https://crad.ict.ac.cn
62(11): 2697−2709, 2025

基于元优化特征解耦的多模态跨域情感分析算法

贾熹滨[1,2] 李 宸[1] 王 珞[1] 张沐晨[1] 刘潇健[1] 张旸旸[3] 温家凯[4]
[1](北京工业大学计算机学院 北京 100124)
[2](北京人工智能研究院(北京工业大学) 北京 100124)
[3](中国电子技术标准化研究院 北京 100007)
[4](广西达译科技有限公司 南宁 530006)
(jiaxibin@bjut.edu.cn)

A Multimodal Cross-Domain Sentiment Analysis Algorithm Based on Feature Disentanglement Meta-Optimization

Jia Xibin[1,2], Li Chen[1], Wang Luo[1], Zhang Muchen[1], Liu Xiaojian[1], Zhang Yangyang[3], and Wen Jiakai[4]
[1](*College of Computer*, *Beijing University of Technology*, *Beijing* 100124)
[2](*Beijing Institute of Artificial Intelligence* (*Beijing University of Technology*), *Beijing* 100124)
[3](*China Electronics Standardization Institute*, *Beijing* 100007)
[4](*Guangxi Daring Technology Co.*, *Ltd.*, *Nanning* 530006)

Abstract Multimodal sentiment analysis aims to utilize the multimodal customer comments and other data to identify users' sentimental tendencies. To realize cross-domain application with the domain bias, commonly used solutions are unsupervised domain adaptation methods. Nevertheless, this type of solutions focuses on the extraction of domain-invariant features, and it neglects the significance of domain-specific features at the target domain. Thus, a meta-optimization based domain-invariant and domain-specific feature disentanglement network is proposed. First, by embedding adapters into the pre-trained large model with fine-tuning fitting, the image-text fused sentiment feature encoder is accordingly constructed. Then, a feature disentanglement module is constructed on the basis of the factorization operation, which utilizes domain adversary and domain classification, together with collaborative independence constraints, respectively, to achieve knowledge-transferable domain-invariant feature embedding while extracting the domain-specific features to enhance the performance of sentiment classification at the target domain. To ensure the consistency of the overall optimization tendency for feature disentanglement and sentiment classification, a meta-learning-based meta-optimization training strategy is put forward to synergistically optimize the sentiment analysis network. Comparative experiments on bidirectional sentiment transfer tasks constructed by MVSA and Yelp datasets demonstrate that compared with other advanced image-text sentiment transfer algorithms, the proposed algorithm achieves superior performance on bidirectional sentiment transfer tasks in terms of three consensus metrics: Precision, Recall and *F*1 score.

Key words multimodal sentiment analysis; unsupervised domain adaptation; cross-domain sentiment classification; feature disentanglement; meta-optimization

摘 要 多模态情感分析旨在利用多模态点评等数据识别用户情感倾向.为实现存在域偏移的跨域应用,

收稿日期: 2024-07-20; **修回日期**: 2025-01-15
基金项目: 国家自然科学基金面上项目(62476015); 北京工业大学研究生教育教学优秀成果培育项目(GER202316)
This work was supported by the General Program of the National Natural Science Foundation of China (62476015) and the Cultivation Program for Excellent Achievements in Graduate Education and Teaching at Beijing University of Technology (GER202316).

常用无监督领域自适应方法. 然而, 该类方法着重于领域不变特征提取, 忽略了目标领域特定特征的重要作用. 为此, 提出基于元优化的领域不变及领域特定特征解耦网络. 首先, 通过嵌入情感适配器对预训练大模型微调, 建立图文融合情感特征编码器. 进而, 构建基于因子分解的特征解耦模块, 分别利用领域对抗及领域分类、协同独立性约束, 实现知识可传递的领域不变特征编码的同时, 提取领域特定特征以增强目标域情感分类性能. 为保证特征解耦与情感分类的总体优化方向一致性, 提出基于元学习的元优化训练策略, 实现情感分析网络的协同优化. 基于 MVSA 和 Yelp 数据集构建的双向情感迁移任务的对比实验表明, 较之其他先进的图文情感迁移算法, 所提算法于双向情感迁移任务的精确率、召回率和 $F1$ 值 3 项评价指标均取得了优异的性能.

关键词 多模态情感分析; 无监督领域自适应; 跨领域情感分类; 特征解耦; 元优化

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202440624 **CSTR:** 32373.14.issn1000-1239.202440624

情感分析[1]是一种语境挖掘技术, 用来识别人们对某种事物的情感倾向, 如对用户点评态度分类, 为商家决策、商品推荐等提供参考等. 随着富媒体形式的广泛存在, 加之图文多模态大模型的快速发展, 有效利用图文等多模态数据实现多模态情感分析有助于克服单模态歧义性, 成为情感分析的主流. 然而, 大量有标签图文数据对的获取费时费力. 此外, 由于不同应用领域的情感态度表达方式亦存在差异性, 在实际应用中存在域偏移现象, 阻碍了源域训练获得的多模态情感分析算法在目标领域的应用性能. 为提高不同场景下的应用性能和降低目标域标注需求, 跨领域情感分析[2]被提出, 其目的是研究如何利用来自具有良好标注的源域可迁移知识来指导无标注或少标注目标域的高性能情感分类模型优化.

其中, 无监督领域自适应[3]因其无目标域数据标注要求, 为跨领域情感分析模型学习带来了更为高效的解决思路. 目前, 基于领域对抗[4-5]和基于分布度量[6-7]的 2 种主流无监督领域自适应方法被广泛应用于跨领域情感分析任务中, 它们的共同原理是尽可能去除领域特定特征以充分提取领域不变的共有特征, 降低领域分布的差异, 实现源域标签监督信息指导下目标域情感分析算法的优化学习, 获得更为精准的目标域情感极性分类. 然而, 情感表达与情感态度判别的界定与特定领域密切相关, 忽略目标域类别可判别的领域特定特征不利于网络更全面理解和表达本领域情感信息[8-9], 因此, 仅依赖领域不变的共有特征不足以实现从源域到目标域的最佳迁移, 尤其是当目标域样本多样性较大时此问题更为明显.

为了解决上述问题, 本文面向图文多模态跨领域情感分析任务, 提出了一种基于元优化的领域不变及领域特定特征解耦网络(meta-optimization-based domain-invariant and domain-specific feature disentanglement network, MDISN). 从基于预训练模型的情感适配调优、基于解耦的领域不变特征和领域特定特征编码器构建以及基于元优化策略的模型训练 3 个方面进行了模型改进和优化. 主要工作和贡献包括 4 个方面:

1)为实现图文多模态情感表示的适配, 建立更为轻量化的多模态特征提取网络以减少所需的训练数据量; 利用适配调优机制, 即在基于 Transformer 预训练模型的特征编码器中引入情感适配器, 在情感数据集上进行训练优化, 提取适用于下游分类的有效情感特征.

2)利用因子解耦构建特征解耦网络, 将融合情感特征解耦为领域不变特征和领域特定特征; 利用领域不变特征实现源域到目标域的共有情感分类知识的迁移, 同时利用领域特定特征实现专有情感分类知识表达, 通过领域不变特征及领域特定特征的协同学习增强目标域的情感分类性能.

3)在模型学习中, 为了缓解解耦目标和分类目标的优化方向不一致问题, 提出利用基于元学习的元优化训练策略, 提升图文情感分类模型的域适应迁移学习总体性能.

4)在 MVSA 和 Yelp 这 2 个图文情感分类数据集上构建的双向迁移实验数据集进行的实验证明, 所提出的算法在图文多模态情感迁移任务上取得了更具竞争力的迁移和情感极性分类性能.

1 相关工作

随着社交媒体的飞速发展, 用户发表的内容数据也更多呈现图片与短文本并用的多模态表达形式.

利用数据样本中不同模态相互补充，可以为基于文本的语义理解和语义消歧提供额外的线索. 因此，多模态情感分析逐渐成为研究热点.

近年来，基于深度学习的多模态情感分析方法围绕情感分类目标，重点研究多模态特征融合算法. 例如，Zadeh 等人[10]提出了一种端到端的张量融合网络 TFN，以动态建模在线视频中的模态内和模态间的交互. 受图像-文本情感相关性的内在特征启发，Truong 等人[11]提出了一种视觉方面注意力网络 VistaNet，将图像作为与文本内容一致的注意力来源. Gui 等人[12]针对社交媒体中抑郁症的检测问题，提出了一种融合文本和视觉信息的多智能体强化学习方法. 近年来，随着基于 Transformer 预训练模型方法的巨大成功，越来越多的研究人员尝试将基于 Transformer 预训练模型的方法引入多模态情感分析中. Ling 等人[13]针对多模态方面级情感分析问题提出了一种特定于任务的视觉语言预训练框架 MABSA，在预训练过程中设计特定的预训练任务，以细粒度方面助力图文的跨模态对齐.

传统的情感分析方法只讨论了模型于单一领域的性能，但情感表达用于的领域差异性，导致情感分析模型跨域应用性能受限. 领域自适应，特别是无监督领域自适应方法的提出使跨领域情感分类的模型迁移成为可能. 无监督领域自适应是一种源域数据有标签、目标域数据无标签以及源任务和目标任务相同，但源域和目标域数据分布不同的迁移学习方法. 主流的无监督领域自适应方法目前主要分为两大类，即基于领域对抗的方法和基于分布度量的方法.

Ganin 等人[14]构建了领域对抗性神经网络 DANN，通过领域判别器混淆领域来源以学习尽可能多的领域不变特征. Zhu 等人[7]提出了一种基于分布度量的领域自适应方法，该方法设计了一个局部最大平均差异 LMMD 度量，通过对齐不同领域间领域特定层的相关子域分布来减小领域差异. 结合以上 2 种方法的优势，Jia 等人[15]从全局和局部 2 个方面对齐了领域分布，提出了一种领域对抗类别对齐网络 DACAN.

广泛的领域自适应研究在计算机视觉和自然语言处理方面取得了巨大的成功，因而被逐渐引入到多模态情感分析领域中. 黄学坚等人[16]提出了一种基于样本内外多模态协同的表示方法，分别学习模态特有与共享特征并实现协同优化，设计了基于注意力机制和门控神经网络的自适应多模态融合方法，在情感数据集上的实验结果优于基线方法. Qi 等人[17]提出了多模态领域自适应神经网络 MDANN，通过构建混合域约束来提高模型在跨领域情感识别任务中的分类性能. Ma 等人[18]引入了多模态对抗网络 MMAN 来解决基于图像-文本数据对中存在的领域自适应问题，所提出的 MMAN 在多模态社会事件识别任务中表现最佳.

综上，多模态情感自适应为解决跨域的高性能情感识别提供了解决路径，但不同模态间的差异使多模态情感跨领域知识迁移变得更为复杂. 近年来，多模态大模型预训练不断完善，能够为下游任务提供大量图文对齐先验知识. 因此，本文利用预训练大模型作为迁移网络的骨干网络，并在其中嵌入情感适配器以实现有效的下游情感任务鲁棒的多模态融合特征提取. 以此模型为基础，本文探讨有效的域适应技术，实现多模态融合特征的跨域迁移学习，进一步挖掘前述工作中忽略的领域特有特征对于情感分类的重要作用，建立同时利用领域不变与领域特有特征的多模态情感分析算法.

2 本文方法

针对跨领域多模态情感分类问题，在将预训练大模型通用知识与下游情感知识相结合，提取图文融合特征的基础上，本文提出显式地解耦领域不变特征和领域特定特征，同时引入元学习的思想协调解耦任务和分类任务的优化一致性，实现源域标签监督下的目标域的多模态情感分类模型的领域适应迁移学习. 具体网络设计及优化训练方法阐述如下.

2.1 总体框架

本文所提的 MDISN 框架如图 1 所示. 总体网络包括特征提取模块、特征解耦模块和分类模块.

特征提取模块以基于 Transformer 块的图文大模型预训练特征编码器为基础，在每层 Transformer 块添加情感适配器. 也就是说，利用情感数据微调的适配层构建图文融合的多模态情感特征编码器. 特征解耦模块包括 2 个并行分支，其中一个分支由领域不变特征编码器和领域判别器构成，另一个分支由领域特定特征编码器和领域分类器构成，实现图文融合特征分解，分别输出解耦后的领域不变特征和领域特定特征. 分类模块为一个情感分类头，利用组合后的图文多模态解耦特征进行分类，实现情感极性预测.

在网络训练过程中，训练数据集由有标注源域数据与无标注目标域数据构成，以所构建网络为基础，采用领域不变特征编码分支的领域对抗约束、领

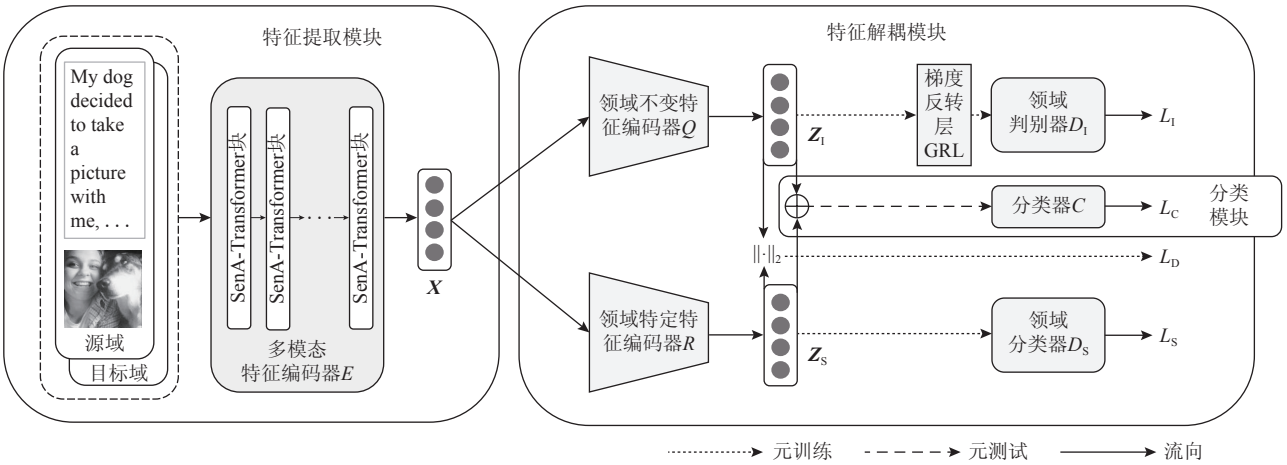


Fig. 1 Meta-optimization-based framework for domain-invariant and domain-specific feature disentanglement networks

图 1 基于元优化的领域不变及领域特定特征解耦网络框架

域特定编码分支的领域分类约束以及 2 分支解耦得到的特征间的特征独立性约束和源域分类标签的一致性监督约束, 实现目标域多模态情感分析网络的迁移学习. 此外, 为保证特征解耦任务和情感分类任务的网络训练优化整体性能, 制定元优化任务无关训练策略, 支持多模态情感分析算法的优化学习.

2.2 基于情感适配 Transformer 块的特征提取模块

如图 1 框架左侧所示, 此模块用于提取多模态融合特征 $\boldsymbol{X}$. 为克服情感图文对多模态标注数据不足下, 应对多模态下领域适应与图文对齐混叠带来的复杂度增加的挑战, 充分利用基于 Transformer 的视觉-语言预训练大模型[19] 对多模态特征的提取能力与对齐能力, 本文以公认的融合前对齐即 ALBEF[20] 模型作为特征提取模块的骨干模型. 考虑到预训练大模型的特征表示通识性强, 但缺少对特定领域的鉴别特征建模, 经典的全量网络调优需要大规模标注数据, 现实应用场景下难以满足. 因此, 借鉴 Adaptformer[21] 的适配调优方法, 本文在原始 Transformer 块中嵌入一种自适应、可拓展的适配器, 建立 SenA-Transformer 块. 本文将所采用的骨干网络 ALBEF[20] 模型中的所有原始 Transformer 块替换为 SenA-Transformer 块, 构建基于情感适配 Transformer 块的多模态特征提取模块.

1) SenA-Transformer 块结构. 如图 2 所示, 左侧表示原始的 Transformer 块结构, 右侧表示本文所引入的轻量级情感适配器. 为每个原始 Transformer 块嵌入情感适配器, 对 SenA-Transformer 块的情感适配优化时, 冻结预训练 Transformer 块的参数, 仅对适配器利用情感数据调优, 建立情感适配 Transformer 块 SenA-Transformer. 该方法有利于在仅添加有限可学习参数的情况下, 结合预训练大模型通用知识和任务相关情感知识, 提取情感图文多模态融合特征.

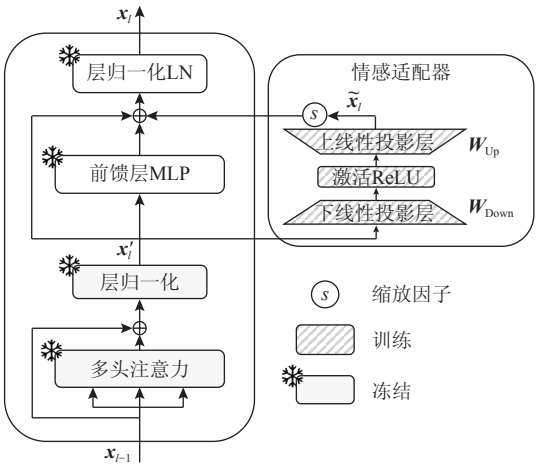


Fig. 2 SenA-Transformer block

图 2 SenA-Transformer 块

多模态特征编码器由多个 SenA-Transformer 块组成, 每层适配后 Transformer 对特征进行变换处理, 具体而言, 图 2 所示为第 l 层 SenA-Transformer 块的计算过程. 其输入为第 $l-1$ 层 SenA-Transformer 块的输出 $\boldsymbol{x}_{l-1}$. 在经由原始 Transformer 块的多头注意力(multi-head attention)与层归一化(layer normalization)处理后, 得到中间特征 $\boldsymbol{x}'_l$. 此时, 将 $\boldsymbol{x}'_l$ 输入情感适配器, 得到情感分支输出 $\tilde{\boldsymbol{x}}_l$:

$$\tilde{\boldsymbol{x}}_l = ReLU\left(\boldsymbol{x}'_l \cdot \boldsymbol{W}_{\mathrm{Down}}\right) \cdot \boldsymbol{W}_{\mathrm{Up}}. \tag{1}$$

情感适配器的网络设计如图 2 右侧所示. 适配器采用限制参数数量的瓶颈结构变换网络, 其中包括 1 个由参数 $\boldsymbol{W}_{\mathrm{Down}}$ 构成的下线性投影层、1 个由参数 $\boldsymbol{W}_{\mathrm{Up}}$ 构成的上线性投影层和 2 个投影层之间的 $ReLU$ 激活层. 适配器加在 Transformer 块的前馈层. 将 $\tilde{\boldsymbol{x}}_l$ 与

缩放因子s相乘，再将$\boldsymbol{x}_l'$输入前馈层MLP，将2项输出与$\boldsymbol{x}_l'$计算求和即得到第l层SenA-Transformer块的输出$\boldsymbol{x}_l$. 最终的SenA-Transformer块输出特征$\boldsymbol{x}_l$如式(2)所示.

$$\boldsymbol{x}_l = LN\left(MLP\left(\boldsymbol{x}_l'\right) + \boldsymbol{x}_l' + s \times \tilde{\boldsymbol{x}}_l\right). \tag{2}$$

2)SenA-Transformer块调优. 在网络训练阶段，对SenA-Transformer块进行训练调优，如图2所示仅对情感适配器网络参数$\boldsymbol{\theta}_G = (\boldsymbol{W}_{\text{Down}}, \boldsymbol{W}_{\text{Up}})$进行训练优化，同时冻结此块其余部分的网络参数. 也就是说，情感适配器的网络参数利用情感数据集进行优化更新，基础Transformer块的参数从预训练模型中加载权重，在训练中保持这些参数冻结. 通过对并行的情感适配器调优训练，提取适用于特定领域下游任务的多模态融合情感特征，同时利用ALBEF[20]预训练模型对图文情感数据对齐，为情感极性分类任务的领域自适应学习提供有效的图文情感特征.

2.3 领域不变及领域特定特征解耦模块

为有效利用源域的标签监督，实现无标注目标域的高性能多模态情感分析，一方面，本文利用领域对抗约束学习源域与目标域共有的领域不变特征，支持跨域知识迁移；另一方面，采用领域分类约束挖掘目标域情感极性判别所需的目标领域专有的领域特定特征. 在双分支特征解耦编码器模块构建基础上，设计多个损失约束，实现跨域情感分析算法的优化学习.

1)领域不变特征损失约束L_{I}. 将特征提取模块得到的多模态融合特征$\boldsymbol{X}$，输入领域不变特征编码器Q，通过对抗性训练框架[22]学习得到领域不变特征$\boldsymbol{Z}_{\text{I}}$. 领域判别器$D_{\text{I}}$用于最小化领域标签的预测损失，通过在$D_{\text{I}}$前设置梯度反转层GRL使领域不变特征编码器$Q$的优化目标转变为最大化领域标签预测损失，通过最小-最大化的对抗关系混淆领域来源，提取领域不变特征. 领域不变特征损失约束L_{I}的计算方法以及对应的特征编码器参数$\boldsymbol{\theta}_Q$和领域分类器参数$\boldsymbol{\theta}_{D_{\text{I}}}$的优化目标函数如式(3)所示.

$$\min_{\boldsymbol{\theta}_Q}\max_{\boldsymbol{\theta}_{D_{\text{I}}}}\left\{L_{\text{I}} := -E_{\boldsymbol{x},\boldsymbol{d}\sim X,D}\left[\boldsymbol{d}\ln D_{\text{I}}(\boldsymbol{Z}_{\text{I}})\right]\right\}, \tag{3}$$

其中，$\boldsymbol{x}$为某一融合特征，$\boldsymbol{X}$为整体融合特征，$\boldsymbol{d}$为某一特征对应的领域标签，D为整体领域标签集合.

2)领域特定特征损失约束L_{S}. 将特征提取模块得到的多模态融合特征$\boldsymbol{X}$，输入领域特定特征编码器R，得到领域特定特征$\boldsymbol{Z}_{\text{S}}$. 与领域不变特征相反，为训练优化$\boldsymbol{Z}_{\text{S}}$，采用领域分类器$D_{\text{S}}$，通过预测领域标签的一致性约束，鼓励领域分类器$D_{\text{S}}$区分领域来源. 领域特定特征损失约束$L_{\text{S}}$的计算方法以及对应的特征编码器参数$\boldsymbol{\theta}_R$和领域分类器参数$\boldsymbol{\theta}_{D_{\text{S}}}$的优化目标函数如式(4)所示.

$$\min_{\boldsymbol{\theta}_R,\boldsymbol{\theta}_{D_{\text{S}}}}\left\{L_{\text{S}} := -E_{\boldsymbol{x},\boldsymbol{d}\sim X,D}\left[\boldsymbol{d}\ln D_{\text{S}}(\boldsymbol{Z}_{\text{S}})\right]\right\}. \tag{4}$$

3)特征独立性约束L_{D}.为保证学习得到的2个解耦特征间的独立性，考虑到最小化2个随机变量之间的协方差相当于最小化它们之间的互信息[23]，因此可将2个特征$\boldsymbol{Z}_{\text{I}}$和$\boldsymbol{Z}_{\text{S}}$之间的解耦约束条件设置为计算协方差矩阵$\boldsymbol{Cov}(\boldsymbol{Z}_{\text{I}}, \boldsymbol{Z}_{\text{S}})$，并使其趋近于0. 基于此，领域不变特征编码器$Q$的参数$\boldsymbol{\theta}_Q$和领域特定特征编码器$R$的参数$\boldsymbol{\theta}_R$的优化目标函数如式(5)所示.

$$\min_{\boldsymbol{\theta}_Q,\boldsymbol{\theta}_R}\left\{L_{\text{D}} := E_{x\sim X}\left[\|\boldsymbol{Cov}(\boldsymbol{Z}_{\text{I}}, \boldsymbol{Z}_{\text{S}})\|_2\right]\right\}, \tag{5}$$

其中$\|\cdot\|_2$为L_2范数.

4)特征解耦损失. 对上述优化目标函数项加权求和，则特征解耦模块的总损失函数如式(6)所示.

$$L_{\text{dise}} = \lambda_{\text{D}}L_{\text{D}} + \lambda_{\text{I}}L_{\text{I}} + \lambda_{\text{S}}L_{\text{S}}, \tag{6}$$

其中λ_{D}，λ_{I}，λ_{S}为超参数，表示特征独立性约束、领域不变特征损失约束和领域特定特征损失约束的权重.

2.4 情感分类模块

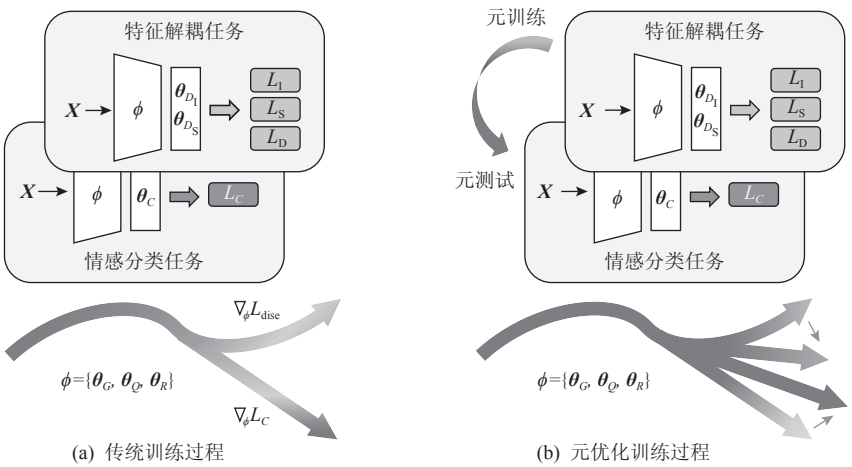
在得到领域不变特征$\boldsymbol{Z}_{\text{I}}$和领域特定特征$\boldsymbol{Z}_{\text{S}}$的基础上，令$\boldsymbol{Z}_{\text{I}} \oplus \boldsymbol{Z}_{\text{S}}$为输入，构建情感分类器. 在网络训练阶段，利用源域的情感标签与情感极性分类预测一致性为约束进行模型训练. 令$\boldsymbol{y}$为真实标签，情感分类器C的优化损失函数L_C如式(7)所示.

$$\min_{\boldsymbol{\theta}_Q,\boldsymbol{\theta}_R,\boldsymbol{\theta}_C}\left\{L_C := -E_{\boldsymbol{x},\boldsymbol{y}\sim X,Y}\left[\boldsymbol{y}\ln C(\boldsymbol{Z}_{\text{I}}, \boldsymbol{Z}_{\text{S}})\right]\right\}. \tag{7}$$

2.5 元优化训练策略

依据典型训练流程，特征解耦任务和情感分类任务的优化各自独立. 一方面，模型利用源域有标注数据和无标注目标域数据，在领域对抗损失、领域判别损失和特征独立性约束下进行特征解耦任务训练，获得图文融合数据的领域不变特征及领域特定特征的解耦表示. 另一方面，模型利用源域有标注数据进行情感分类任务的优化学习，实现情感极性分类. 然而，如图3(a)所示，在该过程中，受任务差异影响，2项任务易出现优化方向不一致问题，即共享网络参数ϕ的梯度下降方向不一致，导致特征编码器优化过程无法向最优目标点收敛，难以学习到可判别情感特征，制约域适应迁移学习下多模态情感分类性能.

受任务无关的MetaAlign[24]元学习方法启发，本文在模型优化过程中引入元学习策略，提出一种基



(a) 传统训练过程 (b) 元优化训练过程

Fig. 3 Interactive optimization training process for feature decoupling task and sentiment classification task

图 3 特征解耦任务和情感分类任务的交互优化训练过程

于元优化的训练方法, 以缓解特征解耦和情感分类任务的优化不一致问题. 具体地, 如图 3(b)所示, 本文将特征解耦任务和情感分类任务视为元学习方案中的 2 个任务, 利用元训练确保域适应策略下领域不变特征提取性能的同时, 结合元测试任务验证同一组样本的优化效果, 深入发掘领域特定特征与情感判别特征的内在联系, 鼓励模型以协调的方式对 2 项任务同时进行优化, 从而在目标域时能快速精准地提取适用于目标域情感分类性能的特征.

通过该元优化训练方法, 实现对本文提出的情感极性分类网络进行特征解耦和情感分类的联合优化训练. 在元训练过程, 模型主要对特征解耦网络进行学习, 优化领域判别器参数$\boldsymbol{\theta}_{D_{\text{I}}}$和领域分类器参数$\boldsymbol{\theta}_{D_{\text{S}}}$. 在元测试阶段, 通过使用情感分类来评估模型优化对元训练任务(特征解耦)的影响, 对情感分类器参数$\boldsymbol{\theta}_C$进行优化. 在整个元优化过程中, 特征提取模块和领域不变、领域特定特征编码器同时参与了 2 项任务, 实现模型参数$\phi=\{\boldsymbol{\theta}_G,\boldsymbol{\theta}_Q,\boldsymbol{\theta}_R\}$的优化. 据此, 总体元优化目标如式(8)所示.

$$\min_{\phi,\boldsymbol{\theta}_C}\max_{\boldsymbol{\theta}_{D_{\text{I}}},\boldsymbol{\theta}_{D_{\text{S}}}} L_{\text{dise}}\left(\phi,\boldsymbol{\theta}_{D_{\text{I}}},\boldsymbol{\theta}_{D_{\text{S}}}\right)+L_C\left(\phi,\boldsymbol{\theta}_C\right)-\alpha\nabla_\phi L_C\left(\phi,\boldsymbol{\theta}_C\right)\nabla_\phi L_{\text{dise}}\left(\phi,\boldsymbol{\theta}_{D_{\text{I}}},\boldsymbol{\theta}_{D_{\text{S}}}\right),\tag{8}$$

其中α表示元学习率. 式(8)的前 2 项为 2.3 节和 2.4 节提出的基于领域不变及领域特定特征解耦模块与情感分类模块的损失项, 而最后 1 项为任务无关的元优化损失项, 使得$\nabla_\phi L_C$和$\nabla_\phi L_{\text{dise}}$的内积最大化, 增强 2 项任务优化时的梯度下降方向的一致性, 保证总体目标向最优点收敛, 由此促进特征解耦和情感分类协同优化.

协同优化过程中, 作为元测试的分类任务, 可以验证元训练的特征解耦效果, 并指导模型将对解耦模块的优化调整至更有利于情感分类的方向. 进而, 模型在泛化到目标域时, 将能够根据优化过程中积累的经验, 调整特征解耦模块的参数. 因此, 在目标域进行应用测试阶段, 分别用 2 个解码器获得类可判别的领域不变特征和领域特定特征, 融合后分类, 有助于提升跨域情感分类性能.

基于 MDISN 的域适应训练和情感预测过程的具体描述如算法 1 所示.

算法 1. 网络的训练和推理过程.

1)训练阶段

输入: 源域样本S和目标域样本T, 可学习参数$\Psi=\{\phi,\boldsymbol{\theta}_{D_{\text{I}}},\boldsymbol{\theta}_{D_{\text{S}}},\boldsymbol{\theta}_C\}$, $\phi=\{\boldsymbol{\theta}_G,\boldsymbol{\theta}_Q,\boldsymbol{\theta}_R\}$, 学习率$\eta,\alpha$;

输出: 优化后的网络参数$\Psi^*=\{\phi^*,\boldsymbol{\theta}^*_{D_{\text{I}}},\boldsymbol{\theta}^*_{D_{\text{S}}},\boldsymbol{\theta}^*_C\}$, $\phi^*=\{\boldsymbol{\theta}^*_G,\boldsymbol{\theta}^*_Q,\boldsymbol{\theta}^*_R\}$.

① for $t=1\rightarrow iterations$ do

② 元训练:

③ 计算特征解耦损失L_{dise}, 如式(6);

④ 通过L_{dise}优化更新ϕ:

$$\phi^{t+1}\leftarrow\phi^t-\alpha\nabla_{\phi^t}L_{\text{dise}}(\phi^t,\boldsymbol{\theta}^t_{D_{\text{I}}},\boldsymbol{\theta}^t_{D_{\text{S}}});$$

⑤ 元测试:

⑥ 计算分类损失$L_C(\phi^{t+1},\boldsymbol{\theta}^t_C)$, 如式(7);

⑦ 元优化:

⑧ 计算网络总体损失, 如式(8);

⑨ 优化更新参数: $\Psi^{t+1}\leftarrow\Psi^t-\eta\nabla_{\Psi^t}L_{\text{total}}$;

⑩ end for

2)推理阶段

输入: 目标域样本T, 优化后的网络参数$\Psi^*=\{\phi^*,\boldsymbol{\theta}^*_C\}$, $\phi^*=\{\boldsymbol{\theta}^*_G,\boldsymbol{\theta}^*_Q,\boldsymbol{\theta}^*_R\}$;

输出：情感极性分类预测结果y'.

3 实验结果与分析

3.1 数据集

本文采用了2个常用的公共多模态情感数据集：MVSA[25]和Yelp[11]. 在MVSA数据集中，积极、中性和消极样本分别为11 899个、4 170个和1 500个. 在Yelp数据集中，积极、中性和消极样本分别为17 722个、8 861个和17 722个. 对数据集进行随机划分，80%的数据用于训练，5%的用于验证，15%的用于测试.

MVSA数据集的每幅图片或文本样本由3个志愿者打上标签. 遵循文献[26]的主流做法，当2个及以上志愿者为某图片给出相同的标签时，该标签视作该图片(或文本)的情感标签，同时筛选具有相同图片和文本情感标签的实例作为实验数据样本.

Yelp数据集的每个图文样本对包含了5个等级的情感量级评分. 本文将原始的Yelp数据集评分标签合并为(1, 2), 3, (4, 5)三组，并将它们重新分配为消极、中性和积极，以保持标签空间的一致性.

利用MVSA和Yelp数据集交替作为源域和目标域，构建了2个多模态跨领域情感分类任务，即MVSA→Yelp和Yelp→MVSA.

3.2 对比方法

为了验证算法的有效性，本文将MDISN与3类情感分类的多个主流先进方法进行了比较. 具体如下：

1)单模态深度分类对比方法. 视觉模态情感分析模型中，VisualSent[27]是在预训练的VGG-16网络[28]基础上建立的逻辑回归模型；ViT[29]是一种基于Transformer架构的经典视觉处理模型. 文本模态情感分析模型中，TextualSent[27]是在带有LSTM单元的BiRNN基础上建立的逻辑回归模型；BERT[30]使用双向的Transformer编码器捕捉文本的上下文信息.

2)多模态融合分类对比方法. EarlySent[27]是一种采用经典早期特征级融合策略方法，整合了VGG-16和BiRNN分别提取的图像和文本特征，并将联合表征输入逻辑回归模型进行情感分类. LateSent[27]采用经典晚期决策级融合策略方法，综合并平均VisualSent[27]和TextualSent[27]对图像和文本的预测结果，生成最终预测的情感标签. TFN[10]采用了张量融合层，将图片和文本特征结合起来进行情感分类. VistaNet[11]应用图像方面注意力来模拟编码后的图像和文本特征之间的相互作用. ALBEF[20], BLIP[31]为图文预训练大模型. ITIN[32]利用图文交互网络探究情感图像区域和文本关联关系，利用自适应跨模态门控实现多模态融合表示，而MDSE[33]采用模态无关对比学习方法建立统一的多模态融合表示，是具有先进性能的多模态情感表示方法.

3)多模态领域自适应对比方法. MMAN[18]采用堆叠注意力模块获取多模态表征，并通过对抗训练减少领域偏移. MDANN[17]由多模态注意力、融合模块和混合域约束组成，通过对抗学习来提取领域不变特征. DACAN[15]构建领域对抗模块和类别对齐模块，从全局和局部2方面对齐领域分布. 为了适应多模态图文分类任务，本文中使用ALBEF[20]作为DACAN[15]的骨干模型. DiSRAN[27]在采用模态交叉注意力计算图文联合表示的基础上，利用解耦去除领域风格，并通过对抗训练获得领域不变特征.

3.3 实施细节

实验中本文提出的基于元优化的领域不变及领域特定特征解耦网络MDISN采用增加情感适配的预训练大模型ALBEF[20]作为特征提取模块的骨干模型.

训练初始阶段，使用Xavier对模型中的权重进行初始化. 训练期间采用AdamW优化器，元学习率α设置为2×10^{-3}，是学习率η的10倍. 损失函数中的超参数设置情况为$\lambda_{\mathrm{D}}=10.0$, $\lambda_{\mathrm{I}}=1.0$, $\lambda_{\mathrm{S}}=1.0$.

所提出的框架使用PyTorch实现，并在NVIDIA TITAN RTX显卡上进行训练和测试.

3.4 评价指标

在实验中，使用精确率、召回率和$F1$值来评估模型性能，并为这3项指标计算宏观平均值. 考虑到数据集存在标签空间不平衡的影响，未利用准确率作为性能评估指标.

精确率反映了预测为正例的样本中预测正确的样本的比例. 精确率的计算如式(9)所示.

$$Pre=TP/(TP+FP), \tag{9}$$

其中TP表示真阳性，FP表示假阳性.

召回率反映了正例样本中预测正确的样本数所占的比例. 召回率的计算如式(10)所示.

$$Rec=TP/(TP+FN), \tag{10}$$

其中，FN表示假阳性.

$F1$值是精确率和召回率的调和平均数，其取值范围为[0, 1]. $F1$值越接近于1，分类模型的预测性能越好. $F1$值的计算如式(11)所示.

$$F1=(2\times Pre\times Rec)/(Pre+Rec). \tag{11}$$

3.5 对比实验及结果分析

表 1 给出了本文所提出方法 MDISN 和对比方法在无监督跨领域情感分类任务中的实验结果. 实验中, 为了进一步验证 MDISN 的有效性, 将多模态特征解码器替换为预训练大模型 BLIP[31]、先进图文融合模型 MDSE[33] 和 ITIN[32] 进行了改进, 记为 MDISN-A, MDISN-B, MDISN-C.

Table 1 Comparative Experiments for the Multimodal Cross-Domain Sentiment Classification Tasks

表 1 在多模态跨域情感分类任务中的对比实验 %

方法	MVSA→Yelp			Yelp→MVSA		
	Pre	*Rec*	*F*1	*Pre*	*Rec*	*F*1
VisualSent[27]	22.1	33.2	25.8	21.3	31.1	22.2
TextualSent[27]	27.8	41.5	32.4	22.1	34.2	24.8
ViT[29]	35.4	34.7	28.4	22.6	33.3	26.9
BERT[30]	38.2	40.3	32.7	28.2	37.2	27.0
EearlySent[27]	13.2	33.3	16.8	19.1	34.1	18.9
LateSent[27]	26.2	36.1	27.6	20.3	33.1	24.5
TFN[10,27]	33.0	33.1	31.1	32.2	31.6	16.4
VistaNet[11,27]	40.3	41.3	33.1	40.4	40.3	35.1
ALBEF[20]	40.8	42.1	35.0	41.1	41.7	35.4
BLIP[31]	57.9	41.2	25.4	55.1	43.4	45.2
ITIN[32]	47.5	39.6	31.3			
MDSE[33]	**56.6**	40.5	26.2	**61.8**	45.7	48.7
MMAN[18,27]	42.9	45.3	40.1	43.1	41.4	36.0
MDANN[17,27]	46.9	46.3	43.9	40.1	40.3	36.9
DACAN[15]	51.7	47.7	44.6	43.7	44.5	39.6
DiSRAN[27]	53.8	51.0	**49.2**	46.7	45.6	42.4
MDISN (本文)	<u>54.0</u>	<u>50.4</u>	<u>47.4</u>	46.7	44.9	41.7
MDISN-A (本文)	51.5	42.0	30.2	<u>56.2</u>	<u>50.6</u>	<u>50.5</u>
MDISN-B (本文)	41.4	40.0	26.0	55.1	**65.8**	**58.4**
MDISN-C (本文)	51.7	**53.1**	45.8			

注: 黑体数值表示最优值; 下划线数值表示次优值.

对于多模态领域自适应方法, 即本文的 MDISN 及改进的 MDISN, DACAN[15] 均遵循无监督领域自适应的惯例, 使用有标签源域数据和无标签目标域数据进行训练, 目标域训练集和测试集无交集. 对于其他方法, 包括单模态方法, 即 ViT[29], BERT[30]; 多模态融合方法, 即 ALBEF[20], BLIP[31], MDSE[33], 在源域应用标准的有监督训练, 然后将学习到的模型直接用到目标域进行测试. ITIN[32] 为论文提供的在 MVSA 数据集训练的开源测试代码, 直接用于目标域进行测试. 多模态领域自适应方法 DiSRAN[27] 及对比方法 VisualSent[27], TextualSent[27], EearlySent[27], LateSent[27], TFN[10], VistaNet[11], MDANN[17], MMAN[18] 的结果来自文献 [27].

从表 1 可以观察到, MDISN 模型及其改进模型在 MVSA→Yelp 和 Yelp→MVSA 这 2 个跨域多模态情感分析任务上均取得了较对比方法更有竞争性的指标值, 证明了 MDISN 在多模态跨领域情感分类任务上的优越性能. 对表 1 实验结果的详细讨论与分析如下.

对于单模态深度分类方法, 由于 Transformer 编码器可以通过自注意力机制捕获上下文依赖关系, 有利于图像和文本的全局语义理解, 因而 ViT[29] 模型和 BERT[30] 模型的性能优于 VisualSent[27] 模型和 TextualSent[27] 模型. 对于多模态融合分类方法, 可以发现 EearlySent[27] 方法和 LateSent[27] 方法在目标域上的精确率甚至不如单模态方法. 这一观察结果表明仅将多模态特征简单拼接会由于模态间存在不一致性与领域差异混叠, 导致有噪迁移并加剧情感极性分类性能的退化. 而 TFN[10], VistaNet[11], ALBEF [20], BLIP[31], ITIN[32], MDSE[33] 模型通过跨模态对齐融合多模态特征, 有利于文本模态与图像模态间的互补信息的利用, 其中图文大模型 ALBEF [20] 和 BLIP[31] 在多个公共数据集进行预训练, 具备了多模态对齐融合先验知识; ITIN[32] 和 MDSE[33] 采用了克服模态异质性影响等先进学习方法, 在单领域上具有相对更佳的多模态融合性能, 在直接跨域应用仍获得了非常优秀的分类性能.

在多模态领域自适应方法中, 无标签的目标域数据也参与训练, 从而极大提高了自适应性能. DACAN[15] 在利用对抗学习对齐全局分布的同时, 通过类别加权方式对齐了类别相关的细粒度领域特定信息, 从而获得了比 MDANN[17] 和 MMAN[18] 更好的迁移性能. 同时, 不同于 DACAN[15] 隐式对齐领域特定信息的方式, MDISN 利用元优化训练策略显式地将情感特征解耦为领域不变特征和领域特定特征, 通过利用领域特定信息, 促进目标领域的情感分类性能, 并取得了更优的自适应性能. 从利用多模态融合先进方法及预训练大模型的直接跨域结果, 与 MDISN 采用了领域适应策略的实验结果进行对比可以发现, MDISN 在 MVSA 上训练, 跨域应用到 Yelp 上的召回率和 *F*1 值更高, 改进的 MDISN-C 相较原多模态图文融合模型直接跨域 3 项指标更高, 且取得了最高召回率. 改进的 MDISN-A 和 MDISN-B 在 Yelp 上训练, 跨域应用到 MVSA 上的 3 项指标略低于最

高的次优性能. 该结果验证了相对跨域直接应用, 本文所提域适应方法有助于克服领域偏差的影响. 进一步分析本文提出的 3 种模型, MDISN 加入情感适配, 有助于融合特征适应源域多模态融合特征学习, 后续可以在 MDISN-A 和 MDISN-B 加入情感适配进一步优化训练, 提升跨域迁移的多模态情感分类整体性能.

与主流先进方法 DiSRAN[27] 相比, 融合领域特定特征与领域不变特征, 采用与有标注训练样本规模接近的 MVSA 数据集上训练, 在 Yelp 上的识别性能精确度更高; 在 Yelp 到 MVSA 的领域适应情感识别任务上, MDISN 在 Yelp 上的有标注训练样本数更少, 且由于 Yelp 数据对关联性弱、领域不变特征表征学习难的条件下, MDISN 仍取得了相同的情感分类精确度, MDISN-A 和 MDISN-B 的 3 项评价指标都远高于 DiSRAN[27]. 与 DiSRAN[27] 在 Yelp 数据集上使用的积极、中性和消极样本各 21 261 个不同, 本文仅选用约 70% 的 Yelp 子集(见 3.1 节所列), 训练集按各类标注样本数的 80% 划分, 积极、中性和消极样本分别为 14 178 个、7 089 个和 14 178 个. 与先进的多模态情感融合方法 MDSE[33] 相比, 在其上加入本文提出的域适应网络后, MDISN-B 在召回率和 $F1$ 值上取得了 10% 左右提升. 究其原因, 与 DiSRAN[27] 采用领域风格解耦及在统一框架下提升领域不变特征学习相比, 本文结合元优化策略学习和解耦思路, 更有助于目标域类判别领域特定特征和融合跨域迁移的领域不变特征, 对目标域具有更好的领域适应性. 因而, 在图文融合领域适应情感识别实验取得了具有竞争力的结果. 后续可考虑采用数据增强方法和样本均衡策略克服数据规模局限性, 在保证预测精确度下进一步提高总体分类性能.

3.6 同域对比实验及结果分析

表 2 给出了 MDISN 在同域, 即无跨域性能对比的实验结果. 实验在 MVSA 和 Yelp 这 2 个独立数据集上(无跨域)分别进行了全监督训练和测试, 训练样本和测试样本来自同一个数据集, 按照 3.1 节比例划分.

同域性能分析实验中, 对比了 MDISN(即 ALBEF[20]+情感适配为骨干)、MDISN-A(即 BLIP[31] 为骨干)与多模态融合 SOTA 方法和基准方法. 其中, 方法 EearlySent[27]、LateSent[27]、TFN[10]、VistaNet[11] 和 DiSRAN[27] 的结果来自文献 [27]. MDISN 去除了域解耦处理, 仅保留分类损失进行模型学习, 训练中预训练大模型骨干的参数冻结, 只训练其余参数.

Table 2 Sentiment Classification Performance Experiment on the Same Domain

表 2 在同域上情感分类性能实验 %

方法	MVSA 数据集			Yelp 数据集		
	Pre	*Rec*	*F*1	*Pre*	*Rec*	*F*1
EearlySent[27]	40.1	38.3	36.4	39.7	36.1	32.4
LateSent[27]	40.7	42.8	32.7	41.6	41.5	39.5
TFN[10,27]	43.3	45.1	41.7	45.6	44.8	43.9
VistaNet[11,27]	41.7	48.5	37.6	62.2	62.9	61.5
DiSRAN[27]	47.6	53.2	46.6	**69.8**	**69.9**	**69.1**
MDSE[33]	**66.8**	69.1	**67.6**	**76.7**	**77.5**	**77.0**
ALBEF[20]	65.3	69.8	65.2	56.5	59.0	54.1
BLIP[31]	65.8	58.6	61.1	47.9	35.4	34.3
MDISN (本文)	66.3	**70.1**	66.4	57.8	60.0	55.7
MDISN-A (本文)	63.8	68.5	64.4	45.7	43.1	43.9

注：黑体数值表示最优值；下划线数值表示次优值.

在同域上的实验结果表明, MDISN 在同域的多模态情感分类任务上, 能够取得具有竞争力的性能. 与基准 ALBEF[20] 相比, 在 MVSA 数据集上, ALBEF 情感适配器模型的 $F1$ 值达到了 66.4%, 相较于基础 ALBEF 模型的 65.2% 有所提升; 在 Yelp 数据集上, $F1$ 值提升至 55.7%, 相较于基准模型 ALBEF 的 54.1% 也有提升.

对比 VistaNet[11], DiSRAN[27], MDSE[33] 模型, 在 MVSA 数据集上, MDISN-A 取得了最优和次优结果, 特别是加上适配微调, 证明了利用了预训练大模型的知识在下游任务具有很好的适应性. 但在 YELP 数据集上的分类性能略低. 剖析原因, MVSA 数据集中的图文对通常围绕同一对象或场景展开, 因而图文之间的情感语义关联性较强; 而 Yelp 数据集的图文对主要是用户对餐厅的评价, 其中图文内容更偏向于主题描述, 情感语义关联性相对较弱. 因此, 与采用了先进的图文对齐融合方法相比, 仅利用预训练大模型构建的多模态融合特征编码器, 虽然未达到最优, 但其同域情感性能具有竞争力.

考虑到本文的元优化领域不变和领域特有解耦方法更多面向目标域任务, 进一步结合跨域实验, 对本文所提方法有效性进行了深入分析. 结合表 1 中单域训练直接跨域的实验结果, 可以发现在 MVSA 数据集上, 单域的 3 项指标远高于无域适应跨域 Yelp→MVSA 的结果. Yelp 数据集与 MVSA→Yelp 的结果也如此. 引入 MDISN 模型, 即通过在情感适配的 ALBEF[20] 的基础上加入领域特征解耦模块, 实验结果表明 MDISN 在跨域任务中的 $F1$ 值分别提升至 47.4%

(MVSA→Yelp)和 41.7%(Yelp→MVSA), 显著优于无域适应的跨域 ALBEF[20] 模型. 对比表 1 与表 2 中本文利用 BLIP[31] 作为图文融合解码器的实验结果, 在同域实验中整体性能更高, 但在 Yelp 数据集上其同域 *F*1 值仅为 34.3%, 反而低于跨域实验中的 45.2%. 而 MDISN-A 模型在跨域实验中取得了更高的性能, Yelp 数据集上的 *F*1 值达到 50.5%, 显著优于原始 BLIP 模型. 究其原因, Yelp 数据集中的图文对数据关联性较弱, 导致 BLIP[31] 在学习领域不变特征时表现出较低的表征能力, 这在同域情感分类任务中成为限制因素. 与之相比, MDISN-A 提取了更高质量图文对数据的目标域领域特定特征, 克服了 BLIP[31] 在领域不变特征学习中的局限性, 如 Yelp→MVSA 跨域任务中, 在源域 Yelp 图文对数据关联性较弱的情况下, 联合所提取的目标域 MVSA 的领域特定特征, 在跨域自适应的情感性能取得了最优结果.

总之, 实验表明跨域域适应学习仍是解决无标注目标域情感分类的有效方法. 通过同域与跨域性能分析实验, 验证了本文利用预训练大模型在同域任务中的有效性, 证明了 MDISN 提升了模型跨域应用的有效性.

3.7 消融实验

为验证所提出 MDISN 方法中各模块和策略的有效性, 对未采用相应模块的网络进行消融实验, 结果如表 3 所示. 其中, MDISN_SA-代表从 MDISN 中去除情感适配器; MDISN_DI-代表只学习领域特定特征及从 MDISN 中去除领域不变编码器和领域判别器; MDISN_DS-代表只学习领域不变特征及从 MDISN 中去除领域特定编码器和领域分类器; MDISN_MO-代表 MDISN 未采用元优化训练.

Table 3 Ablation Study on MDISN for Multimodal Cross-Domain Sentiment Classification Tasks

表 3 MDISN 在多模态跨域情感分类任务上的消融实验 %

方法	MVSA→Yelp			Yelp→MVSA		
	Pre	*Rec*	*F*1	*Pre*	*Rec*	*F*1
MDISN_SA-	51.4	49.1	46.7	39.7	44.7	39.3
MDISN_DI-	52.6	44.7	40.1	44.6	38.0	31.9
MDISN_DS-	50.4	47.9	44.7	40.5	40.4	36.5
MDISN_MO-	52.0	47.2	42.8	41.0	44.0	37.0
MDISN (本文)	**54.0**	**50.4**	**47.4**	**46.7**	**44.9**	**41.7**

注: 黑体数值表示最优值.

从表 3 可以看出, 与 MDISN_SA-和 MDISN_MO-相比, 加入情感适配器及任务无关元学习的 MDISN 性能更优, 证明了情感适配器和元优化训练策略在进行跨领域情感分类时发挥了不可或缺的作用, 增强了网络对多模态情感特征的表征能力, 并在特征解耦任务与情感分类任务的协同优化下, 保证了总体优化目标. 此外, 无论是在 MVSA→Yelp 迁移任务还是 Yelp→MVSA 迁移任务上, MDISN_DI-均获得了最低的 *F*1 值, 这表明领域不变特征的提取对自适应性能的提升起到了重要的作用. 而相较于 MDISN, 仅采用领域特定特征的 MDISN_DI-和领域共有特征的 MDISN_DS-的多模态情感分类性能均有不同程度的降低, 证明了同时利用领域不变特征与领域特定特征有助于提升跨领域自适应情感分类性能.

3.8 情感适配器方案确定实验

为探讨情感适配器在基于 ALBEF[20] 的特征提取模型中采用不同插入位置和不同参数时的调优性能, 确定适配器嵌入调优方案, 仍采用参数量、精确率、召回率和 *F*1 值作为性能指标, 使用完全相同的训练配置进行对比分析实验.

ALBEF[20] 的图像编码器使用 ViT-B/16 进行初始化, 文本编码器则使用 $\text{BERT}_{\text{base}}$ 模型的前 6 层进行初始化, 而多模态编码器使用 $\text{BERT}_{\text{base}}$ 模型的最后 6 层进行初始化. 针对该设计, 表 4 对比了 4 种嵌入调优方法, 其中 Fine-tune 表示对 MDISN 的特征提取模块的骨干网络(即 ALBEF[20] 图文特征提取模型)进行全量微调; SenA-T#L 表示只在 ALBEF[20] 低层, 也即 ViT 前 6 层与 BERT 的编码器部分的 Transformer 块插入情感适配器, 并冻结其余部分; 类似地, SenA-T#H 表示只在 ALBEF[20] 高层, 也即 ViT 后 6 层与多模态融合特征模块部分的 Transformer 块插入情感适配器, 并冻结其余部分; SenA-T 表示在所有 Transformer 块的所有层插入情感适配器.

在基于 ALBEF[20] 的特征提取模型对比情感适配器的不同插入位置的分类性能后, 可以发现 SenA-T

Table 4 Comparative Experiments of Four Optimization Methods for Multimodal Cross-Domain Sentiment Classification Tasks

表 4 4 种调优方法在多模态跨域情感分类任务上的实验对比

方法	参数量	MVSA→Yelp			Yelp→MVSA		
		Pre/%	*Rec*/%	*F*1/%	*Pre*/%	*Rec*/%	*F*1/%
Fine-tune	196.57×10^6	53.6	50.1	47.1	**47.5**	**45.6**	**42.1**
SenA-T#L	1.26×10^6	52.0	49.2	45.4	44.4	42.5	40.3
SenA-T#H	1.26×10^6	53.4	50.1	46.9	45.9	43.8	41.1
SenA-T	2.52×10^6	**54.0**	**50.4**	**47.4**	46.7	44.9	41.7

注: 黑体数值表示最优值, 下划线数值表示次优值.

的性能最好，SenA-T#H 次之，SenA-T#L 最差. 这一观察结果表明，特征提取器较低层表征通用特征，在数据集之间相似度差异较小的情况下，仅训练部分低层 Transformer 块对于下游任务的适应效果提升不大；而较高层表征特定特征与下游情感数据的分类直接相关，因此具有较高的训练价值. 对 ALBEF[20] 模型的所有层均插入情感适配器进行调优可以在较低的计算资源消耗的基础上，兼顾通用表征特征和类判别特征，促进域不变迁移学习和域特有特征挖掘，从而获得下游分类任务的性能提升.

从表 4 还可以发现，全量微调 Fine-tune 的调优方法在 MVSA→Yelp 迁移任务上的性能甚至不如 SenA-T. 其原因在于由于 MVSA 数据集规模相对较小，且类别不平衡问题严重，在需要训练大规模参数的条件下，模型容易拟合到具有高占比的积极类别中，从而导致总体性能不佳. 这也进一步证明了利用适配器调优的 SenA-Transformer 块可以在少量目标域数据下获得使目标任务判别性能更优的特征表示.

4 结　论

利用预训练图文大模型对图文跨模态数据的优秀表征性能，本文首先以图文预训练大模型为基础，利用情感适配器获得良好跨模态情感融合特征. 在此基础上，重点针对跨领域情感分类任务，研究基于解耦的特征提取网络，在提取领域不变特征的同时提取并利用领域特定特征. 为应对特征解耦任务与情感分类任务的优化不一致问题，提出利用任务无关元训练策略，实现特征解耦与情感分类网络的元优化域适应迁移学习.

所提出的基于元优化特征解耦的跨领域情感分析方法面向多模态图文情感数据，利用情感适配 Transformer 块改进预训练大模型，避免了同时进行跨模态对齐与域偏移去除学习带来的复杂性，在少量训练数据下实现下游情感多模态特征提取. 进一步地，将多模态融合情感特征显式解耦为有助于跨领域迁移的领域不变特征和有助于情感分类的领域特定特征，并采用协同特征解耦和情感分类的任务无关元优化策略，实现跨领域知识迁移的性能提升. 于 MVSA 和 Yelp 这 2 个图文情感数据集上构建的双向迁移任务上进行的实验表明，所提出的模型在精确率、召回率和 $F1$ 值上均获得了有竞争力的性能，证明了所提出模型的有效性. 消融实验分析进一步说明，情感表达依赖于特定领域，同时利用领域对抗所提取的领域不变特征和领域分类引导的领域特定特征使模型理解和提取更具判别性的该领域的情感分类特征，提高无标注目标域的多模态情感分析性能.

此外，所提出的元优化域适应学习方法亦可应用于其他的多模态跨域任务，特别是目标域特征空间与源域存在更大属性差异以及领域不变特征对目标域任务表征能力有限的任务场景. 为此，在下一步工作中，从提升多模态融合性能、利用数据增强等角度加强多模态特征表征能力，并在其他多模态任务开展元优化域适应方法研究.

作者贡献声明：贾熹滨指导算法研发和论文撰写；李宸提出算法、完成实验并撰写论文；王珞参与算法研讨并修订论文；张沐晨参与实验并修改论文；刘潇健提供算法优化建议并参与论文修订；张旸旸提出论文修改建议；温家凯对算法优化提供建议.

参　考　文　献

[1] Das R, Singh T D. Multimodal sentiment analysis: A survey of methods, trends, and challenges[J]. ACM Computing Surveys, 2023, 55(13s): 1−38

[2] Chan J Y L, Bea K T, Leow S M H, et al. State of the art: A review of sentiment analysis based on sequential transfer learning[J]. Artificial Intelligence Review, 2023, 56(1): 749−780

[3] Singhal P, Walambe R, Ramanna S, et al. Domain adaptation: Challenges, methods, datasets, and applications[J]. IEEE Access, 2023, 11: 6973−7020

[4] Azuma C, Ito T, Shimobaba T. Adversarial domain adaptation using contrastive learning[J]. Engineering Applications of Artificial Intelligence, 2023, 123: 106394

[5] Zhou Qianyu, Gu Qiqi, Pang Jiangmiao, et al. Self-adversarial disentangling for specific domain adaptation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(7): 8954−8968

[6] Li Jingjing, Chen Erpeng, Ding Zhengming, et al. Maximum density divergence for domain adaptation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(11): 3918−3930

[7] Zhu Yongchun, Zhuang Fuzhen, Wang Jindong, et al. Deep subdomain adaptation network for image classification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(4): 1713−1722

[8] Zhao Han, Des Combes R T, Zhang Kun, et al. On learning invariant representations for domain adaptation[C]//Proc of the 36th Int Conf on Machine Learning. New York: PMLR, 2019: 7523−7532

[9] Johansson F D, Sontag D, Ranganath R. Support and invertibility in domain-invariant representations[C]//Proc of the 22nd Int Conf on Artificial Intelligence and Statistics. New York: PMLR, 2019:

527−536

[10] Zadeh A, Chen Minghai, Poria S, et al. Tensor fusion network for multimodal sentiment analysis[C]//Proc of the 2017 Conf on Empirical Methods in Natural Language Processing. New York: PMLR, 2017: 1103−1114

[11] Truong Q T, Lauw H W. VistaNet: Visual aspect attention network for multimodal sentiment analysis[C]//Proc of the 33rd Association for the Advancement of Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 305−312

[12] Gui Tao, Zhu Liang, Zhang Qi, et al. Cooperative multimodal approach to depression detection in Twitter[C]//Proc of the 33rd Association for the Advancement of Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 110−117

[13] Ling Yan, Yu Jianfei, Xia Rui. Vision-language pre-training for multimodal aspect-based sentiment analysis[C]//Proc of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2022: 2149−2159

[14] Ganin Y, Ustinova E, Ajakan H, et al. Domain-adversarial training of neural networks[J]. Journal of Machine Learning Research, 2016, 17(59): 1−35

[15] Jia Xibin, Li Chen, Zeng Meng, et al. An improved unified domain adversarial category-wise alignment network for unsupervised cross-domain sentiment classification[J]. Engineering Applications of Artificial Intelligence, 2023, 126: 107108

[16] Huang Xuejian, Ma Tinghuai, Wang Gensheng. Multimodal learning method based on intra- and inter-sample cooperative representation and adaptive fusion[J]. Journal of Computer Research and Development, 2024, 61(5): 1310−1324 (in Chinese)
(黄学坚, 马廷淮, 王根生. 基于样本内外协同表示和自适应融合的多模态学习方法[J]. 计算机研究与发展, 2024, 61(5): 1310−1324)

[17] Qi Fan, Yang Xiaoshan, Xu Changsheng. A unified framework for multimodal domain adaptation[C]//Proc of the 26th ACM Int Conf on Multimedia. New York: ACM, 2018: 429−437

[18] Ma Xinhong, Zhang Tianzhu, Xu Changsheng. Deep multi-modality adversarial networks for unsupervised domain adaptation[J]. IEEE Transactions on Multimedia, 2019, 21(9): 2419−2431

[19] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proc of the 31st Int Conf on Neural Information Processing Systems. New York: ACM, 2017: 6000−6010

[20] Li Junnan, Selvaraju R, Gotmare A, et al. Align before fuse: Vision and language representation learning with momentum distillation[J]. Advances in Neural Information Processing Systems, 2021, 34: 9694−9705

[21] Chen Shoufa, Ge Chongjian, Tong Zhan, et al. Adaptformer: Adapting vision transformers for scalable visual recognition[J]. Advances in Neural Information Processing Systems, 2022, 35: 16664−16678

[22] Li Ya, Tian Xinmei, Gong Mingming, et al. Deep domain generalization via conditional invariant adversarial networks[C]//Proc of the 2018 European Conf on Computer Vision (ECCV). Berlin: Springer, 2018: 624−639

[23] Bui M H, Tran T, Tran A, et al. Exploiting domain-specific features to enhance domain generalization[J]. Advances in Neural Information Processing Systems, 2021, 34: 21189−21201

[24] Wei Guoqiang, Lan Cuiling, Zeng Wenjun, et al. MetaAlign: Coordinating domain alignment and classification for unsupervised domain adaptation[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition(CVPR). Piscataway, NJ: IEEE, 2021: 16643−16653

[25] Niu Teng, Zhu Shiai, Pang Lei, et al. Sentiment analysis on multi-view social data[C]//Proc of the 22nd Int Conf on Multimedia Modeling. Berlin: Springer, 2016: 15−27

[26] Liu Qiwei, Li Jun, Gu Beibei, et al. TSAIE: Text sentiment analysis model based on image enhancement[J]. Frontiers of Data & Computing, 2022, 4(3): 131−140 (in Chinese)
(刘琦玮, 李俊, 顾蓓蓓, 等. TSAIE: 图像增强文本的多模态情感分析模型[J]. 数据与计算发展前沿, 2022, 4(3): 131−140)

[27] Zhang Yuhao, Zhang Ying, Guo Wenya, et al. Learning disentangled representation for multimodal cross-domain sentiment analysis[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 34(10): 7956−7966

[28] Baltrušaitis T, Ahuja C, Morency L P. Multimodal machine learning: A survey and taxonomy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(2): 423−443

[29] Dosovitskiy A. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint, arXiv: 2010.11929, 2020

[30] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint, arXiv: 1810.04805, 2018

[31] Li Junnan, Li Dongxu, Xiong Caiming, et al. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//Proc of the 39th Int Conf on Machine Learning. New York: PMLR, 2022: 12888−12900

[32] Zhu Tong, Li Leida, Yang Jufeng, et al. Multimodal sentiment analysis with image-text interaction network[J]. IEEE Transactions on Multimedia, 2022, 25: 3375−3385

[33] Li Jingzhe, Wang Chengji, Luo Zhiming, et al. Modality-dependent sentiments exploring for multi-modal sentiment classification[C]// Proc of the 2024 IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2024: 7930−7934



Jia Xibin, born in 1969. Professor, PhD supervisor. Distinguished member of CCF. Her main research interests include visual computing, multimodal deep learning, affection computing, intelligent medical image, and behaviour understanding and computing.

贾熹滨, 1969 年生. 教授, 博士生导师. CCF 杰出会员. 主要研究方向为视觉计算、多模态深度学习、情感计算、智能医学影像、行为理解与计算.



Li Chen, born in 1998. Master. Her main research interests include multimodal sentiment and transfer learning.

李　宸, 1998 年生. 硕士. 主要研究方向为多模态情感、迁移学习.



Wang Luo, born in 1990. PhD, lecturer. Member of CCF. His main research interests include computer vision and deep learning.

王　珞, 1990 年生. 博士, 讲师. CCF 会员. 主要研究方向为计算机视觉、深度学习.



Zhang Muchen, born in 2002. Master candidate. Her main research interests include multimodal deep learning and computer vision.

张沐晨, 2002 年生. 硕士研究生. 主要研究方向为多模态深度学习、计算机视觉.



Liu Xiaojian, born in 1989. PhD, senior engineer. Member of CCF. His main research interests include software engineering and cybersecurity.

刘潇健, 1989 年生. 博士, 高级工程师. CCF 会员. 主要研究方向为软件工程、网络安全.



Zhang Yangyang, born in 1976. Master. Professorial senior engineer. Member of CCF. Her main research interests include software and systems engineering, and standardization of IT.

张旸旸, 1976 年生. 硕士, 教授级高工. CCF 会员. 主要研究方向为软件与系统工程、信息技术标准化.



Wen Jiakai, born in 1979. Bachelor, senior engineer. Member of CCF. His main research interests include machine translation and big data.

温家凯, 1979 年生. 学士, 高级工程师. CCF 会员. 主要研究方向为机器翻译、大数据.