中, 关键性阈值 ϵ 被用来判别 ID 和 OOD 类别。为了确保 ID 样本的真

阳性率 (TPR) 大于 95%, 需要 ϵ 有一个合适的取值。具体而言, 当样本得分大于 ϵ 时被判定为 ID 样本, 否则判定为 OOD 样本。为了分析 ϵ 对 OOD 样本检测效果的影响, 绘制了 Batc-

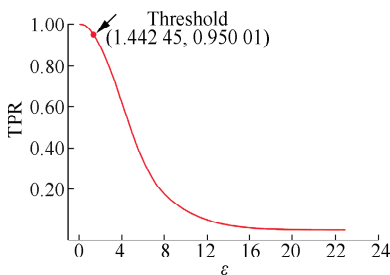
hOOD 在 3 个多标记数据集上的 ϵ -TPR 曲线, 如图 3 所示。由图 3 可知在 MS-COCO, PASCAL-VOC2012 和 NUS-WIDE 数据集上阈值 ϵ 分别为 1.677 90, 1.629 03 和 1.442 45 时 TPR 达到 95%。



(a) 在 MS-COCO 上的 ϵ -TPR 曲线



(b) 在 PASCAL-VOC2012 上的 ϵ -TPR 曲线



(c) 在 NUS-WIDE 上的 ϵ -TPR 曲线

图 3 BatchOOD 在 3 个数据集上的 ϵ -TPR 曲线

Fig.3 BatchOOD ϵ -TPR curves on three datasets

4 结 论

从批处理维度出发研究多标记 OOD 检测问题。以往的算法都在庞大的数据集上通过简单直接的方式去探索样本之间的关联性, 却没有利用深度网络本身探索小批量样本间的关系。针对该问题, 文章提出了在小批量级别上检测 OOD 样本的深度模型 BatchOOD。具体而言, 在多标记设置下, 将每一批次的每个样本视为一个序列的节点, 利用 Transformer 对序列中的每个节点内

的样本进行关系建模, 以达到在各个批量间学习样本关系的目的; 然后将学习到的样本表征送入基于能量的多标记 OOD 检测器判断样本是否为 OOD 样本, 实验结果表明了该方法的有效性。然而文章也存在一定的局限, 例如只按照训练批次建模样本之间相关性的方法还相对简单, 没有考虑到多标记的相关性和稀疏性等独特属性对多标记样本学习的帮助。未来将考虑更合理的多标记技术, 充分探索样本之间的关系以实现更优的 OOD 检测性能。

参考文献:

- [1] NGUYEN A, YOSINSKI J, CLUNE J. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015: 427-436.
- [2] FILOS A, TIGAS P, MCALLISTER R, et al. Can autonomous vehicles identify, recover from, and adapt to distribution shifts? [C]//Proceedings of the 37th International Conference on Machine Learning. Oxford: ACM, 2020: 3145-3153.
- [3] XIA T, DANG T, HAN J, et al. Uncertainty-aware health diagnostics via class-balanced evidential deep learning[J]. IEEE Journal of Biomedical and Health Informatics, 2024, 5(1): 245-256.
- [4] JEZEQUEL L, VU N S, BEAUDET J, et al. Efficient anomaly detection using self-supervised multi-cue tasks[J]. IEEE Transactions on Image Processing, 2023, 32: 807-821.
- [5] ZHENG H T, WANG Q Z, FANG Z, et al. Out-of-distribution detection learning with unreliable out-of-distribution sources[J]. Advances in Neural Information Processing Systems, 2023, 36: 72110-72123.
- [6] BEHPOUR S, DOAN T, LI X, et al. GradOrth: a simple yet efficient out-of-distribution detection with orthogonal projection of gradients[J]. Advances in Neural Information Processing Systems, 2023, 36: 38206-38230.
- [7] YANG J K, ZHOU K Y, LIU Z W. Full-spectrum out-of-distribution detection[J]. International Journal

- of Computer Vision, 2023, 131(10): 2607-2622.
- [8] 刘存, 杨曦晨, 陈天海, 等. 视频质量波动影响下的视频异常检测算法有效性分析[J]. 常州大学学报(自然科学版), 2024, 36(3): 45-58.
- [9] WANG Q Z, FANG Z, ZHANG Y G, et al. Learning to augment distributions for out-of-distribution detection[J]. Advances in Neural Information Processing Systems, 2023, 36: 73274-73286.
- [10] LIN Z Q, ROY S D, LI Y X. MOOD: multi-level out-of-distribution detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021: 15308-15318.
- [11] LEE K, LEE K, LEE H, et al. A simple unified framework for detecting out-of-distribution samples and adversarial attacks[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal: ACM, 2018: 7167-7177.
- [12] GRAHAM M S, PINAYA W H L, TUDOSIU P D, et al. Denoising diffusion models for out-of-distribution detection[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Vancouver: IEEE, 2023: 2948-2957.
- [13] HOU Z, YU B S, TAO D C. BatchFormer: learning to explore sample relationships for robust representation learning[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022: 7246-7256.
- [14] ZUO Z, SHUAI B, WANG G, et al. Learning contextual dependence with convolutional hierarchical recurrent neural networks[J]. IEEE Transactions on Image Processing, 2016, 25(7): 2983-2996.
- [15] SEO S, HUANG J, YANG H, et al. Interpretable convolutional neural networks with dual local and global attention for review rating prediction[C]//Proceedings of the Eleventh ACM Conference on Recommender Systems. Como: ACM, 2017: 297-305.
- [16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. California: ACM, 2017: 6000-6010.
- [17] LIU W T, WANG X Y, OWENS J D, et al. Energy-based out-of-distribution detection[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: ACM, 2020: 21464-21475.
- [18] LECUN Y, CHOPRA S, HADSELL R, et al. A tutorial on energy-based learning [C]//Predicting Structured Data. Cambridge: MIT Press, 2006: 1-59.
- [19] WANG H R, LIU W T, BOCCHIERI A, et al. Can multi-label classification networks know what they don't know? [J]. Advances in Neural Information Processing Systems, 2021, 34: 29074-29087.
- [20] YANG J K, ZHOU K Y, LI Y X, et al. Generalized out-of-distribution detection: a survey [EB/OL]. 2021: arXiv: 2110.11334. <http://arxiv.org/abs/2110.11334>.
- [21] 郭凌云, 李国和, 龚匡丰, 等. 图像分布外检测研究综述[J]. 模式识别与人工智能, 2023, 36(7): 613-633.
- [22] HENDRYCKS D, GIMPEL K. A baseline for detecting misclassified and out-of-distribution examples in neural networks[EB/OL]. 2016: arXiv: 1610.02136. <http://arxiv.org/abs/1610.02136>.
- [23] LIANG S Y, LI Y X, SRIKANT R. Enhancing the reliability of out-of-distribution image detection in neural networks [EB/OL]. 2017: arXiv: 1706.02690. <http://arxiv.org/abs/1706.02690>.
- [24] WANG H Q, LI Z Z, FENG L T, et al. ViM: out-of-distribution with virtual-logit matching[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022: 4911-4920.
- [25] KONER R, SINHAMAHAPATRA P, ROSCHER K, et al. Oodformer: out-of-distribution detection transformer[EB/OL]. 2021: arXiv: 2107.08976. <https://arxiv.org/abs/2107.08976>.
- [26] MACEDO D, REN T I, ZANCHETTIN C, et al. Entropic out-of-distribution detection: seamless detection of unknown examples[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(6): 2350-2364.
- [27] SUN Y Y, MING Y F, ZHU X J, et al. Out-of-distribution detection with deep nearest neighbors[EB/OL]. 2022: arXiv: 2204.06507. <http://arxiv.org/>

- abs/2204.06507.
- [28] RAN X M, XU M K, MEI L R, et al. Detecting out-of-distribution samples via variational auto-encoder with reliable uncertainty estimation[J]. *Neural Networks*, 2022, 145: 199-208.
- [29] LI Y W, WANG C J, XIA X B, et al. Out-of-distribution detection with an adaptive like lihood ratio on informative hierarchical VAE[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 7383-7396.
- [30] LI Z N, WU Q T, NIE F, et al. GraphDE: a generative framework for debiased learning and out-of-distribution detection on graphs[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 30277-30290.
- [31] DINH L, SOHL-DICKSTEIN J, BENGIO S. Density estimation using real nvp[C]//*Proceedings of the International Conference on Machine Learning (ICML)*. Sydney: PMLR, 2017.
- [32] VAN DEN OORD A, KALCHBRENNER N, VINYALS O, et al. Conditional image generation with PixelCNN decoders[EB/OL]. 2016: arXiv: 1606.05328. <http://arxiv.org/abs/1606.05328>.
- [33] HARSHVARDHAN G M, GOURISARIA M K, PANDEY M, et al. A comprehensive survey and analysis of generative models in machine learning[J]. *Computer Science Review*, 2020, 38: 100285.
- [34] HENDRYCKS D, BASART S, MAZEIKA M, et al. Scaling out-of-distribution detection for real-world settings[EB/OL]. 2019: arXiv: 1911.11132. <http://arxiv.org/abs/1911.11132>.
- [35] WANG L, HUANG S, HUANGFU L W, et al. Multi-label out-of-distribution detection via exploiting sparsity and co-occurrence of labels[J]. *Image and Vision Computing*, 2022, 126: 104548.
- [36] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017: 2261-2269.
- [37] NAM H, HA J W, KIM J. Dual attention networks for multimodal reasoning and matching[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017: 2156-2164.
- [38] BELLO I, ZOPH B, LE Q, et al. Attention augmented convolutional networks [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019: 3285-3294.
- [39] EVERINGHAM M, ALI ESLAMI S M, VAN GOOL L, et al. The pascal visual object classes challenge: a retrospective[J]. *International Journal of Computer Vision*, 2015, 111(1): 98-136.
- [40] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context [M]//*Computer Vision: ECCV 2014*. Cham: Springer International Publishing, 2014: 740-755.
- [41] CHUA T S, TANG J H, HONG R C, et al. NUS-WIDE: a real-world web image database from National University of Singapore[C]//*Proceedings of the ACM International Conference on Image and Video Retrieval*. Santorini: ACM, 2009: 1-9.
- [42] ZHU F, LI H S, OUYANG W L, et al. Learning spatial regularization with image-level supervisions for multi-label image classification[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017: 2027-2036.
- [43] HENDRYCKS D, BASART S, MAZEIKA M, et al. A benchmark for anomaly segmentation [EB/OL]. arXiv: 1911.11132. <https://arxiv.org/pdf/1911.11132v1>.
- [44] BREUNIG M M, KRIEGEL H P, NG R T, et al. LOF: identifying density-based local outliers[C]//*Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*. New York: ACM, 2000: 93-104.
- [45] LIU F T, TING K M, ZHOU Z H. Isolation forest [C]//2008 Eighth IEEE International Conference on Data Mining. Pisa: IEEE, 2008: 413-422.
- [46] GE W X, WANG Y B, XU Y T, et al. Causality-driven intra-class non-equilibrium label-specific features learning[J]. *Neural Processing Letters*, 2024, 56(2): 120.

(责任编辑:谭晓荷)