计算机科学
COMPUTER SCIENCE

基于多模态注意力的噪声事件分类模型

吴贺祥, 王中卿, 李培峰

引用本文

吴贺祥, 王中卿, 李培峰.基于多模态注意力的噪声事件分类模型[J]. 计算机科学, 2022, 49(11A): 211000161-7.

WU He-xiang, WANG Zhong-qing, LI Pei-feng. Noise Event Classification Model Based on Multimodal Attention [J]. Computer Science, 2022, 49(11A): 211000161-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211000161

基于多模态注意力的噪声事件分类模型

吴贺祥　王中卿　李培峰
苏州大学计算机科学与技术学院　江苏 苏州 215006
(20205227103@stu.suda.edu.cn)

摘　要　如今,社交媒体因其低成本、易于访问和快速传播而成为人们获取新闻资讯和了解实时事件的主要渠道之一。社交媒体为分析特定事件提供了包含文本和图像等多种模态的信息,这其中包含了大量无关事件和虚假信息。为此,结合文本-图像对来判断文本和图像是否提供了与特定事件相关的信息,从而筛选出与之无关的噪声事件。由于文本中的描述往往与相对应的图像中的情景相关联,因此提出了一个基于多模态注意力的结合文本和图像信息的方法进行事件分类。该方法能很好地关注到文本和图像中的重要信息并促进不同模态的信息交互。在CrisisMMD数据集上的实验结果表明,该方法优于6种强的基线方法,证明了所提多模态注意力模型能够有效融合不同模态的特征,得到更优的联合表示。

关键词:注意力机制;多模态融合;噪声事件分类
中图法分类号　TP391

Noise Event Classification Model Based on Multimodal Attention

WU He-xiang,WANG Zhong-qing and LI Pei-feng
School of Computer Science and Technology,Soochow University,Suzhou,Jiangsu 215006,China

Abstract　Social media is nowadays one of the main channels for people to obtain news and learn about real-time events due to its low cost,easy access and rapid dissemination. Social media provides a variety of modal information including text and images for analyzing specific events,which contains abundant irrelevant events and false information. To this end,this paper combines the text-image pairs to determine whether the text and image provide information related to specific events,so as to find out irrelevant noise events from the sentence-level of the text. Motivated by the observation that the description in the text is often associated with the scene in the corresponding image,this paper proposes a method of combining text and image information to classify events based on attention mechanism,which can effectively attend to the important information in text and image and promote information interaction in different modalities. Experimental results on CrisisMMD show that our model outperforms six strong baselines,and it can effectively fuse features of different modality to obtain a superior joint representation.

Keywords　Attention mechanism,Multimodal fusion,Noise event classification

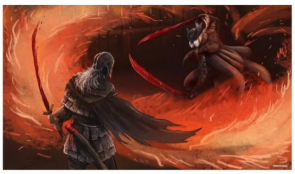
1　引言

事件被定义为在某个特定的时间点或时间片段和地域范围内发生的、由一个或多个角色参与、由一个或多个动作组成的一件事情或状态的改变。随着移动互联网的发展,公共事件被人们在社交媒体上广泛讨论。社交媒体因其低成本、易于访问和快速传播而成为人们获取新闻资讯和了解实时事件的主要渠道之一。但是,社交媒体平台的用户所发布的信息量十分庞大及用户匿名性等问题导致的事件相关信息的庞杂和部分信息较低的可信性,即信息源中存在相当一部分属于噪声的无关事件,对此类信息进行标注滤除既费时又昂贵。

对社交媒体上巨大的信息进行抽取并分析特定问题具有极大的挑战性,而目前研究者大多聚焦于文本信息。本文从分析事件的角度出发,结合社交媒体用户所发布的图像和文本信息对特定事件进行研究。本文所研究的多模态噪声事件分类任务是根据文本-图像对中的信息来判别文本内容是否与特定事件有关。Alam等[1]通过在多个自然灾害期间从推特上收集信息,组建了一个多模态数据集CrisisMMD,该数据集包括数千条人工精心标注的包含文本与图像的样例。Ofli等[2]在该数据集上进行了研究,证明了结合文本和图像相比单模态模型可得到更优的结果。本文主要研究该数据集上的噪声事件分类任务,它将与灾难事件相关信息的样本定义为有信息量的事件(Informative),与灾难事件无关的样本定义为噪声事件(Not Informative)。对于同一事件,不同模态之间能够提供互补的信息,因而多模态学习方法与从单一模态学习相比能从不同模态中学习到互补的信息,从而实现更健壮的推理[2]。图1给出了从CrisisMMD中选取的一些图像与文本样例。若只关注文本传达的信息,则很容易将

基金项目:国家自然科学基金(61806137,61702518,61836007);江苏省高等学校自然科学研究面上项目(18KJB520043);江苏高校优势学科建设工程资助项目
This work was supported by the National Natural Science Foundation of China(61806137,61702518,61836007),Natural Science Foundation of Jiangsu Province(18KJB520043) and A Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.
通信作者:王中卿(wangzq@suda.edu.cn)

图 1(a)错误地判断为与灾难事件相关。因为尽管图 1(a)中文本提到了灾难事件“Hurricane Lady”,但从整个句子分析并无灾难事件发生,而图像展现的是灾难无关场景。图 1(b)可以很直接地结合文本和图像信息推断出它提供了灾难相关信息。因此捕获文本和图像中与灾难事件相关的信息,并排除无关信息的干扰,才能实现正确的推理。



(a) Hurricane Lady # Maria It'll rain burning blood. I hope Puerto Rico knows how to do Visceral Attacks

(b) Help is still needed after the first # Mexico Earthquake please share with others

图 1 CrisisMMD 中噪声事件样例

Fig. 1 Samples of noise events in CrisisMMD

如上所述,样例中图像中的情景和文本内容之间存在着一定的关联。在文本无法提供灾难事件相关信息时,通过分析图像中的信息,如图 1(a)样例所示,能够进一步增加它的“无关性”,从而有利于模型进行正确推理。考虑到文本和图像中存在这种相关信息,本文使用注意力机制来关注这些不同模态中的互补的语义特征,进而进行有效推理。

预训练模型在深度学习领域中起着非常积极的作用,将在庞大语料库下训练的模型迁移到下游任务并进行微调越来越受到关注,并已取得了许多进展。本文采用预训练文本模型 BERT[3]和 VGGNet[4]作为单模态分类和多模态特征提取模块。受 Yang 等[5]研究的启发,VGGNet[4]的最后一个池化层保留着原始图像的空间信息,本文使用该层特征作为图像中各个区域的特征表示。

本文对多模态噪声事件分类进行了研究探索,提出了一个多模态注意力噪声事件分类模型(Noise Event Classification Model on Multimodal Attention, NECMA)。具体地,NECMA 使用 BERT[3]预训练模型来进行文本分类和文本特征的提取,并使用 VGG16[4]作为图像分类和图像特征提取的模型。NECMA 引入了注意力机制对文本和图像中的重要信息进行建模,从而获得更有效的联合表示。NECMA 采用 VGG16 池化层的输出特征矩阵作为对应于原始图像各区域特征的映射矩阵。具体地,对于文本特征,本文选择 BERT 编码器最终隐藏状态的文本特征作为句子中各个单词对应的特征向量,选择池化层的输出作为句子的全局表示。对于图像特征,本文选择 VGG16 分类器输出的向量作为图像全局特征,选择最后一个池化层的输出作为图像各个区域对应的特征向量。以文本或图像的全局特征向量作为查询,对互补模态的特征矩阵应用注意力机制,不仅能够关注到文本和图像中的重要信息,还能实现文本和图像的特征融合。

在 CrisisMMD 数据集上的实验结果验证了合理地利用不同模态的特征优于只使用单种模态特征,本文提出的多模态注意力模型 NECMA 相比 BERT 文本模型在 F1 值上提升了超过 5%。本文与 6 种强的多模态基线方法的对比实验表明,NECMA 模型能够有效融合文本与图像中的重要信息并进一步做出正确推理,相比基准模型得到了更好的实验结果。

2 相关工作

2.1 事件检测

事件检测是事件抽取的一个重要子任务,并且有助于不同的下游自然语言处理的应用。传统的基于特征的方法依靠人工设计的特征来检测事件触发词和事件类型。Liao 等[6]提出了一个使用文档级事件和角色信息来进行句子级事件抽取的方法,从而解决了仅依靠句子局部信息难以处理的问题。Hong 等[7]提出了使用实体间关系来进行事件抽取的跨实体推理,他们将实体类型的一致性作为预测事件提及的关键特征。Li 等[8]通过结构化感知机模型进行事件触发词和事件论元的识别和分类的联合学习,并使用柱搜索策略得到最优结果。

随着神经网络的快速发展,基于神经网络进行事件抽取的方法也不断涌现。Chen 等[9]率先将神经网络用于事件抽取,他们使用卷积神经网络来捕获句子级特征并构建了一个动态多池卷积神经网络来自动学得句子中不同部分的重要信息。Nguyen 等[10]提出了一个基于双向循环神经网络的联合框架来进行事件抽取,这种方法既能够减轻流水线方法的误差传播,还能够自动学习单词的特征表示。Yang 等[11]提出了一个基于预训练语言模型来进行事件抽取的框架,他们通过角色来划分论元预测从而解决角色部分重叠的问题。Sha 等[12]提出了一个依赖性桥循环神经网络来充分利用句子中的句法关系,他们通过在 LSTM 模型[13]上构建依赖性桥来生成句子中各个实体和触发词之间的依赖性解析树。Liu 等[14]针对复杂句子中存在多个事件难以准确识别的问题,提出了一个联合多事件抽取框架,他们通过引入句法短弧来增强信息流,并使用基于注意力机制的图卷积网络对句法短弧进行建模。Wang 等[15]构建了一个覆盖率高的大型事件相关候选集,采用对抗训练机制从候选集中识别出提供有用信息的事件。Tong 等[16]提出了一种丰富知识蒸馏模型,以利用外部开放域触发词知识来减少注释中频现的触发词的内置偏差。

多模态事件抽取是利用与文本对齐的图像特征作为补充信息,近年来也受到越来越多的关注。Zhang 等[17]利用外部图像-字幕对来学习视觉模式,使用视觉信息作为辅助外部知识来消除纯文本模态的歧义并提升事件抽取的性能。Li 等[18]提出了一个多媒体事件抽取任务,他们开发了一个利用现有单模态语料库的弱监督训练框架,利用结构化表示和图神经网络进行多媒体公共空间嵌入的学习。Tong 等[19]为事件检测基准构建了一个图像数据集,并提出了一个双循环多模态模型,以在图像和句子之间进行深度交互从而进行模态特征融合。

2.2 多模态融合

多模态融合将从不同模态中抽取的特征结合在一起形成一个紧凑的多模态表示。通常,多模态特征融合根据融合发生的阶段划分为早期融合、中期融合和晚期融合。早期融合将来自不同模态的特征连接起来作为模型的输入[20]。晚期融合尝试分别对每种模态进行建模,然后在决策层将不同模态的信息进行融合[21]。中期融合则在这二者之间的网络结构中进行。Wang 等[22]提出了一种无参数多模态融合框架,

它动态交换不同模态的子网络的通道，从而实现不同模态的特征的融合。Perez-Rua 等[6]提出了一个根据模型的顺序进行优化的方案，从而在可能的融合结构组成的搜索空间中找到给定数据集的最优的融合结构。Zadeh 等[23]提出了一个张量融合网络，使用 3 次笛卡尔积对单模态、双模态和三模态的相互作用进行建模。Sahu 等[25]提出了一种适应融合技术来对不同模态的上下文进行建模，让模型决定“如何”抽取和有效地组合来自不同模态的特征。

注意力机制在多模态融合中应用普遍，有助于关注图像中某些实体与文本中相对应的语义概念。Hori 等[26]提出了一个多模态注意模型用于视频描述中不同模态特征的融合，证明了对合适的模态进行注意力的关注能够很好地与单词的语义相对齐。Yang 等[5]提出了一个堆叠注意力网络来学习根据图像回答自然语言问题，将经过 LSTM[15]和一种用于文本分类的 CNN 模型[27]编码后的文本信息作为查询对 VGG-Net[4]抽取到的图像特征矩阵应用注意力机制，来寻找图像中与答案相关的区域。除了使用文本特征向量作为查询对图像应用注意力机制之外，还可以采用同时生成关注的图像特征向量和关注的文本特征向量的共同注意力机制。Lu 等[28]提出了并行共同注意力和交替共同注意力的方法。Nam 等[29]提出了一个类似于并行共同注意力的双重注意力网络，它同时对图像和语言的注意力分布进行评价来获得关注的特征向量。

3 基于多模态注意力的噪声事件分类模型

3.1 体系结构

本文提出的噪声事件分类模型 NECMA 的体系结构及采用的多模态注意力模块的实现细节分别如图 2 和图 3 所示。图 2 给出了 NECMA 的总体结构，它分别由语言模型、图像模型和多模态注意力模块构建而成。对于给定的输入文本，将其经过词法单元分词之后的单词序列送入 BERT，从 BERT 编码器可获得单词序列的特征表示（图 2 中 $\boldsymbol{w}^0,\cdots,\boldsymbol{w}^{l-1}$，$l$ 为句子长度），从池化层可获得文本全局特征向量。将输入图像送入图像模型 VGG，从最终的池化层可获得图像中各个区域的特征表示（图 2 中 $\boldsymbol{r}^0,\cdots,\boldsymbol{r}^{m-1}$，$m$ 为图像区域个数）。然后将上述特征送入多模态注意力模块。注意，这里多模态注意力指对文本和图像两种模态应用注意力机制进行融合的实现方法。

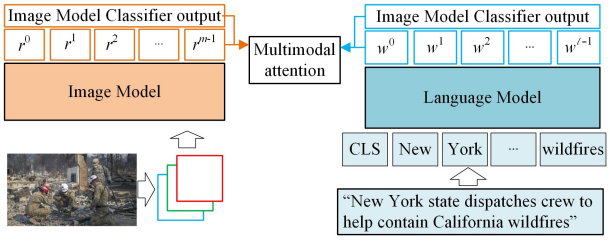


图 2 多模态注意力模型的体系结构

Fig. 2 Architecture of multimodal attention model

图 3 给出了本文提出的两种多模态注意力模型的结构示意图，图中符号“×”为特征矩阵与特征向量之间的点积运算。其中图 3(a)图所示结构除了使用上述单模态模型中输入单模态分类器的向量之外，还使用这两个向量分别作为查询，对图像模型 VGG16 的池化层得到的各个图像区域对应的特征矩阵 $\boldsymbol{I}_a$ 和文本模型 BERT 得到的隐藏状态输出矩阵 $\boldsymbol{T}_h$ 应用注意力机制得到关注的图像注意力向量 $\boldsymbol{a}_{ti}$ 和关注的文本注意力向量 $\boldsymbol{a}_{it}$。

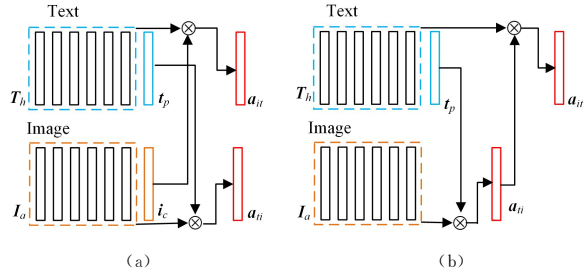


图 3 多模态注意力模型中的注意力模块

Fig. 3 Attention module of multimodal attention model

另外，本文还探索了先使用文本特征向量作为查询对图像特征应用注意力机制得到关注的图像向量 $\boldsymbol{a}_{ti}$，而后将关注的图像向量 $\boldsymbol{a}_{ti}$ 作为查询对文本特征矩阵应用注意力机制来得到关注的文本向量 $\boldsymbol{a}_{it}$，具体的过程如图 3(b)所示。对于这两种模型，本文最后都将关注的图像向量 $\boldsymbol{a}_{ti}$ 和文本向量 $\boldsymbol{a}_{it}$ 与单模态分类向量 $\boldsymbol{t}_p$ 和 $\boldsymbol{i}_c$ 进行融合然后送入多模态分类器。

3.2 基于 BERT 模型的文本表示

本文使用 BERT 作为文本分类模型和多模态分类模型的文本特征提取模块，能够获得预训练的单词嵌入。该模型的输入是词法单元化的单词序列，这些单词通过位置嵌入得到增强，并为每个单词输出一个表示[30]。本文使用 BERT 得到的句子的全局表示作为文本单模态待分类的特征向量，并将该特征向量用作多模态注意力模型中对图像特征矩阵应用注意力机制的查询。另外，把 BERT 模型得到的最后隐藏状态表示作为句子中各个单词对应的特征向量构成的文本特征矩阵。在多模态模型中，使用图像特征向量对其应用注意力机制。

对于多模态注意力模型中的文本模型，本文选择 BERT 来提取特征。具体地，使用 BERT 编码器的输出作为各个单词的特征向量，并使用 BERT 池化层的输出向量 $\boldsymbol{t}_p$ 作为句子的全局表示。本文使用该全局表示作为对图像应用注意力机制的查询向量。

3.3 基于 VGG16 模型的视觉表示

为获得图像特征信息，本文使用在 ImageNet 挑战数据集[31]上预训练的 VGG16 作为图像分类模型和多模态分类模型的图像特征提取部分。在计算机视觉中，迁移预训练卷积神经网络的最终的全连接层是很常见的[32]，它的输出通常是对特征映射进行池化操作后的结果。本文在图像单模态模型中使用 VGG16 作为图像分类模型。具体地，VGG16 原输出维度为 1 000，本文使用的 CrisisMMD 中定义的 informative 任务有 2 个类别，因此将相应的 VGG16 的分类器的输出维度设置为 2。

在本文提出的多模态注意力模型中，本文对 VGG16 不同层输出的特征做了特征选择，选择 VGG16 网络结构中的最后一个池化层的输出特征作为原图像各个区域对应的特征向量构成的图像特征矩阵。因为该层保留了原始图像中的

空间信息，所以有助于注意力机制关注到图像中的重要信息。

对于多模态注意力模型中的图像模型，首先将图像重置大小为 224×224 个像素，然后输入模型，选取模型最后一个池化层的输出特征 $\boldsymbol{f}_I$，它的维度为 512×7×7。7×7 是将原图像划分区域的个数，512 是每个区域的特征向量的维度。各个图像区域的特征向量对应于图 2 中从图像模型得到的 $\boldsymbol{r}^i$，i 是图像区域的索引。这些特征向量整体上就是图像特征矩阵，即图 3 中的 $\boldsymbol{I}_a$。

3.4 基于注意力机制的文本与图像的特征融合

图 2 给出了本文提出的多模态注意力模型的总体体系结构。本文分别使用 BERT 和 VGG16 作为文本模型和图像模型，将从这两个单流模型中抽取得到的特征经由多模态注意力进行融合，从而实现结合文本和图像信息的联合推理。与前面所述的单模态分类模型不同，多模态注意力模型将单模态模型看作特征提取器，并利用注意力机制来关注文本和图像中的重要信息。

本文将上述图像和文本模型抽取得到的特征经过全连接层以获得相同的维度。图 3 给出了本文的多模态注意力模型的具体实现过程和细节。对于给定的输入图像，通过图像模型可获得特征矩阵 $\boldsymbol{I}_a$ 和图像特征向量 $\boldsymbol{i}_c$，其中 $\boldsymbol{I}_a \in \mathcal{R}^{m\times d}$，$m$ 是输入图像的区域个数，d 是每个区域对应的特征向量维度；$\boldsymbol{i}_c \in \mathcal{R}^d$ 是一个 d 维的向量。对于输入句子，可从文本模型获得文本特征矩阵 $\boldsymbol{T}_h$ 和文本特征向量 $\boldsymbol{t}_p$，其中 $\boldsymbol{T}_h \in \mathcal{R}^{l\times d}$，$l$ 为句子长度，d 对应每个单词的特征向量维度；$\boldsymbol{t}_p \in \mathcal{R}^d$ 是一个 d 维的文本全局表示。

本文采用两种方法来对图像和文本特征矩阵应用注意力机制，第一种方法采用文本全局表示和图像模型得到特征向量作为查询，分别对图像区域特征矩阵和文本特征矩阵同时应用注意力机制的并行注意力机制；另一种方法则采用由并行注意力机制得到的文本或图像单种模态的注意力向量作为对互补模态应用注意力机制的查询。下面将对这两种方法的具体细节和思想作进一步的阐述。

3.4.1 基于并行注意力机制的融合方式

首先是并行注意力机制，本文将其称为 $\text{NECMA}_{\text{parallel}}$，它同时使用图像特征向量 $\boldsymbol{i}_c$ 和文本特征向量 $\boldsymbol{t}_p$ 分别对文本特征矩阵 $\boldsymbol{T}_h$ 和图像特征矩阵 $\boldsymbol{I}_a$ 应用注意力机制，从而得到关注的文本注意力向量 $\boldsymbol{a}_{it}$ 和图像注意力向量 $\boldsymbol{a}_{ii}$。以获取关注的图像注意力向量为例，本文首先对文本特征向量 $\boldsymbol{t}_p$ 与图像特征矩阵 $\boldsymbol{I}_a$ 执行乘法运算，得到图像各个区域的注意力分数 $\boldsymbol{h}_I$，然后经由 softmax 函数来生成图像的各个区域的注意力分布 $\boldsymbol{p}_I$，最后根据该注意力分布与图像各个区域对应的特征向量经加权求和即可得到一个关注的图像注意力向量 $\boldsymbol{a}_{ii}$。具体计算过程可用式(1)、式(3)、式(5)描述。以图像特征向量作为查询求解关注的文本注意力向量的过程与上述求解关注的图像注意力向量的过程相同，具体计算过程可用式(2)、式(4)、式(6)描述。

$$\boldsymbol{h}_T = \tanh(\boldsymbol{T}_h \otimes (\boldsymbol{W}_{i,T}\boldsymbol{i}_c + \boldsymbol{b}_T)) \tag{1}$$

$$\boldsymbol{h}_I = \tanh(\boldsymbol{I}_a \otimes (\boldsymbol{W}_{t,I}\boldsymbol{t}_p + \boldsymbol{b}_I)) \tag{2}$$

$$\boldsymbol{p}_T = \text{softmax}(\boldsymbol{W}_{T,p}\boldsymbol{h}_T + \boldsymbol{b}_{T,p}) \tag{3}$$

$$\boldsymbol{p}_I = \text{softmax}(\boldsymbol{W}_{I,p}\boldsymbol{h}_I + \boldsymbol{b}_{I,p}) \tag{4}$$

$$\boldsymbol{a}_{it} = \sum_{i=0}^{d-1}\sum_{n=0}^{l-1} \boldsymbol{p}_T w_i^n \tag{5}$$

$$\boldsymbol{a}_{ii} = \sum_{i=0}^{d-1}\sum_{n=0}^{m-1} \boldsymbol{p}_I r_i^n \tag{6}$$

其中，$\otimes$ 执行一个向量与一个矩阵的相乘的运算，它将该向量与矩阵中每个向量作逐元素乘积并求和。$\boldsymbol{W}_{i,T}, \boldsymbol{W}_{t,I} \in \mathcal{R}^{d\times d}$，$\boldsymbol{b}_T, \boldsymbol{b}_I \in \mathcal{R}^d$，$\boldsymbol{W}_{T,p} \in \mathcal{R}^{l\times l}$，$\boldsymbol{W}_{I,p} \in \mathcal{R}^{m\times m}$ 是模型中的参数和偏置。$\boldsymbol{p}_T \in \mathcal{R}^l$，$\boldsymbol{p}_I \in \mathcal{R}^m$，它们分别是文本和图像的注意力分数和注意力分数的分布；w_i^n 是文本特征矩阵中句子的第 n 个单词的特征向量的第 i 维的取值，r_i^n 是图像特征矩阵中第 n 个区域的特征向量的第 i 维的取值。

3.4.2 基于交替注意力机制的融合方式

图 3(b)给出了本文提出的多模态注意力模块的第二种实现结构，即交替注意力机制，本文将其称为 $\text{NECMA}_{\text{alternate}}$。如图中结构所示，先使用前面所述方法进行图像注意力向量的计算，然后利用该图像注意力向量作为查询计算文本注意力向量。直观上，无论先对哪种模态应用注意力机制，其思想都是统一的，都是利用关注的注意力向量作为对另一种模态应用注意力机制的查询向量。本文采用前面平行注意力机制计算文本或图像注意力向量的方法先计算得到一种模态(文本或图像)的注意力向量，这里令其为 $\boldsymbol{a}_1$；然后用该注意力向量作为查询对互补模态(图像或文本)的特征矩阵应用注意力机制得到关注的另一种模态的注意力向量，这里令其为 $\boldsymbol{a}_2$。该计算过程可用式(7)－式(9)来描述。

$$\boldsymbol{h} = \boldsymbol{a}_1 \otimes \boldsymbol{X} \tag{7}$$

$$\boldsymbol{p} = \text{softmax}(\boldsymbol{h}) \tag{8}$$

$$\boldsymbol{a}_2 = \sum_{i=0}^{d-1}\sum_{n=0}^{N-1} \boldsymbol{p}\, x_i^n \tag{9}$$

其中，$\boldsymbol{a}_1 \in \mathcal{R}^d$ 是先计算得到的一种模态的注意力向量；$\boldsymbol{X} \in \mathcal{R}^{N\times d}$ 是另一种模态的特征矩阵；$\boldsymbol{h}, \boldsymbol{p} \in \mathcal{R}^N$，它们是另一种模态的注意力分数的分布，$N$ 是文本的单词个数或图像的区域个数；x_i^n 是另一种模态特征矩阵中第 n 个特征向量的第 i 维的取值。

通过上述过程，可得到分别对文本和图像应用注意力机制的关注向量。最终，本文将上述两个注意力向量、文本全局表示 $\boldsymbol{t}_p$ 和图像特征向量 $\boldsymbol{i}_c$ 分别进行逐元素乘积的运算，然后再将二者拼接在一起，从而实现不同级别特征的融合，最终将该向量送入多模态分类器。

4 实验

4.1 实验设置

本文使用 CrisisMMD 语料库来评估本文提出的模型 NECMA。实验主要涉及 CrisisMMD 定义的 Informative 任务，目的是判别文本-图像对是否为灾难事件提供有效信息，本文根据该性质对样例是否为灾难事件相关进行区分。本文遵循在该数据集上的先前研究工作[4]的数据划分和预处理方法，使用准确率(P)，召回率(R)和 F 度量($F1$)作为评估指标。

表 1 列出了实验所采用数据集的具体划分和文本与图像的分布。该数据集从社交媒体上收集的信息包含一条文本对应多张图像的情况，因而训练集上文本的数目少于图像的数目。为丰富训练数据，该数据集将包含多张图像的数据简单地进行文本复制，使得该数据下的不同图像对应于相同的文本。

表 1　CrisisMMD 的 Informative 任务的类别和数据集划分

Table 1　Categories and data split for Informative task of CrisisMMD

Categories	Train(70%)		Dev(15%)		Test(15%)		Total	
	Text	Image	Text	Image	Text	Image	Text	Image
Informative	5546	6345	1056	1056	1030	1030	7632	8431
Not informative	2747	3256	517	517	504	504	3768	4277
Total	8293	9601	1573	1573	1534	1534	11400	12708

对于所有模型，本文使用的学习率为 5×10^{-6}，学习的目标函数是交叉熵损失函数，设置批量大小为 32，并使用 Adam[33] 进行优化。本文通过在验证集上设置早停机制来保存最优模型，并在测试集上验证模型的泛化能力。

4.2　基线方法

目前多模态事件分类及事件检测还处于起步阶段，并且针对 CrisisMMD 的研究还较少。本文主要针对文本内容及多模态实验中图像信息对文本特征的辅助作用进行研究，在前人工作[2]在该数据集上取得的最好模型的基础上加入了几种较强的单模态和多模态的基线方法。

对于单模态基线方法，本文选择前人工作[2]中所使用的文本模型 KimCNN[27] 和在自然语言处理领域应用广泛的 LSTM[14] 作为 BERT 预训练模型[3]的对比模型。

对于多模态基线系统，本文实现了基于拼接（对应表 2 中的 BERT-VGG16$_{\text{concat}}$）、加权求和（对应表 2 中的 BERT-VGG16$_{\text{weighted}}$）等方式进行特征融合的方法。同时，本文选择多模态深度学习领域中取得先进性能的如下模态作为对比模型：MMBT[34] 是一个多模态 BiTransformer，它通过将图像嵌入投影到文本特征空间来联合微调在单模态上预训练的文本和图像编码器；Oscar[35] 使用在图像中检测到的对象标签作为连接点来简化文本与图像之间对齐的学习。

表 2　事件分类实验结果

Table 2　Experiment results of event classification

Modality	Model	The Informative task		
		P	*R*	*F*1
Text	KimCNN	82.2	82.3	82.2
	LSTM	83.2	82.1	83.1
	BERT	85.8	85.8	85.8
Image	VGG16	82.6	82.9	82.6
Multimodal	KimCNN-VGG16[2]	83.8	84.1	83.8
	LSTM-VGG16	84.5	84.6	84.5
	BERT-VGG16$_{\text{concat}}$	89.4	89.5	89.4
	BERT-VGG16$_{\text{weighted}}$	89.7	89.7	89.7
	MMBT[31]	85.7	85.4	85.5
	Oscar[32]	89.9	90.0	89.9
	NECMA$_{\text{parallel}}$	90.7	90.7	**90.7**
	NECMA$_{\text{alternate}}$	91.0	91.1	**91.0**

4.3　实验结果

表 2 列出了本文模型 NECMA 与对比模型的性能表现。对于文本分类单模态模型，BERT 预训练模型相比其他文本模型有非常大的提升，这表明 BERT 模型能够有效地学习到语言表示。KimCNN 和 LSTM 这两个文本分类模型使用 Word2Vec 模型预训练的单词嵌入[4]，BERT 相比这两个模型在 *F*1 值上提高了 2%以上，而本文提出的 NECMA 两种方法相比 BERT 都实现了巨大的提升，这表明融合不同模态的特征相比只使用单种模态更易于做出正确推理。对于多模态模型，本文模型 NECMA 相比 KimCNN-VGG16 在 *F*1 值上实现了超过 6%的提升。另外，相比 LSTM-VGG16 提升也十分明显。从表 2 中可看到，NECMA 相比 BERT-VGG16$_{\text{concat}}$ 和 BERT-VGG16$_{\text{weighted}}$ 都实现了相当可观的提升，NECMA 的两种模型相比拼接或加权求和的融合方式在 *F*1 值上提高了超过 1%。MMBT 的结果与 BERT 单模态模型的结果相近，这可能是由于 MMBT 无法将图像嵌入合理地映射到文本特征空间，因而在融合图像特征之后没有得到提升。Oscar 取得了比上述基线方法更优的结果，可能是由于图像中对象标签有助于文本和图像的对齐学习。NECMA 相比 Oscar 仍然实现了较大提升，这证明了对文本模型和图像模型的中间文本特征和图像特征应用注意力机制是十分有效的。

上述实验结果表明注意力机制有助于文本和图像特征的交互，并得到了更优的多模态联合表示。同时，NECMA 的两种方法也有些许性能差别，NECMA$_{\text{alternate}}$ 由 NECMA$_{\text{parallel}}$ 方法先计算得到一种模态的注意力向量，再使用该向量作为关注另一种模态的查询，在 *F*1 值上提高了 0.3%，这可能是由于注意力向量作为查询融合了不同模态的特征，因而有助于特征的融合。

4.4　错误分析

本文选取了 BERT 文本模型、VGG16 图像模型、BERT-VGG16$_{\text{concat}}$ 和本文提出的两种注意力模型的实验结果中的样例，图 4 给出了具体结果。

(a) montana firm whitefish energy hired to rebuild querto rico's power grid hedging
BERT: Informative(×)
VGG16: Informative(×)
BERT-VGG16$_{\text{concat}}$: Informative(×)
NECMA: Not informative(√)

(b) video guy, enormous fish cage aims to save his fish from hurricane iram
BERT: Informative(×)
VGG16: Informative(×)
BERT-VGG16$_{\text{concat}}$: Informative(×)
NECMA: Not informative(√)

(c) harver horror shivering tot found clinging to drowned mom
BERT: Informative(×)
VGG16: Informative(×)
BERT-VGG16$_{\text{concat}}$: Informative(×)
NECMA: Not informative(√)

(d) thick smoke across northern california as seen from nasa visible satellite images monday cafires fire
BERT: Informative(×)
VGG16: Informative(×)
BERT-VGG16$_{\text{concat}}$: Informative(×)
NECMA: Not informative(×)

图 4　不同融合方式的分类结果样例

Fig. 4　Examples of classification results with different fusion methods

对于图4(a)—图4(c)3个样例，只有NECMA得出了正确的分类结果，这证明了本文方法可以进行正确的推理而不受无关信息的干扰。值得注意的是，文本内容和图像场景不完全一致的特点在这些样例中也有所体现，滤除无关信息捕获灾难事件相关信息是正确推理的关键。对于一些情景复杂的图像和相关信息，需要从整个句子分析情况，NECMA也很难推断出正确结果。如图4(d)所示，在图像无法提供有效信息的情况下，文本中的关键信息“cafires”处于句子的末尾，此时只有对整个句子进行良好建模并获悉“ca”所代表的地名这一外部知识才能正确推理。这些样例表明NECMA能够合理利用文本和图像不同模态中的有效信息，从而改进了推理结果，但语言建模仍然是解决该问题的关键。本文的研究对解决社交媒体非正式语言和与之不完全对应的图像之间的多模态学习提供了新的思路。

结束语 本文提出了一个新的基于注意力机制的多模态注意力模型NECMA来进行噪声事件分类任务的研究，它能够有效提升在该任务上的性能表现。通过有效地关注文本和图像中互补的信息并对文本特征和图像特征进行融合，这不仅能够避免过多地关注文本或图像中的不相关信息，并且有助于建立文本和图像之间的跨模态交互。本文验证了合适的多模态融合模型相比单模态文本模型和图像模型有巨大的提升。此外，本文证明了所提多模态注意力模型能够关注到文本和图像中的重要信息并获得更优的多模态联合表示。

本文的研究成果表明图像信息与文本特征的有效交互相比传统仅使用文本单模态有更优的性能，这为未来进行多模态事件分类和事件检测任务的研究提供了一定的参考价值。对于多模态事件分类和事件检测，事件在视觉语义上具有一定的抽象性，这是多模态事件抽取发展缓慢的重要原因，也是未来研究的突破点。

参 考 文 献

[1] ALAM F, OFLI F, IMRAN M. CrisisMMD: Multimodal twitter datasets from natural disasters[C]// Proceedings of the Twelfth International Conference on Web and Social Media. California, USA: 2018: 465-473.

[2] OFLI F, ALAM F, IMRAN M. Analysis of social media data using multimodal deep learning for disaster response[C]// Proceedings of the 17th ISCRAM Conference. Blacksburg, VA, USA: 2020.

[3] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association of Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 4181-4186.

[4] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]// Proceedings of the 3rd International Conference on Learning Representations. San Diego, CA, USA: 2015.

[5] YANG Z, HE X, GAO J, et al. Stacked attention networks for image question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA, 2016: 21-29.

[6] LIAO S, GRISHMAN R. Using document level cross-event inference to improve event extraction[C]// Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden, 2010: 789-797.

[7] HONG Y, ZHANG J, MA B, et al. Using cross-entity inference to improve event extraction[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland, Oregon, USA, 2011: 1127-1136.

[8] LI Q, JI H, HUANG L. Joint event extraction via structured prediction with global features[C]// Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria, 2013: 73-82.

[9] CHEN Y, XU L, LIU K, et al. Event extraction via dynamic multi-pooling convolutional neural networks[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015: 167-176.

[10] NGUYEN T H, CHO K, GRISHMAN R. Joint event extraction via recurrent neural networks[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 300-309.

[11] YANG S, FENG D, QIAO L, et al. Exploring pretrained language models for event extraction and generation[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 5284-5294.

[12] SHA L, QIAN F, CHANG B, et al. Jointly extracting event triggers and arguments by dependency-bridge RNN and tensor-based argument interaction[C]// Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans, Louisiana, USA, 2018: 5916-5923.

[13] HOCHREITER S, SCHMIDHUBER J. Long short-term Memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[14] LIU X, LUO Z, HUANG H. Jointly multiple events extraction via attention-based graph information Aggregation[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 1247-1256.

[15] WANG X, HAN X, LIU Z, et al. Adversarial training for weakly supervised event detection[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, MN, USA, 2019: 998-1008.

[16] TONG M, XU B, WANG S, et al. Improving event detection via open-domain trigger knowledge[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online, ACL, 2020: 5887-5897.

[17] ZHANG T, WHITEHEAD S, ZHANG H, et al. Improving Event Extraction via Multimodal Integration[C]// Proceedings of the 2017 ACM on Multimedia Conference. Mountain View, CA, USA, 2017: 270-278.

[18] LI M, ZAREIAN A, ZENG Q, et al. Cross-media structured common space for multimedia event extraction[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020: 2557-2568.

[19] TONG M, WANG S, CAO Y, et al. Image enhanced event detection in news articles[C]// Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, NY, USA, 2020: 9040-9047.

[20] D'MELLO S K, KORY J M. A review and meta-analysis of multimodal affect detection systems[J]. ACM Computing Surveys, 2015, 47(3): 43: 1-43: 36.

[21] MORVANT E, HABRARD A, AYACHE S. Majority vote of diverse classifiers for late fusion[C]// Structural, Syntactic, and Statistical Pattern Recognition-Joint IAPR International Workshop. Joensuu, Finland, 2014: 153-162.

[22] WANG Y, HUANG W, SUN F, et al. Deep multimodal fusion by channel exchanging[C]// Proceedings of the Thirty-fourth Conference on Neural Information Processing Systems. Virtual, 2020.

[23] PEREZ-RUA J M, VIELZEUF V, PATEUX S, et al. MFAS: Multimodal fusion architecture search[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA, 2019: 6959-6968.

[24] ZADEH A, CHEN M, PORIA S, et al. Tensor fusion network for multimodal sentiment analysis[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 1103-1114.

[25] SAHU G, VECHTOMOVA O. Adaptive fusion techniques for multimodal data[C]// Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021: 3156-3166.

[26] HORI C, HORI T, LEE T Y, et al. Attention-based multimodal fusion for video description[C]// Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 4203-4212.

[27] KIM Y. Convolutional neural networks for sentence classification[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1746-1751.

[28] LU J, YANG J. BATRA D, et al. Hierarchical question-image co-attention for visual question answering[C]// Proceedings of the Thirtieth Annual Conference on Neural Information Processing Systems. Barcelona, Spain, 2016: 289-297.

[29] NAM H, HA J W, KIM J. Dual attention networks for multimodal reasoning and matching[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA, 2017: 2156-2164.

[30] MAJUMDAR A, SHRIVASTAVA A, LEE S, et al. Improving vision-and-language navigation with image-text pairs from the web[C]// Proceedings of the 2020 European Conference on Computer Vision. Glasgow, UK, ECCV, 2020: 259-274.

[31] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database[C]// Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Miami, Florida, USA, 2009: 248-255.

[32] ALI S R, HOSSEIN A, JOSEPHINE S, et al. CNN features off-the-shelf: An astounding baseline for recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2014. Columbus, OH, USA, 2014: 512-519.

[33] KINGMA D P, BA J. Adam: A method for stochastic optimization[C]// Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015. San Diego, Ca, USA.

[34] KIELA D, BHOOSHAN S, FIROOZ H, et al. Supervised multimodal bitransformers for classifying images and text[J]. arXiv: 1909.02950.

[35] LI X, YIN X, LI C, et al. Oscar: Object-semantic aligned pre-training for vision-language tasks[C]// Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK, 2020: 121-137.

WU He-xiang, born in 1998, postgraduate. His main research interests include natural language processing and so on.

WANG Zhong-qing, born in 1987, Ph.D, lecturer, is a member of China Computer Federation. His main research interest is natural language processing.