文章编号：2096-1618(2022)05-0501-07

基于跨模态融合 ERNIE 的多模态情感分析研究

陶全桧， 安俊秀， 陈宏松
(成都信息工程大学软件工程学院，四川 成都 610225)

摘要：针对情感分析主要集中于单模态文本数据，忽略多模态数据融合问题，通过结合屏蔽多模态注意力方式，提出跨模态融合 ERNIE 的情感分析模型(CM-ERNIE)。首先，使用 CNN 和 BiGRU 提取音频数据特征以及词向量提取文本序列特征；其次，通过屏蔽多模态注意力作为 CM-ERNIE 的核心单元动态调整文本和音频数据权重，最后，文本和音频模态的交互作用微调预训练 ERNIE 模型。该模型在多模态电影评论观点数据集 CMU-MOSEI 和 CMU-MOSI 上评估。实验表明，模型在多模态数据集 CMU-MOSEI 和 CMU-MOSI 上评估该模型比单模态情感分析模型准确度高，并且多模态情感分析的研究蕴含巨大的价值，可为多模态场景下的情感分析、舆情分析和意图识别等实际应用问题提供决策支持。

关 键 词：多模态融合；预训练模型；注意力机制；ERNIE；文本分类
中图分类号：TP391.1 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2022.05.003

0 引言

随着新媒体技术的迅速发展，具有丰富情感的多模态数据也日益巨增，例如图片、短视频、音频和文本等，利用大量数据进行多模态情感分析已成为一个新兴领域，并且情感分析的研究有利于疫情防控。新冠肺炎疫情期间，国务院倡议要充分发挥科技支撑作用，运用技术手段积极有效地开展疫情防控，及时加强舆论引导，积极挖掘情感分析的研究价值。

早期情感分析任务主要使用单模态文本数据，首先使用传统的统计学方法提取词语特征进行文本表征，然后使用机器学习算法实现情感分类和预测，随后使用深度学习技术，例如卷积神经网络(convolution neural network，CNN)或词向量提取文本数据特征，解决特征提取困难问题，特别是传统统计机器方法无法解决大数据量的情况。然而目前这些方法只关注单模态文本数据，信息含量有限，数据特征质量低，在如今多媒体时代下很难通过单模态(文本信息)来准确地判断情绪，无法满足多模态的社交网络环境中情感分析问题。

已有的微调预训练模型方法可实现大规模音频与文本的联合表示。然而这类方法不能对上下文相关词加以区分，忽视了构建文本和音频上下词之间语义相关的重要性，导致预训练语言模型无法充分表示所需要的语义信息。最近，微调预训练语言模型 ERNIE(enhanced language representation with informative entities)作为一种高效的预训练语言模型，与传统的预训练语言模型不同，ERNIE 通过对所有层的上下文进行联合调节来生成上下文词特征表示。因此，单词的表征可表达文本上下文内容。ERNIE 在句子级[1]和分词级任务上都取得了较高的结果。然而，大多数微调策略仅基于单模态文本[2]设计，如何将其从单模态扩展到多模态并获得更好的表示，结合多模态信息进行实验研究是一个亟待解决的问题。

本文提出一种跨模态 Cross Modality ERNIE(CM-ERNIE)模型，即通过引入音频模态的信息，以帮助文本模态微调预训练 ERNIE 模型，进而进行多模态情感分析。Masked multi-modal attention 作为 CM-ERNIE 的核心单元，旨在通过跨模态交互动态调整词的权重。实验结果表明，CM-ERNIE 比以前的基线和 ERNIE 等的纯文本微调模型能较显著提高性能。

1 相关工作

1.1 多模态情感分析

多模态情感分析在不同模式之间具有内部相关性以及数据上下文具有时序相关性，多模态融合可以更有效全面地捕获情绪特征，结合不同模态数据的相关性以及互补性来进行情绪分析。多模态融合的关键点是如何有效地融合多模态之间的信息进行互补，目前主要的融合方式为特征层融合和决策层融合两种，特征层融合是通过连接和其他模态数据的有效特征来融合不同模态数据的特征或者补全不同模态之间的特征差异，由于不同特征交互融合，使情感信息更丰富，因

收稿日期：2021-12-29
基金项目：国家自然科学基金面上项目(71673032)；四川省社会科学研究规划项目(22XW043)

此可以显著地提高性能。不同模态融合可明显提高其分类效果,Borth 等[3]提出了利用词性对组合特征补充表达图像包含的语义信息。Guillaumin 等[4]发现图像特征结合文本特征信息(例如文本上下文与时序性)可获得更丰富的情感信息。多模态数据(图像与文本)在处理多模态数据分析可提高准确度[5]。考虑到上下文以及话语之间的关系,Poria 等[6]引入语境长短时期记忆网络,可以利用话语水平的话语情境信息来捕捉更多的情绪特征。随着注意力机制的普及以及它在多模态融合中起着越来越重要的作用,Tsai 等[7]在多模态转换模型中使用定向成对的跨模态注意。文献[8]通过跨时间步长的多模态序列的相互作用,并潜在地从一种模态调整到另一种模态。文献[9]通过对视频弹幕进行聚类分析,实现文本与视频的结合进行多模态情感分析。

1.2 预训练语言模型

微调预训练语言模型两种主要方法为基于特征的方法和基于微调预训练模型。

早期工作[10]专注于采用基于特征的方法,将单词转换为分布式表示。由于这些预训练的词表示捕获语料库中的句法和语义信息,通常用作输入嵌入和各种 NLP 模型的初始化参数,并提供对随机初始化参数的显著改进[11]。由于这些词级模型经常遭受多义词,Peters 等[12]采用序列级模型(ELMo)来捕捉跨不同语言的复杂词特征上下文。

随着人工智能技术的快速发展,Lai 等[13]提出了一种用于中文微博情感分类的图卷积神经网络体系结构,该体系的 F1 值达到了83.32%。Pal 等[14]用基于逻辑回归技术,对文本情绪(喜悦、愤怒、悲伤、悬念)进行分类,准确率为 73%。Puposh 等[15]用支持向量机(svm)对单模态文本进行情感六分类,获得 73% 的准确率。文献[16]用 Elmo 对单模态文本数据进行情感分类。文献[17]通过用 Bert 和 BiLSTM 结合模型,实现文本情感分类。文献[18]使 Bert 和 BiLSTM 相结合,对新媒体时代网络文本情绪趋向进行归类。文献[19]利用 Bert 与 Transformer 相结合,处理名词隐喻识别实现情感分类问题。

尽管基于特征和微调的语言表示模型都取得了很大的成功,但忽略了多模态预训练信息的融合。融合多模态信息可以显著提升原始模型学习能力,例如阅读理解[20]、机器翻译[21]、自然语言推理[22]、知识获取[23]和对话系统[24]。因此,融合信息可以有效地使现有的预训练语言模型受益。事实上,有些工作试图联合词和实体的表示学习,充分利用多模态信息并取得了可观的成果。Yu 等[25]提出了屏蔽语言的知识模型,引入场景图片模态信息增强语言表征。基于此,本文提出利用多模态语料库和多模态融合方式来训练基于 ERNIE 的模型。

2 方法论

提出的跨模态 ERNIE(CM-ERNIE),首先挖掘单模态文本以及音频内部的特征,对单模态文本及音频数据进行特征表示,并提取音频模态信息。然后,采用屏蔽多模态注意作为其核心,通过跨模态交互作用来动态调整单词的权重。结合来自文本和音频模态的信息微调预先训练过的 ERNIE 模型。

2.1 CM-ERNIE 模型

输入字符级别序列长度为 n 的文本序列:$T=[T_1, T_2, \cdots, T_n]$。由于 ERNIE 模型的嵌入层将在输入序列之前附加一个特殊的分类 embedding([CLS]),因此最后一个 encoder 层的输出是一个 $n+1$ 长度的序列,记为 $X_t=[E[\mathrm{CLS}], E_1, E_2, \cdots, E_n]$,为了与文本模态一致,在分词级任务上对齐音频特征之前附加一个零向量,对音频特征进行特征表示:$X_a=[A[\mathrm{CLS}], A_1, A_2, \cdots, A_n]$。其中,$A[\mathrm{CLS}]$是一个零向量,利用$X_t$和$X_a$之间的交互作用来调整每个单词的权重,以便更好地微调预先训练过的 ERNIE 模型,提高情绪分析的性能,模型的总体架构如图 1 所示。

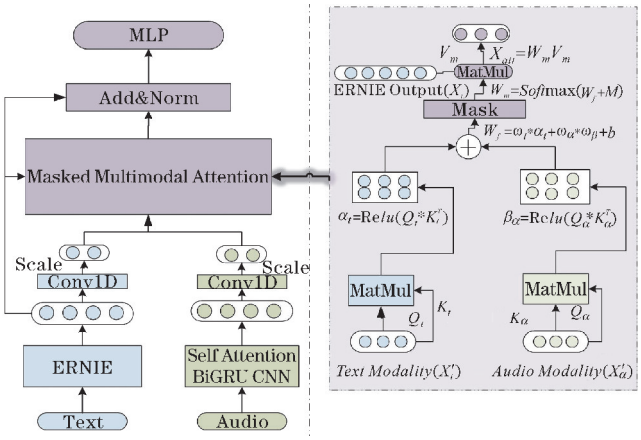

图 1 CM-ERNIE 模型图

2.2 模型细节

2.2.1 模态输入表征

CM-ERNIE 模型的输入包括两部分:字块令牌(word-piece tokens)的文本序列和字级(word-level)对齐音频特征。首先,文本序列将经过 ERNIE 模型,并使用最后一个 Encoder 层的输出作为文本特征,其定义为 $X_t=[E[\mathrm{CLS}], E_1, E_2, \cdots, E_n]$。其次,音频首先经过卷积神经网络:

$$X_a^{\mathrm{conv}}=\mathrm{Conv}(X_a)$$

将 CNN 处理后的音频数据输入到 BiGRU 中,提取与文本对应的音频特征:

$$X_{at}^{\text{conv}} = \text{BiGRU}(X_a^{\text{conv}})$$

由于单词级对齐音频特征 X_a^{conv} 的维度明显小于文本特征 X_t,因此使用了一个一维时间卷积层来控制它们到相同的维度,其中 $k\{t, a\}$ 表示文本和音频模特的卷积内核的大小。

$$\{\hat{X}_t, \hat{X}_a\} = \text{Conv1}D(\{X_t, X_{at}^{\text{conv}}\}, k\{t, a\})$$

因为 X_t 的维数明显高于 X_a^{conv},所以在训练过程中,$\hat{X}_t$ 的值会越来越大于 $\hat{X}_a$,为了防止点积变大,将文本特征 $\hat{X}_t$ 缩放到 $\hat{X}_t'$ 和音频特征 $\hat{X}_a$ 缩放到 $\hat{X}_a'$。

$$\hat{X}_t' = \frac{\hat{X}_t}{\sqrt{||\hat{X}_t||_2}}$$

$$\hat{X}_a' = \frac{\hat{X}_a}{\sqrt{||\hat{X}_a||_2}}$$

2.2.2　屏蔽多模态注意力

在得到 $\hat{X}_t'$ 和 $\hat{X}_a'$ 之后,为了使文本和音频信息充分交互,将它们输入到 Masked multi-modal attention 中,可以通过结合单词在不同模态下的表现来调整单词的权重。获得输出后,在 X_t 和 X_{Att} 上使用残差连接来保持数据的原始结构。然后,加线性层和归一化层。最后,得到一个线性层的输出 $Y_L = [L[\text{CLS}], L_1, L_2, \cdots, L_n]$。因为第一个 token $L[\text{CLS}]$ 表示的是根据其他 token 的信息学习的,所以将其作为聚合表示并输入到一个线性层中以产生最终的预测结果。Masked multi-modal attention 作为 CM-ERNIE 的核心,旨在使用音频模态信息与文本的交互信息动态调整特征词权重和微调预训练的 ERNIE 模型。其详细步骤如下:

首先,评估每个词在不同模态下的权重。Query Q_t 和 Key K_t 文本模态为 $Q_t = K_t = \hat{X}_t'$,其中 $\hat{X}_t'$ 为缩放后文本特征。Query Q_a 和音频模态的 Key K_a 为 $Q_a = K_a = \hat{X}_a'$,其中 $\hat{X}_a'$ 是缩放后的词级对齐音频特征。然后,文本注意力矩阵 $\boldsymbol{\alpha}_t$ 和音频注意力矩阵 $\boldsymbol{\beta}_a$ 定义为:

$$\boldsymbol{\alpha}_t = \text{Relu}(Q_t K_t^T)$$

$$\boldsymbol{\beta}_a = \text{Relu}(Q_a K_a^T)$$

为通过文本和音频模态之间的信息交互来动态调整特征单词权重,对 $\boldsymbol{\alpha}_t$ 和 $\boldsymbol{\beta}_a$ 加权求和,加权融合注意力矩阵 $\boldsymbol{W}_f$ 为

$$\boldsymbol{W}_f = w_t^* \boldsymbol{\alpha}_t + w_a^* \boldsymbol{\beta}_a + b$$

其中,w_t 为文本模态权重,w_a 为音频模态权重,b 为偏差。然后引入 Mask 矩阵 $\boldsymbol{M}$,减少 padding 序列的影响,然后将多模态注意力矩阵 $\boldsymbol{W}_m$ 定义为:

$$\boldsymbol{W}_m = \text{Softmax}(\boldsymbol{W}_f + \boldsymbol{M})$$

得到多模态注意力矩阵后,将 $\boldsymbol{W}_m$ 与屏蔽多模态注意力 V_m 的值相乘,得到注意力 X_{Att} 的输出。其中 V_m 是 ERNIE 最后一个 Encoder 层的输出,定义为 $V_m = X_t$。

$$X_{\text{Att}} = \boldsymbol{W}_m V_m$$

3　实验

在本节中评估了跨模态 ERNIE 在公共多模态情绪分析数据集 CMU-MOSI 和 CMU-MOSEI 上的性能,和在公共数据集(ChnSentCorp)和(Nlpcc2014-Sc)上的准确性。

3.1　数据集与实验设置

实验使用 CMU 多模态观点级情绪强度(CMU-MOSI)和 CMU 多模态意见情绪和情绪强度(CMU-MOSEI)数据集进行评估,并且使用另外两个官方团队提供的文本单模态公共数据集(ChnSentCorp)和(Nlpcc2014-Sc)验证模型的准确性。

(1)CMU-MOSI 是由关于 YouTube 电影评论观点视频组成,视频共包含 93 个观点,共计 2199 条话语,每个话语的标签值由人工注释且标签值在(-3 ~ 3),其中,-3 表示负面最大值,3 表示正面最大值。另外考虑到说话者话语不应同时出现在训练集和测试集中,以及正负数据的平衡,将训练、验证和测试集视频数量拆分为 52、10、31,且对应的话语数量分别对应为 1284、229 和 686。

(2)CMU-MOSEI 由来自 YouTube 的 23454 个电影评论视频剪辑组成。

(3)ChnSentCorp 为情感分析任务的中文句子评论级情感分类数据集。

(4)Nlpcc2014-Sc 是微博短文本情感分析数据集。

为防止预训练 ERNIE 模型过拟合,encoder 层的学习率设置为0.01,其余层的学习率设为 2e-5。为提升实验性能,冻结嵌入层的参数。为训练 CM-ERNIE 模型,将批量大小和最大序列长度分别设置为 24 和 50,epoch 数设置为3。此外,使用 Adam 优化器和均方误差损失函数。

3.2　特征以及模型对齐

为与文本模态一致,在词级对齐音频特征之前附加一个零向量,然后分别对文本与音频进行特征提取。其中,音频提取过程中需重点注意与对应的文本对齐。

3.3　评价指标

实验中,用相同的评价指标来评估基线和提出模型的性能。情绪评分分类任务采用 7 类精度(Acc_7^h),二元情绪分类任务采用 2 类精度(Acc_2^h)和 F1 评分(F_1^h)。指标值越高,模型的性能就越好。为了使实验结果更具准确性,最终的实验结果为随机选择 5 次运

行的平均结果。

3.4 对比实验模型

EF-LSTM:early fusion LSTM (EF-LSTM)是融合早期输入特征,也称前期融合特征,然后送入 LSTM 模型来学习多模态上下文交互相关信息。

LMF:低秩多模态融合(LMF)是一种利用低秩权重张量,在不影响实验性能的情况下,使多模态数据高效融合的方法。

MFN:记忆融合网络(MFN)明确考虑了神经架构中的 LSTM 和增量记忆注意力网络的相互作用,并随着时间的推移,不断对其进行建模。

MARN:multi-attention recurrent network (MARN)使用多头注意力块和长短时混合记忆网络来挖掘不同模式之间的交互信息。

RMFN:循环多级融合网络(RMFN)将多级融合过程与循环神经网络相结合,以对时间和模态数据特征的进行交互建模。

MFM:多模态分解模型(MFM)帮助多模态判别因子和模态特定生成因子中每个因子的提取,专注于从跨多模态数据和标签的联合信息学习表示提取多模态数据特征。

MCTN:多模态循环翻译网络(MCTN)不同模态之间进行转换,联合表示数据特征。

MulT:multimodal transformer (MulT)使用定向成对交叉模式注意力跨不同时间步长的多模式序列之间的交互,并潜在地将数据流进行模式转换,它是 MOSI 数据集上当前最先进的方法。

T-BERT:是改进 Transformers(Bert)的双向 Encoder 表示,仅使用文本模态信息进行微调。

4 结果与讨论

本节展示了实验结果,讨论了提出的方法与前期成果的差异。此外,将屏蔽多模态注意力可视化,以及在单模态数据集上的结果对比,并讨论了引入音频模态信息后注意力矩阵的变化。

4.1 对比实验结果

表1 显示了在 CMU-MOSI 数据集上评估 CM-ERNIE 模型的实验结果。由表 1 知,CM-ERNIE 模型在 MOSI 数据集上创建了一个新的最好的结果,并提高了所有评估指标的性能。在二元情感分类任务中,CM-ERNIE 模型在 Acc_2^h 上达到了83.9%。在情感评分分类任务中,CM-ERNIE 模型的提升效果更加明显。CM-ERNIE 的模型在 Acc_7^h 上达到了42.9%,另外,除 T-BERT 之外的其他基线模型都使用三模态数据信息,但本文提出的模型仅使用双模态数据(文本和音频)取得了新的最好的结果。

表 1 CM-ERNIE 模型在 CMU-MOSI 上的实验结果 单位:%

方法	模态	Acc_7^h	Acc_2^h	F_1^h
EF-LSTM	T+A+V	33.7	75.3	75.2
LMF	T+A+V	32.8	76.4	75.7
MFN	T+A+V	34.1	77.4	77.3
MARN	T+A+V	34.7	77.1	77.0
RMFN	T+A+V	38.3	78.4	78.0
MFM	T+A+V	36.2	78.1	78.1
MCTN	T+A+V	35.6	79.3	79.1
MulT	T+A+V	40.0	83.0	82.8
T-BERT	T	41.5	83.2	83.2
CM-ERNIE	T+A	42.9	83.9	84.0

类似地,在 CMU-MOSEI 数据集上进行了实验。为了便于比较,继之前数据集实验的工作之后,将表 1 中后 3 个模型的 Acc_2^h 和 F_1^h 进行了比较。首先,MulT 在 Acc_2^h 上达到了82.5%,F_1^h 为82.3%。T-BERT 表现出更好的性能,它在 Acc_2^h 上达到了83.0%,F_1^h 为82.7%。但是,CM-ERNIE 在 Acc_2^h 上与 T-BERT 相比,在 Acc_2^h 上达到了83.6%。因此,在 CMU-MOSEI 数据集上的实验结果也说明本文所提的方法在其他多模态数据集上也有不错的泛化性。

为验证所提模型在多模态数据集上的提升,在单模态数据集上进行对比实验,验证模型的准确性,并与 TextCnn、FastText、ERNIE、Bert 模型对比,结果如表 2 所示。

表 2 CM-ERNIE 模型在单模态数据集的实验结果 单位:%

方法	ChnSentCorp			Nlpcc2014-Sc		
	Pr	Re	*F*1	Pr	Re	*F*1
TextCnn	90	90	90	82.2	81.5	82.1
FastText	88.3	88.2	88.2	80.1	79.9	82.3
ERNIE	94.7	94.7	94.7	82.8	82.8	82.8
CM-ERNIE	95.1	95.1	95.1	83.2	83.6	83.3
BERT	94.9	94.9	94.9	82.9	82.9	82.9

从表 2 可以看出,CM-ERNIE 模型将预训练的 ERNIE 模型从单模态扩展到多模态,并引入了音频模态的信息,帮助文本模态有效地调整词的权重。由于

CM-ERNIE 模型可以更全面地反映说话者的情绪状态,并且可以通过文本和音频模态之间的交互来捕捉更多的情感特征,因此它在所有评估指标上的表现都得到了显著的提升。

4.2　多模态屏蔽注意力可视化

为证明屏蔽多模态注意力的效率,分别可视化对比了单模态文本数据注意力矩阵 $\boldsymbol{\alpha}_t$ 和多模态数据注意力矩阵 $\boldsymbol{W}_m$ 中词语权重的差异,并且容易得知在引入多模态音频数据信息后,Masked multimodal attention 可以合理调整词权重。例如从 CMU-MOSI 数据集中选择一个句子,将其单模态文本数据注意力矩阵和多模态数据注意力矩阵可视化,如图 2 所示,颜色梯度代表单词的重要性。

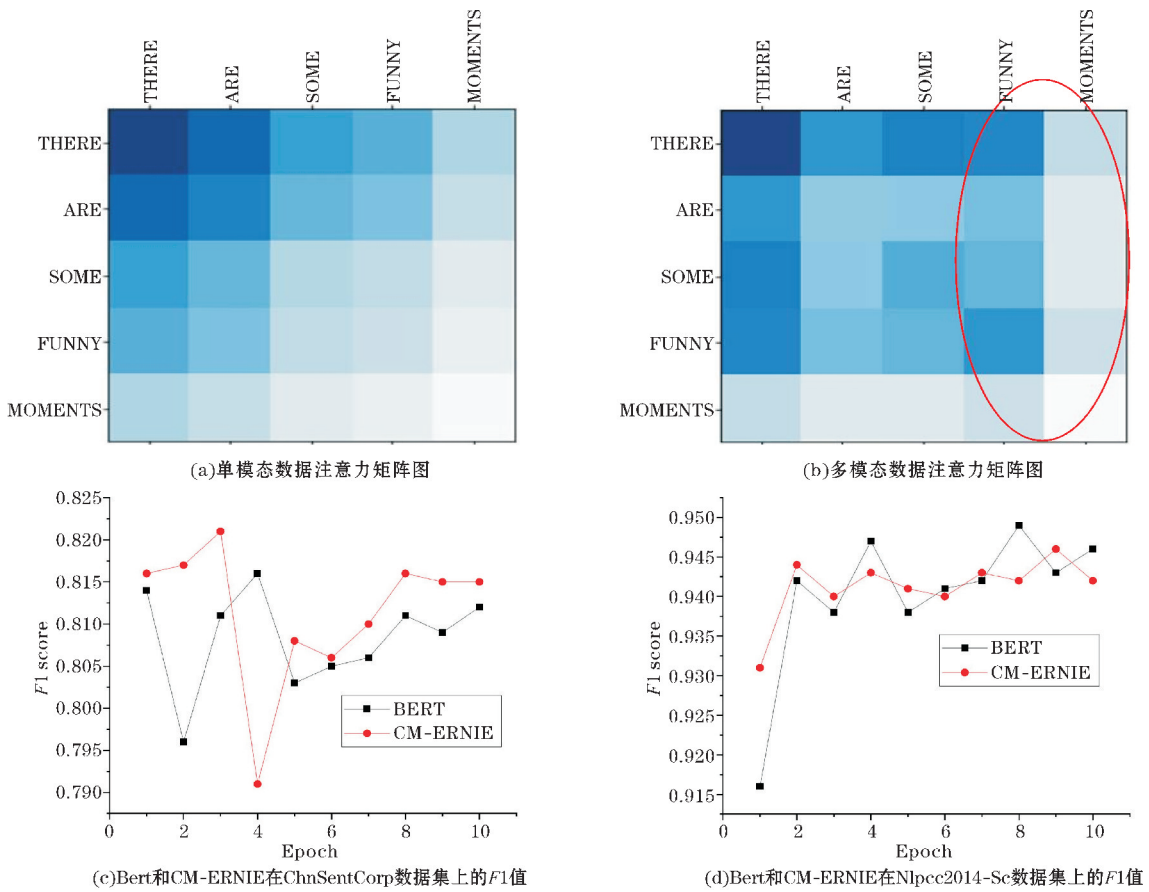

图 2　实验结果可视化

图 2 例句为“THERE ARE SOME FUNNY MOMENTS”,图 2(a)和(b)是对应的注意力矩阵。很明显,图 2(a)和(b)之间存在很多差异。例如,图 2(a)中“FUNNY”这个词在“ARE”这个词上的注意力得分很高。然而,AER 这个词不包含任何情感信息。引入音频信息后的图 2(b),Masked multi-modal attention 降低了“ARE”的分数。相比之下,它更多地关注“SOME”和“MOMENTS”这两个词。为了充分说明 CM-ERNIE 模型的性能,分别统计比较了 Bert 和 CM-ERNIE 模型在两个不同数据集 10 轮结果的加权 $F1$ 值,其性能如图 2(c)和(d)所示。通过实验发现,结合音频的语音语调信息,音频词与文本交互可挖掘更丰富的情感信息,对于情感极性判断结果更准确。

5　结束语

提出一种新颖的多模态情感数据交互分析模型 CM-ERNIE。将预训练的 ERNIE 模型从单模态文本数据扩展到多模态文本加语音数据,引入音频模态信息(例如语音,语调)来辅助文本模态微调预训练模型 ERNIE,通过屏蔽多模态注意力为 CM-ERNIE 的核心单元,动态调整文本和音频跨模态交互数据特征权重。实验结果表明,CM-ERNIE 在多模态数据集上的性能比以前的基线有显著提高,并且在单模态数据集上的性能也超越 ERNIE、Bert、FastText 等。此外,将注意力矩阵可视化,可以清楚地表明在引入音频模态后,能更有效地提升准确度。事实上,CM-ERNIE 也适用于文

本和图片模态,也可应用于两种以上的模态。未来,由于大多数多模态数据通常是未对齐,并且数据具有时序性,将会更多地关注如何对齐不同模态数据,挖掘数据的时序特征以及数据的上下文特征,以及如何使用预训练语言模型从未对齐的多模态数据中学习更好的表示。

参考文献:

[1] Li X, Fu X, Xu G, et al. Enhancing BERT representation with context-aware embedding for aspect-based sentiment analysis[J]. IEEE Access, 2020, 8:46868-46876.

[2] Sun C, Qiu X, Xu Y, et al. How to fine-tune bert for text classification? [C]. China national conference on Chinese computational linguistics. Springer, Cham, 2019:194-206.

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

[4] Guillaumin M, Verbeek J, Schmid C. Multimodal semi-supervised learning for image classification [C]. 2010 IEEE Computer society conference on computer vision and pattern recognition. IEEE, 2010:902-909.

[5] 章荪,尹春勇.基于多任务学习的时序多模态情感分析模型[J].计算机应用,2021,41(6):1631-1639.

[6] Poria Soujanya, Erik Cambria, Devamanyu Hazarika, et al. Context-dependent sentiment analysis in user-generated videos[J]. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics 2017, 1:873-883.

[7] Tsai Y H H, Bai S, Liang P P, et al. Multimodal Transformer for Unaligned Multimodal Language Sequences [C]. Proceedings of the conference. Association for Computational Linguistics. Meeting, 2019:6558-6569.

[8] Ronan C, Jason W. A unified architecture for natural language processing: deep neural networks with multitask learning[C]. In Proceedings of ICML, 2008:160-167.

[9] 王梓懿,安俊秀,王鹏.基于多尺度量子谐振子算法的相空间概率聚类算法[J].计算机应用,2017,37(8):2218-2222.

[10] 杨铠成.基于深度学习的跨模态音频情感分类方法研究[D].石家庄:河北科技大学,2020.

[11] Turian J, Ratinov L, Bengio Y. Word representations: a simple and general method for semi-supervised learning [C]. Proceedings of the 48th annual meeting of the association for computational linguistics. 2010:384-394.

[12] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations. CoRR abs/1802.05365(2018)[J]. arXiv preprint arXiv:2018.

[13] Lai Y, Zhang L, Han D, et al. Fine-grained emotion classification of Chinese microblogs based on graph convolution networks [J]. World Wide Web, 2020, 23(5):2771-2787.

[14] Pal A, Karn B. Anubhuti-An annotated dataset for emotional analysis of Bengali short stories [J]. arXiv e-prints, 2020.

[15] Ruposh H A, Hoque M M. A computational approach of recognizing emotion from Bengali texts [C]. 2019 5th International Conference on Advances in Electrical Engineering (ICAEE). IEEE, 2019:570-574.

[16] Manek A S, Shenoy P D, Mohan M C, et al. Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier[J]. World Wide Web-internet & Web Information Systems, 2017, 20(2):135-154.

[17] Xu J, Xu Y, Xu Y, et al. A Chinese text sentiment classification algorithm framework based on a hybrid of semantic understanding and machine learning[J]. Computer Science, 2015, 42(6):61-66.

[18] Kalchbrenner N, Grefenstette E, Blunsom P. A Convolutional Neural Network for Modelling Sentences[J]. Eprint Arxiv, 2014, 1:35-47.

[19] 李铁飞,生龙,吴迪.BERT-TECNN 模型的文本分类方法研究[J].计算机工程与应用,2021,57(18):186-193.

[20] Mihaylov T, Frank A. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge[J]. arXiv preprint arXiv, 2018.

[21] Zaremoodi P, Buntine W, Haffari G. Adaptive knowledge sharing in multi-task learning: Improving low-resource neural machine translation [C]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2018: 656-661.

[22] Chen Q, Zhu X, Ling Z H, et al. Neural natural language inference models enhanced with external knowledge[J]. arXiv preprint arXiv, 2017.

[23] Han X, Liu Z, Sun M. Neural knowledge acquisition via mutual attention between knowledge graph and text[C]. Thirty-second AAAI conference on artificial intelligence. 2018.

[24] Madotto A, Wu C S, Fung P. Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems[J]. arXiv preprint arXiv, 2018.

[25] Yu F, Tang J, Yin W, et al. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs[C]. Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(4): 3208-3216.

Multi-modal Sentiment Analysis based on Cross-modal Fusion ERNIE

TAO Quanhui, AN Junxiu, CHEN Hongsong

(College of Software Engineering, Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: Aiming at the fact that sentiment analysis mainly focuses on single-modal text data and ignores the problem of multi-modal data fusion, a cross-modal fusion ERNIE sentiment analysis model (CM-ERNIE) is proposed by combining the masked multi-modal attention method. First, use CNN and BiGRU to extract audio data features and word vectors to extract text sequence features; second, dynamically adjust text and audio data weights by masking multimodal attention as the core unit of CM-ERNIE, and finally, text and audio modalities The interaction of fine-tuning the pretrained ERNIE model. The model is evaluated on the multimodal movie review opinion datasets CMU-MOSEI and CMU-MOSI. Comprehensive experiments show that the model is more accurate than the single-modal sentiment analysis model on the multi-modal datasets CMU-MOSEI and CMU-MOSI, and the research of multi-modal sentiment analysis contains great value, which can be used for multi-modal sentiment analysis. It provides decision support for practical application problems such as sentiment analysis, public opinion analysis, and intent recognition in modal scenarios.

Keywords: multimodal fusion; pre-training model; attention mechanism; ERNIE; text classification