计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(2):11−18 [doi: 10.15888/j.cnki.csa.009773] [CSTR: 32024.14.csa.009773]

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

基于感知重构的解耦知识蒸馏①

祝英策, 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: 祝英策, E-mail: 508607035@qq.com

摘　要: 在知识蒸馏 (knowledge distillation, KD) 领域中, 基于特征的方法可以有效挖掘教师模型所蕴含的丰富知识. 然而, 基于 Logit 的方法常面临着知识传递不充分和效率低下等问题. 解耦知识蒸馏 (decoupled knowledge distillation, DKD) 通过将教师模型和学生模型输出的 Logit 划分为目标类和非目标类进行蒸馏. 这种方式虽然提升了蒸馏精度, 但其基于单实例的蒸馏方式使得批次内样本间的动态关系无法被捕捉到, 尤其是当教师模型和学生模型的输出分布存在显著差异时, 仅依靠解耦蒸馏无法有效弥合这种差异. 为了解决 DKD 中存在的问题, 本文提出感知重构的方法. 该方法引入一个感知矩阵, 利用模型的表征能力对 Logit 进行重新校准, 细致分析类内动态关系, 重建更细粒度的类间关系. 由于学生模型的目标是最小化表征差异, 因此将该方法扩展到解耦知识蒸馏中, 把教师模型和学生模型的输出映射到感知矩阵上, 从而使学生模型能够学习到教师模型中更加丰富的知识. 本文方法在 CIFAR-100 和 ImageNet-1K 数据集上进行了一系列的验证, 实验结果表明, 该方法训练的学生模型在 CIFAR-100 数据集上的分类准确率达到了 74.98%, 相较于基准方法提升了 0.87 个百分点, 提升了学生模型的图像分类效果. 此外, 通过对多种方法进行对比实验, 进一步验证了该方法的优越性.

关键词: 模型压缩; 知识蒸馏; 解耦知识蒸馏; 感知重构; 类内关系匹配

引用格式: 祝英策,朱子奇.基于感知重构的解耦知识蒸馏.计算机系统应用,2025,34(2):11–18. http://www.c-s-a.org.cn/1003-3254/9773.html

Decoupled Knowledge Distillation Based on Perception Reconstruction

ZHU Ying-Ce, ZHU Zi-Qi

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

Abstract: In the field of knowledge distillation (KD), feature-based methods can effectively extract the rich knowledge embedded in the teacher model. However, Logit-based methods often face issues such as insufficient knowledge transfer and low efficiency. Decoupled knowledge distillation (DKD) conducts distillation by dividing the Logits output by the teacher and student models into target and non-target classes. While this method improves distillation accuracy, its single-instance-based distillation approach fails to capture the dynamic relationships among samples within a batch. Especially when there are significant differences in the output distributions of the teacher and student models, relying solely on decoupled distillation cannot effectively bridge these differences. To address the issues inherent in DKD, this study proposes a perception reconstruction method. This method introduces a perception matrix. By utilizing the representational capabilities of the model, it recalibrates Logits, meticulously analyzes intra-class dynamic relationships, and reconstructs finer-grained inter-class relationships. Since the objective of the student model is to minimize representational disparity, this method is extended to decoupled knowledge distillation. The outputs of the teacher and student models are mapped onto the perception matrix, enabling the student model to learn richer knowledge from the teacher model. A series of validations on the CIFAR-100 and ImageNet-1K datasets demonstrate that the student model trained

① 基金项目: 公安部科技计划 (2022JSM08)
收稿时间: 2024-07-29; 修改时间: 2024-08-20; 采用时间: 2024-09-10; csa 在线出版时间: 2024-12-19
CNKI 网络首发时间: 2024-12-20

with this method achieves a classification accuracy of 74.98% on the CIFAR-100 dataset, which is 0.87 percentage points higher than that of baseline methods, thereby enhancing the image classification performance of the student model. Additionally, comparative experiments with various methods further verify the superiority of this method.

Key words: model compression; knowledge distillation (KD); decoupled knowledge distillation (DKD); perception reconstruction; intra-class relationship matching

近年来, 深度学习 (deep learning, DL) 技术广泛应用于计算机视觉[1]、自然语言处理[2]和强化学习[3,4]等领域. 然而, 深度神经网络往往面临着庞大的计算成本和存储压力, 解决这一问题比较常用的方法就是知识蒸馏.

知识蒸馏作为一种有效的模型压缩技术, 通过将复杂的教师模型所学到的知识传递给较小的学生模型, 从而在不显著牺牲性能的前提下大幅减少计算资源和存储需求. 其核心目标是使轻量化的学生模型在实际部署时能够接近教师模型的性能, 从而在多种硬件环境中实现高效的模型应用. 该理论最早由 Hinton 等人[5]在 2015 年提出, 其核心思想是通过最小化教师模型和学生模型输出 (即 Logit) 之间的 KL 散度来实现知识转移. 自 Romero 等人[6]之后, 大多数的研究注意力都集中在从中间层的深层特征中提取知识[7−9]. 和基于 Logit 的方法相比, 特征蒸馏虽然精度有所提高, 但其中引入的额外计算和存储成本也较为高昂. 因此, 如何通过 Logit 获取更多的知识逐渐成为众多研究者所关注的话题[10,11].

传统的 Logit 蒸馏在应对诸如语义分割[12,13]、目标检测[14,15]等复杂任务时效果极为受限. 解耦知识蒸馏 (decoupled knowledge distillation, DKD)[16]将 Logit 划分为目标类和非目标类分别进行处理, 对目标类信息注重其在分类任务中的准确性, 而对非目标类信息注重其在区分不同类别中的作用. DKD 巧妙解决了传统 Logit 蒸馏中的高耦合性问题, 进一步挖掘了 Logit 所传递出的知识. 然而, 现有的 DKD 方法主要是针对单实例的蒸馏, 未能充分考虑批次内样本间的动态关系[17]. 此外, DKD 没有充分利用教师模型在类内分布上的差异, 尤其是当教师模型对某个类别具有高置信度时, 通常会导致对其他类别的预测产生较大的偏差或不确定性. DKD 未能有效地将这种类内差异和类别间的相似度信息传递给学生模型, 从而限制了知识传递的全面性和有效性[18].

为了解决上述问题, 本文提出一种基于“感知重构”的解耦知识蒸馏方案. 对于包含多个类别的图像批次, 该方法首先从模型输出的 Logit 中提取每个类别概率的均值和方差, 并利用这些统计信息对类内 Logit 进行标准化处理, 从而构建出该批次所对应的感知矩阵. 这一标准化过程调整了每个类别的 Logit 分布, 使其能够更好地反映固有的统计特性. 教师模型和学生模型的输出都将映射到各自的感知矩阵上以实现 Logit 的重构, 之后再进行解耦蒸馏, 从而确保两者之间的方差尺度一致. 该方法不仅增强了学生模型对实例间上下文关系的理解, 还有效捕捉了教师模型的类内分布差异, 从而进一步提升了解耦知识蒸馏的效果. 本文主要贡献如下.

(1) 本文提出了一种基于感知重构的解耦知识蒸馏方法, 通过扩展现有的 DKD 方法, 强调了实例级分布的重建, 提供了一种新的基于 Logit 蒸馏的方法.

(2) 本文方法基于整个批次 Logit 的类内均值和方差构建出一个感知矩阵, 确保教师模型和学生模型之间的方差一致性, 进一步捕获实例间的动态关系.

(3) 通过对比多种方法, 验证了本文方法在 CIFAR-100 和 ImageNet-1K 数据集上具有良好的表现.

1 相关工作

在知识蒸馏中, 不仅可以从 Logit 中转移知识, 中间层激活值、神经元[19]或中间层特征[20]同样可以作为指导学生模型学习的知识. 因此, 根据蒸馏过程中传递的不同知识, 蒸馏方案可分为基于 Logit 的蒸馏、基于特征的蒸馏和基于结构化关联的蒸馏.

1.1 基于 Logit 的蒸馏

Logit 蒸馏是一种基础的知识蒸馏方法, 通过传递大型教师模型的输出来训练小型学生模型[11]. 在该方案中, 教师模型的输出包含丰富的类别间区分信息, 学生模型通过学习这些输出来进行训练. 这个过程旨在最小化学生模型和教师模型输出之间的距离度量, 使

学生模型能够捕获并模仿教师模型的类别判别能力. 然而, 随着神经网络的层次不断变深, 模型中间层所传递出的信息越来越重要, 因此基于 Logit 的蒸馏方案受到了极大的限制. 解耦知识蒸馏[16]的提出重新引发了研究者对基于 Logit 的知识蒸馏潜力的关注. 解耦知识蒸馏将 Logit 分为目标类和非目标类分别进行蒸馏, 从而解决了传统 Logit 蒸馏中的高耦合性问题, 进一步挖掘了 Logit 中蕴含的知识.

1.2 基于特征的蒸馏

特征蒸馏通过模仿教师模型中间层的特征来训练学生模型, 是一种有效的知识蒸馏方案. 由于教师模型和学生模型的结构可能不同, 因此特征蒸馏又可以进一步分为同构特征蒸馏和异构特征蒸馏. 同构特征蒸馏中比较常见的方法有 Tung 等人提出的 SP[7]和 Passalis 等人提出的 PKT[21]等. 其中, SP 方案通过指导学生网络学习输入对之间在教师网络中的相似性关系, 从而确保在学生网络中语义相似的输入能够产生相似的激活模式, 可以更有效捕捉和传递教师网络的知识. PKT 方案通过建模数据点之间的成对交互信息, 捕捉特征空间的几何结构, 利用核密度估计 (KDE) 来估计教师模型和学生模型特征空间中数据点的联合概率密度, 通过最小化这两个概率分布之间的差异来实现分布的拟合. 在异构特征蒸馏场景中通常需要对中间层做相应的处理使其可用于同构蒸馏, 常见的有 VID[22]和 AB[19]等. Ahn 等人[22]提出的变分信息蒸馏 (VID) 框架通过最大化教师模型和学生模型之间特征的“互信息”来进行知识转移. 该方法引入一种变分下界来近似计算这种“互信息”, 极大程度上保留了教师模型中的关键信息. Heo 等人[19]认为隐藏神经元的激活状态在深度学习模型中形成的关键激活边界在分类决策中起着至关重要作用, 因此提出激活边界 (AB) 框架. 该方法使用一种新的激活转移损失函数对隐藏层神经元形成的激活边界进行蒸馏, 专门优化学生模型以准确再现教师模型的激活边界.

1.3 基于结构化关联的蒸馏

结构化蒸馏不仅传递单个样本的知识, 还包括样本间的类间关系以及类内关系. 它可以基于模型输出层的类别信息, 也可以基于中间层特征来建立知识迁移. Park 等人[23]提出的 RKD 方法通过计算数据样本之间的关系潜能 (如样本间的距离或形成的角度) 来捕捉教师模型中数据样本之间的相对关系. 这种方法特别适合于度量学习、分类和少样本学习等任务中. Liu 等人[24]提出的 IRG 方法则通过计算神经网络之间的实例关系图来提取结构化关联知识, 将每个样本都看作是图中的一个节点, 样本间的相似性或关系则通过图的边来表示. 通过这种方式, IRG 捕捉了教师模型中数据样本之间的关系结构, 并将这种结构性知识传递给学生模型. 此外还有一些方法利用特征内部的上下文关系作为结构化知识进行蒸馏. 例如, Hou 等人[25]提出的区域间亲和度蒸馏方法, 将道路场景图像分解为不同的区域, 并根据特征分布上的相似性建立节点之间的成对关系. Tao 等人[26]通过建立样本间的结构化图表示, 实现了一种新颖的样本增量学习方式.

本文受到了知识蒸馏中结构化知识的启发, 在 DKD 基础之上融入实例间的结构化关联知识, 通过构建实例上下文间的感知矩阵, 重新校准 Logit, 进一步捕获教师模型所传递出的实例间动态关系.

2 本文方法

2.1 回顾解耦知识蒸馏

前文提到, Logit 蒸馏的目标是最小化教师和学生模型输出的概率分布之间的差异. DKD 将蒸馏过程划分为两部分分别进行度量, 即目标类知识蒸馏 (target class knowledge distillation, TCKD) 和非目标类知识蒸馏 (non-target class knowledge distillation, NCKD). 这是因为目标类和非目标类在蒸馏过程中扮演着不同的角色, 分别度量它们之间的差异有助于学生模型更好地捕捉不同类别的信息. 对于输入的一个样本X, 目标类和非目标类损失可分别表示为:

$$\mathcal{L}_{\mathrm{TCKD}} = KL(s[y]\|t[y]) \tag{1}$$

$$\mathcal{L}_{\mathrm{NCKD}} = \sum_{i\neq y} KL(s[i]\|t[i]) \tag{2}$$

其中, y对应真实标签, t和s分别表示教师模型和学生模型的输出. 将目标类和非目标类损失加权求和, 得到最终 DKD 的损失为:

$$\mathcal{L}_{\mathrm{DKD}} = \alpha\cdot\mathcal{L}_{\mathrm{TCKD}} + \beta\cdot\mathcal{L}_{\mathrm{NCKD}} \tag{3}$$

基于 Logit 蒸馏的主要障碍在于任何 Logit 向量$Z_i = f(x_i)$相比于对应的特征向量要更加紧凑, 这使得很难提取嵌入在教师模型中的丰富信息.

DKD 仍存在如下限制: (1) 单实例问题. 尽管 DKD 通过解耦目标类和非目标类提高了蒸馏精度, 但孤立

样本的 Logit 缺乏对批次内相互关系的理解, 换句话说, 它没有在多张图像之间建立联系, 忽略了更广泛的样本间动态关系. (2) 类内依赖性问题. 如图 1 所示, 预测矩阵 (图 1(a)) 的每一行代表样本属于相应类别的预测分数. 通过对列进行平均, 可以计算出一个类内分布 (图 1(b)). 该分布表示一个类别在其他类别中的相似程度. 对任意一个输入, 教师模型对每个单独类别的预测分数都会有一个分布来反映教师模型对数据的正向偏差. 比如, “猫”与其他车辆类别有较大不同, 因此其类内分数较低, 此时, 教师模型可以比其他同时存在的类别更“自信”地对“猫”实例进行分类. 然而, DKD 在蒸馏过程中并没有将这种类内差异从教师模型传递给学生模型.

2.2 基于感知重构的解耦知识蒸馏

当在整个批次中观察实例 x_i 时, 知识蒸馏的过程又可以进一步被丰富. 关于 x_i 的信息度量 K 可以表示为:

$$K(x_i) \propto D(x_i) + \sum_{j \neq i} R(x_i, x_j) \tag{4}$$

其中, D 表示捕捉到的固有暗知识, R 表示捕捉到的实例之间的相互动态关系.

为了实现上述论断, 本文提出一种基于感知重构的解耦知识蒸馏方法. 该方法通过重构类内分布使学生模型能够捕捉到教师模型中更深层次的结构关系. 图 2 展示了感知模块的构建过程. 对于给定的一个样本批次 $B = \{x_1, x_2, \cdots, x_n\}$, 教师模型和学生模型分别为批次中的样本生成相应的 Logit 分数, 分别表示为 $z_i^t = \{z_{i1}^t, z_{i2}^t, \cdots, z_{iC}^t\}$ 和 $z_i^s = \{z_{i1}^s, z_{i2}^s, \cdots, z_{iC}^s\}$. 计算第 j 类在批次中的均值 (U_j^t, U_j^s) 和方差 (V_j^t, V_j^s). 这些数值将被用于批次中预测分数的标准化, 生成新的分数表示如式 (5).

$$h_{ij} = \frac{z_{ij} - U_j}{\sqrt{V_j}} \tag{5}$$

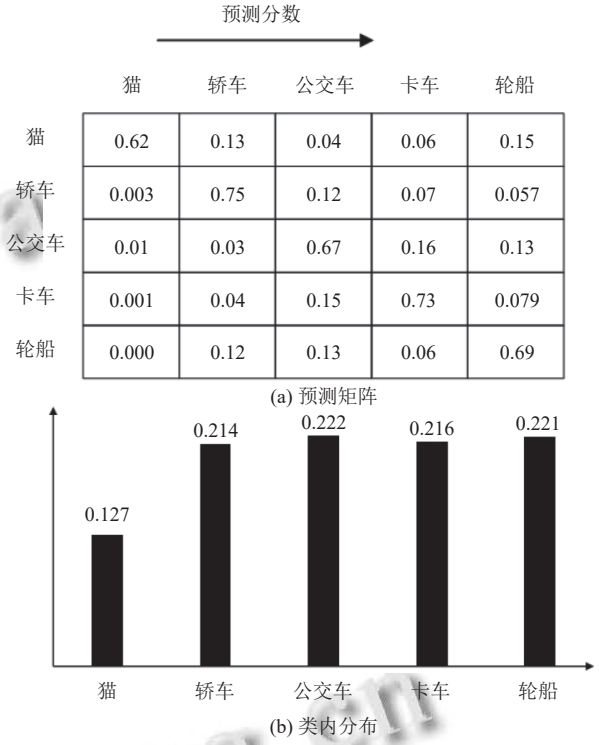

图 1 类内依赖

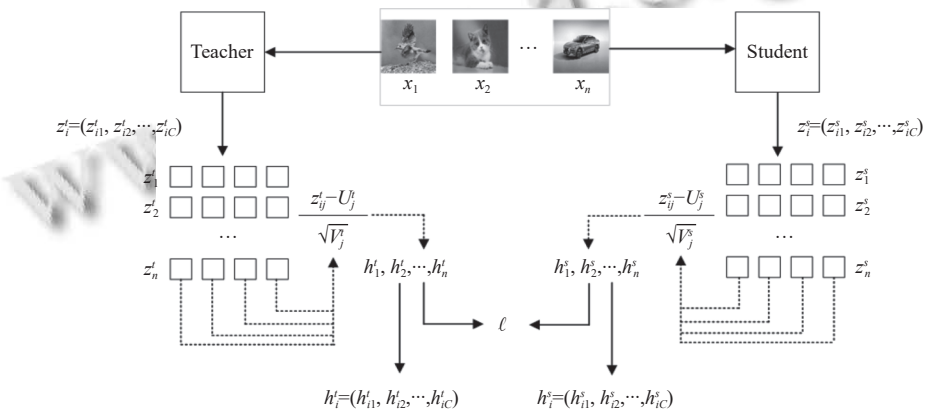

图 2 感知模块构建

因此, 样本 i 经过重构后的 Logit 可以表示为:

$$h_i = \left(\frac{z_{i1} - U_1}{\sqrt{V_1}}, \frac{z_{i2} - U_2}{\sqrt{V_2}}, \cdots, \frac{z_{iC} - U_C}{\sqrt{V_C}}\right) \tag{6}$$

其中, C 表示总的类别数. 在这种情况下, 重构后的 Logit 不仅能够保留每个实例的原始信息, 还能够在批次内捕捉不同实例之间的上下文关系. 这种重构方式使得

教师模型和学生模型能够在相同的尺度上对类内预测进行调整, 从而在保持方差一致性的同时更准确地反映出类别间的相互关系. 这个新的 Logit 集合为每个实例提供了更深入的视角, 本文将其称为“感知”.

整个蒸馏过程如图 3 所示, 对于第t类样本i, 经过分类网络得到原始 Logit 输出, 之后进入到感知模块中对 Logit 进行重构得到式 (6) 中新的预测分数h_i. h_i中的每个元素都可以通过 Softmax 函数和温度因子T进行评估.

$$p_{ij}=\frac{\exp\left(\frac{h_{ij}}{T}\right)}{\sum_{a=1}^{C}\exp\left(\frac{h_{ia}}{T}\right)} \tag{7}$$

其中, p_{ij}表示样本i属于第j个类别的概率. 对于非目标类集合 (不考虑第t类) 单独进行建模, $\overline{P}_i=[\overline{p}_{i1},\cdots,\overline{p}_{i(t-1)},\overline{p}_{i(t+1)},\cdots,\overline{p}_{iC}]\in R^{1\times(C-1)}$, 其中每个类别的概率可表示为:

$$\overline{p}_{ij}=\frac{\exp\left(\frac{h_{ij}}{T}\right)}{\sum_{a=1,a\neq t}^{C}\exp\left(\frac{h_{ia}}{T}\right)} \tag{8}$$

之后, 进入解耦模块, 分别考虑模型预测的目标类别和非目标类别, P_{iT}表示目标类别的概率, P_{iN}表示非目标类别概率:

$$P_{iT}=\frac{\exp\left(\frac{h_{it}}{T}\right)}{\sum_{a=1}^{C}\exp\left(\frac{h_{ia}}{T}\right)} \tag{9}$$

$$P_{iN}=\frac{\sum_{a=1,a\neq t}^{C}\exp\left(\frac{h_{ia}}{T}\right)}{\sum_{b=1}^{C}\exp\left(\frac{h_{ib}}{T}\right)} \tag{10}$$

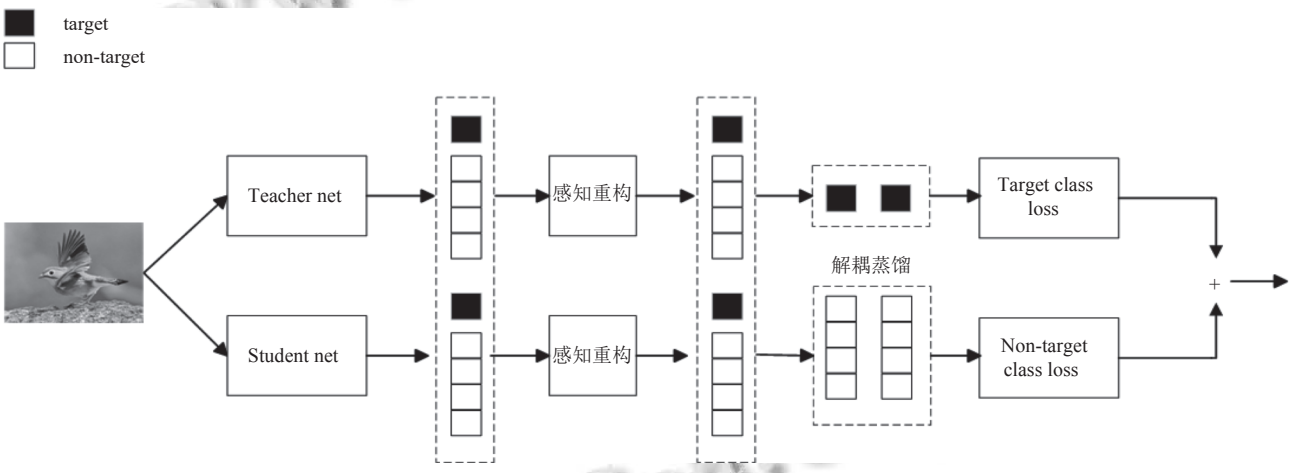

图 3 基于感知重构的解耦知识蒸馏

若T和S分别代表教师和学生, 那么经典 KD 损失可以表示为:

$$L_{\mathrm{KD}}=p_{it}^{(T)}\log\left(\frac{p_{it}^{(T)}}{p_{it}^{(S)}}\right)+\sum_{j=1,j\neq t}^{C}p_{ij}^{(T)}\log\left(\frac{p_{ij}^{(T)}}{p_{ij}^{(S)}}\right) \tag{11}$$

根据式 (7)–式 (10) 可以得到$p_{ij}=\overline{p}_{ij}\times p_{iN}$, 则$L_{\mathrm{KD}}$又可以表示为:

$$\begin{aligned}L_{\mathrm{KD}}=&p_{it}^{(T)}\log\left(\frac{p_{it}^{(T)}}{p_{it}^{(S)}}\right)+p_{iN}^{(T)}\log\left(\frac{p_{iN}^{(T)}}{p_{iN}^{(S)}}\right)\\&+p_{iN}^{(T)}\sum_{j=1,j\neq t}^{C}\overline{p}_{ij}^{(T)}\log\left(\frac{\overline{p}_{ij}^{(T)}}{\overline{p}_{ij}^{(S)}}\right)\\=&KL(b^{(T)}\|b^{(S)})+(1-p_{it}^{(T)})KL(\overline{p}_i^{(T)}\|\overline{p}_i^{(S)})\end{aligned} \tag{12}$$

其中, $KL(b^{(T)}\|b^{(S)})$表示经过重构后的教师和学生 Logit 在目标类上的二值概率相似度, 即 *RTKD*. $KL(\overline{p}_i^{(T)}\|\overline{p}_i^{(S)})$则表示重构后的 Logit 在非目标类上的概率相似度, 即 *RNKD*. 此时将二者进行解耦处理, 引入两个超参数α和β分别作为二者权值, 损失函数进一步表示为:

$$L_{KD}=\alpha\cdot RTKD+\beta\cdot RNKD \tag{13}$$

其中, *RTKD* 通过捕捉并保持教师模型和学生模型之间重构后目标类的二值概率相似度确保学生模型能够准确学习教师模型对目标类的判别能力. *RNKD* 使得学生模型不仅学习如何区分目标类, 还能学习到教师模型在非目标类上的类别区分能力. 该损失函数不仅保留了 DKD 对目标类和非目标类信息解耦的优势, 还通

过感知重构进一步优化了模型对批次内样本动态关系的捕捉, 提高模型在多分类任务中的整体性能.

3 实验数据分析

3.1 数据集介绍

CIFAR-100 数据集[27]由 Krizhevsky 在 2009 年创建, 是广泛用于图像分类任务的标准数据集之一. 该数据集具有较小的分辨率且包含丰富的类别信息, 主要用于评估在有限数据和资源条件下模型的泛化能力和细粒度分类性能. 特别是在小样本和低分辨率场景中, 它能很好地检验本文方法的蒸馏效果.

ImageNet-1K 数据集[28]是由 Russakovsky 等人创建的 ImageNet 项目中的一个子集. 该数据集图片分辨率高且类别丰富, 覆盖了从动物、植物到各种人工物体的丰富类别和现实世界场景, 这种多样性和复杂性使其成为评估大规模图像分类模型性能的标准数据集. 该数据集可以很好验证本文方法在应对复杂、高维度数据分布时的有效性.

本文采用准确率 (*Accuracy*) 和 Top-1 *Accuracy* 来评价学生模型的性能. 这两个指标能够全面、直观地反映模型在不同数据集上的分类效果, 便于与其他研究进行对比.

$$Accuracy = \frac{TP+TN}{TP+TN+TP+TN} \times 100\% \quad (14)$$

3.2 实验参数

本文基于 PyTorch 深度学习框架, 搭配 3 张 NVIDIA GeForce RTX 3090 显卡在 Ubuntu 18.04 平台下进行实验. 本次实验选择 ResNet、VGG、ShuffleNet、MobileNet 和 Wide ResNet 作为基础网络结构, 教师模型和学生模型来自这些网络的不同组合. 在 CIFAR-100 数据集上, Batchsize 大小为 256, 每组实验的训练轮数都为 270, 超参数α和β分别设置为 1.0 和 8.0. 不同的学生模型对应的学习率初始值和衰减系数有所不同, 衰减周期统一从第 150 轮开始每 30 个 Epoch 衰减一次, 具体设置如表 1 所示.

表 1 CIFAR-100 上不同网络的学习率和衰减系数

学生网络	学习率 (lr)	衰减系数
ResNet	0.05	0.01
ShuffleNet	0.01	0.1
MobileNet	0.01	0.1
VGG	0.05	0.1
Wide ResNet	0.05	0.1

3.3 消融实验

为了研究感知模块对蒸馏性能的影响, 本文使用 ResNet110 作为教师网络, ResNet32 作为学生网络, 在 CIFAR-100 数据集上分别对传统 KD 和 DKD 进行消融实验. 在这些实验中, 本文通过移除或保留感知模块来观察学生模型分类准确率的变化, 从而评估感知模块在蒸馏过程中的作用. 经过 270 轮的训练后得到的实验结果如表 2 所示. 在传统 KD 基础上加入感知模块后, 学生模型的分类准确率达到 74.20%, 提升了 1.12 个百分点. 同样, 在 DKD 中提升了 0.87 个百分点. 消融实验的结果表明, 本文提出的感知模块通过对模型输出的 Logit 进行重构, 进一步捕获实例间的上下文关系, 增加了知识传递的深度和广度, 帮助学生模型更好地理解和区分不同类别, 实现更全面的知识传递.

表 2 消融实验结果对比 (%)

方案	学习率 (lr)
KD[5]	73.08
KD+感知重构	74.20
DKD[16]	74.11
DKD+感知重构	**74.98**

3.4 对比实验

为了验证本文方法的优越性和先进性, 本文就 CIFAR-100 和 ImageNet-1K 数据集的分类任务和其他方法进行比较, 结果如表 3 和表 4 所示. 本文的基准方法是解耦知识蒸馏, 加入感知模块后, 分类精度相较于基准方法都有所提升, Δ 表示本文提升量. 实验结果表明, 与 DKD 以及传统的知识蒸馏方法相比, 本文方法在每组教师学生网络蒸馏实验中都取得一定程度的改善. 在基于 CIFAR-100 数据集上的实验中, 本文方法在同系列教师学生网络组合中取得了 0.22–0.87 个百分点的提升, 这一结果表明在相同架构下, 本文方法能够更好地捕捉和传递教师模型的知识, 有效增强了学生模型的学习效果. 在不同系列教师学生网络组合中取得了 0.23–0.74 个百分点的提升, 这说明本文方法在面对教师和学生模型结构具有显著差异的情况下, 同样能够确保知识蒸馏的效果, 充分说明了该方法在多种架构下的广泛适用性和稳定性.

为了进一步验证本文方法的可扩展性, 本文还用了 ResNet34 和 ResNet18 的组合在 ImageNet-1K 数据集上做了对比实验, 如表 5 所示. 结果表明, 本文方法和其他方法相比仍然具有更优的表现.

表 3 同系列教师学生网络对比实验结果 (%)

Teacher	Student	FitNet[6]	RKD[23]	CRD[29]	OFD[30]	ReviewKD[31]	KD[5]	DKD[16]	本文方法	Δ
ResNet56	ResNet20	69.21	69.61	71.16	70.98	71.89	70.66	71.97	**72.47**	**0.50**
ResNet110	ResNet32	71.06	71.82	73.48	73.23	73.89	73.08	74.11	**74.98**	**0.87**
ResNet32×4	ResNet8×4	73.50	71.90	75.51	74.95	75.63	73.33	76.32	**76.90**	**0.58**
WRN-40-2	WRN-16-2	73.58	73.35	75.48	75.24	76.12	74.92	76.24	**76.46**	**0.22**
WRN-40-2	WRN-40-1	72.24	72.22	74.14	74.33	75.09	73.54	74.81	**75.09**	**0.28**
VGG13	VGG8	71.02	71.48	73.94	73.95	74.84	72.98	74.68	**75.33**	**0.65**

表 4 不同系列教师学生网络对比实验结果 (%)

Teacher	Student	FitNet[6]	RKD[23]	CRD[29]	OFD[30]	ReviewKD[31]	KD[5]	DKD[16]	本文方法	Δ
ResNet50	MobileNet-V2	63.16	64.43	69.11	69.04	69.89	67.35	70.35	**70.74**	**0.39**
ResNet32×4	ShuffleNet-V2	73.54	73.21	75.65	76.82	77.78	74.45	77.07	**77.30**	**0.23**
ResNet32×4	ShuffleNet-V1	73.59	72.28	75.11	75.98	77.45	74.07	76.45	**76.85**	**0.40**
WRN-40-2	ShuffleNet-V1	73.73	72.21	76.05	75.85	77.14	74.83	76.70	**77.44**	**0.74**
VGG13	MobileNet-V2	64.14	64.52	69.73	69.48	70.37	67.37	69.71	**70.09**	**0.38**

表 5 ImageNet-1K 数据集对比实验结果 (%)

指标	AT[9]	OFD[30]	CRD[29]	ReviewKD[31]	KD[5]	DKD[16]	本文方法
Top-1	70.69	70.81	71.17	71.61	70.66	71.70	**72.32**

4 结束语

本文针对解耦知识蒸馏存在的单实例和类内依赖等问题, 提出了一种基于感知重构的解耦知识蒸馏方法. 为了对类内 Logit 进行重构, 本文结合类内上下文关系构建了“感知矩阵”, 从而优化了教师模型和学生模型之间的知识传递过程. 通过在不同网络组合的教师模型和学生模型上的训练测试, 本文方法都取得不错的提升, 有效提高了知识传递效率和模型蒸馏性能.

参考文献

1 Dhanya VG, Subeesh A, Kushwaha NL, *et al*. Deep learning based computer vision approaches for smart agricultural applications. Artificial Intelligence in Agriculture, 2022, 6: 211–229. [doi: 10.1016/j.aiia.2022.09.007]

2 Lauriola I, Lavelli A, Aiolli F. An introduction to deep learning in natural language processing: Models, techniques, and tools. Neurocomputing, 2022, 470: 443–456. [doi: 10.1016/j.neucom.2021.05.103]

3 Balhara S, Gupta N, Alkhayyat A, *et al*. A survey on deep reinforcement learning architectures, applications and emerging trends. IET Communications, 2022. [doi: 10.1049/cmu2.12447]

4 Munikoti S, Agarwal D, Das L, *et al*. Challenges and opportunities in deep reinforcement learning with graph neural networks: A comprehensive review of algorithms and applications. IEEE Transactions on Neural Networks and Learning Systems, 2023. [doi: 10.1109/TNNLS.2023.3283523]

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

6 Romero A, Ballas N, Kahou SE, *et al*. Fitnets: Hints for thin deep nets. Proceedings of the 3rd International Conference on Learning Representations. San Diego: OpenReview.net, 2015.

7 Tung F, Mori G. Similarity-preserving knowledge distillation. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 1365–1374.

8 Chen YD, Wang S, Liu JJ, *et al*. Improved feature distillation via projector ensemble. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 878.

9 Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. Proceedings of the 5th International Conference on Learning Representations. Toulon: OpenReview.net, 2017.

10 Müller R, Kornblith S, Hinton G. When does label smoothing help? Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 422.

11 Phuong M, Lampert C. Towards understanding knowledge distillation. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019. 5142–5151.

12 Liu YF, Chen K, Liu C, *et al*. Structured knowledge distillation for semantic segmentation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 2599–2608.

13 He T, Shen CH, Tian Z, *et al*. Knowledge adaptation for efficient semantic segmentation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 578–587.
14 Chen GB, Choi W, Yu X, *et al*. Learning efficient object detection models with knowledge distillation. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 742–751.
15 Tang ST, Feng LT, Shao WQ, *et al*. Learning efficient detector with semi-supervised adaptive distillation. Proceedings of the 30th British Machine Vision Conference. Cardiff: BMVA Press, 2019. 215.
16 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. 11943–11952.
17 Kim Y, Rush AM. Sequence-level knowledge distillation. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin: ACL, 2016. 1317–1327.
18 Wang L, Yoon KJ. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 3048–3068. [doi: 10.1109/TPAMI.2021.3055564]
19 Heo B, Lee M, Yun S, *et al*. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu: AAAI, 2019. 3779–3787.
20 Li YT, Sun LY, Gou JP, *et al*. Feature fusion-based collaborative learning for knowledge distillation. International Journal of Distributed Sensor Networks, 2021, 17(11): 15501477211057037.
21 Passalis N, Tefas A. Learning deep representations with probabilistic knowledge transfer. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 283–299.
22 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, 2020. 9155–9163.
23 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. 3963–3971.
24 Liu YF, Cao JJ, Li B, *et al*. Knowledge distillation via instance relationship graph. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 7089–7097.
25 Hou YN, Ma Z, Liu CX, *et al*. Inter-region affinity distillation for road marking segmentation. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 12483–12492.
26 Tao XY, Hong XP, Chang XY, *et al*. Few-shot class-incremental learning. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 12180–12189.
27 Krizhevsky A. Learning multiple layers of features from tiny images. Technical Report, 2009. http://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf
28 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]
29 Tian YL, Krishnan D, Isola P. Contrastive representation distillation. arXiv:1910.10699, 2022.
30 Heo B, Kim J, Yun S, *et al*. A comprehensive overhaul of feature distillation. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 1921–1930.
31 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. 5006–5015.

(校对责编: 张重毅)