



研究与开发

基于特征校正的多对抗域适应方法

张永^{1,2}, 刘昊双², 章琪², 刘文哲¹

(1. 湖州师范学院信息工程学院, 浙江 湖州 313000;

2. 辽宁师范大学计算机与信息技术学院, 辽宁 大连 116081)

摘要: 领域自适应可以通过对齐源域和目标域的分布将有标签的源域信息迁移到没有标签但相关的目标域。然而, 现有的大多数方法仅对源域和目标域的低层特征分布进行对齐, 无法捕获样本中的细粒度信息。基于此, 提出了一种基于特征校正的多对抗域适应方法。该方法在引入注意力机制以突出可迁移区域的基础上, 通过部署特征校正模块对齐两个域之间的高级特征分布, 进一步缩小域差异。此外, 为了避免单个分类器过度拟合其自身的噪声伪标签, 还提出了双分类器协同训练, 并利用图神经网络特征聚合的特性生成更精准的源域标签。在 3 个迁移学习基准数据集上的大量实验证明所提方法的有效性。

关键词: 领域自适应; 迁移学习; 对抗网络; 注意力机制

中图分类号: TP181

文献标志码: A

doi: 10.11959/j.issn.1000-0801.2024014

Multi-adversarial domain adaptation method based on feature correction

ZHANG Yong^{1,2}, LIU Haoshuang², ZHANG Qi², LIU Wenzhe¹

1. School of Information Engineering, Huzhou University, Huzhou 313000, China

2. School of Computer & Information Technology, Liaoning Normal University, Dalian 116081, China

Abstract: Domain adaptation can transfer labeled source domain information to an unlabeled but related target domain by aligning the distribution of source domain and target domain. However, most existing methods only align the low-level feature distributions of the source and target domains, failing to capture fine-grained information within the samples. To address this limitation, a feature correction-based multi-adversarial domain adaptation method was proposed. An attention mechanism to highlight transferable regions was introduced in this method and a feature correction module was deployed to align the high-level feature distributions between the two domains, further reducing domain discrepancies. Additionally, to prevent individual classifiers from overfitting their own noisy pseudo-labels,

收稿日期: 2023-09-28; 修回日期: 2024-01-07

通信作者: 刘文哲, liuwenzhe@zjhu.edu.cn

基金项目: 国家自然科学基金资助项目 (No.61772252); 辽宁省教育厅科学研究经费资助项目 (No.LJKZ0965); 湖州市科技计划项目 (No.2022GZ08, No.2023ZD2004)

Foundation Items: The National Natural Science Foundation of China (No.61772252), The Scientific Research Foundation of the Education Department of Liaoning Province (No.LJKZ0965), The Huzhou Science and Technology Plan Project (No.2022GZ08, No.2023ZD2004)



dual classifier co-training was proposed and the feature aggregation property of graph neural networks was utilized to generate more accurate source domain labels. Extensive experiments on three benchmark datasets for transfer learning demonstrate the effectiveness of the proposed method.

Key words: domain adaptation, transfer learning, adversarial network, attention mechanism

0 引言

近年来,深度学习在计算机视觉领域得到了广泛应用,极大提高了图像分类的准确性。然而,要训练一个优秀的深度网络需要大量带标签的数据,而在现实世界中获取这些带标签的数据既昂贵又耗时。为解决这个问题,迁移学习应运而生。它通过将源域(即标签丰富的领域)的知识转移到目标域(即标签稀缺的领域)来训练分类器。但是,不同领域的样本具有不同的特征,这使得问题变得更加复杂。当训练和测试数据分布之间存在变化时,学习判别模型就被称为领域自适应^[1-2]。领域自适应是迁移学习的一种特殊场景,旨在学习一种区分模型来减小源域和目标域之间的数据分布差异。先前的领域自适应方法通常使用实例重加权或学习域不变特征来对齐域分布。

研究表明,在深度神经网络特征学习模型中嵌入领域自适应模块可以提取域不变的表示,并且可以学习更多可迁移的特征,从而在深度领域自适应中得到广泛的应用^[3-4]。目前,针对分类任务的深度领域自适应方法主要是通过最小化特征分布之间的差异^[5]或通过对抗性学习特征表示来混淆域鉴别器^[6]。然而,现有的对抗性域自适应方法没有考虑不同区域的复杂分布,而是直接跨域对齐从整个图像中提取的表示。这可能会丢失每个图片的细粒度信息,甚至导致无关知识的负面迁移。为解决这些问题,最新的领域自适应方法表明,对于从不同领域提取的特征进行细粒度对齐,可以在许多迁移学习任务中实现较大的性能提升^[7]。

本文主要围绕细粒度迁移思想构建一种领域自适应网络框架,该框架由3个关键模块组成:

细粒度特征对齐、高级特征对齐和特征聚合。通过提取不同领域的细粒度特征进行对齐,可以显著提升迁移学习任务的性能。在计算机视觉任务中,引入注意力机制可使模型专注于图像中的重要区域,从而实现精准的特征对齐。此外,采用特征校正模块进一步缩小两个域之间的高级特征分布差异。为了学习更加鲁棒的特征表示,在空间上进行特征聚合是必要的。本文借鉴了图卷积网络的思想,将源样本和目标样本表示为图,并利用图神经网络(graph neural network, GNN)的特征聚集能力来融合相似样本的语义信息,以实现准确分类。由于目标域样本的类标签信息是缺失的,本文采用了双分类器协同训练的方法,利用普通分类器为GNN分类器提供伪标签,以减少分类损失的影响。

在此基础上,本文提出了一种用于无监督域自适应的基于特征校正的多对抗域适应方法。首先,考虑同一样本不同区域可迁移性的变化,本文通过多个域级域鉴别器生成的局部注意力来凸显可迁移的区域。其次,引入特征校正块,使目标样本能够更好地适应源样本在高层网络中的特征表示,通过全局域鉴别器生成的全局注意力来凸显可迁移的图像。同时,采用普通分类器和GNN分类器的协同训练,以学习更加鲁棒的特征,并生成更准确的源域标签,从而进一步减少分类损失。本文工作的主要贡献如下。

(1) 提出了一种新的用于细粒度特征对齐的深度神经网络架构。本文在目标域样本特征提取网络中单独添加了特征校正模块,以更好地捕捉到图片局部的细粒度信息,同时进一步减少领域间的特征分布差异。此外,本文为该模块增加了一个正则化,以提升其稳定性和有效性。

(2) 提出了双分类器协同训练的方法, 训练了两个分类器: 一个普通分类器和一个 GNN 分类器, 并由普通分类器为 GNN 分类器提供伪标签。该方法可以进一步减少分类损失, 同时防止单个分类器过度拟合其自身的噪声。该方法利用了图神经网络强大的特征聚合能力, 使得 GNN 分类器能够学习到更稳健的特征。

(3) 在 3 个基准领域自适应图像数据集上进行了实验, 包括 Office Home、Office-31 和 Image CLEF-DA。实验结果表明, 本文方法在无监督领域自适应方面表现了优异的性能。这表明本文方法可以有效处理具有挑战性的领域自适应问题。

1 相关工作

领域自适应方法可以大致分为两类: 基于差异度量最小化的方法和基于对抗性学习的方法。

基于差异度量最小化方法通过减小源域和目标域之间的分布差距来实现领域自适应。其中, 最大均值差异 (maximum mean discrepancy, MMD) 是一种被广泛应用于领域自适应任务中的差异度量方法, 还有一些改进方法, 如多核 MMD 和加权 MMD^[8]。针对源域和目标域的联合分布差异最小化问题, 田磊等^[9]提出了两阶段领域自适应学习方法, 在第一阶段学习一个共享投影变换, 使投影后的共享空间中边缘分布的差异最小, 并在第二阶段学习一个带结构风险的自适应分类器, 最小化源域和目标域的条件分布差异, 同时进一步保持源域和目标域边缘分布的流形一致性。蔡瑞初等^[10]提出了基于类内最大均值差异的分布对齐方法, 借助预测样本的伪标签依次对齐每个类别的领域类内分布。相关性对齐 (correlation alignment, CORAL) 方法也可用于对齐不同领域之间的分布^[11]。研究表明, 深度学习网络能够学习到更多可迁移的特征^[12], 因此越来越多的领域自适应方法将领域自适应模块嵌入深度特征学习网络中, 以实现更好的性能^[7,13]。例如, 对比自适

应网络 (contrastive adaptation network, CAN)^[5]提出了一种优化类内和类间域差异的度量方法。深度适应网络 (deep adaptation network, DAN)^[3]使用多核 MMD 准则来对齐提取到的高层特征。然而, 以往的深度领域自适应方法主要对全局的源域和目标域分布进行对齐, 没有捕获样本细粒度信息, 导致迁移学习性能不理想。为了解决这个问题, 最近的研究提出了深度子域自适应网络 (deep subdomain adaptation network, DSAN)^[14], 该网络通过对齐不同领域之间相关的子域分布, 实现了细粒度的域适应。此外, 域条件适应网络 (domain-conditional adaptation network, DCAN)^[15]利用领域条件通道注意力机制来优化共享特征提取器的结构, 从而提取关键的低层领域特征。

基于对抗性学习的领域自适应方法受生成对抗网络 (generative adversarial network, GAN) 的启发, 通过引入一个领域鉴别器来区分源域特征和目标域特征, 并学习一个能够混淆鉴别器的特征提取器来最小化域差异。其中, 领域对抗网络 (domain-adversarial neural network, DANN)^[16]首次将对抗网络思想引入迁移学习任务中, 它使用梯度反转层来学习领域不变的特征表示。最大分类器差异 (maximum classifier discrepancy, MCD)^[17]通过最大化两个分类器的决策差异进行域适应, 并通过训练特征提取器来最小化差异。条件域对抗网络 (conditional domain adversarial network, CDAN)^[4]通过传递分类器预测的判别信息对对抗网络模型进行调整。这些基于对抗性学习的方法在领域自适应中表现出了强大的迁移能力。

近年来, 注意力机制在计算机视觉任务中获得了显著的成果, 它允许网络对提取的特征进行像素级别的加权。进一步的研究提出了全局和局部注意力机制, 考虑了不同程度的注意力, 并在计算机视觉领域展现良好的性能^[18]。最近, 注意力机制被应用于解决领域自适应问题。深度对抗



性注意对齐 (deep adversarial attention alignment, DAAA) 方法^[19]通过注意力对齐迁移所有卷积层中的数据, 从而实现了领域自适应。领域适应的可迁移注意力 (transferable attention for domain adaptation, TADA)^[20]结合了可迁移的局部和全局注意力, 旨在突出跨领域的可迁移区域和图像特征。

GNN 是一种适用于图结构数据的神经网络模型, 它能够出色地处理非结构化数据, 在自然语言处理、图像处理、轨迹预测和药物医学等领域都取得了新的突破。将卷积神经网络应用到图上, 通过对卷积算子的巧妙转换, 提出了图卷积网络 (graph convolutional network, GCN)^[21], 也被应用于解决领域自适应问题。在开放集自适应中, 渐进式图学习^[22]利用图卷积获取源域和目标域中重叠类之间的关系。在多目标域自适应中, Yang 等^[23]引入了异构图注意网络来学习多个领域之间相似样本的关系, 然后利用基于图的伪标签, 通过对齐具有标签的源质心和具有伪标签的目标质心来学习域不变表示。与以上方法不同, 本文将 GCN 框架应用于更为普遍的单源域自适应, 并结合了注意力机制。

2 基于特征校正模块的多对抗域适应方法

2.1 问题定义与动机

在无监督领域自适应中, 有标签的源域表示为 $D_S = \{(x_1^s, y_1^s), \dots, (x_{n_s}^s, y_{n_s}^s)\}$, 无标签的目标域表示为 $D_T = \{x_1^t, \dots, x_{n_t}^t\}$ 。源域 D_S 有 n_s 个有标签样本, 每个源样本 x_i^s 对应的标签为 y_i^s 。而目标域 D_T 有 n_t 个未标记样本 x_i^t 。本文假设源域和目标域具有相同的标签空间, 即有 C 个公共类。由于两个领域的分布并不相同, 即 $P_S \neq Q_T$, 因此不能直接使用源域数据训练的模型对目标域数据进行分类。本文研究的目标是通过在训练阶段探索有标签的源域数据和无标签的目标域数据得到在目标域上也具有良好性能的模式。

目前, 许多领域自适应方法都采用对抗学习

的思想来实现源域和目标域的特征分布对齐。DANN 通过使用对抗训练的方法来对齐源域和目标域的特征分布, 它由 3 个网络组成: 特征提取器 $F(\cdot)$ 、分类器 $C_s(\cdot)$ 和域鉴别器 $G_d(\cdot)$ 。分类器 $C_s(\cdot)$ 负责将从特征提取器 $F(\cdot)$ 得到的特征分为 n_c 类。一方面, 域鉴别器 $G_d(\cdot)$ 被训练来区分源域特征和目标域特征。另一方面, 特征提取器 $F(\cdot)$ 被训练来混淆鉴别器 $G_d(\cdot)$, 使其无法准确区分源域和目标域的特征。

尽管现有的方法已经显著提高了迁移学习的性能, 但这些方法都没有充分利用样本的局部细粒度结构, 而是将每个样本作为一个整体来考虑其可迁移性。显而易见, 图像不同区域的分布是不相同的, 图像中的某些区域, 如背景等, 尽管可以在特征空间中被跨域对齐, 但可能对领域自适应没有太大的作用。在不同领域特征空间中显著不同的一些图像不应当被强制跨域对齐, 否则可能丢失每个图片的细粒度信息, 甚至造成负迁移。仅在网络顶层部署域差异惩罚项或对抗性损失可能不太有效, 因为任务特定层的损失修改梯度将通过反向传播方案衰减。此外, 对于复杂的视觉跨领域任务, 两个领域的视觉模式存在很大差异, 例如, 对于背景干净的图像和背景复杂的图像, 仍然假设每个图像的不同区域被类似地激活是不合适的。另外, 单一分类器的使用也容易过拟合自身的噪声。因此, 在探索样本细粒度结构的同时, 还应该探索目标域独有的高级特征, 并尝试在空间中学习更鲁棒的特征。

基于以上背景, 本文提出了一种基于特征校正模块的多对抗域适应方法, 其网络架构如图 1 所示, 本文在特征提取网络中引入局部注意力机制, 以凸显可迁移区域。然后, 引入特征校正块, 使目标样本能够更好地适应源样本的高层特征, 从而由全局域鉴别器生成用于突出可迁移图像的全局注意力。同时, 本文采用由普通分类器和 GNN 分类器组成的双分类器协同训练, 以学习更

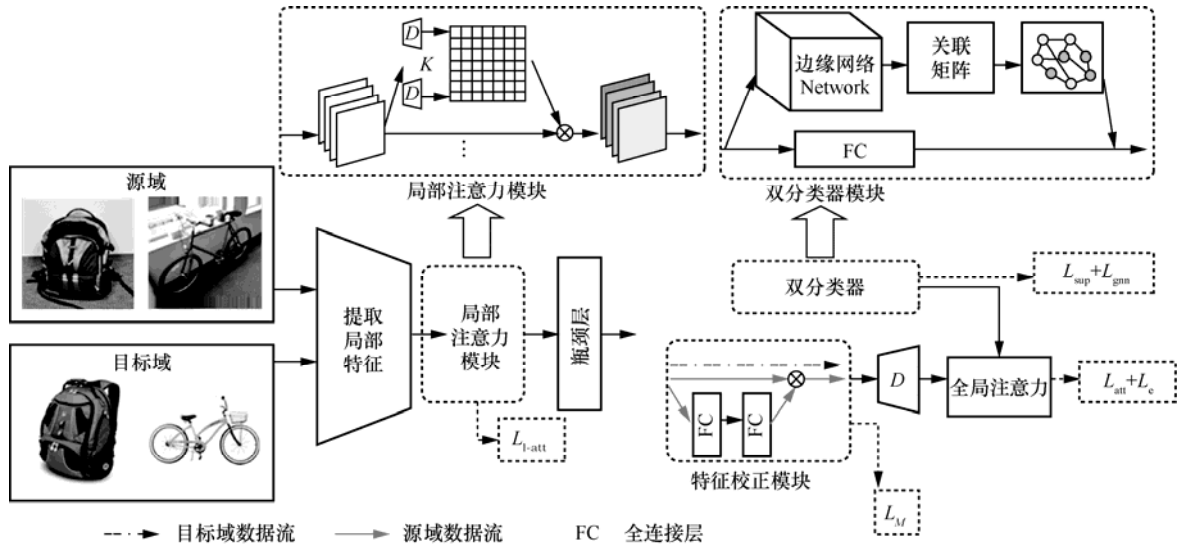


图1 基于特征校正模块的多对抗域适应方法网络架构

鲁棒的特征并生成更准确的源域标签，从而实现
对领域信息更精确的提取和迁移。

2.2 多对抗网络的特征校正

考虑一些图像在不同区域的可迁移性不同，
而强行对齐不可迁移的区域可能会导致负迁移，
本文引入了可迁移注意力^[20]，在传统对抗模型中
添加了局部和全局注意力模块。局部可迁移注意
力是指将域鉴别器 G_d 分割为 K 个区域级域鉴别
器 G_d^k , $k=1,2,\dots,K$ ，每个域鉴别器分别负责混
淆源域和目标域图像中不同的区域。如图 1 所示，
可以得到局部注意力的损失函数：

$$L_{l-att} = \frac{1}{(n_s + n_t)K} \sum_{k=1}^K \sum_{x_i \in D_s \cup D_t} L_d(G_d^k(F(x_i)^k), d_i^k) \quad (1)$$

其中， $F(x_i)^k$ 表示图像 x_i 中的第 k 个区域， G_d^k 输
出的 $\tilde{d}_i^k = G_d^k(F(x_i)^k)$ 表示图像 x_i 中的区域 k 属于
源域的概率。

为了更好地量化不同区域的可迁移性，本文
运用信息论中的熵来度量不确定性，即
 $H(p) = -\sum_j p_j \cdot \lg(p_j)$ ，这里 p_j 表示图像的区域 j
属于源域的概率。另外，可通过添加残差连接来
减轻错误的局部注意力造成的影响^[16]。综合考虑
可迁移性和残差连接，最终样本的局部特征 Z_i^k 可

以表示为：

$$Z_i^k = (1 + (1 - H(\tilde{d}_i^k))) \cdot F(x_i)^k \quad (2)$$

其中， Z_i^k 表示样本 x_i 中第 k 个区域的局部特征，
 $H(\tilde{d}_i^k)$ 表示概率 $\tilde{d}_i^k$ 的熵。通过计算熵值，本文给
具有更强可迁移性的区域更大的权重，从而将域
适应模型的关注点聚焦在这些对训练更有帮助的
区域上。然而，不同领域的高级特征分布仍然存
在较大的差异。为了缓解高级特征的分布不匹配，
本文通过在网络的瓶颈层后引入特征校正模块，
可以对跨领域特征分布的不匹配问题进行修正。

经过瓶颈层后的数据被表示为 $\mu_i = G_b(Z_i)$ ，
其中 G_b 表示瓶颈层网络。在目标域数据通过由两
个全连接（full connection, FC）层和一个 ReLU
层组成的特征校正模块后，可以得到校正后的目
标域特征表示为 $\hat{R}(\mu_t) = R(\mu_t) + \Delta R(\mu_t)$ ，其中
 $R(\mu_t)$ 和 $\Delta R(\mu_t)$ 分别表示校正前的特征和校正增量。

然而，目标域数据通过特征校正模块后可能
会导致源域与目标域之间的过度迁移，解决这一
问题的方法是令源域数据也通过特征校正模块，
但需要保持源域数据经过特征校正模块后的分布
与未经过修正的源域数据分布相似，即 $\hat{R}(\mu_s)$ 和
 $R(\mu_s)$ 的分布应该保持相似。为此，本文引入了一



个正则化损失 L_M 来指导校正过程。本文选择源域的随机子集 M 来代替源域的整体集合，并最小化每个类别和源域子集之间的 MMD 度量。定义为：

$$L_M = \sum_{c=1}^C \frac{1}{n_s^c} \sum_{Z_i^s \in n_s^c} \phi(R(\mu_i^s)) - \frac{1}{|M|} \sum_{Z_i^s \in M} \phi(\hat{R}(\mu_i^s))^2_{\mathcal{H}_k} \quad (3)$$

其中， $|M|$ 是子集 M 的大小， n_s^c 表示属于类别 c 的源域样本数量， ϕ 表示特征映射函数， $\mathcal{H}_k$ 表示核矩阵。通过引入正则化损失 L_M ，可以在一定程度上解决过校正问题，并增强特征校正模块的对齐能力。

将经过特征校正模块处理后的源特征和目标特征 $\hat{\mu}_i = \hat{R}(\mu_s) + \hat{R}(\mu_t)$ 输入域鉴别器中，可以得到以下目标函数：

$$L_{att} = \frac{1}{n_s + n_t} \sum_{i=1}^{n_s+n_t} L_l(G_d(\hat{\mu}_i)) \quad (4)$$

其中， L_l 是全局域鉴别器 G_d 的交叉熵损失。

为了解决分类器在目标域上泛化性能不佳的问题，可以采用条件熵最小化方法。条件熵最小化广泛应用于对抗网络的相关工作中^[4]，最小化目标域数据 D_t 上的类条件分布的熵，鼓励类之间的低密度分离。然而，强制最小化所有图像的熵可能导致负面结果，因为并非所有目标域图像都是可迁移的。例如，目标域中存在与源域特征空间显著差异的图像，可能容易被错误分类，增加这些图像上的分类器预测的置信度可能降低分类准确率。因此，类似于局部注意力机制，可以利用全局域鉴别器的输出来为每个图像生成一个注意值，该值表示图像的可迁移性。注意值越大，表示图像的可迁移性越高。可以将全局注意值 $m_i = 1 + H(\tilde{d}_i)$ 嵌入熵损失中，其中 $\tilde{d}_i = G_d(\hat{\mu}_i)$ 为全局域鉴别器的输出。全局注意熵可以表示为：

$$L_e = -\frac{1}{n_s + n_t} \sum_{i=1}^{n_s+n_t} \sum_{j=1}^C m_i \times p_{i,j} \times \ln p_{i,j} \quad (5)$$

其中， $p_{i,j}$ 表示图像 i 属于类别 j 的概率。

2.3 双分类器协同训练

在无监督领域自适应的背景下，模型训练包括有标签的源域数据和无标签的目标域数据。由于只有源域数据具有标签信息，可以通过最小化分类损失函数来构建源分类器，以保证对源域数据的标签预测的准确性。源分类器的损失函数可以表示为：

$$L_{sup} = -\frac{1}{n_s} \sum_{i=1}^{n_s} L_{sup}(C_s(G_b(Z_i)), y_i) \quad (6)$$

其中， Z_i 表示源域样本的特征， y_i 表示源域样本的真实标签。

为了提取更鲁棒的特征并防止单个分类器过度拟合影响对抗网络的性能，本文提出了双分类器协同训练的方法。该方法在源分类器 $C_s(\cdot)$ 的基础上引入图网络分类器 $C_g(\cdot)$ ，其由边缘网络 F_e 和节点分类器 F_n 组成。具体而言，本文采用图神经网络的思想，在进行分类之前，利用源分类器 $C_s(\cdot)$ 对样本构建了一个无向全连通图 $\Gamma = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ 。其中，每个图像 x_i 在图中表示为单个节点 $v_i \in \mathcal{V}$ 。边 $e_{i,j} \in \mathcal{E}$ 表示节点 v_i 和 v_j 之间的连接关系，而节点 v_i 和 v_j 的语义相似度得分表示为 $a_{i,j}$ 。所有样本之间的相似关系可用相似度矩阵 $\mathbf{A}$ 来表示。

在双分类器协同训练中，将经过瓶颈层处理后的样本特征 $\mu_i = G_b(Z_i)$ 同时输入 $C_s(\cdot)$ 和 $C_g(\cdot)$ 中。在图网络分类器 $C_g(\cdot)$ 中，将 μ_i 输入 F_e 中，利用 $a_{i,j}^L = F_e^L(v_i^{L-1}, v_j^{L-1})$ 来估计样本节点之间的相似性，并输出一个相似度矩阵 $\mathbf{A}$ 。其中， $a_{i,j}$ 的值越大表示样本 i 和 j 属于同一语义类别的可能性越高。这里， F_e 为参数化的非线性相似函数， v_i^{L-1} 表示样本在第 $L-1$ 个图卷积层的特征。为了归一化相似度矩阵，本文在给图中的节点添加自连接的同时，使用式 (1) 将获得的相似性分数归一化：

$$\mathbf{A}^L = \mathbf{M}^{-\frac{1}{2}} (\tilde{\mathbf{A}}^L + \mathbf{I}) \mathbf{M}^{-\frac{1}{2}} \quad (7)$$

其中， $\mathbf{M}$ 为度矩阵， $\mathbf{I}$ 为单位矩阵， $\tilde{\mathbf{A}}$ 为非归一化亲和矩阵。

根据相似度矩阵 $\mathbf{A}$, 本文使用以下传播规则来更新节点特征:

$$\mathbf{v}_i^L = F_n^L \left(\left[\mathbf{v}_i^{L-1}, \sum_{j \in \mu_i} a_{i,j}^L \cdot \mathbf{v}_j^{L-1} \right] \right) \quad (8)$$

其中, $[\cdot, \cdot]$ 表示特征连接, F_n 为参数化的非线性函数, 它根据相似度矩阵来聚合 μ_i 中的特征。 F_n^L 层是输出层, 其输出为 $F_n(\mu_i)$ 。

最后, 为了训练 $C_g(\cdot)$ 中的 F_n , 本文使用标准的交叉熵损失函数来计算样本 x_i 的损失:

$$L_{\text{gmn}} = -\frac{1}{n_s} \sum_{i=1}^{n_s} L_b(F_n(\mu_i), y_i) \quad (9)$$

其中, L_b 为交叉熵损失函数, L_{gmn} 为图网络分类器 $C_g(\cdot)$ 的标准交叉熵损失。

2.4 总训练目标

综合以上内容, 本文所提方法的整体目标函数如下。

$$\min_F L = L_{\text{sup}} + \alpha L_{\text{gmn}} + \beta L_M + \gamma L_e - (L_{\text{l-att}} + L_{\text{att}}) \quad (10)$$

其中, $L_{\text{l-att}}$ 和 L_{att} 分别表示局部和全局注意力损失, L_M 表示特征校正模块的正则化损失, L_{gmn} 和 L_{sup} 分别为图网络分类器 $C_g(\cdot)$ 和分类器 $C_s(\cdot)$ 的标准交叉熵损失, L_e 为注意熵损失。参数 α 、 β 和 γ 用于权衡各个损失项的相对重要性, 具体的参数设置将在第 3 节详细介绍。

3 实验及结果分析

3.1 数据集与实验设置

本文使用 Office-31^[24]、ImageCLEF-DA^[25] 和 Office-Home^[26] 3 个常用的基准图像数据集来评估所提出的方法在无监督领域自适应任务上的性能。图像基准数据集的统计信息见表 1, 列出了这 3 个数据集的相关信息。

Office-31 是一个常用的领域自适应研究数据集, 包含 4 110 张图像。它被分为 3 个不同的领域: Amazon (A)、Webcam (W) 和 DSLR (D)。每个领域都包含 31 个办公环境下的类别。本文评估了 6 组不同的迁移任务: $A \rightarrow W$, $D \rightarrow W$, $W \rightarrow D$, $A \rightarrow D$, $D \rightarrow A$, $W \rightarrow A$ 。

ImageCLEF-DA 是 ImageCLEF 2014 领域自适应挑战的基准数据集。它由来自 4 个公共数据集的 12 个共享类别组成, 每个数据集表示一个领域。这 4 个领域是 Caltech-256 (C)、ImageNet ILSVRC 2012 (I)、Pascal VOC 2012 (P) 和 Bing (B)。每个类包含 50 张图片, 每个域包含 600 张图片。本文选择了其中 3 个领域 (C、I、P) 进行组合, 构建了 6 组迁移任务: $I \rightarrow P$, $P \rightarrow I$, $I \rightarrow C$, $C \rightarrow I$, $C \rightarrow P$, $P \rightarrow C$ 。

Office-Home 是一个包含办公室和家庭环境

表 1 图像基准数据集的统计信息

数据集	子域	样本数/个	类别数
Office-31	Amazon (A)	2 817	31
	DSLR (D)	498	31
	Webcam (W)	795	31
ImageCLEF-DA	Caltech (C)	600	12
	ImageNet (I)	600	12
	Pascal (P)	600	12
Office-Home	Artistic (Ar)	2 427	65
	Clip Art (Cl)	4 365	65
	Product (Pr)	4 439	65
	Real-World (Rw)	4 357	65



图片的数据集,由属于 65 个类别的 15 588 张图像组成。图像来自 4 个不同的领域: Artistic (Ar)、Clip Art (Cl)、Product (Pr)和 Real-World (Rw)。其中, Ar 是艺术绘画相关的图像, Cl 是剪贴画, Pr 是无背景图像, Rw 是由相机采集的真实世界图像。本文在这个数据集上构建了 12 组迁移学习任务: $Ar \rightarrow Cl, \dots, Rw \rightarrow Pr$ 。

本文采用在 ImageNet 数据集^[27]上预训练的 ResNet-50^[28]作为骨干网络,并在 PyTorch 框架下进行实验。为了方便比较,本文所引用的深度领域自适应方法都是在标准无监督领域的自适应实验设置下进行的。在实验中,所有图像都被归一化为 256×256 大小,然后随机裁剪为 224×224 作为网络输入。本文微调了所有的卷积层和池化层,并使用反向传播方法来训练分类器层和所有的域鉴别器。本文采用了随机梯度下降法,设置动量为 0.95,并使用学习率退火策略。在本文所有实验中固定参数 $\alpha = 0.1$ 、 $\beta = 0.2$ 和 $\gamma = 0.1$ 。

3.2 实验结果分析

为了更好地说明本文所提方法的有效性,将其与几种现阶段深度领域的自适应经典方法进行对比实验。在 Office-31 数据集上,本文与

ResNet-50^[28]、DAN^[3]、DANN^[16]、Deep CORAL (D-CORAL)^[13]、联合自适应网络 (joint adaptation network, JAN)^[29]、多对抗域自适应 (multi-adversarial domain adaptation, MADA)^[7]、CDAN^[4]、TADA^[20]、DSAN^[14]、增强传输距离 (enhanced transport distance, ETD)^[30]和类别对比 (category contrast, CaCo)^[31]等方法进行了比较。在 ImageCLEF-DA 数据集上,本文与 ResNet-50、DAN、DANN、MADA 和 DCAN^[15]等方法进行了比较。在 Office-Home 数据集上,本文与 ResNet-50、DAN、DANN、JAN、MADA、CDAN、TADA、DSAN 和域自适应对抗性习得损失 (adversarial-learned loss for domain adaptation, ALDA)^[32]等方法进行了比较。

不同算法在 Office-31 数据集上的分类准确率见表 2。本文方法在大多数迁移学习任务上优于比较方法,在所有迁移学习任务中都得到了次优以上的结果,并且提高了平均准确率。尤其是在困难的 $A \rightarrow D$ 和 $D \rightarrow A$ 这两个任务下,本文方法的准确率比次优方法提高了 0.7%和 0.3%。与 DSAN 相比,本文方法的表现更好,平均准确率提升了 0.4%,验证了本文所提模型的有效性。

表 2 不同算法在 Office31 数据集上的分类准确率

方法	A→W	D→W	W→D	A→D	D→A	W→A	AVG
ResNet-50	68.4%	96.7%	99.3%	68.9%	62.5%	60.7%	76.1%
DAN	83.8%	96.8%	99.5%	78.4%	66.7%	62.7%	80.4%
D-CORAL	77.7%	97.6%	99.7%	81.1%	64.6%	64%	80.8%
DANN	82%	96.9%	99.1%	79.7%	68.2%	67.4%	82.2%
JAN	85.4%	97.4%	99.8%	84.7%	68.6%	70%	84.3%
MADA	90%	97.4%	99.6%	87.8%	70.3%	66.4%	85.2%
CDAN	93.1%	98.2%	100%	89.8%	70.1%	68%	86.6%
TADA	94.3%	98.7%	99.8%	91.6%	72.9%	73%	88.4%
DSAN	93.6%	98.3%	100%	90.2%	73.5%	74.8%	88.4%
ETD	92.1%	100%	100%	88%	71%	67.8%	86.2%
CaCo	89.7%	98.4%	100%	91.7%	73.1%	72.8%	87.6%
本文算法	94%	98.7%	100%	92.4%	73.8%	73.6%	88.8%

不同算法在 ImageCLEF-DA 数据集上的分类准确率见表 3。从表 3 可以看出,与之前的方法相比,本文方法具有显著的改进,在除了任务 $I \rightarrow P$ 外的其他 5 组迁移学习任务中都得到了最高的准确率。平均准确率较 DCAN 提高了 0.4%。与 DCAN 相比,本文方法的平均准确率提高了 0.4%。本文方法可以更好地捕捉每个领域中更细粒度的信息,同时避免了单个分类器过拟合自身的噪声。因此,在这个具有挑战性的数据集上本文获得了更有说服力的结果。

不同算法在 Office-Home 数据集上的分类准确率见表 4。Office-Home 中类别较多,领域差异较大,模型训练难度较大。从表 4 中可以看出,与 DSAN 相比,本文方法在任务 $C \rightarrow A$ 和 $P \rightarrow R$ 上的准确率分别提高了 1.4%和 1.3%,平均准确率为 68.7%,优于所有比较方法。

实验结果表明,本文提出的模型在大多数迁移学习任务上优于比较方法。

3.3 消融实验

消融实验如图 2 所示,给出了本文提出方法的消融实验结果。通过对本文提出的方法进行消融实验,笔者团队探究了不同损失对模型自适应性能的影响。具体来说,本文分别从目标损失中移除了图分类器损失 L_{gmn} 、特征校正损失 L_m 以及注意熵损失 L_e ,并研究了在 ImageCLEF-DA 数据集上 $C \rightarrow I$ 和 $P \rightarrow C$ 两个任务分类准确率的变化。图 2 中,本文提出方法的准确率显著优于移除图分类器损失 L_{gmn} 后的准确率,这说明本文方法中引入的图分类器能够有效防止单个分类器过度拟合,从而能够有效提高对抗网络的性能。而将本文方法与特征校正损失 L_m 移除后的准确率相比,本文方法准确率也更具优势,验证了特征校正模块能够有效缓解高级特征的分布不匹配问题。同时,将图分类器损失 L_{gmn} 和特征校正损失 L_m 移除,结果导致了更低的分类性能。移除注意熵损失 L_e 后,在 $P \rightarrow C$ 任务上的性能急剧下降,表明

表 3 不同算法在 ImageCLEF-DA 数据集上的分类准确率

方法	$I \rightarrow P$	$P \rightarrow I$	$I \rightarrow C$	$C \rightarrow I$	$C \rightarrow P$	$P \rightarrow C$	AVG
ResNet-50	74.8%	83.9%	91.5%	78.0%	65.5%	91.2%	80.8%
DAN	74.5%	82.2%	92.8%	86.3%	69.2%	89.8%	82.5%
DANN	75.0%	86.0%	96.2%	87.0%	74.3%	91.5%	85.0%
MADA	75.0%	87.9%	96.0%	88.8%	75.2%	92.2%	85.9%
DCAN	80.5%	91.2%	95.7%	91.8%	77.2%	93.3%	88.3%
本文算法	78.7%	91.9%	96.5%	92.3%	77.8%	95.0%	88.7%

表 4 不同算法在 Office-Home 数据集上的分类准确率

方法	$A \rightarrow C$	$A \rightarrow P$	$A \rightarrow R$	$C \rightarrow A$	$C \rightarrow P$	$C \rightarrow R$	$P \rightarrow A$	$P \rightarrow C$	$P \rightarrow R$	$R \rightarrow A$	$R \rightarrow C$	$R \rightarrow P$	AVG
ResNet-50	34.9%	50.0%	58.0%	37.4%	41.9%	46.2%	38.5%	31.2%	60.4%	63.9%	41.2%	59.9%	47.0%
DAN	43.6%	57.0%	67.9%	45.8%	56.5%	60.4%	44.0%	43.6%	67.7%	63.1%	51.5%	74.3%	56.3%
DANN	45.6%	59.3%	70.1%	47.0%	58.5%	60.9%	46.1%	43.7%	68.5%	63.2%	51.8%	76.8%	57.6%
JAN	45.9%	61.2%	68.9%	50.4%	59.7%	61.0%	45.8%	43.4%	70.3%	63.9%	52.4%	76.8%	58.3%
CDAN	50.7%	70.6%	76.0%	57.6%	70.0%	70.0%	57.4%	50.9%	77.3%	70.9%	56.7%	81.6%	65.8%
TADA	53.1%	72.3%	77.2%	59.1%	71.2%	72.1%	59.7%	53.1%	78.7%	72.4%	60.0%	82.9%	67.7%
DSAN	54.4%	70.8%	75.4%	60.4%	67.8%	68.0%	62.6%	55.9%	78.5%	73.8%	60.6%	83.1%	67.6%
ALDA	53.7%	70.1%	76.4%	60.2%	72.6%	71.5%	56.8%	51.9%	77.1%	70.2%	56.3%	82.1%	66.6%
本文算法	55.6%	71.3%	78.0%	61.8%	71.8%	73.5%	61.5%	54.5%	80.0%	72.8%	60.9%	83.2%	68.7%



全局注意值嵌入熵损失中,有助于算法准确率的提升。综上,本文方法优于其他变体,这验证了加入图分类器进行双分类器协同训练对于提取更鲁棒的特征的有效性,并证明了特征校正模块在缓解高级特征的分布不匹配方面的重要性。这些消融实验结果进一步证明了本文方法的有效性,并表明多个损失的联合优化可提升模型自适应性能。

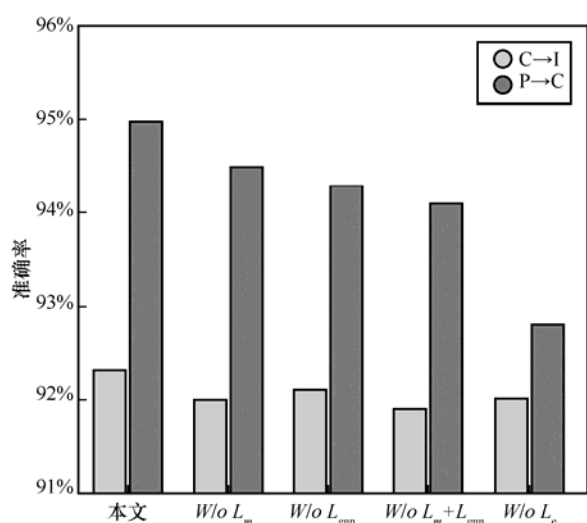


图2 消融实验

3.4 收敛性分析

本节选取了 Office31 数据集上 $A \rightarrow D$ 和 $A \rightarrow W$ 两个任务来进一步分析本文方法的收敛性,收敛性分析如图3所示。在实验中将迭代次数设置为3000,并固定其他超参数。图3中的虚线表示任务 $A \rightarrow W$ 的实验结果,可以看出分类准确率随迭代次数的增加而波动上升,随后逐渐趋于平稳,当迭代到2500次时分类准确率稳定在92%左右。图3中的实线表示迁移学习任务 $A \rightarrow D$ 的实验结果,从图3中可以看出其分类准确率在前600次迭代中迅速提高,随后分类准确率上升的速度逐渐下降,直至迭代至2500次左右其分类准确率逐渐趋近于稳定。这说明本文方法在 $A \rightarrow D$ 任务上也能够有效地进行域自适应训练,并取得了稳定的分类性能。总的来说,通过这些实验结

果可以看出,本文方法在不同任务上都具有良好的收敛性和稳定性。随着迭代次数的增加,模型的性能逐渐提升并趋于稳定,这进一步验证了本文方法的有效性和可靠性。

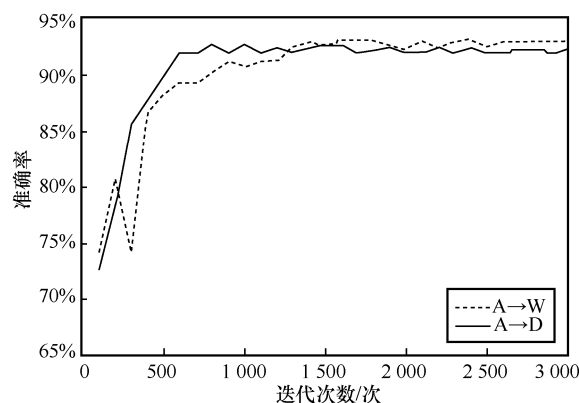


图3 收敛性分析

4 结束语

本文提出了一种基于特征校正模块的多对抗域适应方法,通过考虑同一样本不同区域可迁移性的变化,并利用多个域级域鉴别器生成的局部注意力和全局注意力,突出可迁移区域和图像,使目标样本更好地适应源样本。为了进一步改善分类性能和减少过度拟合,引入了具有强大特征聚合能力的图神经网络,并采用双分类器协同训练的方式学习更鲁棒的特征和生成更准确的源域标签。同时,引入了熵准则以充分利用预测信息和探索数据分布结构。实验结果表明,与其他方法相比,本文方法在多个图像数据集上表现出优越性。总体而言,本文方法是合理且有效的。未来的研究方向可以集中在实现更简便、高效的细粒度深度迁移学习上。

参考文献:

- [1] PAN S J, YANG Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359.
- [2] ZHUANG F, QI Z, DUAN K, et al. A comprehensive survey on transfer learning[J]. Proceedings of the IEEE, 2021, 109(1): 43-76.

- [3] LONG M S, CAO Y, WANG J M, et al. Learning transferable features with deep adaptation networks[C]//Proceedings of the Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37. New York: ACM Press, 2015: 97-105.
- [4] LONG M S, CAO Z J, WANG J M, et al. Conditional adversarial domain adaptation[C]//Proceedings of the Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York: ACM Press, 2018: 1647-1657.
- [5] KANG G L, JIANG L, YANG Y, et al. Contrastive adaptation network for unsupervised domain adaptation[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 4888-4897.
- [6] MA A, LI J J, LU K, et al. Adversarial entropy optimization for unsupervised domain adaptation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(11): 6263-6274.
- [7] PEI Z, CAO Z, LONG M, et al. Multi-adversarial domain adaptation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: ACM Press, 2018: 3934-3941.
- [8] YAN H L, DING Y K, LI P H, et al. Mind the class weight bias: weighted maximum mean discrepancy for unsupervised domain adaptation[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 945-954.
- [9] 田磊, 唐永强, 张文生. 两阶段领域自适应学习[J]. 模式识别与人工智能, 2019, 32(9): 773-784.
TIAN L, TANG Y Q, ZHANG W S. Two stage domain adaptation learning[J]. Pattern Recognition and Artificial Intelligence, 2019, 32(9): 773-784.
- [10] 蔡瑞初, 李嘉豪, 郝志峰. 基于类内最大均值差异的无监督领域自适应算法[J]. 计算机应用研究, 2020, 37(8): 2371-2375.
CAI R C, LI J H, HAO Z F. Unsupervised domain adaptive algorithm with intra-class maximum mean discrepancy[J]. Application Research of Computers, 2020, 37(8): 2371-2375.
- [11] SUN B C, FENG J S, SAENKO K. Return of frustratingly easy domain adaptation[C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. New York: ACM Press, 2016: 2058-2065.
- [12] YOSINSKI J, CLUNE J, BENGIO Y, et al. How transferable are features in deep neural networks[J]. Advances in Neural Information Processing Systems, 2014: 27.
- [13] SUN B C, SAENKO K. Deep CORAL: correlation alignment for deep domain adaptation[C]//European Conference on Computer Vision. Cham: Springer, 2016: 443-450.
- [14] ZHU Y C, ZHUANG F Z, WANG J D, et al. Deep subdomain adaptation network for image classification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(4): 1713-1722.
- [15] LI S, LIU C, LIN Q, et al. Domain conditioned adaptation network[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washington: AAAI Press, 2020: 11386-11393.
- [16] GANIN Y, USTINOVA E, AJAKAN H, et al. Domain-adversarial training of neural networks[J]. Journal of Machine Learning Research, 2016, 17(1): 2096-2130.
- [17] SAITO K, WATANABE K, USHIKU Y, et al. Maximum classifier discrepancy for unsupervised domain adaptation[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 3723-3732.
- [18] WANG F, JIANG M Q, QIAN C, et al. Residual attention network for image classification[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 6450-6458.
- [19] KANG G, ZHENG L, YAN Y, et al. Deep adversarial attention alignment for unsupervised domain adaptation: The benefit of target expectation maximization[C]//Proceedings of the European Conference on Computer Vision. New York: ACM Press, 2018: 420-436.
- [20] WANG X, LI L, YE W, et al. Transferable attention for domain adaptation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washington: AAAI Press, 2019: 5345-5352.
- [21] ESTRACH J B, ZAREMBA W, SZLAM A, et al. Spectral networks and deep locally connected networks on graphs[C]//Proceedings of the 2nd International Conference on Learning Representations. [S.l.:s.n.], 2014.
- [22] LUO Y, WANG Z, HUANG Z, et al. Progressive graph learning for open-set domain adaptation[C]//Proceedings of the 37th International Conference on Machine Learning. [S.l.:s.n.], 2020: 6468-6478.
- [22] LUO Y D, WANG Z J, HUANG Z, et al. Progressive graph learning for open-set domain adaptation[C]//Proceedings of the Proceedings of the 37th International Conference on Machine Learning. New York: ACM Press. [S.l.:s.n.], 2020: 6468-6478.
- [23] YANG X, DENG C, LIU T L, et al. Heterogeneous graph attention network for unsupervised multiple-target domain adaptation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(4): 1992-2003.
- [24] SAENKO K, KULIS B, FRITZ M, et al. Adapting visual category models to new domains[C]//Proceedings of the 11th European Conference on Computer vision: Part IV. New York: ACM Press, 2010: 213-226.
- [25] LONG M, CAO Y, CAO Z, et al. Transferable representation learning with deep adaptation networks[J]. IEEE Transactions



- on Pattern Analysis and Machine Intelligence, 2018, 41(12): 3071-3085.
- [26] VENKATESWARA H, EUSEBIO J, CHAKRABORTY S, et al. Deep hashing network for unsupervised domain adaptation[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 5385-5394.
- [27] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2009: 248-255.
- [28] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 770-778.
- [29] LONG M S, ZHU H, WANG J M, et al. Deep transfer learning with joint adaptation networks[C]//Proceedings of the Proceedings of the 34th International Conference on Machine Learning-Volume 70. New York: ACM Press, 2017: 2208-2217.
- [30] LI M X, ZHAI Y M, LUO Y W, et al. Enhanced transport distance for unsupervised domain adaptation[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 13933-13941.
- [31] HUANG J X, GUAN D Y, XIAO A R, et al. Category contrast for unsupervised domain adaptation in visual tasks[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 1193-1204.
- [32] CHEN M, ZHAO S, LIU H, et al. Adversarial-learned loss for domain adaptation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washington: AAAI Press, 2020: 3521-3528.

[作者简介]



张永（1975-），男，博士，湖州师范学院教授、博士生导师，辽宁师范大学计算机与信息技术学院教授，主要研究方向为数据挖掘、智能计算、情感计算。



刘昊双（1996-），女，辽宁师范大学计算机与信息技术学院硕士生，主要研究方向为数据挖掘、机器学习。



章琪（1999-），女，辽宁师范大学计算机与信息技术学院硕士生，主要研究方向为机器学习、智能计算。



刘文哲（1983-），女，博士，湖州师范学院讲师，主要研究方向为多视角学习、深度学习。