◎模式识别与人工智能◎

多模态混合注意力机制的虚假新闻检测研究

刘华玲,陈尚辉,乔 梁,刘雅欣
上海对外经贸大学 统计与信息学院,上海 201620

摘 要:探索高效的模态表示和多模态信息交互方法一直是多模态虚假新闻检测领域的热门话题,提出了一项新的虚假新闻检测技术(MAM)。MAM方法使用结合位置编码的自注意力机制和预训练的卷积神经网络分别提取文本和图像特征;引入混合注意力机制模块进行文本与图像特征交互,该模块使用了层级特征处理方法来减少多模态交互时产生的冗余信息,又使用了双向的特征融合手段保证训练信息的完整性;加权融合多模态特征并将其输入全连接网络中进行真假新闻分类。对比实验结果表明:相比现有的多模态基准模型,该方法几乎在各个分类指标上都提高3个百分点左右,此外,可视化实验发现混合注意力机制获得的多模态特征具有更强的泛化能力。

关键词:虚假新闻检测;多模态分析;注意力机制;特征融合
文献标志码:A **中图分类号:**TP183 doi:10.3778/j.issn.1002-8331.2202-0204

Multimodal False News Detection Based on Fusion Attention Mechanism

LIU Hualing, CHEN Shanghui, QIAO Liang, LIU Yaxin
School of Statistics and Information, Shanghai University of International Business and Economics, Shanghai 201620, China

Abstract: Exploring efficient modal representation and multimodal information interaction methods has always been a hot topic in the field of multimodal information detection, for which a new fake news detection technology(MAM) is proposed. The MAM method uses a self-attention mechanism combined with position coding and a pre-trained convolutional neural network to extract text and image features respectively. The introduction of a mixed-attention mechanism module for text and image feature interaction, which uses hierarchical feature processing methods to reduce redundant information generated during multimodal interactions. A two-way feature fusion method is used to ensure the integrity of the training information. The multimodal features are weighted and fed into the fully connected network for true and false news classification. The comparative experimental results show that compared with the existing multimodal reference model, the method is almost improved by about 3 percentage points on each classification index, and the visualization experiment finds that the multimodal features obtained by the mixed attention mechanism have stronger generalization ability.

Key words: false new detection; multimodal analysis; attention mechanism; feature fusion

新闻是人们了解外界信息的主要渠道,新媒体时代下,繁杂多样的社交平台促使新闻的传播环境、传播形式和传播内容发生巨大变化,社交媒体在给人们获取信息便利的同时,也成为虚假新闻恣意传播的渠道。虚假新闻是由专业媒体发布,通过操控舆论来达成某种政治或经济目的的手段。虚假新闻的危害极大,如果不能被及时发现并遏止,极易引起经济衰退和社会动荡,据2019年CHEQ和巴尔的摩大学的经济研究报道,全球每年因虚假新闻造成的损失高达780亿美元[1]。在这复杂环境下,为了提高读者对假新闻的防范意识,一些平台,如Twitter、Facebook、新浪微博等,会提供信息检测基站来识别假新闻,但这些平台需要花费大量资金聘请领域专家来应对各类假新闻,耗时又耗力。因此,探索更加智能高效的自动化虚假新闻检测方法具有重要意义。

早期的虚假新闻自动检测方法侧重于从新闻文章中提取特征进行分类,如文章情感特征[2]、主题特征[3]等。虽然新闻的文本内容是影响虚假新闻检测的重要因素,但新闻的附加内容(图像、音频、帖子等)往往包含

基金项目:国家社科基金重大项目(21ZDA105);上海哲学社会科学规划课题(2018BJB023)。
作者简介:刘华玲(1964—),女,博士,教授,研究方向为机器学习、人工智能,E-mail:liuhl@suibe.edu.cn;陈尚辉(1998—),男,硕士研究生。
收稿日期:2022-02-22 **修回日期:**2022-07-05 **文章编号:**1002-8331(2023)09-0095-09

大量的背景信息[4],这些背景信息和新闻文本内容相互补充,可以提高虚假新闻的检测性能。为此,学者们提出了一系列有助于提高检测性能的指标,如新闻页面的可信度、新闻作者特征、帖子中的用户情感特征等。这些研究极大地推进了自动化虚假新闻检测的进程。

然而,假新闻内容非常复杂,依赖这些手工提取的特征很难捕捉其中潜藏的高层语义信息。近年来,由于深度学习模型强大的表示学习能力,许多多模态特征学习任务开始利用深度神经网络学习模态和模态之间的高阶表示。这些研究尝试选取深度语言模型[5]、预训练的卷积神经网络(convolutional neural network,CNN)以及变分自编码器(variational auto-encoders,VAE)[6]等作为多模态特征提取模型,并结合早期融合[7]、晚期融合[8]、跨模态注意力融合[9-10]等不同的多模态特征融合方法构建端到端的多模态检测框架,尽管这些方法在虚假新闻检测任务中有良好的表现效果,但仍然存在以下不足之处:一方面,使用传统的机器学习和深度学习模型进行特征抽取时,会有大量任务需要的信息在模型训练时被丢弃;另一方面,现有的多模态特征融合方法会导致不同模态中任务无关信息的累积,产生大量冗余信息。

针对现有研究存在的问题,本文的设计如下:首先,使用多头注意力机制提取文本上下文信息,这种模型可以关注文本中每个词的信息,有效缓解模型训练时造成的信息损失问题,接着通过预训练的VGG19模型获得新闻图像特征,最后,在多模态特征融合交互部分,借鉴命名实体识别中联合注意力机制的思想,提出了混合注意力模块,该模块通过两次注意力机制交互和残差连接的方法有效地提取新闻中的多模态融合特征,并减少冗余信息的产生。

本文的贡献如下:

(1)使用多头注意力机制深入挖取新闻文本的上下文特征,减少模型在训练时造成的信息损失。

(2)使用混合注意力机制模块进行多模态特征融合交互,有效缓解了多模态特征融合时存在的信息冗余问题。

(3)通过在Weibo A和Weibo B两个数据集上的对比实验和可视化分析,证明了本文提出方法的有效性,一定程度上提高了虚假新闻检测的准确率。

1 相关工作

早期学者们对自动化虚假新闻检测的研究是基于文本内容进行的,这些研究方法可以总结为基于特征构建的技术以及基于深度学习的技术。基于特征构建的方法又可以分为显示特征的构建以及隐式特征的构建[11],显示特征指可以从新闻内容中直接选取出来的特征,包含句子长度、主题词、新闻事件个数等。隐式特征指无法直接获取,需要通过数值计算或关联分析得到的隐藏特征,包含情感特征、用户可信度以及质疑率等。相对于基于特征构建的方法,基于深度学习方法可以自动学习文本中的潜在特征表示,具有强拟合效果,但缺乏可解释能力,主要的模型有循环神经网络[12]、自注意力机制[13]、图神经网络[14]等。

基于图片内容的虚假新闻检测领域发展较晚,早期的图片模态检测依赖于手工构造特征,如颜色、线条[15-16]等,将这些人工构造的特征放入分类器训练。卷积神经网络的发展促使这一领域向前发展了一大步,Shu等[17]构建了MVNN框架,通过基于CNN模型的频域子网络捕获假新闻图像物理层面的特征,通过基于CNN-RNN模型的像素域子网络捕捉假新闻图像语义级别的特征,然后通过融合子网络动态地融合这些特征,得到的模型优于传统模型。然而,受限于人们的情感等信息与图片语义信息的"鸿沟",该领域的发展还存在巨大挑战,现有的研究大多还停留在对篡改和误导性的图片进行识别,作为检验虚假新闻的一个评据。

相比传统的文本新闻,带有图片的新闻容易吸引读者注意力,虚假新闻通常带有极具情感煽动性的文字和视觉冲击力的图片或视频,此外,由于真实信息与虚假新闻混杂,虚假新闻一般很难被人为辨别。相较于单模态虚假新闻检测技术,多模态虚假新闻更具灵活性、真实性,大量研究表明[18-21],在相同数据集下,多模态虚假新闻检测模型性能要高于单模态模型。

Jin等[18]最早将基于深度学习的多模态特征学习方法引入虚假新闻检测中,他们提出一种具有注意力机制的循环神经网络来融合多模态特征,将图片、文本和背景特征进行融合,得到了比单模态虚假新闻检测方法更好的结果;为了提高多模态检测方法在新数据集上的泛化性能,Wang等[19]结合对抗学习的思想提出了EANN模型,该模型在原有的分类任务上,加入事件分类这一子任务,诱导模型学习与事件无关的特征提高了模型的泛化能力;Khattar等[6]提出了一种端到端的多模态变分自编码器模型(MVAE),通过学习多模态特征的潜在分布来探求假新闻的分布规律。还有一些学者考虑通过外部知识来提高模型检测性能,Qian等[20]在传统多模态框架下加入知识图谱模块学习新闻文本实体蕴藏的外部知识;然而在整个新闻文本上构建知识图训练时需要耗费大量时间,针对这一缺陷,Mayank等[21]提出了只针对新闻标题实体的一个虚假新闻检测框架,在该框架下,他们使用双向的长短期记忆网络(long short term memory,LSTM)网络学习新闻标题特征,同时识别并抽取新闻标题实体,构建知识图谱获取外部信息特征,最后将两部分特征拼接融合做二分类任务。

然而,纵观这些学者提出的模型,他们都过多关注于寻找合适的模型提高提取文本和图像特征的精度,导致这些模型在训练时会损失大量任务相关信息。Singhal

等[5]曾提出了使用预训练的双向语言模型(bidirectional encoder representation from transformers,BERT)学习新闻文本特征,最终表现结果比基准的多模态模型更优。受到该方法的启发,本文使用多头注意力机制提取文本特征,结合位置编码的自注意力机制模型可以提取文本序列中的关键信息,而且模型相对简易,训练时造成的信息损失较少。此外,这些研究也都没有考虑过多模态特征融合时造成的信息冗余问题,因此,本文借鉴Zhang等[22]在命名实体识别任务中提出的联合注意力机制,该方法曾在多模态情感分析任务中取得了不错的成果[23],提出使用融合注意力机制来学习多模态交互特征,来减少多模态特征融合时产生的冗余信息。

2　多模态检测模型构建

本章提出的模型包含以下几个部分内容:文本和图像的特征提取、混合注意力机制、多模态特征融合及分类,具体的框架如图1所示。

2.1　文本特征提取

新闻文本特征提取是虚假新闻检测的核心模块,文本特征的好坏直接影响检测精度,本节对分词后的新闻文本使用Word2vec方法进行词嵌入。得到的新闻文本向量为:

$$f_T = f_{\text{word2vec}}(W) \tag{1}$$

由于新闻文本是时间序列数据,文本每个词都包含时间序列信息,为了体现词汇在文本中的时间先后信息,借鉴Transformer中位置编码的思想[24],为每个词添加位置信息。文本序列中词位置编码的由公式(2)定义:

$$\begin{cases} PR(pos,2k) = \sin(pos/10\,000^{2k/d_{\text{model}}}) \\ PR(pos,2k+1) = \cos(pos/10\,000^{2k/d_{\text{model}}}) \end{cases} \tag{2}$$

其中,$pos \in [0,1,\cdots,L_T]$,L_T 表示文本序列长度,$k \in [0, d_T/2]$,d_T 表示词向量维度。最终每条新闻文本的初始特征表示定义为:

$$f_T' = f_T + PR(f_T) \tag{3}$$

针对文本序列,传统模型通常使用时间序列模型、循环神经网络进行文本特征提取,这些模型的使用可能会丢失任务相关的重要信息。利用自注意力机制处理文本序列不仅能有效提取文本中的关键信息,而且模型简洁,不容易在模型训练时丢失信息。因此,本文使用自注意力机制从文本中提取能反映新闻真实性的文本特征。

注意力机制的使用如下:

$$f_T'' = softmax(\frac{\boldsymbol{Q}_T \boldsymbol{K}_T^{\mathrm{T}}}{\sqrt{d_k}})\boldsymbol{V}_T \tag{4}$$

其中,$\boldsymbol{Q}_T = \boldsymbol{W}_Q f_T', \boldsymbol{K}_T = \boldsymbol{W}_K f_T', \boldsymbol{V}_T = \boldsymbol{W}_V f_T', \boldsymbol{W}_Q \in \mathbb{R}^{d_T \times d_k}$,$\boldsymbol{W}_K \in \mathbb{R}^{d_T \times d_k}$,$\boldsymbol{W}_V \in \mathbb{R}^{d_T \times d_v}$ 通过训练学习得到。

相比于单头自注意力机制,多头自注意力能更好地处理上下文信息。对于有 H 个头的多头自注意力机制,通过使用 H 个不同的线性映射,将 $\boldsymbol{Q}_T$、$\boldsymbol{K}_T$、$\boldsymbol{V}_T$ 划分为 H 个不同的子空间。其中第 h 个头的 $\boldsymbol{Q}_T^h$、$\boldsymbol{K}_T^h$、$\boldsymbol{V}_T^h$ 表示为 $\boldsymbol{Q}_T^h = \boldsymbol{Q}_T \times \boldsymbol{W}_Q^h$,$\boldsymbol{W}_Q^h \in \mathbb{R}^{d_T \times \frac{d_k}{H}}$,$\boldsymbol{K}_T^h = \boldsymbol{K}_T \times \boldsymbol{W}_K^h$,$\boldsymbol{W}_K^h \in \mathbb{R}^{d_T \times \frac{d_k}{H}}$,$\boldsymbol{V}_T^h = \boldsymbol{V}_T \times \boldsymbol{W}_V^h$,$\boldsymbol{W}_V^h \in \mathbb{R}^{d_T \times \frac{d_v}{H}}$,由此可得第 h 个头自注意力机制为:

$$f_T^h = soft\,max(\frac{\boldsymbol{Q}_T^h (\boldsymbol{K}_T^h)^{\mathrm{T}}}{\sqrt{d_k/h}})\boldsymbol{V}_T^h \tag{5}$$

多头自注意力表示为:

$$f_{T,H} = (f_T^1 \oplus f_T^2 \oplus \cdots \oplus f_T^H) \times \boldsymbol{W}$$

其中,$f_{T,H} \in \mathbb{R}^{L_T \times d_v}$,$\boldsymbol{W} \in \mathbb{R}^{d_v \times d_v}$,在本文中,$d_T = d_k = d_v$。

最后,对文本序列维度使用平均汇聚的操作得到文本特征:

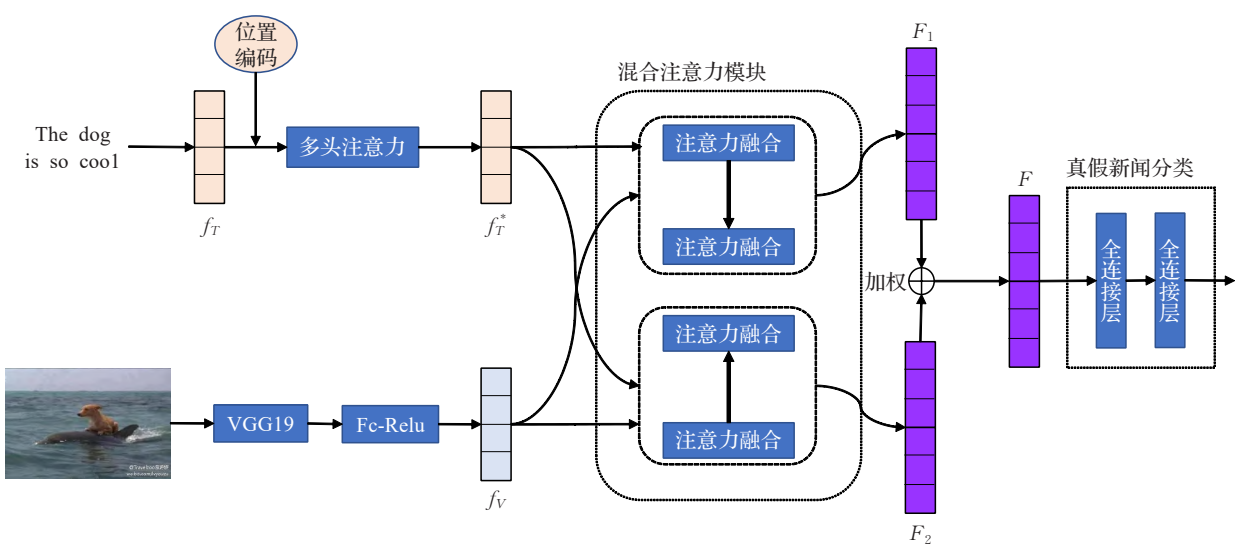

图1　模型框架图

Fig.1　Model frame

$$f_T'' = Avgpool(f_{T,H}) \tag{6}$$

其中，$f_T'' \in \mathbb{R}^{d_v}$。

2.2　图像特征提取

除新闻文本之外，新闻图片可以快速帮助读者理解新闻表达的内容。因此，在新闻的多模态分析中，新闻图像的处理也十分重要。

大量研究表明，卷积神经网络具有强大的图像特征提取能力，在计算机视觉任务中取得巨大的成功。本文使用 VGG19 卷积神经网络作为图像特征提取器，VGG19 事先在包含 1 500 万张图片的 ImageNet 数据集上预训练，取 VGG19 最后一层作为提取的图像特征。为了避免模态之间维度不同造成的信息偏倚问题，在 VGG19 网络后添加一个全连接层进行降维，使得图像特征与文本特征在维度上保持一致。图像特征表示为：

$$f_v = \sigma(W \times VGG_{19}(V) + b) \tag{7}$$

其中，$f_v \in \mathbb{R}^{d_v}$，W，b 是全连接层的参数。

2.3　混合注意力机制模块

通常来说，图像中蕴含的信息只有部分是与文本内容相关的，如果不考虑模态信息的交互，那么最终得到的多模态特征必然包含大量与检测任务无关的冗余信息。为此部分研究通过引入注意力权重的形式提取图像中与文本关联性较强的区域，但是在构建文本和图像特征时是独立进行的，即分别用词引导注意力机制构建图像特征和用图引导注意力机制构建文本特征，这种方式是有效的，但由于操作的独立性仍然不可避免产生冗余信息。为此，联合注意力机制在注意力机制融合的基础上做了改进，首先，基于模型提取得到的文本特征引导生成图片注意力权重分布，再经过加权获得新的图片特征，以此减少图片中不重要区域对检测的影响；然后将新的图片特征引导生成对应文本序列注意力权重分布，经过加权获得新的文本特征。这种方法进一步地削减了文本中的冗余信息，突出了模态之间的交互影响。然而这种处理方式可能弱化模态数据中某些任务相关的信息，据此，本文在联合注意力的基础上提出了混合注意力机制，具体构建如图 2 所示。

混合注意力机制包含两个模块，两个模块的结构相同，第一个模块（图 2 左）和联合注意力机制的处理方式是一致的，但为了防止在多模态交互过程中丢失关键信息，本文引入了残差连接的概念，具体实现过程如下：

对于文本引导的图片特征更新操作，使用经自注意力提取的文本特征 f_T'' 引导完成图片特征的更新，首先将文本特征放入单层神经网络进行激活，然后通过 Softmax 函数生成各个图片区域的注意力权重分布。

$$C_1 = relu(W_{C_1} f_T'' + b_{C_1}) \tag{8}$$

$$\alpha_1 = softmax(W_{\alpha_1} C_1 + b_{\alpha_1}) \tag{9}$$

最后将得到的注意力分数与原图片对应区域进行加权求和，得到的特征为 $\hat{V} = [\hat{v}_1, \hat{v}_2, \cdots, \hat{v}_n]$。

$$\hat{v}_i = \sum \alpha_{1,i} f_{V_i},\ i = 1, 2, \cdots, n \tag{10}$$

$\hat{v}_i$ 表示第 i 张图片的新特征表示。之后再使用残差连接完成图像特征的更新：

$$V' = \hat{V} + f_v$$

然而还需要提取文本内容的重要信息，因此又引入第二步：图引导的注意力机制，首先使用新图片特征 V' 引导生成文本各个部分的注意力权重，注意力机制的作用对象为 f_T''，最终进行加权计算得到特征为 $\hat{T} = [\hat{t}_1, \hat{t}_2, \cdots, \hat{t}_n]$。

$$D_1 = relu(W_{D_1} V' + b_{D_1}) \tag{11}$$

$$\beta_1 = softmax(W_{\beta_1} D_1 + b_{\beta_1}) \tag{12}$$

$$\hat{t}_i = \sum \beta_{1,i} f_{T_i}'',\ i = 1, 2, \cdots, n \tag{13}$$

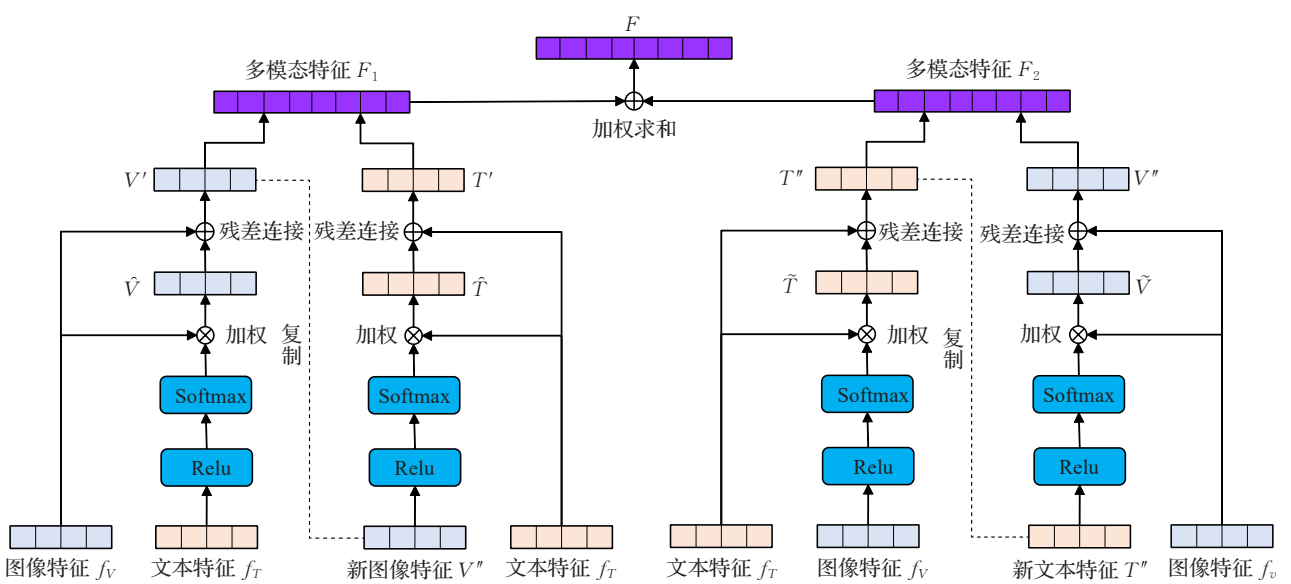

图 2　混合注意力机制构建模块

Fig.2　Hybrid attention mechanism construction module

其中，$W_{C_1}, W_{\alpha_1}, W_{D_1}, W_{\beta_1}, b_{C_1}, b_{\alpha_1}, b_{D_1}, b_{\beta_1}$ 是神经网络中可学的参数，接着通过残差连接的方式得到新文本特征：

$$T' = \hat{T} + f_T'' \tag{14}$$

第二模块（图2右）和第一模块结构是一致的，不同的是，第二模块首先进行的是图引导的文本特征更新操作，其次是词引导的文本特征更新，在更新过程中同样使用了残差连接。

图引导的文本特征更新操作如下：

$$D_2 = relu(W_{D_2} f_v + b_{D_2}) \tag{15}$$

$$\beta_2 = softmax(W_{\beta_2} D_2 + b_{\beta_2}) \tag{16}$$

$$\tilde{t}_i = \sum \beta_{2,i} f_{T_i}'',\ i = 1, 2, \cdots, n \tag{17}$$

得到特征 $\tilde{T} = [\tilde{t}_1, \tilde{t}_2, \cdots, \tilde{t}_n]$，接着使用残差连接更新文本特征：

$$T'' = \tilde{T} + f_T'' \tag{18}$$

词引导的图像特征更新操作如下：

$$C_2 = relu(W_{C_2} T'' + b_{C_2}) \tag{19}$$

$$\alpha_2 = softmax(W_{\alpha_2} C_2 + b_{\alpha_2}) \tag{20}$$

$$\tilde{v}_i = \sum \alpha_{1,i} f_{V_i}, i = 1, 2, \cdots, n \tag{21}$$

得到特征 $\tilde{V} = [\tilde{v}_1, \tilde{v}_2, \cdots, \tilde{v}_n]$，使用残差连接得到新图像特征

$$V'' = \tilde{V} + f_v \tag{22}$$

2.4　特征融合及分类

根据混合注意力机制，完成了对文本和图像特征的更新，其中，混合注意力机制第一模块得到了 T' 和 V'，第二模块得到了 T'' 和 V''；将每个模块的文本和图像特征进行拼接。

$$F_1 = V' \oplus T' \tag{23}$$

$$F_2 = T'' \oplus V'' \tag{24}$$

最终得到的多模态特征为两个模块的加权和。

$$F = \gamma F_1 + (1 - \gamma) F_2 \tag{25}$$

将多模态特征 F 输入含有两层隐藏层的全连接神经网络中进行新闻分类，并使用Softmax函数得到输出的分类结果，损失函数设置为交叉熵损失。

3　实验分析

3.1　数据集和评价指标

本文的实验在Weibo A和Weibo B两个数据集上进行。

Twitter数据集使用的是从Twitter平台上收集的虚假新闻数据集MediaEval2015，该数据集中包含开发集和测试集，每条新闻包含新闻文本、图像和社交信息等，是常用的多模态虚假新闻检测数据集。

Weibo A数据集同样应用在大量的虚假新闻检测研究中，该数据集中的真实新闻取自新华社等权威机构发布的新闻，假新闻是在微博上抓取的，这些新闻已被官方辟谣。新闻的发布时间从2012年5月到2016年1月。

Weibo B数据集是一个互联网虚假新闻检测挑战赛任务3的基准数据集，由Cao等[25]发布，数据集中共有38 471条文本以及34 096张对应的图片，其中，本文对数据集进行简单调整，取出既含有文本又含有图片的新闻。最终处理获得的数据分布如表1所示。

表1　数据集描述

Table 1　Description of datasets

类别	Twitter	Weibo A	Weibo B
真新闻	7 091	3 615	11 064
假新闻	4 140	4 108	10 768
总计	11 231	7 713	21 832

两个数据集中真实新闻与虚假新闻数量接近，不存在样本不平衡问题。实验中，每条新闻都对应一张图片，每个数据集以7∶1∶2划分为训练集、验证集、测试集。

本文选取分类任务中常用的精确率（accuracy）、准确率（precision）、召回率（recall）和F1分数（F1-score）等评价指标分析各个分类模型的效果，首先给出以下几个定义：

（1）TP（true positive）：被模型正确预测的正类样本数量。

（2）FN（false negative）：被模型错误预测为负类的正类样本数量。

（3）FP（false positive）：被模型错误预测为正类的负类样本数量。

（4）TN（true negative）：被模型正确预测的负类样本数量。

精确率表示被模型预测为正类的样本占所有实际正类样本中的比例，其结果可以由公式（13）计算得到：

$$P = \frac{TP}{TP + FP} \tag{26}$$

召回率表示被模型正确预测的正类样本的比例，其结果由公式（14）计算得到：

$$R = \frac{TP}{TP + FN} \tag{27}$$

某些情况下，模型不能同时满足既有高精确率又有高召回率，因此引入F1分数对这两指标进行综合考量，其结果由公式（15）计算得到：

$$F1 = \frac{2 \times P \times R}{P + R} \tag{28}$$

3.2　实验设置

本实验使用的编程语言为Python3.8，使用的深度学习框架为Pytorch1.10.0和Keras2.7.0，本文的新闻文本词嵌入维度为64，文本长度为316，超出部分直接截断，不足部分用0补齐；图片输入大小为224×224×3，激活函数包含Relu、Tanh和Softmax。

在训练过程中，损失函数选择交叉熵损失函数，批

量大小设置为100,训练次数epoch设置为100,为了提高虚假新闻检测模型的鲁棒性并防止过拟合,使用Adam作为目标函数的优化器,学习率为0.001,每个全连接层后连接一个Dropout层,Dropout值为0.5。

3.3 基准模型

Textual:使用Text-CNN模型提取新闻文本特征并分类。

Visual:使用预训练的ResNet50提取新闻图片特征并分类。

EANN:EANN使用Text-CNN模型提取新闻文本特征,同时使用预训练的VGG19模型提取新闻图像特征,最后将两部分特征拼接输入到分类器中,考虑到实验的公平性,本文只保留了假新闻分类模块,去除了新闻事件分类模块。

MVAE:使用BiLSTM和预训练的VGG19模型分别提取文本和图像特征,串联拼接两部分特征并使用变分自编码器学习多模态数据分布,再重构多模态特征用于分类。

JAM:JAM使用CNN-BiLSTM和预训练的VGG16模型分别提取新闻的文本和视觉特征,然后通过联合注意力机制完成多模态交互和融合。

3.4 实验结果分析

本节将本文提出的模型MAM和近些年的基准模型进行对比分析,实验结果如表2所示。

可以看到,本文的模型MAM在多项指标上优于对比模型。其中,在中文两个数据集Weibo A和Weibo B上,基于文本特征的方法有非常好的表现效果,甚至在Weibo B数据集中,基于文本的方法与经典的EANN模型不相上下,表明了文本内容在虚假新闻检测任务中起到核心作用;相比于文本来说,基于图像特征的方法性能较弱,这可能是因为大多数新闻的所附图像质量较差,不能反映新闻的关键信息导致的。在多模态研究方法中,基准模型的表现效果较差,其中,EANN模型使用对抗学习的思想学习事件不变的特征,MVAE模型学习多模态特征的潜在分布,这些模型在获得多模态特征过程中也增加了不必要的冗余信息,而JAM模型使用了联合注意力的思想,有效地减少了冗余信息的产生,在实验结果上都比这些经典的基准模型表现更好。而本文提出的MAM模型在JAM的基础上使用了多头注意力机制提取文本特征,并对联合注意力机制做出改进,增加残差连接来缓解重要信息损失,再通过两个模块双向提取多模态特征,在两个数据集的实验结果上,各个指标都比基准模型提高2~4个百分点。

在英文数据集中,模型的表现与中文数据集类似,JAM和MAM模型效果要明显高于经典的baseline模型,其中,在总准确率上,MAM比最好的模型要高出5个百分点,在假新闻的召回率和F1分数上,MAM比最好的模型提高10个百分点左右,这些优异的表现都证实了MAM模型的合理性和有效性。

3.5 消融实验

为了理清模型中各个模块的功能,本节将MAM各个模块拆解做消融实验,探索每个模块对实验结果的影响,实验结果如表3所示,在本文模型中,“MAM-图像”为去除原始图像内容的实验结果,“MAM-残差”为

表2　两个微博数据集上的模型对比结果

Table 2　Model comparison results on two Weibo datasets

Dataset	Method	Time/s	Accuracy	Fake News			Real News		
				Precision	Recall	F1-Score	Precision	Recall	F1-Score
Twitter	Textual	—	0.526	0.586	0.553	0.569	0.469	0.526	0.496
	Visual	—	0.596	0.695	0.518	0.593	0.524	0.700	0.599
	EANN	2 036.46	0.830	0.786	0.662	0.719	0.874	0.929	0.901
	MVAE	1 937.37	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	JAM	1 793.46	0.882	**0.923**	0.728	0.672	0.841	**0.983**	0.906
	MAM	1 946.61	**0.933**	0.894	**0.833**	**0.813**	**0.973**	0.956	**0.965**
Weibo A	Textual	202.21	0.832	0.860	0.816	0.838	0.804	0.850	0.827
	Visual	997.69	0.668	0.686	0.688	0.687	0.648	0.645	0.646
	EANN	1 689.50	0.836	0.843	0.851	0.847	0.828	0.819	0.824
	MVAE	1 653.74	0.750	0.713	0.864	0.781	0.812	0.629	0.709
	JAM	1 146.28	0.841	0.850	0.854	0.852	0.832	0.827	0.829
	MAM	1 656.91	**0.887**	**0.889**	**0.907**	**0.898**	**0.884**	**0.863**	**0.873**
Weibo B	Textual	1 016.56	0.948	0.959	0.938	0.948	0.939	0.960	0.949
	Visual	2 917.43	0.838	0.848	0.821	0.835	0.828	0.854	0.841
	EANN	4 818.79	0.950	0.967	0.935	0.949	0.935	0.968	0.952
	MVAE	4 514.46	0.872	0.883	0.904	0.892	0.861	0.851	0.862
	JAM	3 270.38	0.961	0.963	0.957	0.960	0.958	0.964	0.961
	MAM	4 469.24	**0.977**	**0.981**	**0.972**	**0.976**	**0.972**	**0.982**	**0.977**

表3 模型消融实验

Table 3 Model ablation experiment

Dataset	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F1-Score	Precision	Recall	F1-Score
Twitter	MAM-图像	0.526	0.586	0.553	0.569	0.469	0.526	0.496
	MAM-残差	0.903	**0.944**	0.703	0.736	0.863	**0.986**	0.920
	MAM-混合	0.873	0.891	0.688	0.708	0.856	0.971	0.910
	MAM	**0.933**	0.894	**0.833**	**0.813**	**0.973**	0.956	**0.965**
Weibo A	MAM-图像	0.862	0.863	0.882	0.872	0.862	0.840	0.851
	MAM-残差	0.870	0.862	0.895	0.878	0.879	0.843	0.861
	MAM-混合	0.867	0.884	0.853	0.868	0.850	**0.882**	0.866
	MAM	**0.887**	**0.889**	**0.907**	**0.898**	**0.884**	0.863	**0.873**
Weibo B	MAM-图像	0.967	0.977	0.956	0.966	0.958	0.977	0.967
	MAM-残差	0.972	0.973	0.971	0.972	**0.972**	0.974	0.973
	MAM-混合	0.969	0.972	0.966	0.969	0.967	0.973	0.970
	MAM	**0.977**	**0.981**	**0.972**	**0.976**	**0.972**	**0.982**	**0.977**

去除残差模块的实验结果,“MAM-混合”为去除混合注意力模块的实验结果,“MAM”为本文完整模型的实验结果。

在这些实验结果中,“MAM-图像”的实验结果较差,说明相比于单模态,多模态的分析方法能有效提高检测性能;以简单拼接方法进行多模态融合的“MAM-混合”模型实验效果要比利用混合注意力进行特征融合的“MAM-残差”模型更差,这表明了与简单拼接相比,通过混合注意力模块进行多模态交互减少了信息冗余;此外,增加了残差模块的混合注意力模型“MAM”可以进一步提高模型检测的准确率,说明了残差模块可以有效地缓解模型训练中的信息损失问题。

3.6 模型复杂度分析

模型复杂度通常包含模型的空间复杂度和时间复杂度,在深度神经网络中,模型的空间复杂度由模型的参数决定,模型的时间复杂度可以用模型的训练时间来表示,关于模型的算法复杂度分析结果可以见表4,其中,模型的训练时间选择在Twitter数据集上实验的训练时间。

表4 模型复杂度对比实验结果

Table 4 Model complexity versus experimental results

模型	准确率	时间/s	参数量/10^6
EANN	0.830	2 036.46	0.795
JAM	0.882	1 793.46	1.651
MAM	0.933	1 946.61	1.680

首先,JAM和MAM模型的参数量接近,但都比EANN模型高1倍左右,说明JAM和MAM在空间复杂度上更高,但是在训练时间上,EANN模型的训练时间更长,说明EANN模型的时间复杂度更高,在这种情况下,MAM模型的准确率相比于JAM提高5.1个百分点,相比于EANN提高10.1个百分点,由此可以证明MAM模型不存在复杂度的问题。

3.7 可视化分析

以Weibo B数据集为例,为了进一步探索混合注意力机制的作用,本节将JAM与MAM获得的多模态特征分布进行可视化比较。针对测试集中的提取到的模态特征,使用t-SNE算法将其映射到二维空间内,并在二维坐标图上做可视化。结果如图3所示,其中,(a)、(b)分别表示联合注意力机制(JAM)提取的多模态特征分布和混合注意力机制(MAM)提取的多模态特征分布,蓝色和棕色的点分别代表真实新闻和虚假新闻。

(a)JAM获取的多模态特征分布 (b)MAM获取的多模态特征分布

图3 t-SNE降维下的真假新闻特征分布

Fig.3 Characteristic distribution of true and false news under t-SNE dimensionality reduction

从图3可以看出,两种方法提取出的特征都能将大部分新闻正确划分;其中,JAM的特征分布相对紧凑,尤其在真假新闻混杂的区域,真假新闻特征很难区分,说明模型的泛化性能较弱;而MAM提取的特征分布相对松散,模型的泛化性能更好,真假新闻特征紧凑在一起的区域也更少,模型检测的性能更高,反映了混合注意力机制的合理性和良好的泛化能力。

3.8 超参数对模型的影响分析

本节以中文数据集Weibo A和Weibo B为例,讨论模型中重要参数对实验结果的影响,其中对模型影响较大的参数为多头自注意力机制的头数 H 和混合注意力机制中两模块的权重 γ。

首先分析多头注意力机制中头数 H 对实验结果的

影响,本文选取的 H 集合为{1,2,4,8},不同头数 H 的实验结果如表5所示。可以看到,相比于单头的自注意力机制,使用多头的方法在准确率上更高,而且随着头数的增加,模型的效果并不是不断变好,当 $H=2$ 时,模型的表现效果最好,之后随着头数的增加,准确率开始下降并保持平稳,这可能是因为头数的增加促使模型关注文本中更多的一般性特征,但也导致了一些文章特定信息的丢失。

表5 多头数量 H 对实验结果(准确率)的影响

Table 5 Influence of number of multiple heads H on experimental results(Accuracy)

Dataset	H			
	1	2	4	8
Weibo A	0.864	0.887	0.880	0.877
Weibo B	0.973	0.977	0.976	0.976

另一个影响模型效果的参数是权重参数 γ,本节中,γ 的取值集合为{0,0.2,0.4,0.6,0.8,1},需要注意的是,根据头数 H 对实验结果的影响,取模型表现最好的 H 做实验,在本文的两个数据集中,模型都取 $H=2$,最后,不同 γ 值得到的实验结果如表6所示。可以看到,随着 γ 取值的不同,模型的准确率会发生波动,而且不同数据集下,模型准确率最好的 γ 取值不一定相同,对于Weibo A,当 $\gamma=0.2$ 时,效果最好,说明混合注意力中第二模块贡献更高,而对Weibo B来说,$\gamma=0.6$ 时的准确率更高,表明混合注意力中两个模块的贡献度相近,这种情况的出现可能是新闻内容中不同模态数据质量不匹配造成的。

表6 参数 γ 对实验结果(准确率)的影响

Table 5 Effect of parameter γ on experimental results(Accuracy)

Dataset	γ					
	0	0.2	0.4	0.6	0.8	1
Weibo A	0.877	0.887	0.884	0.877	0.870	0.880
Weibo B	0.971	0.972	0.969	0.977	0.971	0.973

4 总结与展望

多媒体时代信息交流的便利促使虚假新闻在生活中恣意传播,由于缺乏有效监管,国家社会一直深受其害。为了探索更加高效的自动化虚假新闻检测方法,同时考虑到新闻内容表述的多模态性,本文从新闻文本和图像内容着手,构建一个端到端的多模态虚假新闻检测框架,使用结合位置编码的多头自注意力机制和预训练的VGG19网络提取文本特征和图像特征,并通过混合注意力机制模块进行多模态信息交互和融合,有效缓解了多模态检测模型存在的训练时的信息丢失问题和多模态融合时的信息冗余问题。最后,在两个数据集上进行模型对比实验、消融实验和参数选择等实验,从结果和可视化结果上验证了本文模型的有效性。

此外,本文提出的多模态虚假新闻检测模型还存在以下不足之处:模型只考虑了新闻文本和图像模态信息,后续可以尝试引入更多的模态内容提高模型性能,如新闻评论内容、新闻发布的背景内容、新闻视频内容等;受到现有的虚假新闻数据集限制,本文的模型在其他数据上的拟合效果还有待验证,模型还需要在更大规模且不同领域的真实数据集中训练提高其泛化能力;此外,将虚假新闻检测任务简单地归纳为二分类问题具有不合理性,后续的研究需要对新闻进行更细粒度的分析,如何将其转换为多分类任务甚至是回归任务也是未来工作的一大挑战。

参考文献:

[1] NGADIRON S,ABD AZIZ A,MOHAMED S S.The spread of Covid-19 fake news on social media and its impact among Malaysians[J].International Journal of Law, Government and Communication,2021,6(22):253-260.

[2] WU K,YANG S,ZHU K Q.False rumors detection on sina weibo by propagation structures[C]//2015 IEEE 31st International Conference on Data Engineering,2015:651-662.

[3] YANG F,LIU Y,YU X,et al.Automatic detection of rumor on sina weibo[C]//Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics,2012:1-7.

[4] ALAM F,CRESCI S,CHAKRABORTY T,et al.A survey on multimodal disinformation detection[J].arXiv:2103.12541,2021.

[5] SINGHAL S,SHAH R R,CHAKRABORTY T,et al.Spotfake:a multi-modal framework for fake news detection[C]//2019 IEEE Fifth International Conference on Multimedia Big Data(BigMM),2019:39-47.

[6] KHATTAR D,GOUD J S,GUPTA M,et al.Mvae:multimodal variational autoencoder for fake news detection[C]//The World Wide Web Conference,2019:2915-2921.

[7] BIRADAR S,SAUMYA S,CHAUHAN A.Combating the infodemic:COVID-19 induced fake news recognition in social media networks[J].Complex & Intelligent Systems,2022:1-13.

[8] HAMID A,SHIEKH N,SAID N,et al.Fake news detection in social media using graph neural networks and NLP techniques:a COVID-19 use-case[J].arXiv:2012.07517,2020.

[9] SONG C,NING N,ZHANG Y,et al.A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks[J].Information Processing & Management,2021,58(1):102437.

[10] QIAN S,WANG J,HU J,et al.Hierarchical multi-modal contextual attention network for fake news detection[C]//

Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval,2021:153-162.

[11] 高玉君,梁刚,蒋方婷,等.社会网络谣言检测综述[J].电子学报,2020,48(7):1421-1435.
GAO Y J,LIANG G,JIANG F T,et al.Social network rumor detection:a survey[J].Acta Electronica Sinica,2020,48(7):1421-1435.

[12] OSHIKAWA R,QIAN J,WANG W Y.A survey on natural language processing for fake news detection[J].arXiv:1811.00770,2018.

[13] WANG Y,HAN H,DING Y,et al.Learning contextual features with multi-head self-attention for fake news detection[C]//International Conference on Cognitive Computing.Cham:Springer,2019:132-142.

[14] BENAMIRA A,DEVILLERS B,LESOT E,et al.Semi-supervised learning and graph neural networks for fake news detection[C]//2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining(ASONAM),2019:568-569.

[15] GUPTA M,ZHAO P,HAN J.Evaluating event credibility on twitter[C]//Proceedings of the 2012 SIAM International Conference on Data Mining,2012:153-164.

[16] TIAN D P.A review on image feature extraction and representation techniques[J].International Journal of Multimedia and Ubiquitous Engineering,2013,8(4):385-396.

[17] SHU K,SLIVA A,WANG S,et al.Fake news detection on social media:a data mining perspective[J].ACM SIGKDD Explorations Newsletter,2017,19(1):22-36.

[18] JIN Z,CAO J,GUO H,et al.Multimodal fusion with recurrent neural networks for rumor detection on microblogs[C]//Proceedings of the 25th ACM International Conference on Multimedia,2017:795-816.

[19] WANG Y,MA F,JIN Z,et al.Eann:event adversarial neural networks for multi-modal fake news detection[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,2018:849-857.

[20] QIAN S,HU J,FANG Q,et al.Knowledge-aware multi-modal adaptive graph convolutional networks for fake news detection[J].ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM),2021,17(3):1-23.

[21] MAYANK M,SHARMA S,SHARMA R.DEAP-FAKED:knowledge graph based approach for fake news detection[J].arXiv:2107.10648,2021.

[22] ZHANG Q,FU J,LIU X,et al.Adaptive co-attention network for named entity recognition in tweets[C]//Thirty-Second AAAI Conference on Artificial Intelligence,2018.

[23] 范涛,吴鹏,王昊,等.基于多模态联合注意力机制的网民情感分析研究[J].情报学报,2021,40(6):656-665.
FAN T,WU P,WANG H,et al.Sentiment analysis of online users based on multimodal co-attention[J].Journal of the China Society for Secientific and Technical Information,2021,40(6):656-665.

[24] VASWANI A,SHAZEER N,PARMAR N,et al.Attention is all you need[C]//Advances in Neural Information Processing Systems,2017:5998-6008.

[25] CAO J,GUO J,LI X,et al.Automatic rumor detection on microblogs:a survey[J].arXiv:1807.03505,2018.