计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.009719] [CSTR: 32024.14.csa.009719]

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

基于能量和熵平衡转移的知识蒸馏①

盛自强, 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: 朱子奇, E-mail: 81029530@qq.com

摘　要: 知识蒸馏 (KD) 中的温度在以前的大多数工作中被设置为蒸馏过程的固定值. 然而, 重新研究温度时, 发现固定的温度限制了对每个样本中固有知识的利用. 本文根据能量得分将数据集分为低能量样本和高能量样本, 通过实验证实了低能量样本的置信度得分高, 表明其预测是确定的, 而高能量样本的置信度得分低, 意味着预测是不确定的. 为了通过调整非目标类预测来提取最佳的知识, 本文对低能量样本应用较高的温度以创建更平滑的分布, 并对高能量样本应用较低的温度以获得更清晰的分布. 此外, 为解决学生对突出特征的不平衡依赖和对暗知识的疏忽, 本文引入熵重加权的知识蒸馏, 这是利用教师预测中的熵在样本基础上重新加权能量蒸馏损失的方法. 本文方法可以很容易地应用于其他基于逻辑的知识蒸馏方法中, 并获得更好的性能, 可以更接近甚至优于基于特征的方法. 本文在图像分类数据集 (CIFAR-100、ImageNet) 上进行了广泛的实验, 证明了该方法的有效性.

关键词: 知识蒸馏; 能量; 熵; 暗知识; 蒸馏温度

引用格式: 盛自强,朱子奇.基于能量和熵平衡转移的知识蒸馏.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9719.html

Knowledge Distillation Based on Energy and Entropy Balanced Transfer

SHENG Zi-Qiang, ZHU Zi-Qi

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

Abstract: The temperature in knowledge distillation (KD) is set as a fixed value during the distillation process in most previous work. However, when the temperature is reexamined, it is found that the fixed temperature restricts inherent knowledge utilization in each sample. This study divides the dataset into low-energy and high-energy samples based on energy scores. Through experiments, it is confirmed that the confidence score of low-energy samples is high, indicating that predictions are deterministic, while the confidence score of high-energy samples is low, indicating that predictions are uncertain. To extract the best knowledge by adjusting non-target class predictions, this study applies higher temperatures to low-energy samples to generate smoother distributions and applies lower temperatures to high-energy samples to obtain clearer distributions. In addition, to address the imbalanced dependence of students on prominent features and their neglect of dark knowledge, this study introduces entropy-reweighted knowledge distillation, which utilizes the entropy predicted by teachers to reweight the energy distillation loss on a sample basis. This method can be easily applied to other logic-based knowledge distillation methods and achieve better performance, which can be closer or even better than feature-based methods. This study conducts extensive experiments on image classification datasets (CIFAR-100, ImageNet) to validate the effectiveness of this method.

Key words: knowledge distillation; energy; entropy; dark knowledge; distillation temperature

① 基金项目: 公安部科技计划 (2022JSM08)
收稿时间: 2024-05-29; 修改时间: 2024-06-26; 采用时间: 2024-07-11; csa 在线出版时间: 2024-11-15

近年来, 深度神经网络在图像分类、目标检测、语义分割等任务中取得了巨大成功. 然而, 这些模型通常在计算上非常昂贵并且占用内存很高, 使得它们难以在资源受限的设备上部署. 因此, 模型压缩近年来受到了较大的关注与研究. 其中, 知识蒸馏 (knowledge distillation)[1]以其优越的性能和易于实现的特点脱颖而出, 其通过从复杂的教师模型中提取有意义的信息来训练轻量级的学生模型, 使学生模型能够达到与教师模型相似的性能.

知识蒸馏根据转移知识的类型可以分为基于响应[2–4]、基于特征[5–9]、基于关系[10]这 3 类. 基于 logits 的知识蒸馏是最经典的方法, 知识蒸馏的损失函数通常由两部分组成: 具有硬标签的交叉熵 (CE) 损失和带有软标签的 KL 散度损失. 其中温度对软标签的光滑度有很大的影响, 较高的温度使软标签更光滑, 较小的温度使其柔软标签更清晰. 此外使用温度软化 logits 输出也是获取更多“暗知识”的重要步骤. 通常教师模型的输出过于敏锐, 使学生模型难以学习不正确的类之间的细微差别. 因此, 经常使用由“温度”调节的软标签来提高知识蒸馏的性能. 然而, 以往的大部分工作, 无论各种 logits 输出如何分布, 都将温度设定为一个固定的值, 这将会阻碍知识蒸馏的过程. 如图 1 所示, 当使用固定的温度将会出现软化不均匀的现象. 对于第 1 个样本, 它有一个接近 ground truth 标签的自信预测, 它的软标签包含不充分的“暗知识”, 因为固定的温度对该样本来说太小了. 第 3 个样本则是缺乏自信, 它的最终预测过于接近, 该样本的固定温度明显过大. 因此, 可以发现一个温度没有能力适当的软化所有样本的 logits 输出.

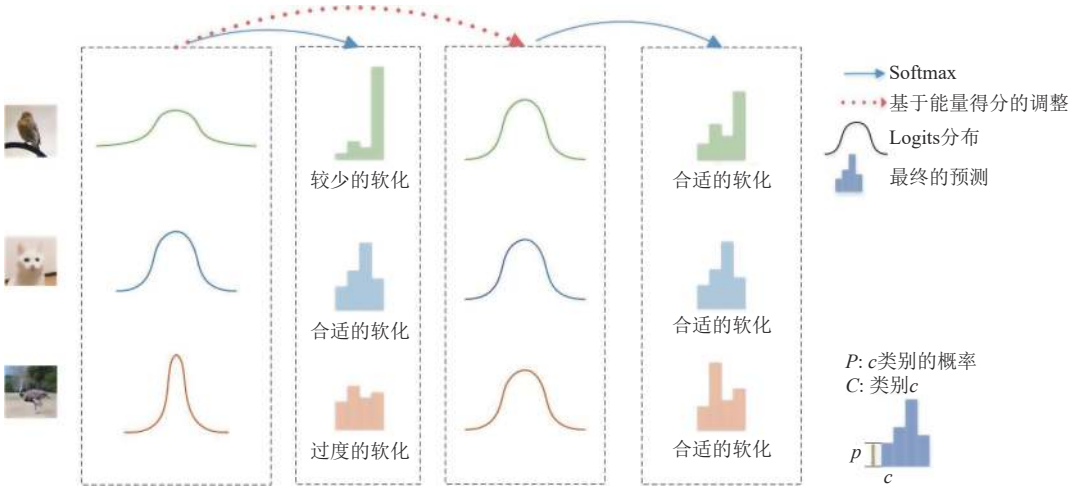

图 1 不同样本在固定温度和基于能量得分调整后的温度下的最终预测示意图

CTKD[11]采用对抗学习模块来预测样本温度, 以适应不同样本的难度. 本文提出了一种新的基于逻辑的蒸馏方法, 该方法通过最大化学生对教师知识的利用, 显著提高其成绩, 并且可以很容易地集成到现有的基于逻辑的蒸馏中. 本文方法对每张图像应用能量函数将整个数据集分为低能量和高能量样本, 然后对分离的样本进行不同温度蒸馏, 低能量样本采用高温, 高能量样本采用低温. 该方法可以使低能量样本的样本分布更平滑, 高能量样本分布更清晰, 有效的调整非目标类预测, 如图 1 所示.

尽管 KD 通过 KL 散度来最小化学生模型的预测和教师模型的预测, 但学生模型与老师模型之间仍有着显著的差距, 这归因于学生过度自信的预测. 为了研究此问题, 本文使用了熵, 这是信息论中的一个概念, 用于量化随机变量的不可预测性, 来测量预测的置信度. 通过实验得出 KD 学生在较低的熵下产生了更大的确定性预测, 这种过度自信意味着对显著特征的过度依赖, 潜在地忽略了复杂的暗知识. 因此, 本文提出熵重加权知识蒸馏, 该方法与焦点损失中的加权方法类似, 利用教师逻辑的熵作为蒸馏损失的动态权重. 本文的方法引入了对 KD 的样本智能适应性, 其中具有高熵的挑战性样本在训练中得到更大的重视, 而具有低熵简单的样本则被降低权重. 因此, 不仅鼓励学生从困难的样本中学习复杂的暗知识, 而且减少对突出知

识的依赖. 本文的主要贡献如下.

1) 为了解决恒定温度限制了对每个样本中固有知识的利用, 本文提出根据能量得分将数据集分为低能量样本和高能量样本, 对低能量样本应用较高的温度来创建更平滑的分布, 并对高能量样本应用较低的温度以获得更清晰的分布.

2) 为了进一步解决学生对突出特征的不平衡依赖和对暗知识的疏忽, 本文引入了熵重加权的知识蒸馏的方法, 利用教师软化逻辑的熵, 在样本基础上重新加权蒸馏损失, 确保更加平衡的知识转移.

3) 通过在 CIFAR-100 和 ImageNet 上对多种师生模型进行广泛的实验, 本文的方法可以很好地与 DKD[2] 和 MLD[4]基于逻辑的方法相结合, 有效地提高了蒸馏性能.

1 相关工作

知识蒸馏这一概念最早由 Hinton 等人[1]于 2015 年提出, 旨在通过将复杂的教师网络中提取的暗知识迁移到轻量级的学生网络中来提高学生网络的性能.

1.1 样本加权

在深度学习中, 已经提出了很多种样本加权方法, Ren 等人[12]通过元学习算法动态地为训练样本分配权重, 来解决样本偏差和标签噪声. 类似的, Lin 等人[13]引入焦点损失来强调较难地实例, 同时降低较容易样本的损失权重, 从而提高目标检测模型的性能.

Lu 等人[14]将样本加权扩展到知识蒸馏, 在自然语言处理任务中使用样本重新加权, 由于之前的工作强调了样本知识的重要性[15,16], 该方法利用元学习方法为每个实例重新加权损失项, 从而改进了蒸馏过程. 然而, 元学习的训练可能需要大量的计算和时间, 本文的熵重新加权 KD 提供了一种效率高的替代方案, 通过教师预测的熵来重新加权 KD 损失, 确保简化和有效的知识转移过程.

1.2 基于能量的学习

基于能量的机器学习模型有着悠久的历史, 其始于玻尔兹曼机[17], 这是整个网络中具有相关能量的单元网络. 基于能量的学习[18]为各种概率和非概率学习方法提供了一个统一的框架. Zhao 等人[19]展示了利用能量函数来训练生成对抗网络 (GAN), 其中鉴别器利用能量值来区分真实图像和生成图像. Liu 等人[20]证明了非概率能量分数可以直接用于评估 out-of-distribution (OOD) 不确定性的分数函数中. 在这些工作的基础上, 本文提出的框架将非概率能量值的使用扩展到知识蒸馏中, 为低能量和高能量样本提供不同的知识.

2 本文方法

2.1 基于能量的模型

基于能量的模型 (EBM)[21]的本质是建立一个函数 $E(x): R^D \to R$, 该函数将输入空间的每个点 x 映射为一个能量的非概率标量. 一个能量值的集合可以通过 Gibbs 分布转化为一个概率 $p(x)$:

$$p(y \mid x) = \frac{\mathrm{e}^{-E(x,y)/T}}{\int_{y'} \mathrm{e}^{-E(x,y')/T}} = \frac{\mathrm{e}^{-E(x,y)/T}}{\mathrm{e}^{-E(x)/T}} \tag{1}$$

而拥有更深层次规模架构的网络会取得更好的表现, 因此会选取一个庞大的教师网络, 这也就导致了教师网络和学生网络规模相差较大. 在这种差距下, 使用传统的 KL 散度来精确的恢复预测变得更为乏力.

式 (1) 中分母 $\int_{y'} \mathrm{e}^{-E(x,y')/T}$ 为配分函数, T 是温度参数. 数据点 $x \in R^D$ 的亥姆霍兹自由能 $E(x)$ 可以表示为对数配分函数的负数:

$$E(x) = -T \cdot \log \int_{y'} \mathrm{e}^{-E(x,y')/T} \tag{2}$$

考虑一个判别神经网络分类器 $f(x): R^D \to R^K$, 它将输入 $x \in R^D$ 映射到 K 个 logits 的实数上:

$$p(y \mid x) = \frac{\mathrm{e}^{f_y(x)/T}}{\sum_{i}^{K} \mathrm{e}^{f_i(x)/T}} \tag{3}$$

其中, $f_y(x)$ 表示 $f(x)$ 的第 y 个类标号的 logits. 通过连接式 (1) 和式 (3) 可以定义给定输入 (x,y) 的能量为 $E(x,y) = -f_y(x)$, 在不改变神经网络 $f(x)$ 参数的情况下, 自由能函数表示:

$$E(x; f) = -T \cdot \log \sum_{i}^{K} \mathrm{e}^{f_i(x)/T} \tag{4}$$

根据能量分数分类的动机, 可以将低似然的输入数据视为高能样本. 这可以通过利用基于能量的模型的表示的数据密度函数 $p(x)$ 来实现[22].

$$p(x) = \frac{\mathrm{e}^{-E(x;f)/T}}{\int_{x} \mathrm{e}^{-E(x;f)/T}} \tag{5}$$

通过式 (5) 可以得到:

$$\log p(x) = -\frac{E(x;f)}{T} - C \tag{6}$$

式 (6) 表明能量函数与对数似然函数成正比. 也就是说能量较低的样本可以很容易地被发现与识别, 但能量较高的样本却不容易被识别. 因此可以有效地利用能量函数的可区分性对样本进行分类, 从而促进知识的最优蒸馏.

2.2 基于能量的知识蒸馏

利用上述的能量分数, 本文提出了一种基于能量的知识蒸馏. 具体来说, 使用式 (4) 通过预训练的教师模型的 logits 获取每个样本的能量分数, 再根据图像的能量分数将其分为低能量和高能量组, 并对每一组应用不同的温度缩放, 从而增强学生模型的学习能力.

知识蒸馏的目标是将封装在教师模型软概率输出中的暗知识转移到学生模型中. 在分类任务中, 软化概率是通过温度缩放的 Softmax 函数计算的, 由式 (7) 计算:

$$p_i(T) = \frac{\exp\left(\frac{y_i}{T}\right)}{\sum_{j=1}^{C}\exp\left(\frac{y_i}{T}\right)} \tag{7}$$

其中, $p_i(T)$是第 i 类经过温度超参数T软化后的概率输出, y_i表示第 i 类的 logits, C 为类的总数.

知识蒸馏的核心思想在于最小化损失函数, 使师生的软逻辑对齐. KD 的损失为:

$$L_{\mathrm{KD}} = T^2 L_{\mathrm{KL}}(\sigma(Z_S/T), \sigma(Z_T/T)) \tag{8}$$

其中, Z_S, Z_T 分别为学生和教师输出的 logits, σ为 Softmax 函数. 本文根据能量分数调整预测的置信度, 使学生获得更广泛的知识, 该调整可以通过简单的缩放温度来实现, 如下所示:

$$L_{\mathrm{our}} = {T_{\mathrm{our}}}^2 L_{KL}(\sigma(Z_S/T_{\mathrm{our}}), \sigma(Z_T/T_{\mathrm{our}})) \tag{9}$$

$$E_e = E(x;Z_T) \tag{10}$$

$$T_{\mathrm{our}} = \begin{cases} T + T_{(+)}, & E_e \leqslant E_e^{\mathrm{low}} = E_e[N\cdot r] \\ T + T_{(-)}, & E_e \geqslant E_e^{\mathrm{high}} = E_e[-N\cdot r] \\ T, & \text{other} \end{cases} \tag{11}$$

其中, x是输入的图像, 每个样本的能量分数都可以使用教师分类器Z_T计算, 这样可以得到所有图像的能量分数, 并将它们按升序排列. E_e^{low}和E_e^{high}是定义低能量和高能量分类范围的常数值. $T_{(+)}$和$T_{(-)}$分别为一个正整数和负整数, 用来增加和降低温度. N为训练样本的总数, 并使用总样本的百分比r来建立能量分类范围, 如图 2 所示.

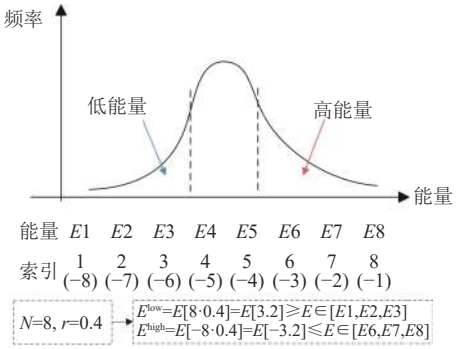

图 2 样本的能量分布

该方法可以增加低能量样本中非目标类的重要暗知识, 同时增加高能样本中目标类的预测.

2.3 基于熵加权的知识蒸馏

在知识蒸馏中, 不同样本会有不同程度的挑战, 学生可能会去学习较多的简单的样本, 使其产生过度自信, 从而对较难暗知识的学习减少. 因此, 本文通过教师对每个样本挑战性的见解, 去指导学生, 减少学生的过度自信.

在信息论中, 熵是对不确定性的度量[23]. 本文将熵作为评估样本难度的指标, 引导学生更多的关注具有挑战的样本. 具体来说, 老师的软化概率预测的熵提供了对每个样本的挑战性的见解, 通过式 (12) 计算:

$$H_n^{\mathcal{T}} = -\sum_{i=1}^{C} p_{n,i}^{\mathcal{T}}(T_{\mathrm{our}}) \log\left(p_{n,i}^{\mathcal{T}}(T_{\mathrm{our}})\right) \tag{12}$$

其中, $H_n^{\mathcal{T}}$ 为教师$\mathcal{T}$ 预测的第n 个样本 logits 的熵, $p_{n,i}^{\mathcal{T}}(T_{\mathrm{our}})$为样本 n 的第i个类别的概率, 用式 (11) 中的温度T_{our}来软化, 来确定它们准确地反映了教师对每个样本所感知到的难度. 本文提出的熵加权能量的损失函数为:

$$L_{\mathrm{EEKD}} = \frac{1}{N}\sum_{n=1}^{N} H_n^{\mathcal{T}} L_{\mathrm{our},n} \tag{13}$$

其中, $L_{\mathrm{our},n}$表示第 n 个样本计算的能量 KD 损失, n 表示数据集中的样本总数, 熵值$H_n^{\mathcal{T}}$作为调节$L_{\mathrm{our},n}$的动态加权因子. 因此, 该加权方法放大了教师认为具有挑战性的样本的蒸馏损失, 同时减少了简单实例的蒸馏损失. 本文提出的 EEKD 如图 3 所示, 算法 1 中提供了类似 PyTorch 风格的伪代码.

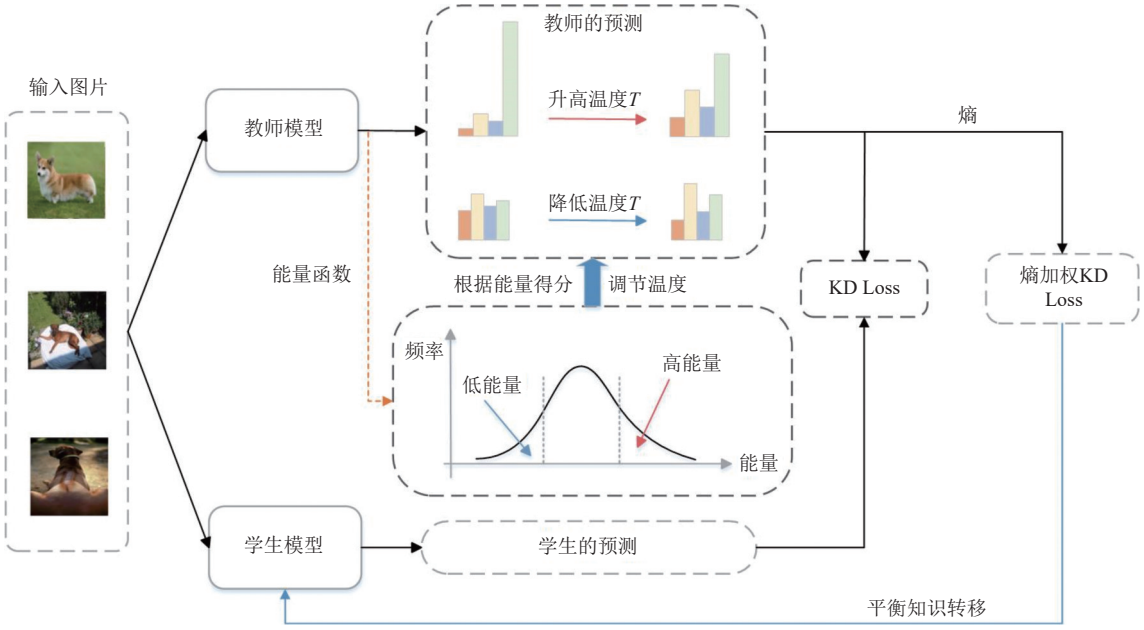

图 3 EEKD 架构图

算法 1. EEKD 算法

```
# x: 输入的图片
# model_s, model_t: 学生和教师的模型
# T: 温度超参数
# T_our: 本文的温度超参数
# E_high, E_low: 高能量阈值和低能量阈值
# E(): 能量函数
y_t=model_t(x) #教师输出的 logits
y_s=model_s(x) #学生输出的 logits
E_t=E(y_t) #教师 logits 的能量值
for i, E_i in enumerate(E_t):
    if E_i<E_low #低能量样本
        T_our[i]=T[i]+T_(+)
    if E_i>E_high #高能量样本
        T_our[i]=T[i]+T_(−)
p_t=F.Softmax(y_t/T_our, dim=1) #教师的预测
p_s=F.Softmax(y_s/T_our, dim=1) #学生的预测
L_our=(T_our)^2×F.kl_div(p_s, p_t) #基于能量的 KD 损失
_p_t=F.Softmax(y_t/T_our, dim=1)
H=−(_p_t×_p_t.log()).sum(1) #教师 logits 的熵
L_EEKD=(L_our.sum(1)×H).mean() #最终的 KD 损失
```

3 实验分析

3.1 数据集

CIFAR-100[24]是一个包含 50 000 张训练图像和 10 000 张测试图像的数据集, 用于图像分类任务和计算机视觉研究. 该数据集由 100 个类别组成, 每个类别有 600 张大小为 32×32 的彩色图像, 其中 500 张作为训练集, 100 张作为测试集.

ImageNet (ILSVRC2012)[25]是一个具有挑战性的数据库, 也是图像分类任务中使用次数较多的大型数据库包含 1 000 个类别的大型数据集, 其中有训练集有 128 万张图片, 验证集有 50 000 张图片, 测试集有 100 000 张图片.

3.2 实验参数

本文实验在 Linux 系统上进行并基于 CUDA 10.2, PyTorch 1.9.0 完成模型的搭建和网络的训练. 实验使用 ResNet[26]、VGG[27]、ShuffleNet[28,29]、MobileNet[30]和 Wide ResNet[31]网络. 对于 CIFAR-100 数据集, 实验的 Batchsize 大小为 64, 进行 360 个 Epoch 的训练, 学习率最初设置为 0.1, 并在 150 个 Epoch 之后, 每 30 个 Epoch 的学习率衰减 0.1. 此外式 (11) 中的超参数 r 设置为 0.4, $T_{(+)}$和$T_{(-)}$设置为 2. 对于 ImageNet (ILSVRC-2012) 数据集, 将 Batchsize 大小设置为 256, 所有模型进行 150 个 Epoch 的训练, 实验将学习率初始化为 0.1, 然后每 30 个 Epoch 衰减为当前大小的 0.1.

3.3 实验结果

为了验证本文方法的先进性, 将本文的方法与其他具有代表性先进的方法进行比较. 在保持原始设计的同时, 将本文的方法集成到基于逻辑的 DKD[2]和 MLD[4]方法中. 如表 1 和表 2 所示, 在 CIFAR-100 数据集中无论是将本文的方法与以前基于逻辑的 KD 结合,

还是与最近先进的基于逻辑的 DKD 和 MLD 结合, 都获得了较高的性能提升, 在同系列师生网络中取得了 0.2–0.6 个百分点的提升, 在不同系列师生网络中取得了 0.2–0.7 个百分点的提升.

表 1 CIFAR-100 数据集同系列教师学生网络实验结果 (%)

Teacher 模块	Student 模块	Teacher	Student	FitNet[5]	RKD[10]	PKT[32]	CTKD[11]	CRD[33]	OFD[7]	ReviewKD[6]	KD[1]	本文方法	DKD[2]	本文方法+DKD	MLD[4]	本文方法+MLD
ResNet56	ResNet20	72.34	69.06	69.21	69.61	70.34	71.19	71.16	70.98	71.89	70.66	**71.34 (+0.68)**	71.97	**72.31 (+0.34)**	72.19	**72.76 (+0.57)**
ResNet32×4	ResNet8×4	79.42	72.50	73.50	71.90	73.64	73.39	75.51	74.95	75.63	73.33	**74.81 (+1.48)**	76.32	**76.83 (+0.51)**	77.08	**77.68 (+0.60)**
WRN-40-2	WRN-16-2	75.61	73.26	73.58	73.35	74.65	75.45	75.48	75.24	76.12	74.92	**75.78 (+0.86)**	76.24	**76.68 (+0.44)**	76.63	**77.17 (+0.54)**
WRN-40-2	WRN-40-1	75.61	71.98	72.24	72.22	73.45	73.93	74.14	74.33	75.09	73.54	**74.30 (+0.76)**	74.81	**75.01 (+0.20)**	75.35	**75.74 (+0.39)**
VGG13	VGG8	74.64	70.36	71.02	71.48	71.62	73.52	73.94	73.95	74.84	72.98	**74.03 (+1.05)**	74.68	**74.96 (+0.28)**	75.18	**75.57 (+0.39)**

表 2 CIFAR-100 数据集不同系列教师学生网络实验结果 (%)

Teacher 模块	Student 模块	Teacher	Student	FitNet[5]	RKD[10]	PKT[32]	CTKD[11]	CRD[33]	OFD[7]	ReviewKD[6]	KD[1]	本文方法	DKD[2]	本文方法+DKD	MLD[4]	本文方法+MLD
ResNet50	MobileNetV2	79.34	64.60	63.16	64.43	66.52	68.47	69.11	69.04	69.89	67.35	**69.33 (+1.98)**	70.35	**70.73 (+0.38)**	71.04	**71.34 (+0.30)**
ResNet32×4	ShuffleNetV2	79.42	71.82	73.54	73.21	74.69	75.37	75.65	76.82	77.78	74.45	**75.89 (+1.44)**	77.07	**77.57 (+0.50)**	78.44	**78.72 (+0.28)**
ResNet32×4	ShuffleNetV1	79.42	70.50	73.59	72.28	74.10	74.48	75.11	75.98	77.45	74.07	**75.23 (+1.16)**	76.45	**77.02 (+0.57)**	77.18	**77.85 (+0.67)**
WRN-40-2	ShuffleNetV1	75.61	70.50	73.73	72.21	75.03	75.78	76.05	75.85	77.14	74.83	**75.86 (+1.03)**	76.70	**77.08 (+0.38)**	77.44	**77.79 (+0.35)**
VGG13	MobileNetV2	74.64	64.60	64.14	64.52	67.35	68.50	69.73	69.48	70.37	67.37	**68.97 (+1.60)**	69.71	**70.17 (+0.46)**	70.57	**70.89 (+0.32)**

表 3 和表 4 展示了本文方法在 ImageNet 数据集上的实验结果, 进一步证明了本文方法的先进性. 本文实验对同系列师生网络组选取了 ResNet34 和 ResNet18, 对不同系列师生网络组选取了 ResNet50 和 MobileNet.

表 3 ImageNet 数据集的同系列师生网络实验结果 (%)

方法	Teacher ResNet34	Student ResNet18	OFD[7]	CRD[33]	ReviewKD[6]	KD[1]	DKD[2]	本文方法+DKD
Top-1	73.31	69.75	70.81	71.17	71.61	70.66	71.70	**72.22**

表 4 ImageNet 数据集的不同系列师生网络实验结果 (%)

方法	Teacher ResNet50	Student MobileNet	OFD[7]	CRD[33]	ReviewKD[6]	KD[1]	DKD[2]	本文方法+DKD
Top-1	76.16	68.87	71.25	71.37	72.56	68.58	72.05	**72.96**

3.4 消融实验

在本文提出的基于能量和熵的蒸馏方法中包含了基于能量的蒸馏和基于熵重加权的蒸馏. 为了进一步了解这两种知识蒸馏对模型性能提升的有效性, 本文通过消融实验探讨了基于能量和熵蒸馏方法的不同情况, 在 CIFAR-100 数据集上使用 ResNet32×4 和 ResNet8×4 分别作为教师和学生模型, 并基于 DKD 作对比, 如表 5 所示.

此外, 为了评估高能量与低能量样本采用不同温度的可行性与熵中的T_{our}能否准确地反映教师对每个样本所感知的难度, 本文进行了温度消融实验, 如表 6 所示. 其中 Low、High 表示只对低能量样本和高能量

样本应用温度缩放. 而 Low+High 表示对两种能量样本都使用温度缩放, 其中 Ours 表示不仅对两种能量样本运用了温度缩放, 而且对熵使用了式 (11) 的温度进行微调. 实验表明调整两种能量类型的温度比只调整一种能量类型的温度产生了更好的结果, 并且使用温度T_{our}可以准确地反映教师对每个样本所感知的难度.

表 5 不同方案在 CIFAR-100 数据集上的实验结果 (%)

方案	基于能量	基于熵	Top-1
A	√	√	76.83
B	√	×	76.49
C	×	√	76.58
D	×	×	76.32

表 6 温度消融在 CIFAR-100 数据集上的实验结果 (%)

方法	Teacher+student	
	ResNet34+ResNet18	ResNet32+ShuffleNet V2
Low	73.28	75.36
High	73.85	75.33
Low+High	74.58	75.85
Ours	74.81	75.89

4 结束语

在本文中, 为解决单一温度无法提取全部知识, 引入了样本的能量得分, 根据能量得分将数据集分为低能量和高能量样本, 并对低能量样本应用较高的温度, 对高能量样本应用较低的温度. 此外, 为了让学生更多的关注暗知识, 本文通过熵重新加权能量蒸馏损失, 利用教师软化逻辑的熵, 在样本基础上重新加权蒸馏损失, 确保更加平衡的知识转移. 通过在同系列教师网络和学生网络以及不同系列教师网络和学生网络上的训练测试, 本文模型的图像分类准确率都取得了不错的进步, 证明了该方法的有效性和泛化性.

参考文献

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

2 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.

3 Chen DF, Mei JP, Zhang HL, *et al*. Knowledge distillation with the reused teacher classifier. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 11933–11942.

4 Jin Y, Wang JQ, Lin DH. Multi-level logit distillation. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 24276–24285.

5 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, 2015.

6 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.

7 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.

8 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.

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

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 Li Z, Li X, Yang LF, *et al*. Curriculum temperature for knowledge distillation. Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington: AAAI Press, 2023. 1504–1512.

12 Ren MY, Zeng WY, Yang B, *et al*. Learning to reweight examples for robust deep learning. Proceedings of the 35th International Conference on Machine Learning. Stockholm: PMLR, 2018. 4334–4343.

13 Lin TY, Goyal P, Girshick R, *et al*. Focal loss for dense object detection. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 2999–3007.

14 Lu P, Ghaddar A, Rashid A, *et al*. RW-KD: Sample-wise loss terms re-weighting for knowledge distillation. Proceedings of the 2021 Findings of the Association for Computational Linguistics. Punta Cana: ACL, 2021. 3145–3152.

15 Tang JX, Shivanna R, Zhao Z, *et al*. Understanding and improving knowledge distillation. arXiv:2002.03532, 2020.

16 Zhou HL, Song LC, Chen JJ, *et al*. Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective. Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021.

17 Salakhutdinov R, Larochelle H. Efficient learning of deep Boltzmann machines. Proceedings of the 13th International Conference on Artificial Intelligence and Statistics. Sardinia: JMLR, 2010. 693–700.

18 Ranzato M, Boureau YL, Chopra S, *et al*. A unified energy-based framework for unsupervised learning. Artificial Intelligence and Statistics, PMLR, 2007: 371–379.

19 Zhao JJ, Mathieu M, LeCun Y. Energy-based generative adversarial networks. Proceedings of the 5th International Conference on Learning Representations. Toulon: OpenReview.net, 2017.

20 Liu WT, Wang XY, Owens JD, *et al*. Energy-based out-of-distribution detection. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1802.

21 LeCun Y, Chopra S, Hadsell R, *et al*. A tutorial on energy-based learning. Bakir G, Hofmann T, Schölkopf B, *et al*. Predicting Structured Data. Cambridge: MIT Press, 2006.

22 Grathwohl W, Wang KC, Jacobsen JH, *et al*. Your classifier is secretly an energy based model and you should treat it like one. Proceedings of the 8th International Conference on Learning Representations. Addis Ababa: OpenReview.net, 2020.

23 Shannon CE. A mathematical theory of communication. Bell System Technical Journal, 1948, 27(3): 379–423. [doi: 10.1002/j.1538-7305.1948.tb01338.x]

24 Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. Handbook of Systemic Autoimmune Diseases, 2009, 1(4): 1–60.

25 Deng J, Dong W, Socher R, *et al*. ImageNet: A large-scale hierarchical image database. Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 248–255.

26 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.

27 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Proceedings of the 3rd International Conference on Learning Representations. San Diego, 2015.

28 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.

29 Ma NN, Zhang XY, Zheng HT, *et al*. ShuffleNet V2: Practical guidelines for efficient CNN architecture design. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 116–131.

30 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.

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

32 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]

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

(校对责编: 孙君艳)