

西安交通大学

本科毕业设计（论文）

社交媒体多模态虚假信息检测技术研究

学院（部、中心）：电子与信息学部

专业：控制科学与工程

班级：自动化 96

学生姓名：梁珉珲

学号：2196213038

指导教师：史橿

2023 年 06 月

摘要

网络虚假信息具有传播速度快、危害程度大、隐蔽性强等特点，因此社交媒体虚假信息检测技术的研究具有重要意义。而随着深度学习技术的发展，多模态检测领域成为了学者研究的热点。然而，现有研究成果在社交媒体多模态虚假信息检测技术上还存在两方面不足：第一，各个模态数据进行融合时指导信息不够充分；第二，进行虚假信息检测时数据源不够丰富。针对这些问题，本文设计了一种以模态间相似度作为指导信息，以文本主题信息与情感信息作为辅助信息的基于深度学习的多模态虚假信息检测神经网络框架。该框架通过模态间相似度判断单模态信息与融合信息在分类时的重要程度，对虚假信息分类任务作出有效指导；与此同时，该框架融合主题信息与情感信息，它们作为虚假信息的重要特征，扩充了虚假信息分类任务的数据来源。另外，该框架使用预训练模型作为特征提取器，注意力机制进行模态融合，充分发挥了迁移学习的优势。本文在微博数据集中设计了横向比较实验与消融实验，证明了使用模态间相似度指导分类与引入主题信息、情感信息在多模态虚假信息检测任务下的有效性。

关 键 词：神经网络；多模态学习；虚假信息检测

ABSTRACT

Fake news on the Internet is characterized by fast propagation, high degree of harm, and strong concealment, so the research of social media fake news detection technology is of great significance. And with the development of deep learning technology, the field of multimodal learning has become a hot spot for scholars' research. However, there are two shortcomings in the existing research results on multimodal fake news detection technology for social media: first, the guidance information is not sufficient when each modal data is fused; second, the data sources are not rich enough for fake news detection. To address these problems, this thesis proposes a deep learning-based multimodal fake news detection neural network framework with cross-modal similarity as the guiding information and textual topic information and sentiment information as the supporting information. The framework uses cross-modal similarity to determine the importance of unimodal information and fused information in classification to effectively guide the fake news classification task; at the same time, the framework fuses topic information and sentiment information, which are important features of fake news, to expand the data sources for the disinformation classification task. In addition, the framework uses pretrained models as feature extractors and attention mechanisms for modality fusion to take full advantage of transfer learning. In this thesis, cross-sectional comparison experiments and ablation experiments are designed in the Weibo dataset to demonstrate the effectiveness of using cross-modal similarity to guide classification with the introduction of topic information and sentiment information under a multimodal fake news detection task.

KEY WORDS: Neural network; Multimodal learning; Fake news detection

目 录

| | |
|-----------------------------|----|
| 1 绪论 | 1 |
| 1.1 社交媒体虚假信息及其检测方法 | 1 |
| 1.2 多模态社交媒体虚假信息检测技术调研 | 1 |
| 1.3 本文采用方法 | 3 |
| 2 神经网络结构与训练 | 5 |
| 2.1 网络功能 | 5 |
| 2.2 网络结构 | 6 |
| 2.2.1 单模态特征提取 | 6 |
| 2.2.2 主题信息与情感信息获得 | 7 |
| 2.2.3 多模态特征融合 | 9 |
| 2.2.4 分步特征聚合 | 10 |
| 2.2.5 多任务训练输出与损失函数 | 10 |
| 2.3 训练细节 | 11 |
| 3 神经网络性能评估 | 13 |
| 3.1 实验设置 | 13 |
| 3.1.1 数据集与对照方法介绍 | 13 |
| 3.1.2 消融实验设置 | 14 |
| 3.2 实验结果 | 14 |
| 3.2.1 横向比较验证结果 | 14 |
| 3.2.2 消融实验结果 | 16 |
| 3.2.3 实验结果分析与总结 | 19 |
| 4 结论与展望 | 21 |
| 致 谢 | 23 |
| 参考文献 | 24 |
| 附录 A 外文文献原文 | 26 |
| 附录 B 外文文献译文 | 37 |
| 附录 C 部分源代码文件 | 47 |
| 附录 D 毕业设计（论文）任务书 | 59 |
| 附录 E 毕业设计（论文）考核评议书 | 61 |
| 附录 F 毕业设计（论文）评审意见书 | 62 |
| 附录 G | 63 |

1 绪论

1.1 社交媒体虚假信息及其检测方法

虚假信息即的错误的，与事实不符的或虚构的信息，主要指广义的虚假新闻，可以由任何人或机构创建并发布。其通过欺骗、误导公众提高社会影响力，从而达到某种目的。随着信息技术的高速发展，各类社交媒体平台逐渐成为人们浏览信息、分享生活的重要途径。然而，由于其操作的简易性与应用的广泛性，虚假信息很容易在社交媒体平台上编辑发布并吸引大量关注，使得政府、各类机构或独立个人名誉受损、财产流失，被视为对民主、言论自由的最大威胁之一，且对社会和谐稳定造成极为不良的影响^[1]。因此，社交媒体虚假信息检测技术成为了学者讨论与研究的热点，对经济与社会健康的发展具有深远的意义。

过往的虚假信息检测研究按照方法可分为基于用户分析、基于内容分析与基于社交上下文三类。其中基于用户分析的分析内容包括用户档案分析、用户活跃程度、发帖时间与行为习惯分析等。由于虚假信息需要在短时间内大规模传播，虚假信息的发帖用户以机器人为主，针对这一点，用户分析方法能够有效鉴别出发帖者是否是机器人，对虚假信息检测具有指向性意义。基于内容分析的方法包含对文本语言的分析、文本知识的分析与文本风格的分析三方面。其中对文本语言的分析主要包括检测文本是否存在前后矛盾与夸张表述；对文本知识的分析主要包括文本内容是否与事实不符；对文本风格的分析主要包括对新闻逻辑性、可读性与复杂性的分析。对文本内容的分析是对社交媒体信息真实性最直接的检测方法。基于社交上下文的分析主要指对于社交媒体信息传播模式的分析。传播模式即社交媒体用户的在线社交行为模式，包括评论模式、转发模式等。异常的信息传播模式往往暗示异常的信息内容。

目前流行的虚假信息检测技术主要是机器学习相关技术，又以深度学习技术为主。虚假信息检测任务可被归类为信息分类任务，而机器学习在信息分类任务中起着至关重要的作用，且操作简单，泛化性能好。然而，传统机器学习虚假信息检测方法需要手工设计特征，在这种方式下模型无法捕捉、挖掘虚假信息中隐藏的深度分类模式与固有特征，且耗费人力成本巨大，模型泛化能力差，对内容、形式复杂的社交媒体信息难以准确预测。故近年来发展迅速的深度学习技术凭借其优秀的特征提取能力、更佳的泛化能力、稳定性、鲁棒性逐渐成为虚假信息检测任务的主流应用技术。

1.2 多模态社交媒体虚假信息检测技术调研

近年来，深度学习技术取得了长足进步，成为学术界与工业界最为活跃的研究领域之一。其对非结构化数据强大的处理能力为自然语言处理、计算机视觉等领域提供了最先进的性能^[2]。因此，如何联合处理不同模态的数据，让模型从例如图像、文本、

语音等不同模态中提取信息，并通过模态融合方式克服其间的语义鸿沟，从而使模型像人类一样综合理解并应用各种模态的信息是深度学习的重要发展方向之一。而现有的多模态学习技术能够使不同模态间的信息相互补充，促进模态间信息的交互与融合，使机器学习到的信息更加立体且全面，从而提升机器在特定任务上的表现。而社交媒体上的信息通常涵盖多种模态，例如文字与配图，且各模态之间具有显著的相关性，例如社交媒体中分享生活的帖子多用文字描述照片内容，表达用户的心情与喜好。因此，使用多模态学习技术进行社交媒体虚假信息分类与检测很好地扩充了数据来源，并且各个模态间的数据能够在分类任务中互相补充与验证。

基于深度学习的多模态学习框架主要分为三个部分：特征提取部分、模态融合部分、任务输出部分。其中，特征提取部分一般由固定权重的预训练模型组成，呈多塔状结构。预训练模型一般较深，具有复杂的结构，并提前在大量训练数据上优化，故使用固定权重与参数的预训练模型能够在较小任务数据集上发挥模型优秀的特征提取能力，并降低了神经网络梯度反向传播的难度。在多模态模型中，作为编码器的预训练模型组对所有模态的数据分别进行特征提取，最终形成代表每一种模态的特征张量以待后续结构处理。而社交媒体虚假信息检测任务是一种特定领域分类任务，任务输出部分为前馈神经网络与分类头构成的分类器。模态融合部分是基于深度学习的多模态学习框架的核心部分，主要方法为围绕注意力机制（Attentional Mechanisms）^[3]进行信息融合，尽量消除不同模态信息间的语义鸿沟，并将其嵌入同一语义空间。在过去的研究中，许多致力于社交媒体虚假信息检测的工作围绕基于深度学习的多模态学习展开。Dhruv 等^[4]使用预训练模型为各个模态数据进行特征提取后，将所有不同模态的特征张量首尾相连作为 VAE（Variational auto-encoder）^[5]的输入张量，利用 VAE^[5]重建各个模态信息并优化编码器概率分布的训练过程促使模态融合，从而得到隐变量，令其作为分类器的输入对虚假信息进行检测。Chen 等^[6]进行模态融合时，利用对比学习方法设计子任务挖掘针对虚假信息检测任务的社交媒体信息各个模态之间的关系，以此辅助分类任务，并比较模态之间分布的距离并以此为权重指导跨模态信息交互与融合。Zhou 等^[7]利用 CLIP（Contrastive Language-Image Pretraining）^[8]提供跨模态信息并指导不同模态融合，从而完善融合后的用作分类的特征。然而，不同模态信息之间的联系对于有效指导模态融合的潜力尚未被完全发掘，不同模态信息之间的相似度、一致性对虚假信息检测的影响仍需进一步研究。

虚假信息检测任务依赖心理学、经济学、社会科学的相关分析与结论，这些理论与信息本身与用户两方面有关。在信息本身的角度而言，虚假信息的篇幅、情绪^[10]、写作风格^[10]可能与真实信息不同，为了使虚假信息显得更加真实，虚假信息中往往含有更高极性、更强烈的正面或负面情感情绪，且具有煽动性；在用户角度而言，用户个人信息肖像、用户发帖的意图以及帖子主题^[11]也是虚假信息检测方法的重要出发点，虚假信息通常针对某几类特定主题，如绯闻主题、社会主题、金融主题等。因此，在多模态学习的基础上引入主题信息与情感信息对虚假信息检测这一特定任务具有积极

有效的指导意义。过去曾有学者通过挖掘文本主题构建辅助模型指导文本分类任务。NTM (Neural Topic Model)^[12]被 Zhang 等^[13]利用进行多任务学习，通过推特平台用户浏览记录，同时提取浏览记录主题特征与文本序列特征预测其发帖标签，取得了极佳效果。而虚假信息检测任务的本质依然使分类任务，故主题信息对该任务的辅助效果值得研究。然而，很少有研究在社交媒体多模态虚假信息检测任务中引入主题信息、情感信息通过多任务学习辅助模型分类。

1.3 本文采用方法

本文设计了一种使用各模态信息之间相似度指导不同模态融合，引入文本主题信息与情感信息辅助模型分类的社交媒体多模态虚假信息检测深度神经网络模型。具体而言，本文使用固定参数的预训练模型 VIT (Vision Transformer)^[14]提取图像特征，使用固定参数的预训练模型 BERT (Bidirectional Encoder Representation from Transformers)^[15]提取文本特征；除此之外，本文使用 CLIP^[8]预训练模型中涵盖跨模态特征的图像编码器与文本编码器分别对图像与文本进行特征提取，作为单模态特征的补充，并获得混合信息，与此同时，CLIP^[8]将计算文本图片对之间的余弦相似度，作为指导模态融合的权重信息，分配融合特征中单模态信息与混合模态信息的比例，使模态融合过程具有指向性；被编码的文本信息、图片信息、混合信息经过投影层后维度被压缩并保持一致，通过跨模态注意力层初步融合后，利用跨模态相似度对其加权求和得到多模态特征。另外，本文采取多任务学习策略，利用文本信息，通过 TF-IDF (Term Frequency-Inverse Document Frequency)^[16]算法构建词袋模型，通过 VAE^[5]重建词袋向量的过程训练并优化编码器，使编码器得出的概率分步趋近于正态分布，利用采样技术得到隐变量，即主题特征；并利用情感分析 API 得到文本情感特征。此时，将所得的多模态特征、文本特征与情感特征按照分步聚合方法按步骤首尾相连，并使用多头自注意力机制 (Multi-Head Self Attention)^[17]对三种特征分步聚合^[18]，最终输入前馈神经网络与分类器，对信息进行分类检测，判别信息的真实性。

本文在 Weibo 数据集上设计了横向比较实验与纵向比较实验（消融实验），采用标准化准确性评估指标，即准确率，以虚假信息为研究对象的精确率、召回率、F1 值，以真实信息为研究对象的精确率、召回率、F1 值对模型效果进行比较与评价，并绘制混淆矩阵热力图。实验结果表明，使用 CLIP^[8]以及模态间相似度信息指导模型模态融合后，任务准确率提升了 4.6%；而引入主题信息与情感信息辅助模型分类后，任务准确率提升了 0.7%，且经过可视化处理后发现，模型自信程度与鲁棒性均得到较大幅度提升。

本文的所解决的问题主要有以下三个方面，即：

第一，验证了 CLIP^[8]预训练模型中含有跨模态特征的编码部分与跨模态相似度指导多模态特征融合的有效性；

第二，验证了文本主题信息、文本情感信息对社交媒体虚假信息检测任务作为辅

助数据来源的有效性；

第三，设计了一种准确性高、鲁棒性强的社交媒体虚假信息检测任务下的深度神经网络框架。

2 神经网络结构与训练

2.1 网络功能

本文设计的社交媒体多模态虚假信息检测模型，即 RDNN 模型（Rumor Detection Neural Network）分为三个部分，分别是特征提取部分、特征融合部分与分类输出部分，三部分之间为串行。其中，特征提取部分可分为单模态特征提取与主题信息、情感信息特征提取两个并行部分。令训练样本为三元组 $\mathbf{x} = (\mathbf{x}_{Txt}, \mathbf{x}_{Img}, \mathbf{y})$ ，其中 \mathbf{x}_{Txt} 是指文本数据， \mathbf{x}_{Img} 是指图像数据， \mathbf{y} 是指标签。则单模态特征提取过程可以表示为式 (2-1) - 式 (2-2)，其中 $x_{Feat-Txt}$ 、 $x_{Feat-Img}$ 分别指文字、图像特征， F_{Txt} 、 F_{Img} 分别指文字、图像提取器；而主题信息、情感信息特征提取可以表示为式(2-3)-(2-4)，其中 $x_{Feat-Topic}$ 是指主题特征， F_{Topic} 是指主题特征提取器。值得注意的是，主题特征提取器接受的输入为词袋向量，因此需要使用预处理函数 F_{Bow} 将文本数据转化为其词袋表示。对于情感特征，本文使用情感分析函数 F_{Emo} 作为特征提取器得到情感特征 $x_{Feat-Emo}$ 。

$$x_{Feat-Txt} = F_{Txt}(x_{Txt}) \quad (2-1)$$

$$x_{Feat-Img} = F_{Img}(x_{Img}) \quad (2-2)$$

$$x_{Feat-Topic} = F_{Topic}(F_{Bow}(x_{Txt})) \quad (2-3)$$

$$x_{Feat-Emo} = F_{Emo}(x_{Txt}) \quad (2-4)$$

特征融合部分可以分为多模态特征融合与分步特征融合两个串行部分。多模态特征融合部分利用跨模态融合函数 $F_{Feat-Multi}$ 对文本特征与图像特征进行聚合，得到多模态特征 $x_{Feat-Multi}$ ，如式 (2-5) 所示。接下来，本文使用分步特征融合方法，对于获得的多模态特征、主题特征与情感特征，先使用起始特征聚合函数 $F_{Feat-Fused-Start}$ 聚合主题特征与情感特征，得到起始聚合特征 $x_{Feat-Fused-Start}$ ，再使用终止特征聚合函数 $F_{Feat-Fused-End}$ 聚合起始聚合特征 $x_{Feat-Fused-Start}$ 与多模态特征，得到最终聚合特征 $x_{Feat-Fused-End}$ 。其过程如式 (2-6) - (2-7) 所示。

$$x_{Feat-Multi} = F_{Feat-Multi}(x_{Feat-Topic}, x_{Feat-Emo}) \quad (2-5)$$

$$x_{Feat-Fused-Start} = F_{Feat-Fused-Start}(x_{Feat-Topic}, x_{Feat-Emo}) \quad (2-6)$$

$$x_{Feat-Fused-End} = F_{Feat-Fused-End}(x_{Feat-Multi}, x_{Feat-Fused-Start}) \quad (2-7)$$

分类输出部分即对融合得到的特征进行投影与分类，其过程可以表示为式 (2-8)，其中 y_{pred} 为预测标签， F_{cls} 为分类函数。本文方法的整体结构由图 2-1 所示。

$$y_{pred} = F_{cls}(x_{Feat-Fused-End}) \quad (2-8)$$

本文的优化方向为，使用基于 CLIP^[8]预训练模型得到的跨模态相似度指导模态信息融合，得到多模态特征 $x_{Feat-Multi}$ ，然后将额外引入的主题特征 $x_{Feat-Topic}$ 与情感特征 $x_{Feat-Emo}$ 与多模态特征分步融合，得到可供分类的特征 $x_{Feat-Fused-End}$ ，以连接分

类器。其中，使用跨模态相似度指导模态信息融合能够有效抑制噪声，增强算法可解释性；使用主题特征与情感特征有效丰富了数据源，能够更好地指导分类任务。

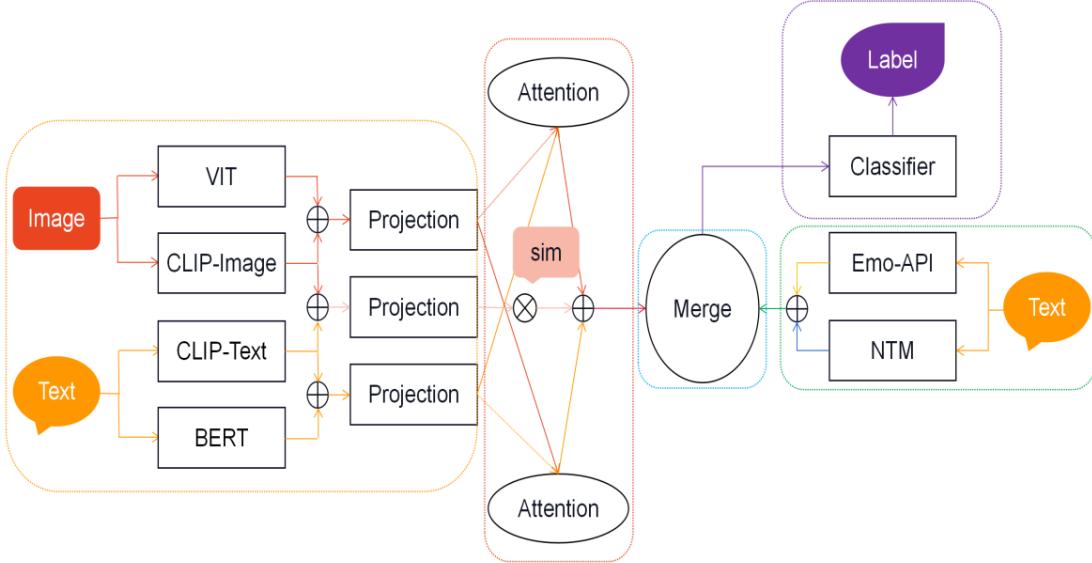


图 2-1 本文所设计神经网络整体结构图。

2.2 网络结构

2.2.1 单模态特征提取

本文使用 BERT^[15]对文本数据进行特征提取，其过程由式（2-9）所示。BERT^[15]是使用 Transformer^[17]模型的 Encoder 部分堆叠实现的语言表示模型。其用字符嵌入、分隔嵌入、位置嵌入编码文本，并在字符间插入特殊标记用于分割或分类。其使用掩码语言建模与下句预测的方式进行预训练，由于 Encoder 没有单向掩膜机制，它能够捕捉到文本中双向的深层次特征。与此同时，Transformer Encoder 中包含的多头自注意力机制有益于 BERT^[15]捕捉段落中较长距离的语义依赖关系。因此，BERT^[15]在自然语言理解领域的多项任务上取得了最高成绩。

$$x_{Feat-Txt-BERT} = BERT(x_{Txt}) \quad (2-9)$$

对于图像数据，本文使用 VIT^[14]进行特征提取，其过程由式（2-10）所示。VIT^[14]是第一个在 ImageNet^[19]数据集中成功训练的 Transformer^[17]，Encoder 类模型，且在大数据集上进行预训练的前提下，其效果超过最先进的 CNN (Convolutional Neural Network)^[20]图像检测与分类模型，克服了缺少归纳偏置的缺点。VIT^[14]在编码图像时，将图像像素拆分为等大小的补丁，并引入位置编码，使图像数据转换为序列数据，以供 Transformer Encoder 进行特征提取。

$$x_{Feat-Img-VIT} = VIT(x_{Img}) \quad (2-10)$$

除此之外，本文额外使用预训练 CLIP^[8]模型的文本数据特征提取器与图像数据特征提取器获得包含跨模态信息的文本与图像特征。CLIP^[8]模型是一个语言-视觉多模态模型，它通过对比学习方法在图片-文本对中构建正例与负例，扩充了数据来源。应用在分类任务上时，CLIP 不受类别数的局限，其通过计算文本-图像对余弦相似度并取最大值进行判别，且在预训练后能够给定文本信息后进行零样本图像分类。因此，在训练过程中，CLIP^[8]的文本数据特征提取器中包含图像模态信息，具有混合模态特征性质，反之亦然。其特征提取过程如式（2-11）-（2-12）所示。

$$x_{Feat-Txt-CLIP} = CLIP_{Txt}(x_{Txt}) \quad (2-11)$$

$$x_{Feat-Img-CLIP} = CLIP_{Img}(x_{Img}) \quad (2-12)$$

上述过程完成后，本文将文本特征 $x_{Feat-Txt-BERT}$ 与 $x_{Feat-Txt-CLIP}$ ，图像特征 $x_{Feat-Img-VIT}$ 与 $x_{Feat-Img-CLIP}$ ，混合特征 $x_{Feat-Txt-CLIP}$ 与 $x_{Feat-Img-CLIP}$ 分别首尾相连，并使用投影层对特征做进一步处理。投影层整合了对单一模态数据位于不同角度的特征，并使各模态特征在尺寸上保持一致。投影过程可表示为式（2-13）-（2-15），其中， cat 是指张量首尾连接操作， $Proj_{Txt}$ 、 $Proj_{Img}$ 、 $Proj_{Mix}$ 分别为文字、图像、混合特征投影头。投影头由全连接层、标准化层与 drop out 层组成，激活函数为 ReLU。其结构由图 2-2（左）所示。

$$x_{Feat-Txt} = Proj_{Txt}(cat(x_{Feat-Txt-BERT}, x_{Feat-Txt-CLIP})) \quad (2-13)$$

$$x_{Feat-Img} = Proj_{Img}(cat(x_{Feat-Img-VIT}, x_{Feat-Img-CLIP})) \quad (2-14)$$

$$x_{Feat-Mix} = Proj_{Mix}(cat(x_{Feat-Txt-CLIP}, x_{Feat-Img-CLIP})) \quad (2-15)$$

2.2.2 主题信息与情感信息获得

本文使用基于 VAE^[5]的 NTM^[12]模型提取文本主题特征，其结构如图 2-2（右）所示。在预处理阶段，本文使用 TF-IDF^[16]算法构建词袋模型，将文本 x_{Txt} 转换为词袋 x_{Bow} 作为 VAE^[5]的输入张量，接着利用编码器采样出隐变量 z 。TF-IDF^[16]算法能够突出一类文本中的特殊关键词，对模型学习分类特征模式具有促进作用，而此处隐变量 z 可视为文本主题特征。对文本重现概率 $p(x)$ 的求解将转化为对 z 的后验分布 $p(z|x)$ 的求解与对 x 的后验分布 $p(x|z)$ 的求解，而编码器投影的过程可视为后验分布 $p(z|x)$ 的求解过程。由于 z 的后验分布 $p(z|x)$ 难以求解，VAE^[5]使用变分方法求解变分后验分布 $q_\theta(z|x)$ ， θ 为编码器神经网络参数，进而求出文本主题特征分布，采样生成文本主题特征。由于对 $p(x)$ 做极大似然估计时，优化目标 $\max \sum p(x)$ 可拆分为下界 $ELBO$ 与 $KL(p(z|x)||q_\theta(z|x))$ 两部分如式（2-16）所示，故优化方向可以是使变分下界最大化，等价于最小化 x_{Bow} 与重建词袋 \hat{x}_{Bow} 的交叉熵损失，且最小化 $q_\theta(z|x_{Bow})$ 与先验分布 $p(z)$ 的 KL 散度，如式（2-17）所示。

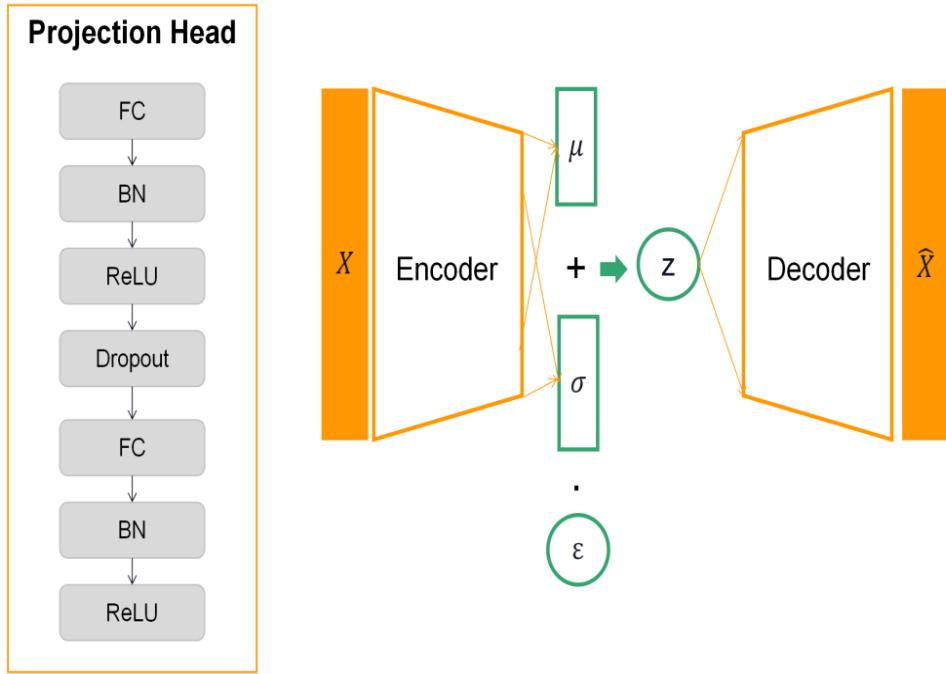


图 2-2 单模态特征提取投影头结构（左）与 NTM 模型结构（右）图

VAE^[5]假设其隐变量服从单位正态分布，即 $p(z) \sim N(0,1)$ 。由此可以通过最大化 $ELBO$ 的方法训练 VAE^[5] 得到 $q_\theta(z|x)$ ，同时 VAE^[5] 解码器的解码过程将被视为重建词袋，计算 \hat{x}_{Bow} 的过程，即求解 $p_\rho(x|z)$ 。而 z 可由 $q_\theta(z|x)$ 在 E 步之后近似服从单位正态分布的条件下由采样得到，采样过程由式 (2-18) - (2-19) 所示，其中 ε 为采样扰动变量，服从单位正态分布。 $x_{Feat-Topic}$ 即主题特征 z 。

$$L = ELBO + KL(p(z|x) || q_\theta(z|x)) \quad (2-16)$$

$$ELBO = -KL(p(z) || q_\theta(z|x)) + \int q(z|x) \log p(x|z) dz \quad (2-17)$$

$$q_\theta(z|x) \sim N(\mu(x), var(x)) \quad (2-18)$$

$$x_{Feat-Topic} = \mu(x) + \varepsilon \cdot \sqrt{var(x)} \quad (2-19)$$

$$x_{Feat-Emo} = Norm(EmoAPI(x_{Txt})) \quad (2-20)$$

对于情感特征，本文使用阿里云文本情感分析 API，将文本情感信息表征为向量。向量由积极分数、消极分数、中立分数、情感极性四维组成，并被规范化，分布在 (0, 1) 区间中。其中，情感极性被建模为离散值，分别由“0”，“0.5”，“1”表示。其计算方法由式 (2-20) 所示。

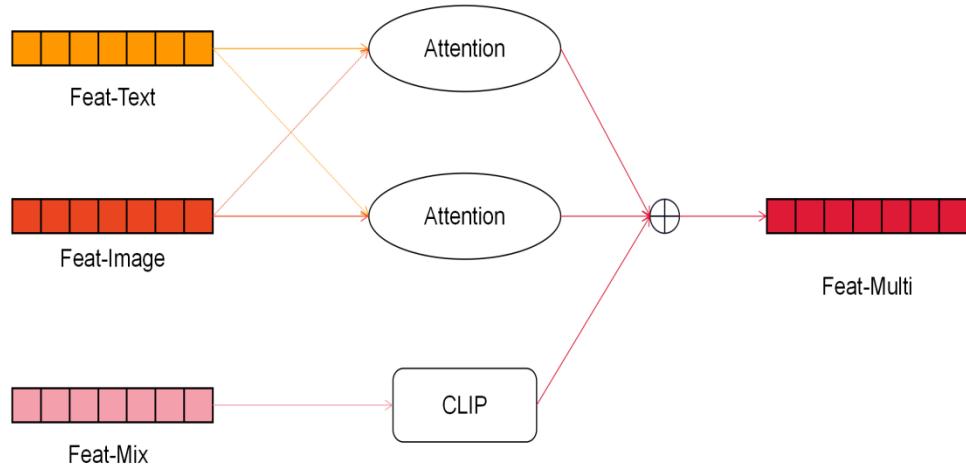


图 2-3 多模态特征融合结构图

2.2.3 多模态特征融合

由于文本特征与图像特征存在语义鸿沟，难以对齐，故本文首先使用交叉多头注意力机制促使文本特征与图像特征相互交融、包含。具体而言，对于文本特征 $x_{Feat-Txt}$ ，本文使用多头注意力机制，以图像特征 $x_{Feat-Img}$ 为查询键，计算其与 $x_{Feat-Txt}$ 的相关性，并使用 Softmax 将其归一化转换为概率分布，接着使用归一化后的相关性分数为权重对文本特征 $x_{Feat-Txt}$ 进行加权求和，得到含有图像信息的文本特征 $x_{Mix-Txt}$ 。对于图像特征，本文使用与上述过程中同样的操作，以文本特征 $x_{Feat-Txt}$ 为查询键经过加权求和获得包含文本特征的图像特征 $x_{Mix-Img}$ 。其计算如式 (2-21) - (2-22) 所示，其中 Q 指查询键 (Query)，K 指键值 (Key)，V 指值 (Value)。

$$x_{Mix-Txt} = \text{Attention}(Q:x_{Feat-Img}, K:x_{Feat-Txt}, V:x_{Feat-Txt}) \quad (2-21)$$

$$x_{Mix-Img} = \text{Attention}(Q:x_{Feat-Txt}, K:x_{Feat-Img}, V:x_{Feat-Img}) \quad (2-22)$$

在以往的工作中，不同模态之间的特征利用交叉多头注意力机制进行信息交融的过程即为多模态任务中模型进行模态特征融合的全部过程。这样的模态特征融合方法是有效的，但欠缺可解释的指导；与此同时，在该方法下形成的多模态特征可能会互相作为噪声，对分类任务进行干扰。因此，寻找解释性强的模态融合指导信息对模态融合部分至关重要。本文使用 CLIP^[8]中模态间余弦相似度作为指导信号，计算由 CLIP^[8]预训练模型文本、图像特征提取器产生的特征 $x_{Feat-Txt-CLIP}$ 、 $x_{Feat-Img-CLIP}$ 的余弦相似度 sim，并令其为混合特征权重，从而指导模态融合，得到加权混合特征 $x_{Weighted-Mix}$ ，其过程由式 (2-23) - (2-24) 所示。使用跨模态余弦相似度作为模态融合指导信息的

依据是，尽管图文之间的相似度与信息的真实性没有明显的关系，但图文之间的相似度正比于图文之间的相关性。若 sim 值高，则图文之间信息相关性高，那么不同模态之间的信息融合时互为噪声的成分小，故在分类时应该提高混合特征的比重；反之，若 sim 值低，则图文并不相关，那么若不对混合特征的比重加以抑制，则图文特征会互为噪声进行干扰，故应降低多模态特征中混合特征的比重，提高单模态特征比重。在模态融合的过程中， sim 值起到门控机制的作用，可用于动态调整混合特征的比例，从而突出混合特征或抑制噪声。得到加权的混合特征后，本文将其与通过投影头后同尺寸的文本特征 $x_{Feat-Txt}$ 、图像特征 $x_{Feat-Img}$ 相加，获得多模态聚合特征 $x_{Feat-Multi}$ ，该过程如图 2-3 与式 (2-25) 所示。

$$\text{sim} = \text{CLIP}(x_{Feat-Txt-CLIP}, x_{Feat-Img-CLIP}) \quad (2-23)$$

$$x_{Weighted-Mix} = \text{sim} \cdot x_{Feat-Mix} \quad (2-24)$$

$$x_{Feat-Multi} = x_{Feat-Txt} + x_{Feat-Img} + x_{Feat-Mix} \quad (2-25)$$

$$x_{Feat-Fused-Start} = \text{Attention}(\text{cat}(x_{Feat-Topic}, x_{Feat-Emo})) \quad (2-26)$$

$$x_{Feat-Fused-End} = \text{Attention}(\text{cat}(x_{Feat-Fused-Start}, x_{Feat-Multi})) \quad (2-27)$$

2.2.4 分步特征聚合

Karpathy^[18]等人设计，使用注意力机制进行特征融合时，慢速分步融合的效果优于直接融合。故本文采用慢速分步融合方法聚合多模态特征 $x_{Feat-Multi}$ 、主题特征 $x_{Feat-Topic}$ 与情感特征 $x_{Feat-Emo}$ 。考虑到 $x_{Feat-Topic}$ 与 $x_{Feat-Emo}$ 语义相近，本文首先聚合主题特征与情感特征。由于主题特征与情感特征之间尺寸相差较大，投影后可能遗漏关键信息或使信息过于稀疏，故本文首先将二者首尾相接，再使用自注意力机制^[17]使主题特征与情感特征对齐，生成初始融合特征 $x_{Feat-Fused-Start}$ 。同理，在第二步融合时，本文首先将初始融合特征 $x_{Feat-Fused-Start}$ 与多模态特征 $x_{Feat-Multi}$ 首尾相接，接着使用自注意力机制^[17]捕捉依赖关系并将其对齐，最终得到用于分类的聚合特征 $x_{Feat-Fused-End}$ 。上述过程由图 2-4、式 (2-26) - (2-27) 所示。

2.2.5 多任务训练输出与损失函数

虚假信息检测任务的实质是二分类任务，故获得聚合特征 $x_{Feat-Fused-End}$ 后，本文使用前馈神经网络对其进行压缩，并在末层后接 Softmax 层归一化以获得分类概率分布。该前馈神经网络结构与图 2-2 所示的投影头结构相同。

为了获得能够辅助检测任务的主题特征，本文采用多任务学习方法训练神经网络。2.2.2 节中提到，为了得到隐变量主题特征，优化网络结构时必须最小化 x_{Bow} 与重建词袋 \hat{x}_{Bow} 的交叉熵损失，且最小化 $q_\theta(z|x_{Bow})$ 与先验分布 $p(z)$ 的 KL 散度。与此同时，对于分类任务，交叉熵被用来度量概率分布与真实标签的距离。因此，本文设计了两个训练目标：最小化二分类交叉熵与最小化 NTM^[12] 损失。

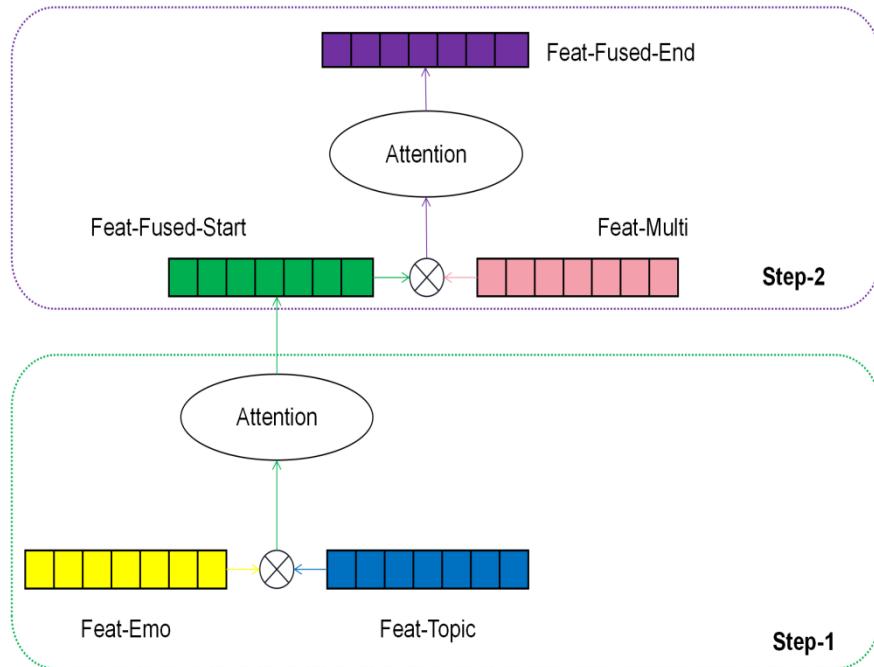


图 2-4 分步特征聚合结构图

两种损失分别由式 (2-28) 与式 (2-29) 表示, $E_*(\cdot)$ 指交叉熵, 其余符号与 2.2.2 节含义一致。

$$L_{cls} = E_{y-P(y)}(p'(\hat{y})) \quad (2-28)$$

$$L_{NTM} = E_{x-P(x)}(p'(\hat{x})) + KL(p(z)||q_\theta(z|x)) \quad (2-29)$$

为了得到可靠的训练结果从而提升检测任务的精度, L_{cls} 与 L_{NTM} 需要得到平衡, 否则在训练时会互相作为干扰项影响。本文使用加权的方法平衡各种任务的损失, 并通过求和得到本文设计方法的损失 L_{RD} 。其表达式如式 (2-30) 所示。其中 μ 为权重系数。

$$L_{RD} = L_{cls} + \mu \cdot L_{NTM} \quad (2-30)$$

2.3 训练细节

本文基于 Hugging Face Transformer 库构建预训练模型。对于 BERT^[15], 本文使用“BERT-base-chinese”模型, 设置文本长度最大值为 197, 将长度超出最大长度者截断, 长度不足最大值者补全; 对于 VIT^[14], 本文使用“google/vit-base-patch16-224-in21k”模型, 预处理图片时, 首先将其转化为 RGB 颜色模式, 再将其压缩并随机剪裁为 224×224 分辨率格式, 逐通道对图像标准化。针对三个通道, 本文使用的均值分别为 0.485、0.456、0.406, 标准差分别为 0.229、0.224、0.225。图像预处理后, VIT^[14] 处理器将每个图分割转化为分辨率为 16×16 的序列, 进入预训练 transformer 进行特征提取; 对于 CLIP^[8], 本文使用“OFA-Sys/chinese-clip-vit-base-patch16”模型。本文固定上述三种预训练模型的

参数，以保持其优秀的特征提取能力。投影头由双层前馈神经网络组成，特征数分别为 512、128。为了获得主题特征，利用 TF-IDF^[16]算法，文本被转化 40535 维词袋向量。NTM^[12]模型的编码器、解码器均由三层前馈神经网路构成，中间两层特征数分别为 768、256，隐变量 z 维度为 100。情感特征由“阿里云自然语言处理-情感分析” API 获得。本文多头自注意力^[17]层头数为 8。

本文其余训练参数如下：使用 Adam 优化器进行梯度下降，学习率为 0.001，训练轮数为 35，批次大小为 64。对于多任务训练损失函数 L_{RD} ，本文令 μ 值为 10^{-4} 。

3 神经网络性能评估

3.1 实验设置

3.1.1 数据集与对照方法介绍

本文采用 Weibo 数据集验证算法有效性。Weibo 数据集集合了微博社交平台中的用户发帖，是经典的虚假信息检测数据集，常被用于虚假信息检测与谣言甄别任务中。微博数据集中的一条数据由发帖 id 编码、发帖内容、发帖时间、用户名称、帖子地址与真实性标签组成。其中，发帖内容中包含文字，部分发帖内容包含文字与图片。本文对 Weibo 数据集进行数据清洗，根据 id 编码配对发帖内容中的文字与图片，过滤不含图片的帖子，再利用停用词词典过滤标点符号与非常用字符，提高数据质量，便于神经网络对多模态信息进行分析。清洗后获得可用数据 7848 条，其中虚假信息 4207 条，真实信息 3641 条。

为了说明本文所设计方法的有效性，本文从横向与纵向两个方面进行比对。对于横向比较，本文利用经典虚假信息检测方法在 Weibo 数据集上的最佳性能与本文方法进行比对，所采用方法如下所示。

EANN^[22]: 多模态虚假信息检测框架。其使用 Text-CNN (Text Convolutional Neural Network)^[21]作为文本特征提取器，VGG-19 (Visual Geometry Group)^[22]作为图像特征提取器，使用多任务学习方法，通过最大化主题分类损失提高模型泛化能力，并促使模态融合。

MVAE^[4]: 多模态虚假信息检测深度神经网络。其使用 Bi-LSTM (Bidirectional Long Short Term Memory)^[23]提取文本特征，VGG-19^[22]提取图像特征，通过多任务学习方法利用 VAE^[5]重建文本信息与图像信息迫使文本与图像模态融合并使用隐变量作为分类特征。

MCNN^[24]: 多模态虚假信息检测深度神经网络。其使用 BERT^[15]与 ResNet (Deep residual network)^[25]分别提取文本与视频特征，同时对文本信息、图像信息进行相似度检测，对图片进行篡改痕迹检测，将其作为辅助信息提高检测任务性能。

CAFE^[6]: 多模态虚假信息检测深度神经网络。其使用预训练模型作为编码器，对各个模态进行特征提取操作，并使用 VAE^[5]获得各个模态信息分布并计算其 KL 散度，利用 KL 散度指导模态融合。另外其设计辅助任务，利用真实信息集通过对比学习方法学习文本、图片相似度，从而辅助模态融合。

MKEMN^[26]: 多模态虚假信息检测深度神经网络。其使用 Bi-LSTM^[23]提取文本特征，使用 VGG-19^[22]提取图像特征，并且为了挖掘文本信息对于虚假信息检测的潜力，额外构建了知识编码模块，利用知识图谱中的事实知识辅助检测任务。

SAFE^[27]: 多模态虚假信息检测深度神经网络。其使用 Text-CNN^[21]作为文本特征提取器，而对于图像特征，其使用一个预训练图像-文本转化器将图像信息转化为文本信息，利用文本之间相似度指导模态融合并用于分类。

att-RNN^[28]: 多模态虚假信息检测深度神经网络。att-RNN^[28]是一个端到端的模型，其使用 LSTM^[23]融合文本信息与社交上下文信息，使用 VGG-19^[22]处理视觉信息，并使用注意力机制^[3]使其与文本、社交上下文复合特征融合，以此作为分类器输入特征。由于本文所设计的模型以及上述参与横向比较的模型均只含有图像、文本两种模态信息，故为消除其他影响因子，实验中除去 att-RNN^[28]中的社交上下文信息部分。

3.1.2 消融实验设置

本文通过设计、实施系列消融实验验证本文所设计方法的可行性，以及优化措施的有效性。消融实验分为三组：基线组，基线+CLIP 指导组，RDNN 模型（本文所设计模型）组。其结构分别如下所述。

基线组：相对于 RDNN 模型，移除 CLIP^[8]指导模态融合部分与主题特征、情感特征部分。文本特征与图像特征被提取后直接进行特征融合，并将融合后的多模态特征直接用于分类。

基线+CLIP 指导组：相对于 RDNN 模型，移除主题特征、情感特征提取与融合部分。保留 CLIP^[8]模型特征提取器与余弦相似度指导信号。文本特征与图像特征被提取后由跨模态余弦相似度指导模态融合，并将多模态特征用于分类。

RDNN 模型组：第二章中所陈述的方法。使用预训练模型进行特征提取，利用 CLIP^[8]模型的跨模态相似度计算方法指导模态融合，得到多模态特征后与主题特征、情感特征进行分步聚合，将聚合特征用于分类，并同时训练 NTM^[12]优化任务。

除模型结构外，模型训练硬件环境、训练设置、训练超参数、使用数据集完全相同，以达到控制变量的目的。

3.2 实验结果

本文使用准确率、精确率、召回率、F1 分数四种指标对社交媒体虚假信息检测结果进行评估。与此同时，对于消融实验，本文使用混淆矩阵热度图对三组模型进行可视化评估，使用 T-SNE (T-distributed Stochastic Neighbor Embedding)^[29]降维方法重现消融实验中用于分类特征的分布图，使模型特征提取与处理效果更加清晰、直观，同时能够判断特征降维后的大致分布。

3.2.1 横向比较验证结果

横向比较验证结果由表 3-1 所示。结果表明，对于标准化准确性评价指标，RDNN 整体准确率达到 87.7%，以虚假信息为研究对象时，精确率、召回率、F1 值分别为 84.3%，91.1%，87.6%；以真实信息作为研究对象时，精确率、召回率、F1 值分别为 91.4%，84.7%，87.9%。为使结果更加清晰，本文将指标以热力图形式呈现，如图 3-1 所示。

表 3-1 横向比较验证结果

| 方法 | 准确率 | 虚假信息 | | | 真实信息 | | |
|-------------------------|--------------|-------|--------------|--------------|--------------|-------|--------------|
| | | 精确率 | 召回率 | F1 值 | 精确率 | 召回率 | F1 值 |
| EANN ^[22] | 0.827 | 0.847 | 0.812 | 0.829 | 0.807 | 0.843 | 0.825 |
| MVAE ^[4] | 0.824 | 0.854 | 0.769 | 0.809 | 0.802 | 0.875 | 0.837 |
| MCNN ^[24] | 0.846 | 0.809 | 0.857 | 0.832 | 0.879 | 0.837 | 0.858 |
| CAFE ^[6] | 0.840 | 0.855 | 0.830 | 0.842 | 0.825 | 0.851 | 0.837 |
| MKEMN ^[26] | 0.814 | 0.823 | 0.799 | 0.812 | 0.897 | 0.851 | 0.873 |
| SAFE ^[27] | 0.816 | 0.818 | 0.815 | 0.817 | 0.816 | 0.818 | 0.817 |
| att-RNN ^[28] | 0.772 | 0.854 | 0.656 | 0.742 | 0.720 | 0.889 | 0.795 |
| RDNN | 0.877 | 0.843 | 0.911 | 0.876 | 0.914 | 0.847 | 0.879 |

与此同时, RDNN 的实验效果相比著名多模态模型的效果有了较大的提升。其中, RDNN 模型在 Weibo 数据集上的整体准确率、以虚假信息为研究对象时的召回率、F1 值、以真实信息为研究对象时的精确率、F1 值在本文所选择进行横向比较模型中最高。这充分说明了 RDNN 模型在虚假信息检测任务上的有效性。

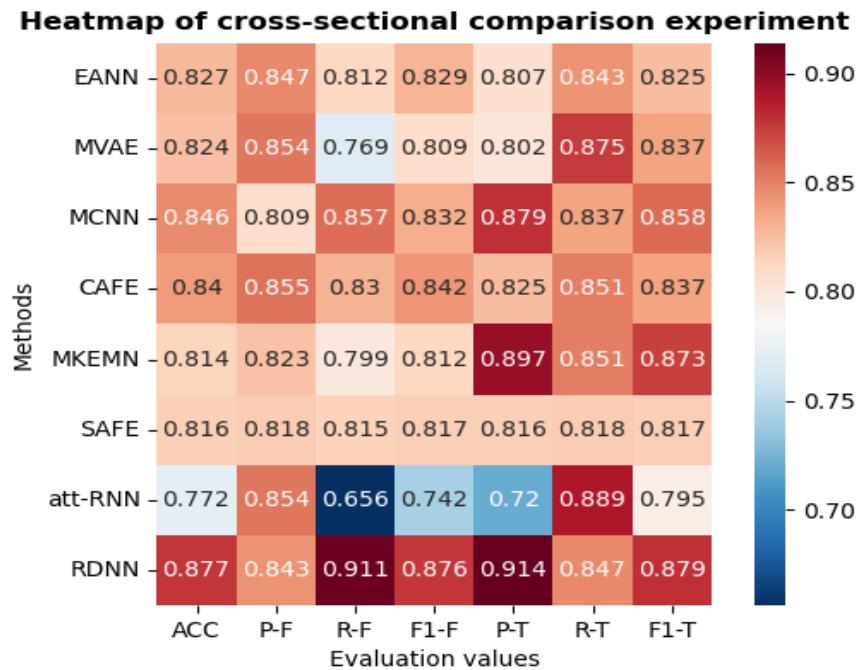


图 3-1 横向对比验证实验标准化评估指标热力图

在所有检测指标中, RDNN 模型针对虚假信息的召回率、针对真实信息的精确率相对横向比较模型提升最大, 分别为 91.1%与 91.4%。这说明 RDNN 模型几乎正确标注了所有虚假信息, 其不太可能错判社交媒体中的谣言, 而将其标注为非谣言。RDNN 模型针对真实信息的召回率相对较低, 为 84.3%, 低于 CAFE^[6]模型针对真实信息的召

回率 85.5%，这说明 RDNN 标注的虚假信息中存在一小部分真实信息，占比为 16% 左右。然而，对于社交媒体虚假信息检测任务，对虚假信息的召回率相较于对真实信息的召回率显得更加重要。对虚假信息的召回率较高，对真实信息的召回率较低的情况下，RDNN 标定的虚假信息可以通过后接检测器、人工评定等方式进行二次筛查，从而使错判的真实信息被剔除；但若模型虚假信息的召回率较低，则会遗漏大量虚假信息，而大量的虚假信息在社交媒体中流通会威胁网络生态与社会安定。

经过分析，RDNN 在虚假信息检测任务上具有良好效果的原因主要有以下三点：第一，RDNN 使用预训练大模型作为特征提取器并固定其内部参数，故尽管训练数据不多，模型的特征提取能力强且稳定，故 EANN^[22]、SAFE^[27]、att-RNN^[28] 虽结构与 RDNN 相似，且使用跨模态相似度指导模态融合或构建辅助任务提高模型泛化能力，但特征提取器结构较为简单，且参数需要优化，使神经网络深度较大，提升了训练难度，故性能弱于 RDNN。第二，RDNN 使用 CLIP^[8] 预训练模型计算模态间相似度指导特征融合，抑制了模态间的噪声，且过程清晰直接，故 MCNN^[24] 虽使用预训练模型提取特征并额外对图像进行篡改检测，MKEMN^[26] 虽引入知识图谱辅助神经网络进行虚假信息检测，但在模态融合部分仍有欠缺，性能弱于 RDNN。第三，RDNN 依据虚假信息的心理学、社会学特征，使用主题特征与情感特征作为额外信息设计多任务学习，辅助模型检测，故 CAFE^[6]、MVAE^[4] 虽使用预训练模型提取特征并利用 VAE^[5] 使各个模态信息得到充分融合，但其缺少额外信息源作为辅助，模型学习到的模式较为单调，故性能弱于 RDNN。

3.2.2 消融实验结果

纵向比较验证，即消融实验结果由表 3-2 所示。实验结果表明，RDNN 模型组的准确率，以虚假信息为研究对象时的精确率、F1 值，以真实信息为研究对象时的召回率、F1 值均为三组最高。基线模型（Baseline）+CLIP 组的以虚假信息为研究对象时的召回率，以真实信息为研究对象时的精确率均高于 RDNN 组，为三组最高，分别达到 91.3%、92.3%。基线组（Baseline）的各项指标均为三组最低。在准确性与 F1 值方面，基线模型（Baseline）+CLIP 组相对于基线组（Baseline）分别提升 3.8%、3.6%、3.6%；RDNN 组相对于基线模型（Baseline）+CLIP 组分别提升 1.7%、2.6%、2.6%。

表 3-2 消融实验结果

| 方法 | 准确率 | 虚假信息 | | | 真实信息 | | |
|---------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | | 精确率 | 召回率 | F1 值 | 精确率 | 召回率 | F1 值 |
| Baseline | 0.817 | 0.762 | 0.862 | 0.809 | 0.878 | 0.780 | 0.826 |
| Baseline+CLIP | 0.869 | 0.829 | 0.906 | 0.858 | 0.910 | 0.837 | 0.872 |
| RDNN | 0.877 | 0.843 | 0.911 | 0.876 | 0.914 | 0.847 | 0.879 |

由实验结果可以看出，使用 CLIP^[8] 预训练模型并利用文本与图像特征之间的余弦相似度指导模态融合能够使模型在虚假信息检测任务上的性能得到全方位的提升，同

时充分证明了本文方法的有效性。其中，以虚假信息为研究对象时的召回率以及以真实信息为研究对象时的精确率较为优秀，分别达到 90.6%、91.0%，且相对基线模型（Baseline）分别提升 4.4%、4.9%。这说明基线模型（Baseline）+CLIP 方法标注为真实信息的数据中绝大部分正确；与此同时，在所有的虚假信息中，基线模型（Baseline）+CLIP 方法仅错判不到 10% 的数据，即仅有很少的虚假数据会被模型遗漏。并且，基线模型（Baseline）+CLIP 方法以虚假信息为研究对象时的精确率以及以真实信息为研究对象时的召回率较基线方法（Baseline）提升较大，分别 6.7%、5.7%，但绝对值仍然较小，仅有 82.9%、83.7%。这说明虽然基线模型（Baseline）+CLIP 方法能够检测出大部分虚假信息，但其标注的虚假信息中含有一部分真实信息，占比约为 18%，模型偏斜程度较大。当模型遇到难以鉴别的信息时，其倾向于将其标注为虚假信息，故模型分别以虚假信息、真实信息为研究对象时 F1 值都相对不高，分别为 85.8%、87.2%。为使实验结果更加清晰直观，本文绘制消融实验标准化评估指标热力图，如图 3-2 所示。

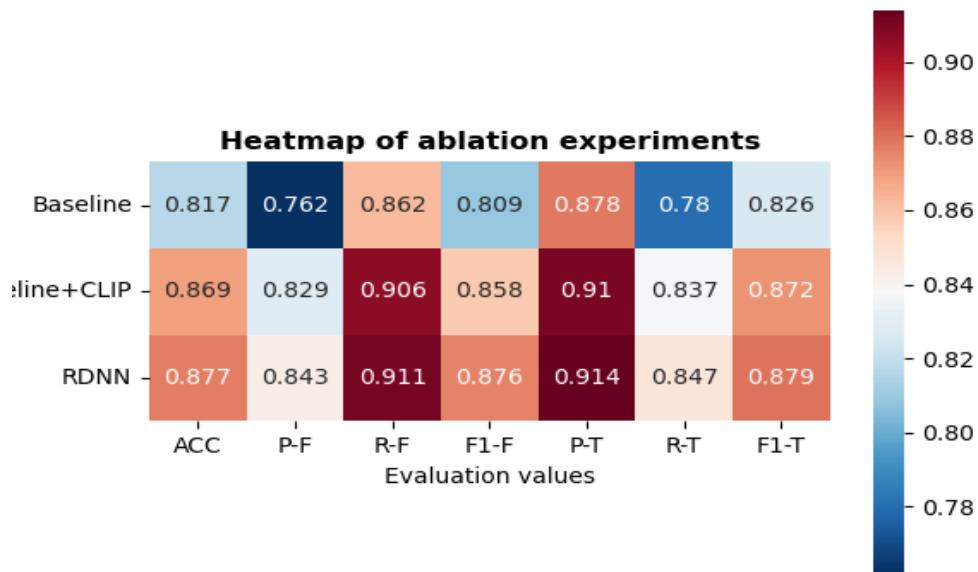


图 3-2 消融实验混淆标准化评估指标热力图

引入主题特征与情感特征后，RDNN 组大部分准确性指标得到提升，总体准确率相对基线模型（Baseline）+CLIP 组上涨 0.8%，并且在所有标准化准确性检验指标上高于基线模型（Baseline）+CLIP 组，因此可以证明，引入主题特征与情感特征在虚假信息检测任务上对模型是有效的。其中，RDNN 组以虚假信息为研究对象时的精确率和以真实信息为研究对象时的召回率相较基线模型（Baseline）+CLIP 组提升最大，分别为 1.4%、1.0%，说明模型偏斜程度得到了纠正。引入主题特征与情感特征后，数据源得到了扩充，在聚合的过程中模型用于分类的特征中包含的模式相较于基线模型

(Baseline) +CLIP 组变得更加丰富,因此模型在决策时能够进行细致的考虑,而新加入的特征有抑制偏斜的作用。与此同时, RDNN 组以虚假信息为研究对象时的召回率以及以真实信息为研究对象时的精确率更优于(Baseline)+CLIP 组,分别达到 91.1%、91.4%。由于 RDNN 模型克服了部分偏斜问题,整体表现相较基线模型(Baseline)+CLIP 组更佳,这体现在 F1 值的增长上: RDNN 组在两种研究情况下的 F1 值的绝对值分别为 87.6%与 87.9%,相较基线模型(Baseline)+CLIP 组分别上涨 1.8%、0.7%。为了使系列消融实验的结果更加直观,三组消融实验的混淆矩阵热力图如图 3-3 所示,由左至右分别为 RDNN 组、基线模型(Baseline)+CLIP 组、基线模型(Baseline)组,图中“0”代指真实信息(负类),“1”代指虚假信息(正类)。

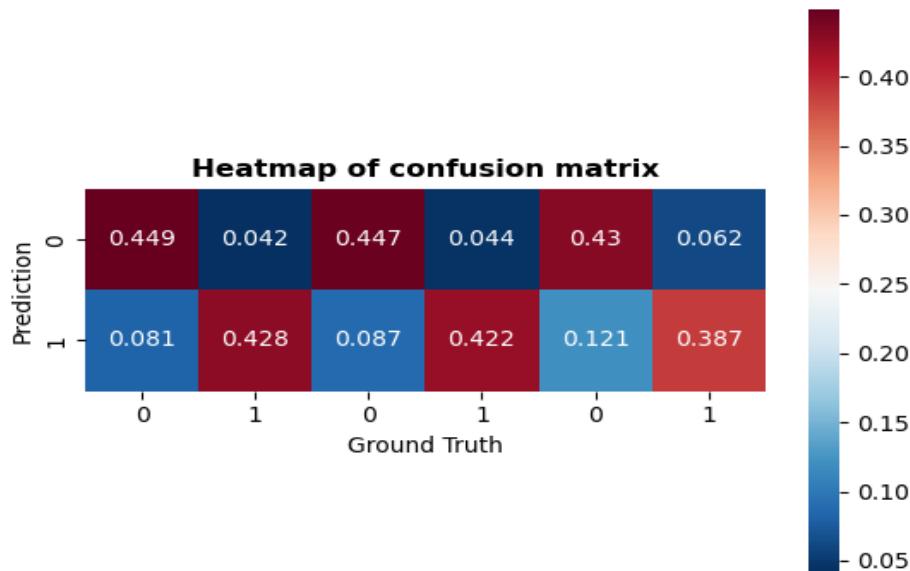


图 3-3 消融实验混淆矩阵热力图

图 3-4 展示了基线模型(Baseline)组、基线模型(Baseline)+CLIP 组、RDNN 组模型分类器之前最后隐藏层特征分布的 T-SNE^[29]降维可视化结果,按从左至右,从上至下顺序排列。图 3-2 表明,三组模型都能够较为有效地将虚假信息与真实信息分开,但基线模型(Baseline)组与基线模型(Baseline)+CLIP 组的最后隐藏层特征较为混乱,虚假信息特征与真实信息特征的投影具有重叠现象。虽然基线模型(Baseline)+CLIP 组中的两类特征中心分离度较优于基线模型(Baseline)组,且重叠部分相对较小,但仍为分类器寻找分类超平面带来较大的难度。而 RDNN 组模型的组后隐藏层特征按照类别几乎完全分离,仅存在微小重叠。且虚假信息特征与真实信息特征中心分离度大,模型分类置信度高,泛化能力与鲁棒性相比基线模型(Baseline)组与基线模型(Baseline)+CLIP 组均得到极大提升,为分类器提供了良好的分类前置条件。

最后隐藏层特征分布的可视化结果图充分体现了引入主题特征与情感特征后

RDNN 的优势，与其在标准指标上的表现互为补充。其不仅在准确度指标的表现上整体优于基线模型（Baseline）+CLIP 组与基线模型（Baseline）+CLIP 组，而且最后隐藏层特征分布更优。这说明通过引入额外的特征信息，模型能够接触的数据源更加丰富，对虚假信息与真实信息的模式与特征把握更加准确，故能够做出较为自信的判断，且检测准确性与置信度更高，并且，使用多任务方法训练优化的过程使模型对文本模态信息的理解更加深刻。与此同时，由于 RDNN 组特征提取与处理能力的提升，其稳定性与鲁棒性高于基线模型（Baseline）+CLIP 组与基线模型（Baseline）组。

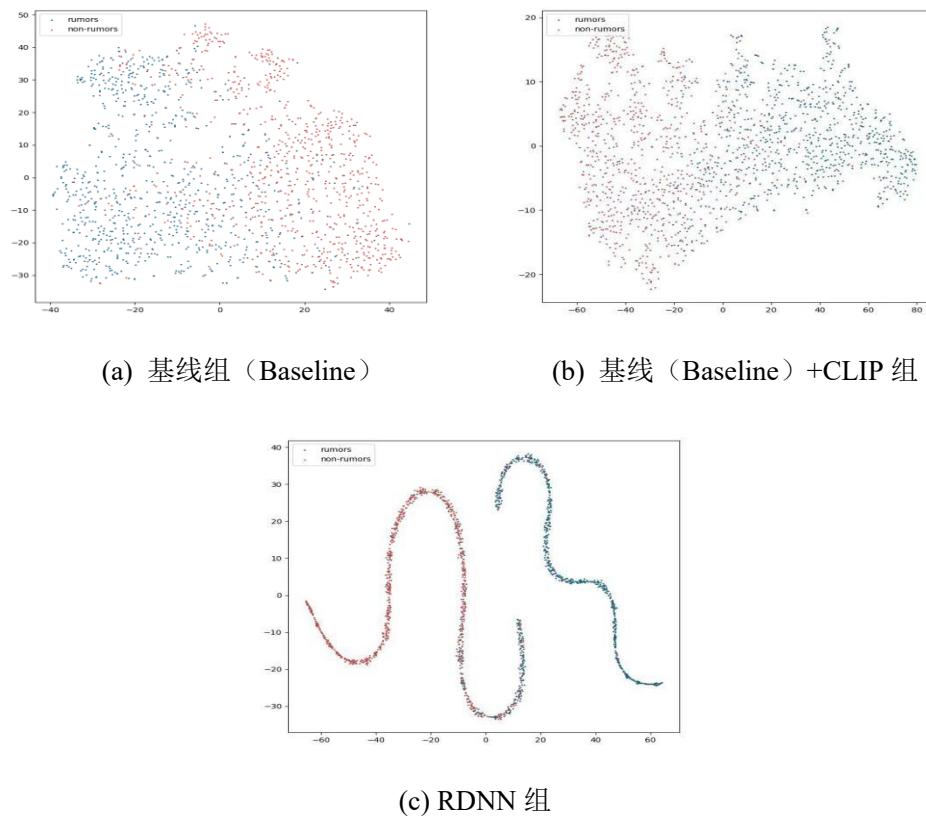


图 3-4 消融实验最后隐藏层特征二维分布图

3.2.3 实验结果分析与总结

本文在 Weibo 数据集上进行了大量的横向比较验证与纵向比较验证实验。横向比较验证指本文所设计的 RDNN 模型与著名多模态社交媒体虚假信息检测模型在统一数据集上就准确性指标进行比较。实验证明，RDNN 模型的准确率与 F1 值高于其他模型，同时对虚假信息的召回率较高。通过分析可知，RDNN 模型效果较优的原因是使用预训练模型作为特征提取器并固定参数权重、使用可解释的指导信号帮助模态融合并抑制噪声、使用主题特征与情感特征扩充信息源。纵向比较验证指消融实验。实验证明，本文所设计的两个优化方向，包括使用 CLIP^[8]预训练模型计算模态间余弦相似度指导模态融合与使用文本特征与情感特征辅助分类任务均有利于提升模型在虚假信息检测任务上的表现。使用 CLIP^[8]指导模态融合后，模型的各项标准化准确性指标得到了全面提升；引入主题特征与情感特征后，在标准化准确性指标上，模型的准确率与 F1 值

得到提升。除此之外，模型的偏斜在一定程度上被纠正，同时特征提取与处理能力得到提升，因此，模型的置信度、稳定性、鲁棒性、泛化能力得到优化。

4 结论与展望

本文设计了一种新的多模态社交媒体虚假信息检测深度神经网络结构，它分为单模态特征提取模块、主题与情感特征提取模块、多模态特征融合模块、分步特征聚合模块与分类输出模块。在单模态特征提取模块中，本文使用预训练模型作为特征提取器，分别提取文本特征与图像特征，同时利用 CLIP^[8]预训练特征提取器提取含有混合模态特征。在主题与情感特征提取模块中，本文利用 NTM^[12]重建文本并优化编码器概率分布采样得到主题特征，同时使用情感分析 API 得到情感特征。在多模态特征融合模块中，本文使用预训练 CLIP^[8]指导模态融合，利用模态间余弦相似度计算各模态信息相关性，对单模态特征与混合模态特征分配权重比例并抑制噪声。在分布特征聚合模块中，本文采用慢速融合方法，保证具有语义鸿沟的三种特征能够充分交融。在分类输出模块中，本文根据多任务学习方法设计损失函数，平衡各种任务的优化方向，利用前馈神经网络压缩特征，得到概率分布后输出分类结果。本文方法的优化方向在于，使用可解释性强的模态间余弦相似度指导模态融合并抑制噪声，以及引入主题特征与情感特征扩充信息源，构造多任务学习方法辅助虚假信息检测。本文在 Weibo 数据集上进行了大量实验，充分验证了所设计方法与优化方向在虚假信息检测任务上的有效性。在横向比较中，本文设计的模型 RDNN 的准确率与 F1 值高于其他著名多模态虚假信息检测模型，充分体现了预训练特征提取器、可解释性模态融合指导信号与额外信息源在多模态社交媒体虚假信息检测任务上的重要性。在纵向比较方面，本文设计了三组消融实验，分别为不含 CLIP^[8]预训练模型且不含额外信息源的基线模型（Baseline）组、含有 CLIP^[8]预训练模型但不含额外信息源的基线模型（Baseline）+CLIP 组、含有 CLIP^[8]预训练模型与额外信息源的 RDNN 组。消融实验表明，通过预训练 CLIP^[8]模型计算模态相似度，以此指导模态融合能够全面地，较大幅度地提升模型在标准化准确性评估指标上的表现；引入额外信息源，即主题特征与情感特征并设计多任务学习方法辅助虚假信息检测能够提升模型的分类准确率与 F1 值，同时部分克服模型偏斜问题。消融实验最后隐藏层特征二维分布图说明引入额外信息源后，模型捕捉虚假信息中利于鉴别的特定模式与固有特征的能力增强，且模型置信度、鲁棒性以及分类稳定性提高。综上所述，本文设计的多模态社交媒体虚假信息检测深度神经网络模型（RDNN）能够以较高的准确度与稳定性检测社交媒体中的虚假信息，与此同时，本文设计的优化方法在此任务上具有一定合理性与有效性。

未来工作的方向分为两个方面。第一方面为持续优化 RDNN，使其在准确性、稳定性、泛化性、鲁棒性上优于目前最先进的虚假信息检测模型。优化的方向有以下三个。第一，调整 RDNN 的超参数，包括训练超参数与结构超参数，提升模型的训练质量与检测能力；第二，调整模型结构，例如使用其他预训练模型作为特征提取器，例如 ResNet^[25]，或者使用其他信号指导模态融合，例如模态间分布的 KL 散度；第三，

引入除文本、图像外的其他模态特征，例如虚假信息在社交媒体上的传播图信息、社交上下文等，而图结构特征与文本、图像特征相比更难被提取，且与其他模态之间具有更深的语义鸿沟，因此引入图结构的多模态融合任务更加具有挑战性。第二方面为可解释方向研究。尽管 RDNN 模型在多模态社交媒体虚假信息检测任务上的效果已经超过了某些著名的或最新的模型，但其输出仍然是“真实”或“虚假”的二值化模式，这并不能解释 RDNN 模型判断该信息为虚假的原因，也不能标注出虚假信息中出现的典型可疑模式。另外，RDNN 模型得出最终结果之前的推理过程也是完全未知的，可能 RDNN 通过训练能够捕捉目前尚未发现的虚假新闻的特定模式或固有特征，但其无法被利用指导虚假信息与谣言在语言学、社会学、心理学或传播学方向的研究。因此，多模态虚假信息检测模型的可解释性研究是十分重要且有价值的。

致 谢

写到此处，本人感慨万千，借完成本科生毕业设计的机会，想要对成长过程中给予我支持与帮助的老师、同学、家人们表达感谢：感谢史梔老师、郑庆华老师对本人毕设从选题、研究计划、实验到论文撰写过程的辛勤指导；感谢罗敏楠老师对本人毕设研究方法、模型选取、训练技巧方面的答疑解惑；感谢周德润同学对实验环境的支持帮助；感谢父母在物质与精神方面对本人的供给教导；感谢自动化科学与工程学院四年来对本人在学习态度、专业知识、学习习惯与科研素养方面的培养；感谢西安交通大学为本人提供的优美安静的学习环境，以及对本人生活上的照顾；感谢党和国家给予我学习知识、磨练技术、提高与完善自己的机会。本人在大学的四年间学习刻苦用功，乐于提问，没有荒废时间，最终收获良多。本人希望自己能够再接再厉，未来成长为对社会与国家有用的人，效仿先贤，为国家科技的发展做出贡献，奉献自己，使母校与自动化学院以本人为骄傲。

参考文献

- [1] Shu K, Wang S, Liu H. Beyond news contents: The role of social context for fake news detection[C]. Proceedings of the twelfth ACM international conference on web search and data mining. 2019: 312-320.
- [2] Ramachandram D, Taylor GW. Deep multimodal learning: A survey on recent advances and trends[J]. IEEE signal processing magazine, 2017, 34(6): 96-108.
- [3] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.
- [4] Khattar D, Goud JS, Gupta M, et al. Mvae: Multimodal variational autoencoder for fake news detection[C]. The world wide web conference. 2019: 2915-2921.
- [5] Kingma DP, Welling M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.
- [6] Chen Y, Li D, Zhang P, et al. Cross-modal ambiguity learning for multimodal fake news detection[C]. Proceedings of the ACM Web Conference 2022. 2022: 2897-2905.
- [7] Zhou Y, Ying Q, Qian Z, et al. Multimodal Fake News Detection via CLIP-Guided Learning[J]. arXiv preprint arXiv:2205.14304, 2022.
- [8] Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]. International conference on machine learning. PMLR, 2021: 8748-8763.
- [9] 地力夏提·阿布都热依木, 马博, 杨雅婷等. 基于注意力机制多特征融合的虚假信息检测 [J]. 厦门大学学报(自然科学版), 2022, 61(04): 608-616.
- [10] McCornack SA, Morrison K, Paik JE, et al. Information manipulation theory 2: A propositional theory of deceptive discourse production[J]. Journal of Language and Social Psychology, 2014, 33(4): 348-377.
- [11] Zhang X, Ghorbani AA. An overview of online fake news: Characterization, detection, and discussion[J]. Information Processing & Management, 2020, 57(2): 102025.
- [12] Miao Y, Grefenstette E, Blunsom P. Discovering discrete latent topics with neural variational inference[C]. International Conference on Machine Learning. PMLR, 2017: 2410-2419.
- [13] Zhang Y, Zhang Y, Xu C, et al. # HowYouTagTweets: Learning User Hashtagging Preferences via Personalized Topic Attention[C]. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 7811-7820.
- [14] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.
- [15] Devlin J, Chang MW, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
- [16] Sparck JK. A statistical interpretation of term specificity and its application in retrieval[J]. Journal of documentation, 1972, 28(1): 11-21.
- [17] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
- [18] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks[C]. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2014: 1725-1732.
- [19] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]. 2009

- IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.
- [20] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
- [21] Kim Y . Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv, 2014.
- [22] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
- [23] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
- [24] Xue J, Wang Y, Tian Y, et al. Detecting fake news by exploring the consistency of multimodal data[J]. Information Processing & Management, 2021, 58(5): 102610.
- [25] 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.
- [26] Zhang H, Fang Q, Qian S, et al. Multi-modal knowledge-aware event memory network for social media rumor detection[C]. Proceedings of the 27th ACM international conference on multimedia. 2019: 1942-1951.
- [27] Zhou X, Wu J, Zafarani R. Similarity-Aware Multi-modal Fake News Detection[C]. Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11 – 14, 2020, Proceedings, Part II. Cham: Springer International Publishing, 2020: 354-367.
- [28] 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.
- [29] Van der Maaten L, Hinton G. Visualizing data using t-SNE[J]. Journal of machine learning research, 2008, 9(11).

附录 A 外文文献原文

Multimodal Fake News Detection via CLIP-Guided Learning

Yangming Zhou
School of Computer Science
Shanghai, China
ymzhou21@fudan.edu.cn

Qichao Ying
School of Computer Science
Shanghai, China
qcying20@fudan.edu.cn

Zhenxing Qian*
School of Computer Science
Shanghai, China
zxqian@fudan.edu.cn

Sheng Li
School of Computer Science
Shanghai, China
lisheng@fudan.edu.cn

Xinpeng Zhang
School of Computer Science
Shanghai, China
zhangxinpeng@fudan.edu.cn

ABSTRACT

Multimodal fake news detection has attracted many research interests in social forensics. Many existing approaches introduce tailored attention mechanisms to guide the fusion of unimodal features. However, how the similarity of these features is calculated and how it will affect the decision-making process in FND are still open questions. Besides, the potential of pretrained multimodal feature learning models in fake news detection has not been well exploited. This paper proposes a FND-CLIP framework, i.e., a multimodal Fake News Detection network based on Contrastive Language-Image Pretraining (CLIP). Given a targeted multimodal news, we extract the deep representations from the image and text using a ResNet-based encoder, a BERT-based encoder and two pairwise CLIP encoders. The multimodal feature is a concatenation of the CLIP-generated features weighted by the standardized cross-modal similarity of the two modalities. The extracted features are further processed for redundancy reduction before feeding them into the final classifier. We introduce a modality-wise attention module to adaptively reweight and aggregate the features. We have conducted extensive experiments on typical fake news datasets. The results indicate that the proposed framework has a better capability in mining crucial features for fake news detection. The proposed FND-CLIP can achieve better performances than previous works, i.e., **0.7%**, **6.8%** and **1.3%** improvements in overall accuracy on Weibo, Politifact and Gossipcop, respectively. Besides, we justify that CLIP-based learning can allow better flexibility on multimodal feature selection.

CCS CONCEPTS

- Computer systems organization → Embedded systems; Redundancy; Robotics;
- Networks → Network reliability.

*Corresponding author. This work is supported by National Natural Science Foundation of China under Grant U20B2051, U1936214.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06.. \$15.00
<https://doi.org/XXXXXXX.XXXXXXXX>

KEYWORDS

datasets, neural networks, gaze detection, text tagging

ACM Reference Format:

Yangming Zhou, Qichao Ying, Zhenxing Qian, Sheng Li, and Xinpeng Zhang. 2018. Multimodal Fake News Detection via CLIP-Guided Learning. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 11 pages. <https://doi.org/XXXXXXX.XXXXXXXX>

1 INTRODUCTION

Online social networks have largely replaced the conventional way of information communication represented by newspapers and magazines. People enjoy the convenience of online social media in seeking friends or sharing viewpoints. However, OSNs have also promoted the wide and rapid spreading of fake news [24, 33, 49]. Online news posts can be more easily manipulated compared to written materials. News forgery can take various forms, for example, replacing a critical object within a picture with another one, or making biased or even misleading comments on the picture. What's worse, the readers are susceptible to well-crafted fake news and further circulate them. In sum, fake news is likely to create panic and misdirect public opinion, which alters society in negative ways.

In the past decades, Fake News Detection (FND) has been the center of data-centric research for decades [1, 36, 45]. While manual observation towards all news and posts on the Internet is both expensive and time-consuming, automatic FND using machine learning is an efficient way to combat the widespread dissemination of fake news. FND has helped news readers identify bias and misinformation in news articles and therefore stop their spreading. Early works on fake news detection merely focused on text-only or image-only content analysis [5, 11]. A pretrained model is usually employed to verify the logical and semantic soundness of the input. Also, trivial clues such as grammatical errors or traces left by image manipulation might be taken into consideration. While unimodal FND schemes are effective, modern news and posts are usually with rich information of several modalities and these methods neglect their correlation. For some fake news, a real image can be combined with total rumors and correct words can be used to describe a tampered image. In that sense, multimodal feature analysis is required to offer complementary benefits to assist FND.

In recent years, there have already been a lot of works that aggregate multimodal features to detect anomalies in news and posts [9, 24, 42]. Besides fusing features from images and texts,



Figure 1: Examples of fake news detection using FND-CLIP. The three attention scores of each news are text score, image score, and fusion score respectively.

comments, up-vote ratio and the spreading graph are mostly preferred by researchers to evaluate the truthfulness of a post. These additional modalities are interactive and change over time. Many previous works prefer using as much modalities as possible. However, interactive modalities are less dependable than images and texts, or static modalities. First, the absence of interactive modalities might be common. A typical example is that no clue can be left in historical behavior in news posted by newly registered users, and nor can it be left in comments or up-votes if we wish to reject fake news shortly after their submission. Second, interactive modalities are less stable and can be changed over time, therefore potentially resulting in varied forensics results. Therefore, we revisit current arts in FND with only static modalities, and find that though many algorithms design well-crafted networks for multimodal feature fusion [38, 42], the mechanisms are largely at a black-box level as to how multimodal features will influence the final decision. Some works try to address this issue by explicitly calculating correlation on generating fused features. For example,

Chen et al. [9] additionally train variational auto encoders (VAE) that first compress the images and texts and contrastively learns to minimize the Kullback-Leibler (KL) divergence for news with correct image-text pairs. The corresponding cross-modal ambiguity score is then used to reweight the multimodal features [9]. Dhruv et al. [24] propose MVAE that trains a decoder to reconstruct the original texts and low-level image features from the fused features. These methods have achieved decent performance in multimodal fake news detection.

However, there are still some issues for multimodal FND to be addressed. First, we find that how the similarity of features from different modalities is to be calculated and how it will affect the decision-making process in FND is still an open question. For [9], we are not sure how efficient the VAEs are so that the KL divergence will be small given matched image-text pairs. For MVAE, though the ability of reconstruction means that the fused features are able to contain more information, the necessity of these auxiliary tasks in the view of FND remains unknown. Besides, we find that more advanced multimodal learning paradigms and pretrained models are not properly applied in FND. For example, CLIP [34] is a multimodal model that combines knowledge of language concepts with semantic knowledge of images. It was trained on a variety of image-text pairs to predict the most relevant text snippet, given an image, and vice versa. CLIP, together with other advanced multimodal technologies can be beneficial in image-text feature fusing, yet their usages in FND still remain ill-posed.

This paper proposes FND-CLIP, a multimodal fake news detection network based on the pretrained Contrastive Language-Image Pretraining (CLIP) model. The CLIP-based learning for fake news detection is to address the issue of cross-modal ambiguity by explicitly measuring the correlation between texts and images of targeted posts, and to guide the feature fusing and decision-making stages. Specifically, we encode the image using a fine-tunable ResNet [16] encoder a pretrained CLIP image encoder. The text is encoded by a fine-tunable BERT [13] encoder as well as a CLIP text encoder. The unimodal features are generated by concatenating the CLIP-generated features with the fine-tunable counterparts. The fused features consist of the two CLIP outputs. We use three projection heads to individually process the unimodal and fused features, which shrinks their sizes in order to distill the most important features for FND. Besides, we calculate the cosine similarity on the CLIP outputs and standardize it as the cross-modal similarity score. The score reweights the fused feature, where we regulate that less information will be provided by the fused features if the image and text show low correlation. Furthermore, we introduce an attention layer that outputs three scores that adaptively measure the significance of these features in their contribution to fake news detection. The classifier finally processes the summarized features to distinguish fake news from real ones.

We have conducted extensive experiments on FND-CLIP on several typical datasets for FND, including a Chinese dataset named Weibo, and two English datasets named Politifact and Gossip. The results show that FND-CLIP achieves **0.7%**, **6.8%** and **1.3%** performance improvement in overall accuracy on the three datasets. Besides, we justify that CLIP-based learning can allow better flexibility on multimodal feature selection. Figure 1 showcases four examples of fake news detection using FND-CLIP, where we see that

the attention score as well as cross-modal similarity vary among different news instances. FND-CLIP is able to pay less attention to the multimodal features when the similarity is low, therefore flexibly aggregating information according to the characteristics of the provided news.

The contributions of this paper are mainly three-folded, namely:

- We propose FND-CLIP, a multimodal fake news detection method with CLIP-based learning, where the CLIP pretrained model is used to measure the cross-modal similarity and guide the mapping and fusion of features.
- We propose a modality-wise attention mechanism to adaptively weight the text, image, and fused features. Given different news instances, we find that the model flexibly learns to pay more attention to useful information in unimodal or multimodal features.
- We have conducted comprehensive experiments on three famous datasets, where the results prove that CLIP-generated features can be important assists to the unimodal features. FND-CLIP outperforms state-of-the-art fake news detection methods.

2 RELATED WORKS

2.1 Unimodal Fake News Detection

Unimodal FND usually works on finding anomalies in either the text or the image of a post. These algorithms often follow the essence of human decision process. For images, Cao et al. [7] jointly study image forensics features, semantic features, statistical features, and context features for fake news detection. It suggests that typical methods for image manipulation detection [8] are useful in unveiling traces for news tampering. Besides, semantic inconsistency regarding the common sense [26] as well as poor image quality [15] can be widely present in fake news. For texts, verifying the logical soundness is essential [14], also accompanied by finding clues such as grammatical errors, writing styles [31] or extracting rhetorical structure [11]. Besides, both linguistic and visual patterns can be highly dependent on specific events and corresponding domain knowledge. Therefore, Nan et al. [29] propose to employ domain gate to aggregate multiple representations extracted by mixture-of-experts, and it deals with multi-domain fake news propagation in the language modality.

Though these unimodal characteristics can be explored and they indeed play key roles in distinguishing fake news, the multimodal characteristics such as correlation and consistency are ignored, which potentially impair the overall performance of these unimodal schemes on multimodal news.

2.2 Multimodal Fake News Detection

In the past literature, many works have been done on mining useful representations from images and texts of the news for fake news detection. Earlier works design sophisticated yet black-box attention mechanisms for multimodal feature fusion [3, 5]. Many other works [9, 24, 42] propose to better align the extracted features from different modalities before sending them into the classifier. Wang et al. [42] propose EANN that further employs an auxiliary task of event classification to aid feature extraction. The event classification branch is designed to better disentangle the mined multimodal

features so that there are both event-specific information and event-agnostic information. Dhruv et al. [24] process the image and text using unimodal feature extractors and further utilize a multimodal VAE to learn a shared representation from them. The sampled representation produced by the VAE is then sent to a decoder which tries to reconstruct the original texts and low-level image features. Besides the focus on network design, other works exploit more information from the datasets. For example, Qi et al. [32] claim that image feature extractors cannot well understand visual entities such as celebrities, landmarks, and texts within the images, and therefore propose to manually extract these kinds of information as linguistic assists. Zhang et al. [47] design a novel dual emotion feature descriptor to measure the emotional gap between the post and its comments and verify that dual emotion is distinctive between fake and real news. Chen et al. [9] use two VAEs to compress the images and texts and contrastively learn to minimize the Kullback-Leibler (KL) divergence for correctly matched image-text pairs. The resultant score is then used to reweight the multimodal features during feature fusing.

Though these methods have achieved decent performance in multimodal FND, there are still issues to be concerned. First, how to explicitly measure the correlation between images and texts within a post still remain unclear. Second, we see that little work in FND consider applying the recently emerged arts in multimodal learning, which motivates us to use the CLIP-based pretraining to further boost the performance.

2.3 Multimodal Learning

Recent years have shown rapid developments in the field of multimodal machine learning [2]. Neural architectures are employed in tasks that go beyond single modalities, for example, Visual Question Answering (VQA) [12], Visual Commonsense Reasoning (VCR) [46], etc. In these tasks and beyond, priors and features from different modalities are required and algorithms or deep networks cannot be effective when provided with only a single modality. Several generic technologies are developed for learning joint representations of image content and natural language. For example, the CLIP model [34] is designed as a bridge between computer vision and natural language processing. It was trained on a variety of image-text pairs to predict the most relevant text snippet, given an image, without directly optimizing for the task. The model consists of two encoders that respectively embed texts and images into a uniform mathematical space. Then, for the matched image-text pair, CLIP is encouraged to maximize the cosine similarity between the embedding of the two modalities. Otherwise, the similarity is minimized for the model to find the most suitable paired images and texts. Multimodal learning has a promising future where the innovation of CLIP has benefitted a lot of down-stream tasks [10, 43]. Other multimodal schemes can be represented by Glide [30] and VilBERT [28] that are respectively for text-to-image generation and multimodal representation learning.

3 METHOD

3.1 Approach Overview

For multimodal fake news detection, we collect the static modalities of the sampled news that includes text and image, and denote each

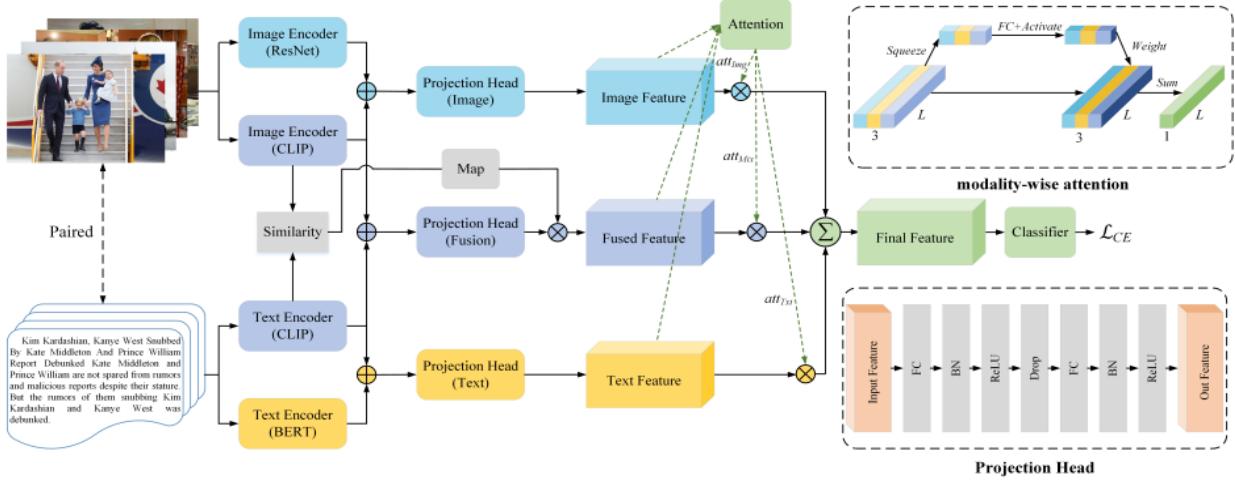


Figure 2: The architecture of the proposed FND-CLIP method. CLIP, BERT, and ResNet are used to extract the features of different modalities of multimodal news. Encoded features of different levels are obtained through projection heads. CLIP similarity score is calculated to determine the importance of fused feature. A modality-wise attention mechanism is further used to reweight different modal features adaptively for the classifier to classify fake news.

sample as $\mathbf{x} = (\mathbf{x}_{Txt}, \mathbf{x}_{Img})$. The ground-truth label is y where $y = 0$ indicates that \mathbf{x} is a real news, otherwise $y = 1$. According to the most traditional multimodal learning paradigm, a rich set of features are first extracted from \mathbf{x}_{Txt} and \mathbf{x}_{Img} that both represents the unimodal characteristics and the multimodal characteristics, which are then further fused and projected into a single value of \hat{y} that should be close to the ground truth.

$$\hat{y} = F_{cls}(F_{Mix}(F_{Txt}(\mathbf{x}_{Txt}), F_{Img}(\mathbf{x}_{Img}))), \quad (1)$$

where F_{Txt} and F_{Img} are unimodal feature extractors, F_{Mix} is the feature fusing model and F_{cls} is the classification head. In order to model F_{Txt} and F_{Img} , most of the previous methods use different pre-trained models to extract text and image features in different semantic spaces, and for F_{Mix} , the proposed mechanisms vary. The crucial point is how to ensure that features provided from both modalities will be utilized in the later stage, otherwise the gap in semantic space makes the fused features unable to accurately represent the correlation between image and text. Instead of applying sophisticated and black-box feature-fusing networks, we employ a simple yet effective method where pretrained networks for multimodal learning is introduced to extract aligned multimodal features and to guide the learning of the classification network. We choose the CLIP model [34] to measure the cross-modal similarity considering that the model is trained to provide the most appropriate language description of a given image and vice versa, and therefore is in line with the above requirements. After feature extraction and alignment, we use a light-weight network to implement L_{cls} which predicts \hat{y} .

3.2 Network Specification

Figure 2 illustrates the network design of FND-CLIP. The whole pipeline consists of four main modules, namely, unimodal feature encoder, CLIP-based encoder, projection and attention module, and finally the classifier.

Unimodal feature generation. We use a pretrained BERT model to obtain the feature $f_{BERT} \in \mathbb{R}^{n_{BERT}}$ of \mathbf{x}_{Txt} . For the image \mathbf{x}_{Img} , we use ResNet [17] to get deep representations $f_{ResNet} \in \mathbb{R}^{n_{ResNet}}$ from the image. Besides f_{BERT} and f_{ResNet} , we use CLIP encoders to encode text and image and obtain the features $f_{CLIP-T} \in \mathbb{R}^{n_{CLIP}}$ and $f_{CLIP-I} \in \mathbb{R}^{n_{CLIP}}$. In order to improve the representation capability of the unimodal branches, embedding concatenation are performed in the text and image unimodal intra-modalities, respectively,

$$\begin{cases} f_{Txt} = concat(f_{BERT}, f_{CLIP-T}) \\ f_{Img} = concat(f_{ResNet}, f_{CLIP-I}), \end{cases} \quad (2)$$

where $f_{Txt} \in \mathbb{R}^{n_{BERT}+n_{CLIP}}$ and $f_{Img} \in \mathbb{R}^{n_{ResNet}+n_{CLIP}}$.

CLIP-guide multimodal feature generation. The text and image features extracted by BERT and ResNet respectively have significant cross-modal semantic gaps, and it is difficult for the network to learn their intrinsic semantic correlation if they are fused directly. Therefore, the two features are only used as unimodal representation, while the multimodal representation is obtained by first concatenating the alignment features of the text-image pair extracted by CLIP and then fine-tuning them to reduce redundancy and introduce attention. The concatenated feature is denoted as $f_{Mix} \in \mathbb{R}^{2 \times n_{CLIP}}$, where

$$f_{Mix} = concat(f_{CLIP-T}, f_{CLIP-I}). \quad (3)$$

The multimodal features reflect the correlation between the two modalities and contain meaningful semantic information. The assistance of the multimodal features to unimodal features is to learn the cross-modal similarity. Previous works often use a single network to mine both coarse and fine features from a modality, which is quite demanding on the learning ability of the model. Here, with the introduction of CLIP model, BERT and ResNet, which is the pre-training models for unimodal tasks, can pay more attention to trivial clues compared to extracting semantic information. For example, BERT can better extract emotional features of texts, and ResNet can identify higher-frequent noise patterns of images. In contrast, the training strategy of CLIP uses large-scale image-text pairs to learn the extraction of semantics, while largely ignoring emotion, noise and other features irrelevant to image and text matching. Therefore, using CLIP for multimodal feature generation can well collaborate with the unimodal features to respectively scrutinize the news from different aspects.

After we get the three features of different modalities, we use three individual projection head P_{Txt} , P_{Img} and P_{Mix} made up of Multi-Layer Perceptrons (MLP) to process the features. The goal is to reduce the dimension of the coarse features provided by the encoders and help filtering out redundant information. These networks share the same architecture but do not share weights. As is shown in Figure 2, every the projection head contains two sets of full connected layer with Batch Normalization [20] layer, a ReLU activation function, and a dropout layer.

Merely combining the CLIP-based features as the multimodal features cannot necessarily provide enough reliable information. The reason is that the authenticity of news is not completely correlated with image-text correlation. Some news posts, no matter real or fake, lack cross-modal relation or even semantic information. In that case, some instances require more emotion, noise, and other features, and the corresponding multimodal features might be noisy when the similarity is low and fully utilizing such information might impair the performance. To address the ambiguity issue between multimodal features, we measure the cosine similarity between the text features and the image features provided by CLIP, to adjust the intensity of fused features. The cosine similarity is calculated as follows.

$$sim = \frac{f_{Txt} \cdot (f_{Img})^T}{\|f_{Txt}\| \|f_{Img}\|}. \quad (4)$$

Then, we apply standardization and a Sigmoid functions to map the similarity into the range $[0 - 1]$. The normalization is done by calculating the running status of mean and standard deviation during training, and subtract the running mean from sim and divide it with the running standard deviation. Compared to the contrastive learning paradigm, the normalization helps to calculate the similarity without comparing the news post with other instances.

Thus, the process of obtaining the projected unimodal and multimodal features is as follows.

$$\begin{cases} m_{Txt} = P_{Txt}(f_{Txt}) \\ m_{Img} = P_{Img}(f_{Img}) \\ m_{Mix} = \text{Sigmoid}(\text{Std}(sim)) \cdot P_{Mix}(f_{Mix}). \end{cases} \quad (5)$$

Feature aggregation using modality-wise attention. We apply an attention mechanism to reweight the projected features before aggregating the features from different modalities using spatial addition. Inspired by the Squeeze-and-Excitation Network (SE-Net) [19], we designed a modality-wise attention module as shown in Figure 2 to weight each feature adaptively. First, the three $L \times 1$ features are concatenated into one $L \times 3$ feature, where L represents the length of the feature. Average pooling and maximum pooling are adopted to squeeze a 1×3 vector via summation, corresponding to the initial weight of each channel. Then, the initial weight obtained in the previous step is sent into the two 3×3 fully connected layers with GELU [18] activation function, and normalized into the range $[0 - 1]$ using Sigmoid functions respectively to obtain the attention weights $att = \{att_{Txt}, att_{Img}, att_{Mix}\}$. Finally, the weights are multiplied respectively on m_{Txt} , m_{Img} and m_{Mix} , and a sum process is performed to obtain the $L \times 1$ aggregated feature m_{Agg} .

$$m_{Agg} = att_{Txt} \cdot m_{Txt} + att_{Img} \cdot m_{Img} + att_{Mix} \cdot m_{Mix}. \quad (6)$$

Classification and objective function. We feed the aggregated representation m_{Agg} into a two-layer fully-connected network as the classifier F_{cls} to predict the label \hat{y} . The objective function of FND-CLIP is to minimize the cross-entropy loss to correctly predict the real and fake news.

$$\mathcal{L}_{CE} = y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}). \quad (7)$$

3.3 Training Detail

On the selection of BERT pretrained models, we respectively use the “bert-base-chinese” model on Chinese data and the “bert-base-uncased” model on English data, perform an attention-based post-processing [21]. The length of the input text is set to 300 words. About the ResNet, we use pre-trained ResNet-101 to extract visual features, setting the size of the input image to 224×224 . The size of the images inputted to CLIP is the same as that to ResNet. Since CLIP has not pre-trained Chinese text model, we use Google Translation API [23] to translate Chinese texts to English. In addition, we use the summary generation model [35] to generate summary statements as the CLIP input for the text with the size longer than 50, to meet the requirements that the input size of the text has an upper bound in CLIP. The used pre-trained CLIP model is “ViT-B/32”. We fine-tune ResNet in training stage, while freezing the weights of BERT and CLIP due to their difficulty in training on small datasets. We implement the projection heads using two fully connected layers with 256 and 64 hidden units, respectively. The hidden sizes of the two fully connected layers in the classifier are 64 and 2, respectively. The batch size is set as 64.

We use Adam optimizer [25] with the default parameters. The learning rate is 1×10^{-3} where weight decay is 12. We trained a model for 50 epochs and chose the epoch getting the best test accuracy among them as the final result to avoid over-fitting.

4 EXPERIMENTS

4.1 Experimental Setup

Dataset. We use three real-world datasets collected from social media, namely, Weibo [22], Gossipcop, and Politifact [36]. During experiments, the unimodal news posts with no image or no text

Table 1: Performance comparison between FND-CLIP and other methods on three datasets. Our method achieves the highest accuracy among these methods, and its precision, recall, and F1-score are also higher than most of the compared methods.

| | Method | Accuracy | Fake News | | | Real News | | |
|------------|-------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | | | Precision | Recall | F1-score | Precision | Recall | F1-score |
| Weibo | EANN [42] | 0.827 | 0.847 | 0.812 | 0.829 | 0.807 | 0.843 | 0.825 |
| | MVAE [24] | 0.824 | 0.854 | 0.769 | 0.809 | 0.802 | 0.875 | 0.837 |
| | Spotfake [40] | 0.892 | 0.902 | 0.964 | 0.932 | 0.847 | 0.656 | 0.739 |
| | MVNN [45] | 0.846 | 0.809 | 0.857 | 0.832 | 0.879 | 0.837 | 0.858 |
| | SAFE [48] | 0.762 | 0.831 | 0.724 | 0.774 | 0.695 | 0.811 | 0.748 |
| | LIIMR [39] | 0.900 | 0.882 | 0.823 | 0.847 | 0.908 | 0.941 | 0.925 |
| | MCAN [44] | 0.899 | 0.913 | 0.889 | 0.901 | 0.884 | 0.909 | 0.897 |
| | CAFE [9] | 0.840 | 0.855 | 0.830 | 0.842 | 0.825 | 0.851 | 0.837 |
| | FND-CLIP | 0.907 | 0.914 | 0.901 | 0.908 | 0.914 | 0.901 | 0.907 |
| Politifact | RoBERTa-MWSS [37] | 0.820 | - | - | - | 0.820 | - | - |
| | SAFE [48] | 0.874 | 0.851 | 0.830 | 0.840 | 0.889 | 0.903 | 0.896 |
| | Spotfake+ [38] | 0.846 | - | - | - | - | - | - |
| | TM [4] | 0.871 | - | - | - | 0.901 | - | - |
| | LSTM-ATT [27] | 0.832 | 0.828 | 0.832 | 0.830 | 0.836 | 0.832 | 0.829 |
| | DistilBert [1] | 0.741 | 0.875 | 0.636 | 0.737 | 0.647 | 0.880 | 0.746 |
| | CAFE [9] | 0.864 | 0.724 | 0.778 | 0.750 | 0.895 | 0.919 | 0.907 |
| | FND-CLIP | 0.942 | 0.897 | 0.897 | 0.897 | 0.960 | 0.960 | 0.960 |
| | RoBERTa-MWSS [37] | 0.800 | - | - | 0.800 | - | - | - |
| Gossipcop | SAFE [48] | 0.838 | 0.758 | 0.558 | 0.643 | 0.857 | 0.937 | 0.895 |
| | Spotfake+ [38] | 0.856 | - | - | - | - | - | - |
| | TM [4] | 0.842 | - | - | - | 0.896 | - | - |
| | LSTM-ATT [27] | 0.842 | 0.845 | 0.842 | 0.844 | 0.839 | 0.842 | 0.821 |
| | DistilBert [1] | 0.857 | 0.805 | 0.527 | 0.637 | 0.866 | 0.960 | 0.911 |
| | CAFE [9] | 0.867 | 0.732 | 0.490 | 0.587 | 0.887 | 0.957 | 0.921 |
| | FND-CLIP | 0.880 | 0.761 | 0.549 | 0.638 | 0.899 | 0.959 | 0.928 |

description were filtered out. If a news post contains a text with multiple associated images, we randomly select one image. Weibo is a widely used Chinese dataset in fake news detection. The training set contains 3,749 real news and 3,783 fake news, and the test set contains 1,996 news. Politifact and Gossipcop datasets are two English datasets collected from the political and entertainment domains of FakeNewsNet [36] repository, respectively. Politifact contains 244 real news and 135 fake news in the training set and 75 real news and 29 news in the test set. Gossipcop contains 10,010 training news, including 7,974 real news and 2,036 fake news. The test set contains 2,285 real news and 545 fake news. Besides, while Twitter [6] is also a well-known multimodal dataset for FND, we find that it contains plenty of duplicated posts and over 10k posts host only 463 images. More importantly, more than 70% of tweets on Twitter dataset are related to a single event, which can easily lead to model overfitting. Therefore, we do not conduct experiments on Twitter.

Baseline Methods. For a fair and reproducible comparison, we have to be selective in choosing the baseline methods. First, we prefer methods that provide pre-trained models or source code publicly available. Second, the methods should follow a common evaluation protocol where the three datasets are used for training

and testing. Accordingly, we compare FND-CLIP with the following methods and provide a quick recap.

EANN [42], which employs an auxiliary task of event classification to improve generalizability.

MVAE [24], which uses a variational autoencoder to model representations between text and images for fake news detection.

Spotfake [40], which uses VGG and BERT to respectively extract image and text features and concatenates them to classify.

MVNN [45], which incorporates textual semantic features, visual tampering features, and similarity of textual and visual information in fake news detection.

SAFE [48], which fed the relevance between news textual and visual information into a classifier to detect fake news.

LIIMR [39], which identifies and suppresses information from weaker modalities and extracts relevant information from the strong modality on a per-sample basis.

MCAN [44], which stacks multiple co-attention layers to fuse the multimodal features.

CAFE [9], which formulates an ambiguity-aware multimodal fake news detection method to adaptively aggregate unimodal features and cross-modal correlations.

RoBERTa-MWSS [37], which exploits multiple weak signals from different sources from user and content engagements.

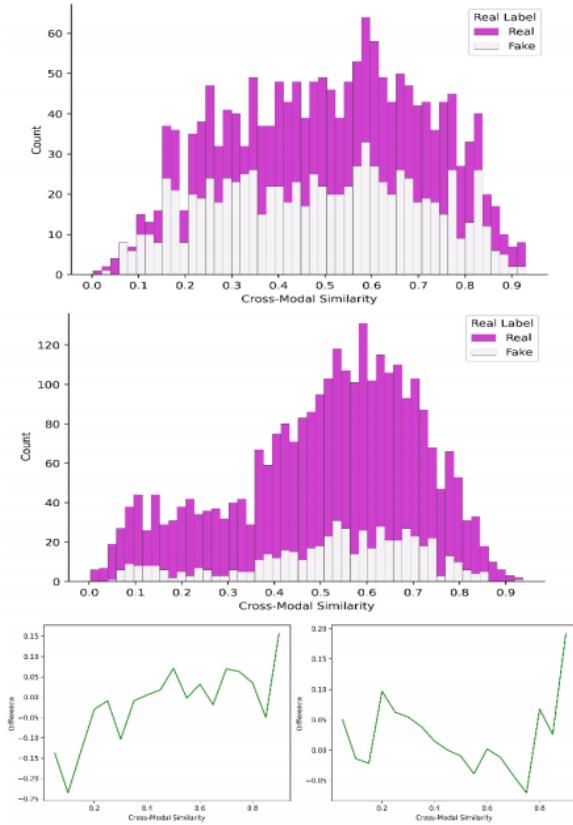


Figure 3: Statistical analysis on cross-modal similarity of different news. The first and second row respectively show the counting of real/fake news according to cross-modal similarity on Weibo and Gossipcop. The third row shows the distance between the real news rate in each bin compared to the average rate on the corresponding dataset (left: Weibo, right: Gossipcop).

Spotfake+ [38], which is an improved version of Spotfake and can detect full length articles.

TM [4], which utilizes lexical and semantic properties of both true and fake news text to detect fake news.

LSTM-ATT [27], which builds a model based on XGBoost to detect full length fake news.

DistilBert [1], which uses latent representations of news articles and user-generated content to guide model learning.

4.2 Performance Analysis

Table 1 shows the average precision, recall, and accuracy of FND-CLIP on three representative datasets. The results are promising, with over 90% average accuracy on Weibo and over 94% on Politifact, which indicates that the proposed method is a dependable

and robust fake news detection algorithm that can detect anomalies given multi-lingual and multi-domain news. Especially, the recall rates of real news on the three datasets are all above 0.9, and therefore FND-CLIP is less likely to classify a real news as fake.

To further conduct statistical analysis on how cross-modal similarity correlates with the attention score and how they vary given different news instances, Figure 3 shows the correlation between the CLIP-based cross-modal similarity score and the fake news ratio on Weibo and Gossipcop dataset. In row 1 and 2, we group all news in each dataset into several bins according to their similarity score, and find that a news is more likely to be real when the similarity score is high. In row 3, we calculate real news rates of each bin and subtract them with the average real news rate of the corresponding dataset. The curves show that real news rate goes up with the increasing ambiguity on Weibo, and first goes down then surges up on Gossipcop. Such statistical characteristics are useful for deep networks to identify fake news.

4.3 Comparison with State-of-the-arts

We further compare FND-CLIP with the above-mentioned state-of-the-art methods and the comparison results are presented in Table 1. '-' means the results are not available from the original paper. As shown in Table 1, FND-CLIP outperforms all the compared methods on the three datasets in terms of Accuracy, and achieves slightly lower than Spot on Weibo in Recall. FND-CLIP achieves the highest accuracy of 90.7%, 94.2%, and 88.0%, which surpasses 0.7%, 6.8%, and 1.3% over the state-of-the-art method, on the three real-world datasets, respectively. Besides, we rank either 1st or 2nd in precision, recall, and accuracy in all tests, which proves the effectiveness of FND-CLIP.

Many fake news detection methods, such as EANN and Spotfake, rely only on the fused features obtained by direct use of concatenating or attention mechanisms. However, these fused features cannot provide sufficient discrimination ability to classify fake news, because the text and image features separately extracted are not in the same semantic space and the correlation information of the text and image is not well-paid attention to during the fusion process. Therefore, the experimental results of these methods are unsatisfactory. CAFE uses cross-modal alignment to train encoders that can map texts and images into the same semantic space. By using the features fused from the aligned text and image features to classify, it achieves good experimental results, especially on the Politifact and Gossipcop datasets. However, due to the limitation of the number of data sets and the rough label method for training labels, the encoding effect of the encoder is not optimal, and the semantic gap between text and image features is still significant. In addition, CAFE designs an ambiguity learning module to calculate a weight used for adaptively adjusting the calculation of different modalities. However, the weights for selecting unimodal or multimodal features are obtained by manual calculation, and cannot be further optimized by reverse gradient propagation, thus affecting the performance of the detection.

FND-CLIP outperforms most of the state-of-the-art methods, mainly due to the following reasons. First, the pre-trained CLIP encoders in FND-CLIP can generate semantically information-rich text and image features in the same semantic space, ensuring the

fused feature correctly reflects the correlation between text and image, and providing complementary information for the unimodal features. The modality-wise attention mechanism adaptively determines the weights of text, image, and fused features, avoiding the influence of invalid features on the representation ability of final features, and further improving the classification accuracy.

4.4 Ablation Studies

We explore the influence of the key components in FND-CLIP by evaluating the performance of the model with varied and partial setups. In each test, we remove different components and train the models from scratch. The compared variants of FND-CLIP are implemented as follows.

- FND-CLIP w/o A. We remove the modality-wise attention module and direct aggregate the three features to obtain final feature;
- FND-CLIP w/o F. We remove the fusion module and use two unimodal features to classify news;
- FND-CLIP w/o C. We remove all CLIP-related modules and only use BERT and ResNet to extract text and image features.
- FND-CLIP multimodal-only: We remove the unimodal feature extractor, BERT and ResNet, and only use CLIP fused feature as final feature;
- FND-CLIP image-only: We remove the all text-related features and only use image feature extracted by ResNet to classify;
- FND-CLIP text-only: We only use BERT-extracting feature to complete the detection task without any visual information.

Effectiveness of Each Component. First, we analyze the impact of different components in FND-CLIP for fake news detection. From the results shown in Table 2, we have the following observations:

1) FND-CLIP outperforms FND-CLIP w/o C, proving that CLIP can effectively provide discernable features for fake news detection task and significantly improve the accuracy of classification. Although only intra-modal features can be used for classification, the lack of interaction between modalities makes the final features lack the ability to represent the intrinsic relationship between images and texts. 2) FND-CLIP outperforms FND-CLIP w/o F, indicating that although the unimodal branches contain the CLIP-coded features, the fused feature reflecting the correlation of text and image provides effective multimodal information for classifier. Meanwhile, FND-CLIP w/o F outperforms FND-CLIP w/o C, indicating that the complement to unimodal features using CLIP-coded features is effective. 3) FND-CLIP outperforms FND-CLIP w/o A on Weibo and Gossipcop, indicating that modality-wise attention can help FND-CLIP adaptively weight useful modalities. FND-CLIP w/o A directly fuses the features of different modalities, which may cause the final feature be affected by invalid information from a modality.

Contributions from Different Modalities. The second set of experiments is to evaluate the classification performance of different modalities in fake news detection. From Table 2, we draw some analysis as follows: 1) FND-CLIP image-only performs worst, especially on Gossipcop dataset, where the F1 score of fake news was almost zero, meaning that all news was judged real and the model had no classification ability at all. This shows that in fake news detection, simple visual information provides fewer classification

clues than other modalities. 2) FND-CLIP multimodal-only achieves accuracy of 81.7%, 90.3%, and 86.2% on Weibo, Politifact, and Gossipcop datasets respectively, but performs worse than FND-CLIP text-only on Weibo and Gossipcop datasets, indicating that the correlation information of images and texts can be used to classify fake news. However, the classification ability of fused feature is limited because news itself has modal irrelevance and ambiguity. In addition, CLIP-based fused features focus on the semantics of the text, while the BERT-based text features also extract emotional features that are helpful for fake news detection. 3) FND-CLIP text-only achieves the second-best results, indicating that only using text feature can basically complete the classification task for fake news. However, FND-CLIP outperforms FND-CLIP text-only, proving the visual feature can supplement classification information and the correct use of multimodal features is superior to using only unimodal features in fake news detection.

4.5 T-SNE Visualizations

In Figure 4, we further analyze the proposed method using t-SNE [41] visualizations of the features before classifier that are learned by FND-CLIP, CAFE, and also the proposed method with partial settings such as FND-CLIP w/o C, FND-CLIP w/o A, FND-CLIP text-only, and FND-CLIP image-only on the test dataset of Weibo in Figure 4.

The dots with the same color mean that they are within the same label. From Figure 4 we can see that the boundary of different label dots in FND-CLIP is more pronounced than that in CAFE, FND-CLIP w/o C, and FND-CLIP w/o A, revealing that the extracted features in FND-CLIP are more discriminative than those in CAFE and the CLIP-related modules and modality-wise attention are useful for improving the classification ability of FND-CLIP.

In addition, by comparing Figure 4a, Figure 4d, Figure 4e, and Figure 4f, we can see that image features alone are not enough for classification, which indicates that the image itself does not have classification ability. The effect of text-only is much better than that of image-only. Proving that the text features play a leading role in fake news detection, but there are still many sample dots that cannot be distinguished. FND-CLIP w/o C, which contains both text and image features, has a more obvious boundary of the dots than FND-CLIP text-only, indicating that different modalities have complementary information. In addition, the separation degree of the sample dots in Figure 4a is higher than that in Figure 4d, indicating that the multimodal features based on CLIP can improve the representation ability of the final features.

5 CONCLUSIONS

In this paper, we present a novel multimodal fake news detection method called FND-CLIP, which uses CLIP to extract aligned multimodal features and guide the learning of network for different modalities. In addition, we introduce modality-wise attention to adaptively determine the weights of text, image, and fused features. It can avoid introducing noisy and redundant features during feature fusion, which further improve the classification accuracy. We conduct comprehensive experiments on several well-known FND datasets. The results show that using CLIP for multimodal feature generation can well collaborate with the unimodal features

Table 2: Ablation study on the architecture design and different features of FND-CLIP on three datasets. The entire FND-CLIP achieved the highest accuracy and F1-score, demonstrating that every module in the architecture of our method is effectiveness and every modality is effectively utilized.

| | Method | Accuracy | Fake News | | | Real News | | |
|------------|--------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | | | Precision | Recall | F1-score | Precision | Recall | F1-score |
| Weibo | FND-CLIP multimodal-only | 0.817 | 0.899 | 0.718 | 0.798 | 0.761 | 0.917 | 0.832 |
| | FND-CLIP image-only | 0.796 | 0.862 | 0.711 | 0.779 | 0.750 | 0.884 | 0.811 |
| | FND-CLIP text-only | 0.872 | 0.906 | 0.833 | 0.868 | 0.842 | 0.911 | 0.875 |
| | FND-CLIP w/o C | 0.874 | 0.895 | 0.851 | 0.872 | 0.855 | 0.898 | 0.876 |
| | FND-CLIP w/o F | 0.893 | 0.925 | 0.857 | 0.890 | 0.864 | 0.929 | 0.895 |
| | FND-CLIP w/o A | 0.897 | 0.936 | 0.855 | 0.893 | 0.863 | 0.940 | 0.900 |
| | FND-CLIP | 0.907 | 0.914 | 0.901 | 0.908 | 0.901 | 0.914 | 0.907 |
| Politifact | FND-CLIP multimodal-only | 0.903 | 0.807 | 0.862 | 0.833 | 0.944 | 0.919 | 0.932 |
| | FND-CLIP image-only | 0.748 | 0.600 | 0.310 | 0.409 | 0.773 | 0.919 | 0.840 |
| | FND-CLIP text-only | 0.903 | 0.913 | 0.724 | 0.808 | 0.900 | 0.973 | 0.935 |
| | FND-CLIP w/o C | 0.893 | 0.875 | 0.724 | 0.793 | 0.899 | 0.960 | 0.928 |
| | FND-CLIP w/o F | 0.903 | 0.880 | 0.759 | 0.815 | 0.910 | 0.960 | 0.934 |
| | FND-CLIP w/o A | 0.942 | 0.926 | 0.862 | 0.893 | 0.947 | 0.973 | 0.960 |
| | FND-CLIP | 0.942 | 0.897 | 0.897 | 0.897 | 0.960 | 0.960 | 0.960 |
| Gossipcop | FND-CLIP multimodal-only | 0.862 | 0.708 | 0.484 | 0.575 | 0.886 | 0.952 | 0.918 |
| | FND-CLIP image-only | 0.814 | 1.000 | 0.033 | 0.064 | 0.813 | 1.000 | 0.897 |
| | FND-CLIP text-only | 0.871 | 0.741 | 0.508 | 0.603 | 0.891 | 0.958 | 0.923 |
| | FND-CLIP w/o C | 0.870 | 0.745 | 0.494 | 0.594 | 0.888 | 0.960 | 0.923 |
| | FND-CLIP w/o F | 0.874 | 0.723 | 0.562 | 0.632 | 0.901 | 0.949 | 0.924 |
| | FND-CLIP w/o A | 0.873 | 0.715 | 0.567 | 0.633 | 0.902 | 0.946 | 0.923 |
| | FND-CLIP | 0.880 | 0.761 | 0.549 | 0.638 | 0.899 | 0.959 | 0.928 |

extracted by ResNet and BERT in mining crucial features for fake news detection. More importantly, FND-CLIP outperforms many of the state-of-the-art methods in multimodal fake news detection.

Aside from the performance gain of FND-CLIP, the outputs are still in the form of binary value that predicts either "real" or "fake", which cannot somehow explain why the news is predicted fake and which elements in the news are most suspicious and abnormal. In future works, we head towards developing more explainable fake news detection systems that can provide reasons why a given news is predicted as real or fake.

REFERENCES

- [1] Liesbeth Allein, Marie-Francine Moens, and Domenico Perrotta. 2021. Like Article, Like Audience: Enforcing Multimodal Correlations for Disinformation Detection. *arXiv preprint arXiv:2108.13892* (2021).
- [2] Tadej Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multi-modal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.
- [3] Gaurav Bhatt, Aman Sharma, Shivam Sharma, Ankush Nagpal, Balasubramanian Raman, and Ankush Mittal. 2018. Combining neural, statistical and external features for fake news stance identification. In *Companion Proceedings of the The Web Conference 2018*. 1353–1357.
- [4] Bimal Bhattacharai, Ole-Christoffer Granmo, and Lei Jiao. 2021. Explainable Tsetlin Machine framework for fake news detection with credibility score assessment. *arXiv preprint arXiv:2105.09114* (2021).
- [5] Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 549–556.
- [6] Christina Boididou, Symeon Papadopoulos, Markos Zampoglou, Lazaros Apostolidis, Olga Papadopoulou, and Yiannis Kompatseris. 2018. Detection and visualization of misleading content on Twitter. *International Journal of Multimedia* *Information Retrieval* 7, 1 (2018), 71–86.
- [7] Juan Cao, Peng Qi, Qiang Sheng, Tianyun Yang, Junbo Guo, and Jintao Li. 2020. Exploring the role of visual content in fake news detection. *Disinformation, Misinformation, and Fake News in Social Media* (2020), 141–161.
- [8] Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. 2021. Image Manipulation Detection by Multi-View Multi-Scale Supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14185–14193.
- [9] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal Ambiguity Learning for Multimodal Fake News Detection. In *Proceedings of the ACM Web Conference 2022*. 2897–2905.
- [10] Marcos V Conde and Kerem Turgutlu. 2021. CLIP-Art: contrastive pre-training for fine-grained art classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3956–3960.
- [11] Nadia K Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the association for information science and technology* 52, 1 (2015), 1–4.
- [12] Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. 2021. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1574–1583.
- [13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [14] Han Guo, Juan Cao, Yazi Zhang, Junbo Guo, and Jintao Li. 2018. Rumor detection with hierarchical social attention network. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 943–951.
- [15] Bing Han, Xiaoguang Han, Hua Zhang, Jingzhi Li, and Xiaochun Cao. 2021. Fighting fake news: two stream network for deepfake detection via learnable SRM. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3, 3 (2021), 320–331.
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

Anonymous et al.

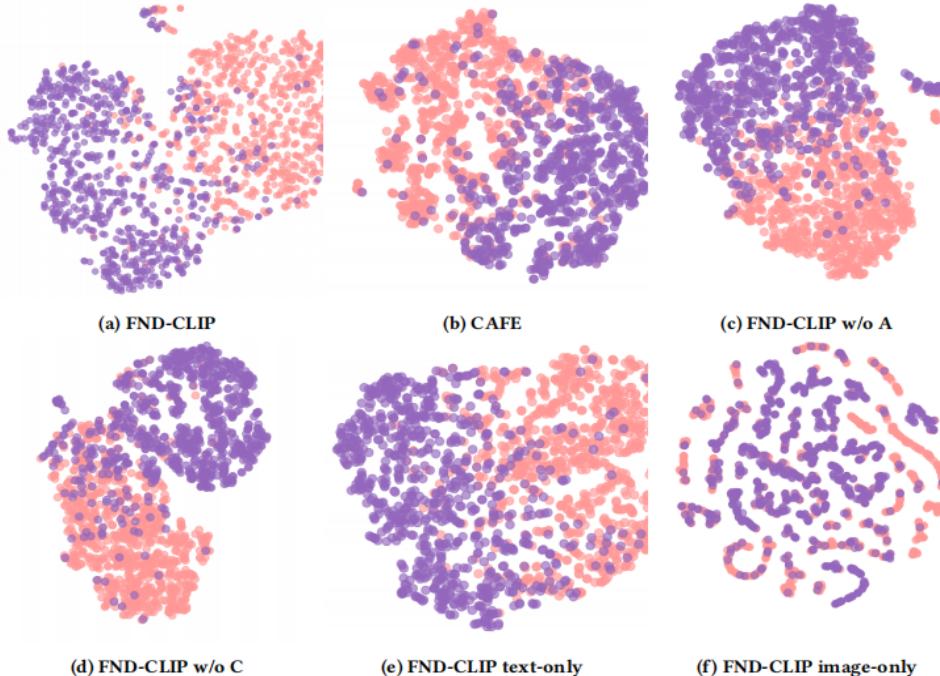


Figure 4: T-SNE visualizations of the features before classifier that are learned by FND-CLIP , CAFE, FND-CLIP w/o C, FND-CLIP w/o A, FND-CLIP text-only, and FND-CLIP image-only on the test dataset of Weibo.

- [18] Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* (2016).

[19] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.

[20] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*. 448–456.

[21] Ganesh Jawahar, Benoit Sagot, and Djamel Seddah. 2019. What does BERT learn about the structure of language?. In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

[22] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*. 795–816.

[23] Gregory Johnson. 2012. Google Translate <http://translate.google.com>. *Technical Services Quarterly* 29, 2 (2012), 165–165.

[24] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*. 2915–2921.

[25] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[26] Peiguang Liu, Xian Sun, Hongfeng Yu, Yu Tian, Fanglong Yao, and Guangluan Xu. 2021. Entity-Oriented Multi-Modal Alignment and Fusion Network for Fake News Detection. *IEEE Transactions on Multimedia* (2021).

[27] Jun Lin, Glenn Tremblay-Taylor, Guanyi Mou, Di You, and Kyumin Lee. 2019. Detecting fake news articles. In *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 3021–3025.

[28] Jiansen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* 32 (2019).

[29] Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. MD-FEND: Multi-domain Fake News Detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3343–3347.

[30] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).

[31] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638* (2017).

[32] Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. 2021. Improving Fake News Detection by Using an Entity-enhanced Framework to Fuse Diverse Multimodal Clues. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1212–1220.

[33] Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. 2019. Exploiting multi-domain visual information for fake news detection. In *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 518–527.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[35] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* (2019).

[36] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data* 8, 3 (2020), 171–188.

[37] Kai Shu, Guoqing Zheng, Yichuan Li, Subhabrata Mukherjee, Ahmed Hassan Awadallah, Scott Ruston, and Huan Liu. 2020. Leveraging multi-source weak social supervision for early detection of fake news. *arXiv preprint arXiv:2004.01732* (2020).

[38] Shivangi Singh, Anubha Kabra, Mohit Sharma, Rajiv Ratn Shah, Tanmoy Chakraborty, and Ponnurangam Kumaraguru. 2020. Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13915–13916.

- [39] Shivangi Singhal, Tanisha Pandey, Saksham Mrig, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2022. Leveraging Intra and Inter Modality Relationship for Multimodal Fake News Detection. (2022).
- [40] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*. IEEE, 39–47.
- [41] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
- [42] Yaqiang Wang, Fenglong Ma, Zhiwei Jin, Yu Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 849–857.
- [43] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Zhentao Tan, Lu Yuan, Weiming Zhang, and Nenghai Yu. 2021. Hairclip: Design your hair by text and reference image. *arXiv preprint arXiv:2112.05142* (2021).
- [44] Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal Fusion with Co-Attention Networks for Fake News Detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2560–2569.
- [45] Junxiao Xue, Yabo Wang, Yichen Tian, Yafei Li, Lei Shi, and Lin Wei. 2021. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management* 58, 5 (2021), 102610.
- [46] Keren Ye and Adriana Kovashka. 2021. A case study of the shortcut effects in visual commonsense reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 3181–3189.
- [47] Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the Web Conference 2021*. 3465–3476.
- [48] Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: Similarity-Aware Multi-modal Fake News Detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 354–367.
- [49] Arkaitz Zubia, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)* 51, 2 (2018), 1–36.

附录 B 外文文献译文

基于CLIP引导学习的多模态假新闻检测

Yangmin Zhou

计算机科学学院

中国上海

ymzhou21@fudan.edu.cn

Qichao Ying

计算机科学学院

中国上海

qcying20@fudan.edu.cn

Zhenxing Qian

计算机科学学院

中国上海

zxqian@fudan.edu.cn

Sheng Li

计算机科学学院

中国上海

lisheng@fudan.edu.cn

Xinpeng Zhang

计算机科学学院

中国上海

zhangxinpeng@fudan.edu.cn

摘要

多模态假新闻检测吸引了社会取证领域的许多研究兴趣。许多现有的方法引入定制的注意机制来指导单峰特征的融合。然而，如何计算这些特征的相似性，以及它将如何影响FND的决策过程，仍然是公开的问题。此外，预训练的多模态特征学习模型在假新闻检测中的潜力还没有得到很好的开发。提出了一种FND-CLIP框架，即基于对比语言-图像预训练的多模态假新闻检测网络。给定目标多模态新闻，我们使用基于ResNet的编码器、基于BERT的编码器和两个成对CLIP编码器从图像和文本中提取深度表示。多模态特征是由两种模态的标准化跨模态相似性加权的CLIP生成的特征的串联。在将提取的特征馈送到最终分类器之前，对它们进行进一步处理以减少冗余。我们引入了一个基于模态的注意力模块来自适应地重新加权和聚集特征。我们在典型的假新闻数据集上进行了大量的实验。结果表明，该框架能够更好地挖掘假新闻检测的关键特征。与之前的作品相比，提出的FNDCLIP可以实现更好的性能，即在微博、Politifact和Gossipcop上的整体准确率分别提高了0.7%、6.8%和1.3%。此外，我们证明了基于CLIP的学习可以允许更好的多模态特征选择的灵活性。

CCS概念

- 计算机系统组织→嵌入式系统；重复冗余；机器人技术；网络→网络可靠性。

通讯作者。这项工作得到了国家自然科学基金资助U20B2051, U1936214。

允许免费制作本作品全部或部分的数字或硬拷贝供个人或课堂使用，前提是不得以盈利或商业利益为目的制作或分发拷贝，并且拷贝第一页带有本声明和完整引用。ACM以外的其他人拥有的本作品组成部分的版权必须得到尊重。允许带信用摘要。以其他方式复制，或重新发布，张贴在服务器上或重新分发到列表，需要事先明确的许可和/或费用。向请求权限permissions@acm.org。大会缩写'XI, 2018年6月3日至5日，纽约伍德斯托克
2018计算机机械协会 美国计算机学会国际标准书
号978-1-4503-XXXX-X/18/06。 \$15.00
<https://doi.org/XXXXXX.XXXXXXX>

关键词

数据集、神经网络、凝视检测、文本标记

ACM参考格式：

周阳明、应启超、钱振兴、张新鹏。2018. 基于CLIP引导学习的多模态假新闻检测。在过程中，请确保从您的权利确认电子邮件中输入正确的会议标题（会议缩写为“XX”）。美国纽约州纽约市ACM, 11页。
<https://doi.org/XXXXXX.XXXXXXX>

1 介绍

在线社交网络已经在很大程度上取代了以报纸和杂志为代表的传统信息交流方式。人们享受在线社交媒体在寻找朋友或分享观点方面的便利。然而，OSNs也促进了假新闻的广泛和快速传播[24, 33, 49]。与书面材料相比，在线新闻帖子更容易被操纵。新闻伪造可以采取各种形式，例如，用另一个对象替换图片中的关键对象，或者对图片进行有偏见甚至误导的评论。更糟糕的是，读者容易受到精心制作的假新闻的影响，并进一步传播它们。总之，假新闻可能会制造恐慌，误导公众舆论，从而对社会产生负面影响。在过去的几十年里，假新闻检测(FND)一直是以数据为中心的研究中心[1, 36, 45]。虽然人工观察互联网上的所有新闻和帖子既昂贵又耗时，但使用机器学习的自动FND是打击假新闻广泛传播的有效方式。FND帮助新闻读者识别新闻文章中的偏见和错误信息，从而阻止它们的传播。关于假新闻检测的早期工作仅仅集中在仅文本或仅图像的内容分析上[5, 11]。预先训练的模型通常被用来验证输入的逻辑和语义的合理性。此外，琐碎的线索，如语法错误或图像处理留下的痕迹可能会被考虑在内。虽然单峰FND模式是有效的，但现代新闻和帖子通常具有多种模式的丰富信息，这些方法忽略了它们的相关性。对于一些假新闻，一个真实的形象可以与全部谎言结合在一起，正确的词语可以用来描述一个被破坏的形象。从这个意义上说，需要多模态特征分析以提供补充福利来帮助FND。

近年来，已经有很多工作综合多模态特征来检测新闻和帖子中的异常[9, 24, 42]。除了融合图像和文本的特征之外，



图1：使用FNDCLIP检测假新闻的例子。每条新闻的三个关注度分值分别是文本分值、图像分值、融合分值。

评论、高票率和传播图是研究者评价帖子真实性的常用方法。这些额外的模式是互动的，并随着时间的推移而变化。许多以前的作品喜欢使用尽可能多的模态。然而，交互式模态不如图像和文本或静态模态可靠。首先，缺乏交互模式可能是常见的。一个典型的例子是，在新注册用户发布的新闻中，不能在历史行为中留下任何线索，如果我们希望在他们提交后不久就拒绝假新闻，也不能在评论或投票中留下任何线索。第二，交互模式不太稳定，并且会随着时间的推移而改变，因此可能会导致不同的取证结果。因此，我们重新考察了FND目前仅有静态模态的技术，发现尽管许多算法为多模态特征融合设计了精心制作的网络[38, 42]，关于多模态特征将如何影响最终决策，这些机制在很大程度上处于黑盒级别。一些工作试图通过生成融合特征时显式计算相关性来解决这个问题。举个例子，

陈等[9]，额外训练变分自动编码器(VAE)，其首先压缩图像和文本，并对比学习以最小化具有正确图像-文本对的新闻的Kullback-Leibler (KL) 散度。相应的跨模态模糊度分数然后被用于重新加权多模态特征[9]。Dhruv等人[24]提出了训练解码器从融合的特征中重建原始文本和低级图像特征的MVAE。这些方法在多模态假新闻检测中取得了不错的性能。

然而，多式FND仍然有一些问题需要解决。首先，我们发现如何计算来自不同模态的特征的相似性以及它将如何影响FND的决策过程仍然是一个公开的问题。对于[9]，我们不确定VAEs的效率如何，以便在给定匹配的图像-文本对的情况下KL发散会很小。对于MVAE来说，虽然重建的能力意味着融合的特征能够包含更多的信息，但在FND看来，这些辅助任务的必要性仍然未知。此外，我们发现，更先进的多模态学习范式和预训练模式在FND没有得到适当的应用。例如，CLIP[34]是一个结合了语言概念知识和图像语义知识的多模态模型。它在各种图像-文本对上进行训练，以预测给定图像的最相关的文本片段，反之亦然。CLIP和其他先进的多模态技术有利于图像-文本特征融合，但它们在FND的使用仍然是不稳定的。

本文提出了一种基于预先训练的对比语言-图像预训练(CLIP)模型的多模态假新闻检测网络FND-CLIP。用于假新闻检测的基于CLIP的学习通过明确地测量目标帖子的文本和图像之间的相关性来解决跨模态歧义问题，并指导特征融合和决策阶段。具体来说，我们使用可微调的ResNet [16] encoder预训练的CLIP图像编码器。文本由可微调的BERT [13] 编码器以及CLIP文本编码器。单峰特征是通过将CLIP生成的特征与可微调的对应物连接起来而生成的。融合特征由两个CLIP输出组成。我们使用三个投影头分别处理单峰和融合特征，缩小它们的尺寸，以便提取FND最重要的特征。此外，我们计算CLIP输出的余弦相似性，并将其标准化为跨模态相似性得分。分数重新加权融合的特征，其中我们规定，如果图像和文本显示低相关性，则融合的特征将提供较少的信息。此外，我们引入了一个注意层，它输出三个分数，自适应地测量这些特征在假新闻检测中的重要性。分类器最终处理总结的特征以区分假新闻和真新闻。

我们在FND的几个典型数据集上进行了大量的实验，包括中文数据集Weibo和英文数据集PolitiFact和Gossip。实验结果表明，FND-CLIP在三个数据集上的整体准确率分别提高了0.7%、6.8%和1.3%。此外，我们证明了基于CLIP的学习可以在多模态特征选择上提供更好的灵活性。数字1展示了四个使用FNDCLIP检测假新闻的例子，我们看到注意力分数以及跨模态相似性在不同的新闻实例中是不同的。当相似性较低时，FNDCLIP能够较少关注多模态特征，因此根据所提供的新闻的特征灵活地聚集信息。

本文的贡献主要有三个方面，即：

- 我们提出了一种基于CLIP学习的多模态假新闻检测方法FND-CLIP，其中CLIP预训练模型用于度量跨模态相似性并指导特征的映射和融合。

Multimodal Fake News Detection via CLIP-Guided Learning

- 我们提出了一种基于模态的注意力机制来对文本、图像和融合特征进行自适应加权。给定不同的新闻实例，我们发现该模型灵活地学习更多地关注单峰或多峰特征中的有用信息。
- 我们在三个著名的数据集上进行了全面的实验，结果证明CLIP生成的特征可以是单峰特征的重要辅助。FND-CLIP优于最先进的假新闻检测方法。

2 相关作品

2.1 单峰假新闻检测

单峰FND通常致力于发现文章文本或图片中的异常。这些算法往往遵循人类决策过程的本质。对于图片，曹等[7]共同研究用于假新闻检测的图像取证特征、语义特征、统计特征、上下文特征。它表明用于图像处理检测的典型方法[8]有助于揭露篡改新闻的痕迹。此外，关于常识的语义不一致[26]以及较差的图像质量[15]可以广泛存在于假新闻中。对于文本来说，验证逻辑合理性是必不可少的[14]，还伴随着发现语法错误、写作风格等线索[31]或提取修辞结构[11]。此外，语言和视觉模式可能高度依赖于特定的事件和相应的领域知识。因此，南等人[29]提出用领域门来聚合混合专家提取的多种表征，并以语言形式处理多领域假新闻传播。

虽然这些单峰特征可以被探究，并且它们确实在区分假新闻中起着关键作用，但是多峰特征如相关性和一致性被忽略了，这潜在地损害了这些单峰模式在多峰新闻中的整体性能。

2.2 多模态假新闻检测

在过去的文献中，已经有很多工作致力于从新闻的图像和文本中挖掘有用的表示来检测假新闻。早期的工作为多模态特征融合设计了复杂的黑盒注意机制[3, 5]。许多其他作品[9, 24, 42]建议在将它们发送到分类器之前更好地对齐从不同模态提取的特征。王等[42]提出了进一步采用事件分类辅助任务来帮助特征提取的EANN。事件分类分支被设计成更好地地理清挖掘的多模态功能，因此既有特定于事件的信息，也有与事件无关的信息。Dhruv等人[24]使用单峰特征提取器处理图像和文本，并进一步利用多峰VAE从它们那里学习共享表示。由VAE产生的采样表示然后被发送到解码器，该解码器试图重建原始文本和低级图像特征。除了关注网络设计之外，其他工作也从数据集中挖掘更多信息。例如，齐等人[32]声称图像特征提取器不能很好地理解图像中的视觉实体，如名人、地标和文本，因此建议手动提取这些类型的信息作为语言辅助。张等[47]设计一个新颖的双重情绪特征描述符来测量帖子及其评论之间的情绪差距，并验证双重情绪在虚假和真实新闻之间是有区别的。陈等[9]使用两个vae来压缩图像和文本，并对比学习最小化正确匹配的图像-文本对的Kullback- Leibler (KL) 散度。然后，在特征融合期间，所得分数用于重新加权多模态特征。

尽管这些方法在多模态FND中取得了不错的性能，但仍有一些问题需要关注。首先，如何明确衡量一篇文章中图像和文本之间的相关性仍然不清楚。其次，我们看到在FND很少有工作考虑在多模态学习中应用最近出现的艺术，这促使我们使用基于CLIP的预训练来进一步提高性能。

2.3 多模态学习

近年来，多模态机器学习领域发展迅速[2]。神经结构被用于超

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

越单一模态的任务，例如，视觉问题回答(VQA) [12]，视觉常识推理(VCR) [46]等。在这些任务和其他任务中，需要来自不同模态的先验和特征，并且当只提供单一模态时，算法或深度网络不能有效地工作。开发了几种通用技术来学习图像内容和自然语言的联合表示。例如，CLIP模型[34]旨在成为计算机视觉和自然语言处理之间的桥梁。它在各种图像-文本对上进行训练，以预测给定图像的最相关的文本片段，而不直接优化任务。该模型由两个编码器组成，分别将文本和图像嵌入一个统一的数学空间。然后，对于匹配的图文对，鼓励CLIP以最大化两种模态嵌入之间的余弦相似性。否则，模型的相似性被最小化，以找到最合适成对图像和文本。多维学习有着广阔的前景，其中CLIP的创新已经使许多下游任务受益[10, 43]。其它多模态方案可用Glide [30]和维尔Bert [28]分别用于文本到图像生成和多模态表示学习。

3 方法

3.1 方法概述

对于多模态假新闻检测，我们收集包括文本和图像在内的新闻样本的统计模态，并对每种模态进行表示。样本为 $x = (x_{Txt}, x_{Img})$ ，事实标签为 y ， $y=0$ 说明新闻真实，反之亦然。根据最传统的多模态学习范式，首先从 x_{Txt} 和 x_{Img} 中提取丰富的特征。首先从 x_{Txt} 和 x_{Img} 中提取丰富的特征，这些特征既代表单模态特征，也代表多模态特征。的单模态特征和多模态特征。然后进一步融合并投射到一个单一的值中，该值应接近于事实标签。

$$\hat{y} = F_{cls}(F_{Mix}(F_{Txt}(x_{Txt}), F_{Img}(x_{Img}))), \quad (1)$$

其中 F_{Txt} 和 F_{Img} 是单模态特征提取器， F_{Mix} 是特征融合模型， F_{cls} 是分类头。为了对 F_{Txt} 和 F_{Img} 进行建模，以前的方法大多使用不同的训练过的模型来提取不同语义空间中的文本和图像特征。对于 F_{Mix} ，提议的机制各不相同。最关键的一点是如何保证特征的准确性，克服语义空间的差距使融合后的特征在后期得到利用，否则语义空间的差距使得融合后的特征无法准确表达图像和文本之间的相关性。不同于应用复杂的黑箱特征融合网络，我们采用了一种简单而有效的方法，即引入多模态学习的预训练网络来提取对齐的多模态特征并指导分类网络的学习。我们选择我们选择CLIP模型[34]来测量跨模态的相似性，考虑到该模型被训练为提供最合适的语言描述给定的图像，反之亦然，因此是符合上述要求的。在特征提取和我们使用一个轻量级网络来实现分类，预测标签。

3.2 网络结构

图2举例说明了FND-CLIP的网络设计。整个流水线由四个主要模块组成，即单峰特征编码器、基于CLIP的编码器、投影和注意模块，最后是分类器。

单模态特征提取：我们使用Bert预训练模型对样本 x_{Txt} 得到文本特征 $f_{Txt} \in \mathbb{R}^{n_{BERT}}$ ，使用ResNet[17]得到图像 x_{Img} 的深层特征表示 $f_{ResNet} \in \mathbb{R}^{n_{ResNet}}$ 。除此之外我们使用CLIP的编码器分别获得文本与图像特征 $f_{CLIP-T} \in \mathbb{R}^{n_{CLIP}}$ 和 $f_{CLIP-I} \in \mathbb{R}^{n_{CLIP}}$ 。为了提高单模态分支的表示能力，在文本和图像的单模态内分别进行了嵌入串联。

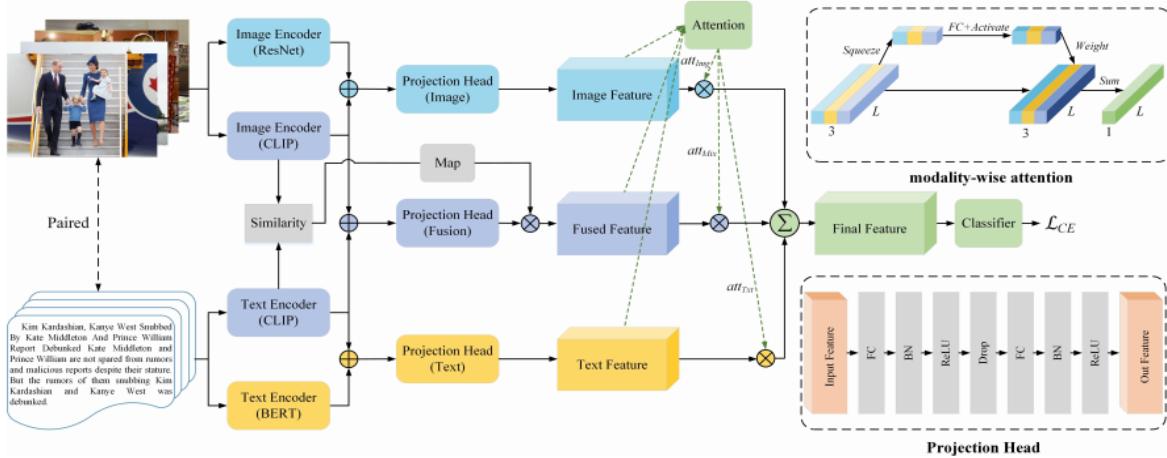


图2:提出的FNDCLIP方法的架构。使用CLIP、BERT和ResNet来提取多模态新闻的不同模态的特征。通过投影头获得不同级别的编码特征。计算片段相似性得分以确定融合特征的重要性。一种基于模态的注意机制被进一步用于自适应地重新加权不同的模态特征，以便分类器对假新闻进行分类。

$$\left\{ \begin{array}{l} f_{Txt} = concat(f_{BERT}, f_{CLIP-T}) \\ f_{Img} = concat(f_{ResNet}, f_{CLIP-I}), \end{array} \right. \quad (2)$$

CLIP指导多模态特征生成：由BERT和ResNet分别提取的文本和图像特征具有显著的跨模态语义差距，如果直接融合，网络很难学习它们内在的语义相关性。因此，这两个特征只被用作单模态表示，而多模态表示则是通过以下方式获得的：融合CLIP提取的文本-图像对的对齐特征，然后对它们进行微调，以减少冗余并引入注意力。串联的特征被表示为：

$$f_{Mix} = concat(f_{CLIP-T}, f_{CLIP-I}). \quad (3)$$

多模态特征反映了两种模态之间的相关性，并包含有意义的语义信息。多模态特征对单峰特征的帮助是学习跨模态相似性。以往的工作往往使用单一网络从一个模态中同时挖掘粗特征和细特征，这对模型的学习能力要求相当高。这里，随着CLIP模型的引入，用于单峰任务的预训练模型BERT和ResNet与提取语义信息相比，可以更加关注琐碎的线索。例如，BERT可以更好地提取文本的情感特征，ResNet可以识别图像的高频噪声模式。相比之下，CLIP的训练策略使用大规模的图文对来学习语义的提取，而很大程度上忽略了情感、噪声等与图文匹配无关的特征。因此，使用CLIP进行多模态特征生成可以很好地与单峰特征协作，分别从不同的角度对新闻进行审查。

在我们获得不同模态的三个特征之后，我们使用三个独立的投影头 P_{Txt} 、 P_{Img} 和 P_{Mix} 组成的多层次感知器(MLP)来处理这些特征。目标是减少由编码器提供的粗略特征的维度，并帮助过滤掉冗余信息。这些网络共享相同的架构，但不共享权重。如图所示2，每个投影头包含两组批量归一化的全连接层[20]层、ReLU激活功能和丢弃层。

仅仅组合基于CLIP的特征作为多模态特征不一定能够提供足够可靠的信息。原因在于新闻的真实性与图文相关性并不完全相关。一些新闻帖子，不管是真的还是假的，都缺乏跨模态关系甚至语义信息。在这种情况下，一些实例需要更多的情感、噪声和其他特征，并且当相似性较低时，相应的多模态特征可能

会有噪声，并且完全利用这种信息可能会损害性能。为了解决多模态特征之间的不确定性问题，我们通过测量文本特征和CLIP提供的图像特征之间的余弦相似度来调整融合特征的强度。余弦相似度计算如下。

$$sim = \frac{f_{Txt} \cdot (f_{Img})^T}{\|f_{Txt}\| \|f_{Img}\|}. \quad (4)$$

然后，我们应用标准化和Sigmoid函数来映射，将相似度映射到[0 - 1]的范围内。归一化是通过计算平均数和标准差的运行状态来完成的。训练期间，从模拟中减去运行中的平均值，再除以运行中的标准差。与对比性学习范式相比，归一化有助于计算相似性，而不需要将新闻帖子与其他实例进行比较。因此，获得预测的单模态和多模态特征的过程如下。

$$\left\{ \begin{array}{l} m_{Txt} = P_{Txt}(f_{Txt}) \\ m_{Img} = P_{Img}(f_{Img}) \\ m_{Mix} = Sigmoid(Std(sim)) \cdot P_{Mix}(f_{Mix}). \end{array} \right. \quad (5)$$

使用模式化的注意力进行特征聚合：我们运用注意力机制，在使用空间加法将不同模态的特征聚合起来之前，对预测的特征进行重新加权。受挤压和激发网络(SE-Net)的启发[19]。我们设计了一个模式化的注意力模块，如图2所示。对每个特征进行自适应加权。首先，三个 $L \times 1$ 特征被串联成一个 $L \times 3$ 的特征，其中 L 代表特征的长度。采用平均池化和最大池化方法，通过求和挤压出一个 1×3 的向量，对应于每个通道的初始权重。然后，上一步得到的初始权重被送入两个 3×3 的全连接层，使用GELU[18]激活函数，并使用Sigmoid函数将其归一化为[0-1]的范围，从而得到注意力权重 $att = \{att_{Txt}, att_{Img}, att_{Mix}\}$ 。最后，这些权重分别乘以 m_{Txt} 、 m_{Img} 与 m_{Mix} ，并进行求和处理。

得到 $L \times 1$ 的聚合特征 m_{Agg} 。

$$m_{Agg} = att_{Txt} \cdot m_{Txt} + att_{Img} \cdot m_{Img} + att_{Mix} \cdot m_{Mix}. \quad (6)$$

分类和目标函数：我们将汇总后的表示 m_{Agg} 馈送到一个两层全连接的网络中，作为分类器 F_{cls} 来预测标签。FND-CLIP的目标函数是使交叉熵损失最小，以正确预测真实和虚假新闻。

$$\mathcal{L}_{CE} = y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}). \quad (7)$$

3.3 训练细节

在BERT预训练模型的选择上，我们分别在中文数据上使用“bert-base-chinese”模型，在英文数据上使用“bert-baseuncased”模型，进行基于注意力的处理[21]。输入文本的长度被设定为300字。关于ResNet，我们使用预先训练好的ResNet-101来提取视觉特征。特征，将输入图像的大小设置为 224×224 。输入到CLIP的图像与输入到ResNet的图像相同。CLIP没有预先训练的中文文本模型，我们使用谷歌翻译API[23]将中文文本翻译成英文。此外。我们使用摘要生成模型[35]来生成摘要语句作为CLIP的输入，以满足大小超过50的文本的要求。使用的预训练的CLIP模型是“ViTB/32”。我们在训练阶段对ResNet进行了微调，同时冻结了BERT和CLIP的权重，因为它们很难在小数据集上进行训练。我们使用两个全连接层来实现投影头，分别有256个和64个隐藏单元。这两个全连接层的分类器中两个全连接层的隐藏大小分别为64和2。批量大小被设定为64。我们使用默认参数的Adam优化器[25]。学习率为 1×10^{-3} ，权重衰减为12。我们训练了50个轮，并选择其中获得最佳测试精度的历时作为最终结果，以避免过度拟合。

4 实验

4.1 实验设置

数据集：我们使用了三个从社交媒体上收集的真实数据集，即微博[22]、Gossipcop和Politifact[36]。在实验中，没有图片或没有文字的单模态新闻帖子被过滤掉了。如果一篇新闻文章包含一个带有多个相关图像的文本，我们会随机选择一个图像。微博是假新闻检测中广泛使用的中文数据集。训练集包含3,749条真实新闻和3,783条虚假新闻，测试集包含1,996条新闻。Politifact和Gossipcop数据集是从FakeNewsNet的政治和娱乐领域收集的两个英文数据集[36]存储库。Politifact在训练集中包含244条真实新闻和135条虚假新闻，在测试集中包含75条真实新闻和29条新闻。Gossipcop包含10,010条训练新闻，包括7,974条真实新闻和2,036条虚假新闻。测试集包含2285条真实新闻和545条虚假新闻。此外，虽然Twitter[6]也是FND著名的多模态数据集，我们发现它包含大量重复的帖子，超过10k的帖子仅包含463张图片。更重要的是，Twitter数据集上超过70%的推文与单一事件相关，这很容易导致模型过拟合。因此，我们不在推特上进行实验。

基线方法：为了公平和可重复的比较，我们必须有选择性地选择基线方法。首先，我们更喜欢提供预先训练好的模型或公开源代码的方法。第二，这些方法应该遵循一个通用的评估协议，其中三个数据集用于训练和测试。因此，我们比较了FNDCLIP与以下方法，并提供了一个快速回顾。

EANN[42]，它采用事件分类的辅助任务来提高可推广性。

MVAE[24]，它使用一个可变的自动编码器来模拟文本和图像之间的表示，用于假新闻检测

SpotFake[40]，分别用VGG和Bert来表示跟踪图像和文本特征并将它们连接起来进行分类。

MVNN[45]，该模型结合了文本语义特征、视觉篡改特征以及文本和视觉信息的相似性，用于假新闻检测。

SAFE[48]，将新闻文本和视觉信息之间的相关性输入到分类器中，以检测假新闻。

LIIMR[39]，它识别和抑制来自较弱模态的信息，并在每个样本的基础上从强模态提取相关信息。

MCAN[44]，其堆叠多个共同关注层以融合多模态特征。

CAFE[9]，它制定了一种模糊感知的多模态假新闻检测方法，以自适应地聚集单峰特征和跨模态相关性。

RoBERTa-MWSS[37]，它利用来自用户和内容约定的不同来源的多个弱信号。

Spotfake+[38]，这是Spotfake的改进版本，可以检测全长文章。TM[4]它利用真实和虚假新闻文本的词汇和语义属性来检测虚假新闻。

LSTM-ATT[27]，它建立了一个基于XGBoost的模型来检测完整长度的假新闻。

DistilBert[1]，它使用新闻文章和用户生成内容的潜在表示来指导模型学习。

表1显示了FNDCLIP在三个代表性数据集上的平均精度、召回率和准确度。实验结果表明，该方法在微博上的平均准确率达到90%以上，在politifact上的平均准确率达到94%以上，是一种可靠的，以及鲁棒的假新闻检测算法，可以检测给定多语言和多领域的新闻的异常。特别地，真实新闻在三个数据集上的召回率都在0.9以上，因此FNDCLIP不太可能将真实新闻归类为虚假新闻。

为了进一步进行统计分析，了解跨模态相似性如何与注意力得分相关联，以及它们在不同的新闻实例中如何变化，图3显示了基于CLIP的跨模态相似性得分与微博和Gossipcop数据集上假新闻比率之间的相关性。在行1和2中，我们根据相似性得分将每个数据集中的所有新闻分组到几个箱中，并且发现当相似性得分高时，

Table 1: Performance comparison between FND-CLIP and other methods on three datasets. Our method achieves the highest accuracy among these methods, and its precision, recall, and F1-score are also higher than most of the compared methods.

| Method | Accuracy | Fake News | | | Real News | | |
|-------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | | Precision | Recall | F1-score | Precision | Recall | F1-score |
| EANN [42] | 0.827 | 0.847 | 0.812 | 0.829 | 0.897 | 0.843 | 0.825 |
| MVAE [24] | 0.824 | 0.854 | 0.769 | 0.809 | 0.802 | 0.875 | 0.837 |
| Spotfake [40] | 0.892 | 0.902 | 0.964 | 0.932 | 0.847 | 0.656 | 0.739 |
| MVNN [45] | 0.846 | 0.809 | 0.857 | 0.832 | 0.879 | 0.837 | 0.858 |
| SAFE [48] | 0.762 | 0.831 | 0.724 | 0.774 | 0.695 | 0.811 | 0.748 |
| LiMB [39] | 0.900 | 0.882 | 0.823 | 0.847 | 0.968 | 0.941 | 0.925 |
| MCAN [44] | 0.899 | 0.913 | 0.889 | 0.901 | 0.884 | 0.909 | 0.897 |
| CAFE [9] | 0.840 | 0.855 | 0.830 | 0.842 | 0.825 | 0.851 | 0.837 |
| FND-CLIP | 0.907 | 0.914 | 0.901 | 0.908 | 0.914 | 0.901 | 0.907 |
| RoBERTa-MWSS [37] | 0.820 | - | - | - | 0.820 | - | - |
| SAFE [48] | 0.874 | 0.851 | 0.830 | 0.840 | 0.859 | 0.903 | 0.896 |
| Spotfake+ [38] | 0.846 | - | - | - | - | - | - |
| TM [4] | 0.871 | - | - | - | 0.901 | - | - |
| Politifact | 0.832 | 0.828 | 0.832 | 0.830 | 0.836 | 0.832 | 0.829 |
| LSTM-ATT [27] | 0.741 | 0.875 | 0.636 | 0.737 | 0.647 | 0.880 | 0.746 |
| DistilBert [1] | 0.864 | 0.724 | 0.778 | 0.750 | 0.895 | 0.919 | 0.907 |
| FND-CLIP | 0.942 | 0.897 | 0.897 | 0.897 | 0.960 | 0.960 | 0.960 |
| RoBERTa-MWSS [37] | 0.809 | - | - | 0.800 | - | - | - |
| SAFE [48] | 0.838 | 0.758 | 0.558 | 0.643 | 0.857 | 0.937 | 0.895 |
| Spotfake+ [38] | 0.856 | - | - | - | - | - | - |
| TM [4] | 0.842 | - | - | - | 0.896 | - | - |
| Gossipcop | 0.842 | 0.845 | 0.842 | 0.844 | 0.839 | 0.842 | 0.821 |
| LSTM-ATT [27] | 0.857 | 0.805 | 0.527 | 0.637 | 0.866 | 0.960 | 0.911 |
| DistilBert [1] | 0.867 | 0.732 | 0.490 | 0.587 | 0.887 | 0.957 | 0.921 |
| FND-CLIP | 0.880 | 0.761 | 0.549 | 0.638 | 0.899 | 0.959 | 0.928 |

表1显示了FNDCLIP在三个代表性数据集上的平均精度、召回率和准确度。实验结果表明，该方法在微博上的平均准确率达到90%以上，在politifact上的平均准确率达到94%以上，是一种可靠的，以及鲁棒的假新闻检测算法，可以检测给定多语言和多领域的新闻的异常。特别地，真实新闻在三个数据集上的召回率都在0.9以上，因此FNDCLIP不太可能将真实新闻归类为虚假新闻。

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

Anonymous et al.

以上，在politifact上的平均准确率达到94%以上，是一种可靠的，以及鲁棒的假新闻检测算法，可以检测给定多语言和多领域的新闻的异常。特别地，真实新闻在三个数据集上的召回率都在0.9以上，因此FNDCLIP不太可能将真实新闻归类为虚假新闻。

为了进一步进行统计分析，了解跨模态相似性如何与注意力得分相关联，以及它们在不同的新闻实例中如何变化，图3显示了基于CLIP的跨模态相似性得分与微博和Gossipcop数据集上假新闻比率之间的相关性。在行1和2中，我们根据相似性得分将每个数据集中的所有新闻分组到几个箱中，并且发现当相似性得分高时，新闻更有可能是真的。在第3行，我们计算每个箱的真实新闻率，并将其减去平均真实新闻率。用相应数据集的平均真实新闻率减去它们。曲线显示，真实新闻率随着微博上模糊性的增加而上升。曲线显示，微博上的真实新闻率随着模糊性的增加而上升，而在Gossipcop上先是下降，然后激增。在Gossipcop上则是先下降后上升。这样的统计特征有益于深度网络来识别假新闻。

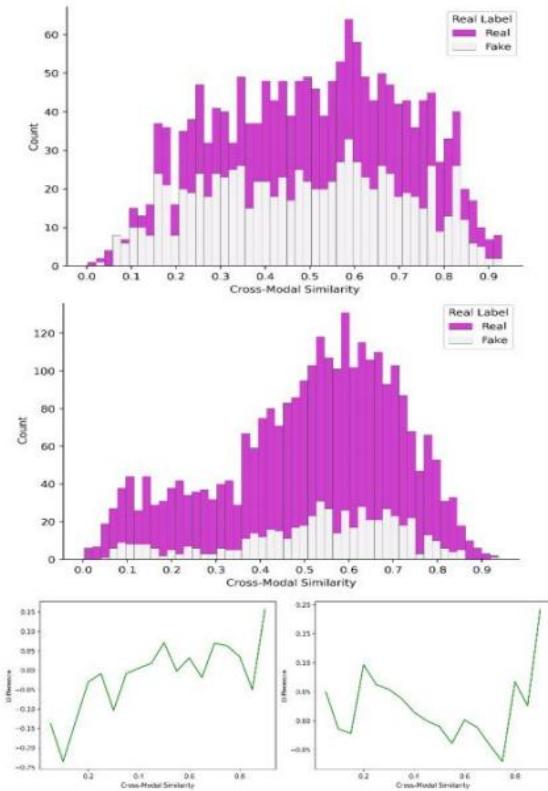


图3：对不同新闻的跨模式相似性的统计分析。第一行和第二行分别显示了根据微博和Gossipcop的跨模态相似度对真/假新闻进行统计。第三行显示了每个分类中的真实新闻率与相应数据集上的平均率之间的距离。与相应数据集上的平均比率的距离（左：微博；右：Gossipcop）。

新闻更有可能是真的。在第3行，我们计算每个箱的真实新闻率，并将其减去平均真实新闻率。用相应数据集的平均真实新闻率减去它们。曲线显示，真实新闻率随着微博上模糊性的增加而上升。曲线显示，微博上的真实新闻率随着模糊性的增加而上升，而在Gossipcop上先是下降，然后激增。在Gossipcop上则是先下降后上升。这样的统计特征有益于深度网络来识别假新闻。

4.3 与先进技术的比较

我们进一步将FND-CLIP与上述最先进的方法进行比较，比较结果列于表1。'-'表示原始论文中没有这些结果。如表1所示，FND-CLIP在三个数据集上的准确度方面优于所有被比较的方法，而在微博上在召回率方面略低于微博上的Spot。FND-CLIP实现了最高的准确率分别为90.7%、94.2%和88.0%，超过了最先进的方法0.7%、6.8%和1.3%。分别比最先进的方法高出0.7%、6.8%和1.3%，在三个真实世界的数据集，分别超过了0.7%、6.8%和1.3%。此外，在所有的测试中，我们的精度、召回率和准确率都排名第一或第二，这证明了FND-CLIP的有效性。许多假新闻检测方法，如EANN和Spotfake。

只依靠直接使用连接或关注机制获得的融合特征。然而，这些融合的特征不能提供足够的辨别能力来对假新闻进行分类，因为单独提取的文本和图像特征并不在同一个语义空间内，而且相关性也不高。融合过程中，文本和图像的相关信息没有得到很好的关注。因此，这些方法的实验结果并不令人满意。CAFE使用跨模态对齐来训练编码器，这些编码器可以将文本和图像映射到相同的语义空间。可以将文本和图像映射到相同的语义空间。通过使用融合的文本和图像的特征来进行分类。它取得了良好的实验结果，特别是在Politifact和Gossipcop数据集。然而，由于数据集数量的限制和训练标签的粗略方法编码器的编码效果并不理想，而且文本和图像特征之间的语义差距很大。此外CAFE设计了一个模糊性学习模块来计算自适应调整不同模式的计算的权重。然而，选择单模态或多模态特征的权重是通过手工计算得到的，不能通过反向梯度传播进一步优化。从而影响了检测的性能。

FND-CLIP优于大多数最先进的方法。主要是由于以下原因。首先，FND-CLIP中预先训练好的CLIP编码器可以在同一语义空间内生成语义信息丰富的文本和图像特征，确保融合后的特征正确反映文本和图像之间的关联性，并为文本和图像提供补充信息。模式明智的注意机制自适应地决定了文本、图像和融合特征的权重，避免了无效特征对表示的影响。进一步提高了分类精度。

4.4 消融研究

我们通过评估具有不同和部分设置的模型的性能来探索FNDCLIP中的关键组件的影响。在每个测试中，我们移除不同的组件并从头开始训练模型。FND消融的比较变体实现如下。

- FNDCLIP w/o A。我们删除模态方式的注意模块，并直接聚合三个特征，以获得最终的特征；
- FNDCLIP w/o F。我们去掉融合模块，使用两个单峰特征对新闻进行分类；
- FNDCLIP w/o C。我们删除所有CLIP相关的模块，只使用BERT和ResNet来提取文本和图像特征。
- 仅FND-CLIP多模态：我们去除了单峰特征提取器，BERT和ResNet，并且仅使用CLIP融合特征作为最终特征；

Multimodal Fake News Detection via CLIP-Guided Learning

- 仅FNDCLIP图像：去除所有与文本相关的特征，仅使用ResNet提取的图像特征进行分类；
- 仅FNDCLIP文本：我们仅使用Bert提取特征来完成检测任务，不需要任何视觉信息。

每个组件的有效性。首先，我们分析了FNDCLIP中不同成分对假新闻检测的影响。从表中所示的结果来看²，我们有以下观察结果：

1) FND-CLIP优于FND-CLIP w/o C，证明CLIP可以有效地为假新闻检测任务提供可辨别的特征，并显著提高分类的准确性。虽然只有模态内特征可以用于分类，但是模态之间缺乏交互使得最终的特征缺乏表示图像和文本之间内在关系的能力。2) andCLIP的性能优于FND-CLIP，表明虽然单峰分支包含CLIP编码的特征，但反映文本和图像相关性的融合特征为分类器提供了有效的多峰信息。同时，FND-CLIP w/o F优于FND-CLIP w/o C，表明使用CLIP编码特征对单峰特征的补充是有效的。3) FNDCLIP在微博和Gossipcop上的表现优于FNDCLIP，这表明模态式注意可以帮助FNDCLIP自适应地加权有用的模态。FNDCLIP w/o A直接融合不同模态的特征，这可能导致最终特征受到来自模态的无效信息的影响。不同模式的贡献。第二组实验是评估不同模态在假新闻检测中的分类性能。来自表格2，我们得出如下一些分析：1) FND-仅CLIP图像-表现最差，特别是在Gossipcop数据集上，其中假新闻的F1分几乎为零，这意味着所有新闻都被判断为真实的，并且该模型根本没有分类能力。这表明，在假新闻检测中，简单的视觉信息提供的分类较少。

2) FND-CLIP multimodal-only 在微博、Politifact 和 Gossipcop 数据集上分别达到 81.7%、90.3% 和 86.2% 的准确率，但在微博和Gossipcop数据集上的表现不如FND-CLIP text-only，表明图像和文本的相关性信息可以用于假新闻的分类。然而，由于新闻本身具有模态无关性和模糊性，融合特征的分类能力是有限的。此外，基于CLIP的融合特征侧重于文本的语义，而基于BERT的文本特征也提取了有助于假新闻检测的情感特征。3) FND-CLIP text-only 取得了第二好的结果，表明仅使用文本特征就可以基本完成对假新闻的分类任务。然而，FNDCLIP优于FNDCLIP的纯文本，证明视觉特征可以补充分类信息，正确使用多模态特征优于只使用单峰特征的假新闻检测。

4.5 T-SNE可视化

在图中4，我们使用t-SNE进一步分析了所提出的方法⁴¹在图中的微博测试数据集上，FNDCLIP、CAFE以及所提出的具有部分设置的方法在分类器之前学习的特征的可视化，所述部分设置诸如FNDCLIPw/o C、FNDCLIPw/o A、FNDCLIP纯文本以及FNDCLIP纯图像。

具有相同颜色的点意味着它们在同一标签内。从图中4 我们可以看到，FNDCLIP中不同标签点的边界比CAFE、FNDCLIPw/o C和FNDCLIPw/o A更明显，这表明FNDCLIP中提取的特征比CAFE中提取的特征更具区分性，CLIP相关的模块和模态注意有助于提高FNDCLIP的分类能力。

此外，通过对比图4a, 图4d, 图4e, 和图4f, 我们可以看到，光靠图像特征是不足以进行分类的，这说明图像本身不具备分类能力。纯文本的效果比纯图像的效果好得多，证明文本特征在假新闻检测中起主导作用，但仍有很多样本点无法区

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

分。包含文本和图像特征的FNDCLIPw/o C比仅包含文本的FNDCLIP具有更明显的点边界，这表明不同的模态具有互补的信息。此外，图中样本点的分离度4a 比图中的要高4d, 表明基于CLIP的多模态特征能够提高最终特征的表示能力。

5 结论

本文提出了一种新的多模态假新闻检测方法，称为FND-CLIP，该方法使用CLIP提取对齐的多模态特征，并指导不同模态的网络学习。此外，我们还引入了基于模态的注意力来自适应地确定文本、图像和融合特征的权重。它可以避免在特征融合过程中引入噪声和冗余特征，进一步提高分类精度。我们在几个著名的FND数据集上进行了全面的实验。结果表明，使用CLIP进行多模态特征生成可以很好地与单峰特征协作由ResNet和BERT在挖掘假新闻检测的关键特征时提取。更重要的是，在多模态假新闻检测方面，FNDCLIP优于许多最先进的方法。除了FNDCLIP的性能增益之外，输出仍然是预测“真实”或“虚假”的二进制值的形式，这无法以某种方式解释为什么新闻被预测为虚假，以及新闻中的哪些元素最可疑和异常。在未来的工作中，我们将致力于开发更多可解释的假新闻检测系统，它可以提供给定新闻被预测为真或假的原因。

| Method | Accuracy | Fake News | | | Real News | | | |
|------------|--------------------------|-----------|--------|----------|-----------|--------|----------|-------|
| | | Precision | Recall | F1-score | Precision | Recall | F1-score | |
| Weibo | FND-CLIP multimodal-only | 0.817 | 0.899 | 0.718 | 0.798 | 0.761 | 0.917 | 0.832 |
| | FND-CLIP image-only | 0.796 | 0.862 | 0.711 | 0.779 | 0.750 | 0.884 | 0.811 |
| | FND-CLIP text-only | 0.872 | 0.906 | 0.833 | 0.868 | 0.842 | 0.911 | 0.875 |
| | FND-CLIP w/o C | 0.874 | 0.895 | 0.851 | 0.872 | 0.855 | 0.898 | 0.876 |
| | FND-CLIP w/o F | 0.893 | 0.925 | 0.857 | 0.890 | 0.864 | 0.929 | 0.895 |
| | FND-CLIP w/o A | 0.897 | 0.936 | 0.855 | 0.893 | 0.863 | 0.940 | 0.900 |
| | FND-CLIP | 0.907 | 0.914 | 0.901 | 0.908 | 0.901 | 0.914 | 0.907 |
| Politifact | FND-CLIP multimodal-only | 0.905 | 0.807 | 0.862 | 0.833 | 0.944 | 0.919 | 0.932 |
| | FND-CLIP image-only | 0.748 | 0.600 | 0.310 | 0.409 | 0.773 | 0.919 | 0.840 |
| | FND-CLIP text-only | 0.905 | 0.913 | 0.724 | 0.808 | 0.900 | 0.973 | 0.935 |
| | FND-CLIP w/o C | 0.893 | 0.875 | 0.724 | 0.793 | 0.899 | 0.960 | 0.928 |
| | FND-CLIP w/o F | 0.903 | 0.880 | 0.759 | 0.815 | 0.910 | 0.960 | 0.934 |
| | FND-CLIP w/o A | 0.942 | 0.926 | 0.862 | 0.893 | 0.947 | 0.973 | 0.960 |
| | FND-CLIP | 0.942 | 0.897 | 0.897 | 0.897 | 0.960 | 0.960 | 0.960 |
| Gossipcop | FND-CLIP multimodal-only | 0.862 | 0.708 | 0.484 | 0.575 | 0.886 | 0.952 | 0.918 |
| | FND-CLIP image-only | 0.814 | 1.000 | 0.033 | 0.064 | 0.813 | 1.000 | 0.897 |
| | FND-CLIP text-only | 0.871 | 0.741 | 0.508 | 0.603 | 0.891 | 0.958 | 0.923 |
| | FND-CLIP w/o C | 0.870 | 0.745 | 0.494 | 0.594 | 0.888 | 0.960 | 0.923 |
| | FND-CLIP w/o F | 0.874 | 0.723 | 0.562 | 0.632 | 0.901 | 0.949 | 0.924 |
| | FND-CLIP w/o A | 0.873 | 0.715 | 0.567 | 0.633 | 0.902 | 0.946 | 0.923 |
| | FND-CLIP | 0.880 | 0.761 | 0.549 | 0.638 | 0.899 | 0.959 | 0.928 |

表2：在三个数据集上对FND-CLIP的架构设计和不同特征的消融研究。FND-CLIP取得了最高的准确率和F1分数，表明我们的方法架构中的每个模块都是有效的。

REFERENCES

- [1] Liesbeth Allein, Marie-Francine Moens, and Domenico Perrotta. 2021. Like Article, Like Audience: Enforcing Multimodal Correlations for Disinformation Detection. *arXiv preprint arXiv:2108.13892* (2021).
- [2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multi-modal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.
- [3] Gaurav Bhatt, Aman Sharma, Shivam Sharma, Ankush Nagpal, Balasubramanian Raman, and Ankush Mittal. 2018. Combining neural, statistical and external features for fake news stance identification. In *Companion Proceedings of the The Web Conference 2018*. 1353–1357.
- [4] Bimal Bhattacharjee, Ole-Christoffer Granmo, and Lei Jiao. 2021. Explainable Tsetlin Machine framework for fake news detection with credibility score assessment. *arXiv preprint arXiv:2105.09114* (2021).
- [5] Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 549–556.
- [6] Christina Boididou, Symeon Papadopoulos, Markos Zampoglou, Lazaros Aposto Iidis, Olga Papadopoulou, and Yiannis Kompatsiaris. 2018. Detection and visualization of misleading content on Twitter. *International Journal of Multimedia Information Retrieval* 7, 1 (2018), 71–86.
- [7] Juan Cao, Peng Qi, Qiang Sheng, Tianyun Yang, Junbo Guo, and Jintao Li. 2020. Exploring the role of visual content in fake news detection. *Disinformation, Misinformation, and Fake News in Social Media* (2020), 141–161.
- [8] Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. 2021. Image Manipulation Detection by Multi-View Multi-Scale Supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14185–14193.
- [9] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal Ambiguity Learning for Multimodal Fake News Detection. In *Proceedings of the ACM Web Conference 2022*. 2897–2905.
- [10] Marcos V Conde and Kerem Turgutlu. 2021. CLIP-Art: contrastive pre-training for fine-grained art classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3956–3960.
- [11] Nadia K Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the association for information science and technology* 52, 1 (2015), 1–4.
- [12] Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. 2021. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1574–1583.
- [13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [14] Han Guo, Juan Cao, Yazi Zhang, Junbo Guo, and Jintao Li. 2018. Rumor detection with hierarchical social attention network. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 943–951.
- [15] Bing Han, Xiaoguang Han, Hua Zhang, Jingzhi Li, and Xiaochun Cao. 2021. Fighting fake news: two stream network for deepfake detection via learnable SRM. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3, 3 (2021), 320–331.
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [18] Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* (2016).
- [19] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [20] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*. 448–456.
- [21] Ganesh Jawahar, Benoit Sagot, and Djamel Seddah. 2019. What does BERT learn about the structure of language?. In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.
- [22] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*. 795–816.
- [23] Gregory Johnson. 2012. Google Translate <http://translate.google.com>. Technical Services Quarterly 29, 2 (2012), 165–165.
- [24] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*. 2915–2921.
- [25] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [26] Peigang Li, Xian Sun, Hongfeng Yu, Yu Tian, Fanglong Yao, and Guangluan Xu. 2021. Entity-Oriented Multi-Modal Alignment and Fusion Network for Fake News Detection. *IEEE Transactions on Multimedia* (2021).
- [27] Jun Lin, Glenna Tremblay-Taylor, Guanyi Mou, Di You, and Kyumin Lee. 2019. Detecting fake news articles. In *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 3021–3025.
- [28] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* 32 (2019).
- [29] Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. MD-FEND: Multi-domain Fake News Detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3343–3347.
- [30] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).
- [31] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638* (2017).
- [32] Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. 2021. Improving Fake News Detection by Using an Entity-enhanced Framework to Fuse Diverse Multimodal Clues. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1212–1220.
- [33] Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. 2019. Exploiting multi-domain visual information for fake news detection. In *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 518–527.
- [34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.
- [35] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* (2019).
- [36] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnets: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data* 8, 3 (2020), 171–188.
- [37] Kai Shu, Guoqing Zheng, Yichuan Li, Subhabrata Mukherjee, Ahmed Hassan Awadallah, Scott Ruston, and Huan Liu. 2020. Leveraging multi-source weak social supervision for early detection of fake news. *arXiv preprint arXiv:2004.01732* (2020).
- [38] Shrivangi Singhal, Anubha Kabra, Mohit Sharma, Rajiv Ratn Shah, Tanmoy Chakraborty, and Ponnurangam Kumaraguru. 2020. Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13915–13916.
- [39] Shrivangi Singhal, Tanisha Pandey, Saksham Mrig, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2022. Leveraging Intra and Inter Modality Relationship for Multimodal Fake News Detection. (2022).
- [40] Shrivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Ku

Multimodal Fake News Detection via CLIP-Guided Learning

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

- maraguru, and Shin’ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*. IEEE, 39–47.
- [41] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
- [42] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 849–857.
- [43] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Zhentao Tan, Lu Yuan, Weiming Zhang, and Nenghai Yu. 2021. Hairclip: Design your hair by text and reference image. *arXiv preprint arXiv:2112.05142* (2021).
- [44] Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal Fusion with Co-Attention Networks for Fake News Detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2560–2569.
- [45] Junxiao Xue, Yabo Wang, Yichen Tian, Yafei Li, Lei Shi, and Lin Wei. 2021. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management* 58, 5 (2021), 102610.
- [46] Keren Ye and Adriana Kovashka. 2021. A case study of the shortcut effects in visual commonsense reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 3181–3189.
- [47] Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the Web Conference 2021*. 3465–3476.
- [48] Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: Similarity-Aware Multi-modal Fake News Detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 354–367.
- [49] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)* 51, 2 (2018), 1–36.

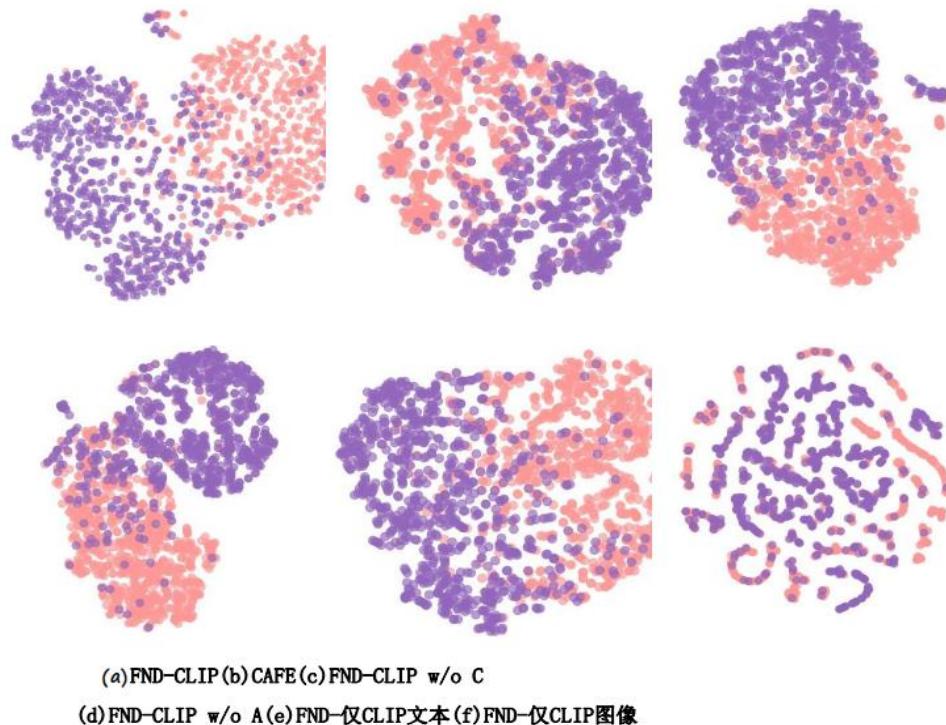


图4:在微博的测试数据集上,由FNDCLIP、CAFE、FNDCLIP w/o C、FNDCLIP w/o A、FNDCLIP纯文本和FNDCLIP纯图像学习的分类器之前的特征的T-SNE可视化。

附录 C 部分源代码文件

```
import torch

from transformers import BertModel, BertConfig

device = 'cuda' if torch.cuda.is_available() else 'cpu'

pretrained = BertModel.from_pretrained('bert-base-chinese')
pretrained.to(device)

for param in pretrained.parameters():
    param.requires_grad_(False)

class Bert_pretrain(torch.nn.Module):

    def __init__(self):
        super().__init__()

    def forward(self, input_ids, attention_mask, token_type_ids):
        with torch.no_grad():
            out = pretrained(input_ids=input_ids, attention_mask=attention_mask,
                             token_type_ids=token_type_ids)
        return out.last_hidden_state[:, 0]

from transformers import ChineseCLIPVisionConfig, ChineseCLIPVisionModel
import torch
```

```
device = 'cuda' if torch.cuda.is_available() else 'cpu'
print('device=', device)

pretrained =
ChineseCLIPVisionModel.from_pretrained("OFA-Sys/chinese-clip-vit-base-patch16")
pretrained.to(device)

for param in pretrained.parameters():
    param.requires_grad_(False)

class ClipImgFeat(torch.nn.Module):

    def __init__(self):
        super(ClipImgFeat, self).__init__()

    def forward(self, Img):
        with torch.no_grad():
            out = pretrained(Img)
        return out.last_hidden_state[:, 0]

import requests
from PIL import Image
from transformers import ChineseCLIPModel, ChineseCLIPProcessor
import torch

device = 'cuda' if torch.cuda.is_available() else 'cpu'
print('device=', device)

pretrained =
ChineseCLIPModel.from_pretrained("OFA-Sys/chinese-clip-vit-base-patch16")
pretrained.to(device)

for param in pretrained.parameters():
    param.requires_grad_(False)

class ClipSim(torch.nn.Module):
```

```

def __init__(self):
    super(ClipSim, self).__init__()

    def forward(self, inputs):
        with torch.no_grad():
            out = pretrained(**inputs)
            logits_per_image = out.logits_per_image
            probs = logits_per_image.softmax(dim=1)
        return probs

from transformers import CLIPTokenizer, ChineseCLIPTextModel,
ChineseCLIPTextConfig
import torch

device = 'cuda' if torch.cuda.is_available() else 'cpu'
print('device=', device)

pretrained =
ChineseCLIPTextModel.from_pretrained("OFA-Sys/chinese-clip-vit-base-patch16")
pretrained.to(device)

for param in pretrained.parameters():
    param.requires_grad_(False)

class ClipTextFeat(torch.nn.Module):

    def __init__(self):
        super(ClipTextFeat, self).__init__()

        def forward(self, input_ids, attention_mask, token_type_ids):
            with torch.no_grad():
                out = pretrained(input_ids=input_ids, attention_mask=attention_mask,
token_type_ids=token_type_ids, output_hidden_states=True)
            return out.last_hidden_state[:, 0]
        import torch

        import torch.nn as nn

        import torch.nn.functional as F

```

```
topic_num = 100
vec_size = 40535

device = 'cuda'

class VAE(nn.Module):

    def __init__(self):
        super(VAE, self).__init__()

        #encoder:
        self.fc1 = nn.Sequential(nn.Linear(vec_size, 768), nn.BatchNorm1d(768))
        self.fc2 = nn.Sequential(nn.Linear(768, 256), nn.BatchNorm1d(256))
        self.mean = nn.Linear(256, topic_num)
        self.var = nn.Linear(256, topic_num)

        #decoder:
        self.fc4 = nn.Linear(topic_num, 256)
        self.fc5 = nn.Sequential(nn.Linear(256, 768), nn.BatchNorm1d(768))
        self.fc6 = nn.Linear(768, vec_size)

    def encoder(self, x):
        x = F.relu(self.fc1(x))
        x = F.relu(self.fc2(x))
```

```
return self.mean(x), self.var(x)

def decoder(self, z):
    z = F.relu(self.fc4(z))
    z = F.relu(self.fc5(z))
    out = F.sigmoid(self.fc6(z))

    return out

def reparameterize(self, mu, log_var):
    std = torch.exp(log_var/2)
    eps = torch.randn_like(std)
    return mu + eps * std

def forward(self, inputs):
    mean, log_var = self.encoder(inputs)

    z = self.reparameterize(mean, log_var)

    inputs_hat = self.decoder(z)

from transformers import ViTModel, ViTConfig
import torch

device = 'cuda' if torch.cuda.is_available() else 'cpu'
print('device=', device)

pretrained = ViTModel.from_pretrained("google/vit-base-patch16-224-in21k")
```

```
pretrained.to(device)

for param in pretrained.parameters():
    param.requires_grad_(False)

class ViTImgFeat(torch.nn.Module):

    def __init__(self):
        super(ViTImgFeat, self).__init__()

    def forward(self, Img):
        with torch.no_grad():
            out = pretrained(Img)
        return out.last_hidden_state[:, 0]

import torch

import torch.nn as nn

import torch.nn.functional as F

device = 'cuda' if torch.cuda.is_available() else 'cpu'

print('device=', device)

class RumorDetectionModel(torch.nn.Module):

    def __init__(self):
        super(RumorDetectionModel, self).__init__()

        self.text_projection_1 = nn.Sequential(nn.Linear(768 * 2, 512),
                                              nn.BatchNorm1d(512))
```

```
    self.text_projection_2 = nn.Sequential(nn.Linear(512, 128),  
nn.BatchNorm1d(128))
```

```
    self.img_projection_1 = nn.Sequential(nn.Linear(768 * 2, 512),  
nn.BatchNorm1d(512))
```

```
    self.img_projection_2 = nn.Sequential(nn.Linear(512, 128),  
nn.BatchNorm1d(128))
```

```
    self.fused_projection_1 = nn.Sequential(nn.Linear(768 * 2, 512),  
nn.BatchNorm1d(512))
```

```
    self.fused_projection_2 = nn.Sequential(nn.Linear(512, 128),  
nn.BatchNorm1d(128))
```

```
self.cross_att = nn.MultiheadAttention(embed_dim=128, num_heads=8)
```

```
self.ahead_att = nn.MultiheadAttention(embed_dim=128 + 4, num_heads=4)
```

```
    self.self_att = nn.MultiheadAttention(embed_dim=128 + 4 + 100,  
num_heads=8)
```

```
    self.pre_att = nn.MultiheadAttention(embed_dim=768*2, num_heads=8)
```

```
self.mlp = nn.Sequential(nn.Linear(100 + 4, 128), nn.BatchNorm1d(128))
```

```
self.fc1 = nn.Sequential(nn.Linear(128 + 4 + 100, 64), nn.BatchNorm1d(64))
```

```
self.fc2 = nn.Sequential(nn.Linear(64, 1))
```

```
self.fc3 = nn.Sequential(nn.Linear(64, 2))
```

```
self.drop_out = nn.Dropout(0.2)
```

```
def forward(self, bert_text, vit_img, clip_text, clip_img, clip_sim, ntm,  
bert_input_ids, bert_attention_mask, bert_token_type_ids, clip_input_ids,
```

clip_attention_mask, clip_token_type_ids, vit_imgs_tensor, clip_imgs_tensor, clip_sim_feat, bow_tensor, senti_vec):

#text models:

```
bert_tensor = bert_text(input_ids=bert_input_ids,  
attention_mask=bert_attention_mask, token_type_ids=bert_token_type_ids) #[16,768]
```

```
clip_text_tensor = clip_text(input_ids=clip_input_ids,  
attention_mask=clip_attention_mask, token_type_ids=clip_token_type_ids) #[16,768]
```

#image models:

```
vit_tensor = vit_img(vit_imgs_tensor) #[16,768]
```

```
clip_image_tensor = clip_img(clip_imgs_tensor) #[16,768]
```

#fused models:

```
sim_tensor = clip_sim(clip_sim_feat) #[16,16]
```

```
sim_weight, _ = sim_tensor.max(1) #[16,1]
```

```
sim_weight = sim_weight.reshape((64, 1))
```

```
sim_weight = sim_weight.expand(64, 128)
```

#weights:

```
unimodal_weight = 1
```

#ntm

```
inputs_hat, mean, log_var, z = ntm(bow_tensor) #[16,40535]
```

#senti

```
sentiment = senti_vec #[16,4]
```

```
#concat

text_feat = torch.cat((bert_tensor, clip_text_tensor), dim=1) #[16,768*2]

img_feat = torch.cat((vit_tensor, clip_image_tensor), dim=1) #[16,768*2]

fused_feat = torch.cat((clip_text_tensor, clip_image_tensor), dim=1)

#[16,768*2]

#projection

text_feat = F.relu(self.text_projection_1(text_feat)) #[16,128]

text_feat = self.drop_out(text_feat)

text_feat = F.relu(self.text_projection_2(text_feat))

img_feat = F.relu(self.img_projection_1(img_feat)) #[16,128]

img_feat = self.drop_out(img_feat)

img_feat = F.relu(self.img_projection_2(img_feat))

fused_feat = F.relu(self.fused_projection_1(fused_feat)) #[16,128]

fused_feat = self.drop_out(fused_feat)

fused_feat = F.relu(self.fused_projection_2(fused_feat))

#get_weighted_fused

fused_weighted_feat = fused_feat * sim_weight #[16,128]

#Transpose_n_reshape

text_feat = text_feat.reshape((64, 1, 128))

img_feat = img_feat.reshape((64, 1, 128))

text_feat_trans = torch.transpose(text_feat, 0, 1)
```

```
img_feat_trans = torch.transpose(img_feat, 0, 1)
```

#text_image_cross_attention

```
txt_attn_feat, txt_attn_weights = self.cross_att(text_feat_trans, img_feat_trans,
```

```
img_attn_feat, img_attn_weights = self.cross_att(img_feat_trans,
```

```
text_feat_trans, text_feat_trans)
```

#Transpose

```
txt_attn_feat = torch.transpose(txt_attn_feat, 0, 1)
```

```
img_attn_feat = torch.transpose(img_attn_feat, 0, 1)
```

```
txt_attn_feat = txt_attn_feat.squeeze() #[16, 128]
```

```
img_attn_feat = img_attn_feat.squeeze() #[16, 128]
```

#weighted_sum

```
feat = fused_weighted_feat + txt_attn_feat * unimodal_weight +
```

```
img_attn_feat * unimodal_weight #[16,128]
```

two-step-attention

```
feat_n_senti = torch.cat((feat, sentiment), dim=1) #[64, 128 + 4]
```

```
feat_n_senti = feat_n_senti.reshape((64, 1, 128 + 4))
```

```
feat_n_senti = torch.transpose(feat_n_senti, 0, 1)
```

```
feat_n_senti, _ = self.ahead_att(feat_n_senti, feat_n_senti, feat_n_senti)
```

```
feat_n_senti = torch.transpose(feat_n_senti, 0, 1) #[16, 128 + 4]
```

```
feat_n_senti = feat_n_senti.squeeze()
```

```
feat_inte = torch.cat((feat_n_senti, z), dim=1) #[16, 128 + 4 + 100]
feat_inte = feat_inte.reshape((64, 1, 128 + 4 + 100))
feat_inte = torch.transpose(feat_inte, 0, 1)
feat_attn_inte, feat_attn_inte_weights = self.self_att(feat_inte, feat_inte,
feat_inte)
feat_attn_inte = torch.transpose(feat_attn_inte, 0, 1) #[16, 128 + 4 + 100]
feat_attn_inte = feat_attn_inte.squeeze()
"""

# let senti_n_topic be query

senti_n_topic = torch.cat((senti, z), dim=1) #[64, 100 + 4]
senti_n_topic = senti_n_topic.reshape((64, 1, 100 + 4))
senti_n_topic = torch.transpose(senti_n_topic, 0, 1)
senti_n_topic, _ = self.ahead_att(senti_n_topic, senti_n_topic, senti_n_topic)
senti_n_topic = torch.transpose(senti_n_topic, 0, 1) #[64, 100 + 4]
senti_n_topic = senti_n_topic.squeeze()

senti_n_topic = F.relu(self.mlp(senti_n_topic)) #[64, 128]

senti_n_topic = senti_n_topic.reshape((64, 1, 128))
senti_n_topic = torch.transpose(senti_n_topic, 0, 1)
feat_attn_inte, feat_attn_inte_weights = self.self_att(senti_n_topic, feat, feat)
feat_attn_inte = torch.transpose(feat_attn_inte, 0, 1) #[16, 128 + 4 + 100]
feat_attn_inte = feat_attn_inte.squeeze()
"""

"""


```

```
#integrate_with_senti_n_topic  
feat_inte = torch.cat((feat, sentiment, z), dim=1) #[16, 128 + 4 + 100]  
  
#transform_n_reshape  
feat_inte = feat_inte.reshape((64, 1, 128 + 4 + 100))  
feat_inte = torch.transpose(feat_inte, 0, 1)  
  
#self_attentoin  
feat_attn_inte, feat_attn_inte_weights = self.self_att(feat_inte, feat_inte,  
feat_inte)  
  
#Transpose_n_squeeze  
feat_attn_inte = torch.transpose(feat_attn_inte, 0, 1) #[16, 128 + 4 + 100]  
feat_attn_inte = feat_attn_inte.squeeze()  
"  
  
#MLP  
out = F.relu(self.fc1(feat_attn_inte))  
out = self.drop_out(out)  
#out = F.sigmoid(self.fc2(out))  
out = self.fc3(out)  
  
return out, mean, log_var, inputs_hat, feat_attn_inte
```

附录 D 毕业设计（论文）任务书

西安交通大学

系(专业) 0504 自动化

系(专业)主任 张爱民

批准日期 2022年10月26日

毕业设计(论文)任务书

电子与信息学部 学院 0504 自动化 专业 自动化 96 班 学生 梁珉珲

毕业设计(论文)课题 社交媒体多模态虚假信息检测技术研究

毕业设计(论文)工作自 2022 年 9 月 26 日起至 2023 年 6 月 14 日止

课题的背景、意义及培养目标

随着信息技术的发展，各类社交媒体平台逐渐流行，成为人们浏览信息、分享生活的重要工具。然而，由于其操作的简易性与应用的广泛性，虚假信息很容易在社交媒体中编辑发布并吸引大量关注，使得政府、各类机构或个人名誉受损，财产流失，对社会和谐稳定造成极为不良的影响。因此，社交媒体虚假信息检测技术成为了学者讨论与研究的热点，对经济与社会健康的发展具有深远的意义。目前主流的社交媒体虚假信息检测方法可大体分为两类：基于文本内容的虚假信息检测与基于社交上下文的虚假信息检测。其中，基于文本内容的虚假信息检测通过对虚假信息文本进行挖掘把握虚假信息的写作风格，基于社交上下文的虚假信息检测通过分析源用户信息以及传播链条把握虚假信息的转发特征。本课题综合利用上文所述的两种主流社交媒体虚假信息检测方法构建检测模型，同时在虚假信息文本中引入图片信息使多模态信息互相补充融合，便于神经网络更好地提取利于分类的特征。综上所述，基于深度学习的社交媒体多模态虚假信息检测技术研究具有极高的研究价值与实际意义。

设计(论文)的原始数据与资料

1) 编程语言与框架等实现技术的书籍;2) 文本挖掘、多模态融合相关算法。

课题的主要任务

本课题主要工作内容分为以下三部分：1 数据获取与处理：计划使用 MediaTwitter、MediaWeibo 或 Weibo20 等包含多模态信息的数据集，并进行数据清洗等预备工作。

2 模型构建与调试：使用 CLIP 等预训练模型对文本与图片信息进行特征提取与模态融合，使用图神经网络建模社交上下文信息，对虚假信息特征进行全方面提取，最终利用分类器进行预测。成功构建模型后在验证集中试验最佳参数。
3 设计实验并撰写论文：设计基线方法证明模型有效性，设计消融实验证明模块有效性，撰写论文总结模型设计思路、方法与实验结果。

课题的基本要求(工程设计类题应有技术经济分析要求)

1) 检索并阅读相关文献；2) 论文和外文翻译要符合原意，条理清晰，文句通顺；3) 严禁抄袭、剽窃他人成果，树立和养成高尚文明的科研风气及道德修养；4) 按照本科毕业设计（论文）要求，按时提交各阶段文档、最终代码及毕业论文。

完成任务后提交的书面材料要求(图纸规格、数量，论文字数，外文翻译字数等)

提交合格、高质量的毕业设计论文（15000 字以上），包含以下主要内容：1) 研究背景、现状；2) 本论文完成的工作、解决的问题；3) 算法的设计思想与测试结果；4) 技术创新方面等。提交 3000 字外文文献翻译。

主要参考文献

社交媒体虚假信息检测技术的相关文献

指导教师 _____ 史施

接受设计(论文)任务日期 _____ 2022 年 10 月 26 日

(注：由指导教师填写) 学生签名 _____

附录 E 毕业设计（论文）考核评议书

西 安 交 通 大 学

毕业设计(论文)考核评议书

电子与信息学部 学院 0504 自动化 专业 自动化 96 班

指导教师对学生 梁珉珲 所完成的课题为 社交媒体多模态虚假信息检测技术研究 的毕业设计(论文)进行的情况，完成的质量及评分的意见： 梁珉珲同学在整个毕业设计期间学习认真、工作努力。通过自学大量相关开发工具并阅读相关参考文献，了解了虚假信息检测技术研究，设计了一种多模态社交媒体虚假信息检测深度神经网络结构，通过在 Weibo 数据集上的实验验证了其有效性。设计合理，实验数据准确，较好地完成了本科毕业设计任务。有一定的实际动手能力。

论文写作认真、格式规范，翻译准确。建议成绩：B+。

指导教师建议成绩： B+

指导教师 史施

2023 年 06 月 06 日

附录 F 毕业设计（论文）评审意见书

西安交通大学
毕业设计（论文）评审意见书（一）

评审意见：论文针对社交媒体多模态虚假信息检测技术展开讨论。验证了 CLIP 预训练模型中含有跨模态特征的编码部分与跨模态相似度指导多模态特征融合的有效性；验证了文本主题信息、文本情感信息对社交媒体虚假信息检测任务作为辅助数据来源的有效性；提出了一种准确性高、鲁棒性强的社交媒体虚假信息检测任务下的深度神经网络框架，通过实验验证了在虚假信息检测任务上的有效性。

论文写作认真，条理清晰，格式规范。论文工作量饱满。同意组织答辩。

评阅结论：同意答辩

评阅人建议成绩：A-

评阅人 葛思擘 职称 副教授

2023 年 06 月 09 日

西安交通大学
毕业设计（论文）评审意见书（二）

评审意见：本文研究了多模态社交媒体虚假信息检测深度神经网络结构，一种以模态间相似度作为指导信息，以文本主题信息与情感信息作为辅助信息的基于深度学习的多模态虚假信息检测神经网络框架。该框架通过模态间相似度判断单模态信息与融合信息在分类时的重要程度，对虚假信息分类任务作出指导；该框架使用预训练模型作为特征提取器，注意力机制进行模态融合。论文在 Weibo 数据集上进行实验，给出了实验结果，验证了论文方法在虚假信息检测任务上的有效性。

论文写作认真，格式基本规范。

评阅结论：同意答辩

评阅人建议成绩：B

评阅人 刘小勇 职称 副教授

2023 年 06 月 09 日

附录 G

毕业设计(论文)答辩结果

电子与信息学部

0504 自动化专业

毕业设计(论文)答辩组对学生梁珉珲所完成的课题为社交媒体多模态虚假信息检测技术研究的毕业设计(论文)经过答辩,其意见为该同学在毕设期间,态度比较认真,工作较努力,组织纪律较好。论文写作认真,格式规范,工作量饱满。答辩过程中讲述清楚,回答问题正确。答辩组一致同意通过毕设论文答辩,并确定成绩为:良。

并确定成绩为 85 分

毕业设计(论文)答辩组负责人

史施

答辩组成员

胡怀中

葛思璧

杨清宇

刘小勇

安豆

2023年06月11日

本科毕业设计（论文）工作进展情况记录表

(学生填写)

毕业设计（论文）题目：社交媒体多模态虚假信息检测技术研究

学生姓名：梁珉珲 学号：2196213038 专业班级：0504 自动化自动化 96 指导教师：史施

| 次 | 日期 | 具体工作内容 | 指导教师指导情况 |
|---|------------|--|-------------------------------|
| 1 | 2022-12-16 | 完成对数据集的清洗工作，将微博数据集中的多模态数据整理为（标签，文字，图片）三元组。 | 指导教师对数据集的选择进行了指导。 |
| 2 | 2023-01-13 | 完成了对文本词袋的构建，完成了文本情感分析任务。 | 指导教师指导词袋与情感分析任务所用的工具。 |
| 3 | 2023-02-12 | 基于 HuggingFace 构建出 Bert 预训练模型，Vit 预训练模型，CLIP 预训练模型及其多模态特征提取器。构建出 NTM 神经主题模型。 | 指导教师提供潜在能够使用的模型与论文供学生学习试验。 |
| 4 | 2022-11-30 | 确定选题，梳理任务难点与今后工作计划。 | 指导教师帮助学生确定选题，指定研究目标，梳理任务核心技术。 |
| 5 | 2023-02-19 | 完成数据批处理格式转换与馈送，模块化测试模型保证模型有效性。 | 指导教师指导学生模块化编程的重要性以及测试的重要性。 |
| 6 | 2023-02-28 | 完成模型训练过程以及测试过程编码，抽取模型最后一层特征进行降维并可视化，直观观测模型效果。 | 指导教师指导学生了解降维方法与可视化方法。 |
| 7 | 2023-03-05 | 将项目部署在服务器上， | 指导教师指导学生 debug。 |

附录 G

| | | | |
|----|------------|---|------------------|
| | | debug, 成功运行代码。 | |
| 8 | 2023-04-05 | 设计消融实验，分别得出各个对照组实验结果。 | 指导了消融实验的设计方法。 |
| 9 | 2023-04-14 | 删除消融实验中的 attention 组。 | 指导消融实验进行。 |
| 10 | 2023-04-17 | 调整 baseline 与 baseline+clip 组模型 attention 结构。 | 指导消融实验。 |
| 11 | 2023-05-02 | 绘制 tsne 降维可视化结果。 | 指导降维使用方法与绘图注意事项。 |
| 12 | 2023-05-15 | 发现 RDNN 中 NTM 投影层维度有误，修改后重新训练。 | 指导实验过程。 |
| 13 | 2023-05-22 | 重构代码，进行参数管理，设置命令行接口。 | 指导代码编写工作。 |
| 14 | 2023-06-01 | 完成论文初稿。 | 给出论文格式修改意见。 |
| 15 | 2023-06-04 | 论文初稿修改。 | 监督论文初稿修改工作。 |
| 16 | 2023-06-06 | 提交论文终稿。 | 审核通过，等待答辩。 |

注：此表由学生如实填写，毕业设计（论文）工作完成后，此表交院（系）教学秘书存档。

学 生 签

名：梁珉珲 指导教师签名：史橈

本科毕业设计（论文）选题审核表

| | | | | | | | |
|---|---------|-------|------------|------|-----------|------|-----------|
| 学院 | 电子与信息学部 | 系(专业) | 自动化科学与工程学院 | 指导教师 | 史楠 郑庆华 | 职称 | 副教授 教授 |
| 题目名称： 社交媒体多模态虚假信息检测技术研究 | | | | | 题目类别 | 题目来源 | 学生人数 |
| | | | | | 专题研究 | 自选 | 1 |
| 一、本题目目的、意义： 随着信息技术的发展，各类社交媒体平台逐渐流行，成为人们浏览信息、分享生活的重要工具。然而，由于其操作的简易性与应用的广泛性，虚假信息很容易在社交媒体中编辑发布并吸引大量关注，使得政府、各类机构或个人名誉受损，财产流失，对社会和谐稳定造成极为不良的影响。因此，社交媒体虚假信息检测技术成为了学者讨论与研究的热点，对经济与社会健康的发展具有深远的意义。目前主流的社交媒体虚假信息检测方法可大体分为两类：基于文本内容的虚假信息检测与基于社交上下文的虚假信息检测。其中，基于文本内容的虚假信息检测通过对虚假信息文本进行挖掘把握虚假信息的写作风格，基于社交上下文的虚假信息检测通过分析源用户信息以及传播链条把握虚假信息的转发特征。本课题综合利用上文所述的两种主流社交媒体虚假信息检测方法构建检测模型，同时在虚假信息文本中引入图片信息使多模态信息互相补充融合，便于神经网络更好地提取利于分类的特征。综上所述，基于深度学习的社交媒体多模态虚假信息检测技术研究具有极高的研究价值与实际意义。 | | | | | | | |
| 二、主要工作内容： 本课题主要工作内容分为以下三部分：1 数据获取与处理：计划使用 MediaTwitter、MediaWeibo 或 Weibo20 等包含多模态信息的数据集，并进行数据清洗等预备工作。2 模型构建与调试：使用 CLIP 等预训练模型对文本与图片信息进行特征提取与模态融合，使用图神经网络建模社交上下文信息，对虚假信息特征进行全面提取，最终利用分类器进行预测。成功构建模型后在验证集中试验最佳参数。3 设计实验并撰写论文：设计基线方法证明模型有效性，设计消融实验证明模块有效性，撰写论文总结模型设计思路、方法与实验结果。 | | | | | | | |
| 三、前期工作及具备条件（含实验风险评估）： (1)了解项目背景、目的、意义，阅读相关文献资料；(2)熟悉文本挖掘、多模态融合相关算法；(3)学习相关编程语言与框架等实现技术。 | | | | | | | |
| 指导教师签名： 史楠 郑庆华 2022 年 10 月 24 日 | | | | | | | |
| 系(专业)意见： 通过 | | | | | | | |

附录 G

| | |
|------------------------------------|-----------------------------------|
| 系(专业)主任签名: 张爱民 2022 年 10 月 24 日 | |
| 学院审核意见: 同意 | 主管教学院长签名: 罗新民 2022 年 10 月 24 日 |

- 注: 1、题目类别: 工程设计、专题研究、综合实验。
- 2、题目来源: 国家项目、省部级或学校科研任务、校外协作项目、实验室建设、就业所在单位项目、自选。
- 3、系(专业)意见主要是指对题目难度是否适中、题目内容能否达到毕业设计(论文)的教学要求和目的、完成本题目的条件是否满足等方面给出审核意见。
- 4、本表由指导教师填写, 学院存档保存。

本科毕业设计（论文）中期检查表

专业班级：0504 自动化自动化 96

| | | | | | |
|---|-------------------|----|-----|---------------------------------------|----------------------------|
| 学院 | 电子与信息学部 | 导师 | 史橚 | 学生姓名 | 梁珉珲 |
| | | | | 学号 | 2196213038 |
| 课题名称 | 社交媒体多模态虚假信息检测技术研究 | | | 有无关于指导的文字记录 | |
| | | | | <input checked="" type="checkbox"/> 有 | <input type="checkbox"/> 无 |
| 初稿计划完成时间 | 修改稿计划完成时间 | | | 定稿计划完成时间 | |
| | | | | | |
| <p>1. 指导情况（指导方式和内容、学生执行情况）：</p> <p>主要采取每周工作进展汇报的方式进行毕业设计工作指导；对学生毕业设计中具体出现的问题集中问答，启发并鼓励学生自主思考解决；本课题对于学生来说内容较新，引导学生阅读技术文档，启发学生找到正确的实现方法；学生能有序独立地完成相应的阶段任务，计划性强。</p> <p>2. 对毕业设计（论文）目标与方案的执行情况：</p> <p>学生能够独立有计划地完成中期阶段工作内容；学生对毕业工作有良好的设计和把控；后期须进一步对实验进行完善，并完成论文撰写的工作。</p> <p>3. 存在的问题、拟采取解决问题的方案及措施：</p> <p>进一步完善实验中的模型参数优化，以获得更高的精度。</p> | | | | | |
| 导师签名： | | | 史橚 | | |
| 2023年04月17日 | | | | | |
| <p>系（专业）意见：</p> <p>同意</p> | | | | | |
| 系（专业）主任签名： | | | 张爱民 | | |
| 2023年04月18日 | | | | | |

注：理工医艺术类不填“初稿计划完成时间、修改稿计划完成时间、定稿计划完成时间”。

主管教学院长签名：罗新民 2023 年 04 月 18 日