doi：10.13413/j.cnki.jdxblxb.2023330

基于双层数据增强的监督对比学习文本分类模型

吴 量，张方方，程 超，宋诗楠
（长春工业大学 计算机科学与工程学院，长春 130012）

摘要：针对 DoubleMix 算法在数据增强时的非选择性扩充及训练方式的不足，提出一种基于双层数据增强的监督对比学习文本分类模型，有效提高了在训练数据稀缺时文本分类的准确率. 首先，对原始数据在输入层进行基于关键词的数据增强，不考虑句子结构的同时对数据进行有选择增强；其次，在 BERT 隐藏层对原始数据与增强后的数据进行插值，然后送入 TextCNN 进一步提取特征；最后，使用 Wasserstein 距离和双重对比损失对模型进行训练，进而提高文本分类的准确率. 对比实验结果表明，该方法在数据集 SST-2，CR，TREC 和 PC 上分类准确率分别达 93.41%，93.55%，97.61%和 95.27%，优于经典算法.

关键词：数据增强；文本分类；对比学习；监督学习
中图分类号：TP39 **文献标志码**：A **文章编号**：1671-5489(2024)05-1179-09

Supervised Contrastive Learning Text Classification Model Based on Double-Layer Data Augmentation

WU Liang, ZHANG Fangfang, CHENG Chao, SONG Shinan
(*College of Computer Science and Engineering, Changchun University of Technology, Changchun* 130012, *China*)

Abstract: Aiming at the non-selective expansion and training deficiencies of the DoubleMix algorithm during data augmentation, we proposed a supervised contrastive learning text classification model based on double-layer data augmentation, which effectively improved the accuracy of text classification when training data was scarce. Firstly, keyword-based data augmentation was applied to the original data at the input layer, while selectively enhancing the data without considering sentence structure. Secondly, we interpolated the original and augmented data in the BERT hidden layers, and then send them to the TextCNN for further feature extraction. Finally, the model was trained by using Wasserstein distance and double contrastive loss to enhance text classification accuracy. The comparative experimental results on SST-2, CR, TREC, and PC datasets show that the classification accuracy of the proposed method is 93.41%, 93.55%, 97.61%, and 95.27% respectively, which is superior to classical algorithms.

Keywords: data augmentation; text classification; comparative learning; supervised learning

收稿日期：2023-08-04.

第一作者简介：吴 量(1987—)，男，汉族，博士，副教授，从事机器视觉技术和空间光学测量导航的研究，E-mail：wuliang@ccut.edu.cn. **通信作者简介**：宋诗楠(1990—)，女，汉族，博士，讲师，从事深度神经网络和边缘计算的研究，E-mail：songshinan@163.com.

基金项目：吉林省发展和改革委员会项目(批准号：2022C047-7)和长春市科技发展计划项目(批准号：21GD05).

文本分类是自然语言处理(NLP)的基本任务之一，在新闻过滤、论文分类、情感分析等方面应用广泛[1-2]. 深度学习模型在文本分类中已取得了巨大成功，其通常建立在大量高质量的训练数据上，而这些数据在实际应用中并不容易获得. 因此，为提高文本分类模型的泛化能力，当训练数据有限时，数据增强技术得到广泛关注[3]. 文本分类要获得较好的分类精度，好的特征表示和分类器的训练也至关重要[4].

在自然语言处理领域中，存在标记级别增强(token-level augment)、句子级别增强(sentence-level augment)、隐藏层增强(hidden-level augment)等类型[5]. EDA[6](easy data augmentation)是最常见的标记级别数据增强，通过对句子中的单词进行随机替换、删除、插入等操作实现数据增强. 句子级别的增强通过修改句子的语法或结构实现，最常见的是反向翻译技术[7]. 隐藏层数据增强的方法是基于对数据插值(interpolation)实现的. Mixup[8]是最早出现的一种基于插值的增强方式，TMix[9](interpolation in textual hidden space)是在其基础上发展的线性插值数据增强方式. Ssmix[10](saliency-based span mixup)是一种输入级的混合插值方式. 上述几种插值方式都伴随伪标签(soft label)生成，会限制数据增强的有效性. DoubleMix[5]增强方法的提出避免了伪标签生成，首先利用EDA与回译技术从原始数据中生成几个扰动样本，然后在隐藏空间中混合扰动样本与原始样本，最后采用JSD(Jensen-Shannon divergence)散度为正则项与交叉熵损失一起训练. 但DoubleMix生成扰动样本的方式有的对句子结构要求较高，有的对文本进行非选择性的补充. 低资源条件下，会限制数据增强的有效性，且不易短时间生成大量高质量的增强数据. 并且在训练时JSD散度会出现梯度为零的情况，交叉熵损失也存在噪声等问题[11].

针对上述问题，本文提出一种基于双层数据增强(TDACL)的有监督对比学习文本分类模型. 首先，在输入层对原始数据进行基于关键词的数据增强；其次，在BERT[12]编码层中插值原始数据和输入层增强后的数据以生成新的样本特征表示，并将每一层的空间特征和分类器特征送到TextCNN[13]中进一步提取；最后，利用Wasserstein距离和双重对比损失DualCL[14]进行训练. 在公开的文本分类数据集上进行多次重复实验，实验结果表明，该模型能提升文本分类性能，尤其是在低资源情况下.

1 模型设计

DoubleMix数据增强生成扰动样本的方式有EDA和回译. EDA通过对句子中的单词进行随机操作生成增强样本，但其未考虑关键词的作用，进行操作时可能会删掉关键词，导致不理想的扩充. 回译的方式也不宜在短时间内获得大量高质量的样本，且其在训练时JSD散度会出现梯度为零的情况，交叉熵损失也存在噪声等问题. 因此，本文提出一种基于双层数据增强的有监督对比学习文本分类模型，在输入层进行基于关键词的数据增强. 该方法不需要考虑句子的结构，能快速生成大量高质量的样本，有选择地对样本进行扩充，进而提高数据增强的有效性；将样本经过BERT和TextCNN共同提取特征后，利用Wasserstein距离与双重对比损失DualCL进行训练，从而解决梯度消失问题，并最小化增强样本与输入样本的差异，进而学到更利于分类的特征表示，最终提高文本分类的准确率.

1.1 模型的主要框架

本文基于双层数据增强的有监督对比学习文本分类模型总体架构如图1所示. 给定BERT语言模型和监督数据集D，在数据集D上对BERT进行微调，以获得D的多样化特征表示，更适合下游分类任务. 下面首先介绍该方法的总体框架，然后描述输入层和BERT编码层的数据增强策略，最后结合双重监督对比损失和Wasserstein距离对模型进行训练. 本文模型框架主要由以下四部分组成：

1) 输入层的数据增强模块，通过对样本进行基于关键词的数据增强，生成大量高质量的样本；

2) 一个共享的BERT编码器，对原始样本和增强的高质量样本进行插值提高模型的鲁棒性；

3) TextCNN特征提取层，将BERT得到的向量表示输入TextCNN进行进一步特征提取，得到更好的句子向量表示和特征向量表示；

4) 使用双重对比度损失和Wasserstein距离对模型进行训练，最小化原始数据和增强数据之间的

差距，使提取的特征表示更紧凑，更好服务于文本分类任务.

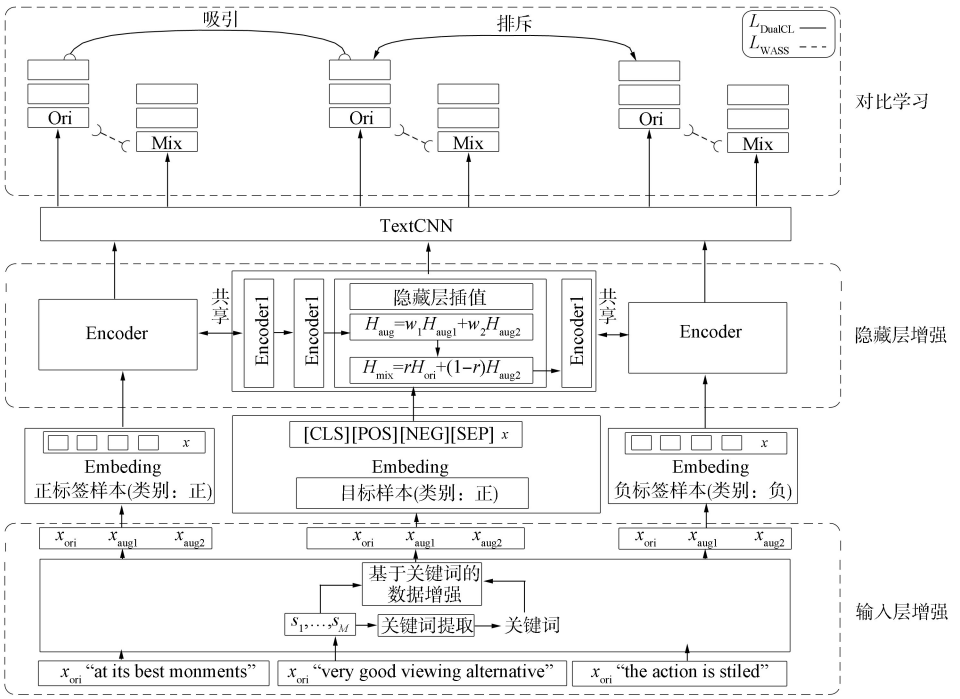

图 1　TDACL 的总体架构

Fig. 1　Overall architecture of TDACL

1.2　输入层基于关键词的数据增强方法

基于输入层关键词操作的数据增强方法可在不考虑句子结构的情况下，在短时间内生成大量高质量的样本. EDA 是一种简单而有效的数据增强方法，该方法通过替换同义词、删除输入、随机替换和随机插入生成增强的数据. 但 EDA 未考虑关键词的作用，随机删除等操作会对增强后的文本产生很多不利影响. 本文基于关键词的数据增强方法，首先提取原始数据的关键词，然后对数据进行基于关键词的增强.

KeyBERT 提取关键词的方法不需要针对特定的文档集进行培训，也不依赖于文本大小、领域或语言，它使用 BERT 嵌入和余弦相似度在文本中找到与文本本身最相似的单词或短语. 因此，本文使用 KeyBERT 提取文本中的关键词，并执行基于关键词的文本数据增强. 增强步骤为：首先利用 BERT 模型进行文本嵌入，得到文本的表示形式；然后采用词嵌入模型提取关键词；最后使用余弦相似度查找与文本最相似的单词作为样本的关键词. 之后对文本进行基于关键词的增强. 基于关键词的替换(KRE)：选择原句子中的 n 个关键词，并将其替换为同义词. 基于关键词的插入(KIN)：根据原句子的长度比例选择插入的单词数量，在原文本的任意位置插入关键词的同义词. 基于关键词的交换(KSW)：在句子中选择 n 个关键词，根据句子长度和概率交换它们的位置. 基于关键词的选择(KSE)：只选择句子中的关键词生成新句子. 这种方法通过添加标点使句子更自然. n 的选择由句子的长度 l 和超参数 p 定义，即 $n=p\times l$. p 的强度可根据任务变化，本文设 $p=0.1$.

基于关键词选择的数据增强会生成不同的增强样本，如图 2 所示. KRE 保证增强的原始语义不会改变. 为确保增强后的数据不与原数据过于相似，KIN 采用插入关键词同义词的方法. 因为关键词在文本分类的拟合中起决定性作用，KSE 可被认为是选择性地过滤掉有噪声的单词，在对文本进行分类时帮助学习最能代表文本特征的单词. 本文的关键词提取算法是一种动态算法，不需要将所有的训练数据和标签信息都投入到词频统计等计算中[15]. 本文方法具有提取关键词速度快，并保证在短时间内生成大量高质量样本的优点.

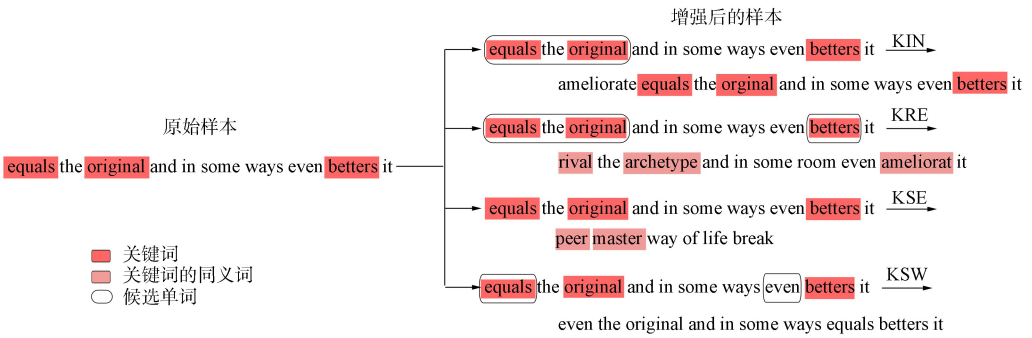

图 2 基于关键词的数据增强

Fig. 2 Data augmentation based on keyword

1.3 隐藏层数据增强

在对数据进行基于关键词的增强后，将原始数据与增强后的数据在隐藏层进行空间插值，实现隐藏层的数据增强. 不同于 DoubleMix 数据增强方式，本文将句子的标签与输入拼接到一起送入到编码器中(见图 1). 例如，当数据集的标签为“positive”和“negative”时，拼接的输入变成“[CLS] positive negative [SEP] I like this movie”. 本文使用编码样本对应标签的 token 值作为分类器特征. 如果一个标签包含多个单词，则将 token 特征的平均池化作为分类器特征[16]. 这种处理方法便于后续提取空间特征和分类器特征，从而进行有监督对比学习.

实验结果表明，BERT 编码层的{9,10,12}层插值效果最好. 将原始样本与输入层增强后的样本送入编码器，各编码层之间共享参数. 采用 Dirichlet 分布进行增强数据的权重组合. 定义权重 $(w_1,w_2,\cdots,w_n)\sim \mathrm{Dir}(\alpha)$，在输入层增强后的数据与原始数据在{9,10,12}层对其插值，得到插值后的增强表示(H_{aug})记为

$$H_{\mathrm{aug}}=\{w_1 h_{\mathrm{aug}}^{\mathrm{KIN}},w_2 h_{\mathrm{aug}}^{\mathrm{KSE}},w_3 h_{\mathrm{aug}}^{\mathrm{KSW}},w_n h_{\mathrm{aug}}^{\mathrm{KSE}}\}. \tag{1}$$

然后对第一步插值后的结果(H_{ori})与原始数据的隐藏状态进行权重服从 Beta 分步的方式进行插值，插值后的结果为

$$H_{\mathrm{mix}}=\beta\cdot H_{\mathrm{ori}}+(1-\beta)\cdot H_{\mathrm{aug}}. \tag{2}$$

这种插入方式可保证原始的数据集占据主要地位，同时又混入增强的数据集，增强其扰动性.

1.4 基于 TextCNN 的特征提取

本文将 BERT 编码层每层提取的样本空间特征和分类器特征输入 TextCNN 进行进一步处理. 文献[17]研究表明，在处理下游任务时，直接从 BERT 获得的最后一层的向量表示并不能给出最好的结果. BERT 的编码层越浅，句子的语义信息层次越低；层越深，表示的语义信息层次越高. TextCNN 的核心思想是捕捉局部特征. TextCNN 的优势是其可自动组合并过滤 N-gram 特征，从而获得不同抽象层次的语义信息. 当 TextCNN 应用于文本分类任务时，使用几个不同大小的核提取句子中的关键信息，可以更好地捕捉本地相关性. 处理后的空间特征包含语义特征和关键词信息.

本文从 BERT 的每个编码器层(不包括第一层输入)中提取样本空间特征 f_{CLS} 和分类器特征 f_{POS}，f_{NEG}(假设是二分类)，并将其作为 TextCNN 的输入进行进一步的特征提取. BERT 可以更好地对模糊样本进行分类，而 TextCNN 可以更关注关键词信息. 因此，经过两步特征提取，本文获得的空间特征包括句子的语义信息和关键词信息. 该方法还尝试从编码器的最后一层提取空间特征和分类器特征，但结果没有每层都提取的结果好.

1.5 模型的训练

将 TextCNN 提取到的 $f_{\mathrm{CLS}},f_{\mathrm{CLSmix}}$ 使用 Wasserstein 距离作一致正则项，将空间特征 f_{CLS} 与分类器特征 $f_{\mathrm{POS}},f_{\mathrm{NEG}}$ 使用双重对比损失，一同对模型进行训练. 这样可以解决梯度消失问题，并最小化增强样本与输入样本的差异，双重对比损失的加入可以更好地学习特征表示，服务于下游分类任务.

将经过空间插值后获得的样本 x_i 特征记为 f_i，f_i 为输入样本 x_i 的一个锚点，$\{\Phi_j^*\}_{j\in C_i}$ 为正样本

集合，$\{\Phi_j^*\}_{j\in R_i/C_i}$为负样本集合. 关于空间特征的对比损失定义为

$$L_f=\frac{1}{N}\sum_{i\in I}\frac{1}{|C_i|}\sum_{c\in C_i}-\log\frac{\exp\{\Phi_c^*\cdot f_i/\tau\}}{\sum_{r\in R_i}\exp\{\Phi_r^*\cdot f_i/\tau\}}. \tag{3}$$

同理，Φ_i^* 为输入样本 x_i 的一个锚点，$\{f_j\}_{j\in C_i}$为正样本集合，$\{f_j\}_{j\in R_i/C_i}$为负样本集合. 关于分类器参数的对比损失定义为

$$L_\Phi=\frac{1}{N}\sum_{i\in I}\frac{1}{|C_i|}\sum_{c\in C_i}\left(-\log\frac{\exp\{\Phi_c^*\cdot f_c/\tau\}}{\sum_{r\in R_i}\exp\{\Phi_i^*\cdot f_r/\tau\}}\right). \tag{4}$$

为充分利用监督信号，交叉熵损失定义为

$$L_{\mathrm{CE}}=\frac{1}{N}\sum_{i\in I}\left(-\log\frac{\exp\{\Phi_i^*\cdot f_i/\tau\}}{\sum_{r\in R_i}\exp\{\Phi_i^*\cdot f_i/\tau\}}\right). \tag{5}$$

在进行隐藏空间的插值时，为最小化原始数据与混合后增强数据之间的差异，并解决梯度消失问题，使用 Wasserstein 距离作为一致正则项与双重对比损失和交叉熵损失一起训练. Wasserstein 距离定义为

$$W(P_1,P_2)=\min_{\gamma\sim\Pi(P_1,P_2)}E_{(x,y)\sim\gamma}(\|x-y\|), \tag{6}$$

其中 $\Pi(P_1,P_2)$为 P_1 和 P_2 分布组合所有可能的联合分布的集合. 对每个可能的联合分布 γ，可以从中采样$(x,y)\sim\gamma$得到样本 x 和 y，并计算出这对样本的距离 $\|x-y\|$，所以可以计算该联合分布 γ 下，样本对距离的期望值 $\min\limits_{\gamma\sim\Pi(P_1,P_2)}E_{(x,y)\sim\gamma}(\|x-y\|)$. 在所有可能的联合分布中能对该期望值取得的下界即为 Wasserstein 距离.

Wasserstein 距离相比 KL(Kullback-Leibler)散度和 JSD 散度的优势：即使两个分布的支撑集没有重叠或重叠非常少，仍能反映两个分布的远近. 而 JSD 散度在这种情况下是常量，KL 散度可能无意义，其定义为

$$L_{\mathrm{Wass}}=W(p_{\mathrm{mix}},\overline{p})+W(p_{\mathrm{ori}},\overline{p}), \tag{7}$$

其中 p_{mix}，p_{ori}分别表示增强数据的概率和原始数据的预测概率，$\overline{p}=\frac{1}{2}(p_{\mathrm{mix}}+p_{\mathrm{ori}})$. 因此，总体损失函数为

$$L_{\mathrm{ALL}}=L_{\mathrm{CE}}+\xi L_{\mathrm{DCL}}+\zeta L_{\mathrm{Wass}}. \tag{8}$$

2　实　验

2.1　实验数据集

为充分评估本文模型在文本分类任务中的性能，实验选择不同大小的基准数据集进行验证. SST-2[18] 是美国斯坦福大学情感分析电影评论数据集，可预测消极和积极情绪；数据集 SUBJ[19]将电影评论分为主观评论和客观评论；CR[20]是客户评论数据集，其中评论被分类为积极和消极；数据集 TREC[21]是 6 个不同领域的六分类问题，包括描述、实体、缩写、人、位置和数字. PC[22]是一个情绪数据集，包含了正反两种情绪. 各数据集的统计信息列于表 1.

表 1　实验采用的数据集信息

Table 1　Dataset information used for experiment

数据集	类别	平均长度	训练集	测试集	数据集	类别	平均长度	训练集	测试集
SST-2	2	17	67 349	1 821	TREC	6	9	5 452	500
SUBJ	2	21	9 000	1 000	PC	2	7	32 097	13 759
CR	2	18	3 394	376					

2.2　参数设置

使用 BERT-base-uncase 作为微调模型，隐藏层尺寸为 768 维. 考虑过滤器的大小会影响实验结果，本文将 TextCNN 过滤器的大小设为[2,3,4]，使用二维卷积. 随机选取训练集的 20%作为验证

集. 本文选取在测试集上的准确率作为评价指标. 实验使用的主要配置参数为 Epoch=10，优化器为 Adam，最大层数设置为[9,10,12]，β 插值为 0.75，基线模型学习率为 1×10^{-5}，分类器学习率为 0.01，$\xi=0.5$，$\zeta=5$，Batch_size=32.

2.3 对比基线

将本文模型与 BERT 融合经典数据增强分类方法和双重对比学习文本分类方法进行比较(配置设备和数据集均相同)，以验证本文分类模型的有效性.

BERT+CE[12]：BERT 模型被认为是 NLP 里程碑式的进步. BERT+DualCL[14]：DualCL 是一种双对比度损失文本分类模型，利用标记数据增强，是一种有监督的对比度损失. BERT+EDA+DualCL[6]：EDA 是一种简单的记号级数据扩充方法，通过对原始样本执行同义词替换、随机插入、随机删除和随机交换 4 种方式扩充数据. BERT+TMix+DualCL[9]：挖掘未标记数据与已标记数据之间的隐藏关系，并将未标记数据应用于已标记数据上，通过对隐藏空间中的不同训练样本进行线性插值，生成大量新的训练数据. BERT+SSMix[10]：在输入层对原始样本进行增强运算，而不是对隐藏空间中的隐藏向量进行增强运算；SSMix 通过基于广度的 Mixup 保持两个原始文本的局部性，并根据显著性信息保留更多与预测相关的标记. BERT+DoubleMix+DualCL[5]：一种简单的基于插值的数据增强方法，首先将扰动数据混合到合成样本中，然后将原始数据与扰动数据混合；DoubleMix 通过学习隐藏空间中的“移位”特征增强模型的鲁棒性.

3 实验结果分析

在 4 种常用的文本分类任务数据集上评估本文方法，并展示在低资源场景下基于关键词的数据增强性能. 为更好地说明有监督对比学习与数据增强的优越性，本文验证了在低资源场景下分类的效果，并设计了消融实验验证各模块的作用.

3.1 实验结果

表 2 列出了本文分类模型与基线模型在 4 个文本分类数据集上的对比结果. 本文还将其与常用的数据增强方法和双重监督对比损失结合训练进行比较. 由表 2 可见，虽然相对于仅用交叉熵损失训练的 BERT，所有的文本增强方法都能提高准确性，但本文模型的准确率最高. 与交叉熵训练的 BERT 模型相比，本文方法的平均改进率为 1.32%. 引入双重对比度损失可改善仅使用交叉熵训练的 BERT 模型鲁棒性低的缺点. 与仅使用双重对比损失训练相比，本文方法的平均改良率为 0.76%. 这是因为在对样本进行双层数据增强后，增加了样本的多样性. 通过 BERT 和 TextCNN 的共同抽取，可得到更有利于文本分类的句子向量和特征向量. 与 DoubleMix 算法相比，本文方法的平均改善率为 0.70%. 基于关键词的数据增强生成扰动样本的方式对文本进行选择性的补充，以保证生成样本的质量. 使用双重对比度损失和 Wasserstein 距离对模型进行训练，在保证梯度不消失的情况下最小化原始数据与增强数据之间的差距，使提取的特征表示更紧凑，更有利于文本分类任务.

表 2 本文算法与其他数据增强算法性能对比结果

Table 2 Performance comparison results of proposed algorithm with other data augmentation algorithms

算法	分类精度/%			
	SST-2	CR	TREC	PC
BERT+CE	91.03	91.83	96.90	94.79
BERT+DualCL	91.87	92.42	97.40	95.10
BERT+EDA+DualCL	91.98	92.50	97.30	95.15
BERT+TMix+DualCL	91.60	91.89	96.97	94.90
BERT+SSmix	91.45	92.30	97.40	95.00
BERT+DoubleMix+DualCL	92.20	92.58	97.40	95.10
本文	93.41	93.55	97.61	95.27

为考察本文方法是否能产生更好的特征表示，使用 DoubleMix 与本文方法进行训练，在 PC 训练数据集($N=12\ 000$)上绘制了学习表征的 tSNE 图，如图 3 所示. 由图 3 可见，本文方法可以学习到的

每一类别内样本的特征表示更接近，从而更有利于模型分类.

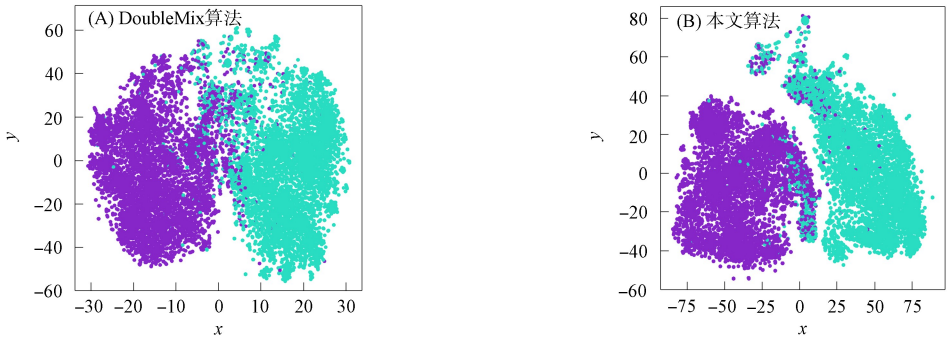

图 3　在 PC 训练数据集中学习到的表征 tSNE 图

Fig. 3　tSNE plots of learned representations on PC training dataset

3.2　基于关键词的数据增强性能分析

为验证基于关键词数据增强方法的有效性，对低资源条件下基于关键词的数据增强的文本分类性能进行评估. 基线是无数据增强、EDA 方法和 STA 数据增强方法. 数据集为 SST-2，CR 和 SUBJ，样本数量 $N=100$. 使用 DistilBERT-base，从 BERT 中提取的轻量级转换器模型作为文本分类器的主干. 图 4 为不同算法在低资源场景下对不同数据集的测试精度. 由图 4 可见，在数据量相对少的情况下，本文算法比基于 token 级别的数据增强算法平均准确率提高了 0.8 个百分点. 由于 EDA 在扩充句子时未考虑关键词的影响，可能会删除与分类任务相关的关键词信息. STA 基于 4 个角色词扩展数据，为使用静态提取关键词的算法，因此提取角色词所用的时间较长，不易在短时间内生成大量样本. 表 3 列出了本文算法提取关键词和 STA 提取角色词在不同数据集上所用的时间. 由表 3 可见，本文方法在保证分类结果基本一致的情况下，提取关键词的时间较短.

表 3　本文算法提取关键词和 STA 提取角色词在不同数据集上所用的时间

Table 3　Time spent in keyword extraction using proposed method and role words extraction by STA on different datasets

算法	t/s		
	SST-2	CR	SUBJ
STA-extract keyword	30.9	34.8	39.2
本文	4.9	4.3	5.3

3.3　低资源条件下性能分析

为证明本文模型在低资源场景下可以取得更好的效果，实验选择 SST-2，CR 和 PC 3 个数据集进行验证. 分类数据大小分别为 $N=100$ 和 $N=500$. 表 4 列出了低资源场景下不同算法在数据集 CR 上的测试精度. 由表 4 可见，本文算法优于 BERT＋Dual 和 BERT＋DualCL＋DoubleMix. 图 5 为不同算法在数据集 CR 上当 $N=\{10,20,30,40,50,60,70,80,90,100\}$ 时分类准确率的折线图. 由图 5 可

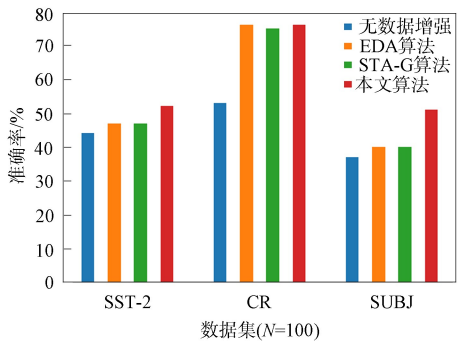

图 4　不同算法在低资源场景下对不同数据集的测试精度

Fig. 4　Testing accuracy of different algorithms on different datasets in low resource scenarios

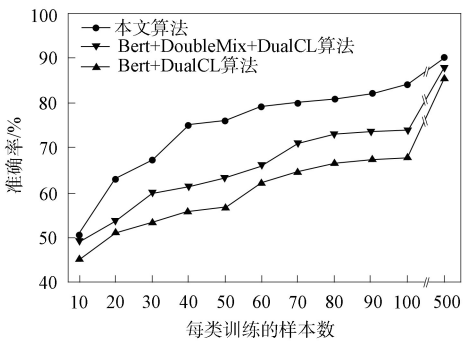

图 5　低资源场景下不同算法在数据集 CR 上的测试精度对比

Fig. 5　Comparison of testing accuracy of different algorithms on CR dataset in low resource scenarios

见，随着输入样本数量的增加，本文算法分类准确率始终是最高的. 当数据增强量达到500后，其分类准确率提高速度缓慢. 因此，将数据增强服务于有监督对比学习可有效提高低资源条件下文本分类任务的准确率.

表4 低资源场景下不同算法在数据集CR上的测试精度

Table 4 Testing accuracy of different algorithms on CR dataset in low resource scenarios %

算法	$N=100$			$N=500$		
	SST-2	CR	PC	SST-2	CR	PC
BERT+CE	60.30	67.65	62.50	69.57	85.57	92.32
BERT+DualCL	53.30	64.71	75.03	63.04	86.60	92.90
BERT+DualCL+DoubleMix	53.30	73.53	76.50	80.43	87.63	93.40
本文	76.00	88.64	78.94	86.71	90.18	94.44

3.4 消融实验

为验证每个模块都能对本文模型发挥作用，本文进行了消融实验. 实验在数据集SST-2上进行. 实验设置：无输入层基于关键词的数据增强，无TextCNN特征提取，无隐藏层插值(训练时不使用Wasserstein距离)，训练时不加入双重对比损失. 实验结果列于表5.

表5 模型的不同模块在完整训练样本的数据集SST-2上测试的准确率

Table 5 Testing accuracy on SST-2 dataset with complete training samples of different modules of model %

网络结构	准确率	网络结构	准确率
无输入层增强	92.55	无DualCL训练	91.89
无TextCNN提取	92.70	所有模块	93.41
无隐藏层插值与Wasserstein距离训练	92.03		

由表5可见，改变实验策略后，模型性能有所下降. 表明模型的每一部分模块都影响模型的训练结果. 首先移除输入层数据增强模块，即使用原始数据本身在隐藏层进行插值增强，模型的分类准确率下降0.86个百分点. 然后，去掉TextCNN信息提取模块，此时分类准确率下降0.71个百分点，结果表明，与直接使用BERT向量表示相比，将得到的每一层的向量表示馈送到TextCNN中会有积极效果. 模型训练时不使用隐藏层插值(无Wasserstein距离)及训练时不加入双重对比损失，平均准确率降低1.45个百分点. 因此，模型中每个模块在文本分类任务中都具有积极作用.

综上所述，针对DoubleMix现有算法存在的问题，本文提出了一种基于双层数据增强的监督对比文本分类模型. 该方法通过基于关键词的数据增强对数据进行更有效、简便地扩充，并使用新的训练方式提取到更利于文本分类的特征表示. 在常用的文本分类数据集上进行验证，实验结果表明该方法能在提高样本多样性的同时学习到紧凑的特征表示，最终提高文本分类的准确率，尤其是低资源情况下的文本分类. 消融实验也说明了模型的每一部分都起到了积极作用. 因此，该模型通过对数据进行有效地数据增强并对模型更好地训练，有效提高了模型的分类能力.

参 考 文 献

[1] 高云龙，吴川，朱明. 基于改进卷积神经网络的短文本分类模型[J]. 吉林大学学报(理学版)，2020，58(4)：923-930.(GAO Y L，WU C，ZHU M. Short Text Classification Model Based on Improved Convolutional Neural Network[J]. Journal of Jilin University (Sicence Edition)，2020，58(4)：923-930.)

[2] 王进，徐巍，丁一，等. 基于图嵌入和区域注意力的多标签文本分类[J]. 江苏大学学报(自然科学版)，2022，43(3)：310-318.(WANG J，XU W，DING Y，et al. Multi-label Text Classification Based on Graph Embedding and Regional Attention[J]. Journal of Jiangsu University (Natural Science Edition)，2022，43(3)：310-318.)

[3] 车颖，冯晶，郑宏亮. 基于卷积神经网络的超声造影图像去噪方法[J]. 吉林大学学报(理学版)，2021，59(5)：1256-1259.(CHE Y，FENG X，ZHENG H L. Ultrasonography Image Denoising Method Based on Convolutional Neural Network[J]. Journal of Jilin University (Science Edition)，2021，59(5)：1256-1259.)

[4] 王进，陈重元，邓欣，等. 多状态图神经网络文本分类算法[J]. 重庆邮电大学学报(自然科学版)，2023，35(2)：193-201.(WANG J，CHEN C Y，DENG X，et al. Multi-state Graph Neural Network Text Classification

Algorithm [J]. Journal of Chongqing University of Posts & Telecommunications (Natural Science Edition), 2023, 35(2): 193-201.)

[5] CHEN H, HAN W, YANG D Y, et al. DoubleMix: Simple Interpolation-Based Data Augmentation for Text Classification [C]//Proceedings of the 29th International Conference on Computational Linguistics. [S.l.]: ACL, 2022: 4622-4632.

[6] WEI J, ZOU K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks [C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. [S.l.]: ACL, 2019: 6382-6388.

[7] SERGEY E, MYLE O, MICHAEL A, et al. Understanding Back-Translation at Scale [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. [S.l.]: ACL, 2018: 489-500.

[8] YIN W P, WANG H, QU J, et al. BatchMixup: Improving Training by Interpolating Hidden States of the Entire Mini-batch [C]//Findings of the Association for Computational Linguistics. [S.l.]: ACL, 2021: 4908-4912.

[9] CHEN J A, YANG Z C, YANG D Y. MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-supervised Text Classification [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [S.l.]: ACL, 2020: 2147-2157.

[10] YOON S Y, KIM G, PARK K. SSMix: Saliency-Based Span Mixup for Text Classification [C]//Findings of the Association for Computational Linguistics. [S.l.]: ACL, 2021: 3225-3234.

[11] ZHANG Z L, MERT R S. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels [EB/OL]. (2018-05-20)[2023-02-10]. https://arxiv.org/abs/1805.07836.

[12] JACOB D, CHANG M W, KENTON L, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S.l.]: ACL, 2019: 4171-4186.

[13] YOON K. Convolutional Neural Networks for Sentence Classification [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.]: ACL, 2014: 1746-1751.

[14] CHEN Q B, ZHANG R C, ZHENG Y W, et al. Dual Contrastive Learning: Text Classification via Label-Aware Data Augmentation [EB/OL]. (2022-01-21)[2023-02-10]. https://arxiv.org/abs/2201.08702.

[15] GUO B Y, HAN S Q, HUANG H L. Selective Text Augmentation with Word Roles for Low-Resource Text Classification [EB/OL]. (2022-09-04)[2023-01-01]. https://arxiv.org/abs/2209.01560.

[16] XIONG Y J, FENG Y K, WU H, et al. Fusing Label Embedding into BERT: An Efficient Improvement for Text Classification [C]//Findings of the Association for Computational Linguistics. [S.l.]: ACL, 2021: 1743-1750.

[17] GANESH J, BENOIT S, DJAME S. What Does BERT Learn about the Structure of Language [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. [S.l.]: ACL, 2019: 3651-3657.

[18] RICHARD S, JOHN B, CHRISTOPHER D, et al. Parsing with Compositional Vector Grammars [C]// Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. [S.l.]: ACL, 2013: 455-465.

[19] BO P, LILLIAN A. Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts [C]//Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. [S.l.]: ACL, 2004: 271-278.

[20] DING X W, LIU B, YU P S. A Holistic Lexicon-Based Approach to Opinion Mining [C]//Proceedings of the 2008 International Conference on Web Search and Data Mining. New York: ACM, 2008: 231-240.

[21] LI X, ROTH D. Learning Question Classifiers [C]//The 19th International Conference on Computational Linguistics. New York: ACM, 2022: 1-7.

[22] MURTHY G, LIU B. Mining Opinions in Comparative Sentences [C]//Proceedings of the 22nd International Conference on Computational Linguistics. [S.l.]: ACL, 2008: 241-248.

(责任编辑：韩　啸)