计算机科学
COMPUTER SCIENCE

基于多阶段多生成对抗网络的互学习知识蒸馏方法

黄仲浩，杨兴耀，于炯，郭亮，李想

引用本文

黄仲浩，杨兴耀，于炯，郭亮，李想. 基于多阶段多生成对抗网络的互学习知识蒸馏方法[J]. 计算机科学, 2022, 49(10): 169-175.

HUANG Zhong-hao, YANG Xing-yao, YU Jiong, GUO Liang, LI Xiang. Mutual Learning Knowledge Distillation Based on Multi-stage Multi-generative Adversarial Network[J]. Computer Science, 2022, 49(10): 169-175.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210800250

基于多阶段多生成对抗网络的互学习知识蒸馏方法

黄仲浩　杨兴耀　于　炯　郭　亮　李　想
新疆大学软件学院　乌鲁木齐 830008
(hzhao@stu.xju.edu.cn)

摘　要　针对传统的知识蒸馏方法在图像分类任务中对知识蒸馏的效率不高、阶段训练方式单一、训练过程复杂且难收敛的问题,设计了一种基于多阶段多生成对抗网络(MS-MGANs)的互学习知识蒸馏方法。首先,将整个训练过程划分为多个阶段,得到不同阶段的老师模型,用于逐步指导学生模型,获得更好的精度效果;其次,引入逐层贪婪策略取代传统的端到端训练模式,通过基于卷积块的逐层训练来减少每阶段迭代过程中需优化的参数量,进一步提高模型蒸馏效率;最后,在知识蒸馏框架中引入生成对抗结构,使用老师模型作为特征辨别器,使用学生模型作为特征生成器,促使学生模型在不断模仿老师模型的过程中更好地接近甚至超越老师模型的性能。在多个公开的图像分类数据集上对所提方法和其他流行的知识蒸馏方法进行对比实验,实验结果表明所提知识蒸馏方法具有更好的图像分类性能。

关键词:互学习知识蒸馏;逐层贪婪策略;生成对抗网络;模型压缩;图像分类
中图法分类号　TP391

Mutual Learning Knowledge Distillation Based on Multi-stage Multi-generative Adversarial Network

HUANG Zhong-hao, YANG Xing-yao, YU Jiong, GUO Liang and LI Xiang
School of Software, Xinjiang University, Urumqi 830008, China

Abstract　Aiming at the problems of insufficient knowledge distillation efficiency, single stage training methods, complex training processes and difficult convergence of traditional knowledge distillation methods in image classification tasks, this paper designs a mutual learning knowledge distillation based on multi-stage multi-generative adversarial networks (MS-MGANs). Firstly, the whole training process is divided into several stages, teacher models of different stages are obtained to guide student models to achieve better accuracy. Secondly, the layer-wise greedy strategy is introduced to replace the traditional end-to-end training mode, and the layer-wise training strategy based on convolution block is adopted to reduce the number of parameters to be optimized in each iteration process, and further improve the distillation efficiency of the model. Finally, a generative adversarial structure is introduced into the knowledge distillation framework, with the teacher model as the feature discriminator and the student model as the feature generator, so that the student model can better follow or even surpass the performance of the teacher model in the process of continuously imitating the teacher model. The proposed method is compared with other advanced knowledge distillation methods on several public image classification data sets, and the experimental results show that the new knowledge distillation method has better performance in image classification.

Keywords　Mutual learning knowledge distillation, Layer-wise greedy strategy, Generative adversarial network, Model compression, Image classification

1　引言

近年来,深度学习技术在计算机视觉、自然语言处理、语音识别、推荐系统等领域中得到了广泛的应用[1-4]。深度学习的优异性依赖于设计更深和更宽的网络架构,可拥有多层和数百万的参数,同时借助相关技术如残差连接、批处理归一化等实现在强大的 GPU 或 TPU 集群上训练。然而,在计算和存储资源有限的平台上部署百万级参数模型几乎是不可能

到稿日期:2021-08-30　返修日期:2022-03-07
基金项目:国家自然科学基金(61862060,61966035,61562086);新疆维吾尔自治区教育厅项目(XJEDU2016S035);新疆大学博士科研启动基金项目(BS150257)
This work was supported by the National Natural Science Foundation of China(61862060,61966035,61562086),Education Department Project of Xinjiang Uygur Autonomous Region(XJEDU2016S035) and Doctoral Research Start-up Foundation of Xinjiang University(BS150257).
通信作者:杨兴耀(yangxy@xju.edu.cn)

的，如移动机器人，自动驾驶汽车等；同时，笨重的模型会因参数量过大导致推理时间较长，延迟高，无法执行相关实时任务。Han 等[5]的研究表明，对于一个完全收敛的笨重模型，大约有 85%以上的权值不会因消失而对模型的性能产生明显的影响，这也说明了笨重模型中含有大量的冗余信息。为了解决上述问题，许多研究者对深度神经网络（Deep Neural Networks，DNNs）结构展开了研究，这些研究主要分为 4 类：模型量化、模型剪枝、张量分解和知识蒸馏。

模型量化是通过减少用于表示预训练模型权值参数的比特值来压缩模型的大小；模型剪枝通过某一标准剪枝预训练模型中的冗余和非信息权值来降低模型的复杂度；张量分解是矩阵分解的高阶泛化，通过将高维度张量拆解为多个低维度张量来降低模型的计算量。然而，这些方法中的绝大部分研究都无法充分利用流行的深度学习框架和先进的 GPU 实现，且它们的加速需要特定的硬件支持。

知识蒸馏是一种从笨重的老师模型中提取知识来指导轻量的学生模型训练的方法。这种基于知识重用的方法减轻了学生模型训练和存储的负担，提高了目标模型的性能。Hinton 等[6]利用老师模型前一层输出的 logits 值作为软目标来指导学生模型训练；Romero 等[7]利用注意力机制将模型的中间层输出结果构造为注意图来减小老师-学生模型之间的差距；Ye 等[8]基于堆栈双 GAN 网络（Group-Stack Dual-GAN）来实现无数据的蒸馏过程。Guo 等[9]提出一种基于协作学习的在线知识蒸馏方法（Knowledge Distillation via Collaborative Learning，KDCL），该方法引入 Ensemble logits 来实现多个学生模型的相互学习。

然而，现有的知识蒸馏方法存在几点局限：

（1）老师模型的强收敛性。现有的大多数方法直接使用训练好的老师模型用于蒸馏学生模型，但由于学生-老师模型之间的性能差距过大，导致知识在传递的过程中会产生一定的损失。

（2）模型训练方式复杂且难以收敛。现有的大多数方法以增大计算复杂度为代价来提高目标模型的精度，然而实现过程复杂且难以收敛，例如，Ye 等[8]引入的 Group-Stack Dual-GAN 方法要不断对生成器和辨别器的相关参数和迭代次数进行调试，训练过程十分复杂。

（3）学生模型和老师模型之间的结构差异性。不同结构的学生-老师模型之间的网络结构、通道数量以及初始条件等方面都存在差异，老师模型直接给出相关输出结果而没有提供相关解释给学生模型。就好比一个老师给一个学生提出一个问题，老师只给出答案是不够的，还需要根据解题思路逐步引导学生解决问题，学生才更容易理解和提升。

基于上述问题，本文提出一种基于多阶段多生成对抗网络的互学习知识蒸馏方法（Mutual Learning Knowledge Distillation based on Multi-Stage Multi-Generative Adversarial Networks，MS-MGANs）。与传统的知识蒸馏方法相比，所提方法一方面结合了在线蒸馏和逐层贪婪思想，将老师模型训练划分为多个阶段，与学生模型共同学习，同时学生模型也分为多步逐块学习，老师逐步引导学生进行训练，以解决模型之间性能差距过大导致的性能损失问题；另一方面，结合生成对抗思想，使用学生模型的每个卷积块作为生成器，老师模型作为辨别器，将学生模型对应的输出结果输入老师的对应块中，实现学生模仿老师形成结果的思路过程，同时也解决了传统 GAN 难以训练且模型之间的结构差异问题。最后，运用互学习的训练方法，使学生模型和老师模型相互学习，进一步提高学生模型的训练精度，解决了学生模型的性能超过老师模型后无法再提升的问题。综上所述，本文的主要贡献如下：

（1）结合在线蒸馏和逐层贪婪方法，提出了一种多阶段学习策略用于解决老师模型强收敛性导致的蒸馏精度损失问题。

（2）运用生成对抗方法，结合逐层贪婪策略，构造出多阶段多生成对抗网络模型，解决传统 GAN 模型难以训练和梯度消失等问题。

（3）运用互学习的训练方法进一步提升学生和老师的性能。同时引入中间层损失、空间转换损失、软目标和硬目标损失，优化了模型在线蒸馏过程。

2 相关工作

2.1 课程学习

课程学习的基本思想是按照难度的升序组织样本或任务，除了基于先验知识设计难易课程外，还努力融入学习过程，对课程进行动态调整。Bengio 等[10]首次提出了课程学习概念，通过将实例进行有意义排序，有效提高了模型的训练效率。Pentina 等[11]认为，在多任务的学习过程中，不同任务之间学习的顺序将会严重影响模型的精度，因此基于泛化界准则来选择任务顺序，并按照任务顺序来训练模型。Zhang 等[12]通过将无标记领域的实例与目标域进行相似度比较，采用概率课程学习策略使更多相似的样本在训练过程中被更早、更频繁地使用。Guo 等[13]提出一种课程神经体系结构搜索方法（Curriculum Neural Architecture Search，CNAS），不同于传统的 NAS 算法需要从非常大的搜索空间中寻找最优候选框架，该算法从小搜索空间开始，逐步引导至大搜索空间，极大地提高了搜索效率。

2.2 生成对抗网络

博弈论亦称竞赛论或对策论，是研究具有斗争或竞争性质的数学理论和方法。Goodfellow 等[14]基于博弈论的思想首次提出了生成对抗网络（Generative Adversarial Network，GAN），它由生成网络和辨别网络构成，两者通过相互对抗来实现无数据的模型训练。许多研究者将 GAN 引入模型压缩领域以实现对模型的无数据压缩。Chen 等[15]引入生成器随机生成数据，使用老师-学生模型作为辨别器来实现模型的无数据蒸馏。Wang 等[16]提出了一种新的生成模型知识转移方法，该方法从单个或者多个 GAN 中提取知识到特定的目标领域。Li 等[17]提出了一种压缩 GAN 的方法，通过训练协议将老师-学生的结果结合起来，采用知识蒸馏和神经结构搜索方法降低训练的不稳定性，提高模型效率。Viazovetskyi 等[18]提出了一种提取 StyleGAN2 的特定图像到成对训练的 image-to-image 网络方法，从而实现了将 StyleGAN2 的特定

图像向单个 image-to-image 的平移。

2.3 模型加速

笨重模型的计算代价大且推理时间长，难以部署在计算资源有限的平台上。为此，一些研究者提出了各种方法来解决此问题。模型加速方法主要分为模型量化、模型剪枝、张量分解和知识蒸馏。在模型量化方面，Gong 等[19]提出了一种可微软量化方法（Differentiable Soft Quantization，DSQ），以弥补低比特网络和全精度网络之间的差距。Boo 等[20]提出了一种随机精度集成训练方法，该方法在每一层随机改变激活的位精度，从而获得软标签，用于训练学生模型。在模型剪枝方面，Guo 等[21]提出一种 DMCP 方法，该方法将通道剪枝的过程建模为一个马尔可夫模型，每一个状态表示在剪枝过程中应保留的通道数，状态之间的转换则是剪枝的过程。Lin 等[22]提出了一种 HRank 方法，此方法通过对包含较少信息的低秩特征图进行剪枝，从而实现模型压缩的目的。在张量分解方面，Tai 等[23]提出了一种低阶张量分解算法，通过寻找精确的全局优化器，从卷积核中提取冗余。Wu 等[24]提出将 Hierarchical Tucker（HT）引入神经网络压缩，实验证明 HT 在权值矩阵方面效果更好，而 Tensor-Train（TT）格式更适合用于卷积核压缩。

2.4 知识蒸馏

知识蒸馏首先由 Hinton 等[6]提出，其目的是从笨重的老师模型（下文简写为老师）中提取知识来训练一个轻量的学生模型（下文简写为学生），在深度模型压缩中有着广泛的应用。Romero 等[7]在预训练阶段引入了暗示层和引导层，使用老师的中间层作为暗示层来初始化学生的引导层参数，从而使学生达到更好的收敛效果。Yim 等[25]基于 Gramian 矩阵提出了一种 FSP 方法，通过学生-老师模型的 FSP 矩阵相似度实现知识的传递。Jin 等[26]认为学生很难充分吸收训练完成的老师的知识，因此提出了 anchor point 来将老师的训练过程分为多个阶段，逐步指导学生进行训练。Kulkarni 等[27]认为分块地训练学生能够达到更好、更快的收敛效果；Liu 等[28]认为传统的知识蒸馏过程往往忽略了对实例空间关系的计算，因此提出了实例关系图使学生在空间分布上近似于老师。Wang 等[29]提出了一种多出口结构协调密集知识蒸馏方法，鼓励每一层的出口都能灵活地从后面的所有出口进行学习。

本文提出了一种基于多阶段多生成对抗网络的互学习知识蒸馏方法，该方法引入课程学习思想将模型训练过程分为多个阶段，以解决老师模型的强收敛性对蒸馏过程的影响的问题。同时结合逐层贪婪策略，对学生-老师进行基于卷积块的多层次训练，有效减少了单一阶段需要优化的参数量。然后，结合生成对抗框架来促使学生不断模仿甚至超越老师的性能。最后，通过互学习策略使学生-老师的身份相互转换，促进学生-老师进一步地提升精度性能。

3 MS-MGANs 方法

采用多阶段多生成对抗框架训练良好的目标网络可分为两步：第一步，将老师划分为 M 个阶段，在每个阶段得到经过训练的老师；第二步，利用每个阶段获得的老师来引导学生进行逐块训练。其中学生的每个块被用作生成器，其输出被输入到相应的老师（辨别器）中产生“假”结果，并与辨别器输出的“真”结果进行最小化损失计算。因此，学生可以在假定自己是老师的过程中继续接近甚至超过老师的性能，总体框架如图 1 所示。

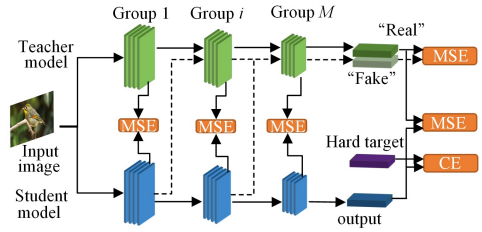

图 1 MS-MGANs 框架图

Fig. 1 Block diagram of MS-MGANs model

3.1 学生-老师的知识蒸馏机制

为了更好地说明，本文将 W_T 表示为老师的参数，将 W_S 表示为学生的参数。$P_T=\mathrm{Softmax}(Z_T)$，$P_S=\mathrm{Softmax}(Z_S)$ 分别代表老师和学生的预测输出，Z_T 和 Z_S 代表老师和学生输出的 logits 值。知识蒸馏（Knowledge Distillation，KD）的实质是以老师的输出为软目标来指导学生的训练。同时学生的输出也与硬标签 $\boldsymbol{y}$ 进行计算，因此 KD 的损失计算为：

$$L_{\mathrm{KD}}=(1-\boldsymbol{\delta})L_{\mathrm{CE}}(\boldsymbol{y},P_S)+2\boldsymbol{\delta}T^2L_{\mathrm{CE}}\left(\frac{P_S}{\boldsymbol{t}}+\frac{P_T}{\boldsymbol{t}}\right) \tag{1}$$

其中，$L_{\mathrm{CE}}(\cdot,\cdot)$ 表示交叉熵损失，$\boldsymbol{y}$ 为硬标签的热向量，$\boldsymbol{t}$ 为温度超参数，$\boldsymbol{\delta}$ 为权衡超参数。式（1）中的第一项是用真实标签定义的交叉熵损失，第二项是用来鼓励学生模仿老师的软化类别分数。算法常用符号如表 1 所列。

表 1 MS-MGANs 算法的常用符号

Table 1 Common symbols of MS-MGANs

符号	含义
$L_{\mathrm{CE}}(\cdot,\cdot)$	交叉熵损失函数
$\|\cdot\|_2^2$	均方误差损失函数
α,β	权衡超参数
P_Z	模型 Z 最后的 Softmax 层输出
$P_{i,j}^Z$	模型 Z 的第 i 阶段第 j 个卷积块的输出
N	第一阶段训练总迭代数
M	第二阶段训练总迭代数，同时表示模型内卷积块的个数
$stage_{i,j}$	模型第 i 个阶段第 j 个卷积块的训练过程
G^i	生成器/学生模型的第 i 个卷积块
D^i	辨别器/老师模型的第 i 个卷积块
F_{gan}^j	模型第 j 个卷积块的输出

3.2 多阶段多生成对抗网络

3.2.1 老师模型的损失函数

首先，由于老师和学生之间存在着巨大的性能差距，一个小而浅的学生很难模仿甚至超越老师。其次，模型训练通常采用随机梯度下降法使损失函数最小。由于损失函数的高度非凸性，在训练过程中存在许多局部最优值。当网络收敛到某个局部最小值时，无论初始化方式如何，其训练损失都会收敛到某个（或类似）值[26]。

为了解决上述问题，本文将老师的训练分为多个阶段。将图像作为老师的输入，老师得到的输出与硬标签 $\boldsymbol{y}$ 计算交叉熵损失，然后对老师进行反向传播。为了更好地解释，本文

将老师的训练过程分为 N 个阶段，记为 $stage_i(i=1,\cdots,N)$，对老师的参数进行随机初始化，每一阶段学习率衰减，以更好地达到较低的收敛值。因此，老师 L_{Teacher} 的训练损失可以表示为：

$$L_{\text{Teacher}}=\sum_{i}^{N}L_{\text{CE}}(P_T,y) \tag{2}$$

其中，$stage_i$ 阶段训练结束后，训练完成的老师 T_i 的参数被冻结，用于后续学生的训练过程。通过这种阶段性的训练策略，可以使学生更容易、更快地模仿甚至超越老师。

3.2.2 学生模型的损失函数

传统的 KD 方法直接以老师输出作为软目标，最大限度地减少学生输出的损失，然而老师传授的知识可能并不总是对学生有帮助。理想情况下，学生正确地学习重要的细节，而省略不必要的细节。因此在对学生的训练中，基于卷积块将学生的训练过程划分为 M 个步骤，对于第 i 个老师训练学生的第 j 块，定义为 $stage_{i,j}$。同时，学生的第 j 块作为生成器 G_j，将其生成的特征图输入到老师对应的块 D_{j+1} 中，并继续向后直到得到一个“假”结果。最后，最小化老师的“真”结果和学生的“假”结果损失，使学生继续模仿甚至超越老师。

(1)逐层贪婪训练。本文采用分段训练的方法，即一次训练一个模块。图像作为输入传递给老师和学生，从第一个卷积块中获得特征图的输出结果，并计算老师和学生第一个卷积块的均方误差。在对第一个卷积块进行一定迭代训练之后，停止对第一个卷积块的训练。在下一阶段，图像输入仍然传递给老师和学生，而来自第二个卷积块的特征被取走，然后按照与第一阶段相同的步骤，即使用反向传播，最小化老师和学生的第二块输出之间的 MSE 损失，这个过程对所有的块重复。老师 T_i 用于指导学生的中间层损失 L_{Mid}。

$$L_{\text{Mid}}=\sum_{i}^{N}\sum_{j}^{M}\parallel P_{i,j}^{T}-P_{i,j}^{S}\parallel_2^2 \tag{3}$$

其中，$P_{i,j}^{T}$ 和 $P_{i,j}^{S}$ 分别表示老师/学生的第 i 阶段第 j 块的输出特征，$\parallel\cdot\parallel_2^2$ 表示 L-2 损失函数。这样学生可以逐步了解老师的重要细节，且可以更好地模仿老师。但是由于不同模型之间存在一定的差异，模型训练的过程也不完全相同。分层贪婪策略直接计算中间层结果的损失，却忽略了不同模型之间实例转换的差异以及硬标签对模型训练的重要性。因此本文引入多生成对抗网络来解决这个问题。

(2)多生成对抗网络。考虑到 GAN 的难收敛性和梯度消失的问题，本文利用模型内部的卷积块代替额外引入的生成器和辨别器。假设学生的前 b 个卷积块(1,2,…,b)作为生成器($b\geqslant 0$)，相对应的老师后 $M-b$ 个块($b+1,\cdots,M$) 作为辨别器，本文方法通过在多阶段训练模式中引入多生成网络和对抗网络来提高模型训练的效率。

首先从任意的 Vanilla GAN 开始。GAN 是在生成器 G_x 和辨别器 D_x 之间的极大极小博弈，目标函数 L_{GAN} 可表示为：

$$L_{\text{GAN}}=E_{x\sim P_{\text{data}(I)}}[\log D(I)]+E_{z\sim P_{z(z)}}[\log D(1-D(G(z)))] \tag{4}$$

由于生成器缺乏噪声数据，传统的如式(4)所示的训练方法是不可行的。因此本文从以下两个方面对其进行改进。

(1)多生成网络。本文设计的生成器不同于其他方法直接引入生成器模型，而是巧妙利用多阶段训练方法，将学生的前 b 块定义为特征生成器。假设 M 同样是生成器的总组号，老师和学生的总组号相同。这样，生成器可以用 $\{G^1,G^2,G^3,\cdots,G^M\}$ 表示，则输入图像 I 和第 j 组的输出用 F_{gan}^{j} 表示：

$$\begin{aligned}&F_{\text{gan}}^{1}=G^1(I)\\&F_{\text{gan}}^{j}=G^j(F_{\text{gan}}^{j-1}),1\leqslant j\leqslant M\end{aligned} \tag{5}$$

当 j 等于 M 时，F_{gan}^{j} 是学生最后一块的输出结果。

(2)对抗网络。由于生成器是由学生的多个模块组成，因此可以用老师的多个模块来构造辨别器。M 是辨别器的总组数，则该辨别器可表示为 $\{D^1,D^2,D^3,\cdots,D^M\}$，从而得到 j 块的最优生成器 G^{j*}：

$$G^{j*}=\arg\min_{G^j}E_{z\sim P_{\text{data}(z)}}[\log(1-D^{j*}(G(F_{\text{gan}}^{j-1})))] \tag{6}$$

其中，$1\leqslant j\leqslant M$，D^{j*} 为第 j 块的最优辨别器。由于缺乏噪声数据，传统的方法无法对该辨别器进行训练。因此，本文将这个最大值和最小值博弈转换为生成样本和真实样本之间的损失最小化。因此，在训练第 j 组时，只优化了 G^j，$\{D^{j+1},D^{j+2},\cdots,D^M\}$ 固定，则将 G^j 的输出输入到 $\{D^{j+1},D^{j+2},\cdots,D^M\}$ 中得到“假”输出。但需要注意的是，对于第 j 组的训练，生成器对应 $\{G^1,G^2,G^3,\cdots,G^j\}$，定义为 $G^{1,j}$，辨别器对应 $\{D^{j+1},D^{j+2},\cdots,D^M\}$，定义为 $D^{j+1,M}$。

对于训练生成器 G^j，可以得到改进的多生成对抗网络的训练损失：

$$L_{\text{GAN}}=\parallel D^{j+1,M}(F_{\text{gan}}^{j}),D(I)\parallel_2^2+L_{\text{CE}}(D^{j+1,M}(F_{\text{gan}}^{j}),y) \tag{7}$$

计算最小化 $D^{j+1,M}(F_{\text{gan}}^{j})$ 得到的“假”结果和 $D(I)$ 得到的“真”结果的损失，鼓励学生模仿老师的实例转换过程。另一方面，对 $D^{j+1,M}(F_{\text{gan}}^{j})$ 和硬标签 $\boldsymbol{y}$ 进行交叉熵损失计算，以消除老师的限制，从而使学生有更大的可能超越老师的表现，总体训练框架如图 2 所示。

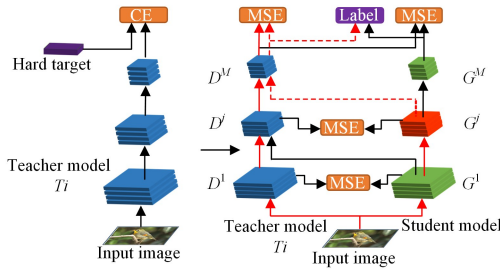

图 2 MS-MGANs 在 i^{th} 阶段对生成器 G^j 的训练过程

Fig. 2 Training process of generator G^j in the i^{th} stage of MS-MGANs

3.2.3 学生-老师的互学习策略

由于迭代训练到一定程度时，学生模型的性能将超越老师模型，那么提供软目标的老师模型将会失去指导学生模型的效果。因此，借鉴互学习的思想，当学生模型的性能超越老师模型的性能时，学生和老师的身份将会互换，此时学生将去指导老师进行训练，直到老师模型的性能再次超越学生模型。这种互学习的思想将会促进学生和老师的共同进步，进一步

提升学生的精度性能，后面的实验结果证明了这一点。

3.3 MS-MGANs

由式(6)完成对学生的多个生成器的训练后，得到由多个生成器组成的目标网络。在训练学生的最后一个卷积块G^M时，本文只计算了学生-老师中间层损失和硬标签损失，得到最后的损失函数如下：

$$Loss=L_{\text{Mid}}+\alpha L_{\text{GAN}}+\beta L_{\text{GT}} \tag{8}$$

其中，α，β 代表了权衡参数，L_{Mid}表示 $\{D^j,G^j\}$ 输出的均方误差，L_{GAN}表示最小化多生成对抗网络损失，L_{GT}表示学生与硬标签 $\boldsymbol{y}$ 的交叉熵损失。需要注意的是，本文只用 L_{Mid}和 L_{GT}来计算学生最后一个块 G^M的损失。

本文提出的 MS-MGANs 将学生-老师分为多阶段多个块逐层训练，$G^{1,j}$作为生成器，其输出F^j_{gan}作为老师 $D^{j+1,M}$的输入，得到“假”的结果和“真”的结果并进行最小化损失计算，最终获得目标网络。算法 1 给出了 MS-MGANs 互学习知识精馏方法的整个训练过程。

算法 1 MS-MGANs 的随机梯度训练过程

输入：由$\{G^1,G^2,G^3,\cdots,G^M\}$组成的学生模型 S，由$\{D^1,D^2,\cdots,D^M\}$组成的老师模型 T，数据集 D，训练总阶段 N 和训练卷积块总数 M

输出：学生模型，老师模型

```
repeat
    转换 T 为 train();
repeat
    随机选择一个 batch{x_p^D}_{p=1}^Z;
    根据式(2)优化L_Teacher;
until 得到训练好的老师模型T^i;
转换T^i为 eval()，转换 S 为 train();
repeat
    if 学生模型 S 的性能优于老师模型T^i;
      转换 T 为 train();
      转换 S 为 eval();
    随机选择一个 batch {x_p^D}_{p=1}^Z;
    冻结学生模型中除了生成器G^j以外的其他参数
    根据式(6)优化 S 的G^j;
    根据式(7)最小化损失函数 Loss;
until G^M完成训练
until 完成模型 N 个阶段的训练
```

4 实验

4.1 实验设定

本文在多个公共图像分类数据集上对所提方法进行了评价。1)CIFAR10/100。数据集包含 60 000 张 32×32 像素的 RGB 图像，10/100 个类，每个类由 1 000 张图像组成，其中 50 000 张用于训练图像，10 000 张用于测试图像。CIFAR 图像的每个边缘填充 4 个像素，图像被随机裁剪为 32×32 像素进行训练和测试。2)ImageNette/Imagewoof。这两个数据集是 ImageNet 数据集的子集，仅在难度方面有所不同。前者是一个比较容易分类的数据集，而后者分类相对较难。由于本文工作的主要目的是确保所提方法允许学生无限接近甚至超过老师，因此评价标准较少关注模型的准确性。

参考文献[26]，本实验中老师默认采用 ResNet34，学习率为 5×10^{-2}，重量衰减为 5×10^{-4}，动量为 0.9。SGD 优化器的学习率在每个阶段都会衰减。学生默认采用 ResNet18，学习率为 1×10^{-4}，采用 Adam 作为优化器。将 ResNet 的第一卷积层 Conv1 修改为 3×3，padding 为 1，stride 为 1，去掉第一个 maxpool，其他结构保持不变。对于 ImageNette，没有修改其 ResNet 结构。批处理大小默认设置为 64，λ 为 0.2。所有实验均使用 Intel Core i7-9750H 和两个 NVIDIA GTX3080 图形处理器。

4.2 实验结果及分析

4.2.1 消融分析

本节将分析本文方法的各个组成部分，并解释它们对最终性能产生的影响。由于 CIFAR10 的训练效果较好，且不同方法之间的性能差距较小，因此消融实验基于 CIFAR10 和 CIFAR100 联合完成。所有的学生都在一位老师那里学习，实验结果如表 2 所列。

表 2 不同损失函数组合对模型性能的影响

Table 2 Effects of different loss function composition on model performance

（单位：%）

Dataset	Loss composition	*Accuracy*	*Gap*
CIFAR10	L_{T}	95.10	0.00
	L_{sw}	94.88	0.22
	L_{tg}	91.56	3.54
	$L_{\text{sw}}+L_{\text{tg}}$	94.92	0.18
	$L_{\text{sw}}+L_{\text{tg}}+L_{\text{IRG}}$	94.80	0.30
	$L_{\text{MS-MGANs}}$	**95.35**	**−0.25**
	$L_{\text{MS-MGANs}}+L_{\text{IRG-t}}$	**95.02**	**0.08**
CIFAR100	L_{T}	78.50	0.00
	L_{sw}	76.56	1.94
	L_{tg}	67.72	10.78
	$L_{\text{sw}}+L_{\text{tg}}$	77.84	0.66
	$L_{\text{sw}}+L_{\text{tg}}+L_{\text{IRG}}$	77.82	0.68
	$L_{\text{MS-MGANs}}$	**79.30**	**−0.80**
	$L_{\text{MS-MGANs}}+L_{\text{IRG-t}}$	79.02	**−0.52**

在表 2 中，本文比较了所提方法中不同组成对模型性能的影响。其中 L_{T}表示老师基线，L_{ST}表示学生-老师逐层蒸馏，L_{TG}表示学生-老师共同学习，$L_{\text{SW}}+L_{\text{TG}}$表示多阶段在线学习，$L_{\text{IRG-t}}$表示实例空间信息的损失，$L_{\text{MS-MGANs}}$表示本文提出的方法。本文的 $L_{\text{IRG-t}}$是参考文献[28]中提出的L_{MTK}，但由于计算成本较高，本文仅参考L_{MTK}中的$L_{\text{IRG-t}}$进行损失最小化计算。可以看出，本文提出的 MS-MGANs 性能明显优于其他相关组成部分，学生的准确率超过了老师的准确率。但在 MS-MGANs 中加入 $L_{\text{IRG-t}}$损失后，精度降低，这与分阶段训练方法缺乏整体空间信息有关；另一方面，不同模型体系结构的空间结构是多样化的，不同模型转换实例的空间信息也有很大的不同，因此$L_{\text{IRG-t}}$未被引入到本文方法当中。

4.2.2 基于 CIFAR10 和 CIFAR100 的评估

(1)对不同比例的数据进行分类。将本方法与目前流行的蒸馏方法 DT(学生直接训练)、KD[6]、AT[7]、FSP[25]、TG[26]、SW[27]、IRG-t[28]进行了比较，实验结果如表 3 所列。

表 3 在不同规模的 CIFAR10/CIFAR100 数据集上比较本文方法和当前流行的方法

Table 3 Comparison of proposed method and state-of-the-art methods on CIFAR10/CIFAR100 datasets with different sizes

(单位:%)

	CIFAR10-10%	CIFAR10-20%	CIFAR10-100%	CIFAR100-10%	CIFAR100-20%	CIFAR100-100%
Baseline(T)	79.16	87.33	95.10	37.72	55.67	78.50
DT(S)	71.71	78.59	92.06	32.35	40.53	68.37
KD	78.18	85.65	93.72	39.13	51.15	72.87
FSP	77.84	84.81	94.40	39.41	53.79	74.21
AT	73.51	80.48	93.94	31.17	44.06	73.76
SW	79.17	86.93	94.88	38.57	55.45	76.56
SW+TG	79.47	87.34	94.92	38.85	56.23	77.84
MS-MGANs	**80.02**	**87.91**	**95.35**	**39.23**	**57.29**	**79.30**
MS-MGANS+IRG-t	79.56	87.86	95.02	38.96	56.85	79.02

从表 3 可以看出,本文提出的 MS-MGANs 在 10%,20%,100%的不同比例数据上都表现出明显的优势,并且随着数据集比例的减小,本文方法与流行方法之间的精度差距逐渐增大。由此可见,MS-MGANs 同样适用于小规模数据的训练。其次,所提方法使得学生的最终准确率超过了老师。在 CIFAR100-10%中,学生甚至比老师的准确率高 1.31%,这是从一定比例的硬标签中学习所得到的效果。值得注意的是,在数据集比例较低的情况下,KD 的准确率优于 FSP 和 AT。主要原因是 KD 方法在学生-老师性能差距小、数据集小的情况下具有较好的泛化效果。

(2)对不同的学生-老师对进行分类。本文基于 CIFAR100 测试了不同师生的精度,实验结果如表 4 所列。

表 4 在不同学生-老师组合上比较本文方法和当前流行的方法

Table 4 Comparison of proposed method and state-of-the-art methods on different student-teacher pairs

(单位:%)

Teacher / Student	Resnet34(85.5m)				Resnet20
	ResNet34	ResNet18	ResNet10	MobileNet (9.29m)	ResNet10 (19.8m)
Baseline(T)	78.50	78.50	78.50	78.50	77.51
DT(S)	68.57	68.37	64.68	53.27	64.68
KD	71.08	72.87	70.40	55.17	70.91
FSP	73.99	74.21	73.39	53.82	72.89
AT	69.10	73.76	66.37	54.86	66.03
SW	78.06	76.56	74.40	55.74	73.95
SW+TG	78.13	77.84	74.83	55.95	73.93
MS-MGANs	**79.39**	**79.30**	**76.84**	56.21	**76.58**
MS-MGANS+IRG-t	78.89	78.99	76.03	**57.70**	76.02

通过不同的学生-老师对的比较可以看出,本文方法表现出了较好的优越性。随着老师与学生差距的增大,MS-MGANs 与 SW+TG 的准确率差距也逐渐增大。在 ResNet34 和 ResNet10 中,MS-MGANs 与 SW+TG 的准确度差距可达 1.9%。由此说明本文方法适用于不同的师生对,并解决了不同模型对蒸馏过程的影响。对于 ResNet34/MobileNet,各种方法精度差距减小的主要原因是 MobileNet 的结构限制导致模型精度受到限制。总的来说,本文提出的方法在不同的老师和学生对中具有良好的模型泛化能力。

4.2.3 基于 ImageNette 和 ImageWoof 的评估

在本实验中,由于实验设备的限制,批处理量调整为 32,老师学习率调整为 5×10^{-4},学生学习率仍然为 1×10^{-4},其他设置与前文实验一致,实验结果如图 3 所示。

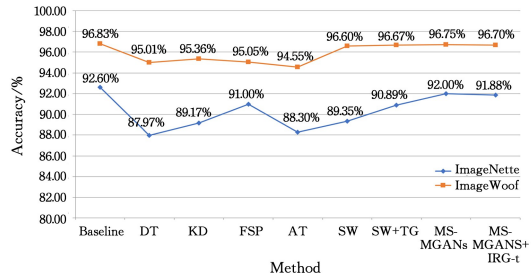

图 3 在 ImageNette/ImageWoof 数据集上比较本文方法和当前流行的方法

Fig. 3 Comparison of proposed method and state-of-the-art methods on ImageNette/ImageWoof datasets

从图 3 可以看出,MS-MGANs 也适用于大规模数据集。在多阶段蒸馏过程中,学生逐渐学习和模仿老师,学生最终在一定程度上超越老师。

虽然本文在 ImageNette 和 ImageWoof 中提出的方法并没有超越老师,但它们之间的差距非常接近,最小差距为 0.08%,实验结果表明 MS-MGANs 在大数据集上仍然有效。

4.2.4 讨论

比较 KD, FSP, AT, SW, SW+TG, MS-MGANs, MS-MGANs+IRG-t 等方法发现,KD 只使用模型最后一层的知识来实现蒸馏操作,而其他方法使用的是中间层的信息。具体来说,FSP 采用模型中的不同层来构建相似矩阵,AT 引入了注意力机制,但它们都忽略了师生之间的性能差距对蒸馏过程的影响。SW 将训练过程以卷积块的形式分为多阶段训练,虽然提高了学生收敛速度,但忽略了不同模型实例转换过程的差异以及硬标签对不同卷积块训练的促进作用。GAN 的引入可以激发学生更好地模拟老师实例变换,但训练过程复杂,容易导致梯度爆炸。相反,本文选择模型内部的卷积块作为生成器/辨别器,将传统的 GAN 极小极大博弈转化为学生与老师之间输出结果的"假""真"最小化,从而降低了模型训练的复杂性,能更好地激发学生模仿老师实例转换过程。

结束语 本文提出了一种基于多阶段多生成对抗网络的互学习知识蒸馏方法。该方法将训练分为多阶段逐块对抗训练,比传统的端到端训练策略更加有效,且减少了在某一阶段需要优化的参数量。实验结果表明,所提方法在小数据集和大数据集上都具有良好的泛化能力。此外,本文方法具有通用性,因此不局限于图像分类,还可用于目标检测、像素级图像分割等。在未来工作中,将探索更有效的互学习策略来

实现不同场景下的蒸馏压缩。

参 考 文 献

[1] WANG R Z, GAO J, HUANG S H, et al. Malicious Code Family Detection Method Based on Knowledge Distillation[J]. Computer Science, 2021, 48(1): 280-286.

[2] LIU J, CHEN Y, LIU K. Exploiting the Ground-Truth: An Adversarial Imitation Based Knowledge Distillation Approach for Event Detection[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 6754-6761.

[3] TAN K, WANG D. Towards model compression for deep learning based speech enhancement[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2021, 29: 1785-1794.

[4] CHEN X, ZHANG Y, XU H, et al. Adversarial Distillation for Efficient Recommendation with External Knowledge[J]. ACM Transactions on Information Systems, 2019, 37(1): 12. 1-12. 28.

[5] HAN S, MAO H, DALLY W. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding[J]. arXiv: 1510. 00149, 2015.

[6] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. arXiv: 1503. 02531, 2015.

[7] ROMERO A, BALLAS N, KAHOU S, et al. FitNets: Hints for Thin Deep Nets[J]. arXiv: 1412. 6550, 2014.

[8] YE J, JI Y, WANG X, et al. Data-free Knowledge Amalgamation via Group-stack Dual-GAN[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 12516-12525.

[9] GUO Q, WANG X, WU Y, et al. Online Knowledge Distillation via Collaborative Learning[C] // 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11017-11026.

[10] BENGIO Y, LOURADOUR J, COLLOBERT R, et al. Curriculum learning[C]// Proceedings of the 26th Annual International Conference on Machine Learning. 2009: 41-48.

[11] PENTINA A, SHARMANSKA V, LAMPERT C H. Curriculum learning of multiple tasks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 5492-5500.

[12] ZHANG X, SHAPIRO P, KUMAR G, et al. Curriculum learning for domain adaptation in neural machine translation[J]. arXiv: 1905. 05816, 2019.

[13] GUO Y, CHEN Y, ZHENG Y, et al. Breaking the curse of space explosion: Towards efficient nas with curriculum search[C]// International Conference on Machine Learning. 2020: 3822-3831.

[14] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. arXiv: 1406. 2661, 2014.

[15] CHEN H, WANG Y, XU C, et al. Data-free learning of student networks[C] // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 3514-3522.

[16] WANG Y, GONZALEZ-GARCIA A, BERGA D, et al. Minegan: effective knowledge transfer from gans to target domains with few images[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 9332-9341.

[17] LI M, LIN J, DING Y, et al. Gan compression: Efficient architectures for interactive conditional gans[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 5284-5294.

[18] VIAZOVETSKYI Y, IVASHKIN V, KASHIN E. Stylegan2 distillation for feed-forward image manipulation[C]// European Conference on Computer Vision. 2020: 170-186.

[19] GONG R, LIU X, JIANG S, et al. Differentiable soft quantization: Bridging full-precision and low-bit neural networks[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 4852-4861.

[20] BOO Y, SHIN S, CHOI J, et al. Stochastic precision ensemble: Self-knowledge distillation for quantized deep neural networks [J]. arXiv: 2009. 14502, 2020.

[21] GUO S, WANG Y, LI Q, et al. Dmcp: Differentiable Markov channel pruning for neural networks[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 1539-1547.

[22] LIN M, JI R, WANG Y, et al. Hrank: Filter pruning using high-rank feature map[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 1529-1538.

[23] TAI C, XIAO T, ZHANG Y, et al. Convolutional neural networks with low-rank regularization [J]. arXiv: 1511. 06067, 2015.

[24] WU B, WANG D, ZHAO G, et al. Hybrid tensor decomposition in neural network compression[J]. Neural Networks, 2020, 132: 309-320.

[25] YIM J, JOO D, BAE J, et al. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4133-4141.

[26] JIN X, PENG B, WU Y, et al. Knowledge distillation via route constrained optimization[C] // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 1345-1354.

[27] KULKARNI A, PANCHI N, CHIDDARWAR S. Stagewise knowledge distillation[J]. arXiv: 1911. 06786, 2019.

[28] LIU Y, CAO J, LI B, et al. Knowledge distillation via instance relationship graph[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 7096-7104.

[29] WANG X, LI Y. Harmonized dense knowledge distillation training for multi-exit architectures[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 10218-10226.

HUANG Zhong-hao, born in 1997, postgraduate. His main research interests include data compression and recommendation system.

YANG Xing-yao, born in 1984, Ph.D, associate professor, is a member of China Computer Federation. His main research interests include recommender system and trust computing.

（责任编辑：何杨）