软件学报 ISSN 1000-9825, CODEN RUXUEW
2024,35(11):5196−5209 [doi: 10.13328/j.cnki.jos.007054] [CSTR: 32375.14.jos.007054]

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

基于最优传输理论的深度半监督学习伪标签生成算法*

翟德明[1], 沈斯娴[1], 周 雄[1], 江俊君[1], 刘贤明[1], 季向阳[2]

[1](哈尔滨工业大学 计算学部, 黑龙江 哈尔滨 150001)

[2](清华大学 自动化系, 北京 100084)

通信作者: 翟德明, Email: zhaideming@hit.edu.cn

摘 要: 目前, 深度学习广泛应用于各个领域并取得了优异的表现, 这通常需要大量标注数据的支持, 而大量标注数据的获取往往意味着高昂的成本与苛刻的应用条件. 因此, 随着深度学习的发展, 如何在实际场景下突破数据限制, 成为目前重要的研究目标, 而半监督学习正是其中一大研究方向. 半监督学习通过利用大量的未标记数据辅助少量的标记数据进行学习, 很好地减轻了深度学习的数据需求压力. 伪标签生成方法是当前半监督学习的重要组成部分, 所生成的伪标签质量的优劣会很大程度影响半监督学习的最终效果. 聚焦半监督学习中的伪标签生成问题, 提出基于最优传输理论的伪标签生成方法. 所提方法在将有标签信息作为生成过程引导的同时引入类别均衡约束, 在此基础上将半监督学习的伪标签生成过程转换成最优传输优化问题, 给出新的求解伪标签生成问题的形式. 为求解该优化问题, 引入 Sinkhorn-Knopp 算法进行近似快速求解, 避免不可计算问题. 所提伪标签生成方法作为半监督学习中的独立过程可结合当前一致性正则等半监督学习技巧构成完整的半监督学习过程. 最终, 在 CIFAR-10、SVHN、MNIST、FashionMNIST 这 4 大公共经典图像分类数据集上进行实验, 验证方法的有效性. 实验结果显示, 所提方法与当前先进的半监督学习方法相比, 均取得更优异的结果, 尤其是在标签情况较少的情况下提升显著.

关键词: 半监督学习; 伪标签生成; 最优传输; 图像分类; 深度学习

中图法分类号: TP18

中文引用格式: 翟德明, 沈斯娴, 周雄, 江俊君, 刘贤明, 季向阳. 基于最优传输理论的深度半监督学习伪标签生成算法. 软件学报, 2024, 35(11): 5196–5209. http://www.jos.org.cn/1000-9825/7054.htm

英文引用格式: Zhai DM, Shen SX, Zhou X, Jiang JJ, Liu XM, Ji XY. Pseudo-labeling Algorithm Based on Optimal Transport for Deep Semi-supervised Learning. Ruan Jian Xue Bao/Journal of Software, 2024, 35(11): 5196–5209 (in Chinese). http://www.jos.org.cn/1000-9825/7054.htm

Pseudo-labeling Algorithm Based on Optimal Transport for Deep Semi-supervised Learning

ZHAI De-Ming[1], SHEN Si-Xian[1], ZHOU Xiong[1], JIANG Jun-Jun[1], LIU Xian-Ming[1], JI Xiang-Yang[2]

[1](Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China)

[2](Department of Automation, Tsinghua University, Beijing 100084, China)

Abstract: Deep learning has been widely employed in many fields and yields excellent performance. However, this often requires the support of large amounts of labeled data, which usually means high costs and harsh application conditions. Therefore, with the development of deep learning, how to break through data limitations in practical scenarios has become an important research problem. Specifically, as one of the most important research directions, semi-supervised learning greatly relieves the data requirement pressure of deep learning by conducting learning with the assistance of abundant unlabeled data and a small number of labeled data. The pseudo-labeling method plays a significant role in semi-supervised learning, and the quality of its generated pseudo labels will influence the final

* 基金项目: 国家自然科学基金 (6207115, 61922027)
收稿时间: 2023-03-03; 修改时间: 2023-05-29; 采用时间: 2023-09-07; jos 在线出版时间: 2024-04-03
CNKI 网络首发时间: 2024-04-09

results of semi-supervised learning. Focusing on pseudo-labeling in semi-supervised learning, this study proposes the pseudo-labeling method based on optimal transport theory, which introduces the pseudo-labeling procedure constraint with labeled data as generation process guidance. On this basis, the pseudo-labeling procedure is converted to the optimization problem of optimal transport, which offers a new form for solving pseudo-labeling. Meanwhile, to solve this problem, this study introduces the Sinkhorn-Knopp algorithm for approximate fast solutions to avoid the heavy computation burden. As an independent module, the proposed method can be combined with other semi-supervised learning tricks such as consistency regularization for complete semi-supervised learning. Finally, this study conducts experiments on four classic public image classification datasets of CIFAR-10, SVHN, MNIST, and FashionMNIST to verify the effectiveness of the proposed method. The experimental results show that compared with the state-of-the-art semi-supervised learning methods, this method yields better performance, especially under fewer labeled data.

Key words: semi-supervised learning; pseudo-labeling; optimal transport; image classification; deep learning

1 引 言

在计算机视觉领域, 深度学习技术取得了令人瞩目的成功, 广泛应用于图像分类[1]、目标检测[2]、图像分割[3]等任务并取得超越传统方法的优异效果. 然而, 深度学习模型训练往往需要大规模标注数据的支持, 而在现实应用中, 标注数据的获取需要昂贵的时间和经济成本投入, 且获取难度大, 尤其是在医学、军事等特殊领域. 因此, 当前计算机视觉领域一个热点研究问题是如何克服深度学习对于大量标注数据的依赖. 在领域内, 众多研究者提出了一些有效策略, 如: 半监督学习[4–7]、小样本学习[8–10]、自监督学习[11–13]等. 其中, 半监督学习采取基于部分有标记和剩余无标记的数据集进行训练的方式缓解深度学习对于标注数据的大量需求, 成为领域内重要的研究方向.

伪标签生成是半监督学习的重要组成部分. 伪标签生成即为无标记部分的数据生成标签, 这种非原有数据集中人工标注的标签被叫作伪标签 (pseudo label). 半监督学习使用生成的伪标签作为训练数据的一部分, 并可结合其他半监督学习方法进行后续的模型训练. 大多数半监督学习方法使用自训练 (self-training) 的形式进行训练, 即单个模型使用自身生成的伪标签进行训练. 该训练方式会面临确认偏置问题, 即在模型为无标记数据生成的伪标签噪声过大的情况下, 由于模型无法纠正自身产生标签的错误, 随着训练过程进行错误逐渐累积, 最终导致半监督学习模型的退化. 因此, 本文将着重研究如何为无标记数据生成更加准确的伪标签, 以提升半监督学习的训练性能.

为提高生成伪标签的质量, 当前半监督学习研究多采用置信度筛选的方式, 如 FixMatch[14], 对模型预测概率结果使用阈值进行筛选, 以保留符合条件的高置信度标签. 该类方法本质上直接依赖模型本身对于无标记样本的概率预测结果进行标签判定. 我们希望引入更多的其他先验信息, 以生成更加优质的标签. 具体而言, 对于半监督学习问题场景, 我们考虑在伪标签生成的过程中引入有标记数据的标签信息作为引导, 以生成更加准确的伪标签.

除此之外, 当前半监督学习方法直接使用模型预测类别概率作为伪标签, 而忽略了最终的标签生成结果的类别分布与原始数据的类别分布的一致性. 具体来说, 由于在训练学习过程中, 必然会存在简单类别和困难类别之分, 其中简单类别更容易被学习且更易获得高置信度的类别概率预测结果, 而高类别概率的更容易通过置信度筛选, 从而导致最终得到的标签生成结果有简单类别样本多、困难类别样本少这样不平衡分布的情况. 进而, 由于训练数据类别分布与测试数据类别分布不一致, 使用这些生成的伪标签进行半监督学习训练获得的模型最终的测试分类结果也会不甚理想. 为了缓解该问题, 我们将类别分布作为先验知识引入标签生成的过程中, 鼓励生成类别均衡的伪标签信息.

最后, 受到无监督表示学习工作 self-labelling[15]的启发, 我们将半监督学习中的无标记数据进行标注的过程形式化为最优传输优化问题的求解过程, 并引入上述有标签信息以及类别均衡信息作为约束条件. 由此, 我们可以实现在结合上述多种先验信息的同时, 仅通过求解一个优化式来得到理想的伪标签结果. 除此之外, 为了适应具有大规模训练样本的深度学习场景, 我们还将考虑更加高效的求解方法来得到上述基于最优传输的优化问题的解, 以达到节省计算资源的目的.

本文的贡献包含以下几个方面.

● 提出基于最优传输理论的伪标签生成方法, 使用求解最优传输优化问题的新形式来生成伪标签.

● 在伪标签生成过程中引入有标签信息以及类别均衡约束引导, 以提高最终生成的伪标签的质量.

● 在 CIFAR-10、SVHN 等经典图像分类数据集上验证了所提方法的有效性, 对比当前主流半监督学习方法均取得一定优势.

本文第 2 节回顾当前主流的半监督学习方法. 第 3 节介绍本文相关背景知识. 第 4 节详细介绍本文所提出的基于最优传输理论的伪标签生成方法. 第 5 节给出本文方法的具体实验设置以及在多个经典图像分类数据集上的测试与对比结果. 第 6 节对本文工作进行全面总结.

2 相关工作

当前已有大量工作研究如何使用半监督学习提升图像分类性能. 参考 Yang 等人[16]对半监督学习工作的总结, 其中包括基于图的方法[17–19], 生成建模[20–22]等的工作, 本文主要关注其中两种先进方法: 基于伪标签生成和一致性正则的半监督学习方法.

2.1 伪标签生成

该类方法利用有标记数据集, 通过预测函数本身或其他启发式方法, 为未标记数据集样本生成标签. 然后使用这些伪标签与其对应样本和标记数据集一起用作训练信息, 进行深度学习模型训练.

Lee 等人[7]提出了一种简单且高效的方法, 使用模型对无标签数据的分类概率预测伪标签. Iscen 等人[23]结合图的相关知识, 使用标签传播的知识进行伪标签的生成, 根据样本特征之间的相似性, 参考相邻节点数据的标签进行标签生成. 为了解决普通伪标签生成方法会对错误的标签进行过拟合并产生确认偏置的问题, Arazo 等人[24]提出可以结合数据增广方法 Mixup[25]并设置每个训练批次中有标签数据的最少数量作为正则手段来减少偏置. Pham 等人[26]则引入了元学习的思想, 构造学生-老师的训练框架结构, 教师模型基于元伪标签算法, 生成标签并促进学生模型的提升, 在基于学生模型在外部验证集上的结果进一步更新改善教师模型, 重复上述过程以得到最终的训练结果. 该类使用单一模型的自训练方式由于训练初期模型本身质量较差且很难更正自身产生的错误, 容易存在确认偏置问题. 为缓解该问题, 通常需要筛选出高置信度的伪标签结果. 当前工作通常会选择设置固定阈值[14,27]或使用启发式方法设置动态的阈值[28,29]来进行伪标签筛选. 而也有工作采用基于分歧 (disagreement-based) 的训练模式, 通过多模型联合训练, 利用分歧带来的互补性提升性能, 缓解确认偏置. Co-training[30,31]基于数据的两个不同视图进行分歧训练. 而 Dong 等人[32]提出的 Tri-Net 利用 3 个网络进行伪标签生成, 根据两个模型一致认同的结果得到最终的伪标签.

在当前的半监督方法中, 该类方法已成为较为重要的组成部分与其他半监督学习技巧一起组合使用构成更高效的半监督训练策略[4,14,33]. 本文的方法研究正是聚焦该部分, 旨在提高生成的伪标签质量, 为半监督学习改善性能.

2.2 一致性正则

一致性正则方法主要靠利用未标记样本来强制训练模型符合聚类假设, 即学习到的决策边界必须位于低密度区域, 从而达到半监督学习的目的. 该类方法的核心思想是对样本添加扰动不会明显的影响模型预测, 比如同一样本的不同增广具有一致的预测等. Rasmus 等人[34]提出在使用 Ladder Networks[35]的基础上额外添加编、解码器, 计算干净样本和带噪样本的译码器特征之间的 MSE 损失作为无标签样本的损失. Π 模型[36]利用神经网络自然的随机性, 如: 数据增广、网络随机丢弃技术 (dropout), 对输入数据造成干扰, 以构造一致性正则损失. 为避免训练不稳定性, Laine 等人[36]采用了时间集成 (temporal ensembling) 的策略来改进 Π 模型, 记录历史的预测结果的指数移动平均值 (EMA) 作为无标签样本的目标. Mean teacher[5]通过引入教师网络并方法通过计算学生模型参数的指数移动平均值作为教师网络的参数避免单一模型造成的确认偏置问题. Miyato 等人[6]提出了虚拟对抗训练 (virtual adversarial training, VAT), 引入对抗样本作为无标签样本的一致性目标使模型具有更好的平滑性. 双学生模型 (dual-student)[37]将 Mean teacher 的教师模型部分用学生模型替代, 构造了两个学生互相学习的框架结构, 学生模型具有独立的初始状态和优化路径避免教师与学生模型耦合造成性能瓶颈. Xie 等人[38]提出引入无监督数据增广

(unsupervised data augmentation, UDA), 如 AutoAugment[39]等高质量的数据增广替代噪声作为扰动进行一致性训练. 总的来说, 该类方法主要通过约束在干扰下的模型输出一致性来达到对于无监督信息的利用, 通常从输入数据、网络结构、训练过程等方面入手研究设计更好用的扰动提高模型鲁棒性和泛化性. 本文方法并依赖该模式的半监督训练方法, 但是由于本文方法与该类方法并不冲突, 且可互相补充, 因此为达到更好的半监督训练效果, 将该方法与本文方法进行组合使用进一步提升网络性能.

3 预备知识

对于半监督问题, 训练数据集表示为 D, 其中包括: 有标记子集 D_s 和无标记子集 D_u. 我们定义 $D_s=\{(x_i,y_i)|i=1,\dots,M\}$ 表示具有 M 个样本的有标记子集, 其中 x_i 为输入样本, $y_i=[y_i^1,\dots,y_i^K]\subseteq\{0,1\}^K$ 是 K 个分类的对应独热码(one-hot) 标签. $D_u=\{(x_i)|i=M+1,\dots,N\}$ 表示具有 $N-M$ 个样本的无标记子集, 该部分数据集并没有对应输入的标签. 在训练集 D 中, 有标记部分数据量远小于无标记部分数据量, 即 $M\ll N-M$.

当前主流半监督学习方法多采用结合有标记部分损失与无标记部分损失进行模型优化的方法, 即最终的损失函数 $\mathcal{L}$ 可以表示为两者的线性加权函数:

$$\mathcal{L}=\mathcal{L}_s+\lambda_u\mathcal{L}_u \tag{1}$$

其中, $\mathcal{L}_s$ 为有标记部分损失, $\mathcal{L}_u$ 为无标记部分损失, λ_u 为两者折中的权重参数.

对于无标记数据的损失函数, 通常会引入一致性正则与伪标签生成技术结合进行优化计算. 一致性正则基于相同的输入图像即使面对少量扰动模型也会输出相似结果的原则, 鼓励在数据增广、模型扰动等方式干扰前后的同一输入样本的输出相同. 如 Π-Model[36]、Mean teacher[5]、MixMatch[4], 可以将无标记损失表示为:

$$\frac{1}{\mu B}\sum_{i=1}^{\mu B}\|p(w(x_i))-p(w(x_i))\|_2^2 \tag{2}$$

其中, $p(x)$ 表示模型对于输入样本 x 的预测类别分布, $w(\cdot)$ 为增广函数, 两者均为随机函数. μB 为无标记数据的批大小, μ 为固定常数系数, 用以控制有标记数据与无标记数据批大小比.

FixMatch[14]在此基础上引入强、弱两种增广方法, 可分别表示为 $\mathcal{A}(\cdot)$ 和 $\alpha(\cdot)$, 以此为同一输入样本生成两个不同的视图进行一致性正则, 进一步提高了半监督学习性能. 因此无标记数据部分损失可以表示为:

$$\mathcal{L}_u=\frac{1}{\mu B}\sum_{i=1}^{\mu B}\mathbb{1}(\max(q_i)\geqslant\tau)H(\tilde{y}_i,p(\mathcal{A}(x_i))) \tag{3}$$

其中, $\tilde{y}_i=\arg\max(q_i)$, q_i 为对样本弱增广生成的伪标签, $q_i=p(\alpha(x_i))$ 这里定义为使用模型预测分类结果, $\mathbb{1}$ 为对应维度大小的值全为 1 的向量. 通过函数 $H(\cdot)$ 计算样本强增广的模型输出与上述生成的伪标签交叉熵损失值, 即为基础的无标记部分损失函数. 除此之外, 为了确保模型在训练过程中, 使用的伪标签结果具有一定的置信度, 这里使用置信度阈值 τ, 筛选出标签最大类别概率大于 τ 的样本进行损失计算.

对于有标记部分损失则正常计算标签与样本输出的交叉熵损失:

$$\mathcal{L}_s=\frac{1}{B}\sum_{i=1}^{B}H(y_i,p(\alpha(x_i))) \tag{4}$$

其中, $p(\alpha(x_i))$ 为有标签样本的弱增广的模型输出的类别概率值, 这里用该概率值与其对应的标签 y_i 进行交叉熵损失计算.

4 所提出的方法

本节将详细介绍本文提出的伪标签生成方法. 针对第 3 节中定义的半监督学习问题, 为其中的无标记数据生成高质量的伪标签. 值得注意的是, 为了方便后续的半监督学习计算, 这里我们希望得到的伪标签为软标签形式(soft-label), 即对于每个样本生成的标签 Q_i, Q_i 为 K 维向量, $Q_i=[Q_i^1,\dots,Q_i^K]$, $0\leqslant Q_i^K\leqslant 1$, Q_i^K 表示样本 i 被归类为类别 k 的概率, 而非独热码形式, 即 $Q_i^K\in[0,1]$.

第 4.1 节和第 4.2 节给出本文提出的伪标签生成优化问题的具体定义以及求解过程, 第 4.3 节介绍如何结合我们的伪标签生成方法进行完整的半监督学习训练.

4.1 伪标签生成优化问题定义

对于一般的有监督学习问题, 模型的学习优化问题可以定义为最小化通过网络模型得到的预测分类概率结果与样本真实标签的交叉熵损失, 具体可表示为:

$$\min_{p}\left(-\sum_{j=1}^{K}\sum_{i=1}^{N}y_i^j\log p\left(y_i^j\mid x_i\right)\right) \tag{5}$$

其中, 分类预测概率 $p(y=\cdot\mid x_i)$ 为对参数为 θ 的深度模型 $f_\theta(\cdot)$ 的输出结果, 并使用 *Softmax* 函数进行激活映射到 0–1 范围内, 具体可表示为: $p(y=\cdot\mid x_i)=Softmax(f_\theta(x_i))$.

而对于半监督学习问题, 其中无标记部分数据的真实标签我们无法获得, 也就无法使用上式直接进行求解. 因此, 这里我们将标签用后验分布 $q(j\mid x_i)$ 替代. 优化问题可改写为:

$$\min_{p,q}\left(-\sum_{j=1}^{K}\sum_{i=1}^{N}q(j\mid x_i)\log p(j\mid x_i)\right) \tag{6}$$

为利用有标记部分的标签信息, 这里我们引入标签数据约束条件 $q(j\mid x_i)=y_i^j,\ \forall i\in[1,\ldots,M]$, 表示有标记部分数据的后验概率应该等于其本身的标签结果. 除此之外, 为了鼓励训练得到的标签生成结果能够趋于类别平衡, 这里增加约束项: $\sum_{i=1}^{N}q(j\mid x_i)=\dfrac{N}{K},\ \forall j\in[1,\ldots,K]$, 即表示 N 个数据节点的后验概率应该均匀地分布在 K 个分类上. 在通过使用惩罚函数, 将上述有标记部分数据的约束转换成原优化问题的一部分后, 原半监督学习的优化问题可以重新写为以下形式:

$$\begin{cases}\min\limits_{p,q}\left(-\sum\limits_{j=1}^{K}\left(\sum\limits_{i=1}^{N}q(j\mid x_i)\log p(j\mid x_i)+\sum\limits_{i=1}^{M}q(j\mid x_i)\log y_i^j\right)\right)\\ \text{s.t.}\ \sum\limits_{i=1}^{N}q(j\mid x_i)=\dfrac{N}{K},\ \forall j\in[1,\ldots,k]\end{cases} \tag{7}$$

其中, 第 1 项为原优化项, 第 2 项为有标记部分数据约束项转换成的惩罚项.

然而, 上式最优化求解困难. 幸运的是, 当 $p(\cdot)$ 固定时, 可以将其看作最优传输问题的一种形式化表达, 进而可以用相对高效的方式解决该组合优化问题. 具体来说, 我们令 P' 为一个 $K\times N$ 的矩阵, 具体定义为:

$$P'_{ji}=\begin{cases}p(j\mid x_i)\cdot y_i^j, & i\in[1,\ldots,M]\\ p(j\mid x_i), & i\in[M+1,\ldots,N]\end{cases} \tag{8}$$

即当下标 i 指示属于有标记数据部分时, 只保留对于真实标签 y_i 指示为正确类别的概率, 其余为 0; 对于下标 i 指示属于无标记部分数据时, 保留由网络模型输出得到的原有的类别概率向量即可.

令 Q 为 $K\times N$ 的矩阵, 即所求的所有样本的伪标签信息, 具体定义为 $Q_{ji}=q(j\mid x_i)$. 根据文献 [40] 中定义, 将 Q 松弛为传输多面体 (transportation polytope) 组成元素:

$$U(r,c)=\{Q\in R_+^{K\times N}\mid Q\mathbb{1}=r, Q^{\mathrm{T}}\mathbb{1}=c\} \tag{9}$$

其中, r 和 c 分别表示矩阵 Q 在行与列上的投影. 根据公式 (7) 条件约束, 生成的标签应保证类别均衡性. 因此 r 和 c 可定义为: $r=\dfrac{N}{K}\cdot\mathbb{1},\ c=\mathbb{1}$.

因此, 原半监督优化问题 (7) 可等价于:

$$\min_{Q\in U(r,c)}\langle Q,-\log P'\rangle \tag{10}$$

其中, $\langle\cdot\rangle$ 为矩阵内积计算. 经过上述问题的转换, 我们将半监督问题的伪标签生成过程形式化为求解一个最优传输优化问题, 并巧妙地引入了有标签信息引导以及类别平衡约束. 上式的求解过程可直观理解为, 得到在传输代价为 $-\log P'$ 的情况下的最优传输方案 Q, 即我们所需的伪标签结果矩阵. Q_i 为单个样本的伪标签, Q_{ji} 即表示样本 i

为类别 j 的概率.

4.2 最优化问题求解

本节将给出上述最优传输问题的求解方法. 当涉及数百万的数据样本以及更多类别数目时, 上述的最优化问题若使用传统的线性规划求解算法计算复杂度将难以承受. 因此, 这里我们采用 Sinkhorn-Knopp 算法[40]进行快速近似求解.

Sinkhorn-Knopp 算法为当前较为流行的一种最优传输问题求解方法. 它使用熵正则化项将原最优传输问题变成一个强凸性的近似问题. 经过正则化后的最优传输问题可以使用一系列矩阵乘法来进行求解, 因而最优传输问题的求解可以充分利用 GPU 进行矩阵计算加速, 提高求解速度. 具体来说, 就是求解如下问题: $\min\limits_{Z\in U(r,c)}\langle Z,M\rangle-\eta^{-1}H(Z)$, 其中 $\eta>0$, $H(Z)$ 为熵函数. 可以理解为: 通过该正则项最优传输问题的求解鼓励利用数量多但小流量的传输路径, 而不是数量少但是流量大的路径进行传输, 从而达到减少计算复杂度的目的. 基于以上经过熵正则化后的最优传输问题, 使用 Sinkhorn-Knopp 迭代求解得到最终的解. 其解可以写成以下形式 $Z_{i,j}=u_iK_{i,j}v_j$, 其中 $K_{i,j}=\mathrm{e}^{-\eta M_{i,j}}$. 由于 M 已知, 因此只需要通过迭代得到 u, v 的收敛最优解, 即可得到最优传输策略 Z. u, v 的迭代步骤可简述为如下过程, 其中 t 标记迭代次数: $u^{t+1}=\dfrac{r}{Kv^t}$, $v^{t+1}=\dfrac{c}{K^{\mathrm{T}}u^{t+1}}$.

基于以上的基础理论, 对于最优化目标函数 (7), 我们引入正则项 $KL(Q\,\|\,rc^{\mathrm{T}})$, $KL(\cdot)$ 为 KL 散度计算, $1/\lambda$ 为正则项系数, 原优化问题改写为:

$$\min_{Q\in U(r,c)}\langle Q,-\log P'\rangle-\frac{1}{\lambda}KL(Q\,\|\,rc^{\mathrm{T}}) \tag{11}$$

通过正则项的引入, 原本的最优传输问题的近似解可以表示为:

$$Q=\mathrm{diag}(\alpha)\mathrm{e}^{-\lambda\left(-\log P'\right)}\mathrm{diag}(\beta)=\mathrm{diag}(\alpha)P'^{\lambda}\mathrm{diag}(\beta) \tag{12}$$

即 $q(j\mid x_i)=\alpha_jP'^{\lambda}_{ji}\beta_i$, 具体元素计算可表示为: $q(j\mid x_i)=\begin{cases}\alpha_j\cdot y_i^j\cdot\beta_i, & i\in[1,\ldots,M]\\ \alpha_j\cdot P^{\lambda}_{ji}\cdot\beta_i, & i\in[M+1,\ldots,M+N]\end{cases}$, $\alpha\in R^N$ 并且 $\beta\in R^k$. 通过以下表达式:

$$\forall j:\alpha_j\leftarrow\left[P'^{\lambda}\beta\right]_j^{-1},\ \forall i:\beta_i\leftarrow\left[\alpha^{\mathrm{T}}P'^{\lambda}\right]_i^{-1} \tag{13}$$

迭代更新 Q 的行列边界值 α 和 β, 当迭代收敛后, 即可得到近似解.

当 λ 趋于无穷大时, 优化公式 (11) 将完全等价于公式 (10). 即当 λ 越大公式 (11) 的求解结果将越接近原优化问题的解, 但是这需要更长的迭代时间. 因此需要权衡准确性和计算速度, 选择恰当的参数 λ. 详细的伪标签生成算法过程可见算法 1. 实验中收敛的条件是根据错误最小阈值 *err_lower* 和最大迭代次数 *maxIterNum* 两个条件共同决定的.

算法 1. 基于最优传输的伪标签生成算法.

输入: 无标记数据分类概率 P_u, 有标记数据分类概率 P_l 及标签 Y, 近似算法正则项参数 λ, 错误最小阈值 *err_lower*, 最大迭代次数 *maxIterNum*;

输出: 无标记数据伪标签矩阵 Q.

// 第 1 阶段: 初始化数据

1. $P=cat(P_u,P_l*Y)$
2. $P=P^{\lambda}$
3. $\alpha=0_n,\ \beta=0_k$
4. $r=N/K,\ c=1$
5. $err=IFF,\ iterNum=0$

// 第 2 阶段: 迭代优化行列边界 α、β

6. **WHILE** $err<err\_lower$ **OR** $iterNum>maxIterNum$:
7. $\quad\alpha=r/(P\cdot\beta)$

8. $new\_beta = c/(\alpha.T \cdot P).T$

9. $err = sum(\beta/new\_beta - 1)$

10. $\beta = new\_beta$

11. **END WHILE**

// 第 3 阶段: 计算最终的伪标签结果

12. $Q = P \cdot \beta$

13. $Q = Q.T \cdot \alpha$

14. **RETURN** Q

4.3 半监督学习方法

该部分主要阐述整个半监督学习网络训练过程, 介绍如何结合使用上述伪标签生成方法的结果进行完整的半监督学习训练.

我们使用的半监督训练方法与当前主流半监督方法基本一致, 在标签生成的基础上, 融入一致性正则方法[14]. 最终损失函数计算采用与公式 (1) 相同的加权有标记与无标记部分损失的方式, 其中有标记损失使用公式 (4) 进行计算. 而对于无标记损失部分, 我们使用本文提出的伪标签生成方法为弱增广数据进行标签预测, 即形式上将公式 (3) 改写为:

$$\mathcal{L}_u = \frac{1}{\mu B}\sum_{i=1}^{\mu B}\mathbb{1}(\max(Q_i) \geqslant \tau) H(\tilde{y}_i, p(j \mid \mathcal{A}(x_i))) \tag{14}$$

其中, $\tilde{y}_i = \arg\max(Q_i)$, Q 即上述研究的伪标签生成方法产生的 $K \times N$ 伪标签矩阵. 公式 (14) 中 $Q_i = Q(X = \alpha(X))[i]$ 表示对于样本的弱增广视图 $\alpha(x_i)$ 的生成的软标签 (soft-label) 形式的伪标签. 通过 $\arg\max(\cdot)$ 函数将该样本概率向量转换成独热码 (one-hot) 形式.

对于半监督学习整体的训练方案, 可以概括为对伪标签结果与模型参数的迭代优化. 核心步骤主要包括以下两步.

1) 使用伪标签矩阵 Q, 计算损失函数 $\mathcal{L}$, 并使用梯度下降法优化模型参数 θ.

2) 给定模型 θ, 使用伪标签生成方法得到优化后的伪标签 Q.

重复上述 1)、2) 步骤, 直至算法收敛或达到结束条件.

5 实验验证

本节将对上述提出的算法在图像分类数据集上进行实验验证, 验证使用上述基于最优传输理论的伪标签生成方法进行半监督学习的有效性, 并与当前主流半监督学习方法进行对比. 此外, 将着重分析伪标签生成部分算法作用, 以及对其中关键参数作用进行评估以及衡量算法代价等.

5.1 实验实现与评价指标

本节将详细介绍实验的实现细节, 包括超参设置、网络训练与具体实现的细节, 以及与现有典型方法进行性能对比的评估准则等.

● 数据集

实验主要涉及 4 种图像分类数据集, 包括: CIFAR-10、SVHN、MNIST、FashionMNIST. 对于半监督学习问题背景, 将参考普遍使用的划分比例对数据集进行划分, 只保留部分标签信息, 以测试性能. 具体介绍如下.

CIFAR-10[41]: 数据集为具有 10 个分类的自然图像数据集, 并且包含 50 000 训练数据和 10 000 的测试数据. 数据集中的图像均为 32×32 RGB 图像. 按照普遍设置, 我们将从数据集中分别随机抽取 40、250、4 000 张图像作为有标记数据, 剩余作为无标记数据, 由此构成半监督数据集进行后续实验验证.

SVHN[42]: 街景房屋编号数据集, 每张图像均为 32×32 RGB 图像且均包含一组 0–9 的阿拉伯数字. 数据集包含

73257 个训练数据和 26032 个测试数据, 另有 531131 个附加训练数据. 按照普遍设置, 我们分别随机抽取 40、250、1000 张图像作为有标记数据集, 其余为无标记部分.

MNIST[43]: 手写数字图像分类数据集, 图像内容为 0–9 之间的数字, 共有 60000 张训练图像和 10000 张测试图像, 每张图像为 28×28 的黑白图像. 为得到半监督学习设置, 我们将从中随机抽取 40、250、1000 张图像作为有标记部分数据集, 其余将舍弃标签作为无标记部分.

FashionMNIST[44]: 数据集为具有 10 个分类服装图像分类数据集, 总共包含 60000 张训练集图像和 10000 张测试图像, 图像均为 28×28 的黑白图像. 我们将从训练数据集中随机抽取 40、250、1000 张图像作为有标记部分数据集, 其余作为无标记部分数据集.

● 实现细节

网络结构: 为了对比的公平性, 对于 MNIST 和 FashionMNIST 上的测试方法我们使用 4 层 CNN 网络结构, 结构具体为: $Conv(1,32,3) \to BatchNorm$ [45] $\to ReLU$ [46] $\to MaxPool(2,2) \to Conv(32,64,3) \to BatchNorm \to ReLU \to MaxPool(2,2) \to Linear \to BatchNorm \to ReLU \to Linear$; 对于 CIFAR-10 与 SVHN 上的测试方法, 使用 WideResNet (WRN) 28-2[47]结构来进行网络训练.

参数设置: 对于在 CIFAR-10、SVHN 上测试的半监督方法, 为了更加公平的比较, 我们这里使用与文献 [14] 相同的参数设置. 具体来说, 实验使用动量参数为 0.9 的标准的随机梯度下降 (SGD) 方法作为优化器; 学习率设置上, 使用余弦学习率调整策略 $\eta = \eta_0 \cos\left(\frac{7\pi t}{16T}\right)$, η_0 为初始学习率, 实验时设置为 0.03, t 为当前训练迭代次数, T 为训练迭代总次数, 实验中总的迭代次数为 2^{20} (共包括 2^{10} 个迭代轮次, 每个迭代轮次批量访问数据集并训练模型 2^{10} 次). 网络的训练模型的权重衰减值为 10^{-3}, 除此之外为了优化实验性能, 这里使用指数加权平均策略并设置其动量参数为 0.999. 所有数据集的有标记数据部分批大小设置为 64, 而无标记数据集的批量大小为有标记数据集的 μ 倍, 在实验中 μ 设置为 7, 即批大小为 448. 方法中为提高标签置信度的固定阈值 τ 设置为 0.95. 而有标记部分损失和无标记部分损失的加权参数 $\lambda_u = 1$. 基于最优传输算法的标签生成方法算法实现中, 正则项系数 λ 设为 500. 在实验测试中, 迭代更新 Q 的参数 $maxIterNum$=1000, err_lower=1E–5. 此外, 本节实验中涉及一致性正则化的均使用相同的增广方法, 具体来说: 弱增广策略包括 4 个像素的随机平移和概率为 0.5 的随机水平翻转 (仅对 CIFAR-10 数据集进行, 不包括 SVHN); 强增广则使用 RandAugment 策略[48]. 对于在 MNIST、FashionMNIST 上进行的标签生成方法的比较, 则在上述设置的基础上进行调整. 具体为基于 66 余弦学习率调整策略的初始学习率 η_0 设置为 0.01, 训练总迭代次数为 38400 (共包括 150 个迭代轮次, 每个迭代轮次批量访问数据集并训练模型 256 次). 除此之外, 该部分基于标签生成方法的训练方法并不需要对伪标签基于置信度进行筛选, 因此不设置固定阈值.

● 对比方法与评判标准

我们将与目前典型的半监督算法进行分别在上述数据集上进行比较, 以验证算法的有效性. 为了方便比较, 这里将方法分为以下两类.

基于综合性的半监督方法: 该类方法将在 CIFAR-10、SVHN 数据集上使用 WideResNet 进行训练测试, 主要包括 Π-Model[36]、Pseudo-label[7]、Mean teacher[5]、UDA[38]、MixMatch[4]、ReMixMatch[33]以及目前最佳性能工作 FixMatch[14].

基于标签生成的半监督方法: 该类方法将在 MNIST、FashionMNIST 数据集上使用 CNN 进行训练测试, 对比方法为 Pseudo-label[7].

评判标准: 对于每个数据集, 我们使用官方训练/测试分区并使用 Top-1 准确率作为评估标准. 对于每项指标, 通过 3 次实验, 给出平均值和标准差.

5.2 半监督学习方法对比

该部分将所提出的方法在 CIFAR-10 和 SVHN 数据集上与现在主流的半监督方法进行对比. 该部分使用 WideResNet 28-2 网络, 使用第 3.4 节中结合一致性正则的综合性半监督学习方法. 为了比较的公平性, 这里对比

的方法也是当前 SOTA 综合性半监督方法, 其中由于 Pseudo-label、Π-model、Mean teacher 在数据集具有 250 个标签情况下表现不佳, 这里不再给出其在 40 个标签情况下的测试结果.

如表 1 所示, 本文所提出的方法无论是在 CIFAR-10 还是 SVHN 数据集上, 均较其他方法有一定程度的性能提升, 尤其是在具有 40 个与 250 个有标签数据的情况下提升较为明显. 例如, 我们的方法在 CIFAR-10 具有 40 个标签数据的情况下, 相较 FixMatch[14]的实验结果提升约 9%, 基本上与该方法在具有 250 个标签数据的情况下的实验效果接近. 同样, 在 SVHN 数据集上也有相似实验结果, 本文方法在具有 40 个标签数据的情况下达到了 97.82% 的准确率, 超过 FixMatch 在 250 个标签情况下的结果. 我们方法在 SVHN 数据集具有 1 000 个有标签数量的划分情况下, 对比现有的其他半监督学习方法也有提升但提升程度不大, 在 CIFAR-10 数据集具有 4 000 个样本情况下比之 FixMatch 准确度有所减少. 这里我们认为该结果可能是由于得到的准确度已经接近全监督设置下的预测准确度, 导致这些标签划分下整体提升空间有限, 但是从多次实验的方差结果来看, 本文方法比较其他方法更具有稳定性. 因而, 总体来说, 我们所提出的方法对于稀疏半监督学习尤其有效.

表 1　标签量不同时, 本文算法及其他半监督学习算法在数据库 CIFAR-10, SVHN 上的测试准确率对比 (%)

方法	CIFAR-10			SVHN		
	40	250	4 000	40	250	1 000
Π-model[36]	—	45.74±3.97	58.99±0.38	—	81.04±1.92	92.46±0.36
Pseudo-label[7]	—	50.22±0.43	83.91±0.28	—	79.79±1.09	90.06±0.61
Mean teacher[5]	—	67.68±2.30	90.81±0.19	—	96.43±0.11	93.58±0.07
MixMatch[4]	52.46±11.50	88.95±0.86	85.58±0.10	57.45±14.53	96.02±0.23	96.50±0.28
UDA[38]	70.95±5.93	91.18±1.08	95.12±0.18	47.37±20.51	94.31±2.76	97.54±0.24
ReMixMatch[33]	80.90±9.64	94.56±0.05	95.28±0.13	96.66±0.20	97.08±0.48	97.35±0.08
FixMatch[14]	93.01±3.30	94.99±0.60	**95.82±0.10**	96.06±2.29	97.77±0.25	97.88±0.16
Ours	**94.68±0.54**	**95.19±0.27**	95.79±0.03	**97.82±0.05**	**98.12±0.03**	**98.13±0.02**

5.3　半监督训练过程伪标签生成情况分析

为了进一步研究本文提出的伪标签生成方法在半监督学习过程中的作用, 我们对半监督学习过程中对无标记部分数据生成的伪标签情况进行了统计分析, 分别记录了在 CIFAR-10, SVHN 两种数据集具有 40 个标签数据的情况下, 训练过程中生成标签的正确率、召回率、在测试集上准确率这 3 种数据, 同时将我们的方法与半监督学习方法 FixMatch 的标签生成进行对比. 图 1 展示了最终的统计数据结果.

如图 1(a)(b) 所示, 这里我们对比了方法在训练过程中生成的伪标签的正确率, 即对于无标记数据进行标注 (即超过置信度阈值) 的标签数量中预测正确的标签占比. 可以看到相比较 FixMatch 方法, 除了在 CIFAR-10 (40 个标签) 数据集上中间训练部分正确率略低以外, 整体上本文方法的正确率略高于 FixMatch 方法.

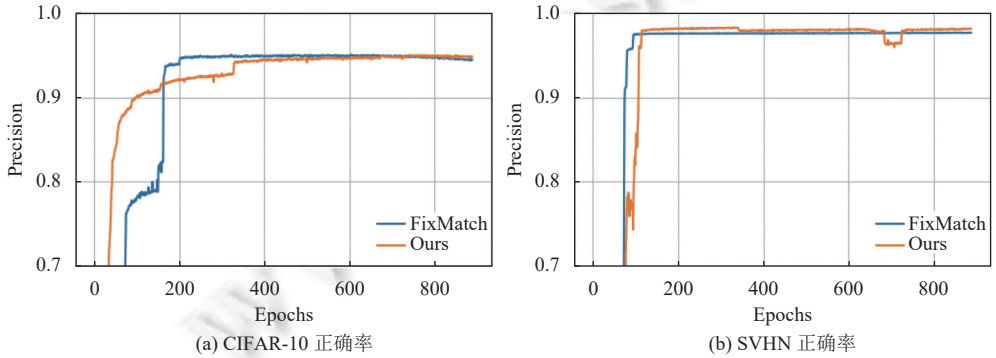
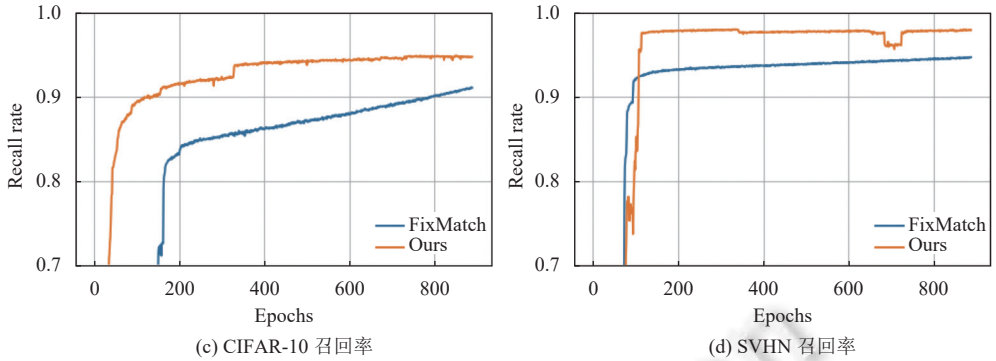
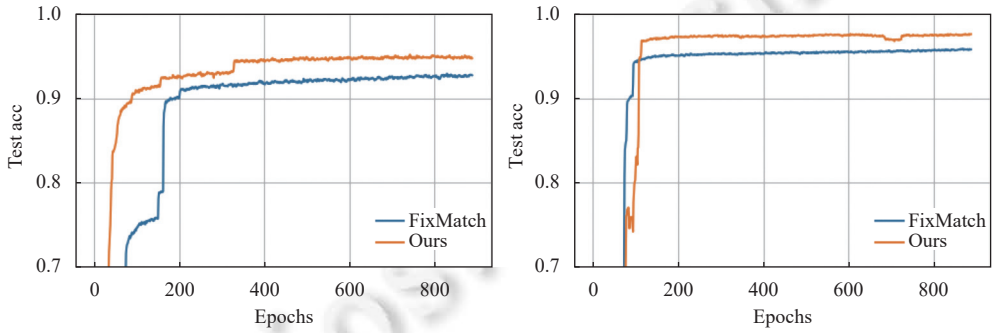


图 1　在 40 个标签数据时, 半监督方法训练过程中生成伪标签情况对比 (数据集 CIFAR-10 和 SVHN)



(c) CIFAR-10 召回率 (d) SVHN 召回率



(e) CIFAR-10 准确率 (f) SVHN 准确率

图 1 在 40 个标签数据时, 半监督方法训练过程中生成伪标签情况对比 (数据集 CIFAR-10 和 SVHN)(续)

图 1(c)(d) 对比方法正确标签的召回率, 即对于无标记数据生成的准确的标签数目与无标记数据集总数的比值. 从召回率的对比结果来看, 本文方法在召回率方面有着明显优势, 相比 FixMatch 方法可以相对生成更多的正确标签数据, 从而能够提高整体半监督学习的效果. 也正如我们的预期, 如图 1(e)(f) 所示, 使用我们的方法进行半监督训练的模型在测试集上具有更高的准确率.

综上所述, 就训练过程的生成的伪标签情况来看, 使用我们的方法进行伪标签生成, 可以生成更多的正确标签数据, 且生成的所有伪标签的准确率不低, 从而可以促使模型能够学习到更多的信息, 以达到更好的测试效果.

5.4 伪标签生成方法对比

本节在 MNIST 和 FashionMNIST 数据集上进行实验, 将我们的方法与传统的伪标签生成方法进行对比, 以单独验证本文中伪标签生成方法的有效性. 为了保证实验的公平性, 所有实验统一使用 CNN 网络框架, 且统一方法框架仅使用生成标签进行训练, 不包括一致性正则等约束优化. 实验中直接使用本文标签生成优化方法替代框架中的标签生成部分.

如表 2 所示, 本文的标签生成方法相对于传统的标签生成方法确实有所提升, 在 MNIST、FashionMNIST 上的实验结果均优于普通的标签生成方法. 例如, 对于 MNIST 数据集, 在仅具有 40 个标签数据的情况下, 本文的方法的最终结果比较提高了约 3%, 接近 Pseudo-label 在具有 250 个有标签数据的情况下的实验结果.

表 2 仅伪标签生成进行半监督学习在 MNIST、FashionMNIST 上测试准确率对比 (%)

数据集	方法	标签量		
		40	250	1 000
MNIST	Pseudo-label	94.68±0.21	98.07±0.08	98.21±0.05
	Ours	**97.86±0.07**	**98.25 ±0.10**	**98.42 ±0.07**
FashionMNIST	Pseudo-label	66.16 ±0.05	82.03 ±0.11	84.99 ±0.07
	Ours	**73.04 ±0.29**	**82.12 ±0.07**	**85.21±0.17**

5.5 伪标签生成方法结果可视化分析

图 2 展示了对 Pseudo-label 和我们的基于最优传输的伪标签生成方法的可视化结果, 图 2(a) 为 Pseudo-label 方法, 图 2(b) 为基于最优传输理论的伪标签生成算法. 该部分使用 t-SNE[49]对在 MNIST 数据集具有 40 个有标签数据的情况下通过上述方法进行半监督训练的模型在测试集上的特征进行可视化展示, 每个颜色代表 1 个类别. 从可视化结果可以看到, 传统的 Pseudo-label 方法在该情况下提取的表示在某几类上区分性较差, 各类之间存在连接. 而我们的方法可以得到更高质量的表示, 各表示更加聚集且能较为清晰地区分各类.

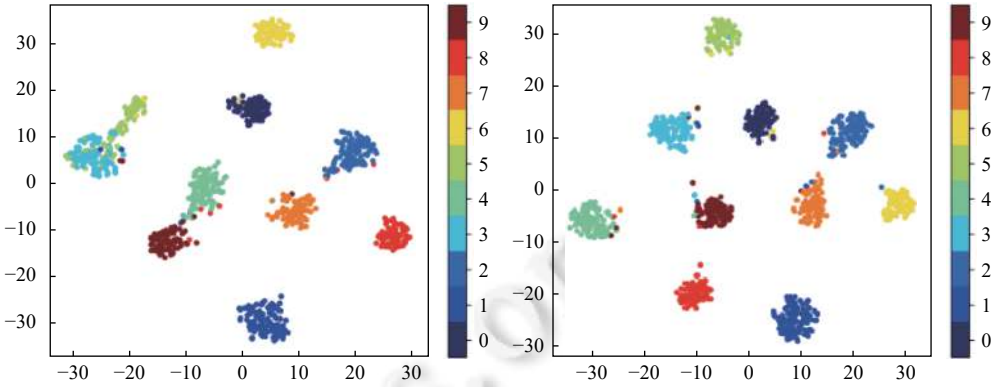


图 2 不同伪标签生成半监督学习算法在 MNIST (40 个标签) 上训练得到网络提取的测试数据的二维表示的 t-SNE 可视化结果

5.6 近似性与计算量分析

(1) Sinkhorn 算法近似性

Sinkhorn-Knopp 算法中公式 (11) 中正则参数 λ 的值越大, 算法得到的近似解就越接近原最优传输问题的解, 但是这会导致 Sinkhorn-Knopp 算法迭代次数的增加, 从而引入额外的计算时间. 因此, 为选择一个恰当的 Sinkhorn 正则项参数, 我们在 CIFAR-10 数据集具有 250 个标签数据的情况下进行对比实验, 比较了使用我们的基于最优传输的标签生成进行半监督学习的模型, 分别在 Sinkhorn 正则项参数 λ 设置为 5、100、500、1 000 的情况下, 模型最终的预测准确. 最终的实验结果如图 3 所示, 为权衡计算量与准确度, 本文在实验中选择 $\lambda = 500$ 作为 Sinkhorn-Knopp 算法的正则项参数.

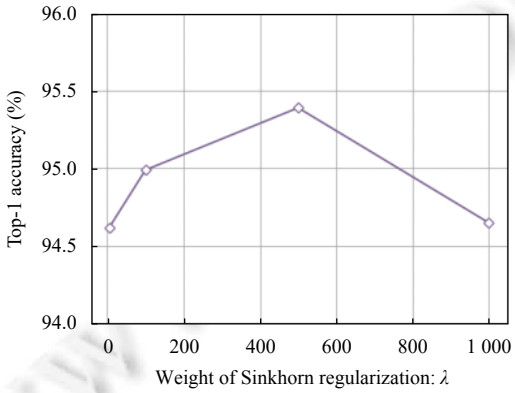


图 3 在 CIFAR-10 具有 250 个标签数据情况下, 不同 Sinkhorn-Knopp 正则参数设置的训练结果测试准确率对比

(2) 计算量研究

本文提出的基于最优传输的标签生成方法, 相比于直接使用模型预测值作为伪标签, 可以很好地优化半监督

学习过程中为无标记数据生成的伪标签的质量. 但是, 由于在得到最终标签结果前需要额外增加对于最优传输问题的求解, 必然会引入额外的计算时间. 因此, 为了研究算法的计算代价, 我们对使用该算法在不同数据集上进行半监督学习训练的计算时间进行统计计算, 并与直接使用模型预测结果进行计算的训练时间进行对比.

我们统计了本文方法平均超出直接使用模型预测结果进行半监督学习的时间占比. 本文方法在 CIFAR-10 数据集上需要额外使用 21.06% 的时间, SVHN 则需 13.73%. 本方法通过牺牲一定的计算时间, 换取更加优质的伪标签.

6 总 结

对于半监督学习方法来说, 伪标签生成已经成为一个重要的组成部分, 本文聚焦半监督学习过程中的伪标签生成方法研究, 旨在能够通过优化伪标签生成方法, 为无标记数据生成更加准确、优质的伪标签, 以辅助整体半监督学习方法的性能提升. 我们提出了基于最优传输理论的伪标签生成方法. 该方法引入类别均衡的类别分布约束, 以及额外引入有标记信息在标签生成时进行引导, 以期能够优化伪标签的质量. 除此之外, 我们希望能够将上述的所有约束融合成为一个优化求解过程, 而最优传输理论正好得以适用. 通过问题转换, 在类别均衡约束以及有标签信息约束下的伪标签生成问题可以十分巧妙地转换成最优传输的优化问题. 考虑到优化问题的计算可行性, 我们进一步引入 Sinkhorn-Knopp 算法进行求解最优传输优化问题, 由此可以快速得到最终的伪标签结果. 此外, 本文也给出了基于上述伪标签生成方法进行完整半监督学习的具体过程方法. 最后, 我们在常用的图像分类数据集上进行实验测试了上述方法, 并单独对伪标签生成部分方法进行实验分析, 实验结果与主流的半监督学习方法进行比较, 验证了所提算法的有效性.

References:

[1] Vasuki A, Govindaraju S. Deep neural networks for image classification. Advances in Parallel Computing, 2017, 31: 27–49. [doi: 10.3233/978-1-61499-822-8-27]

[2] Szegedy C, Toshev A, Erhan D. Deep neural networks for object detection. In: Proc. of the 26th Int'l Conf. on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2013. 2553–2561.

[3] Bojja AK. Deep neural networks for semantic segmentation [MS. Thesis]. Victoria: University of Victoria, 2020.

[4] Berthelot D, Carlini N, Goodfellow I, Oliver A, Papernot N, Raffel C. MixMatch: A holistic approach to semi-supervised learning. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 454.

[5] Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 1195–1204.

[6] Miyato T, Maeda SI, Koyama M, Ishii S. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2019, 41(8): 1979–1993. [doi: 10.1109/TPAMI.2018.2858821]

[7] Lee DH. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Proc. of the 30th Int'l Conf. on Machine Learning. Atlanta: JMLR, 2013. 1–6.

[8] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: Proc. of the 34th Int'l Conf. on Machine Learning. Sydney: JMLR.org, 2017. 1126–1135.

[9] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 4080–4090.

[10] Vinyals O, Blundell C, Lillicrap T, Kavukcuoglu K, Wierstra D. Matching networks for one shot learning. In: Proc. of the 30th Int'l Conf. on Neural Information Processing Systems. Barcelona: Curran Associates Inc., 2016. 3637–3645.

[11] Doersch C, Gupta A, Efros AA. Unsupervised visual representation learning by context prediction. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 1422–1430. [doi: 10.1109/ICCV.2015.167]

[12] Noroozi M, Favaro P. Unsupervised learning of visual representations by solving jigsaw puzzles. In: Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 69–84. [doi: 10.1007/978-3-319-46466-4_5]

[13] Chen T, Kornblith S, Norouzi M, Hinton GE. A simple framework for contrastive learning of visual representations. In: Proc. of the 37th Int'l Conf. on Machine Learning. PMLR, 2020. 1597–1607.

[14] Sohn K, Berthelot D, Li CL, Zhang ZZ, Carlini N, Cubuk ED, Kurakin A, Zhang H, Raffel C. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 51.

[15] Asano YM, Rupprecht C, Vedaldi A. Self-labelling via simultaneous clustering and representation learning. arXiv:1911.05371v3, 2020.

[16] Yang XL, Song ZX, King I, Xu ZL. A survey on deep semi-supervised learning. IEEE Trans. on Knowledge and Data Engineering, 2023, 35(9): 8934–8954. [doi: 10.1109/TKDE.2022.3220219]

[17] Zhu XJ, Ghahramani Z, Lafferty J. Semi-supervised learning using Gaussian fields and harmonic functions. In: Proc. of the 20th Int'l Conf. on Machine Learning. Washington: AAAI Press, 2003. 912–919.

[18] Liu B, Wu ZR, Hu H, Lin S. Deep metric transfer for label propagation with limited annotated data. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision Workshops. Seoul: IEEE, 2019. 1317–1326. [doi: 10.1109/ICCVW.2019.00167]

[19] Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Proc. of the 14th Int'l Conf. on Neural Information Processing Systems: Natural and Synthetic. Vancouver: MIT Press, 2001. 585–591.

[20] Lasserre JA, Bishop CM, Minka TP. Principled hybrids of generative and discriminative models. In: Proc. of the 2006 IEEE Conf. on Computer Vision and Pattern Recognition. New York: IEEE, 2006. 87–94. [doi: 10.1109/CVPR.2006.227]

[21] Kingma DP, Rezende DJ, Mohamed S, Welling M. Semi-supervised learning with deep generative models. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2014. 3581–3589.

[22] Pu YC, Gan Z, Henao R, Yuan X, Li CY, Stevens A, Carin L. Variational autoencoder for deep learning of images, labels and captions. In: Proc. of the 30th Int'l Conf. on Neural Information Processing Systems. Barcelona: Curran Associates Inc., 2016. 2360–2368.

[23] Iscen A, Tolias G, Avrithis Y, Chum O. Label propagation for deep semi-supervised learning. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5065–5074. [doi: 10.1109/CVPR.2019.00521]

[24] Arazo E, Ortego D, Albert P, O'Connor NE, McGuinness K. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: Proc. of the 2020 Int'l Joint Conf. on Neural Networks (IJCNN). Glasgow: IEEE, 2020. 1–8. [doi: 10.1109/IJCNN48605.2020.9207304]

[25] Zhang HY, Cisse M, Dauphin YN, Lopez-Paz D. mixup: Beyond empirical risk minimization. arXiv:1710.09412v2, 2018.

[26] Pham H, Dai ZH, Xie QZ, Le QV. Meta pseudo labels. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 11552–11563. [doi: 10.1109/CVPR46437.2021.01139]

[27] Rizve MN, Duarte K, Rawat YS, Shah M. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In: Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2021.

[28] Zhang BW, Wang YD, Hou WX, Wu H, Wang JD, Okumura M, Shinozaki T. FlexMatch: Boosting semi-supervised learning with curriculum pseudo labeling. In: Proc. of the 35th Conf. on Neural Information Processing Systems. 2021. 18408–18419.

[29] Xu Y, Shang L, Ye JX, Qian Q, Li YF, Sun BG, Li H, Jin R. Dash: Semi-supervised learning with dynamic thresholding. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 11525–11536.

[30] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In: Proc. of the 11th Annual Conf. on Computational Learning Theory. Madison: ACM, 1998. 92–100. [doi: 10.1145/279943.279962]

[31] Qiao SY, Shen W, Zhang ZS, Wang B, Yuille A. Deep co-training for semi-supervised image recognition. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 142–159. [doi: 10.1007/978-3-030-01267-0_9]

[32] Chen DD, Wang W, Gao W, Zhou ZH. Tri-net for semi-supervised deep learning. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence. Stockholm: AAAI Press, 2018. 2014–2020.

[33] Berthelot D, Carlini N, Cubuk ED, Kurakin A, Sohn K, Zhang H, Raffel C. ReMixMatch: Semi-supervised learning with distribution matching and augmentation anchoring. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020.

[34] Rasmus A, Valpola H, Honkala M, Berglund M, Raiko T. Semi-supervised learning with ladder networks. In: Proc. of the 28th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2015. 3546–3554.

[35] Valpola H. From neural PCA to deep unsupervised learning. In: Bingham E, Kaski S, Laaksonen J, Lampinen J, eds. Advances in Independent Component Analysis and Learning Machines. London: Academic Press, 2015. 143–171. [doi: 10.1016/B978-0-12-802806-3.00008-7]

[36] Laine S, Aila T. Temporal ensembling for semi-supervised learning. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017.

[37] Ke ZH, Wang DY, Yan Q, Ren J, Lau R. Dual student: Breaking the limits of the teacher in semi-supervised learning. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 6727–6735. [doi: 10.1109/ICCV.2019.00683]

[38] Xie QZ, Dai ZH, Hovy E, Luong MT, Le QV. Unsupervised data augmentation for consistency training. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 525.

[39] Cubuk ED, Zoph B, Mané D, Vasudevan V, Le QV. AutoAugment: Learning augmentation strategies from data. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 113–123. [doi: 10.1109/CVPR.2019.00020]

[40] Cuturi M. Sinkhorn distances: Lightspeed computation of optimal transport. In: Proc. of the 26th Int'l Conf. on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2013. 2292–2300.

[41] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proc. of the 25th Int'l Conf. on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2012. 1097–1105.

[42] Netzer Y, Wang T, Coates A, Bissacco A, Wu B, Ng AY. Reading digits in natural images with unsupervised feature learning. In: Proc. of the 2011 Conf. and Workshop on Deep Learning and Unsupervised Feature Learning. Granada, 2011. 4.

[43] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc. of the IEEE, 1998, 86(11): 2278–2324. [doi: 10.1109/5.726791]

[44] Xiao H, Rasul K, Vollgraf R. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. arXiv:1708.07747, 2017.

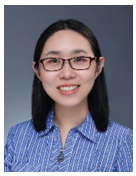
[45] Ma XJ, Huang HX, Wang Y, Romano S, Erfani SM, Bailey J. Normalized loss functions for deep learning with noisy labels. In: Proc. of the 37th Int'l Conf. on Machine Learning. PMLR, 2020. 6543–6553.

[46] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. In: Proc. of the 14th Int'l Conf. on Artificial Intelligence and Statistics. Fort Lauderdale: JMLR.org, 2011. 315–323.

[47] Zagoruyko S, Komodakis N. Wide residual networks. In: Proc. of the 2016 British Machine Vision Conf. York: BMVA Press, 2016. 1–12.

[48] Cubuk ED, Zoph B, Shlens J, Le QV. Randaugment: Practical automated data augmentation with a reduced search space. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops. Seattle: IEEE, 2020. 3008–3017. [doi: 10.1109/CVPRW50498.2020.00359]

[49] van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research, 2008, 9(86): 2579–2605.



翟德明(1984－), 女, 博士, 副教授, 博士生导师, 主要研究领域为机器学习, 数据挖掘及其在计算机视觉和图像处理中的应用.

江俊君(1986－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为计算机视觉, 图像处理, 深度学习.

沈斯娴(1998－), 女, 硕士生, 主要研究领域为深度学习, 鲁棒学习.

刘贤明(1983－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为可信人工智能, 计算成像, 生物医学信号压缩, 3D 信号处理和分析.

周雄(1996－), 男, 博士生, 主要研究领域为图像处理, 计算机视觉, 鲁棒机器学习.

季向阳(1976－), 男, 博士, 教授, 博士生导师, 主要研究领域为工智能, 计算成像.