基于注意力机制的多模态融合谣言检测方法

戚力鑫,万书振,唐 斌,徐义春
三峡大学 计算机与信息学院,湖北 宜昌 443000

摘 要:谣言会对社会生活造成不利影响,同时具有多种模态的网络谣言比纯文字谣言更容易误导用户和传播,这使得对多模态的谣言检测不可忽视。目前关于多模态谣言检测方法没有关注词与图片区域对象之间的特征融合,因此提出了一种基于注意力机制的多模态融合网络AMFNN应用于谣言检测,该方法在词-视觉对象层面进行高级信息交互,利用注意力机制捕捉与关键词语相关的视觉特征;提出了基于自注意力机制的自适应注意力机制Adapive-SA,通过增加辅助条件来约束内部的信息流动,使得模态内的关系建模更有目标性和多样性。在两个多模态谣言检测数据集上进行了对比实验,结果表明,与目前相关的多模态谣言检测方法相比,AMFNN能够合理地处理多模态信息,从而提高了谣言检测的准确性。

关键词:深度学习;注意力机制;多模态融合;谣言检测
文献标志码:A **中图分类号:**TP391 doi:10.3778/j.issn.1002-8331.2102-0229

Multimodal Fusion Rumor Detection Method Based on Attention Mechanism

QI Lixin, WAN Shuzhen, TANG Bin, XU Yichun
College of Computer and Information, Three Gorges University, Yichang, Hubei 443000, China

Abstract: Rumors have an adverse impact on social life, and online rumors with multiple modalities are more likely to mislead users and spread better than pure text rumors, which makes the detection of multimodal rumors not negligible. At present, the multimodal rumor detection method does not pay attention to the feature fusion between the word and the image area object. Therefore, this paper proposes an attention-based multimodal fusion neural network (AMFNN) for rumor detection. The method carries out high-level information interaction between word and visual object, using the attention mechanism to capture visual features related to the key feature of words. An adaptive attention mechanism (Adaptive-SA) based on the self-attention mechanism is proposed, which restricts the internal information flow by adding auxiliary conditions to make the relationship modeling within the modality is more targeted and diverse. The paper conducts comparative experiments on two datasets about multimodal rumor detection. Experimental results show that, compared with the current related multimodal rumor detection methods, AMFNN can reasonably process multimodal information, thereby improving the accuracy of rumor detection.

Key words: deep learning; attention mechanism; multimodal fusion; rumor detection

近年来,多媒体技术突飞猛进,新闻的形式逐渐向多媒体转变。根据*We Are Social*[1]的统计报告,2020年的互联网用户比2019年同期增长了7%以上,达到了45.7亿人。每天都有大量推文在Twitter、微博等网络社交平台上发布,导致虚假信息也很容易传播,对社会造成负面的影响。例如,在2016年美国总统大选前一个月,每个选民平均接触到1~3条政治假新闻[2]。因此,准确的识别谣言,防止其广泛传播势在必行。

目前,新闻或者推文的内容不再只是使用纯文本,文本-视觉结合的内容形式正变得流行。与纯文本相比,图像和视频具有很好的视觉效果,不仅可以传播信息,也能够抓住人们的注意力,提供了丰富的视觉信息[3]。谣言中的图片、视频正是利用这个特点,使用篡改或者完全伪造的夸张图片传播负面信息,所以,结合视觉特征进行谣言检测可能有助于对谣言的区分。

目前的谣言检测方法模式正从单纯基于文本内容

基金项目:NSFC-新疆联合基金重点项目(U1703261);水电工程智能视觉监测湖北省重点实验室(三峡大学)开放基金(2017SDSJ06)。
作者简介:戚力鑫(1995—),男,硕士研究生,CCF会员,主要研究方向为自然语言处理;万书振(1976—),男,博士,副教授,主要研究方向为演化算法、云计算、机器学习等,E-mail:wanshuzhen@163.com;唐斌(1998—),男,硕士研究生,主要研究方向为自然语言处理;徐义春(1970—),男,教授,CCF会员,主要研究方向为模式识别与人工智能。
收稿日期:2021-02-22 **修回日期:**2021-04-06 **文章编号:**1002-8331(2022)19-0209-09

转向多模式检测的模式转变。在关注文本内容、社交信息或者传播结构的研究方法中,部分工作采用以文本内容和用户信息为主要特征的基于特征的结构检测方法,而另一部分工作则考虑了传输时间、传输结构和语言特征等因素。在利用视觉特征联合文本特征进行的谣言检测方法中,Jin等[4]利用循环神经网络(recurrent neural network)[5]提取文本、社会上下文特征,并用预训练的VGG-16提取图片特征,二者拼接后放入分类器中进行检测;Wang等[6]添加了一个事件标识符来理解事件之间的共同特征,这有助于检测未预见的新事件;Khattar等[7]提出了一种多模态变分自编码器来发现模态之间的相关性。以上多模态谣言检测方法的融合部分只是利用了拼接操作将特征提取器得到的文本特征和图片特征融合在一起,没有考虑到两种模态之间的有效融合。

本文提出了一种基于注意力的多模态融合神经网络AMFNN(attention-based multimodal fusion neural network),以有效完成词语-图片对象之间的信息交互。它主要包括3个模块:多模态特征提取器、多模态特征融合模块和谣言检测器。对于多模态特征提取器的文本特征提取器,采用预先训练的词向量作为门控循环单元GRU(gate recurrent unit)[8]的输入来获取时间语义特征用于更新视觉特征,并采用预训练的BERT(bidirectional encoder representations from transformers)[9]来提取句子特征用于和更新后的视觉特征进行融合以得到深度多模态特征表示。对于视觉特征提取器,采用预先训练的Faster RCNN(faster regions with CNN features)[10]来捕捉视觉对象特征。多模态特征融合模块基于注意力机制(attention mechanism)[11]实现,包括基于交叉注意力机制(cross-attention)的模态间信息交互模块(inter-modality information interaction module)和基于自注意力机制(self-attention)的模态内信息增强模块(intra-modality information enhancement module),其中模态间信息交互模块应用于词-视觉对象之间的信息交互,模态内信息增强模块应用于模态内部的关系建模。另外,当使用朴素自注意力进行模态内部的关系建模时,模态内部元素之间的关联始终相同,缺少对外界因素的考虑,针对于此,在朴素自注意力上进行了改进,增加了对方模态的关键信息作为约束条件,使得模态内的关系建模受对方模态的辅助调整,我们将其定义为自适应自注意力Adaptive-SA(adaptive self-attention)。由特征提取器获得的词特征和视觉对象特征会放入多模态融合模块中,完成信息的传递和融合,最后句子特征和由词特征更新的视觉对象特征用于谣言检测器中进行谣言检测。

本文的主要贡献有如下:

(1)提出了一种基于注意力的多模态融合神经网络(AMFNN)用于谣言检测,该模型基于注意力机制来发现单词与视觉对象之间的相关性,并实现模态内部和模态之间的信息交互。

(2)提出了一种基于自适应自我注意的内部关系建模方法(Adaptive-SA),该方法将另一种模式的重要特征用作条件约束来动态捕获重要特征。

(3)在收集自Twitter和微博的多模态谣言检测数据集进行了大量实验来验证本文方法的有效性。实验结果表明,AMFNN在多模态谣言检测的效果上优于其他基线模型。

1 相关工作

在本章中,回顾了提出的AMFNN模型所涉及的相关工作,主要涉及谣言检测任务和关于多模态之间的特征融合。

1.1 谣言检测

谣言检测任务与许多任务类似,如垃圾邮件检测[12]、讽刺文章检测[13]。基于之前研究者的相关工作,将可以被官方证实的虚假信息定义为谣言。

早期的研究主要集中在手工设置的有效特征提取,包括信息本身的语言特征,交际过程中的上下文特征,以及训练用于谣言检测的分类器。Castillo等[14]通过统计如特殊字符、链接数量等文本信息来检测谣言;Qazvinian等[15]研究了推文的主题等特征,并利用贝叶斯网络作为分类器识别谣言;Kwon等[16]发现,虚假信息发布量具有更显著的周期波动,使用随机森林来拟合时间序列特征;Wu等[17]将传播树引入到谣言检测模型中,使用核化随机游走算法构造分类器。这些研究方法主要是基于人工提取特征进行建模,需要复杂的特征工程。

为了克服人工特征提取的诸多缺点,一些学者将深度学习应用于谣言检测研究中。基于谣言内容及传播结构的相关研究中,Ma等[18]引入了循环神经网络用于捕捉潜在的时空语义特征;Chen等[19]引入了一种递归神经网络来关注注意机制对有效时间的表征;Yu等[20]使用卷积神经网络对与事件相关的时间序列文本进行分割。对于谣言的传播结构,Ma等[21]提出了一种树结构神经网络模型来学习推文的表示;Zhang等[22]提出了一种深度扩散模型来捕捉推文的内容、作者和话题之间的联系;Liu等[23]将神经网络引入到谣言的传播结构,建立了时间序列分类器来捕捉传播路径中用户特征的变化。

现如今的社交媒体新闻的形式通常由文本和图片或者视频组成,已有研究证明,视觉内容可以提供有效的信息[24]。基于文本-图片的多模态谣言检测方法研究中,文献[4]提出了一种带有注意机制的递归神经网络att-RNN来融合多模态特征;文献[6]提出了端到端的谣言检测模型EANN(event adversarial neural network),增加了事件判别器,消除了对特定事件的依赖,提高了未知事件的泛化能力;文献[7]提出了一种多模态变分

自编码器MVAE(multimodal variational autoencoder),来发现各模态之间的相关性,从而获得更好的多模态共享表示用于谣言检测;文献[25]提出了模型BDANN(BERT-based domain adaptation neural network),通过添加领域分类器来识别陌生谣言,将不同事件的多模态特征映射到同一空间。

目前的多模态谣言检测方法仅在融合阶段进行拼接操作,导致模态信息冗余,丢失有效信息。

1.2　多模态特征融合

多模态特征融合指将多个不同类型的模态特征进行有效地处理,模态之间合理地完成信息互补并解决可能存在的冗余问题,以获得更加丰富的深度融合特征。Zadeh等[26]提出了一种基于矩阵的多模态张量融合网络(TFN),通过计算不同模态间的张量外积来获得模态相关性;Hou等[27]考虑了模态局部关系,建立了多项式张量池块,通过高阶矩阵整合了多模态特征;Xu等[28]利用注意力机制捕捉模态间的交互信息。上述方法验证了可以通过组合多种模态的特征来获得具有更丰富内容的多模态的表示。

为了克服关于谣言检测多模态融合研究的局限性,提出了一种基于注意力的多模态融合网络(AMFNN),该融合网络实现了词-视觉对象层次的信息交互,并使用文本特征中的关键字来指导视觉内的关系建模以获得有利于谣言检测的特征。内容丰富的句子特征和更新的视觉特征与自适应权重结合在以突出显示最重要的特征,这些特征被谣言检测器接收以进行谣言检测。

2　基于注意力机制的多模态融合网络

基于注意力机制的多模态融合网络AMFNN的工作流程如图1所示,输入由文本-图像对组成,文本提取为单词特征和句子特征,图片根据图片对象区域获得视觉对象特征。为了捕捉与文本内容相关的有效视觉特征,将单词特征与视觉对象特征结合起来,实现形式内和形式间的信息流动,以充分挖掘有效地视觉对象信息。AMFNN由以下三部分组成。

(1)特征提取器:从由文本和附加图像的组成的多模态输入中提取潜在特征。

(2)多模态融合模块:它在文字与视觉对象之间传递信息,实现模态间的信息交互。

(3)谣言检测器:它使用深度融合特征来确定一条博文是否是谣言。

2.1　特征提取器

AMFNN的单个输入由文本和图像组成,通过特征提取器可以获得关于文本和图像的潜在特征。特征提取器根据输入内容类别分为文本特征提取器和视觉特征提取器。

2.1.1　文本特征提取器

文本特征提取器的输入由许多长度不相等的句子组成,首先对所有句子进行分词,然后将每个句子转换成由 m 个单词组成的列表:

$$T=[w_1,w_2,\cdots,w_m] \tag{1}$$

使用预训练的GloVe[29]词向量作为GRU的输入对推文的词特征进行编码。为了捕获潜在的语义和上下文含义,采用了BERT来提取句子特征,BERT本质上是基于文献[30]中所述的原始实现的多层双向变压器编码器。

$$E=GRU(T;\theta_{\mathrm{GRU}}) \tag{2}$$

$$S=BERT(T;\theta_{\mathrm{pretrained}}) \tag{3}$$

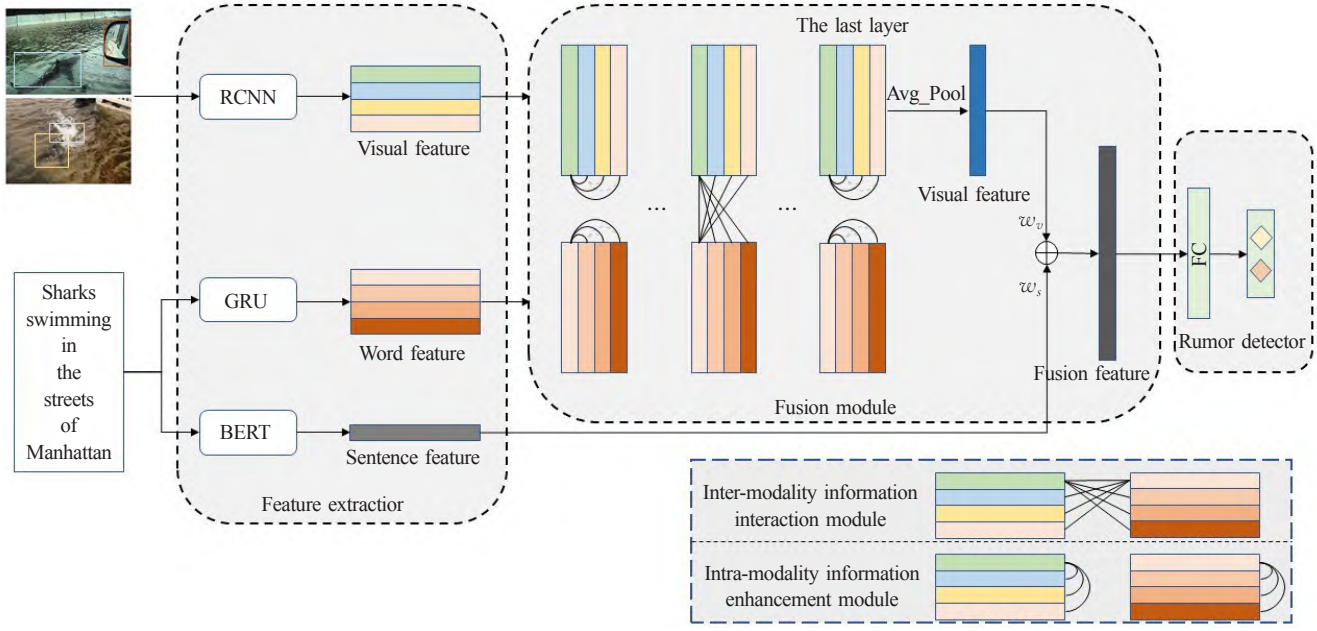


图1　AMFNN模型结构图

Fig.1　Structure diagram of AMFNN model

输入文本 T，输出关于词的特征矩阵 $E\in\mathbb{R}^{m\times d}$ 和句子向量 $S\in\mathbb{R}^{\dim}$，其中，第 i 个词表示为 $\boldsymbol{w}_i\in\mathbb{R}^d$，$d$ 为从GRU提取的词向量维度，dim为句向量的维度，θ_{GRU} 为GRU的训练权重，$\theta_{\text{pretrained}}$ 为BERT官方提供的预训练权重。

经过文本特征提取器得到的词特征矩阵 T 用于与视觉特征进行词-对象级别的信息交互，句子特征 S 用于与更新后的视觉进行特征融合。

2.1.2　视觉特征提取器

每篇推文的图片被处理成由一组图片对象组成的特征矩阵，这些对象由预先训练好的目标检测模型提取，每个对象代表图片中的一个人、动作、对象或两两间的交互。

输入图片 V，使用Faster RCNN[10]捕捉图片中的 n 个对象，然后通过全连接层(full connected layer)将维度调整成和词特征矩阵相同的数值，最后输出由图片对象特征组成的视觉矩阵 $O\in\mathbb{R}^{n\times d}$，每行向量相当于一个图片对象的潜在特征。

$$O=RCNN(V;\theta_{\text{pretrained}}) \tag{4}$$

其中，$\theta_{\text{pretrained}}$ 为Faster RCNN在数据集PASCAL VOC上的预训练权重。

2.2　多模态融合模块

考虑到每个词和视觉对象之间可能存在潜在联系，使用多头缩放点积注意力(multi-head scaled dot-product attention)[11]来实现词和视觉对象之间的信息流动。受Gao等[30]工作启发，构建了两个基本模块：模态间信息交互模块(inter-modality information interaction module)和模态内信息增强模块(intra-modality information enhancement module)。

2.2.1　模态间信息交互模块

模态间信息交互模块(inter-modality information interaction module)基于交叉注意力机制实现，用于每个词与视觉对象之间的信息传递，从而交叉捕捉词与视觉对象之间的关系。给定词级特征矩阵 E 和视觉特征矩阵 V，将每个词特征和视觉对象特征转换为query、key、value：

$$\begin{cases}E_Q=\sigma(E;\theta_Q),E_K=\sigma(E;\theta_K),E_V=\sigma(E;\theta_V)\\O_Q=\sigma(O;\theta_Q),O_K=\sigma(O;\theta_K),O_V=\sigma(O;\theta_V)\end{cases} \tag{5}$$

其中 $E_Q,E_K,E_V\in\mathbb{R}^{m\times\dim}$，$O_Q,O_K,O_V\in\mathbb{R}^{n\times\dim}$，dim表示变换后的特征维度，$\theta$ 为激活函数为 σ 的全连接层的权重。

当使用文本信息来更新视觉特征时，首先计算每个视觉对象和词之间的相似性，以获得从单词特征到每个视觉特征的注意力权重矩阵，注意权重矩阵的每一行代表视觉对象与所有单词的关联，并在归一化后使用它来更新视觉特征。采用多头注意力来执行多个不同的线性映射以获得多维注意力权重。

$$\begin{cases}\boldsymbol{atten}_i^{OE}=softmax\dfrac{\boldsymbol{O}_Q\boldsymbol{E}_K^{\mathrm{T}}}{\sqrt{d_i}},\ d_i=d/h\\\boldsymbol{Atten}^{OE}=[atten_1^{OE},atten_2^{OE},\cdots,atten_h^{OE}]\boldsymbol{W}_{mh}\\\boldsymbol{O}^{\text{update}}=\boldsymbol{Atten}^{OE}\times\boldsymbol{E}_V\end{cases} \tag{6}$$

其中，$\boldsymbol{atten}_i^{OE}\in\mathbb{R}^{1\times m}$ 表示第 i 个头的注意力权重，d_i 为第 i 个头的输出特征维度，$\boldsymbol{Atten}^{OE}\in\mathbb{R}^{h\times m}$，$\boldsymbol{W}_{mh}\in\mathbb{R}^{m\times m}$。

更新后的视觉特征利用残差与原始视觉特征拼接，并利用全连接层调整维度。

$$O'=\sigma_O([\boldsymbol{O}^{\text{update}},\boldsymbol{O}];\theta_O) \tag{7}$$

其中，σ_O 表示全连接层的激活函数，θ_O 为训练权重。

通过视觉特征引导文本特征更新的过程类似于以上通过文本特征引视觉特征更新的过程。

$$\begin{cases}\boldsymbol{atten}_i^{EO}=softmax\dfrac{\boldsymbol{E}_Q\boldsymbol{O}_K^{\mathrm{T}}}{\sqrt{d_i}},\ d_i=d/h\\\boldsymbol{Atten}^{EO}=[atten_1^{EO},atten_2^{EO},\cdots,atten_h^{EO}]\boldsymbol{W}_{mh}\\\boldsymbol{E}^{\text{update}}=\boldsymbol{Atten}^{EO}\times\boldsymbol{O}_V\\\boldsymbol{E}'=\sigma_E([\boldsymbol{E}^{\text{update}},\boldsymbol{E}];\theta_E)\end{cases} \tag{8}$$

2.2.2　模态内信息增强模块

模态内信息增加模块(intra-modality information enhancement module)基于自注意力机制实现，与模态间信息交互模块不同的是query、key、value均取自于自身模态。

$$\boldsymbol{atten}^{EE}=softmax\frac{\boldsymbol{E}_Q\boldsymbol{E}_K^{\mathrm{T}}}{\sqrt{d}} \tag{9}$$

$$\boldsymbol{atten}^{OO}=softmax\frac{\boldsymbol{O}_Q\boldsymbol{O}_K^{\mathrm{T}}}{\sqrt{d}} \tag{10}$$

模态间信息交互模块的设计是为了将视觉对象与词联系起来，例如图1的模型结构图中，视觉内容中的“鲨鱼”与文字中的“Sharks”相对应，这有助于模型理解图片和文字的内容。模态内信息增强模块是用于捕捉模态内的重点信息，但是如果使用上述朴素自注意力机制计算词与词、图片对象与图片对象之间的相似度，由其确定的词与词之间的关系总是相同的，图片对象之间的关系也总是相同的。认为模态内部的关系建模需要另外一种模态的辅助推理，对于视觉模态内部的关系建模，相同视觉对象之间的关系需要根据不同的词有不同的权重，例如，计算图1中的视觉对象“鲨鱼”与其他视觉对象的相关性，给予文字中的“swimming”和“in”来辅助匹配，则与之相关的视觉对象“水”和“浪花”可以具有很高的权重，达到提升匹配速度、多样化重要特征的效果。视觉特征辅助文本特征内部关系建模同理。

在此基础上，认为模态内信息增强模块在进行模态内关系建模的同时关注另一种模态关键特征之间的关系，可以进一步提高特征关系的建模效果，因此提出了自适应自注意力机制Adaptive-SA，如图2所示。以视觉模态内的关系建模为例，基本思想相似于门控机制，视

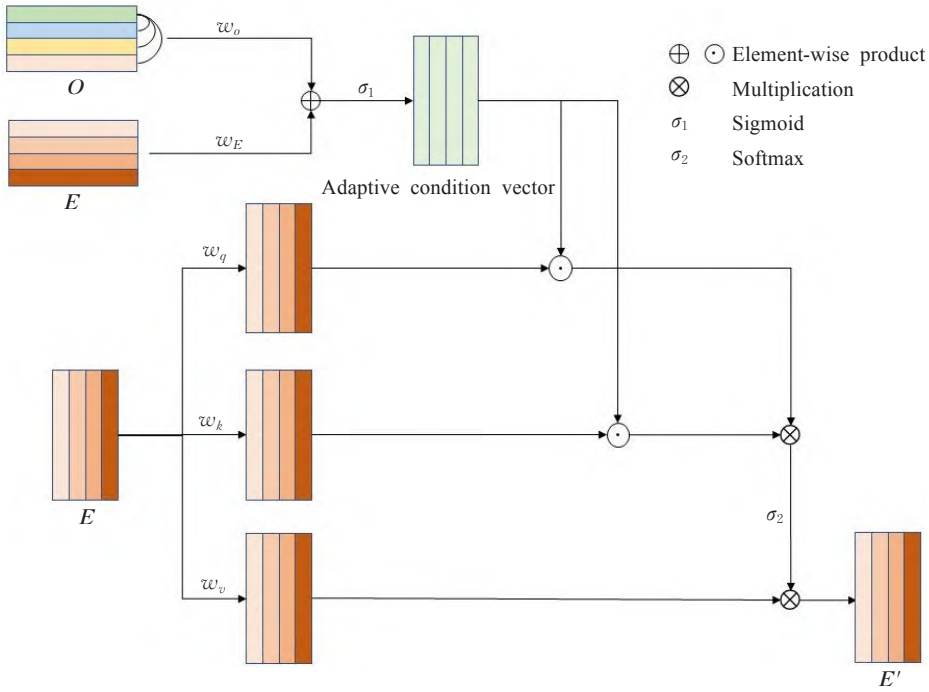


图2 Adaptive-SA结构图

Fig.2 Illustration of Adaptive SA

觉特征联合基于自注意力机制获取的有效词特征,首先探索视觉特征中与有效词特征相关的特征,然后在激活函数Sigmoid的作用下生成自适应条件向量:

$$G^E=\sigma_G([Self\_atten(\boldsymbol{E}),O];\theta_E) \tag{11}$$

视觉模态内部的相似度计算受自适应权重向量约束,与有效词特征相关的视觉对象间的关系建模程序被激活,以自适应地根据条件向量对不同视觉对象之间给予不同程度的关注:

$$O'_Q=(1+G^E)\odot O_Q \tag{12}$$

$$O'_K=(1+G^E)\odot O_K \tag{13}$$

后续的操作基于微调的 O'_Q , O'_K :

$$\begin{cases}\boldsymbol{atten}_i^{OO}=softmax\dfrac{O'_Q\left(O'_K\right)^{\mathrm{T}}}{\sqrt{d_i}},\ d_i=d/h\\ \boldsymbol{Atten}^{OO}=[atten_1^{OO},atten_2^{OO},\cdots,atten_h^{OO}]\boldsymbol{W}_{mh}\\ O^{\mathrm{update}}=\boldsymbol{Atten}^{OO}\times O_V\\ O'=\sigma_O([O^{\mathrm{update}}+O];\theta_O)\end{cases} \tag{14}$$

其中, σ_G 表示激活函数Sigmoid, θ_G 表示 $Self\_atten(\boldsymbol{E})$ 和 O 的自适应权重, $\boldsymbol{Atten}^{OO}$ 为多个 $\boldsymbol{atten}_i^{OO}$ 组成的关于视觉特征之间的注意力权重矩阵, O^{update} 为更新后的视觉特征, O' 为经过残差连接得到的视觉特征。

2.2.3 特征融合

在给定词语特征 E 和视觉对象特征 V 的情况下,模态间间信息流模块对文本-图像的跨通道关系进行建模,模态内信息流模块分别对两种模态之间的关系进行建模。两个模块可以按不同的顺序堆叠,将更新后的图像对象特征定义为 O' 。将更新后的视觉特征 O' 和句子特征 S 拼接在一起作为融合特征,考虑到并非所有的模态对分类都有相同的贡献。对视觉特征和句子特征自适应的训练注意力权重,以增强某类模态的表示,同时为了保证不丢失模态的特征表示,注意权重增加1。

$$\boldsymbol{\alpha}=\sigma_2(\sigma_1([O',\boldsymbol{S}];\theta_1);\theta_2) \tag{15}$$

$$F=(\boldsymbol{\alpha}^{\mathrm{T}}+1)[O',\boldsymbol{S}] \tag{16}$$

其中, θ_1 、 θ_2 为权重矩阵, σ_1 为激活函数Tanh, σ_2 为激活函数Softmax。另外,将特征的提取和融合过程定义为 B :

$$F=B([T,\boldsymbol{V}];\theta_B) \tag{17}$$

θ_B 表示特征提取器和多模态融合块的训练参数集, B 为映射函数, F 为融合后的多模态特征表示。

2.3 谣言检测器

谣言检测器输入融合的多模态特征表示来区分推文是否是谣言,它由两个全连接层组成,分别具有激活函数ReLU和Softmax。将谣言检测器定义为 D :

$$D(F;\theta_D) \tag{18}$$

其中 θ_D 为谣言检测器的参数,谣言检测器的输出为该推文是谣言的概率。

$$\hat{y}=D(F;\theta_D) \tag{19}$$

把假新闻标记为1,真实新闻标记为0。使用交叉熵来计算分类损失:

$$L_D(\theta_B,\theta_D)=-E_{(x,y)\sim(X,Y)}[y\log(\hat{y})+(1-y)\log(1-\hat{y})] \tag{20}$$

其中 X 代表一组由文本和附加图像组成的帖子, Y 代表相应的标签。通过最小化分类损失来优化提取、融合和检测时的参数 θ_B 、 θ_D :

$$(\theta_B,\theta_D)=\arg\min_{\theta_B,\theta_D}L_D \tag{21}$$

3 实验

在本章中,首先介绍多模态谣言检测的相关数据集,然后对有关模型进行了详细描述,最后通过实验结果对本文提出的AMFNN的模型效果进行了分析讨论。

3.1 数据集

为了验证本文提出的AMFNN的有效性,在来自Twitter和微博的两个多模态谣言数据集上进行了丰富的实验,表1统计了两个数据集的详细信息。

表1 多模态谣言数据集的内容统计

Table 1 Content statistics of multimodal rumor datasets

Label	Twitter	Weibo
rumor	7 898	4 749
non-rumor	6 026	4 779
image	514	9 528

3.1.1 Twitter数据集

用于多模态谣言检测的Twitter数据集由文献[31]提供。它收集了大量发布在Twitter上的推文,每条推文的内容都是由短文本和附加的图片/视频组成,并附有标签来表明是否为假新闻。数据集分为两个部分:开发集(9 000条谣言推文,6 000条真实新闻推文)和测试集(2 000条)。由于本文工作主要集中在文本和图像信息上,因此带有视频的推文将从数据集中删除。

3.1.2 微博数据集

数据集使用文献[4]中使用的多模态谣言检测数据集。数据集中,真实数据来源于中国权威新闻来源(如新华社),虚假数据采集时间为2012年5月至2016年1月,并经官方谣言检测系统验证。遵循文献[4]中数据预处理的方法,删除重复的低质量图像,然后按4∶1的比例划分为训练集和测试集。

3.2 实验模型

为了衡量所提出的AMFNN模型在多模态谣言检测中的性能,根据输入类型将比较模型分为两类:单模态谣言检测模型和多模态谣言检测模型。

3.2.1 单模态谣言检测模型

单模态谣言检测模型只使用单一的模态特征,包括纯文本和纯图像。

(1)Only-Text。使用文本内容用于谣言检测,包括Text-GRU、Text-BERT。Text-GRU使用预训练的词向量作为输入,最后一个时间步的特征作为文本特征输出到分类器;Text-BERT使用原始文本作为输入,输出对应的特征向量用于分类。

(2)Only-Image。与Only-Text模型类似,Only-Image模型只使用推文的附加图像作为输入,将图像送入预训练的VGG-19、Faster RCNN模型中,通过平均池化层获取视觉特征,然后送入分类器进行谣言检测。

3.2.2 多模态谣言检测模型

多模态谣言检测模型的输入为文本内容和附加的图片包括最先进的模型及变体和本文提出的模型。

(1)VQA[32]:视觉问答(visual question answering)任务是通过图像回答有关问题。为了采用VQA来检测假新闻,将文本和图像之间的逐元素乘法修改为特征级联,并将多类层修改为二分类层。

(2)NeuralTalk[33]:NeuralTalk旨在生成描述句子的自然句。遵循NeuralTalk的主网络结构,将RNN的每个时间步长输出平均作为联合特征,然后输入到分类器中。

(3)att-RNN[4]:att-RNN采用LSTM提取文本和社会语境的特征并通过预先训练的VGG-19模型提取视觉特征。然后使用注意力机制捕获视觉特征与文本/社会语境特征之间的关联,最后将二者合并进行分类。不使用社会语境特征的att-RNN定义为att-RNN−。

(4)EANN[6]:EANN中的文本特征通过TextCNN模型提取,图片特征使用预训练的VGG-19模型提取,图片特征和文本特征拼接后输入事件判别器和谣言分类器,事件判别器用于学习事件的不变表示。不带事件判别器的EANN模型被定义为EANN−。

(5)MVAE[7]:MVAE采用一种编码-解码的方法构造了一个多模态特征表达式。通过对多模态变分自编码器的训练,可以从学习到的共享表示中重构出两个模态,从而找到交叉模态之间的相关性。

(6)BDANN[25]:BDANN采用预训练的BERT模型提取文本特征,采用预训练VGG-19模型提取视觉特征。通过添加域分类器,将不同事件的多模态特征映射到同一空间,消除对特定事件的依赖。不带域分类器的BDANN记为BDANN−。

(7)AMFNN:该神经网络模型由3个模块组成:特征提取器、特征融合块和谣言检测器。基于注意力机制,AMFNN实现词语-视觉对象层面的信息交互和模态内的信息流动,从而获得内容更丰富的融合特征。无融合块的AMFNN表示为AMFNN-1,模态内使用朴素自我注意力的模型表示为AMFNN-2。

3.3 实验设置

实验环境:Linux×86_64,GTX1080TI,使用Python语言编写实验程序,在Tensorflow平台上搭建神经网络模型。

对于词的特征,采用GloVe[29]来训练词的表示,每个词的嵌入维数为100,并由隐藏维数为512的GRU编码为词的特征。维度为768的句子特征向量由BERT模型提取。对于视觉特征,采用预先训练的Faster RCNN来捕捉2 048维的视觉对象特征,并将图片对象的数量设置为与文本的长度相同,即每张图片提取100个视觉对象,然后使用尺寸为512的全连接层来嵌入视觉特征。多模态融合模块中,多头注意力的数量设置为8个,每个注意头的尺寸为64个。多模式混合融合模块按照模

态间信息交互模块-模态间信息增强的顺序堆叠两层。

模型的批处理大小(batch_size)设置为32,训练100次,并在10个误差范围内模型精度不再提升则提前停止训练并报告训练结果。学习速率初始化为10^{-3},当模型训练效果不再上升时,学习率自适应下降。利用Adam优化器来寻找神经网络的最优参数。

4 实验结果及分析

4.1 谣言的检测效果及分析

为了验证本文提出的模型AMFNN的谣言检测效果,将AMFNN与其他基线模型在Twitter和微博两个多模态谣言检测数据集上进行比较,采用分类问题中常用的评价指标[35]:精确率(Precision)、召回率(Recall)、*F*1分值(*F*1-Score)和准确率(Accuracy)。实验对比结果如表2、3所示。

表2 Twitter数据集上各模型实验结果

Table 2 Experimental rusults of each model on Twitter dataset

Method	Accuracy	Precision	Recall	*F*1
Text-GRU	0.536	0.540	0.510	0.520
Text-BERT	0.713	0.690	0.630	0.660
Image-VGG	0.596	0.695	0.518	0.593
Image-RCNN	0.590	0.720	0.470	0.570
VQA	0.631	0.765	0.509	0.611
NeuralTalk	0.610	0.728	0.504	0.595
att-RNN−	0.664	0.749	0.615	0.676
att-RNN	0.682	0.780	0.615	0.689
EANN−	0.648	0.810	0.498	0.617
EANN	0.719	0.642	0.474	0.545
MVAE	0.745	0.801	**0.719**	**0.758**
BDANN−	0.821	0.790	0.610	0.690
BDANN	0.830	0.810	0.630	0.710
AMFNN-1	0.823	0.790	0.640	0.710
AMFNN-2	0.832	0.800	0.620	0.700
AMFNN	**0.841**	**0.820**	0.660	0.730

观察表2、3,在谣言检测的文本模型中,Text-BERT可以提取有效的文本特征,获得更好的准确率和*F*1分值,谣言检测结果接近EANN和MVAE。BDANN和我们提出的AMFNN都受益于此,取得了较大效果提升。Text-GRU由于难以在预处理阶段对文本进行有效的分割,因此效果不明显。

Twitter数据集中的图片数量只有500左右,存在一张图片对应多条推文的现象,导致Twitter数据集中的图片特征不足,而微博数据集中的每条推文都由文本和对应的图片组成。这种差异可能会影响文本和图像对象的信息交互,这点可以观察BDANN-和AMFNN-1的实验统计。Twitter数据集上,与使用VGG-19提取全局图像特征的BDANN-相比,提出的AMFNN-1的准确率提升不明显,但是在微博数据集上有较突出的提升幅度:

表3 微博数据集上各模型实验结果

Table 3 Experimental rusults of each model on Weibo dataset

Method	Accuracy	Precision	Recall	*F*1
Text-GRU	0.730	0.780	0.630	0.700
Text-BERT	0.807	0.810	0.830	0.820
Image-VGG	0.608	0.610	0.605	0.607
Image-RCNN	0.689	0.690	0.700	0.690
VQA	0.736	0.797	0.634	0.706
NeuralTalk	0.726	0.794	0.613	0.692
att-RNN−	0.772	0.854	0.656	0.742
att-RNN	0.788	**0.862**	0.686	0.764
EANN−	0.794	0.790	0.820	0.800
EANN	0.816	0.820	0.820	0.820
MVAE	0.824	0.854	0.769	0.809
BDANN−	0.814	0.800	0.860	0.830
BDANN	0.842	0.830	**0.870**	0.850
AMFNN-1	0.827	0.830	0.850	0.840
AMFNN-2	0.849	0.840	0.860	0.840
AMFNN	**0.862**	0.860	0.850	**0.850**

AMFNN-1的准确率和*F*1分值分别提升了3.2个百分点、1.2个百分点。说明与使用图像全局特征的BDANN-相比,AMFNN-1通过捕捉图像对象和单词之间的潜在关系,有效地提高了模型的检测效果。

可以观察到,两个数据集上的多模态谣言检测方法的表现都优于单模态方法,证明了图片特征包含文本特征中不存在的但是跟谣言相关的有效特征,二者的结合实现了有效信息的互补,得到了更加丰富的深度特征,提高了检测的效果。

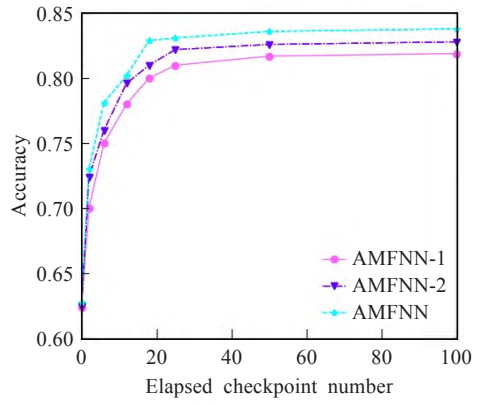
观察表2、3中AMFNN-2与AMFNN的实验结果对比,其中AMFNN-2和AMFNN分别基于朴素的自注意力机制、Adaptive-SA实现模态内部的信息增强,与朴素的自注意力机制相比,Adaptive-SA额外提供了约束向量,探索与其他模态关键特征相关的模态内有效对象,使得模态内的对象关系建模过程有明显的目标,相关的有效对象组合更加多样。AMFNN-2在Twitter和微博数据集上的表现均没有AMFNN的检测效果优秀,在Twitter数据集上AMFNN的精度和*F*1分值提升了1.1个百分点、4.3个百分点,AMFNN在微博数据集上同样取得了领先(1.5个百分点、1.2个百分点),说明在自注意力机制中,使用另外模态的关键信息作为自模态内的约束条件,以辅助推理自模态内的对象关系进行关系建模这种方法是有效的。

目前关于多模态谣言检测的研究方法中,关于模态特征的组合方式基本是在谣言分类器之前拼接,没有考虑到模态之间的交互。MVAE[7]例外,它通过训练多模式变分自动编码器来学习文本和图像之间的共享表示,并且实验结果表明,挖掘模态之间的相关性可以得到更为突出的结果。AMFNN基于注意力机制实现了模态间的有效交互以及模态内自适应的关键信息提取,充分

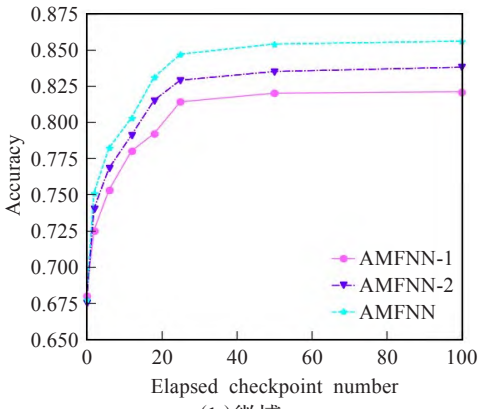
挖掘了与文本的关键内容有关的视觉对象特征,并与经过有效提取的句向量进行了自适应权重分配,进一步增强有效模态的表示。AMFNN在两个数据集的实验效果均优于MVAE,其中在Twitter数据集上模型精度大幅提升((0.841−0.745)/0.745=12.9%),在微博数据集上模型精度和$F1$分值分别了提高了4.5个百分点、5.1个百分点,证明了AMFNN充分挖掘了词语与图像对象之间的潜在联系,有效捕捉到了视觉对象中与关键词语相关的特征,视觉特征向量与句向量进行了有效融合,提升了谣言检测的效果。

4.2　谣言检测过程的可视化及分析

在4.1节中通过模型的谣言检测效果证明了所提出的Adaptive-SA可以使得模态内的对象关系建模过程有明显的目标,相关的有效对象组合更加多样化。本节中通过可视化模型训练期间的谣言检测过程,使得Adaptive-SA的效果能够更加形象化的表示。谣言检测过程的可视化结果如图3所示,在Twitter和微博两个关于多模态谣言检测的数据集上进行实验,分别记录为图3(a)、(b)。AMFNN-1作为未实现多模态融合模型的代表,根据谣言检测对应模型检查点的效果来衡量检测效果。给定检查点,使用对应的模型检查点权重在测试集上所获得的谣言检测准确率来评估模型性能。



(a)Twitter



(b)微博

图3　模型的谣言检测性能根据检查点变化的可视化过程

Fig.3　Visualization process of rumor detection performance of model according to checkpoint changes

图3显示,尽管AMFNN-1、AMFNN-2最终也会达到最佳检测效果,但是在相同的检查点上,AMFNN达到的检测效果最佳,性能的提升最大。AMFNN在Twitter数据集上大约在检查点为16时可达到AMFNN-2的最佳性能,在微博数据集上也可以在检查点为18左右时达到AMFNN-2的最佳性能。证明了Adaptive-SA在条件向量的辅助下可以捕捉模态内有效的元素对,赋予较大的目标权重予以快速激活,同时根据对方模态的关键信息,使得模态内的元素对组合更加丰富,有利于获取更多有效特征。

5　结论

在本文中,提出一个基于注意力的多模态混融合网络AMFNN应用于谣言检测,从词-图片对象的角度处理模态之间的信息交互,并提出了Adaptive-SA,对自注意力机制添加了辅助条件,以便内部信息的流动是约束于其他模式的关键信息。更新后的视觉特征和句子特征被赋予自适应权值,以增强重要特征的贡献。在两个多模态数据集上进行了广泛实验,实验结果验证了词-图片对象交互角度的可行性,模态内的关系建模受其他模态约束的有效性,以及探索多模态特征关联的必要性。

参考文献:

[1] KEMP S.Digital around the world in April 2020[EB/OL].(2020-03-30).https://wearesocial.com/digital-2020.

[2] ALLCOTT H,GENTZKOW M.Social media and fake news in the 2016 election[J].Journal of Economic Perspectives,2017,31(2):211-236.

[3] WU K,YANG S,ZHU K Q.False rumors detection on sina weibo by propagation structures[C]//Proceedings of 2015 IEEE 31st International Conference on Data Engineering,April 13-17,2015:651-662.

[4] 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,October 23-27,2017:795-816.

[5] ZAREMBA W,SUTSKEVER I,VINVALS O.Recurrent neural network regularization[J].arXiv:1409.2329,2014.

[6] 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,August 19-23,2018:849-857.

[7] KHATTAR D,GOUD J S,GUPTA M,et al.MVAE:Multimodal variational autoencoder for fake news detection[C]//Proceedings of World Wide Web Conference,May 13-17,2019:2915-2921.

[8] CHUNG J,GULCEHRE C,CHO K,et al.Empirical evaluation of gated recurrent neural networks on sequence

modeling[J].arXiv:1412.3555,2014.
[9] DEVLIN J,CHANG M,LEE K,et al.BERT:Pre-training of deep bidirectional transformers for language understanding[J].arXiv:1810.04805,2018.
[10] REN S,HE K,GIRSHICK R,et al.Faster R-CNN:Towards real-time object detection with region proposal networks[J].arXiv:1506.01497,2015.
[11] VASWANI A,SHAZEER N,PARMAR N,et al.Attention is all you need[J].arXiv:1706.03762,2017.
[12] ZHU Y,WANG X,ZHONG E,et al.Discovering spammers in social networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence,July 22,2012:171-177.
[13] RUBIN V L,CONROY N,CHEN Y,et al.Fake news or truth? using satirical cues to detect potentially misleading news[C]//Proceedings of the 2nd Workshop on Computational Approaches to Deception Detection,June 17,2016:7-17.
[14] CASTILLO C,MENDOZA M,POBLETE B.Information credibility on twitter[C]//Proceedings of the 20th International Conference on World Wide Web,March 28-April 1,2011:675-684.
[15] QAZVINIAN V,ROSENGREN E,RADEV D,et al.Rumor has it:Identifying misinformation in microblogs[C]//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing,July 27-31,2011:1589-1599.
[16] KWON S,CHA M,JUNG K,et al.Prominent features of rumor propagation in online social media[C]//Proceedings of the 2013 IEEE 13th International Conference on Data Mining,December 7-13,2013:1103-1108.
[17] WU K,YANG S,ZHU K Q.False rumors detection on sina Weibo by propagation structures[C]//Proceedings of the 2015 IEEE 31st International Conference on Data Engineering,April 13-17,2015:651-662.
[18] MA J,GAO W,MITRA P,et al.Detecting rumors from microblogs with recurrent neural networks[C]//Proceedings of the 25th International Joint Conference on Artificial Intelligence(IJCAI-16),July 9-15,2016:3818-3824.
[19] CHEN T,LI X,YIN H,et al.Call attention to rumors:Deep attention based recurrent neural networks for early rumor detection[C]//Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining,June 3-6,2018:40-52.
[20] YU F,LIU Q,WU S,et al.A convolutional approach for misinformation identification[C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence,August 19-25,2017:3901-3907.
[21] MA J,GAO W,WONG K.Rumor detection on twitter with tree-structured recursive neural networks[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics,July 15-20,2018:1980-1989.
[22] ZHANG J,DONG B,PHILIP S Y.Deep diffusive neural network based fake news detection from heterogeneous social networks[C]//Proceedings of the 2019 IEEE International Conference on Big Data(Big Data),December 9-12,2019:1259-1266.
[23] LIU Y,WU Y F.Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence,February 2-7,2018:354-361.
[24] JIN Z,CAO J,ZHANG Y,et al.Novel visual and statistical image features for microblogs news verification[J].IEEE Transactions on Multimedia,2016,19(3):598-608.
[25] ZHANG T,WANG D,CHEN H,et al.BDANN:BERT-based domain adaptation neural network for multi-modal fake news detection[C]//Proceedings of the 2020 International Joint Conference on Neural Networks(IJCNN),July 19-24,2020:1-8.
[26] ZADEH A,CHEN M,PORIA S,et al.Tensor fusion network for multimodal sentiment analysis[J].arXiv:1707.07250,2017.
[27] HOU M,TANG J,ZHANG J,et al.Deep multimodal multilinear fusion with high-order polynomial pooling[C]//Advances in Neural Information Processing Systems,2019:12136-12145.
[28] XU N,MAO W,CHEN G.Multi-interactive memory network for aspect based multimodal sentiment analysis[C]//Proceedings of the AAAI Conference on Artificial Intelligence,Jan 27-Feb 1,2019:371-378.
[29] PENNINGTON J,SOCHER R,MANNING C D.Glove:Global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP),October 25-29,2014:1532-1543.
[30] GAO P,JIANG Z,YOU H,et al.Dynamic fusion with intra-and inter-modality attention flow for visual question answering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,June 15-20,2019:6639-6648.
[31] BOIDIDOU C,ANDREADOU K,PAPADOPOULOS S,et al.Verifying multimedia use at media Eval 2015[C]//Proceedings of the MediaEval 2015 Multimedia Benchmark Workshop,September 14-15,2015.
[32] ANTOL S,AGRAWAL A,LU J,et al.VQA:Visual question answering[C]//Proceedings of the IEEE International Conference on Computer Vision,December 7-13,2015:2425-2433.
[33] VINYALS O,TOSHEY A,BENGIO S,et al.Show and tell:A neural image caption generator[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,June 7-12,2015:3156-3164.
[34] LIN T Y,GOYAL P,GIRSHICK R,et al.Focal loss for dense object detection[C]//Proceedings of the IEEE International Conference on Computer Vision,October 22-29,2017:2980-2988.
[35] GODBOLE S,SARAWAGI S.Discriminative methods for multi-labeled classification[C]//Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining,May 26-28,2004:22-30.