基于多模态融合提升的文本分类方法

刘德志[1,2]，何柳[3]，刘幼峰[1,2]，韩德纯[2]
1. 北京航空航天大学计算机学院，北京 100191；
2. 北京航空航天大学大数据与脑机智能高精尖创新中心，北京 100191；
3. 中国航空综合技术研究所，北京 100028

摘要
尽管基于多模态的文本分类技术在应用到具体场景中具有潜力，但仍存在局限性。现有多模态融合模型要求输入数据模态对齐，因此大量不完整的多模态数据被直接浪费，从而限制了推理时可用数据的规模和灵活性。为了解决这个问题，提出了一种基于多模态融合提升的文本分类模型和不充分多模态资源训练方法。与传统方法相比，提出的模型在标准数据集上的性能平均提高了约4.25%。此外，在除文本输入模态外的其他模态缺失率为50%的情况下，不充分多模态资源训练方法的性能比传统多路由策略提高了约4%。这表明所提出的模型和训练方法具有明显的优势和有效性。

关键词
文本分类；交叉注意力；多模态融合；不充分多模态资源训练方法

中图分类号：TP183 文献标志码：A doi: 10.11959/j.issn.2096-0271.2023067

A text classification method based on multimodal fusion enhancement

LIU Dezhi[1,2], HE Liu[3], LIU Youfeng[1,2], HAN Dechun[2]
1. School of Computer Science and Engineering, Beihang University, Beijing 100191, China
2. Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100191, China
3. AVIC China Aero-Polytechnology Establishment, Beijing 100028, China

Abstract
Although multimodal text classification techniques have potential when applied to specific scenarios, there are still some limitations. Existing multimodal fusion models require modal alignment in the input data, resulting in a large amount of incomplete multimodal data being directly discarded, thus limiting the scale and flexibility of available data for inference. To address this problem, we proposed a text classification model based on multimodal fusion enhancement and an insufficient multimodal resource training method. Compared with traditional methods, our model had shown an improved performance of an average of 4.25% on a standard dataset. Furthermore, when the missing rate of other modalities except for text input was 50%, using the insufficient multimodal resource training method improved

the performance by about 4% compared with traditional multi-route strategies. The experimental results demonstrate the effectiveness of the proposed model and training method.

Key words
text classification, cross attention, multimodal fusion, insufficient multimodal resource training method

0 引言

随着网络技术的发展，信息呈现形式逐渐向多模态化发展。文本作为人类信息表达的主要载体，具有直接鲜明、容易理解、信息密度大等优点。但随着其他模态信息（如音轨、视觉信息）的不断增多，特别是在“短文本+图片”或“短文本+小视频”类媒体的高度发展中，图片和音视频中往往会隐含一些文本中没有的信息。因此，需要通过多模态技术捕捉图片和音视频中蕴含的信息，补充于文本研究中。

近年来，基于多模态的任务研究越来越热门，其中多模态融合是重要的技术之一。多模态融合应用的场景十分广泛，包括人类行为检测、人机交互、智能手机持续认证、情感识别、健康检测和多模态词向量等。这些具体场景或多或少是分类任务或其变形，其主要流程为对收集到的多模态数据进行编码融合，并分配到预设类别中。在多模态技术早期研究中，由于缺少标准合适的数据集、方法集和基线，如何统计分析互相联系的不同模态源信息成为研究重点。2016年Zadeh A等[1]和2017年Poria S等[2]做了一些多模态数据集准备研究工作，为本文使用的数据集打下了基础。

模态融合[3]是一个重要且具有挑战性的任务。对于多模态表征，主要考虑在不同模态内各自的时空特征。基于卷积神经网络[4]（convolutional neural network，CNN）、循环神经网络[5]（recurrent neural network，RNN）和深度神经网络[6]（deep neural network，DNN）的方法是具有代表性的3种抽取单模态特征的方法。多模态融合应用于分类任务主要按照融合所处的阶段进行分类，主要包括输入层融合（early fusion），即特征融合；输出层融合（decision fusion），即决策融合；中间层融合（immediate fusion），即隐层融合，如**图1**所示。输入层数据融合方法一般指对采集数据的原始信息进行预处理和简单的特征抽取后，直接将特征各个模态的特征向量组合在一起；输出层融合是指对从分布与各自模态训练的多个分类器作出的决策进行聚合，输出层融合会将来自各个分类器的决策以多种不同现存的策略来进行组合；中间层融合是指将不同的模态数据先转化为高维特征表达，与模型的中间层进行融合。

在现在的多模态研究中，使用的多模态资源大多拥有对齐输入的特点，多数做法是将规范的多模态输入注入一个拥有特征子网络的规则网络；而现实场景中，多模态信息存在不对称性。例如科普文章，多数信息来自于文本，配图经常存在缺失或者与文章本身相关性很差的问题。构建的多模态规则网络与现实场景不完全相符，输入数据要经过预处理与规范化后才能输入网络之中，这样的策略导致判断出现了不一致性，即最后的分类不是通过同一模型产生的，不能很好地解决信息存在不对

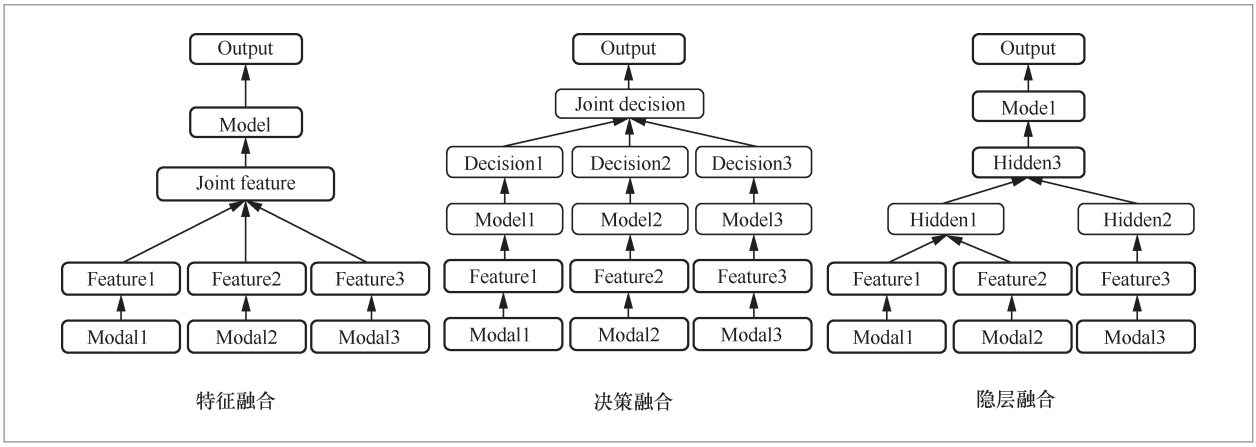


图 1　3 种不同时期多模态融合示意图 [3]

称性的问题。

为了解决以上问题，本文提出了基于多模态融合提升的文本分类模型，旨在通过捕捉多模态信息的不对称性，并利用不充分多模态资源的训练方法，提高模型在真实场景中的适应性和泛用性。经实验证明，本文提出的方法能够显著提高模型在多模态数据集上的表现，为多模态技术的应用提供了有效的思路和方法。

1 特征提取

1.1 文本特征提取

在文本特征提取方面，本文采用预训练语言模型对文本语义进行建模，以更好地提取文本的语义特征。针对预训练模型的选择，本文实验了多组模型，包括Bert-base[7]、百度的ERNIE[8]、Bert-wwm[9]和RoBERTa-wwm[10]。其中，ERNIE使用Transformer的Encoder部分，其结构与Bert相同，但采用了不同的预训练技巧。在Bert中，预训练的掩码语言模型任务（mask language model，MLM）掩盖的是词片（wordpiece），是通过BPE编码方法[11]获得的进一步的小的单位；而ERNIE是对知识级别进行掩盖，包含实体级别和短语级别的掩盖。**图2**展示了ERNIE模型不同级别的掩码策略。

Bert-wwm[9]采用全词掩码，是谷歌发布的BERT的升级版本，主要改进

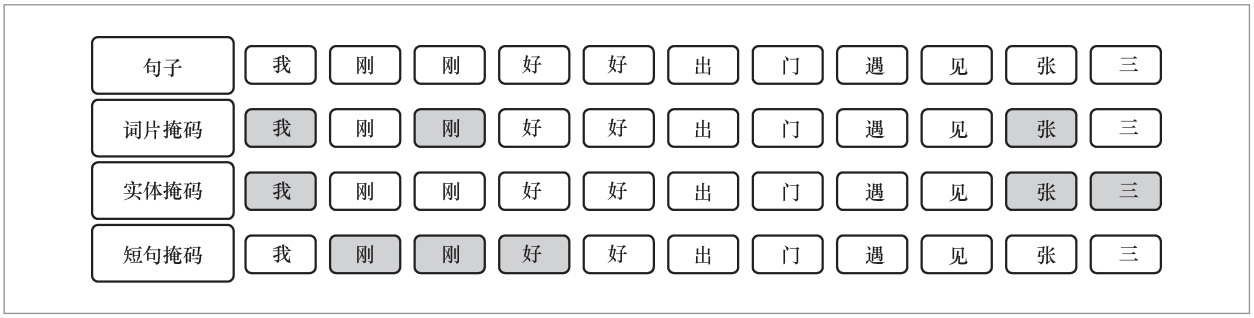


图 2　ERNIE 模型不同级别的掩码策略 [8]

了预训练阶段的训练样本生成策略。区别在于，原有基于词片的分词方式会把一个完整的词切分成若干个子词，在生成训练样本时，这些被分开的子词会随机被掩码。在全词掩码中，如果一个完整的词的部分词片子词被掩码，则同属该词的其他部分也会被掩码，即全词掩码。

RoBERTa-wwm[10]除了使用全词掩码，还采用了动态掩码技巧。在MLM任务中，每个训练周期的掩码策略动态变化，以期模型学到更加全面的信息，提高模型的表现效果。

1.2 图像特征提取

在图像特征提取部分，本文选用了ResNet152[12]进行特征提取。深度神经网络由于参数量庞大，很难训练至收敛，但残差神经网络可以有效地解决这个问题。通常来说，神经网络越深，就越具备学习复杂模型的能力，但随之梯度消失或梯度爆炸问题也会越来越明显，这会导致训练的精确度停滞或急剧下降。残差神经网络则通过添加短连接来解决这个问题，具体结构如**图3**所示。

此外，ResNet也经过了大量图像数据集的预训练，这使它能够熟练地在各种任务中提取对应图片的特征。预训练这种迁移学习的思想本身始于计算机视觉领域，之后被借鉴到自然语言处理领域，进而发展出了后来的ELMo[13]、GPT[14]以及Bert[7]等。

2 多模态联合建模方法

2.1 模态间交叉注意力

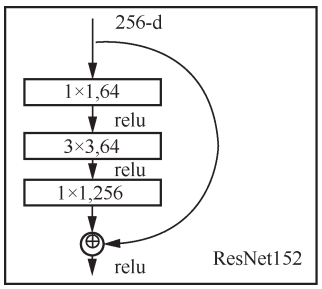


图 3 残差神经网络的一个块 [13]

参考文献[15]提出了注意力机制，该机制被应用于Transformer中，使其成为自然语言处理领域里重要的编码器之一。注意力机制的核心在于特征的交互。对于一个序列，分别用$[q_1,q_2,\cdots,q_N]$、$[k_1,k_2,\cdots,k_N]$、$[v_1,v_2,\cdots,v_N]$、$[k_1,k_2,\cdots,k_N]$表示这个序列embeddings的3个维度：询问者的查询向量（**query**）；代表自己应对来访的钥匙向量（**key**）；代表自己的价值向量（**value**）。在使用注意力机制时，首先每个词去查询所有词的钥匙，得到每个词与其他所有词的关系，然后再将这个关系作为权重，将这个词用其他所有词的注意力来表示，具体的过程可以用式（1）表示：

$$\begin{aligned} \mathrm{attn}_{ij} &= q_i \cdot k_j \\ \mathrm{attn}_{ij} &= \mathrm{softmax}\left(\frac{\mathrm{attn}_{ij}}{\dim\_k}\right) \\ o_i &= \sum_j \mathrm{attn}_{ij} v_j \end{aligned} \tag{1}$$

其中，$\dim\_k$ 为3组维度向量对的单独维度，第二个计算式中的除法用来控制 attn_{ij} 的取值范围，以免在softmax时带来过大的负担，从而使差异化更加显著。

本文的多模态融合提升模型借鉴了这种思想，提出了模态间交叉注意力机制。当多模态特征被抽取完毕后，可以将得到的核心特征分为两类：一类是经过Transformer编码获得的文本特征向量，

另一类是其他模态特征向量。

首先通过两个不同的全连接层将其他模态特征转换为相同维度的特征空间，然后将文本特征和其他模态特征分别作为查询模态和值模态。通过将值模态特征投影为两组**key**、**value**向量，然后将查询模态特征都通过全连接层映射为一个**query**向量，称之为模态**query**向量，通过对**query**、**key**、**value**向量进行同样的操作，可以实现模态间交叉注意力。如果选用其他模态作为查询模态，则可以将文本特征表示的其他模态特征进行拼接作为最终的交互特征。如果选用文本模态作为查询模态，则可以使用[CLS]对应的表征向量作为最终的交互特征。通过这种方法，可以集中图片特征和手工特征与文本特征之间的交互，并提取特征以供最后的决策。具体过程如图4所示。

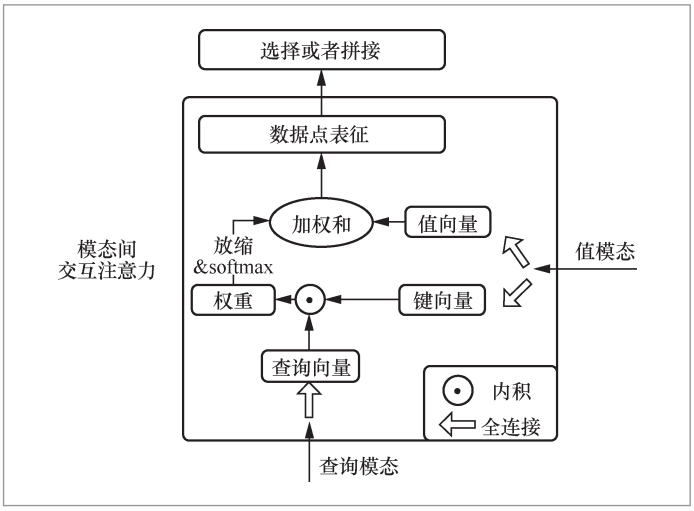


图 4 模态间交叉注意力

这样的做法可以集中图片特征以及手工特征与文本特征之间的交互，并提取特征以供最后的决策，本文对使用其他模态作为查询模态时，实验结果中图片和手工特征与文字之间的注意力做了可视化处理，以此更加充分说明模态间交叉注意力的有效性。

图5描述了文本序列和一张图片的模态间交叉注意力热力值分布。当使用图片作为**query**时，这条微博中的“我”“图”和“已经”等词被高亮了出来，这些词具有强烈的主观性特征，在谣言检测领域，这类特征能够为判断这条微博是否为谣言带来较多的信息，这说明模型着重于对比图片和文本的语义信息，并提取出了相应的特征。

图6描述了文本序列和手工特征交叉注意力热力值分布。当使用手工特征作为

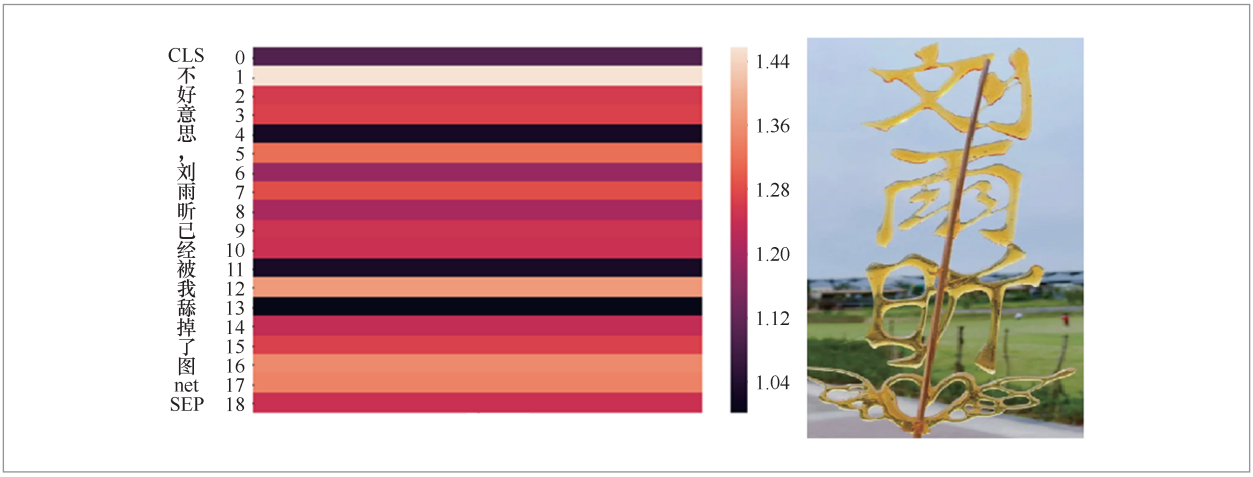
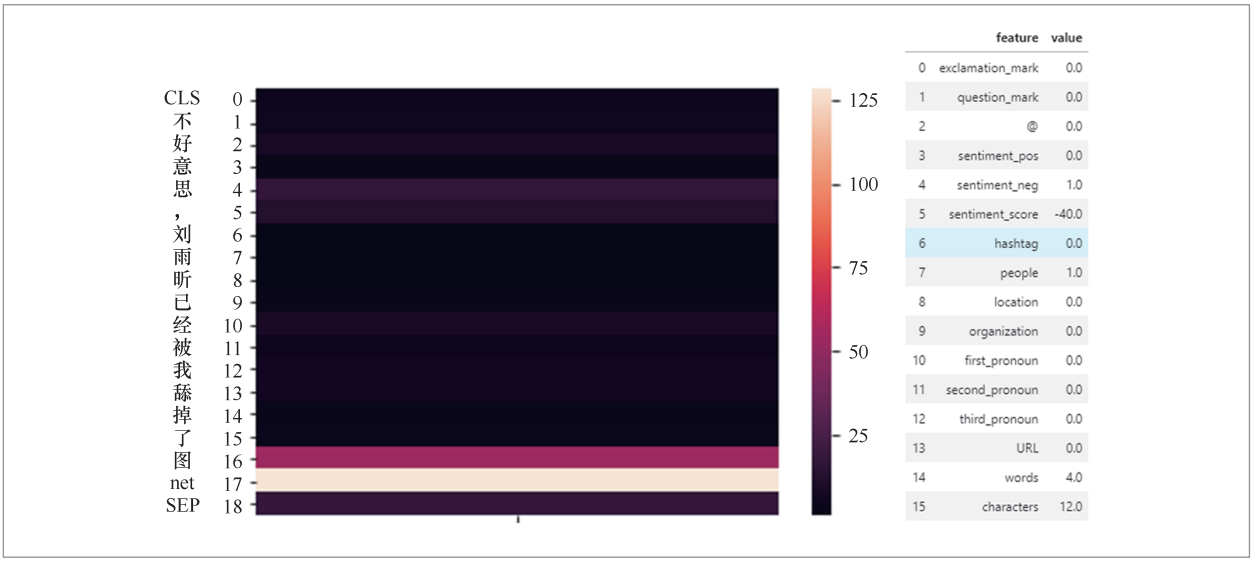


图 5 图片文本交叉注意力可视化热力图



	feature	value
0	exclamation_mark	0.0
1	question_mark	0.0
2	@	0.0
3	sentiment_pos	0.0
4	sentiment_neg	1.0
5	sentiment_score	-40.0
6	hashtag	0.0
7	people	1.0
8	location	0.0
9	organization	0.0
10	first_pronoun	0.0
11	second_pronoun	0.0
12	third_pronoun	0.0
13	URL	0.0
14	words	4.0
15	characters	12.0

图 6　手工特征文本交叉注意力可视化热力图

query时，注意力的权重较为接近，此时“图”和“net”两个词被高亮出来，结合右图手工特征中的负面情感分数，文本中的主观色彩和随意性被模型提取出来，为推理谣言提供信息。

2.2　基于多模态融合提升的文本分类模型

（1）模态交互编码器

本文采用Transformer块的设计思想来构建模态交互编码器。在实现模态间交叉注意力之后，引入了残差和层归一化结构，分别应用于模态间交叉注意力和前向全连接层之后。层归一化是一种归一化方法，主要用于解决内部协变量平移（internal covariate shift，ICS）[16]问题。本文设计的模态交互编码器如**图7**所示。

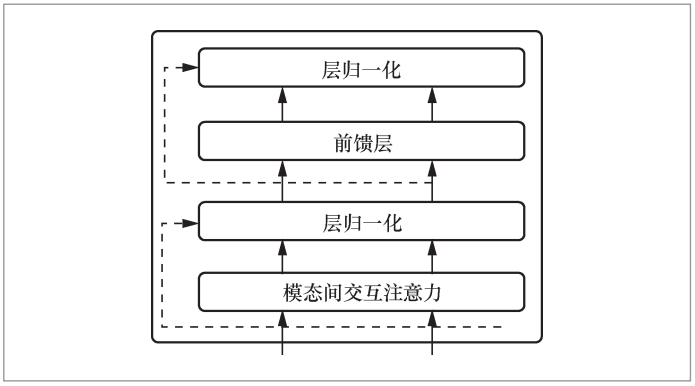


图 7　模态交互编码器

（2）多模态特征层融合

特征层的多模态融合提升模型如**图8**所示。首先，对文本使用预训练模型进行编码，获得最后一层所有词的深度表征；图片经过ResNet152进行编码，得到原始结构的1 000维特征向量；手工特征包括点赞数、评论数、转发数、文章发表时间的规范化表示、发布源等，有任意一项特征的缺失即视为手工特征缺失。由于手工特征比较短且易于提取，因此直接从数据源中提取原始数据，并采用特征工程方法将原始数据转换为特征矢量。接下来，网络分为两个分支：一条是直接的文本通路，作为主干路；另一条是多模态融合支路，为主干路的文本分类提供额外信息。

对于主干路，先对文本特征使用传统随机舍弃来提供更高的鲁棒性，然后经

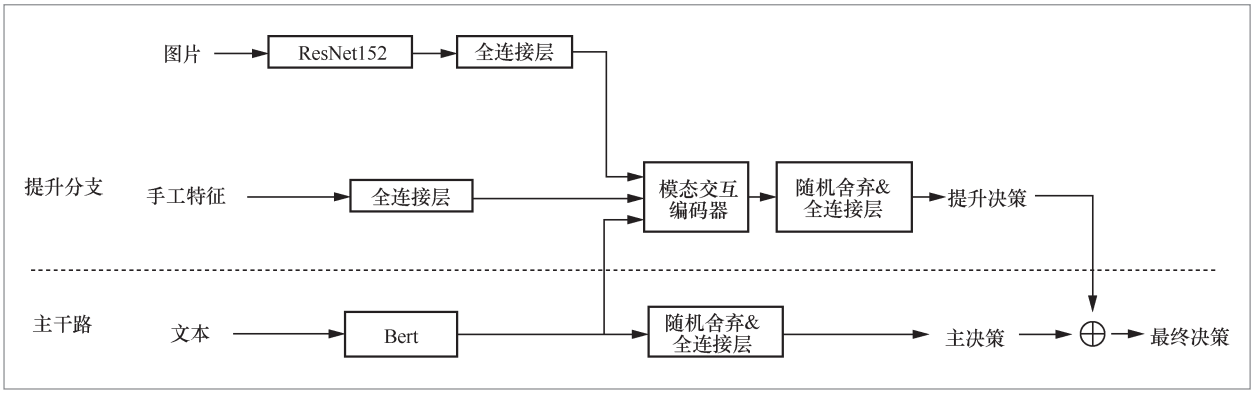


图 8 基于多模态融合提升的文本分类模型

过一层隐层，最后得出主干路的决策信息。对于支路，由于使用的图片特征和手工特征的维数与注意力维度不一致，首先使用两个全连接层将其映射为注意力维度（实验中使用的是64维），然后使用前文所说的模态间交互注意力机制，得到图片和手工特征分别与文本交互得到的特征，采取同样的先dropout后加全连接层的策略，给出多模态融合的最后支路提升决策。

（3）主干路决策层融合

由于采用主干路作为文本分类、支路作为多模态融合分类的结构，在最后主干路和支路汇合的时候，需要对两者的决策进行融合。将主干路的决策值使用按位加法来做出决策。这样的决策构造有助于直观解释，能够更好地理解文本决策的决定性作用，此外，可以使对多模态融合决策实施一定的惩罚和衰减（根据信号位）更加简单易行。

3 不充分多模态资源训练方法

不充分多模态资源训练方法可以有效地解决传统场景下的资源不平衡的问题。该方法的主要思想是：一方面，通过构造默认输入来代替缺失的模态输入，使模型在遇到模态缺失的情况下仍能正常训练以及推理；另一方面，对模型结构进行适应性改动，使模型能够以直接或者间接的方式获得缺失的模态信息。

笔者使用随机掩码算法来构造不充分多模态资源数据集。具体而言：0表示完整的数据，1表示缺失图片模态，2表示缺失手工特征模态，3表示仅包含文本模态。在训练和推理过程中，根据标志位对输入的资源进行掩码。由于不充分状态仅由标志位决定，无须对数据集进行更多的预处理。

对于默认输入，使用全0的张量作为默认输入。笔者设计了以下几种适应性结构来使模型能够适应模态不充分的训练和推理场景。

（1）标志位全连接

在获得多模态提升支路的决策向量后，将标志位**flag**向量和支路决策向量拼接，通过全连接层再次做出支路决策，这时提升支路的决策受到了标志位的间接影响，具体的过程如式（2）和式（3）所示：

$$\boldsymbol{h}_2 = [\boldsymbol{h}_1; \mathbf{flag}] \tag{2}$$

$$\boldsymbol{d} = \mathrm{GeLU}\left(\mathrm{FC}\left(\boldsymbol{h}_2\right)\right) \tag{3}$$

其中，$\boldsymbol{h}_1$是原本的提升支路决策，FC代表全连接层，*EeLU*[17]是激活函数。

(2)自学习标志位隐层掩码

对于模态间交互特征的隐层使用直接的掩码，掩码来源于标志位决定的嵌入向量。这种改动是半直接的，由于它以掩码的方式作用于隐层计算，使用按位乘法而不是像标志位全连接一样的加法，表达能力更强大。由于掩码仅作用于支路而不是主干路，且作用于隐层而非最终决策，因此具有更好的泛化性。这样的掩码仅作用于提升支路，对主干路不加干预，与本文支路仅作提升作用的想法一致。

具体的过程可以用式(4)、式(5)以及**图9**表示。其中embedding为隐层标志位掩码嵌入层，$\boldsymbol{w}$为掩码向量，$\boldsymbol{h}$为隐层特征。

$$\boldsymbol{w}=\text{embedding}(\textbf{flag}) \tag{4}$$

$$\boldsymbol{h}=\boldsymbol{w}\otimes\boldsymbol{h} \tag{5}$$

(3)自学习标志的双路决策融合

笔者对双路融合的决策进行加权融合，权重由模型学习，且权重跟数据的标志位直接相关。为了获得每条路的权重，将设置单独的嵌入层，具体的权重向量由标志位查找获得，并且两个嵌入层的参数为可学习的。提升支路和主干路的权重计算分别如式(6)、式(7)所示，决策如式(8)所示。

$$\boldsymbol{w}_1=\text{embedding}_1(\textbf{flag}) \tag{6}$$

$$\boldsymbol{w}_2=\text{embedding}_2(\textbf{flag}) \tag{7}$$

$$\boldsymbol{d}_3=\boldsymbol{w}_1\otimes\boldsymbol{d}_1+\boldsymbol{w}_2\otimes\boldsymbol{d}_2 \tag{8}$$

其中，embedding_1和embedding_2分别为提升支路和主干路的权重嵌入层，不同标志位会查找获得不同的融合权重，$\boldsymbol{w}_1$和$\boldsymbol{w}_2$分别表示提升支路和主干路的权重，$\boldsymbol{d}_1$和$\boldsymbol{d}_2$分别为提升支路和主干路的决策，$\boldsymbol{d}_3$为最终决策。使用按位乘法处理这些决策，然后将它们相加。自学习标志的双路融合通过标志位来影响最终决策。这种方法的优点是比较直接，但由于它会对每条路的最终决策产生影响，而最终决策的维度较低，因此表达能力会受到限制。

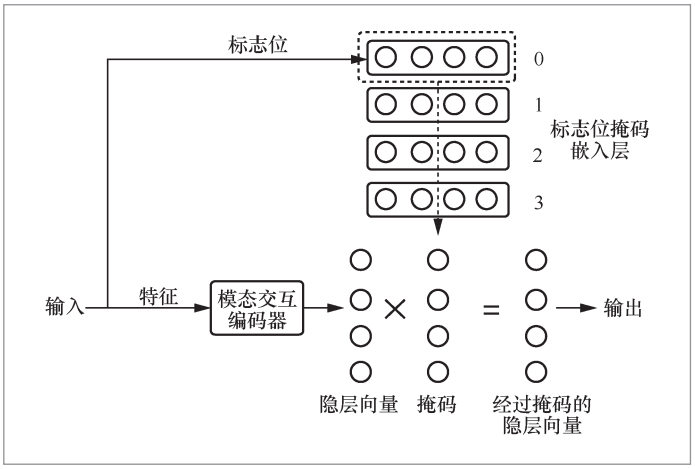


图9 自学习标志位隐层掩码

4 实验分析

4.1 数据集及评价指标

本文选用了两个谣言检测数据集，这是典型的二分类问题。所选数据集分别为来自Cao J、Qi P[18-19]的中文微博数据集和来自MediaEval2015[20]的英文推特数据集。这两个谣言数据集拥有多模态数据资源，能更好地反映实际社会需要。数据集的具体相关信息见**表1**。

本文没有对Weibo数据集进行进一步的预处理，而对于Twitter数据集，则将多种多样的短URL替换为标识符[URL]，将数字统一处理为[NUM]，并过滤了非正常数字、字母和标点符号等其他表情符号。

本文使用准确率(accuracy)和宏F1

表1 文本分类多模态融合提升方法数据集

数据集	训练集条目/条	测试集条目/条	平均长度
Weibo	7 481	1 917	110.8
Twitter	9 307	1 387	79.3

（macro-F1）分数作为算法的评估指标，使用预测logits与真实值的交叉熵作为整个多模态融合提升文本分类模型的损失函数，如式（9）：

$$H(p,q)=-\frac{1}{N}\sum_{i=1}^{K}p(x_i)\ln q(x_i) \tag{9}$$

其中，N为每个批次的数据量大小，K为类别数。$p(x_i)$是真实值的分布，实际上只有一个类别的$p(x_i)$为1，而$q(x_i)$为预测的概率分布。神经网络的输出是一个类别数维度的向量，经过softmax之后可以变成对类别分布的概率的预测。由于真实分布只有一个值为1，当预测概率与真实标签越接近时，损失函数的值将趋近于更小。

4.2 实验结果及分析

（1）基于多模态融合提升的文本分类模型

笔者使用多种不同的模型方法在Weibo以及Twitter数据集上进行测试，其结果见**表2**和**表3**。其中，Textual使用BOW特征并通过逻辑回归模型来建模；Visual使用在预训练好的VGG图像分类模型上前向传播得到的4 096维特征，使用多张图片的均值向量方法，并在逻辑回归模型上进行建模；Early Fusion是将3种模态分支的特征拼接后，在逻辑回归模型上进行建模，从而训练一个恶意信息识别的分类器；Late Fusion为本文使用上述3种模型的结果进行的晚期融合，使用概率融合（blending）和模型堆叠（stacking）两种经典的集成方法。此外，也参考了现在的一些较为先进的多模态方法，包括att-RNN[19]、VQA[21]、Neural Talk[22]、EANN[23]、MSRD[24]以及DCNN[25]。

MBN-ot（multimodal boost net-othermodal-textmodal）指的是使用其他模态作为查询模态、文本模态作为值模态，并使用模态间交叉注意力的多模态融合文本分类提升模型。最后，使用其他模态特征

表2 Weibo数据集上结果比较

方法	准确率	真实内容			谣言内容		
		精确率	召回率	宏F1	精确率	召回率	宏F1
Textual	0.592	0.605	0.531	0.566	0.581	0.653	0.615
Visual	0.608	0.61	0.605	0.607	0.607	0.611	0.609
Socal Content	0.65	0.672	0.591	0.629	0.634	0.71	0.67
Early Fusion	0.603	0.612	0.567	0.589	0.595	0.639	0.616
Late Fushion	0.669	0.693	0.611	0.649	0.651	0.728	0.687
VQA[21]	0.736	0.797	0.634	0.706	0.695	0.838	0.76
NeuralTalk[22]	0.726	0.794	0.613	0.692	0.684	0.84	0.754
att-RNN[19]	0.772	0.854	0.656	0.742	0.72	0.889	0.795
EANN[23]	0.782	0.827	0.697	0.756	0.752	0.863	0.804
MSRD[24]	0.794	0.854	0.716	0.779	–	–	–
DCNN[25]	0.803	0.799	0.801	0.809	–	–	–
MBN-ot	0.803	0.894	0.666	0.763	0.753	0.928	0.832
MBN-to	0.823	0.887	0.721	0.795	0.783	0.916	0.844

表 3 Twitter 数据集上结果比较

方法	准确率	真实内容			谣言内容		
		精确率	召回率	宏F1	准确率	召回率	宏F1
Textual	0.532	0.598	0.541	0.568	0.462	0.52	0.489
Visual	0.596	0.695	0.518	0.593	0.524	0.7	0.599
Socal Content	0.509	0.566	0.589	0.577	0.426	0.403	0.414
Early Fusion	0.619	0.727	0.528	0.612	0.542	0.738	0.625
Late Fushion	0.594	0.661	0.589	0.623	0.526	0.602	0.561
VQA[21]	0.631	0.765	0.509	0.611	0.55	0.794	0.65
NeuralTalk[22]	0.61	0.728	0.504	0.595	0.534	0.752	0.625
att-RNN[19]	0.664	0.749	0.615	0.676	0.589	0.728	0.651
EANN[23]	0.648	0.810	0.498	0.617	0.584	0.759	0.66
MSRD[24]	0.685	0.725	0.636	0.678	–	–	–
DCNN[25]	–	–	–	–	–	–	–
MBN-ot	0.720	0.830	0.794	0.812	0.427	0.486	0.455
MBN-to	0.750	0.837	0.825	0.831	0.507	0.528	0.517

向量的拼接来作为模态交叉的表征。

MBN-to(multimodal boost net-textmodal-othermodal)指的是使用文本模态作为查询模态、其他模态作为值模态，并使用模态间交叉注意力的多模态融合文本分类提升模型。最后，使用[CLS]位置的文本特征向量作为模态交叉的表征。

实验结果表明，MBN-to模型的效果最佳，对比性能最优的基线提升平均4.25%。为了进一步提升效果，笔者使用不同的预训练模型，包括ERNIE、Bert-wwm和RoberTa-wwm模型，具体结果见**表4**，表明不同预训练模型都取得了较好的效果。

其次，笔者还基于不同的对抗扰动策略进行了实验，其中FGM指的是使用了FGM对抗扰动策略[26]，PGD指的是使用了PGD对抗扰动策略[27]，具体数据见**表5**。结果表明，使用对抗扰动可以使模型效果得到进一步的提升，使用了FGM对抗策略的模型取得了最佳效果。

表 4 不同预训练模型下的效果

方法(MBN-to+)	Weibo		Twitter	
	准确率	宏F1	准确率	宏F1
Base(bert)	0.823	0.818	0.750	0.674
ERNIE	0.830	0.826	0.782	0.680
Bert-wwm	0.827	0.824	–	–
RoBERTa-wwm	0.841	0.838	–	–

表 5 不同对抗扰动下的效果

方法(MBN-to+Best-pretrain)	Weibo		Twitter	
	准确率	宏F1	准确率	宏F1
Base(no adv)	0.841	0.838	0.782	0.680
FGM	0.847	0.844	0.807	0.731
PGD	0.841	0.838	0.808	0.694

(2)不充分多模态资源训练方法

本文使用两种不充分多模态资源的分类方法进行了实验，并对多模态融合文本分类提升方法及各种改进进行了测试，验证了不充分训练方法相对于多路由方法的

有效性。**表6**展示了不充分多模态数据的分布情况，而**表7**则列出了使用不同不充分训练策略的效果。

具体而言，S1 Multi-router是采用多路由策略，使用已有数据中对齐好的多模态资源训练模型，并对测试的对齐多模态资源数据点进行推理；同时，对已有所有数据的文本模态进行训练，并对测试时缺失模态的数据点进行文本推理。S2则采用不充分的多模态训练方法，其中，padding策略仅使用填充；flag策略则使用填充和flag提示；flag_mask策略则使用自学习标志位掩码；而weight_fusion策略则使用自学习双路融合。在实验中，使用了图片模态和特征模态各自掩码50%的数据集，以比较各个模型策略的效果。在基于多模态融合提升的文本分类模型实验部分，由于使用的数据集是完全对齐的，因此此处使用的测试集与前面的测试集不同，本部分的实验对比主要针对设置的多路由策略基线。

实验结果表明，使用了不充分训练方法的测试结果比使用多路由策略提升了2%～3%，而使用flag位的不充分训练方法略优于使用padding的方法。此外，本文对策略1中多模态通道和文本单通道的表现进行了分析，并与本文的不充分训练方法的效果进行了比较，充分说明了本文方法的有效性。

以上结果表明，仅仅使用部分多模态资源进行训练的多模态融合模型效果较差，而全面地使用不充分多模态资源训练方法较多路由策略平均提升了约4%。

表 6 不充分多模态数据分布情况

数据集	Weibo		Twitter	
	训练集	测试集	训练集	测试集
Complete	1880	491	2327	335
No Img	1829	468	2323	359
No Craft	1829	468	2323	359
Only txt	1880	490	2326	334

表 7 使用不同不充分训练策略的效果

方法 (MBN+)	Weibo		Twitter	
	准确率	宏F1	准确率	宏F1
S1 Multi-router	0.800	0.795	0.711	0.634
S2 padding	0.820	0.816	0.748	0.680
S2 flag	0.827	0.823	0.764	0.646
S2 flag_mask	0.829	0.827	0.766	0.676
S2 weight_fusion	0.824	0.821	0.742	0.648

5 结束语

本文针对多模态信息的不对称性和模态资源不充分性，提出了一种基于多模态融合提升的文本分类模型和不充分多模态资源训练方法。同时，本文设计了一种模态间交互注意力和双路融合机制来处理模态之间的不对称性，主干路以文本为主，支路通过比较文本和其他模态的模式来提供支持。为了解决模态缺失的问题，笔者在现有数据集上构造了掩码来模拟真实场景，并通过适应性调整模型，使其可以应对各种情况的模态缺失。最后，使用微博和推特的谣言数据集测试了本文提出的方法。实验结果表明，相比原有的多模态融合方法，基于多模态融合提升的文本分类模型平均提升了约4.25%。在模态不充分率为50%的情况下，不充分多模态资源训练方法较多路由策略平均提升了约4%，进一步验证了基于多模态融合提升的文本分类方法以及不充分多模态资源训练方法的有效性。

参考文献:

[1] ZADEH A, ZELLERS R, PINCUS E, et al.

Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos[EB]. arXiv preprint, 2016, arXiv:1606.06259.

[2] PORIA S, CAMBRIA E, HAZARIKA D, et al. Multi-level multiple attentions for contextual multimodal sentiment analysis[C]//Proceedings of 2017 IEEE International Conference on Data Mining (ICDM). Piscataway: IEEE Press, 2017: 1033-1038.

[3] GUO W, WANG J, WANG S. Deep multimodal representation learning: a survey[J]. IEEE Access, 2019, 7: 63373-63394.

[4] CAMBRIA E, HAZARIKA D, PORIA S, et al. Benchmarking multimodal sentiment analysis[M]//Computational linguistics and intelligent text processing. Cham: Springer, 2018: 166-179.

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

[6] ZADEH A, LIANG P P, MAZUMDER N, et al. Memory fusion network for multi-view sequential learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.:s.n.], 2018.

[7] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[EB]. arXiv preprint, 2018, arXiv:1810.04805.

[8] SUN Y, WANG S, LI Y, et al. Ernie: enhanced representation through knowledge integration[EB]. arXiv preprint, 2019, arXiv:1904.09223.

[9] CUI Y M, CHE W X, LIU T, et al. Revisiting pre-trained models for Chinese natural language processing[C]// Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg: Association for Computational Linguistics, 2020: 657-668.

[10] LIU Y, OTT M, GOYAL N, et al. Roberta: a robustly optimized bert pretraining approach[EB]. arXiv preprint, 2019, arXiv:1907.11692.

[11] SENNRICH R, HADDOW B, BIRCH A. Neural machine translation of rare words with subword units[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2016: 1715-1725.

[12] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.

[13] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[EB]. arXiv preprint, 2018, arXiv:1802.05365.

[14] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[Z]. OpenAI, 2018.

[15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000-6010.

[16] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]//Proceedings of the 32nd International Conference on International Conference on Machine Learning. New

York: ACM, 2015: 448-456..
[17] HENDRYCKS D, GIMPEL K. Gaussian error linear units (GELUs)[EB]. arXivpreprint, 2016, arXiv: 1606.08415.
[18] QI P, CAO J, YANG T Y, et al. Exploiting multi-domain visual information for fake news detection[C]//Proceedings of 2019 IEEE International Conference on Data Mining. Piscataway: IEEE Press, 2020: 518-527.
[19] JIN Z W, 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. New York: ACM, 2017: 795-816.
[20] BOIDIDOU C, PAPADOPOULOS S, KOMPATSIARIS Y, et al. Challenges of computational verification in social multimedia[C]//Proceedings of the 23rd International Conference on World Wide Web. New York: ACM, 2014: 743-748.
[21] ANTOL S, AGRAWAL A, LU J, et al. Vqa: visual question answering[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2015: 2425-2433.
[22] VINYALS O, TOSHEV 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. Piscataway: IEEE Press, 2015: 3156-3164.
[23] WANG Y Q, MA F L, JIN Z W, 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. New York: ACM, 2018: 849-857.
[24] JINSHUO L, KUO F, PAN J Z, et al. MSRD: multi-modal web rumor detection method[J]. Journal of Computer Research and Development, 2020, 57(11): 2328-2336.
[25] JIANA M, XIAOPEI W, TING L, et al. Cross-modal rumor detection based on adversarial neural network[J]. Data Analysis and Knowledge Discovery, 2023, 6(12): 32-42.
[26] MIYATO T, DAI A M, GOODFELLOW I. Adversarial training methods for semi-supervised text classification[EB]. arXiv preprint, 2016, arXiv:605.07725.
[27] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[EB]. arXiv preprint, 2017, arXiv:1706.06083.

作者简介

刘德志(1996-),男,北京航空航天大学计算机机学院博士生,主要研究方向为知识消歧、信息抽取。

何柳（1988– ），男，中国航空综合技术研究所高级工程师，主要研究方向为人工智能、计算机视觉、多模态机器学习。

刘幼峰（1996– ），男，北京航空航天大学计算机学院硕士生，主要研究方向为多模态融合、知识图谱。

韩德纯（1980– ），男，北京航空航天大学大数据与脑机智能高精尖创新中心首席架构师，主要研究方向为大数据应用、网络安全系统。

收稿日期： 2023-02-22