资源受限场景下的虚假信息识别技术研究

武成龙, 胡明昊, 廖劲智, 杨慧, 赵翔

引用本文

武成龙, 胡明昊, 廖劲智, 杨慧, 赵翔. 资源受限场景下的虚假信息识别技术研究[J]. 计算机科学, 2024, 51(11): 15-22.

WU Chenglong, HU Minghao, LIAO Jinzhi, YANG Hui, ZHAO Xiang. Study on Fake News Detection Technology in Resource-constrained Environments [J]. Computer Science, 2024, 51(11): 15-22.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240700099

资源受限场景下的虚假信息识别技术研究

武成龙[1]　胡明昊[2]　廖劲智[3]　杨　慧[4]　赵　翔[1]
1 国防科技大学大数据与决策实验室　长沙 410073
2 军事科学院信息研究中心　北京 100036
3 国防大学军事管理学院　北京 100000
4 中国电子科技集团公司第三十研究所　成都 610041
(wuchenglong13@163.com)

摘　要　近年来,社交媒体因其开放性和便捷性,为虚假信息的扩散和泛滥提供了温床。相较于单模态虚假信息,多模态虚假信息通过融合文本和图片等多种信息形式,创造出更具迷惑性的虚假内容,造成更深远的影响。现有的多模态虚假信息识别方法大多基于小模型,而多模态大模型的快速发展为多模态虚假信息的识别提供了新思路。然而,这些模型通常参数众多、计算资源消耗大,无法直接部署在计算和能量资源受限的场景中。为了解决以上问题,提出一种基于多模态大模型 Long-CLIP 的多模态虚假信息识别模型。该模型能够处理长文本,关注更多粗粒度和细粒度细节。同时,利用高效多粒度分层剪枝进行模型压缩,得到一个更加轻量化的多模态虚假信息识别模型,以适应资源受限场景。最后,在微博数据集上,通过与微调前后的当前流行的多模态大模型和其他剪枝方法进行对比,验证了该模型的有效性。结果显示,基于 Long-CLIP 的多模态虚假信息识别模型在模型参数和推理时间方面远少于当前流行的多模态大模型,但检测效果更佳。模型压缩后,在检测效果仅下降 0.01 的情况下,模型参数减少 50%,推理时间减少 1.92s。

关键词:虚假信息识别;多模态大模型;资源受限;模型压缩;剪枝
中图分类号　TP391

Study on Fake News Detection Technology in Resource-constrained Environments

WU Chenglong[1],HU Minghao[2],LIAO Jinzhi[3],YANG Hui[4] and ZHAO Xiang[1]
1 Laboratory for Big Data and Decision,National University of Defense Technology,Changsha 410073,China
2 Center of Information Research,Academy of Military Science,Beijing 100036,China
3 College of Military Management,National Defense University,Beijing 100000,China
4 The 30th Research Institute of China Electronics Technology Group Corporation,Chengdu 610041,China

Abstract　In recent years,social media has become a fertile ground for the spread and proliferation of fake news due to its openness and convenience. Compared to unimodal fake news,multimodal fake news,which combines various forms of information such as text and images,creates more confusing false content and has a more far-reaching effects. Existing methods for multimodal fake news detection predominantly rely on small models. However,the rapid development of multimodal large models offers new perspectives for addressing this issue. These models,though,are typically parameter-intensive and computationally demanding,making them challenging to deploy in environments with limited computational and energy resources. To address these challenges,this study proposes a multimodal fake news detection model based on the multimodal large model Long-CLIP. This model is capable of processing long texts and attending to both coarse-grained and fine-grained details. Additionally,by employing an efficient coarse-to-fine layer-wise pruning method,a more lightweight multimodal fake news detection model is obtained to adapt to resource-constrained scenarios. Finally,on the Weibo dataset,the proposed model is compared with current popular multimodal large models before and after fine-tuning and other pruning methods,and its effectiveness is verified. Results indicate that the Long-CLIP-based multimodal fake news detection model significantly reduces model parameters and inference time compared to current popular multimodal large models,while maintaining superior detection performance. After compression,the model achieves a 50% reduction in parameters and a 1.92s decrease in inference time,with only a 0.01 drop in detection accuracy.

到稿日期:2024-07-16　返修日期:2024-08-30
基金项目:国家重点研发计划(2022YFB3102600);国家自然科学基金(72301284,62376284)
This work was supported by the National Key R&D Program of China(2022YFB3102600) and National Natural Science Foundation of China(72301284,62376284).
通信作者:赵翔(xiangzhao@nudt.edu.com)

Keywords Fake news detection, Multimodal large models, Resource-constrained, Model compression, Pruning

1 引言

在数字化时代，社交媒体已经成为获取新闻、知识和生活资讯的重要渠道[1]。然而，社交媒体开放且便捷的高效传播机制为虚假信息的扩散和泛滥提供了温床[2]，虚假信息的快速传播不仅可能引发公众恐慌、歪曲舆论导向，也能对政治选举、公共安全和市场经济造成显著干扰[3]。相较于单模态虚假信息，含有文本和图像的多模态虚假信息在传播范围和影响力上更为显著[4]。多模态虚假信息通过融合文本、图片、音频和视频等多种信息形式，能够创造出更具迷惑性和说服力的虚假内容[5]。因此，需要综合分析多模态特征，以提高虚假信息识别的准确性[6]。

现有的多模态虚假信息识别方法大多基于小模型，而多模态大模型的快速发展为多模态虚假信息识别提供了新思路。预训练多模态大模型，如 CLIP[7]，BLIP[8] 和 BLIP2[9] 等，通过预先在大规模数据集上学习到丰富特征，展现了其在多模态信息处理方面的强大潜力[10]。然而，以上模型在长文本与复杂关系建模上的能力有所不足。Long-CLIP 通过保留知识的位置编码扩充，加入核心成分对齐的微调策略，能够关注更多粗粒度和细粒度内容，从而增强了处理长文本的能力[11]。同时，大模型能力的发现凸显了扩大模型规模的重要性[12]，DeepSeek-VL[13]，Baichuan2[14] 和 Qwen-VL[15] 等更大规模的多模态大模型为多模态虚假信息识别提供了更多可能。

然而，这些模型通常参数众多、计算资源消耗大，直接部署到计算和能量资源受限的场景（如处理器速度低、内存和存储容量小以及网络带宽有限等）面临诸多挑战[16]。在资源受限的环境中，快速准确地识别虚假信息不仅要求模型高效、准确，还要足够轻量。因此，如何在资源受限场景下优化虚假信息识别技术，利用模型压缩减少计算和存储需求，成为当前研究的关键方向。

针对以上问题，本研究提出了一种基于多模态大模型 Long-CLIP 的多模态虚假信息识别模型。该模型能够处理长文本，关注更多粗细粒度细节。同时，面向资源受限场景，研究利用高效多粒度分层剪枝的模型压缩方法。首先，在给定模型总压缩比的情况下，利用零阶梯度获得每层的压缩比，然后通过逐层的方式去除不关键的权重，得到一个更加轻量化的多模态虚假信息识别大模型。本方案不仅有助于识别多模态虚假信息，还能在降低模型复杂度的同时尽量维持识别效率，为多模态虚假信息识别大模型在资源受限场景下的部署提供了可行性。本研究的创新点总结如下：

1）基于多模态大模型 Long-CLIP，开展多模态虚假信息识别模型的设计与实现。

2）面向资源受限场景的现实应用需求，利用高效多粒度分层剪枝方法，开展多模态大模型压缩的实现。

3）通过与微调前后的当前流行的多模态大模型和其他剪枝方法进行对比，验证了该模型的有效性。

2 相关工作

考虑多模态大模型在资源受限场景中的应用，当前多模态虚假信息识别任务的研究可以从虚假信息识别、基于视觉-语言基座的多模态大模型，以及多模态大模型剪枝 3 个角度出发。

2.1 虚假信息识别

根据信息的模态形式，虚假信息识别技术可分为单模态和多模态两类。单模态虚假信息识别研究主要侧重于统计文本特征[17]或图像特征[18]，利用 CNN[19]、RNN[20]、注意力机制[21]和图结构[22]等深度神经网络提取语义、风格、情感和图像信息等特征，以识别虚假信息。除此之外，一些研究也考虑了元数据，旨在捕捉基于评论[23]、平台[24]、用户资料[25]和传播结构[26]等特征进行虚假信息识别。多模态假信息识别的研究主要集中在一致性对齐和交互融合。一致性对齐通常使用相似性比较[27]、语义匹配[28]、实体对齐[29]以及其他对齐策略[30]进行检测；而交互融合机制大致可分为早期融合[31]和晚期融合[27]两类。近年来，大模型凭借其卓越的信息处理能力，在解决多模态任务方面展现了令人印象深刻的效果[12]，被视为有潜力的通用解决方案[32]。

2.2 基于视觉－语言基座的多模态大模型

大模型能力的发现凸显了扩大模型规模的重要性[33]。在视觉语言等多模态任务中，使用自回归语言模型作为解码器，利用跨模态转移的优势，允许在语言和多模态领域之间共享知识[33]。以视觉-语言大模型为代表的多模态大模型相继实现了惊人的性能。例如，CLIP 利用大规模对比学习同时理解图像与文本，从而增强了其在多种视觉任务中的泛化性能[7]。BLIP 通过自我监督学习进一步加强了图文协同理解，展示了自回归语言模型在理解复杂图文关系方面的高效性[8]。随后，BLIP-2 采用 Flan-T5 和 Q-Former，有效地将视觉特征与语言模型对齐[9]。再后来，具有长文本处理能力的 Long-CLIP 弥补了在长文本和复杂关系建模上的重大短板[11]。此外，DeepSeek-VL[12]，Baichuan2[13] 和 Qwen-VL[14] 等多种更大参数的多模态大模型的出现也标志着视觉-语言大模型迎来了发展的新纪元。

2.3 多模态大模型剪枝

通过剪枝进行模型压缩是大模型出现后长期受到研究者们关注的热点之一[34]，并且随着基于 Transformer 的模型规模增大脱颖而出[35]。剪枝通常分为结构化剪枝和非结构化剪枝。结构化剪枝通过牺牲模型性能来提升推理速度和吞吐量[36]；非结构化剪枝[37]即使在人工智能加速软件或稀疏矩阵计算方案的高模型稀疏度下，也能保持较优的性能[38]。Frantar 等[39]提出了一种一次性剪枝技术 SparseGPT，其以非结构化方式实现了 GPT 系列的大规模预训练 Transformer 结构模型的快速压缩。Sun 等[37]提出了一种针对大模型的基于量值的非结构化剪枝方法 Wanda，它根据重要性促进分层权重稀疏化，通过同时考虑权重和激活来确定被移除的权重。Shi 等[40]提出了统一渐进式剪枝技术来压缩视觉-语言

Transformer 模型，实现了对可压缩模态与结构的自动剪枝比例分配，通过逐步搜索和重新训练子网络，获得了更高的压缩比。Sung 等[41]提出高效多粒度分层剪枝方法，利用近似的全局重要性得分找到每层的自适应稀疏度，并利用零阶优化仅通过前向传递获得梯度，采用逐层剪枝方式调整视觉语言模型的优化剪枝率。

总体来说，多模态大模型在多模态虚假信息识别领域具有巨大潜力，而通过剪枝实现多模态大模型压缩，有利于在资源受限场景下实现快速部署和高效推理。

3 模型介绍

本研究旨在利用预训练多模态大模型在多模态信息处理方面的强大潜力，设计多模态虚假信息识别模型。同时，为满足资源受限场景的现实应用需求，通过剪枝技术实现多模态大模型的压缩，得到更加轻量化的模型。图 1 给出了基于 Long-CLIP 的多模态虚假信息识别模型架构，该模型包括图像特征提取、文本特征提取、特征融合和分类等模块。

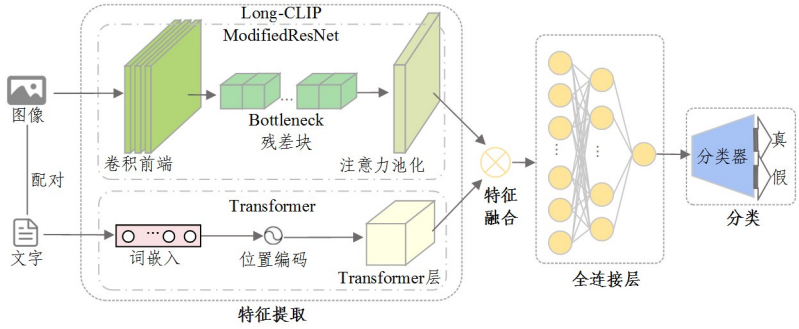


图 1 基于 Long-CLIP 的多模态虚假信息识别模型架构

Fig. 1 Framework of multimodal fake news detection model based on Long-CLIP

3.1 多模态虚假信息识别网络

3.1.1 图像特征提取

图像通过提供直观的视觉证据，在识别和分析虚假信息中发挥着至关重要的作用。本研究使用 ModifiedResNet 进行图像特征提取，其结合了传统的 ResNet 特性和一些关键改进。

给定多模态虚假信息 $\mathcal{D}=\{\boldsymbol{x}_k^v,\boldsymbol{x}_k^t\}_{k=1}^K$。首先采用一个由 3 个卷积层组成的卷积前端结构，每层卷积核大小均为 3×3。对于第 m 层：

$$\boldsymbol{x}_{m+1}^v=AvgPool(\mathrm{ReLU}(BN(Conv(\boldsymbol{x}_m^v)))) \tag{1}$$

其中，$\boldsymbol{x}_m^v$ 表示第 m 层卷积的图像数据，$Conv(\cdot)$表示卷积操作，$BN(\cdot)$表示批量归一化操作，$\mathrm{ReLU}(\cdot)$表示 ReLU 激活函数，$AvgPool$ 表示平均池化操作。

其次是 Bottleneck 单元，这是一种残差网络结构，其结构如图 2 所示。最后，采用多头注意力池化层，突出更重要的视觉信号。设输入的图像数据 $\boldsymbol{x}^v$ 提取到的图像特征为 $\boldsymbol{e}^v$。

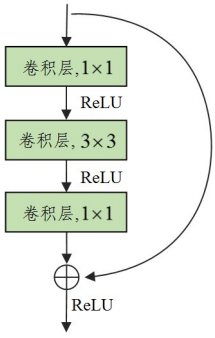


图 2 Bottleneck 单元

Fig. 2 Bottleneck unit

3.1.2 文本特征提取

通过深入挖掘文本特征，能够捕捉到虚假信息中的不一致性、逻辑漏洞、情感偏向等重要线索。本研究主要依赖 Transformer 实现文本特征提取，其流程如图 3 所示。

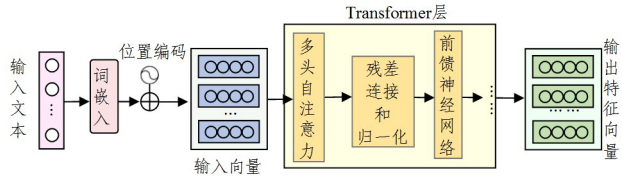


图 3 文本特征提取流程

Fig. 3 Process of text feature extraction

输入的文本数据 $\boldsymbol{x}^t$ 通过词嵌入层，实现文本信息的向量化表示。同时，使用位置编码表示序列中的位置信息。将词嵌入向量与位置编码向量相加，得到输入向量。

多头自注意力机制计算查询向量 $\boldsymbol{Q}$、键向量 $\boldsymbol{K}$ 和值向量 $\boldsymbol{V}$，并通过点积、缩放和 softmax 操作得到注意力权重，最终将注意力权重与值向量相乘并求和。前馈神经网络由两个全连接层组成，通过激活函数进行非线性转换，进一步提取和转换特征。以上两个子层通过残差连接和层归一化进行连接。通过堆叠多个 Transformer 层，模型能够逐层地对输入的文本序列进行特征提取和转换，从而获得更丰富、更高级的文本表示。设经过 Transformer 层处理后得到图像特征 $\boldsymbol{e}^t$。

3.1.2 特征融合

图像可能显示具有误导性的视觉信息，而文本可能包含不准确或误导性的描述。通过融合这两种模态的特征，可以显著提高识别虚假信息的准确性。本研究采用哈达玛积实现特征融合。哈达玛积融合通过将对应元素相乘产生新的特征向量，强调了输入特征之间的相互作用，不会增加特征的维度，相比串联融合更加轻量级。融合后得到的特征如下：

$$\boldsymbol{c}=\boldsymbol{e}^v\odot\boldsymbol{e}^t \tag{2}$$

其中，$\odot$表示对应元素相乘。

3.1.4 分类

在得到融合特征后，将特征向量依次输入两层全连接神经网络进行处理。其流程可表示为：

$$\boldsymbol{h}_j = Dropout(\mathrm{Swish}(BN(\boldsymbol{W}_j\boldsymbol{c}+\boldsymbol{b}_j))) \tag{3}$$

其中，$\boldsymbol{W}_j$ 表示第 j 层全连接层权重矩阵，$\boldsymbol{b}_j$ 表示第 j 层全连接层偏置向量，Swish(·)表示 Swish 激活函数，$Dropout(\cdot)$ 表示 Dropout 操作。Swish 激活函数公式如下：

$$\mathrm{Swish}(x)=\frac{x}{1+\mathrm{e}^{-x}} \tag{4}$$

由于多模态虚假信息识别本质上是一个二分类问题，因此在分类模块中，本研究采用交叉熵损失函数作为模型损失的计算函数：

$$\mathscr{L}=-\sum_{k=1}^{K}\sum_{g=1}^{G}\boldsymbol{y}_{kg}\log(\hat{\boldsymbol{y}}_{kg}) \tag{5}$$

其中，K 表示样本数量，G 表示类别数量，$\boldsymbol{y}_{kg}$ 表示第 k 个样本第 g 类的真实标签独热编码，$\hat{y}_{kg}$ 表示第 k 个样本为第 g 类的预测概率，由 softmax 函数计算得出。其计算过程如下：

$$\hat{y}_{kg}=\frac{\mathrm{e}^{h_{kg}}}{\sum_{g=1}^{G}\mathrm{e}^{h_{kg}}} \tag{6}$$

其中，h_{kg} 为最后一个全连接层的输出向量 $\boldsymbol{h}_2$ 中的对应值。

3.2 高效多粒度分层剪枝

基本的剪枝方法主要分为全局剪枝和分层剪枝两类。迭代全局剪枝作为全局剪枝的一种改进方法，在模型参数重要性排序的基础上进行优化。在每次剪枝之后，模型都会经历微调以恢复部分性能，然后再进行下一轮剪枝，一直到达到预设的压缩比。由于需要多次进行剪枝和微调，因此迭代全局剪枝计算开销较大，训练时间较长。逐层一次性剪枝是分层剪枝的经典方法，每一层的参数根据其重要性度量，例如权重的绝对值，被独立地评估和排序。然后，根据预设的固定剪枝比例，剪除重要性较低的参数。剪枝是逐层进行的，每层压缩比被设置为相同的固定值，这可能导致无法达到全局最优的模型压缩效果。

本研究采用高效多粒度分层剪枝解决了前两种剪枝方法存在的问题。如图 4 所示，高效多粒度分层剪枝首先执行高效的粗粒度步骤，在给定模型总压缩比的情况下，利用零阶梯度获得每层的压缩比，然后在细粒度步骤中逐层去除不关键的权重。

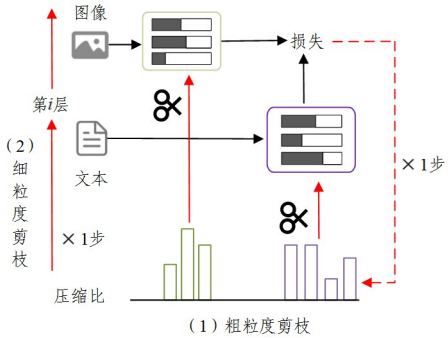


图 4 高效多粒度分层剪枝步骤

Fig. 4 Steps of efficient coarse-to-fine layer-wise pruning

高效多粒度分层剪枝本质上是一种分层剪枝方法。传统的分层剪枝方法旨在通过相应的局部目标 $\mathscr{L}_s$ 找到第 s 层的稀疏权重 $\hat{\boldsymbol{W}}_s$：

$$\hat{\boldsymbol{W}}_s=\mathrm{argmax}\,\mathscr{I}(\boldsymbol{W}_s|\hat{\boldsymbol{W}}_{s-1},D,\mathscr{L}_s) \tag{7}$$

其中，$\mathscr{I}(\cdot)$ 表示计算权重重要性的得分函数，$\hat{\boldsymbol{W}}_s$ 表示 $\boldsymbol{W}_s$ 的剪枝权重，$p_s\%$ 为第 s 层的稀疏度。在逐层剪枝方法中，所有层的 p_s 通常都是相同的。

在高效多粒度分层剪枝中，通过计算模型权重的零阶梯度作为全局重要性来估计视觉-语言模型中每一层的最佳稀疏度。获得层稀疏度的一个直接方向是利用全局损失目标 $\mathscr{L}$ 对模型进行全局剪枝，以获得目标全局稀疏度 p。

$$\hat{\boldsymbol{W}}=\mathrm{argmax}\,\mathscr{I}(\boldsymbol{W}|\mathscr{D},\mathscr{L}) \tag{8}$$

然后，通过 $p_s=\hat{\boldsymbol{W}}_s/\boldsymbol{W}_s$ 估计层稀疏度。但是，这种方法需要大量的操作来获取大模型中所有权重的梯度，并通过 argmax 提取修剪后的权重，会消耗大量的内存和计算资源。因此，本研究引入一种替代的数值方法，该方法根据重要性得分线性计算保留比(1−稀疏比)，以避免估计全局重要性得分的大量操作。首先，通过全局目标函数获得每个权重的重要性得分，即 $\boldsymbol{s}=\mathscr{I}(\boldsymbol{W}|\mathscr{D},\mathscr{L})$，然后通过 3 个步骤将得分转换为稀疏性：1)基于 p 进行选择找出需要的总参数；2)对分数进行归一化；3)根据要选取的参数数量和该层的参数获得每层的稀疏度。确定模型第 s 层稀疏率的示例如下所示：

$$normalize(s_s,\boldsymbol{s})=\frac{\boldsymbol{s}_s}{\sum\boldsymbol{s}} \tag{9}$$

$$N_{\mathrm{select}}=(1-p)\cdot|\boldsymbol{W}| \tag{10}$$

$$p_s=1-\frac{(normalize(s_s,\boldsymbol{s})\cdot N_{\mathrm{select}})}{|\boldsymbol{W}_s|} \tag{11}$$

算法 1 高效多粒度分层剪枝算法

输入：视觉模型权重 $\mathbf{W}_{\mathrm{v}}$，语言模型权重 $\mathbf{W}_{\mathrm{l}}$，多模态数据集 $\mathscr{D}$，目标稀疏度 p，每层最大稀疏度 $\mathrm{p}_{\max}$

输出：剪枝权重 $\hat{\mathbf{W}}_{\mathrm{s}}^{\mathrm{v}}$ 和 $\hat{\mathbf{W}}_{\mathrm{s}}^{\mathrm{l}}$

1. $\mathbf{s}\leftarrow\mathscr{I}(\mathbf{W}^{\mathrm{v}},\mathbf{W}^{\mathrm{l}},\mathscr{D},\mathscr{L})$//获取全局重要性得分
2. $\mathrm{p}\leftarrow\mathrm{getSparityFromScores}(\mathbf{s},\mathbf{W}^{\mathrm{v}},\mathbf{W}^{\mathrm{l}},\mathrm{p},\mathrm{p}_{\max})$// 通过式(9)获得稀疏度
   //细粒度步骤：进行分层剪枝
3. $\mathbf{W}^{\mathrm{v}},\mathbf{W}^{\mathrm{l}}=\{\},\{\}$//将第 0 个权重初始化为单位矩阵，第一层的输入是原始输入
4. $\mathrm{W}_0^{\mathrm{l}}\leftarrow\mathrm{I}$
5. for s=1 to M do
6. $\quad\hat{\mathbf{W}}_{\mathrm{s}}^{\mathrm{v}}=\mathrm{argmax}\,\mathscr{I}(\mathbf{W}_{\mathrm{s}}^{\mathrm{v}}|\hat{\mathbf{W}}_{\mathrm{s-1}}^{\mathrm{v}},\mathscr{D},\mathscr{L}_{\mathrm{s}}^{\mathrm{v}},\mathbf{p}_{\mathrm{s}}^{\mathrm{v}})$
   //逐层修剪视觉模型权重
7. $\quad\hat{\mathbf{W}}^{\mathrm{v}}=\hat{\mathbf{W}}^{\mathrm{v}}+\{\hat{\mathbf{W}}_{\mathrm{s}}^{\mathrm{v}}\}$
8. $\quad\mathbf{W}_0^{\mathrm{l}}\leftarrow\mathbf{W}_{\mathrm{M}}^{\mathrm{v}}$
9. for s=1 to L do
10. $\quad\hat{\mathbf{W}}_{\mathrm{s}}^{\mathrm{l}}=\mathrm{argmax}\,\mathscr{I}(\mathbf{W}_{\mathrm{s}}^{\mathrm{l}}|\hat{\mathbf{W}}_{\mathrm{s-1}}^{\mathrm{l}},\mathscr{D},\mathscr{L}_{\mathrm{s}}^{\mathrm{l}},\mathbf{p}_{\mathrm{s}}^{\mathrm{l}})$//逐层修剪语言模型权重
11. $\quad\hat{\mathbf{W}}^{\mathrm{l}}=\hat{\mathbf{W}}^{\mathrm{l}}+\{\hat{\mathbf{W}}_{\mathrm{s}}^{\mathrm{l}}\}$
12. return $\hat{\mathbf{W}}_{\mathrm{s}}^{\mathrm{v}},\hat{\mathbf{W}}_{\mathrm{s}}^{\mathrm{l}}$

然后，将导出的稀疏率代入式(7)，以逐层修剪模型。同时，引入一个超参数 $p_{\max}$ 来控制每层的最大稀疏度，以避免层崩溃。为了将此超参数合并到式(9)，只需为每一层预先选择

参数以满足最大稀疏条件，用预先选择的参数数量减去 N_{select}，然后启动算法。逐层剪枝，即将层的输入激活与权重大小结合起来作为局部重要性得分，根据获得的层稀疏率删除不关键的参数。算法 1 详细描述了高效多粒度分层剪枝方法。

4 实验设计与分析

本章主要介绍实验所使用的数据集、实验的基本设置，以及对实验结果的深入分析。

4.1 数据集

本研究使用了由 Jin 等创建的多模态虚假信息微博数据集[42]。该数据集包含了微博官方辟谣系统在 2012 年 5 月至 2016 年 1 月期间核实的所有虚假信息，以及由中国权威通讯社新华社核实的非虚假信息。数据集由原始推文文本、附加图像以及相关的社会背景信息组成，本研究对所使用的数据进行图文配对和预处理。具体信息如表 1 所列。

表 1 微博数据集统计信息

Table 1 Weibo dataset statistics

数据集划分	真实信息	虚假信息	总计
训练集	2898	2517	5415
验证集	389	454	843
测试集	709	756	1465

4.2 实验设置

为了提高模型的普适性，本研究参考当前的研究方法，并选择实验参数的通用尺寸。在图像特征和文本特征提取阶段，特征向量均编码为 768 维。全连接层部分包含两个隐藏层，分别包含 256 和 128 个隐藏单元。为增加模型的鲁棒性和训练效率，采用了 AdamW 优化器。激活函数选取了 Swish 和 ReLU，其中 Swish 用于全连接层部分，而其他地方使用 ReLU。模型迭代次数为 100，批处理大小为 64，学习率设定为 0.001，权重衰减率为 0.001。为防止过拟合，采用 dropout 技术，其值为 0.5。训练过程中采用了早停法，设置耐心值为 5，即若连续 5 个 epoch 在验证集上性能没有改善，则触发早停。在迭代剪枝过程中，将迭代次数和微调次数均设置为 3。在高效多粒度分层剪枝过程中，将 p_{max} 设置为比目标稀疏度大 0.1，噪声数量设置为 1，ε 设置为 0.001。为了评估算法性能，本研究采用了准确率、精确率、召回率、F1 值、推理时间和参数数量等广泛使用的评估指标，重点关注准确率、F1 值、推理时间和参数数量。准确率衡量了模型在整体样本中的正确分类比例，而 F1 值综合考虑了精确度和召回率。推理时间是衡量模型压缩效果的重要指标，而参数数量直接影响模型的规模和复杂度，同时也会影响模型的存储需求和计算资源消耗。本模型基于 PyTorch 2.2.1 构建，运行在 Linux 系统上，使用 NVIDIA A800-SXM4-80GB GPU 进行模型训练。

4.3 对比实验

4.3.1 多模态大模型对比实验

为了验证基于 Long-CLIP 的多模态虚假信息识别模型的优越性，本研究选取了 4 个当下流行的多模态大模型进行对比，分别是 DeepSeek-VL-7B-Chat，Baichuan2-13B-Chat，Yi-34B-Chat 以及 Qwen-VL-Chat。

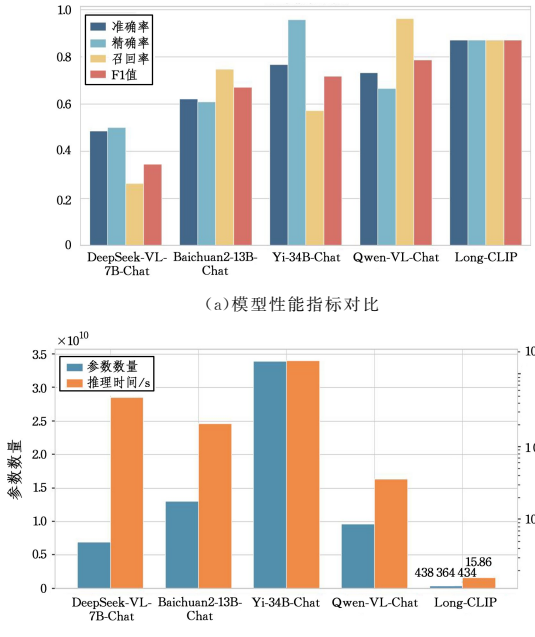
DeepSeek-VL 是于 2024 年 3 月发布的开源视觉-语言模型，旨在应用于实际的视觉和语言理解任务中。其在相同模型规模下实现了一系列视觉-语言基准测试中的最优性能。

Baichuan 2 是百川智能于 2023 年 12 月发布的第二代多模态大语言模型，包括 7B 和 13B 的 Base 和 Chat 版本。

Yi-34B 是由零一万物于 2023 年 11 月开源的双语大型语言模型，该模型在多项评测中取得了全球领先地位，并达到了多项 SOTA 指标表现。

Qwen-VL 系列是阿里云于 2023 年 9 月开源的大型视觉语言模型，主要用于感知和理解文本和图像，在图像描述、通用视觉问答、面向文本的视觉问答等任务中展现出卓越性能。

对比实验结果如图 5 所示。从图中可以看出，基于 Long-CLIP 的多模态虚假信息识别模型在所有性能指标上均表现优异，其准确率和 F1 值均高于其他模型，且 4 项指标较为接近，说明其在处理正负样本时具有较好的平衡性。相比之下，DeepSeek-VL-7B-Chat 的表现较为逊色，其准确率和精确率均在 0.5 左右，召回率和 F1 值则更低。Yi-34B-Chat 和 Qwen-VL-Chat 的表现接近，虽然 Yi-34B-Chat 在精确率上表现出色，但其召回率较低，导致 F1 值略低于 Qwen-VL-Chat。Qwen-VL-Chat 在召回率方面表现尤为突出，能够在较高的召回率下保持相对较高的精确率和准确率。



(a)模型性能指标对比

(b)模型参数数量和推理时间对比

图 5 多模态大模型对比实验结果

Fig. 5 Comparison results of multimodal large models

在参数数量和推理时间方面，基于 Long-CLIP 的多模态虚假信息识别模型的参数数量最少，仅为 428 364 434，而且其推理时间最短，仅为 15.86s。这一结果表明，与其他模型相比，本研究设计的基于 Long-CLIP 的多模态虚假信息识别模型在保持高性能的同时，具备极高的计算效率和资源利用率。虽然 Yi-34B-Chat 和 Qwen-VL-Chat 在模型性能指标上表现接近，但 Yi-34B-Chat 的参数数量高达 340 亿左右，推理时间长达 15 000 s 左右。

总的来说，基于 Long-CLIP 的多模态虚假信息识别模型

在模型参数和推理时间方面远少于当前流行的多模态大模型，但检测效果最佳。综合来看，Qwen-VL-Chat 表现较好。为进一步凸显本研究所设计模型的优越性，接下来将对 Qwen-VL-Chat 进行微调，并进行对比实验。

4.3.2 多模态大模型微调对比实验

微调是通过在预训练的基础上进一步调整模型参数，使大模型保持原有的优势，同时针对目标任务进行优化，提高模型的准确性和有效性。LORA(Low-Rank Adaptation)微调是一种轻量级的模型微调方法，旨在高效地调整预训练模型。本研究将 LORA 微调前后的 Qwen-VL-Chat 模型与基于 Long-CLIP 的多模态虚假信息识别网络模型进行对比实验，结果如表 2 所列。

表 2 多模态大模型微调对比实验结果

Table 2 Fine-tuning comparison of multimodal large models

模型	准确率/%	精确率/%	召回率/%	F1 值/%	推理时间/s	参数数量
微调前	0.732	0.666	**0.963**	0.788	357.461	9 656 935 168
微调后	0.745	0.845	0.787	**0.960**	375.630	9 656 935 168
Long-CLIP	**0.871**	**0.872**	0.871	0.871	**15.860**	**428 364 434**

微调后 Qwen-VL-Chat 的准确率由 0.732 提升到 0.745，精确率由 0.666 提升到 0.845，召回率下降至 0.787，F1 值显著提高到 0.960，表明模型在综合考虑精确率和召回率后的表现更加平衡。虽然微调后的 Qwen-VL-Chat 性能有所提升，但在准确率、精确率、召回率方面仍不及基于 Long-CLIP 的多模态虚假信息识别模型。虽然微调后的 Qwen-VL-Chat 的 F1 值较高，但 Long-CLIP 模型的推理时间和模型参数远少于 Qwen-VL-Chat 模型。

总的来说，Long-CLIP 模型效果优于微调后的 Qwen-VL-Chat 模型，其在保持高性能的同时，显著减少了参数数量和推理时间，在进行多模态虚假信息识别时展示出了更高的效率和可行性。接下来，将考虑资源受限的情况，使用高效多粒度分层剪枝方法进行模型压缩的对比实验。

4.3.3 全局剪枝与分层剪枝对比实验

为了验证使用高效多粒度分层剪枝方法实现基于 Long-CLIP 的多模态虚假信息识别模型压缩的优越性，本研究选取了全局幅度剪枝、迭代全局剪枝、迭代 OBD 剪枝这 3 种全局剪枝方法和 1 种分层剪枝方法 Wanda 进行对比实验。

全局幅度剪枝通过计算权重的大小来评估参数的重要性，根据重要性得分进行剪枝，以减少模型参数数量并保持性能。迭代全局剪枝是一种反复执行全局剪枝的策略，通过多次剪枝和再训练，逐步逼近最优的稀疏模型，同时尽可能地保持模型性能。迭代 OBD 剪枝本质上是一种迭代全局剪枝方法，它使用 OBD 方法来计算参数的重要性得分，该得分基于参数的权重和梯度的乘积。Wanda 是一种分层剪枝方法，其利用权重大小与输入激活范数的乘积作为局部重要性得分。

本实验结果主要关注准确率、F1 值和推理时间 3 个指标，5 种剪枝方法在不同压缩比下的准确率和 F1 值的变化情况如图 6 所示，推理时间变化情况如图 7 所示。

由图 6 可知，全局幅度剪枝方法在低压缩比下准确率和 F1 值表现较差，这可能与剪枝大量参数后模型性能严重下降有关，但在高压缩比下性能有所改善。迭代 OBD 剪枝和迭代全局剪枝方法总体上比全局幅度剪枝方法的效果好很多，这可能是因为迭代训练与微调恢复了模型的部分性能。同时，迭代 OBD 剪枝和迭代全局剪枝方法在较高压缩比下表现稳定，但在低压缩比下模型性能会有一定损失。分层剪枝方法总体上优于全局剪枝方法。高效多粒度分层剪枝方法在不同压缩比下准确率和 F1 值表现最佳。

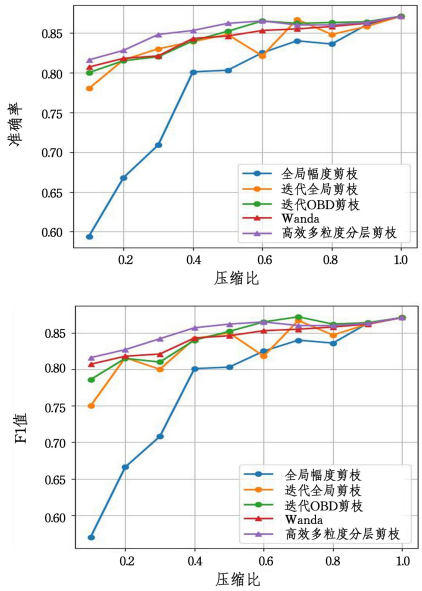


图 6 准确率和 F1 值实验结果

Fig. 6 Results of accuracy and F1 score

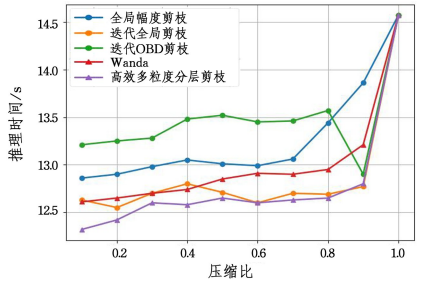


图 7 推理时间实验结果

Fig. 7 Results of inference time

由图 7 可知，在相同的压缩比下，各方法的推理时间表现出明显的差异。整体来看，推理时间由长到短依次是迭代 OBD 剪枝、全局幅度剪枝、Wanda、迭代全局剪枝和高效多粒度分层剪枝。当压缩比接近 1 时，所有方法的推理时间急剧增加，这说明基于 Long-CLIP 的多模态虚假信息识别模型中存在较多冗余参数，有必要进行剪枝以提升推理效率。

总的来说，高效多粒度分层剪枝方法在不同压缩比下具有明显的时间和效率优势，同时还能保持较高的准确率和 F1 值，可以实现基于多模态大模型 Long-CLIP 的多模态虚假信息识别模型轻量化。在检测效果下降 0.01 的情况下，模型

参数减少 50%，推理时间减少 1.92s，是 5 种方法中实现模型压缩的最优选择，其在实际应用中具有巨大潜力和优越性。

4.4 消融实验

为了进一步研究基于 Long-CLIP 的多模态虚假信息识别模型各个组件的有效性，进行了两组消融实验，分别为仅提取图像特征和仅提取文字特征。在实验中，仍关注准确率、F1 值和推理时间 3 个指标，实验结果如表 3 所列。

表 3 消融实验结果

Table 3 Ablation study results

模型	准确率/%	F1 值/%	推理时间/s
仅提取图像特征	0.665	0.673	15.64
仅提取文字特征	0.714	0.653	**15.53**
原模型	**0.871**	**0.871**	15.86

由表 3 可知，尽管仅考虑单模态特征可以在一定程度上减少模型的推理时间，但对虚假信息识别的效果影响显著。因此，每个模态的特征提取都是必要的。

5 局限和不足

本研究的局限性和不足如下：

1）在特征融合的过程中，未考虑不同模态特征的重要性程度。

2）采取的指标为准确率、精确率、召回率、F1 值、推理时间和模型参数，但评价模型压缩效果还包括功效、能耗、性能-功耗比和 FLOPs 等指标。

结束语 本研究针对多模态虚假信息检测任务，考虑资源受限场景的现实应用需求，提出了一种基于多模态大模型 Long-CLIP 的多模态虚假信息识别模型。该模型能够处理长文本，关注更多粗粒度和细粒度细节。同时，利用高效多粒度分层剪枝进行模型压缩，得到一个更加轻量化的多模态虚假信息识别模型，以适应资源受限场景。通过在微博数据集上进行实验，比较微调前后的当前流行的多模态大模型和其他剪枝方法的性能，并进行消融实验，验证了该模型及其各模块的有效性。结果显示，基于 Long-CLIP 的多模态虚假信息识别模型在模型参数和推理时间方面远少于当前流行的多模态大模型，且检测效果最佳。模型压缩后，在检测效果仅下降 0.01 的情况下，模型参数减少 50%，推理时间减少 1.92 s。

未来的研究工作将聚焦于构建更精准的多模态虚假信息识别大模型，并深入探索高性能模型压缩技术。首先，引入模态注意力机制增强模型对关键信息的捕捉能力，同时结合大模型的优势，利用检索增强技术与多任务学习策略，进一步提高对虚假信息的识别效果。此外，探索除剪枝外的其他模型压缩技术，如知识蒸馏，以在保持模型高性能的同时降低计算资源的消耗。还将考虑功效、能耗、性能-功耗比和 FLOPs 等因素，全面评估和优化模型压缩的效果。

参 考 文 献

[1] DUAN Y X, HU Y L, GUO H, et al. Research on improved cross-modal association ambiguity learning for fake news detection [J]. Computer Science, 2024, 51(4): 307-313.

[2] WANG J, WANG Y C, HUANG M J. Fake News in Social Networks: Definition, Detection, and Control [J]. Computer Science, 2021, 48(8): 263-277.

[3] LI Z Y, LI J. Fake news detection method based on multimodal attention network using contrastive learning [J]. China Science Papers, 2023, 18(11): 1192-1197.

[4] LIANG Y, TUO H T, AIMUDULA A. Multimodal fake news detection based on multi-layer CNN feature fusion and multi-classifier hybrid prediction [J]. Computer Engineering & Science, 2023, 45(6): 1087-1096.

[5] LAO A, ZHANG Q, SHI C, et al. Frequency spectrum is more effective for multimodal representation and fusion: a multimodal spectrum rumor detector [J]. arXiv: 2312.11023, 2023.

[6] YING Q, HU X, ZHOU Y, et al. Bootstrapping multi-view representations for fake news detection [C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2023, 37(4): 5384-5392.

[7] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision [C]// Proceedings of the International Conference on Machine Learning. PMLR, 2021: 8748-8763.

[8] LI J, LI D, XIONG C, et al. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation [C]// Proceedings of the International Conference on Machine Learning. PMLR, 2022: 12888-12900.

[9] LI J, LI D, SAVARESE S, et al. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models [C]// Proceedings of the International Conference on Machine Learning. PMLR, 2023: 19730-19742.

[10] ALAYRAC J B, DONAHUE J, LUC P, et al. Flamingo: a visual language model for few-shot learning [J]. Advances in Neural Information Processing Systems, 2022, 35: 23716-23736.

[11] ZHANG B, ZHANG P, DONG X, et al. Long-CLIP: unlocking the long-text capability of CLIP [J]. arXiv: 2403.15378, 2024.

[12] WEI J, TAY Y, BOMMASANI R, et al. Emergent abilities of large language models [J]. arXiv: 2206.07682, 2022.

[13] LU H, LIU W, ZHANG B, et al. DeepSeek-VL: towards real-world vision-language understanding [J]. arXiv: 2403.05525, 2024.

[14] YANG A, XIAO B, WANG B, et al. Baichuan 2: open large-scale language models [J]. arXiv: 2309.10305, 2023.

[15] BAI J, BAI S, YANG S, et al. Qwen-vl: a versatile vision-language model for understanding, localization, text reading, and beyond [J]. arXiv: 2308.12966, 2023.

[16] SMITH S, PATWARY M, NORICK B, et al. Using DeepSpeed and Megatron to train Megatron-Turing NLG 530B, a large-scale generative language model [J]. arXiv: 2201.11990, 2022.

[17] QI P, CAO J, LI X, et al. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues [C]// Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 1212-1220.

[18] VERMA P K, AGRAWAL P, AMORIM I, et al. WELFake: word embedding over linguistic features for fake news detection

[J]. IEEE Transactions on Computational Social Systems, 2021, 8(4): 881-893.
[19] SHU K, CUI L, WANG S, et al. Defend: explainable fake news detection[C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2019: 395-405.
[20] WU L, RAO Y, ZHANG C, et al. Category-controlled encoder-decoder for fake news detection [J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 35(2): 1242-1257.
[21] HU L, YANG T, ZHANG L, et al. Compare to the knowledge: graph neural fake news detection with external knowledge[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg: ACL, 2021: 754-763.
[22] ABDALI S, GURAV R, MENON S, et al. Identifying misinformation from website screenshots[C]// Proceedings of the International AAAI Conference on Web and Social Media. Palo Alto: AAAI, 2021: 2-13.
[23] QI P, BU Y, CAO J, et al. Fakesv: a multimodal benchmark with rich social context for fake news detection on short video platforms[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2023: 14444-14452.
[24] LI S, YAO T, LI S, et al. Semantic-enhanced multimodal fusion network for fake news detection [J]. International Journal of Intelligent Systems, 2022, 37(12): 12235-12251.
[25] BAZMI P, ASADPOUR M, SHAKERY A. Multi-view co-attention network for fake news detection by modeling topic-specific user and news source credibility [J]. Information Processing & Management, 2023, 60(1): 103146.
[26] DAVOUDI M, MOOSAVI M R, SADREDDINI M H. DSS: a hybrid deep model for fake news detection using propagation tree and stance network [J]. Expert Systems with Applications, 2022, 198: 116635.
[27] SINGHAL S, DHAWAN M, SHAH R R, et al. Inter-modality discordance for multimodal fake news detection[C]// Proceedings of the 3rd ACM International Conference on Multimedia in Asia. New York: ACM, 2021: 1-7.
[28] WU L, LIU P, ZHANG Y. See how you read? Multi-reading habits fusion reasoning for multimodal fake news detection [C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2023: 13736-13744.
[29] LI P, SUN X, YU H, et al. Entity-oriented multi-modal alignment and fusion network for fake news detection [J]. IEEE Transactions on Multimedia, 2022, 24: 3455-3468.
[30] LUVEMBE A M, LI W, LI S, et al. Dual emotion based fake news detection: a deep attention-weight update approach [J]. Information Processing & Management, 2023, 60(4): 103354.
[31] BOULAHIA S Y, AMAMRA A, MADI M R, et al. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition [J]. Machine Vision and Applications, 2021, 32(6): 121.
[32] MA Y, CAO Y, HONG Y C, et al. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! [J]. arXiv: 2303.08559, 2023.
[33] DRIESS D, XIA F, SAJJADI M S M, et al. Palm-e: an embodied multimodal language model [J]. arXiv: 2303.03378, 2023.
[34] ZHANG Q, ZUO S, LIANG C, et al. Platon: pruning large transformer models with upper confidence bound of weight importance[C]// Proceedings of the International Conference on Machine Learning. PMLR, 2022: 26809-26823.
[35] FANG G, MA X, SONG M, et al. DepGraph: towards any structural pruning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 16091-16101.
[36] MCCARLEY J S, CHAKRAVARTI R, SIL A. Structured pruning of a BERT-based question answering model [J]. arXiv: 1910.06360, 2019.
[37] SUN M, LIU Z, BAIR A, et al. A simple and effective pruning approach for large language models [J]. arXiv: 2306.11695, 2023.
[38] DAS A B, RAMAMOORTHY A. Coded sparse matrix computation schemes that leverage partial stragglers [J]. IEEE Transactions on Information Theory, 2022, 68(6): 4156-4181.
[39] FRANTAR E, ALISTARH D. SparseGPT: massive language models can be accurately pruned in one-shot[C]// Proceedings of the International Conference on Machine Learning. PMLR, 2023: 10323-10337.
[40] SHI D, TAO C, JIN Y, et al. UPOP: unified and progressive pruning for compressing vision-language transformers[C]// Proceedings of the International Conference on Machine Learning. PMLR, 2023: 31292-31311.
[41] SUNG Y L, YOON J, BANSAL M. EcoFLAP: efficient coarse-to-fine layer-wise pruning for vision-language models [J]. arXiv: 2310.02998, 2023.
[42] 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. New York: ACM, 2017: 795-816.

WU Chenglong, born in 2001, postgraduate, is a member of CCF (No. U8270G). His main research interests include fake news detection and model compression.

ZHAO Xiang, born in 1986, Ph.D, professor, Ph.D supervisor, is a member of CCF (No. 39960D). His main research interests include big data knowledge engineering and network content security.

(责任编辑: 李亚辉)