基于多级融合的多模态谣言检测模型

王 壮，隋 杰+

（中国科学院大学 工程科学学院，北京 100049）

摘 要：针对当前多模态谣言检测模型存在的模态间信息融合不足和过于依赖各模态信息完整度的问题，提出一种基于多级融合的多模态谣言检测模型。分别利用 TextCNN 和 Resnet18 网络对文本和图片编码并进行特征级融合，对纯文本模型、纯图片模型和特征级融合模型进行决策级融合并对决策级融合进行改进。多级融合框架加深各模态间的信息融合程度，改进后的决策级融合有效缓解了传统模型对各模态信息完整度要求过高的问题。实验结果表明，该模型在微博数据集上的 F1 值和准确率均高于传统的多模态谣言检测模型，进一步提升了谣言检测效果。

关键词：多模态；谣言检测；神经网络；模态融合；深度残差网络

中图法分类号：TP391 **文献标识号**：A **文章编号**：1000-7024（2022）06-1756-06

doi：10.16208/j.issn1000-7024.2022.06.033

Multimodal rumor detection model based on multilevel fusion

WANG Zhuang, SUI Jie+

(School of Engineering Sciences, University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Aiming at the problems of insufficient information fusion among modes and over-dependence on the integrity of information of each mode in current multimodal rumor detection models, a multimodal rumor detection model based on multilevel fusion was proposed. TextCNN and Resnet18 networks were used to encode text and image respectively, and feature level fusion was carried out. Decision level fusion was carried out on the pure text model, pure image model and feature level fusion model, and the decision level fusion was improved. The multi-level fusion framework deepened the degree of information fusion among various modes, and the improved decision-level fusion effectively alleviated the problem that the traditional model requires too much information integrity of each mode. Experimental results show that the F1 value and accuracy of the proposed model are higher than that of the traditional multimodal rumor detection model on the microblog dataset, which further improves the rumor detection effect.

Key words: multimodal; rumor detection; neural network; modal fusion; Resnet

0 引 言

网络谣言检测本质上是一个二分类问题，即实验者通过使用特定模型对一条或者一系列的帖子进行判别，将其分类为谣言或者非谣言。

近年来，随着神经网络技术不断在自然语言处理领域取得进展，越来越多的学者将其应用到谣言检测领域[1]。Ma 等[2]将循环神经网络（recurrent neural network，RNN）运用到网络谣言检测中，相比于传统的机器学习方法，大大提高了谣言检测的效率。刘政等[3]提出了一种改进的卷积神经网络（convolutional neural networks，CNN）模型用于微博谣言检测，该模型结构简单，易于实现。Chen 等[4]将注意力机制与 RNN 模型相结合用于谣言检测，在一定程度上解决了文本信息过度冗余、远程间信息联系薄弱的问题。然而，上述模型只关注于谣言的文本信息而忽略了其附带的图片和社会信息，限制了模型的检测效果。针对于此，Jin 等[5]利用神经网络分别对待测事件中的图片、文本和社会特征等信息进行编码并使用注意力机制将其结合，提升了图片信息的利用价值。Dhruv 等[6]通过自动编码器对融合后的多模态向量进行约束，从而更好地学习多模态

收稿日期：2020-12-25；**修订日期**：2021-04-20

基金项目：国家重点研发计划基金项目（2017YFB0803001）；国家自然科学基金面上基金项目（61572459）

作者简介：王壮（1995-），男，安徽宿州人，硕士研究生，研究方向为自然语言处理；+通讯作者：隋杰（1976-），女，山东日照人，博士，副教授，研究方向为社交网络分析、应急管理。E-mail：suijie@ucas.ac.cn

之间的联合表示。刘金硕等[7]则通过提取图像中隐藏的文字来提高模型的检测效果。

目前，多模态谣言检测模型已成为谣言检测领域的一大发展趋势，但现有的该类模型仍存在着各模态间信息融合不足和模型泛化能力较差等问题，这也是本模型着重解决的。

1　相关工作

现有的谣言检测模型大多数只关注于谣言的传播途径或者文本内容，而忽略了事件相关的图片信息。有研究表明[8]，带有图片内容的新闻转发次数是纯文本新闻的 11 倍以上，其具有更强的迷惑性和传播性。目前，仅有少数工作关注到了新闻中的图片信息，但这些多模态谣言检测模型普遍只是将图像信息与文本信息进行简单的特征级融合后进行分类，而实际上各模态间的语义信息在特征空间是异构的，这可能会导致以下两个问题：①多模态之间的信息融合不够充分；②模型过于依赖各模态间的信息完整度（可能有的事件只存在文本信息，而有的事件只存在图片信息）。

针对上述问题，本文提出了将特征级融合与改进后的决策级融合相结合的多模态谣言检测模型 MFCD。本模型通过多级融合框架对视觉特征和文本内容之间的区别性特征和相关性进行学习，在保留各模态原始信息的基础上进一步提升了各模态间的信息融合程度。同时，根据不同模态信息的实际缺失情况采取不同的决策级融合策略，在一定程度上解决了现有的多模态谣言检测模型过于依赖各模态间信息完整度的问题。

2　MFCD 模型

MFCD 模型总体框架如图 1 所示，主要由纯文本模型 Textual、纯图片模型 Visual 和深度特征级融合模型（feature-level fusion model，FFM）等 3 个部分组成。首先，分别利用文本-卷积神经网络（text convolutional neural network，TextCNN）和深度残差网络（residual neural network，Resnet）[9]对文本内容和图片内容进行编码，构建 Textual 模型和 Visual 模型；然后将两者的语义映射进行特征级的融合，得到深度特征级融合模型 FFM，其可以挖掘不同模态信息间的非线性关系，在剔除多模态间冗余信息的同时学习互补信息；最后，将 3 个模型的各自决策输入改进后的决策级融合层得到最终的决策结果。

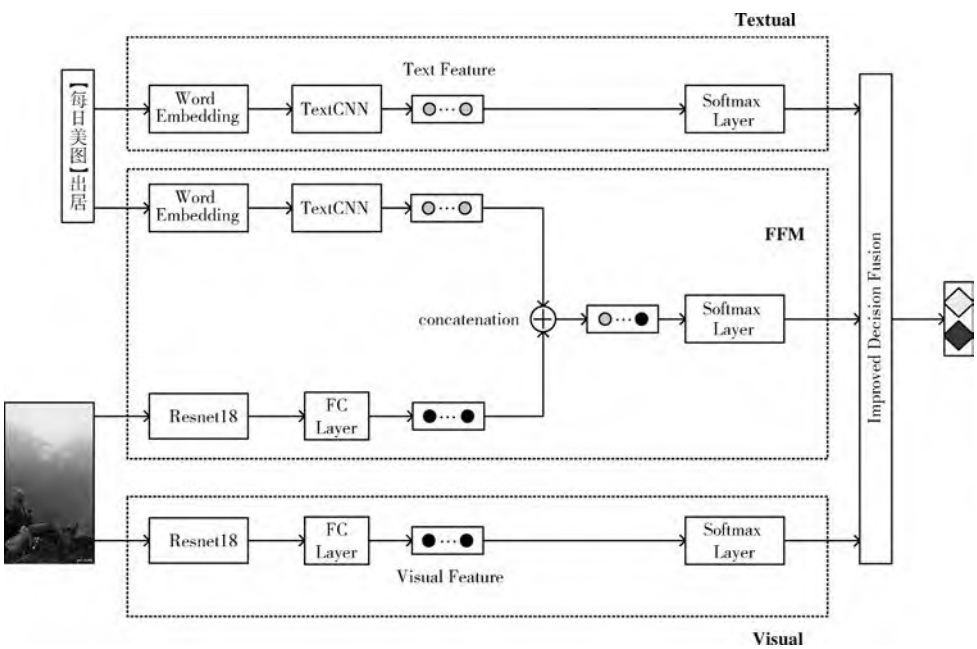

图 1　MFCD 模型框架

2.1　文字特征提取器

相比于传统的 RNN 模型，TextCNN 模型能够更好地扑捉到文本中的局部语义关系，已在短文本分类领域取得一定的成功[10]。其工作原理为利用各种形状的卷积核分别提取文本中不同粒度的特征并加以拼接，进而对文本进行分类。模型主体结构如图 2 所示，主要由预处理层、卷积层、池化层、融合层和输出层等 5 个部分组成。

（1）预处理层

使用预处理模型对分词后的中文文本进行编码，得到代表该文本的矩阵。其中，n 代表该文本分词后的词语数量，d 代表每个词语的向量表示维度。

（2）卷积层

卷积核 c 是一个形状为 $d*h$ 的矩阵。其中，h 为该卷积核的长度，d 为该卷积核的宽度，该宽度与词语的向量

维度保持一致。单位向量 A 中的第 i 个节点 $a(i)$ 可表示为

$$a(i)=\mathrm{f}(c\cdot W_{i:i+h-1}+b_i) \tag{1}$$

其中，$W_{i:i+h-1}$ 表示从第 i 个词语向量到第 $i+h-1$ 个词语向量的连续序列；·表示两个矩阵的点乘之积；b_i 表示该节点对应的偏置向量；f 表示模型中使用的激活函数。则对应的单位向量 A 可表示为：$A=[a_1,a_2,\cdots\cdots,a_{n-h+1}]$。

(3) 池化层

池化操作是从每个单位向量 A 中选取出最能代表该单位向量的某个节点。一般采用最大值池化方法，即选取其中值最大的节点代表该单位向量 A。

(4) 融合层

将各个池化层得到的特征进行拼接即得到对应的融合层。

(5) 输出层

一般使用 Softmax 函数对融合层进行处理得到输出层，输出结果即为各个类别所对应的概率大小。

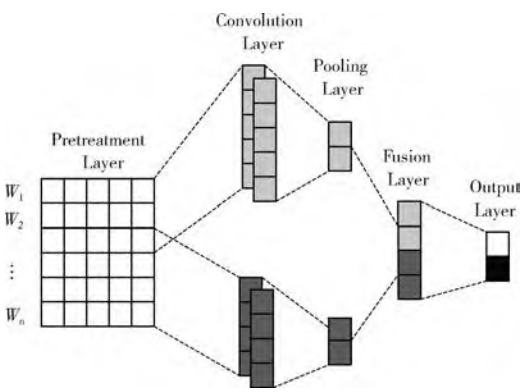

图 2 TextCNN 模型结构

2.2 图像特征提取器

现有的大多数多模态谣言检测模型[5-7]使用深度卷积神经网络模型 VGG19（visual geometry group，VGG）对图片进行特征提取，在提升了图片信息利用率的同时也造成了模型参数过多、易过拟合等问题。针对这个问题，本文提出采用基于迁移学习的深度残差网络 Resnet18 模型代替 VGG19。相比之下，Resnet18 模型的参数量更小，训练速度更快，准确率更高。

Resnet 模型最初由 Kaiming He 等提出，被广泛应用于图像处理和计算机视觉领域。其主要思想是采用了多级的残差模块进行连接，有效地缓解了传统深度卷积神经网络模型因层数过多而导致的反向传播梯度消失和模型性能退化等问题。

每个残差单元由一个残差学习分支和一个恒等映射分支组成，具体结构如图 3 所示。其中，x 表示输入，$F(x)$ 表示残差学习分支的结果，relu 表示激活函数，则残差单元的输出可表示为 $H(x)=F(x)+x$。当残差学习分支不工作时，$H(x)=x$。残差分支中的两个 $1*1$ 层作用分别为降维和升维，以保证 $F(x)$ 的维度与 x 的维度保持一致，进而进行后续操作。

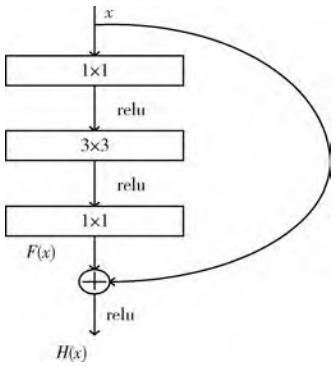

图 3 Resnet 单元结构

2.3 多级融合

根据目前已有的研究，多模态信息间的融合方式大致可分为数据级融合[11]、特征级融合[12,13]和决策级融合[14]。数据级融合是指直接将多个数据源集成到单个特征向量中再进行后续处理。该种融合方式的优点在于在一定程度上保持了数据的完整性，避免了数据的丢失和污染。但是其缺点也很明显，由于各种模态间的信息表现形式差异较大，此种融合方式很难利用各种模态信息间的互补性，甚至会造成很大的信息冗余。相比数据级融合，特征级融合和决策级融合的应用范围更加广泛，应用方式更加灵活。

近年来，将多种融合模式相结合使用的多模态模型已在多个领域取得进展[15,16]，受此启发，本文提出了结合特征级融合与改进后的决策融合的多模态谣言检测模型，在保留了各模态信息独立性的同时使其充分互补，其融合原理如图 4 所示。

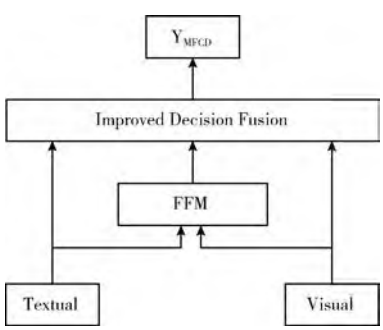

图 4 多级融合框架

2.3.1 特征级融合

特征级融合又称中期融合，是指分别使用不同的特征提取器对各个模态信息提取后再进行拼接、按位加或者按位乘的操作过程。本模型采用直接拼接的特征级融合方案，具体过程如式 (2) 所示

$$F_i = T_i \oplus V_i \tag{2}$$

式中：T_i 表示第 i 个样本中文本信息的特征映射，V_i 表示该样本中图片信息的特征映射，$\oplus$表示连接操作，F_i 则表示该样本的图文联合信息映射。

相比于数据级融合方式，特征级融合对不同模态信息采用不同的特征提取器，更加能挖掘各模态信息的潜在语义特征。但是单独的特征级融合方式对各模态的信息完整度要求很高，如果某个模态的信息缺失，只能使用对应模态数据的平均值或者其它数据进行填充，将会对模型的决策结果造成很大的不利影响。

2.3.2　改进后的决策级融合

决策级融合又称为后期融合，其首先分别提取各个模态的特征，然后输入对应的模型中并得到各自的分类结果，最后将各个模态的分类结果进行整合计算，以得到最终的分类结果。

决策级融合可以较好处理不同模态间的数据异步性问题，其融合规模也可以随着模态个数的增加而进行简单的扩展，对于融合不同性质特征的数据分类结果是十分有效的。但是当出现各个模态信息不完整的情况时，传统的决策级融合方案不能避免其带来的影响。针对本模型而言，当某一模态信息缺失时，不仅会影响到该模态对应模型的决策结果，还会对特征级融合模型 FFM 的结果产生影响，从而使总决策结果产生巨大偏差，导致分类错误。

针对于此，本模型采用一种改进后的后期融合方案，对于不同的模态信息缺失情况采用不同的决策级策略，从而避免了因某个模态信息缺失而影响到最终模型的分类结果。具体来说，当某一模态信息缺失时，则使用另一模态的分类结果作为最终分类结果；当各个模态信息完整时，使用三者与对应的自适用权重之积的和作为最终分类结果。具体见式（3）

$$Y_{MFCD} = \begin{cases} Y_{Textual} & input:Text \\ Y_{Visual} & input:Visual \\ \alpha \cdot Y_{Textual} + \beta \cdot Y_{Visual} + \gamma \cdot Y_{FFM} & input:Text\ and\ Visual \end{cases} \tag{3}$$

式中：$Y_{Textual}$、Y_{Visual}、Y_{FFM} 和 Y_{MFCD} 分别为各自模型对应的决策结果；α、β 和 γ 分别为各个模型的所占权重，且满足 $\{\alpha,\beta,\gamma \mid \alpha,\beta,\gamma \in [0,1], \alpha+\beta+\gamma=1\}$。

3　实验结果及分析

为验证本模型的可行性及检测效果，在微博数据集上进行了实验。

3.1　实验数据集

为公平比较本模型与各基线模型的检测性能，本文采用目前多模态谣言检测领域较为常用的微博数据集进行实验，训练集与测试集划分比例为 4∶1。该数据集由 Jin 等在文献［5］中使用，其中谣言事件来自于微博官方辟谣网站，真实事件来自于权威媒体新华社验证后的推文。该数据集的具体情况见表 1。

表 1　微博数据集

	数量	图片项缺失样本数	文本项缺失样本数
谣言事件	4748	1146	119
非谣言事件	4779	469	2
总事件	9527	1615	121

从表 1 中可以看出，样本中模态信息不完整的情况是切实存在的，这也从侧面反映出本模型所采用的改进后的决策级融合方案是具有一定现实意义的。

3.2　模型参数设置

本模型采用小批量随机梯度下降方法训练数据，每个批次的样本数量为 32，初始学习率设置为 0.001，共训练 60 次循环。采用交叉熵损失函数和 Adam 优化器进行反向传播优化。同时，为防止模型过拟合，使用 Dropout 和 L2 正则化对模型参数进行约束。

3.3　基线模型

为公平比较各模型性能，以下模型均在上述数据集进行实验，且训练集与测试集的划分比例相同。

（1）Textual 模型

该模型仅利用样本中的文本信息进行实验。首先使用结巴分词将中文文本进行分词，然后使用 Word2vec 技术将分词后的文本进行编码，编码后的单词维度为 32。对于文本长度不足或者文本缺失的样本，使用 0 元素进行填充。TextCNN 模型共有高度分别为 1、2、3、4 等 4 种形状不同的卷积核，每种卷积核的个数为 8，故输出向量维度为 32。将输出向量输入分类器，即得到该样本的最终分类结果。

（2）Visual 模型

该模型仅利用样本中的图片信息进行实验，对于图片信息缺失的样本用 0 进行像素级的填充，即利用一张纯黑图片代替该样本中的图片信息。将图片编码后输入 Resnet18 模型，后接一个维度为 32 全连接层，最后输入分类器得到样本分类结果。为增强模型泛化能力和减少训练时间，Resnet18 网络采取迁移学习的方式，选用已在大型数据集 Image1000 上训练完毕的模型参数且不参与反向传播，仅对后接线性层进行微调。

（3）FFM 模型

将 Textual 模型和 Visual 模型提取的特征向量进行特征级融合后输入分类器进行分类。

（4）DFM 模型

将 Textual 模型和 Visual 模型分别得到的分类结果进行决策级的融合。

(5) att-RNN 模型[5]

该模型利用注意力机制将文本、图片及社会特征等信息进行融合后输入分类器进行判断。为公平比较，本文采用其删除了社会特征后的模型，其余参数与文献[5]中所述一致。

(6) MVAE 模型[6]

该模型利用 VAE (variational autoencoder, VAE) 模块对多模态特征融合后的向量进行约束，然后利用该特征向量进行分类。

(7) MRSD 模型[7]

该模型首先将图片中的文本提取出来，然后将其与样本中的文本内容进行连接，最后将图片与连接后的文本进行特征级的融合并分类。

(8) MFCD-模型

使用一般的决策级融合代替改进后的决策级融合，其余部分与 MFCD 模型一致。

3.4 模型对比结果及分析

本文采用 F1 值、准确率 (Accuracy, A)、精确率 (Precision, P) 和召回率 (Recall, R) 等 4 个常用指标对各模型进行评价。各模型结果见表 2。

表 2 实验结果

模型	F1 值	Accuracy	Precision	Recall
Textual	0.775	0.779	0.794	0.777
Visual	0.714	0.715	0.715	0.714
FFM	0.808	0.811	0.828	0.809
DFM	0.811	0.810	0.822	0.810
att-RNN	0.769	0.772	0.786	0.775
MVAE	0.823	0.824	0.827	0.823
MSRD	0.779	0.794	**0.854**	0.716
MFCD-	0.824	0.823	0.838	0.823
MFCD	**0.830**	**0.829**	0.834	**0.829**

由表 2 可以看出，MFCD 模型在最重要的性能指标 F1 值和准确率上分别达到了 0.830 和 0.829，均高于目前主流的多模态谣言检测模型，充分验证了本模型的先进性能。精确率与召回率在一般情况下是相互矛盾的，难以做到双高。MSRD 模型在精确率指标上最高，达到了 0.854，但其召回率却只有 0.716，这可能是由于该模型对正负样本的判别能力相差较大而导致的。

(1) 单模态与多模态的性能对比

多模态模型 FFM 和 DFM 在 F1 值上分别达到了 0.808 和 0.811，均高于纯文本模型 Textual 的 F1 值 0.775 和纯图片模型 Visual 的 F1 值 0.714，说明特征级融合和决策级融合都能够有效地提高谣言检测效果。

(2) 单级融合与多级融合的性能对比

多级融合模型 MFCD-的 F1 值达到了 0.824，分别比 FFM 和 DFM 高出了 1.6%和 1.3%，且高于目前主流的多模态谣言检测模型，说明通过构建多级融合框架进一步提高了多模态间的信息互补能力，剔除了冗余信息。

(3) 决策级融合改进前后的性能对比

MFCD 模型在 F1 和准确率指标上均高于 MFCD-模型，验证本文提出的改进后的决策级融合方案对于缓解样本中模态信息缺失情况是切实有效的。针对模态信息不全的样本单独进行统计，MFCD 模型在其上的准确率为 0.831，而 MFCD-模型的准确率仅为 0.812，进一步验证了改进后决策级融合方案的效果。

MFCD 模型效果优于 MFCD-模型的原因主要可以归纳为以下两点：

一方面在于 MFCD-模型对于样本中缺失的模态信息需要使用 0 元素进行填充。如在谣言事件“今天下午位于深圳东门发生特大暴力恐怖事件，前往东门的朋友要注意安全！请互相转告！怕二次事件再次发生!!”中，因为该事件缺少图片信息，MFCD-模型使用了大量的无意义的 0 元素填充出缺失的图片信息，增加了无用的干扰信息，从而导致了最终的分类结果错误。而 MFCD 模型直接利用纯文本进行判断，避免了图片信息缺失带来的负面影响，更容易得到正确的分类结果。

另一方面得益于 MFCD 模型对不同的子模型给予了不同的权重，更好地发挥了各自的性能。通过对超参数进行迭代，发现当 $\alpha=0.11$、$\beta=0.39$、$\gamma=0.5$ 时 MFCD 模型取得最佳分类效果。

4 结束语

本文针对目前多模态谣言检测领域存在的模态间信息融合不充分、过于依赖各模态信息完整度等问题提出了 MFCD 模型，该模型将特征级融合与决策级融合相结合并对决策级融合进行了改进，在一定程度上解决了上述的两个问题。实验结果表明，本模型性能在 F1 值和准确率指标上均优于相关基线模型。下一步将重点研究如何构建更加高效合理的特征级融合方案，进一步剔除冗余信息，提高各模态间的信息互补能力。

参考文献：

[1] LIU Yahui, JIN Xiaolong, SHEN Huawei, et al. A survey on rumor identification over social media [J]. Chinese Journal of Computer, 2018, 41 (7): 1536-1558 (in Chinese). [刘雅辉，靳小龙，沈华伟，等. 社交媒体中的谣言识别研究综述[J]. 计算机学报，2018，41 (7)：1536-1558.]

[2] 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，2016：3818-3824.

[3] LIU Zheng，WEI Zhihua，ZHANG Renxian. Rumor detection based on convolutional neural network [J]. Journal of Computer Applications，2017，37 (11)：3053-3056 (in Chinese). [刘政，卫志华，张韧弦. 基于卷积神经网络的谣言检测 [J]. 计算机应用，2017，37 (11)：3053-3056.]

[4] Chen T，Li X，Yin H，et al. Call attention to rumors：Deep attention based recurrent neural networks for early rumor detection [C] //Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer，Cham，2018：40-52.

[5] 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.

[6] Dhruv K，JaiPal S，Manish G，et al. MVAE：Multimodal variational autoencode for fake news detection [C] //Proc of the World Wide Web Conf ACM，2019：2915-2921.

[7] LIU Jinshuo，FENG Kuo，Jeff Z Pan，et al. MSRD：Multimodal web rumor detection method [J]. Journal of Computer Research and Development，2020，57 (11)：2328-2336 (in Chinese). [刘金硕，冯阔，Jeff Z Pan，等. MSRD：多模态网络谣言检测方法 [J]. 计算机研究与发展，2020，57 (11)：2328-2336.]

[8] Jin Z，Cao J，Zhang Y，et al. Novel visual and statistical image features for microblogs news verification [J]. IEEE Transactions on Multimedia，2017，19 (3)：598-608.

[9] He K，Zhang X，Ren S，et al. Deep residual learning for image recognition [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2016：770-778.

[10] ZENG Fanfeng，LI Yuke，XIAO Ke. Sentence-level fine-grained news classification based on convolutional neural network [J]. Computer Engineering and Design，2020，41 (4)：978-982 (in Chinese). [曾凡锋，李玉珂，肖珂. 基于卷积神经网络的语句级新闻分类算法 [J]. 计算机工程与设计，2020，41 (4)：978-982.]

[11] Donahue J，Hendricks LA，Rohrbach M，et al. Long-term recurrent convolutional networks for visual recognition and description [J]. IEEE Trans on Pattern Analysis and Machine Intelligence，2017，39 (4)：677-691.

[12] Poria S，Chaturvedi I，Cambria E，et al. Convolutional MKL based multimodal emotion recognition and sentiment analysis [C] // Proceedings of the IEEE 16th International Conference on Data Mining，2016：439-448.

[13] Haghighat M，Abdelmottaleb M，Alhalabi W. Discriminant correlation analysis：Real-time feature level fusion for multimodal biometric recognition [J]. IEEE Transactions on Information Forensics and Security，2016，11 (9)：1984-1996.

[14] Wu P，Liu H，Li X，et al. A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion [J]. IEEE Transactions on Multimedia，2016，18 (3)：326-338.

[15] Huang F，Zhang X，Zhao Z，et al. Image-text sentiment analysis via deep multimodal attentive fusion [J]. Knowledge-Based Systems，2019，167 (MAR. 1)：26-37.

[16] MIAO Yuqing，WANG Junhong，LIU Tonglai，et al. Joint visual-textual approach for microblog sentiment analysis [J]. Computer Engineering and Design，2019，40 (4)：1099-1105 (in Chinese). [缪裕青，汪俊宏，刘同来，等. 图文融合的微博情感分析方法 [J]. 计算机工程与设计，2019，40 (4)：1099-1105.]