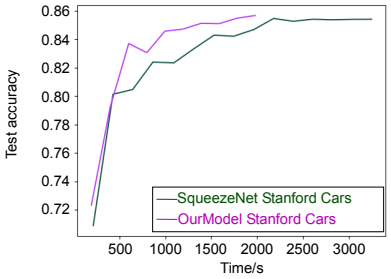
Opto-Electronic Engineering
光 电 工 程

Article

2019 年，第 46 卷，第 6 期

DOI: 10.12086/oee.2019.180416

基于深度迁移学习的微型细粒度图像分类



汪荣贵，姚旭晨，杨　娟*，薛丽霞

合肥工业大学计算机与信息学院，安徽 合肥 230601

摘要：现有的细粒度分类模型不仅利用图像的类别标签，还使用大量人工标注的额外信息。为解决该问题，本文提出一种深度迁移学习模型，将大规模有标签细粒度数据集上学习到的图像特征有效地迁移至微型细粒度数据集中。首先，通过衔接域定量计算域间任务的关联度。然后，根据关联度选择适合目标域的迁移特征。最后，使用细粒度数据集视图类标签进行辅助学习，通过联合学习所有属性来获取更多的特征表示。实验表明，本文方法不仅可以获得较高精度，而且能够有效减少模型训练时间，同时也验证了进行域间特征迁移可以加速网络学习与优化这一结论。

关键词：迁移学习；细粒度分类；深度学习；卷积神经网络

中图分类号：TP18　　**文献标志码：**A

引用格式：汪荣贵，姚旭晨，杨娟，等. 基于深度迁移学习的微型细粒度图像分类[J]. 光电工程，2019，**46**(6): 180416

Deep transfer learning for fine-grained categorization on micro datasets

Wang Ronggui, Yao Xuchen, Yang Juan*, Xue Lixia

School of Computer and Information, Hefei University of Technology, Hefei, Anhui 230601, China

Abstract: Existing fine-grained categorization models require extra manual annotation in addition to the image category labels. To solve this problem, we propose a novel deep transfer learning model, which transfers the learned representations from large-scale labelled fine-grained datasets to micro fine-grained datasets. Firstly, we introduce a cohesion domain to measure the degree of correlation between source domain and target domain. Secondly, select the transferrable feature that are suitable for the target domain based on the correlation. Finally, we make most of perspective-class labels for auxiliary learning, and learn all the attributes through joint learning to extract more feature representations. The experiments show that our model not only achieves high categorization accuracy but also economizes training time effectively, it also verifies the conclusion that the inter-domain feature transition can accelerate learning and optimization.

Keywords: transfer learning; fine-grained categorization; deep learning; convolutional neural network

Citation: Wang R G, Yao X C, Yang J, *et al.* Deep transfer learning for fine-grained categorization on micro datasets[J]. *Opto-Electronic Engineering*, 2019, **46**(6): 180416

收稿日期：2018-08-02；　收到修改稿日期：2018-12-20

作者简介：汪荣贵(1966-)，男，教授，博士生导师，主要从事智能视频处理与分析、视频大数据与云计算、智能视频监控与公共安全、嵌入式多媒体技术等的研究。E-mail：wangrgui@hfut.edu.cn

通信作者：杨娟(1983-)，女，博士，讲师，主要从事视频信息处理、视频大数据处理技术、深度学习与二进神经网络理论与应用研究等。E-mail：yangjuan6985@163.com

1 引 言

近年来，深度卷积神经网络在通用对象分类任务方面取得重大突破[1]。然而，在实际应用中，不仅需要从图像中识别出目标的基本类型，而且需要进一步识别出目标的子类型或者对目标进行更加精细的分类，通常称为细粒度图像分类。有别于通用对象分类问题，细粒度图像分类不仅可以对某一类别下的子类别进行分类，而且能够区分相似度极高的同一物种，例如对狗类的子类“哈士奇”和“爱斯基摩犬”进行区分。

同类别物种的不同子类往往仅在某些部件的外观上存在细微差别，一般需要依靠对象的姿态特征作为分类的先决条件。因此通常情况下训练细粒度分类模型首先从图像中提取特征，然后据此进一步构造出相应的派生特征来训练多级分类器。Donahue 等人[2]借助细粒度图像中物体标注框和部件标注训练出检测模型，并对得到的检测框添加位置几何约束。Branson 等人[3]提出利用部件标注的预测点来获取物体级别和部件级别的检测框，采用姿态对齐操作和不同层特征融合的方式。这些方法都取得了较好的细粒度分类效果，然而，由于物体部件标注点等信息需要大量的人工参与，这在一定程度上限制了上述强监督模型算法的实际应用。

目前，细粒度图像分类的一个明显趋势是使用弱监督细粒度分类模型，在模型训练时仅使用图像级别标签信息，而不再使用额外的部件标注信息。Simon 等人[4]利用卷积网络特征产生关键点，再利用这些关键点来提取局部区域信息，但是弱监督模型的分类精度与强监督模型的精度相比依然存在差距。由于多个任务可以通过使用包含在相关任务监督信号中的领域知识来改善泛化性能，即使对于优化目标只有一个的特殊情形，辅助任务仍然可以改善主任务的学习性能，所以本文拟用额外的样本标签对细粒度图像原属性进行分类以解决由于姿态、视点等差异引起的细粒度数据集类内差异较大的问题。

基于深度卷积神经网络的分类任务通常需要使用大量的训练样本，以此来避免过拟合，而且细粒度数据的类别标注通常需要专业知识。因此，获取大量有标记样本成本巨大，很多数据集仅有几千张，有些数据集甚至只有几百张图片。本文着重研究稀缺样本下的细粒度图像分类问题，将样本量稀缺的细粒度数据集称为微型细粒度数据集。

研究表明[5]，利用学习目标和已有知识之间的相关性，可以把知识从已有的模型和数据中迁移到要学习的目标上。迁移学习能够学习到领域无关的特征表达，这与深度学习不谋而合，将两者结合可以充分利用神经网络图像表征的能力，学习域不变的特征表示。Tzeng 等人[6]通过共享卷积神经网络进行特征自适应，将源域中的类关系迁移到目标域。Ge 等人[7]使用低级别特征在大规模标记数据集中搜索最近邻，进行特殊类型的传输学习。深层网络较低层内核提取低级别特征，而这些低级别特征的质量却决定了网络特征的质量。又有研究表明[8]，卷积神经网络前几层所提取的基本上是一般性特征，进行特征的迁移学习效果比较好。所以本文设计将微型细粒度数据集放入模型的目标域中，将源域网络中间层提取的特征迁移至目标域用于微型细粒度数据集的训练。值得注意的是，虽然此种迁移学习效果取决于源域与目标域之间任务的相似程度，但细粒度数据集之间却具有特征相似性。因此，本文充分利用这种细粒度数据集之间的关联性，引入衔接域的概念，提出一种基于相似度匹配的方法对域间任务的关联程度进行度量。本文工作具体如下：

1) 本文提出一种基于相似度匹配的方法，对于源域和目标域的关联程度进行度量；

2) 本文冻结部分网络层提取细粒度样本的特征表示进行域间迁移，从而达到加速网络学习与优化的目的；

3) 本文使用额外的样本标签对细粒度图像原属性进行分类以解决由于姿态、视点等差异引起的细粒度数据集类内差异较大的问题。

2 本文方法

传统的机器学习一般通过人类先验知识将未加工数据预处理成特征，再对特征进行分类。由于分类结果取决于特征的好坏，所以长期以来机器学习专家将大部分时间花费在设计特征上。而深度学习是多层次的特征提取器与识别器统一训练和预测的网络，因此，端到端的训练显得尤为重要，本文模型的设计理念就是充分发挥卷积神经网络自身能够进行端到端处理的优势。本文旨在进行端到端的联合调整，通过对嵌入跨数据集图像信息的学习，捕捉更难的跨领域知识，以尽量减少源域任务和目标域任务中原始函数的损失。对于域间迁移网络问题，本文将网络层数较少的卷积神经网络的部分层冻结，部分层权值共享来提取源域细粒度数据集中明确的属性特征以迁移至目标

域。由于姿态、视点等差异引起的细粒度数据集类内变化较大的问题，本文拟用额外的样本标签对细粒度原属性进行分类。

针对迁移学习中源域任务与目标域任务关联程度度量和细粒度图像分类任务中样本量稀缺这两个问题，本文提出了一个基于关联度度量的深度迁移学习模型，模型包括两个部分：1) 域间任务关联度度量阶段，详细论述将在 2.1 节给出；2) 特征迁移适应性调节阶段，详细论述将在 2.2 节给出。

2.1 域间任务关联度度量

迁移学习研究最大局限之一是其学习能力很大程度上取决于源域任务与目标域任务的关联程度，若两个领域之间没有任何的相似部分则对其进行知识的迁移显然是无效的。但一般情况下，源域与目标域之间直接共享少量特征，迁移学习通过一系列的辅助概念将两个域连接起来[5]。不失一般性，本论文引入衔接域的概念，如图 1 所示，衔接域作为源域与目标域之间的桥梁实现领域间任务的关联度度量。直观地说，若源域中分类任务比目标域更难，那么从源域数据中学习到的模型具有高度的预测性，并且能在目标域上实现高性能。另一方面，若衔接域能扩大源域和目标域之间的距离，那么源域和目标域之间知识迁移的过程将会减少信息丢失。

故对于定量计算源域任务与目标域任务之间的关联程度，本文制定了一个用于学习图片相似性度量的判别式训练方法。该方法无需输入所有样本进行训练，适用于源域数据集类别数量较多的情况，其相似性度量可用于匹配未在先前类别中出现的样本。从当前目标域所取用的数据集中随机选一定比例的样本输入衔接域，与此同时，从源域的每个数据集中选取同样数量的样本输入衔接域。

首先，对目标域中每个类别的数据集选取相同数量的 c 个样本。然后，根据目标域中每个类别的数据集样本总量，选取一定样本量的源域样本输入衔接域进行两域之间关联程度的判断性训练，样本量为 $c\times$(源域每类数据集样本量/目标域每类数据集样本量)。将成对输入衔接域训练的源域样本和目标域样本分别定义为 $X_{\rm S}$ 和 $X_{\rm T}$，定义一个二元标签 P，当 P=0 时，$X_{\rm S}$ 和 $X_{\rm T}$ 属于同一类型样本为正对；否则 $X_{\rm S}$ 和 $X_{\rm T}$ 为负对，P=1。设 $W_{\rm I}$ 是学习的共享参数，设 $G_{W_{\rm I}}(X_{\rm S})$ 和 $G_{W_{\rm I}}(X_{\rm T})$ 为低维空间中由映射 $X_{\rm S}$ 和 $X_{\rm T}$ 产生的特征向量，衔接域中系统可以看作是一个用来衡量 $X_{\rm S}$ 和 $X_{\rm T}$ 之间兼容性的标量能量函数 $E_{W_{\rm I}}(X_{\rm S},X_{\rm T})$，定义为

$$E_{W_{\rm I}}(X_{\rm S},X_{\rm T})=\|G_{W_{\rm I}}(X_{\rm S})-G_{W_{\rm I}}(X_{\rm T})\| \ 。 \tag{1}$$

给定训练集中正对 $(X_{\rm S},X_{\rm T})$ 与负对 $(X'_{\rm S},X_{\rm T})$，当其满足以下条件时衔接域中系统合理化运行：存在一个大于零的数 e，使得

$$E_{W_{\rm I}}(X_{\rm S},X_{\rm T})+{\rm e}<E_{W_{\rm I}}(X'_{\rm S},X_{\rm T}) \ , \tag{2}$$

成立，其中正数 e 为余量。

本文定义 $L_{\rm p}$ 和 $L_{\rm N}$ 分别为正对损失和负对损失，

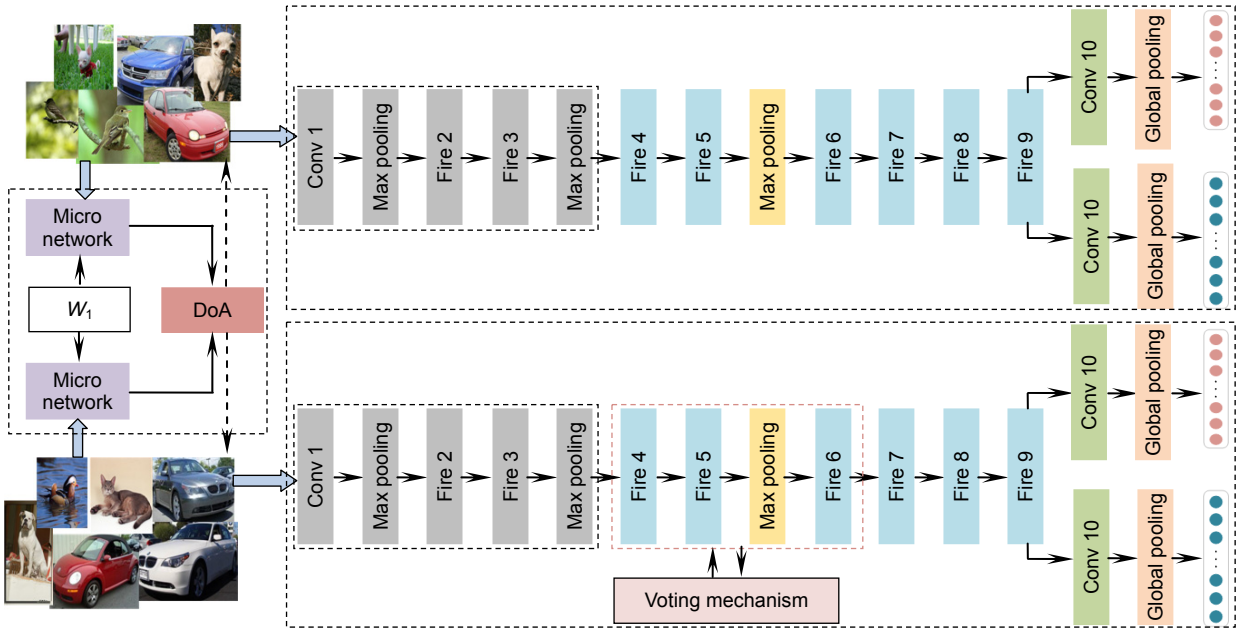


图 1 本文总体框架

Fig. 1 Overall view of network architecture

最小化损失时期望总体损失可以减少正对能量，增加负对能量[9]，即损失 L_{P} 单调增加，损失 L_{N} 单调减小，所以定义总体损失为

$$L(E_{W_1}(X_{\mathrm{S}},X_{\mathrm{T}}),E_{W_1}(X'_{\mathrm{S}},X_{\mathrm{T}})) = L_{\mathrm{P}}(E_{W_1}(X_{\mathrm{S}},X_{\mathrm{T}}))+L_{\mathrm{N}}(E_{W_1}(X'_{\mathrm{S}},X_{\mathrm{T}}))\ 。 \quad (3)$$

由于只使用小样本进行特征相似性的度量，所以图像的局部特征显得尤为重要。传统的卷积神经网络利用单独的线性滤波器来检测相同概念的不同变化，然而对于单个概念而言，拥有过多的滤波器需要考虑前层所有变化的组合，此种做法会给下一层造成负担。较高层的滤波器能映射到原始输入的更大区域，可以通过结合底层的低级概念产生更高层次的概念，所以微型网络将每个位置的块结合到更高层次的概念之前先将每个块做一个很好的抽象。此时特征提取的过程相当于在一个普通的卷积上级联交叉通道参数，每个输出层都在输入特征图上执行加权线性重组后通过整形线性单元，跨通道汇集的特征图在下一层中反复汇集。本文使用非线性替代线性模型，将上述微型网络作为衔接域中的特征提取器，结构如图 2 所示。微型网络由两层的多层感知卷积层和一个全局平均池化层组成。使用多层感知器主要是因为其反向传播训练与卷积神经网络的结构相一致，并且多层感知器本身是一个深层模型可以进行特征重用。多层感知器对成对输入的样本卷积过程如下：

$$f^1_{a,b,k_1} = \max(\omega^{1\ \mathrm{T}}_{k_1}x_{a,b}+b_{k_1},0)\cdots f^n_{a,b,k_n} = \max(\omega^{n\ \mathrm{T}}_{k_n}f^{n-1}_{a,b}+b_{k_n},0)\ , \quad (4)$$

其中：n 表示多层感知器的层数，$x_{a,b}$ 表示在位置 (a,b) 的输入块，k 表示特征图的通道。网络输出特征向量后进行相似性(用 S_{Sim} 表示)计算，其函数定义为

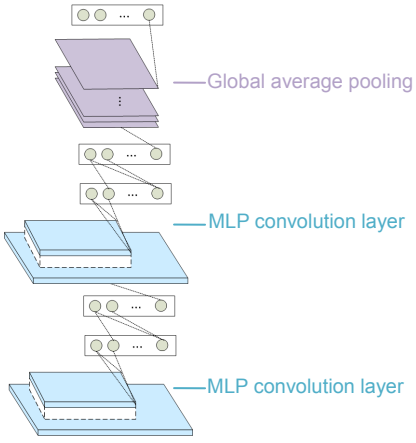


图 2 衔接域中微型网络示意图

Fig. 2 Micro network in cohesion domain

$$S_{\mathrm{Sim}}(D_{\mathrm{S}},D_{\mathrm{T}}) = \mu\frac{\sum_{j=1}^{n}(C_{\mathrm{S}}(j)\times C_{\mathrm{T}}(j))}{\sqrt{\sum_{j=1}^{n}(C_{\mathrm{S}}(j))^2\times\sum_{j=1}^{n}(C_{\mathrm{T}}(j))^2}} + \rho\frac{\sum_{g=1}^{m}(K_{\mathrm{S}}(g)\times K_{\mathrm{T}}(g))}{\sqrt{\sum_{g=1}^{m}(K_{\mathrm{S}}(g))^2\times\sum_{g=1}^{m}(K_{\mathrm{T}}(g))^2}}\ , \quad (5)$$

其中：μ，ρ 分别表示细粒度特征与视图类特征权重，给定包括源域、目标域的一个二元组 $A_{\mathrm{tr}}=\{\mathrm{S},\mathrm{T}\}$，本论文从两个域中分别提取三个视图类特征和三个细粒度特征，记为 $F=\{(C_{\mathrm{S}}(1),C_{\mathrm{T}}(1)),\cdots,(C_{\mathrm{S}}(3),C_{\mathrm{T}}(3))$，$(K_{\mathrm{S}}(1),K_{\mathrm{T}}(1)),\cdots,(K_{\mathrm{S}}(3),K_{\mathrm{T}}(3))\}$，再将各域中同类型特征两两做交叉域计算，共 36 个特征，前 12 个特征总结了单个域内特征，后 24 个特征捕获了成对交叉域距离，这些特征共同影响着本论文模型中的关联度学习算法。

然而，设计通用的领域关联程度判断标准是不现实的，因为对于不同的问题可能对这些特征具有不同的权重。所以本论文提出了对领域内图像特征的统一表达式：

$$Cov(A_{\mathrm{tr}}) = \varphi(\sum_i^{n=36}\alpha f_i+\beta)\ , \quad (6)$$

其中：$\varphi(x)=\dfrac{1}{1+\exp(-x)}$，$\alpha$，$\beta$ 表示常参数。

则可以将领域间的关联度(degree of association，DoA，用 D_{DoA} 表示)定义为

$$D_{\mathrm{DoA}} = \sum_{i=1}^{n}I^{(i)}\log Cov(A_{\mathrm{tr}}) +(1-I^{(i)})\log(1-Cov(A_{\mathrm{tr}}))\ , \quad (7)$$

其中：$I^{(i)}$ 是一个二进制标签，当视图类特征与细粒度特征同时大于阈值时，$I^{(i)}=1$；否则 $I^{(i)}=0$。衔接域是源域与目标域之间的一个桥梁，不仅增强了领域间的关联程度，而且使得目标域中的分类任务不只是由目标域内的特征来决定。关联度的输出会反馈给模型的特征迁移适应性调节阶段，第二阶段会根据第一阶段的反馈调整迁移网络的权重。

2.2 特征迁移适应性调节

本文使用属性标签在源域数据集中训练整个网络。由于卷积神经网络前几层提取的基本上是一般性特征，进行特征迁移学习效果比较好[8]，所以本论文选择固定网络的前三层所提取的特征来创建学习目标，这样做的目的是为了节约训练时间的同时保证特定特征的提取。为了兼顾模型的计算速度，本文选用 SqueezeNet[10]作为最终的特征迁移学习网络，该网络结构中大量采用 1×1 和 3×3 卷积核来提升速度，对于类似 caffe[11]这样的深度学习框架，在卷积层的前向计

算中，采用 1×1 卷积核直接利用通用矩阵乘法进行矩阵加速运算，可避免对索引图像块重排列为矩阵列的额外操作，在保证精度的同时使用最少的参数。

卷积神经网络在相邻层上以复杂且易受影响的方式进行交互[8]，这就使得相邻层间的神经元在训练时会产生共适应，然而这种共适应不仅仅是由上层学习得到，即冻结网络的某一层时，网络无法重新获取这种共适应。所以在深度卷积神经网络中进行特征迁移学习不仅要考虑网络高层中所提取特征通用性的下降，还要考虑中间层的相邻层间神经元共同适应性的下降。因此本文选取网络的中间层 4、5、6 作为特征迁移网络的冻结最高层，并制定了一个投票机制来对不同层的特征迁移进行适应性调节。

在模型第一阶段衔接域反馈损失为目标域选择相匹配的迁移网络之后，投票机制根据目标域中微型细粒度数据集的迁移网络，再适应性调节出适合每个数据集进行学习的分层参数。冻结 SqueezeNet[10]的 1~4 层、1~5 层、1~6 层作为竞争投票的候选者，每个候选者为同质个体且相互之间不存在依赖性关系，因此可以并行生成。进行预测的候选者类别为 $\{L_A, L_B, L_C\}$，对于任意一个目标域任务，给出分层预测的结果分别为 $\{h_A, h_B, h_C\}$。本文使用的投票机制是相对多数投票法，即三个候选者对目标域任务的预测结果中，数量最多的候选者作为最终的选择。若不止一个候选者获得最高票，则随机选取一个最高票候选者作为最终选择。在每个目标域任务执行过程中，循环迭代对候选者进行投票选择，聚合各数据集间完整的训练得到最终的迁移特征效果。表 1 给出了投票机制的候选者作为单一个体进行独立迁移学习的结果。

表 1 单一任务与增加辅助任务对比结果

Table 1 Categorization result comparisons between single-task and auxiliary-task

	Single-task	Auxiliary-task
BMVC[15]	97.4	97.7
BMW-10[16]	80	80.32
Oxford-IIIT Pet[17]	83.6	84.2
Birds[18]	99.8	99.8

Xie 等人[12]通过利用细粒度识别模型和超类标签识别模型之间的正则化来建立新的学习模型。受此启发，本论文利用来源于数据集本身且易于标注的视图类信息，视图类的类别标签如图 3 所示。给定细粒度与视图类有标签数据，通过共享共同特征和学习分类器来训练多个任务深度卷积神经网络，每个任务提供一个对应于其来源的函数，各域内任务本论文定义一个损失函数。

假设有 n 个训练样本 $\{(x_{1,j}^{(r)}, y_{1,j}^{(r)}), (x_{2,j}^{(r)}, y_{2,j}^{(r)}), \cdots, (x_{n,j}^{(r)}, y_{n,j}^{(r)})\}$，其输入特征量为 $x_{i,j}^{(r)} \in \Re^{m+1}$，细粒度类标记为 $y_{i,j}^{(r)} \in \{0,1,\cdots,c\}$，其中 $r \in \{\mathrm{S},\mathrm{T}\}$ 是一个双变量元组表示领域归属，S：源域，T：目标域。视图类样本表示为 $\{(x_{1,g}^{(r)}, v_{1,g}^{(r)}), (x_{2,g}^{(r)}, v_{2,g}^{(r)}), \cdots, (x_{n,g}^{(r)}, v_{n,g}^{(r)})\}$，视图类标记为 $v_{i,g}^{(r)} \in \{0,1,\cdots,k\}$。首先通过预测图像真实值 $G_{i,g}^{(r)}$ 在第 g 个视图类类别的属性后验概率，建立每个分类类别的属性值之间的互斥关系：

$$P(v_{i,g}^{(r)} = G_{i,g}^{(r)} | x_{i,g}^{(r)}; \theta) = \frac{\exp(\theta_g^{\mathrm{T}} x_{i,g}^{(r)})}{\sum_{l=1}^{K} \exp(\theta_l^{\mathrm{T}} x_{i,g}^{(r)})}, \quad (8)$$

其中 θ 表示视图类分类模型的权重，由于视图类可被

图 3 从四个不同视角对三个细粒度数据集检索的视图类标签

Fig. 3 Each category in perspective class

视为细粒度类别的一个隐变量，所以可以得出如下的预测函数：

$$P(y_{i,j}^{(r)}|x_{i,j}^{(r)})=\sum_{g=1}^{K}P(y_{i,j}^{(r)}\,|\,v_{i,g}^{(r)},x_{i,g}^{(r)})P(v_{i,g}^{(r)}\,|\,x_{i,g}^{(r)}) \quad , \quad (9)$$

其中：$P(v_{i,g}^{(r)}|x_{i,g}^{(r)})$ 表示输入图像属于任意一个视图类类别的概率，$P(y_{i,j}^{(r)}|v_{i,g}^{(r)},x_{i,g}^{(r)})$ 表示指定一个特定类别的视图类时该输入图像属于任意细粒度类别的概率，因此用一个分类器对其进行建模：

$$P(y_{i,j}^{(r)}=G_{i,j}^{(r)}|v_{i,g}^{(r)},x_{i,j}^{(r)};\theta')=\frac{\exp(\theta_j'^{\mathrm{T}}x_{i,j}^{(r)})}{\sum_{l'=1}^{C}\exp(\theta_{l'}'^{\mathrm{T}}x_{i,j}^{(r)})} \quad , \quad (10)$$

其中 θ' 表示特定视图类类别时细粒度模型的权重，将式(8)，式(10)整合后，可以得出：

$$P(y_{i,j}^{(r)}=G_{i,j}^{(r)}|x_{i,j}^{(r)})=\sum_{v=1}^{K}\left(\frac{\exp(\theta_j'^{\mathrm{T}}x_{i,j}^{(r)})}{\sum_{l'=1}^{C}\exp(\theta_{l'}'^{\mathrm{T}}x_{i,j}^{(r)})}\frac{\exp(\theta_g^{\mathrm{T}}x_{i,g}^{(r)})}{\sum_{l=1}^{K}\exp(\theta_l^{\mathrm{T}}x_{i,g}^{(r)})}\right) 。 \quad (11)$$

在损失函数中引入指示函数 $Z\{\cdot\}$，其具体形式为

$$Z\{\delta\}=\begin{cases}0, & \text{if} \quad \delta=\text{false}\\ 1, & \text{if} \quad \delta=\text{true}\end{cases} 。 \quad (12)$$

将域内任务数据集的总体损失函数表示为相等权重的损失平均的加法和计算：

$$L_{\theta,\theta'}=-\sum_{i=1}^{N}\sum_{g=1}^{K}Z\{v_{i,g}^{(r)}=G_{i,g}^{(r)}\}\log P(v_{i,g}^{(r)}=G_{i,g}^{(r)}|x_{i,g}^{(r)};\theta)$$
$$-\sum_{i=1}^{N}\sum_{j=1}^{C}Z\{y_{i,j}^{(r)}=G_{i,j}^{(r)}\}\log P(y_{i,j}^{(r)}=G_{i,j}^{(r)}\,|\,x_{i,j}^{(r)};\theta') 。 \quad (13)$$

原始的细粒度数据因为无法推断基于视点的类别，不足以学习每个视图类别的分类器。然而，研究表明权重 θ' 可以捕获与视图类型分类器相似的高级视图类特征，所以本文在参数 θ 与参数 θ' 之间引入正则化的表示：

$$R(\theta,\theta')=\frac{\sigma}{2}\sum_{g=1}^{K}\sum_{j=1}^{C}||\theta'-\theta||_2^2 。 \quad (14)$$

正则化有利于将知识迁移到每个视图类类别的分类器，从而有助于调整细粒度任务中的类内差异。

3 实 验

为了证明本文模型的普遍适用性，本文对细粒度分类图像样本进行了广泛搜集并进行了充分地实验。将 Stanford Cars[16]，Stanford Dogs[13]，CUB-200-2011[14] 三个标准的细粒度数据集作为源域数据集。将 BMVC[15]，BMW-10[16]，Oxford-IIIT Pet[17]，birds[18]作为目标域数据集，上述四个数据集均为微型细粒度图像数据集，其样本量相对于源域数据集来说较少，最少为几百张，最多为 Oxford-IIIT Pet[17]训练集(共 3680 张图片)。如图 3 所示，视图类标签从四个不同的视角对三个细粒度数据集进行检索：Stanford Cars[16]，车的头部、尾部、左侧、右侧；Stanford Dogs[13]，狗的正面、左面、右面、其它；CUB-2011-200[14]，鸟的正面、左面、右面、其它。

3.1 衔接域设置

如本文 2.1 节所述，对目标域中每种类的数据集各选取同等数量的 c 个样本，选取 $c\times$(源域每类数据集样本量/目标域每类数据集样本量)，数量的源域样本输入衔接域进行关联程度的判断性训练。在实际应用中，由于标注数据花费大量的人力和物力，很多图像分类问题并没有足够的标注数据，因此本文利用小样本进行特征相似性度量。小样本问题是统计量性质的一种刻画，它研究样本容量固定时，各种统计量的性质及由此进行的统计推断。利用小样本进行特征相似性度量具有良好的可操控性和普遍适用性。根据目标域中微型细粒度数据集数量，实验将 c 的值设置为 50，在目标域的四个数据集中分别取 50 张样本，在源域中取对应量的样本同样输入衔接域中进行判断性训练。在每个类别的目标域数据集进入特征迁移适应性调节之前，随机选取该数据集样本总量的 5%的数据输入衔接域，取相同数量的源域所有类别的数据集同时输入进行相似性的判断。

关联度度量后，将衔接域中损失反馈给第二阶段特征迁移适应性调节来为目标域选择相匹配的迁移网络参数。本文选取两层的多层感知器卷积网络作为判断性学习的网络，图像数据成对输入，利用一个二进制标签说明两类是否属于同一个类别，在 slice 层将两个图片分开各自得到一个输出向量。网络的最后是一个全局平均池化层，作为结构调整器明确执行特征映射以确定置信度。与全连接层相比，全局平均池化无需优化参数，这样可以有效防止过拟合，并且全局池化总结了空间信息，对输入的空间转换更加稳健。

3.2 源域设置

本文首先在 caffe[11]框架下利用 ImageNet ILSVRC 数据集对模型进行预训练，用于获取良好的初始化参数。利用基础网络固定前三层的权重参数分别训练源域细粒度数据集。为了获取源域中的视图类标签，本文在三个细粒度数据集四个不同的视角对图像进行检索分类，分类结果如图 3 所示。

本文模型选用 SqueezeNet[10]作为最终的迁移网络，挤压比(squeeze 卷积层滤波器数量/expand 层中滤波器数量)设置为 0.125，将 3×3 滤波器在 expand 层中

比例设置为 0.5。与文献 SqueezeNet[10]网络中的不同之处是，本文将池化层放在了第一层卷积和 Fire3、Fire5 之后，为进一步减少计算量，第一层从原来的 96 个 7×7 卷积核变为 64 个 3×3 卷积核。实验表明处理每张图片时网络计算量可以减少 2.4 倍。本文网络同样移除全连接层，用平均池化层代替全连接层以实现稀疏连接，但依然保留 Dropout，应用于 Fire9 之后比率为 50%。从时间性能方面进行比较的结果，如图 4 至图 6 所示。

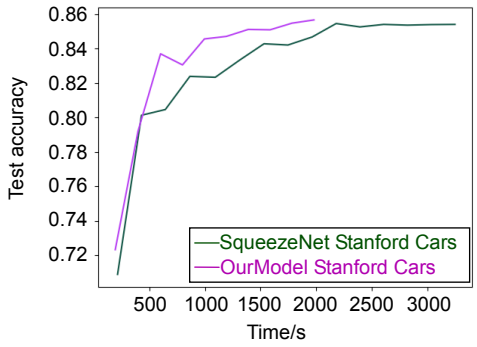


图 4 使用本文方法和未用本文方法在源域数据集 Stanford Cars 上时间性能比较

Fig. 4 Time performance on the source-domain dataset Stanford Cars with and without our method

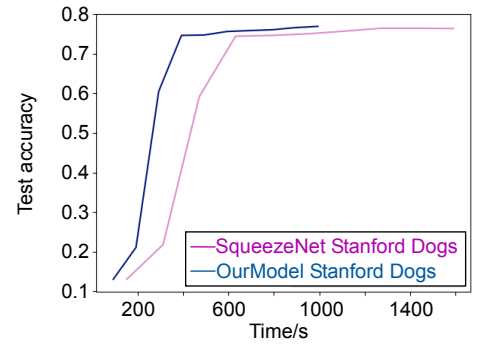


图 5 使用本文方法和未用本文方法在源域数据集 Stanford Dogs 上时间性能比较

Fig. 5 Time performance on the source-domain dataset Stanford Cars with and without our method

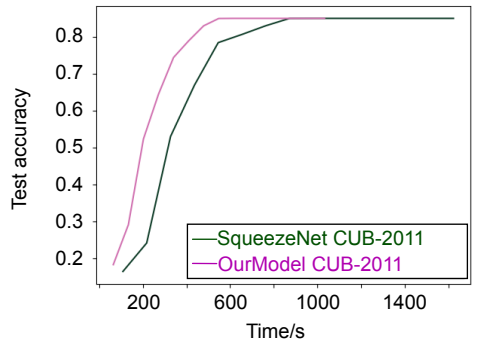


图 6 使用本文方法和未用本文方法在源域数据集 CUB-200-2011 上时间性能比较

Fig. 6 Time performance on the source-domain dataset CUB-200-2011 with and without our method

3.3 目标域设置

通过模型第一阶段的域间任务关联度度量的结果可知，当源域数据集为 Stanford Cars[16]时，数据集 BMVC[15]和 BMW-10[16]作为其目标域数据集最合理。同理，针对于目标域数据集选取 Oxford-IIIT Pet[17]，源域数据集选取 Stanford Dogs[13]，针对于目标域数据集选取 birds[18]，源域数据集选取 CUB-2011-200[14]。本文方法不仅可以使数据集很好地完成其本身的源域内任务，而且充分地利用到了域间数据集的关联性，提高了网络提取图片特征的泛化能力。

由于数据集 Stanford Dog[13]中皆为狗类图片，不失一般性，本论文将 Oxford-IIIT Pet[17]中类别为狗的 25 类图片提取出来单独进行实验。由于猫类与狗类图片具有很高的相似性，为了保证实验的全面性，将类别为猫的 7 类图片同样进行单独实验。由于相邻层间的神经元在训练时产生共适应，当网络的某一层被冻结时，整个网络就无法重新获取这种共适应。因此本文针对网络中间层的相邻层间神经元共同适应性问题，冻结 SqueezeNet[10]的 1~4、1~5 和 1~6 层作为特征迁移网络的独立个体分别进行迁移学习，在目标域数据集上进行实验，特征迁移效果如表 2 所示。

表 2 投票机制的候选者作为单一个体进行独立迁移学习的结果对比

Table 2 Comparison candidates of voting mechanism as a single individual to transfer learning independently

	frozen 1-4	frozen 1-5	frozen 1-6
BMVC[15]	97.4	97.7	97.2
BMW-10[16]	76.7	80.23	74
Oxford-IIIT Pet[17]	83.7	84.2	83.55
Birds[18]	99.8	99.8	99.6

4 结果分析

从图 4 至图 6 中可以看出，通过本文方法可以在保证准确率的同时有效减少训练时间。

BMVC[15]：表 3 中给出了本文方法和一些其他方法的结果。LLC[31]分类准确率为 84.5%，该利用局部性约束将每个描述符投影到其局部坐标系中，并且通过最大池来集成投影坐标以生成最终表示。LLC 方法首先进行 K-最近邻搜索，然后求解约束最小二乘拟合问题，使用特征描述算子 HoG 和简单线性 SVM 作为分类器；PHOW[2]分类准确率为 89.0%，该方法通过利用自下而上区域提议计算的深度卷积特征，学习整个目标和部件检测器，在它们之间强制学习几何约束，并

表 3 各方法在目标域微型数据集上分类结果

Table 3 Categorization result comparison on micro fine-grained datasets with advanced methods

Method	BMVC[15]	Method	BMW-10[16]	Method	Oxford-IIIT Pet[17]	Method	Birds[18]
PHOW[2]	89.0	KDES[19]	46.5	GMP+XColor[20]	56.8	CoCount[21]	55.22
BoT[22]	96.6	BB[23]	58.7	DDTF[24]	57.5	GP[25]	58.06
LLC[31]	84.5	LLC[31]	52.8	Zernike+SCC[26]	59.5	Low-rank[27]	74.5
StructDPM[15]	93.5	structDPM[15]	29.1	BW-FMP[28]	69.6	SNAK[29]	81.33
BB-3D-G[16]	94.5	BB-3D-G[16]	66.1	MsML+[30]	81.18	MEF-PB[18]	92.33
AlexNet	94.95		57.2		82.5		98.67
SqueezeNet	96.65		74		83.38		99.8
Ours	**97.7**		**80.23**		**84.2**		**99.8**

从姿势规范化表示中预测细粒度类别；StructDPM 方法准确率为 93.5%，该方法在本地特征外观和位置的层面上将两个最先进方法的 2D 对象表示提升为 3D，3D 空间池的目标为表征局部特征相对于对象的 3D 几何的位置，利用 3D 几何估计作为基础。这些方法在 BMVC 数据集上显示出非常有竞争力的结果，证明了它们对公平比较的有效性。本文模型分类准确率为 97.7%，验证了深度迁移学习对提高细粒度分类性能的可行性。

BMW-10[16]：表 3 给出了各种方法在 10 辆宝马轿车的细粒度数据集的结果。部件布局 PB[15]为 29.1%，区分性局部特征的方法表现更好(KDES[19]46.5%，LLC[31]52.8%，BB[23]58.7%)，利用 3D 图像表征 BB-3D-G[16](66.1%)提高了 7.4%。由表 3 可以看出针对于迁移学习中的数据集 BMW-10，训练准确率提升明显，因为数据集 BMW-10 不仅仅是一个简单的细粒度车型识别数据集，其十个类别都是宝马品牌的子系列车型，各类皆为普通轿车且类别间相似度较高，所以经过迁移后的网络对此数据集进行训练准确率的提升效果较其他训练集更为明显。

birds[18]：表 3 给出了一些高性能的分类方法结果。基于有效凸优化的低秩双线性分类器(low-rank)[27]为 74.5%，该方法通过最小化分类器的跟踪范数来优化分类器，在多核学习和跨模态学习方面提出了两个新的双线性分类器扩展，通过对双线性方法进行核化；基于概率部件(MEF-PB)的方法[18]为 92.33%，使用判别性最大熵框架来学习类别标签的后验分布。本文方法相对于上述传统方法中最高的分类准确率提高了 7.47%，达到了目前所有方法中的最高精度 99.8%。

Oxford-IIIT Pet[17]：由表 4 可以看出狗类与猫狗类别混合的训练结果提升效果不显著，实验时将 SqueezeNet[10]替换为 AlexNet[1]使用同样的方法在单独的猫类和狗类图片上进行实验。在猫类图片上分类准确率与直接使用 AlexNet[1]相比提高 3.3%，提升效果显著。这也验证了源域与目标域的迁移学习效果取决于两个域之间任务的相似程度这一结论，可见本文所提出的域间关联度度量取得了良好的效果。

表 4 将数据集 Oxford-IIIT Pet[17]中的猫与狗数据分开后单独实验结果

Table 4 Categorization results for separate statistics of cats and dogs in Oxford-IIIT Pet[17]

	Oxford-IIIT Pet-dog	Oxford-IIIT Pet-cat
AlexNet	59.45	40.3
Ours-AlexNet	59.75	43.6
SqueezeNet	59.50	44.4
Ours-SqueezeNet	59.95	44.7

5 总 结

本文引入衔接域提出了一种关于细粒度分类的深度迁移学习模型，利用源域大量有标签的细粒度训练数据来解决目标域中标签样本数量相对较少的细粒度分类任务。该方法与之前直接为目标域学习任务增加额外训练数据的工作不同，本文方法仅使用源域中非大规模有标签数据集的视图类标签进行辅助分类，并且对于目标域与源域数据集的相似性进行基于匹配的关联度度量。实验表明，本文的深度迁移学习模型在微型细粒度数据集上分类性能良好，深度卷积神经网络进行特征迁移可以加速网络的学习和优化。然而，如何为特定的目标域学习任务寻找更为合适的源域仍然是未来研究的一个悬而未决的问题。

参考文献

[1] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]//*Proceedings of 2012 International Conference on Neural Information Processing Systems*, Lake Tahoe, USA, 2012: 1097–1105.

[2] Zhang N, Donahue J, Girshick R, *et al*. Part-based R-CNNs for

fine-grained category detection[C]//*Proceedings of the 13th European Conference on Computer Vision*, Zurich, Switzerland, 2014: 834–849.

[3] Branson S, Van Horn G, Belongie S, *et al*. Bird species categorization using pose normalized deep convolutional nets[OL]. arXiv preprint arXiv:1406.2952[cs.CV].

[4] Simon M, Rodner E. Neural activation constellations: unsupervised part model discovery with convolutional networks[C]//*Proceedings of 2015 IEEE International Conference on Computer Vision*, Santiago, Chile, 2015: 1143–1151.

[5] Tan B, Song Y Q, Zhong E H, *et al*. Transitive transfer learning[C]//*Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Sydney, NSW, Australia, 2015: 1155–1164.

[6] Tzeng E, Hoffman J, Darrell T, *et al*. Simultaneous deep transfer across domains and tasks[C]//*Proceedings of 2015 IEEE International Conference on Computer Vision*, Santiago, Chile, 2015: 4068–4076.

[7] Ge W F, Yu Y Z. Borrowing treasures from the wealthy: deep transfer learning through selective joint fine-tuning[C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, 2017: 10–19.

[8] Yosinski J, Clune J, Bengio Y, *et al*. How transferable are features in deep neural networks?[C]//*Proceedings of 2014 International Conference on Neural Information Processing Systems*, Montreal, Canada, 2014: 3320–3328.

[9] Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with application to face verification[C]//*Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005, **1**: 539–546.

[10] Iandola F N, Han S, Moskewicz M W, *et al*. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size[OL]. arXiv preprint arXiv:1602.07360 [cs.CV].

[11] Jia Y Q, Shelhamer E, Donahue J, *et al*. Caffe: convolutional architecture for fast feature embedding[C]//*Proceedings of the 22nd ACM International Conference on Multimedia*, Orlando, USA, 2014: 675–678.

[12] Xie S N, Yang T B, Wang X Y, *et al*. Hyper-class augmented and regularized deep learning for fine-grained image classification[C]//*Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, 2015: 2645–2654.

[13] Deng J, Dong W, Socher R, *et al*. ImageNet: A large-scale hierarchical image database[C]//*Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009: 248–255.

[14] Wah C, Branson S, Welinder P, *et al*. The Caltech-UCSD birds-200–2011 dataset[R]. California: California Institute of Technology, 2011.

[15] Stark M, Krause J, Pepik B, *et al*. Fine-grained categorization for 3D scene understanding[J]. *International Journal of Robotics Research*, 2011, **30**(13): 1543–1552.

[16] Krause J, Stark M, Deng J, *et al*. 3D object representations for fine-grained categorization[C]//*Proceedings of 2013 IEEE International Conference on Computer Vision Workshops*, Sydney, Australia, 2013: 554–561.

[17] Parkhi O M, Vedaldi A, Zisserman A, *et al*. Cats and dogs[C]//*Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, USA, 2012: 3498–3505.

[18] Lazebnik S, Schmid C, Ponce J. A maximum entropy framework for part-based texture and object recognition [C]//*Proceedings of the 10th IEEE International Conference on Computer Vision*, Beijing, China, 2005, **1**: 832–838.

[19] Bo L F, Ren X F, Fox D. Kernel descriptors for visual recognition [C]//*Proceedings of the 23rd International Conference on Neural Information Processing Systems*, Vancouver, Canada, 2010: 244–252.

[20] Murray N, Perronnin F. Generalized max pooling [C]//*Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014: 2473–2480.

[21] Khamis S, Lampert C H. CoConut: co-classification with output space regularization[C]//*Proceedings of 2014 British Machine Vision Conference*, Nottingham, UK, 2014.

[22] Wang Y M, Choi J, Morariu V I, *et al*. Mining discriminative triplets of patches for fine-grained classification [C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, 2016: 1163–1172.

[23] Deng J, Krause J, Li F F. Fine-grained crowdsourcing for fine-grained recognition[C]//*Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition*, Portland, USA, 2013: 580–587.

[24] Wu X M, Mori M, Kashino K. Data-driven taxonomy forest for fine-grained image categorization[C]//*Proceedings of 2015 IEEE International Conference on Multimedia and Expo*, Turin, Italy, 2015: 1–6.

[25] Escalante H J, Ponce-López V, Escalera S, *et al*. Evolving weighting schemes for the bag of visual words[J]. *Neural Computing and Applications*, 2017, **28**(5): 925–939.

[26] Iscen A, Tolias G, Gosselin P H, *et al*. A comparison of dense region detectors for image search and fine-grained classification[J]. *IEEE Transactions on Image Processing*, 2015, **24**(8): 2369–2381.

[27] Kobayashi T. Low-rank bilinear classification: efficient convex optimization and extensions[J]. *International Journal of Computer Vision*, 2014, **110**(3): 308–327.

[28] Hang S T, Aono M. Bi-linearly weighted fractional max pooling. An extension to conventional max pooling for deep convolutional neural network[J]. *Multimedia Tools and Applications*, 2017, **76**(21): 22095–22117.

[29] Ionescu R T, Popescu M. Have a SNAK. Encoding spatial information with the spatial non-alignment kernel [C]//*Proceedings of 18th International Conference on Image Analysis and Processing*, Genoa, Italy, 2015: 97–108.

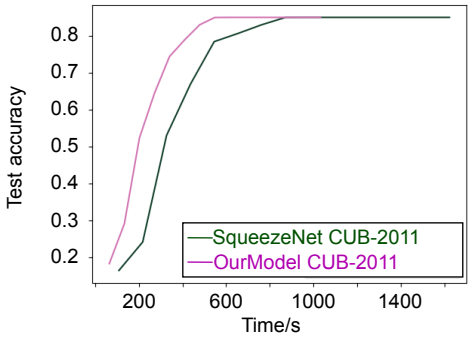
[30] Qian Q, Jin R, Zhu S H, *et al*. Fine-grained visual categorization via multi-stage metric learning[C]//*Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, 2015: 3716–3724.

[31] Wang J J, Yang J C, Yu K, *et al*. Locality-constrained linear coding for image classification[C]//*Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, UAS, 2010: 3360–3367.

Deep transfer learning for fine-grained categorization on micro datasets

Wang Ronggui, Yao Xuchen, Yang Juan*, Xue Lixia

School of Computer and Information, Hefei University of Technology, Hefei, Anhui 230601, China



Time performance on the source-domain dataset CUB-200-2011 with and without our method

Overview: Fine-grained categorization is challenging due to its small inter-class and large intra-class variance. Moreover, requiring domain expertise makes fine-grained labelled data much more expensive to acquire. Existing models predominantly require extra information such as bounding box and part annotation in addition to the image category labels, which involves heavy human manual labor. To solve this problem, we propose a novel deep transfer learning model, which transfers the learned representations from large-scale labelled fine-grained datasets to micro fine-grained datasets. While the network in deep learning is a unified training and prediction framework that combines multi-level feature extractors and recognizers, end-to-end processing is particularly important. The design concept for our model is to take full advantage of the ability that the convolutional neural network itself can perform end-to-end processing. As is known that feature transfer learning can use the existing data to rapidly construct the corresponding network parameters for new data through end-to-end training, which assumes that the source domain and the target domain contains some common cross-features, data from each domain can be transformed into the same feature space for the following learning. We present a novel discriminative training method that is used to learn similarity measurement, introducing the cohesion-domain quantitative calculation for the correlation between the two domains. Firstly, we introduce a cohesion domain to measure the degree of correlation between source domain and target domain. Secondly, selecting the transferrable feature that are suitable for the target domain based on the correlation. Finally, we make most of perspective-class labels for auxiliary learning, and learn all the attributes through joint learning to extract more feature representations. Our model aims to make joint adjustments from end to end, we expect to explore abundant source-domain attributes through cross-domain learning and capture more complex cross-domain knowledge by embedding cross-dataset information, in order to minimize the original function loss for the learning tasks in two domains as much as possible. For the problem of inter-domain transition network, we freeze part of the network layers to extract relatively more well-defined representations of labelled fine-grained samples for transferring to target domain. Since feature learning has the ability to collect hierarchical information which is not affected by the training data. In this way, the problem of high non-convex model optimization is not only simplified, but also can be modified from a more local perspective. So that subsequent incremental learning can limit the switching task to its own domain, and it is also conducive for multi-task parallel training to share the learned representation from different tasks. The experiments show that our model not only achieves high categorization accuracy but also economizes training time effectively, it also verifies the conclusion that the inter-domain feature transition can accelerate learning and optimization.

Citation: Wang R G, Yao X C, Yang J, *et al*. Deep transfer learning for fine-grained categorization on micro datasets[J]. *Opto-Electronic Engineering*, 2019, **46**(6): 180416

* E-mail: yangjuan6985@163.com