计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(2):94−104 [doi: 10.15888/j.cnki.csa.009385]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于注意力融合网络的方面级多模态情感分类①

冼广铭, 招志锋, 阳先平

(华南师范大学 软件学院, 佛山 528225)
通信作者: 冼广铭, E-mail: xgm20011@163.com

摘　要: 方面级多模态情感分类任务的一个关键是从文本和视觉两种不同模态中准确地提取和融合互补信息, 以检测文本中提及的方面词的情感倾向. 现有的方法大多数只利用单一的上下文信息结合图片信息来分析, 存在对方面和上下文信息、视觉信息的相关性的识别不敏感, 对视觉中的方面相关信息的局部提取不够精准等问题, 此外, 在进行特征融合时, 部分模态信息不全会导致融合效果一般. 针对上述问题, 本文提出一种注意力融合网络 AF-Net 模型去进行方面级多模态情感分类, 利用空间变换网络 STN 学习图像中目标的位置信息来帮助提取重要的局部特征; 利用基于 Transformer 的交互网络对方面和文本以及图像之间的关系进行建模, 实现多模态交互; 同时补充了不同模态特征间的相似信息以及使用多头注意力机制融合多特征信息, 表征出多模态信息, 最后通过 *Softmax* 层取得情感分类的结果. 在两个基准数据集上进行实验和对比, 结果表明 AF-Net 能获得较好的性能, 提升方面级多模态情感分类的效果.

关键词: 多模态; 情感分类; 空间变换网络; 交互网络; 相似信息; 注意力融合网络

引用格式: 冼广铭,招志锋,阳先平.基于注意力融合网络的方面级多模态情感分类.计算机系统应用,2024,33(2):94–104. http://www.c-s-a.org.cn/1003-3254/9385.html

Aspect-level Multimodal Sentiment Classification Based on Attention Fusion Network

XIAN Guang-Ming, ZHAO Zhi-Feng, YANG Xian-Ping

(School of Software, South China Normal University, Foshan 528225, China)

Abstract: One of the key tasks of aspect-level multimodal sentiment classification is to accurately extract and fuse complementary information from two different modals of text and vision, so as to detect the sentiment orientation of the aspect words mentioned in the text. Most of the existing methods only use single context information combined with image information for analysis, revealing the problems such as insensitive to the recognition of the correlation between aspect-, context- and visual-information, and imprecise in local extraction of aspect-related information in vision. In addition, when performing feature fusion, insufficient partial modal information will lead to mediocre fusion effect. To solve the above problems, an attention fusion network AF-Net model is proposed to perform aspect-level multimodal sentiment classification in this study. The spatial transformation network (STN) is used to learn the location information of objects in images to help extract important local features. The Transformer based interaction network is used to model the relationship between aspects, texts and images, and realize multi-modal interaction. At the same time, the similar information between different modal features is supplemented and the multi-feature information is fused by multi-attention mechanism to represent the multi-modal information. Finally, the result of sentiment classification is obtained through *Softmax* layer. Experiments and comparisons carried out on the two benchmark datasets show that AF-Net can achieve better performance and improve the effect of aspect-level multimodal sentiment classification.

① 基金项目: 国家自然科学基金 (61070015)
收稿时间: 2023-08-01; 修改时间: 2023-09-01; 采用时间: 2023-09-15; csa 在线出版时间: 2023-12-25
CNKI 网络首发时间: 2023-12-27

Key words: multimodal; sentiment classification; spatial transformation network (STN); interaction network; similar information; attention fusion network

1 引言

方面级情感分类 (aspect-based sentiment classification, ABSC) 是情感分类中的一项细粒度分类任务, 其研究目的是判断句子在不同方面词下对应的情感倾向 (积极、消极和中性), 而这些方面词一般是来自于句子中的目标实体[1]. 例如, “Former top John McCain aide says he’d back Hillary Clinton over Donald Trump.” 该句子来自于 Twitter 2017 数据集, 句子中对于“Hillary Clinton”方面词表达积极的情绪, 而对于“Donald Trump”方面词表达消极的情绪.

随着众多社交媒体平台的发展, 用户在社交媒体平台上发布的内容越来越多元化, 越来越多的用户更倾向于在文本的基础上, 附加一些图片来表达自己的情感和意见. 因此, 把文本和附加的图片结合起来进行融合提取信息, 能够更好地分析用户表达的情感极性. 现在社交媒体上用户发表的评论观点几乎都是文本和图片的多模态结合内容, 所以, 情感分类任务不仅仅是针对纯文本的单模态内容, 还需要考虑图片等视觉模态数据, 里面同样包含了丰富的情感信息. 到目前为止, 基于方面的情感分类任务主要集中于文本, 而在多模态的情感分类任务上所做的研究有限, 不同模态之间的信息不是简单的冗余关系, 而可能是互补、对比、增强等关系. 例如图 1(a) 中, 上下文信息的“smile”和图像中灿烂的笑容都明示了积极的情绪, 判断实体“Kanye”的情感极性时将两个模态的数据共同结合可以有效提高准确率. 在图 1(b) 中, 文本中对实体“Geordi”并没有表达明显的情感, 但可以通过捕捉对应图像中的笑容特征来判断出对该实体表达了积极的情感. 在图 1(c) 中, 图像信息并未明显表达情感, 但文本中“greate”则明示了对实体“Eric Afable”的情感极性是积极.

基于方面的多模态情感分类任务中, 一个关键问题是如何从不同模态中准确地提取和融合互补信息, 但是现有的方法只利用单一的上下文信息结合图片信息来分析基于方面的多模态情感分类情况, 对方面和上下文信息、视觉信息的相关性的识别不敏感, 具体来说, 对于视觉信息的利用, 我们应该着重关注于图片中有效的情感信息部分, 例如图 1(a) 和图 1(b) 人物的笑容信息局部特征是要着重被提取的, 而图像中的无关背景信息属于噪声, 我们在细粒度上进行模态交互时就需要充分考虑方面词与有效模态信息之间的关联性, 最后在特征融合层面把经过细粒度交互的不同模态信息进行结合、对齐并表征出多模态信息. 为此, 本文提出一种基于注意力融合网络的情感分类模型 (attention fusion network, AF-Net) 用于方面级多模态情感分类任务.

(a) RT@warriors: steph curry’s so good, even $[\text{Kanye}]_{\text{positive}}$ smiles. #MVP

(b) Sherlock data, $[\text{Geordi}]_{\text{positive}}$ and I at star trek Las Vegas. #STLV#STLV2014#cosplay

(c) $[\text{Eric Afable}]_{\text{positive}}$ looking great for his wedding day today

图 1 方面级多模态情感分类示例

具体来说, 我们使用嵌入位置信息的 *BERT* 模型来提取上下文和方面词的特征, 并使用 GRU 进一步处

理短文本序列的方面词. 为了重点捕捉图像中可能包含的一些与方面相关的情感信息, 例如人的表情, 我们使用 STN 空间变换网络进行视觉模态的深度局部特征提取, 通过学习图像的空间位置信息来提取重要的局部特征, 然后我们使用带有多头注意力机制的 Transformer 交互网络来建模方面词和上下文信息以及视觉信息之间的关联, 生成方面相关的文本特征和方面相关的视觉特征. 考虑到相似信息可以通过增强模态之间的一致性和互补性, 我们在特征融合阶段通过使用标准欧氏距离计算两种特征的相似度, 将其作为相似信息和前面得到的特征信息一起经过注意力机制的动态整合, 从而得到多模态联合表示, 有效提升方面级多模态情感分类的效果.

本文的主要贡献如下.

(1) 本文提出了一种注意力融合网络, 该模型创新性地使用空间变换网络 (STN) 来学习图像中目标的空间位置信息, 以进一步提取重要的局部视觉特征, 放大感兴趣的视觉线索. 同时, 该网络实现了模态内和模态间的交互, 能够生成方面词关联的文本和视觉特征.

(2) 为了整合多个特征信息, 网络中设计了一个多特征融合模块, 由相似信息计算和多头注意力机制组成, 该模块重点补充了不同模态间的相似信息, 有效融合文本和视觉的相关特征, 得到更具表达能力和综合性的多模态特征表示.

(3) 在两个 Twitter 基准数据集上的实验表明, 与一些基线模型相比, 本文提出的模型在方面级多模态情感分类任务中具有一定的竞争力.

2 相关工作

方面级多模态情感分类任务是方面级情感分类和多模态情感分类两方面研究的结合.

2.1 方面级情感分类

作为情感分析中的一项重要任务, 近年来, 方面级情感分类得到了广泛研究. 一项工作侧重于利用外部资源手动设计一组特定于任务的特征, 然后将传统的统计学习方法应用于情感预测的特征. 另一项工作方向是将目标信息纳入各种神经网络 (NN) 模型, 包括基于递归神经网络的方法[2]、基于循环神经网络的方法[3]和基于卷积神经网络的方法[4]. 受其他 NLP 任务中注意力机制优势的启发, 许多研究设计了不同的基于注意力机制的方法来模拟方面词和上下文之间的交互作用[5]. 然而, 这些研究没有考虑到视觉特征可能会促进这些基于文本的方法.

2.2 多模态情感分类

随着社交媒体的迅猛发展, 情感分析已经超越了传统的仅基于纯文本的研究范畴, 目前可以利用不同来源的模态信息来补充文本信息, 进行更全面的情感表达分析[6]. 对于图文多模态情感分类的研究, 早期主要采用基于特征的方法, 例如, Borth 等人[7]提取图像中的 1 200 个形容词和名词对来生成视觉特征, 并根据英语语法和拼写风格计算文本的情感分数来生成文本特征, 然后将两者结合起来进行情感分类. Xu 等人[8]提出从图像中提取场景和物体特征, 然后利用这些视觉语义特征去建模图像对文本的影响, 这种利用图像信息的方式能够更好发挥图像语义信息的作用. 随着深度学习技术的发展, 许多基于神经网络的多模态情感分类模型被提出, 并取得了显著的成功. Yu 等人[9]使用卷积神经网络和文本卷积神经网络分别从图像和文本中提取特征, 然后将两种模态特征直接拼接起来训练一个逻辑回归模型用于情感分类. 实际上, 在情感分类中, 文本信息与图像信息是相辅相成、相互作用的. Ju 等人[10]最近完成的一项端到端 MABSA 任务中也观察到, 文本通常比图像发挥更重要的作用, 而图像可能为文本提供重要线索[11]. 因此, Xu 等人[12]提出一种共注意力机制, 能够建模文本和图像之间的交互作用. Zadeh 等人[13]提出了一种张量融合网络 (TFN), 主要用于捕获单模态内部信息和跨模态交互信息. 林敏鸿等人[14]则针对多模态情感分类任务提出了一种基于注意力神经网络的模型 (ANNM), 该模型使用注意力机制提取图文的情感特征, 并且强调与情感信息最关联的部分, 然后再使用张量融合的方法将图文特征进行融合, 从而完成后续的多模态情绪分析. 除了张量融合的方法, 还有一些工作采用特征拼接、相加、相乘或注意力机制等方法来融合多模态信息. Huang 等人[15]提出了一种深度多模态注意力融合的方法用于图像和文本的情感分析 (DMAF), 该方法在 4 个真实数据集上的性能优于最先进的基线模型. Truong 等人[16]提出了一种视觉方面注意网络, 研究人员认为在这项任务中图像信息对文本信息是有辅助作用的, 具体是指它并不独立于文本来表达情感, 而是可以突出显示文本中某些实体, 所以该模型将视觉信息作为对齐方式, 利用注意力指示出文档中的重要句子, 进行情感分类[17]. Han 等人[18]基

于注意力机制提取两两模态间的互补信息, 并通过门控机制对单模态表示进行了改进. 宋云峰等人[19]使用跨模态注意力机制融合两两模态, 并使用自注意力机制提取不同层次的显著特征. 胡慧君等人[20]针对多模态情感分类任务还提出了一种基于图文语义相关的分类方法 (MSSA-SC), 在融合阶段的第 1 步先判断图像和文本是否存在关联关系, 如果发现图文语义不相关则只需要对文本信息进行情感分类, 否则进行的就是多模态情感分类.

2.3 方面级多模态情感分类

作为多模态情感分类的一项重要细粒度任务, 基于图文的方面级多模态情感分类旨在识别一对句子和图像中提到的每个目标方面的情感极性[21]. 现有的方法主要强调建模方面、文本和图像之间的成对交互[11]. Xu 等人[22]提出一个多交互记忆网络, 以构建跨模态和单模态之间的交互, 并充分捕捉方面对文本和图像的影响, 以及文本和图像之间的多种交互. Yu 等人[23]提出了一种基于 *BERT* 架构的多模态情感分类模型 (TomBERT), 该模型使用了多个基于 *BERT* 的模块, 其中底部的两个基于 *BERT* 的模块用于捕获模态内的动态特征, 实现方面、文本对齐和方面、图像对齐, 而另一个基于 *BERT* 的模块位于顶部, 用于捕获模态间的动态特征, 实现文本、图像对齐. 此外, Yu 等人[24]还提出建立一个实体敏感注意和融合网络 (ESAFN), 利用有效的注意力机制来生成实体敏感的文本表示和实体敏感的视觉表示, 然后引入门控机制来消除嘈杂的视觉上下文. Zhou 等人[25]提出了一个多模态极性预测网络, 将图像、文本和方面特征表示进行矩阵相乘操作来实现多模态交互. Zhang 等人[26]采用一对记忆网络来捕获模态内信息和提取不同模态之间的交互信息, 然后设计判别矩阵来监督模态信息的融合. Wang 等人[27]设计了一个用于方面级多模态情感分类任务的注意力胶囊提取和多头融合网络 (EF-Net), 通过多头注意力机制和胶囊网络的集成捕捉多模态输入之间的相互作用. 现有的模型偏向于利用图像的整体信息, 而忽略了一些局部特征, 例如面部情绪这种视觉情绪线索[28]. 基于以上研究, 文本提出了一种注意力融合网络, 能够充分利用图像的空间位置信息深度提取视觉信息的局部特征, 更加准确地捕获在细粒度层面上不同模态与方面词的相关性, 动态挖掘上下文全局特征、方面特征和视觉特征之间的互补增强关系. 此外, 还通过计算补充不同特征间的相似度信息以及配合多头注意力机制来更加精确地融合多模态特征, 挖掘多模态信息之间的深度交互, 最后获取多模态表示用于情感分类.

3 AF-Net

AF-Net 模型结构如图 2 所示, 主要由特征提取层、模态交互层、模态融合层、情感分类输出层 4 个部分组成. 其中, 特征提取层是对上下文、方面词和图片进行编码, 分别提取上下文特征表示、方面词特征表示和视觉特征表示, 然后模态交互层将上下文表示、方面词表示和视觉表示相结合, 使用多头注意力机制分别学习方面词与上下文信息的模态内交互和方面词与视觉信息的模态间交互. 为了能充分利用文本特征和视觉特征的互补关系, 模态融合层补充不同模态特征的相似信息, 并将其和之前得到的包括方面词-上下文关联表示以及方面词-视觉关联表示在内等多特征信息通过注意力机制进行动态融合, 得到多模态表示, 最后通过 *Softmax* 层来预测对方面词的一个情感极性.

3.1 任务定义

首先我们得到一组多模态样本 D, 对于每个样本 $c \in D$, 它包含一个带有 n 个单词$(w_1,\cdots,w_n)$的句子 S 和一个相关图像 I, 以及一个方面实体 A (来自 S 的单词或短语, 可以是多个), 对于方面实体 A, 它与情绪标签 y 相关, $y \in \{-1,0,1\}$, -1 代表消极情感, 0 代表中性情感, 1 代表积极情感. 整个问题就可以表述为: 以(S,I)对以及方面实体 A 之一作为输入, 学习到一个方面级的情感分类器, 去预测方面实体 A 的情绪取向 y.

3.2 特征提取层

对于文本输入, 我们将文本分为两部分, 一部分是上下文, 记为T_{context}, 一部分为要进行情感分类的方面实体T_{aspect}, 将上下文、方面词经过加“[CLS]”“[SEP]”的预处理.

3.2.1 上下文特征

我们直接用嵌入了位置信息的 *BERT* 模型处理上下文部分, 获取它们的隐藏表示, $H_c \in R^{d\times|C|}$为上下文特征表示:

$$H_c = BERT\,(T_{\text{context}}) \tag{1}$$

3.2.2 方面词特征

我们首先用 *BERT* 模型获得方面词的基本特征, 然后由于方面词都是短文本序列, 我们再将其馈送至

一个标准的 *GRU* 深度提取序列语义信息, 获得最终的方面词表示. 通过将 *BERT* 和 *GRU* 结合, 可以充分利用 *BERT* 对语义信息的提取能力以及 *GRU* 对时序信息的建模能力, 从而提高短文本处理的效果.

$$T_a = BERT\left(T_{\text{aspect}}\right) \tag{2}$$

$$H_a = GRU\left(T_a\right) \tag{3}$$

其中, $H_a \in R^{d\times|A|}$ 为方面词特征表示, d 是隐藏层维度, $|C|$ 和 $|A|$ 分别是输入的上下文、方面词的最大长度.

3.2.3 视觉特征

首先将输入图片 I 调整为 224×224 像素大小, 采用 ResNet-152 的残差网络将图像分割成个 7×7 的区域, 每个区域由一个 2048 维的向量表示, 具体公式如下:

$$R = ResNet(I) \tag{4}$$

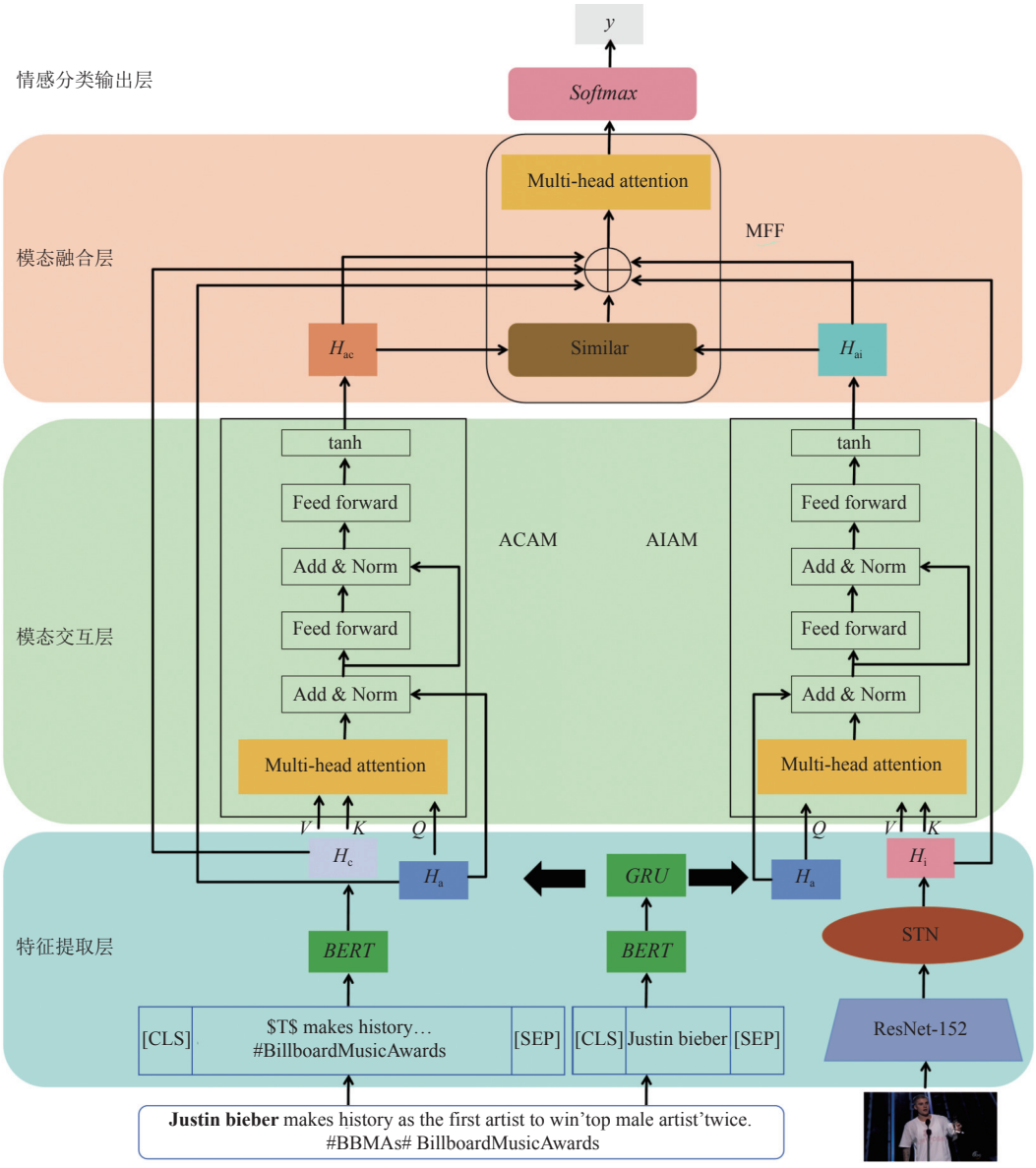

图 2 AF-Net 模型图

由于 *ResNet* 提取的图像特征没有处理图像中目标, 比如一些人物表情的位置信息, 因此 R 被馈送到一层空间变换网络 (STN) 中去学习图像中目标的位置信息, 并对特征进行深度的局部提取:

$$H_i = STN(R) \tag{5}$$

3.3 模态交互层

为了充分识别方面词与上下文信息和视觉信息之

间的相关性, 我们在模态交互层用两个基于 Transformer 结构的方面关联模块 ACAM 和 AIAM, 利用多头注意力机制建模方面词-上下文关联表示, 动态学习到与方面词最相关的文本特征信息, 此外, 还有建模方面词-视觉关联表示, 去除与方面词无关的视觉区域, 例如图像背景, 并保留最相关的部分, 整个模态交互过程实现了动态减轻无关上下文、视觉噪声的效果.

3.3.1 方面词-上下文关联表示

我们将 H_a 设置为 ACAM 的查询 Q, 将 H_c 设置为 ACAM 的键和值, 也就是 K=V 的情况下, 以方面词作为引导, 在语义信息与位置信息层面上去动态选择上下文中最重要的单词部分, 生成方面词-上下文关联表示, 整个过程是在完成方面信息与上下文信息的特定注意力交互, 具体用以下公式计算:

$$ATT_i(H_a,H_c)=Softmax\left(\frac{[Q_iH_a]^{\mathrm{T}}[K_iH_c]}{\sqrt{d/m}}\right)[V_iH_c]^{\mathrm{T}} \quad (6)$$

$$MATT(H_a,H_c)=W_m[ATT_1(H_a,H_c),\cdots]^{\mathrm{T}} \quad (7)$$

$$Z=LN(H_a+MATT(H_a,H_c)) \quad (8)$$

$$H_{ac}=\tanh(LN(FFN(Z)+Z)) \quad (9)$$

其中, ATT_i 是单个头的注意力机制输出, $\{Q_i,K_i,V_i\}\in R^{d/m\times d}$ 分别为查询、键和值的权重矩阵, $MATT(H_a,H_c)$ 是经过多头注意力机制处理后的输出, $W_m\in R^{d\times d}$ 是该过程随机生成的权重矩阵, LN 是层归一化, Z 是经过残差连接和层归一化处理输出的中间值, FFN 是一层前馈神经网络, tanh 是非线性激活函数, 是经过一系列处理得到的方面词-上下文关联表示.

3.3.2 方面词-视觉关联表示

同样, 我们也是将 H_a 设置为 AIAM 的查询 Q, 将 H_i 设置为 AIAM 的键和值, 也就是 K=V 的情况下, 以方面词作为引导, 动态选择视觉信息中最重要的局部特征, 生成方面词-视觉关联表示, 整个过程是在完成方面信息与视觉信息的特定注意力交互, 具体用以下公式计算:

$$ATT_i(H_a,H_i)=Softmax\left(\frac{[Q_iH_a]^{\mathrm{T}}[K_iH_i]}{\sqrt{d/m}}\right)[V_iH_i]^{\mathrm{T}} \quad (10)$$

$$MATT(H_a,H_i)=W_m[ATT_1(H_a,H_i),\cdots]^{\mathrm{T}} \quad (11)$$

$$Z=LN(H_a+MATT(H_a,H_i)) \quad (12)$$

$$H_{ai}=\tanh(LN(FFN(Z)+Z)) \quad (13)$$

其中, H_{ai} 是方面词-视觉关联表示.

3.4 模态融合层

经过前面的处理, 我们通过方面和文本交互以及方面和图像交互提取了具有针对性的与方面词关联的重要文本表示和视觉表示, 为了更好地将文本表示和视觉表示相结合, 我们设计了一个多特征融合模块 MFF 来对不同模态特征进行融合. 为了避免忽略一些特征之间的非关键互补信息, 我们考虑用标准化欧氏距离计算两个与方面关联的特征向量的相似度信息, 生成文本-视觉相似特征, 然后将上下文特征、方面词特征、视觉特征、方面词-上下文关联的文本特征、方面词-视觉关联的视觉特征以及文本-视觉相似特征这几部分特征信息在全局上进行一个整合, 从而提高特征的整体质量, 然后将整合后的特征馈送到多头自注意力机制模块中去动态学习每部分信息的权重, 最后输出一个更具表达能力和综合性的特征表示, 作为最终的多模态特征表示. 这样可以更好地捕捉到不同特征信息之间的相互作用和依赖关系, 能够更全面地利用多个特征之间的信息. 具体计算公式如下:

$$H_s=\sqrt{\sum_{k=1}^{N}\frac{\left(H_{ac_k}-H_{ai_k}\right)^2}{s_k^2}} \quad (14)$$

$$H=concat(H_c,H_a,H_{ac},H_{ai},H_s) \quad (15)$$

$$H_{mm}=MultiHead\text{-}Attn(H,H,H) \quad (16)$$

其中, H_s 是通过标准欧氏距离计算出来的相似特征信息, H 是多特征信息联合表示, H_{mm} 是经过多头自注意力机制处理输出的多模态表示, 用于最后的情感分类.

3.5 情感分类输出层

上述的最终表示被馈送到 *Softmax* 分类器中, 得到最终的情感极性标签分类, 即:

$$P=Softmax\left(W^{\mathrm{T}}H_{mm}+b\right) \quad (17)$$

其中, $W\in R^{d\times 3}$, $b\in R^3$ 分别是可训练的权重矩阵和偏差, P 为预测的情感值. 模型的损失函数采用标准的交叉熵损失函数, 并引入 L2 正则化:

$$Loss=-\frac{1}{N}\sum_{j=1}^{N}\log P^j+\lambda\|\theta\|_2 \quad (18)$$

其中, N 是情感极性的数量, y 是真实标签值, P 是预测标签值, θ 表示所有可训练的参数, λ 表示 L2 正则化的系数.

4 实验

4.1 数据集

AF-Net 主要采用来自于 Yu 等人[23]的 Twitter15 和 Twitter17 两个基准数据集来评估模型. Twitter15 和 Twitter17 分别包括 2014–2015 年和 2016–2017 年发布的推文. 这些数据集提供了每条推特的各个方面以及 3 种情绪极性标签. 表 1 汇总了数据集的统计数据.

表 1 两个 Twitter 数据集的统计数据

数据分类	Twitter15			Twitter17		
	Train	Dev	Test	Train	Dev	Test
Positive	928	303	317	1 508	515	493
Neutral	1883	670	607	1 638	517	573
Negative	368	149	113	416	144	168
Total aspects	3 179	1 122	1 037	3 562	1 176	1 234

4.2 实验评价指标及参数设置

实验采用方面级情感分类任务中常用的指标准确率 (*ACC*) 和宏平均 *F*1 值 (*Macro-F*1) 作为模型的最终的性能评价指标. 具体公式如下:

$$ACC = \frac{TP}{N} \times 100\% \tag{19}$$

$$F1 = 2 \times \frac{P \times R}{P + R} \times 100\% \tag{20}$$

$$Macro\text{-}F1 = \frac{1}{n} \sum_{i=1}^{n} F1_i \tag{21}$$

其中, *TP* 是正确预测的样本数量, *N* 是样本总数, *P* 表示精确率, 是预测出来为正类中真正的正类所占的比例, *R* 表示召回率, 是预测出来正确的正类占所有真实正类的比例, *F*1 是精确率和召回率的调和平均数, *n* 是情感类别数, 而 *Macro-F*1 是所有类别的 *F*1 值的平均[29].

我们基于 PyTorch 框架来实现所提出的 AF-Net 模型, 并在 NVIDIA A30 GPU 计算资源上进行训练和测试. 实验参数设置如表 2 所示.

表 2 实验参数设置

参数	设置值	说明
Embed_dim	768	文本嵌入向量维度
Hidden_dim	768	隐藏层维度
max_seq_length	64	文本的最大长度
max_entity_length	16	方面的最大长度
Dropout	0.1	Dropout
batch_size	32	批处理大小
learning_rate	3E–5	学习率
attention_head	12	注意力头数
image_dim	2048	图片嵌入维度
Optimizer	Adam	优化器

4.3 基线方法

为了验证我们模型的优越性, 考虑将我们的模型与经典的方面级情感分类方法和代表性的方面级多模态情感分类方法进行比较.

Res-Aspect: 该模型将使用 *ResNet* 提取的视觉特征和方面词嵌入向量做简单的拼接馈送到 *Softmax* 层进行情感分类.

Res-Target[23]: Yu 等人提出的一个对比变体, 将 *ResNet* 提取的图像特征和方面词嵌入向量拼接后输入到 *BERT* 用于方面级情感分类.

TD-LSTM[2]: 该模型使用两个 LSTM 结构学习方面词嵌入的左上下文和右上下文表示, 然后将其连接.

MemNet[30]: 该模型是使用方面词作为引导的记忆模型. 在词嵌入和位置嵌入的基础上, 使用多跳注意力机制更新存储的记忆, 得到交互后的文本表示和视觉表示.

IAN[31]: 该模型在 LSTM 的基础上设计一种交互式注意力机制, 可以获得基于上下文的方面词表示以及基于方面词的上下文表示, 并使用隐藏状态, 通过池化过程计算注意力分数.

MGAN[32]: 该模型设计了一种多粒度自注意力网络来以不同粒度捕捉方面和上下文的交互.

RAM[33]: 使用 Bi-LSTM 进行方面表示学习并在输出上构建多层注意力框架, 框架中每一层的注意力输出采用递归神经网络进行非线性组合.

Res-IAN、Res-MGAN、Res-RAM 这 3 个模型是前面 3 个处理文本的模型的变体, 使用了 *ResNet* 作为图像编码器提取视觉特征并将其分别与经过 IAN、MGAN 和 RAM 处理得到的文本特征做联合用于方面级多模态情感分类.

MIMN[22]: 该模型采用双向 LSTM 分别获取文本、方面词和图片的隐藏信息, 并设计一种多跳记忆神经网络对文本和图片之间的交互进行建模.

ESAFN[24]: 该模型设计了一种实体敏感的注意力与融合网络, 采用双向 LSTM 分别对方面短语的左右上下文进行建模, 并利用标准注意力机制和双线性融合整合得到文本表示和视觉表示. 此外, 还增加了一个门控机制以消除视觉模态中的噪声. 最后, 采用低秩双线性方法对多模态特征进行融合.

TomBERT[23]: 该模型由 4 个 *BERT* 组成, 第 1 个和第 2 个 *BERT* 用于提取文本特征, 第 3 个 *BERT* 用于

捕获方面和图像之间的交互, 第 4 个 *BERT* 用于多模态信息的融合.

EF-Net[27]: 该模型用多头自注意力机制、残差网络以及胶囊网络集成来处理文本和图像, 捕捉多模态输入之间的相互作用.

4.4 实验结果

我们在 Twitter15 和 Twitter17 这两个数据集上进行了模型的对比实验, 实验结果如表 3 所示.

表 3 AF-Net 和其他模型的对比结果 (%)

Modality	Model	Twitter15		Twitter17	
		ACC	*Macro-F*1	*ACC*	*Macro-F*1
Visual	Res-Aspect	60.08	33.18	59.89	54.63
	Res-Target	59.88	46.48	58.59	53.98
Text	TD-LSTM	68.30	61.43	60.67	56.97
	MemNet	70.11	61.76	64.18	60.90
	IAN	70.90	63.32	64.61	61.20
	RAM	70.68	63.05	64.42	61.01
	MGAN	71.17	64.21	64.75	61.46
Text+Visual	Res-IAN	71.85	63.84	66.53	63.35
	Res-MGAN	71.65	63.88	66.37	63.04
	Res-RAM	71.55	64.68	65.40	62.23
	MIMN	71.84	65.69	65.88	62.99
	ESAFN	73.38	67.37	67.83	64.22
	TomBERT	74.06	66.69	67.59	65.36
	EF-Net	73.65	67.90	67.77	65.32
	AF-Net	**76.37**	**71.55**	**70.91**	**69.48**

首先我们可以发现 Res-Aspect 和 Res-Target 的性能非常有限, 才大约 60% 的准确率, 这表明文本内容对于方面级情感分类是非常重要, 不能被忽略的. 其次, 通过比较基于文本的方法, 可以看出 TD-LSTM 的性能最差, 模型只将方面左右上下文嵌入连接起来, 没有考虑上下文与方面词之间的交互作用, 而 IAN 模型使用交互式注意力机制对基于上下文的方面词表示和基于方面词的上下文表示进行建模从而构造其内部关系, 这表明了建模上下文和方面之间的交互关系的重要性. 而表现最好的是 MGAN 模型, 说明多头自注意力机制比 LSTM 具有更强的捕捉上下文特征的能力, 以及多头自注意力机制与交互注意力机制相比, 可以更有效地捕获上下文和方面的关联特征, 即具有更强的交互信息建模的能力.

通过比较单模态和多模态方法, 可以看到结合图片信息的多模态方法比仅使用图片信息或文本信息的单模态方法都要好, 说明在方面级情感分类任务中, 融入视觉信息可以和文本信息进行互补, 从而增强情感分类的效果. 实验结果也证明, 多模态模型的性能明显高于基于纯图片或纯文本的单一模态模型. 例如, 从表 3 中可以看出, Res-IAN、Res-RAM、Res-MGAN 优于仅使用文本信息的 IAN、RAM、MGAN, 并且优于表中使用文本信息的 MemNet 模型, 这意味着关联图像确实能够为文本提供互补信息. 此外, 对于其他的多模态方法, 可以看出 MIMN 的效果也不错, 这表明将方面引导的文本和方面引导的视觉进行交互是有效的. ESAFN 模型根据方面词的位置对左右文本进行细分本, 效果明显优于 MIMN 模型, 充分说明了一般文本特征的提取不利于方面词和上下文相关性的识别, 需要从不同角度对文本进行细分和提取全局文本信息, 但它缺乏细粒度的多模态之间的交互, 所以它的效果并不理想. TomBERT 模型通过堆叠 *BERT* 来对方面和图像进行对齐和捕捉模态内的动态和模态间的交互. 然而 TomBERT 只将方面引导图像, 并没有将方面引导文本. EF-Net 模型总体上优于大多数基线方法, 这表明多头注意力机制在方面级多模态情感分类中的交互作用是更强的.

本文提出的 AF-Net 模型不需要区分有关方面的左上下文和右上下文, 使用 *BERT* 预训练模型并且嵌入位置信息, 比 LSTM 具有更好的捕获上下文特征的能力, 因为 LSTM 通常具有一定程度的远距离依赖性, 而我们的模型能够提取更全面且更具语义优势的文本特征信息. 而且我们还使用空间变换 STN 网络深度处理图像特征, 学习位置信息, 提取到更有利于情感分析的局部视觉特征, 比如人物表情, 一定程度上减轻了无关背景信息的噪声影响. 另外, 我们的模型使用基于 Transformer 的交互网络, 利用多头注意力机制来建模方面和文本以及方面和图像的交互, 并且增加多特征信息融合模块 MFF, 其中相似度信息计算模块 Similar 把文本特征和视觉特征的相似信息作为补充增强, 让其和之前获取到的多角度的特征信息先做整合, 之后再利用多头注意力机制将这些信息进一步融合得到多模态特征表示.

从表 3 可以看出, 与 EF-Net 模型相比, 我们的 AF-Net 模型在 Twitter15 数据集上准确率提升了 2.72%, *Macro-F*1 值提升了 3.65%, 在 Twitter17 数据集上准确率提升了 3.14%, *Macro-F*1 值提升了 4.16%. 总体而言, 结果表明我们提出的 AF-Net 模型是一种合理且有效的方面级多模态情感分类方法.

4.5 消融实验

为了深入分析 AF-Net 模型的不同组件对整个模型性能的影响, 我们设计了 4 组消融实验, 所使用的数据集仍然是 Twitter15 和 Twitter17, 评估指标是 *ACC* 和 *Macro-F*1, 实验的结果如表 4 所示.

表 4 消融实验 (%)

Ablations	Twitter15		Twitter17	
	ACC	*Macro-F*1	*ACC*	*Macro-F*1
w/o STN	75.22	70.13	69.29	66.15
w/o AIAM	74.83	69.42	69.45	66.56
w/o ACAM	75.12	69.43	70.01	67.61
w/o MFF	73.67	66.62	68.48	65.57
AF-Net	**76.37**	**71.55**	**70.91**	**69.48**

从表 4 可以看出, 当去除空间变换网络 STN 时, 模型的效果有所下降, 说明通过学习图像位置信息帮助提取局部的视觉特征对模型性能有一定的贡献. 此外, 去除任意一个基于 Transformer 的交互网络时, 模型的性能也会有所下降, 其中去除方面-视觉关联模块 AIAM 时性能下降得更多, 说明方面关联视觉的特征信息在该模型中提供了更多情感判断依据. 最后, 去除多特征融合模块 MFF, 模型的效果大大下降, 说明直接将特征信息级联融合的效果没有使用结合相似信息的注意力机制融合效果好, 我们提出的融合方式提高了模态的融合效果.

综上所述, AF-Net 模型中的各部分组件都对模型的效果有一定程度的影响和贡献.

4.6 模型参数分析

为了进一步提高模型预测的准确性, 我们对两个数据集训练时的批处理大小进行了研究, 批处理大小设置为 32 时, 训练效果较好, 如果超过该值, 训练的效率会降低, 小于该值会影响模型的收敛, 所以可以认为两个数据集的批处理大小设置成 32 能够使模型的训练达到较好的效果.

另外, 我们也研究了学习率变化对模型的影响, 以 Twitter17 数据集为例, 如图 3 所示, 从整体上看, 随着学习率的增加, *ACC* 值存在波动, *Macro-F*1 值呈现下降趋势, 当学习率为 3E–5 时模型的性能最好, *ACC* 和 *Macro-F*1 值达到最优, 学习率超过 5E–5 时, 对模型的负面影响开始出现, *Macro-F*1 值不断下降. 由此可见选择适当的学习率, 可以提高模型的性能和稳定性.

5 结论与展望

本文提出了一种注意力融合网络 AF-Net 模型用于方面级多模态情感分类任务. 该模型能通过 *BERT* 以及 *BERT* 联合 GRU 分别有效提取具有语义优势的文本的上下文特征信息和短文本序列方面词信息, 使用空间变换网络 (STN) 学习图像中目标的空间位置信息深度提取重要的视觉特征, 使用基于 Transformer 的交互网络充分建模不同模态与方面词之间的相关性, 此外, 利用不同模态的相似信息增强文本与图像之间的互补关系和注意力机制动态融合多个特征信息, 提升了最后的情感分类效果. 实验在 Twitter15 和 Twitter17 这两个数据集上的准确率和 *Macro-F*1 值均取得了较好的结果.

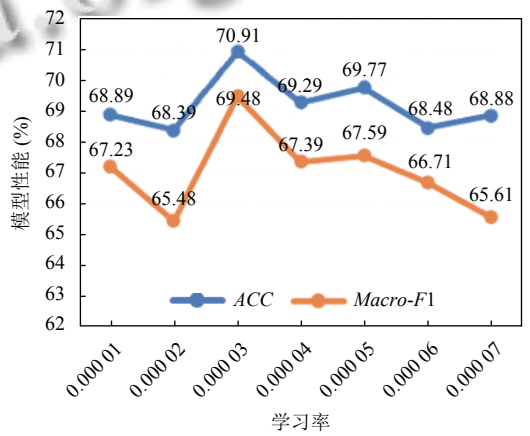

图 3 Twitter17 数据集上不同学习率对模型性能的影响

此外, 模型仍然存在可以继续改进的地方, 例如没有考虑到特征信息之间存在的语义差距问题对融合的影响, 接下来我们计划引入其他领域的知识来减轻这一问题的影响, 具体可以通过设计相应的数学公式来弥合模态间的语义差距并且将其与损失函数关联起来做进一步优化, 提升方面级多模态情感分类的效果.

参考文献

1 王家乾, 龚子寒, 薛云, 等. 基于混合多头注意力和胶囊网络的特定目标情感分析. 中文信息学报, 2020, 34(5): 100–110.

2 Dong L, Wei FR, Tan CQ, *et al*. Adaptive recursive neural network for target-dependent twitter sentiment classification. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore: ACL, 2014. 49–54.

3 Tang DY, Qin B, Feng XC, *et al*. Effective LSTMs for target-dependent sentiment classification. Proceedings of the 26th International Conference on Computational Linguistics:

Technical Papers. Osaka: ACL, 2016. 3298–3307.
4 Li X, Bing LD, Lam W, *et al*. Transformation networks for target-oriented sentiment classification. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: ACL, 2018. 946–956.
5 Wang S, Mazumder S, Liu B, *et al*. Target-sensitive memory networks for aspect sentiment classification. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: ACL, 2018. 957–967.
6 李丽, 李平. 基于交互图神经网络的方面级多模态情感分析. 计算机应用研究, 1–8. [doi: 10.19734/j.issn.1001-3695.2022.10.0532]
7 Borth D, Ji RR, Chen T, *et al*. Large-scale visual sentiment ontology and detectors using adjective noun pairs. Proceedings of the 21st ACM International Conference on Multimedia. Barcelona: ACM, 2013. 223–232.
8 Xu N, Mao WJ. MultiSentiNet: A deep semantic network for multimodal sentiment analysis. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. Singapore: ACM, 2017. 2399–2402.
9 Yu YH, Lin HF, Meng JN, *et al*. Visual and textual sentiment analysis of a microblog using deep convolutional neural networks. Algorithms, 2016, 9(2): 41. [doi: 10.3390/a9020041]
10 Ju XC, Zhang D, Xiao R, *et al*. Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana: ACL, 2021. 4395–4405.
11 Yang L, Na JC, Yu JF. Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis. Information Processing & Management, 2022, 59(5): 103038.
12 Xu N, Mao WJ, Chen GD. A co-memory network for multimodal sentiment analysis. Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. Ann Arbor: ACM, 2018. 929–932.
13 Zadeh A, Chen MH, Poria S, *et al*. Tensor fusion network for multimodal sentiment analysis. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: ACL, 2017. 1103–1114.
14 林敏鸿, 蒙祖强. 基于注意力神经网络的多模态情感分析. 计算机科学, 2020, 47(11A): 508–514, 548.
15 Huang FR, Zhang XM, Zhao ZH, *et al*. Image-text sentiment analysis via deep multimodal attentive fusion. Knowledge-based Systems, 2019, 167: 26–37. [doi: 10.1016/j.knosys.2019.01.019]
16 Truong QT, Lauw HW. VistaNet: Visual aspect attention network for multimodal sentiment analysis. Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu: AAAI Press, 2019. 305–312.
17 袁景凌, 丁远远, 盛德明, 等. 基于视觉方面注意力的图像文本情感分析模型. 计算机科学, 2022, 49(1): 219–224. [doi: 10.11896/jsjkx.201000074]
18 Han W, Chen H, Gelbukh A, *et al*. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. Proceedings of the 2021 International Conference on Multimodal Interaction. Montréal: ACM, 2021. 6–15.
19 宋云峰, 任鸽, 杨勇, 等. 基于注意力的多层次混合融合的多任务多模态情感分析. 计算机应用研究, 2022, 39(3): 716–720. [doi: 10.19734/j.issn.1001-3695.2021.08.0357]
20 胡慧君, 冯梦媛, 曹梦丽, 等. 基于语义相关的多模态社交情感分析. 北京航空航天大学学报, 2021, 47(3): 469–477. [doi: 10.13700/j.bh.1001-5965.2020.0451]
21 Yu JF, Wang JM, Xia R, *et al*. Targeted multimodal sentiment classification based on coarse-to-fine grained image-target matching. Proceedings of the 31st International Joint Conference on Artificial Intelligence. Vienna: IJCAI.org, 2022. 4482–4488.
22 Xu N, Mao WJ, Chen GD. Multi-interactive memory network for aspect based multimodal sentiment analysis. Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu: AAAI Press, 2019. 371–378.
23 Yu JF, Jiang J. Adapting BERT for target-oriented multimodal sentiment classification. Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao: IJCAI.org, 2019. 5408–5414.
24 Yu JF, Jiang J, Xia R. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 429–439. [doi: 10.1109/TASLP.2019.2957872]
25 Zhou J, Zhao JB, Huang JX, *et al*. MASAD: A large-scale dataset for multimodal aspect-based sentiment analysis. Neurocomputing, 2021, 455: 47–58. [doi: 10.1016/j.neucom.2021.05.040]
26 Zhang Z, Wang Z, Li XN, *et al*. ModalNet: An aspect-level sentiment classification model by exploring multimodal data with fusion discriminant attentional network. World Wide Web, 2021, 24(6): 1957–1974. [doi: 10.1007/s11280-021-00955-7]

27 Wang JQ, Gu DH, Yang C, *et al*. Targeted aspect based multimodal sentiment analysis: An attention capsule extraction and multi-head fusion network. arXiv:2103.07659, 2021.

28 Yang H, Zhao YY, Qin B. Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi: ACL, 2022. 3324–3335.

29 孟甜甜, 韩虎, 吴渊航. 面向方面抽取与情感分类的多任务联合建模. 计算机科学与探索, 2023, 17(7): 1669–1679.

30 Tang DY, Qin B, Liu T. Aspect level sentiment classification with deep memory network. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin: ACL, 2016. 214–224.

31 Ma DH, Li SJ, Zhang XD, *et al*. Interactive attention networks for aspect-level sentiment classification. Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne: AAAI Press, 2017. 4068–4074.

32 Fan FF, Feng YS, Zhao DY. Multi-grained attention network for aspect-level sentiment classification. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: ACL, 2018. 3433–3442.

33 Chen P, Sun ZQ, Bing LD, *et al*. Recurrent attention network on memory for aspect sentiment analysis. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: ACL, 2017. 452–461.

(校对责编: 牛欣悦)