计算机科学
COMPUTER SCIENCE

基于因果关系的领域泛化长尾学习

吕佳豪, 刘进锋

引用本文

吕佳豪, 刘进锋. 基于因果关系的领域泛化长尾学习[J]. 计算机科学, 2024, 51(11A): 240300041-8.

LYU Jiahao, LIU Jinfeng. Domain Generalization and Long-tailed Learning Based on Causal Relationships [J]. Computer Science, 2024, 51(11A): 240300041-8.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240300041

基于因果关系的领域泛化长尾学习

吕佳豪　刘进锋
宁夏大学信息工程学院　银川 750021
(12022131956@stu.nxu.edu.cn)

摘　要　以深度学习为代表的机器学习方法已经得到了广泛的应用并取得了很多成就,数据集分布偏移和长尾分布问题会使传统的深度学习方法性能出现显著下降,而这两个问题也常常存在于真实场景的数据集中。虽然领域泛化和长尾学习研究已经使这两个问题单独得到了较好的解决,但在分布偏移和长尾分布相结合(LT-DS)的复杂场景下,单一的领域泛化和长尾学习方法效果并不太好。针对LT-DS问题,可以从因果关系出发,统一地去解决这两个问题。对于分布偏移,通过傅里叶变换进行因果干预及因果分解,并通过去相关加权来获得一个跨域不变的因果特征表示。对于长尾分布,通过去混淆训练构建一个因果效应分类器来消除动量所带来的偏差,并通过Balanced Softmax和logit调整来进一步消除长尾分布带来的影响。实验结果表明,该方法在LT-DS问题上比现有最好方法在AWA2-LTS数据集和ImageNet-LTS数据集上分别平均高出了8%和5%,表现出有竞争力的结果。

关键词:深度学习;领域泛化;长尾学习;因果推断
中图分类号　TP391.41

Domain Generalization and Long-tailed Learning Based on Causal Relationships

LYU Jiahao and LIU Jinfeng
School of Information Engineering,Ningxia University,Yinchuan 750021,China

Abstract　Deep learning,as a representative of machine learning methods,has been widely applied and achieved many successes. However,problems such as dataset distribution shift and long-tailed distribution can significantly degrade the performance of traditional deep learning methods,and these two issues often exist in real-world datasets. Although domain generalization and long-tailed learning research have provided good solutions to these two problems separately,the effect of a single domain generalization or long-tailed learning method is not satisfactory in the complex scenario of combining distribution shift and long-tailed distribution(LT-DS). To address the LT-DS problem,a unified approach can be taken from a causal perspective to solve both issues simultaneously. For distribution shift,causal intervention and decomposition can be achieved through Fourier transform,and cross-domain invariant causal feature representations can be obtained through decorrelation weighting. For long-tailed distribution,a causal effect classifier can be constructed through debiasing training to eliminate momentum-induced biases,and further eliminate the impact of long-tailed distribution through Balanced Softmax and logit adjustment. Experimental results show that this method outperforms the best existing methods by an average of 8% and 5% on the AWA2-LTS dataset and ImageNet-LTS dataset,respectively,demonstrating competitive results on the LT-DS problem.

Keywords　Deep learning,Domain generalization,Long-tailed learning,Causal inference

1　引言

近些年来,随着机器学习的发展,以深度学习为代表的各种机器学习方法已经在各个领域取得了令人瞩目的成就。然而这些方法都经常依赖于独立同分布的假设[1],但是在真实世界中很难满足这样的假设,因为可获得的源域数据与未知的目标域数据之间常常会存在分布偏移。当分布偏移发生时,机器学习模型在目标域数据上性能往往会显著下降。这严重妨碍了机器学习模型在医疗健康、无人驾驶、国防军事等精密行业的部署应用。领域泛化(Domain Generalization)研究就是为了解决这个问题,提高模型在不可见的目标域上的泛化性能。

长尾分布(Long-tailed Distribution)问题也是真实世界数据集常常存在的一个问题。一个长尾分布的图像分类数据集中,少数类别会占据大量的样本,称为头部类;而多数其他类别样本只占据少量的样本,称为尾部类[2]。由于数据集的长尾分布,训练得到的模型容易在预测时偏向头部类,也就是说容易把尾部类识别为头部类,从而使得分类精度相比平衡数据集显著下降。而长尾学习的目标就是提高模型在面对长尾分布的数据集时的学习能力。

分布偏移和长尾分布一直是真实世界数据集中的两个主要问题,虽然解决分布偏移问题的领域泛化研究和解决长尾分布问题的长尾学习研究目前已经提出不少单独解决这些问题的有效方法,但很少有研究考虑到分布偏移和长尾分布共存(LT-DS)的复杂情况。设 X 代表样本,Y 代表样本所属类别,由于领域泛化的先验假设会要求平衡的分布 $P(Y)$,而长

基金项目:宁夏自然科学基金(2023AAC03126)
This work was supported by the Natural Science Foundation of Ningxia,China(2023AAC03126).
通信作者:刘进锋(jfliu@nxu.edu.cn)

尾学习也会先验假设相同的分布 $P(X|Y)$，因此目前单独的领域泛化方法或长尾学习方法并不能有效地解决 LT-DS 问题的复杂情况。而在现实场景中，这两种问题相结合的复杂情况并不少见。

近年来，因果推断方法由于具有良好的可解释性和效果，在与生物、医学、经济相关的领域得到了越来越广泛的应用[3]。如何将因果推断方法应用于计算机视觉的研究也受到了越来越多研究者的关注。有学者指出，领域泛化任务本质上就是一个因果发现任务，因果关系可以被视为一种分布稳健性的形式[4]。同时，通过因果推断方法分析因果关系也可用于消除数据集的偏差。而目前无论是领域泛化研究还是长尾学习研究中，因果推断的应用还较少，并未充分发挥因果推断方法的理论优势。因此，使用因果推断的方法将因果关系的思想引入领域泛化和长尾学习当中是解决这两种问题的一个很好的研究思路。对此本文结合因果推断理论，提出了一种有效的方法来解决 LT-DS 问题。首先利用傅里叶变换进行因果干预，通过相关性矩阵构建出一个因果分解模块，并通过结合去相关样本加权来学习一个具有跨域不变性的因果特征表示；然后构建了一个因果效应分类器并利用 Balanced Softmax[5] 和 logit 调整[6] 对损失函数进行了优化改进来消除长尾分布带来的影响；随后在两个结合了分布偏移特性和长尾分布特性的数据集[7] AWA2-LTS 和 ImageNet-LTS 上进行了实验。结果表明，本文方法取得了有竞争力的结果。

本文的主要贡献如下：

(1)提出了一种基于因果关系的领域泛化长尾学习框架，并取得了较好的结果。

(2)所提方法从因果关系角度出发使领域泛化和长尾分布问题得到了统一的解决，并有着较好的可解释性。

(3)对传统的交叉熵损失函数进行了优化改进，使其更适应于长尾学习。

2 相关工作

近年来，为解决领域泛化问题，研究者们已经提出了不少方法。这些方法根据研究手段的不同可以分为数据调整、表示学习、学习策略三大类[8]。数据调整类方法注重于调整模型的输入数据来促进学习过程，分为数据增强和数据生成这两条技术路线。例如 zhou 等[9] 提出一种 MixStyle 的方法，该方法通过混合特征信息在特征空间中合成新领域，从而实现数据增强；而 CuMix[10] 通过混合训练期间可用的多个源领域和类别来模拟测试时的领域和语义转移，以生成越来越复杂的训练样本进行数据生成来提高领域泛化能力。但这些数据调整方法并不能有效处理 LT-DS 场景中的长尾分布问题。对于表示学习类的方法，由于领域泛化任务的关键就是学习一个跨域不变的表示，所以该方法是领域泛化中最为流行的方法。例如 Epi-FCR[11] 通过一种情节训练策略来构建一个特征提取器从而学习鲁棒的域不变特征表示。但单纯的学习域不变特征表示并不能解决 LT-DS 中长尾分布导致的分类器向头部类偏移的问题。学习策略类方法注重于用一般性的学习策略来促进泛化能力的提升。例如 DAML[12] 提出一种域增强元学习方法，该方法通过设计的元学习任务和损失来进行跨域的元学习，以同时保留域的独特知识并跨域泛化。这些基于学习策略的方法由于只专注于改善领域泛化性能，并事先依赖平衡数据集的假设，因此并不能很好地解决 LT-DS 中的长尾分布问题。由此可知，单一的领域泛化方法并不能解决 LT-DS 问题。

现有的长尾学习方法根据技术特点的不同，可分为三大类，即类再平衡、信息增强、模块改进。类再平衡是一种主流的长尾学习方法，该类方法的主要思路是寻求重新平衡训练样本数中类不平衡所带来的负面影响，主要有重采样、重加权等方法。如 cRT[13] 通过类别平衡采样重新训练分类器和采用分类器权重归一化方法来调整分类器的决策边际从而提高其长尾学习能力。信息增强旨在在模型训练中引入额外的信息，从而提高长尾学习的模型性能，最常用的方法是数据增强。例如 Remix[14] 通过解耦特征和标签的混合因子来混合样本进行数据增强从而利于尾部类。模块改进主要是改进网络模块的方法，比如改进分类器设计或者改进损失函数等。如 BSCE[5] 通过对传统的 Softmax 函数进行改进，提出了 Balanced Softmax 对交叉熵损失函数进行改进，从而最小化了 Softmax 回归的泛化界以适应长尾分布的变化；而 Equal[15] 提出了一种梯度驱动的训练机制对训练过程进行改进，该方法引入了一类新的梯度驱动损失函数，从梯度不平衡的角度改善了长尾学习问题。大多数现有的长尾学习方法并没有考虑到分布偏移带来的影响，并遵循着独立同分布的假设，因此这些长尾学习方法难以在 LT-DS 问题中获得良好的表现。

Judea Pearl 作为因果推断方法的开创者，提出了基于因果关系的方法论，创造了因果图、后门调整、前门调整、do 算子等概念[16]。这为因果推断方法的应用奠定了基础。而因果推断方法在领域泛化和长尾学习这两个领域也有着一些应用。Lv 等[17] 提出了一种因果启发的表示学习方法来解决领域泛化问题，该方法通过因果推断方法来利用因果特征表示的不变性进行领域泛化，并取得了不错的表现。而 Tang 等[18] 就长尾学习问题设计了一个因果推理框架，该方法通过因果干预的方法消除了动量作为混淆因子的影响，从而提高了模型的长尾学习能力。这些基于因果关系的方法在领域泛化和长尾学习中的应用为解决 LT-DS 问题提供了新的研究视角。

目前专门针对 LT-DS 问题的研究还很少，Zhang 等的研究[19] 虽然为了提高领域泛化的难度，在领域泛化数据集中进行了长尾分布设置，但并没有对此问题进入深入研究和探讨。Gu 等的研究[7] 首次正式定义了 LT-DS 问题，并提出了两个相关的数据集 AWA2-LTS 和 ImageNet-LTS。在此基础上还提出了一个针对 LT-DS 问题的强基线方法 ML-LTDG，该方法通过一个元学习框架来集成分布校准分类损失、视觉语义映射和语义相似性引导增强 3 个核心功能块来解决 LT-DS 问题。综上所述，大多数现有的领域泛化方法或者长尾学习方法并不能很好地应对 LT-DS 问题，而基于因果关系的方法在领域泛化和长尾学习中的应用为本文的研究提供了思路。因此，本文就 LT-DS 问题提出一种基于因果关系的领域泛化长尾学习方法。

3 实验方法

3.1 基于因果关系的领域泛化

针对领域泛化中图像分类任务的过程，根据样本数据与标签之间的因果关系对其进行分析。首先假设数据样本 X 中与类别标签 Y 相关的信息为因果因素 X_C，其与标签的关系独立于域分布，具体表现为图像中目标物体的形状、纹理等特征。而与类别无关的信息为非因果因素 X_U，具体表现为图像中的背景、风格等特征。然后对其建立结构因果模型[17]，

如图 1 所示。每个原始的数据样本 X 由因果因素 X_C 和非因果因素 X_U 混合构成，其中因果因素 X_C 才会影响类别标签 Y 的判别。想要从原始输入样本 X 中提取因果因素 X_C，然后重建不变的因果机制，可以借助因果干预 $P(Y|\mathrm{do}(X_U),X_C)$ 来完成，$\mathrm{do}(X_U)$ 表示对非因果因素的干预。

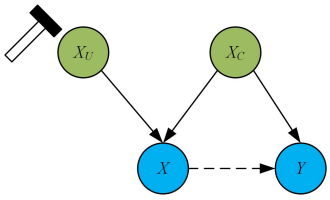


图 1　领域泛化因果图

Fig. 1　Domain generalization causal graph

为了获得因果特征表示，首先要通过因果干预将因果因素 X_C 从非因果因素 X_U 中分离出来。虽然因果因素 X_C 的映射函数 $g(\cdot)$ 通常是未知的，但能确定干预非因果因素 X_U 即 $P(S|\mathrm{do}(X_U))$，因果因素 X_C 会保持不变。傅里叶变换是一种常用的图像处理手段，傅里叶变换频谱图有一个特性：其相位部分保留了原始信号的高级语义，而振幅部分包含低级别统计量[20]。因此傅里叶变换频谱图的相位部分可以和因果因素建立联系，而振幅部分可以和非因果因素建立联系。

对于一张原始图像 x^o，它的傅里叶变换可以表示为：

$$\mathcal{F}(x^o)=\mathcal{A}(x^o)\times \mathrm{e}^{-j\times\mathcal{P}(x^o)} \tag{1}$$

其中，$\mathcal{A}(x^o)$ 表示振幅部分，$\mathcal{P}(x^o)$ 表示相位部分，通过在原始图像 x_i^o 和随机采样的图像 x_j^o 之间进行线性插值并按一定比例进行混合，从而对振幅信息进行扰动干预，如下式所示：

$$\hat{\mathcal{A}}(x^o)=(1-\lambda)\mathcal{A}(x_i^o)+\lambda\mathcal{A}(x_j^o) \tag{2}$$

其中，$\lambda\sim U(0,\eta)$，η 控制了扰动的强度。然后将混合振幅和原始的相位进行结合，通过逆傅里叶变换 $\mathcal{F}^{-1}$ 得到增强图像，如下式所示：

$$\mathcal{F}(x^a)=\hat{\mathcal{A}}(x^o)\times e^{-j\times\mathcal{P}(x^o)},x^a=\mathcal{F}^{-1}(\mathcal{F}(x^a)) \tag{3}$$

综上，整个因果干预的效果如图 2 所示，从图中可以看出，得到的增强图像虽然产生了颜色上的一些畸变，但是还是可以清晰地看出马还是马，房子还是房子。由此可知，通过对振幅进行扰动干预模拟了对非因果因素的干预，而非因果因素的变化并不会影响类别的判别，这为捕获到因果因素创造了条件。

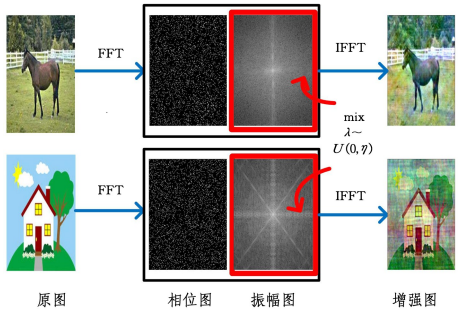


图 2　因果干预生成增强图像

Fig. 2　Causal intervention for generating augmented images

在上述因果干预的基础上，为了捕获到因果因素，需要通过一个因果分解模块将因果因素与非因果因素分离开来。因果因素之间应该是相互独立的，也就是说它们之间互不依赖各自的信息。这里通过图像的特征表示的维度信息来模拟因果因素的分解。具体来说，将表示生成器表示为 $\hat{g}(\cdot)$，则特征表示 $\boldsymbol{r}=\hat{g}(x)\in\mathbb{R}^{1\times N}$，这里 N 表示维度的维数。COR 函数可以用来计算两个变量之间的相关性系数。为了模拟对 Xu 的干预保持不变的因果因素，通过 COR 函数来优化 $\hat{g}(\cdot)$ 最大化原始图像和增强图像的特征表示中同一维度的相关性，使生成的特征表示在干预下，各维度能具有不变性，如式(4)所示：

$$\max_{\hat{g}}\frac{1}{N}\sum_{i=1}^{N}COR(\tilde{\boldsymbol{r}}_i^o,\tilde{\boldsymbol{r}}_i^a) \tag{4}$$

其中，$\tilde{\boldsymbol{r}}_i^o$ 和 $\tilde{\boldsymbol{r}}_i^a$ 代表原始图像的特征表示集合 $\boldsymbol{R}^o=[(\boldsymbol{r}_1^o)^{\mathrm{T}},\cdots,(\boldsymbol{r}_B^o)^{\mathrm{T}}]^T\in\mathbb{R}^{B\times N}$ 和增强图像的特征表示集合 $\boldsymbol{R}^a=[(\boldsymbol{r}_1^a)^{\mathrm{T}},\cdots,(\boldsymbol{r}_B^a)^{T}]^{\mathrm{T}}\in\mathbb{R}^{B\times N}$ 第 i 列的 Z 标准化分数，其中 B 表示批量大小，$\boldsymbol{r}_B^o$ 和 $\boldsymbol{r}_B^a$ 分别表示一个批量中最后一个样本由 $\hat{g}(x)$ 得到的特征表示。通过 COR 函数来衡量干预前后特征表示的相关性对表示生成器 $\hat{g}(x)$ 优化，可以有效地将因果因素和非因果因素分离开来。此外，为了使代表因果因素分解因子的各维度能够保持一定的独立性，需要最小化原始图像和增强图像的特征表示中不同维度的相关性，如式(5)所示：

$$\min_{\hat{g}}\frac{1}{N(N-1)}\sum_{i\neq j}COR(\tilde{\boldsymbol{r}}_i^o,\tilde{\boldsymbol{r}}_j^a),i\neq j \tag{5}$$

为了能够综合优化式(4)和式(5)，如式(6)所示建立了一个相关性矩阵 $\boldsymbol{M}$：

$$\boldsymbol{M}_{ij}=\frac{\langle\tilde{\boldsymbol{r}}_i^o,\tilde{\boldsymbol{r}}_j^a\rangle}{\|\tilde{\boldsymbol{r}}_i^o\|\ \|\tilde{\boldsymbol{r}}_j^a\|},i,j\in1,2,\cdots,N \tag{6}$$

其中，$\langle\cdot\rangle$ 表示内积操作。把特征表示 $\boldsymbol{R}^o$ 和 $\boldsymbol{R}^a$ 中的相同维度视为正对，不同维度视为负对，通过该相关性矩阵 $\boldsymbol{M}$ 来最大化正对的相关性，最小化负对的相关性。在此基础上构建一个因果分解损失来进行优化，如式(7)所示：

$$\mathcal{L}_{cf}=\frac{1}{2}\|\boldsymbol{M}-\mathbf{I}\|_{\mathrm{F}}^2 \tag{7}$$

其中，$\mathbf{I}$ 表示单位矩阵。因为相关性最大值为 1，最小值为 0，所以通过式(7)进行优化时，相关性矩阵上代表正对的对角线元素会趋近于 1，这体现了对非因果因素干预后因果因素的不变性，同时也将非因果因素和因果因素从混合中分离开来。另外，代表负对的非对角线元素趋近于 0，体现了特征表示各维度之间的独立性。通过最小化 $\mathcal{L}_{cf}$，有助于得到干净独立的特征表示，从而满足因果因素的性质。

由于特征表示中可能还存在一些与标签和因果因素相关联的关联特征，比如牛常常伴随着草这种关联事物，这种关联特征作为一种混淆因子容易产生一种虚假的相关性[19]。这种混淆因子的存在，使得上述方法得到的因果因素还无法拥有充分的因果关系来解释分类任务 $X\to Y$ 的分类判别，也不能够保证因果因素之间的独立性。通过重新加权数据进行协变量平衡来估计目标特征的影响，并通过变量去相关正则化器构造样本权重来减少协变量之间的相关性，可以消除混淆因子导致的虚假相关性而实现稳定预测[21]。假设有两个一维随机变量 A 和 B，根据希尔伯特-施密特独立准则(HSIC)[22]，可以得到：

$$\Sigma_{AB}=0\Leftrightarrow A\perp B \tag{8}$$

弗罗比尼乌斯范数在欧几里得空间中可以对应希尔伯特施密特范数[23]，因此独立检验统计量可以基于弗罗比尼乌斯范数。设部分交叉协方差矩阵如式(9)所示：

$$\hat{\Sigma}_{AB}=\frac{1}{n-1}\sum_{i=1}^{n}\left[(\boldsymbol{u}(A_i)-\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{u}(A_j))^{\mathrm{T}}\cdot(\boldsymbol{v}(B_i)-\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{v}(B_j))\right]$$

$$\boldsymbol{u}(A)=(u_1(A),u_2(A),\cdots u_{n_A}(A)),u_j(A)\in\mathcal{T}_{\mathrm{RFF}},\forall j$$
$$\boldsymbol{v}(B)=(v_1(B),v_2(B),\cdots v_{n_B}(B)),v_j(B)\in\mathcal{H}_{\mathrm{RFF}},\forall j \quad (9)$$

这里我们分别从 $\mathcal{H}_{\mathrm{RFF}}$ 中采样 n_A 和 n_B 函数，$\mathcal{H}_{\mathrm{RFF}}$ 表示随机傅里叶特征的函数空间，其形式如下式所示：

$$\mathcal{H}_{\mathrm{RFF}}=\{h:x\rightarrow\sqrt{2}\cos(wx+\phi)\mid w\sim N(0,1),\phi\sim \mathrm{Uniform}(0,2\pi)\} \quad (10)$$

其中，w 从标准正态分布中采样，ϕ 从均匀分布中采样。然后，将独立性检验统计量 I_{AB} 定义为部分交叉协方差矩阵的弗罗比尼乌斯范数，则有 $I_{AB}=\|\hat{\Sigma}_{AB}\|_{\mathrm{F}}^2$。当 I_{AB} 减小到零时，两个变量 A 和 B 趋于独立。如果这里的 A 和 B 是特征表示的话，那么 A 和 B 就进行了去相关，消除了关联特征的影响，使特征表示之间的独立性和具有的因果关系更强。为了达到这个效果，这里使用样本加权的方式进行迭代优化，消除表示空间中特征之间的依赖性，并通过 *RFF* 来度量独立性，用 ω 来表示样本权重，并且 $\sum_{i=1}^{n}\omega_i=n$。经过加权后，式(9)中随机变量 A 和 B 的部分交叉协方差矩阵可以表示如下：

$$\hat{\Sigma}_{AB;\omega}=\frac{1}{n-1}\sum_{i=1}^{n}\left[\left(\omega_i\boldsymbol{u}(A_i)-\frac{1}{n}\sum_{j=1}^{n}\omega_j\boldsymbol{u}(A_j)\right)^{\mathrm{T}}\cdot\left(\omega_i\boldsymbol{v}(B_i)-\frac{1}{n}\sum_{j=1}^{n}\omega_j\boldsymbol{v}(B_j)\right)\right] \quad (11)$$

其中，$\boldsymbol{u}$ 和 $\boldsymbol{v}$ 是式(9)中的 RFF 映射函数。对于特征表示 R_i 和 R_j、表示生成函数 g，以及预测函数 y，使用下列式子迭代地优化样本权重：

$$g^{(t+1)},f^{(t+1)}=\arg\min_{g,f}\sum_{i=1}^{n}\omega_i^{(t)}\mathcal{L}(f(g(X_i)),y_i) \quad (12)$$

$$\omega^{(t+1)}=\arg\min_{\omega\in\Delta_n}\sum_{1\leqslant i<j\leqslant m_R}\|\hat{\Sigma}_{R_i^{(t+1)}R_j^{(t+1)};\omega}\|_{\mathrm{F}}^2 \quad (13)$$

其中，$\boldsymbol{R}^{(t+1)}=g^{(t+1)}(\boldsymbol{X})$，$\mathcal{L}(\cdot,\cdot)$ 表示损失函数，t 表示时间，初始化权重 $\omega^{(0)}=(1,1,1,\cdots,1)^{\mathrm{T}}$。

通过式(12)和式(13)可以不断迭代更新权重 ω，由于图像样本 X_i 可分为原始图像样本和傅里叶变换得到的增强图像样本，因此可以分别迭代优化得到其权重 ω_{ori} 和 ω_{aug}，并将其乘到分类损失上得到 $\mathcal{L}_{\mathrm{cls}}^{\mathrm{ori}}$ 和 $\mathcal{L}_{\mathrm{cls}}^{\mathrm{aug}}$。

综上所述，基于因果关系的领域泛化方法最终优化的损失函数如式(14)所示：

$$\min_{\hat{g},\hat{h}_1,\hat{h}_2}\alpha\mathcal{L}_{\mathrm{cls}}^{\mathrm{ori}}+\beta\mathcal{L}_{\mathrm{cls}}^{\mathrm{aug}}+\gamma\mathcal{L}_{cf} \quad (14)$$

其中，α,β,γ 是权衡参数。$\mathcal{L}_{\mathrm{cls}}^{\mathrm{ori}}$ 代表了原始图像 x^o 得到的分类损失，$\mathcal{L}_{\mathrm{cls}}^{\mathrm{aug}}$ 代表通过傅里叶变换得到的增强图像 x^a 的分类损失，$\mathcal{L}_{cf}$ 如式(7)所示。通过样本权重对分类损失 $\mathcal{L}_{\mathrm{cls}}^{\mathrm{ori}}$ 和 $\mathcal{L}_{\mathrm{cls}}^{\mathrm{aug}}$ 的优化，加之因果分解损失的优化，便可促进模型学习一个跨域不变的因果特征表示，从而提高模型的领域泛化能力。

3.2 基于因果关系的长尾学习

动量是随机梯度下降(SGD)算法中一个重要的因素，一方面，动量确实显著提高了训练的稳定性，使得模型更容易收敛到理想的状态。在平衡数据集上，这是一个明显的优点。然而，另一方面，带有动量的 SGD 算法在求解时会考虑过去一段时间内的样例，并进行移动平均计算。在长尾分布数据集中，由于大部分的样本来自数据丰富的头部类别，这种移动平均计算会导致训练方向偏向头部类别，优化求解的方向更倾向于头部类别，因此，如何在长尾分类中平衡这两方面的影响，是解决长尾学习问题的一个关键点。为从动量的因果效应方面解释动量对长尾学习的影响，本节对长尾学习的具体过程建立了结构因果模型[18]，其中包括 4 个变量：动量(M)、特征(X)、特征对头部大类的偏移量(H)、模型预测(Y)之间的因果关系。其因果图如图 3 所示，节点 M 作为 X 和 Y 之间的混淆因子，通过后门路径 $X\leftarrow M\rightarrow H\rightarrow Y$ 会使 X 和 Y 之间产生虚假的相关性，具体表现为 X 与预测的 Y 无关，即尾部类会被错误地分类到头部类，产生一种坏的偏差影响。H 作为中介因子会通过 $M\rightarrow H\rightarrow Y$ 使 M 对 Y 产生间接影响，误判为头部类的尾部类和头部类也有一些相似之处，这是应该保留的一种好的偏差影响。

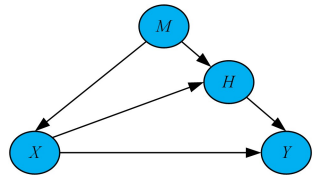


图 3 长尾学习因果图

Fig. 3 Causal diagram of long-tailed learning

通过对动量效应的因果图的分析可知，由于混淆因子 M 和中介因子 H 的存在，X 对 Y 产生的效应在长尾分布中会向头部类进行偏移，因此需要获得 X 和 Y 之间的直接因果效应。为此需要通过去混淆训练干预 M 来控制 M 对 X 的影响并利用反事实方法来消除中介因子的影响，从而获得如式(15)所示的总的直接因果效应(TDE)[24]。其中 x_0 表示值为 0 的空白输入，$TDE(Y_i)$ 表示第 i 类预测的因果效应，下标 h 表示在 $\mathrm{do}(X=x)$ 时，H 的取值为 h。do 算子表示对 X 进行因果干预，相当于对因果图中 $M\rightarrow X$ 这条路径进行了调整。因此在计算最终的 TDE 之前需要通过去混淆训练来估计 $\mathrm{do}(X=x)$ 调整后的值，从而消除混淆因子的坏影响，而保留中介因子的好影响。

$$\arg\max_{i\in C}TDE(Y_i)=[Y_h=i\mid\mathrm{do}(X=x)]-[Y_h=i\mid\mathrm{do}(X=x_0)] \quad (15)$$

为了通过因果干预 $\mathrm{do}(X=x)$ 从动量中保留其对特征学习的优化的好处，而消除其通过混杂效应带来的坏处，需要通过后门调整来切断后门路径 $X\leftarrow M\rightarrow H\rightarrow Y$ 带来的影响，其过程可以用下列式子来描述：

$$P(Y=i\mid\mathrm{do}(X=x))=\sum_m P(Y=i\mid X=x,M=m)P(M=m) \quad (16)$$

$$P(Y=i\mid\mathrm{do}(X=x))=\sum_m\frac{P(Y=i,X=x\mid M=m)P(M=m)}{P(X=x\mid M=m)} \quad (17)$$

因为在该式中 $M=m$ 有无穷种可能，所以想通过式(16)实现后门调整是很困难的。因此需要通过式(17)的逆概率加权[25]来逼近无穷的采样 $(i,x)\mid m$。然后通过后门调整可以将去混淆和未去混淆的状态之间的等价性联系起来，这就使的从前者状态中收集的样本与从后者提取的样本类似。因此式(16)可以简化为下式：

$$P(Y=i\mid\mathrm{do}(X=x))\approx\tilde{P}(Y=i,X=x\mid M=m) \quad (18)$$

其中，$\tilde{P}$ 表示逆概率加权，这里就可以得到去混淆模型概率分布 P 的 logit 表达式[26]：

$$P(Y=i\mid\mathrm{do}(X=x))=\tau\frac{(\boldsymbol{w}_i)^{\mathrm{T}}(\ddot{\boldsymbol{x}}+\boldsymbol{h})}{(\|\boldsymbol{w}_i\|+\gamma)\|\boldsymbol{x}\|}=\tau\frac{(\boldsymbol{w}_i)^{\mathrm{T}}\boldsymbol{x}}{(\|\boldsymbol{w}_i\|+\gamma)\|\boldsymbol{x}\|} \quad (19)$$

其中，τ 是一个类似于吉布斯分布中的逆温度的正比例因子，分子部分为正常的无偏差项的线性分类器，对应于特征向量 $\boldsymbol{x}$ 分解，而分母则是一个归一化项，起到对数值归一化的作用。结合该标准化分类器，再通过在训练过程中统计一个移

动平均特征$\boldsymbol{x}$就可以进行去混淆训练，去混淆训练切断了$M\to X$的路径，去除了混淆因子M的作用，如图4所示。经过去混淆训练后，已经消除了混淆因子M带来的坏影响。接下来需要消除中介因子H带来的坏影响，获得$X\to Y$的直接因果效应。根据反事实一致性原则[27]，可以得到以下关系：$[Y_h=i|\mathrm{do}(X=x)]=[Y=i|\mathrm{do}(X=x)]$。这使得可以使用式(19)来计算式(15)中等式右边的第一项。在计算$[Y_h=i|\mathrm{do}(X=x)]$时，可以用一个代表空白输入的零向量x_0来代替而保持其他不变，从而达到一个“欺骗”模型的效果，捕获到中介因子的效应。然后通过式(20)就可以计算出$X\to Y$的直接因果效应(TDE)：

$$TDE(Y_i)=\tau\left(\frac{(\boldsymbol{w}_i)^{\mathrm{T}}\boldsymbol{x}}{(\|\boldsymbol{w}_i\|+\gamma)\|x\|}-\alpha\cdot\frac{\cos(\hat{\boldsymbol{x}},\hat{\boldsymbol{h}})\cdot(\boldsymbol{w}_i)^{\mathrm{T}}\hat{\boldsymbol{h}}}{\|\boldsymbol{w}_i\|+\gamma}\right)\tag{20}$$

其中，α是一个权衡参数，用于控制直接因果效应和间接效应之间的比例。通过该式计算出的直接因果效应反映在因果图上如图5所示。

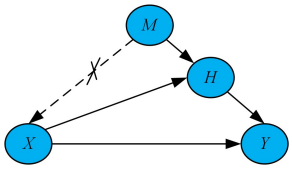


图4　去混淆训练因果图

Fig. 4　Causal diagram of removing confuse training

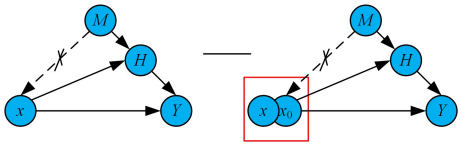


图5　得到TDE的因果图

Fig. 5　Causal diagram of obtaining TDE

通过因果干预后的值$\mathrm{do}(X=x)$与空白输入$\mathrm{do}(X=x_0)$的差值就可以捕获到$X\to Y$的直接因果效应，从而建立一个因果效应分类器。通过该因果效应分类器便可有效消除长尾分布中动量带来的偏差影响，建立分类判别$X\to Y$之间的直接联系，从而提高模型的长尾学习能力。

3.3　Balanced Softmax和logit调整相结合

为了进一步在保持领域泛化能力的同时提高长尾学习能力，这里使用Balanced Softmax和logit调整对交叉熵损失函数进行了优化改进，从而通过增大头部类和尾部类的相对边际以及最小化长尾分布下的泛化错误上界来抑制长尾分布的影响，提高原来领域泛化方法的长尾学习能力。将Balanced Softmax和logit调整相结合，可以得到最终的优化的损失函数，如下式所示：

$$\mathscr{L}(y,f(x))=-\log\frac{n_y\mathrm{e}^{f_y(x)+k\cdot\log\pi_y}}{\sum\limits_{y'\in[L]}n_{y'}\mathrm{e}^{f_{y'}(x)+k\cdot\log\pi_{y'}}}\tag{21}$$

其中，$f(x)$表示模型的输出结果，y代表真实标签，该式通过乘上一个权重和加上偏移量，在最小化泛化错误上界的同时，增大了头部类和尾部类的相对边际，从而减少了数据集长尾分布带来的影响。权重和偏移量来自训练样本的先验估计。用式(20)来计算输出结果$f(x)$，用该式作为分类损失就可以计算出式(14)的$\mathscr{L}_{\mathrm{cls}}^{\mathrm{ori}}$和$\mathscr{L}_{\mathrm{cls}}^{\mathrm{aug}}$，从而对整个模型进行优化。

3.4　总体方案

综合第3节所描述的各个模块，可以得到本文的总体方案——基于因果关系的领域泛化长尾学习方法。该方法的过程如图6所示，具体来说，首先随机取一对图片，利用傅里叶变换混合它们的振幅部分来生成增强图像对非因果因素进行干预，此时经过特征提取器得到的特征表示是嘈杂的和相互依赖的表示；然后通过因果分解和去相关加权模块的作用可以得到因果分解损失$\mathscr{L}_{cf}$和权重ω_{ori}及ω_{aug}。通过它们对分类损失的优化，可以得到干净的、相互独立的表示；之后，原始图像和增强图像的特征表示各自进入一个因果效应分类器，通过该因果效应分类器可以获得分类判别的直接因果效应而消除长尾分布中动量带来的偏差影响；最后通过Balanced Softmax和logit调整改进分类损失函数，进一步消除长尾分布的影响。通过这一系列过程可以使模型在保持良好领域泛化能力的同时，又有着较好的长尾学习能力。

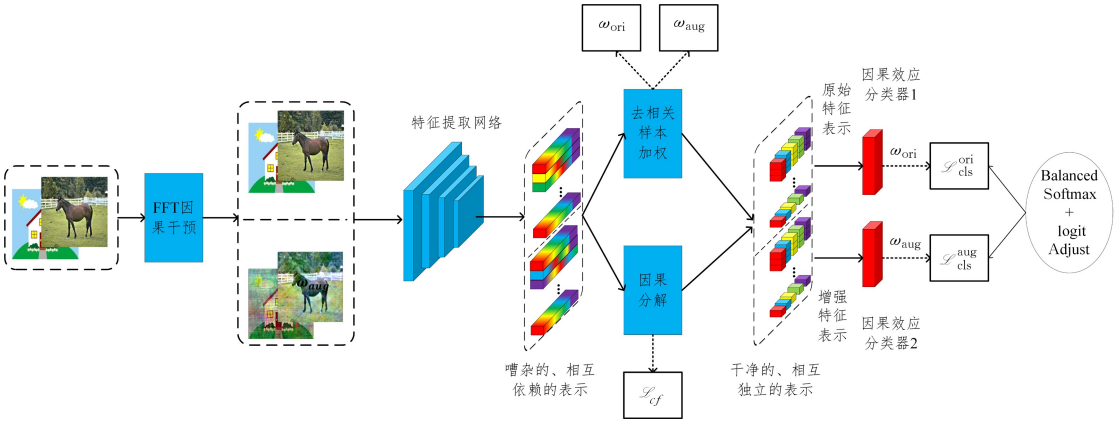


图6　方法流程示意图

Fig. 6　Diagram of method flow

4　实验结果及分析

4.1　数据集

本文使用AWA2-LTS和ImageNet-LTS这两个数据集来评估所提方法，两个数据集同时满足了领域泛化的分布偏移特性和数据集长尾分布的特性，因此能同时体现出所提方法的领域泛化性能和长尾学习性能。AWA2-LTS数据集总共有约8 000张大小不一的图像，包含Original(O)，Hayao(H)，Shinkai(S)，Vangogh(V)，Ukiyoe(U) 5种风格域的图像，用于模拟分布偏移并有50个类别。ImageNet-LTS数据集总共有约100 000张大小不一的图像，同样有Original(O)，Hayao(H)，Shinkai(S)，Vangogh(V)，Ukiyoe(U) 5种风格域

的图像以及1000个类别。这两个数据集中每个风格域的样本数目都是长尾分布的。最大不平衡比例是反映长尾分布数据集头部类与尾部类差距的属性，代表了样本最多的类与样本最少的类的比值。AWA2-LTS的最大不平衡比例是78，ImageNet-LTS的最大不平衡比例是256。数据集训练集、验证集和测试集的划分遵循先前研究[7]的设置。

4.2 实验设置

本文方法的实验均在NVIDIA GeForce RTX 3090上执行。使用ResNet10作为特征提取网络，在训练时每张图像都会经过预处理设置为224×224大小，使用的优化器为SGD优化器，批量大小为32，学习率为0.1，momentum为0.9，weight decay为1×10^{-4}，总共训练100个epoch，每过40个epoch学习率衰减为原来的十分之一。本文采用的评价指标有两个：其中一个是AccU，代表在未知的测试域上图像分类的准确率(%)；另一个是Acc，代表所有的5个风格域上的测试域上准确率的平均值(%)。为了体现本文方法能在保持领域泛化能力的同时提高长尾学习能力，本文方法对比了领域泛化中一些方法，包括Epi-FCR[11]，CuMix[10]，DAML[12]，MixStyle[9]；还对比了一些长尾学习方法，包括cRT[13]，BSCE[5]，Equal[15]，Remix[14]；此外还对比了这两种类型方法的一些组合。同时也对比了开创性研究LT-DG问题的方法ML-LTDG[7]。实验结果使用了该研究里报道的结果。

4.3 实验结果

本文所提出的方法与其他方法在AWA2-LTS数据集上的对比结果如表1所列。由结果可知，本文方法相比其他方法在5个风格域上的两个评价指标上均取得了最优结果。其中AccU是最能体现出领域泛化能力的指标，其最差在O这个风格域比次优方法ML-LTDG高出了4.1%，整体比次优方法ML-LTDG平均高出了8%左右。Acc是最能体现长尾学习能力的指标，在该指标上，本文方法整体比次优方法平均高出了6%左右，同时从结果上看，单独的领域泛化方法和长尾学习方法表现都比较差，这说明单独的领域泛化或者长尾学习方法很难应对领域泛化和长尾学习相结合下的复杂情况。将一些典型的领域泛化方法和长尾学习方法相结合，如MixStyle+BSCE，Epi-FCR+BSCE，虽然两个评价指标都有一些提升，但也差于本文提出的方法。这充分说明了本文方法在领域泛化和长尾学习相结合的复杂情况下的有效性。

表1 AWA2-LTS数据集上的结果

Table 1 Results on AWA2-LTS dataset

方法	Original		Hayao		Shinkai		Vangogh		Ukiyoe	
	AccU	Acc	AccU	Acc	AccU	Acc	AccU	Acc	AccU	Acc
cRT[13]	30.4	29.1	23.5	33.6	34.7	35.8	28.6	35.8	28.4	36.7
BSCE[5]	41.8	35.9	24.7	36.1	30.2	35.8	29.0	37.7	25.9	33.6
Equal[15]	34.1	32.9	24.3	35.3	33.5	36.2	28.8	35.8	27.3	34.7
Remix[14]	32.7	30.3	16.9	30.7	27.6	32.0	26.9	31.8	26.5	32.0
Epi-FCR[11]	34.0	33.1	23.3	34.0	29.7	35.5	27.5	36.1	27.0	35.7
MixStyle[9]	36.7	34.0	27.1	36.2	32.0	36.2	28.4	36.0	28.8	36.2
CuMix[10]	36.1	33.8	24.7	35.3	30.2	35.1	28.2	35.1	26.5	34.7
DAML[12]	42.2	35.3	25.7	35.2	31.2	36.8	29.4	37.5	28.6	36.0
MixStyle+ BSCE	40.0	36.8	28.8	39.7	32.4	38.3	30.8	38.2	29.8	38.9
Epi-FCR+ BSCE	41.3	36.9	24.0	35.9	32.0	39.2	30.1	38.5	26.6	35.9
ML-LTDG[7]	49.4	42.1	29.8	42.4	34.3	42.6	32.7	40.3	32.9	42.4
ours	**53.5**	**47.5**	**39.2**	**48.6**	**41.6**	**48.7**	**45.3**	**46.7**	**39.2**	**47.2**

表2列出了本文方法和其他方法在ImageNet-LTS数据集上的结果。对于ImageNet-LTS数据集，由于该数据集比较复杂且庞大，所以整体的准确率都偏低，但本文方法相比其他方法在5个风格域及两个指标上也取得了最优的结果。从AccU来看，本文方法最差在O这个风格域比次优方法ML-LTDG高出了3.2%，整体比次优方法平均高出了5%左右。从Acc看，本文方法整体比次优方法ML-LTDG平均高出了4%左右，这进一步说明了本文方法有效性和优越性。

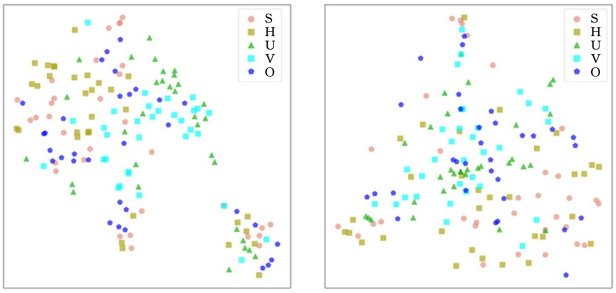
两个数据集的实验结果表明，本文所提出的方法具有显著的优势，分类精度明显优于现有方法，表现出有竞争力的结果。这证明基于因果关系的方法有效地提高了模型的领域泛化能力和长尾学习能力。对于领域泛化，基于因果关系的方法能捕获到具有因果不变性的特征表示，从而提高模型的领域泛化能力。对于长尾学习，基于因果关系的方法能有效地消除动量带来的对数据集长尾分布时向头部类偏移带来的坏影响而保留其好影响。同时，基于因果关系的方法也从因果推断角度为方法提供了一定的可解释性。

表2 ImageNet-LTS数据集上的结果

Table 2 Results on ImageNet-LTS dataset

方法	Original		Hayao		Shinkai		Vangogh		Ukiyoe	
	AccU	Acc	AccU	Acc	AccU	Acc	AccU	Acc	AccU	Acc
cRT[13]	20.7	18.7	13.8	19.0	16.2	18.8	14.0	18.4	13.6	18.7
BSCE[5]	20.8	19.1	14.3	19.4	16.5	19.1	14.7	18.8	14.1	19.1
Equal[15]	16.3	15.4	10.7	15.2	13.2	16.4	10.5	14.8	10.9	15.8
Remix[14]	14.8	13.8	10.1	14.1	11.3	14.1	11.1	13.5	10.5	14.8
Epi-FCR[11]	19.2	18.8	13.5	19.0	15.0	20.0	12.2	18.0	13.0	17.5
MixStyle[9]	17.7	16.4	12.1	16.7	13.6	16.5	11.8	16.0	11.5	16.2
CuMix[10]	18.2	17.2	13.2	17.5	14.2	17.1	12.1	16.8	12.1	17.1
DAML[12]	14.7	12.7	10.5	13.0	11.5	13.2	9.7	12.8	10.1	13.0
ML-LTDG[7]	24.3	20.8	16.3	21.3	17.4	20.9	14.3	20.3	15.4	20.3
ours	**27.5**	**25.1**	**21.3**	**25.3**	**22.0**	**25.4**	**23.3**	**24.3**	**19.9**	**22.5**

t-SNE 可以把高维的特征降维到 2 维或 3 维进行可视化，具有相似性质的特征点会被聚类到一起。图 7 展示了相同类别但不同域的图像通过本文方法提取到的特征表示与只用 ResNet10 提取到的特征表示的 t-SNE 特征降维可视化之间的区别。



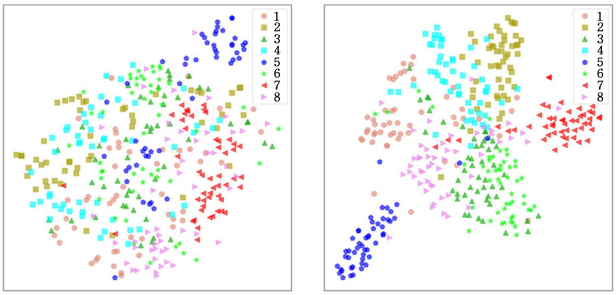
(a) ResNet　　(b) Ours

图 7　领域泛化的 t-SNE 特征降维可视化分析

Fig. 7　Visualization analysis of t-SNE feature dimensionality reduction in domain generalization

从图中可以看出本文方法提取的图像特征表示在 5 个域上没有被区分开来，全部聚集在一起，而只用 ResNet10 提取的图像特征表示在一些域上有区分聚集的现象，这说明通过本文方法提取的图像特征表示具有更多跨域不变性，体现了本文方法在一定程度上学到了具有不变性的因果特征表示，从而能具有较好的领域泛化能力。

为了进一步说明本文方法的有效性，使用 t-SNE 特征降维可视化对本文方法的长尾学习能力进行了分析。图 8 展示了不同类别图像的特征表示的 t-SNE 特征降维可视化效果，1－8 代表了不同的分类类别，且数字越小的类是样本数越多的头部类，数字越大的类是样本数越少的尾部类。从图中可以看出通过本文方法提取到的特征表示聚类效果要明显好于只用 ResNet10 提取到的特征表示。特别是对于 7 和 8 这些尾部类，本文方法也体现出了较好的聚类效果。这种聚类效果反映出了特征表示的可分类性，说明本文方法在长尾学习下也有着较好的分类效果。



(a) ResNet　　(b) Ours

图 8　长尾学习的 t-SNE 特征降维可视化分析

Fig. 8　Visualization analysis of t-SNE feature dimensionality reduction in long-tailed learning

t-SNE 特征降维可视化的结果有力地说明了本文方法能取得优秀性能的原因。对于领域泛化性能，本文方法由于学习到了一个跨域不变的因果特征表示，而这种因果特征表示不会因为域分布偏移而发生变化，因此能保持着良好的领域泛化能力。对于长尾学习，本文方法有效地保持了尾部类样本特征表示的可分性，因此本文方法在长尾分布下也能保持着良好的分类性能。

4.4　消融实验

为了说明本文方法各个模块的有效性，进行了消融实验。消融实验在 AWA2-LTS 数据集上进行，其特征提取网络都采用 ResNet10，且保持实验参数设置一致，消融的设置和结果分别如表 3、表 4 所列。

表 3　消融实验设置

Table 3　Ablation experiment setup

方法	因果领域泛化	因果效应分类器	Balanced Softmax 和 logit 调整损失
a	—	—	—
b	√	—	—
c	√	√	—
ours	√	√	√

表 4　消融实验结果

Table 4　Results of ablation experiment

方法	Avg	
	AccU	Acc
a	28.0	33.4
b	38.1	43.3
c	43.2	46.6
ours	**43.8**	**47.7**

从结果上看，在加上基于因果关系的领域泛化模块后以及用于长尾学习的因果效应分类器模块后，各个指标在 5 个域上的平均值都逐步上升，因此可以证明从因果关系出发的因果推断方法在领域泛化和长尾学习中的有效性。同时，在加入 Balanced Softmax 和 logit 调整模块后，模型性能整体上也有所提升，这说明了 Balanced Softmax 和 logit 调整对长尾学习过程的协同促进作用。

结束语　领域泛化和长尾学习相结合的复杂场景是真实场景数据集经常会遇到的情况，而要在这种场景下保持深度学习模型良好的性能，对模型的领域泛化能力和长尾学习能力都有着较高的要求。为此，本文从因果关系出发，构建了一个基于因果关系的领域泛化方法来学习一个跨域不变的因果特征表示，并在此基础上运用因果推断原理从 SGD 优化器中动量的因果关系出发构建了一个因果效应分类器，来消除动量带来的在长尾学习中向头部类偏移的缺点而保留其对学习过程优化的优点，同时通过 Balanced Softmax 和 logit 调整相结合的方法进一步对传统交叉熵损失函数进行改进来抑制长尾分布带来的影响。实验结果表明，基于因果关系的领域泛化长尾学习方法在领域泛化和长尾分布相结合的这个复杂场景下表现出了有竞争力的结果，并且有着良好的可解释性。虽然本文所提的方法在 LT-DS 问题上取得了一些进步，但目前所取得结果还是偏低，难以满足真实场景下的精度需求，因此对于 LT-DS 问题还需要进行更多的研究和探索。

参 考 文 献

[1]　XU H, WANG Y, WU Z, et al. Embedding-based complex feature value coupling learning for detecting outliers in non-iid categorical data[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 5541-5548.

[2]　ZHANG Y, KANG B, HOOI B, et al. Deep long-tailed learning: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9): 10795-10816.

[3] YAO L, CHU Z, LI S, et al. A survey on causal inference[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2021, 15(5): 1-46.
[4] MEINSHAUSEN N. Causality from a distributional robustness point of view[C]// 2018 IEEE DataScience Workshop (DSW). IEEE, 2018: 6-10.
[5] REN J, YU C, MA X, et al. Balanced meta-softmax for long-tailed visual recognition[J]. Advances in neural information processing systems, 2020, 33: 4175-4186.
[6] MENON A K, JAYASUMANA S, RAWAT A S, et al. Long-tail learning via logit adjustment[C]// International Conference on Learning Representations. 2020.
[7] GU X, GUO Y, LI Z, et al. Tackling long-tailed category distribution under domain shifts[C]// European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 727-743.
[8] WANG J, LAN C, LIU C, et al. Generalizing to unseen domains: A survey on domain generalization[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(8): 8052-8072.
[9] ZHOU K, YANG Y, QIAO Y, et al. Mixstyle neural networks for domain generalization and adaptation[J]. International Journal of Computer Vision, 2024, 132(3): 822-836.
[10] MANCINI M, AKATA Z, RICCI E, et al. Towards recognizing unseen categories in unseen domains[C]// European Conference on Computer Vision. Cham: Springer International Publishing, 2020: 466-483.
[11] LI D, ZHANG J, YANG Y, et al. Episodic training for domain generalization[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 1446-1455.
[12] SHU Y, CAO Z, WANG C, et al. Open domain generalization with domain-augmented meta-learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 9624-9633.
[13] KANG B, XIE S, ROHRBACH M, et al. Decoupling Representation and Classifier for Long-Tailed Recognition[C]// International Conference on Learning Representations. 2019.
[14] CHOU H P, CHANG S C, PAN J Y, et al. Remix: rebalanced mixup[C]// Computer Vision-ECCV 2020 Workshops: Glasgow, UK, Part VI 16. Springer International Publishing, 2020: 95-110.
[15] TAN J, WANG C, LI B, et al. Equalization loss for long-tailed object recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11662-11671.
[16] PEARL J. Direct and indirect effects [M]// Probabilistic and causal inference: the works of Judea Pearl. 2022: 373-392.
[17] LV F, LIANG J, LI S, et al. Causality inspired representation learning for domain generalization [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 8046-8056.
[18] TANG K, HUANG J, ZHANG H. Long-tailed classification by keeping the good and removing the bad momentum causal effect [J]. Advances in Neural Information Processing Systems, 2020, 33: 1513-1524.
[19] ZHANG X, CUI P, XU R, et al. Deep stable learning for out-of-distribution generalization[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 5372-5382.
[20] XU Q, ZHANG R, ZHANG Y, et al. A fourier-based framework for domain generalization[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 14383-14392.
[21] KUANG K, XIONG R, CUI P, et al. Stable Prediction with Model Misspecification and Agnostic Distribution Shift[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 4485-4492.
[22] GRETTON A, FUKUMIZU K, TEO C, et al. A kernel statistical test of independence[C]// Proceedings of the 20th International Conference on Neural Information Processing Systems. 2007: 585-592.
[23] STROBL E V, ZHANG K, VISWESWARAN S. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery[J]. Journal of Causal Inference, 2019, 7(1): 20180017.
[24] VANDERWEELE T J. A three-way decomposition of a total effect into direct, indirect, and interactive effects[J]. Epidemiology (Cambridge, Mass.), 2013, 24(2): 224.
[25] PEARL J, GLYMOUR M, JEWELL N P. Causal inference in statistics: A primer[M]. John Wiley & Sons, 2016.
[26] LECUN Y, CHOPRA S, HADSELL R, et al. A tutorial on energy-based learning[C]// Predicting Structured Data. 2006.
[27] PEARL J, MACKENZIE D. The book of why: the new science of cause and effect [M]// Basic books, 2018.

LYU Jiahao, born in 1996, master. His research interests include image classification and domain generalization.

LIU Jinfeng, born in 1971, Ph.D, professor, master supervisor. His main research interests include image processing and heterogeneous computing.