计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(5): 1190−1204, 2025

基于类别对抗联合学习的跨提示自动作文评分方法

张春云[1,2] 赵洪焱[1,2] 邓纪芹[1,2] 崔超然[1,2] 董晓琳[1,2] 陈竹敏[3]
[1](山东财经大学计算机科学与技术学院 济南 250014)
[2](山东省数字媒体重点实验室(山东财经大学) 济南 250014)
[3](山东大学计算机科学与技术学院 山东青岛 266237)
(zhangchunyun1009@126.com)

Category Adversarial Joint Learning Method for Cross-Prompt Automated Essay Scoring

Zhang Chunyun[1,2], Zhao Hongyan[1,2], Deng Jiqin[1,2], Cui Chaoran[1,2], Dong Xiaolin[1,2], and Chen Zhumin[3]
[1](*School of Computer Science and Technology*, *Shandong University of Finance and Economic*, *Jinan* 250014)
[2](*Key Laboratory of Digital Media of Shandong Province* (*Shandong University of Finance and Economic*), *Jinan* 250014)
[3](*School of Computer Science and Technology*, *Shandong University*, *Qingdao*, *Shandong* 266237)

Abstract Automated essay scoring (AES) can effectively alleviate the burden on teachers when evaluating student essays and provide students with objective and timely feedback. It is a crucial application of natural language processing in the field of education. Cross-prompt AES aims to develop a transferable automated scoring model that performs well on essays from a target prompt. However, existing cross-prompt AES models primarily operate in scenarios where target prompt data are available. These models align feature distributions between source and target prompts to learn invariant feature representations for transferring to the target prompt. Unfortunately, such methods cannot be applied to scenarios where target prompt data are not available. In this paper, we propose a cross-prompt AES method based on category adversarial joint learning (CAJL). First, we jointly model AES as classification and regression tasks to achieve combined performance improvement. Second, unlike existing methods that rely on prompt-agnostic features to enhance model generalization, our approach introduces a category adversarial strategy. By aligning category level features across different prompts, we can learn invariant feature representations of different prompts and further enhance model generalization. We evaluate our proposed method on ASAP (automated student assessment prize) and ASAP++ datasets, predicting both overall essay scores and trait scores. Experimental results demonstrate that our method outperforms six classical methods in terms of the QWK (quadratic weighted Kappa) metric.

Key words automated essay scoring (AES); cross-prompt; category adversarial; joint modeling; domain generalization

摘 要 自动作文评分(automated essay scoring, AES)能够有效减轻教师的作文评阅负担并为学生提供客观、及时的反馈,是自然语言处理在教育领域的一项重要应用.跨提示AES旨在学习一个可迁移的自动评分模型,使之能够有效为目标提示的作文评分.然而,现有的跨提示AES大都是面向目标提示数据可见

收稿日期: 2024-04-18; **修回日期**: 2024-09-20
基金项目: 国家自然科学基金项目(62077033); 山东省自然科学基金项目(ZR2020KF015); 山东省泰山学者计划项目(tsqn202211199)
This work was supported by the National Natural Science Foundation of China (62077033), the Shandong Provincial Natural Science Foundation (ZR2020KF015), and the Taishan Scholar Program of Shandong Province (tsqn202211199).
通信作者: 崔超然(crcui@sdufe.edu.cn)

的场景，通过将源提示和目标提示的特征分布进行对齐，学习提示不变特征表示来学习可迁移到目标提示的评分模型，但是这类方法无法应用于目标提示数据不可见的场景. 面向目标提示数据不可见的场景，提出一种基于类别对抗联合学习的跨提示 AES 方法. 一方面，通过对分类和回归联合任务进行联合建模来学习 2 个任务的共享特征，从而实现二者性能的相互促进；另一方面，不同于现有方法采用提示无关特征来提升模型泛化性能，针对不同提示的类别分布差异引入类别对抗策略，通过对不同提示进行类别级特征对齐，学习不同提示间的细粒度不变特征表示，从而提升模型泛化性能. 将所提出方法用于自动学生评估奖（ASAP）和 ASAP++数据集，分别对作文的总体评分和属性评分进行预测. 实验结果表明，与 6 种经典方法相比，在平方卡帕（QWK）指标上取得最好的实验效果.

关键词 自动作文评分；跨提示；类别对抗；联合建模；领域泛化

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202440266 **CSTR:** 32373.14.issn1000-1239.202440266

近几年，在教育改革的推动下，语言类学科的核心地位凸显. 作文作为一种评估学生语言综合运用能力的重要手段，成为语言类考试中分数占比最高的题型. 随着教育评估体系的现代化，自动作文评分（automated essay scoring, AES）技术应运而生. AES 是指通过自动识别作文中的技术错误和理解作文陈述、连贯性及其他方面等要素来评估学生作文质量的技术. AES 可以有效减轻教师作文评分的负担，同时也可以为学生提供客观、及时的反馈，是自然语言处理（natural language processing, NLP）在教育领域的一项重要应用[1-2].

针对 AES 的研究开始于 20 世纪 60 年代[3]，AES 通常被定义为回归、分类或排序任务[1]. 早期的 AES 方法主要依赖于手工设计的特征，采用基于浅层机器学习的方法进行评分，这些特征包括语法错误、文章长度和连贯性等[4-5]. 近几年，得益于深度学习网络强大的建模复杂模式和学习重要特征的能力，卷积神经网络（convolutional neural network, CNN）[6]、循环神经网络（recurrent neural network, RNN）[7] 和 Transformer[8] 成为深度 AES 方法的骨干架构，取得了不错的评分效果.

通常而言，在语言类考试中学生需根据给定的写作提示（prompt）撰写作文，这些写作提示仅对写作的主题和要求进行约束，不对标题或体裁进行限制. 因此，根据不同写作提示撰写的文章在词汇选用、篇章结构和写作风格等方面往往存在较大的差异，即存在领域漂移. 早期的 AES 大都是基于特定提示的方法，即评分模型的训练和测试数据都来自于同一写作提示. 然而，在现实场景中，难以收集大量的给定提示（目标提示）的已评分作文，甚至无法获取该提示的已评分作文. 与此同时，存在大量属于其他不同提示（源提示）的已评分作文可以利用. 因此，近年来，跨提示 AES 成为一个热点研究课题[9-13].

跨提示 AES 根据目标提示的数据在训练阶段是否可见分为目标提示数据可见和目标提示数据不可见 2 种场景. 其中目标提示数据不可见的场景更贴合实际的作文评分环境，即在模型训练阶段无法访问到任何目标提示的数据，只能依靠从多个源提示中获取的数据来训练一个具有较好泛化能力的评分模型，使之能够对任意未见的目标提示作文进行较好的评分. 现有跨提示 AES 方法大都是面向目标提示数据可见且目标提示数据无标签的场景，这些方法尝试通过将源提示作文和目标提示作文进行分布对齐或学习二者的共享特征的方式对目标提示作文进行评分[9-10,13]. 然而，在目标提示数据不可见的场景中，由于模型无法利用目标提示数据，难以学习源提示和目标提示的共享特征. 因此，面向未见提示场景的跨提示 AES 研究比较少，尚处于探索阶段. 例如，Ridley 等人[14] 提出了一种针对目标提示数据不可见场景的跨提示 AES 方法 PAES，该方法假设基于通用语言特征训练的评分模型具有较好的泛化性能，将词性 POS（part-of-speech）特征作为输入，并将其特征表示与手工设计的提示无关特征进行拼接来进行作文表示，从而提升模型泛化性能；随后，针对实际场景中对作文属性评分的反馈需求，Ridley 等人[12] 对 PAES 框架进行扩展，提出面向未见提示的多属性自动作文评分方法 CTS，采用基于注意力机制的多任务学习方法实现对作文整体和多属性的联合评分.

虽然现有面向未见提示的跨提示 AES 方法取得了初步的进展，但是这些方法仅采用提示无关的通用特征来进行跨提示 AES，无法从根本上提升模型的泛化性能. 为解决上述问题，本文尝试从迁移学习

的角度学习不同源提示之间的不变特征表示, 从而提升模型的泛化性能. 领域对抗训练是一种被广泛应用于领域自适应中学习领域不变特征表示的方法. Ganin 等人[15]提出领域对抗网络 DANN, 通过特征提取器和领域判别器之间的对抗训练学习领域之间可迁移的不变特征. 然而, 直接将领域对抗训练应用到跨提示自动作文评分, 存在 2 个问题: 1)领域对抗训练方法大都用于分类任务, 由于分类任务对学习到的域不变特征对尺度变化具有较好的鲁棒性, 但是改变特征尺度对回归任务影响较大[16], 因此, 不能直接将领域对抗训练应用于回归任务; 2)现有的领域对抗训练专注于从全局上对源提示和目标提示的分布进行对齐, 没有考虑不同提示之间的类别分布差异. 为了直观地说明这种差异, 我们将 AES 中广泛使用的自动学生评估奖(automated student assessment prize, ASAP)数据集[17]中 8 个提示数据的类别分布①进行了可视化. 如图 1 所示, 各个不同提示的类别分布存在较大差异. 在这种情况下, 若直接在提示级别上对源提示数据进行分布对齐, 可能会导致不同类别分布之间错位对齐的问题, 进而降低跨提示 AES 的性能.

为了解决上述问题, 本文提出了一种基于类别对抗联合学习的跨提示 AES 方法. 该方法将 AES 任务建模为结合分类和回归的联合学习框架, 一方面通过学习分类和回归 2 个任务的共享特征, 实现二者性能的联合提升; 另一方面基于分类任务目标引入对抗训练方法, 学习不同提示之间的不变特征表示. 此外, 针对不同目标提示之间存在较大的类别差异的问题, 本文采用类别对抗网络对源提示和目标提示进行类别级的特征对齐, 从而避免在提示级别上进行特征对齐可能产生的错位对齐问题. 本文将所提方法用于 ASAP 和 ASAP++数据集[17], 分别对作文的整体评分和多属性的评分进行预测, 实验结果表明, 经与现有的 6 个经典方法相比, 本文方法在平方卡帕(quadratic weighted Kappa, QWK)指标上取得最好的实验效果. 本文的创新点有 3 点:

1)提出一种新的面向未见目标提示场景的跨提示 AES 方法, 通过将 AES 联合建模为分类任务和回归任务, 学习 2 个任务之间的共享特征, 实现二者性能的联合提升.

2)与现有方法依赖于提示无关的通用特征来增强模型泛化性能不同, 本文基于作文分类任务引入类别对抗训练策略对源提示和目标提示特征进行类别级特征对齐, 从而学习不同源提示之间细粒度的不变特征表示, 进而实现模型泛化性能的提升.

3)将本文方法用于广泛使用的 AES 数据集 ASAP 和 ASAP++, 对作文进行整体评分和多属性评分, 实验结果表明, 经与现有的 6 个经典方法相比, 本文方法在 QWK 指标上取得最好的实验效果.

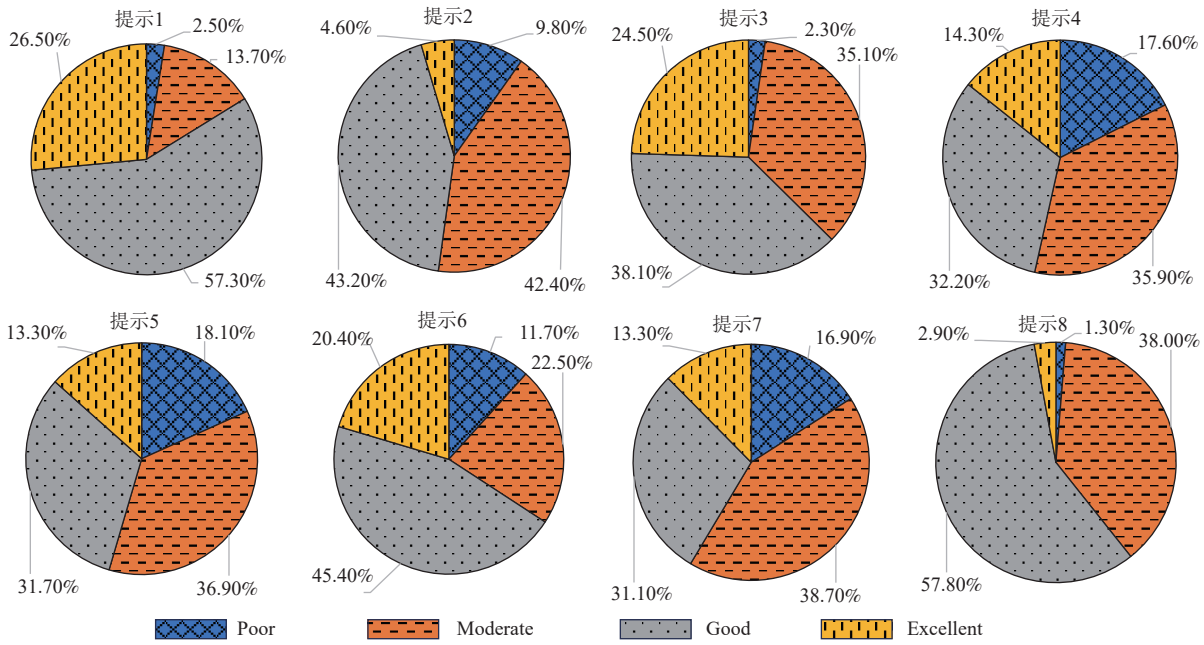


Fig. 1 Distribution of essay categories for 8 prompts in ASAP dataset

图 1 ASAP 数据集 8 个提示的作文类别分布

① 我们将 ASAP 中每篇作文分数根据其所属区间分别划分为优、良、中、差 4 个等级, 具体的类别标注方式见 3.1 节.

1 相关工作

本节首先对作文自动评分的相关工作进行概述，然后介绍与本文密切相关的领域泛化的相关工作.

1.1 AES

早期的 AES 主要采用手工定义的特征并基于传统的机器学习方法来进行评分[4-5,18-20]. 虽然手工特征中蕴含丰富的专家知识，但是对手工特征的设计和抽取耗费大量的人力成本. 随着深度学习技术的发展，基于深度学习的 AES 方法被提出，并取得了不错的评分效果. 通常，基于深度学习的 AES 方法根据其训练数据和测试数据是否来自于同一提示，可分为基于特定提示的 AES 和跨提示的 AES.

1.1.1 特定提示的 AES

特定提示的 AES 旨在基于来自同一提示的训练和测试样本学习一个自动评分模型，因此特定提示的自动评分模型可以充分利用数据中的特定提示信息提升模型的评分性能.T&N 模型[21] 是最早的基于深度学习的 AES 方法，该方法采用 CNN 和长短期记忆(long short term memory, LSTM)网络分别来建模局部和长距离依赖特征. 同年，Alikaniotis 等人[22] 提出了一种增强的 CBOW 模型，通过联合建模语言模型和 AES 来学习用于评分的词嵌入表示. 王耀华等人[23] 提出了一种通过分析句子向量与整体作文向量之间的关系来衡量语义差异性的深度神经网络模型，用于作文评分预测. 随后，2018 年，Tay 等人[24] 提出了 SKIPFLOW 机制，使用 LSTM 显式捕获相关属性，并利用神经张量层进行连贯的文本表示. 而付瑞吉等人[25] 则基于 CNN-LSTM 模型建模优美句识别，并将学习的特征用于提升 AES 的性能. 2021 年，周险兵等人[26] 提出一种基于多层次语义特征的神经网络，采用 CNN 捕捉局部语义特征的同时，采用混合神经网络捕捉全局语义特征.

随着注意力机制的广泛使用[8]，许多 AES 方法采用注意力机制来提升 AES 方法的性能. 典型的方法之一是 Hi-att[27]，它将自注意力引入到经典的 T&N 框架中，从而对 AES 中不同单词的重要程度进行建模. 另一种广泛采用的范式是基于 BERT 的方法. 受益于其在各种自然语言处理任务上的卓越成就，BERT 被用来解决 AES[28-31]. Rodriguez 等人[30] 直接采用预训练的 BERT 和 XLNet 模型来预测作文分数，验证了使用预训练模型的有效性. 随后，Yang 等人[32] 通过联合优化回归和排序损失来微调 BERT 模型. 此外，Ormerod 等人[33] 通过集成预训练模型来增强 AES 的性能. 随着预训练模型参数的增加，直接对模型进行微调不太现实，因此，Sethi 等人[34] 引入了一种参数高效微调方法，通过使用适配器模块来最大限度地减少训练参数的数量，作为微调整个预训练模型的替代方法. 2024 年，于明诚等人[35] 提出基于多尺度上下文的 AES 方法，通过利用 XLNet 预训练模型来提取文本中的词嵌入和句子嵌入并将其进行融合用于评分预测.

1.1.2 跨提示 AES

尽管特定提示 AES 方法取得不错的效果，但是在实际场景中难以收集大量特定提示的已评分作文，甚至根本无法获取目标提示的作文. 与此同时，存在大量属于其他不同提示(源提示)的已评分作文可以利用. 因此，近年来，跨提示 AES 成为一个热点研究课题.

跨提示 AES 旨在学习一个可迁移的自动评分模型，该模型可以很好地对目标提示的作文进行评分. 其主要包括：目标提示可见和目标提示不可见 2 种场景.

1)目标提示可见的跨提示 AES

现有的目标提示可见的 AES 方法大都是采用 2 阶段的方法：在第 1 阶段，基于带标签的源提示数据训练一个提示无关的评分模型；在第 2 个阶段，基于提示相关的信息提升该模型的性能. 一个具有代表性的 2 阶段方法是 Jin 等人[36] 提出的 TDNN 模型. 该模型首先基于多个源提示训练一个提示无关模型，并基于该模型为目标提示作文生成伪标签. 随后，利用伪标签中评分最高和评分最低的目标提示作文，训练一个针对目标提示的评分模型. 在此基础上，Li 等人[10] 进一步提出了 SEDNN 模型，该模型在起始阶段通过采用对抗网络学习提示无关的模型，并在后续阶段加入提示信息，以此来提升跨提示作文评分模型的性能.

另外一种 2 阶段的跨提示 AES 方法主要是基于预训练模型，采用预训练和微调的方式. 预训练模型允许基于 Transformer 的模型整合从各种自然语言文本中获取的通用知识. Song 等人[11] 提出一种基于预训练模型的中文 AES 方法，采用 1 个预训练阶段和 2 个微调阶段来提升目标提示的评分性能. 类似地，Cao 等人[9] 首先通过多个源提示和目标提示构建 2 个提示无关的自监督任务来学习提示之间的共享特征，随后采用对抗训练实现对目标提示作文的评分. 此外，Chen 等人[13] 采用了一种基于对比学习的技术，

将源提示和目标提示进行特征映射来实现源提示和目标提示的特征对齐, 从而提升模型的泛化性能.

2) 目标提示不可见的跨提示 AES

目标提示可见的方法尝试利用目标提示样本提升跨提示 AES 模型的性能, 然而, 在目标提示作文不可见的场景中, 由于模型无法利用目标提示数据, 难以学习源提示和目标提示之间的共享特征. 因此, 面向未见提示的跨提示 AES 研究比较少, 尚处于探索阶段.

目前, 有些目标提示不可见的跨提示 AES 方法假设基于通用语言特征(如词性特征)训练的模型具有一定的泛化能力, 能较容易地迁移到目标提示. Ridley 等人[14]将词性特征作为输入, 并将其特征表示与手工设计的提示无关特征进行拼接来提升模型的泛化性能. 随后, 针对现实场景中对作文属性评分的反馈需求, Ridley 等人[12]对 PAES 框架进行扩展, 提出面向未见提示的分属性跨提示 AES 方法 CTS, 采用基于注意力机制的多任务学习方法实现整体和多属性的联合评分. 进一步地, Do 等人[37]在 CTS 模型基础上, 探索了将提示信息融入到模型中, 以提升模型针对未见目标提示作文的评分能力.

不同于现有面向未见提示的跨提示 AES 方法采用提示无关的通用特征来提升模型泛化性能, 本文提出基于类别对抗联合学习的跨提示 AES, 对回归和分类进行联合建模. 一方面通过 2 个任务的联合建模学习二者的共享特征, 实现二者性能的联合提升; 另一方面, 基于分类任务引入类别级对抗网络来实现各源提示的类别级特征对齐, 从而学习到提示间的不变特征表示, 提升模型泛化性能.

1.2 领域泛化

领域泛化(domain generalization, DG)旨在基于 1 个或多个不同但相关领域的样本, 学习一个能够推广到新领域的模型[38]. 现有的领域泛化方法大致可以分为 3 种: 基于数据操作的方法、基于表示学习的方法和基于学习策略的方法[39]. 其中基于表示学习的方法是目前应用最为广泛的领域泛化方法.

大部分的基于表示学习的方法致力于学习领域间的不变特征表示. 一个具有代表性的学习领域不变特征方法是基于对抗网络 DANN[15]来对不同领域进行隐式对齐. DANN 方法最早应用于解决领域自适应问题, 后来被引入到领域泛化场景[40-43]. 该方法在分类任务中引入领域判别器, 通过特征提取器和领域判别器之间的对抗隐式地实现域分布对齐, 从而学习领域不变特征表示, 实现模型泛化性能的提升[44].

不同于现有的基于域对抗训练的方法, 本文模型基于分类任务, 通过考虑不同源提示之间的类别差异, 引入类别判别器, 通过类别判别器与特征提取器之间的对抗训练实现类别级分布对齐, 从而提升模型泛化性能.

2 基于类别对抗联合学习的跨提示 AES 方法

在本文中, 我们提出了一种基于类别对抗联合学习的跨提示 AES 方法. 本节首先对面向未见提示的跨提示 AES 进行定义. 然后, 介绍各提示共用的特征提取器架构. 随后, 介绍回归和分类的建模, 并介绍对抗网络. 最终, 介绍综合各个损失函数的模型联合训练过程. 本文模型的框架在图 2 中展示.

2.1 问题定义

AES 通常被定义为回归、分类或排序任务, 本文将其联合建模为分类和回归多任务学习框架.

假设给定 N 个源提示 $S=\{S_1,S_2,\cdots,S_p,\cdots,S_N\}$, 其中 S_p 表示源提示的作文集合 $S_p=\{(x_1^p,Y_1^p),(x_2^p,Y_2^p),\cdots,(x_m^p,Y_m^p)\}$, m 为该提示下的作文数量. 对于每个源提示 S_p, x_j^p,Y_j^p 分别代表第 j 篇作文及其评分. Y_j^p 是包括整体评分和属性评分的集合, $Y_j^p=\{y_{0,j}^p,y_{1,j}^p,\cdots,y_{M,j}^p\}$, 其中 $y_{0,j}^p$ 是整体评分, 其余 M 个评分结果则是其 M 个属性的评分, 如内容、句子流畅度等属性. 同时, 存在 1 组未评分的目标提示作文 $T=\{x_1,x_2,\cdots,x_{|T|}\}$. 其中 x_t 表示目标提示下第 t 篇作文, 目标提示的作文在训练阶段不可见.

面向未见提示的跨提示 AES 的目标是仅基于已评分的源提示作文集合 S, 学习到一个鲁棒的泛化性能强的 AES 模型 $\phi(S;\theta)$, 使得目标提示 T 的预测评分误差最小, 其中 θ 是需要学习的模型参数.

2.2 特征提取器

使用特征提取器的目的是对输入的作文进行特征表示. 本文采用被广泛认可的 Hi-att 网络[27]作为特征提取器, 该网络对作文的层次结构进行建模, 从词汇级到句子级, 再到文章级, 分层次地学习作文的嵌入表示. 假设一篇作文 x 包含 l 个句子 $s_1,s_2,\cdots,s_l$, 其中句子 s_i 包含 $|w|$ 个单词 $word_1,word_2,\cdots,word_{|w|}$. 首先通过 CNN 获取每个句子的单词嵌入:

$$z_k=f(\boldsymbol{W}_n*[\boldsymbol{w}_k:\boldsymbol{w}_{k+n-1}]+b_n), \tag{1}$$

其中 $\boldsymbol{w}_k$ 代表第 k 个词的词嵌入, f 代表激活函数, n 代表 CNN 中滑动窗口的大小, $\boldsymbol{W}_n$ 和 b_n 分别代表权重矩阵和偏置.

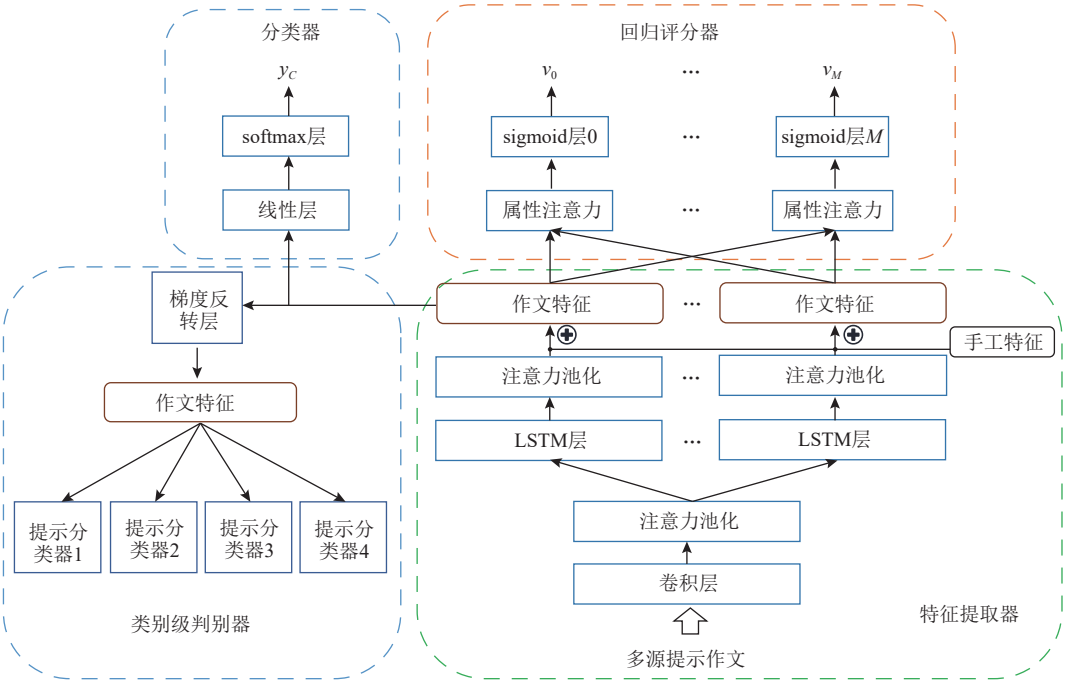


Fig. 2 The architecture of our proposed model

图 2 本文模型架构

由于在 AES 中不同的词语贡献不同，我们利用词注意力池化来增强句子表征 $\boldsymbol{s}$，其表达式为

$$\boldsymbol{s}=\sum_{k=1}^{m}a_k\boldsymbol{z}_k, \tag{2}$$

$$\tilde{\boldsymbol{z}}_k=\tanh(\boldsymbol{W}_a\boldsymbol{z}_k+b_a), \tag{3}$$

$$a_k=\frac{\exp(\boldsymbol{W}_z\tilde{\boldsymbol{z}}_k)}{\sum_{k=1}^{m}\exp(\boldsymbol{W}_z\tilde{\boldsymbol{z}}_k)}, \tag{4}$$

其中 $\boldsymbol{W}_a$ 和 $\boldsymbol{W}_z$ 表示权重矩阵，b_a 表示偏置，$\tilde{\boldsymbol{z}}_k$ 和 a_k 分别表示第 k 个词的注意力向量和注意力权重. 随后，采用 LSTM 单元聚合句子的上下文信息：

$$\boldsymbol{h}_q=f_{\mathrm{LSTM}}(\boldsymbol{s}_q,\boldsymbol{h}_{q-1}), \tag{5}$$

其中 $\boldsymbol{s}_q$ 和 $\boldsymbol{h}_q$ 分别是第 q 个输入句子及其隐状态. 之后，经过注意力池化层得到作文的特征表示 $\boldsymbol{e}$，具体的公式为

$$\tilde{\boldsymbol{h}}_q=\tanh(\boldsymbol{W}_{ha}\boldsymbol{h}_q+b_{ha}), \tag{6}$$

$$ha_q=\frac{\exp(\boldsymbol{W}_h\tilde{\boldsymbol{h}}_q)}{\sum_{q=1}^{l}\exp(\boldsymbol{W}_h\tilde{\boldsymbol{h}}_q)}, \tag{7}$$

$$\boldsymbol{e}=\sum_{q=1}^{l}ha_q\boldsymbol{h}_q, \tag{8}$$

这里，$\boldsymbol{W}_{ha}$ 和 $\boldsymbol{W}_h$ 分别表示注意力权重矩阵，b_{ha} 表示偏置项，$\tilde{\boldsymbol{h}}_q$ 和 ha_q 分别为第 q 个句子的注意力向量和注意力权重.

2.3 回归评分器

针对在实际场景中对评分反馈的需求，本文的回归评分器通过采用多任务框架对作文的整体评分和多属性评分进行联合预测.

具体的，获得作文特征表示 $\boldsymbol{e}$ 后，为了进一步提升评分性能，我们将其与手工特征 $\boldsymbol{e}_{\mathrm{h}}$ 进行拼接，即 $[\boldsymbol{e};\boldsymbol{e}_{\mathrm{h}}]$，其中本文采用与 Ridley 等人[12]相同的手工特征，包括如作文长度、可读性和文本复杂性等. 将融合手工特征的作文表示通过一个非线性特征变换层得到用于第 i 个评分任务的作文特征表示：

$$\boldsymbol{c}_i=\tanh(\boldsymbol{W}_{\boldsymbol{e}}^{i}[\boldsymbol{e};\boldsymbol{e}_{\mathrm{h}}]+b_{\boldsymbol{e}}^{i}), \tag{9}$$

其中 $\boldsymbol{W}_{\boldsymbol{e}}^{i}$ 和 $b_{\boldsymbol{e}}^{i}$ 分别为权重矩阵和偏置项参数. 因此，针对整体评分和多属性评分任务，分别将 $[\boldsymbol{e};\boldsymbol{e}_h]$ 输入不同任务的非线性特征变换层可得到其相应的特征表示集合 $\{\boldsymbol{c}_0,\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_M\}$，即 $\boldsymbol{c}_0$ 表示用于整体评分的特征表示，$\{\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_M\}$ 表示用于 M 个属性任务的特征表示.

直观地，在作文评分任务中，作文整体评分之间或各属性评分之间存在一定的关联性，为了充分考虑各属性之间及整体评分之间的关系，本文采用文献 [12] 定义的属性注意力机制来建模属性间及整体与各属性间相互影响的重要程度，从而获得用于第 m 个属性评分的最终作文特征表示 $\boldsymbol{t}_m$.

$$\boldsymbol{F}=(\boldsymbol{f}_0,\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_M), \tag{10}$$

$$u_k^m=\frac{\exp(score(\boldsymbol{f}_m,\boldsymbol{F}_{-m,k}))}{\sum_q \exp(score(\boldsymbol{f}_m,\boldsymbol{F}_{-m,q}))},\tag{11}$$

$$\boldsymbol{p}_m=\sum_k u_k^m \boldsymbol{F}_{-m,k},\tag{12}$$

$$\boldsymbol{t}_m=(\boldsymbol{f}_m;\boldsymbol{p}_m),\tag{13}$$

其中$score(\cdot)$是计算2个属性相关性的函数，u_k^m是第k个属性相对于第m个属性的注意力权重，$\boldsymbol{F}_{-m,k}$表示在$\boldsymbol{F}$中去掉$\boldsymbol{f}_m$后的第k个属性的作文特征表示.

最后，将面向整体评分和所有属性评分的作文特征表示分别送入其各自的激活函数为sigmoid的非线性回归层，预测其相应的评分值.其中第m个任务的评分值v_m为

$$v_m=\sigma(\boldsymbol{W}_m\boldsymbol{t}_m+b_m),\tag{14}$$

其中$\boldsymbol{W}_m$和b_m分别为第m个评分任务预测层的权重矩阵和偏置项.

针对每一篇作文的评分预测值和真实值，我们计算其均方误差损失，因此，第p个源提示的损失为

$$L_p=-\frac{1}{(M+1)\times N_p}\sum_{i=1}^{N_p}\sum_{m=0}^{M}(v_{m,i}^p-y_{m,i}^p)^2.\tag{15}$$

值得注意的是，由于每个提示涉及不同的体裁且不同体裁包含的评分属性不同，因此，并非每个属性都会在每个提示中出现.考虑到有些属性分数缺失的情况，因此，我们在损失函数中引入掩码函数.具体的，对于第i个训练样本，其属性标签集合Y_C^i为标签集合Y的一个子集，掩码函数为

$$\forall m\in Y, mask_{m,i}=\begin{cases}1, & \text{if}\quad Y_m\in Y_C^i,\\ 0, & \text{其他},\end{cases}\tag{16}$$

$$y_i=y_i\otimes mask_i,\tag{17}$$

$$v_i=v_i\otimes mask_i,\tag{18}$$

其中$mask_{m,i}\in[0,1]$表示第i篇作文的第m个属性掩码取值.如果$y_{m,i}^p$值为缺失的，则经过逐个掩码操作后其对应的$(v_{m,i}^p-y_{m,i}^p)^2=0$.

由于共有N个源提示，因此，总体的回归损失为

$$L_{\text{reg}}=\frac{1}{N}\sum_{p=1}^{N}L_p.\tag{19}$$

2.4 分类评分器

不同于现有AES方法仅将AES单独建模为分类或回归任务，本文对回归和分类任务进行联合建模，通过学习2个任务之间的共享特征，实现二者性能的联合提升.

针对作文分类任务，由于每篇作文都有具体的分数，我们首先将作文根据其评分所在的区间分为优、良、中和差共4个等级(具体划分标准在3.1节进行详述).分类评分器旨在预测每一篇作文的所属等级.具体的，给定作文特征表示$\boldsymbol{e}$后，将其与手工特征表示$\boldsymbol{e}_{\text{h}}$进行拼接，即$[\boldsymbol{e};\boldsymbol{e}_{\text{h}}]$，将其通过一个特征映射层得到用于作文分类的特征表示：

$$\boldsymbol{e}_c=\tanh(\boldsymbol{W}_{\boldsymbol{e}}^c[\boldsymbol{e};\boldsymbol{e}_{\text{h}}]+b_{\boldsymbol{e}}^c),\tag{20}$$

其中$\boldsymbol{W}_{\boldsymbol{e}}^c$和$b_{\boldsymbol{e}}^c$分别为该特征映射层的权重矩阵和偏置项.针对第$p$个源提示中的第$i$个样本$\boldsymbol{e}_{c,i}^p$，将其输入到作文等级分类器中，计算其预测值$C(\boldsymbol{e}_{c,i}^p)$与真实类别标签$y_{c,i}^p$之间的交叉熵损失：

$$L_{\text{cls}}^p=-\frac{1}{N_p}\sum_{i=1}^{N_p}y_{c,i}^p\ln C(\boldsymbol{e}_{c,i}^p),\tag{21}$$

其中C为评分分类器，$C(\boldsymbol{e}_{c,i}^p)$和$y_{c,i}^p$分别为$\boldsymbol{e}_{c,i}^p$的预测值和真实类别标签.由于共有$N$个源提示，因此，总体分类损失为

$$L_{\text{cls}}=\frac{1}{N}\sum_{p=1}^{N}L_{\text{cls}}^p.\tag{22}$$

2.5 类别对抗网络

为学习不同源提示之间的不变特征表示，提升模型的泛化性能，本文通过联合建模回归和分类任务，并基于分类任务引入对抗训练策略来学习不同源提示之间的不变特征表示.由于不同源提示的类别分布存在较大差异(如图1所示)，直接对各源提示进行提示级特征对齐会导致各类别分布错位对齐的问题.因此，本文考虑不同源提示之间的分布差异，对不同源提示进行类别级特征对齐.

具体地，引入4个类别判别器，将所有源提示的作文分别输入到相应的类别判别器来判别样本来自于N个源提示的哪一个.一方面，判别器被训练判别输入样本来自于哪个源提示，另一方面，特征提取器混淆判别器使其难以准确判断样本的源提示归属，通过特征提取器和判别器之间的对抗训练学到不同源提示之间的不变特征表示.给定特征提取器输出的作文表示$\boldsymbol{e}$，将其输入相应的类别判别器D_k，该判别器为N分类器，计算其预测标签和真实标签的交叉熵损失：

$$L_{\text{cls}}^k=-\frac{1}{N_k}\sum_{i=1}^{N_k}y_{k,i}^{'}\ln D_k(\boldsymbol{e}_{k,i}^{'}),\tag{23}$$

其中N_k表示所有源提示中的真实标签为第k个类别的作文数量，$D_k(\boldsymbol{e}_{k,i}^{'})$和$y_{k,i}^{'}$分别表示来自任意一个提示$(')$的第$k$个类别中第$i$个样本的预测标签和真实类别标签.由于有$C$个类别判别器，因此对抗损失为

$$L_{\text{cadv}} = -\frac{1}{C}\sum_{k=0}^{C-1} L_{\text{cls}}^{k}. \tag{24}$$

2.6 联合学习

综上所述，本文模型的最终损失由回归损失、分类损失和类别对抗损失 3 部分构成. 回归损失用来保证模型对作文进行准确评分，分类损失是用来确保模型能准确预测输入作文的所属等级，对抗损失用于混淆 C 个类别判别器使其难以区分样本属于哪一个提示. 其中在对抗过程中，借鉴 Ganin 等人[15]的工作，通过采用梯度反转层（gradient reversal layer，GRL）自动反转对抗损失的梯度方向来实现梯度的反向传播，达到对抗训练的效果. 因此，本文模型的总损失是 3 个损失的线性组合：

$$L = L_{\text{reg}} + \alpha L_{\text{cls}} + \beta L_{\text{cadv}}, \tag{25}$$

其中 α 和 β 是用来平衡不同损失项重要性的超参数.

3 实　　验

3.1 数据集和评估指标

1）数据集

本文对被广泛使用的自动学生评估奖数据集（ASAP）开展跨提示的 AES 实验. 该数据集最初是为 Kaggle 平台上的自动学生评奖竞赛构建，它包括了总计 12 987 篇由 7~10 年级学生撰写的作文. 该数据集包含 8 个不同的提示，每个提示下的作文涉及不同的体裁和评分范围. 其中涉及的体裁有议论文（argumentative）、依托文献文（source-dependent）和记叙文（narrative）三种. 议论文的目的是说服读者接受作者对某特定话题的立场，依托文献文要求针对一个给定的文献提出的问题展开论述，而记叙文则涉及围绕相应提示开展讲述或描述.

在 AES 的分类任务中，我们将不同的作文分数划分为 4 个不同的类别. 首先将作文分数标准化到 [0, 1] 范围内. 随后，我们将处于区间 (0, 0.4)，(0.4, 0.6)，(0.6, 0.8)，(0.8, 1) 的分数分别赋予对应的等级：差、中、良和优，并由类别 0, 1, 2, 3 表示. 其中 ASAP 数据集的详细信息包括提示、体裁、分数范围和各类别的数量，均在表 1 中列出.

为了给学生提供更多的评分反馈，本文除了对作文进行整体评分，还对其多个不同的属性进行评分. 然而，在 ASAP 的数据集中，仅有提示 7 和提示 8 提供了人工标注的属性评分. ASAP++数据集在原有 ASAP 数据集上，对提示 1~6 进行了属性标注. 值得注意的是，由于不同提示的作文涉及多个不同的体裁，其属性的种类及个数也有所不同. 关于每个提示下属性的详细信息已罗列在表 2.

Table 1　Detailed Statistics of ASAP Dataset

表 1　ASAP 数据集详细统计信息

提示编号	体裁	分数范围	作文总数	优的数量	良的数量	中的数量	差的数量
1	议论文	2~12	1 783	472	1 021	245	45
2	议论文	1~6	1 800	82	778	763	177
3	依托文献文	0~3	1 726	423	657	607	39
4	依托文献文	0~3	1 772	254	570	636	312
5	依托文献文	0~4	1806	258	572	649	326
6	依托文献文	0~4	1 800	367	817	405	211
7	记叙文	0~30	1 569	96	488	717	268
8	记叙文	0~60	723	21	418	275	9

Table 2　Attribute Information for ASAP and ASAP++ Datasets

表 2　ASAP 和 ASAP++数据集属性信息

提示编号	体裁	属性
1	议论文	内容、组织结构、用词选择、句子流畅度、书写规范性
2	议论文	内容、组织结构、用词选择、句子流畅度、书写规范性
3	依托文献文	内容、主题依从性、语言运用性、叙事性
4	依托文献文	内容、主题依从性、语言运用性、叙事性
5	依托文献文	内容、主题依从性、语言运用性、叙事性
6	依托文献文	内容、主题依从性、语言运用性、叙事性
7	记叙文	内容、组织结构、书写规范性
8	记叙文	内容、组织结构、用词选择、句子流畅度、书写规范性

此外，本文将整合后的 ASAP 及 ASAP++数据集依据 80%, 10%, 10% 的数量比例划分为训练集、验证集和测试集.

2）评估指标

我们采用在 AES 领域广泛使用的评估指标——平方加权卡帕（QWK）. 该指标可以测量 2 个评分器（人工评分器和自动评分器）之间的一致性. 假设训练数据有 C 个评分，QWK 值按下面的公式计算：

$$v_{\text{QWK}} = 1 - \frac{\sum_{ij} W_{ij} Q_{ij}}{\sum_{ij} W_{ij} E_{ij}}, \tag{26}$$

其中 $W_{ij} = \dfrac{(i-j)^2}{(C-1)^2}$，$i$ 和 j 分别代表人类评分员标记的

结果和自动评分器预测的标签. Q_{ij}表示被人类评分员标记为 i 但预测为 j 的作文数量. $\boldsymbol{E}$是一个 $N\times N$矩阵, 代表 2 个评分向量的外积, $\boldsymbol{E}=(E_{ij})$. v_{QWK}值越接近 1, 表示自动评分结果和人类评定分数越一致.

3.2 实现细节

在实验过程中, 所有基线模型使用其相应论文中推荐的参数. 在本文模型中, 初始化词嵌入时使用了 50 维的词嵌入. 在配置通用特征提取器 CNN 时, 我们设定窗口长度为 5 并采用 100 个卷积核. LSTM 的隐藏层状态维度为 100. 损失函数的重要性权重参数为$\alpha=5$和 $\beta=1$. 其中关于此参数的选择将在 3.7 节中详述.

本文模型的训练和测试过程是在 PyTorch 深度学习库中进行的. 我们使用小批量随机梯度下降优化器来训练. 在训练期间, 批量大小、初始学习率和最大训练周期数分别被设置为 32, 0.01, 50. *dropout* 率设定为 0.5.

3.3 基线模型

为了证明本文方法的有效性, 我们将本文方法与 6 个经典的跨提示自动作文模型进行了比较, 具体的模型如下.

1)Hi-att[27] 采用基于注意力机制的 CNN 和 LSTM 网络建模作文的层次结构. 本文将所有源提示作为训练数据学习模型, 并将其对未见目标提示进行评分. 该方法包括 Hi-att 和 Hi-att-POS 这 2 个版本, 分别表示采用词嵌入和 POS 嵌入作为输入特征的方法.

2)DANN[15] 是一种经典的基于领域对抗训练学习不变特征表示的领域自适应模型, 该方法通过引入域判别器学习不同域之间的不变特征. 本文基于该模型对作文分类进行建模, 采用提示级判别器学习各源提示之间的共享特征.

3)MAN[45] 是一种用于文本分类的改进的多项对抗网络框架. 它通过引入多项判别器来从不同的域中提取出域不变特征.

4)PAES[14] 是一种基于通用特征的面向未见提示的跨提示 AES 方法. 该模型将作文的 POS 作为输入, 并将其特征表示与提示无关的特征进行拼接从而基于所有源提示学习一个通用评分模型.

5)CTS[12] 是一种结合整体评分和属性评分的跨提示 AES 模型. 该模型在 PAES 基础上联合建模整体评分和属性评分, 采用注意力机制考虑不同属性之间的关系.

6)ProTACT[37] 是一种基于提示和属性关系感知的跨提示 AES 模型. 该模型对 CTS 框架进行改进, 通过捕捉提示和属性之间的关系, 联合建模整体评分和属性评分.

7)CAJL 是我们提出的基于类别对抗联合学习的跨提示 AES 模型.

3.4 整体评分及分类实验结果和分析

表 3 和表 4 分别列出了本文模型 CAJL 以及所有基线方法在 QWK 值上的分类和整体回归评分表现. 在实验中, 我们分别将每个提示设置为目标提示, 其余 7 个提示作为源提示来学习跨提示 AES 模型.

针对分类任务, 本文将所有的基线方法都设置为分类模型, 从表 3 可以观察到, 本文模型的分类性能优于所有的基线模型. 其中本文模型在 8 个提示上的平均 QWK 值超过最好基线模型 3.47%. 说明了本文模型采用的回归和分类联合建模和类别对抗训练策略的有效性, 本文将在消融实验部分对其有效性进行详细阐述.

此外, 根据表 3 的实验结果, 我们发现 Hi-att 模型的性能较差. 这表明由于目标提示和源提示之间存在明显的领域漂移, 直接将基于源提示数据学习的 Hi-att 模型对目标提示的作文进行评分会导致性

Table 3 Comparison of Classification Results of Different Methods on ASAP Dataset

表 3 不同方法在 ASAP 数据集上的分类结果比较

方法	提示 1	提示 2	提示 3	提示 4	提示 5	提示 6	提示 7	提示 8	平均值
Hi-att	0.265 7	0.118 8	0.162 2	0.443 9	0.427 2	0.172 8	0.354 4	0.196 2	0.267 7
Hi-att-POS	0.270 9	0.214 5	0.427 2	0.537 8	0.487 2	0.437 8	0.383 8	0.211 5	0.371 3
DANN	0.511 3	0.510 1	0.373 1	0.366 7	0.570 1	0.535 4	0.470 0	0.344 4	0.460 1
MAN	0.495 5	0.508 0	0.493 2	0.493 8	0.568 7	0.541 9	0.500 3	0.302 2	0.488 0
PAES	0.469 4	0.565 7	0.510 8	0.606 6	0.681 6	0.553 7	0.535 4	0.274 3	0.524 7
CTS	0.506 9	0.569 9	0.498 5	0.619 9	0.695 2	0.564 5	0.565 8	0.274 3	0.536 9
ProTACT	**0.516 4**	0.583 5	0.470 5	0.648 3	0.704 5	**0.571 1**	0.553 7	0.325 9	0.546 7
CAJL（本文）	0.498 5	**0.618 4**	**0.626 3**	**0.703 2**	**0.748 7**	0.536 4	**0.568 8**	**0.350 9**	**0.581 4**

注：黑体数值表示最优的结果，带下划线的为次优的结果.

Table 4 Comparison of Regression Results of Different Methods on ASAP Dataset

表 4 不同方法在 ASAP 数据集上的回归结果的比较

方法	提示 1	提示 2	提示 3	提示 4	提示 5	提示 6	提示 7	提示 8	平均值
PAES	0.780 3	0.603 8	0.582 8	0.566 9	0.684 1	0.546 5	0.608 0	0.395 3	0.596 0
CTS	0.790 1	0.601 2	0.595 0	0.570 3	0.701 0	**0.549 0**	0.620 2	0.400 7	0.603 4
ProTACT	**0.792 3**	0.608 1	0.604 5	**0.583 4**	0.723 9	0.541 3	**0.623 7**	0.390 5	0.608 5
CAJL（本文）	0.781 6	**0.608 8**	**0.609 8**	0.575 8	**0.735 4**	0.547 0	0.622 1	**0.422 7**	**0.612 9**

注：黑体数值表示最优的结果，带下划线的为次优的结果.

能严重下降. 值得注意的是, Hi-att-POS 相比于 Hi-att 性能有一定提升, 这与 PAES[14] 中结论一致, 即采用领域无关的通用特征, 如词性特征, 可以提升模型的泛化性能. DANN 和 MAN 等方法是目前经典的领域自适应方法, 本文将其直接作为用于跨提示 AES, 通过采用提示级判别器学习不同源域之间的不变特征表示. 然而, 它们的性能并不尽如人意. 其主要原因是 DANN 和 MAN 没有考虑到不同源提示的类别分布差异, 直接对各源提示的特征进行提示级对齐, 可能会引起各个类别的分布错位对齐, 从而导致性能不佳.

相比之下, PAES, CTS, ProTACT 这 3 种方法作为面向未见目标提示的跨提示 AES 方法, 取得了相对不错的分类效果. 但是这些方法性能仍低于 CAJL 的性能, 这说明虽然采用作文通用特征, 如词性特征作为输入, 并将提示无关的手工特征考虑进来可以在一定程度上提升模型的泛化性能. 然而, 这些方法仅以词性特征为输入, 将不同提示的作文转换为词性序列表示, 丢失了作文的语义信息, 使得模型无法利用提示相关信息, 从而在与提示高度相关的评分维度(如扣题度)上表现不佳. 相对而言, 本文采用的基于对抗训练的方法, 通过学习提示之间不变的特征表示, 有效地提高了模型对未见目标提示的泛化能力.

表 4 列出了本文方法与 PAES, CTS, ProTACT 这 3 种方法在回归任务上的性能比较. 通过表 4 可以发现, 本文方法在回归任务上的平均 QWK 值均超过其他 3 种方法. 尤其是在提示 5 和提示 8 的评分上, 本文方法分别超过最好的评分基线模型 1.15% 和 2.20%, 进一步验证了本文方法的有效性.

3.5 属性评分实验结果和分析

仅对作文进行整体评分难以满足学生对评分反馈的需求, 本文用 CAJL 分别对各个提示中的属性进行了评分. 由于现有跨提示 AES 方法中只有 CTS 和 ProTACT 这 2 个模型可以实现属性评分, 因此, 表 5 仅列出了 CTS, ProTACT, CAJL 在 8 个属性上的平均性能表现. 其中属性取值为该属性在所有提示上的平均 QWK 值.

从表 5 可以看出, CAJL 在 8 个属性上的平均 QWK 值分别超过 CTS 和 ProTACT 模型 1.60% 和 1.17%. 此外, CAJL 在 5 个属性评分上明显优于最好的基线模型 ProTACT. 这说明, 相比于 CTS 和 ProTACT, 虽然 CAJL 采用作文词向量而不是词性特征作为输入, 仍能取得最好的属性评分性能. 此外, CAJL 采用与 CTS 相同的属性评分网络, CAJL 除在叙事性这一属性上性能与 CTS 相当外, 在其他 6 个属性上均超过 CTS, 进一步说明了 CAJL 在属性评分任务上的有效性.

3.6 消融实验

为了进一步验证 CAJL 关键模块的有效性, 我们开展了详细的消融实验. 这些实验旨在通过逐个移除 CRJL 中的关键模块, 来说明每一模块对整个系统性能的贡献大小.

1)移除分类任务 (w/o classification). 将模型的分类任务模块去除, 相应的依赖分类任务的类别对抗模块也去除, 仅对回归任务进行建模.

2)移除对抗训练(w/o adversarial training). 对回

Table 5 Comparison of Attribute Score Results of Different Methods on All Prompts

表 5 不同方法在所有提示上的属性评分结果比较

模型	总分	内容	组织结构	单词选择	句子流畅性	规范性	主题依从度	语言规范性	叙事性	平均值
CTS	0.603 4	0.460 7	0.377 1	0.501 0	0.538 5	0.422 2	0.541 5	0.459 7	**0.575 9**	0.497 8
ProTACT	0.608 5	0.472 8	0.403 9	0.462 6	0.545 1	**0.438 1**	0.547 3	**0.474 3**	0.565 9	0.502 1
CAJL（本文）	**0.612 9**	**0.484 7**	**0.433 1**	**0.521 7**	**0.550 9**	0.426 3	**0.548 3**	0.474 1	0.572 8	**0.513 8**

注：黑体数值表示最优的结果，带下划线的为次优的结果.

归和分类任务进行联合建模，但是去除类别判别器模块.

3)采用 BERT 作为特征提取器(w/o BERT as feature extractor). 将模型采用的 Hi-att 特征提取器更换为 BERT 特征提取器，在模型优化过程中仅优化 BERT 模型的最后 2 层参数.

表 6 列出了去除各个模块的实验结果. 通过表 6 可以看出，去掉分类任务的建模后，各个模型性能相比 CAJL 有明显下降，这说明对分类和回归任务进行联合建模可以实现对 AES 任务性能的提升. 大部分的作文评分任务是将其建模为回归任务，预测作文的真实得分. 本文对分类和回归任务进行联合建模，通过学习能够区分不同评分等级的判别特征来实现对回归任务的促进. 此外，去掉类别对抗模块后，模型性能相比于 CAJL 性能也有一定程度的下降，这说明了类别对抗模块的有效性. 这主要得益于基于分类任务引入了类别判别器，通过特征提取器和类别判别器的对抗训练，可以学到不同源提示之间的不变特征表示，从而提升模型对未见提示的评分性能. 总的来说，对分类和回归任务进行联合建模和采用类别对抗训练都能对模型性能提升起到促进作用.

此外，我们发现去掉分类任务建模后，提示 1 和提示 6 的性能略有提升，其可能的原因为：提示 1 要求学生向本地报纸写一封阐述其观点的信，对学生来说写作要求相对简单. 相比之下，提示 6 是由 10 年级的学生撰写的依托文献文体裁作文，其相应的写作要求相对较难. 因此，提示 1 的作文分数集中分布在较高的分数段而提示 6 的作文集中分布在中等的分数段(如表 1 所示)，这就导致虽然很多学生的分数比较相近，但是却被映射到不同的评分等级中，此时，对分类任务的建模会使得模型性能下降. 同时，我们发现，去掉对抗训练模块后，提示 7 和提示 8 的性能有一定提升，其可能原因为：提示 7 的评分属性仅包含内容、组织结构、书写规范性，与其他 7 个提示的评分属性有较大的差异(如表 2 所示)；提示 8 的作文数量较少且各评分等级的分布与其他 7 个提示的评分等级的分布有较大不同(如表 1 所示)，因此，对于提示 7 和提示 8 来说，直接将基于其他 7 个提示学习到的不变特征表示进行评分会导致评分性能的下降.

最后，通过表 6 我们可以看到，CAJL 采用 BERT 作为特征提取器后，其平均 QWK 值比采用 Hi-att 作为特征提取器的模型性能提升了 5.22%. 这主要是由于 BERT 模型经过大规模文本的预训练蕴含丰富的知识，仅对 BERT 模型的最后 2 层参数进行优化可以既保留模型蕴含的丰富知识又能很好地适用于下游的评分和分类任务. 这也进一步说明 CAJL 可以适配于不同的特征提取器，尤其是可以结合现有预训练模型的优势，实现跨提示 AES 性能的进一步提升.

3.7 可视化与参数敏感性实验

为了更直观地理解 CAJL 的实验结果，我们对 CTS 和 CAJL 模型进行了特征可视化. 由于 ASAP 数据集中 8 个提示共涵盖议论文、依托文献文和记叙文 3 种体裁，因此，我们从每个体裁中选一个提示进行可视化，即分别对提示 1、提示 4 和提示 7 的基于 CTS 和 CAJL 模型学到的特征分布进行可视化分析，来说明模型的有效性.

图 3 展示了使用 CTS 和 CAJL 方法学习的特征分布. 其中红色表示源提示的特征，绿色表示目标提示的特征. 左图为通过 CTS 学习的特征分布，右图为基于 CAJL 模型学习到的特征分布. 我们采用 t-SNE 技术，将源提示和目标提示的特征投影到 2 维空间以实现可视化. 从图 3 可见，CTS 学到的特征分布显得较为分散，缺乏关联性，源提示特征与目标提示特征之间的分布差异性较大. 相比之下，CAJL 学习的特征分布呈现出一定的连续性，由于作文得分集中在某些分数值或分数段，其特征点呈现出一定的重合或连接在一起的现象. 这种连续和重合的特征分布有利于回归问题的建模，因为它能较好地体现不同分数段之间的过渡.

此外，源提示特征与目标提示特征的连续对齐证明了 CAJL 能够有效学习不同提示之间的不变特

Table 6 Effects of Key Components of the CAJL on Regression Score

表 6 CAJL 模型关键模块对回归评分结果的影响

方法	提示 1	提示 2	提示 3	提示 4	提示 5	提示 6	提示 7	提示 8	平均值
完整 CAJL（本文）	0.781 6	0.608 8	0.609 8	0.575 8	**0.735 4**	0.547 0	0.622 1	0.422 7	0.612 9
移除分类任务	**0.790 1**	0.601 2	0.595 0	0.570 3	0.701 0	0.549 0	0.620 2	0.400 7	0.603 4
移除对抗训练	0.764 8	0.587 4	0.602 9	**0.579 1**	0.701 7	0.545 1	0.624 3	0.443 1	0.606 1
采用 BERT 作为特征提取器	0.777 8	**0.634 5**	**0.624 1**	**0.579 1**	0.729 8	**0.583 4**	**0.732 1**	**0.657 0**	**0.664 7**

注：黑体数值表示最优的结果.

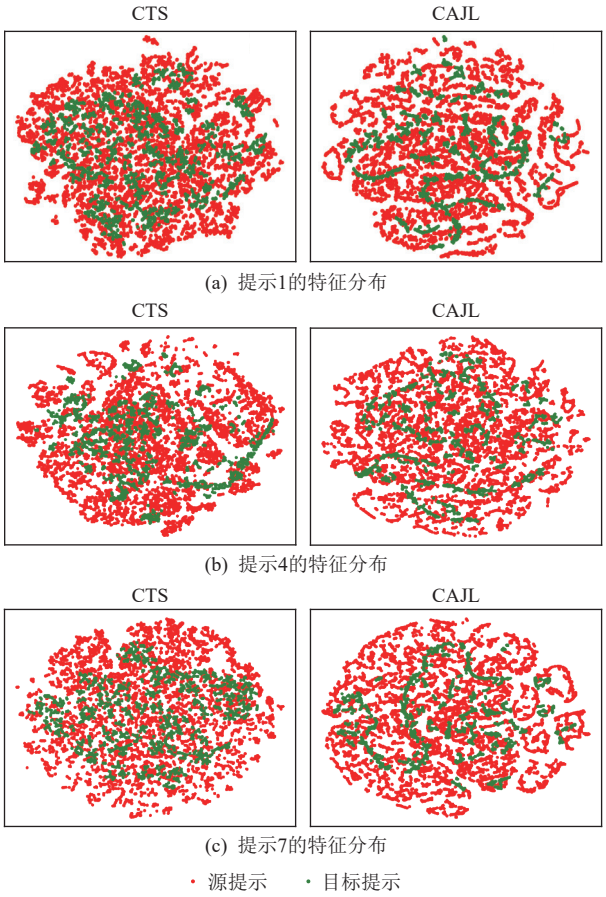


Fig. 3 Visualization results of the distribution of features learned through CTS and CAJL

图 3 通过 CTS 与 CAJL 模型学习得到的特征分布可视化结果

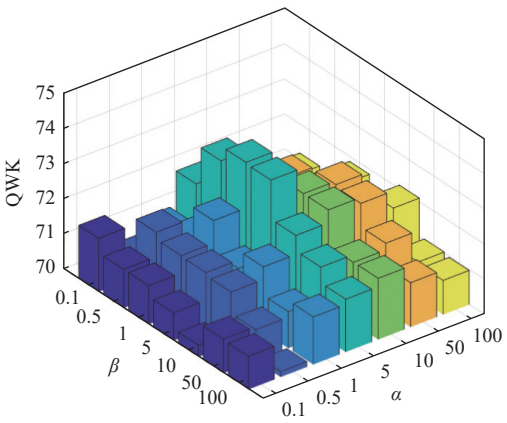
征表示.

进一步地, 图 3 也展示了基于 CAJL 学习的 3 种体裁中, 议论文(图 3(a))和记叙文(图 3(c))的特征分布呈现出一定的规律性, 而依托文献文(图 3(b))的特征分布则相对分散. 这一差异可能源于议论文和记叙文具有较为固定的撰写模式及结构, 大部分作文遵循其模式, 使得它们的特征在空间上聚集得更紧密, 从而形成了较为清晰的分布规律. 相较而言, 依托文献文主要是围绕特定文献提出的问题进行分析, 作文写作呈现出较大的差异, 其分布相对分散. 这种观察指出了不同文体特征分布的内在差异, 进一步验证了 CAJL 在捕捉和区分不同文体特征方面的有效性.

图 4 展示了超参数 α 和 β 对模型的性能影响. 这 2 个超参数分别调节回归损失和分类损失与类别感知对抗损失之间的相对权重. 为了研究 α 和 β 的敏感性, 我们采用了网格搜索策略, 选取了 1 组数值 0.1, 0.5, 1, 5, 10, 50, 100 来进行测试. 从图 4 可以观察到, 随着 α 和 β 数值的增长, 模型评分表现呈现出先上升后下降的趋势, 并在 $\alpha=5$ 和 $\beta=1$ 时达到最优的表现. 此外, 我们也注意到, 尽管 α 和 β 的取值有较大范围的变化, 模型性能的波动却相对较小, 这一发现说明了 CAJL 模型在不同参数设置下的鲁棒性和稳定性.

4 结 论

本文提出了一种基于类别对抗联合学习的跨提示 AES 方法. 该方法将作文评分建模为结合分类与回归的多任务联合学习框架. 通过学习分类和回归任务的共享特征, 实现二者性能的联合提升, 同时, 基于分类任务引入对抗训练策略学习不同提示间的不变特征表示. 此外, 为了避免进行提示级特征对齐时可能出现的错位对齐问题, 本文采用类别对抗网络进行类别级的特征对齐. 通过将本文方法应用于



(a) QWK值的变化趋势

β \ α	0.1	0.5	1	5	10	50	100
100	0.709 4	0.701 5	0.713 5	0.715 0	0.717 4	0.712 6	0.709 7
50	0.709 7	0.708 8	0.710 4	0.719 5	0.717 3	0.719 2	0.710 3
10	0.703 3	0.714 9	0.718 5	0.723 9	0.727 6	0.725 9	0.721 0
5	0.707 9	0.715 8	0.713 4	0.735 0	0.727 6	0.726 6	0.714 5
1	0.711 0	0.716 0	0.722 4	0.735 4	0.713 3	0.713 7	0.718 4
0.5	0.711 0	0.718 0	0.712 5	0.731 2	0.710 8	0.719 3	0.711 6
0.1	0.715 2	0.706 1	0.710 4	0.720 0	0.709 0	0.712 2	0.711 6

(b) QWK值分布

Fig. 4 Effect of hyperparameters on performance of models in ASAP dataset

图 4 在 ASAP 数据集上超参对模型性能的影响

ASAP和ASAP++数据集进行实验验证,实验结果表明本文方法优于现有的跨提示AES方法.由于本文仅采用比较简单的Hi-att模型进行作文特征表示,在未来的工作我们将研究如何结合预训练模型采用基于提示学习学习不同提示间的不变特征表示.

作者贡献声明:张春云提出了算法思路和实验方案;赵洪焱负责完成实验并撰写论文;邓纪芹和董晓琳辅助完成实验;崔超然和陈竹敏提出指导意见并修改论文.

参 考 文 献

[1] Ke Z, Ng V. Automated essay scoring: A survey of the state of the art [C] //Proc of the 28th Int Joint Conf on Artificial Intelligence. Palo Alto, CA: IJCAI, 2019: 6300−6308

[2] Ding Gejian, Liu Chang. Survey of automated essay scoring technology[J]. Journal of Computer Applications, 2022, 42(S1): 386−390 (in Chinese)
(丁革建, 刘畅. 作文自动评分技术综述[J]. 计算机应用, 2022, 42(S1): 386−390)

[3] Page E B. The imminence of grading essays by computer[J]. The Phi Delta Kappan, 1966, 47(5): 238−243

[4] Landauer T K. Automated scoring and annotation of essays with the intelligent essay assessor [J]. Automated Essay Scoring: A Cross-Disciplinary Perspective, 2003: 87−112

[5] Attali Y, Burstein J. Automated essay scoring with e-rater V.2[J]. Journal of Technology, Learning, and Assessment, 2006, 4(3): 1−30

[6] Kim Y. Convolutional neural networks for sentence classification [C] // Proc of the 2014 Int Conf on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA: ACL, 2014: 1746−1751

[7] Yang Zichao, Yang Diyi, Dyer C, et al. Hierarchical attention networks for document classification [C] //Proc of the 2016 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2016: 1480−1489

[8] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 5998−6008

[9] Cao Yue, Jin Hanqi, Wan Xiaojun, et al. Domain-adaptive neural automated essay scoring [C] //Proc of the 43rd Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2020: 1011−1020

[10] Li Xia, Chen Minping, Nie Jianyun. SEDNN: Shared and enhanced deep neural network model for cross-prompt automated essay scoring[J]. Knowledge-Based Systems, 2020, 210: 106491

[11] Song Wei, Zhang Kai, Fu Ruiji, et al. Multi-stage pre-training for automated Chinese essay scoring [C] //Proc of the 2020 Conf on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA: ACL, 2020: 6723−6733

[12] Ridley R, He Liang, Dai Xinyu, et al. Automated cross-prompt scoring of essay traits [C] //Proc of the 35th Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 13745−13753

[13] Chen Yuan, Li Xia. PMAES: Prompt-mapping contrastive learning for cross-prompt automated essay scoring [C] //Proc of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2023: 1489−1503

[14] Ridley R, He Liang, Dai Xinyu, et al. Prompt agnostic essay scorer: A domain generalization approach to cross-prompt automated essay scoring [J]. arXiv preprint, arXiv: 2008.01441, 2020

[15] 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

[16] Chen Xinyang, Wang Sinan, Wang Jianmin, et al. Representation subspace distance for domain adaptation regression [C] //Proc of the 38th Int Conf on Machine Learning. New York: ACM, 2021: 1749−1759

[17] Mathias S, Bhattacharyya P. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores [C] //Proc of the 11th Int Conf on Language Resources and Evaluation (LREC 2018). Stroudsburg, PA: ACL, 2018: 1169−1173

[18] Larkey L S. Automatic essay grading using text categorization techniques [C] //Proc of the 21st Int 47th ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 1998: 90−95

[19] Uto M, Xie Yikuan, Ueno M. Neural automated essay scoring incorporating handcrafted features [C] //Proc of the 28th Int Conf on Computational Linguistics. Stroudsburg, PA: ACL, 2020: 6077−6088

[20] Peng Xingyuan, Ke Dengfeng, Zhao Zhi, et al. Automated Chinese essay scoring based on word scores[J]. Journal of Chinese Information Science, 2012, 26(2): 102−108 (in Chinese)
(彭星源, 柯登峰, 赵知, 等. 基于词汇评分的汉语作文自动评分[J]. 中文信息学报, 2012, 26(2): 102−108)

[21] Taghipour K, Ng H T. A neural approach to automated essay scoring [C] //Proc of the 2016 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2016: 1882−1891

[22] Alikaniotis D, Yannakoudakis H, Rei M. Automatic text scoring using neural networks [C] //Proc of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2016: 715−725

[23] Wang Yaohua, Li Zhoujun, Chao Wenhan, et al. Research on key technology of automatic essay scoring based on text semantic dispersion[J]. Journal of Chinese Information Science, 2016, 30(6): 173−181 (in Chinese)
(王耀华, 李舟军, 巢文涵, 等. 基于文本语义离散度的自动作文评分关键技术研究[J]. 中文信息学报, 2016, 30(6): 173−181)

[24] Tay Y, Phan M C, Tuan L A, et al. SKIPFLOW: Incorporating neural coherence features for end-to-end automatic text scoring [C] //Proc of

the 32nd Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 5948−5955

[25] Fu Ruiji, Wang Dong, Wang Shijin, et al. Elgart sentence recognition for automated essay scoring[J]. Journal of Chinese Information Science, 2018, 32(06): 88−97 (in Chinese)
(付瑞吉, 王栋, 王士进, 等. 面向作文自动评分的优美句识别[J]. 中文信息学报, 2018, 32(06): 88−97)

[26] Zhou Xianbing, Fan Xiaochao, Ren Ge, et al. Automated English essay scoring method based on multi-level semantic features[J]. Journal of Computer Applications, 2021, 41(08): 2205−2211 (in Chinese)
(周险兵, 樊小超, 任鸽, 等. 基于多层次语义特征的英文作文自动评分方法[J]. 计算机应用, 2021, 41(08): 2205−2211)

[27] Dong Fei, Zhang Yue, Yang Jie. Attention-based recurrent convolutional neural network for automatic essay scoring [C] //Proc of the 21st Conf on Computational Natural Language Learning. Stroudsburg, PA: ACL, 2017: 153−162

[28] Xie Jiayi, Cai Kaiwei, Kong Li, et al. Automated essay scoring via pairwise contrastive regression [C] //Proc of the 29th Int Conf on Computational Linguistics. Stroudsburg, PA: ACL, 2022: 2724−2733

[29] Lin Szu-Yin, Kung Yun-Chi, Leu Fang-Ye. Predictive intelligence in harmful news identification by BERT-based ensemble learning model with text sentiment analysis[J]. Information Processing & Management, 2022, 59(2): 102872

[30] Rodriguez P U, Jafari A, Ormerod C M. Language models and automated essay scoring [J]. arXiv preprint, arXiv: 1909.09482, 2019

[31] Briskilal J, Subalalitha C N. An ensemble model for classifying idioms and literal texts using BERT and RoBERTa[J]. Information Processing & Management, 2022, 59(1): 102756

[32] Yang Ruosong, Cao Jiannong, Wen Zhiyuan, et al. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking [C]//Proc of the 58th Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 1560−1569

[33] Ormerod C M, Malhotra A, Jafari A. Automated essay scoring using efficient transformer-based language models[J]. arXiv preprint, arXiv: 2102.13136, 2021

[34] Sethi A, Singh K. Natural language processing based automated essay scoring with parameter-efficient transformer approach [C] //Proc of the 6th Int Conf on Computing Methodologies and Communication (ICCMC). Piscataway, NJ: IEEE, 2022: 749−756

[35] Yu Mingcheng, Dang Yagu, Wu Qilin, et al. Research on automatic scoring for English essay based on multi-scale context[J]. Journal of Computer Engineering, 2024, 50(3): 259−266 (in Chinese)
(于明诚, 党亚固, 吴奇林, 等. 基于多尺度上下文的英文作文自动评分研究[J]. 计算机工程, 2024, 50(3): 259−266)

[36] Jin Cancan, He Ben, Hui Kai, et al. TDNN: A two-stage deep neural network for prompt-independent automated essay scoring [C] //Proc of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2018: 1088−1097

[37] Do H, Kim Y, Lee G G. Prompt-and trait relation-aware cross-prompt essay trait scoring[C] //Proc of the 61st Association for Computational Linguistics. Stroudsburg, PA: ACL. 2023: 1538−1551

[38] Farahani A, Voghoei S, Rasheed K, et al. A brief review of domain adaptation [J/OL]. Advances in Data Science and Information Engineering, 2021: 877−894 [2021-10-30]. https://link.springer.com/chapter/10.1007/978-3-030-71704-9_65

[39] Wang Jindong, Lan Cuiling, Liu Cang, et al. Generalizing to unseen domains: A survey on domain generalization[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(8): 35−147

[40] Li Haoliang, Pan S J, Wang Siqi, et al. Domain generalization with adversarial feature learning [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 5400−5409

[41] Matsuura T, Harada T. Domain generalization using a mixture of multiple latent domains [C] //Proc of the 34th Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 11749−11756

[42] Zhu Kaijie, Hu Xixu, Wang Jindong, et al. Improving generalization of adversarial training via robust critical fine-tuning [C] //Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 4424−4434

[43] Singh N D, Croce F, Hein M. Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models[J]. Advances in Neural Information Processing Systems, 2024, 36: 1−14

[44] Feng Jun, Shi Yichen, Gao Yuhao, et al. Domain adaptation for face anti-spoofing based on dual disentanglement and liveness feature progressive alignment[J]. Journal of Computer Research and Development, 2023, 60(8): 1727−1739 (in Chinese)
(封筠, 史屹琛, 高宇豪, 等. 二次解耦与活体特征渐进式对齐的域自适应人脸反欺诈[J]. 计算机研究与发展, 2023, 60(8): 1727−1739)

[45] Chen Xilun, Cardie C. Multinomial adversarial networks for multi-domain text classification [C] //Proc of the 2018 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Stroudsburg, PA: ACL, 2018: 1226−1240



Zhang Chunyun, born in 1986. PhD, associate professor. Member of CCF. Her main research interests include natural language processing and machine learning.
张春云, 1986 年生. 博士, 副教授. CCF 会员. 主要研究方向为自然语言处理、机器学习.



Zhao Hongyan, born in 2000. Master. His main research interests include natural language processing and machine learning.
赵洪焱, 2000 年生. 硕士. 主要研究方向为自然语言处理、机器学习.



Deng Jiqin, born in 1996. Master. Her main research interests include natural language processing and machine learning.

邓纪芹, 1996 年生. 硕士. 主要研究方向为自然语言处理、机器学习.



Cui Chaoran, born in 1987. PhD, professor, Young Taishan scholar of Shandong province. Member of CCF. His main research interests include machine learning and data mining.

崔超然, 1987 年生. 博士, 教授, 山东省青年泰山学者. CCF 会员. 主要研究方向为机器学习、数据挖掘.



Dong Xiaolin, born in 1998. Master. Her main research interest includes machine learning.

董晓琳, 1998 年生. 硕士. 主要研究方向为机器学习.



Chen Zhumin, born in 1978. PhD, professor. Member of CCF. His main research interests include machine learning and natural language processing.

陈竹敏, 1978 年生. 博士, 教授. CCF 会员. 主要研究方向为机器学习、自然语言处理.