软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(10):4794−4805 [doi: 10.13328/j.cnki.jos.006999]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于监督对比学习的文本情绪类别表示*

王祥宇[1], 宗成庆[1,2]

[1](模式识别国家重点实验室 (中国科学院 自动化研究所), 北京 100190)

[2](中国科学院大学 人工智能学院, 北京 100049)

通信作者: 宗成庆, E-mail: cqzong@nlpr.ia.ac.cn

摘　要: 揭示情绪之间的关系是认知心理学的一项重要基础研究. 从自然语言处理的角度来说, 探讨情绪之间的关系的关键在于得到合适的情绪类别的嵌入式表示. 最近, 在情感空间中获得一个可以表征情绪关系的类别表示已经引起了一些关注. 然而, 现有的情绪类别嵌入方法存在以下几个缺点. 比如固定维度, 情绪类别表示的维度依赖于所选定的数据集. 为了取得一个更好的情绪类别表示, 引入监督对比学习的表示方法. 在之前的监督对比学习方法中, 样本之间的相似性取决于样本所标注的标签的相似性. 为了更好地反映出不同情绪类别之间的复杂关系, 进一步提出部分相似的监督对比学习表示方法, 认为不同情绪类别 (比如情绪 anger 和 annoyance) 的样本之间也可能是部分相似的. 最后, 组织一系列实验来验证所提方法以及其他 5 个基准方法在表述情绪类别之间关系的能力. 实验结果表明, 所提方法取得了理想的情绪类别表示结果.

关键词: 情感分析; 情绪表示; 情绪空间; 情绪类别

中图法分类号: TP18

中文引用格式: 王祥宇, 宗成庆. 基于监督对比学习的文本情绪类别表示. 软件学报, 2024, 35(10): 4794–4805. http://www.jos.org.cn/1000-9825/6999.htm

英文引用格式: Wang XY, Zong CQ. Supervised Contrastive Learning for Text Emotion Category Representations. Ruan Jian Xue Bao/Journal of Software, 2024, 35(10): 4794–4805 (in Chinese). http://www.jos.org.cn/1000-9825/6999.htm

Supervised Contrastive Learning for Text Emotion Category Representations

WANG Xiang-Yu[1], ZONG Cheng-Qing[1,2]

[1](National Laboratory of Pattern Recognition (Institute of Automation, Chinese Academy of Sciences), Beijing 100190, China)

[2](School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Revealing the complex relations among emotions is an important fundamental study in cognitive psychology. From the perspective of natural language processing, the key to exploring the relations among emotions lies in the embedded representation of emotional categories. Recently, there has been some interest in obtaining a category representation in the emotion space that can characterize emotion relations. However, the existing methods for emotion category representations have several drawbacks. For example, fixed dimensionality, the dimensionality of the emotion category representation, depends on the selected dataset. In order to obtain better representations for the emotion categories, this study introduces a supervised contrastive learning representation method. In the previous supervised contrastive learning, the similarity between samples depends on the similarity of the annotated labels of the samples. In order to better reflect the complex relations among different emotion categories, the study further proposes a partially similar supervised contrastive learning representation method, which suggests that samples of different emotion categories (e.g., anger and annoyance) may also be partially similar to each other. Finally, the study organizes a series of experiments to verify the ability of the proposed method and the other five benchmark methods in representing the relationship between emotion categories. The experimental results show that the proposed method achieves satisfactory results for the emotion category representations.

Key words: sentiment analysis; emotion representation; emotion space; emotion category

* 收稿时间: 2022-12-12; 修改时间: 2023-03-06, 2023-05-29; 采用时间: 2023-07-09; jos 在线出版时间: 2023-09-27
CNKI 网络首发时间: 2023-10-07

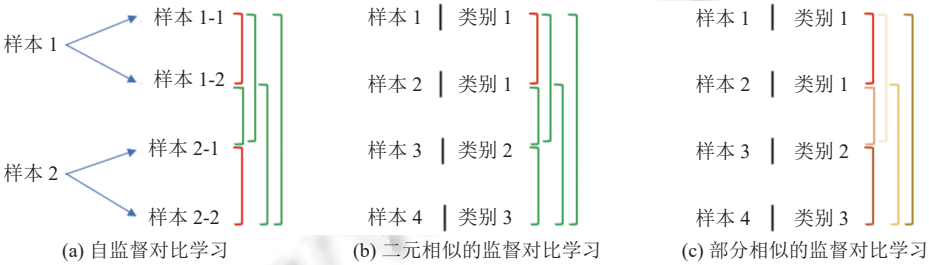
情感分析是自然语言处理领域的一项重要研究课题. 情绪分类任务作为其中最基础的研究问题[1], 自提出以来就得到了广泛的应用和发展. 基于不同的研究场景, 研究人员提出了很多更加契合环境的情感分析任务. Lee 等人[2]提出了情绪原因抽取任务, 旨在从文本中抽取出给定情绪对应的原因. 他们基于语言学规则开发出了一个基于规则的情绪原因检测系统, 并提出了一个两阶段的评价体系. 为了更加贴合实际应用场景, Xia 等人[3]提出了情绪原因对抽取任务, 旨在从文本中同时抽取出潜在的所有情绪及其对应的原因. Mohammad 等人[4]提出了情绪强度检测任务, 旨在预测文本中给定情绪的强度值. 他们同时构建了 4 类情绪的强度数据集. 为了获取文本中更细粒度的情绪, Jiang 等人[5]提出了目标依赖的推特情感分类任务.

为了在目标任务上取得更好的效果, 研究人员提出了大量的数据集[6−8]及具有针对性的模型[9−11]. 然而, 上述所讨论的情感分析任务, 数据集和模型都将不同的情绪看作是独立的维度, 并用元向量来表示这些情绪. 这种表示方法认为不同情绪类别之间是彼此正交的. 实际上, 不同情绪类别之间的边界并不清晰, 正交的表示方法与情绪类别之间错综复杂的关系并不吻合.

值得说明的是, 情绪类别表示在多个领域都有着重要的应用. 得到一个合理的情绪类别表示对于多个领域的后续研究也有着积极的意义. 比如在语言学领域, 很多人类语言都有着类似于“开心”和“悲伤”等描述情感的词汇. Jackson 等人[12]从语言学的角度探讨了不同语言下情感词汇以及情绪结构的差异和原因. 而一个合适的情绪类别表示可以实现从自然语言处理的角度定量地去分析情绪结构的跨语言差异. 再比如在文本分类领域, 探讨类别之间的关系可以有效改善模型的过自信等缺陷[13,14], 而学习一个合适的类别表示可以更好地去探讨类别之间的关系.

为了得到情绪空间下各情绪类别的分布式表示, Wang 等人[15]提出了一种基于软标签的学习方法. 他们将训练好的神经网络模型输出的软标签看作是输入样本在情绪空间中的分布式表示. 并进一步基于聚类的思想求解出情绪类别的表示. 然而, 这种表示方法仍然存在着一些问题. (1) 维度受限: 模型输出的软标签的维度和类别数的维度是一致的. 而类别数的维度又取决于特定的数据集. (2) 各向异性[16]: 对于模型输出的软标签, 每个维度上的值介于 0–1. 这意味着该方法得到的类别表示的覆盖范围为空间中的超立方体.

为了解决上述问题, 得到一个维度可以预定义且在空间中分布是各向同性的表示, 我们在本文中引入了对比表示学习的方法. 特别地, 考虑到情绪类别之间关系特殊性, 即不同情绪之间并非是完全独立的, 我们针对性地提出部分相似的监督对比学习方法. 在相似与不相似关系之外, 我们认为不同类别样本之间的关系也可能是部分相似的. 图 1 显示了自监督对比学习, 二元相似的监督对比学习和我们的方法之间的关系. 红线表示样本之间的关系是相似的, 绿线表示不相似关系, 黄线表示部分相似关系, 黄色的深度表示相似关系的大小. 对于自监督对比学习, 源于相同样本的两个增强样本是相似的, 源于不同样本的增强样本是不相似的. 对于二元相似的监督对比学习, 标注为相同类别的样本之间是相似的, 而标注为不同类别的样本之间是不相似的. 对于部分相似的监督对比学习, 标注为相同类别的样本之间是相似的, 标注为不同类别的样本是部分相似的.



(a) 自监督对比学习 (b) 二元相似的监督对比学习 (c) 部分相似的监督对比学习

图 1 对比学习中样本相似关系示意图

本文第 1 节介绍情绪表示理论模型, 情绪嵌入模型和对比学习的研究内容和发展现状. 第 2 节具体介绍我们的方法: 包括方法的出发点, 损失函数的构建, 表示模型的选取和最终情绪类别表示的计算方法. 第 3 节介绍我们的方法在 GoEmotions 数据集上的实验结果. 最后总结全文.

1 相关工作

1.1 情绪表示理论模型

情绪表示理论模型主要可以划分为情绪类别模型和情绪维度模型两大类. Ekman 基本情绪理论[17]是目前应用最广泛的情绪类别表示模型. 该文将人类情绪状态划分为 6 个独立地情绪类别 (anger, disgust, fear, joy, sadness 和 surprise). Plutchik 圆盘模型[18]是另外一个重要的基础情绪理论模型. Plutchik 将人类情绪划分为圆盘上的八组. 每组由初级, 次级和第 3 级情绪组成. 通过分析自我报告的情绪体验的种类, Cowen 等人[19,20]将人类情感划分为 27 个细粒度的情绪类别. 他们证明了所有的情感从根本上来说都是相同的, 并形成了统一的连续统一体. 另外一类维度模型是用预先定义的若干个属性 (通常是 2–4 个) 来定量表示人类的情感状态. VAD 模型[21−23]是使用最为广泛的维度模型. 在 VAD 模型中, 人类情绪状态被 3 个维度定量的描述: valence (正面-负面), arousal (冷静-激动) 和 dominance (可控-服从). 然而, 维度模型的属性难以定义, 且由于其属性值的连续性而难以标注. 因此, 类别模型在情感分析的下游任务中得到了更加广泛地引用.

1.2 情绪嵌入模型

“情绪嵌入”并不是一个新鲜的词汇, 在不同的工作中可能有着不同的含义. 为了增强词嵌入中的情感信息, Xu 等人[24]在不同情感相关任务地模型之间共享嵌入层来学习情绪的向量化表示. 同样地, Wang 等人[25]提出了一种基于向量情感值的相似度得重排序方法, 并称之为“情绪嵌入”. 然而上述两种方法都是在词向量空间中从不同角度尝试给予词汇表示更多的情感信息, 以适应于不同的下游任务. 在多模态领域, Han 等人[26]聚焦于学习一个共享的嵌入空间, 以更好地融合不同模态下的信息. 在多语言领域, Buechel 等人[27]提出了一个多语言共享的情绪识别模型, 将不同语言的情绪数据集融合到一个空间中进行表征. 然而上述两种方法都是从不同角度去融合不同领域的信息 (模态或者语言) 到同一个特征表示空间中去. Wang 等人[16]提出了一种学习情绪类别表示的方法, 并将情绪类别看作是情绪空间中文本情绪状态的聚类中心, 进而将情绪空间中情绪类别和文本情绪状态视为情绪嵌入.

1.3 对比学习

近年来, 基于深度学习模型的自监督表示方法在自然语言处理领域[28,29]和计算机视觉领域[30,31]都得到了广泛的应用. 其中, 自监督对比学习自提出以来得到了巨大的发展, 并在多个领域取得了先进的效果[32]. 自监督对比学习方法通常学习一个判别模型来判断不同样本对表示之间的相似性. 自监督对比学习通常使用一个正样本对和多个负样本对来构建损失函数, 其中正样本对可以是来自不同样本[33,34], 也可以是基于数据增强的同一样本的不同形式[35]. 在自然语言处理领域, Gao 等人[36]提出了一个简单高效的框架以学习句子的嵌入表示. 对于同一个句子, 他们对句子的嵌入表示做随即丢弃, 并参照 Chen 等人[35]的思想构建了无监督条件的正负样本对.

当数据集提供了额外的标签时, 这些标签也可以被整合进既有的对比框架中去. Khosla 等人[37]提出了监督对比学习. 他们将对比学习的框架从自监督表示扩展到监督表示, 进而可以充分利用数据集中的标签信息. 在他们的工作中, 属于同一类别的样本对应的嵌入空间中的点将被拉近. 反之, 属于不同类别的样本在嵌入空间中将被拉地更远. Gunel 等人[38]将监督对比学习引入预训练模型领域, 通过修改训练的目标损失函数使得预训练模型在小样本领域取得了更好的效果.

2 本文方法

在本文中, 我们提出了部分相似的对比学习方法. 下面我们对本文方法进行详细地介绍.

2.1 部分相似

在之前的对比学习方法中, 两个样本只被区分为相似或者不相似. 但是在很多场景下, 相似或者不相似的二元描述不足以表达样本之间的相似关系. 比如在情绪关系中, 两个样本所蕴含的情绪之间的关系可能是多样的. 如表 1 所示, 第 1 对示例的两个样本都被标注为开心类别, 因而认为这两个样本所蕴含的情绪是相似的. 对于第 2 对示例, 一个样本被标注为开心类别而另一个样本被标注为悲伤类别. 开心和悲伤是两个截然不同的情绪, 一般认为

这两个情绪是两个独立的维度. 因而认为这两个样本所蕴含的情绪是不相似的. 对于第 3 对示例, 尽管两个样本被标注为不同的类别 (开心和支持), 但是这两个类别都是正面情绪, 他们之间的相似性应该介于第 1 对示例和第 2 对示例. 在本文中, 我们对部分相似概念做进一步简化处理. 当且仅当两个样本的标签同来自于正面情绪 (或者同来自负面情绪, 中性情绪) 时, 我们认为这两个样本所蕴含的情绪是部分相似的.

表 1　相似关系示例

样例1 (文本|标签)	样例2 (文本|标签)	样例1与样例2的相似关系
好的天气真让人心情愉悦!|开心	努力工作的感觉可真好.|开心	相似
好的天气真让人心情愉悦!|开心	哎, 我点的外卖又丢了.|悲伤	不相似
好的天气真让人心情愉悦!|开心	我相信皇马会赢下这场比赛的!|支持	部分相似

2.2　损失函数

我们在本节中详细地讨论从自监督领域到监督领域的对比损失. 我们用 $B=\{x_k,y_k\}_{k=1,2,\cdots,N}$ 表示批数据中的 N 个随机采样的样本. $I=\{x'_k,y'_k\}_{k=1,2,\cdots,2N}$ 是对应的 $2N$ 条数据, 其中, x'_{2k-1} 和 x'_{2k} 是从初始样例 x_k 中生成的两条增强样例.

在自监督对比学习中, 数据集中的标签是缺失的 (比如 B 中的 y 以及 I 中的 y'). 在增强的批数据 I 中, 对于样例 x'_{2k-1} 来说, 唯一的正例来自于同一个原始样本生成的样例 x'_{2k} . 自监督对比学习的损失函数可以写成如下形式:

$$\mathcal{L}^{\text{self}}=\sum_{i\in I}\mathcal{L}_i^{\text{self}}=-\sum_{i\in I}\log\frac{\exp\left(f_i\cdot\frac{f_{p(i)}}{T}\right)}{\sum\limits_{k\in I\setminus\{i\}}\exp\left(f_i\cdot\frac{f_k}{T}\right)} \tag{1}$$

其中, f_i 表示第 i 个样例的特征表示; 点“·”表示两个矢量之间的点积; 参数 T 表示温度, 更大的 T 值使得概率分布更加均匀. $p(i)$ 表示批数据中第 i 个样本对应的唯一的正例, 其余的 $2N-1$ 个样本则是负例.

在监督数据集中, 为了更好地使用数据集中标签的信息, 两个被标注为相同类别的样本被认为是正样本对, 来自于不同标签的两个样本则被认为是负样本. 因此, 在监督对比学习中, 对于批数据中的一个样本, 可能存在多个正例. 监督对比学习的损失函数可以写成如下形式:

$$\mathcal{L}^{\text{sup}}=\sum_{i\in I}\mathcal{L}_i^{\text{sup}}=-\sum_{i\in I}\sum_{p\in P(i)}\frac{1}{|P(i)|}\log\frac{\exp\left(f_i\cdot\frac{f_p}{T}\right)}{\sum\limits_{k\in I\setminus\{i\}}\exp\left(f_i\cdot\frac{f_k}{T}\right)} \tag{2}$$

其中, $P(i)$ 表示批数据中第 i 个样例对应的正例的集和.

特别地, 如果不对原始的数据集中的样本进行数据增强, 而是直接采用原始的数据集. 那么不使用数据增强的监督对比学习函数可以写成如下形式:

$$\mathcal{L}^{\text{sup}}=\sum_{i\in B}\mathcal{L}_i^{\text{sup}}=-\sum_{i\in B}\sum_{p\in P(i)}\frac{1}{|P(i)|}\log\frac{\exp\left(f_i\cdot\frac{f_p}{T}\right)}{\sum\limits_{k\in B\setminus\{i\}}\exp\left(f_i\cdot\frac{f_k}{T}\right)} \tag{3}$$

在上述的损失函数中, 相似的样本对的权重被设置为 1, 而不相似的样本对的权重被设置为 0. 然而情绪类别之间的关系是错综复杂的. 不同类别的情绪之间也可能有着一定的相似性. 为了让损失函数更加贴近标签之间的相关性, 我们提出部分相似的概念. 我们使用 $\lambda(i,j)$ 来表示第 i 个样例和第 j 个样例之间的相似性.

$$\mathcal{L}^{\text{our}}=\sum_{i\in B}\mathcal{L}_i^{\text{our}}=-\sum_{i\in B}\sum_{p\in B(i)}\frac{\lambda(i,p)}{\sum\limits_{j\in B(i)}\lambda(i,j)}\log\frac{\exp\left(f_i\cdot\frac{f_p}{T}\right)}{\sum\limits_{k\in B\setminus\{i\}}\exp\left(f_i\cdot\frac{f_k}{T}\right)} \tag{4}$$

然而, 准确地得到任意两个样本之间的相似性是困难的. 在本文中, 我们认为标注为相同的标签的两个样本是相似的, 同标注为正面情绪 (或负面情绪) 的两个样本是部分相似的, 标注为正面情绪的样本和标注为负面情绪的样本是不相似的. 即:

$$\lambda = \begin{cases} 1, \text{ if similar} \\ 0, \text{ if dissimilar} \\ t, \text{ if partly similar} \end{cases} \tag{5}$$

在本文中, 我们设置 t 为 0.1. 即给同为正面情绪 (或负面情绪) 的样本之间一个较小的相似性. 更一般性地根据不同的类别先验地设置不同的相似性值, 我们留待未来工作完成.

在既往的二元相似关系的监督对比学习中, 不同情绪类别之间只能是相似或者不相似的关系. 而在本文中, 对于标注为相同标签的两个样本, 我们认为它们是相似的. 对于标注同为正面情绪 (或者负面情绪) 但不同类别的样本, 我们认为它们是部分相似的. 比起基于交叉熵或者基于二元监督对比学习的损失函数, 本文事实上在损失函数构建中像模型引入了更多的信息.

2.3 特征表示模型

本文采用的特征表示模型如图 2 所示. D 表示模型输出的特征维度数, 在本文中等于 64. 最后一步采用的正则化是 L2 正则化. 输入一批数据到模型中去, 模型会输出对应的特征表示 $f_{b\times D}$. 我们对模型输出的每个特征都进行了 L2 正则化. 因此, $f_{b\times D}$ 每一行的 L2 范数都是 1.

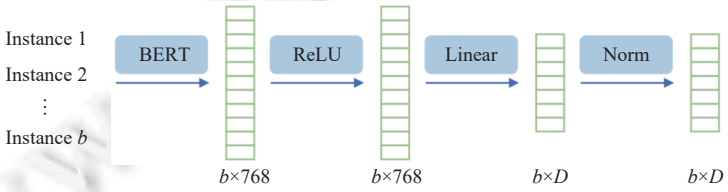


图 2 本文方法选用的特征表示模型

2.4 计算情绪类别的嵌入表示

在模型训练完毕以后, 我们将测试集所有数据输入到模型中去并得到对应的输出. 我们假定测试集有 N_t 条数据, D 为每条数据对应的特征维度. 对于 BERT-base 模型, D 等于 768. 我们使用 $f_{N_t\times D}$ 来表示 N_t 条数据对应的模型输出. 在本文中, 我们将模型输出的特征表示看作是对应样例在情绪空间中的表示.

我们进一步定义 $L_{N_t\times C}$ 为测试数据集对应的标签结果. 其中 C 是数据集中标注的情绪类别的数量. $L_{N_t\times C}$ 中的每一行之和为 1. 比如, 对于 C 等于 5 的数据集, 第 i 条样本的标注类别为 $(0,1,0,0,1)$, 那么 L_i (也就是 $L_{N\times C}$ 的第 i 行) 就是 $(0,0.5,0,0,0.5)$. 为了得到情绪类别的嵌入表示, 我们首先计算表示矩阵 $E_{C\times D}$.

$$E_{C\times D} = L^{\mathrm{T}}_{N_t\times C} f_{N_t\times D} \tag{6}$$

其中, $E_{C\times D} = [e_1; e_2; \ldots; e_C]$. 由于不同情绪类别对应标注样本的数量并不相同, $E_{C\times D}$ 中标注数据更多的类别对应的行的向量长度也会更大. 因此, 我们需要对 $E_{C\times D}$ 的每一行做归一化处理.

$$e_i = \frac{e_i}{\| e_i \|_2}, \ i = 1,2,\ldots,C \tag{7}$$

其中, e_i 就是我们最终得到的每 i 个情绪类别对应的嵌入表示.

3 实验分析

3.1 数据集

本实验基于 GoEmotions 数据集[39]进行验证. 我们选取 GoEmotions 计算出对应的情绪类别的分布式表示, 并

进一步验证这些类别表示的内在质量. 该数据集包含两个版本: 初始版本和精选版本. 其中初始版本为标注者初始的标注结果, 未进行交叉验证, 数据量较大但含有较多的噪声. 精选版本为经过后续处理的标注数据, 经过交叉验证, 质量较高. 本文选取的是精选版本. GoEmotions 包含 27 类情绪类别. 该数据集作者将 27 类情绪划分为正面情绪, 负面情绪和中性情绪. 详细的情绪类别及划分结果参见表 2.

表 2　27 类情绪类别及其划分结果

情绪类别	划分结果
正面情绪	admiration (钦佩), amusement (娱乐), approval (赞成), caring (关心), desire (渴望), excitement (兴奋), gratitude (感谢), joy (快乐), love (爱), optimism (乐观), pride (骄傲), relief (宽慰)
负面情绪	anger (生气), annoyance (恼怒), disappointment (失望), disapproval (反对), disgust (厌恶), embarrassment (尴尬), fear (害怕), grief (悲伤), nervousness (紧张), remorse (悔恨), sadness (难过)
中性情绪	confusion (困惑), curiosity (好奇), realization (明白), surprise (惊讶)

3.2 基准方法

在本实验中, 我们选取了 5 个不同的基准模型进行比较, 以全面评估本文所提方法的性能. 上述 5 种方法及我们方法的说明及比较详见表 3. 表 3 总结了各方法在得到情绪类别表示时是否使用了数据集的文本及标签, 以及是否需要对模型进行额外的训练. 基准模型的详细信息如下.

表 3　本文所用方法的说明及比较

方法	说明	使用数据集文本	使用数据集标签	训练模型
GloVe	词向量	否	否	否
EWE	引入情绪信息的词向量	否	否	否
GloVe_ave	基于词向量的句子向量	是	是	否
BERT_cls	基于BERT的句子向量	是	是	是
DREC	基于软标签的类别表示	是	是	是
ours	基于监督学习的类别表示	是	是	是

GloVe: GloVe 是由 Pennington 等人[40]提出来的全局词向量. 对于此方法, 我们直接使用情绪词汇对应的训练好的词向量作为此次实验的情绪表示. 比如, 对于情绪 joy, 我们直接使用 joy 对应的词向量作为 joy 的类别表示.

EWE: EWE 是由 Agrawal 等人[41]提出来的蕴含了情绪信息的情绪增强词向量. 同 GloVe 一样, 我们使用情绪词汇对应的 EWE 向量作为该词汇对应的情绪类别表示.

GloVe_ave: 在本方法中, 我们基于本文所采用 GoEmotions 数据集并结合 GloVe 来构造情绪类别表示. 对于数据集中的每条样本, 我们用该条样本的所有文本的平均词向量作为该样本所标注类别的一次采样. 对于某个情绪类别, 我们将该数据集下标注为该情绪类别所有的样本的平均词向量加权求和得到该类别的表示. 特别地, 本方法只使用数据集的文本及其标注结果, 而不需要去额外训练任何模型.

BERT_cls[28]: 在本方法中, 我们使用 BERT-base 模型在 GoEmotions 数据集上进行微调. 对于数据集中测试集的每条样本, 我们将微调后的 BERT-base 模型输出的 [CLS] 对应的向量作为该样本的表示, 并基于这些样本表示来进一步计算情绪类别表示. 本方法不仅使用了数据集的文本和标注结果, 同时也对模型进行了微调.

DREC[15]: 本方法采用训练好的分类模型输出的软标签作为样本在情绪空间中的表示, 并基于数据集中的样本表示来进一步计算出情绪类别的表示. 本文选用 BERT-base 模型作为基准分类模型. 同 BERT_cls 方法一样, 我们采用测试集的样本计算最终的类别表示.

3.3 参数设置

本文使用 BERT-base 模型作为特征抽取模型. 初始学习率设置为 2E–5, 批数据大小为 128.

3.4 Valence 维度

对于维度模型, 研究人员认为 valence (情绪的正负面程度) 是人类情感的最重要的维度[21−23]. 因此, 一个好的

表示应该可以较好地体现出情感的正面-负面维度.

为了定量衡量不同方法得到的情绪类别表示与 valence 维度的一致性. 我们引入类内相似度和类间相似度两个概念. 前者衡量同一类别下两个情绪类别 (比如两个正面情绪) 的相似性. 后者衡量不同类别下两个情绪类别 (比如一个正面情绪和一个负面情绪) 的相似性. 他们的计算公式如下:

$$s_{\text{sim}} = \frac{1}{N_{\text{sim}}} \sum_{c\in\{pos,neg,neu\}} \sum_{i<j} sim\left(e_i^c, e_j^c\right) \tag{8}$$

$$s_{\text{dis}} = \frac{1}{N_{\text{dis}}} \sum_{\substack{c_1\in\{pos,neg,neu\}\\ c_2\in\{pos,neg,neu\}\\ c_1\neq c_2}} \sum_{i,j} sim\left(e_i^{c_1}, e_j^{c_2}\right) \tag{9}$$

我们选择余弦相似度作为两个向量之间相似度的度量. 我们进一步定义本实验下得分如下:

$$score_{\text{valence}} = 1 - \frac{s_{\text{dis}}}{s_{\text{sim}}} \tag{10}$$

该得分分布在 0–1 之间, 且正面情绪类别表示和负面情绪表示区分度越高, 得分越高. 我们按照公式 (10) 计算不同方法的 valence 维度的一致性得分. 结果参见表 4.

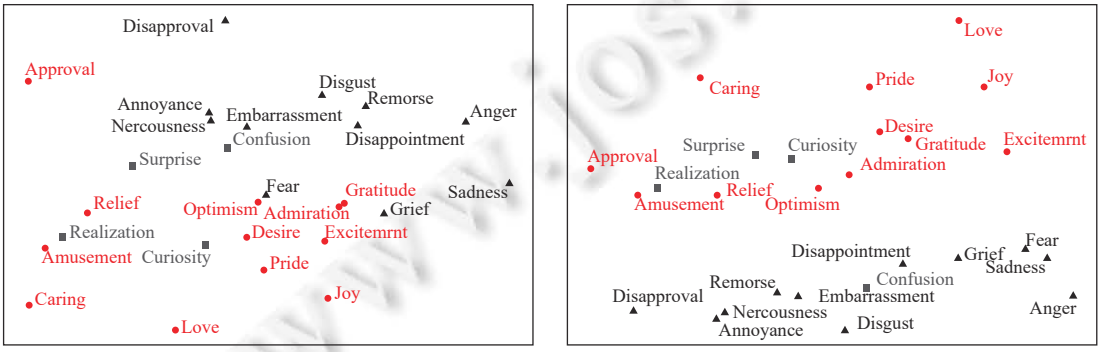
表 4　不同方法与 valence 维度的一致性

方法	GloVe	EWE	GloVe_ave	BERT_cls	DREC	ours (t=0)	ours
得分	0.285	0.246	0.420	0.519	0.558	0.552	**0.637**

如表 4 结果所述, 可以发现, 对数据集的使用越多, 得到的情绪类别表示越能反应情绪的正负面维度. 对于 GloVe 和 EWE, 这两个方法直接使用的初始词向量, 对应的一致性得分均不高于 0.3. 而剩下 4 个方法都使用了数据集中的文本, 对应的一致性得分均显著高于 GloVe 和 EWE. 对于 GloVe_ave, 该方法仅使用上下文平均词向量作为样本的情绪表示, 对应的一致性得分也小于微调后的预训练语言模型输出的 [CLS] 对应的表示方法. 更进一步地, 基于软标签的 DREC 的一致性得分也比 768 维向量的 BERT_cls 更好. 最后, 本文所提出方法最能有效的体现出情绪类别的正负面维度.

特别地, 为了分析部分相似在实验中的作用. 在本实验中, 我们额外分析了 t 取值为 0 时, 即二元相似的监督对比学习的结果. 从表 4 中可以看到, 当不采用部分相似时, 得到的类别表示对正负面情绪的区分程度甚至比不上 DREC 方法, 和 BERT_cls 方法类似. 而加上了部分相似得到的类别表示, 在 valence 维度取得了最好的结果, 这也说明本文方法确实学习到了不同类别情绪之间部分相似的关系.

为了更好地展现情绪类别表示在空间中的分布情况. 我们使用奇异值分解将高维的情绪类别表示降至两维. 如上所述, 情绪的正负面是情绪类别最重要的维度, 而奇异值分解则提取了情绪类别表示特征值绝对值最大的两个维度. 因此, 一个良好的情绪类别表示在图 3 中的正面, 负面和中性情绪应该有着较好的区分度.



(a) GloVe　　(b) EWE

图 3　不同方法下情绪类别表示的可视化结果

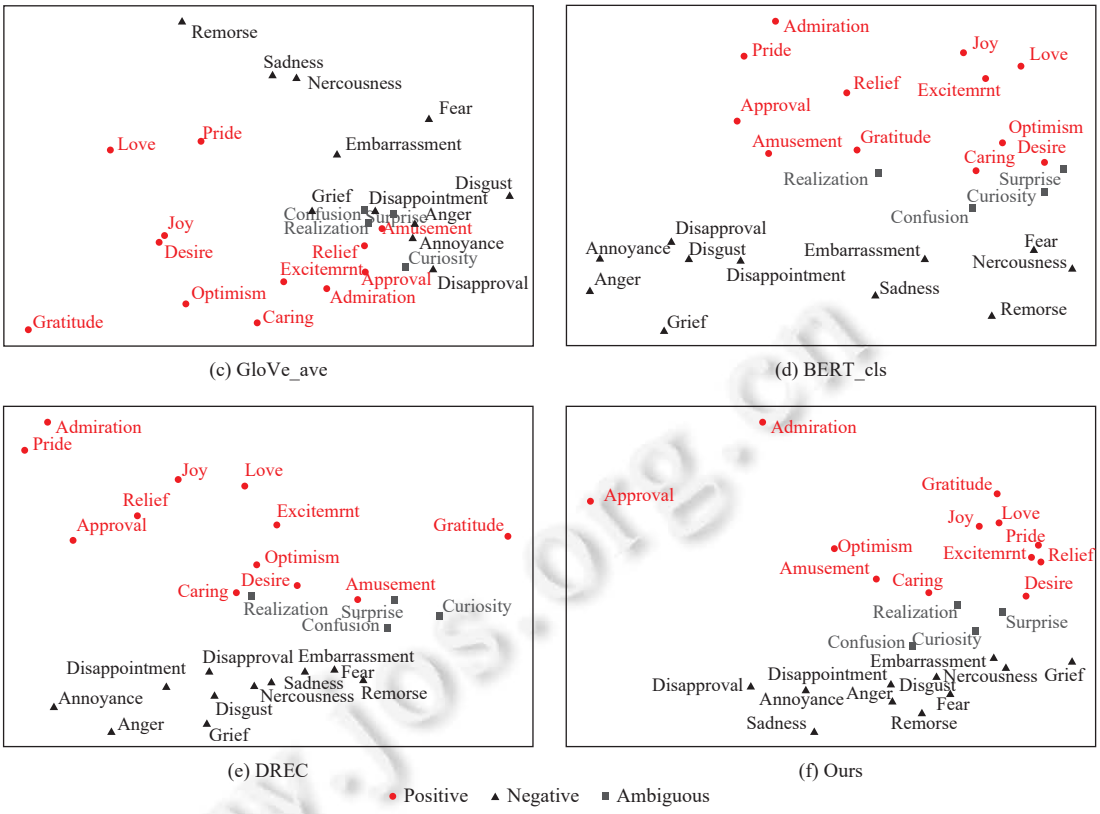


图 3 不同方法下情绪类别表示的可视化结果 (续)

如图 3 所示, 红色圆形, 黑色三角形和灰色正方形分别对应正面, 负面和中性情绪. 其中对于 GloVe, EWE 和 GloVe_ave 这 3 个方法, 正负面情绪区分度在不断增大, 但是正面情绪和负面情绪之间均未有明显的线性边界. 而对于 BERT_ave, DREC 和本文方法, 正面情绪和负面情绪之间都有着一定的间隔, 并且中性情绪刚好位于正面情绪和负面情绪之间. 图 3 的结果也与表 4 结果相互呼应.

3.5 映射实验

Demszky 等人[39]曾将 27 类情绪映射到 Ekman 提出的基本情绪[17]. 在本实验中, 我们将基于情绪类别表示之间的相似性来重构出这种映射关系, 并将结果与人类结果进行比较. 本文将 6 类基本情绪看作是源情绪, 剩下的 21 类情绪看作是目标情绪. 对于 21 类情绪中的每个情绪, 计算 6 类情绪中与其相似性最高的情绪作为映射结果. 具体的计算公式如下:

$$e = \arg\max_{e_t} sim(e_s, e_t) \tag{11}$$

其中, e_s 是源情绪, e_t 是目标情绪, 本文选取余弦相似度作为相似函数.

实验结果如表 5 所示. 我们将人类结果作为标准结果, 各方法与人类分类不一致的情绪已被红色斜体标注. 如表 5 所示, 未使用数据集信息的 GloVe 和 EWE 仅正确匹配了 3 和 7 个情绪类别. 基础词向量表示的结果远低于其他方法, 这与上面的实验是一致的. GloVe_ave 正确匹配了 14 个情绪, 这说明基于数据集的平均词向量可以在一定程度上捕捉到数据集所标注情绪类别之间的关系. 尽管 BERT_cls 在一致性得分上高于 GloVe_ave, 但是在映射实验上却和 GloVe_ave 一致, 并低于 DREC 和本文所提方法. DREC 正确匹配了 18 个情绪 (85.7%), 而本文所提方法更是正确匹配了 19 个情绪 (90.5%), 仅分类错误了 caring 和 embarrassment. 下面我们来分析这两个情绪.

表 5 各方法在映射实验上的结果

情绪	人类结果	GloVe	EWE	GloVe_ave	BERT_cls	DREC	ours
admiration	joy	*disgust*	*anger*	*disgust*	joy	joy	joy
amusement	joy	*anger*	*disgust*	*anger*	joy	joy	joy
annoyance	anger	anger	anger	anger	anger	anger	anger
approval	joy	*fear*	*fear*	*surprise*	joy	joy	joy
caring	joy	*anger*	*anger*	joy	*fear*	*sadness*	*sadness*
confusion	surprise	*anger*	*anger*	surprise	surprise	surprise	surprise
curiosity	surprise	*fear*	surprise	surprise	surprise	surprise	surprise
desire	joy	*fear*	joy	joy	*disgust*	joy	joy
disappointment	sadness	*fear*	*anger*	*disgust*	*disgust*	sadness	sadness
disapproval	anger	*disgust*	*disgust*	anger	*disgust*	*disgust*	anger
embarrassment	sadness	*disgust*	*fear*	sadness	*disgust*	*disgust*	*disgust*
excitement	joy	*anger*	joy	joy	joy	joy	joy
gratitude	joy	joy	joy	joy	joy	joy	joy
grief	sadness	*anger*	sadness	*anger*	*anger*	sadness	sadness
love	joy	joy	*surprise*	joy	joy	joy	joy
nervousness	fear	*anger*	*sadness*	*sadness*	fear	fear	fear
optimism	joy	*anger*	*anger*	joy	*disgust*	joy	joy
pride	joy	*anger*	*anger*	joy	joy	joy	joy
realization	surprise	*sadness*	*joy*	surprise	surprise	surprise	surprise
relief	joy	*anger*	*anger*	*fear*	joy	joy	joy
remorse	sadness	*disgust*	sadness	sadness	sadness	sadness	sadness
得分	—	3	7	14	14	18	**19**

对于 embarrassment, 本文方法认为 embarrassment 和 disgust 的相似性更为接近. 而 Demszky 等人[39]认为 embarrassment 和 sadness 相似性更高. 实际上, Scherer[42]曾从心理学角度探讨过这些情绪之间的关系, 他发现 embarrassment 和 disgust 和 sadness 同时都有着较高的相似度. 从这个角度来说, 将 embarrassment 划分为 sadness 未必完全正确, 而划分为 disgust 未必完全错误.

对于 caring, 本文方法认为 caring 应该归类为 sadness, 而 Demszky 等人[39]认为应该归类为 joy. 看上去完全相反的结论是由 caring 情绪的特殊性引起的. 一方面, caring 情绪的内核是正面的, 表达的是对人或事物的一种关心及担忧. 而另一方面, caring 情绪的产生往往是起源于负面事件的发生 (或者可能发生)[43]. 对于模型, 往往更关注文本本身所伴随的负面事件, 而 Demszky 等人[39]更关注于情绪自身的内核意义. 模型和人类关注的视角不同, 最终的分类结果自然也不尽相同. 比如对于样例“突然下雨了, 好担心他出门会被淋湿了”, 该样本表达的是说话的关心的情绪, 但是文本中反映的却是“淋湿”这么一个负面的伴随事件.

3.6 实验总结

在实验环节, 我们设计了两个实验来衡量不同情绪类别表示方法的内在质量. 在第 3.3 节 (valence 维度) 的实验中, 我们关注所得的情绪类别表示能否在粗粒度上较好地体现出情绪类别的正负面属性.3.3 节 (valence 维度) 的实验结果表明, 本文所提方法能最好体现出情绪中最重要的正负面维度. 在第 3.4 节 (映射实验) 中, 我们关注情绪类别表示能否在细粒度上较好地体现出不同情绪类别之间的相似程度. 第 3.4 节 (映射实验) 的实验结果表明, 本文所提方法能够最好地将 21 类情绪映射到 6 类基本情绪中去, 且映射结果与人类分类结果比肩.

同时本文所提方法得到的情绪类别表示也不再等价于类别数量, 而是可以人为预定义维度, 解决了维度受限的问题. 此外, 新的情绪类别表示也不再局限于各向异性的超立方体中, 而是分布在各向同性的超球面上.

4 结束语

本文针对情绪类别在情绪空间中的分布式表示的问题, 引入了监督对比学习的表示方法. 为了生成的表示能

够更好地反映出不同情绪类别之间的复杂关系, 我们进一步提出部分相似的监督对比学习方法. 部分相似方法的提出, 丰富了既有监督对比学习中不同样本之间的相似性关系, 将样本之间的相似性关系从相似或不相似的二元离散相似关系拓展至 0–1 分布的连续相似关系. 实验部分, 我们将本文方法和 5 个其他基准方法进行比较. 实验结果表明, 本文方法不管是在 valence 维度的一致性还是情绪类别之间的相似性, 均较其他方法取得了提升. 特别地, 本文方法所得到的情绪类别表示的维度不受限于特定数据集下的类别维度. 并且每个维度下的特征值也不再强制介于 0–1 之间, 即每个情绪类别表示不再受限于超立方体中.

在下一步的工作中, 一方面, 我们将进一步研究如何更好地融合既有知识 (比如心理学领域的研究), 以先验的构造出不同情绪类别之间更加细致的相似关系, 并进一步生成更加全面更加符合人类情感状态的情绪类别表示. 另一方面, 现有方法仍然基于特定的数据集, 得到的情绪类别表示也局限于数据集中的标注类别, 未来我们将探索如何融合广泛的现有数据集来得到更为全面的情绪类别表示结果.

References:

[1] Zong CQ, Xia R, Zhang JJ. Text Data Mining. Beijing: Tsinghua University Press, 2019 (in Chinese).

[2] Lee SYM, Chen Y, Huang CR. A text-driven rule-based system for emotion cause detection. In: Proc. of the 2010 NAACL HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text. Los Angeles: Association for Computational Linguistics, 2010. 45–53.

[3] Xia R, Ding ZX. Emotion-cause pair extraction: A new task to emotion analysis in texts. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 1003–1012. [doi: 10.18653/v1/P19-1096]

[4] Mohammad S, Bravo-Marquez F. Emotion intensities in tweets. In: Proc. of the 6th Joint Conf on Lexical and Computational Semantics. Vancouver: Association for Computational Linguistics, 2017. 65–77. [doi: 10.18653/v1/S17-1007]

[5] Jiang L, Yu M, Zhou M, Liu XH, Zhao TJ. Target-dependent Twitter sentiment classification. In: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland: Association for Computational Linguistics, 2011. 151–160.

[6] Gui L, Xu RF, Wu DY, Lu Q, Zhou Y. Event-driven emotion cause extraction with corpus construction. In: Social Media Content Analysis: Natural Language Processing and Beyond. Singapore: World Scientific Publishing, 2017. 145–160. [doi: 10.1142/9789813223615_0011]

[7] Gui L, Xu RF, Lu Q, Wu DY, Zhou Y. Emotion cause extraction, a challenging task with corpus construction. In: Proc. of the 5th National Conf. on Social Media Processing. Nanchang: Springer, 2016. 98–109. [doi: 10.1007/978-981-10-2993-6_8]

[8] Gui L, Yuan L, Xu RF, Liu B, Lu Q, Zhou Y. Emotion cause detection with linguistic construction in Chinese Weibo text. In: Proc. of the 3rd CCF Conf. on Natural Language Processing and Chinese Computing. Shenzhen: Springer, 2014. 457–464. [doi: 10.1007/978-3-662-45924-9_42]

[9] Wang K, Shen WZ, Yang YY, Quan XJ, Wang R. Relational graph attention network for aspect-based sentiment analysis. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 3229–3238. [doi: 10.18653/v1/2020.acl-main.295]

[10] Chen X, Sun CL, Wang JJ, Li SS, Si L, Zhang M, Zhou GD. Aspect sentiment classification with document-level sentiment preference modeling. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020. 3667–3677. [doi: 10.18653/v1/2020.acl-main.338]

[11] Li YG, Zhou XG, Sun Y, Zhang HG. Research and implementation of Chinese microblog sentiment classification. Ruan Jian Xue Bao/Journal of Software, 2017, 28(12): 3183–3205 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5283.htm [doi: 10.13328/j.cnki.jos.005283]

[12] Jackson JC, Watts J, Henry TR, List JM, Forkel R, Mucha PJ, Greenhill SJ, Gray RD, Lindquist KA. Emotion semantics show both cultural variation and universal structure. Science, 2019, 366(6472): 1517–1522. [doi: 10.1126/science.aaw8160]

[13] Müller R, Kornblith S, Hinton GE. When does label smoothing help. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver, 2019. 4694–4703.

[14] Zhang CB, Jiang PT, Hou QB, Wei YC, Han Q, Li Z, Cheng MM. Delving deep into label smoothing. IEEE Trans. on Image Processing, 2021, 30: 5984–5996. [doi: 10.1109/TIP.2021.3089942]

[15] Wang XY, Zong CQ. Distributed representations of emotion categories in emotion space. In: Proc. of the 59th Annual Meeting of the

Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing (Vol. 1: Long Papers). Association for Computational Linguistics, 2021. 2364–2375. [doi: 10.18653/v1/2021.acl-long.184]

[16] Wang TZ, Isola P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: Proc. of the 37th Int'l Conf. on Machine Learning. JMLR.org, 2020. 9929–9939.

[17] Ekman P. An argument for basic emotions. Cognition & Emotion, 1992, 6(3–4): 169–200. [doi: 10.1080/02699939208411068]

[18] Plutchik R. A general psychoevolutionary theory of emotion. In: Plutchik R, Kellerman H, eds. Emotion: Theory, Research, and Experience. New York: Academic Press, 1980. 3–33.

[19] Cowen AS, Keltner D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. Proc. of the National Academy of Sciences of the United States of America, 2017, 114(38): E7900–E7909. [doi: 10.1073/pnas.1702247114]

[20] Cowen A, Sauter D, Tracy JL, Keltner D. Mapping the passions: Toward a high-dimensional taxonomy of emotional experience and expression. Psychological Science in the Public Interest, 2019, 20(1): 69–90. [doi: 10.1177/1529100619850176]

[21] Russell JA. A circumplex model of affect. Journal of Personality and Social Psychology, 1980, 39(6): 1161–1178. [doi: 10.1037/h0077714]

[22] Russell JA. Core affect and the psychological construction of emotion. Psychological Review, 2003, 110(1): 145–172. [doi: 10.1037/0033-295X.110.1.145]

[23] Bakker I, Van Der Voordt T, Vink P, de Boon J. Pleasure, arousal, dominance: Mehrabian and Russell revisited. Current Psychology, 2014, 33(3): 405–421. [doi: 10.1007/s12144-014-9219-4]

[24] Xu P, Madotto A, Wu CS, Park JH, Fung P. Emo2Vec: Learning generalized emotion representation by multi-task training. In: Proc. of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Brussels: Association for Computational Linguistics, 2018. 292–298. [doi: 10.18653/v1/W18-6243]

[25] Wang S, Maoliniyazi A, Wu XL, Meng XF. Emo2Vec: Learning emotional embeddings via multi-emotion category. ACM Trans. on Internet Technology, 2020, 20(2): 13. [doi: 10.1145/3372152]

[26] Han J, Zhang ZX, Ren Z, Schuller B. EmoBed: Strengthening monomodal emotion recognition via training with crossmodal emotion embeddings. IEEE Trans. on Affective Computing, 2021, 12(3): 553–564. [doi: 10.1109/TAFFC.2019.2928297]

[27] Buechel S, Modersohn L, Hahn U. Towards label-agnostic emotion embeddings. arXiv:2012.00190, 2021.

[28] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]

[29] Yang ZL, Dai ZH, Yang YM, Carbonell J, Salakhutdinov R, Le QV. XLNet: Generalized autoregressive pretraining for language understanding. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver, 2019. 5753–5763.

[30] Doersch C, Gupta A, Efros AA. Unsupervised visual representation learning by context prediction. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 1422–1430. [doi: 10.1109/ICCV.2015.167]

[31] Zhang R, Isola P, Efros AA. Colorful image colorization. In: Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 649–666. [doi: 10.1007/978-3-319-46487-9_40]

[32] Le-Khac PH, Healy G, Smeaton AF. Contrastive representation learning: A framework and review. IEEE Access, 2020, 8: 193907–193934. [doi: 10.1109/ACCESS.2020.3031549]

[33] Hénaff OJ, Srinivas A, de Fauw J, Razavi A, Doersch C, Ali Eslami SM, van den Oord A. Data-efficient image recognition with contrastive predictive coding. In: Proc. of the 37th Int'l Conf. on Machine Learning. JMLR.org, 2020. 4182–4192.

[34] Hjelm RD, Fedorov A, Lavoie-Marchildon S, Grewal K, Bachman P, Trischler A, Bengio Y. Learning deep representations by mutual information estimation and maximization. arXiv:1808.06670, 2019.

[35] Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: Proc. of the 37th Int'l Conf. on Machine Learning. JMLR.org, 2020. 1597–1607.

[36] Gao TY, Yao XC, Chen DQ. SimCSE: Simple contrastive learning of sentence embeddings. In: Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing. Punta Cana: Association for Computational Linguistics, 2021. 6894–6910.

[37] Khosla P, Teterwak P, Wang C, Sarna A, Tian YL, Isola P, Maschinot A, Liu C, Krishnan D. Supervised contrastive learning. arXiv: 2004.11362, 2021.

[38] Gunel B, Du JF, Conneau A, Stoyanov V. Supervised contrastive learning for pre-trained language model fine-tuning. arXiv:2011.01403, 2021.

[39] Demszky D, Movshovitz-Attias D, Ko J, Cowen A, Nemade G, Ravi S. GoEmotions: A dataset of fine-grained emotions. arXiv: 2005.00547, 2020.

[40] Pennington J, Socher R, Manning C. GloVe: Global vectors for word representation. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing. Doha: Association for Computational Linguistics, 2014. 1532–1543. [doi: 10.3115/v1/D14-1162]

[41] Agrawal A, An AJ, Papagelis M. Learning emotion-enriched word representations. In: Proc. of the 27th Int'l Conf. on Computational Linguistics. Santa Fe: Association for Computational Linguistics, 2018. 950–961.

[42] Scherer KR. What are emotions? And how can they be measured. Social Science Information, 2005, 44(4): 695–729. [doi: 10.1177/0539018405058216]

[43] Scherer KR, Shuman V, Fontaine JRJ, Soriano C. The GRID meets the wheel: Assessing emotional feeling via self-report. In: Fontaine JJR, Scherer KR, Soriano C, eds. Components of Emotional Meaning: A Sourcebook. Oxford: Oxford University Press, 2013. 281–298.

附中文参考文献:

[1] 宗成庆, 夏睿, 张家俊. 文本数据挖掘. 北京: 清华大学出版社, 2019.

[11] 李勇敢, 周学广, 孙艳, 张焕国. 中文微博情感分析研究与实现. 软件学报, 2017, 28(12): 3183–3205. http://www.jos.org.cn/1000-9825/5283.htm [doi: 10.13328/j.cnki.jos.005283]

王祥宇(1998－), 男, 博士, 主要研究领域为自然语言处理, 情感分析.

宗成庆(1963－), 男, 博士, 研究员, 博士生导师, CCF 会士, 主要研究领域为自然语言处理, 机器翻译, 情感分析.