科学报告

Check for updates

开放

基于注意力和特征迁移的知识蒸馏

杨国良、干帅英 *、盛洋洋和杨浩

现有的知识蒸馏（KD）方法主要基于特征、逻辑或注意力，其中特征和逻辑代表卷积神经网络不同阶段的推理结果，而注意力图则象征推理过程。由于这两者在时间上具有连续性，仅将其中之一迁移到学生网络会导致不理想的结果。我们对师生网络之间的知识迁移进行了不同程度的研究，揭示了同时将与推理过程和推理结果相关的知识迁移到学生网络的重要性，为 KD 研究提供了新的视角。在此基础上，我们提出了基于注意力和特征迁移的知识蒸馏方法（AFT-KD）。首先，我们利用转换结构将中间特征转换为同时包含推理过程信息和推理结果信息的注意力与特征块（AFB），并迫使学生学习 AFB 中的知识。为了节省学习过程中的计算量，我们采用分块操作来对齐师生网络。此外，为了平衡不同损失之间的衰减率，我们设计了一种基于损失优化率的自适应损失函数。实验表明，AFT-KD 在多个基准测试中取得了最先进的性能。

卷积神经网络（CNN）的成功给计算机视觉领域带来了深远的变革 1。然而，为了实现性能提升，CNN 不断扩大规模，随之而来的是计算和存储需求的增加，这对于在资源有限的边缘设备上部署而言是一个巨大的挑战。为了获得高效且紧凑的网络，诸如低秩分解 [2–6]、网络剪枝 [7–11]、量化 [12–14]、高效网络架构设计 [15–17] 以及知识蒸馏 [18–20] 得到了快速发展。其中，知识蒸馏（KD）是一种很有前景的方法。

知识蒸馏（KD）迫使学生网络从更大的教师网络中提取知识，以提高性能。KD 最早在文献 21 中提出，是一种基于逻辑的知识蒸馏方法。近年来，出现了基于特征和基于注意力的 KD 方法。基于特征的 KD 使学生网络能够从教师网络的中深层特征中提取知识，其特点是学生网络的精度较高。在基于逻辑的 KD 方法中，知识从教师网络的逻辑输出中提取，学生网络同时受到真实标签和教师逻辑的监督。基于注意力的 KD 方法在文献 30 中提出，该方法迫使学生网络模拟从教师网络传递过来的注意力图，以学习在分类过程中应该关注哪些有价值的内容。

现有方法根据知识转移的类型将知识蒸馏分为三类：基于特征的、基于逻辑的和基于注意力图的，并着重研究如何利用单一特征来提升网络性能。与现有方法不同，我们的目标是探究如何利用更全面的推理信息来提高蒸馏性能。为此，我们重新审视卷积神经网络（CNN）架构，发现 CNN 中的卷积层不仅在空间上具有连续性，在数据的前向传播过程中还表现出时间连续性。在从输入特征到输出类别预测的连续推理过程中，特征图和逻辑预测分别代表中间推理阶段和整个推理过程的推理结果，而注意力图则反映了卷积层在推理过程中最关注的内容，其中包含与推理过程相关的知识。基于上述发现，我们将现有的知识蒸馏（KD）方法重新分为两类：基于推理过程的和基于推理结果的，传统分类中基于特征和逻辑的蒸馏方法统称为基于推理结果的蒸馏方法。当从这一新视角审视知识蒸馏方法时，我们发现现有方法仅使用了推理过程信息或推理结果信息中的一种，而忽略了两者之间的关联性。针对这一现象，我们提出了一种基于注意力和特征转移的知识蒸馏方法（AFT-KD，图 1），该方法

江西理工大学电气工程与自动化学院，中国江西省赣州市，341000。* 邮箱：6720210550@mail.jxust.edu.cn

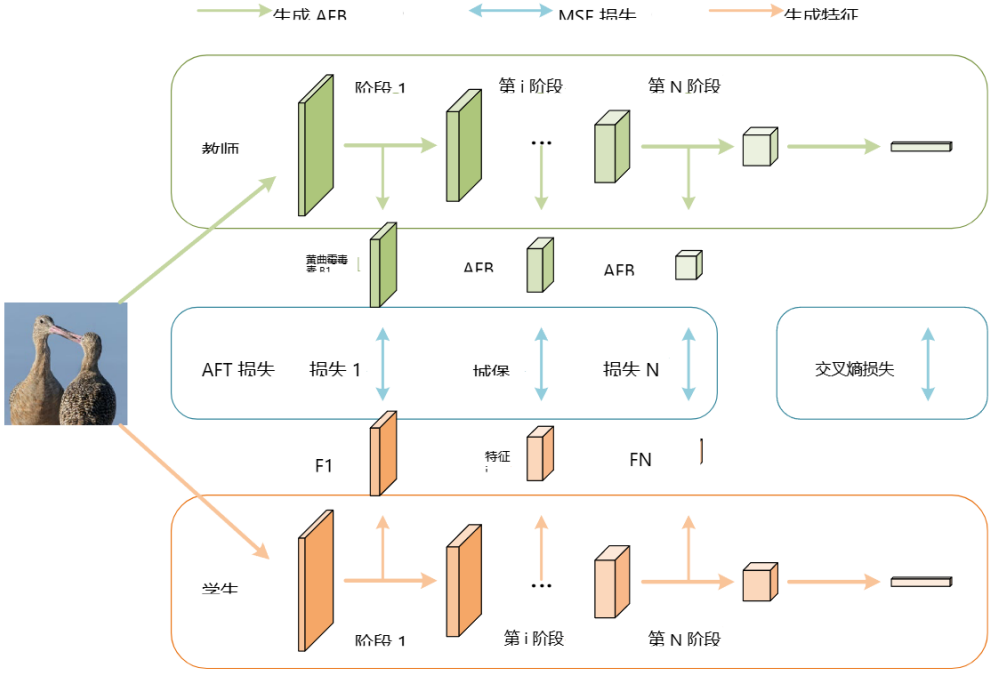

图 1. AFT-KD 示意图。为了对齐教师网络和学生网络，我们将其划分为 N 个推理阶段。N 的值在不同的师生结构中有所不同。例如，在 “AFB 的探索” 部分的实验中，N 等于 3。

同时将中间特征及其相应的注意力尝试转移到学生网络，以获得更好的性能。

我们的工作主要包括两部分。第一部分是关于如何获取推理信息。首先，我们发现了文献 19 中的可扩展性，即通过将输出层的权重投影回卷积特征图来生成类激活图（CAM）的操作可以轻松扩展到其他卷积层。受文献 30 的启发，我们使用 1×1 点卷积生成与中间特征图对应的注意力图，并在二值化后将其叠加到原始特征图上（图 2），我们将其称为注意力与特征块（AFB）。然后，根据卷积神经网络（CNN）的结构，我们将其划分为不同的阶段，以模拟不同的推理时刻，并迫使学生网络逼近所有阶段的 AFB。这样做的好处是学生能够学习完整的推理过程，并且块操作减少了所需的计算量。第二部分是关于如何平衡损失函数。我们将逼近 AFB 所产生的误差称为知识蒸馏损失（KD Loss），将预测输出与真实标签之间的误差称为交叉熵损失（CE Loss）。为了平衡 KD Loss 和 CE Loss 之间的优化速率，并防止在其中一个损失收敛后继续训练导致精度损失，我们设计了一种自适应损失函数，利用两种损失衰减率与预期衰减率的比值来调整损失权重。

总体而言，我们的贡献总结如下：

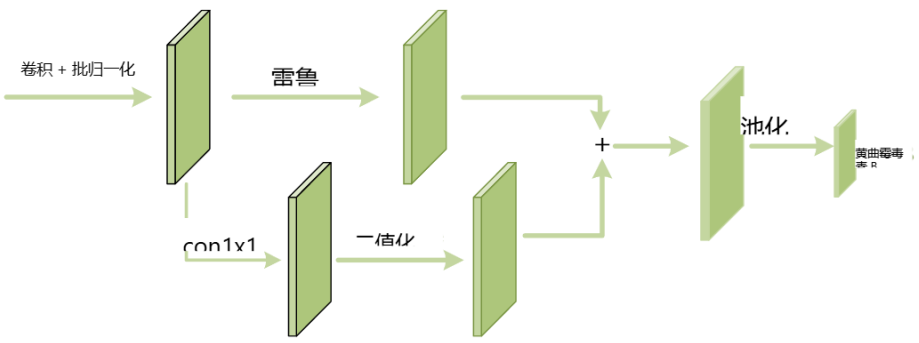

图 2. AFB 生成方法示意图。

- 我们将现有的知识蒸馏方法重新划分为基于推理过程的方法和基于推理结果的方法，揭示了现有知识蒸馏方法之间的关系，并为知识蒸馏研究提供了一个新视角。
- 我们设计了一种高效的策略，用于平衡蒸馏损失和交叉熵损失之间的最优比例，并且该策略易于扩展到多任务学习场景。
- 我们提出了一种基于注意力和特征迁移的知识蒸馏方法，名为 AFT-KD，该方法在多个基准测试中实现了最先进的性能。

最后，我们共使用五章来安排文章内容。第二章按照传统分类方法介绍了各种知识蒸馏方法的相关工作，并分析了不同方法之间的联系与不足。第三章详细介绍了所提出的方法，包括 CAM 回顾、AFT-KD 理论分析以及自适应损失的实现。第四章包含所有实验内容。首先，我们介绍了实验中使用的数据集，然后分析了 AFB 所包含的信息对蒸馏性能的影响，并进一步验证了通过 AFB 学习的 AFB 方法在性能上的优越性。最后，我们对自适应损失的实际性能进行了验证与分析。在最后一章中，我们总结了所提出的方法，并分析了其局限性以及未来的工作方向。

相关工作

知识蒸馏（KD）的概念由 Hinton 等人提出 21，它迫使学生网络从教师提供的软标签和真实标签中提取知识。为了充分利用软标签中包含的“暗知识”，引入了温度的概念。现有的 KD 方法主要可分为三类：基于逻辑的 20,21,31–34、基于特征的 18,22–29,35 以及基于注意力图的 19,30。

逻辑蒸馏将教师模型输出逻辑中隐含的知识转移到学生网络中。RAN^{32}通过使用与教师模型相同的参数化网络，获得了优于教师模型的性能。DKD20 将 KD 损失重新表述为目标类知识蒸馏（TCKD）和非目标类知识蒸馏（NCKD），揭示了 KD 的耦合公式限制了知识转移的有效性和灵活性。CrossKD34 将学生网络的中间特征传递到教师的检测头，产生交叉预测，然后强制这些预测模仿教师的预测。此外，有几篇关于逻辑蒸馏方法的文章 21,33,34。

基于特征的知识蒸馏方法往往具有更好的性能，它迫使学生网络从教师网络的中间特征中提取有效内容，但其代价是比逻辑蒸馏需要更多计算量。RKD^{25}能够转换数据样本之间的关系，以惩罚教师网络和学生网络在相关性上的差异。类似于教师网络和学生网络之间样本相关性研究的迁移 26,27，PKT^{35}对其进行建模。

教师的知识被视为一种概率分布，并使用 KL 散度来衡量距离。RKD25 利用多案例关系指导学生学习。CRD^{22}将对比学习与知识蒸馏相结合，并使用对比目标进行知识转移。ReviewKD18 采用跨层连接路径，整合不同层次特征所蕴含的知识。

基于注意力图的知识蒸馏（KD）方法能指导学生网络在推理时应关注哪些信息。AT^{30}验证了转移注意力图的有效性，该方法利用类激活图向学生网络传递知识。CAT-KD19 揭示，区分类别区域的能力是网络分类的关键，并证明这种能力可通过传递类激活图（CAM）来获得和增强。CAT-KD 能够通过转换结构来传递知识以获取注意力，这使得基于注意力的知识蒸馏具备良好的竞争力。

基于迁移逻辑和特征的知识蒸馏（KD）方法性能良好，而基于迁移注意力图的知识蒸馏方法具有较高的可解释性。以往的研究忽略了这两个特征之间的联系。本文旨在通过提出一种基于注意力和特征迁移的知识蒸馏方法来解决这一问题，该方法在多个基准测试中表现优异。

我们的方法

在本节中，我们首先回顾了 CAM 并分析了其可扩展性，然后进一步提出了注意力 - 特征融合模块（AFB）并将其应用于知识蒸馏，最后提出了结合知识蒸馏损失（KD Loss）和交叉熵损失（CE Loss）衰减率的自适应损失函数。

回顾类别激活图

首先，我们考虑一种常用的 CNN 结构，其最后一个卷积层的输出特征为$A \in \mathbb{R}^{C\times H\times W}$。

tional 层，其中 C 表示通道数，H 和 W 分别表示特征的高度和宽度。$A_i \in \mathbb{R}^{H\times W}$表示第 i 个通道的特征。$A_i(x,y)$表示通道 i 上空间位置（x,y）$(x,y)$处的激活值。此时，常规 CNN 生成预测结果的过程可表示如下。

$$
\begin{aligned}
P_j &= \sum_{1\le i\le C} W_i^j GAP(A_i) \\
&= \frac{1}{W\times H}\sum_{x,y}\sum_{1<i<C} W_i^j A_i(x,y)
\end{aligned}
$$

其中P_i表示第 j 类的逻辑预测，w_i^j表示全连接层中与第 j 类对应的权重。In^{30}，类激活图（CAM）是通过将 CNN 中全连接层的权重投影回卷积特征图而生成的，因此我们可以得到第 j 类对应的 CAM 计算公式：

$$CAM_i = \sum_{1<i<C} W_i^j A_i (2)$$

公式（2）也可以重写为：

$$CAM_i(x,y) = \sum_{1<i<C} W_i^j A_i(x,y)$$

根据式（1）和式（3），逻辑输出P_i与逻辑输出 CAMj 之间的关系如下：

$$\begin{aligned} P_j &= \frac{1}{W \times H} \sum_{x,y} CAM_j(x,y) \\ &= GAP\left(CAM_i(x,y)\right) \end{aligned}$$

如公式（4）所示，类别 j 的逻辑输出可通过计算对应类别的 CAM 平均值得到。这种通过特征映射获取进一步推理结果的操作在卷积神经网络（CNN）中十分常见。例如，卷积核生成的特征图经过正则化和激活处理后，可得到最终结论。因此，我们能够轻松地将 CAM 计算过程推广到 CNN 中的其他位置，生成的类 CAM 图与该层的输出特征图相对应。

AFT-KD

分别为输入特征的数量、高度和宽度，那么$k+1$卷积层的输入特征是：为了获取与推理过程相关的知识，我们将 CAM 扩展到 CNN 的所有卷积层。首先，其中$ACT_k(a)$、$Norm_k(a)$、$Conv_k(a)$分别代表全连接层的激活函数、正则化函数和卷积权重，那么公式 3 可以重写为：第 k 层的操作。Pre_F_k代表预激活输出特征。我们将预激活输出特征的激活过程近似为从推理过程到推理结果的映射。为了提取F_k中关于推理过程的知识，我们使用预激活输出特征 Pr e_F_k来计算与F_k对应的类 CAM。受此启发，我们使用点卷积来替代，其中 convj (a)代表与输出特征对应的第 j 个通道的卷积核。结合式（6），与F_k对应的注意力图可以表示为：其中CAM_k^L代表与F_k对应的 CAM 类别。$Bin(a)$代表二值化，使用二值化操作可以进一步突出卷积在推理时所关注的内容。最后，我们将CAM_k^L和F_k叠加，得到融合了推理过程和推理结果的输出特征，称为注意力与特征块（AFB）。第 k 个卷积层的注意力与特征块的计算过程如下：

$$\begin{aligned} F_k &= ACT_k\left(Norm_k\left(Conv_k\left(F_k-1\right)\right)\right) \\ &= ACT_k\left(Pre\_F_k\right) \end{aligned}$$

$$CAM_i = conv_i(A) (6)$$

$$CAM_k^L = Bin(conv_k(\mathrm{Pr}\, e\_F_k)) \backslash \mathrm{eqno}(7)$$

$$AFB_k = CAM_k^L + F_k (8)$$

在此基础上，我们提出了 AFT-KD，它迫使学生模拟教师传递的 AFB 中的知识。为了节省计算量并对齐学生网络和教师网络，我们将 CNN 划分为 N 个推理阶段，其中 AFT 损失可以定义为：其中，C_k表示调整后 k 阶段输出特征图的通道数。$\phi(a)$是用于统一AFB_k和F_k的通道数和分辨率的调整函数。F_k是 k 阶段学生网络的输出特征图。

$$\begin{aligned} L_{AFT} &= \sum_{1\le k\le N} L_{AFT}^k \\ &= \sum_{1\le k\le N} \sum_{1\le j\le C_k} \frac{1}{C_k} \left\| \frac{\phi\left(AFB_k^j\right)}{\phi\left(AFB_k^j\right)_2} - \frac{\phi\left(F_k^j\right)}{\phi\left(F_k^j\right)_2} \right\|_2^2 \end{aligned}$$

自适应损失

为了平衡不同损失之间的优化率，我们设计了一种基于损失优化率的自适应损失函数。在确定 AFT 损失后，我们将 AFT-KD 的总损失定义为：其中L_{CE}代表标准交叉熵损失，α、β 是用于平衡L_{CE}和L_{AFT}的动态超参数。当L_{CE}和L_{AFT}的优化率不同时，训练中会出现某一损失过早收敛的现象。在这种情况下，继续训练网络可以降低未收敛的损失函数，但会以牺牲已收敛任务的精度为代价。为了避免这种现象，我们引入损失优化率来平衡这两个损失函数的衰减率。

$$L_{KD} = \alpha L_{CE} + \beta L_{AFT} \quad (10)$$

首先，我们需要在训练开始时记录初始损失L^0_{CE}和L^0_{AFT}，并在任意迭代中记录当前损失L_{CE}和L_{AFT}。此时，我们可以计算该迭代中的损失衰减率：其中Dr_{CE}和Dr_{AFT}分别是本轮迭代中L_{CE}和L_{AFT}相对于初始损失的衰减率。平均衰减率$\bar{Dr} = (Dr_CE + Dr_AFT)/2$.α、β 可表示为：

$$\begin{cases} Dr\_CE = L_{CE}/L^0_{CE} \\ Dr\_AFT = L_{AFT}/L^0_{AFT} \end{cases}$$

$$\begin{cases} \alpha = Dr\_CE/\bar{Dr} \\ \beta = Dr\_AFT/\bar{Dr} \end{cases}$$

根据公式（11 和 12），对于优化速度更快的任务，其本轮的动态系数是一个小于 1 的正数，且优化速度越快，该系数越小，这会降低其本轮的优化效率，从而达到平衡另一项任务的目的。实验表明，自适应损失能够有效缩短Dr_{CE}与Dr_{AFT}之间的距离。

实验

数据集与实现细节

数据集

我们的实验主要在两个图像分类数据集上进行：

CIFAR-10036 包含 100 个类别的共 60,000 张 32×32 像素图片，其中训练集和验证集分别包含 50,000 张和 10,000 张图片

ImageNet37 是一个大规模数据集，包含 1000 个分类对象，其中有 120 万张训练图像和 5 万张验证图像。

实现细节

我们在 CIFAR100 和 ImageNet 上的实验设置严格遵循文献 19、20。在 CIFAR100 的实验中，我们训练了

使用 SGD 优化器，批量大小设为 64，进行 240 个轮次的训练，初始学习率为 0.05（对于 为 0.01）

ShuffleNet38,39 和 MobileNet17）在 150、180 和 210 次迭代时除以 10。关于 Ima 的实验

在 geNet 中，我们以 512 的批次大小训练了 100 个轮次。初始学习率为 0.2，每 30 个轮次衰减至原来的十分之一。此外，我们在多种具有代表性的 CNN 网络上进行了实验：VGG40、ResNet41、WideResNet42、MobileNet17 以及 ShuffleNet38.39。表 1 对这些网络进行了简要概述。

为了确保实验结果的公平性，现有方法的结果要么被报告

在他们的论文 18–20 中，或者使用他们提供的代码并采用完全相同的设置获得。所有结果

CIFAR100 的结果是 5 次试验的平均值，而 ImageNet 的结果是 3 次试验的平均值

AFB 的探索

在本节中，我们首先探讨了迁移 AFB 的有效性，然后提出学习连续推理知识对学生有更大益处。

VGG[40]	VGG is constructed by stacking convolutional layers and connecting fully connected layers, and uses the relu activation function in a unified way. Common models include VGG16/VGG19
ResNet[41]	ResNet introduced residual structure to alleviate the problem of gradient disappearance/explosion, commonly used structure Resnet 34/50/101 and so on
WideResNet[42]	On the basis of ResNet, the network width is increased to further improve the training speed of the network. Common structures include WRN16-2/WRN40-2, etc
MobileNet[17]	MobileNet uses deep separable convolution instead of normal convolution to make the model more lightweight. Common structures include MobileNetv2/v3
ShuffleNet[38,39]	ShuffleNet uses point-by-point grouping convolution and channel rearrangement to reduce the model size. The common structure is ShuffleNetv1/v2

表 1. 简要介绍了几种常见的神经网络。

AFB 包含比 CAM 和输出特征更完整的推理信息

由于推理过程与推理结果密切相关，教师仅传递与推理结果相关的特征可能会导致学生模仿错误的推理过程。AFB 包含完整的推理信息，理论上能为学生带来益处，这种益处可以通过学生在分类任务中的表现直接观察到

我们将不同的信息从 ResNet32×4 迁移到 ResNet8×4，包括（1）仅输出特征、（2）仅 CAM 以及（3）AFB，并在 CIFAR-100 上观察其性能。为了对齐教师网络和学生网络，我们根据 ResNet 的层分组将推理过程分为三个阶段。如表 2 所示，单独迁移 CAM 或输出特征也能达到较好的分类准确率，但迁移 AFB 能进一步提升性能，这表明 AFB 比 CAM 或输出特征承载更完整的推理知识，且这种知识直接有助于学生网络性能的提升。此外，将 AFB 迁移到不同位置会给子网络的性能带来不同程度的提升。

迁移连续的 AFB 对学生更有利

数据在网络中的前向传播具有时间连续性。从输入图片到输出预测逻辑的过程可以分为不同阶段，前一阶段的推理结果会作为后一阶段的推理输入。因此，相邻阶段的推理信息（包括推理过程和推理结果）也密切相关。我们通过实验证明，向学生提供连续且完整的 AFB 能带来更显著的成绩提升。

同样，我们使用 ResNet32×4 作为 AFB 生成器，并使用 ResNet8×4 来学习 AFB 的不同阶段。我们将实验分为三组：（1）仅研究由 Layer1 生成的 AFB；（2）学习由 Layer1-2 生成的 AFB；（3）学习由所有三层生成的 AFB。Layer1-3 对应图 1 中的 Stage1-3，其中 N 等于 3。如表 3 所示，仅学习部分推理知识具有相近的分类精度（高于基准网络），但学习完整的 AFB 能显著提升网络性能。

AFT-KD 的评估

在本节中，我们将 AFT-KD 与几种主流的知识蒸馏方法进行比较，包括特征

基于知识蒸馏的方法 18,22–25、基于逻辑的知识蒸馏方法 20,21 以及基于注意力图的知识蒸馏方法 19,30。

CIFAR-100 上的结果

与文献 19、20 的研究类似，我们在 CIFAR-100 上使用相同的师生框架进行了实验。

以及不同的师生框架，并将结果分别报告在表 4 和表 5 中。我们的方法在相同的师生架构以及一些不同师生架构的实验中都实现了新的最优结果。其中，与基于逻辑的方法相比，我们的方法在所有不同架构的实验中都取得了更好的结果。然而，当学生模型

是 MobileNetV2，AFT-KD 的性能略低于文献 18、19，我们推测这是由于较大的

师生架构与我们所使用的简单对齐方式之间的差异。在相同架构的实验中，通过我们的方法训练的学生模型的准确率超过了教师模型。特别是当教师模型为 ResNet32×4 时，AFT-KD 的准确率达到 77.14%，比教师模型高出 4.64%。

ImageNet 上的结果

表 6 和表 7 给出了在ImageNet上图像分类的 top-1 和 top-5 准确率。由于教师模型的能力限制，我们的方法并未达到最佳性能，但仍优于大多数知识蒸馏（KD）方法

最后，我们在 CIFAR-100 上比较了几种 SOTA 方法的性能，其中训练集以不同比例衰减，这与文献 18 中的实践一致，通过这种方式评估它们对训练数据量的依赖性。如图 3 所示，AFT-KD 受训练数据量的影响最小，这表明我们的方法具有出色的蒸馏效率。

Knowledge	Feature	CAMs	AFB
Acc	75.34	73.89	77.14

表 2. 使用不同知识在 cifar-100 上训练的 ResNet8×4 的准确率（%）。迁移知识由 ResNet32×4 生成

AFB Producer	Layer-1	Layer1-2	Layer1-3
Acc	75.47	75.68	77.14

表 3. 使用不同层知识在 cifar100 上训练的 ResNet8×4 的准确率（%）。

Distillation manner	Teacher	ResNet32×4	WRN40-2	ResNet32×4	ResNet50	VGG13
	Acc	79.42	75.61	79.42	79.34	74.64
	Student	ShuffleNetV1	ShuffleNetV1	ShuffleNetV2	MobileNetV2	MobileNetV2
	Acc	70.5	70.5	71.82	64.6	64.6
Logits	KD[21]	74.07	74.83	74.45	67.35	67.37
	DKD[20]	76.45	76.7	77.07	70.35	69.71
Features	CRD[22]	75.11	76.05	75.65	69.11	69.73
	OFD[23]	75.98	75.85	76.82	69.04	69.48
	FitNet[24]	73.59	73.73	73.54	63.16	64.14
	RKD[25]	72.28	72.21	73.21	64.43	64.52
	ReviewKD[18]	77.45	77.14	77.78	69.89	**70.37**
Attention	AT[30]	71.73	73.32	72.73	58.58	59.4
	CAT-KD[19]	78.26	77.35	78.41	**71.36**	69.13
Ours	AFT-KD	**78.39**	**77.49**	**78.47**	70.48	69.22

表 4. CIFAR-100 上的结果。教师模型和学生模型具有不同的架构。显著值在 [粗体] 中。

Distillation manner	Teacher	ResNet56	ResNet110	ResNet32×4	WRN-40–2	WRN-40–2	VGG13
	Acc	72.34	74.31	79.42	75.61	75.61	74.64
	Student	ResNet20	ResNet32	ResNet8×4	WRN-16–2	WRN-40–1	VGG8
	Acc	69.06	71.14	72.5	73.26	71.98	70.36
Logits	KD[21]	70.66	73.08	73.33	74.92	73.54	72.98
	DKD[20]	**71.97**	74.11	76.32	**76.24**	74.81	74.68
Features	CRD[22]	71.16	73.48	75.51	75.48	74.14	73.94
	OFD[23]	70.98	73.23	74.95	75.24	74.33	73.95
	FitNet[24]	69.21	71.06	73.5	73.58	72.24	71.02
	RKD[25]	69.61	71.82	71.9	73.35	72.22	71.48
	ReviewKD[18]	71.89	73.89	75.63	76.12	**75.09**	**74.84**
Attention	AT[30]	70.55	72.31	73.44	74.08	72.77	71.43
	CAT-KD[19]	71.62	73.62	76.91	75.6	74.82	74.65
Ours	AFT-KD	71.92	**74.23**	**77.14**	75.86	74.68	74.7

表 5. CIFAR-100 上的结果。教师模型和学生模型具有相同的架构。显著值用 [粗体] 表示。

Distillation manner			Features			Logits		Attention		Ours
Acc	Teacher	Student	OFD	CRD	ReviewKD	KD	DKD	AT	CAT-KD	AFT-KD
Top-1	73.31	69.75	70.81	71.17	71.61	70.66	**71.7**	70.69	71.26	71.47
Top-5	91.41	89.07	89.98	90.13	90.51	89.88	90.41	90.01	90.45	**90.61**

表 6. ImageNet 上的结果。我们将 ResNet34 设为教师模型，ResNet18 设为学生模型。显著值在 [粗体] 中。

Distillation manner			Features			Logits		Attention		Ours
Acc	Teacher	Student	OFD	CRD	ReviewKD	KD	DKD	AT	CAT-KD	AFT-KD
Top-1	76.16	68.87	71.25	71.37	**72.56**	68.58	72.05	69.56	72.24	72.44
Top-5	92.86	88.76	90.34	90.41	91.00	88.98	91.05	89.33	91.13	**91.35**

表 7. ImageNet 上的结果。我们将 ResNet50 设为教师模型，MobileNet 设为学生模型。显著值在 [粗体] 中。

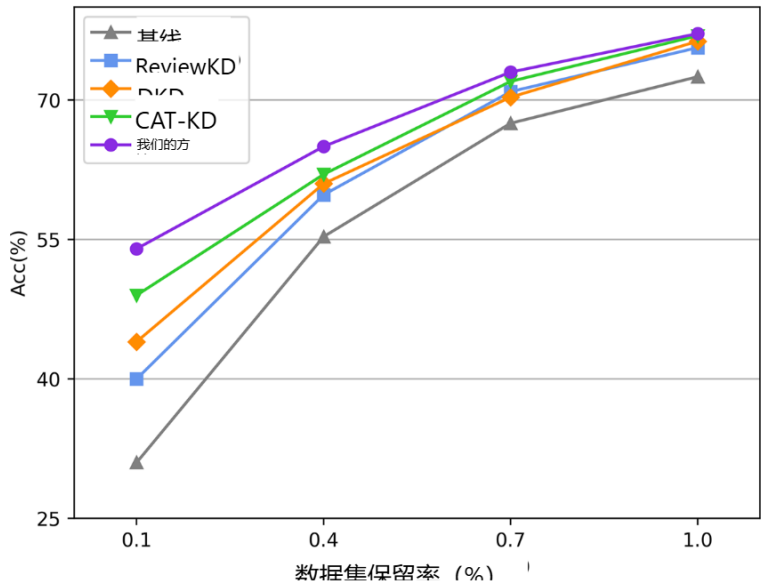

图 3. 学生模型在 CIFAR-100 上使用多种 SOTA 方法训练后的准确率（%）。我们将 ResNet32×4 设为教师模型，ResNet8×4 设为学生模型，并以不同比例缩减训练集。

自适应损失的探索

从“自适应损失”部分的分析可以看出，多任务学习中损失函数的优化速率不平衡会导致训练后期准确率下降的问题。自适应损失函数会根据 AFT 损失和交叉熵损失的优化速率自动调整损失值，有效缓解了上述问题。这种改进可以通过多个任务的损失衰减速率曲线之间的距离来衡量，图 4 展示了采用自适应损失函数前后优化速率曲线的对比。我们对优化速率的原始数据进行了采样并计算均值，得到了平滑的对比曲线。可以观察到，使用自适应损失函数平衡两种损失的优化速率后，损失衰减曲线之间的距离减小，且两种损失的优化更加同步。这一改进也有效提升了模型的分类性能。

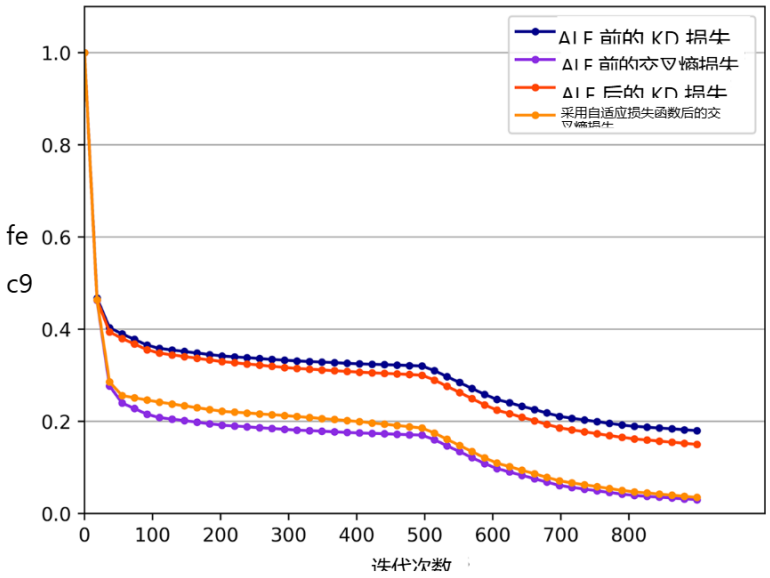

图 4. 使用自适应损失函数前后的优化率曲线对比。

结论

我们对现有的知识蒸馏（KD）方法进行了分析，并将这些方法重新归类为基于推理过程的方法和基于推理结果的方法，这为知识蒸馏研究提供了一个新的视角。在此基础上，我们提出了带有注意力和特征迁移的 AFT-KD，该方法在多个常用基准测试上取得了具有竞争力的结果。最后，为了平衡 AFT-KD 中的损失优化率，我们提出了一种基于损失衰减率的自适应损失函数，以进一步提升 AFT-KD 的性能。然而，在教师模型和学生模型采用不同架构的实验中，AFT-KD 的性能并非所有方法中最优的，我们推测这是由于师生模型架构差异较大，且我们所使用的对齐方法过于简单所致。这一局限性可以通过为网络结构设计特定的对齐规则来改善。此外，与基于特征的知识蒸馏方法相比，AFT-KD 在运算成本方面带来了一定改善，这也是我们在未来工作中需要继续探索的内容。

数据可用性

请联系干帅英（邮箱：6720210550@mail.ixust.edu.cn）获取数据和代码。

收到日期：2023 年 7 月 23 日；接受日期：2023 年 10 月 1 日
在线发表时间：2023 年 10 月 26 日

参考文献

刘，Y. 等。面向旅游企业连续兴趣点推荐的交互增强与时序感知图卷积网络。《IEEE 工业信息汇刊》19 (1)，635–643（2022）。

Grabek, J. 与 Cyganek, B. 基于张量的数据压缩方法对深度神经网络精度的影响。《计算机科学与信息系统年报》26，3-11（2021）。

Hameed, M. G. A. 等。通过广义克罗内克积分解压缩卷积神经网络. 《AAAI 人工智能大会论文集》36 (1), 771–779 (2022).

Hua, W. 等。通道门控神经网络。《神经信息处理系统进展》32, 1 (2019).

古萨克（J.）、霍利亚夫琴科（M.）、波诺马廖夫（E.）等。《神经网络的自动化多阶段压缩》。收录于《IEEE/CVF 国际计算机视觉会议研讨会论文集》（2019 年）

Phan A. H.、Sobolev, K.、Sozykin, K. 等人。用于压缩卷积神经网络的稳定低秩张量分解。见《计算机视觉 ——ECCV 2020：第 16 届欧洲会议，英国格拉斯哥，2020 年 8 月 23-28 日，论文集，第 XXIX 卷》16 522-539（施普林格国际出版社，2020 年）。

林 M、季 R、王 Y 等。Hrank：使用高秩特征图的滤波器剪枝。见《IEEE/CVF 计算机视觉与模式识别会议论文集》，1529–1538（2020）。

侯子（音译）、秦敏（音译）、孙飞（音译）等。Chex：用于 CNN 模型压缩的通道探索。见《IEEE/CVF 计算机视觉与模式识别会议论文集》，第 12287–12298 页（2022 年）。

方，马，宋等人。Depgraph：迈向任意结构剪枝。arXiv 预印本 arXiv:2301.12900（2023）

任安、张涛、叶帅等人。《ADMM-NN：一种基于交替方向乘子法的深度神经网络算法 - 硬件协同设计框架》。发表于《第 24 届编程语言和操作系统架构支持国际会议论文集》，第 925-938 页（2019 年）。

罗俊辉等。Thinet：剪枝 CNN 滤波器以构建更精简的网络。《IEEE 模式分析与机器智能汇刊》41（10），2525-2538（2018）。

蔡宇、姚哲伟、董震宇等。ZeroQ：一种新颖的零样本量化框架。《IEEE/CVF 计算机视觉与模式识别会议论文集》，13169–13178 页，2020 年。

徐思（音译）、李华（音译）、庄兵（音译）等。生成式低位宽无数据量化。见《计算机视觉 ——ECCV 2020：第 16 届欧洲会议，英国格拉斯哥，2020 年 8 月 23-28 日，论文集，第 XII 部分 16》，第 1-17 页（施普林格出版社，2020 年）。

李华、吴霞、吕峰等。难样本在零样本量化中至关重要。《IEEE/CVF 计算机视觉与模式识别会议论文集》，24417–24426（2023）。

霍华德（A.）、桑德勒（M.）、朱（G.）等。《MobileNetV3 探索》。收录于《IEEE/CVF 国际计算机视觉会议论文集》，第 1314–1324 页（2019 年）。

Howard, A. G., Zhu, M., Chen, B., 等. Mobilenets：适用于移动视觉应用的高效卷积神经网络. arXiv 预印本 arXiv:1704.04861 (2017).

Sandler, M., Howard, A., Zhu, M., 等. MobileNetV2：倒置残差与线性瓶颈。《IEEE 计算机视觉与模式识别会议论文集》，4510–4520（2018）。

陈鹏、刘思、赵华等。通过知识回顾提炼知识。《IEEE/CVF 计算机视觉与模式识别会议论文集》，第 5008-5017 页（2021 年）。

郭志、闫浩、李浩等。基于类别注意力迁移的知识蒸馏。《IEEE/CVF 计算机视觉与模式识别会议论文集》，第 11868-11877 页（2023）。

赵博、崔强、宋然等。解耦知识蒸馏。《IEEE/CVF 计算机视觉与模式识别会议论文集》，11953–11962（2022）。

辛顿（Hinton）、维尼尔斯（Vinyals）、迪恩（Dean）。《提炼神经网络中的知识》。arXiv 预印本 arXiv:1503.02531（2015）。

田宇、克里希南（D.）、伊索拉（P.）。对比表征蒸馏。arXiv 预印本 arXiv:1910.10699（2019）

Heo, B.、Kim, J.、Yun, S. 等人。《特征蒸馏的全面改进》。收录于《IEEE/CVF 国际计算机视觉会议论文集》，第 1921–1930 页（2019 年）

罗梅罗（A. Romero）、巴拉斯（N. Ballas）、卡胡（S. E. Kahou）等。Fitnets：瘦深网络的提示。arXiv 预印本 arXiv:1412.6550（2014）。

Park, W., Kim, D., Lu, Y., 等。关系知识蒸馏。见：IEEE/CVF 计算机视觉与模式识别会议论文集，3967–3976 (2019).

彭博、金鑫、刘军等。知识蒸馏的相关性一致性。《IEEE/CVF 国际计算机视觉会议论文集》，第 5007-5016 页（2019 年）。

Tung, F., & Mori, G. 保相似性知识蒸馏。《IEEE/CVF 国际计算机视觉会议论文集》，1365-1374（2019）。

吉（Ji）、申（Shin）、黄（Hwang）等人。通过自我教学来完善自我：基于自知识蒸馏的特征优化。《IEEE/CVF 计算机视觉与模式识别会议论文集》，10664–10673（2021）

陈磊、王丹、甘臻等。《瓦瑟斯坦对比表示蒸馏》。见《IEEE/CVF 计算机视觉与模式识别会议论文集》，第 16296–16305 页（2021 年）

Komodakis, N.，& Zagoruyko, S. 更多关注注意力：通过注意力迁移提升卷积神经网络的性能。国际学习表征大会（ICLR）。2017 年。

Cho, J. H., 与 Hariharan, B. 论知识蒸馏的有效性。见《IEEE/CVF 国际计算机视觉会议论文集》，第 4794–4802 页（2019 年）。

Furlanello, T., Lipton, Z., Tschannen, M., 等。重生神经网络。国际机器学习会议. PMLR, 1607–1616 (2018).

Mirzadeh, S. I. 等。通过教师助手改进知识蒸馏。《AAAI 人工智能大会论文集》34 (04)，5191-5198（2020）。

王、陈、郑等人。CrossKD：用于密集目标检测的跨头知识蒸馏。arXiv 预印本 arXiv:2306.11369（2023）。

帕萨利斯，N. 与特法斯，A. 用于深度表征学习的概率知识迁移。

克里热夫斯基（Krizhevsky, A.）和辛顿（Hinton, G.）。从微小图像中学习多层特征。《自身免疫性疾病系统手册》1（4），1（2009）。

马楠、张晓、郑海涛等。Shufflenet v2：高效 CNN 架构设计的实用指南。《欧洲计算机视觉会议论文集》（ECCV），116–131（2018）。

张，X.，周，X.，林，M.，等。Shufflenet：一种适用于移动设备的高效卷积神经网络。《IEEE 计算机视觉与模式识别会议论文集》，6848–6856（2018）。

Simonyan, K., & Zisserman, A. 用于大规模图像识别的非常深的卷积网络。arXiv 预印本 arXiv:1409.1556（2014）。

何凯明、张祥雨、任少卿等。用于图像识别的深度残差学习。《IEEE 计算机视觉与模式识别会议论文集》，第 770-778 页（2016 年）。

Zagoruyko, S.，& Komodakis, N. 宽残差网络。arXiv 预印本 arXiv:1605.07146（2016）。

作者贡献

方法与概念构思：G.Y.；初稿撰写：G.Y. 和 S.Y.；审阅与编辑：G.Y.、S.Y. 和 Y.S.；软件与验证：S.Y.、Y.S. 和 H.Y.；数据整理与项目管理：Y.S. 和 H.Y.。所有作者均已阅读并同意发表本稿件的最终版本。

资金

本研究得到江西省教育厅科技项目（编号：GJJ190450）和江西省教育厅科技项目（编号：GJJ180484）的资助。

利益冲突

作者声明不存在利益冲突

补充信息

通信和材料请求应寄给 S.Y.

重印和许可信息可在www.nature.com/reprints获取