计 算 机 研 究 与 发 展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202220024
60(7): 1615−1627, 2023

基于双生成器网络的 Data-Free 知识蒸馏

张 晶 鞠佳良 任永功
(辽宁师范大学计算机与人工智能学院 辽宁大连 116081)
(zhangjing_0412@lnnu.edu.cn)

Double-Generators Network for Data-Free Knowledge Distillation

Zhang Jing, Ju Jialiang, and Ren Yonggong
(*School of Computer Science and Artificial Intelligence*, *Liaoning Normal University*, *Dalian*, *Liaoning* 116081)

Abstract Knowledge distillation (KD) maximizes the similarity of output distributions between teacher-network and student-network to achieve network compression and the large-scale network proximal-end deployment and application. However, the privacy protection and transmission problems result in that the training data are difficultly collected. In the scenario of training data shortage that is called data-free, improving the performance of KD is a meaningful task. Data-free learning (DAFL) builds up teacher-generator to obtain pseudo data that are similar as real samples, and then pseudo data are utilized to train student-network by distilling. Nevertheless, the training process of teacher-generator will produce both problems: 1) Absolutely trusting the discrimination outputs of teacher-network maybe include incorrectly information from unlabeled pseudo data, moreover, teacher-network and student-network have different learning targets. Therefore, it is difficult to obtain the accuracy and coincident information for training student-network. 2) Over-dependences loss values originated from teacher-network, which induces pseudo data with un-diversity damaging the generalization of student-network. Aim to resolve above problems, we propose a double generators network framework DG-DAFL for data-free by building up double generators. In DG-DAFL, student-network and teacher-network obtain the same learning tasks by optimizing double generators at the same time, which enhances the performance of student-network. Moreover, we construct the distribution loss between student-generator and teacher-generator to enrich sample diversity and further improve the generalization of student-network. According to the results of experiments, our method achieves the more efficient and robust performances in three popular datasets. The code and model of DG-DAFL are published in https://github.com/LNNU-computer-research-526/DG-DAFL.git.

Key words deep neural network; knowledge distillation; data-free knowledge distillation; generative adversarial network; generator

摘 要 知识蒸馏(knowledge distillation, KD)通过最大化近似输出分布使“教师网络”指导“学生网络”充分训练,成为大规模深度网络近端迁移、部署及应用的重要技术.然而,隐私保护意识增强与传输问题加剧使网络训练数据难以获取.如何在Data-Free的自由环境下,保证压缩网络准确率成为重要的研究方向.Data-Free学生网络学习(data-free learning of student networks, DAFL)模型,建立“教师”端生成器获得与预训练网络分布近似的伪数据集,通过知识蒸馏训练“学生网络”.然而,该框架中生成器构建及优化仍存在2个

收稿日期: 2022-01-04; **修回日期**: 2022-08-08
基金项目: 国家自然科学基金项目(61902165, 61976109); 大连市科技创新基金项目(2018J12GX047); 教育部人文社会科学研究规划基金项目(21YJC880104)
This work was supported by the National Natural Science Foundation of China (61902165, 61976109), the Dalian Science and Technology Innovation Fund (2018J12GX047), and the Social Science Foundation of Ministry of Education of China(21YJC880104).
通信作者: 任永功(ryg@lnnu.edu.cn)

问题: 1) 过度信任“教师网络”对缺失真实标签伪样本的判别结果, 同时, “教师网络”与“学生网络”优化目标不同, 使“学生网络”难以获得准确、一致的优化信息; 2) 仅依赖于“教师网络”训练损失, 导致数据特征多样性缺失, 降低“学生网络”泛化性. 针对这 2 个问题, 提出双生成器网络架构 DG-DAFL(double generators-DAFL), 分别建立“教师”与“学生”端生成器并同时优化, 实现网络任务与优化目标一致, 提升“学生网络”判别性能. 进一步, 增加双生成器样本分布差异损失, 利用“教师网络”潜在分布先验信息优化生成器, 保证“学生网络”识别准确率并提升泛化性. 实验结果表明, 该方法在 Data-Free 环境中获得了更为有效且更鲁棒的知识蒸馏效果. DG-DAFL 方法代码及模型已开源: https://github.com/LNNU-computer-research-526/DG-DAFL.git.

关键词 深度神经网络; 知识蒸馏; 无数据环境知识蒸馏; 对抗生成网络; 生成器

中图法分类号 TP183

深度学习凭借对样本高维特征的非线性表达及数据信息的抽象表示, 极大地推进了语音识别、计算机视觉等人工智能方法在工业中的应用. 1989 年 LeCun 等人[1] 提出深度卷积网络 LeNet 模型, 在手写体图像识别领域取得了突破性进展, 为深度学习的发展提供了前提和基础. 为进一步提升深度神经网络模式识别及图像处理精度, 推广其在工业中的应用, 国内外学者不断优化及改进网络结构. 随着模型层数逐步增加, 模型参数和架构愈加庞大, 算法对存储、计算等资源的需求不断增长, 导致大模型网络失效等问题[2], 例如 Resnet50, VGG16 等大型神经网络, 尽管在图像分类应用上表现出卓越性能, 但其冗余参数导致较高计算成本和内存消耗. 同时, 多媒体、5G 技术、移动终端的快速发展, 边缘计算设备广泛部署, 使网络应用需求逐步增加. 手机、平板电脑、移动摄像机等便携式近端设备相比于固定设备存在数十倍的计算、存储等能力差距, 为大规模网络近端迁移与运行带来困难. 如何提升边缘设备计算、识别及分类能力, 实现大规模深度学习网络的近端部署成为有意义的工作. 基于此, Buciluǎ等人[3] 提出神经网络模型压缩方法, 将信息从大模型或模型集合传输到需要训练的小型模型, 而不降低模型精度. 同时, 大规模神经网络模型中包含的大量参数存在一定功能稀疏性, 使网络结构出现过参数化等问题, 即使在网络性能敏感的大规模场景中, 仍包含产生重复信息的神经元与链接. 知识蒸馏 (knowledge distillation, KD) 将高性能大规模网络作为教师网络指导小规模学生网络[4], 实现知识精炼与网络结构压缩, 成为模型压缩、加速运算、大规模网络近端部署的重要方法.

然而, 随着人们对隐私保护意识的增强以及法律、传输等问题的加剧, 针对特定任务的深度网络训练数据往往难以获取, 使 Data-Free 环境下的神经网络模型压缩, 即在避免用户隐私数据泄露的同时得到一个与数据驱动条件下压缩后准确率相似的模型, 成为一个具有重要实际意义的研究方向. Chen 等人[5] 提出 Data-Free 环境知识蒸馏框架 DAFL (data-free learning of student networks, DAFL), 建立教师端生成器, 生成伪样本训练集, 实现知识蒸馏并获得与教师网络性能近似的小规模学生网络. 然而, 该方法在复杂数据集上将降低学生网络识别准确率, 其主要原因有 3 个方面:

1)判别网络优化目标不同. 模型中教师网络优化生成器产生伪数据, 实现学生网络知识蒸馏, 使学生网络难以获得与教师网络一致的优化信息构建网络模型.

2)误差信息优化生成器. 教师端生成器的构建过度信任教师网络对伪数据的判别结果, 利用误差信息优化并生成质量较差的伪训练样本, 知识蒸馏过程学生网络难以有效利用教师网络潜在先验分布信息.

3)学生网络泛化性低. 模型中生成数据仅依赖于教师网络训练损失, 导致生成数据特征多样性缺失, 降低学生网络判别性.

如图 1 所示, MNIST 数据集中类别为 1 和 7 时图像特征有较大差异, 而图 1 右侧中 DAFL 方法的学生网络得到的 2 类数据统计特征直方图相当近似, 该模型训练得到的小规模学生网络针对特征相似图像难以获得更鲁棒的判别结果. 为提升 DAFL 模型中学生的网络准确率及泛化性, 提出新的双生成器网络架构 DG-DAFL(double generators-DAFL, DG-DAFL), 图 1 右侧中由 DG-DAFL 框架训练得到学生网络判别器特征统计直方图对比, 即 1 类和 7 类特征统计结果有一定差距, 为后续分类提供了前提.

为解决 Data-Free 环境知识蒸馏、保证网络识别精度与泛化性, 本文提出双生成器网络架构 DG-DAFL, 学生端生成器在教师端生成器的辅助下充分利用教师网络潜在先验知识, 产生更适合学生网络训练的伪训练样本, 利用生成器端样本分布差异, 避免 DAFL 学生网络对单一教师网络端生成器样本依赖, 保证生成器样本多样性, 提升学生网络判别器识

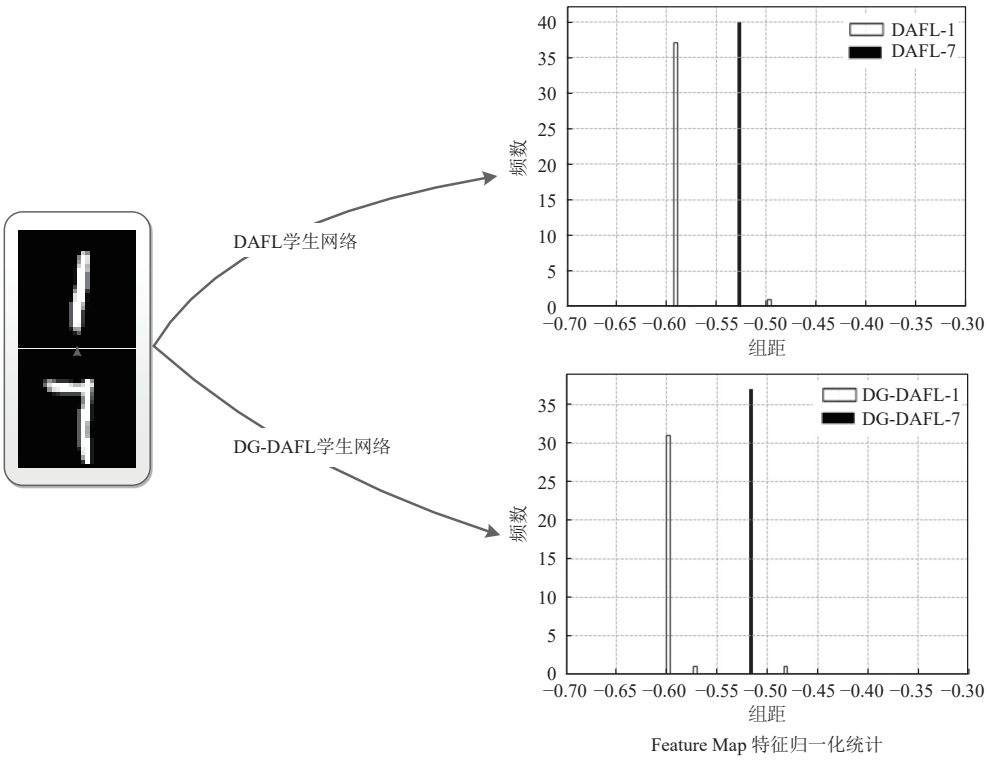

Fig. 1 Comparison of normalized statistical results for approximate sample characteristics

图 1 近似样本特征归一化统计结果对比

别泛化性. 本文贡献有 3 方面：

1)针对 Data-Free 知识蒸馏问题提出双生成器网络架构 DG-DAFL, 建立教师生成器网络与学生生成器网络, 生成伪样本. 优化教师生成器网络的同时, 学生网络判别器优化学生生成器网络, 实现生成器与判别器分离, 避免误差判别信息干扰生成器构建. 同时, 使网络任务及优化目标一致, 提升学生网络性能. 该结构可被拓展于解决其他任务的 Data-Free 知识蒸馏问题.

2)通过增加教师网络及学生网络生成器端样本分布差异度量, 避免单生成器网络结构中学生网络训练过度依赖教师生成器网络样本, 产生泛化性较低等问题. 同时, 该差异度量可使得学生网络生成数据在保证分布近似条件下的样本多样性, 进一步提升学生网络识别鲁棒性.

3)所提出框架在流行分类数据集 Data-Free 环境下, 学生网络参数量仅为教师网络的 50% 时, 仍取得了令人满意的识别性能. 同时, 进一步验证并分析了近似样本数据集的分类问题, 取得了更鲁棒的结果.

1 相关工作

针对大规模神经网络的近端部署与应用, 网络模型压缩及加速成为人工智能领域的研究热点. 目前的模型压缩方法包括网络剪枝[6]、参数共享[7]、量化[8]、网络分解[9]、紧凑网络设计, 其中知识蒸馏凭借灵活、直观的知识抽取及模型压缩性能受到学者广泛关注. 2015 年, Hinton 等人[4] 提出知识蒸馏模型, 构建教师网络、学生网络及蒸馏算法 3 部分框架, 引入温度(temperature, T)系数, 使卷积神经网络 softmax 层的预测标签由硬标签(hard-label)转换为软标签(soft-label), 利用庞大、参数量多的教师网络监督训练得到体量、参数量更少且分类性能与教师网络更近似的学生网络[3-4,10-11]. 根据知识蒸馏操作的不同部分, 分为目标(logits)蒸馏[12-16] 与特征图蒸馏[17-22] 两类. logits 知识蒸馏模型主要目标集中在构建更为有效的正则化项及优化方法, 在硬标签(hard-label)监督训练下得到泛化性能更好的学生网络. Zhang 等人[16] 提出深度互学习(deep mutual learning, DML)模型, 利用交替学习同时强化学生网络与教师网络. 然而, 教师网络与学生网络的性能差距使蒸馏过程难以收敛. 基于此, Mirzadeh 等人[14] 提出助教知识蒸馏(teacher assistant knowledge distillation, TAKD)模型, 引入中等规模助教网络, 缩小教师网络和学生网络之间过大的性能差距, 达到逐步蒸馏的目的. 特征图知识蒸馏模型通过直接将样本表征从教师网络迁移至学生网络[17-18,20],

或将训练教师网络模型样本结构迁移至学生网络[19,21-22]，实现知识抽取. 该类方法充分利用大规模教师网络对样本的高维、非线性特征表达及样本结构，获得更高效的学生网络.

Data-Free 环境中用于训练模型的真实数据往往难以获取，使知识蒸馏模型失效. 对抗生成网络(generative adversarial network, GAN)技术的发展，激发了该类环境下知识蒸馏领域方法的进步. 2014 年，Goodfellow 等人[23]提出 GAN 模型，通过模型中生成器与鉴别器的极大极小博弈，二者相互竞争提升各自生成和识别能力[24]，可用于生成以假乱真的图片[25]、影片[26]等的无监督学习方法. GAN中的生成器可合成数据直接作为训练数据集，或用于训练数据集增广及生成难样本支持学生网络训练. Nguyen 等人[27]利用预训练的 GAN 生成器作为模型反演的先验，构建伪训练数据集. Bhardwaj 等人[28]利用 10% 的原始数据和预训练教师模型生成合成图像数据集，并将合成图像用于知识蒸馏. Liu 等人[29]与 Zhang 等人[30]均利用无标签数据提升模型效果，分别提出无标签数据蒸馏的光流学习(learning optical flow with unlabeled data distillation, DDFlow)模型[29]与图卷积网络可靠数据蒸馏(reliable data distillation on graph convolution network, RDDGCN)模型[30]. 其中 RDDGCN 模型利用教师网络对所生成的未标注数据给予新的训练注释，构建训练数据集训练学生网络. 有研究借助大规模预训练数据集提升模型效果，Yin 等人[31]提出的 DeepInversion 方法将图像更新损失与教师、学生之间的对抗性损失结合，教师网络通过对 Batch Normalization 层中所包含通道的均值和方差进行推导，在大规模 ImageNet 数据集上预训练深度网络后合成图像作为训练样本集. Lopes 等人[32]进一步利用教师网络先验信息，通过教师网络激活层重构训练数据集以实现学生网络知识蒸馏. 文献 [28−32] 所述方法均利用少量训练数据或常用的预训练数据集信息，在 Data-Free 环境中仍难以解决无法直接获取真实且可用于训练小规模学生网络的先验信息等问题.

基于此，DAFL 框架借助 GAN 学习模型，将预训练好的教师网络作为判别器网络，构建并优化生成器网络模型，生成更加接近真实样本分布的伪数据，为高精度、小规模学生网络的知识蒸馏与网络压缩提供有效先验信息，框架如图 2 所示. 首先，通过函数 *one_hot* 获得伪标签，利用损失函数将 GAN 中判别器的输出结果从二分类转换为多分类，以实现多分类任务的知识蒸馏；其次，采用信息熵损失函数、特征图激活损失函数、类别分布损失函数优化生成器，为学生网络训练提供数据；最终，实现在没有原始数据驱动条件下，通过知识蒸馏方法使学生网络参数减少一半，且具有与教师网络近似的分类准确率. 然而，DAFL 框架中生成器优化过程完全信任判别器针对 Data-Free 环境中初始生成伪样本的先验判别，忽略了伪样本所构造伪标签带来的误差，干扰生成器优化，直接影响学生网络性能. 同时，教师网络与学生网络执行不同任务时存在学生网络过度依赖教师网络生成器样本，降低 Data-Free 环境下模型学习泛化性.

为了提升生成样本质量，Fang 等人[33]提出无数据对抗蒸馏(data-free adversarial distillation, DFAD)模型，通过训练一个外部生成器网络合成数据，使学生网络和教师网络输出差异最大化图像. Han 等人[34]提出鲁棒性和多样性的 Data-Free 知识蒸馏(robustness and diversity seeking data-free knowledge distillation, RDSKD)方法在生成器训练阶段引入指数惩罚函数，

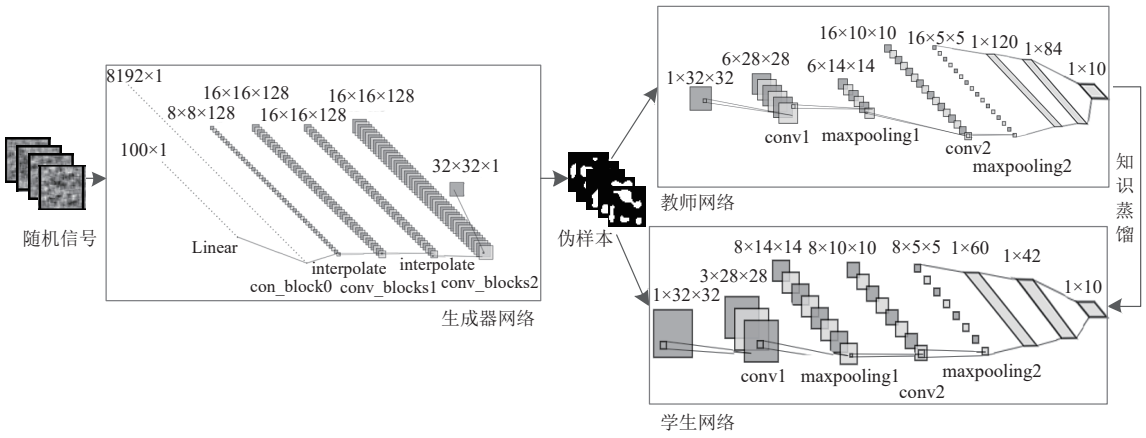

Fig. 2 Architecture of DAFL

图 2 DAFL 架构

提升生成器生成图像的多样性. Nayak 等人[35]提出零样本知识蒸馏模型，仅利用教师网络参数对 softmax 层空间建模生成训练样本. 同时，Micaelli 等人[36]提出零样本对抗性信息匹配模型，利用教师网络特征表示的信息生成训练样本. 为避免零样本学习中先验信息缺失降低学生网络学习准确率等问题，Kimura 等人[37]与 Shen 等人[38]分别提出伪样本训练模型与网络嫁接模型，二者均借助少量确定性监督样本，将知识从教师模型提取到学生神经网络中. 为充分利用教师网络先验信息，Storkey 等人[39]提出 zero-shot 知识蒸馏方法，将教师网络同时定义为样本鉴别器. 同时，Radosavovic 等人[40]提出全方位监督学习模型.

文献 [5, 33−40] 所述的 Data-Free 环境中知识蒸馏模型所需的训练数据通常由已训练教师模型的特征表示生成，该类数据包含部分教师网络先验信息，在无数据可用的情况下显示出了很大的潜力. 然而，Data-Free 知识蒸馏仍是一项非常具有挑战性的任务，主要集中在如何生成高质量、多样化、具有针对性的训练数据，进而获得更高精度、高泛化性的小规模学生网络.

2 双生成器网络

针对提升 Data-Free 环境中知识蒸馏方法有效性与泛化性，本文受 DAFL 模型的启发，提出 DG-DAFL 网络架构，如图 3 所示. 包括 4 部分网络结构：教师端生成器网络G_{T}、学生端生成器网络G_{S}、教师端判别器网络N_{T}、学生端判别器网络N_{S}. DG-DAFL 利用教师端与学生端判别器网络N_{T}与N_{S}，同时优化生成器网络G_{T}与G_{S}，保证学生网络与教师网络优化目标一致，避免真实样本标签类别先验信息缺失时生成器过度信任教师网络判别结果，产生质量较低的伪样本，降低学生网络判别性能. 同时，通过增加生成器端伪样本分布损失，保证学生端生成器网络训练样本多样性，提升学生网络学习泛化性. DG-DAFL 框架的训练过程可总结为 3 个步骤：教师端辅助生成器G_{T}构建、最优化学生端生成器G_{S}构建、学生网络N_{S}与教师网络N_{T}知识蒸馏.

2.1 教师端辅助生成器G_{T}构建

本文构建双生成器网络架构G_{T}与G_{S}，通过教师网络提取训练样本先验信息，训练教师端生成器网络G_{T}，使生成的伪样本分布更近似于真实样本. 由于真实样本标签缺失，G_{T}难以得到来自于N_{T}准确、充分的样本分布先验信息，实现最优化训练. 因此，本文仅利用教师端生成器网络G_{T}作为训练学生端生成器网络G_{S}的辅助网络，强化生成伪样本质量，提升学生网络判别准确率.

随机样本$\boldsymbol{Z}^{(\mathrm{T})}$作为教师端生成器网络$G_{\mathrm{T}}(\boldsymbol{Z}^{(\mathrm{T})};\boldsymbol{\theta}_{\mathrm{g}})$的初始输入，经网络计算后得到伪样本$\boldsymbol{x}_i^{(\mathrm{T})}$, $i=1,2,\cdots,N$，其中$\boldsymbol{\theta}_{\mathrm{g}}$为$G_{\mathrm{T}}$网络参数. 同时，伪样本集$\boldsymbol{X}^{(\mathrm{T})}$作为教师网络判别器$N_{\mathrm{T}}(\boldsymbol{X}^{(\mathrm{T})};\boldsymbol{\theta}_{\mathrm{d}})$的输入，可得到该网络判别结果，结合先验信息构造损失函数$\mathcal{L}_{G_{\mathrm{T}}}$，反馈训练生成器网络$G_{\mathrm{T}}$，得到更真实样本分布的伪训练样本集，用于学生网络知识蒸馏. 为获得优化反馈信息，$\mathcal{L}_{G_{\mathrm{T}}}$由 3 部分构成：

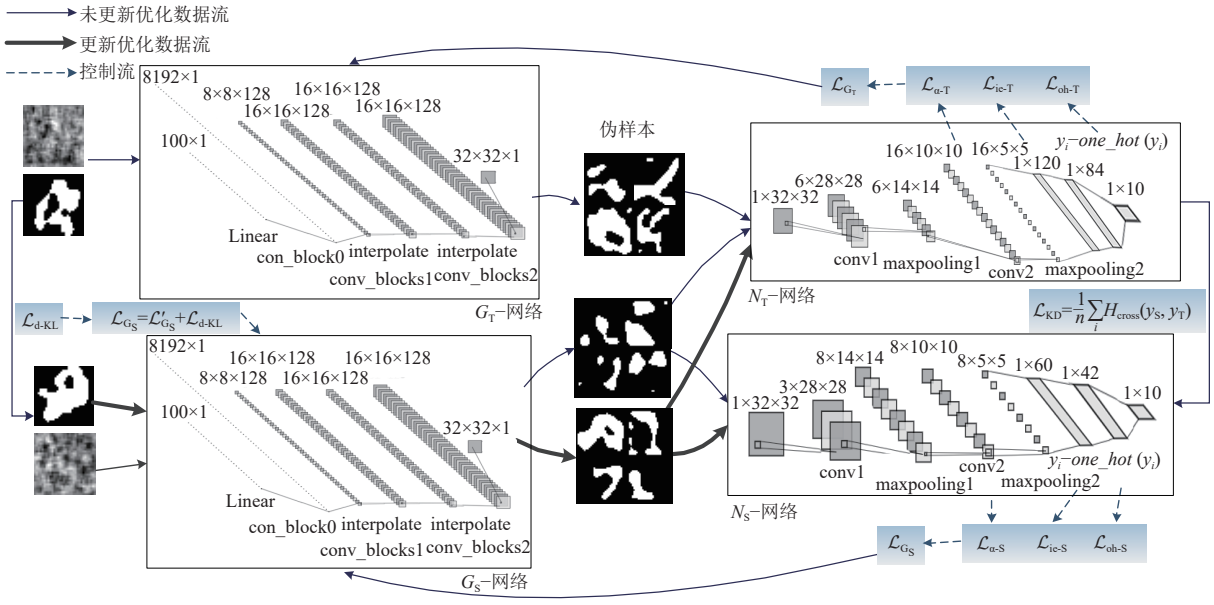

Fig. 3 Architecture and learning process of DG-DAFL

图 3 DG-DAFL 架构及学习过程

1) 伪样本集可计算得到网络输出向量 $\boldsymbol{y}_i^{(\mathrm{T})}=N_{\mathrm{T}}(\boldsymbol{X}^{(\mathrm{T})};\boldsymbol{\theta}_{\mathrm{d}})$, 由于伪样本 $\boldsymbol{x}_i^{(\mathrm{T})}$ 缺少真实标签信息, 可求解输出向量的伪标签 $\boldsymbol{t}_i=\arg\max_j(\boldsymbol{y}_i^{\mathrm{T}})_j$, 其中 $j=1,2,\cdots,k$. $\boldsymbol{t}_i$ 为包含 k 类向量中最大值位置, 构建经验损失函数 $\mathcal{L}_{\text{oh-T}}$:

$$\mathcal{L}_{\text{oh-T}}=\frac{1}{n}\sum_i H_{\text{cross}}(\boldsymbol{y}_i^{(\mathrm{T})},\boldsymbol{t}_i). \tag{1}$$

最小化预测标签与真实标签交叉熵值, 学习教师网络判别器先验信息, 使 G_{T} 生成与真实样本分布更为接近的伪样本集.

2) 借助 DAFL 中模型训练过程, N_{T} 网络中多卷积层所提取的特征向量中更具判别性的神经元将被激活, 即伪样本 $\boldsymbol{X}^{(\mathrm{T})}$ 经预训练网络 N_{T} 逐层非线性特征计算后得到特征向量 $\boldsymbol{f}_i^{(\mathrm{T})}$, 其中更大激活值可包含更多的真实样本特征先验信息, 特征图激活损失函数可被表示为

$$\mathcal{L}_{\alpha\text{-T}}=-\frac{1}{N}\sum_i\|\boldsymbol{f}_i^{(\mathrm{T})}\|_1, \tag{2}$$

该损失在生成器优化过程中减小伪样本经卷积滤波器后激活值更大的特征, 得到更接近真实样本特征表达.

3) 为充分利用预训练教师网络样本分布及类别先验信息, 构建预训练集样本类平衡分布损失 $\mathcal{L}_{\text{ie-T}}$. 定义 $p=\{p_1,p_2,\cdots,p_k\}$ 为 k 类样本集中的每类样本出现的概率, 当各类样本为均匀分布时, 即 $p_k=\frac{1}{K}$, 所含信息量最大. 为保证教师网络判别结果的均衡性、多样性, 充分利用预训练样本分布信息, 以教师网络优化生成器在该类数据集下等概率生成各类样本, 构建信息熵损失函数:

$$\mathcal{L}_{\text{ie-T}}=-H_{\text{info}}\left(\frac{1}{N}\sum_i\boldsymbol{y}_i^{\mathrm{T}}\right). \tag{3}$$

结合式(1)~(3), 可得到用于优化辅助生成器 G_{T} 的目标函数为

$$\mathcal{L}_{G_{\mathrm{T}}}=\mathcal{L}_{\text{oh-T}}+\alpha\mathcal{L}_{\alpha\text{-T}}+\beta\mathcal{L}_{\text{ie-T}}, \tag{4}$$

其中 α 和 β 为平衡因子. 利用式(4)保证 G_{T} 优化过程充分利用教师网络保存的训练样本分布等先验信息, 即可获得更近似于真实数据的高质量伪样本数据集.

2.2 最优化学生端生成器 G_{S} 的构建

根据 2.1 节所述的教师端生成器 G_{T} 的优化过程, 借助教师端判别器网络 N_{T} 包含的真实样本先验信息. 然而, 由于函数 *one_hot* 所构建的伪样本标签将带来大量噪音, 当 G_{T} 对 N_{T} 完全信任时, 其优化过程将引入错误信息, 使学生端判别器网络 N_{S} 训练阶段难以生成与真实样本分布近似的伪样本集, 影响学生网络判别准确率. 同时, 当 N_{S} 的训练将完全依赖于网络 G_{T} 生成伪样本时将降低模型 N_{S} 的泛化性.

为解决上述问题, 本文在学生网络端引入生成器 G_{S}, 如图 2 所示. 利用 G_{T} 信息辅助 G_{S} 优化, 生成更接近真实分布且更具多样性的训练样本. 首先, 双生成器 G_{T} 与 G_{S} 通过随机初始样本同时生成伪样本矩阵 $\boldsymbol{X}^{(\mathrm{T})}$ 与 $\boldsymbol{X}^{(\mathrm{S})}$, 其中, $\boldsymbol{X}^{(\mathrm{T})}$ 通过 N_{T} 计算并由式(4)构建损失反馈训练生成器 G_{T}, 生成新的教师端伪样本集 $\boldsymbol{X}'^{(\mathrm{T})}$; 其次, $\boldsymbol{X}^{(\mathrm{S})}$ 同时经 N_{T} 与 N_{S} 计算, 为充分借助教师网络先验数据分布信息度量分布差异, 利用式(5)优化 N_{S}:

$$\mathcal{L}_{\text{oh-S}}=\frac{1}{n}\sum_i H_{\text{cross}}\left(N_{\mathrm{T}}\left(\boldsymbol{X}^{(\mathrm{T})};\boldsymbol{\theta}_{\mathrm{d}}\right),N_{\mathrm{S}}\left(\boldsymbol{X}^{(\mathrm{S})};\boldsymbol{\theta}_{\mathrm{d}}^{(\mathrm{S})}\right)\right). \tag{5}$$

此时, 利用初步训练得到的 N_{S} 结合当前生成伪样本集 $\boldsymbol{X}^{(\mathrm{S})}$ 与式(4), 构建反馈损失函数 $\mathcal{L}'_{G_{\mathrm{S}}}=\mathcal{L}_{\text{oh-S}}+\alpha\mathcal{L}_{\alpha\text{-S}}+\beta\mathcal{L}_{\text{ie-S}}$, 优化当前学生网络生成器 G_{S}. 该模型可保证教师网络与学生网络执行相同任务, 提升学生网络学习能力. 同时, 通过对学生网络优化避免对缺失真实标签判别结果的过分信任, 降低生成器优化效果. 最后, G_{S} 生成新的学生端伪样本集 $\boldsymbol{X}'^{(\mathrm{S})}$. 为使 G_{S} 获得更多样本先验信息保证生成样本与真实样本分布一致性, 同时, 保证生成伪样本多样性, 提升学生网络模型泛化性, 本文采用 KL 散度获得 2 个优化得到的伪样本集 $\boldsymbol{X}'^{(\mathrm{T})}$ 与 $\boldsymbol{X}'^{(\mathrm{S})}$ 随分布差异, 如式(6)所示:

$$\mathcal{L}_{\text{d-KL}}=\sum_{i=1}^{N}G_{\mathrm{S}}(\boldsymbol{x}'^{(\mathrm{S})}_i)\mathrm{lb}\frac{G_{\mathrm{S}}(\boldsymbol{x}'^{(\mathrm{S})}_i)}{G_{\mathrm{T}}(\boldsymbol{x}'^{(\mathrm{T})}_i)}. \tag{6}$$

本文仅期望学生网络生成器 G_{S} 所得的样本集 $\boldsymbol{X}'^{(\mathrm{S})}$ 在分布上与先验样本分布更为接近. 此时, 构建学生网络生成器优化损失表达, 如式(7)所示, 实现最优化生成器 G_{S} 的构建.

$$\mathcal{L}_{G_{\mathrm{S}}}=\mathcal{L}'_{G_{\mathrm{S}}}+\gamma\mathcal{L}_{\text{d-KL}}, \tag{7}$$

其中, γ 为平衡因子.

2.3 学生网络与教师网络知识蒸馏

本文利用优化得到的学生端生成器 G_{S}, 更新伪样本集 $\boldsymbol{X}'^{(\mathrm{S})}$ 作为训练数据辅助学生网络构建.

教师网络 N_{T} 与学生网络 N_{S} 同时接受学生端生成器获得的优化为样本集 $\boldsymbol{X}'^{(\mathrm{S})}$, 由于模型差异, 网络结构相对复杂的教师网络输出结果优于网络结构相对简单的学生网络. 为提升模型压缩效果, 借助知识蒸馏技术, 将二者 softmax 层上输出结果进行交叉熵函数计算, 使学生网络的输出 $\boldsymbol{y}_i^{(\mathrm{S})}$ 更近似教师网络的输出 $\boldsymbol{y}_i^{(\mathrm{T})}$, 提升学生网络 N_{S} 的性能. 知识蒸馏损失函数为

$$\mathcal{L}_{\mathrm{KD}}=\frac{1}{N}\sum_{i=1}^{N}H_{\mathrm{cross}}(\boldsymbol{y}_i^{(\mathrm{S})},\boldsymbol{y}_i^{(\mathrm{T})}). \tag{8}$$

结合伪样本训练，在此损失函数约束下，实现在相同任务下较为稀疏的大规模网络到紧凑小规模网络的压缩及知识蒸馏.

3 实验结果与分析

本文在 3 个流行图像数据集上验证了所提出方法的有效性，并与近年 Data-free 环境下较为流行的知识蒸馏模型，包括 DAFL, DFAD, RDSKD 模型在精度、鲁棒性、泛化性上进行对比与分析. 同时，通过对模型消融实验结果的统计，讨论模型框架结构设计的合理性. 本文进一步设置实验数据，验证 DG-DAFL 模型的泛化性. 实验运行在 Intel Core i7-8700 及 NVIDIA Geforce RTX 2070 硬件环境，及 Windows10 操作系统、Python3 语言环境、Pytorch 深度学习框架上.

本文为了更全面地验证模型效果，采用 4 种评价指标：准确率(*Accuracy*)、精确率(*Precision*)、召回率(*Recall*)、特异度(*Specificity*).

准确率(*Accuracy*)指分类模型中正确样本量占总样本量的比重，其计算公式为

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN}. \tag{9}$$

精确率(*Precision*)指分类结果预测为阳性的正确比重，计算公式为

$$Precision=\frac{TP}{TP+FP}. \tag{10}$$

召回率(*Recall*)指真实值为阳性的正确比重，其计算公式为

$$Recall=\frac{TP}{TP+FN}. \tag{11}$$

特异度(*Specificity*)指真实值为阳性的正确比重，其计算公式为

$$Specificity=\frac{TN}{TN+FP}. \tag{12}$$

式(9)~(12)中，*TP* 为模型正确预测为正例样本量，*TN* 为模型正确预测为反例样本量，*FP* 为模型错误预测为正例样本量，*FN* 为模型错误预测为反例样本量.

本文引入双生成器端损失在充分利用教师网络先验样本分布信息条件下，保证生成样本多样性，如式(7)所示，其中 γ 为平衡因子. 为保证实验的公平性，γ 值的选取采用确定范围{0.01, 0.1, 1, 10, 100}内值遍历选取方法，如图 4 中所示，γ 取值将对学生网络模型识别结果产生较大影响. 当 $\gamma=10$ 时，MNIST 与 USPS 数据集均达到 *Accuracy* 统计的最高值. 因此，本文验证实验中的所有数据集，均设置 $\gamma=10$.

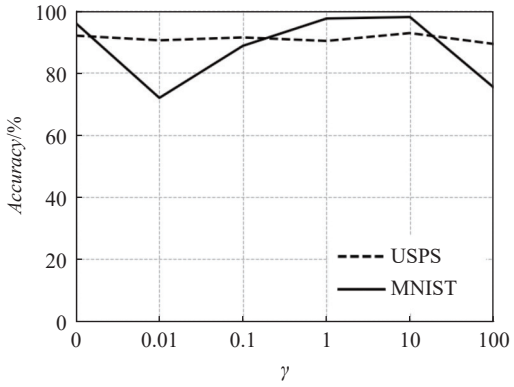

Fig. 4 Effect of γ on model performance

图 4 参数 γ 值对模型性能的影响

3.1 实验结果对比

1）MNIST 手写体数据集

MNIST 数据集为 10 分类手写体数据集，由像素大小为 28×28 的 70 000 张图像组成，本文中随机选取 60 000 张图像为训练数据集，10 000 张图像为测试数据集，部分样本可视化结构如图 5 所示.

Fig. 5 Sample visualization of MNIST dataset

图 5 MNIST 数据集中样本可视化

本数据集实验中，利用 LeNet-5 作为教师网络实现该数据集分类模型训练. 构建学生网络 LeNet-5-half，其网络结构与教师网络相同，每层通道数相比教师网络少一半，计算成本相比教师网络少 50%，可实现网络压缩. 表 1 中统计并对比了所提算法在 MNIST 数据集上的 *Accuracy* 值.

表 1 中对 10 次实验统计的均值可见，利用真实数据训练得到教师网络的 *Accuracy*=0.989 4. 由噪声数据随机生成伪样本作为训练集，在教师网络指导下，利用知识蒸馏可得到 *Accuracy*=0.867 8 的学生网

Table 1 Classification Results on MNIST Dataset

表 1 MNIST 数据集上的分类结果

算法	*Accuracy*
教师网络	0.989 4
KD[4]	0.867 8
DAFL[5]	0.968 7±0.001
DFAD[33]	0.959 6±0.002 1
RDSKD[34]	0.975 5±0.002 4
DG-DAFL	**0.980 9±0.000 9**

注：加粗为最优结果；“±”后的数值为多次实验的标准差.

络，该状态下仅利用教师网络前期训练得到的判别信息，不借助样本分布信息，难以达到满意的蒸馏效果. DAFL 方法中，通过教师网络模型判别结果回传损失，优化生成器网络，生成与真实样本分布更为接近的伪样本数据，训练学生网络，模型 *Accuracy* 值可达到 0.968 7. 本文提出的 DG-DAFL 方法相比 DAFL 方法，避免了单一生成器网络对教师网络在无标签伪样本集上判别结果过度信任所产生的无效先验优化失败问题，同时，学生网络端生成器在教师端生成器的辅助下产生更适合学生端生成器的训练样本，保证生成样本的多样性，提升识别泛化性. 同时，RDSKD 模型通过增加正则化项提升样本多样性，针对不同类样本特征较为近似的 MNIST 数据集取得了比 DAFL 与 DFAD 模型更好的分类性能. DG-DAFL 模型中，学生网络 *Accuracy* 值提升至 0.980 9，其网络性能十分接近教师网络，同时，根据 10 次实验运行结果的均值与方差可知 DG-DAFL 模型获得了更好的鲁棒性.

2）AR 人脸数据集

AR 数据集为包含 100 类的人脸数据集，由图像尺寸为 120×165 的 2 600 张图片组成，其中前 50 类为男性样本，后 50 类为女性样本，每类包含 26 张人脸图，包括不同的面部表情、照明条件、遮挡情况，是目前使用最为广泛的标准数据集. 在实验中，本文将每类的 20 张图片作为训练集，剩余的 6 张作为测试集，通过此方式对网络性能进行评价. AR 数据集可视化结果如图 6 所示.

本数据集实验中，利用 ResNet34 作为教师网络，ResNet18 作为学生网络. ResNet34 与 ResNet18 采用相同的 5 层卷积结构，ResNet34 在每层卷积结构中的层数更多，其所消耗的计算成本更高；ResNet34 的 Flops 计算量为 3.6×10^9，ResNet18 的 Flops 计算量为 1.8×10^9. 表 2中统计并对比了所提方法在 AR 数据集上的 *Accuracy*结果.

Fig. 6 Sample visualization results of AR dataset

图 6 AR 数据集的可视化结果

Table 2 Classification Results on AR Dataset

表 2 AR 数据集上的分类结果

算法	*Accuracy*
教师网络	0.865
DAFL[5]	0.676 7±0.001 3
DFAD[33]	0.52±0.003 2
RDSKD[34]	0.52±0.002 6
DG-DAFL	**0.718 3±0.001**

注：加粗为最优结果；“±”后的数值为多次实验的标准差.

实验统计结果如表 2 所示. 教师网络经包含真实标签数据集训练后 *Accuracy*=0.865. Data-Free 环境下，DAFL 模型中经知识蒸馏后学生网络的 *Accuracy*=0.676 7. AR 数据集相比 MNIST 数据集，图像类别数量提升，图像复杂度及细节增加，不同类别间样本特征分布更为近似，难以判别. DAFL 模型中生成器优化过程完全依赖教师网络判别结果，导致生成大量用于训练学生网络的噪音样本，使学生网络判别准确率与鲁棒性下降. DFAD 模型忽略教师网络对样本生成所提供的先验信息，难以获得与原训练样本分布更为近似的生成样本，极大影响学生网络识别准确率. RDSKD 模型面对的复杂特征样本集同样面临未充分利用预训练教师网络样本先验信息，导致知识蒸馏效果下降，学生网络的 *Accuracy* 仅为 0.52. 本文通过构建双生成器模型 DG-DAFL，在充分利用教师网络的潜在样本先验知识的同时，构造生成器端损失，避免对误差样本信息过学习，生成更有效且与

真实样本分布一致的伪样本. 在 AR 较为复杂的数据集上，本文所提出的 DG-DAFL 模型的 *Accuracy*=0.718 3.

3）USPS 手写体数据集

USPS 数据集为 10 类别分类数据集，由像素大小为 16×16 的 9 298 张灰度图像组成，该数据集相比于 MNIST 数据集包含的样本量更多，样本尺寸更小，且样本表达更为模糊、抽象，为识别带来了困难，USPS 数据集可视化结果如图 7 所示. 本文实验中，随机选取 7 291 张与 2007 张图像分别构建教师网络的训练集与测试集.

Fig. 7 Sample visualization results of USPS dataset

图 7 USPS 数据集的可视化结果

教师网络选择与 MNIST 数据集下相同的网络结构 LeNet-5，学生网络结构为 LeNet-5-half. 表 3 中统计并对比了所提出方法在 USPS 数据集上的 *Accuracy* 结果.

Table 3 Classification Results on USPS Dataset

表 3 USPS 数据集上的分类结果

算法	*Accuracy*
教师网络	0.96
DAFL[5]	0.926 7±0.002 1
DFAD[33]	0.889 9±0.002 4
RDSKD[34]	0.907 3±0.001 7
DG-DAFL	**0.930 2±0.001 2**

注：加粗为最优结果；“±”后的数值为多次实验的标准差.

由表 3 可知，教师网络分类 *Accuracy*=0.96，在此基础上实现 DAFL 模型. 学生网络的 *Accuracy*=0.926 7. DFAD 模型在 USPS 数据集上的 *Accuracy*=0.889 9，由于教师网络过度信任生成样本集中包含的噪音等样本，影响知识蒸馏效果及模型鲁棒性. RDSKD 模型同样存在忽略生成样本质量等问题，降低学生网络准确率. DG-DAFL 通过引入学生端生成器的双生成器方法，解决单生成器网络结构中学生网络训练过度依赖教师生成器网络样本产生的泛化性较低等问题. 同时，学生网络生成器所生成的数据在保证分布近似条件下的样本多样性，进一步提升学生网络识别泛化性的基础上，学生网络在 USPS 数据集下获得了更高的准确率及鲁棒性.

3.2 实验分析

1）DG-DAFL 消融分析

为进一步讨论所提 DG-DAFL 模型中学生端生成器G_S优化过程的合理性及损失函数各部分的必要性，本节在 MNIST 数据集上实现消融实验并分析实验结果. 表 4 统计并对比了不同损失函数部分对 Data-Free 环境下模型准确率的影响.

Table 4 Ablation Experiment Results on MNIST Dataset

表 4 MNIST 数据集上消融实验结果

伪标签损失	信息熵损失	特征损失	伪样本 KL 散度损失	*Accuracy*
√				0.868 7
				0.233 6
	√			0.114 0
		√		0.216 7
			√	0.871 1
√	√	√		0.975 8
√	√	√	√	0.980 0

注：√表示该项存在.

在消融实验中，利用真实数据训练的教师网络分类 *Accuracy*=0.983 9；学生端生成器G_S在没有任何损失函数优化的情况下，利用随机生成样本并结合教师网络知识蒸馏，*Accuracy* 达到 0.868 7. 若仅利用对随机伪样本判别结果所构造的任一损失函数，包括伪标签损失、信息熵损失、特征损失，优化学生网络生成器G_S，均难以得到满意的判别结果，其主要原因在于学生网络判别器未经过真实样本训练不包含真实先验信息，难以指导生成器训练. 若仅利用双生成器端 KL 散度作为优化信息，教师端生成器G_T经教师网络优化包含部分真实样本先验信息，可对G_S生成样本产生一定的先验监督作用，辅助生成器G_S生成相近的输出分布，在 KL 散度损失单独优化下，学生网络性能有小幅度提升. 当 3 种损失函数与生成器损失结合后，生成器G_S获得更多样本先验信息，保证生成样本与真实样本的分布一致性，并保证生成伪样本的多样性，提升学生网络模型的准确率.

2）DG-DAFL 泛化性分析

为验证所提出的 DG-DAFL 模型具有更好的泛化性，本文基于 MNIST 数据集，构建实验数据集

MNIST-F(训练集 Tra 与测试集 Te). 其中 0~9 为类别编号, 由于样本类别编号 1 和 7、0 和 8、6 和 9 等具有判别特征上的相似性, 将混淆分类模型, 为识别带来难度. 本文缩小易混淆类别训练样本规模, 具体将原始数据集中的训练样本类别编号为 1, 6, 8 的样本量减半, 测试数据量保持不变, 其详细描述如表 5 所示, 表 5 中 nTra 与 nTe 分别为原始训练集与原始测试集.

Table 5 Description of Generalizability Test Dataset

表 5 泛化性测试数据集描述

类别编号	nTra	nTe	Tra	Te
0	5 923	5 923	980	980
1	3 371	6 742	1 135	1 135
2	5 958	5 958	1 032	1 032
3	6 131	6 131	1 010	1 010
4	5 842	5 842	982	982
5	5 421	5 421	892	892
6	2 959	5 918	958	968
7	6 265	6 265	1 028	1 028
8	2 925	5 851	974	974
9	5 949	5 949	1 009	1 009

数据集 MNIST-F 实验中, 教师网络结构为 LeNet-5, 学生网络结构为 LeNet-5-half. 本文分别统计及对比了 DAFL 模型与所提出 DG-DAFL 模型的分类 *Accuracy*, 结果如表 6 所示.

Table 6 Classification Results on MNIST-F Dataset

表 6 MNIST-F 数据集上的分类结果

算法	*Accuracy*
教师网络	0.989 7
DAFL[5]	0.942 5
DG-DAFL	0.969 5

表 6 所示的是不同算法在 MNIST-F 数据集下的泛化性测试结果. DAFL 算法的 *Accuracy*=0.942 5, DG-DAFL 算法的 *Accuracy*=0.969 5, 相比在 MNIST 数据集下的测试结果, DAFL 算法的 *Accuracy* 值下降 0.026 2, DG-DAFL 的算法 *Accuracy* 值下降 0.011 4, 当在易混淆类别训练不足的情况下, 本文所提出的 DG-DAFL 模型相比 DAFL 模型具有更好的泛化性和鲁棒性. DG-DAFL 模型中的学生网络 N_S的训练数据不完全依赖于教师端生成器G_T, 避免在 DAFL 模型下由于函数 *one_hot* 构建的伪样本标签带来的大量噪声, 解决学生网络N_S鲁棒性的问题. 为便于观察与分析, 本文统计并对比了 DAFL 与 DG-DAFL 模型在 MNIST-F 数据集上的其他评价标准结果, 如表 6 和表 7 所示.

由表 7 与表 8 可知, 泛化性测试下 DG-DAFL 模型总体上比 DAFL 模型在精确率、召回率、特异度指标上均有所提升. 类别 1, 6, 8 中训练样本量减少为一半的情况下, 本文所提出的模型 DG-DAFL 在这 3 类上均获得了更好的性能. 原因在于 DG-DAFL 模型下, 训练数据由双生成器生成, 其更具多样性, 避免了单一生成器容易导致生成数据泛化性低的问题.

Table 7 Statistical Results of DAFL Model for Different Categories

表 7 DAFL 模型针对不同类别统计结果

类别编号	*Accuracy*	*Recall*	*Specificity*
0	0.993	0.957	0.999
1	**0.988**	**0.989**	**0.998**
2	0.928	0.987	0.991
3	0.910	0.988	0.989
4	0.986	0.987	0.998
5	0.908	0.952	0.991
6	**0.995**	**0.956**	**0.999**
7	0.972	0.979	0.997
8	**0.989**	**0.860**	**0.999**
9	0.973	0.966	0.997

注: 加粗为最优结果.

Table 8 Statistical Results of DG-DAFL Model for Different Categories

表 8 DG-DAFL 模型针对不同类别统计结果

类别编号	*Accuracy*	*Recall*	*Specificity*
0	0.989	0.989	0.999
1	**0.982**	**0.996**	**0.998**
2	0.954	0.995	0.994
3	0.937	0.994	0.993
4	0.989	0.983	0.999
5	0.968	0.964	0.997
6	**0.996**	**0.973**	**1.000**
7	0.982	0.982	0.998
8	**0.997**	**0.878**	**1.000**
9	0.958	0.983	0.995

注: 加粗为最优结果.

图 8~10 通过 MNIST-F 数据集下各类别的分类结果样本量及误分类样本量的混淆矩阵, 可更为清晰地观察到 DG-DAFL 模型的效果更加接近教师网络, 分类效果较优. 在真实标签为 0, 5, 6, 8, 9 上的分

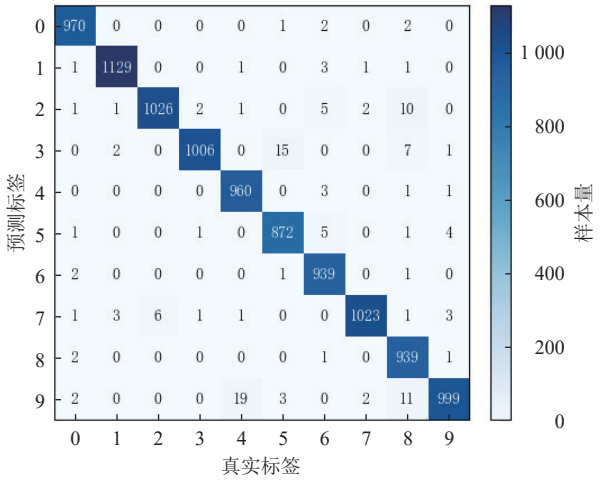

Fig. 8 Confusion matrix for teacher network generalization test

图 8 教师网络泛化性测试的混淆矩阵

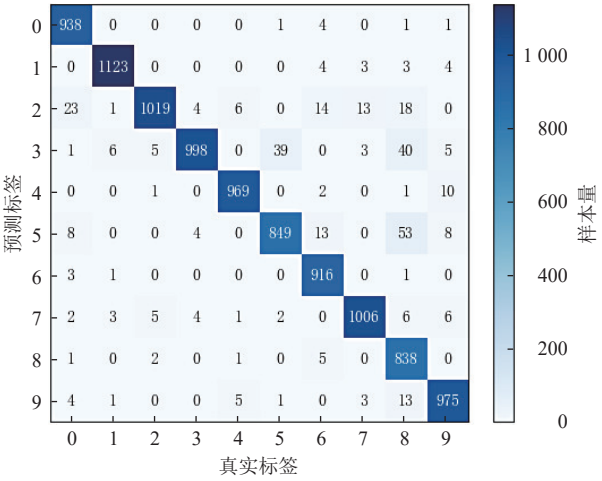

Fig. 9 Confusion matrix for DAFL generalization test

图 9 DAFL 模型泛化性测试的混淆矩阵

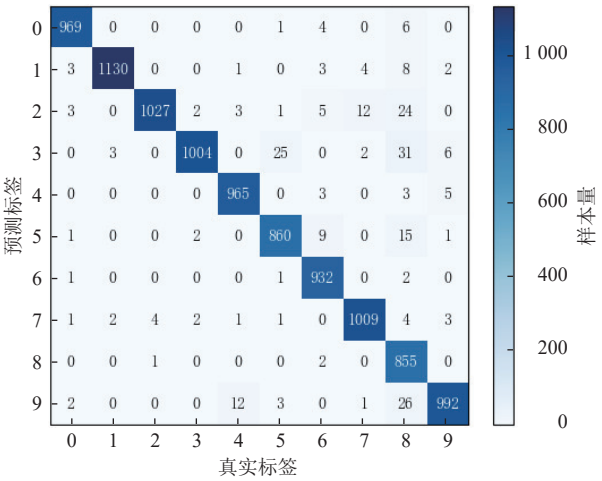

Fig. 10 Confusion matrix for DG-DAFL generalization test

图 10 DG-DAFL 模型泛化性测试的混淆矩阵

类中，DAFL 模型比 DG-DAFL 模型出现更多错误分类，其原因为 DAFL 模型的训练数据仅依赖于教师网络，教师网络生成的伪标签带来大量噪声影响生成器性能，降低学生网络性能. DG-DAFL 模型中学生网络的训练数据取决于教师端生成器和学生端生成器 2 方面的影响，避免过度依赖教师网络端生成器的情况，使得在 DG-DAFL 模型的训练过程中，生成训练数据更加接近真实数据，且保证生成图像的多样性. 同时，可观察到 DAFL 模型在易混淆的类别中将 1 类样本被误分类为 7 类样本，0, 6, 8 类样本由于模型泛化性较低而被互相混淆，产生错误的分类.

4 结　　论

本文针对 Data-Free 环境中网络压缩及知识蒸馏问题，借助 DAFL 模型通过构建生成器获得伪训练样本的学习方式，提出 DG-DAFL 网络框架. 该框架设计双生成器网络结构，保证教师网络与学生网络完成一致学习任务，并实现样本生成器与教师网络分离，避免 DAFL 模型中生成器完全信任教师网络判别结果，产生失效优化问题. 同时，在学生网络生成器训练过程中，构造双生成器端伪样本分布损失，在充分利用教师网络潜在样本分布先验信息的同时避免过度依赖，生成更具多样性的伪样本集. 本文在 3 个流行的数据集上验证了算法的有效性，并构造数据集进一步分析了算法的泛化性及鲁棒性. 然而，Data-Free 环境中生成的伪训练样本的质量将影响学生网络性能，接下来本文工作将围绕充分挖掘教师网络预训练样本结构特征等先验知识，构建更高质量的学生网络训练样本集. DG-DAFL 方法代码及模型已开源：https://github.com/LNNU-computer-research-526/DG-DAFL.git.

作者贡献声明：张晶主要负责模型提出、算法设计及论文撰写；鞠佳良负责算法实现、实验验证及论文撰写；任永功负责模型思想设计及写作指导.

参　考　文　献

[1] LeCun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural Computation, 1989, 1(4): 541−551

[2] Neill J O. An overview of neural network compression[J]. arXiv preprint, arXiv: 2006.03669, 2020

[3] Buciluǎ C, Caruana R, Niculescu-Mizil A. Model compression[C]//Proc of the 12th ACM SIGKDD Int Conf on Knowledge

Discovery and Data Mining. New York: ACM, 2006: 535−541

[4] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. Computer Science, 2015, 14(7): 38−39

[5] Chen Hanting, Wang Yunhe, Xu Chang, et al. Data-free learning of student networks[C]//Proc of the 17th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 3514−3522

[6] Anwar S, Hwang K, Sung W. Structured pruning of deep convolutional neural networks[J]. ACM Journal on Emerging Technologies in Computing Systems, 2017, 13(3): 1−18

[7] Yang Yingzhen, Yu Jiahui, Jojic N, et al. FSNet: Compression of deep convolutional neural networks by filter summary[J]. arXiv preprint, arXiv: 1902.03264, 2019

[8] Gong Yunchao, Liu Liu, Yang Ming, et al. Compressing deep convolutional networks using vector quantization[J]. arXiv preprint, arXiv: 1412.6115, 2014

[9] Jaderberg M, Vedaldi A, Zisserman A. Speeding up convolutional neural networks with low rank expansions[J]. arXiv preprint, arXiv: 1405.3866, 2014

[10] Ba L J, Caruana R. Do deep nets really need to be deep?[J]. Advances in Neural Information Processing Systems, 2014, 3: 2654−2662

[11] Urban G, Geras K J, Kahou S E, et al. Do deep convolutional nets really need to be deep and convolutional?[J]. arXiv preprint, arXiv: 1603.05691, 2016

[12] Cho J H, Hariharan B. On the efficacy of knowledge distillation[C]//Proc of the 17th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 4794 − 4802

[13] Furlanello T, Lipton Z, Tschannen M, et al. Born again neural networks[C]//Proc of the 35th Int Conf on Machine Learning. New York: ACM, 2018: 1607−1616

[14] Mirzadeh S I, Farajtabar M, Li A, et al. Improved knowledge distillation via teacher assistant[C]//Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020, 34(04): 5191−5198

[15] Yang Chenlin, Xie Lingxi, Su Chi, et al. Snapshot distillation: Teacher-student optimization in one generation[C]//Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 2859−2868

[16] Zhang Ying, Xiang Tao, Hospedales T M, et al. Deep mutual learning[C]//Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 4320−4328

[17] Heo B, Kim J, Yun S, et al. A comprehensive overhaul of feature distillation[C]//Proc of the 17th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 1921−1930

[18] Heo B, Lee M, Yun S, et al. Knowledge transfer via distillation of activation boundaries formed by hidden neurons[C]//Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 3779 − 3787

[19] Park W, Kim D, Lu Yan, et al. Relational knowledge distillation[C]//Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 3967−3976

[20] Romero A, Ballas N, Kahou S E, et al. Fitnets: Hints for thin deep nets[J]. arXiv preprint, arXiv: 1412.6550, 2014

[21] Tian Yonglong, Krishnan D, Isola P. Contrastive representation distillation[J]. arXiv preprint, arXiv: 1910.10699, 2019

[22] Tung F, Mori G. Similarity-preserving knowledge distillation[C]//Proc of the 17th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 1365−1374

[23] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in Neural Information Processing Systems, 2014, 27: 2672−2680

[24] Wang Yang. Survey on deep multi-modal data analytics: Collaboration, rivalry, and fusion[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2021, 17(S1): 1−25

[25] Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training GANs[J]. Advances in Neural Information Processing Systems, 2016, 29: 2234−2242

[26] Vondrick C, Pirsiavash H, Torralba A. Generating videos with scene dynamics[J]. Advances in Neural Information Processing Systems, 2016, 29: 613−621

[27] Nguyen A, Dosovitskiy A, Yosinski J, et al. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks[C]//Proc of the 17th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 4794−4802

[28] Bhardwaj K, Suda N, Marculescu R. Dream distillation: A data-independent model compression framework[J]. arXiv preprint, arXiv: 1905.07072, 2019

[29] Liu Pengpeng, King I, Lyu M R, et al. DDFlow: Learning optical flow with unlabeled data distillation[C]//Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 8770−8777

[30] Zhang Wentao, Miao Xupeng, Shao Yingxia et al. Reliable data distillation on graph convolutional network[C]//Proc of the 45th ACM SIGMOD Int Conf on Management of Data. New York ACM, 2020: 1399−1414

[31] Yin Hongxu, Molchanov P, Alvarez J M, et al. Dreaming to distill: Data-free knowledge transfer via deepinversion[C]//Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 8715−8724

[32] Lopes R G, Fenu S, Starner T. Data-free knowledge distillation for deep neural networks[J]. arXiv preprint, arXiv: 1710.07535, 2017

[33] Fang Gongfan, Jie Song, Shen Chenchao, et al. Data-free adversarial distillation[J]. arXiv preprint, arXiv: 1912.11006, 2019

[34] Han Pengchao, Park J, Wang Shiqiang, et al. Robustness and diversity seeking data-free knowledge distillation[C]//Proc of the 46th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2021: 2740−2744

[35] Nayak G K, Mopuri K R, Shaj V, et al. Zero-shot knowledge distillation in deep networks[C]//Proc of the 36th Int Conf on

Machine Learning. New York: ACM, 2019: 4743−4751

[36] Micaelli P, Storkey A J. Zero-shot knowledge transfer via adversarial belief matching[J]. Advances in Neural Information Processing Systems, 2019, 32: 9551−9561

[37] Kimura A, Ghahramani Z, Takeuchi K, et al. Few-shot learning of neural networks from scratch by pseudo example optimization[J]. arXiv preprint, arXiv: 1802.03039, 2018

[38] Shen Chengchao, Wang Xinchao, Yin Youtan, et al. Progressive network grafting for few-shot knowledge distillation[C]//Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021, 35(3): 2541−2549

[39] Micaelli P, Storkey A J. Zero-shot knowledge transfer via adversarial belief matching[J]. Advances in Neural Information Processing Systems, 2019, 32: 9551−9561

[40] Radosavovic I, Dollár P, Girshick R, et al. Data distillation: Towards omni-supervised learning[C]//Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 4119−4128

Zhang Jing, born in 1984. PhD, associate professor. Her main research interests include machine learning and reinforcement learning.

张　晶, 1984 年生. 博士, 副教授. 主要研究方向为机器学习与强化学习.

Ju Jialiang, born in 1998. Master. His main research interests include deep learning and machine learning.

鞠佳良, 1998 年生. 硕士. 主要研究方向为深度学习与机器学习.

Ren Yonggong, born in 1972. PhD, professor. His main research interests include data mining and artificial intelligence.

任永功, 1972 年生. 博士, 教授. 主要研究方向为数据挖掘和人工智能.