融合注意力协同和对比学习的跨模态突发事件识别方法

黄少年[1,2]，彭永涛[1]，文沛然[3]，刘耀[1]
（1. 湖南工商大学计算机学院，长沙，410205；
2. 统计学习与智能计算湖南省重点实验室，长沙，410205；
3. 湖南工商大学数字媒体工程与人文学院，长沙，410205）

摘　要：针对跨模态突发事件识别中图像复杂、文本有限以及模态间误导信息干扰等问题，提出了一种融合注意力协同和对比学习的跨模态融合方法。首先，通过构建融合多支路自注意力的图像特征编码器和融合外部知识的文本特征编码器，分别提取图像的深层语义特征和文本的上下文关联信息，以克服单一模态信息表达的局限性。其次，为有效捕捉图像-文本特征间的复杂关联，设计跨模态注意力协同网络增强模态间特征的交互一致性，减少误导信息的传递。最后，引入对比学习机制，采用联合监督策略优化图像-文本特征的贡献，实现模态间信息的动态平衡。在公开数据集CrisisMMD和自建数据集上开展的实验证明了提出方法的优越性。

关键词：突发事件识别；跨模态；注意力协同；对比学习
中图分类号：TP391
文献标志码：A

A Cross-Modal Emergency Recognition Method Integrating Attentional Collaboration and Contrastive Learning

Huang Shaonian[1,2], Peng Yongtao[1], Wen Peiran[3], Liu Yao[1]
1. School of Computer Science, Hunan University of Technology and Business, Changsha 410205, China
2. Key Laboratory of Statistical Learning and Intelligent Computing in Hunan Province, Changsha 410205
3. School of Digital Media Engineering and Humanities, Hunan University of Technology and Business, Changsha 410205, China

Abstract: To address the challenges of image complexity, limited textual information, and inter-modal misleading data in cross-modal emergency recognition, this paper proposes a novel fusion recognition method that integrates attentional collaboration and contrastive learning. First, a multi-branch self-attention mechanism is integrated into an image feature encoder, while a text feature encoder is augmented with external knowledge. These encoders are designed to extract rich semantic features from images and contextual information from text, respectively, for overcoming the limitations of single-modal representations. Next, a cross-modal attentional collaboration network is introduced to capture intricate relationships between image-text features, enhancing inter-modal consistency and mitigating the influence of misleading information. Finally, a contrastive learning mechanism, that is optimized through a joint supervision strategy, is employed to dynamically balance the contributions of image and text features. Extensive experiments conducted on the public CrisisMMD dataset and the self-constructed dataset demonstrate the superior performance of the proposed method.

Key words: Emergency recognition, Cross-modal, Attention collaboration, Contrastive learning

收稿日期：XXXX-XX-XX；**修回日期：**XXXX-XX-XX
通信作者：刘耀 yliu@hutb.edu.cn
基金项目：国家社科基金一般项目（21BTJ026），湖南省社会科学成果评审委员会课题（XSP25YBC164）
Foundation Items: The National Social Science Foundation of China (21BTJ026), The Project of Hunan Social Science Achievement Appraisal Committee (XSP25YBC164)

1 引言

随着移动互联网的迅速发展，Twitter、Facebook 和微博等社交媒体平台作为突发事件信息传播的关键途径，已经成为实时获取信息的主要来源[1]。在突发事件发生时，这些平台以文本、图片和视频等多种形式提供了丰富的事件相关信息。突发事件识别任务旨在从社交媒体数据中提取与突发事件相关的图像和文本信息，并根据其事件类型、信息性和损失程度进行分类。因此，社交媒体突发事件识别受到越来越多的关注。已有研究表明[2, 3]，有效地对社交媒体突发事件进行分析与识别，将有助于各级政府部门的应急救援和决策管理工作，为人民群众的生命财产安全提供必要保障。

传统的突发事件识别通常从文本模态提取事件特征，这种方法实现简单有效，但忽略了多模态信息对事件识别的积极作用。与仅基于文本特征的突发事件识别不同，跨模态突发事件识别基于两个或多个模态的特征进行事件分析，其重点是如何融合不同模态的特征。已有研究[4, 5]的关注点主要集中于设计不同深度神经网络模型，加强突发事件图像-文本的语义特征表示。最近，一些基于 Transformer[6, 7]和预训练模型[8, 9]的方法被提出，用于实现突发事件多模态特征融合。然而，在社交媒体突发事件识别的研究中，仍然存在两个主要挑战。一是社交媒体数据中图像的复杂性、文本的有限性。广泛使用的卷积神经网络架构或预训练语言模型，在应用到社交媒体数据时，仍存在特征语义相关性、特征提取效率等问题[10]。如何有效提取复杂图像和有限文本中的语义相关特征，仍值得深入研究。二是模态间噪声信息的传递。基于不同模态特征相关性设计的早期融合、中期融合或后期融合策略，实现了不同阶段的模态关系建模。然而，由于社交媒体图像、文本数据中通常存在一些噪声，特征融合中噪声的传递可能形成非信息性或误导性特征[11]。减少噪声信息传递，可以提升模态间的有效特征交互，增强图像-文本的一致性特征表示。

此外，现有方法大多基于突发事件类型进行识别[6, 12, 13]，由于突发事件的分类粒度较为宽泛，其结果对应急救援的实际参考价值有限。事实上，在社交媒体图像-文本数据中，图像和文本从不同的方面展示了丰富的突发事件相关信息，体现了不同模态数据对突发事件信息贡献的不平衡性。如图 1 所示的洪水事件中，图像展示了洪水场景，而文本信息中则还包含了人员损伤、建筑物损坏等具体信息。因此，平衡图像、文本数据在突发识别中的贡献，将有助于提取应急救援相关信息，提升突发事件识别效果。本文将突发事件识别任务划分成三类子任务：事件类型识别任务、信息性识别任务和损失程度识别任务，三类子任务共同实现了从社交媒体数据中提取突发事件关键信息的完整流程。首先识别突发事件的类型，提取关于突发事件类型的细粒度信息。信息性识别任务主要关注社交媒体图像-文本数据中是否具备与应急救援相关的有价值信息，实现应急救援信息的快速筛选。损失程度识别则关注突发事件造成的影响，为后续的应急救援措施提供依据。

4 killed in Manipur, 140 houses destroyed in Mizoram.
https://t.co/dMlEngezZ4

图 1 突发事件图像-文本示例

鉴于此，本文提出了一种融合注意力协同和对比学习的跨模态突发事件识别方法。首先设计了融合多支路注意力的图像特征编码器和融合外部知识的文本特征编码器，来解决图像模态的语义复杂性和文本模态的文本有限性；然后，构建了跨模态注意力协同网络处理多模态融合中噪声传递的问题，该部分包含交叉注意力模块和权重注意力模块，以增强各模态特征之间的信息交互，从而减少非有效信息的传递；最后，设计对比学习网络，引入对比损失和交叉熵损失进行联合学习，减少因各模态贡献度不同而造成的特征损失，提取图-文一致性特征表示。在公开数据集和自建数据集上开展的实验证明了本文方法的优越性。

2 相关工作

2.1 基于单模态的方法

基于单模态的突发事件识别主要包括基于文本模态和图像模态的方法。针对文本模态，Sreenivasulu 等人[14]提取文本中与突发事件相关的词汇、句

法、频率等特征，采用随机森林分类器实现突发事件识别。Madichetty 等人[15]基于突发事件的词汇嵌入表示，采用堆叠卷积神经网络识别具有医疗资源需求的社交媒体突发事件。Abhishek 等人[16]提出了基于自注意的混合深度学习框架，旨在从突发事件文本数据中提取分层文本特征的关键信息。针对图像模态，Alam 等人[17]采用 VGG16 进行突发事件图像分类，实现突发事件识别，并辅助突发事件应急救援。Valdez 等人[18]提出了轻量级 CNN 传输学习模型，用于识别突发事件的强度等级，该方法在灾害类突发事件识别中达到较高的识别率。Kyrkou 等人[19]面向无人机应急响应需求，提出了一种具有残差连接的航空图像分类模型，该方法在减少内存需求的情况下实现了较高的准确率。基于单模态的方法仅利用突发事件中文本或图像特征，忽略了其它模态中包含的语义信息。

2.2　基于多模态的方法

突发事件中的多模态内容可以提供更丰富的事件相关信息，其研究重点集中在如何进行多模态内容的融合。根据特征融合方式的不同，现有方法可分为三类：输入级的融合、高级特征融合及注意力融合。输入级融合采用单一网络从各模态中分别提取特征，然后直接拼接各模态特征实现多模态特征融合[20, 21]。此类方法在一定程度上利用了多模态数据，但没有充分考虑不同模态的数据特征差异。基于高级特征融合的方法采用独立的骨干网络处理各个模态，并基于池化、点积、张量融合等聚合操作，实现各模态高级特征的融和[22-27]。Wang 等人[28]提出了一种基于模态对齐和分层思维的分层多模态特征融合方法，实现层次化多模态突发事件摘要生成。Lv 等人[29]提出了一种基于对抗生成的多模态信息融合模型，用于社交媒体突发事件识别。然而，由于各模态中存在无意义甚至误导性的信息，导致融合后的有效特征损失，影响了此类方法的识别准确率。随着 Transformer 模型的兴起，自注意力机制被用来计算模态特征的相关性，实现基于注意力分数的跨模态融合。Abavisani 等人[30]首次将注意力机制用于多模态突发事件分类中。Wu 等人[31]设计了一种捕获模态间相关性和独立性的识别模型，实现跨模态的图-文事件识别。Gupta 等人[32]提出了一个知识注入和可解释的多模态注意力网络，实现突发事件的分类。鉴于注意力机制在多模态融合中展现出的优势，本文通过设计细粒度的注意力融合机制，实现突发事件图-文一致性特征融合。

3　跨模态突发事件识别方法

本文提出了一种融合注意力协同和对比学习的跨模态突发事件识别方法，整体框架如图 2 所示。该框架包含图像特征编码器、文本特征编码器、跨模态注意力协同网络、跨模态对比学习网络。

3.1　融合多支路自注意力的图像特征编码器

在图像特征提取方面，本文在 DenseNet[33]的基础上设计了融合多支路自注意力的图像特征编码器。多支路注意力层能更好地捕捉图像的局部特征，这些局部特征经过 DenseNet 的深层密集连接，有效地促进了特征重用和信息流动，有利于学习图像的复杂全局特征。

首先对原始特征图进行批量处理，将其大小调整为 224 × 224，再对特征图进行去中心化和归一

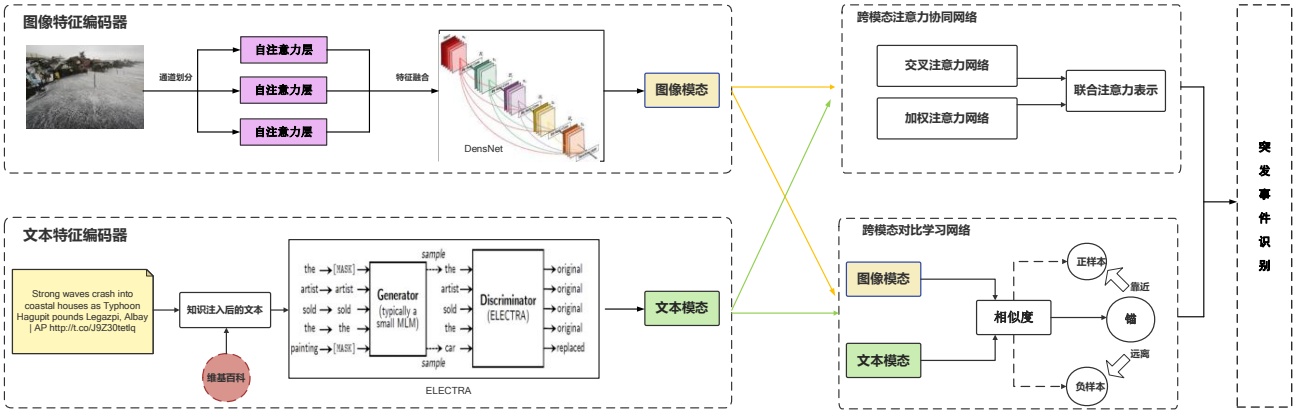

图 2　跨模态突发事件识别框架图

化，并将特征图沿RGB通道切分为三条支路。各支路的特征图经过线性变换映射成查询向量$\boldsymbol{Q}_i$、键向量$\boldsymbol{K}_i$和值向量$\boldsymbol{V}_i$。然后，各支路分别进行自注意力计算，第i条支路的自注意力计算公表示为

$$\boldsymbol{b}_i = \text{softmax}\left(\frac{\boldsymbol{Q}_i \boldsymbol{K}_i^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_i \tag{1}$$

其中，$Q_i \in R^{1\times d_k}, K_i \in R^{1\times d_k}, V_i \in R^{1\times d_v}$，$d_k$表示键向量的特征维度，$d_v$表示值向量的特征维度，$d_k = d_v = 512$。

将各支路自注意力特征图进行直接拼接，输入DenseNet网络后提取图像特征。该过程表示为

$$\boldsymbol{F}_{\boldsymbol{I}} = \text{DenseNet}(\text{Concat}(\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3)) \in R^{c\times h\times w} \tag{2}$$

其中，$\boldsymbol{b}_1$、$\boldsymbol{b}_2$和$\boldsymbol{b}_3$分别表示每条支路的特征图，c, h, w分别为输入图像的通道数、高度和宽度。

3.2 融合外部知识的文本特征编码器

现有的多模态突发事件检测方法通常基于事件原始文本直接提取文本特征，但由于社交媒体文本的长度限制，直接提取文本特征的方式可能导致文本重要信息的丢失。因此，本文设计了融合外部知识的文本特征提取模块，通过更多的外部语境理解上下文语义信息，从而生成社交媒体事件的文本特征表示。

为了减少有限文本的约束、丰富文本的语义表示，本文融合Wikipedia库作为外部知识。通过WikiAPI获取社交媒体文本中的外部语境信息，将文本实体的摘要信息追加到社交媒体事件文本中。事件文本中单词与外部知识的关联度ρ的计算如式(3)所示。

$$\rho = \text{wiki}(w_i), w_i \in T_i \tag{3}$$

其中，w_i是事件文本中单词，函数wiki表示维基百科中w_i的关联性，T_i表示事件文本。如果ρ的值大于指定阈值时，说明wiki百科知识与该文本事件相关联。

然后，通过TagMe库进行文本注释，并从注释结果中提取实体的名称。采用Wikipedia库获取实体对应的摘要信息，注入到实体中，从而得到实体数据w_i^{wiki}。将实体数据连接起来表示为

$$T^{wiki} = \text{Concat}(w_1^{wiki}, w_2^{wiki} \cdots w_i^{wiki}) \tag{4}$$

最后，采用基于Transformer架构的语言模型ELECTRA[34]提取融入外部知识后的突发事件文本特征表示，以提高预训练的效率和性能。对于突发事件文本T^{wiki}，其文本特征表示如式(5)所示。

$$\boldsymbol{F}_{\boldsymbol{T}} = \text{ELECTRA}(T^{wiki}) \tag{5}$$

3.3 跨模态注意力协同网络

在图像-文本跨模态任务中，各模态的输入常包含一些噪声，导致模态间会产生非信息性甚至误导性信息的传递。为此，本文设计了一种跨模态注意力协同网络，由交互注意力模块和加权注意力模块构成，网络结构如图3所示。交互注意力模块实现图像模态与文本模态之间的信息交互，增强事件的一致性特征表示。加权注意力模块基于不同模态间权重的动态匹配，学习不同模态间的相关性特征。

将提取的图像和文本特征向量投影到固定维数，作为交互注意力模块的和加权注意力模块的输入，图像特征表示为$\boldsymbol{F}_I$，文本特征表示为$\boldsymbol{F}_T$。交互注意力模块包括图像-文本注意力网络(Image-Correlation-Text Attention network, ICTA)和文本-图像注意力网络(Text-Correlation-Image Attention network, TCIA)。在ICTA网络中，图像模态的查询向量$\boldsymbol{Q}_I$从文本特征空间中选择性地关注相关文本键向量$\boldsymbol{K}_T$和值向量$\boldsymbol{V}_T$，得到图像-文本视觉特征，其计算如式(6)所示。

$$\mathbf{Att}\boldsymbol{n}_{\mathbf{ICT}} = \text{softmax}\left(\frac{\boldsymbol{Q}_I \times \boldsymbol{K}_T}{\sqrt{d_k}}\right) \times \boldsymbol{V}_T \tag{6}$$

其中，$\boldsymbol{Q}_I \in R^{1\times l_I}$，$\boldsymbol{K}_T \in R^{l_T\times d_k}$，$\boldsymbol{V}_T \in R^{l_T\times d_v}$，$d_k$、$d_v$分别表示文本键向量和文本值向量的维度，$l_T$表示文本模态特征的维度，$l_I$表示图像模态特征的维度。

同理，在TCIA网络中，文本模态的查询向量$\boldsymbol{Q}_T$从图像特征空间中选择性地关注相关图像键向量$\boldsymbol{K}_I$和值向量$\boldsymbol{V}_I$，得到文本-图像视觉特征，其计算如式(7)所示。

$$\mathbf{Att}\boldsymbol{n}_{\mathbf{TCI}} = \text{softmax}\left(\frac{\boldsymbol{Q}_T \times \boldsymbol{K}_I}{\sqrt{d_k}}\right) \times \boldsymbol{V}_I \tag{7}$$

其中，$\boldsymbol{Q}_T \in R^{1\times l_T}$，$\boldsymbol{K}_I \in R^{l_I\times d_k'}$，$\boldsymbol{V}_I \in R^{l_I\times d_v'}$，$d_k'$、$d_v'$分别表示图像键向量和图像值向量的维度。

最后，将$\boldsymbol{Attn}_{ICT}$和$\boldsymbol{Attn}_{TCI}$进行直接拼接，得到交互注意力特征表示，表示为

$$\boldsymbol{F}_{\mathbf{cross}} = \text{Concat}(\mathbf{Att}\boldsymbol{n}_{\mathbf{ICT}}, \mathbf{Att}\boldsymbol{n}_{\mathbf{TCI}}) \tag{8}$$

为了使模型更加灵活地关注各模态的全局特征，动态学习不同模态之间的相关性，本文设计了

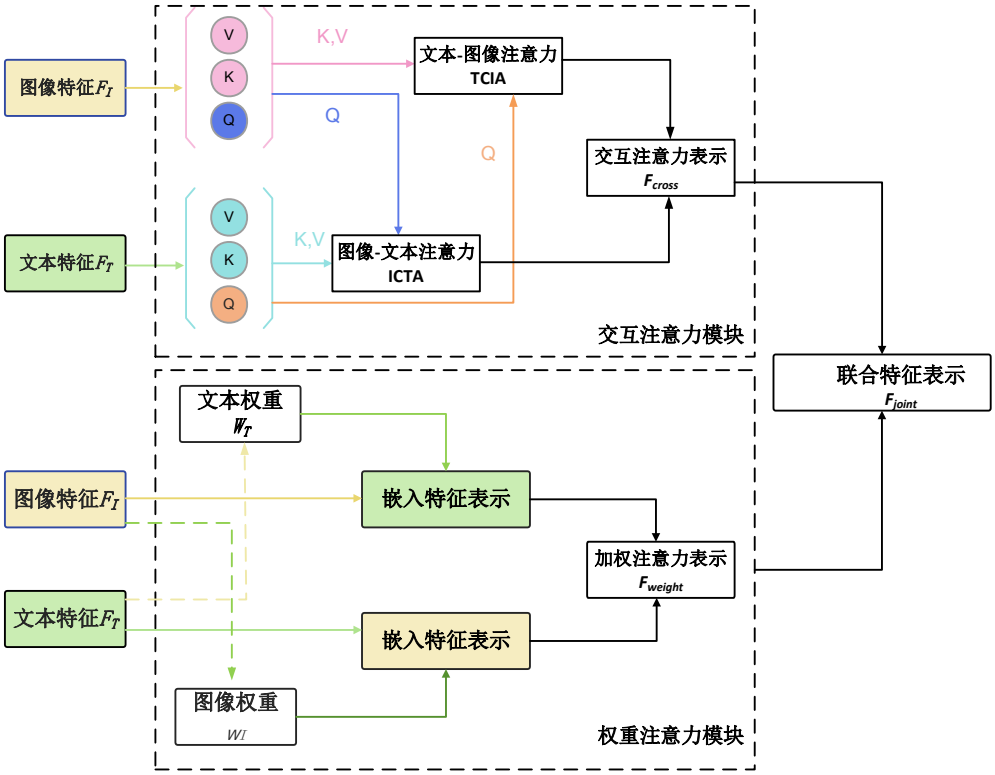

图3　跨模态注意力协同网络

加权注意力模块。加权注意力模块分别计算每个模态的注意权重，然后将图像模态的注意权重与文本模态匹配，而图像模态的权重则与文本模态匹配，这种交叉匹配的方式有利于增强关联性，减少误导性或者冗余信息的传输。具体过程可以表示为

$$\boldsymbol{W}_I'=\sigma(\boldsymbol{W}_1\boldsymbol{F}_I+b_T) \tag{9}$$

$$\boldsymbol{W}_T'=\sigma(\boldsymbol{W}_T\boldsymbol{F}_T+b_I) \tag{10}$$

其中，σ是Sigmoid激活函数，$\boldsymbol{W}_I$和$\boldsymbol{W}_T$分别表示图像模态和文本模态的初始权重矩阵，b_T和b_I分别表示其偏置项。得到图像模态权重$\boldsymbol{W}_I'$和文本模态$\boldsymbol{W}_T'$的权重后，分别与不同模态的嵌入特征表示逐元素相乘，具体表示为

$$\boldsymbol{F}_I'=\boldsymbol{F}_I\odot\boldsymbol{W}_T' \tag{11}$$

$$\boldsymbol{F}_T'=\boldsymbol{F}_T\odot\boldsymbol{W}_I' \tag{12}$$

最后，通过特征拼接得到加权注意力特征表示，具体表示为

$$\boldsymbol{F}_{\text{weight}}=\text{Concat}(\boldsymbol{F}_I',\boldsymbol{F}_T') \tag{13}$$

3.4　跨模态对比学习网络

对于基于图像-文本的突发事件识别任务，图像语义和文本语义并非总是一致的，这导致了图像特征和文本特征的重要性和贡献度不平衡问题。本文设计了跨模态对比学习网络，通过计算图像特征和文本特征的相似度和贡献程度，自动调整模态的权重，平衡不同模态的特征表示。图像-文本特征相似度的训练，促使模型学习到语义匹配的特征表示，更好地捕捉到突发事件中图像特征与文本关键词句之间的相关性与一致性，学习更具判别性的突发事件特征表示。

训练过程中，首先计算每个训练样本中图像特征与文本特征的相似度，若相似度大于某一阈值λ，则视为正样本对，反之为负样本对。训练图像-文本对的相似度计算如式(14)所示。

$$Sim=\frac{\boldsymbol{F}_T\boldsymbol{F}_I}{\sqrt{\boldsymbol{F}_I^2}\sqrt{\boldsymbol{F}_T^2}} \tag{14}$$

通过对比损失函数的设计，将正样本之间的相似度最大化，同时最小化负样本之间的相似度，从而促进模型学习到更好的嵌入表示，对比损失函数如式(15)所示。

$$Loss_{contract}=-\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp(Sim_+)}{\sum_{j=1}^{N}\exp(Sim_+)+\sum_{j=1}^{N}\exp(Sim\_)}\right) \tag{15}$$

其中，Sim_+为图-文正样本的相似度，$Sim_$为图-文

负样本的相似度，N表示训练样本数目。

3.5 跨模态突发事件识别

将交叉注意力联合表示$\boldsymbol{F}_{cross}$和权重注意力联合表示$\boldsymbol{F}_{weight}$做特征增强得到联合特征表示$\boldsymbol{F}_{joint}$，并送入全连接层进行预测，其计算如式(16)-(17)所示。

$$\boldsymbol{F}_{joint}=\boldsymbol{F}_{cross}\odot\boldsymbol{F}_{weight} \tag{16}$$

$$y'=Fc(\boldsymbol{F}_{joint}) \tag{17}$$

其中，y'表示预测的标签值，$\odot$表示逐元素对应相乘，Fc为全连接层。然后，使用Softmax交叉熵损失进行分类，交叉熵损失为

$$Loss_{\text{entropy}}=-\frac{1}{N}\sum_{N}[y\log y'+(1-y)\log(1-y')] \tag{18}$$

其中，y是真实的标签值，N表示样本数目。模型总的损失函数表示为

$$Loss=\alpha*Loss_{entropy}+\beta*Loss_{contract} \tag{19}$$

其中，α和β分别表示交叉熵损失和对比损失的权重系数。

4 实验与结果分析

4.1 数据集

本文分别在公开数据集CrisisMMD[35]和自建数据集上进行了多种突发事件识别实验，来验证本文提出方法的有效性。

CrisisMMD数据集是公认的多模态突发事件数据集，其数据来源于Twitter推文。该数据集包含发生在世界不同地区的地震、飓风、火灾和洪水事件。根据事件信息对突发事件应急救援影响的不同，该数据集将事件分为受影响的个人、基础设施和公用设施损坏、受伤或死亡、失踪或被发现、救援和捐赠、车辆损坏、其他相关信息以及非人道主义事件八类。该数据集的事件标签分布如表1所示。

为了进一步验证模型在中文突发事件识别上的有效性，根据社交媒体突发事件的概念和类别，本文还自建了中文社交媒体突发事件数据集。自建数据集数据来源于微博等社交媒体平台，首先筛选出突发事件相关的数据样本，去除重复的图文和无法获取元数据和图像的样本，得到不同类型突发事件的图像-文本样本对。然后对图像-文本样本对进行突发事件类型、信息性、损失程度的标注，形成本文的自建数据集。自建数据集包括受影响的个人、基础设施和公用设施损坏、警告和建议、需求和供给、反应、其他和非人道主义信息七大类，其事件标签分布如表2所示。

表1 CrisisMMD数据集的标签分布

标签	训练样本	验证样本	测试样本	全部样本
受影响的个人	424	71	86	581
基础设施和公用设施损坏	1905	319	251	1665
受伤或死亡的个人	244	42	41	327
失踪或被发现的个人	24	4	5	33
救援和捐赠	2323	353	340	3016
车辆损坏	134	22	19	175
其他	3294	605	578	4477
非人道	5260	889	849	6998

表2 自建数据集的标签分布

标签	训练样本	验证样本	测试样本	全部样本
受影响的个人	236	24	23	283
基础设施和公用设施损坏	857	117	94	1068
警告和建议	559	74	58	691
需求和供给	302	36	40	378
反应	807	100	96	1003
其他	175	24	18	217
非人道主义信息	630	76	60	766

4.2 实验设置和评价标准

为识别社交媒体突发事件的类型及与应急救援相关的具体信息，本文设置了三类任务在CrisiMMD和自建数据集上进行实验，不同数据集在不同任务下的数据分布如表3所示。三类任务的具体情况如下：

(1)任务1：社交媒体突发事件识别任务，即识别突发事件所属的类型。

(2)任务2：信息性任务，即根据突发事件的图像-文本信息，判断该事件信息是否对突发事件应急救援有所帮助。

(3)任务3：损害程度任务。即根据突发事件的图像-文本信息，将事件损失分为严重损害、轻微损害或者无损害三类，为后期的应急救援服务提供依据。

实验采用pytorch1.9.0+cuda111框架在ubuntu18.04.5LTS操作系统下完成。模型训练中参数设置如表4所示。实验采用分类准确度Auccu-

表3不同任务和设置下的训练、验证和测试数据分布

数据集	任务	训练集	验证集	测试集	全部数据
CrisisMMD	任务1	13608	2237	2237	18082
	任务2	13608	2237	2237	18082
	任务3	2468	529	529	3526
自建数据集	任务1	3566	451	389	4406

racy、宏平均F1分数以及加权F1分数来评估模型的性能。

表4　　参数设置

参数名	数值
学习率	2×10^{-3}
批处理量	8
学习率衰减	10X
优化器	Adam
迭代次数	60
交叉熵损失权重α	0.5
对比损失权重β	0.5
相似度阈值λ	0.1

4.3　实验结果分析

为了验证本文提出的方法的有效性，本文在CrisisMMD数据集和自建数据集上分别与基线模型进行了对比。基线模型主要分为三类：

(1) 基于单模态的方法: 包括模型DenseNet[33]、BERT[36]和MMBT[21]。

(2) 基于高级特征融合的方法: 包括模型CentralNet[22]、GMU[23]、CBP[27]、VisualBERT[24]、PixelBERT[25]和VilT[26]。

(3) 和基于自注意力融合的方法: 包括模型DMCC[37]、SSE-Cross-BERT-DenseNet[30]、CrisisKAN[32]、ME Cross-attention[38]和GAFNet[39]。

4.3.1　CrisisMMD数据集实验结果分析

表5-表7展示了本文方法与其它方法在突发事件3种信息识别任务上的定量比较结果。

表5展示了本文方法与基线模型在社交媒体突发事件类型识别任务上的定量比较结果，总体上来说，本文方法各个指标上的识别结果，明显优于基于单模态的方法和基于高级特征融合的方法，这表明了本文方法在跨模态识别任务上的有效性。具体来说，本文模型在CrisisMMD数据集上的分类精度和宏平均F1分数都达到了SOTA。相较于次优的方法，本文模型比GAFNet模型在分类精度上提升了0.3%，而宏平均F1分数提升21.7%。在加权F1分数上，本文模型略低于GAFNet和ME Cross-attention模型，但GAFNet和ME Cross-attention模型均采用了基于预训练的方法。

表5　　CrisisMMD数据集上任务1的实验结果

Method	Accuracy(%)	Macro F1(%)	Weighted F1(%)
基于单模态的方法			
DenseNet[33]	83.4	60.4	86.9
BERT[36]	86.1	66.8	87.7
MMBT[21]	85.8	64.8	88.7
基于高级特征融合的方法			
CentralNet[22]	89.3	64.7	89.8
GMU[23]	88.7	64.3	89.1
CBP[27]	90.2	66.1	89.8
VisualBERT[24]	87.5	64.7	86.1
PixelBERT[25]	89.1	66.5	88.9
VilT[26]	86.7	61.2	87.2
基于自注意力融合的方法			
DMCC[37]	88.0	-	87.7
SSE-Cross-BERT-DenseNet[30]	91.1	68.4	91.8
CrisisKAN[32]	90.1	65.3	90.9
ME Cross-attention[38]	93.5	85.6	93.6
GAFNet[39]	93.9	73.8	**95.3**
Ours	**94.2**	**94.1**	93.2

表6展示了本文方法与基线模型在信息性任务上的定量比较结果。从表6可以看出，本文模型在所有指标都达到了SOTA，相较于各项次优的GAFNet模型，本文模型在分类精度、宏平均F1分数、加权F1分数方数上分别提升了4.3%、7.2%和4.9%。信息性任务要求快速地提取各种事件的重点，并进行判断是否能提供信息性帮助。本文提出的融合外部知识的文本特征提取能提供了更多的上下文信息，加权注意力模块动态地调整每个模态的重要性，从而整合、理解不同模态的数据特征。

表7展示了本文方法与其它方法在损失程度任务上的定量比较结果，本文提出的方法在F1分数上高居前列，分类精度略低于ME Cross-attention模型，其原因在于损失程度任务上的训练数据较少，而ME Cross-attention模型则基于预训练模型进行损失程度分类。

图4描述了本文模型在CrisisMMD执行各项任务的损失函数变化，损失函数值在15轮左右的训练下趋于稳定，分类结果证明模型在较短的训练时

表 6 CrisisMMD 数据集上任务 2 的实验结果

Method	Accuracy（%）	MacroF1（%）	WeightedF1（%）
基于单模态的方法			
DenseNet[33]	81.6	79.1	81.2
BERT[36]	84.9	81.2	83.3
MMBT[21]	82.4	81.2	82.1
基于高级特征融合的方法			
CentralNet[22]	87.8	85.3	86.1
GMU[23]	87.2	84.6	85.7
CBP[27]	87.9	86.7	87.3
VisualBERT[24]	88.1	86.7	88.6
PixelBERT[25]	88.7	86.4	87.1
ViLT[26]	87.6	85.1	88.0
基于自注意力融合的方法			
DMCC[37]	92.2	-	92.2
Method	Accuracy（%）	MacroF1（%）	WeightedF1（%）
SSE-Cross-BERT-DenseNet[30]	89.3	88.1	89.3
CrisisKAN[32]	91.7	90.3	91.2
ME Cross-attention[38]	92.0	91.2	91.3
GAFNet[39]	94.7	92.0	94.3
Ours	**99.0**	**99.2**	**99.2**

表 7 CrisisMMD 数据集上任务 3 的实验结果

Method	Accuracy（%）	MacroF1（%）	WeightedF1（%）
基于单模态的方法			
DenseNet[33]	62.9	52.3	66.1
BERT[36]	68.2	45.0	61.1
MMBT[21]	65.3	52.1	69.3
基于高级特征融合的方法			
CentralNet[22]	71.1	57.4	68.7
GMU[23]	70.6	57.1	68.2
CBP[27]	65.8	60.4	69.3
VisualBERT[24]	66.3	56.7	62.1
PixelBERT[25]	65.2	57.3	63.7
ViLT[26]	67.6	58.4	65.0
基于自注意力融合的方法			
SSE-Cross-BERT-DenseNet[30]	72.6	59.7	70.4
CrisisKAN[32]	73.1	63.1	72.2
ME Cross-attention[38]	**76.5**	63.8	**75.7**
Ours	72.6	**79.2**	72.1

间内就能取得很好的效果，整体表现良好且稳定。图 5 展示了在每类突发事件上的具体识别结果。

表 8 展示不同突发事件类型的具体识别效果。从表 8 可以看出，在飓风事件中，所有类别的精度

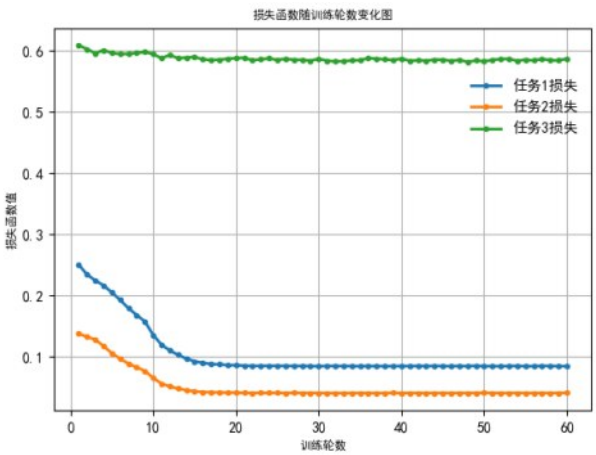

图 4 CrisisMMD 数据集上各项任务的损失函数变化曲线图

均有 90% 以上，说明飓风事件相对容易被识别。在火灾事件中，“非人道”类和“受伤或死亡的个人” 类精度高达 99.7% 和 99.6%。“失踪或被发现的个人”类相对于其他的类别的火灾精确度略低，是因为该类别的训练数据量很少。在地震事件中，由于“非人道”类事件数据量占该事件的 25% 以上，该类别的类分类精度最高。在洪水事件中，“救援和捐赠”类和“基础设施和公用设施损坏”比起其他类更容易被识别，它们的分类精度有 99% 和 98%，这可能说明在洪水的影响下基础设施方面

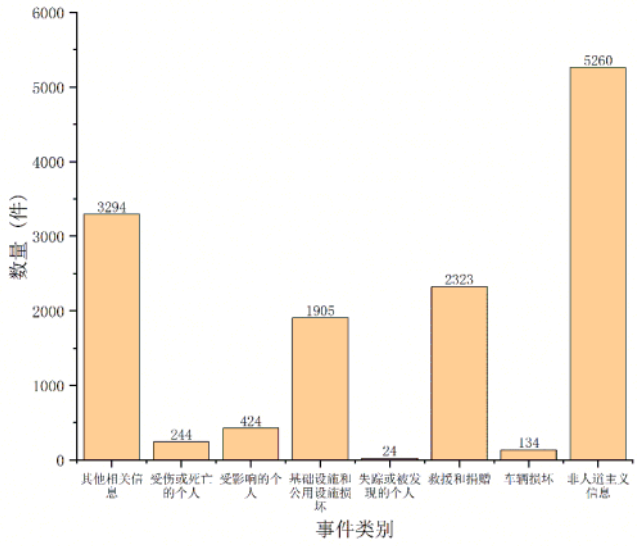

图5　CrisisMMD数据集任务1-事件识别结果图

影响较大，救援和捐赠行为的需求更大。

表8 CrisisMMD不同类型突发事件的信息识别效果

类别	事件			
	飓风	火灾	地震	洪水
受影响的个人	99.6	90.9	92.8	95.4
基础设施和公用设施损坏	95.4	97.7	94.3	98.0
受伤或死亡的个人	92.2	99.6	97.4	87.0
失踪或被发现的个人	98.2	75.0	87.5	96.7
救援和捐赠	92.3	98.7	99.4	99.0
车辆损坏	90.6	95.2	86.2	-
其他相关信息	96.4	81.8	98.3	96.2
非人道	96.9	99.7	100.0	94.2

4.3.2　自建数据集的实验结果分析

表9展示了本文方法在自建数据集上进行任务1的对比实验结果，实验结果表明，本文提出方法在各项指标上均取得了最优性能。

表9　　自建数据集上任务1的实验结果

Method	Accu-racy(%)	Macro F1(%)	Weighted F1(%)
SSE-Cross-BERT-DenseNet[30]	94.3	95.7	93.6
CrisisKAN[32]	97.6	97.6	96.4
Ours	**99.3**	**98.3**	**96.9**

在自建数据集上，基于注意力融合的多模态识别模型整体表现出了高准确率和稳定的性能，三种基于注意力融合的模型在该数据集上均都达到了90%以上的准确率。这表明本文的自建数据集提供了足够的信息和特征以支持模型的有效学习和泛化。与CrisisKAN相比，本文的模型准确率提高了约1.7%，达到了99.37%，取得了较好的识别结果。

自建数据集上的训练损失函数的变化曲线图和事件识别结果如图6、图7所示。图6描述了本文模型在自建数据集执行任务1的损失函数变化，图7描述了本文方法在自建数据集上的事件识别结果。

根据损失图和结果图可知，损失函数值在20轮左右的训练下趋于稳定，并且分类结果证明模型在较短的训练时间内就能取得很好的效果，整体表现良好且稳定。

4.4　消融实验

为验证各模块对模型性能的影响，本文在自建数据集和CrisisMMD数据集上设计了多种消融实验，验证各模块的有效性。

4.4.1　多支路自注意力增强的影响

为评估多支路自注意力增强模块对模型性能的影响，本文在与原始DenseNet相同参数配置下进行对比实验，实验结果如表10所示。

表10的实验结果表明，与原始的DenseNet相比，使用了多支路注意增强后的DenseNet精度在自建数据集上获得了5.1%的提升；在CrisisMMD数据集上获得了3.5%的提升，这说明融合多支路

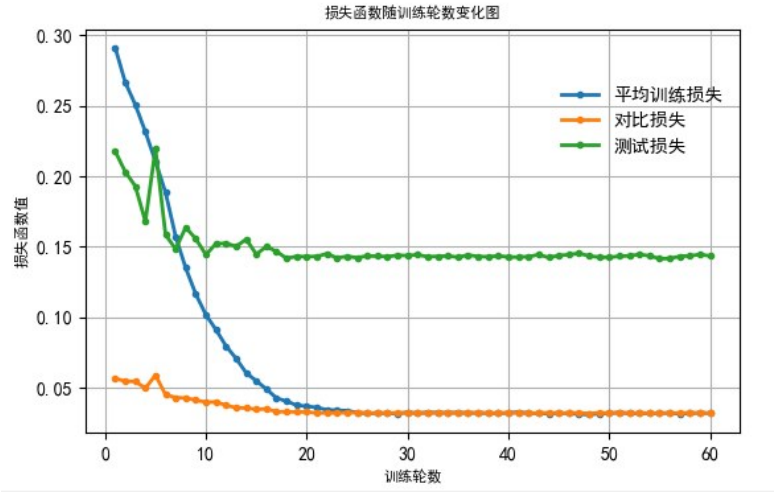

图6　自建数据集上任务1的损失函数变化曲线图

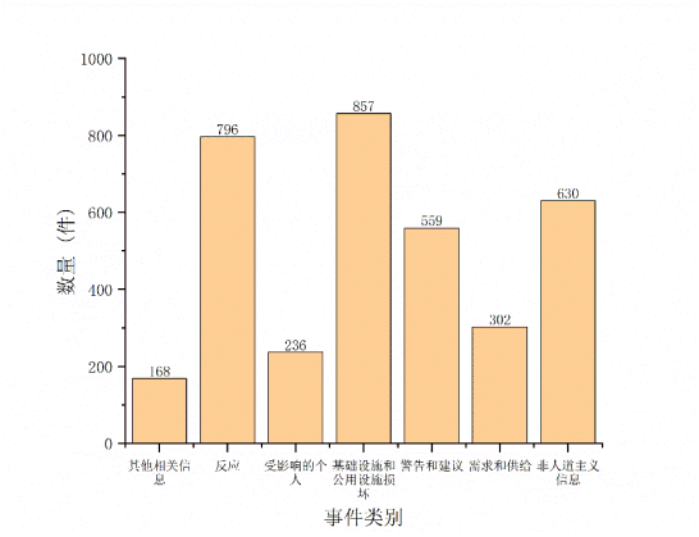

图7　自建数据集任务1-事件识别结果图

表10　　**多支路注意力增强对模型性能的影响**

图像特征提取模块	自建数据集			CrisisMMD数据集		
	Accuracy	MacroF1	Weighted F1	Accuracy	MacroF1	Weighted F1
原始DenseNet	94.2	90.6	88.1	90.7	93.8	92.8
多支路自注意增强	**99.3**	**98.3**	**96.9**	**94.2**	**94.1**	**93.2**

注意力的图像特征编码器更能提取图像的复杂全局特征，有利于跨模态的事件识别。

4.4.2　跨模态注意力协同模块的影响

为评估跨模态注意力协同模块对模型性能的影响，对交互注意力模块和加权注意力模块在相同配置下进行消融实验，结果见表11。

表11的实验结果表明，在联合使用交互注意力和加权注意力时，模型的性能最佳。与直接连接、仅使用交互注意力和仅使用加权注意力相比，在自建数据集上分类精度上分别提高了7.7%、49.7%和3%；在CrisisMMD数据集上分类准确率分别提高了2.8%、42%和3.4%。。当图像和文本特征本身具有高度的区分性和信息丰富性，直接连接就能够充分利用这些特征实现较高准确率的识别。而仅使用交互注意力模块时，交互注意力机制对图像-文本间复杂交互的捕捉会强化噪声特征，影响

表 11　　**跨模态注意力协同模块对性能的影响**

方法	自建数据集			CrisisMMD 数据集		
	Accuracy	Macro F1	Weighted F1	Accuracy	Macro F1	Weighted F1
直接连接	91.6	92.6	89.7	91.4	88.2	84.2
仅有交互注意力	49.6	56.3	48.4	52.2	49.2	44.6
仅有加权注意力	96.3	93.7	96.4	90.8	91.1	88.4
交互注意力+加权注意力	**99.3**	**98.3**	**96.9**	**94.2**	**94.1**	**93.2**

识别准确率。而加权注意力和交互注意力的联合使用则可以弥补这一缺陷，基于不同模态数据的权重平衡进行特征融合，从而达到较高的识别准确率。

4.4.3　联合损失对性能的影响

为评估联合损失对模型性能的影响，将对交叉熵损失和对比损失在相同参数配置下做对比实验，其结果见表 12。

表 12 的实验结果表明，联合损失在各项指标上取得最优值。在自建数据集上相对于仅使用交叉熵损失，联合损失取得了 7.5% 的准确率提升；在 CrisisMMD 数据集上相对于仅使用交叉熵损失，联合损失取得了 3.3% 的准确率提升。在多模态数据不平衡的情况下，本文采用对比损失来最小化相似样本对的距离，并最大化不相似样本对的距离。交叉熵损失强调类别之间的区分度，对比损失则强调相似样本之间的距离。在仅使用对比损失的情况下，对比损失难以有效区分不同类间的特征差异，因此分类准确率较低。而交叉熵损失能直接优化分类边界，在联合使用对比损失和交叉熵损失时，模型能在最小化类内距离的同时最大化类间距离，从而达到较高的识别准确率。

5　结论

本文构建了融合注意力协同和对比学习突发事件识别模型，通过多支路自注意力增强的 DenseNet 模型和注入外部知识的 Electra 模型，缓解图像模态的语义复杂性和文本模态的有限约束性。同时，提出了跨模态注意力协同网络增强，增强图像-文本文模态的相关一致性，减少了无效或误导信息的传递。此外，设计了跨模态对比学习网络，解决了不同模态之间特征不平衡的问题。在多个数据集上进行了大量实验证明了本文算法的有效性和优越性。在处理大规模数据集时，注意力机制和监督算法的使用显著依赖计算资源和数据标签。因此，本文下一步将进一步优化注意力机制，减少计算资源依赖。同时，考虑采用无监督的学习方法，使模型适应不同的数据分布和场景，以应对海量突发事件数据的挑战。

表 12　　**联合损失对性能的影响**

损失函数	自建数据集			CrisisMMD 数据集		
	Accuracy	Macro F1	Weighted F1	Accuracy	Macro F1	Weighted F1
仅使用对比损失	18.7	50.1	16.5	15.2	16.9	20.4
仅使用交叉熵损失	91.8	96.8	95.4	89.9	89.7	86.3
联合损失	**99.3**	**98.3**	**96.9**	**94.2**	**94.1**	**93.2**

参考文献：

[1] 管泽礼, 杜军平, 薛哲等 . 基于强化联邦GNN的个性化公共安全突发事件检测 [J]. 软件学报, 2024, 35(4): 1774-1789.
Guan Zeli, Du Junping, Xue Zhe, et al. Personalized public Security Emergency Detection based on enhanced Federal GNN [J]. Journal of Software, 2024, 35(4): 1774-1789.

[2] Abhina Kumar, Jyoti Prakash Singh, Nripendra P. Rana, et al. Multichannel convolutional neural network for the identification of eyewitness tweets of disaster[J]. Information Systems Frontiers, 2023, 25(4): 1589-1604.

[3] 周红磊, 张海涛, 栾宇等 . 基于文本-图像增强的突发事件识别及分类方法研究 [J]. 情报理论与实践, 2024, 47(4): 181-188.
Zhou Honglei, ZHANG Haitao, Luan Yuet al. Research on emergency recognition and classification based on text-image enhancement [J]. Information Theory & Practice, 2024, 47(4): 181-188.

[4] Wang Zhihong, Guo Yi, Wang Jiahui. Empower Chinese event detection with improved atrous convolution neural networks[J]. Neural Computing and Applications, 2021, 33(11): 5805-5820.

[5] Li Jiankai, Wang Yuhong, Li Weixin. MGMP: Multimodal graph message propagation network for event detection[C]//International Conference on Multimedia Modeling. Cham: Springer International Publishing, 2022: 141-153.

[6] 陈宏, 钱胜胜, 李章明等 . 基于多模态掩码Transformer网络的社会事件分类 [J]. 北京航空航天大学学报, 2024, 50(2): 579-587.
Chen Hong, QIAN Shengsheng, Li Zhangming, et al. Social event classification based on Multimodal Mask Transformer Network [J]. Journal of Beijing University of Aeronautics and Astronautics, 2024, 50(2): 579-87.

[7] Rani Koshy, Sivasankar Elango. Multimodal tweet classification in disaster response systems using transformer-based bidirectional attention model[J]. Neural Computing and Applications, 2023, 35(2): 1607-1627.

[8] Muhammad Mujahid, Khadija Kanwal, Furqan Rustam, et al. Arabic ChatGPT tweets classification using RoBERTa and BERT ensemble model[J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2023, 22(8): 1-23.

[9] Prakash Babu Yandrapati, R. Eswari. Classifying informative tweets using feature enhanced pre-trained language model[J]. Social Network Analysis and Mining, 2024, 14(1): 48.

[10] Zhou Han, Yin Hongpeng, Zheng Hengyi, et al. A survey on multimodal social event detection[J]. Knowledge-Based Systems, 2020, 195: 105695.

[11] Mathilde Brousmiche, Jean Rouat,, Stéphane Dupont. Multimodal attentive fusion network for audio-visual event recognition[J]. Information Fusion, 2022, 85: 52-59.

[12] Abhinav Kumar, Jyoti Prakash Singh, Yogesh K. Dwivedi,et al. A deep multi-modal neural network for informative Twitter content classification during emergencies[J]. Annals of Operations Research, 2022: 1-32.

[13] Iustin Sirbu, Tiberiu Sosea, Cornelia Caragea, et al. Multimodal semi-supervised learning for disaster tweet classification[C]//Proceedings of the 29th international conference on computational linguistics. 2022: 2711-2723.

[14] Sreenivasulu Madichetty, Sridevi M . A novel method for identifying the damage assessment tweets during disaster[J]. Future Generation Computer Systems, 2021, 116: 440-454.

[15] Sreenivasulu Madichetty, Sridevi M. A stacked convolutional neural network for detecting the resource tweets during a disaster[J]. Multimedia tools and applications, 2021, 80(3): 3927-3949.

[16] Abhishek Upadhyay, Yogesh Kumar Meena, Ganpat Singh Chauhan. SatCoBiLSTM: Self-attention based hybrid deep learning framework for crisis event detection in social media[J]. Expert Systems with Applications, 2024, 249: 123604.

[17] Firoj Alam, Ferda Ofli, Muhammad Imran. Processing social media images by combining human and machine computing during crises[J]. International Journal of Human – Computer Interaction, 2018, 34(4): 311-327.

[18] Daryl B. Valdez, Rey Anthony G. Godmalin. A deep learning approach of recognizing natural disasters on images using convolutional neural network and transfer learning[C]//Proceedings of the international conference on artificial intelligence and its applications. 2021: 1-7.

[19] Christos Kyrkou, heocharis Theocharides. Deep-Learning-Based Aerial Image Classification for Emergency Response Applications Using Unmanned Aerial Vehicles[C]//CVPR workshops. 2019: 517-525.

[20] Ganesh Nalluru, Rahul Pandey, Hemant Purohit. Relevancy classification of multimodal social media streams for emergency services[C]// 2019 IEEE International Conference on Smart Computing (SMARTCOMP). IEEE, 2019: 121-125.

[21] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, et al. Supervised multimodal bitransformers for classifying images and text[J]. arXiv preprint arXiv:1909.02950, 2019.

[22] Valentin Vielzeu, Alexis Lechervy, Stephane Pateux, et al. Centralnet: a multilayer approach for multimodal fusion[C]//Proceedings of the European Conference on Computer Vision (ECCV) Workshops. 2018: 575-589.

[23] John Arevalo, Thamar Solorio, Manuel Montes-y-Gómez, et al. Gated multimodal networks[J]. Neural Computing and Applications, 2020, 32: 10209-10228.

[24] Liunian Harold Li, Mark Yatskar, Da Yin, et al. Visualbert: A simple and performant baseline for vision and language[J]. arXiv preprint arXiv:1908.03557, 2019.

[25] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, et al. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers[J]. arXiv preprint arXiv:2004.00849, 2020.

[26] Wonjae Kim, Bokyung Son, Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision[C]//International conference on machine learning. PMLR, 2021: 5583-5594.

[27] Akira Fukui, Dong HukPark, Daylen Yang, et al. Multimodal compact bilinear pooling for visual question answering and visual grounding[J]. arXiv preprint arXiv:1606.01847, 2016.

[28] Wang Jing, Yang Shuo, Zhao Hui. Crisis event summary generative model based on hierarchical multimodal fusion[J]. Pattern Recognition, 2023, 144: 109890.

[29] Lv Jiandong, Wang Xingang, Shao Cuiling. AMAE: Adversarial multimodal auto-encoder for crisis-related tweet analysis[J]. Computing, 2023, 105(1): 13-28.

[30] Mahdi Abavisani, Wu Liwei, Hu Shengli, et al. Multimodal categorization of crisis events in social media[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 14679-14689.

[31] Wu Xuehua, Mao Jin, Xie Hao, et al. Identifying humanitarian information for emergency response by modeling the correlation and independence between text and images[J]. Information Processing & Man-

agement, 2022, 59(4): 102977.

[32] Shubham Gupta, Nandini Saini, Suman Kundu, et al. CrisisKan: Knowledge-infused and explainable multimodal attention network for crisis event classification[C]//European Conference on Information Retrieval. Cham: Springer Nature Switzerland, 2024: 18-33.

[33] Mao Yudong, Jiang Qiuping, Cong Runmin, et al. Cross-modality fusion and progressive integration network for saliency prediction on stereoscopic 3D images[J]. IEEE Transactions on Multimedia, 2021, 24: 2435-2448.

[34] Kevin Clark, Minh-Thang Luong, Quoc V. Le, et al. Electra: Pre-training text encoders as discriminators rather than generators[J]. arXiv preprint arXiv:2003.10555, 2020.

[35] Firoj Alam, Ferda Ofli, Muhammad Imran. Crisismmd: Multimodal twitter datasets from natural disasters[C]//Proceedings of the international AAAI conference on web and social media. 2018, 12(1).

[36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, 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 for computational linguistics: human language technologies, volume 1 (long and short papers). 2019: 4171-4186.

[37] Mariham Rezk, Noureldin Elmadany, Radwa K. Hamad,et al. Categorizing crises from social media feeds via multimodal channel attention [J]. IEEE Access, 2023, 11: 72037-72049.

[38] Liang Tao, Lin Guosheng, Wan Mingyang, et al. Expanding large pre-trained unimodal models with multimodal information injection for image-text multimodal classification[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 15492-15501.

[39] Onkar Susladkar,Gayatri Deshmukh, Dhruv Makwana, et al. Gafnet: A global fourier self attention based novel network for multi-modal downstream tasks[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023: 5242-5251.

[作者简介]

黄少年（1977-），女，博士，副教授，主要研究方向为机器学习、计算机视觉、多媒体信息处理等。

彭永涛（2002-），男，硕士研究生，主要研究方向为自然语言处理、突发事件识别等。

文沛然（2000-），男，硕士研究生，主要研究方向为图像识别、多模态信息处理等。

刘耀（1976-），男，博士，湖南工商大学副教授，主要研究方向为情感计算、边缘智能。