计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications [doi: 10.15888/j.cnki.csa.009615]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于扩散模型的解耦知识蒸馏①

王鹏宇, 朱子奇

(武汉科技大学 计算机科学与技术学院, 武汉 430065)
通信作者: 王鹏宇, E-mail: 741003446@qq.com

摘　要: 知识蒸馏 (KD) 是一种将复杂模型 (教师模型) 的知识传递给简单模型 (学生模型) 的技术, 目前比较受欢迎的蒸馏方法大多停留在基于中间特征层, 继解耦知识蒸馏 (DKD) 提出后基于响应的知识蒸馏又重新回到 SOTA 行列, 这种使用强一致性约束条件的策略, 将经典的知识蒸馏拆分为两个部分, 解决了高度耦合的问题. 然而, 这种方法忽略了师生网络架构差距较大所引起的表征差距过大, 进而导致学生模型由于体量较小无法更有效的学习到教师模型的知识的问题. 为了解决这个问题本文提出了使用扩散模型来缩小师生模型之间的表征差距, 这种方法将教师特征传输到扩散模型中训练, 然后通过一个轻量级的扩散模型对学生模型进行降噪从而缩小了师生模型的表征差距. 大量的实验表明这种方法对比于基准方法在 CIFAR-100、ImageNet 数据集上均有较大的提升, 在师生网络架构差距较大时依然能够保持较好的性能.

关键词: 知识蒸馏; 解耦知识蒸馏; 扩散模型; 表征差距; 师生网络

引用格式: 王鹏宇,朱子奇.基于扩散模型的解耦知识蒸馏.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9615.html

Decoupled Knowledge Distillation Based on Diffusion Model

WANG Peng-Yu, ZHU Zi-Qi

(School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China)

Abstract: Knowledge distillation (KD) is a technique that transfers knowledge from a complex model (teacher model) to a simpler model (student model). While many popular distillation methods currently focus on intermediate feature layers, response-based knowledge distillation (RKD) has regained its position among the SOTA models after decoupled knowledge distillation (DKD) was introduced. RKD leverages strong consistency constraints to split classic knowledge distillation into two parts, addressing the issue of high coupling. However, this approach overlooks the significant representation gap caused by the disparity in teacher-student network architectures, leading to the problem where smaller student models cannot effectively learn knowledge from teacher models. To solve this problem, this study proposes a diffusion model to narrow the representation gap between teacher and student models. This model transfers teacher features to train a lightweight diffusion model, which is then used to denoise the student model, thus reducing the representation gap between teacher and student models. Extensive experiments demonstrate that the proposed model achieves significant improvements over baseline models on CIFAR-100 and ImageNet datasets, maintaining good performance even when there is a large gap in teacher-student network architectures.

Key words: knowledge distillation (KD); decoupled knowledge distillation; diffusion model; representation gap; teacher-student network

① 基金项目: 公安部科技计划 (2022JSM08)
收稿时间: 2024-03-13; 修改时间: 2024-04-10; 采用时间: 2024-04-23; csa 在线出版时间: 2024-07-24

近些年深度神经网络改变了人类的生活, 无论是在工业、科研、日常生活中都有着很高的应用需求, 然而这种深度神经网络模型对设备的要求苛刻, 通常需要大量的计算量和内存, 解决这一问题目前比较受欢迎的方法是知识蒸馏 (knowledge distillation, KD)[1].

知识蒸馏作为一种模型压缩的技术, 通过将复杂模型 (教师模型) 的知识传递给简单模型 (学生模型) 来实现, 目的是在花费最小化代价的情况下尽量使其效果向复杂模型不断靠近, 以满足大多数设备的性能. 自 vanilla-KD 提出之后基于响应的 (response-based)[2]知识蒸馏走近人们的视野, 随后基于特征的 (feature-based)[3–7]、基于关系的 (relation-based)[8–10]的知识蒸馏也被相继提出, 目前受到大家关注的方法大多数都是基于中间特征层的, 这些方法忽略掉了 logit 层的语义信息更高的特点, 解耦知识蒸馏 (decoupled knowledge distillation, DKD)[11]的提出使得基于响应的知识蒸馏重新回到了 SOTA 行列.

解耦知识蒸馏 (DKD) 是一种基于响应的方法, 传统的知识蒸馏 logit 部分存在高度耦合, 这是导致 logit 蒸馏局限性的根本原因, 该方法通过将 logit 中的信息拆分为目标类 (target class) 和非目标类 (non-target class) 两部分, 来达到解耦的效果, 将经典的 KD 改为 TCKD 和 NCKD, 从实验结果来看, TCKD 只是起到了转移训练样本的“困难”知识, NCKD 中转移了大量的暗知识, 这一部分才是 logit 蒸馏的主角, 将学生模型输出的 logit 不断地向教师模型的 logit 靠近, 然而, 这种方法忽略了一个问题, 当师生网络模型差距较大的时候, 即使通过上述方法, 将经典 KD 进行解耦, 效果也无法达到理想状态[12], 因为教师网络模型和学生网络模型之间的表征差距过大, 学生模型的特征无法对齐教师模型的特征.

为了解决上述的问题, 本文从一个新的角度出发, 认为解决这种体量差距较大的知识蒸馏的关键点是在于如何摒弃掉无用的噪声, 学生模型可以视作一个低配版的教师模型, 由于其体量较小, 学习能力相对较弱, 在训练过程中不能准确地分辨出哪些是需要学习的知识, 这种对噪声的提炼会导致训练效果退化. 本文采用扩散模型[13,14]来执行去噪模块, 先消除掉学生模型内部的噪声信息, 再进行蒸馏. 在这种方法下的学生模型更加贴合教师模型.

本文主要贡献如下.

1) 本文提出了一个噪声自适应模块, 能够更精确的指定学生特征的噪声水平.

2) 使用了一个由 ResNet[15]中两个瓶颈块组成的轻量级扩散模型, 从实际情况出发搭建更加符合知识蒸馏的扩散模型.

3) 解决了以前知识蒸馏中师生特征无法对齐的问题, 在利用解耦知识蒸馏在 logits 上可以获取低纬度语义信息的优势, 来优化模型效果.

1 相关工作

1.1 知识蒸馏

知识蒸馏这一理论最早出现是在 Hinton 等人[1]所发表的 vanilla-KD 论文中, 采用的是一种师生模型网络, 将预训练的教师模型的知识通过网络传递给学生模型, 输出一般为模型的预测和中间特征. 在这个过程中学生模型的输出为$F^{(s)}$, 教师模型的输出为$F^{(t)}$, 最后得到的知识蒸馏损失为:

$$L_{\text{kd}} := d(F^{(t)},F^{(s)}) \tag{1}$$

其中, d为距离函数, 通过该函数测量师生模型输出间的差距, 例如在概率分布任务中使用的 Kullback-Leibler (KL) 散度, 一般分类任务中的交叉熵 (cross entropy), 中间特征和回归输出使用的均方误差 (MSE).

1.2 扩散模型

扩散模型是一种对图像增强和图像恢复的非线性的处理方法, 正向扩散过程遵循马尔可夫链的概念, 通过控制时间步 t 不断向图像中添加噪声, 然后利用预测和去噪逆转这一过程. 用y表示样本, 下角标为时间步 $t \in \{0,1,2,\cdots,T\}$, 扩散模型的过程可表示为:

$$q(y_t \mid y_0) = \mathcal{N}\left(y_t; \sqrt{\bar{\alpha}_t}y_0, (1-\bar{\alpha}_t)I\right) \tag{2}$$

其中, 定义$\alpha_t = 1-\beta_t, \bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$, β_t是一个超参数序列, 为每一时间步中应添加的噪声量, 通常会随着时间的增长而增长. 当z是从一个均值为μ方差为σ^2的高斯分布中采样得到的[16], 可以引入一个服从标准高斯分布的随机变量$\mathcal{E}$:

$$z \sim \mathcal{N}\left(z,\mu,\sigma^2 I\right) \to z = \mu + \sigma\cdot\mathcal{E}, \mathcal{E} \sim \mathcal{N}(0,I) \tag{3}$$

因此可以将y_t表示为y_0与噪声$\mathcal{E}$的线性组合:

$$y_t = \sqrt{\bar{\alpha}_t y_0} + \sqrt{1-\bar{\alpha}_t}\mathcal{E} \tag{4}$$

在逆向扩散过程中需要训练一个神经网络$\Phi_\theta(y_t,t)$, 通过缩小$\mathcal{E}$与$\Phi_\theta(y_t,t)$的L_2损失来预测$y_t \cdots y_0$中的噪声,

即:

$$\mathcal{L}_{\text{diff}} := \|\mathcal{E}_t - \Phi_\theta(y_t, t)\|_2^2 \tag{5}$$

在推理过程中, 初始噪声为Y_t, 通过训练好的$\Phi_\theta(y_t,t)$, 来完成对数据样本Y_0的迭代去噪过程:

$$p_\theta(y_{t-1} \mid y_t) := \mathcal{N}\left(y_{t-1}; \Phi_\theta(y_t,t), \sigma_t^2 I\right) \tag{6}$$

其中, σ_t^2为 DDIM 中的过渡方差.

本文受到上述方法的启发, 利用扩散模型对学生模型进行去噪, 如图 1 DDKD 架构图所示, 学生模型的特征在初始阶段可视作带有噪声的教师特征, 然后利用教师特征训练扩散模型, 最后用训练好的扩散模型对学生特征进行去噪, 在得到无噪声的学生特征后进行蒸馏.

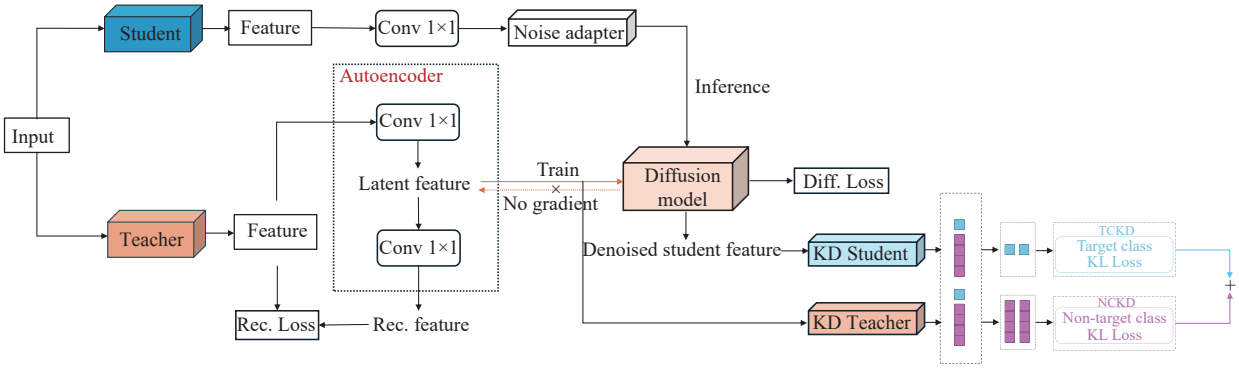

图 1 DDKD 架构图

2 本文方法

在本节中将介绍本文所提出的基于扩散模型的解耦知识蒸馏, 首先 KD 中的特征对齐任务会被转换为扩散模型的去噪过程, 通过对齐师生模型的特征来获得更有效的蒸馏.

为了更好地提高计算效率, 本文引入了特征自编码器[17]来降低特征映射的维数, 从而简化了扩散过程. 此外, 本文还提出了自适应噪声匹配模块, 以提高学生特征的去噪性能. 在最后将特征的高语义信息进行蒸馏.

2.1 基于扩散模型的解耦知识蒸馏

一般来说不同的模型有着不同的架构, 在特征的提取上也会有不同的关注点, 即使是在相同的数据集上训练也是如此. 教师模型体量较大, 在特征提取上比学生模型的效果好, 当教师模型与学生模型之间体量相差越大的时候他们的差距就会越加明显. 所以解决知识蒸馏问题的关键在于如何缩小这种师生差距. 在以前的论文中[18]也调查过教师模型和学生模型之间的差异. 教师的预测概率分布比学生的预测概率分布更加清晰, 而且教师模型的预测错误答案的方差也比学生模型要小[19], 以上这些说明, 教师模型的输出相对于学生模型来说要更加显著. 学生模型的预测比教师模型的预测包含了更多的噪声.

由于教师模型和学生模型之间的差距, 这些噪声无法仅凭通过简单的模仿蒸馏中的教师模型来消除. 但可以将教师模型和学生模型关注点更多地放在有价值的信息上面, 本文受到扩散模型的启发[20], 将学生模型视为教师模型的带噪版本, 先用教师特征去训练一个扩散模型, 然后再用它对学生特征进行去噪.

对于一个样本在蒸馏过程中教师特征表示为$F^{(\text{Tea})}$学生特征表示为$F^{(\text{Stu})}$, 在前向噪声过程$q\left(F_t^{(\text{Tea})} \mid F^{(\text{Tea})}\right)$式 (2) 中使用$F^{(\text{Tea})}$来训练$\mathcal{L}_{\text{diff}}$式 (5) 的扩散模型.

然后将学生特征输入到学习扩散模型的迭代去噪过程中, 即式 (6) 中的$p_\theta\left(F_{t-1}^{(\text{Stu})} \mid F_t^{(\text{Stu})}\right)$, 其中$F^{(\text{Stu})}$为该过程的初始噪声特征. 经过这个去噪过程, 我们得到一个去噪的学生特征$F^{(\text{Stu})}$, 用它来计算式 (1) 中与原始教师特征$F^{(\text{Tea})}$的 KD 损失.

根据经验来说 logits 在语义水平上要比深层特征更高级, 所以本文中提到的方法是在整个模型架构的最后一部分使用解耦知识蒸馏, 将去噪后的学生输出$F^{(\text{Stu})}$再进行 logits 部分的解耦. DDKD 的架构如图 1 所示.

2.2 具有线性自编码器的扩散模型

由于教师特征尺寸相对较大, 模型在去噪过程会耗费大量的计算资源, 需要转发 T 次 (在本文方法中使用 $T=5$) 噪声预测网络Φ_θ来去噪学生特征和训练 1 次带有教师特征的噪声预测网络. 当教师特征的维度较大时, 这 T+1 次转发会导致计算成本很高. 考虑到这个

问题, 本文从 ResNet 中提取了一种由两个瓶颈块叠加到一起的轻量级扩散模型. 然后, 遵循着扩散模型的方法, 并提出使用线性自编码器模块压缩通道数量. 将压缩后的特征作为扩散模型的输入. 线性自编码器由两个卷积神经网络组成, 一个是编码器用于减少通道数量, 另一个是解码器用来重构教师特征. 编码器的输出特性被用来训练扩散模型和学生模型 (式 (2)).

自编码器只使用重构损失进行训练, 重构损失是原始教师 $F^{(\mathrm{Tea})}$ 与重构教师特征 $\tilde{F}_{\mathrm{ae}}^{(\mathrm{Tea})}$ 之间的均方误差, 即:

$$\mathcal{L}_{\mathrm{ae}} := \left\| \tilde{F}^{(\mathrm{Tea})} - F^{(\mathrm{Tea})} \right\|_2^2 \tag{7}$$

用于训练扩散模型的潜在教师特征 $Z^{(\mathrm{Tea})}$ 与扩散模型是分离的, 并且扩散模型没有向后梯度.

本文还使用卷积层将学生特征投影到与教师潜在特征 $Z^{(\mathrm{Tea})}$ 相同的维度, 表示为 $Z^{(\mathrm{Stu})}$. 然后将其传递给扩散模型执行反向去噪过程如式 (6), 并生成去噪的学生特征 $Z^{(\mathrm{Stu})}$. 然后用它们来计算 KD 损失并监督学生, 即:

$$\mathcal{L}_{\mathrm{DDKD}} := d\left(\hat{Z}^{(\mathrm{Stu})}, Z^{(\mathrm{Tea})}\right) \tag{8}$$

本文使用简单的 MSE 损失和 KL 散度损失作为距离函数 d 来计算去噪的学生特征和教师特征的差异, 然后利用基准的方法, 在 logits 部分将经典的 KD 损失分解为两个部分, 一个是目标类和非目标类的二分类预测, 另一个是非目标类的多分类预测.

2.3 自适应噪声匹配

如前文所述, 本文将学生特征视为教师特征的嘈杂版本. 然而, 表示教师和学生特征之间差距的噪声水平是未知的, 并且可能因不同的训练样本而变化. 因此, 不能直接确定应该从哪个初始时间步长开始扩散过程. 为了解决这个问题, 本文引入了一个自适应噪声匹配模块, 将学生特征的噪声水平与预定义的噪声水平相匹配.

如图 2 所示, 将构建一个简单的卷积模块来学习一个融合了学生输出和高斯噪声的权值 γ, 这有助于学生输出能更快匹配与初始时间步长 t 的噪声特征的相同噪声水平. 因此, 去噪过程中的初始噪声特征变为:

$$Z_T^{(\mathrm{Stu})} = \gamma Z^{(\mathrm{Stu})} + (1-\gamma) \in T \tag{9}$$

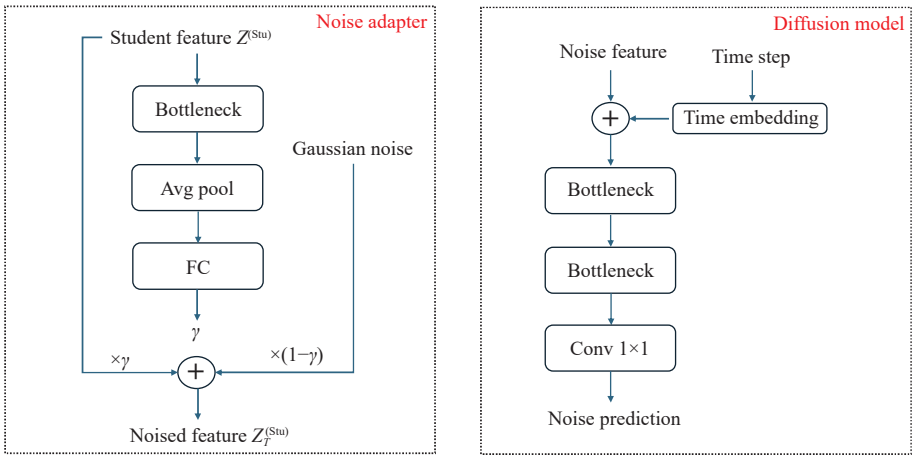

图 2 自适应噪声模块

这种噪声适应可以自然地通过 KD 损失 L_{kd} 进行优化, 因为在去噪过程中, 当学生特征与适当的噪声水平匹配时, 这时去噪后的学生特征与教师特征的差异才会最小.

2.4 总体损失函数

DDKD 的整体损失函数由原始任务损失、优化扩散模型的扩散损失、学习自编码器的重建损失以及对教师特征和去噪学生特征进行蒸馏的 KD 损失组成, 即:

$$\mathcal{L}_{\mathrm{train}} = \mathcal{L}_{\mathrm{task}} + \lambda_1 \mathcal{L}_{\mathrm{diff}} + \lambda_2 \mathcal{L}_{\mathrm{ae}} + \lambda_3 \mathcal{L}_{\mathrm{DDKD}} \tag{10}$$

其中, $\lambda_1, \lambda_2, \lambda_3$ 是用来平衡损失的损失权值. 在所有实验中, 简单地设 $\lambda_1 = \lambda_2 = 1$.

3 实验数据分析

为了验证本文方法的有效性, 我们在 ImageNet[21] 和 CIFAR-100[22] 上进行实验, 首先会介绍本文的实验细节和相关数据集, 在实验部分会介绍我们的消融实验, 来展示本文方法的有效性.

3.1 数据集介绍

ImageNet 数据集, 是目前最大的图像识别数据库, 在分类、定位和检测任务中使用较多, 其中包含了

14 197 122 张图像, 共有 21 841 个类别, 图像覆盖了生活中大多数能见场景, 包含了更多与图像无关的噪声和变化, 可以更好地测试模型的鲁棒性.

在本文中我们使用的是 ImageNet 的子数据集, 具有相同的效果, 训练集有 128 167 张图像, 包含 1 000 个类别, 每个类别包含约 1 300 张图片, 验证集有 50 000 张图像, 每个类别包含 50 张图像, 测试集有 100 000 张图像, 每类 100 张图像.

CIFAR-100 是一个常用的图像数据集, 用于图像分类任务和计算机视觉研究. 图像内容丰富多样, 涵盖了各种日常物体和动物, 可以更好地检验算法的泛化能力, 该数据集总共包含 60 000 张彩色图像, 其中训练集包含 50 000 张图像, 测试集包含 10 000 张图像. 每张图像的尺寸为 32×32 像素, 分为 100 个类别. 每个类别包含 500 张训练图像和 100 张测试图像.

3.2 实验细节

为了验证本文方法的可行性, 将在 ResNet[15]、ShuffleNet[23]、MobileNet[24]、Wide ResNet (WRN)[25]和 VGG[26]网络中进行对比.

对于 ImageNet 数据集, 本文的训练策略将 Epoch 设为 100, Batchsize 设为 256, 学习率 (LR) 初始值设为 0.1, 每 30 个 Epoch 再衰减为当前的 0.1, 优化器 (optimizer) 使用的是随机梯度下降算法 (SGD), 超参数权重衰退 (weight decay) 为 0.000 1.

对于 CIFAR-100 数据集, 本文的训练策略将 Epoch 设为 240, Batchsize 设为 64, 学习率 (LR) 初始值设为 0.05, 优化器选择随机梯度下降算法, 权重衰退为 0.000 5, 学习率会在从第 150 个 Epoch 之后, 每 30 个 Epoch 衰减为当前大小的 0.1.

3.3 消融实验

为了验证不同方法对实验效果的影响, 本文进行了消融实验, 在基线设置上, 使用了 ResNet-18 和 MobileNet V1 作为学生模型, 教师模型分别为 ResNet-34 和 ResNet-50 网络. 表 1 为在 ImageNet 数据集上的实验结果. 可以发现 vanilla-KD 的效果无论是在同构网络还是异构网络中, 在 TOP-1 和 TOP-5 上的准确率都是最低的, 当使用 DKD 方法时, 准确率会有明显提升, 在 MobileNet V1 和 ResNet50 这种设置下更为明显, 这种方法解决了传统知识蒸馏中的目标类与非目标类的高耦合问题, 比传统知识蒸馏 TOP-1 的准确率提高了 1.37%. 本文的方法相对于目前最先进的方法 (DKD) 在 ResNet-18 和 ResNet-34 这种网络设置下 TOP-1 的准确率提升了 0.64%, 在 MobileNet V1 和 ResNet-50 设置下 TOP-1 的准确率提升了 1.82%, 这是因为在蒸馏前利用扩散模型对学生模型进行了去噪过程, 进行了特征对齐. 这个过程减少了学生与教师模型的差距.

表 1 ImageNet 上的评估结果 (%)

Student (teacher)		Teacher	Student	KD[1]	DKD[11]	DDKD
ResNet-18	TOP-1	73.31	69.76	70.66	71.70	**72.34**
(ResNet-34)	TOP-5	91.42	89.08	89.88	90.41	**90.82**
MobileNet V1	TOP-1	76.16	70.13	70.68	72.05	**73.87**
(ResNet-50)	TOP-5	92.86	89.49	90.30	91.05	**91.75**

3.4 对比实验

为了验证本文方法的有效性, 本文还与其他方法进行对比, 如表 2、表 3 所示, 分别为同构网络和异构网络准确率对比的结果.

表 2 在 CIFAR-100 数据集上同构网络中师生实验结果 (%)

网络设置		Teacher	Student	FitNet[3]	VID[27]	RKD[10]	PKT[28]	CRD[29]	KD[1]	DIST[12]	DKD[11]	DDKD
教师模块	学生模块											
WRN-40-2	WRN-16-2	75.61	73.26	73.58	73.49	73.35	74.65	75.48	74.92	75.24	76.24	**76.38**
WRN-40-2	WRN-40-1	75.61	71.98	72.24	73.30	72.22	73.45	74.14	73.54	74.73	**74.81**	74.75
ResNet-56	ResNet-20	72.34	69.06	69.21	70.38	69.61	70.34	71.16	70.66	71.75	71.97	**72.24**
ResNet-32×4	ResNet-8×4	79.42	72.50	73.50	73.09	71.90	73.64	75.51	73.33	76.31	76.32	**76.93**
VGG13	VGG8	74.64	70.36	71.02	72.15	71.48	71.62	73.94	72.98	73.57	74.68	**74.73**

表 3 在 CIFAR-100 数据集上异构网络中师生实验结果 (%)

网络设置		Teacher	Student	FitNet[3]	VID[27]	RKD[10]	PKT[28]	CRD[29]	KD[1]	DIST[12]	DKD[11]	DDKD
教师模块	学生模块											
WRN-40-2	ShuffleNetV1	75.61	70.50	73.73	74.28	72.21	75.03	76.05	74.83	75.11	76.70	**76.78**
ResNet-50	MobileNetV2	79.34	64.60	63.16	67.57	64.43	66.52	69.11	67.35	68.66	68.82	**69.18**
ResNet-32×4	ShuffleNetV1	79.42	70.50	73.59	73.38	72.28	74.10	75.11	74.07	76.34	75.93	**76.62**
ResNet-32×4	ShuffleNetV2	79.42	71.82	73.54	73.40	73.21	74.69	75.65	74.45	77.35	77.29	**77.71**
VGG13	MobileNetV2	74.64	64.6	64.14	63.98	64.52	67.35	69.73	67.37	70.08	69.71	**70.79**

从结果上来看, 本文方法在异构网络中有更为显著的效果, 考虑到其他方法是通过添加中间网络, 将知识过渡给学生模型, 本文提出了一个更为可靠的方法从根本上解决师生的差距问题, 在蒸馏开始前利用轻量级的扩散模型解决了在训练前师生特征对齐的问题, 然后再利用解耦知识蒸馏的方法来解决传统知识蒸馏中目标类与非目标类高度耦合的问题.

在 CIFAR-100 数据集上, 同系列网络架构设置的情况下最好的结果比基准方法提升了 0.61%, 准确率提高了 2.6 个百分点. 在不同系列的网络架构设置下, 最好的结果比基准方法提升了 0.71%, 准确率提高了 3.4 个百分点.

4 结束语

本文研究了教师和学生在知识蒸馏方面的差异. 从使用到的基准方法来看, 解耦知识蒸馏在 logits 部分将其分为目标类和非目标类确实有很大的提升, 但本文从开始就将学生与教师特征对齐, 从本质上缩小了师生之间的差距, 为了减少差异, 提高蒸馏性能, 本文从一个新的角度出发, 提出用扩散模型显式地消除学生特征中的噪声. 在此基础上, 进一步引入了一个带有线性自编码器的轻量级扩散模型来降低该方法的计算成本, 并引入了一个自适应噪声匹配模块来将学生特征与正确的噪声水平相匹配, 从而提高了去噪性能. 在图像分类任务上的大量实验验证了本文的有效性和泛化性.

参考文献

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

2 Gou JP, Yu BS, Maybank SJ, *et al*. Knowledge distillation: A survey. International Journal of Computer Vision, 2021, 129(6): 1789–1819. [doi: 10.1007/s11263-021-01453-z]

3 Romero A, Ballas N, Kahou SE, *et al*. FitNets: Hints for thin deep nets. arXiv:1412.6550, 2014.

4 Guo ZY, Yan HN, Li H, *et al*. Class attention transfer based knowledge distillation. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 11868–11877.

5 Chen ZH, Shamsabadi EA, Jiang S, *et al*. Robust feature knowledge distillation for enhanced performance of lightweight crack segmentation models. arXiv:2404.06258, 2024.

6 Tian YL, Krishnan D, Isola P. Contrastive representation distillation. Proceedings of the 8th International Conference on Learning Representations. Addis Ababa: OpenReview.net, 2020.

7 Chen PG, Liu S, Zhao HS, *et al*. Distilling knowledge via knowledge review. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 5008–5017.

8 Yim J, Joo D, Bae J, *et al*. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 4133–4141.

9 Lee SH, Kim DH, Song BC. Self-supervised knowledge distillation using singular value decomposition. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 335–350.

10 Park W, Kim D, Lu Y, *et al*. Relational knowledge distillation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3967–3976.

11 Zhao BR, Cui Q, Song RJ, *et al*. Decoupled knowledge distillation. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 11953–11962.

12 Huang T, You S, Wang F, *et al*. Knowledge distillation from a stronger teacher. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, 2022. 33716–33727.

13 Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 574.

14 Ponglertnapakorn P, Tritrong N, Suwajanakorn S. DiFaReli: Diffusion face relighting. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 22646–22657.

15 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.

16 Song JM, Meng CL, Ermon S. Denoising diffusion implicit models. arXiv:2010.02502, 2020.

17 Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science, 2006, 313(5786): 504–507. [doi: 10.1126/science.1127647]

18 Kundu S, Sun QR, Fu Y, *et al*. Analyzing the confidentiality of undistillable teachers in knowledge distillation. Proceedings of the 35th Conference on Neural Information Processing Systems. 2021. 9181–9192.

19 Li XC, Fan WS, Song SM, *et al*. Asymmetric temperature scaling makes larger networks teach well again. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 277.

20 Rombach R, Blattmann A, Lorenz D, *et al*. High-resolution image synthesis with latent diffusion models. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 10684–10695.

21 Russakovsky O, Deng J, Su H, *et al*. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115(3): 211–252. [doi: 10.1007/s11263-015-0816-y]

22 Krizhevsky A. Learning multiple layers of features from tiny images. Techinical Report, University of Toronto. 2009. http://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf.

23 Zhang XY, Zhou XY, Lin MX, *et al*. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6848–6856.

24 Sandler M, Howard A, Zhu ML, *et al*. MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4510–4520.

25 Zagoruyko S, Komodakis N. Wide residual networks. arXiv:1605.07146, 2016.

26 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014.

27 Ahn S, Hu SX, Damianou A, *et al*. Variational information distillation for knowledge transfer. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 9163–9171.

28 Passalis N, Tzelepi M, Tefas A. Probabilistic knowledge transfer for lightweight deep representation learning. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(5): 2030–2039. [doi: 10.1109/TNNLS.2020.2995884]

29 Tian YL, Krishnan D, Isola P. Contrastive representation distillation. arXiv:1910.10699, 2019.

(校对责编: 张重毅)