第 9 卷 第 3 期
2020 年 5 月
集 成 技 术
JOURNAL OF INTEGRATION TECHNOLOGY
Vol. 9 No. 3
May 2020

引文格式：
陈晨, 王亚立, 乔宇. 任务相关的图像小样本深度学习分类方法研究 [J]. 集成技术, 2020, 9(3): 15-25.
Chen C, Wang YL, Qiao Y. Task-relevant few-shot image classification [J]. Journal of Integration Technology, 2020, 9(3): 15-25.

任务相关的图像小样本深度学习分类方法研究

陈 晨[1,2] 王亚立[1] 乔 宇[1]

[1](中国科学院深圳先进技术研究院 深圳 518055)
[2](中国科学院大学深圳先进技术学院 深圳 518055)

摘 要 传统基于度量学习的图像小样本分类方法与任务无关，这导致模型对新查询任务的泛化能力较差。针对该问题，该研究提出一种任务相关的图像小样本深度学习方法——可以根据查询任务自适应地调整支持集样本特征，从而有效形成任务相关的度量分类器。同时，该研究通过引入多种正则化方法，解决了数据量严重不足所带来的过拟合问题。基于 miniImageNet 和 tieredImageNet 两个常用标准数据集，在特征提取网络相同的前提下，所提出方法在 miniImageNe 中 1-shot 上获得了 66.05% 的准确率，较目前最好的模型提高了 4.29%。

关键词 任务相关；特征嵌入；正则化；度量学习；小样本分类
中图分类号 TP 399 **文献标志码** A doi: 10.12146/j.issn.2095-3135.20200402001

Task-Relevant Few-Shot Image Classification

CHEN Chen[1,2] WANG Yali[1] QIAO Yu[1]

[1](*Shenzhen Institutes of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)
[2](*Shenzhen College of Advanced Technology*, *University of Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)

Abstract The traditional metric learning based few-shot image classification methods are task independent, which leads to poor generalization performance of the model on new query tasks. To solve this problem, a task-relevant image few-shot learning method was proposed in this paper, which can adaptively adjust the feature of support samples according to the query task. Moreover, a variety of regularization methods to address the overfitting problem under severely-limited data scenarios were also investigated. We conduct comprehensive experiments on two popular benchmarks, i.e., miniImageNet and tieredImageNet. The result of 1-shot task on the miniImageNet by the proposed method was 66.05%, and it outperforms the SOTA (state of the art)

收稿日期：2020-04-02 **修回日期**：2020-04-23
基金项目：国家重点研发计划项目(2016YFC1400704)；国家自然科学基金-深圳机器人基础研究中心项目(U1713208)
作者简介：陈晨，硕士研究生，研究方向为计算机视觉、深度学习和小样本图像分类等；王亚立，博士，副研究员，硕士研究生导师，研究方向为计算机视觉、深度学习和行为识别等；乔宇，博士，研究员，博士研究生导师，研究方向为计算机视觉、深度学习、行为识别、场景识别、人脸识别和目标检测等，E-mail：yu.qiao@siat.ac.cn。

approaches by 4.29% under the same backbones.

Keywords task-relevant; feature embedding; regularization; metric learning; few-shot classification

1 引 言

近年来，深度学习(Deep Learning)[1]在计算机视觉领域的发展不断获得突破和成功，如其不断刷新图像识别、目标检测等领域的最优结果。深度学习是一项数据驱动的技术，其性能严重依赖标注数据的数量。然而，大量标注数据的收集存在多种挑战。一方面，在诸如医疗、安全等特定领域，由于涉及隐私或国家安全等问题，数据的采集受到严格限制；另一方面，大规模数据的采集、清洗和标注需要耗费大量的人力和物力。因此，如何使用少量样本训练深度网络，成为亟待解决的问题。受到人类从少量数据中快速学习能力的启发，Li 等[2]提出小样本学习(Few-Shot Learning)的概念。其目的是使在已知类别(Seen Class)中训练的分类模型，面对只有少量标注数据的未知类别(Unseen Class)依然具有较好的性能。

目前，小样本学习已成为深度学习领域中非常重要的前沿研究问题，在医疗图像分析等数据采集难度较大的领域具有十分广阔的应用前景。如何在图像小样本的数据上训练得到一个泛化性能较好的模型，是一个既有理论意义又具有实际应用价值的研究课题。国内外研究学者提出的常见解决方法主要有两种：一种是基于元学习(Meta Learning)[3-6]的方法，另一种是基于度量学习(Metric Learning)[7-13]的方法。

元学习是机器学习的一个子领域，其目标为使模型学会学习。例如，Santoro 等[3]提出使用记忆增强的方法来解决小样本识别问题。该方法利用权重更新来调节偏差，使模型学会通过将表达快速缓存到记忆中来调节输出。由于传统的梯度下降方法(如 Adagrad[14]、Adadelta[15]、Adam[16]等)需要选取众多超参数，无法在几步内完成优化，Finn 等[4]提出 MAML(Model-Agnostic Meta-Learning)，通过找到一个模型参数的更加敏感状态，使模型能快速地迁移到新的任务上。Rusu 等[5]提出 LEO(Learning Embedding Optimization)方法，通过构造隐层空间解决 MAML 不能很好处理高维数据的问题。Ravi 等[6]提出一个基于长短期记忆网络(Long Shot Term Memory，LSTM)[17]的元学习器(Meta Learner)模型，利用 LSTM 来代替梯度下降算法的更新规则。该方法通过学习一个通用的初始化方式，使得模型在新任务上可以从一个好的初始状态开始训练。以上方法的共同点是通过在训练集上学习到的元知识，帮助模型很好地泛化到新的任务上。

度量学习则是将分类问题转化为样本间的相似性度量问题。其主要思路是将图像映射到更具有区分性的特征空间中，然后通过比较待分类样本和已标注样本在特征空间中距离的远近，来预测待分类样本的类别。一个更有区分性的特征空间应该具备这样的性质[8]：同类之间的图像特征嵌入距离较近，而不同类之间的图像特征嵌入距离较远。其中，距离的度量方式包括欧氏距离和余弦距离等。

度量学习的目标是学习一个具有较好泛化能力的图像到特征空间的映射。例如，Vinyals 等[7]提出匹配网络(Matching Networks)和任务片段式训练(Task Episode Training)方法，同时使用余弦距离对样本进行分类。Snell 等[8]提出原型网络(Prototypical Networks)，用原型作为一个类别在特征空间中的表示，并使用欧式距离来

进行分类。Sung 等[9]提出一个可学习的度量表示——关系网络(Relation Network)，与简单的欧式距离或余弦距离相比，它能够更好地表示样本之间的相似关系。Liu 等[10]提出传导传播网络(Transductive Propagation Network，TPN)，通过图构造模块来表征新类别数据的流形结构，学习如何将标签从已标记的支持集样本传播到未标记的查询集样本。Gidaris 等[11]提出去噪自编码器图神经网络(Denoising Autoencoders Graph Neural Network，wDAE-GNN)，基于已知类别的分类参数的数据分布，利用降噪编码器同时重建已知类别的分类参数和小样本未知类别的参数分布。除任务片段式训练方法外，也可以先用整个训练集做预训练，再迁移到小样本类别上。例如，Qiao 等[12]提出 PFA(Predicting Parameters from Activations)，使用激活函数输出层的倒数生成小样本类别对应的参数。Gidaris 等[13]提出 DFVL(Dynamic Few-shot Visual Learning)，通过小样本权重生成器生成对应类别的参数。其中，小样本权重生成器由训练集类别权重和余弦相似度相乘得到。值得一提的是，这些基于度量学习的方法都是任务无关的。由于训练样本量过少，任务无关带来的后果是模型容易过拟合已知类别，而对新类别上的查询任务泛化能力不足[7-8]。

本文在传统度量学习方法的基础上，提出任务相关的特征嵌入模块来抑制过拟合，引导模型充分地利用任务的信息，可根据查询任务自适应地调整支持集样本的特征嵌入，使得从已知类别上学习的从图像到特征的映射在未知类别上也具有很好的泛化性。其中，任务相关的特征嵌入模块没有引入庞大的参数或复杂的计算，很好地控制了模型的复杂度，避免过拟合问题。同时该模块具有很强的扩展性，可以在大部分基于度量学习的方法中方便地引入，以提高特征嵌入的可区分性。同时，本文还引入了多种正则化方法，解决数据量较少带来的过拟合问题，提高小样本图像分类的性能。最终通过对不同方法的结果进行对比和分析，验证了本文所提出方法的有效性。

2　任务相关的小样本深度学习方法

小样本学习存在两个重要的问题：(1)已知类别和未知类别之间没有交集，导致它们的数据分布差别很大，不能直接通过训练分类器和微调的方式得到很好的性能；(2)未知类别只有极少量数据(每个类别仅 1 个或 5 个训练样本)，导致分类器学习不可靠。根据小样本图像分类的特点，本文的小样本图像分类采用了任务片段式训练与测试。具体而言，分为元训练和元测试两个阶段。

在元训练阶段，本文从训练数据集中采样出多个 C-way K-shot 的训练任务。具体地，从所有类别中随机选取 C 类，且每类选取 K 个样本及对应标签作为支持集 $S_{\text{train}}=\{s_i,l_i\}_{i=1}^{C\times K}$。其中，$s_i$ 为样本，$l_i\in\{0,1\}^C$ 为独热编码(One-Hot Encoding)的标签。随后，从这 C 类中除了上述 $C\times K$ 张图像以外，再选取 N 个样本和对应的标签作为查询集 $Q_{\text{train}}=\{q_i,y_i\}_{i=1}^{N}$。其中，$q_i$ 为样本，$y_i\in\{0,1\}^C$ 为独热编码的标签。

在元测试阶段，虽然会面对与元训练阶段完全不同的类别，但仍采用与元训练阶段类似的方法——从测试数据集中采样出多个 C-way K-shot 的测试任务(本文为 600 个)。其中，每个测试任务包含支持集 $S_{\text{test}}=\{s_i,l_i\in\{0,1\}^C\}_{i=1}^{C\times K}$ 和查询集 $Q_{\text{test}}=\{q_i,y_i\in\{0,1\}^C\}_{i=1}^{N}$。最终的测试准确率均值和 95% 置信区间由所有测试任务的准确率计算而得。

2.1　任务相关的特征嵌入模块

度量学习的核心问题是寻找一种更优的映

射，将图像嵌入到一个更有区分性的空间中。本研究希望这种映射不仅适用于已知类别，还要适用于未知类别。传统基于度量的小样本学习方法是任务无关的——支持集样本的特征嵌入只与该样本自身有关，而与查询任务无关。在已知类别上训练的模型，面对未知类别的查询任务时，无法自适应地调整支持集样本的特征嵌入方式。因此，模型的泛化性能会受到很大影响。

为解决上述问题，本文提出一种任务相关的特征嵌入模块(如图 1)，引导网络快速地从查询任务中获取有用信息，调整支持集样本在特征空间中的特征嵌入，从而使得特征空间更有区分性。将查询集样本和支持集样本输入基于卷积神经网络的特征提取网络 $\varphi(\cdot)$，可以得到相应的特征向量。本文将第 i 个支持集样本的特征向量记为 $\varphi(s_i)$，同时将第 j 个查询集样本的特征向量记为 $\varphi(q_j)$。任务相关的特征嵌入 $e_{i,j}$ 的计算公式如下：

$$e_{i,j}=\sigma\left[\varphi(s_i)\cdot\boldsymbol{U}+\varphi(q_j)\cdot\boldsymbol{W}\right] \tag{1}$$

其中，$\boldsymbol{U}$ 和 $\boldsymbol{W}$ 均为参数矩阵；$\sigma(\cdot)$为非线性激活函数。通过上述公式，网络可以学习到任务相关的支持集样本特征嵌入。

如图 1 所示，传统的任务无关的度量学习方法得到了支持集样本和查询集样本的任务无关的特征嵌入。在该特征空间中，查询集样本和各支持集样本间的距离很接近，因此没有很好的区分性。而在本文中，所提出的任务相关的特征嵌入模块，通过将支持集样本特征嵌入和查询集样本特征嵌入的拼接，随后接入卷积层，网络通过学习可以根据不同的查询集样本自适应地改变支持集样本的特征嵌入方式，使其面对未知类别的待分类样本也具有较好的泛化性能。

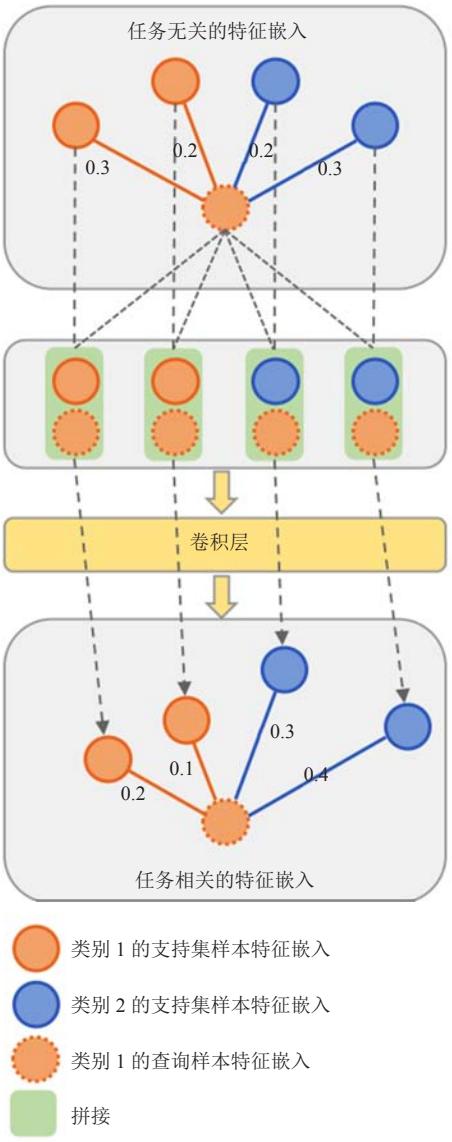

图 1 任务相关的特征嵌入模块

Fig. 1 Task-relevant feature embedding module

2.2 正则化模块

2.2.1 Mixup

Mixup[18]使用两个不同样本和对应标签的凸组合来训练深度神经网络。在传统的图像分类任务中，使用 Mixup 可以有效控制模型的复杂度，提高网络泛化能力。本文将其引入小样本分类任务中，以解决数据量严重缺乏带来的过拟合问题。与传统的图像分类不同的是，在基于度量学习的小样本分类中，查询样本是待分类的对象，而支持集样本是度量分类器的组成部分。因此，Mixup 的引入需要对查询样本的标签及度量分类器都做出适当的调整。

训练时，每次采样两组类别相同的 C-way

K-shot 任务：

$$T_1=\left\{S_1=\left\{s_i^1,l_i^1\right\}_{i=1}^{C\times K},\ Q_1=\left\{q_j^1,y_j^1\right\}_{j=1}^{N}\right\} \tag{2}$$

$$T_2=\left\{S_2=\left\{s_i^2,l_i^2\right\}_{i=1}^{C\times K},Q_2=\left\{q_j^2,y_j^2\right\}_{j=1}^{N}\right\} \tag{3}$$

其中，S 为支持集；s 为支持集样本图片数据；l 为 s 对应的独热编码标签；Q 为查询集；q 为查询集样本图片数据；y 为 q 对应的独热编码标签。

将两个任务的样本打乱后进行 Mixup，可以得到新的任务 T，其中的样本和标签定义如下：

$$T=\left\{S=\left\{s_i,l_i\right\}_{i=1}^{C\times K},Q=\left\{q_j,y_j\right\}_{j=1}^{N}\right\} \tag{4}$$

$$s_i=\lambda_i s_i^1+\left(1-\lambda_i\right)s_i^2 \tag{5}$$

$$l_i=\lambda_i l_i^1+\left(1-\lambda_i\right)l_i^2 \tag{6}$$

$$q_j=\lambda_j q_j^1+\left(1-\lambda_j\right)q_j^2 \tag{7}$$

$$y_j=\lambda_j y_j^1+\left(1-\lambda_j\right)y_j^2 \tag{8}$$

$$\lambda_i,\lambda_j\sim Beta\left(\alpha,\beta\right) \tag{9}$$

其中，λ_i 和 λ_j 为 Mixup 加权系数；α 和 β 为 *Beta* 分布的参数，通常取 $\alpha=\beta=1$。若给定特征提取网络 $\varphi(\cdot)$、查询集样本的特征嵌入和对应的 Mixup 标签为 $\left\{\varphi\left(q_j\right),y_j\right\}$，则支持集样本的任务相关特征嵌入及对应的 Mixup 标签为$\left\{e_{i,j},l_i\right\}$。其中，$e_{i,j}=\sigma\left[\varphi\left(s_i\right)\cdot\boldsymbol{U}+\varphi\left(q_j\right)\cdot\boldsymbol{W}\right]$。通过某种距离度量方式 $f(\cdot)$，如欧氏距离、余弦距离等，可以计算出 $e_{i,j}$ 与 $\varphi\left(q_j\right)$之间的距离 $d_{i,j}=f\left[e_{i,j},\varphi\left(q_j\right)\right]$。由于 Mixup 的引入，度量分类器的预测公式如下：

$$p_j=\sum_{i=1}^{C\times K}d_{i,j}l_i \tag{10}$$

其中，$p_j\in N^C$表示度量分类器预测的 q_j 属于每一类的概率分布。相应的交叉熵(Cross Entropy，CE)损失函数为：

$$L_{\mathrm{CE}}=-y_j^T\cdot\log\left(p_j\right) \tag{11}$$

Mixup 的引入相当于引入了多个度量分类器。这种集成学习的思想可以有效增强小样本分类的鲁棒性。

2.2.2 标签平滑

传统的图像分类任务标签是独热编码的，这种方式容易导致过拟合。在数据十分匮乏的小样本场景中，这种过拟合问题会变得更加严重。为缓解过拟合问题，本文使用标签平滑[19]的方法软化标签的编码。给定一个 C-way K-shot 的小样本图像分类任务，查询集样本 q_i 的独热编码标签可以表示为 $y_i\in\{0,1\}^C$，其中第 c 位定义如下：

$$y_i(c)=\begin{cases}1 & c\text{ 为正确类别}\\ 0 & \text{其他}\end{cases} \tag{12}$$

记平滑后的标签向量为 $\tilde{y}_i$，则其第 c 位为：

$$\tilde{y}_i(c)=\alpha y_i(c)+\frac{1-\alpha}{C} \tag{13}$$

其中，α 为标签平滑(Label Smoothing，LS)引入的超参数。假设模型预测查询样本 q_i 属于各类的概率分布为 p_i，则引入了标签平滑的交叉熵损失函数公式如下：

$$L_{\mathrm{LS-CE}}=-\tilde{y}_i^T\cdot\log\left(p_i\right) \tag{14}$$

3　实　验

3.1　数据集

本文选择在两个小样本学习领域中被广泛使用的标准数据集——miniImageNet[7]和tieredImageNet[20]上进行实验。miniImageNet 数据集是 ImageNet[21]数据集的一个子集，包含 100 类图片数据，每类包含 600 张图片。其中，训练集、验证集和测试集分别包含 64 类、16 类和 20 类。tieredImageNet 也是 ImageNet 数据集的一个子集，包含 608 类，共 779 165 张图片。其中，训练集、验证集和测试集分别包含 351 类、97 类和 160 类。由此可见，tieredImageNet 的训练集、验证集和测试集划分更加谨慎，从而确保每个集合中的类别差异更大。

3.2　模型结构

近年来出现了许多基于度量学习的小样本图

像分类方法，本文选取其中一个十分具有代表性的方法——原型网络[8]作为基准方法。该方法首先使用特征提取网络提取支持集样本和查询集样本的特征嵌入；然后，根据支持集样本的类别对支持集样本的特征向量求出均值向量，以此作为该类别的特征嵌入，接着计算查询集样本的特征嵌入与各类别的特征嵌入的欧式距离，并使用距离的相反数作为预测的类别分数；最后，在训练阶段使用带有指数归一化函数(Softmax)的交叉熵损失函数作为目标函数对模型进行优化，在测试阶段通过选取分数最大的类别进行预测。

在此基础上，本文引入了类别相关的特征嵌入模块。在使用特征提取网络提取支持集样本和查询集样本的特征嵌入后，将支持集样本的特征嵌入和查询集样本的特征嵌入拼接起来，输入卷积层得到新的特征嵌入。其余部分的处理与基准方法一致。

3.3 特征提取网络

为了更加公平地与当前最好的模型进行比较，本文使用了 3 种不同的特征提取网络——ConvNet[8]、ResNet[22]和 WideResNet[23]。

(1) ConvNet：ConvNet 由四层卷积模块组成，其中每个卷积模块由卷积层、批归一化层(Batch Normalization Layer)、Leakly ReLU 层和最大池化层(Max Pooling Layer)顺序连接组成。

(2) ResNet：深度残差网络(Deep Residual Networks)提出残差学习的概念，通过在层与层之间引入一个恒等连接，将上一层的输入与后面层的输出直接相加，很好地解决了卷积神经网络(CNN)随着层数加深而出现的性能退化问题。

(3) WideResNet：WideResNet(WRN)通过实验发现增加网络宽度(网络中每一层的卷积核个数)并减少网络深度(网络层数)，可以有效地提升模型性能并提升训练速度。

3.4 网络的训练和测试

首先，使用训练集的全部数据对特征提取网络进行预训练。然后，使用预训练模型作为初始化参数，进行元训练。其中，使用 Adam 优化器，初始学习率为 10^{-3}，每 15 000 次迭代学习率减半，权重衰减为 10^{-6}，标签平滑的超参数 $\alpha=0.2$。本文在训练中采用了随机改变大小、随机裁剪和随机水平翻转、颜色抖动(明度、对比度、饱和度和色相等变化)等数据增强方式。最后，在元测试阶段，采样 600 组 C-way K-shot 的测试任务，每组中查询集包含 15 个样本。并通过上述 600 组任务的准确率来计算准确率均值和 95% 置信区间。实验代码使用了深度学习框架 Pytorch[24]，并在单张 NVIDIA GeForce GTX Titan X GPU 上运行。

4 结果与讨论

本文在 miniImageNet 和 tieredImageNet 两个数据集上完成了 5-way 1-shot 和 5-way 5-shot 两种任务的实验，并将结果和目前最好的方法进行了充分对比，在测试集上的分类准确率结果如表 1 和表 2 所示。由于小样本分类的结果在相同的特征提取网络下才具有可比性，下面将针对各特征提取网络下的实验结果分别进行分析。

4.1 ConvNet

在使用 ConvNet 作为特征提取网络时，本文提出的方法在 miniImageNet 数据集上 1-shot 的分类准确率为 55.63%，5-shot 的为 71.87%；在 tieredImageNet 数据集上 1-shot 的结果为 62.32%，5-shot 的为 78.45%。与采用相同特征提取网络的 PFA 相比，在 miniImageNet 数据集上 1-shot 的结果提高了 1.10%，5-shot 的提高了 4.00%。PFA 先用整个训练集做预训练，然后固定前面特征提取层参数，通过仅训练最后的分类器层的参数来防止过拟合现象的发生。PFA 在训练中使用了 miniImageNet 80 类的数据，包括训练集 64 类、验证集 16 类。而本文方法仅使用 64 类训练集的数据，就明显超过了 PFA，表明任务

相关的特征嵌入模块以及多种正则化方法有利于减轻小样本数据下网络的过拟合现象。

与同样特征提取网络的 TPN 方法相比，本文方法在 miniImageNet 数据集上 1-shot 的结果提高 1.88%，5-shot 的提高 2.44%；在 tieredImageNet 数据集上 1-shot 的提高 4.79%，5-shot 的提高 5.60%。TPN 将全部无标签数据和有标签数据一起建立无向图连接，通过标签传播的方式得到无标签数据的标签。本文方法在类别差异更大的 tieredImageNet 上提升效果更加明显，说明本文方法对新类别泛化性能更好。

4.2　ResNet-12

在使用 ResNet-12 作为特征提取网络时，本文所提出方法在 miniImageNet 数据集上 1-shot 的结果为 63.39%，5-shot 的为 77.88%；在 tieredImageNet 数据集上 1-shot 的结果为 67.81%，5-shot 的为 83.26%。与同样特征提取网络的 DFVL 相比，在 miniImageNet 数据集上 1-shot 的结果提高 7.94%，5-shot 的提高 7.75%。DFVL 和 PFA 类似，两种方法都是先训练特征提取器，再固定特征提取器训练分类器。相比于本文端到端的训练过程，这种分两阶段的训练不利于效果的提升。

与同样特征提取网络的元优化网络支持向量机（MetaOptNet-SVM）[25]相比，本文在 miniImageNet 数据集上的 1-shot 任务结果提高 0.75%；在 tieredImageNet 数据集上 1-shot 的结果提高 1.82%，5-shot 的提高 1.70%。MetaOptNet-SVM 采用梯度下降联合特征提取一起联合训练，并把最后基于距离的分类器改进成线性分类器。本文提出的任务相关的特征嵌入模块实现简单，在 miniImageNet（1-shot）和 tieredImageNet（1-shot、5-shot）上具有更好的性能。

4.3　WRN-28-10

在使用 WRN-28-10 作为特征提取网络时，本文方法在 miniImageNet 数据集上 1-shot 的结果为 66.05%，5-shot 的为 81.72%。与同样特征提取网络的 LEO 相比，在 miniImageNet 数据集上 1-shot 的结果提高 4.29%，5-shot 的提高 4.13%。LEO 通过构造隐层空间把图像编码得到隐空间向量再解码得到分类参数，并把分类参数和隐空间向量使用 MAML 的方式进行训练。这种训练方法存在一定的不可控性，如梯度下降的步数很难选取，其中步数过多容易过拟合，步数过少则效果不够好。本文使用基于度量的方法，在得到任务相关的特征嵌入后，通过度量可以很简单地使用距离表达样本的相似性，最终得到的分类器也具有足够强的自适应能力。

本文方法在 tieredImageNet 上 1-shot 的结果为 68.96%，5-shot 的为 84.17%。与同样特征提取网络的 wDAE-GNN 方法相比，在 miniImageNet 数据集 1-shot 上的结果提高 0.78%，5-shot 上提高 1.08%。wDAE-GNN 的缺点在于，降噪编码器需要同时重建已知类别的参数分布和小样本未知类别的参数分布，而模型并未考虑到已知类别和未知类别的任务差异，因此在重建过程中小样本未知类别的参数分布和真实的参数分布会产生偏差。从实验结果对比可以发现，本文提出的任务相关的特征嵌入模块可以有效解决上述问题，从而显著提高模型性能。

综上，在特征提取网络相同的前提下，本文方法在 miniImageNet 和 tieredImageNet 上的结果都超过目前的其他方法。这验证了本文所提出的任务相关的特征嵌入模块以及多种正则化方法在提升模型的泛化性能方面的有效性。

4.4　消融实验

为了进一步探究任务相关的特征嵌入模块与各种正则化方法本身的效果，本文还使用 ResNet-12 作为特征提取网络，在 miniImageNet 数据集上进行充分的消融实验，结果如表 3 和表 4 所示。

从表 3 可以看出，不包含任何本文所提出

表 1 MiniImageNet 数据集上的结果对比

Table 1 Comparison with SOTA (state of the art) on miniImageNet dataset

方法	主干网络	准确率（%）	
		1-shot	5-shot
Matching Net[7]	32-32-32-32	43.44±0.77	55.31±0.73
Reptile[26]	32-32-32-32	47.07±0.26	62.74±0.37
MAML[4]	32-32-32-32	48.70±1.84	63.10±0.92
Prototypical Net[8]	64-64-64-64	46.61±0.78	65.77±0.70
Spot and Learn-CS[27]	64-64-64-64	51.03±0.78	67.96±0.71
TPN(trans)[10]	64-64-64-64	53.75±0.00	69.43±0.00
PFA[12]	64-64-64-64	54.53±0.40	67.87±0.20
RelationNet[9]	64-96-128-256	50.40±0.80	65.30±0.70
R2-D2[28]	96-192-384-512	51.80±0.20	68.40±0.20
DFVL[13]	ResNet-12	55.45±0.89	70.13±0.68
SNAIL[29]	ResNet-12	55.71±0.99	68.88±0.92
TADAM[30]	ResNet-12	58.50±0.30	76.70±0.30
MTL[31]	ResNet-12	61.20±1.80	75.50±0.80
Variational[32]	ResNet-12	61.23±0.26	77.69±0.17
MetaOptNet-SVM[25]	ResNet-12	62.64±0.61	78.63±0.46
TEAM(trans)[34]	ResNet-18	60.07±0.00	75.90±0.00
CTM[35]	ResNet-18	62.05±0.55	78.63±0.06
PFA[12]	WRN-28-10	59.60±0.41	73.74±0.19
wDAE-GNN[11]	WRN-28-10	61.07±0.15	76.75±0.11
LEO[5]	WRN-28-10	61.76±0.08	77.59±0.12
Ours	64-64-64-64	55.63±0.67	71.87±0.55
Ours	ResNet-12	63.39±0.55	77.88±0.37
Ours	WRN-28-10	66.05±0.52	81.72±0.31

注：主干网络一栏中 64-64-64-64 表示在第 1、2、3、4 层卷积层中的滤波器的数量

表 2 TieredImageNet 数据集上的结果对比

Table 2 Comparison with SOTA on tieredImageNet dataset

方法	主干网络	准确率（%）	
		1-shot	5-shot
MAML[4]	32-32-32-32	51.67±1.81	70.30±0.08
Prototypical Net[8]	64-64-64-64	53.31±0.89	72.69±0.74
TPN(trans)[10]	64-64-64-64	57.53±0.00	72.85±0.00
RelationNet[9]	64-96-128-256	54.48±0.93	71.32±0.78
MetaOptNet-SVM[25]	ResNet-12	65.99±0.72	81.56±0.53
CTM[35]	ResNet-18	64.78±0.11	81.05±0.52
LEO[5]	WRN-28-10	66.33±0.05	81.44±0.09
wDAE-GNN[11]	WRN-28-10	68.18±0.16	83.09±0.12
Ours	64-64-64-64	62.32±0.75	78.45±0.58
Ours	ResNet-12	67.81±0.61	83.26±0.40
Ours	WRN-28-10	68.96±0.57	84.17±0.36

注：主干网络一栏中 64-64-64-64 表示在第 1、2、3、4 层卷积层中的滤波器的数量

方法的基准实验 1-shot 和 5-shot 的结果分别为 55.27%、71.45%，明显较低，说明在这种情况下模型存在较为严重的过拟合问题。数据增强模块的引入给 1-shot 和 5-shot 任务分别带来了 4.55% 和 3.78% 的性能提升。这表明，本文针对小样本学习训练数据不足的问题所采用的随机改变大小、随机裁剪和随机水平翻转、颜色抖动等数据增强方式，有效扩大了训练数据的规模，同时增加了训练样本的多样性，提高模型网络对同一类样本不停变换的适应性，从而使模型学习到更加本质的特征。可见，数据增强从数据层面解决过拟合问题，提高模型的泛化能力。标签平滑的策略将 1-shot 和 5-shot 的结果分别提升了 0.92% 和 0.63%。可以看出，相比于独热编码，平滑的标签可以有效提高模型的泛化能力。Mixup 训练方式的引入进一步将准确率提高 1.34% 和 1.05%，证明使用训练样本的线性插值进行训练，可以约束模型的复杂度，减轻数据稀少所带来的过拟合问题。任务相关的特征嵌入模块将结果提升了 1.31% 和 0.97%，最终使 1-shot 和 5-shot 的准确率达到 63.39% 和 77.88%。这表明根据查询任务主动调整支持集样本的特征嵌入，可以帮助模型使用在已知类别上学到的元知识，快速地迁移到新的任务中。

为了更进一步探究每个部分单独剥离后对网络性能的影响，本文进行了充分的对比实验，结果如表 4 所示。首先，探究了数据增强、标签平滑以及 Mixup 三种不同的正则化方式对网络性能的影响。从表 3 可以看出，相比于不包含任何本文所提出方法的基准实验，数据增强模块给 1-shot 和 5-shot 任务分别带来了 4.55% 和 3.78% 的性能提升。而在表 4 的实验结果中，单独剥离数据增强模块在 1-shot 和 5-shot 任务上性能分别降低 1.25% 和 1.03%。显然，单独剥离数据增强模块造成的性能损失程度并没有表 3 实验中在基准模型上引入该模块所带来的增益程度高。这说明，标签平滑和 Mixup 两个模块起到了明显的正则化作用，弥补了剥离数据增强模块所带来的负面影响。类似地，单独剥离标签平滑模块会造成模型在 1-shot 和 5-shot 任务上性能分别降低 0.66% 和 0.33%，单独剥离 Mixup 模块会造成模型在 1-shot 和 5-shot 任务上性能分别降低 0.97% 和 0.74%。可以发现，当缺失某一种正则化方式时，网络性能并没有表 3 中增加模块引起的变化那样剧烈，说明这些正则化方式之间存在一定的互补性。其次，进行了单独剥离任务相关的特征嵌入模块的实验。正如对表 3 中结果的分析，去除任务相关的特征嵌入模块使得模型在 1-shot 和 5-shot 任务上的结果降低了 1.31% 和 0.97%。这表明任务相关的特征嵌入模块对于网络的泛化性能至关重要。它可以有效地引导模型在已知类别上学习到有用的元知识，使其在新的任务中可以快速利用元知识对支持集样本特征进行调整。

表 3　依次引入各模块对性能的影响

Table 3　Effect on performance after introduce each module in sequence

网络模块				准确率（%）	
数据增强	标签平滑	Mixup	任务相关	1-shot	5-shot
/	/	/	/	55.27±0.58	71.45±0.39
√	/	/	/	59.82±0.52	75.23±0.37
√	√	/	/	60.74±0.53	75.86±0.36
√	√	√	/	62.08±0.58	76.91±0.40
√	√	√	√	63.39±0.55	77.88±0.37

注：“√”为包含该模块，“/”为不包含该模块

表 4 剥离各模块对性能的影响

Table 4 Effect on performance after remove each module

网络模块				准确率（%）	
数据增强	标签平滑	Mixup	任务相关	1-shot	5-shot
√	√	√	√	63.39±0.55	77.88±0.38
/	√	√	√	62.14±0.57	76.85±0.35
√	/	√	√	62.73±0.53	77.55±0.33
√	√	/	√	62.42±0.59	77.14±0.36
√	√	√	/	62.08±0.56	76.91±0.35

注：“√”为包含该模块；“/”为不包含该模块

5 结 论

针对现有基于度量的图像小样本深度学习方法与任务无关，容易造成模型过拟合已知类别，而对新类别上的查询任务泛化能力不足等问题。本文提出一种新颖的任务相关的小样本深度学习方法，帮助模型根据查询任务，自适应地调整支持集样本的特征，从而有效形成任务相关的度量分类器。同时，本文引入多种正则化方法，进一步地提升了模型的泛化性能。实验结果表明，这些方法可以有效地解决网络的过拟合问题，提升小样本图像分类的准确率。在实际的数据样本中，除了少量已标注好的样本外，还有很多未标注的样本。这是因为在实际的医疗、安全等领域中，完全标注需要很大的人力成本。因此，小样本半监督分类具有很强的应用价值和实际价值，未来可以将本文提出的方法扩展到半监督的小样本分类问题中。

参 考 文 献

[1] Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets [J]. Neural Computation, 2006, 18(7): 1527-1554.

[2] Li FF, Fergus R, Perona P. One-shot learning of object categories [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(4): 594-611.

[3] Santoro A, Bartunov S, Botvinick M, et al. One-shot learning with memory-augmented neural networks [Z/OL]. arXiv:1605.06065, 2016.

[4] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks [C] // Proceedings of the 34th International Conference on Machine Learning, 2017: 1126-1135.

[5] Rusu AA, Rao D, Sygnowski J, et al. Meta-learning with latent embedding optimization [C] // Proceedings of the 7th International Conference on Learning Representations, 2019.

[6] Ravi S, Larochelle H. Optimization as a model for few-shot learning [C] // Proceedings of the International Conference on Learning Representations, 2017, https://openreview.net/pdf?id=rJY0-Kcll.

[7] Vinyals O, Blundell C, Lillicrap T, et al. Matching networks for one shot learning [M] // Advances in Neural Information Processing Systems, 2016: 3630-3638.

[8] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning [M] // Advances in Neural Information Processing Systems, 2017: 4077-4087.

[9] Sung F, Yang YX, Zhang L, et al. Learning to compare: relation network for few-shot learning [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 1199-1208.

[10] Liu YB, Lee J, Park M, et al. Learning to propagate labels: transductive propagation network for few-shot learning [Z/OL]. arXiv:1805.10002v5, 2019.

[11] Gidaris S, Komodakis N. Generating classification weights with GNN denoising autoencoders for few-shot learning [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 21-30.

[12] Qiao SY, Liu CX, Shen W, et al. Few-shot image recognition by predicting parameters from activations [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7229-7238.
[13] Gidaris S, Komodakis N. Dynamic few-shot visual learning without forgetting [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4367-4375.
[14] Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization [J]. The Journal of Machine Learning Research, 2011, 12: 2121-2159.
[15] Zeiler MD. ADADELTA: an adaptive learning rate method [Z/OL]. arXiv:1212.5701v1 2012, 2012.
[16] Kingma DP, Ba J. Adam: a method for stochastic optimization [C] // Proceedings of the 3rd International Conference on Learning Representations, 2015.
[17] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.
[18] Zhang HY, Cissé M, Dauphin YN, et al. Mixup: beyond empirical risk minimization [C] // Proceedings of the 6th International Conference on Learning Representations, 2018.
[19] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2818-2826.
[20] Ren M, Triantafillou E, Ravi S, et al. Meta-learning for semi-supervised few-shot classification [C] // Proceedings of the 6th International Conference on Learning Representations, 2018.
[21] Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009: 248-255.
[22] He KM, Zhang XY, Ren SQ, et al. Deep residual learning for image recognition [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.
[23] Zagoruyko S, Komodakis N. Wide residual networks [C] // Proceedings of the British Machine Vision Conference, 2016.
[24] Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library [M] // Advances in Neural Information Processing Systems 32, 2019: 8024-8035.
[25] Lee KJ, Maji S, Ravichandran A, et al. Meta-learning with differentiable convex optimization [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 10657-10665.
[26] Nichol A, Achiam J, Schulman J. On first-order meta-learning algorithms [Z/OL]. arXiv:1803.02999v3, 2018.
[27] Chu WH, Li YJ, Chang JC, et al. Spot and learn: a maximum-entropy patch sampler for few-shot image classification [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 6251-6260.
[28] Bertinetto L, Henriques JF, Torr PHS, et al. Meta-learning with differentiable closed-form solvers [C] // Proceedings of the 7th International Conference on Learning Representations, 2019.
[29] Santoro A, Raposo D, Barrett DG, et al. A simple neural network module for relational reasoning [M] // Advances in Neural Information Processing Systems 30, 2017: 4967-4976.
[30] Oreshkin B, López PR, Lacoste A. TADAM: task dependent adaptive metric for improved few-shot learning [M] // Advances in Neural Information Processing Systems 31, 2018: 721-731.
[31] Sun QR, Liu YY, Chua TS, et al. Meta-transfer learning for few-shot learning [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 403-412.
[32] Zhang J, Zhao CL, Ni BB, et al. Variational few-shot learning [C] // Proceedings of the IEEE International Conference on Computer Vision, 2019: 1685-1694.
[33] Lee KJ, Maji S, Ravichandran A, et al. Meta-learning with differentiable convex optimization [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 10657-10665.
[34] Qiao LM, Shi YM, Li J, et al. Transductive episodic-wise adaptive metric for few-shot learning [C] // Proceedings of the IEEE International Conference on Computer Vision, 2019: 3603-3612.
[35] Li HY, Eigen D, Dodge S, et al. Finding task-relevant features for few-shot learning by category traversal [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 1-10.