基于拓扑一致性对抗互学习的知识蒸馏

赖 轩 [1] 曲延云 [1] 谢 源 [2] 裴玉龙 [1]

摘 要 针对基于互学习的知识蒸馏方法中存在模型只关注教师网络和学生网络的分布差异, 而没有考虑其他的约束条件, 只关注了结果导向的监督, 而缺少过程导向监督的不足, 提出了一种拓扑一致性指导的对抗互学习知识蒸馏方法 (Topology-guided adversarial deep mutual learning, TADML). 该方法将教师网络和学生网络同时训练, 网络之间相互指导学习, 不仅采用网络输出的类分布之间的差异, 还设计了网络中间特征的拓扑性差异度量. 训练过程采用对抗训练, 进一步提高教师网络和学生网络的判别性. 在分类数据集 CIFAR10、CIFAR100 和 Tiny-ImageNet 及行人重识别数据集 Market1501 上的实验结果表明了 TADML 的有效性, TADML 取得了同类模型压缩方法中最好的效果.

关键词 互学习, 生成对抗网络, 特征优化, 知识蒸馏

引用格式 赖轩, 曲延云, 谢源, 裴玉龙. 基于拓扑一致性对抗互学习的知识蒸馏. 自动化学报, 2023, **49**(1): 102−110

DOI 10.16383/j.aas.c200665

Topology-guided Adversarial Deep Mutual Learning for Knowledge Distillation

LAI Xuan[1] QU Yan-Yun[1] XIE Yuan[2] PEI Yu-Long[1]

Abstract The existing mutual-deep-learning based knowledge distillation methods have the limitations: the discrepancy between the teacher network and the student network is only used to supervise the knowledge transfer neglecting other constraints, and the result-driven supervision is only used neglecting process-driven supervision. This paper proposes a topology-guided adversarial deep mutual learning network (TADML). This method trains multiple classification sub-networks of the same task simultaneously and each sub-network learns from others. Moreover, our method uses an adversarial network to adaptively measure the differences between pairwise sub-networks and optimizes the features without changing the model structure. The experimental results on three classification datasets: CIFAR10, CIFAR100 and Tiny-ImageNet and a person re-identification dataset Market1501 show that our method has achieved the best results among similar model compression methods.

Key words Mutual learning, generative adversarial network, feature optimization, knowledge distillation

Citation Lai Xuan, Qu Yan-Yun, Xie Yuan, Pei Yu-Long. Topology-guided adversarial deep mutual learning for knowledge distillation. *Acta Automatica Sinica*, 2023, **49**(1): 102−110

图像分类是计算机视觉领域的一个经典任务, 有广泛的应用需求, 例如机场和车站闸口的人脸识别、智能交通中的车辆检测等, 图像分类的应用在一定程度上减轻了工作人员的负担, 提高了工作效率. 图像分类的解决方法也为目标检测、图像分割、场景理解等视觉任务奠定了基础. 近年来, 由于 GPU 等硬件和深度学习技术的发展, 深度神经网络 (Deep neural network, DNN)[1] 在各个领域取得了长足的进展, 比如, 在 ImageNet 大规模视觉识别挑战赛 ILSVRC 比赛库上的图像分类, 基于深度学习的图像分类方法已经取得了与人类几乎相同甚至超越人类的识别性能. 然而, 这些用于图像分类的深度学习模型往往需要较高的存储空间和计算资源, 使其难以有效的应用在手机等云端设备上. 如何将模型压缩到可以适应云端设备要求, 并使得性能达到应用需求, 是当前计算机视觉研究领域一个活跃的研究主题. 轻量级模型设计是当前主要的解决途径, 到目前为止, 模型压缩方法大致分为基于模型设计的方法[2]、基于量化的方法[3]、基于剪枝的方法[4]、基于权重共享的方法[5]、基于张量分解的方法[6] 和基于知识蒸馏的方法[7] 六类.

本文主要关注知识蒸馏方法. 知识蒸馏最初被用于模型压缩[8]. 不同于剪枝、张量分解等模型压缩方法, 知识蒸馏 (Knowledge distillation, KD) 的方法, 先固定一个分类性能好的大模型作为教师网络,

收稿日期 2020-08-18 录用日期 2020-12-23
Manuscript received August 18, 2020; accepted December 23, 2020
国家自然科学基金 (61876161, 61772524, 61671397, U1065252, 61772440), 上海市人工智能科技支撑专项 (21511100700) 资助
Supported by National Natural Science Foundation of China (61876161, 61772524, 61671397, U1065252, 61772440) and Shanghai Science and Technology Commission (21511100700)
本文责任编委 张敏灵
Recommended by Associate Editor ZHANG Min-Ling
1. 厦门大学信息学院 厦门 361005 2. 华东师范大学计算机科学与技术学院 上海 200064
1. School of Informatics, Xiamen University, Xiamen 361005
2. Department of Computer Science & Technology, East China Normal University, Shanghai 200064

然后训练一个轻量级模型作为学生网络学习教师网络蒸馏出来的知识, 在不增加参数量的情况下提升小模型的性能. 基于知识蒸馏的模型压缩方法, 将教师网络输出的预测分布视为软标签, 用于指导学生网络的预测分布, 软标签反映了不同类别信息间的隐关联, 为新网络的训练提供了更丰富的信息, 通过最小化两个网络预测的 Kullback-Leibler (KL) 散度差异, 来实现知识迁移. Romero 等[9] 认为让小模型直接在输出端模拟大模型时会造成模型训练困难, 从而尝试让小模型去学习大模型预测的中间部分, 该方法提取出教师网络中间层的特征图, 通过一个卷积转化特征图大小来指导学生网络对应层的特征图. Yim 等[10] 使用 FSP (Flow of solution procedure) 矩阵计算卷积层之间的关系, 让小模型去拟合大模型层与层之间的关系. Peng 等[11] 和 Park 等[12] 同时输入多个数据, 在原知识蒸馏模型的基础上通过学习样本之间的相关性进一步提升学生网络性能.

考虑到知识蒸馏的本质是知识的迁移, 即将知识从一个模型迁移到另一个模型, Zhang 等[13] 提出了深度互学习 (Deep mutual learning, DML) 方法, 设计了一种蒸馏相关的相互学习策略, 在训练的过程中, 学生网络和教师网络可以相互学习, 知识不仅从教师网络迁移到学生网络, 也从学生网络迁移到教师网络.

协同学习也是常见的迁移学习方法之一, 多用于半监督学习. 在协同学习中, 不同的模型或者在不同分组的数据集上学习, 或者通过不同视角的特征进行学习, 例如识别同一组物体类别, 但其中一种模型输入 RGB 图像, 而另一种模式输入深度图像. 协同属性学习[14] 就是通过属性矩阵的融合进行属性的挖掘, 从而指导两个模型的分类. 而深度互学习方法中所有模型在同一数据集上训练完成相同的任务.

尽管现有的知识蒸馏的方法已经取得了长足的进展, 但仍存在以下问题: 1) 现有的深度互学习方法仅关注教师网络和学生网络输出的类分布之间的差异, 没有利用对抗训练来提升模型的判别能力; 2) 现有的深度互学习仅关注结果监督, 忽视了过程监督. 特别是没有考虑高维特征空间中拓扑关系的一致性. 针对问题 1), 本文设计对抗互学习框架, 生成器使用深度互学习框架, 通过对抗训练, 提高教师和学生网络的判别性; 针对问题 2), 本文在教师网络和学生网络互学习模型中, 增加过程监督, 即对中间生成的特征图, 设计了拓扑一致性度量方法, 通过结果和过程同时控制, 提高模型的判别能力.

总之, 本文提出了一种基于拓扑一致性的对抗互学习知识蒸馏方法 (Topology-guided adversarial deep mutual learning, TADML), 在生成对抗[15] 网络架构下, 设计知识蒸馏方法, 教师网络和学生网络互相指导更新, 不仅让教师网络的知识迁移到学生网络, 也让学生网络的知识迁移到教师网络. 本文的模型框架可以推广到多个网络的对抗互学习. TADML 由深度互学习网络构成的生成器和一个判别器组成. 生成器的每个子网络都是分类网络. 类似于知识蒸馏, 任一子网络都可以看作是其余网络的教师网络, 对其他网络训练更新, 进行知识迁移. 为方便计算, 本文将所有子网络组视为一个大网络同时优化更新. 每个被看作生成器的子网络, 生成输入图像的特征. 判别器更新时判断生成器的输出特征属于哪一个类别、来源于哪一个子网络, 而生成器更新时尽量混淆判别器使其无法准确判断特征来源于哪一个生成器, 进而拟合网络中隐含的信息.

1 本文方法

本节介绍如何通过对抗训练框架实现网络间的知识转移. 首先概述 TAMDL 网络结构, 然后讨论所提的损失函数的构成, 最后描述模型的训练过程.

1.1 网络结构

如图 1 所示, 给出了基于拓扑一致性的对抗互学习知识蒸馏 (TADML) 框架, 该框架由生成器和判别器两部分组成:

1) 生成网络. 该部分由两个或多个分类子网络组成, 生成器中的分类网络执行相同的分类任务, 可以选取不同的模型结构, 彼此间无需共享参数. 不失一般性, 现有的深度分类模型都可作为生成器中的分类网络, 例如 ResNet 和 Wide-ResNet[16]. 由于所有的生成网络使用相同的数据集执行相同的分类任务, 对于输入图像 x, 定义第 i 个网络的激活函数层 Softmax 的类别分布概率值为 $f_i(x, \omega_i)$, 其中 ω_i 是相应的分类模型网络参数.

2) 判别器. 在 TADML 架构中, 将两个或多个分类网络看作生成器, 而判别器只有一个. 由于常见的判别器容易陷入过早收敛或难以训练两种极端情况, 本文设计了一个能较好平衡判别器稳定性和辨别能力的判别器, 相对于常见的多层感知器[17] 更加稳定. 如图 2 所示, 提出的判别器由三个全连接的层 (128fc-256fc-128fc) 组成, 且判别器的第一层和与最后一层没有批标准化处理 (Batch normalization, BN) 与 LeakyRelu 激活函数操作. 与常见的判别器不同, 本文所设计判别器的输出不是简单的

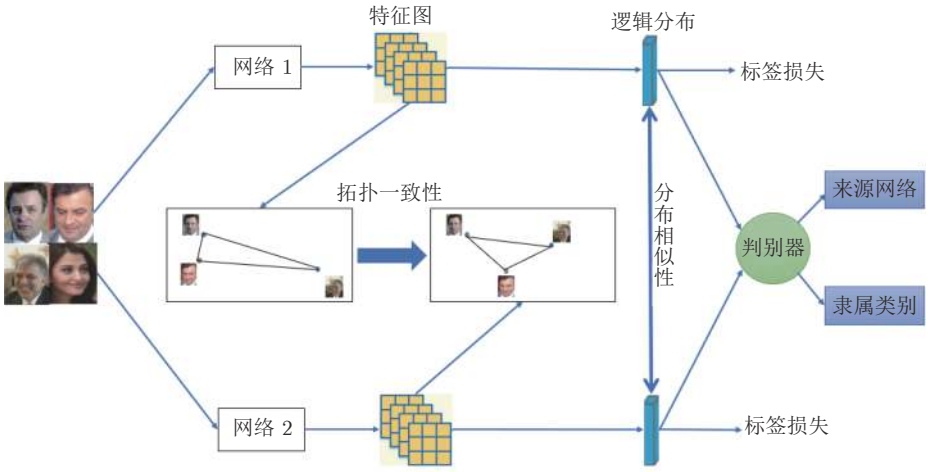

图 1 本文方法框架

Fig. 1 The framework of the proposed method

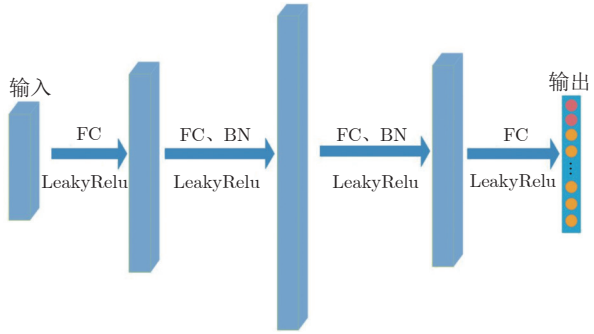

图 2 判别器结构图

Fig. 2 The structure of discriminator

真假 (自然图像/伪造图像), 而是判断输入来源于哪个网络且隶属于哪个类别. 受到条件 GAN (Conditional-GAN, C-GAN[18]) 在图像恢复领域中的启发, 本文根据 C-GAN 的对判别器的输入进行改造, 在后续的消融实验部分对判别器的输入进行不同程度的约束.

1.2 损失函数

所提方法考虑四种损失: 标签监督损失 L_S, 对抗损失 L_{adv}, 分布一致性损失 L_b, 拓扑一致性损失 L_T. 标签监督损失 L_S 是广泛用于图像分类中带注释数据分类任务的监督损失, 这对提取知识起着至关重要的作用. 分布一致性损失 L_b 是直接匹配所有分类子网络的输出的显式损失, 而对抗性损失 L_{adv} 表示隐式损失, 该损失将所有分类子网络的逻辑分布之间经过分类器判断的差异最小化. 换句话说, 对抗性损失提供了一些通过传统分布相似性度量而丢失的信息. 拓扑一致性损失 L_T 是样本实例间隐藏的高阶结构信息.

在训练对抗生成抗网络时, 为指导网络的学习, 尽可能迁移分类网络之间的知识, 总的损失函数定义为:

$$L_{total} = \alpha L_S + \beta L_b + L_{adv} + L_T \tag{1}$$

式中, α和β 分别表示四项损失所占的权重, 在本文中分别设定为 $\alpha = 0.6$, $\beta = 0.4$. 下面依次对这四个部分进行详细说明.

1) 标签监督损失. 该损失为常用的监督分类交叉熵损失. 对于给定的图像标签对 $(x; l)$, 优化模型参数使得预测类别与标签的交叉熵降至最低, 以正确预测每个训练实例的真实标签:

$$L_S(f) = \frac{1}{N}\sum_{i=1}^{N} L_{CE}(l_i, f(x_i)) \tag{2}$$

2) 分布一致性损失. 考虑到互学习模型中的知识迁移, 与之前的蒸馏网络不同, 本文没有固定一个预训练网络作为教师网络进行单向指导, 所提方法中任意一个网络都接受其余网络的监督指导, 最小化分类网络输出特征的类别分布差异, 输出越相似则表示迁移效果越好. 受到 Knowledge squeezed adversarial network compression (KSANC)[19] 的启发, 本文考虑从结果导向和过程导向两个方面同时进行知识迁移. 过程导向约束仅针对最后一个全连接层的输出. 最终输出的逻辑分布作为结果导向, 即各个网络之间只保留网络输出之间的实例级对齐.

考虑到网络输出的类别分布的差异性度量, 本文使用 Jensen-Shannon (JS) 散度衡量输出分布的相似性:

$$L_b = JS(f_i, f_j) = KL(f_i, f_j) + KL(f_j, f_i) \tag{3}$$

式中, f_i 表示由第 i 个网络预测的逻辑分布. KL 散度定义为:

$$KL(f_i,\, f_j) = f_i \ln \frac{f_i}{f_j} \tag{4}$$

3) 对抗性损失. 在 TADML 的模型中, 采用对抗学习 (GAN) 的方法, 将从每个网络中提取的知识转移到另一个网络中. 在知识蒸馏中, 学生网络通过模仿教师网络从而学习教师网络中的知识, 直到最后学生网络的输出与教师网络相近则视为指导完成. TADML 网络整体框架分为生成器和判别器两个部分, 多个分类网络构成生成器. 对于一个输入的样本, 经过生成网络得到多个类别概率, 每一个分类网络都对应输出一个概率分布 (也可以视为图像经过这个网络表征的特征编码). 这些概率分布作为判别器的输入, 判别器判断类别概率分布是由哪个分类网络产生. 生成器与判别器交替迭代更新, 固定判别器更新生成器时, 尽量生成相似的特征编码, 使得判别器无法分辨特征编码来自于生成器的哪一个子网络; 而在固定生成器更新判别器时, 尽量训练判别网络, 使其可以轻易的分辨输入来源于生成器中哪个分类子网络. 二者交替迭代直到动态平衡, 则视为收敛.

到目前为止, 基于 GAN 的方法已在很多领域取得了显著的效果, 在 TADML 方法中, 每个分类子网络都被视为 GAN 中的生成器, 并提供逻辑分布作为另一个分类子网络的真实标签. 相较于原始的 GAN 网络只输出一个布尔值, 即真或假, 本文判别器判断其输入来源于哪个分类子网络:

$$L_{adv}^{o} = \min_{f_i(x)} \max_{D} \sum_{i} L_{CE}(g_n(mi),\, mD^{o}(mf_i(x))) \tag{5}$$

式中, $g_n(i)$ 是第 i 个元素为 1, 其余元素为 0 的向量, 表示生成器 n 个分类子网络的第 i 个分类网络的输出作为判别器的输入, $D^{o}(f_j(x))$ 表示判别器输出的 n 位向量, 代表判别器预测输入来源于哪个网络, n 为分类子网络数.

此外, 如果判别器仅仅区分输入来自生成器的哪个子网络, 则缺少类别信息可能导致错误的关联. 为此, 引入辅助分类来预测输入所属类别. 即本文所提的判别器不仅需要判断输入来源于哪个分类子网络, 还需要判断输入属于哪一个类别标签, 损失函数表示为:

$$L_{adv}^{C} = \min_{f_i(x)} \max_{D} \sum_{i} L_{CE}(g_N(C),\, D^{C}(f_i(x))) \tag{6}$$

式中, $g_N(C)$ 表示真实的类别分布, $D^{C}(f_i(x))$ 表示判别器输出的类别分布, N 是类别总数.

鉴于 GAN 网络的判别器容易在极少的迭代次数后收敛和过度拟合. 本文设计了惩罚项作为对模型的正则化处理, 定义如下:

$$L_{adv}^{reg} = -\mu(\|\omega_D\|_2^2 - L_{CE}(g(0),\, D(f(x)))) \tag{7}$$

式中, μ 权重参数设为 0.7, ω_D 是判别器的网络参数, $g(0)$ 表示元素全为 0 的向量, 负号表示该项仅在式 (5) 最大化步骤中更新, 前一项迫使判别器的权重缓慢增长, 后一项则是对抗性样本正则化.

本文设计的对抗损失为:

$$L_{adv} = L_{adv}^{o} + L_{adv}^{C} + L_{adv}^{reg} \tag{8}$$

4) 拓扑一致性损失. 在过程导向的监督学习中, 考虑样本组间的拓扑结构相似性, 本文选择计算样本在高维空间嵌入特征的距离及其角度的一致性. 对于输入的样本组 $\{x_1, x_2, x_3, \cdots, x_n\}$, 经过第 i 个分类网络的最后一层全连接输出的特征映射看作高维嵌入特征 $\{h_i(x_1), h_i(x_2), h_i(x_3), \cdots, h_i(x_n)\}$, 则两个网络间基于特征距离的拓扑一致性损失可以表示为:

$$L_T^D = \sum_{i \neq j} \left| \varphi_D^{(1)}(x_i,\, x_j) - \varphi_D^{(2)}(x_i,\, x_j) \right| \tag{9}$$

式中, $\varphi_D^{(s)}(x_i,\, x_j)$ 表示样本 $x_i,\, x_j$ 在第 s 个网络中高维嵌入特征的位置距离, 本文采用归一化的二范式表示:

$$\varphi_D(x_i,\, x_j) = \frac{1}{\phi} \|h(x_i) - h(x_j)\|_2 \tag{10}$$

式中, ϕ 是距离归一化系数, 表达式为:

$$\phi = \sum_{i \neq j} \|h(x_i) - h(x_j)\|_2 \tag{11}$$

同理, 两个网络间基于特征角度的一致性损失可以表示为:

$$L_T^A = \sum_{i \neq j} \max\left(\varphi_A^{(1)}(x_i,\, x_j) - \varphi_A^{(2)}(x_i,\, x_j),\, 0\right) \tag{12}$$

式中, $\varphi_A^{(s)}(x_i,\, x_j)$ 表示样本 $x_i,\, x_j$ 在第 s 个网络中高维嵌入特征的夹角, 本文采用特征向量的内积表示:

$$\varphi_A(x_i,\, x_j) = \frac{h^{\mathrm{T}}(x_i) h(x_j)}{\|h(x_i)\|_2 \|h(x_j)\|_2} \tag{13}$$

则两个网络之间的拓扑一致性损失可以表示为:

$$L_T = \lambda_1 L_T^D + \lambda_2 L_T^A \tag{14}$$

式中, λ_1 和 λ_2 分别表示两项损失所占的权重, 在本文中分别设定 $\lambda_1 = 1$ 和 $\lambda_2 = 2$.

1.3 训练步骤

在训练过程中, 本文交替更新判别器和生成器. 在更新生成器参数时, 固定判别器不动, 将生成器

的所有分类网络视为一个整体, 通过最小化式 (1) 同时更新生成器中所有的分类网络参数. 在更新判别器参数时, 所有的生成网络都是固定的, 以提供稳定的输入, 通过最大化式 (8) 更新. 交替迭代更新, 每输入一组数据交替一次, 直至迭代次数满足终止条件. 在测试阶段, 本文仅考虑作为生成器的分类子网络, 并将每个分类子网络视为一个完整的分类网络来对输入图像分别进行分类.

2 实验设置

2.1 数据集

本文在 3 个公开的分类数据集 CIFAR10、CIFAR100 和 Tiny-ImageNet 上进行训练和测试, 进一步在行人重识别数据集 Market1501 上验证所提方法的有效性. 其中, CIFAR100 和 CIFAR10 数据集都包含 60 000 张 32 × 32 像素大小的图像, 分别由 100 个类和 10 个类组成, 50 000 张用于训练, 10 000 张用于验证. Tiny-ImageNet 源于 ImageNet dataset (1 000 个类别), 从中抽取 200 个类别, 每个类别有 500 个训练图像, 50 个验证图像和 50 个测试图像, 且所有图片都被裁剪放缩为 64 × 64 像素大小. Market1501 是常用的行人重识别数据集, 包含 12 936 张训练图像 (751 个不同的行人) 和 19 732 张测试图像 (750 个不同的行人), 图像大小为 64 × 128 像素.

2.2 实现细节

本文算法使用 Torch0.4 在 NVIDIA GeForce GTX 1 080 GPU 上实现. 对于所有分类数据集, 均使用随机梯度下降法进行优化, 将权重衰减设置为 0.0001, 动量设置为 0.9. 对于 CIFARs 的实验, 批量大小设置为 64, 生成网络和判别器的初始学习率分别设置为 0.1 和 0.001, 每隔 80 次迭代两者都缩小为 0.1 倍, 总共训练了 200 次迭代. 对于 Tiny-ImageNet 的实验, 批量大小设置为 128, 总迭代次数为 330 代, 生成网络初始学习率设为 0.1, 每隔 60 代学习率乘以 0.2, 判别网络初始学习率为 0.001, 每隔 120 代乘以 0.1. 对于 Market1501 的实验, 采用与 DML 相同的实验设置: 使用 Adam 优化器, 学习率为 0.0002, β_1 设为 0.5, β_2 设为 0.999, 批量大小设置为 16, 图像输入大小为 64 × 160 像素, 共迭代 100 000 次. 尽管使用预训练模型能得到更高的精度, 在实验中, 所有网络都采用随机初始化的. 由于训练前期网络变化较大, 仅在总迭代次数过半的时候才加入拓扑一致性损失更新网络, 且用上一次迭代时分类精度高的网络指导精度低的网络, 而不是互相指导学习.

2.3 消融实验

关于损失函数的选择, 本文尝试不同损失组合的效果. 表 1 展示了在 CIFAR10 和 CIFAR100 上, 将两个 ResNet32 设置为生成器中的教师网络和学生网络, 遵循相同的实验方案进行训练, 并选择这两个子网络的平均精度作为最终结果. 其中, L_S 表示标签损失, L_p $(p = 1, 2)$ 表示两个网络输出分布之间的 l_1, l_2 范数损失, L_{JS} 表示两个网络输出分布的 L_{JS} 散度相似性, L_{adv} 表示本章提出的对抗损失. 从表中可知, 单独使用类别标签监督损失 L_S 在所有组合中结果最差, 增加任意一种知识迁移的损失都能增加预测的精度, $L_S + L_{JS} + L_{adv}$ 取得最高的平均分类精度, 在 CIFAR10 和 CIFAR100 上增幅分别为 0.62% 和 2.28% 在固定类别标签监督损失 L_S 和对抗损失 L_{adv} 的情况下, 对比增加 L_2 和 JS 损失, 前者增加 L_{JS} 比增加 L_2 使得分类性能有所提升, 在两个数据集上的增幅分别为 0.48% 和 0.78%. 综上所述, 在后续的实验中, 单独使用 L_{JS} 差异来计算 L_b.

表 1 损失函数对分类精度的影响比较 (%)
Table 1 Comparison of classification performance with different loss function (%)

损失构成	CIFAR10	CIFAR100
L_S	92.90	70.47
$L_S + L_{JS}$	93.18	71.70
$L_S + L_{JS} + L_{adv}$	93.52	72.75
$L_S + L_1 + L_{adv}$	93.04	71.97
$L_S + L_2 + L_{adv}$	93.26	72.02
$L_S + L_1 + L_{JS} + L_{adv}$	92.87	71.63
$L_S + L_2 + L_{JS} + L_{adv}$	92.38	70.90
$L_S + L_{JS} + L_{adv} + L_T$	93.05	71.81

进一步讨论判别器结构对 TAMDL 性能的影响. 在 CIFAR100 上进行实验, 在分类子网络固定为 ResNet32 的情况下, 讨论判别器采用不同的架构对最终网络的分类误差的影响. 由表 2 可以看出, 不同结构的判别器对结果的影响不大. 尝试了两层到四层不同容量的全连接层模型, 且为了尽可能保留输入数据的差异性, 仅在全连接层之间进行 BN 与 LeakyReLU 操作. 实验表明四层全连接层的效果普遍会略低于三层的效果, 三层结构的判别器取得了略优的分类性能, 128fc-256fc-128fc在 CIFAR100 上取得了最好的分类性能, 相比最差的四层结构的判别器 128fc-256fc-256fc-128fc 分类精度仅

表 2 判别器结构对分类精度的影响比较 (%)
Table 2 Comparison of classification performance with different discriminator structures (%)

结构	CIFAR100
256fc-256fc	71.57
500fc-500fc	72.09
100fc-100fc-100fc	72.33
128fc-256fc-128fc	72.51
64fc-128fc-256fc-128fc	72.28
128fc-256fc-256fc-128fc	72.23

提高了 0.28. 为此, 在后续实验中, TAMDL采用三层结构的判别器.

本节讨论判别器的输入对 TAMDL 性能的影响. 在 2 个 ResNet32 构成的网络上进行了实验. 对比了不同的判别器的输入: 1) Conv4 表示图像经过第 4 组卷积得到的特征; 2) FC 表示单张图像经过全连接层转化但未经 Softmax 的特征; 3) DAE 表示原始图像经过深度自编码器得到的压缩特征; 4) Label 表示分类标签的热编码; 5) Avgfc 表示一组图像经过全连接层转化但未经 Softmax 的特征的平均值. 表 3 对比了针对不同判别器输入网络的最终结果, 表中的结果是经过分类网络输出的平均值. 由表 3 可以看出, FC 得到的特征作为判别器的输入取得了最好的判别性能, 增加的条件约束信息对最终结果没有正面的促进, 如 FC + Conv4 判别器的性能并没有提升, 反而下降了 0.44%. FC + Label 作为输入, 判别器性能仅次于 FC 作为输入得到的结果.

表 3 判别器输入对分类精度的影响比较 (%)
Table 3 Comparison of classification performance with different discriminator inputs (%)

输入约束	CIFAR100
Conv4	72.33
FC	72.51
Conv4 + FC	72.07
FC + DAE	71.97
FC + Label	72.35
FC + Avgfc	71.20

进一步讨论采样数量对 TAMDL 分类性能的影响. 在训练过程中通常采用从训练数据集中随机采样来训练网络. 不加限制的随机采样器可能会导致所有样本都来自不同类别的情况. 尽管它是对实例一致性的真实梯度的无偏估计, 但是在本节提出的样本组间结构相似性损失计算中, 过多的样本类别数容易导致组间关系过于复杂难以学习优化, 且过少的样本类别数又容易导致类间相关性偏差较大. 为了正确的传递样本组间的真实相关信息, 采样策略十分重要. 在批量输入大小固定为 64 的情况下, 对样本组中的类别数目进行了限定. 表 4 给出了在 CIFAR100 数据集上, 学生和教师网络为 ResNet32 和 ResNet110 时的分类结果, 其中每个样本组中类别总数为 K 且每类的样本数目为 $64/K$, Random 表示不进行采样约束的互学习结果, Vanila 表示原始网络精度. 由表 4 可知, 当类别总数 K 取值过小时, 网络无法正常训练或过早陷入过拟合状态. 如 $K = 2$, TADML 取得最低的分类性能. 当 K 取值刚好等于类别总数时, 即每个类别样本仅出现一次, 网络的性能与随机采样效果基本保持一致. 在 $K = 8$, 16, 32 时, TAMDL 的性能均优于随机采样的方式, 增幅分别为 0.31%、0.72%、0.38%. 由此可知, 样本组的类别数在平衡类间内相关一致性中有很重要的作用, 选取适当的类别数, 后续实验采用 $K = 16$.

2.4 TAMDL 与 DML 比较实验

本节讨论 TAMDL 与 DML 的性能对比. 为了说明 TAMDL 的鲁棒性和优越性, 实验设置不同结构的分类网络作为生成器, 并与原始分类网络和深度互学习方法 (DML) 进行比较. 对比实验的优化器参数设置与本文提出算法保持一致, DML 算法优化步骤按照原文的设置, 使用 KL 散度进行知识迁移并交替训练子网络. 为了进一步说明本文所提两个损失模块的有效性, 把仅加上对抗损失模块的网络 (损失函数未加拓扑一致性损失度量) 定义为 ADML. 实验部分列出了 ADML 算法与同时使用对抗性损失模块、拓扑一致性损失模块的 TADML 算法的测试结果. 由表 5 可以看出, 本文方法在 ResNet32, ResNet110 和 Wide-ResNet (WRN) 之间的几乎所有组合中, 都比 DML 表现更好, 无论

表 4 采样数量对分类精度的影响比较 (%)
Table 4 Comparison of classification performance with different sampling strategies (%)

网络结构	Vanila	Random	$K = 2$	$K = 4$	$K = 8$	$K = 16$	$K = 32$	$K = 64$
Resnet32	71.14	72.12	31.07	60.69	72.43	72.84	72.50	71.99
Resnet110	74.31	74.59	22.64	52.33	74.59	75.18	75.01	74.59

两个网络是同等大小, 还是一大一小, 大网络几乎都可以从小网络中进一步获益, 从而达到更高的精度. 换句话说, ADML 进一步提升了所有网络的能力. 表 5 中除第 1 行外, 第 2 ~ 5 行所有的教师和学生网络结构模型, ADML 的性能都优于 DML. 学生网络 (第 1 列) 的第 2 ~ 5 行增幅分别为 1.04%、0.49%、0.71%、1.03%, 教师网络 (第 2 列) 的第 2 ~ 5 行增幅分别为 0.1%、0.55%、0.74%、0.32%. 当在 CIFAR10 上重复相同的实验时, 由于生成网络的输出过于简单导致基于 GAN的优化难以收敛, 提出的 ADML 的性能几乎等于 DML.

由表 5 可以看出, TADML 在所有的网络结构试验中几乎都达到了最优的结果, 最优值用黑体标记, 次优值用下划线标记. 相对于 DML, TADML 在所有设置的网络结构中都优于 DML, 学生网络的增幅分别为 1.21%、1.52%、0.93%、0.91% 和 1.52%, 教师网络的增幅分别为 1.24%、0.78%、1.16%、1.07% 和 1.01%. 进一步可以发现, 当 2 个分类子网络大小不一致时, 较大网络的提升效果远没有较小网络明显.

将本文方法用于行人再识别, 用平均识别精度 mAP 进行度量. 为公平比较起见, 采用了与 DML[13] 在行人在识别实验中相同的网络设置, 设置了 2 组不同网络学生和教师的架构: 网络 1(InceptionV, MobileNetV1)、网络 2 (MobileNetV1, MobileNetV1). 对比 DML、ADML 和 TADML, 结果如表 6 所示. 在行人重识别数据集上的性能进一步表明了, 本文算法的有效性和优越性. ADML 相对于 DML, 2 组师生网络性能分别提升了 0.26% 和 0.35%、0.47% 和 1.01%; TADML 相对于 DML, 两组师生网络性能分别提升了 0.59% 和 1.04%、0.89% 和 1.39%. 实验结果表明, ADML 和 TADML 方法在 Market1501 数据集上的 mAP 普遍高于 DML.

2.5 主流方法对比

将本文 TAMDL 方法与当前流行的方法进行比较, 为比较公平, 将模型压缩的性能作为比较指标, 在三个常见的分类数据集 CIFAR10、CIFAR100、Tiny-ImageNet 上进行比较. 对比了 9 种方法, 分别为 2 种广泛使用的基于量化的模型压缩方法: Quantization[20]、Binary Connect[21], 4 种常见的知识蒸馏方法: 解过程流方法 (Flow of solution procedure, FSP)[10]、模拟浅层神经网络的 SNN-MIMIC 方法[22]、KD[8]、用浅而宽的教师网络训练窄而深的学生网络的 FitNet[9], 3 种对抗训练的蒸馏方法: 对抗网络压缩方法 (Adversarial network compression, ANC[23]、用条件对抗学习加速训练学生网络的 TSANC 方法[24]、用知识挤压进行对抗学习的 KSANC 方法[19]. 其中 Quantization[20] 将网络权重的进行三值化, Binary Connect[21] 在前向和后向传递期间对权重进行二值化. SNN-MIMIC[22] 模拟学习 L_2 损失, KD[8] 通过 KL 散度进行软目标的知识转移, Yim 等[10]使用 FSP 矩阵进行蒸馏, FitNet[9] 使用更深但更薄的网络尝试迁移模型中间层的知识. ANC[23] 首次将生成对抗网络融入到知识蒸馏中对学生网络的逻辑分布层进行指导, TSANC[24] 在此基础上对判别器的输入进行了条件约束, KSANC[19] 进一步加入了网络中间层的监督指导.

在对比实验中, 教师网络使用 ResNet164, 学生网络使用 ResNet20. 其中 Tiny-ImageNet 的实

表 5 网络结构对分类精度的影响比较 (%)

Table 5 Comparison of classification performance with different network structures (%)

网络结构		原始网络		DML[13]		ADML		TADML	
网络 1	网络 2	网络 1	网络 2	网络 1	网络 2	网络 1	网络 2	网络 1	网络 2
ResNet32	ResNet32	70.47	70.47	71.86	71.89	**72.85**	**72.89**	**73.07**	**73.13**
ResNet32	ResNet110	70.47	73.12	71.62	74.08	**72.66**	**74.18**	**73.14**	**74.86**
ResNet110	ResNet110	73.12	73.12	74.59	74.55	**75.08**	**75.10**	**75.52**	**75.71**
WRN-10-4	WRN-10-4	72.65	72.65	73.06	73.01	**73.77**	**73.75**	**73.97**	**74.08**
WRN-10-4	WRN-28-10	72.65	80.77	73.58	81.11	**74.61**	**81.43**	**75.11**	**82.13**

表 6 网络结构对行人重识别平均识别精度的影响比较 (%)

Table 6 Comparison of person re-identification mAP with different network structures (%)

网络结构		原始网络		DML[13]		ADML		TADML	
网络 1	网络 2	网络 1	网络 2	网络 1	网络 2	网络 1	网络 2	网络 1	网络 2
InceptionV1	MobileNetV1	65.26	46.07	65.34	52.87	**65.60**	**53.22**	**66.03**	**53.91**
MobileNetV1	MobileNetV1	46.07	46.07	52.95	51.26	**53.42**	**53.27**	**53.84**	**53.65**

验结果由复现的代码运行得到, 表中的其余结果均来自自文献[19], 一些对比方法未给出实验结果, 则标记为 "-". 如表 7 所示, 第 1 行 ResNet20 为学生网络的分类性能, 第 2 行 ResNet164 为教师网络的性能. 从第 2 行至最后一行为在相同的教师和学生网络设置下, 对比方法仅使用学生网络进行分类达到的分类性能. 第 1 列为对比方法, 第 2 列为模型大小. 最优值使用黑色粗体标记, 次优值使用下划线粗体标记. 本文方法 TAMDL 在 3 个数据集上均取得了最高的分类精度, 与最新的对比方法 KSANC 比较, 在 CIFAR10、CIFAR100 和 Tiny-ImagNet 上增幅分别为 0.37%、2.23% 和 0.34%.

表 7 本文算法与其他压缩算法的实验结果

Table 7 Experimental results of the proposed algorithm and other compression algorithms

对比算法	参数量 (MB)	CIFAR10 (%)	CIFAR100 (%)	Tiny-ImageNet (%)
ResNet20	0.27	91.42	66.63	54.45
ResNet164	2.6	93.43	72.24	61.55
Yim 等[10]	0.27	88.70	63.33	-
SNN-MIMIC[22]	0.27	90.93	67.21	-
KD[8]	0.27	91.12	66.66	57.65
FitNet[9]	0.27	91.41	64.96	55.59
Quantization[20]	0.27	91.13	-	-
Binary Connect[21]	15.20	91.73	-	-
ANC[23]	0.27	91.92	67.55	58.17
TSANC[24]	0.27	92.17	67.43	58.20
KSANC[24]	0.27	**92.68**	68.58	**59.77**
DML[13]	0.27	91.82	69.47	57.91
ADML	0.27	92.23	**69.60**	59.00
TADML	0.27	**93.05**	**70.81**	**60.11**

由表 7 可以看出, 学生网络都没能达到教师网络的性能. 对于 CIFAR10, 在相同规模下采用对抗学习后, 学生网络的性能得到改善, ANC、TSANC、KSANC、AMDL、TAMDL 的增幅分别为 0.5%、0.75%、1.26%、0.81% 和 2.63%. 对于类别复杂的 CIFAR100, 增幅更为明显, 以上 5 种方法的增幅分别为 0.92%、0.80%、1.95%、2.97% 和 4.81%. 对于更为复杂的 Tiny-ImageNet 数据集, 以上五种方法的增幅分别为 3.72%、3.75%、5.32%、4.55% 和 5.66%. 比较实验表明, 数据集越复杂, 对抗训练的提升效果越明显, 本文方法 TAMDL 相对于其他对比方法优势越明显.

2.6 模型复杂性分析

本节以 ResNet164/ResNet20 做为教师网络/学生网络为例, 来分析 TAMDL 模型的复杂性. 在训练阶段, 先固定判别器, 此时优化生成器—两个分类网络 ResNet164 和 ResNet20, 两个模型的参数量分别为 2.61 MB 和 0.27 MB, 即生成器参数量为 2.88 MB, 耗时与传统互学习网络一致; 优化判别器时, 生成器固定不动, 此时优化的是一个多层感知器—三个全连接层 128-256-128, 参数量为 0.59 MB. 在训练时生成器和判别器以 1:1 的轮次交替迭代, 在数据集 CIFAR100 使用 Pytorch0.4 进行实验, 生成器为 ResNet164 + ResNet20, 判别网络为三个维度为 128-256-128 的全连接层, 批尺寸 Batchsize 设为 64, 即每个训练轮次 Epoch 将训练集划分为 781个 Batch, 平均每训练轮次 Epoch 耗时 82 s, 其中每个 Batch平均耗时 0.1045 s, 优化生成器反向传播耗时 0.0694 s, 优化判别器反向传播耗时 0.0016 s. 采用对抗训练, 并没有带来太大的时间开销.

3 结束语

本文提出了一种拓扑一致性指导的对抗互学习知识蒸馏方法. 该方法在 GAN 框架下, 对轻量级的学生网络进行知识迁移, 所提方法设计了样本组间拓扑一致性度量, 依此设计的损失函数结合常规的实例级别的分布相似性, 以及对抗损失及标号损失, 作为训练模型的总损失. 文中评估了不同损失函数和不同模型架构对分类精度的影响. 在 3 个公开的数据集上验证了本文方法 TAMDL 的有效性. 本文方法效果稳定且提升明显, 而且在压缩模型的性能比较中, 取得最好的结果.

References

1 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA: IEEE, 2016. 770−778

2 Zhang X Y, Zhou X Y, Lin M X, Sun J. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake, USA: IEEE, 2018. 6848−6856

3 Guo Y W, Yao A B, Zhao H, Chen Y R. Network sketching: Exploiting binary structure in deep CNNs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA: IEEE, 2017. 4040−4048

4 Tai C, Xiao T, Wang X G,E W N. Convolutional neural networks with low-rank regularization. In: Proceedings of the 4th International Conference on Learning Representations, San Juan, Puerto Rico, 2016

5 Chen W, Wilson J T, Tyree S, Weinberger K Q, Chen Y X. Compressing neural networks with the hashing trick. In: Proceedings of the 32nd International Conference on Machine Learning, Lille, France: 2015. 37: 2285−2294

6 Denton E L, Zaremba W, Bruna J, LeCun Y, Fergus R. Exploiting linear structure within convolutional networks for efficient

evaluation. In: Proceedings of the 27th Annual Conference on Neural Information Processing Systems, Montreal, Canada: 2014. 1269−1277

7 Li Z, Hoiem D. Learning without forgetting. In: Proceedings of the 14th European Conference on Computer Vision, Amsterdam, Netherlands: 2016. 614−629

8 Hinton G E, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint, 2015, arXiv: 1503.02531

9 Romero A, Ballas N, Kahou S E, Chassang A, Gatta C, Bengio Y. Fitnets: Hints for thin deep nets. In: Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA, 2015

10 Yim J, Joo D, Bae J H, Kim J. A Gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA: IEEE, 2017. 7130−7138

11 Peng B Y, Jin X, Li D S, Zhou S F, Wu Y C, Liu J H, et al. Correlation congruence for knowledge distillation. In: Proceedings of the IEEE International Conference on Computer Vision, Seoul, South Korea: IEEE, 2017. 5006−5015

12 Park W, Kim D, Lu Y, Cho M. Relational knowledge distillation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA: IEEE, 2019. 3967−3976

13 Zhang Y, Xiang T, Hospedales T M, Lu H C. Deep mutual learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA: IEEE, 2018. 4320−4328

14 Batra T, Parikh D. Cooperative Learning with Visual Attributes. arXiv preprint, 2017, arXiv: 1705.05512

15 Zhang H, Goodfellow I J, Metaxas D N, Odena A. Self-attention generative adversarial networks. In: Proceedings of the 36th International Conference on Machine Learning, Long Beach, USA: 2019. 7354−7363

16 Zagoruyko S, Komodakis N. Wide residual networks. In: Proceedings of the British Machine Vision Conference, York, UK: 2016. 1−12

17 Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, 2009, **1**(4): 1−54

18 Mirza M, Osindero S. Conditional generative adversarial nets. arXiv preprint, 2014, arXiv: 1411.1784

19 Shu C Y, Li P, Xie Y, Qu Y Y, Kong H. Knowledge Squeezed Adversarial Network Compression. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence, New York, NY, USA: 2020. 11370−11377

20 Zhu C Z, Han S, Mao H Z, Dally W J. Trained ternary quantization. In: Proceedings of the 5th International Conference on Learning Representations. Toulon, France, 2017.

21 Courbariaux M, Bengio Y, David J P. Binaryconnect: Training deep neural networks with binary weights during propagations. In: Proceedings of the 27th Annual Conference on Neural Information Processing Systems. Montreal, Canada: 2015. 3123−3131

22 Ba J, Caruana R. Do deep nets really need to be deep? In: Proceedings of the 27th Annual Conference on Neural Information Processing Systems, Montreal, Canada: 2014. 2654−2662

23 Belagiannis V, Farshad A, Galasso F. Adversarial network compression. In: Proceedings of the European Conference on Computer Vision, Munich, Germany: 2018. 11132: 431−449

24 Xu Z, Hsu Y C, H J W. Training student networks for acceleration with conditional adversarial networks. In: Proceedings of British Machine Vision Conference, Newcastle, UK: 2018. 61

赖 轩 厦门大学信息学院硕士研究生. 主要研究方向为计算机视觉与图像处理.
E-mail: laixuan@stu.xmu.edu.cn
(**LAI Xuan** Master student at the School of Informatics, Xiamen University. His research interest covers computer vision and image processing.)

曲延云 厦门大学信息学院教授.主要研究方向为模式识别, 计算机视觉和机器学习. 本文通信作者.
E-mail: yyqu@xmu.edu.cn
(**QU Yan-Yun** Professor at the School of Informatics, Xiamen University. Her research interest covers pattern recognition, computer vision and machine learning. Corresponding author of this paper.)

谢 源 华东师范大学计算机科学与技术学院教授. 主要研究方向为模式识别, 计算机视觉和机器学习.
E-mail: yxie@cs.ecnu.edu.cn
(**XIE Yuan** Professor in the Department of Computer Science & Technology, East China Normal University. His research interest covers pattern recognition, computer vision and machine learning.)

裴玉龙 厦门大学信息学院硕士研究生. 主要研究方向为计算机视觉与图像处理. E-mail: 23020181154279@stu.xmu.edu.cn
(**PEI Yu-Long** Master student at the School of Informatics, Xiamen University. His research interest covers computer vision and image processing.)