计算机科学
COMPUTER SCIENCE

基于视觉语义与提示学习的多模态情感分析模型

莫书渊, 蒙祖强

引用本文

莫书渊, 蒙祖强. 基于视觉语义与提示学习的多模态情感分析模型[J]. 计算机科学, 2024, 51(9): 250-257.

MO Shuyuan, MENG Zuqiang. Multimodal Sentiment Analysis Model Based on Visual Semantics and Prompt Learning [J]. Computer Science, 2024, 51(9): 250-257.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

融合多模态物联网设备指纹与集成学习的物联网设备识别方法

IoT Device Recognition Method Combining Multimodal IoT Device Fingerprint and Ensemble Learning

计算机科学, 2024, 51(9): 371-382. https://doi.org/10.11896/jsjkx.230800076

基于文本和图像门控融合机制的多模态方面级情感分析

Text-Image Gated Fusion Mechanism for Multimodal Aspect-based Sentiment Analysis

计算机科学, 2024, 51(9): 242-249. https://doi.org/10.11896/jsjkx.230600117

基于半监督学习的域适应实体解析算法

Domain-adaptive Entity Resolution Algorithm Based on Semi-supervised Learning

计算机科学, 2024, 51(9): 214-222. https://doi.org/10.11896/jsjkx.230800102

基于双编码器的多模态融合方法

Multi-modal Fusion Method Based on Dual Encoders

计算机科学, 2024, 51(9): 207-213. https://doi.org/10.11896/jsjkx.230700212

基于多模态注意力网络的红外人体行为识别方法

Infrared Human Action Recognition Method Based on Multimodal Attention Network

计算机科学, 2024, 51(8): 232-241. https://doi.org/10.11896/jsjkx.230600143

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230600047

基于视觉语义与提示学习的多模态情感分析模型

莫书渊　蒙祖强
广西大学计算机与电子信息学院　南宁 530004
(msygxu2023@163.com)

摘　要　随着深度学习技术的发展,多模态情感分析已成为研究热点之一。然而,大多数多模态情感分析模型或从不同模态中提取特征向量并简单地进行加权求和,导致数据无法准确地映射到统一的多模态向量空间中,或依赖图像描述模型将图像转化为文本,导致提取到过多不包含情感信息的视觉语义,造成信息冗余,最终影响模型的性能。为了解决这些问题,提出了一种基于视觉语义与提示学习的多模态情感分析模型 VSPL。该模型将图像转化为精确简短、蕴含情感信息的视觉语义词汇,从而缓解信息冗余的问题;并基于提示学习的方法,将得到的视觉语义词汇与针对情感分类任务而提前设计好的提示模板组合成新文本,实现模态融合,这样做既避免了由加权求和导致的特征空间映射不准确的问题,又能借助提示学习的方法激发预训练语言模型的潜在性能。对多模态情感分析任务进行了对比实验,结果表明所提模型 VSPL 在 3 个公开数据集上的性能超越了先进的基准模型。此外,还进行了消融实验、特征可视化和样例分析,验证了 VSPL 的有效性。

关键词:多模态;视觉语义;提示学习;情感分析;预训练语言模型
中图分类号　TP391

Multimodal Sentiment Analysis Model Based on Visual Semantics and Prompt Learning

MO Shuyuan and MENG Zuqiang
School of Computer and Electronic Information,Guangxi University,Nanning 530004,China

Abstract　With the development of deep learning technology,multimodal sentiment analysis has become one of the research highlights. However,most multimodal sentiment analysis models either extract eigenvector from different modalities and simply use weighted sum method,resulting in data that cannot be accurately mapped into a unified multimodal vector space,or rely on image description models to translate image into text,resulting in the extraction of too many visual semantics without sentimental information and information redundancy,and ultimately affecting the performance of the model. To address these issues,a multimodal sentiment analysis model VSPL based on visual semantics and prompt learning is proposed. This model translates images into precise,concise,and sentimentally informative visual semantic vocabulary to alleviate the problem of information redundancy. Based on prompt learning,the obtained visual semantic vocabulary is combined with pre-designed prompt templates for sentiment classification tasks to form new text,achieving modal fusion. It not only avoids the problem of inaccurate feature space mapping caused by weighted sum method,but also stimulates the potential performance of pre-trained language model through prompt learning methods. Comparative experiments are conducted on multimodal sentiment analysis tasks,and the proposed model VSPL outperforms advanced baseline models on three public datasets. In addition,ablation experiments,feature visualization,and sample analysis are conducted to verify the effectiveness of VSPL.

Keywords　Multimodal,Visual semantics,Prompt learning,Sentiment analysis,Pre-trained language model

1　引言

随着移动设备和社交网络的蓬勃发展,人们逐渐倾向于将图片和文本一起发布,以表达自己的情感或观点。近年来,为了从不同模态的数据中分析与识别情感,多模态情感分析吸引了越来越多研究者的关注[1]。多模态情感分析具有各种潜在的应用,包括情感识别[2]、情感跨模态检索[3]、意见挖掘[4]、决策[5]等。相对于单模态情感分析,对来自不同模态的数据进行处理和分析,既是机遇,也面临着挑战。随着深度学习技术的发展,深度神经网络也逐渐被应用于多模态情感分析领域,并且表现不俗。现有的多模态情感分析模型的方法大致可以分为两类:第一类是基于多模态联合表示的

到稿日期:2023-06-06　返修日期:2023-12-25
基金项目:国家自然科学基金(62266004)
This work was supported by the National Natural Science Foundation of China(62266004).
通信作者:蒙祖强(zqmeng@126.com)

方法[6]，这类方法使用深度神经网络分别提取不同模态的特征向量，并进行简单的加权求和，最后再进行分类；第二类是基于多模态转化的方法[7]，这类方法使用图像描述模型将图像转化为相应的描述，并与文本进行拼接，再进行分类。然而，基于多模态联合表示的方法由于使用简单的加权求和操作，会出现映射到向量空间中的数据不准确的问题；基于多模态转化的方法由于依赖跨模态生成模型，生成的描述可能不准确，并且生成的描述中会包含一些与情感分析任务无关的冗余信息。这些问题最终都会影响模型的性能。

为了解决上述问题，本文提出了一种基于视觉语义与提示学习的多模态情感分析模型 VSPL（Visual Semantics Prompt Learning），该模型专注于将图像转化为精确简短、蕴含情感信息的视觉语义词汇，并结合提示学习的方法来提升模型的分类性能。

本文的主要贡献包括 3 个方面：

（1）提出了一种用于多模态情感分析的新模型 VSPL。相对于生成模型，该模型使用更准确的分类模型将图像模态转化为精确简短、蕴含情感信息的视觉语义词汇，并与提示模板、文本模态组合，输入预训练语言模型进行情感分类。

（2）使用分类模型将图像模态转化为精确简短、蕴含情感的视觉语义词汇的方式，既能为使用提示学习的方法创造前提条件，又能精炼地表示图像模态中的情感要素，从而改善信息冗余的问题；将图像模态处理得到的提示模板与原始文本组合再输入预训练语言模型的方式避免了加权求和操作导致的数据映射不准确的问题。

（3）在多模态情感分析领域中使用了提示学习的方法，激发了预训练语言模型的潜在性能，所提出的 VSPL 模型在 3 个公开数据集上的性能超越了先进的基准模型。

2 相关工作

本章将简要回顾视觉情感分析、提示学习，以及多模态情感分析的相关研究。

2.1 视觉情感分析

早期对视觉情感分析的研究主要利用设计手工特征的方式来表示视觉模态情感。受心理学和艺术理论的启发，Machajdik 等[8]从图像中提取低层次特征，例如构图、纹理等，以预测视觉情感。Zhao 等[9]根据艺术原则设计了由平衡、强调、和谐、多样、渐变和运动等多种特征组成的中层特征。Borth 等[10]提出了一个新概念，称为形容词与名词对（ANPs），以同时保存图像中对象的情感和位置信息。Li 等[11]基于 1 200 个 ANPs 构建的图像特征向量充分分析了 ANPs 的情感值对图像情感分析结果的影响。Zhao 等[12]将不同级别的特征与多图学习相结合，包括基于艺术的低级特征的通用元素、基于艺术的中级特征的属性和原则，以及基于语义概念和面部表情的高级特征。

近年来，深度神经网络已被广泛应用于视觉情感分析。You 等[13]通过注意力机制研究了情感相关的局部图片区域，并基于这些局部特征训练情感分类器。Yang 等[14]引入了弱监督耦合卷积网络来检测情感图，该网络利用整体和局部特征进行视觉情感预测。Zhang 等[15]通过多层次情感区域相关性分析来进行视觉情感分析。受心理学中 S-O-R（Stimuli-Organism-Response）模型的启发，Yang 等[16]提出了一种基于刺激感知的视觉情感分析网络来模拟人类情感的唤起过程，进行视觉情感分析。现有趋势是将视觉情感分析与文本相结合来进行多模态情感分析的研究。

2.2 提示学习

提示学习的概念源自于自然语言处理领域，目的是将预训练语言模型，如 BERT[17]或 GPT[18]等视为知识库，并从中提取出对特定下游任务有用的信息。具体来说，给定一个预训练语言模型，下游任务通常被设计为“填空题”的形式，例如要求模型预测“这部电影我很喜欢，这让人感到[MASK]”中的[MASK]标记，给出“积极”或“消极”等词语进行情感分类。提示学习的关键在于如何设计文本中下划线的部分，即提示模板，以此激发预训练语言模型的潜在性能。Jiang 等[19]使用文本挖掘等方法来生成一组候选提示，选择最佳提示以使模型具有最高的训练精度。Shin 等[20]提出了自动提示，这是一种基于梯度的方法，该方法根据标签可能性从词汇表中选择导致梯度变化最大的最佳标记。在计算机视觉中，提示学习是一个新兴的研究方向，最近才有相关研究提出[21-22]。CoOp[22]是最早将连续提示学习引入视觉领域以适应预训练视觉语言模型的工作。而 Zhou 等[23]解决了 CoOp 的弱可推广性问题，提出了一种基于条件提示学习的简单思想。近年来，提示学习在各领域中的研究取得了诸多进展，例如文本情感分析领域[24]。因此，将多模态情感分析与提示学习相结合的相关研究是可行的。

2.3 多模态情感分析

随着互联网的发展，社交媒体的用户们越来越倾向于同时使用文本和图像来表达自己的观点，使得多模态情感分析成为一种有吸引力的选择[25-26]。Morency 等[27]首先考虑了多模态情感分析的任务，并提出分析音频、视频和文本数据。研究人员还发布了一些多模态情感分析资源，如 MOSEI[28]，Muse Toolbox[29]，MOSI[30]和 MVSA[31]。这些数据集中的数据大部分集中在视频、文本和图像上。本文主要考虑具有图像和文本信息的多模态情感分析数据集（如 MVSA）。Twitter15 和 Twitter17[32]是面向目标的多模态情感分析任务的两个最流行的数据集。Xu 等[33]提出了一个名为 MultiSentiNet 的深度网络，该网络利用图像的场景和对象特征，基于注意力指出重要的句子单词。Xu 等[34]考虑了视觉和文本信息的相互关系，并提出了共记忆网络 CoMN，通过图像模态和文本模态之间的相互作用构建模型，进行多模态情感分析。Cai 等[35]提出了 CNN-Multi 网络，将两个单独的卷积神经网络结构用于学习文本特征和视觉特征。Khan 等[7]使用对象感知 Transformer 在输入空间中对图像生成描述，并使用单通道非自回归文本生成方法，最后利用图像描述构建一个辅助句子，为语言模型提供多模态信息。Zhu 等[6]提出了 ITIN 网络，引入一个跨模态对齐模块来捕获区域词对应关系，并通过自适应跨模态门控模块来融合多模态特征。Chochlakis 等[36]提出了 VAuLT，将预训练语言模型（如 BERT）的输出表示传播到

ViLT[37]的语言部分的输入中。然而，大多数多模态情感分析模型或从不同模态中提取特征向量并简单地进行加权求和，导致数据无法准确地映射到统一的多模态向量空间中，或依赖图像描述模型将图像转化为文本，提取到过多不包含情感信息的视觉语义，造成信息冗余，最终影响模型的性能。

3 VSPL 模型

如图 1 所示，本文提出的基于视觉语义与提示学习的多模态情感分析模型 VSPL 主要由 3 部分组成：视觉语义提取、提示模板组合，以及情感分类模型。

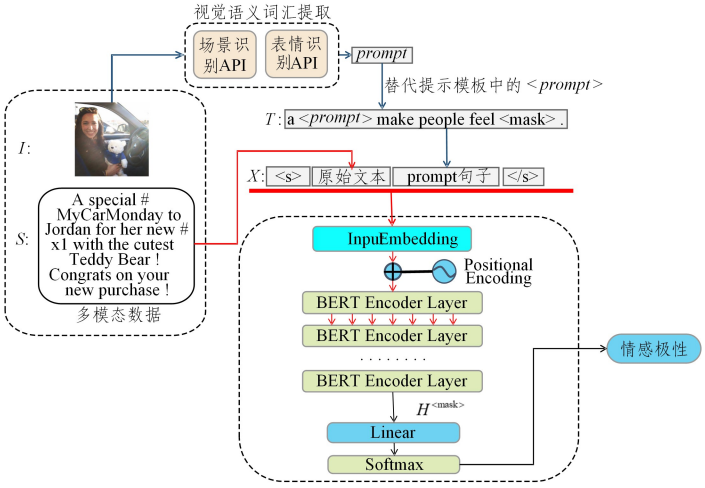


图 1 模型处理流程图

Fig. 1 Flow chart of model processing

3.1 问题定义

每个样本由一个句子、一张图片和对应的标签组成。句子表示为 $S=(w_1,w_2,\cdots,w_n)$，其中 w 代表单词，n 是句子的单词数量。图片表示为 I，标签表示为 y，标签 $y\in${消极，中立，积极}。

目标是学习函数 $f:(S,I)\rightarrow y$。简而言之，在一个多模态样本中，比如句子 S 为“上班累了，想休息一会儿”，对应的图片 I 为办公室的场景，则该模型会预测这个多模态样本的标签 y 为消极。

3.2 视觉语义提取

如图 1 所示，首先需要将图片 I 的信息提取为可能蕴含情感信息的视觉语义词汇 *prompt*。从图片中提取的视觉语义词汇是否正确，会直接影响分类的结果。因此，本文选择将图片 I 输入阿里云视觉智能开放平台的场景识别 API 中，通过准确的场景及实体识别，能较为准确地提取图片中的视觉语义词汇，例如办公室、演唱会、日出以及日落等场景词汇，或者猫、狗、花以及人物等实体词汇。特别地，当场景识别 API 提取到的视觉语义词汇为人物时，则需要进一步调用表情识别 API，以识别图片中人物的表情，提取对应的视觉语义词汇，并设置为最终的 *prompt*。

具体的算法如算法 1 所示。

算法 1 视觉语义提取算法

输入：图像 I

输出：视觉语义词汇 prompt

```
1. Senne_Class=Api_Senne(I);
2. prompt=Senne_Class;
3. if Senne_Class==‘human’:Expression_Class=Api_Expression
   (I)
   prompt=Expression_Class;
   /* 如果 Senne_Class 的值为‘human’，则将 Expression_Class 设为
   最终的 prompt */
4. end if。
```

其中，I 代表图片，*Api_Senne* 为场景识别 API，*Senne_Class* 表示场景识别的结果；*Api_Expression* 为表情识别 API，*Expression_Class* 表示表情识别的结果，得到的最后结果设置为 *prompt*。

3.3 提示模板设计

由于本模型借鉴了提示学习的思想，因此需要设计一个与情感分析任务相关的提示模板。目前，提示模板的设计形式有手工模板、自动离散模板等，本文提出的方法将采用手工模板的形式进行设计。如图 1 所示，本文研究的是情感分析任务，因此本实验中的提示模板可设计为：“a 〈*prompt*〉 make people feel 〈mask〉”。提示模板中的〈*prompt*〉需要替换为上一步提取到的视觉语义词汇 *prompt*。〈mask〉是掩蔽位，将在下一节具体介绍。例如，提取到的语义词汇 *prompt* 是“happiness”，则需要用“happiness”替换模板中的“〈*prompt*〉”，得到对应的文本“a happiness make people feel 〈mask〉”，并将这个文本设为 T，到此完成了对图像模态 I 的处理。得到的文本 T 再与原始文本 S 拼接，则得到形如“〈*s*〉 S. a happiness make people feel 〈mask〉. 〈/*s*〉”的新文本 X，其中〈*s*〉和〈/*s*〉分别代表开始位和结束位。文本模态的 S 与图像模态提取到的 *prompt* 在下一步可以对〈mask〉掩蔽位起作用，到此完成了图像模态与文本模态的融合。这种提示学习的方式相当于提示模型要进行情感分类的任务，因此可以利用预训练语言模型蕴含的信息，充分发挥预训练语言模型的潜能。

3.4 情感分类

基于掩蔽的预训练语言模型在预训练的过程中会将输入文本中的单词进行随机掩蔽，并在训练过程中要求模型恢复这些被随机掩蔽的单词[17]，也就是〈mask〉位置上的单词。这种训练方式使得模型能够捕捉词与词之间的关系，包括句子

内部的依赖关系和上下文之间的关系；且由于在预训练时使用了大规模的无标注数据，因此模型可以学习到更加丰富的语言表示。为了充分利用预训练语言模型所蕴含的信息，需要与基于掩蔽的预训练过程保持一致，即上一步本文设计的与情感分类任务相关的提示模板，其中〈mask〉标记代表情感标签，可以强制模型更关注情感内容。因此，首先对输入进行拼接得到新文本 X，将在〈mask〉位置上的单词作为关键字将其掩蔽。然后，基于掩蔽的预训练语言模型可以计算未知掩蔽词的表示。如图 1 所示，本文选择使用 BERT 模型。BERT 模型是一种常用的基于掩蔽的预训练语言模型，由嵌入层以及多个编码器组成。

上一步得到的包含图像模态和文本模态所有信息的新文本 X 可重新表示为：

$$X=\{x_1,x_2,\cdots,\langle\text{mask}\rangle,\cdots,x_n\} \tag{1}$$

其中，X 中的每个元素是组成新文本 X 的单词，〈mask〉是掩蔽位。将 X 输入 BERT 模型，经过嵌入层映射到向量空间，并经过多个编码层处理，可以得到最终的隐层向量 $\boldsymbol{H}$。提示学习的常规做法是在微调之后取得〈mask〉处预测的单词，再使用提前人工设计的标签映射字典找到对应的预测结果。而人工设计映射关系是繁琐的，因此本文改进了相关步骤，选出最后一层隐层向量中与新文本 X 中的〈mask〉位置处所对应的向量，并使用全连接层自动地学习映射关系。设 $\boldsymbol{H}^{\langle\text{mask}\rangle}\in\mathbb{R}^{768}$ 是最后一层隐层向量中与新文本 X 中的〈mask〉位置所对应的向量输出，然后可以计算 y 属于消极、中立或积极情感的概率，表示为：

$$p(y|\boldsymbol{H}^{\langle\text{mask}\rangle})=\text{softmax}(\theta_{\text{Linear}}\text{Dropout}(\boldsymbol{H}^{\langle\text{mask}\rangle})) \tag{2}$$

其中，$\theta_{\text{Linear}}\in\mathbb{R}^{3\times768}$，并通过反向传播算法进行参数的优化。本文使用交叉熵损失函数微调 BERT 编码器和学习 θ_{Linear}，计算损失值的公式可表示为：

$$loss=-\log p(y|\boldsymbol{H}^{\langle\text{mask}\rangle}) \tag{3}$$

4 实验结果

为了验证文本提出的模型 VSPL 的有效性，将 VSPL 与多个多模态情感分析模型进行了比较，并进行了消融实验、特征可视化以及样例分析。

4.1 数据集

本次实验使用 3 个公共多模态情感分析数据集，包括 Twitter15 和 Twitter17[32] 以及 MVSA Single[31]，这 3 个数据集都包含对推文图片和文本情感的标注，代表积极、消极以及中立 3 种情感。

Twitter15 数据集包含 1 548 个积极、630 个消极和 3 169 个中立样本。Twitter17 数据集包含 2 516 个积极、728 个消极和 2 728 个中立样本，本文使用数据集的作者提供的拆分，其他处理与文献[36]相同。

MVSA Single 只有一个标注人员进行标注，处理后的 MVSA Single 数据集包含 2 683 个积极、460 个消极和 1 358 个中立样本。由于 MVSA Single 数据集没有提供标准的拆分，评价指标也没有明确定义，因此本文遵循先前的工作，以 8∶1∶1 的比例随机拆分数据集[6]。

4.2 评价指标

本文的实验使用准确率和 F1 作为评价指标，可表示为：

$$Accuracy=\frac{TP+TN}{TP+FN+TN+FP} \tag{4}$$

$$F1\ score=\frac{2*Precision*Recall}{Precision+Recall} \tag{5}$$

4.3 实验细节

本实验使用预训练模型 twitter-roberta-base-sentiment[38] 进行参数初始化，该模型的嵌入层词典大小为 50 265，编码器部分由 12 个自注意力子层组成，每个子层有 12 个注意力头，隐层维度为 768，序列最大长度为 40。学习率设置为 2×10^{-5}，训练轮次为 20 轮，批次大小为 16，优化器使用 Adam，使用交叉熵损失函数。本实验的软硬件实验环境如表 1 所列。

表 1 软硬件实验环境

Table 1 Hardware and software experiment environment

GPU	RTX 3060
CPU	AMD Ryzen 7 5800H with Radeon Graphics 3.20 GHz
操作系统	Windows 10
开发环境	Python 3.7
编程语言	Python
深度学习框架	Pytorch 1.11.0
并行计算架构	CUDA 11.3

4.4 基准模型

为了验证 VSPL 模型的有效性，本文将其与 9 个多模态基准模型进行比较，其中包括最先进的(state-of-the-art，SOTA)基准模型。

(1)MultiSentiNet[33] 是一种用于多模态情感分析的语义网络。

(2)CoMN[34] 提出了一种迭代地对图片和文本之间的相互影响进行建模的共记忆网络，以改进多模态情感分类。

(3)CNN-Multi[35] 使用两个独立的 CNN 架构来学习文本和图片的特征，并将其用作另一个 CNN 的输入，用于多模态情感分析。

(4)ALBEF[39] 是一种多模态预训练模型，可获得多模态任务的相关结果。

(5)ViLT[37] 是一种流行的多模态 Transformer。对于文本信息，将输入标记嵌入查找表中；对于视觉信息，将每个图片分割成不重叠的正方形面片，并使用线性投影将其映射到输入空间。

(6)VAuLT[36] 是 ViLT 的扩展，它提高了视觉和语言任务的性能，这些任务涉及比图片字幕更复杂的文本输入，同时对训练和推理效率的影响最小。

(7)SMP[40] 包含一个基于视觉和文本信息之间的交互的跨模式对比学习模块，并引入了额外的情感感知预训练目标(如细粒度情感标记)，以从情感丰富的数据集中捕获细粒度情感信息，用于多模态情感分析。

(8)EF-CaTr-BERT[7] 使用对象感知 Transformer 在输入空间中生成图像描述，并使用单通道非自回归文本生成方法，最后利用得到的描述构建一个辅助句子，为语言模型提供多模态信息。

(9)ITIN[6]引入了一个跨模态对齐模块来捕获区域词对应关系,并通过自适应跨模态门控模块来融合多模态特征,以实现多模态情感分析。

4.5 结果分析

表 2—表 4 分别列出了不同模型在 3 个多模态数据集上的实验结果。总体来说,本文提出的模型 VSPL 在 3 个实验数据集上的表现都优于其他模型。

表 2 Twitter15 数据集上的对比实验结果

Table 2 Comparative experimental results on Twitter15 dataset

(%)

Model	*ACC*	*F*1
ViLT[37]	70.50	62.60
ALBEF[39]	74.25	69.65
VAuLT[36]	77.50	72.90
SMP[40]	77.53	72.24
EF-CaTr-BERT[7]	77.90	73.90
VSPL	78.88	74.04

表 3 Twitter17 数据集上的对比实验结果

Table 3 Comparative experimental results on Twitter17 dataset

(%)

Model	*ACC*	*F*1
ViLT[37]	62.60	58.10
ALBEF[39]	67.91	65.37
VAuLT[36]	71.00	69.50
SMP[40]	71.15	69.47
EF-CaTr-BERT[7]	72.30	70.20
VSPL	73.34	72.26

表 4 MVSA Single 数据集上的对比实验结果

Table 4 Comparative experimental results on MVSA Single dataset

(%)

Model	*ACC*	*F*1
CNN-Multi[35]	61.20	58.37
MultiSentiNet[33]	69.84	69.83
CoMN[34]	70.51	70.01
ViLT[37]	74.40	73.70
ITIN[6]	75.19	74.97
VAuLT[36]	78.00	77.40
VSPL	78.71	77.62

具体来说,对于 Twitter15 数据集以及 Twitter17 数据集,ViLT 和 ALBEF 是两个普通的多模态预训练模型,它们只捕获普通的语义多模态信息而忽略了情感信号,因此性能相对较差。VAuLT 和 SMP 都使用了 CNN 或者注意力机制等深度神经网络来提取各模态的特征向量,并使用了简单的加权求和,因此映射到向量空间中的数据不准确,干扰了分类结果。目前这两个数据集的 SOTA 模型,即 EF-CaTr-BERT 使用模型生成的图片描述作为辅助句子替代图片,与文本模态融合,而图片描述会包含一定的不包含情感的词汇,造成了信息冗余,因此干扰了分类结果。对于 MSVA Single 数据集的 SOTA 模型,即 VAuLT 模型使用深度神经网络提取特征向量,使用了简单的加权求和,导致映射到向量空间中的数据不准确,因此干扰了分类结果。

在 Twitter15 数据集上,本文模型 VSPL 相比 SOTA 模型分别实现了 0.98%和 0.14%的性能提升。在 Twitter17 数据集上,VSPL 相对于 SOTA 模型分别实现了 1.04%和 2.06%的性能提升。在 MVSA Single 数据集上,VSPL 在准确率和 F1 得分方面分别以 0.71%和 0.22%的优势优于现有的 SOTA 模型。

结果表明,本文提出的模型 VSPL 利用阿里云场景识别 API 和表情识别 API 将图像模态转化为精确简短、蕴含情感的视觉语义词汇,相比将图像模态转化图片描述的方法更精炼;然后与精心设计的提示模板组成新文本,而提示模板相对于图片描述的辅助句子能更精炼、有效地输出与情感相关的词汇,进一步与文本模态组成纯文本的形式;最后输入经过大量情感文本数据集训练得到的预训练语言模型进行分类,既能避免加权求和操作导致的数据映射不准确的问题,又能利用提示学习的方法,激发预训练语言模型的潜力,以进行更准确的情感预测。

4.6 消融实验

为了进一步验证模型 VSPL 的有效性,本文在两个 Twitter 数据集上进行了消融实验,分别基于 VSPL 模型删除了视觉语义以及提示模板。这两个实验在图 2 中分别表示为"del VS"和"del PT",实验结果如图 2 所示。

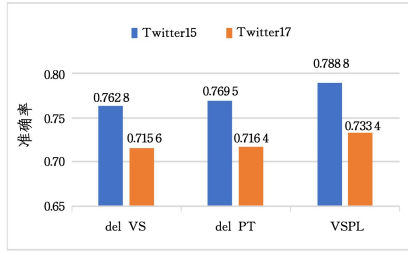


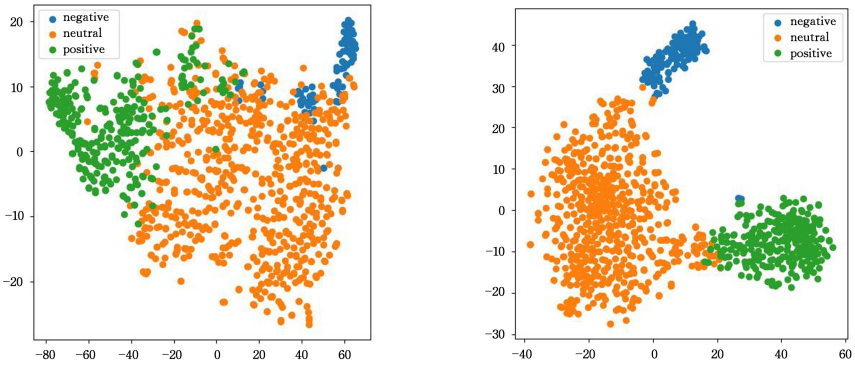
图 2 消融实验结果

Fig. 2 Results of ablation experiment

从这些结果中可以观察到:(1)由所有模块组成的方法,即 VSPL,在 3 个数据集上实现了最佳性能,移除任何一个模块都将导致预测结果下降。(2)del VS 的准确率比 VSPL 低,这表明了提取图像模态中包含情感要素的视觉语义词汇并与原始文本进行拼接的有效性。图片的情感要素直接作为视觉语义词汇,与原始文本拼接后能更合理地在特征空间进行数据映射,进而提升分类性能。(3)与 VSPL 相比,del PT 准确率较低,表明利用提示学习的方法,设计与情感分类任务相关的模板,可以更有效地激发预训练语言模型对于情感分析任务的潜力,进而进行更准确的情感预测。消融实验结果表明,VSPL 模型中的每个模块都是不可或缺的,它们共同为模型的最终性能做出贡献。

4.7 t-SNE 可视化特征

为了证明 VSPL 模型在数据映射到特征空间方面的有效性,我们将 Twitter15 数据集的同一组测试集分别输入训练好的 VAuLT 模型和 VSPL 模型,并使用 t-SNE 算法[41]可视化了对应的特征。对于 VAuLT 模型,取加权求和之后的特征;对于 VSPL 模型,取掩蔽位〈mask〉处的特征。如图 3 所示,VSPL 模型相对于 VAuLT 模型在特征空间中能够更好地将同一类别的数据点聚集在一起,对于不同类别的数据点则能将它们更清晰地分开。因此,VSPL 模型能有效避免加权求和导致的特征空间映射不准确的问题,进而提升模型的分类性能。



(a)VAuLT (b)VSPL

图 3 使用 t-SNE 可视化数据

Fig. 3 Data visualization using t-SNE

4.8 样例分析及可视化

为了更好地理解本文方法的合理性，选择了几个样例进行可解释分析，其中借鉴了先进的可解释性方法[42]。获取训练好的 VSPL 模型后，首先输入多模态数据，获取视觉语义词汇，并嵌入提示模板，组成新文本，最后再输入模型得到预测结果。使用可解释方法对新文本进行解释可以生成热力图，通过热力图的方式，可以直观地看到新文本中各个部分对预测结果造成的影响，其中颜色越深代表关注度越高，红色代表积极，绿色代表消极。样例细节以及可视化结果如表 5 所列。为了区分图像模态和文本模态，在热力图中使用黑色框体框出了将图像模态提取为视觉语义词汇并与提示模板组合的部分，其余为原始文本。

表 5 样例分析(电子版为彩图)

Table 5 Sample analysis

示例	图像	*prompt*	情感预测	热力图
1		happiness	积极	¡s¿ A special My Car Monday to Jordan for her new with the cute Teddy Bear ! Congrats on your new purchase ! . a happiness make people feel ¡mask¿ . ¡/s¿
2		seaside	积极	¡s¿ We are thrilled to announce the plans for our newest ultra luxury resort , One amp Only Bahrain . a seaside make people feel ¡mask¿ . ¡/s¿
3		train	消极	¡s¿ absolute disgrace two car riages from Bang or half way there standing room only # disgrace . a train make people feel ¡mask¿ . ¡/s¿
4		other	消极	¡s¿ # dep ressed # dep ression qu otes # death # dep ression . a other make people feel ¡mask¿ ¡/s¿
5		other	积极	¡s¿ Awesome record . Keep in the night energetic . a other make people feel ¡mask¿ . ¡/s¿
6		other	积极	¡s¿ # February # Winter # Rain y # Storm y # Wind y # Wednesday # Even ing # Love # Happy # Pos itive # Passion ate # Cal m # M # C off ee . a other make people feel ¡mask¿ ¡/s¿

例如，示例 1 中，VSPL 模型的预测结果为积极，图片提取得到的视觉语义词汇 *prompt* 为“happiness”。对于图片，显然人的笑容是判断情感极性的重要依据，对应新文本中的“happiness”，还有原始文本中的“congrats”，VSPL 模型关注到了这些单词。示例 2 中，图片提取得视觉语义词汇 *prompt* 为“seaside”。“seaside”意为海滨，是一种景点，会隐性地给人一种休闲舒适的感觉，因此 VSPL 模型对“seaside”这个单词给予了部分关注。但是 VSPL 模型更关注文本中的单词“thrilled”，“thrilled”意为兴奋，所以 VSPL 模型判定情感为积极是合理的。示例 3 中，图片提取得到的视觉语义词汇 *prompt* 为“train”。“train”的情感极性为中性，但也可以给人旅行途中的积极印象，或者给人混乱拥挤的消极印象。文本中的单词“disgrace”赋予了消极的印象，因此 VSPL 模型对“train”这个单词给予了部分关注，但是模型更关注文本中的单词“disgrace”——耻辱，故 VSPL 模型判定情感为消极是合理的。示例 4 中的图像给人一种阴郁的印象，但提取得到的视觉语义词汇 *prompt* 为“other”，因此情感极性主要由文本判断。再比如包含许多实体的图像示例 5 与示例 6 中，图像描述的方法可能会提取到许多不包含情感信息的视觉语义词汇，造成信息冗余，而 VSPL 模型将这些不包含情感的图像归纳为代表中立的语义词汇“other”，以此降低图像模态对分类结果的影响，并将分类重心转移到文本模态。

结束语 现有的多模态情感分析模型或从不同模态中提取特征向量并粗糙地进行加权求和，导致数据无法准确地映射到统一的多模态向量空间中，或依赖图像描述模型将图像转化为文本，提取到过多不包含情感信息的视觉语义，造成

信息冗余，最终影响模型的性能。为了解决上述问题，本文提出了一种基于视觉语义与提示学习的多模态情感分析模型VSPL，该模型专注于将图像转化为精确简短、蕴含情感信息的视觉语义词汇，并结合提示学习的方法来提升模型的分类性能，在3个公开数据集上的性能超越了基准模型。近年来，诸多研究发现提示工程可以用来改进大规模预训练模型的性能，例如ChatGPT，在未来的工作中，将设计有更效的文本提示来引导模型进行情感分析。

参 考 文 献

[1] YUE L, CHEN W, LI X, et al. A survey of sentiment analysis in social media [J]. Knowledge and Information Systems, 2019, 60: 617-663.

[2] GAO Y, ZHEN Y, LI H, et al. Filtering of brand-related microblogs using social-smooth multiview embedding [J]. IEEE Transactions on Multimedia, 2016, 18(10): 2115-2126.

[3] PANG L, ZHU S, NGO C W. Deep multimodal learning for affective analysis and retrieval [J]. IEEE Transactions on Multimedia, 2015, 17(11): 2008-2020.

[4] CAMBRIA E, SCHULLER B, XIA Y, et al. New avenues in opinion mining and sentiment analysis [J]. IEEE Intelligent Systems, 2013, 28(2): 15-21.

[5] GUO W, ZHANG Y, CAI X, et al. LD-MAN: Layout-driven multimodal attention network for online news sentiment recognition [J]. IEEE Transactions on Multimedia, 2020, 23: 1785-1798.

[6] ZHU T, LI L, YANG J, et al. Multimodal sentiment analysis with image-text interaction network [J]. IEEE Transactions on Multi-media, 2023, 40(1): 1-27.

[7] KHAN Z, FU Y. Exploiting BERT for multimodal target sentiment classification through input space translation [C]// Proceedings of the 29th ACM International Conference on Multimedia. 2021: 3034-3042.

[8] MACHAJDIK J, HANBURY A. Affective image classification using features inspired by psychology and art theory [C]// Proceedings of the 18th ACM International Conference on Multimedia. 2010: 83-92.

[9] ZHAO S, GAO Y, JIANG X, et al. Exploring principles-of-art features for image emotion recognition [C]// Proceedings of the 22nd ACM International Conference on Multimedia. 2014: 47-56.

[10] BORTH D, JI R, CHEN T, et al. Large-scale visual sentiment ontology and detectors using adjective noun pairs [C]// Proceedings of the 21st ACM International Conference on Multimedia. 2013: 223-232.

[11] LI Z, FAN Y, LIU W, et al. Image sentiment prediction based on textual descriptions with adjective noun pairs [J]. Multimedia Tools and Applications, 2018, 77: 1115-1132.

[12] ZHAO S, YAO H, YANG Y, et al. Affective image retrieval via multi-graph learning [C]// Proceedings of the 22nd ACM International Conference on Multimedia. 2014: 1025-1028.

[13] YOU Q, JIN H, LUO J. Visual sentiment analysis by attending on local image regions [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2017.

[14] SHE D, YANG J, CHENG M M, et al. Wscnet: Weakly supervised coupled networks for visual sentiment classification and detection [J]. IEEE Transactions on Multimedia, 2019, 22(5): 1358-1371.

[15] ZHANG J, LIU X, CHEN M, et al. Image sentiment classification via multi-level sentiment region correlation analysis [J]. Neurocomputing, 2022, 469: 221-233.

[16] YANG J, LI J, WANG X, et al. Stimuli-aware visual emotion analysis [J]. IEEE Transactions on Image Processing, 2021, 30: 7432-7445.

[17] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv: 1810.04805, 2018.

[18] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners [J]. OpenAI blog, 2019, 1(8): 9.

[19] JIANG Z, XU F F, ARAKI J, et al. How can we know what language models know? [J]. Transactions of the Association for Computational Linguistics, 2020, 8: 423-438.

[20] SHIN T, RAZEGHI Y, LOGAN IV R L, et al. Autoprompt: Eliciting knowledge from language models with automatically generated prompts [J]. arXiv: 2010.15980, 2020.

[21] ZHANG R, GUO Z, ZHANG W, et al. Pointclip: Point cloud understanding by clip [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 8552-8562.

[22] ZHOU K, YANG J, LOY C C, et al. Learning to prompt for vision-language models [J]. International Journal of Computer Vision, 2022, 130(9): 2337-2348.

[23] ZHOU K, YANG J, LOY C C, et al. Conditional prompt learning for vision-language models [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 16816-16825.

[24] WU H, SHI X. Adversarial soft prompt tuning for cross-domain sentiment analysis [C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022: 2438-2447.

[25] KAUR R, KAUTISH S. Multimodal sentiment analysis: A survey and comparison [J]. International Journal of Service Science Management Engineering & Technology, 2019, 10(2): 38-58.

[26] SOLEYMANI M, GARCIA D, JOU B, et al. A survey of multimodal sentiment analysis [J]. Image and Vision Computing, 2017, 65: 3-14.

[27] MORENCY L P, MIHALCEA R, DOSHI P. Towards multimodal sentiment analysis: Harvesting opinions from the web [C]// Proceedings of the 13th International Conference on Multimodal Interfaces. 2011: 169-176.

[28] ZADEH A A B, LIANG P P, PORIA S, et al. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph [C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 2236-2246.

[29] STAPPEN L, SCHUMANN L, SERTOLLI B, et al. Muse-toolbox: The multimodal sentiment analysis continuous annotation

fusion and discrete class transformation toolbox [C]// Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge. 2021:75-82.

[30] LIANG C,XU J,ZHAO J,et al. Deep Learning-based Construction and Processing of Multimodal Corpus for IoT Devices in Mobile Edge Computing[J]. Computational Intelligence and Neuroscience,2022,2022(1):2241310.

[31] NIU T,ZHU S,PANG L,et al. Sentiment analysis on multi-view social data [C]// 22nd International Conference MultiMedia Modeling(MMM 2016),Miami,FL,USA,Part II 22. Springer International Publishing,2016:15-27.

[32] YU J,JIANG J. Adapting BERT for target-oriented multimodal sentiment classification [C]// IJCAI,2019.

[33] XU N,MAO W. Multisentinet:A deep semantic network for multimodal sentiment analysis [C]// Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2017:2399-2402.

[34] XU N,MAO W,CHEN G. A co-memory network for Multimodal sentiment analysis [C]// The 41st international ACM SIGIR Conference on Research & Development in Information Retrieval. 2018:929-932.

[35] CAI G,XIA B. Convolutional neural networks for multimedia sentiment analysis [C]// 4th CCF Conference Natural Language Processing and Chinese Computing(NLPCC 2015). Nanchang,China:Springer International Publishing,2015:159-167.

[36] CHOCHLAKIS G,SRINIVASAN T,THOMASON J,et al. VAuLT: Augmenting the Vision-and-Language Transformer with the Propagation of Deep Language Representations [J]. arXiv:2208.09021,2022.

[37] KIM W,SON B,KIM I. Vilt:Vision-and-language transformer without convolution or region supervision [C]// International Conference on Machine Learning. PMLR,2021:5583-5594.

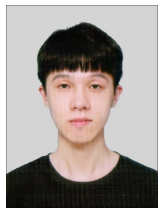
[38] LIU Y,OTT M,GOYAL N,et al. Roberta:A robustly optimized bert pretraining approach [J]. arXiv:1907.11692,2019.

[39] LI J,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.

[40] YE J,ZHOU J,TIAN J,et al. Sentiment-aware multimodal pretraining for multimodal sentiment analysis [J]. Knowledge-Based Systems,2022,258:110021.

[41] VAN DER MAATEN L,HINTON G. Visualizing data using t-SNE [J]. Journal of Machine Learning Research,2008,9(11):2579-2605.

[42] CHEFER H,GUR S,WOLF L. Transformer interpretability beyond attention visualization [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021:782-791.



MO Shuyuan, born in 1997, postgraduate. His main research interest is multimodal deep learning.

MENG Zuqiang, born in 1974, Ph. D, professor, is a senior member of CCF (No. 06312S). His main research interests include multimodal deep learning and granular computing.

(责任编辑:何杨)