浙江科技学院学报，第33卷第3期，2021年6月
Journal of Zhejiang University of Science and Technology
Vol. 33 No. 3, Jun. 2021

doi：10.3969/j.issn.1671-8798.2021.03.006

基于知识蒸馏的差异性深度集成学习

张锡敏，钱亚冠，马丹峰，郭艳凯，康　明
（浙江科技学院 理学院，杭州 310023）

摘　要：深度神经网络模型在图像识别、语音识别等领域表现出了优异的性能，但高性能的模型对计算资源提出了更高的要求，存在难以部署于边缘设备的问题，对此提出一种基于知识蒸馏的差异性深度集成学习。首先对成员模型进行知识蒸馏，然后使用余弦相似度作为损失函数的正则化项对成员模型进行集成，最后得到训练好的模型。在MNIST（Mixed National Institute of Standards and Technology）和CIFAR10（Canadian Institute for Advanced Research）数据集上的试验结果表明，基于知识蒸馏的差异性深度集成学习在压缩模型的同时将模型的分类准确率提升至83.58%，相较于未经蒸馏的原始模型，分类准确率提高了4%，在压缩模型的同时提高模型的泛化性能。基于知识蒸馏的差异性深度集成学习打破了模型的压缩必然以泛化性能为代价这一认知，为模型集成提供了新的研究思路。

关键词：知识蒸馏；差异性集成；深度神经网络
中图分类号：TP183　　**文献标志码**：A　　**文章编号**：1671-8798(2021)03-0220-07

Differential deep ensemble learning based on knowledge distillation

ZHANG Ximin, QIAN Yaguan, MA Danfeng, GUO Yankai, KANG Ming
(School of Sciences, Zhejiang University of Science and Technology, Hangzhou 310023, Zhejiang, China)

Abstract: Deep Neural Networks model has achieved significant progress in many tasks, such as image recognition and speech recognition. However, the high-performance model raises higher requirements for computing resources, and is difficult to deploy on edge devices. For this reason, the differential deep ensemble learning was proposed on the basis of knowledge distillation. Firstly the member model was distilled by knowledge, then the cosine similarity was used as the regularization term of the loss function for ensemble training, and finally the trained model was obtained. The experimental results on MNIST(Mixed National Institute of Standards and Technology) and CIFAR10(Canadian Institute for Advanced Research) data sets show that the differential deep ensemble learning based on knowledge distillation can compress the model and increase the classification accuracy of the model to 83.58%,

收稿日期：2020-08-20
基金项目：国家自然科学基金项目(61902082)；浙江省自然科学基金项目(LY17F020011)；浙江省公益技术应用研究计划项目(LGG19F030001)

通信作者：钱亚冠(1976—　)，男，浙江省嵊州人，副教授，博士，主要从事机器学习与大数据处理、对抗性机器学习研究。E-mail：Qianyaguan@zust.edu.cn。

4% higher than that of the original model without distillation, which means that the differential deep ensemble learning can compress the model and improve the generalization performance of the model. Differential deep ensemble learning based on knowledge distillation breaks the stereotype that model compression is inevitable at the cost of generalization performance, which provides a new research idea for model ensemble.

Keywords: knowledge distillation; differential ensemble; Deep Neural Networks

随着深度学习的兴起，深度神经网络模型在很多领域得到了应用，如图像识别[1-3]、语音识别[4]、自然语言处理[5-6]等。借助残差连接[2,7]及批处理归一化[8]等新兴算法来配合云计算中心强大的图形处理器(graphic processing unit，GPU)，训练出的深度神经网络模型的层级可以达到数千层。模型规模的扩大虽然能够带来运算性能上的提升，但是难以训练，需要占用大量的数据和计算资源。计算量的大幅度提升意味着对硬件计算能力、内存带宽及数据存储的要求更高。这给深度神经网络模型在边缘智能设备上的部署带来了极大的挑战，使得深度神经网络模型难以应用到嵌入式设备中(如人脸识别系统[9]、自动驾驶汽车[10]等)。为了适应移动计算、边缘计算，缩小模型规模，研究人员开始研究模型的压缩方法与策略。

Hinton 等[11]提出了知识蒸馏方法，通过将结构复杂的深度学习模型(教师模型)知识，迁移到结构简单的深度学习模型(学生模型)上，以实现深度学习模型的压缩。但是，学生模型的分类准确率低于教师模型，这意味着模型压缩以模型的分类性能为代价。如何在缩小模型规模的同时保持模型泛化能力成为一个具有挑战性的问题。为了解决这个问题，笔者提出一种基于知识蒸馏的差异性深度集成学习，将知识蒸馏与差异性集成相结合，在知识蒸馏压缩模型的基础上，使用集成来提高模型的分类性能。由于集成性能取决于成员模型间的差异性[12]，故从模型层面提出新的集成训练方法，构造新的损失函数，即添加余弦相似度作为衡量模型差异性的正则化项。此外，在模型集成的过程中，通过最小化损失函数来增强成员模型间的预测差异性。

1　蒸馏集成相关知识

1.1　知识蒸馏

深度神经网络模型是一种模仿人脑神经网络结构的深层学习模型，模型的输出可用 $\boldsymbol{f}(x,\theta)$ 表示。x 是输入变量，θ 是模型参数。$\boldsymbol{f}(x,\theta)\in\mathbb{R}^m$ 是一个 m 维概率向量，表示 m 个类的置信度。模型的最后一层采用 Softmax 层，定义为 $f_i=\exp(\boldsymbol{z}_i)/\sum_{j=1}^{m}\exp(\boldsymbol{z}_j)\mid_{i=1\cdots m}$，其中 $\boldsymbol{z}$ 是最后一个隐藏层的输出向量。模型最后的预测标签 $t=\underset{i=1\cdots m}{\operatorname{argmax}}\boldsymbol{f}_i$。Hinton[11] 认为模型获得的知识不仅存储于 K 个模型的模型参数集 $\{\theta^i\}_{i\in K}$ 中，还表现在模型的输出概率分布 $(\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_m)$ 中。基于这个观点，蒸馏就是从教师模型中提取类之间的概率结构，生成软标签，通过训练迁移知识到学生模型中。

知识蒸馏使用温度 T 调节 x 在教师模型 $\boldsymbol{f}_1(x,\theta)$ 上的输出概率向量 $\boldsymbol{y}_{\text{soft}}$，称为 x 的软标签：

$$\boldsymbol{y}_{\text{soft}}=(p_1,p_2,\cdots,p_m)。\tag{1}$$

式(1)中：$p_i=\exp(\boldsymbol{z}_i/T)/\sum_{j=1}^{m}\exp(\boldsymbol{z}_j/T)\mid_{i=1\cdots m}$，$T>0$。利用 x 的硬标签 $\boldsymbol{y}$ 与软标签 $\boldsymbol{y}_{\text{soft}}$ 同时对学生模型 $\boldsymbol{f}_2(x,\theta)$ 进行蒸馏训练：

$$\underset{\theta}{\operatorname{argmin}}(\alpha J(y,\boldsymbol{f}_2(x,\theta))+(1-\alpha)J(\boldsymbol{y}_{\text{soft}},\boldsymbol{f}_2(x,\theta))T^2)。\tag{2}$$

式(2)中：$J(\cdot)$为代价函数；$\alpha\in[0,1]$。可设置不同的温度值进行训练，T 的值越大意味着每个类之间的概率差异性越小。

1.2　模型集成

在单个模型的泛化能力提升有限的情况下，把多个弱模型集成起来。由于多个局部极小值的存在，

多个不同的模型会实现不同的概率分布，因此将单独训练的模型输出组合起来可提高性能，更好地泛化数据[13]，可获得比单个模型性能更强的分类模型[14]。常见的模型集成方法有简单平均法和带权平均法。简单平均法中的集成模型

$$F(x)=\frac{1}{K}\sum_{i}^{K}\boldsymbol{f}^{(i)}(x)\boldsymbol{f}^{(i)}(x)。\tag{3}$$

带权平均法中的集成模型

$$F(x)=\sum_{i}^{K}w_i\boldsymbol{f}^{(i)}(x)。\tag{4}$$

式(3)～(4)中：$\boldsymbol{f}^{(i)}(x)$为成员模型；K 为成员模型个数；$w_i\in[0,1]$。参考文献[15-16]的研究结果，本研究采用简单平均法进行集成。

2 基于知识蒸馏的神经网络模型集成

2.1 直接蒸馏集成

知识蒸馏作为典型的模型压缩方法，其最终目的是缩小模型规模，从而减少对计算资源的需求。但由于学生模型的规模变小，与教师模型相比分类性能更低，这意味着模型压缩以减弱模型的分类性能为代价。为了解决这个问题，考虑将多个学生模型进行直接集成，以更好地泛化数据。一方面，由于集成性能取决于成员模型间的差异性，成员模型间的差异性大就能获得比单个成员模型性能更强的分类模型[17]；另一方面，蒸馏后的学生模型会减小成员模型间的差异性[15]。可见，提升单个学生模型的蒸馏效果与提升直接蒸馏集成模型的泛化性能之间产生了矛盾[13]。对此，我们提出一种基于差异性蒸馏集成的训练方法，通过进一步增加模型间的预测差异性来实现更好的蒸馏模型集成效果。

2.2 差异性蒸馏集成

模型集成的性能主要取决于成员模型之间的差异性[12]。因此，增大成员模型之间的差异性是提高模型集成性能的有效方法。本文提出差异性蒸馏集成方法，保持单个模型蒸馏效果的同时，提高集成效果。为了保持模型的准确性，每个成员模型必须得到正确的输出。

使用余弦相似度来度量成员模型之间的差异性。假设 $\boldsymbol{f}^{(i)}=(\boldsymbol{f}_j^{(i)})_{j=1}^{m}$ 为第 i 个成员模型的输出概率分布，$\hat{\boldsymbol{f}}^{(i)}=(\boldsymbol{f}_j^{(i)})_{j=1,j\neq t}^{m}$ 为 $\boldsymbol{f}^{(i)}$ 除正确类 t 以外的概率分布。梯度$\nabla_x\hat{\boldsymbol{f}}^{(i)}$ 为$\hat{\boldsymbol{f}}^{(i)}$ 对输入变量 x 的预测方向。如果$\nabla_x\hat{\boldsymbol{f}}^{(1)}$ 与$\nabla_x\hat{\boldsymbol{f}}^{(2)}$ 的方向一致，那么成员模型$\hat{\boldsymbol{f}}^{(1)}$ 与$\hat{\boldsymbol{f}}^{(2)}$ 会以相似的方式变化，即$\hat{\boldsymbol{f}}^{(1)}$ 与$\hat{\boldsymbol{f}}^{(2)}$ 对x有相似的预测。显然$\hat{\boldsymbol{f}}^{(1)}$ 与$\hat{\boldsymbol{f}}^{(2)}$ 之间是一种低效集成。为了增强成员模型之间的差异性，需增大$\nabla_x\hat{\boldsymbol{f}}^{(1)}$ 与$\nabla_x\hat{\boldsymbol{f}}^{(2)}$ 间的差异性。

对于两个成员模型，使用余弦相似度来度量两者的差异性：

$$\cos(\nabla_x\hat{\boldsymbol{f}}^{(1)},\nabla_x\hat{\boldsymbol{f}}_j^{(2)})=\frac{\langle\nabla_x\hat{\boldsymbol{f}}^{(1)},\nabla_x\hat{\boldsymbol{f}}^{(2)}\rangle}{|\nabla_x\hat{\boldsymbol{f}}^{(1)}|\cdot|\nabla_x\hat{\boldsymbol{f}}^{(2)}|}。\tag{5}$$

式(5)中：$\cos(\cdot)\in[-1,1]$。若 $\cos(\nabla_x\hat{\boldsymbol{f}}^{(1)},\nabla_x\hat{\boldsymbol{f}}^{(2)})=-1$，则$\nabla_x\hat{\boldsymbol{f}}^{(1)}$ 与$\nabla_x\hat{\boldsymbol{f}}^{(2)}$ 完全不一致，$\boldsymbol{f}^{(1)}$ 与 $\boldsymbol{f}^{(2)}$ 对除正确类 t 以外的预测差异性达到最大。对于多个成员模型，可以取最大的余弦相似度：

$$M(\{\nabla_x\hat{\boldsymbol{f}}^{(i)}\}_{i=1}^{K})=\max_{i,j\in1,\cdots,K}\cos(\nabla_x\hat{\boldsymbol{f}}^{(i)},\nabla_x\hat{\boldsymbol{f}}^{(j)})。\tag{6}$$

最大化 $M(\{\nabla_x\hat{\boldsymbol{f}}^{(i)}\}_{i=1}^{K})$能够增大成员模型对不同类别的预测差异性。然而，式(6)是一个非光滑函数，无法使用一阶优化方法，因此我们对式(6)采用 LogSumExp 函数进行光滑逼近：

$$L(M(\{\nabla_x\hat{f}^{(i)}\}_{i=1}^{K}))=\log_{10}\Big(\sum_{1\leqslant i\leqslant j\leqslant K}\exp(M(\{\nabla_x\hat{f}^{(i)}\}_{i=1}^{K}))\Big)。\tag{7}$$

如果 K 个成员模型的$L(M(\{\nabla_x\hat{\boldsymbol{f}}^{(i)}\}_{i=1}^{K}))$较小，那么成员模型对不同类别的预测差异性就较大，所集成训练的模型 $F(x)$则具有更强的泛化能力。为了使集成模型具有较小的 $L(M(\{\nabla_x\hat{\boldsymbol{f}}^{(i)}\}_{i=1}^{K}))$，将$L(M(\{\nabla_x\hat{\boldsymbol{f}}^{(i)}\}_{i=1}^{K}))$作为正则化项加入集成模型的交叉熵损失函数中：

$$L=-\sum_{K}\log_{10}\boldsymbol{f}^{(i)}+\lambda L(M(\{\nabla_x\hat{\boldsymbol{f}}^{(i)}\}_{i=1}^{K}))。\tag{8}$$

式(8)中：λ 为超参数控制在训练过程中 $L(M(\{\nabla_x \hat{\boldsymbol{f}}^{(i)}\}_{i=1}^{K}))$ 的权重，λ 值越大，$\{\boldsymbol{f}^i\}_{i\in K}$ 的差异性就越大。因此，差异性集成训练转化为如下优化过程：

$$\min_{\theta}\left[-\sum_{K}\log_{10}\boldsymbol{f}(i)+\lambda L(M(\{\nabla_x \hat{\boldsymbol{f}}^{(i)}\}_{i=1}^{K}))\right]。\tag{9}$$

式(9)中：$\theta=\{\theta^k\}_k\in K$ 为所有成员模型的参数。差异性蒸馏集成训练流程如图 1 所示。

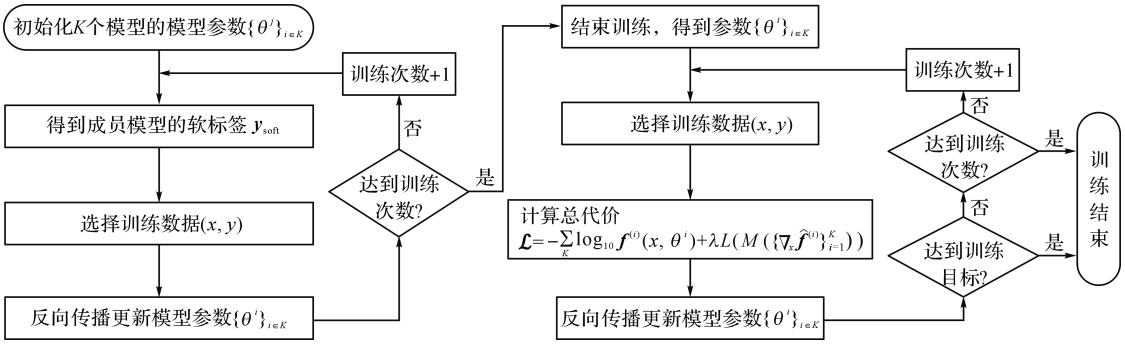


图 1　差异性蒸馏集成训练流程

Fig. 1　Differential distillation ensemble training process

3　试验结果

3.1　数据集与网络结构

本文试验数据集采用 MNIST(Mixed National Institute of Standards and Technology)[18] 和 CIFAR10(Canadian Institute for Advanced Research)[19]。MNIST 是一个被广泛应用于机器学习性能测试的手写体数据集，由从数字 0 到 9 的 10 个类组成，共计 70 000 张手写数字图像，包括 60 000 张图像作为训练数据及 10 000 张图像作为测试数据，每个图像为 2 828 像素的单通道灰度图像。CIFAR10 数据集由 10 个类 60 000 张大小为 3 232 像素的三通道彩色图像组成，包括 50 000 张训练图像和 10 000 张测试图像。

在 MNIST 数据集上训练 LeNet[20] 教师模型，设置最小批次大小为 128，经过 50 个批次的训练后，达到 98.79%的分类准确率。在 CIFAR-10 数据集上训练 AlexNet[21] 教师模型，设置最小批次大小为 40，经过 70 个批次的训练后达到 76.97%的分类准确率。教师模型和学生模型的 LeNet 和 AlexNet 网络结构分别见表 1 和表 2。另外，两个模型使用 RMSProp(root mean aquare prop)优化算法[22] 训练网络，初始学习率 η 设置为 0.000 1，迭代次数设置为 3 000 次。

表 1　教师模型和学生模型的 LeNet 网络结构

Table 1　LeNet network structure of teacher model and student model

网络层类型	神经元个数	
	LeNet 教师模型	LeNet 学生模型
卷积层 1	556	3 316
池化层 1	22	22
卷积层 2	5 516	—
池化层 2	22	—
全连接层 1	120	100
全连接层 2	84	100
全连接层 3	10	—
Softmax 层	10	10

表 2　教师模型和学生模型的 AlexNet 网络结构

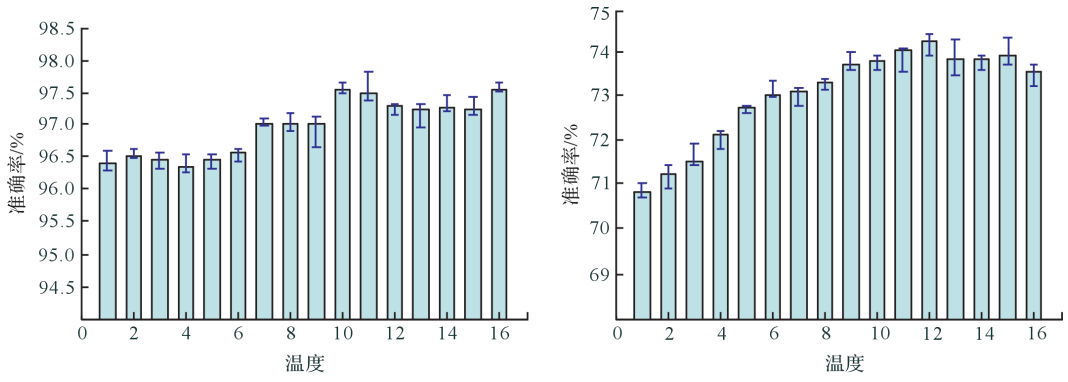
Table 2　AlexNet network structure of teacher model and student model

网络层类型	神经元个数	
	AlexNet 教师模型	AlexNet 学生模型
卷积层 1	111 148	3 364
卷积层 2	55 128	3 364
卷积层 3	33 192	—
卷积层 4	33 192	—
卷积层 5	33 128	—
全连接层 1	2 048	256
全连接层 2	2 042	—
Softmax 层	10	10

3.2　试验对比

3.2.1　直接蒸馏集成训练的有效性

首先蒸馏训练不同温度下的单个 LeNet 学生模型和 AlexNet 学生模型。图 2 为 LeNet 学生模型和 AlexNet 学生模型的不同温度下直接蒸馏集成训练效果。在不同温度下蒸馏出 10 个学生模型，记录 10 个分类准确率的最高值、最低值及平均值。温度为 0 时为正常训练的学生模型。



(a) LeNet学生模型 (b) AlexNet学生模型

图 2 不同温度下直接蒸馏集成训练

Fig. 2 Direct distillation ensemble training at different temperatures

已知 LeNet 教师模型的分类准确率为 98.79%，AlexNet 教师模型的分类准确率为 76.97%。由图 2 可知，在不同的温度下，LeNet 学生模型与 AlexNet 学生模型的分类准确率都低于各自对应的教师模型，其原因是学生模型网络结构更简单，表达能力弱于教师模型，无法容纳教师模型迁移过来的全部知识。因此，需考虑集成多个学生模型以提升分类准确率。

在 AlexNet 模型上采用不同蒸馏设置进行集成，将集成分为四类：1)未蒸馏集成，对多个教师模型进行集成；2)同源同温集成，对同一教师模型蒸馏的学生模型进行集成($T=22$)；3)同源异温集成，对同一教师模型蒸馏的学生模型进行集成($T=1,20,\cdots,100$)；4)异源同温集成，对不同教师模型蒸馏的学生模型进行集成($T=22$)。

在 AlexNet 上进行的不同蒸馏设置的直接蒸馏集成试验结果如图 3 所示。由图 3 可知，随着模型个数的增加，未蒸馏的模型集成的分类准确率持续上升，其原因是未蒸馏的模型集成由容量更大的教师模型集成得到。与此相比，随着模型个数的增加，直接蒸馏集成模型的分类准确率虽有提高，但始终低于未蒸馏的集成模型。

为了解直接蒸馏集成模型分类准确率不高的原因，需从学生模型差异性角度进行研究。将 $T=1$、20、30 蒸馏得到的 3 个 AlexNet 学生模型作为观测对象，以 CIFAR-10 中的一张图片为例，记录其在不同学生模型中的输出。图 4 为 CIFAR10 数据集上不同学生模型的概率输出。试验结果表明，随着温度的升高，蒸馏得到的学生模型的概率分布变平滑，学生模型间的差异性被减弱。可见，模型之间的差异性影响了模型集成的效果，直接蒸馏集成训练无法提高分类准确率。对此，我们提出差异性蒸馏集成训练来增强学生模型间的差异性，以增强集成性能。

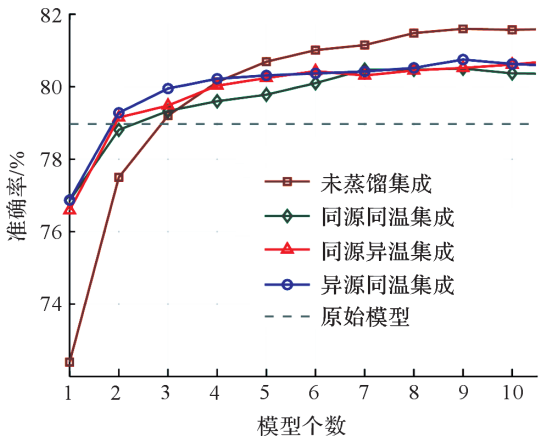


图 3 AlexNet 不同蒸馏设置的直接蒸馏集成训练试验结果

Fig. 3 Results of direct distillation ensemble training with AlexNet different distillation settings

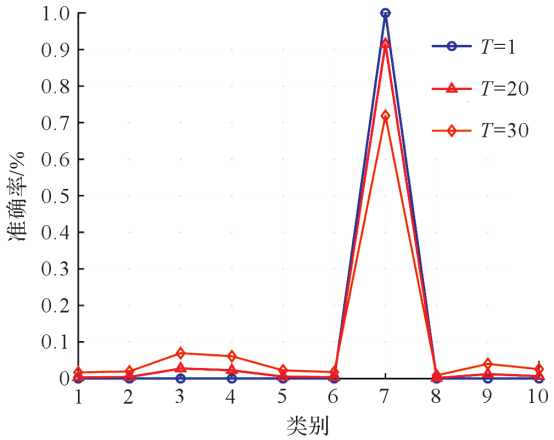


图 4 CIFAR10 数据集上不同学生模型的概率输出

Fig. 4 Probability output of different student models on CIFAR10 dataset

3.2.2 差异性蒸馏集成训练的有效性

通过试验验证差异性蒸馏集成训练的有效性，图 5 为 3 种情况的差异性蒸馏集成训练结果。由图 5 可知，差异性蒸馏集成训练可以有效增强模型的泛化性能。随着模型个数的增加，直接蒸馏集成的分类

准确率趋于平稳且低于教师模型集成，而差异性蒸馏集成的分类准确率则不断上升且显著高于教师模型集成和直接蒸馏集成。这说明差异性蒸馏集成不但能缩小模型规模，而且能提升模型的泛化性能。

考虑到成员模型的分类准确率和成员模型间的差异性决定集成性能[12]，我们对成员模型的集成准确率 A 和集成差异性 D 进行量化。集成准确率定义为被所有成员模型分类正确的样本占样本总量的比例。模型间集成差异性采用 Lu 等[22]提出的双误差(double fault，DF)作为衡量成员模型间差异性的指标，定义为所有成员模型分类错误的样本占样本总量的比例。A 和 D 的定义分别如下：

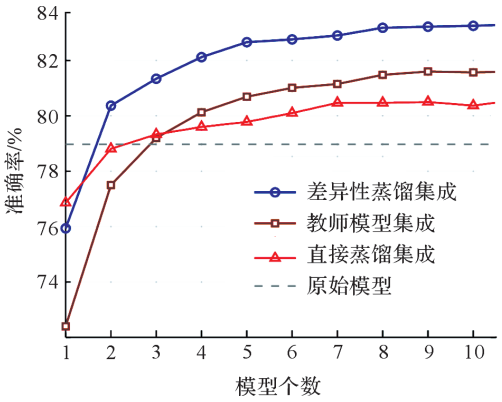


图 5　差异性蒸馏集成训练结果

Fig. 5　Result of differential distillation ensemble training

$$A:\frac{a}{N}; \tag{10}$$

$$D:=1-\frac{b}{N}。 \tag{11}$$

式(10)和式(11)中：a 为所有成员模型分类正确的样本数量；b 为所有成员模型分类错误的样本数量；N 为样本总量。不同集成设定下模型集成准确率和集成差异性见表 3。其中，同源同温集成、同源异温集成、异源同温集成都属于直接蒸馏集成。以教师模型的分类准确率为基线可以发现，差异性蒸馏集成的分类准确率及集成差异性都高于教师模型集成和直接蒸馏集成。因此，我们可以得出结论，差异性蒸馏集成可以在缩小模型规模的同时提升集成准确率。

表 3　不同集成设定下模型集成准确率和集成差异性

Table 3　Model ensambal accuracy rate and ensemble difference under different ensemble settings

模型集成方法	集成准确率/%	集成差异性
教师模型集成	74.82	0.559
同源同温集成	74.12	0.465
同源异温集成	74.09	0.472
异源同温集成	74.06	0.478
差异性蒸馏集成	75.32	0.599

上述试验都建立在相同规模的模型上。进一步地，我们在不同规模的模型上验证差异性蒸馏集成训练的有效性：采用的模型为不同层数的 ResNet，设置通道大小为 16、32、64；教师模型使用 Res26，学生模型使用 Res20、Res16、Res8；最小批次都设置为 256。教师模型在 CIFAR-10数据集上训练 80 个批次，学习率设置为 0.01，动量为 0.9，权重衰减为 0.000 1。在此基础上对批次进行调整，教师模型在 CIFAR-10 数据集上训练 60 个批次，学习率设置为 0.01，动量为 0.9，权重衰减为 0.000 1。图 6 为 Res20、Res16 和 Res8 的差异性蒸馏集成训练结果。试验发现，随着学生模型个数的增多，3 种学生模型的差异性蒸馏集成准确率逐渐升高，规模越大的模型分辨准确率越高，并且都高于原始教师模型。由此可以得出结论，差异性蒸馏集成训练模型不受模型规模的约束，能在小规模学生网络的情况下得到优于教师网络的集成模型。

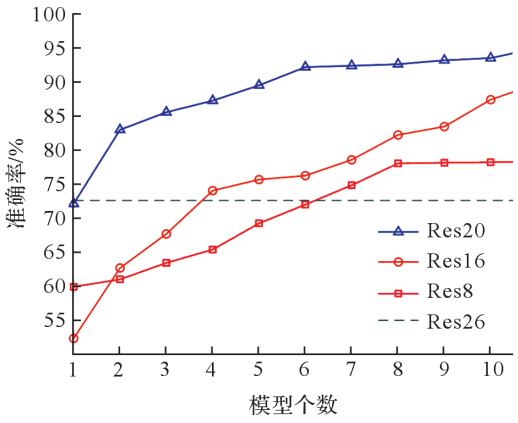


图 6　Res20、Res16 和 Res8 的差异性蒸馏集成训练

Fig. 6　Differential distillation ensemble training of Res20, Res16 and Res8

4　结　语

本文提出的基于知识蒸馏的差异性深度集成学习，通过在集成过程中利用余弦距离增大成员模型的预测差异性，在缩小模型规模的同时提升了成员模型集成后的分类准确率。在集成训练过程中，我们使用新的损失函数在同一个数据集上同时交互式地训练所有的成员模型。在 MNIST 和 CIFAR10 数据集

上的试验结果表明,本文提出的基于知识蒸馏的差异性深度集成学习优于将教师模型直接集成或将学生模型直接蒸馏集成。可见,基于知识蒸馏的差异性深度集成学习更适用于计算资源有限的边缘计算环境,有助于边缘智能系统的应用推广。未来的工作中,我们将进一步探索模型压缩对泛化性能的影响,以提升集成模型的泛化能力。

参考文献:

[1] LI Y C, ZHOU R G, XU R Q, et al. A quantum deep convolutional neural network for image recognition[J]. Quantum Science and Technology,2020,5(4):4375.

[2] 孙佳佳,吕飞,雷晨曦,等.基于图像识别的有害生物检疫鉴定探索研究[J].植物检疫,2020,34(5):42.

[3] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04)[2020-08-20]. https://arxiv.org/abs/1409.1556.

[4] HÜLSMEIER D, WARZYBOK A, KOLLMEIER B, et al. Simulations with FADE of the effect of impaired hearing on speech recognition performance cast doubt on the role of spectral resolution[J]. Hearing Research,2020,39(15):107995.

[5] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C]//Advances in Neural Information Processing Systems. Montreal: NIPS,2014:3104.

[6] WANG Z W, SHE Q, SMEATON A F, et al. Synthetic-neuroscore: using a neuro-AI interface for evaluating generative adversarial networks[J]. Neurocomputing,2020,405(10):26.

[7] 翟翔宇,杨风暴,吉琳娜,等.标准化全连接残差网络空战目标威胁评估[J].火力与指挥控制,2020,45(6):39.

[8] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift [EB/OL]. (2015-02-11)[2020-08-20]. https://arxiv.org/abs/1502.03(6).

[9] 强宇佶,申双琴.智能家居嵌入式人脸识别门禁系统的设计与实现[J].科学技术创新,2020(26):112.

[10] 张施晶,唐天宇,张志明,等.基于CNN-PID的竞赛机器人赛道识别与自主驾驶[J].试验技术与管理,2020,37(9):39.

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

[12] HANSEN L K, SALAMON P. Neural network ensembles[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1990,12(10):993.

[13] ISLAM M M, YAO X, MURASE K. A constructive algorithm for training cooperative neural network ensembles [J]. IEEE Transactions on Neural Networks,2003,14(4):820.

[14] TRAMÈR F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: attacks and defenses[EB/OL]. (2020-05-19)[2020-08-20]. https://arxiv.org/abs/1705.07204.

[15] FURLANELLO T, LIPTON Z C, TSCHANNEN M, et al. Born again neural networks[C]//Proceedings of Machine Learning Research. Stockholm: PMLR,2018:1607.

[16] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: CVPR,2018:9185.

[17] PANG T Y, XU K, DU C, et al, Improving adversarial robustness via promoting ensemble diversity[EB/OL]. (2020-05-29)[2020-08-20]. https://arxiv.org/pdf/1901.08846.

[18] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE,1998,86(11):2278.

[19] KRIZHEVSKY A. Learning multiple layers of features from tiny images[J]. Handbook of Systemic Autoimmune Diseases,2009,1(4):13.

[20] MUKKAMALA M C, HEIN M. Variants of RMSProp and adagrad with logarithmic regret bounds[C]// Proceedings of Machine Learning Research. Sydney: PMLR,2017:2545.

[21] ZHU X H, NI Z W, NI L P, et al. Spread binary artificial fish swarm algorithm combined with double-fault measure for ensemble pruning[J]. Journal of Intelligent and Fuzzy Systems,2019,36(5):4375.

[22] LU H J, AN C L, ZHENG E H, LU Y. An adaptive and momental bound method for stochastic learning[J]. Neurocomputing,2014,128:22.