文章编号：2095-2163(2022)12-0153-06 中图分类号：TP391.4 文献标志码：A

深度交流学习模式

张仁斌[1,2,3]，王 龙[1]，周泽林[1]，左艺聪[1]，谢 昭[1]
（1 合肥工业大学 计算机与信息学院，合肥 230601；2 合肥工业大学 大数据知识工程教育部重点实验室，合肥 230601；
3 合肥工业大学 工业安全与应急技术安徽省重点实验室，合肥 230601）

摘 要：本文针对深度神经网络如何更快速和充分学习的问题，提出一种基于知识传递的深度交流学习（Deep Communication Learning，DCL）模式。该模式中多个神经网络在各自独立学习的同时将网络参数作为知识进行交流，单个神经网络在训练中将自身所学到的知识分享给其他网络，同时从其他网络上吸纳一定比例的学习成果，交替进行独自学习和在集体中的知识交流。基于多个公开数据集的实验结果表明，相对于单独学习，仅用 2 个网络进行 DCL 就可获得学习效果最高 3.44%的提升；增加进行 DCL 的网络个数至 6 个，学习效果可进一步得到最高 2.74%的提升。DCL 模式有利于训练出效果更好的神经网络。

关键词：神经网络；深度学习；知识传递；网络参数；交流学习

Deep communication learning

ZHANG Renbin[1,2,3], WANG Long[1], ZHOU Zelin[1], ZUO Yicong[1], XIE Zhao[1]
（1 School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China; 2 Key Laboratory of Knowledge Engineering with Big Data, Hefei University of Technology, Hefei 230601, China; 3 Anhui Province Key Laboratory of Industry Safety and Emergency Technology, Hefei University of Technology, Hefei 230601, China）

【Abstract】 The paper proposes a Deep Communication Learning (DCL) pattern based on knowledge transfer to deal with the problem of how deep neural networks can learn more quickly and adequately. In the pattern, multiple neural networks communicate network parameters as knowledge while learning, and a single neural network shares its learned knowledge with other networks during training, while absorbing a certain percentage of learning results from other networks, alternately learning alone and exchanging knowledge in the group. Experimental results based on several publicly available datasets show that DCL with only two networks can achieve up to 3.44% improvement in learning compared to independent learning. The number of networks performing DCL is increased to 6 further, which increases the learning by up to 2.74%. DCL is beneficial for training better neural networks.

【Key words】 neural network; deep learning; knowledge transfer; network parameter; communication learning

0 引 言

基于知识传递的知识蒸馏和参数迁移学习分别被广泛使用于模型压缩[1-2]和迁移学习[3]领域中，本文的目标是基于知识传递实现网络间的交流学习，使网络学习更加快速和充分。

尽管复杂庞大的网络具有很高的性能，但是计算缓慢和网络庞大不利于存储的不足使其难以满足在便携设备上的应用需求。模型压缩是解决这个问题的方法之一。Hinton 等人[4]通过知识蒸馏（Knowledge Distillation，KD），首先利用大规模数据训练一个大模型作为教师网络，然后将小模型学生网络向大模型学习，知识从教师网络传递到学生网络上，以此得到的小网络也具有大网络相当的泛化能力，实现模型压缩。

在知识蒸馏的研究中，Zagoruyko 等人[5]提出将注意力作为知识从一个网络转移到另一个网络中的学习方法，并且与将教师网络的输出作为学习对象的知识蒸馏方法进行结合。Chen 等人[6]提出交叉样本的相似性作为网络间可转移的知识，并在多个图像任务中进行验证，转移这种知识使行人识别任务相对基线取得明显提升。Cho 等人[7]进一步探索知识蒸馏的有效性，得出了教师网络的效果越好并非意味着学生网络效果就会越好的结论，这与 Mirzadeh 等人[8]的实验结论相同。Heo 等人[9]将隐藏层特征作为知识进行蒸馏，并在图形分类、检测和

基金项目：国家重点研发计划专项资助项目（2016YFC0801804，2016YFC0801405）；中央高校基本科研业务费专项资金资助项目（PA2019GDPK0074）。

作者简介：张仁斌（1971-），男，博士，副教授，主要研究方向：工业互联网安全、人工智能；王 龙（1996-），男，硕士研究生，主要研究方向：人工智能；周泽林（1997-），男，硕士研究生，主要研究方向：网络安全；左艺聪（1999-），男，硕士研究生，主要研究方向：网络安全；谢 昭（1980-），男，博士，教授，主要研究方向：计算机视觉、人工智能。

通讯作者：王 龙 Email：wanglong.email@foxmail.com

收稿日期：2022-03-24

分割三种任务上进行实验,验证了特征蒸馏的有效性。不同于分类任务,Saputra 等人[10]在回归任务中成功应用了知识蒸馏。Phuong 等人[11]从多个角度解释了为什么知识蒸馏能够成功地将知识在网络间进行转移。近期,Facebook 团队提出的 Deit[12]方法,探索了使用多种其他类型的网络来对图像分类网络 ViT[13]进行注意力的教学,达到了非常理想的效果。Deit 方法中,在训练时将基于 Transformer[14]的 ViT 作为学生网络,将其他类型的网络,如以 CNN 为基础的 ResNet[15]、EfficientNet[16]作为教师网络,借鉴知识蒸馏的方法,通过将学生网络和教师网络的输出计算损失值并进行反向传播,实现将知识从教师传递给学生,以此显著提高作为学生的 ViT 网络的性能。实验结果表明,相对于需要在大量数据集上进行预训练的 ViT,Deit 不需要额外的数据做预训练,且用更少的计算资源生成更高性能的图像分类模型。Deit 通过将不同网络的知识进行传递,达到很好的学习效果。Lu 等人[17]分别在高分辨率和多分辨率模型中运用知识蒸馏提炼知识,通过交叉特征融合和多尺度训练等方式获得了更优的学生分辨率模型。Chen 等人[18]把神经网络实例的特征和节点的关系作为编码知识从教师网络传递给学生网络,在物体检测的任务中取得了更好的模型效果。

把网络的参数作为知识进行转移也有着非常经典的应用。Pan 等人[3]将源领域中模型的参数迁移到目标领域的模型中的方法归类为参数迁移(Parameter-transfer)学习。Fan 等人[19]将少样本检测任务上学习到的知识迁移到检测模型的最后一层,检测效果相对基线得到了稳定的提高。Jing 等人[20]通过知识参数迁移,把多个教师图神经网络的知识传递给同一个学生图神经网络,以此得到的学生网络在多个任务上取得了与教师网络相当的效果。在 Mean Teachers[21]中,教师网络通过将学生网络的参数进行组合得到教师自身的网络参数,以此实现知识从学生网络向教师网络的传递。

Mean Teachers 作为半监督的学习方法,同样包括了教师网络和学生网络两种结构。其中,学生网络的参数是通过梯度下降进行更新,而教师网络的参数则是仅仅通过组合学生网络所学到的知识参数进行更新,而不进行梯度下降。在 Mean Teachers 中,知识通过网络参数的形式从学生网络流向教师网络。进一步地,教师网络的输出结果作为学生网络的学习目标,进行对学生网络的教学。

深度互助学习[22](Deep Mutual Learning ,DML)中,K 个网络中每一个网络既有学生的身份,也有教师的身份。当对其中某个网络传递知识时,其他所有 $K-1$ 个网络都作为教师。在每一轮互助中,每个网络都会接收到其他 $K-1$ 个网络传递的知识。在知识蒸馏的方法中,小的学生模型通过将大的教师模型输出作为学习的软目标计算交叉熵进行梯度下降,进而完成知识从大模型向小模型的传递。DML 不以模型压缩为目的,而是通过将学生网络与其他 $K-1$ 个教师网络的输出结果的 KL 散度(Kullback-Leibler divergence, KL)取均值,并作为损失的一部分进行反向传播,依托多个网络输出结果的互相借鉴,以此达到更高的鲁棒性,实现共同进步。

利用知识蒸馏可以加快小模型的训练速度和效果,但是具有一定的局限性。比如蒸馏的前提是拥有一个性能足够好的教师网络,且蒸馏的主要目的在于更好地训练出一个小模型,并不能够提升教师网络自身的性能。DML 中每个网络都会利用其他网络的知识来提高自己,但是 DML 中实现互助的方式是利用网络之间的差异度作为损失值进行梯度下降,模型性能受梯度下降方法局限性的影响,如梯度消失和梯度爆炸等导致互助失败。

针对以上问题,本文提出一种基于深度交流学习(Deep Communication Learning, DCL)的网络训练模式。在 DCL 中,多个神经网络在各自独立学习的同时将网络参数作为知识进行交流,单个神经网络在训练中将自身所学到的知识分享给其他网络,同时从其他网络上吸纳一定比例的学习成果,独自学习和在集体中的知识交流是交替进行的。

DCL 和 Mean Teachers 都将网络所学到的参数作为知识,并将这些知识进行传递。不同的是,Mean Teachers 中教师网络的目的在于对无标签数据进行标记,且最终学生网络向教师网络的学习方式同样类似于知识蒸馏,是通过计算学生网络和教师网络输出结果之间的差异度进行反向传播实现的。Mean Teachers 中教师网络和学生网络的主体是固定的,而 DCL 中每个网络既会作为知识的传授方,也会作为知识的接收方,这些网络的身份是等同的,并且 DCL 各个网络间互相学习的策略与 Deit 和 DML 完全不同。Deit 和 DML 借鉴知识蒸馏,以教师模型的输出结果为目标,让学生向教师模仿和学习,而 DCL 则是将各个网络所学到的网络参数作为知识进行吸纳和融合,交流的过程不使用梯度下降,而是对所学知识的直接交流。

本文利用经典、成熟的图像分类神经网络来验

证所提出的学习模式，使用 Inception[23]，ResNet，WRN[24]，DenseNet[25]，MobileNet[26]，ResNeXt[27] 和 EfficientNet 等 7 种经典网络在 Fashion-MNIST[28]，CIFAR-10 和 CIFAR-100[29] 等多个数据集上进行实验。结果表明，利用 DCL，这些网络获得了学习效果最高 3.44%的提升。

论文内容安排如下：本文第 1 节提出了一种基于知识交流的深度神经网络学习方式-DCL，并对该方法进行了详细说明；第 2 节通过使用多种网络和数据集对 DCL 进行了实验，验证了 DCL 学习模式的有效性；第 3 节对全文进行总结并展望未来工作。本文将代码和模型进行了开源[30]。

1　深度交流学习

深度交流学习模式如图 1 所示。深度交流学习是对人类社会学习进步的一个仿照。正如人类在个体单独学习后进入集体进行知识的交流，并经独自的学习把从集体获得的知识进行消化和吸收，利用集体的知识提高自己，同时在独自学习中探索和获取新的知识，再于此后的交流中对其进行分享。

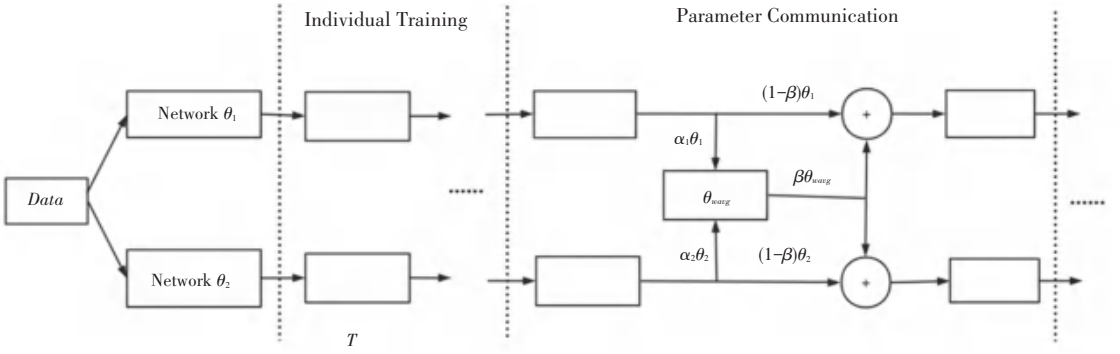

图 1　深度交流学习模式

Fig. 1　The process of Deep Communication Learning

网络学习到的知识存在于网络的参数之中，让深度神经网络在学习的同时进行知识的交流是深度交流学习的核心。DCL 的具体策略是，网络在独自学习一定的迭代轮次 T 后，各个网络把自己学习到的知识贡献到集体中，并以一定的比例 β_i 收纳来自集体的知识。随后各个网络再独自学习一段时间 T，以适应和吸收集体的知识经验，再独自探索新知识用于下次和其他网络的交流。DCL 用这样的方式让所有网络不断地在互相交流中进行学习和进步。

即使是同一类型的网络，不同的初始化参数也会使网络变得互不相同。虽然数据集和网络结构都一样，但是额外的知识存在于不同的初始化参数之中。针对于 Independent 学习中网络知识量有限的问题，DCL 模式中采用知识交流的方法，支持每个网络拥有额外的知识量。

1.1　网络交流方式

设 DCL 中的网络初始化数量为 K，此 K 个网络表示为：$\theta_1, \theta_2, \cdots\cdots, \theta_K$。设具有 N 个样本且分为 M 类的数据集为：

$$X = \{x_i\}_{i=1}^{N} \tag{1}$$

$$Y = \{y_i\}_{i=1}^{N}, \quad y_i \in \{1,2,\cdots,M\} \tag{2}$$

初始化 DCL 后，每个网络进行随机采样和反向传播学习。在经过 T 次独自学习的迭代后，所有网络进行一次知识交流。交流中，每个网络首先将自己所学的参数知识以 α_i 的比例贡献到集体中，并存储于 θ_{wavg} 中。对此过程可用式(3)进行描述：

$$\theta_{wavg} = \sum_{i=1}^{K} \alpha_i \theta_i \tag{3}$$

其中，对于每个网络所贡献的比例，具有以下约束：

$$\sum_{i=1}^{K} \alpha_i = 1 \tag{4}$$

然后，每个网络从集体的知识中吸纳比例为 β 的知识量，实现总体网络的知识向每个网络的传递：

$$\theta_i = (1-\beta)\,\theta_i + \beta\,\theta_{wavg} \quad 0 < \beta < 1 \tag{5}$$

对于个体向集体贡献的知识量比例 α_i 数值的确定，本文借鉴正则化(regularization)的思想，即如果网络参数的绝对值 $|\theta_i|$ 越小，则让其对集体贡献更大比例的知识，从而让这些网络在表现效果相当的情况下，更多地向参数小的网络学习，以此增加自身的鲁棒性。本文采用的策略是令贡献的比例与自身参数的绝对值大小成反比，即：

$$\alpha_i \propto \frac{1}{|\theta_i|} \tag{6}$$

根据式(3)~(5)，可以得出每次交流中单个网络的参数对总体的贡献比例为：

$$\alpha_i = \frac{\prod_{j=1, j\neq i}^{K} |\theta_j|}{\sum_{i=1}^{K} \prod_{j=1, j\neq i}^{K} |\theta_j| + \varepsilon} \tag{7}$$

其中，ε 为一个极小数，用来避免当网络某层的参数全为 0 时出现分母非法的情况，在实际使用中，本文对 ε 的取值为 1×10^{-18}。

1.2 深度交流学习流程

算法 1 描述了 DCL 的具体流程。

算法 1 Deep Communication Learning

输入 数据集 X，标签 Y

输出 $\theta_1, \theta_2, \cdots\cdots, \theta_K$

1：Initialize：Initialize $\theta_1, \theta_2, \cdots\cdots, \theta_K$ to different conditions, T, β, E, γ_0

2：for $t \leftarrow 1$ to E do

3：Randomly sample data x from X

4：Compute γ_t fromlearning rate decay

5：Learn and update θ_i：

6： $\theta_i \leftarrow \theta_i + \gamma_t \frac{\partial L_{\theta_i}}{\partial \theta_i}$

7： if $t\%\ T$ equals 0：

8： Compute α_i by (6)

9： Compute θ_{wavg} by (2)

10： Updata θ_i：

11： $\theta_i \leftarrow (1-\beta)\ \theta_i + \beta\ \theta_{wavg}$

12：return $\theta_1, \theta_2, \cdots\cdots, \theta_K$

代码中，E 为训练的总迭代次数，E 和学习率衰减策略的具体设置见本文的 2.3 节。

在本算法中，T 的设置是关键之一，因为来自于其他网络的知识参数，未必会在吸纳后立即就能很好地适应自身的网络参数。因此，如果让网络一直交流而不给予足够的独自学习和适应时间，很容易会出现这些网络由于无法适应其他网络的知识参数，而一直处于欠拟合状态。

单个网络每次从集体中吸纳的知识比例 β 是一个超参数。如果 β 的取值过小，会使网络向集体学习的知识量很少，这种情况下一方面会相对保持网络的独特性，即网络之间不会非常相像，另一方面会降低交流学习给网络所带来的收益。在 β 取极值为 0 时，网络之间停止交流，个体不再向集体学习。同样，如果 β 的取值过大，则会使网络之间随着交流次数增多而变得更加相像，在一定程度上丧失自身的独特性。因此将 β 取一个适当大小的值是非常重要的，本文第 2 节实验中将 0.1 作为 β 的取值。

2 实验

本文使用多种网络在多个数据集上进行实验，所有源代码、模型和实验结果均已开源[30]。

2.1 数据集

本文使用 3 个数据集进行实验。CIFAR-10 和 CIFAR-100 数据集由大小为 32×32 的 RGB 图像组成，分别包含 10 个和 100 个类别的物体。两者都被划分为 50 000 张图像作为训练集和 10 000 张图像作为测试集。Fashion-MNIST 是一个包含 10 种服饰类别的图像数据集，图像大小为 28×28，并且以 60 000 张图片作为训练集，10 000 张图片作为测试集。本文将图像分类的正确率作为这 3 个数据集的评价指标。

2.2 网络

本文使用 7 种具有不同原理和参数量大小的经典神经网络进行实验。包括经典卷积网络 Inception-V1 以及深度残差网络 ResNet-18，以及以残差为基础进一步发展而来的 WRN-16-4、DenseNet-121 和 MobileNet-V2。作为 ResNet 和 Inception 的结合，ResNeXt-50 也被用在本文的实验中。兼顾速度与精度的 EfficientNet-B3 在图像分类领域有着优秀的表现，本文也采用这个网络作为实验对象之一。

2.3 实验设置

本文使用 PyTorch 实现了所有网络，并且以 NVIDIA Tesla V100 GPU 作为加速进行实验。实验采用 Nesterov 动量设置为 0.9 的 SGD 作为模型优化器，*batch size* 设置为 128，并且设置了0.000 1的 L_2 正则损失。在训练时，对于在 ImageNet 进行过预训练的模型，学习率被初始化为 0.001，而没有预训练过的模型，学习率被初始化为 0.1。学习率每迭代 60 个 *epoch* 会衰减为原来的 0.1，并且 200 个 *epoch* 被作为训练的总迭代次数。实验中，数据增强方法包括对图像的随机翻转和每边填充 4 个像素后进行的随机裁剪，裁剪后缺失的像素被填充为 0。

2.4 实验结果与分析

表 1～表 3 分别比较在 3 个数据集上多种网络在 $K = 2$ 时，通过 Independent 学习和 DCL 学习达到的 Top-1 正确率。结果分析表明：

(1)相对于 Independent 学习，所有这些网络都可以通过 DCL 来提高自己的学习效果，这些提高体现在 DCL-Ind 一列中的数据都是正数。

(2)没有经过预训练的网络结构，通过 DCL 则更显著地提升了学习效果。3 个数据集上最大的提升都来自于 WRN-16-4 和 EfficientNet-B3 这 2 个未进

行预训练的网络，分别是 3.44%，2.79%和 1.16%。

（3）相对于单通道且图片尺寸更小的 Fashion-MNIST，DCL 的学习方式在三通道且图片尺寸更大的 CIFAR 数据上对学习效果的提升更加明显。

表 1　CIFAR-100 数据集 $K=2$ 的 Top-1 正确率

Tab. 1　Top-1 accuracy for CIFAR-100 dataset when $K=2$　%

Network	Pretraining	Independent	DCL	DCL-Ind
ResNet-18	yes	77.68	79.82	**2.14**
WRN-16-4	no	65.88	69.32	**3.44**
DenseNet-121	yes	80.47	82.59	**2.12**
ResNeXt-50	yes	82.04	83.82	1.78
EfficientNet-B3	no	70.48	72.39	1.91

表 2　CIFAR-10 数据集 $K=2$ 的 Top-1 正确率

Tab. 2　Top-1 accuracy for CIFAR-10 dataset when $K=2$　%

Network	Pretraining	Independent	DCL	DCL-Ind
ResNeXt-50	yes	95.84	97.32	1.48
EfficientNet-B3	no	87.28	90.07	**2.79**
WRN-16-4	no	89.84	91.89	2.05
Inception-V1	yes	93.29	94.28	0.99
MobileNet-V2	yes	90.67	91.22	0.55

表 3　Fashion-MNIST 数据集 $K=2$ 的 Top-1 正确率

Tab. 3　Top-1 accuracy for Fashion - MNIST dataset when $K=2$　%

Network	Pretraining	Independent	DCL	DCL-Ind
EfficientNet-B3	no	90.46	91.62	**1.16**
ResNeXt-50	Yes	93.73	94.49	0.76
MobileNet-V2	Yes	92.83	93.56	0.73
WRN-16-4	No	92.49	93.04	0.55

在 DML 中，本文通过使用 *KL* 散度作为损失的一部分，让不同网络的输出更集中（不离群），而让网络各自都取得更好的学习效果。DML 能够生效的原因是这种方式使输出结果更加集中，提高网络的鲁棒性。在蒸馏学习中，通过让小网络把大网络的输出做软目标进行学习，实现大网络向小网络的知识传递。在 Deit 中，通过让 ViT 网络在某些层的输出向和卷积网络或者混合学习，实现把知识向 ViT 的传递，因而让 ViT 在图像分类中取得了更加优秀的表现。

本文 DCL 模式中，不同网络通过分享一部分权重来进行知识的沟通，在一定程度上使得这些网络的最终输出更加集中、即不离群，也有利于网络获得更高的鲁棒性。

图 2 和图 3 比较 EfficientNet-B3 和 ResNet-18 使用不同数量的网络进行 DCL 学习的结果。学习效果对比表明：

（1）DCL 的正确率曲线总是高于 Independent 的正确率曲线，表明在相同的迭代学习次数下，DCL 比 Independent 学习得更加充分，并且在训练结束后，DCL 达到了 Independent 所没有达到的学习效果。

（2）K 的值越大，图中的正确率曲线就处在更高的位置，表明增加进行交流和沟通的网络个数 K，将提高整体的学习效果。在 DCL 中，进行交流的学习者越多，集体会倾向于取得更加优秀的学习表现。

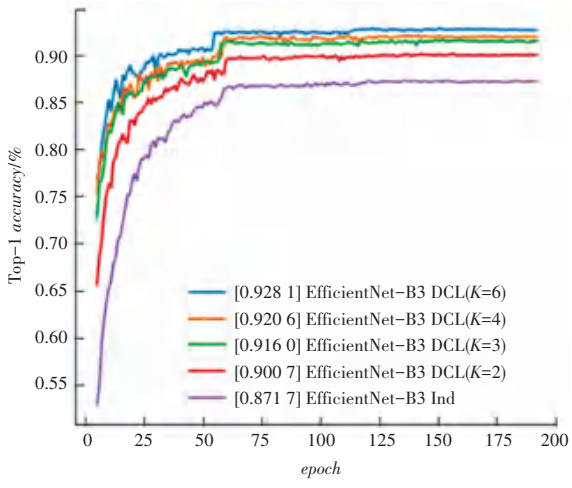

图 2　CIFAR-10 上 EfficientNet-B3 使用不同数量网络进行 DCL 学习的 Top-1 正确率结果

Fig. 2　Top-1 accuracy of DCL learning using different number of networks for EfficientNet-B3 on CIFAR-10

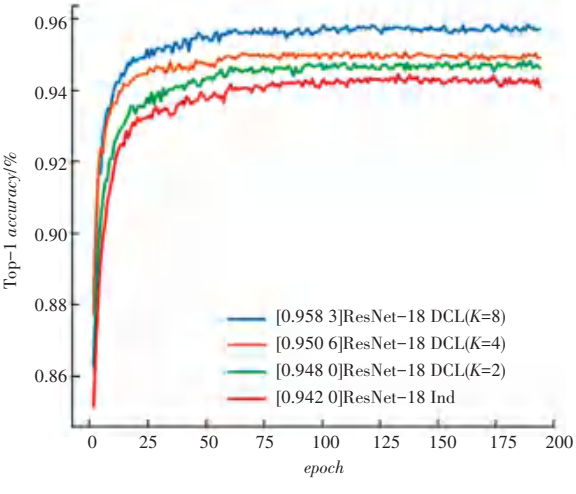

图 3　CIFAR-10 上 ResNet-18 不同数量网络进行 DCL 学习的 Top-1 正确率结果

Fig. 3　Top-1 accuracy of DCL learning using different number of networks for ResNet-18 on CIFAR-10

3　结束语

本文提出了一种让深度神经网络在学习中进行互相交流的训练模式，利用经典、成熟的图像分类神

经网络对所提出的学习模式的验证结果表明,该模式使多种深度神经网络的学习效果获得了明显提高。利用 DCL,深度神经网络学习的效果更好。实验结果证明了 DCL 模式对多类神经网络都有效,且增加交流的网络个数,能进一步提高学习效果。未来的工作将对分布式训练的交流方式进行探索,以提高多个网络进行交流训练的时间效率。

参考文献

[1] HAN Song, MAO Huizi, DALLY W J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding[C]//International Conference on Learning Representations(ICLR).Puerto Rico:dblp,2016:1-14.

[2] 高晗,田育龙,许封元,等. 深度学习模型压缩与加速综述[J]. 软件学报,2021,32(01):68-92.

[3] PAN S J, YANG Qiang. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10):1345-1359.

[4] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. Computer Science,2015,14(7):38-39.

[5] ZAGORUYKO S, KOMODAKIS N. Paying more attention to Attention: Improving the performance of convolutional neural networks via attention transfer[C]//International Conference on Learning Representations(ICLR). Toulon:dblp, 2017:1-13.

[6] CHEN Yuntao, WANG Naiyan, ZHANG Zhaoxiang. DarkRank: Accelerating deep metric learning via cross sample similarities transfer [C]//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence(AAAI-18). New Orleans, Louisiana, USA: dblp,2018:2852-2859.

[7] CHO J H, HARIHARAN B. On the efficacy of knowledge distillation [EB/OL]. [2019]. https://arxiv.org/abs/1910.01348. DOI: 10.1109/ICCV.2019.00489.

[8] MIRZADEH S I, FARAJTABAR M, LI Yuan, et al. Improved knowledge distillation via teacher assistant: Bridging the gap between student and teacher[J]. arXiv preprint arXiv: 1909.11723,2019.

[9] HEO B, KIM J, YUN S, et al. A comprehensive overhaul of feature distillation [C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South):IEEE, 2019:1921-1930.

[10] SAPUTRA M, GUSMAO P, Y ALMALIOGLU, et al. Distilling knowledge from a deep pose regressor network[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019: 263-272.

[11] PHUONG M, LAMPERT C. Towards understanding knowledge distillation[C]//International Conference on Machine Learning (ICML). Long Beach, California:IEEE,2019:5142-5151.

[12] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers & distillation through attention[J]. arXiv preprint arXiv:2012.12877, 2020.

[13] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words:Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Advances in Neural Information Processing Systems. Long Beach,USA:NIPS Foundation,2017:5998-6008.

[15] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//2016 Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE,2016:770-778.

[16] TAN Mingxing, LE Q V. EfficientNet:Rethinking model scaling for Convolutional Neural Networks [C]// International Conference on Machine Learning (ICML). Long Beach, California:IEEE,2019:1-11.

[17] LU Qi, KUEN J, GU Jiuxiang, et al. Multi-scale aligned distillation for low-resolution detection[C]//Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA:IEEE, 2021:1-11.

[18] CHEN Yixin, CHEN Pengguang, LIU Shu, et al. Deep structured instance graph for distilling object detectors[J]. arXiv preprint arXiv:2109.12862, 2021.

[19] FAN Zhibo, MA Yuchen, LI Zeming, et al. Generalized few-shot object detection without forgetting[C] //Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, 2021:1-13.

[20] JING Yongcheng, YANG Yiding, WANG Xinchao, et al. Amalgamating knowledge from heterogeneous graph neural networks[C] // Computer Vision and Pattern Recognition (CVPR).Nashville, TN, USA: IEEE, 2021:15704-15713.

[21] TARVAINEN A, VALPOLA H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results[J]. arXiv preprint arXiv:1703.01780,2017.

[22] ZHANGG Y, XIANG T, HOSPEDALES T M, et al. Deep mutual learning[C]//IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City:IEEE,2018: 4320-4328.

[23] SZEGEDY C, LIU Wei, JIA Yangqing, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston:IEEE,2015:1-9.

[24] ZAGORUYKO S, KOMODAKIS N. Wide residual networks[J]. arXiv preprint arXiv:1605.07146, 2016.

[25] HUANG G, LIU Z, LAURENS V, et al. Densely connected convolutional networks[C]// arXiv preprint arXiv:1608.06993, 2016.

[26] SANDLER M, HOWARD A, ZHU M, et al. Inverted residuals and linear bottlenecks:Mobile networks for classification,detection and segmentation[J]. arXiv preprint arXiv:1801.04381,2018.

[27] XIE Saining, GIRSHICK R, DOLLáR P, et al. Aggregated residual transformations for deep neural networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, Hawaii:IEEE Computer Society, 2017:1-10.

[28] XIAO Han, RASUL K, VOLLGRAF R. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms [J]. arXiv preprint arXiv: 1708.07747, 2017.

[29] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images[R]. Toronto: University of Toronto, 2009.

[30] WANG Long.Deep-Communication-Learning[EB/OL].[2021]. https://github.com/mlimwxxnn/Deep-Communication-Learning.