

# 北方工业大学

## 硕士 学位 论 文



### 基于深度学习的多模态虚假新闻检测技术研究

学 生 姓 名 \_\_\_\_\_ 葛红 \_\_\_\_\_

学 号 \_\_\_\_\_ 2020316210168 \_\_\_\_\_

学科(专业学位) \_\_\_\_\_ 计算机技术 \_\_\_\_\_

研 究 方 向 \_\_\_\_\_

导 师 \_\_\_\_\_ 李晋宏, 郭颖 \_\_\_\_\_

校 外 导 师 \_\_\_\_\_ 赵桂芬 \_\_\_\_\_

2023 年 4 月 5 日

论文密级	
保密时限	

北方工业大学

硕士 学位 论文



基于深度学习的多模态虚假新闻检测技术研究

学 生 姓 名 \_\_\_\_\_ 葛红 \_\_\_\_\_

学 号 \_\_\_\_\_ 2020316210168 \_\_\_\_\_

学科(专业学位) \_\_\_\_\_ 计算机技术 \_\_\_\_\_

研 究 方 向 \_\_\_\_\_

导 师 \_\_\_\_\_ 李晋宏, 郭颖 \_\_\_\_\_

校 外 导 师 \_\_\_\_\_ 赵桂芬 \_\_\_\_\_

2023 年 4 月 5 日

# **Multimodal Fake News Detection Technology**

## **Based on Deep Learning**

**By**

**Ge Hong**

**A Dissertation Submitted to**  
**North China University of Technology**  
**In partial fulfillment of the requirement**  
**For the degree of**  
**Master of Engineering**

**North China University of Technology**  
**April, 2023**

## 北方工业大学学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期： 年 月 日

## 学位论文使用授权书

学位论文作者完全了解北方工业大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属北方工业大学。学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许学位论文被查阅和借阅；学校可以公布学位论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存、汇编学位论文（保密的学位论文在解密后适用于本授权书）。

- 保密论文注释：经本人申请，学校批准，本学位论文定为保密论文，密级： ，期限： 年，自 年 月 起至 年 月 日止，解密后适用本授权书。
- 非保密论文注释：本学位论文不属于保密范围，适用本授权书。

本人签名： \_\_\_\_\_ 日期： \_\_\_\_\_

导师签名： \_\_\_\_\_ 日期： \_\_\_\_\_

# 基于深度学习的多模态虚假新闻检测技术研究

## 摘要

伴随科技的发展，虚假新闻引发社会危机的事件不时发生。传媒介质的改变使得虚假新闻逐渐由文本形式向图文并存的多模态形式转变。多模态虚假新闻承载着异构形式的文本和图像信息，对虚假新闻的多模态内容进行研究可以提高虚假新闻检测的效果。

现有的多模态虚假新闻检测方法存在以下不足。首先，在特征提取方面，大多数方法没有考虑自身的有效特征，且现阶段特征并不具有普适性。其次，在特征融合方面，以往利用图像特征联合文本特征的虚假新闻检测方法中，新闻中的特征信息只是简单地拼接，没有考虑到两种模态之间的交互信息。最后，在模型结构方面，少有模型能够验证提取特征的有效性。根据目前研究遇到的问题，本文基于深度学习进行多模态虚假新闻检测研究，进一步提高了虚假新闻检测的准确率，主要内容如下：

(1) 在特征提取方面，本文在深入研究了多模态特征提取技术的基础上，对虚假新闻的内容特性进行分析，提出了一种基于双分支对抗网络的多模态虚假新闻检测模型。该模型分别从不同层次提取模态的特征信息，并加入领域对抗网络来提高特征的普适性。

(2) 在特征融合方面，本文在研究了多模态的特征融合技术的基础上，提出了基于组合式融合机制的虚假新闻检测算法。该算法使用多模双线性池化方法进行模态间信息交互以补充特征的丰富性，使用自注意力机制进行模态内部信息增强以提高自身特征的有效性，实现异构模态特征信息的交互和增强。

(3) 在模型结构方面，本文为提高模型的可解释性，提出基于变分自编码器进行多任务学习的虚假新闻检测算法。加入变分自编码器用于重构特征，进而利用多任务学习的特性对损失函数进行改进优化，使得模型具有泛化性。

(4) 本文基于Weibo数据集和Twitter数据集两个权威数据集进行了大量实验，验证了本文所提模型和算法的有效性，并设计虚假新闻检测系统进行辅助检测。

**关键词：**深度学习，虚假新闻检测，特征提取，多模态融合

# **Multimodal Fake News Detection Technology Based on Deep Learning**

## **Abstract**

Along with the development of technology, incidents of social crisis caused by fake news occur. The change of media medium makes the fake news gradually change from textual form to multimodal form with the coexistence of image and text. Multimodal fake news carries heterogeneous forms of text and image information, and research on multimodal content of fake news can improve the effectiveness of fake news detection.

The existing multimodal fake news detection methods have the following shortcomings. First, in terms of feature extraction, most methods do not consider their own effective features, and the features are not universal at this stage. Second, in terms of feature fusion, in the previous fake news detection methods using image features combined with text features, the feature information in the news is simply concatenated together without considering the interaction information between the two modalities. Finally, in terms of model structure, few models can verify the effectiveness of extracted features. According to the problems encountered in the current research, this thesis conducts research on multimodal fake news detection based on deep learning to further improve the accuracy of fake news detection, which is mainly as follows:

(1) In terms of feature extraction, this thesis analyzes the content characteristics of fake news based on the deep research of multimodal feature extraction techniques and proposes a multimodal fake news detection model based on a two-branch adversarial network. The model extracts modal feature information from different levels separately, and adds domain adversarial network to improve the generalizability of features.

(2) In terms of feature fusion, this thesis proposes a fake news detection algorithm based on a combinatorial fusion mechanism. The algorithm uses a

multimodal bilinear pooling method for inter-modal information interaction to complement feature richness, and a self-attention mechanism for intra-modal information enhancement to improve the effectiveness of its own features, to achieve interaction and enhancement of heterogeneous modal feature information.

(3) In terms of model structure, this thesis proposes a fake news detection algorithm based on Variational Auto-Encoder for multi-task learning in order to improve the interpretability of the model. The Variational Auto-Encoder is added to reconstruct the features, and then the loss function is improved and optimized by using the multi-task learning feature to make the model generalizable.

(4) In this paper, we conducted extensive experiments based on two authoritative datasets, Weibo dataset and Twitter dataset, to verify the effectiveness of the model and algorithm proposed in this paper, and design a fake news detection system to assist in the detection.

**Key words:** Deep Learning, Fake News Detection, Feature Extraction, Multimodal Fusion

# 目 录

摘 要 .....	I
ABSTRACT .....	II
第一章 绪论 .....	1
1.1 研究背景与意义 .....	1
1.2 国内外研究现状 .....	2
1.2.1 深度学习 .....	2
1.2.2 虚假新闻检测 .....	3
1.3 研究内容 .....	5
1.4 论文组织架构 .....	7
第二章 相关技术与理论 .....	9
2.1 虚假新闻研究相关理论基础 .....	9
2.1.1 虚假新闻定义 .....	9
2.1.2 虚假新闻内容特征 .....	10
2.2 文本特征表示 .....	10
2.2.1 词袋模型表示方法 .....	10
2.2.2 TF-IDF 表示方法 .....	11
2.2.3 Word2Vec 词向量表示方法 .....	11
2.2.4 BERT 表示方法 .....	12
2.3 视觉特征表示 .....	13
2.3.1 VGGNet .....	15
2.3.2 ResNet .....	15
2.4 多模态特征融合 .....	17
2.4.1 数据级融合 .....	17
2.4.2 特征级融合 .....	18
2.4.3 决策级融合 .....	19

2.5 本章总结 .....	19
<b>第三章 基于双分支对抗网络的多模态虚假新闻检测算法 .....</b>	<b>20</b>
3.1 引言 .....	20
3.2 相关工作 .....	21
3.2.1 领域对抗网络 .....	21
3.3 基于双分支对抗网络的特征提取改进方法 .....	22
3.3.1 文本特征提取模块 .....	23
3.3.2 图像特征提取模块 .....	24
3.3.3 领域对抗模块 .....	25
3.3.4 虚假检测模块 .....	26
3.3.5 损失函数 .....	26
3.4 实验设计 .....	27
3.4.1 数据集 .....	27
3.4.2 评价标准 .....	28
3.4.3 实验环境 .....	28
3.4.4 对比实验模型 .....	29
3.5 实验结果分析 .....	30
3.5.1 对比实验结果分析 .....	30
3.5.2 消融实验结果分析 .....	31
3.6 本章总结 .....	32
<b>第四章 基于组合式融合机制的多模态虚假新闻检测算法 .....</b>	<b>33</b>
4.1 引言 .....	33
4.2 相关工作 .....	34
4.2.1 多模双线性池化 .....	34
4.2.2 自注意力机制 .....	36
4.3 基于组合式融合机制的特征融合改进方法 .....	37
4.3.1 多模态特征提取模块 .....	38
4.3.2 模态间信息交互模块 .....	38

4.3.3 模态内信息增强模块 .....	39
4.3.4 领域对抗模块 .....	40
4.3.5 虚假检测模块 .....	40
4.4 实验设计 .....	40
4.5 实验结果分析 .....	40
4.5.1 对比实验结果分析 .....	41
4.5.2 消融实验结果分析 .....	42
4.6 本章小结 .....	43
<b>第五章 基于变分自编码器进行多任务学习的多模态虚假新闻检测算法 .....</b>	<b>44</b>
5.1 引言 .....	44
5.2 相关工作 .....	45
5.2.1 变分自编码器 .....	45
5.3 基于变分自编码器的模型结构改进方法 .....	47
5.3.1 编码器模块 .....	47
5.3.2 解码器模块 .....	48
5.3.3 领域对抗模块 .....	49
5.3.4 虚假检测模块 .....	49
5.3.5 多任务学习损失函数 .....	49
5.4 实验设计 .....	51
5.5 实验结果分析 .....	53
5.5.1 权重对比实验结果 .....	53
5.5.2 对比实验结果分析 .....	54
5.6 本章小结 .....	56
<b>第六章 虚假新闻检测系统设计实现 .....</b>	<b>57</b>
6.1 系统功能分析与设计 .....	57
6.1.1 系统功能需求分析 .....	57
6.1.2 系统概要设计 .....	57

6.1.3 系统总体设计 .....	58
6.2 系统实现及测试.....	58
6.2.1 系统功能实现 .....	58
6.2.2 系统功能测试 .....	59
6.3 本章小结 .....	61
<b>第七章 结论与展望 .....</b>	<b>62</b>
7.1 主要结论 .....	62
7.2 研究展望 .....	63
<b>参考文献 .....</b>	<b>64</b>
<b>附录 A .....</b>	<b>70</b>
<b>在学期间的研究成果 .....</b>	<b>71</b>

# 第一章 绪论

## 1.1 研究背景与意义

中国社会科学院新闻传播研究所于 2022 年 8 月发布的《新媒体蓝皮书：中国新媒体发展报告 No.13(2022)》显示：在内容生产主体多元化、信息技术变革及传播环境重构的背景下，社交平台已成为公众获取信息的重要渠道<sup>[1]</sup>。社交网络具有实时性、开放性、便捷性和双向性等特点，给公众提供了一个言论自由空间的同时，也成为了虚假新闻的滋生地<sup>[2]</sup>。目前，新闻传播的方式发生了巨大的变化，虚假新闻带来的危害也如同瘟疫一样肆意扩散<sup>[3]</sup>。第一，虚假新闻破坏了新闻生态系统的真实性平衡。虚假、猎奇的新闻更容易吸引公众的注意力和关注度，从而形成以讹传讹的恶性循环<sup>[4]</sup>。第二，虚假新闻有意引导读者接受有偏见或错误的观点。在舆论斗争中，虚假新闻是一种常用的手段。不法分子通过散布虚假消息来混淆视听、煽动情绪，从而导致了不良的舆论趋势<sup>[5]</sup>。第三，虚假新闻影响了公众对真实新闻的认知。由于社交媒体平台上的虚假新闻泛滥，导致公众在看到新闻时无法辨别真假，严重损害了公众对新闻的信任度，损害了媒体的权威性和公信力<sup>[6]</sup>。在虚假新闻泛滥的严峻形势下，提出有效的虚假新闻检测技术，具有十分重要的现实意义。

随着传媒环境的变迁，新闻内容逐渐多元化，虚假新闻的结构形式由纯文本向图文并茂的多模态形式转变<sup>[7]</sup>。多模态新闻与纯文本新闻相比，携带的信息更加丰富直观，容易吸引公众的关注从而在短时间内快速传播<sup>[8]</sup>。因此，对虚假新闻中的文本和图像内容进行深入研究有助于提高虚假新闻检测的准确率。

鉴于虚假新闻普遍具有多种模态的表现形式，文本和图像为虚假新闻检测的实现带来了各自侧重、相互补充的信息，仅从单模态数据提取特征已经不足以解决当下的虚假新闻检测问题。因此目前主流方法正从基于单模态内容检测转向基于多模态内容检测。由于文本和图像两个跨模态内容的差异性，如何从不同模态中提取涵盖各自领域的有效特征是一个关键问题。对于提取的不同模态的特征如何进行有效的特征融合是现阶段的一个难点。对于融合后的多模态特征使用哪种验证方式来验证其有效性是一个需要解决的问题。因此需要研究

出一种在多模态数据的基础上能够有效提取、充分融合文本信息和图像信息的方法来检测新闻的真实性。

## 1.2 国内外研究现状

在当前社会环境下，虚假新闻检测成为了一个热门研究方向。深度学习的创新发展为虚假新闻检测的研究提供了基础、平台和方法。本节将从深度学习、虚假新闻检测两个方面介绍近些年的相关工作。

### 1.2.1 深度学习

2022 年 3 月 16 日，斯坦福大学发布了《2022 年人工智能指数报告》<sup>[9]</sup>，该研究报告指出：中国和美国对人工智能领域的投资不断加大，相继涌现出各类新的人工智能创新企业，诞生了大量的人工智能产品，给人们的生活提供了巨大的便捷。其中，深度学习是目前人工智能的重点支撑研究之一，应用于人工智能的众多领域，包括计算机视觉、自然语言处理、语音处理等<sup>[10]</sup>。

深度学习是机器学习的一个分支。与传统机器学习方法相比，深度学习具有成本优势和更好的学习效果，通过不断地优化迭代，可以挖掘出数据中的隐藏信息，提高特征提取能力。随着互联网数据的不断积累，计算机硬件性能的快速提升，深度学习成为了一个重要研究方向。

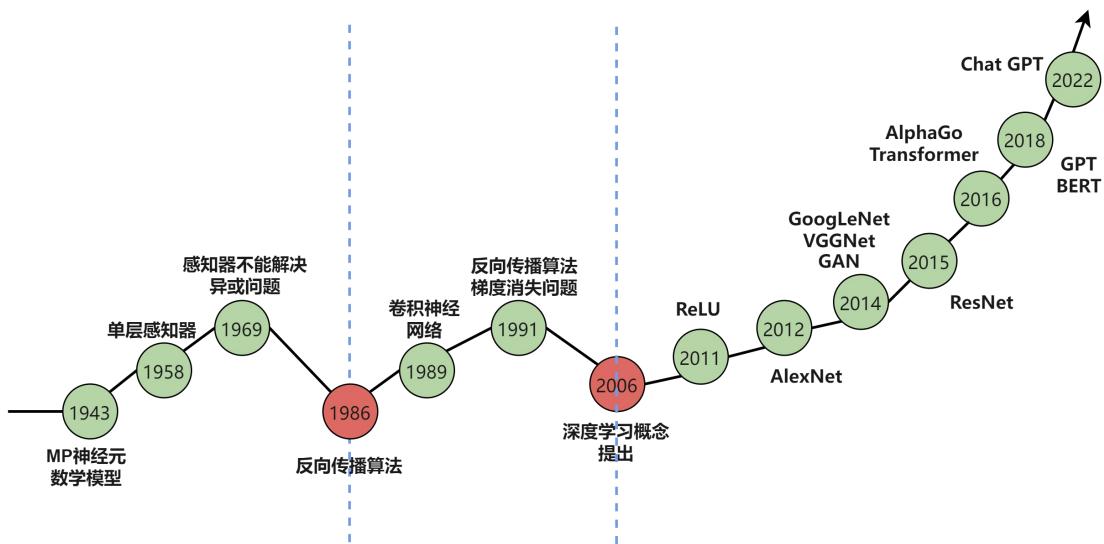


图 1-1 深度学习发展历史

深度学习发展历史如图 1-1 所示。1943 年，MP 神经元数学模型的提出，开启了用元器件和计算机程序实现人工神经网络的研究道路<sup>[11]</sup>。1958 年，

Rosenblatt F 提出了一个只有一层神经元构成的单层感知器，通过训练实现了基本的分类功能<sup>[12]</sup>。1969 年，Marvin M 和 Papert S 指出了单层感知器不仅在处理能力上具有局限性，并且对异或问题的处理上也有困难<sup>[13]</sup>。由于这个缺陷，此后的数十年里，人工神经网络进入了第一个瓶颈期，研究神经网络的高潮逐渐褪去。

1986 年，Rumelhart D E 和 Hinton G E 共同提出了基于多重感知器的误差反向传播算法（Back Propagation，BP），在非线性分类中展现出不错的效果<sup>[14]</sup>。1989 年，LeCun Y 等人提出一种具有卷积层和池化层的卷积神经网络结构，并使用 BP 算法进行优化参数，在各类比赛中取得了优异成绩。该结构奠定了卷积神经网络的基础架构<sup>[15]</sup>。但在 1991 年，BP 算法被指出存在梯度消失问题，神经网络的研究因此被搁置。

2006 年，Hinton G E 等人首次提出了深度学习这一概念<sup>[16]</sup>，基于神经网络以降维方式处理高维数据，开启了人们对深层神经网络的研究道路。2011 年，Glorot X 等人提出 ReLU 激活函数，该函数可以抑制在训练中梯度消失问题的发生<sup>[17]</sup>。2012 年，在 ImageNet 图像识别竞赛中，Hinton G E 等人提出了 AlexNet，利用该神经网络显著提高了图像识别的正确率，获得了该竞赛的冠军<sup>[18]</sup>。2014 年，Szegedy C 设计了卷积神经网络 GoogLeNet，该神经网络引入了 Inception 结构，并添加了两个辅助分类器，刷新了 ImageNet 图像识别竞赛的最好成绩<sup>[19]</sup>。同年，VGGNet 深度卷积网络<sup>[20]</sup>、生成对抗网络（Generative Adversarial Networks，GAN）也出现在公众视野，并迅速成为了深度学习研究领域的热点课题<sup>[21]</sup>。2015 年，He K 等人构建了含有残差块的卷积神经网络 ResNet<sup>[22]</sup>，缓解了层数加深带来的梯度消失。

2016 年，谷歌团队开发的人工智能围棋程序 AlphaGo 击败了世界围棋冠军李世石，将深度学习的研究推向了一个崭新的高度。同年，Vaswani A 等人提出的完全基于注意力机制的 Transformer 模型开启了深度学习的再一次研究热潮<sup>[23]</sup>。2018 年，Radford A 等人提出的 GPT<sup>[24]</sup>和 Devlin J 等人提出的 BERT<sup>[25]</sup>预训练模型，在自然语言领域大放异彩。2022 年，Open AI 陆续发布的 Chat GPT 聊天机器人模型，不仅能够流畅地与人类进行对话，像人类一样根据对话的上下文来交流，甚至还能完成“舞文弄墨”、“吟诗作对”、撰写报告、编写代码等任务，给人工智能领域带来了无限可能。

### 1.2.2 虚假新闻检测

虚假新闻的内容、传播方式和用户类型是检测虚假新闻较为关键的指标。

鉴于此，目前的虚假新闻检测研究通常采用基于内容、传播和用户的方法。虚假新闻的内容是传播信息的主要载体，因此基于内容的虚假新闻检测方法是当前研究的主要方向。基于新闻内容的模态数量，当前的研究方法可以分为两类：基于单模态的虚假新闻检测方法和基于多模态的虚假新闻检测方法。

### （1）基于单模态的虚假新闻检测方法

基于单模态的虚假新闻检测方法可以分为基于文本内容的虚假新闻检测方法和基于图像内容的虚假新闻检测方法。早期基于文本的检测方法主要是提取各种人工特征，结合机器学习方法检测虚假新闻。Castilo C 等人通过统计特殊字符、链接数目等单模态文本信息来检测虚假新闻<sup>[26]</sup>。Qazvinian V 等人通过对文本主题特征的研究，利用贝叶斯网络作为分类器对虚假新闻进行识别<sup>[27]</sup>。然而，这种单一的内容在研究虚假新闻的内容层面上缺乏全面性和灵活性。随着深度学习技术的不断发展，Ma J 等人利用循环神经网络对文本内容进行特征信息的提取，进而检测虚假新闻<sup>[28]</sup>。Ma J 等人基于 GAN 网络，提出了一种新的虚假新闻检测模型，通过对抗性训练能够获取低频但具有较强判别能力的特征<sup>[29]</sup>。除了基于文本内容的检测方法以外，基于图像内容的检测方法也是检测虚假信息的重要手段<sup>[30]</sup>。早期的研究主要是利用图像基本统计特征对虚假新闻进行检测，但是这种方法不能充分提取图像中所包含的语义信息。目前，对图像特征提取的研究多采用深度卷积神经网络。Qi P 等人利用注意力机制结合神经网络动态融合图像频域特征和像素域特征来检测新闻，取得了不错的效果<sup>[31]</sup>。

目前，基于单模态的虚假信息检测方法已取得一些研究成果，但仅从文本内容或图像内容的角度来研究，导致多模态的信息效用不高，检测效果不佳。同时文本内容和图像内容之间存在一定的互补性，只单独检测一种模态内容会损失不同模态内容之间的关联性信息。因此联合建模多模态特征能够更好地学习到新闻语义特征从而提升虚假新闻检测的准确率。

### （2）基于多模态的虚假新闻检测方法

多模态虚假新闻检测即同时使用文本内容和图像内容进行鉴别新闻。目前，新闻发布方式主要以文本和图像相结合的方式为主，所以基于多模态内容的虚假新闻检测成为了人们的关注热点。早期的多模态虚假新闻检测主要通过人工抽取文本和图像的特征，将提取到的文本特征和图像特征进行拼接，输入到训练模型中<sup>[32]</sup>。这种方法虽然可以兼顾文本特征和图像特征，但无法获取到两种模态间的深层次语义，导致虚假新闻检测很难获得较好的效果。

伴随着深度神经网络技术的不断成熟，更多的专家学者尝试利用深度神经网络来解决多模态虚假新闻检测的问题<sup>[33]</sup>。Jin Z 等人利用循环神经网络和

VGG19 卷积神经网络模型分别获取文本特征和图像特征，利用注意力机制简单融合得到的特征并用于虚假新闻检测<sup>[34]</sup>。Wang Y 等人为了排除特定事件对新闻真假判别的干扰，提出一种利用事件对抗神经网络来检测虚假信息的方法，该方法可以引导模型学习与新闻事件的无关特征来识别虚假新闻<sup>[35]</sup>。Zhang H 等人在虚假新闻检测模型中引入了事件记忆网络，通过获取具体新闻中的潜在共性特征，提高了对新事件领域中新闻检测的迁移能力<sup>[36]</sup>。Khattar D 等人在虚假信息检测模型中添加了分类器和变分自编码器，提高了获取潜在共性特征的能力<sup>[37]</sup>。Singhal S 等人利用预训练模型 BERT 提取文本特征，利用卷积神经网络 VGG19 提取图像特征，最终将两种特征拼接作为联合表征进行分类，取得了良好的检测效果<sup>[38]</sup>。

综合之前的研究内容，已有研究人员将多模态内容作为虚假新闻检测的研究方向，但当前的多模态虚假新闻检测方法大多都是在融合阶段做简单地拼接操作。仅通过对两种模态特征向量进行简单拼接的融合方式，难以充分利用新闻多种模态之间的互补性及差异性，造成虚假新闻的检测效果不佳。因此，亟需研究一种基于多模态数据，能够有效提取并充分融合文本与图像两种模态信息的新闻真实性检测方法。

### 1.3 研究内容

在当今环境下，新型社交媒体的兴起，改变了传统的新闻传播方式以及人们获取新闻的途径。由于缺乏对虚假新闻的有效检测，造成了虚假新闻的肆意传播，严重危害公共空间的安全。因此，对公共空间中的新闻进行快速检测，遏制虚假新闻的传播，具有十分重要的现实意义。尽管现有的检测方法基本实现了虚假新闻检测功能，但大部分的检测方法是基于单模态文本或图像数据，无法提取不同模态内容的语义特征，尽管有一些多模态的检测方法，但却无法充分地进行多模态内容之间的交互融合，并且特征的有效性不能得到验证和解释。

因此本文以新闻多模态内容特征的提取、融合、以及验证为切入点来开展研究工作。首先需要调查虚假新闻的主要特点，新闻内容中的每个模态都具有其独有的特性，需要对新闻不同模态的内容进行特征的有效提取，其次经过提取的各个模态的特征需要在跨模态的前提下进行特征融合以获得其隐藏的补充特征，最终需要改变现有模型的框架结构对融合后的特征进行验证，提高模型的可解释性。综上，本文主要的研究内容如下：

### (1) 多模态特征提取问题

在虚假新闻检测任务中，所涉及的模态有文本和图像，需要对两种模态信息分别进行特征提取。由于文本和图像表现形式和描述方式的差异性，如何从新闻内容中提取不同模态的有效特征？其次虚假新闻事件种类繁多，每天有数以万计的新闻被发布，大多数现有的工作倾向于学习特定新闻事件的特征，不能转换到未曾见过的新闻事件上，如何从新出现的新闻事件中提取有效特征？

针对以上两个特征提取的问题，本文提出在使用深度预训练模型的前提下，使用双分支网络进行深层和浅层的特征提取，以获得不同层级的特征向量。同时提出使用领域对抗网络进行对抗训练，以获得不同新闻领域中的共性特征解决模型的泛化性问题。

### (2) 多模态特征融合问题

现阶段，在利用文本特征联合图像特征的虚假新闻检测方法中，仅通过将两种模态特征向量简单拼接的融合方式，难以充分利用新闻多种模态之间的互补性内容信息及差异性内容信息，从而造成虚假新闻检测效果不佳。因此如何更好的融合多种模态特征以补充其互补性及差异性是一个主要研究问题。

针对特征融合问题，本文将提出一种基于组合式融合机制的多模态融合方法，使用不同的融合策略从模态间和模态内分别进行交互增强特征。对于模态间内容，使用多模双线性池化方法充分结合文本与图像每一位置的独特维度信息。对于每个模态内部，使用自注意力机制来进行自身内容的加强，从而保持特征的完整性和多样性。

### (3) 模型结构问题

在深度学习的背景下，虚假新闻检测模型抽取的高维特征缺失可解释性，无法了解其具体效果。同时在文本特征和视觉特征融合后直接进入检测器进行检测，缺少一个模块来验证融合特征的效果，补充模型的可解释性。

针对模型的可解释性问题，本文提出使用变分自编码器进行重现高维特征，增加重构模块验证融合特征的效果，并在损失函数上进行改进，使用多任务训练的方式使得提取的特征更具有有效性，提高模型的可解释性。

### (4) 虚假新闻检测系统

本文将对研究的虚假新闻检测模型进行封装，设计了一个虚假新闻检测系统，辅助进行验证虚假新闻检测的效果。

## 1.4 论文组织架构

本文对基于深度学习的多模态虚假新闻检测算法进行研究，论文的整体组织架构如图 1-2 所示：

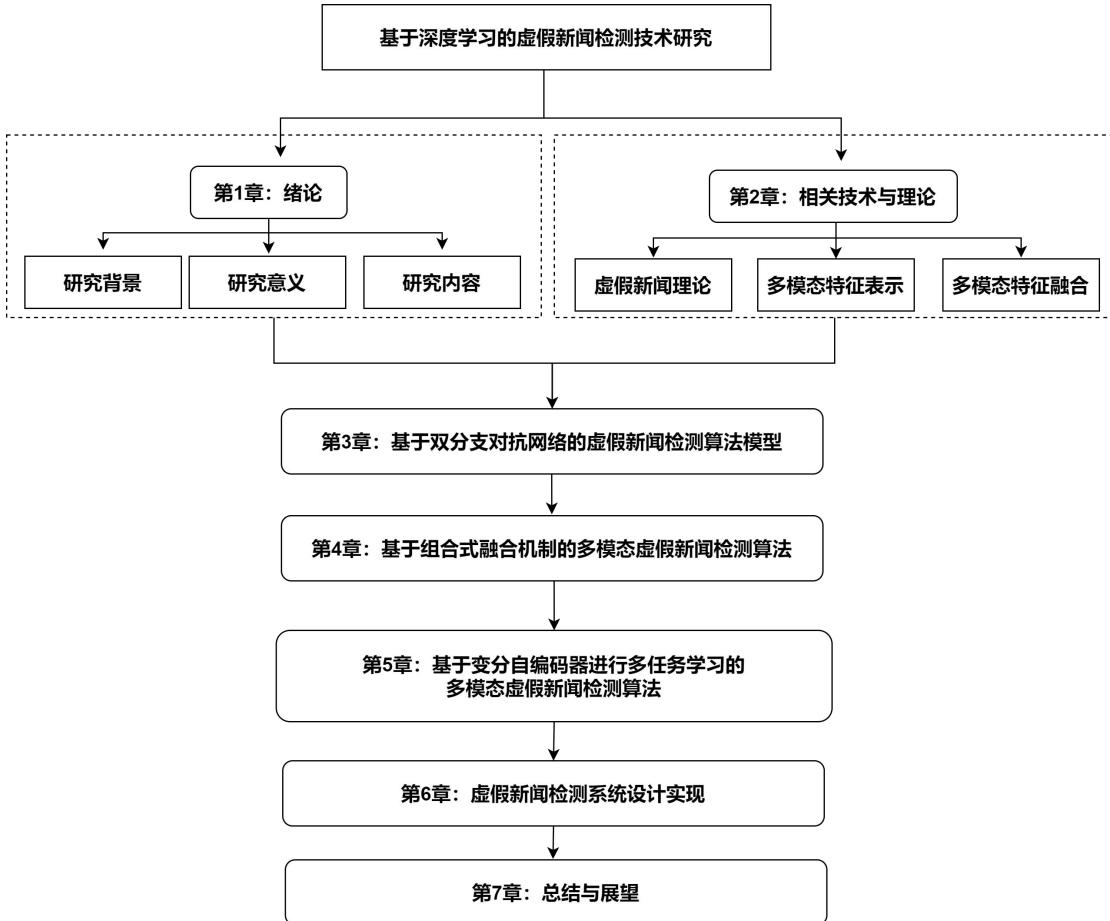


图 1-2 论文组织架构

**第 1 章：绪论。** 绪论部分首先阐述了研究问题的背景和意义，其次介绍相关的国内外研究现状以及研究方法，并引入了本文的主要研究内容。

**第 2 章：相关技术与理论。** 相关技术与理论部分首先阐述了虚假新闻的定义和特点。在此基础上，对虚假新闻检测所需要的技术进行了理论分析。

**第 3 章：基于双分支对抗网络的虚假新闻检测算法模型。** 主要研究对不同模态内容特征的提取问题。本文采用预训练模型从深层和浅层双分支分别提取不同模态的特征信息。为了保证特征的普适性，加入领域对抗网络训练以得到泛化性特征。最后通过实验证明了该模型可以提高特征提取的能力。

**第 4 章：基于组合式融合机制的多模态虚假新闻检测算法。** 此章在第 3 章的基础上提出了对模态间信息和模态内信息进行交互和增强。在第 3 章使用模

型分别提取高阶的文本及图像特征后，在文本模态和图像模态之间使用多模双线性池化的方式进行交互融合，在文本模态内部和图像模态内部分别使用自注意力机制进行自模态信息的增强。经过实验分析，该框架可以有效融合不同模态的特征。

第 5 章：基于变分自编码器进行多任务学习的多模态虚假新闻检测算法。此章基于第 3 章和第 4 章，提出一种可以验证多模态特征有效性的虚假新闻检测算法。该算法使用变分自编码器重现高维特征，同时对不同任务进行整合学习，使得模型能够兼顾多个任务，从而能够更全面地处理多模态虚假新闻检测任务。

第 6 章：虚假新闻检测系统设计。此章在前面几个章节的基础上，按照软件工程的思想，实现了一个完整的虚假新闻检测系统。

第 7 章：总结与展望。对本文所做出的研究工作进行分析总结，同时展望未来工作中可以进行的相关研究内容。

## 第二章 相关技术与理论

本章主要介绍了虚假新闻检测的相关技术理论。首先，阐述了虚假新闻的相关理论基础，介绍了虚假新闻的定义以及特征。其次，引入了虚假新闻中文本特征和图像特征的表示方法，并梳理了多模态特征融合方式。

### 2.1 虚假新闻研究相关理论基础

#### 2.1.1 虚假新闻定义

社会自有新闻传播活动以来，就一直面对着与虚假新闻的抗争。但时至今日，对于“虚假新闻”这一术语，学界仍未达成共识。目前的研究一般将虚假新闻归纳为狭义虚假新闻和广义虚假新闻<sup>[39]</sup>。狭义的虚假新闻定义为那些可证实为假并且意图误导读者的新闻<sup>[40]</sup>。该定义有两大特征：真实性与目的性。首先，在真实性上，虚假新闻一般为可核实为假的消息。第二，在目的性上，虚假新闻一般是以欺骗读者为目的。对于广义虚假新闻的界定，多关注于其报道的真伪。比如，一些报纸将讽刺新闻视为虚假新闻，虽然讽刺新闻通常是为了提高娱乐效果，并且将它的欺骗行为暴露给新闻读者，但是内容却是假的<sup>[41]</sup>。

为了将当前研究的虚假新闻与一些其他术语和概念相区别，例如：错误新闻（false news）<sup>[42]</sup>、虚假新闻（fake news）<sup>[43]</sup>、讽刺新闻（satire news）<sup>[44]</sup>、虚假信息（disinformation）<sup>[45]</sup>、错误信息（misinformation）<sup>[46]</sup>和谣言（rumor）<sup>[47]</sup>等，通过信息真实性、意图和是否是新闻三个属性来区分它们之间的差异。表 2-1 列出了具体的差异。在以上分析的基础上，本文采用了狭义的虚假新闻概念，将虚假新闻定义为新闻网站发布的可证实为假的、意图为坏的新闻。

表 2-1 相关概念表

类型	真实性	意图	是否是新闻
虚假新闻	假	坏	是
错误新闻	假	未知	是
讽刺新闻	未知	未知	是
虚假信息	假	坏	否
错误信息	假	未知	否
谣言	未知	未知	否

### 2.1.2 虚假新闻内容特征

传统新闻报道仅包含新闻自身内容，社交媒体中的多模态新闻内容可以增强虚假新闻的表现力和传播力<sup>[48]</sup>。虚假新闻一般会具有吸引眼球的文本标题以及具有强烈视觉冲击力的配图<sup>[49]</sup>，以此达到吸引读者注意的目的，诱导读者对新闻事件做出错误判断。通过对社交媒体中出现的虚假新闻案例的分析，归纳出部分虚假新闻的内容特点。

#### (1) 文本内容特点

在社交媒体上，由于每条新闻有字数限制，媒体报道的主体内容通常仅由能够表明作者观点和立场的文字描述所组成。在语言层面上，虚假新闻中的语言往往具有偏见和不良诱导等特点，所以可以根据不同层次的情感特征、整体句子特征、词汇特有特征等来对文本内容进行判断，因此可以利用不同的分支，提取不同层次的特征来检测新闻。

#### (2) 图像内容特点

由于文本内容的限制，图像作为一种辅助的信息传播媒介被广泛应用。图像所包含的信息，不仅有图像颜色、清晰度和统计特性，同时具有深层次的语义信息，可以辅助文本发现有用的信息特征。因此可以分别抽取出图像自身的不同层面的特征，来提升虚假新闻的检测效果。同时，虚假新闻的创作者还会通过修改其它新闻中的图像，将其伪装为另一个新闻事件的图像，这使得很多读者很难分辨图像的真实度。

## 2.2 文本特征表示

文本特征表示的含义是将文本信息中的特征抽取出来，进而转换成数学中向量或矩阵的形式。下面对文本特征表示方法分别展开介绍。

### 2.2.1 词袋模型表示方法

词袋模型是一种早期的文本表示方法，可以直接将文本转化为向量表示，对于词袋模型而言，文本中的每个词是相互独立的，不关心每个词的出现顺序和词义，只关注每个词在文本中的使用频次<sup>[50]</sup>。词袋模型忽略了每个词的位置信息，但词的出现位置对文本语义的影响较大，因此该方法难以衡量每个词的重要性。

### 2.2.2 TF-IDF 表示方法

TF-IDF 为词频-逆向文档频率，该模型是词袋模型的一种优化模型<sup>[51]</sup>。TF-IDF 同样以词频衡量在文本中的重要程度。该算法主要有 TF (Term Frequency) 和 IDF (Inverse Document Frequency) 两部分组成，如式(2-1)所示：

$$TF - IDF_{ij} = TF_{ij} \times IDF_{ij} \quad (2-1)$$

其中，TF 表示词频，即某词在文档中出现的频率。过程如式(2-2)所示：

$$TF_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \quad (2-2)$$

其中  $n_{ij}$  分子为关键词  $i$  在文档  $j$  中出现的次数，分母是文档  $j$  中所有关键词的出现次数之和。

IDF 表示逆向文档频率，指的是对关键词普遍重要性的度量。计算 IDF 的方法如式(2-3)所示：

$$IDF_{ij} = \lg \frac{|D|}{|\{j : t_i \in d_j\}|} \quad (2-3)$$

其中  $|D|$  表示整个语料库中的文档数， $|\{j : t_i \in d_j\}|$  表示包含关键词  $i$  的文档数目。

TF-IDF 具有原理简单、可解释性强等特点，解决了词袋模型无法区分常用词和专有名词对文本重要性的问题，可以衡量关键词在文本中的重要性。但由于 TF-IDF 结构简单，单词的重要性和词的多义问题不能得到有效解决，同时 TF-IDF 不关心单词的位置关系，无法为不同位置的单词分配不同的权重。

上述两种方法都是文本的离散表示方法，均不关心文本的语序和语义信息，具有缺乏词序信息、特征稀疏、可能造成维度灾难等缺点。

### 2.2.3 Word2Vec 词向量表示方法

文本的分布式表示将词转换成低维的连续稠密向量<sup>[52]</sup>，能够包含更多的语义信息，可以有效解决离散表示方法的不足。

Word2vec 是 Mikolov T 在 2013 年首次提出的一种基于文本的分布式表达方式<sup>[53]</sup>。基于 Word2Vec 产生的词向量能够较好地衡量词汇间的相似程度，有效地实现了词向量的语义表达。Word2vec 包含两种模型，CBOW (Continuous Bags-of-Words Model) 和 Skip-Gram (Skip Gram Model)。CBOW 模型用于对给定的语境词汇进行预测中心词。模型图如图 2-1 (a) 所示。Skip-Gram 模型根据中心词作为输入来预测周围的上下文，模型图如图 2-1 (b) 所示。由于

Word2vec 词向量方法会用到上下文语句，因此相对于离散型的文本表示方式，它将会有一些语义特性。但是，在不同的上下文语境中词汇与向量之间存在着一一对应的关系，因此 Word2vec 词向量方法不能很好地解决一词多义的问题。

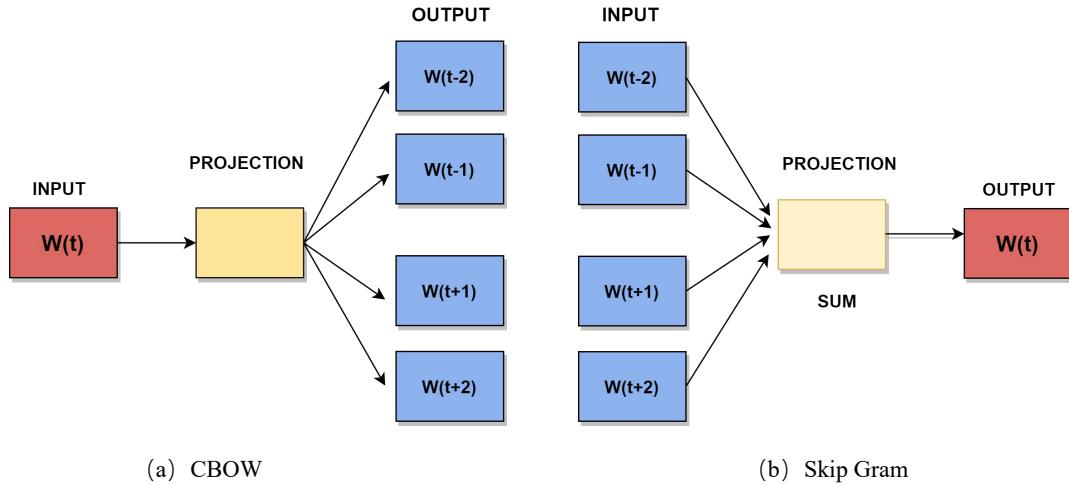


图 2-1 Word2Vec 模型

#### 2.2.4 BERT 表示方法

在不同的上下文中若一个单词只有一种语义表示，那么多义词的问题将会影响句子整体的表达含义。预训练模型 BERT ( Bidirectional Encoder Representation from Transformers ) 诞生，解决了一词多义的问题。

BERT 模型的预训练方法和下游任务微调方式，使得一词多义问题迎刃而解。从结构上分析，BERT 以 Transformer 的编码器作为基本单元建立模型，而 Transformer 模型完全基于注意力机制。其模型结构如图 2-2 所示。由于编码器单元中的自注意机制在编码一个词语时同时关注其上下文的词语，使得任意位置的两个单词的距离都为一，因此可以很好地解决 NLP 中的冗余依赖问题。

BERT 的输入信息是由词嵌入 (Token Embeddings)、位置嵌入 (Position Embeddings) 和分割嵌入 (Segment Embeddings) 三部分共同组成，同时其中含有 [CLS] 和 [SEP] 两个符号信息，其整体结构如图 2-3 所示。模型的预训练任务包括两个：掩码语言模型任务，下句预测任务。其中掩码语言模型任务可以在预训练过程中通过上下文预测提前遮蔽文本。下句预测任务可以通过训练得到上一句文本最可能出现的下一句文本，进而得到句子之间的关系。

BERT 提供了不同的迁移策略应对具体的下游任务。因此本文选取 BERT 作为文本特征抽取的基本模型。

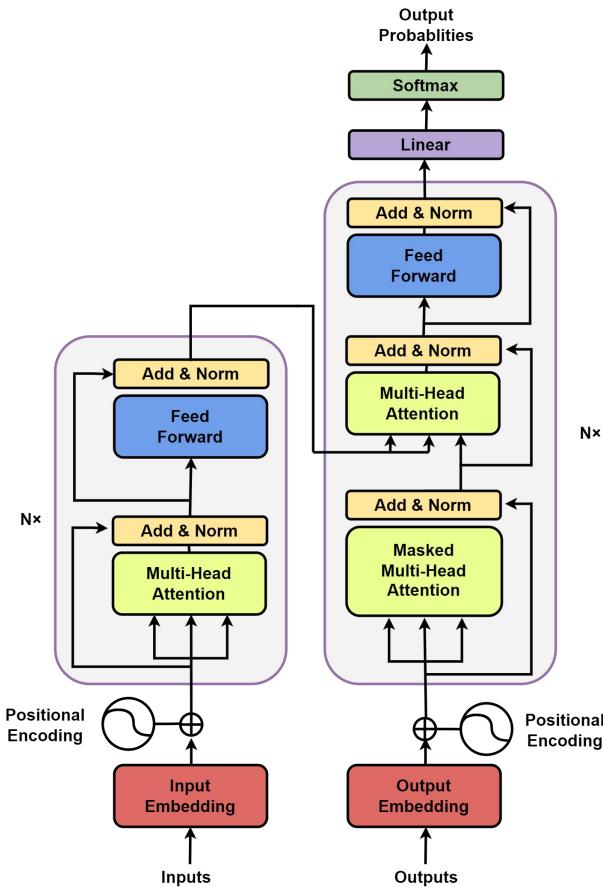


图 2-2 Transformer 模型结构

Input	[CLS]	my	cat	is	cute	[SEP]	he	likes	play	##ing	[sep]
Token Embeddings	E[CLS]	E_my	E_cat	E_is	E_cute	E[SEP]	E_he	E_likes	E_play	E##ing	E[sep]
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
Position Embeddings	E <sub>0</sub>	E <sub>1</sub>	E <sub>2</sub>	E <sub>3</sub>	E <sub>4</sub>	E <sub>5</sub>	E <sub>6</sub>	E <sub>7</sub>	E <sub>8</sub>	E <sub>9</sub>	E <sub>10</sub>

图 2-3 BERT 模型输入特征构成

## 2.3 视觉特征表示

视觉信息对于虚假新闻的检测同样重要。卷积神经网络（Convolutional Neural Networks, CNN）是现阶段视觉特征建模的主要方法，可以实现自动化抽取图像的低层次风格和高层次语义特征。卷积神经网络是一类包含卷积层和池化层的神经网络模型。LeNet-5 作为早期的卷积神经网络模型，其详细结构如

图 2-4 所示。主要结构包括两层卷积层、两层池化层以及三层全连接层组成。

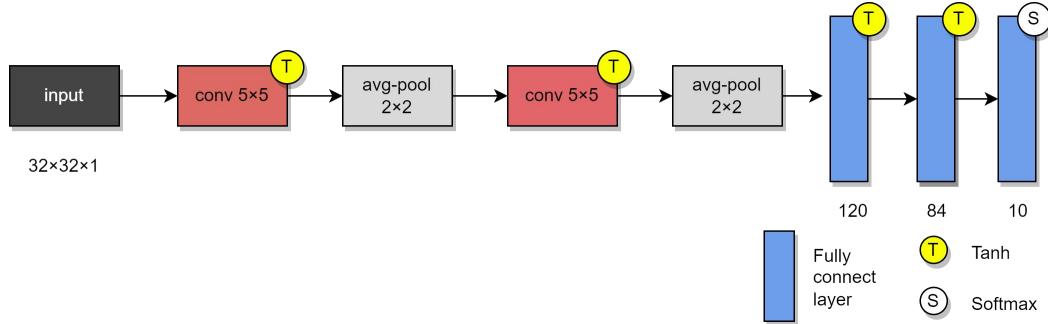


图 2-4 LeNet-5 模型结构

其中卷积层由多个卷积单元构成，其作用是抽取出输入图像的特征信息。在提取特征的过程中神经网络通过卷积核（filter）对输入图像数据进行计算。如图 2-5 为一个卷积运算实例，其中使输入数据的尺寸为  $3 \times 3$ ，使用  $2 \times 2$  的卷积核对矩阵数据进行卷积，让卷积核与输入矩阵进行滑动覆盖后进行点积运算。

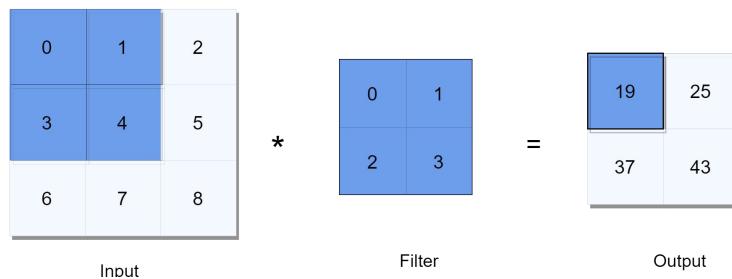


图 2-5 卷积运算实例

池化层的目的是特征选择，降低特征中的冗余信息，从而减少参数的数量，抑制过拟合影响。池化方式一般有平均池化与最大池化两种方式。最大池化是选择每个特征矩阵区域中的最大值作为最终特征。而均值池化策略选择特征区域中的均值作为最终输出。如图 2-6 为最大池化和平均池化的实例。最后的全连接层将特征信息展开为特征向量，用于给出最后的分类结果。

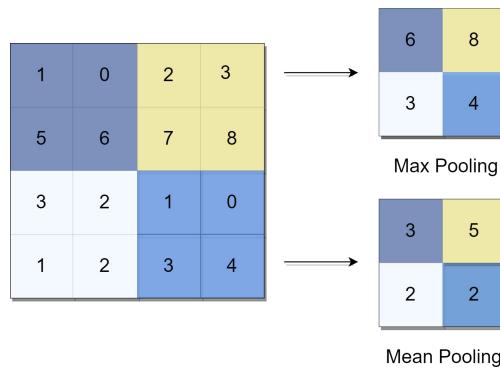


图 2-6 池化运算实例

### 2.3.1 VGGNet

2014 年，VGGNet 由牛津大学视觉几何课题组提出，并以牛津大学研究小组的名字命名。在整个 VGGNet 卷积神经网络中，全部采用  $3 \times 3$  的卷积核和  $2 \times 2$  的池化核。通过将小卷积核进行叠加，来实现大卷积核的效果。如图 2-7 所示，用两个  $3 \times 3$  的卷积叠加来代替  $5 \times 5$  的卷积，能够在保证计算结果的基础上，尽可能地减少计算量。

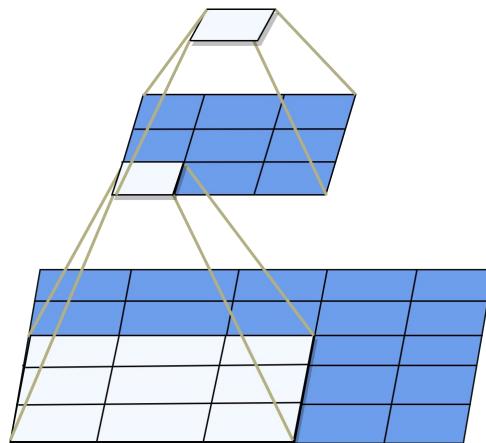


图 2-7 卷积叠加实例

VGGNet 中常用的模型是 VGGNet-16 与 VGGNet-19。VGGNet-16 网络结构如图 2-8 所示。每段卷积后都会有最大池化层，模型最后会有 3 个全连接层。VGGNet 证明了增加卷积网络的深度对于性能的提升有帮助。

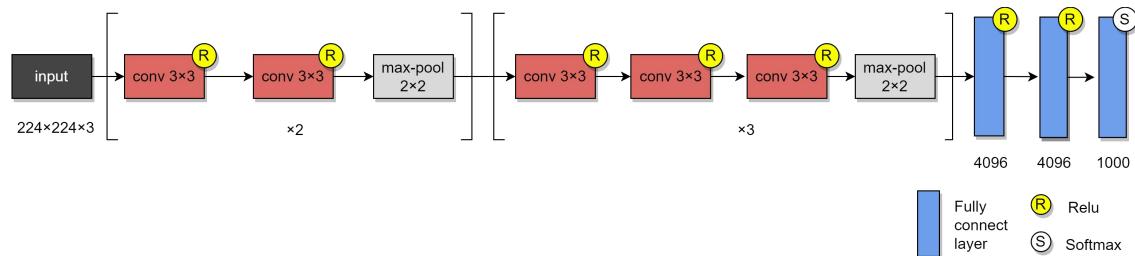


图 2-8 VGGNet-16 模型结构

### 2.3.2 ResNet

从 VGGNet 模型的经验来看，网络深度越深意味着能够提取到的特征越丰富。但事实证明，持续增加卷积网络的深度时，准确率会开始下降。因为随着网络层数的增多，会出现梯度爆炸或衰减现象<sup>[54]</sup>。

残差网络（Residual Networks，ResNet）通过在卷积神经网络结构中引入残差块来解决此类问题，基本残差块如图 2-9 所示。当新添加的一些层学习效果

非常差时，可以通过残差块将这些层的权重参数设置为0，从而直接跳过这一部分。假设卷积神经网络的输入是 $x$ ，在网络中使用恒等映射，将学习目标的函数 $H(x)$ 等价变换为 $F(x)+x$ ，于是ResNet相当于将学习目标改变为残差 $H(x)=H(x)-x$ 。

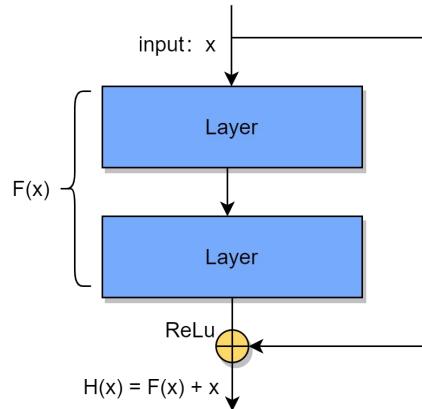


图 2-9 基本残差块

ResNet作为特征提取网络共存在5种结构。在本文中采用的是ResNet-50的网络结构，如图2-10结构所示，残差块结构由Conv block模块和Identity block模块组成。首先通过 $1\times 1$ 的卷积层使得输入的向量降低维度，然后通过 $3\times 3$ 的卷积层学习隐藏特征，最后再通过 $1\times 1$ 的卷积层将特征向量恢复至原维度，这种残差块结构可以减少网络的参数量，并且可以有效地解决深层网络退化问题。

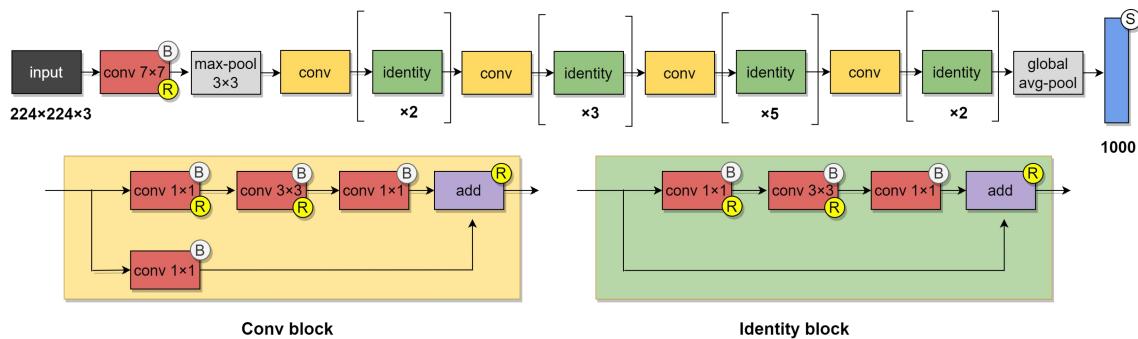


图 2-10 ResNet-50 模型结构

残差网络模型的出现，使模型深度在一定范围内不受限制。能够有效的降低卷积神经网络训练的复杂度，加速网络收敛，促进了深度卷积神经网络的发展。

## 2.4 多模态特征融合

在大数据时代，多媒体技术的迅速发展和应用造就了各种多模态数据。多模态数据可以是通过多领域或多视角得到的数据信息<sup>[55]</sup>。例如，图 2-11 (a) 一条视频中的元素可以被拆分成音频、图像和字幕等多模态信息；如图 2-11 (b) 所示，在新闻传播中，一条新闻可以拆分成本文和图像两种模态数据描述。



图 2-11 多模态数据样例

多模态融合（Multimodal Fusion）核心思想是将异构模态的信息进行交互融合，同时利用模态之间的互补性特点补充相关的细节。本文根据融合层次的不同将融合的种类分为三种：数据级融合、特征级融合以及决策级融合<sup>[56]</sup>。下面分别就这三种融合方式展开具体介绍。

### 2.4.1 数据级融合

数据级别的融合方式，是将多个模态的数据从形式层次进行统一。将统一成一种形式的数据使用深度学习方法或其他模型进行其他任务。使用本文研究内容中涉及的文本、图像两种模态为例，首先对数据集中的文本和图像两种模态的数据进行数据级融合，然后经过算法模型进行特征向量的抽取，将特征向量送入分类器中得到决策结果<sup>[57]</sup>，数据级融合过程如图 2-12 所示。

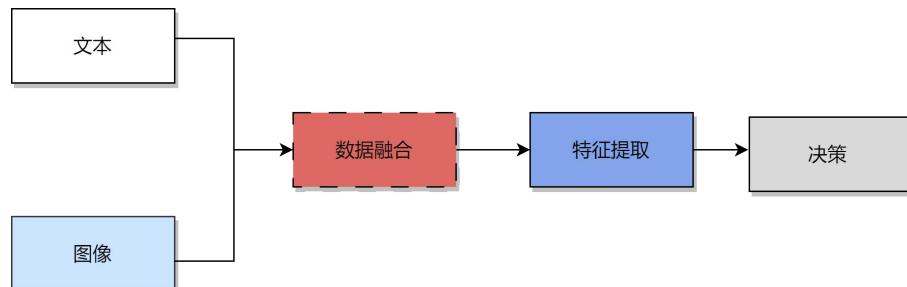


图 2-12 数据级融合

由于数据级融合方式仅适用于同一大类数据的融合。对于本文所涉及的虚假新闻检测的不同模态数据，可能会造成模态内容信息丢失的问题，因此本文不采用该方式。

### 2.4.2 特征级融合

特征级融合的核心思想是先对每个模态的数据进行特征提取，特征提取完成后得到特征向量，然后特征向量之间再进行互相融合，以得到更有效特征。以本文虚假新闻检测的研究内容为例，首先使用合适的模型提取出文本向量和图像向量。而后将特征最后输入到分类器中进行决策分类，图 2-13 展示了特征级融合的具体过程。

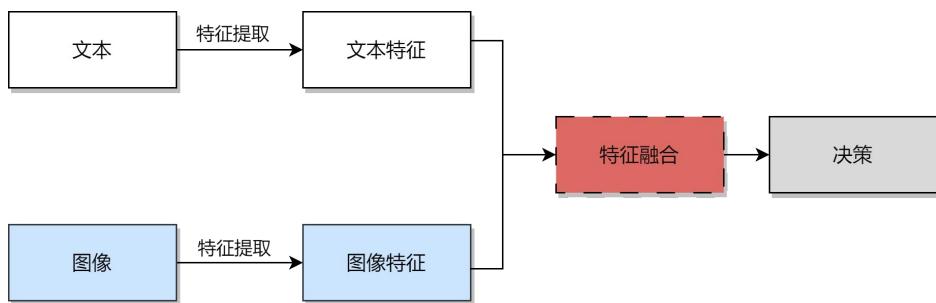


图 2-13 特征级融合

在特征级融合方法中，由于融合方式的不同，可以将特征级融合分为三种方法：基于拼接和线性组合的融合方法、基于注意力机制的融合方法<sup>[58]</sup>和基于双线性池化的融合方法<sup>[59]</sup>。三种特征融合方法的都是对特征向量进行计算操作，使得特征信息具有更丰富。

#### (1) 基于拼接和线性组合融合

简单拼接和线性组合方式使用一些类似于拼接或加权求和的方法对不同模态的特征信息进行融合。这种方法关联参数较少。

#### (2) 基于注意力机制的融合

注意力机制在每个时间步动态生成的一组标量权重向量的加权和。通常使用多个输出生成多组动态权重以进行求和。因此最终在拼接时候可以保存额外的权重信息以补充模态之间的信息。

#### (3) 基于双线性池化的融合

双线性池化通过创建不同模态数据之间的联合表示空间，利用外积的方式对不同模态之间的信息进行融合，充分利用不同模态向量元素间的独有特征。

### 2.4.3 决策级融合

决策级融合的层次在决策阶段，因此也被称为晚期融合。决策级融合的流程结构首先是要针对不同模态的数据训练不同的模型，然后利用一些常见的融合规则对不同模型的结果进行归纳，得到融合的输出。其中决策级融合规则一般包括最大值、平均值、投票法、贝叶斯规则等<sup>[57]</sup>。在多模态虚假新闻的背景下，决策级融合的流程结构如图 2-14 所示，首先对文本和图像两种模态分别搭建虚假新闻检测模型，从而得到两个决策结果，最后使用融合规则得到最终结果。

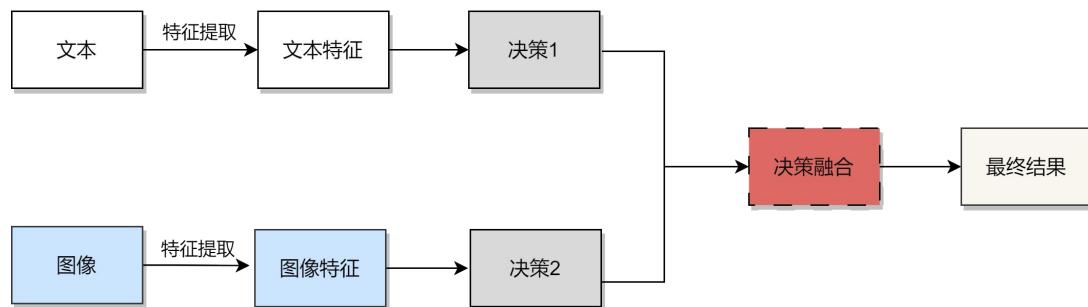


图 2-14 决策级融合

经过分析可知，所描述事件一致的情况下，模态间数据往往是相互支撑的，而决策级融合方式通过对比单一模态的结果，然后将结果进行规则组合，从而丧失了模态之间的交互信息。

## 2.5 本章总结

本章主要对虚假新闻检测的基础定义和相关技术进行分析。首先阐述了虚假新闻的定义和特征。接着详细介绍了文本表示方法和视觉表示方法。最后分析了不同多模态特征融合方式的特点。

## 第三章 基于双分支对抗网络的多模态虚假新闻检测算法

### 3.1 引言

互联网的发展与普及使得公众通过社交网络进行信息获取与交流。与此同时，社交网络上的虚假新闻日益增加，对社会产生了严重的负面影响。为此，应采取措施对虚假新闻进行识别和打击，加强对社会网络中虚假信息的监管和审查，同时加强对公众的媒体素养教育，提高公众识别虚假新闻的能力。

从内容上分析，虚假新闻通常会利用虚假或误导性的文本和图像来欺骗读者。虚假新闻的文本通常会包含错误、夸张、误导性或不准确的陈述，用来操纵读者的情绪或者误导他们的判断。虚假新闻的图像通常会被篡改或者被取出用于其他新闻使用，以达到误导或者欺骗的目的。鉴于此，深入分析虚假新闻的文本特征和图像特征，对现阶段的虚假新闻检测非常关键。

传统的虚假新闻检测方法主要针对单模态文本，主要利用文字本身的语言特征和人为特征，而近年来的研究逐渐发展到综合利用文本与图像两种多模态信息来检测假新闻。然而，在当前的多模态虚假新闻检测研究中，尽管一些方法已经使用了文本和图像模态的信息，但是缺乏有效的特征提取，信息得不到充分应用，从而损失了丰富的自身特征信息，导致检测性能不稳定、准确率低。其次，现有的工作通常都是在特定的新闻领域中进行特征提取，这些特征可能不适用于新的新闻领域。这可能会导致特征提取方法在应用于新领域时的性能下降，因为这些方法没有考虑到不同领域之间的差异。因此，确保特征提取方法具有普适性，使得该方法可以应用于不同领域，捕捉到不同领域虚假新闻的共性特征。

基于当前研究遇到的问题，本章提出一种基于双分支对抗网络的多模态虚假新闻检测模型（Multimodal Fake News Detection based on Two Branch Adversarial, MFNDTBA）。其中双分支网络用于有效地提取文本与图像的深层和浅层特征，领域对抗网络用于提取具有普适性的特征。本章的研究内容主要包括以下几点：

- (1) 本章提出使用双分支网络分别提取深层和浅层的文本和图像特征，挖掘出隐含在不同层次的特征信息。
- (2) 本章使用领域对抗网络提取不同新闻领域的通用特征，提高模型泛化

性能。

(3) 本章使用 Weibo 数据集和 Twitter 数据集验证本章模型的效果。根据大量实验结果和对比分析，证明该模型提高了虚假新闻检测的准确率。

## 3.2 相关工作

### 3.2.1 领域对抗网络

生成对抗网络模型（Generative Adversarial Networks, GAN）包含生成器和判别器。生成器的目标是要对真实数据的分布进行学习，生成与真实样本数据分布尽可能相似的合成样本，使其真实性不会被判别器所识别。相反地，判别器需要尽最大努力检测出合成样本的真实性。生成器可以被视为假钞生产者，其目标是生成更“真”的假钞。判别器可以被视为假钞鉴定者，其目标是识别出钞票的真伪。假钞生产者和假钞鉴定者互相达到其目标的过程可以视为一个博弈的过程。GAN 结构如图 3-1 所示。

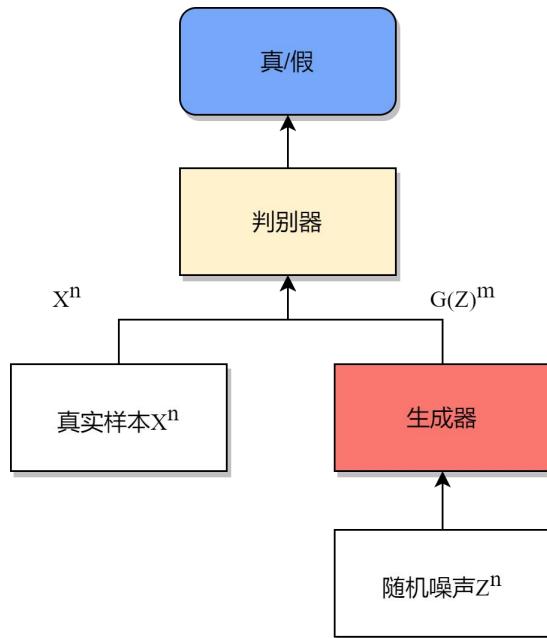


图 3-1 生成对抗网络（GAN）模型结构

在虚假新闻检测领域，由于新闻领域的多样性，导致不在同一领域的新闻所包含的特征具有互异性。若在特定新闻领域数据训练后的虚假新闻检测模型，对于其他领域一般会无效。因此为了得到兼顾多领域新闻数据的检测模型，大量且优质的数据是前期的必要工作。然而优质的数据大多数需要人工标注，这

个过程复杂且昂贵。因此模型的可迁移性是虚假新闻研究中需要解决的重要问题。

借鉴 GAN 的对抗思想，领域对抗网络 (domain-adversarial training of neural networks, DANN) 将目标域样本视为生成的数据，从而省去了 GAN 模型中生成的过程。生成器的角色转变为特征提取器。因此 DANN 的目标为提取源域中的特征，并让判别器识别不出提取的特征来自源域还是目标域。

DANN 是一种基于对抗训练策略的无监督领域自适应模型。该模型包含三部分：特征提取器（Feature extractor）、标签分类器（Label predictor）以及领域分类器（Domain predictor）。DANN 的结构示意图如图 3-2 所示。其训练过程如下：首先，利用特征提取器提取源域的特征并训练标签分类器。其次，将提取的特征作为是生成的数据，并训练领域分类器。最后，采用反馈机制更新参数达到均衡<sup>[60]</sup>。

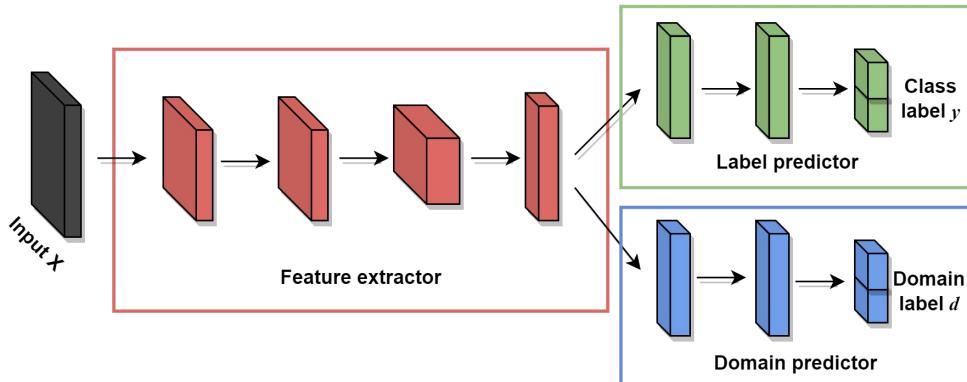


图 3-2 领域对抗网络 (DANN) 模型结构

DANN 可以提取领域不变性的普适性特征，通过最小化标签分类器的损失来提取表征能力强的深度特征。同时，最大化领域分类模块的损失使得源域和目标域特征不可区分，即得到两域的公共特征<sup>[60]</sup>。

### 3.3 基于双分支对抗网络的特征提取改进方法

多模态虚假新闻检测技术涉及文本、图像两种模态。本章将对文本和图像数据的特征进行分析，并使用合适的模型分别提取两种模态的特征。本章的 MFNDTBA 模型包括四个模块，文本特征提取模块（Text Feature Extractor）、图像特征提取模块（Image Feature Extractor）、领域对抗模块（Domain Adversarial）和虚假检测模块（Fake News Detector）。整体的结构图如图 3-3 所示。下文将对每一个模块进行具体介绍。

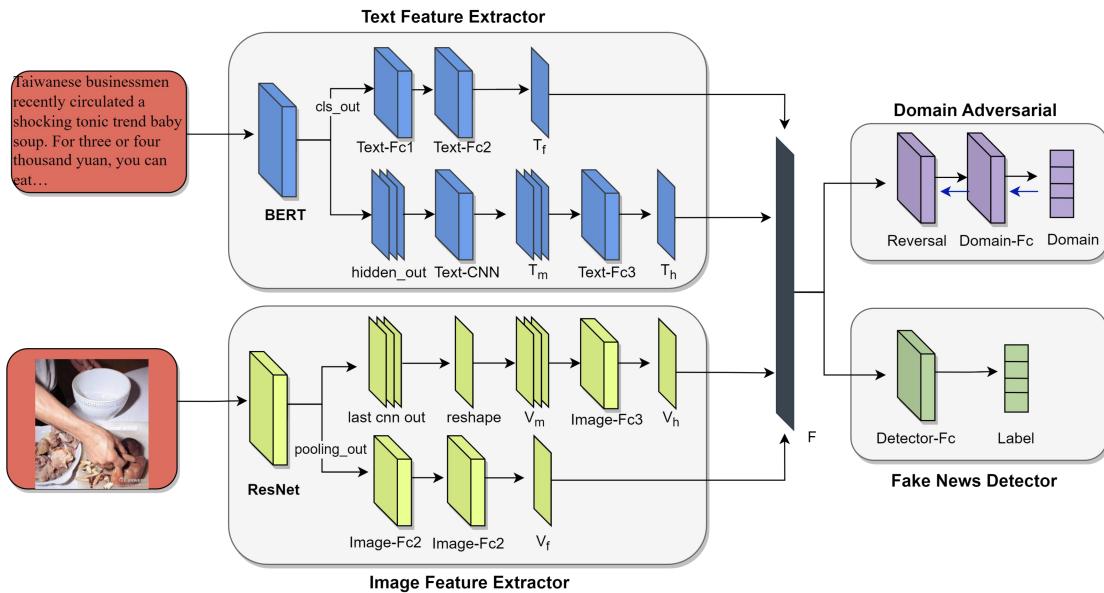


图 3-3 基于双分支对抗网络的多模态虚假新闻检测模型（MFNDTBA）

### 3.3.1 文本特征提取模块

文本模态内容从早期虚假新闻检测开始就起着不可或缺的作用。而目前多模态虚假新闻检测研究中，主要聚焦于更加有效地融合多模态之间的特征，而忽视单模态数据本身特征的有效提取。但是，提取出优质的文本语义特征是虚假新闻检测重要的前期工作。

新闻文字的限制导致语言层面通常会带有固执己见和煽动性的语言内容，可以根据不同维度上的情感特征、写作风格以及文字、词汇等来判断文本内容的虚假程度。因此可以使用双分支从整体句向量层面和细节词语层面进行特征的提取，同时借助深度学习的方法实现自动化提取流程。

双向预训练模型 BERT 具有优秀的语义建模能力，能够实现文本上下文信息的有效提取。该模型基于 Transformer 的编码器单元，并具有多层网络结构，可以学到多个层级的优质语义表示，为下游任务的开展做好充分准备。因此，BERT 方法是一种可行的方法。

基于上述思路，为了捕获文本上下文不同层次信息。使用双分支网络分别生成句向量纬度特征和词纬度特征，其结构如图 3-3 中的 Text Feature Extractor 蓝色部分所示。使用 BERT 对虚假新闻文本特征提取：首先，预处理新闻的文本内容，接着将清理后的文本数据送入 BERT 中。第一个分支提取模型最后一层输出的 *cls\_out* 特征，其维度为  $batchsize \times 768$ ，通过归一化层调整分布，同时使用 Dropout 层防止过拟合。最后通过两个全连接层调整大小得到最终文本句

向量  $T_f$ ，其过程如式(3-1)所示。

$$T_f = W_{t2}(W_{t1}(\text{Dropout}(\text{BN}(\text{cls\_out})))) \quad (3-1)$$

其中， $T_f$  为最终获取到的文本特征句向量，Dropout 为丢弃层，BN 为标准化层， $W_{t1}$ ， $W_{t2}$  为文本全连接层的权重矩阵。

第二个分支提取文本中隐含的词特征信息。使用最后四个隐藏词向量层输出的特征进行拼接，每个隐藏词特征的维度为  $\text{batchsize} \times \text{seq\_len} \times 768$ ，输入到具有 2、3、4、5 窗口大小的 Text-CNN 中，以提取不同隐藏层次特征  $T_m$ 。然后将  $T_m$  展平后并送入全连接层中调整形状大小获取隐含特征，最终获得特征  $T_h$ 。其过程如式(3-2)所示：

$$T_h = W_{t3}(\text{flatten}(\text{Text-CNN}(\text{hidden\_out}))) \quad (3-2)$$

其中  $T_h$  为隐藏词特征输出，Text-CNN 为文本卷积神经网络， $W_{t3}$  为全连接层的权重矩阵。

### 3.3.2 图像特征提取模块

图像模态在新闻文本描述的基础上提供了重要的辅助性信息传递方式。社交媒体上的新闻为了使新闻更直观易懂，一般会使用丰富的图像信息进行补充。并且发布虚假新闻的造谣者为了引发读者的共情能力，会使用图像误导读者。因此获取具有深层语义的图像特征对于多模态检测来说具有重要意义。

与真实新闻图像相比，虚假新闻图像从浅层角度来分析通常质量较差，且由于一些修改合成操作，使得虚假图像具有多尺度、较复杂的特征信息。因此对于这些不同粒度的特征信息，可以从深层和浅层提取图像本身的不同层次特征，实现对图像内容特征的全面挖掘。

当前，大部分的图像关联研究都是利用卷积神经网络来进行特征抽取。卷积神经网络能够对图像中的纹理、颜色、空间关系等信息进行有效地建模。在一些深度卷积神经网络中，会存在梯度消失问题。在 ResNet 模型中，采用“捷径连接技术”的方式，可以将输入的特征信息跨层传递，有效地解决梯度丢失和特征降质问题。因此本文拟采用 ResNet-50 残差网络构建图像特征抽取模型，有效避免因特征缺失和语义缺失而造成的错误识别，增强模型的稳健性。

基于上述思路，使用双分支结构在图像不同层次上得到特征信息，具体来说其结构如图 3-3 中的 Image Feature Extractor 黄色部分所示。将图像数据输入到 ResNet-50 模型中，第一个分支在 ResNet-50 网络的最后一层添加两个全连接层，得到输出向量，然后调整分布并避免过拟合，得到最终的图像特征表示  $V_f$ ，其过程如式(3-3)所示。

$$V_f = W_{v2}(W_{v1}(\text{Dropout}(\text{BN}(\text{pooling\_out})))) \quad (3-3)$$

其中,  $V_f$  为最终获取到的文本特征句向量, Dropout 为丢弃层, BN 为标准化层,  $W_{v1}$ ,  $W_{v2}$  为文本全连接层的权重矩阵。

第二分支提取最后一个卷积层的输出，将特征维度 reshape 为二维张量，生成图像区域特征  $V_m$ 。然后将  $V_m$  展平并送入全连接层调整形状，最终获得特征  $V_h$ ，其过程如式(3-4)所示。

$$V_h = W_{v3}(\text{flatten}(\text{reshape}(last\_cnn\_out))) \quad (3-4)$$

其中  $V_h$  为隐藏层特征输出,  $W_{v_3}$  为全连接层的权重矩阵。

将文本特征  $T_f$ 、 $T_h$  与图像特征  $V_f$ 、 $V_h$  拼接起来得到最终的特征  $F$ 。其过程如式(3-5)所示：

$$F = concat([T_f, T_h, V_f, V_h]) \quad (3-5)$$

### 3.3.3 领域对抗模块

对于新闻来说，其领域涉及社会中的各行各业，如图 3-4 列举了 Weibo 数据集中聚类后部分主题的词云效果，可以看出本文中的新闻数据涉及了多个新闻领域，包括社会、国家、健康等领域。



图 3-4 Weibo 数据词云主题

如果直接使用不同新闻领域的数据对模型进行训练，模型仅获取特定领域知识且泛化能力较差，只能对特定领域相关的新闻进行检测。因此必须去除面向某一特殊领域的特定特征，以获取到事件不变的特征表示。

领域对抗网络研究的重点正是提取源领域和目标领域之间的共性特征信息，通过博弈的方式优化目标。因此本章基于领域对抗网络的思想，将领域对抗网络引入到模型训练阶段对领域类别进行预测，并通过损失来估计不同领域间特征表达的差异，从而使模型能够学习到非领域依赖性的特征。

在本文中，领域对抗模块是由两个具有对应激活函数的全连接层构成。其结构如图 3-3 中的 Domain Adversarial 部分所示。目的是利用多模态的特征表示，将新闻分类为 K 类领域类别之一。在对抗过程中，由于领域分类损失越大证明不同新闻领域之间的特征表示越具有一致性，因此特征提取模块总是尝试蒙骗

领域对抗模块，从而达到最大化领域分类损失的目的。也就暗示着模型训练得到的特征是领域之间的不变特征。同时领域对抗模块则试图通过挖掘多模态特征表示中所蕴含的特征信息，实现对新闻领域的准确判断。总结来说，多模态特征提取模块与领域对抗模块建立了一个对抗流程：多模态特征提取模块尝试通过最大化领域对抗损失来提取领域间不变特征，而领域对抗模块企图从多模态特征中发现领域特定信息进而最小化领域分类损失。领域对抗模块使用交叉熵函数作为衡量标准。损失函数定义如式(3-6)所示：

$$L_{dom} = -E_{(x, y_e) \sim (X, Y_e)} \left[ \sum_{k=1}^K y_e \log(dom(F; \theta_{dom})) \right] \quad (3-6)$$

其中  $dom$  是领域对抗模块， $\theta_{dom}$  是领域对抗模块参数， $dom(F; \theta_{dom})$  是模型预测的领域类别概率。 $Y_e$  表示领域类别集合， $X$  为新闻集， $y_e$  表示新闻  $x$  的实际领域类别分类。

多模态特征提取模块提取到的特征表示被视为生成的数据，使用领域对抗网络对多模态特征表示进行分析研判，逐步迭代更新模型中特征提取模块的参数信息。损失  $L_{dom}$  被用来估算各领域分布之间的差值。损失变大意味着不同领域表征的分布是相似的，并且学到的特征是共性特征。因此为了消除领域多变化的特征，需要最大化区分损失，以寻求最佳的参数。

### 3.3.4 虚假检测模块

虚假检测模块通过输入多模态特征来判断新闻的真假，使用激活函数为  $softmax$  的全连接层作为检测模块的组成部件，最后得到二分类结果。其结构如图 3-3 中的 Fake News Detector 部分所示。虚假检测模块的计算流程如式(3-7)所示。

$$\hat{y} = fnd(F; \theta_{fnd}) \quad (3-7)$$

其中  $fnd$  为虚假判别模块， $\theta_{fnd}$  为新闻检测模块参数， $\hat{y}$  为模型预测真假分类概率。同样的虚假检测模块的损失函数为交叉熵函数。如式(3-8)所示。

$$L_{fnd} = -E_{(x, y) \sim (X, Y)} [y \log(\hat{y}) + (1-y) \log(1-\hat{y})] \quad (3-8)$$

其中  $Y$  为新闻真假标签类别集合， $X$  为新闻集合， $y$  是指新闻  $x$  的实际分类标签。

### 3.3.5 损失函数

为了提高虚假检测模块的检测准确率，需要将虚假检测模块损失最小化。在整个模型训练过程中，多模态特征提取模块、领域对抗模块和虚假检测模块

共同参与到损失函数中。虚假检测模块用来提高检测新闻的准确率，领域对抗模块和特征提取模块形成对抗性质用来生成通用特征，模型总损失定义如公式(3-9)所示。

$$L = L_{fnd}(\theta_{dom}) - \lambda L_{dom}(\theta_{dom}) \quad (3-9)$$

其中  $\lambda$  为平衡系数，用于平衡虚假检测模块和领域对抗模块的损失。

模型的目标为最小化损失函数，以找到最优解参数。在多模态特征提取模块与领域对抗模块中间加入反向梯度层实现特征提取模块和领域对抗模块的对抗效果。在前向传播期间，反向梯度层充当恒等变换即不改变数据的值。反向传播过程中，到达该层时将导数乘以负系数得到的结果作为一个新的梯度传递到下一层。因为反向梯度层的影响，特征提取模块的参数会顺着减小领域对抗损失函数的负方向训练，最终达到对抗性。

## 3.4 实验设计

在本小节中，首先对多模态虚假新闻检测所需的数据集进行介绍，然后阐述多模态新闻检测所需的相关对比模型，最后通过实验来分析所提模型的效果。

### 3.4.1 数据集

在多模态虚假新闻检测中使用到两个权威公开数据集：Weibo 数据集、Twitter 数据集。两个数据集的统计信息如表 3-1 所示。下面将分别进行介绍。

表 3-1 多模态数据集

统计	Weibo	Twitter
虚假新闻数量	4749	7021
真实新闻数量	4779	5794
图片数量	9528	517

#### (1) Weibo 数据集

Weibo 数据集来自中国的权威消息来源，其中的虚假新闻是经过微博的官方验证系统进行验证后的数据<sup>[34]</sup>。在预处理操作中，删除数据集中的重复图像和质量不高的图像，以确保数据集的质量。按照 7:1:2 的比例将数据分为三个部分，即训练集、验证集和测试集。

#### (2) Twitter 数据集

Twitter 数据集来自文献<sup>[7]</sup>，在 Twitter 社交平台上发布的新闻数据大部分被该数据集进行汇集。每条新闻都是由文字内容和附加的图片或视频内容组成。

本文研究的内容是基于文字和图像内容，因此附有视频的新闻将被删除。

### 3.4.2 评价标准

为了验证模型的有效性，本文采用分类中使用的准确率（accuracy）、精确率（precision）、召回率（recall）和 F1 值这四个指标来对实验产生的结果进行有效地评估。关于分类问题，有四个基本的指标，分别为 TP（True Positives）、TN（True Negatives）、FP（False Positives）和 FN（False Negatives）<sup>[61]</sup>。其中，TP 表示在虚假类别中标签为虚假新闻，实际预测也被预测为虚假新闻；TN 表示在虚假类别中标签为真实新闻，实际预测也被预测为真实新闻；FP 表示在虚假类别中标签为真实新闻，在实际预测分类中被预测为虚假新闻；FN 表示在虚假类别中标签为虚假新闻，在实际预测分类中被预测为真实新闻。评价模型的四个指标：准确率、精确率、召回率和 F1 值的计算方法如式(3-10)~(3-13)所示：

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3-10)$$

$$\text{precision} = \frac{TP}{TP + FP} \quad (3-11)$$

$$\text{recall} = \frac{TP}{TP + FN} \quad (3-12)$$

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (3-13)$$

### 3.4.3 实验环境

本章的实验基本环境是在 Linux 操作系统和 Python3.8 编程环境下进行，使用 GPU 进行加速训练，整体的环境设置如表 3-2 所示：

表 3-2 实验环境设置

实验环境	配置
操作系统	Linux
GPU	RTX2080
内存	15GB
编程语言	Python3.8
深度学习框架	Pytorch1.5.1

同时在整个模型的训练中，超参数的设置如表 3-3 所示。对于实验中提取

文本内容的特征，采用预训练的 BERT-base 模型。文本最终向量维度大小设置为 32。对于图像内容特征，采用预先训练的 Resnet-50 网络进行提取特征，并设置最终图像特征向量的特征维度设置为 32。本章中的总损失目标函数里面的平衡系数  $\lambda$  设置为 1。

表 3-3 超参数设置

模型参数	具体配置
Batch Size	32
Epoch	100
优化器	Adam
学习率	0.001
丢弃率	0.2

### 3.4.4 对比实验模型

为了衡量本文模型的性能，本小节将根据输入模态数将对比模型分为：单模态虚假新闻检测模型、多模态虚假新闻检测模型。

- (1) 单模态虚假新闻检测模型
  - 1) Text-GRU：将经过预训练的词嵌入输入 Bi-GRU 模型中以提取文本特征向量，然后将特征向量输入全连接层进行分类新闻。
  - 2) Image-VGG：将新闻中的图像内容送入 VGG-19 网络得到图像特征，接着将特征输入全连接层检测新闻真实性。
- (2) 多模态虚假新闻检测模型
  - 1) att-RNN<sup>[34]</sup>：对文本内容和社会语境内容采用 LSTM 模型进行抽取特征信息，对于图像特征使用 VGG-19 模型进行提取。利用注意力机制对三种特征进行融合，最终将特征拼接并送入分类器进行分类。
  - 2) EANN<sup>[35]</sup>：对于文本特征的抽取，采用 Text-CNN 模型。对于图像特征抽取，使用 VGG-19 模型。图像特征和文本特征拼接后输入事件判别器和新闻分类器进行检测。
  - 3) MVAE<sup>[37]</sup>：应用变分自编码器模型得到文本和图像之间关联性信息，将得到的特征送入新闻分类器以检测虚假信息。
  - 4) Spotfake<sup>[38]</sup>：使用预训练语言模型 BERT 来学习文本特征，使用预训练的 VGG19 模型学习图像特征。将特征拼接送入虚假检测模块进行检测。
  - 5) BDANN<sup>[62]</sup>：对于文本特征信息，使用 BERT 模型提取。对于图像特征信息，使用 VGG-19 模型提取。同时应用域分类器，消除特征依赖，最后进行检测新闻。

6) MFNDTBA: 本章提出的模型, 使用 BERT 模型从不同层次抽取文本特征信息, 使用 ResNet-50 从深层和浅层提取图像内容信息, 通过领域对抗网络提取领域间共性特征, 最后送入虚假检测模块中检测。

## 3.5 实验结果分析

本小节将对本章提出的双分支特征提取方法以及领域对抗方法进行多次实验, 通过对比实验结果分析和消融实验分析, 验证所提出的特征提取方法、领域对抗方法的有效性。

### 3.5.1 对比实验结果分析

将本文模型 MFNDTBA 与多个对比模型进行大量的实验分析, 实验结果数据如表 3-4 和图 3-5 所示。从表 3-4 和图 3-5 的实验结果可以得出, Weibo 数据集和 Twitter 数据集上的多模态模型的准确率均高于单模态模型, 说明本文使用多模态数据进行特征抽取优于单模态数据, 证明了多模态特征的有效性。

表 3-4 MFNDTBA 与对比模型实验结果

数据集	方法	Accuracy	真新闻			假新闻		
			Precision	Recall	F1	Precision	Recall	F1
Weibo	Text-GRU	0.643	0.662	0.578	0.617	0.662	0.578	0.617
	Image-VGG	0.633	0.630	0.500	0.550	0.630	0.750	0.690
	att-RNN	0.772	0.797	0.713	0.692	0.684	0.840	0.754
	EANN	0.816	0.820	0.820	0.820	0.810	0.810	0.810
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	BDANN	0.842	0.830	0.870	0.850	0.850	0.820	0.830
	Spotfake	0.892	0.902	0.964	0.932	0.847	0.656	0.739
	<b>MFNDTBA</b>	0.896	0.923	0.874	0.898	0.868	0.920	0.893
Twitter	Text-GRU	0.526	0.586	0.553	0.569	0.469	0.526	0.496
	Image-VGG	0.596	0.695	0.518	0.593	0.550	0.700	0.599
	att-RNN	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN	0.719	0.642	0.474	0.545	0.771	0.870	0.817
	MVAE	0.745	0.801	0.719	0.758	0.686	0.777	0.730
	BDANN	0.830	0.810	0.630	0.710	0.830	0.930	0.880
	Spotfake	0.777	0.751	0.900	0.820	0.832	0.606	0.701
	<b>MFNDTBA</b>	0.847	0.821	0.684	0.746	0.856	0.927	0.890

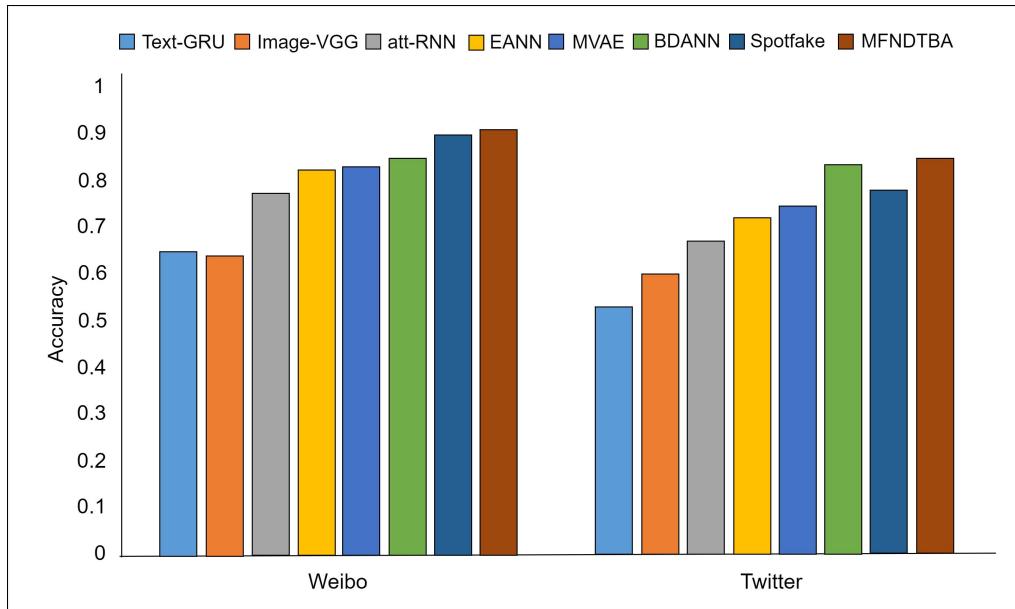


图 3-5 MFNDTBA 与对比模型准确率结果

从双分支的角度分析实验结果，由于 EANN 模型和 MVAE 模型并未从双分支多角度进行特征提取，在 Weibo 数据集上的实验结果显示，MFNDTBA 的准确率比 EANN 高出 9.8%，比 MVAE 高出 8.7%，证明了 MFNDTBA 模型从双分支网络层面增强了文本和图像本身的深层和浅层的特征信息，提取更广泛更具有语义的文本特征和图像特征，从而提高了虚假新闻检测的有效性。

从领域对抗的角度分析实验结果，att-RNN 模型和 Spotfake 模型都是从新闻多领域中直接提取特征，都不含有消除特定领域特征的功能模块，在 Twitter 数据集上实验结果也同样显示，MFNDTBA 模型的准确率比 att-RNN 模型高出 27.6%，比 Spotfake 模型高出 9%，证明领域对抗模块可以对来自不同领域的新闻数据特征进行特征筛选，使得提取的特征更具有通用性，增加模型的泛化性能和鲁棒性能，从而提高模型的性能表现。

因此，本章提出的基于领域对抗的双分支网络首先有效地提取了文本内容和图像内容深层和浅层的特征状态信息，从文本和图像本身提取出了更有效的特征。其次使用领域对抗模块解决了领域特征通用性的问题，使得提取的特征更具有普适性，从而提高了虚假新闻检测整体检测的性能。

### 3.5.2 消融实验结果分析

为了更好地验证本章模型中双分支模块和领域对抗模块对实验结果的影响，分别在 Weibo 和 Twitter 两个关于多模态虚假新闻检测的数据集上进行消融实验，设计两种对比模型进行分析。

(1) 去除双分支模块。使用单分支网络从模型中提取文本特征和图像特征输出，然后拼接特征，将拼接后的特征输入到虚假检测模块和领域对抗模块。该模型定义为 MFNDTBA-1。

(2) 去除领域对抗模块。使用双分支网络提取文本和图像内容特征，这些特征被连接起来并送入虚假检测模块。该模型定义为 MFNDTBA-2。

基于双分支对抗网络的模型消融实验结果如图 3-6 所示。根据实验结果分析，去除双分支模块后的 MFNDTBA-1 模型准确率在 Weibo 数据集和 Twitter 数据集上均比原模型低，证明了没有双分支网络模块时存在一定程度文本特征信息和图像特征信息的丢失，再次验证了双分支网络对特征提取的有效性。其次当模型中存在领域对抗模块时，模型的准确率在两个数据集上均提升了 1% 左右，证明领域对抗机制的存在确实更有助于寻找通用性特征，从而使得模型更具有泛化性能。

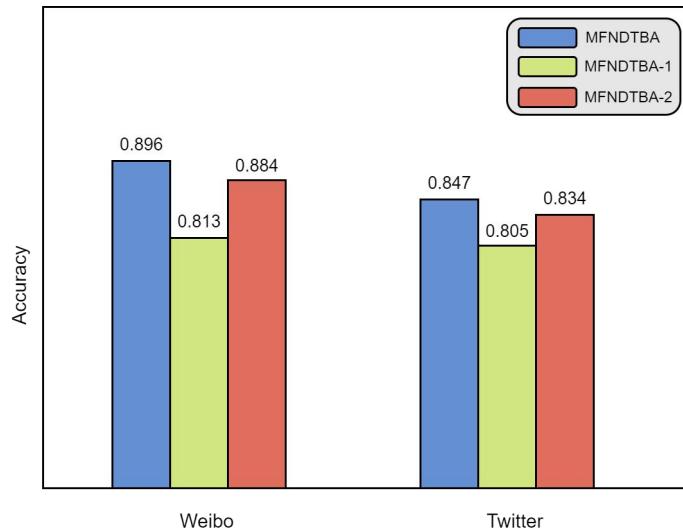


图 3-6 MFNDTBA 消融实验结果

## 3.6 本章总结

本章提出了一种基于双分支对抗网络的虚假新闻检测模型。对新闻内容中的图像信息和文本信息进行深度挖掘，提取不同层次的特征，得到更全面有效的特征表示。其次，该模型使用领域对抗网络来捕获通用特征以提高模型迁移性能。本章为了证明双分支特征提取方法以及领域对抗模块的效果，进行多组对比实验分析，实验证明了该模型特征提取的有效性。

## 第四章 基于组合式融合机制的多模态虚假新闻检测算法

### 4.1 引言

以文本为单一载体的新闻表现形式已经成为过去式，文本和图像并存的多模态形式是现阶段新闻的主流呈现方式。由于文本和图像模态之间的数据信息可以互相进行补充说明，鉴于此，对多模态数据进行研究势在必行。

在第三章的工作中，文本和图像模态内部的有效特征被充分提取。但多个模态间数据的语义信息是可以互相进行铺垫和补充的。就虚假新闻而言，文本信息和图像信息都是用来更全面地描述新闻的核心表达信息。新闻图像一般情况下注重于视觉效果。新闻文本则更注重于语言表达。同时文本可以补充描述图像所呈现的视觉信息，而图像则可以提供文本中没有提及的物体、场景、情感等信息。二者既有密切的联系，又有各自独特的特点。同时在多模态融合中，不同模态内部的某些关键单词或区域会提供更重要的信息，因此应当为这些单词或区域分配更大的权重，以提高其对整个模态的贡献。因此，充分融合文本和图像的信息可以挖掘出两种模态间的关联和互补信息，进一步提升信息的理解和表达能力<sup>[63]</sup>。然而当前研究在利用图像特征联合文本特征的虚假新闻检测方法中，模态间信息只是简单地拼接并没有考虑到两种模态之间的有效融合，也无法充分发挥多模态信息的优势。因此，进行有效的多模态融合，是当前虚假新闻检测研究的重要工作。

综合之前的研究内容，本章开展多模态特征融合方法调研，提出基于组合式融合机制的虚假新闻检测算法（Multimodal Fake News Detection based on Combinatorial Fusion Mechanism, MFNDCFM）来检测新闻。在第三章的基础上增加特征融合模块，包括以多模双线性池化为基础的模态间信息交互模块和以自注意力机制为基础的模态内信息增强模块。本章研究内容主要包括三方面：

- (1) 本章提出了一种基于组合式融合机制的多模态虚假新闻检测算法，可以更好地融合多模态特征数据。
- (2) 本章的组合式融合机制使用了多模线性池化进行模态间信息交互以补充特征的丰富性，使用自注意力机制进行模态内部信息增强以加强自身特征的有效性，实现异构模态数据内外的交互，有效提高了特征的表示能力。
- (3) 基于 Weibo 数据集和 Twitter 数据集进行实验分析，验证了多模态特

征融合方法的有效性。

## 4.2 相关工作

### 4.2.1 多模双线性池化

多模双线性池化源于双线性池化。双线性池化模型是解决细粒度图像识别任务的一项工作，这种模型主要由两路网络组成。双线性池化模型的两路网络在学习图像数据的特征向量时，同时完成两个阶段的工作，然后在网络末端将获得的结果进行融合从而获得更有效的特征<sup>[64]</sup>。其实质是对两个特征进行计算，其过程如下所示：对于图像  $I$  在位置  $l$  的两个特征  $f_x(l, I)$  和  $f_y(l, I)$ ，将特征  $f_x(l, I)$  与  $f_y(l, I)$  进行向量相乘之后得到矩阵  $m(l, I, f_x, f_y)$ ，而后对矩阵进行 Sum pooling 操作过程后得到矩阵结果  $\xi(I)$ ，接着将矩阵张成一个双线性向量  $p = \text{vec}(\xi(I))$ ，接着对向量做矩归一化操作后得到向量  $q = \text{sign}(x)(\sqrt{|p|})$ ，再进行 L2 归一化得到最终融合后的特征  $z = q / \|q\|_2$ 。这种方法能够提取到更丰富的特征，提高识别性能。

表 4-1 Algorithm of Count Sketch

---

#### Algorithm: Count Sketch

---

```

1: input:  $v \in R^n$ 
2: output:  $y \in R^d$ 
3: if  $h, s$  not initialized then
4:   for  $i \leftarrow 1 \dots n$  do
5:     sample  $h[i]$  from  $\{1, \dots, d\}$ 
6:     sample  $s[i]$  from  $\{-1, 1\}$ 
7:    $v = \Psi\{v, h, s, n\}$ 
8: procedure  $\Psi\{v, h, s, n\}$ 
9:    $y = [0, \dots, 0]$ 
10:  for  $i \leftarrow 1 \dots n$  do
11:     $y[h[i]] = y[h[i]] + s[i] \cdot v[i]$ 
12:  return  $y$ 

```

---

在双线性池化过程中两个特征向量进行外积相乘操作时，得到的特征维度过高，计算非常复杂。因此研究者们针对降维问题进行研究并提出解决方法，利用 Count Sketch<sup>[65]</sup>方法从外积结果中得到低维度特征。这种方法使得双线性池化可以用于多模态融合中，Count Sketch 具体降维过程的伪代码如表 4-1 所示，其目的是对向量  $v \in R^n$  进行降维，输出  $y \in R^d$  的过程：首先随机初始化两个向量

$s \in \{-1, 1\}^n$ ,  $h \in \{1, \dots, d\}^n$ , 用于降维过程中使用。 $s$  向量包含  $v$  向量每个索引随机的 1 或 -1 值,  $h$  将输入的  $v$  向量索引  $i$  映射至  $y$  向量索引  $j$  的值。首先  $y$  被初始化为零向量。对于需要被降维的元素  $v[i]$ , 使用  $h \in \{1, \dots, d\}^n$  查找  $y$  向量目标索引  $j = h[i]$ , 并将  $s[i] \cdot v[i]$  添加到  $y[j]$ 。

为了避免明确计算外积, Pham 证明两个向量外积后的值进行 Count Sketch 降维的操作, 等价于两个向量先分别 Count Sketch 降维操作后的结果的卷积<sup>[66]</sup>。则计算过程如式(4-1)所示。

$$\Psi(x \otimes q, h, s) = \Psi(x, h, s) * \Psi(q, h, s) \quad (4-1)$$

其中,  $x$  和  $q$  分别表示两种特征,  $h$  与  $s$  分别表示初始化向量  $h \in \{1, \dots, d\}^n$  和  $s \in \{-1, 1\}^n$ , 用于高维空间特征  $x$  和  $q$  向低维空间的映射, 进而将两个特征映射到低维空间进行卷积计算, \* 代表卷积操作。

因此可以先将两个特征分别降维至低维度特征, 再进行卷积计算。但是计算过程所花费的成本较大。依据卷积定理的定义, 在时域上的卷积过程等价于在频域上的点积过程。故可以通过傅里叶变换将两个降维后的特征向量变换至频域, 并在频域作向量内积, 然后将内积结果作傅里叶逆变换至时域空间求得最后结果。式(4-2)和式(4-3)代表两个特征降维后的低维特征, 式(4-4)代表变换至频域做点积操作, 然后逆变换到时域得到结果。

$$x' = \Psi(x, h, s) \quad (4-2)$$

$$q' = \Psi(q, h, s) \quad (4-3)$$

$$x' * q' = FFT^{-1}(FFT(x') \odot FFT(q')) \quad (4-4)$$

其中  $\odot$  表示点积操作。

上述整个过程为多模双线性池化模块 (Multimodal Compact Bilinear Pooling, MCBP) 的执行流程。MCBP 模块的构造如图 4-1 所示。

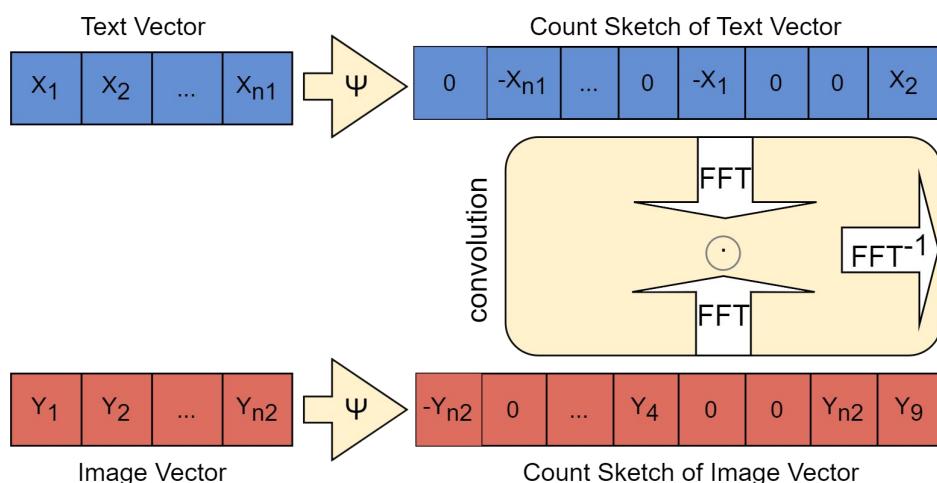


图 4-1 多模双线性池化模块 (MCBP)

### 4.2.2 自注意力机制

自注意力机制（Self-Attention Mechanism）是一种基于注意力机制的方法，能够让模型更好地关注输入序列中的相关信息，从而提高模型的性能。在图像模态中，可以使用自注意力机制来捕捉不同区域之间的关联性，提取图像特征的空间分布；在文本模态中，自注意力机制可以用来捕捉单词之间的关联性，提取文本特征的序列关系。

注意力机制（Attention Mechanism, AM）的灵感来自人类的视觉感知系统。人类在观察图像时，会根据目标的重要性和上下文信息进行选择性地关注。因此注意力机制是一种能够对输入的不同位置或特征的重要性进行不同程度的加权处理的机制。一般来说，注意力机制包括三部分：查询向量（ $Query, Q$ ）键向量（ $Key, K$ ）和值向量（ $Value, V$ ）。其中， $Query$  用于指示模型应该关注哪些位置或特征， $Key$  和  $Value$  则分别用于表示输入序列中每个位置或特征的信息，假设其输入长度为  $n$ 。通过给定的查询向量  $Query$ ，计算  $Query$  与所有  $Key$  的注意力分数，得到  $Key$  相对应  $Value$  的权值，将注意力分数视为权重系数，然后对  $Value$  加权求和，以得到最终的 Attention Value。图 4-3 展示了注意力机制的步骤分解细节，每个阶段的过程描述如下。

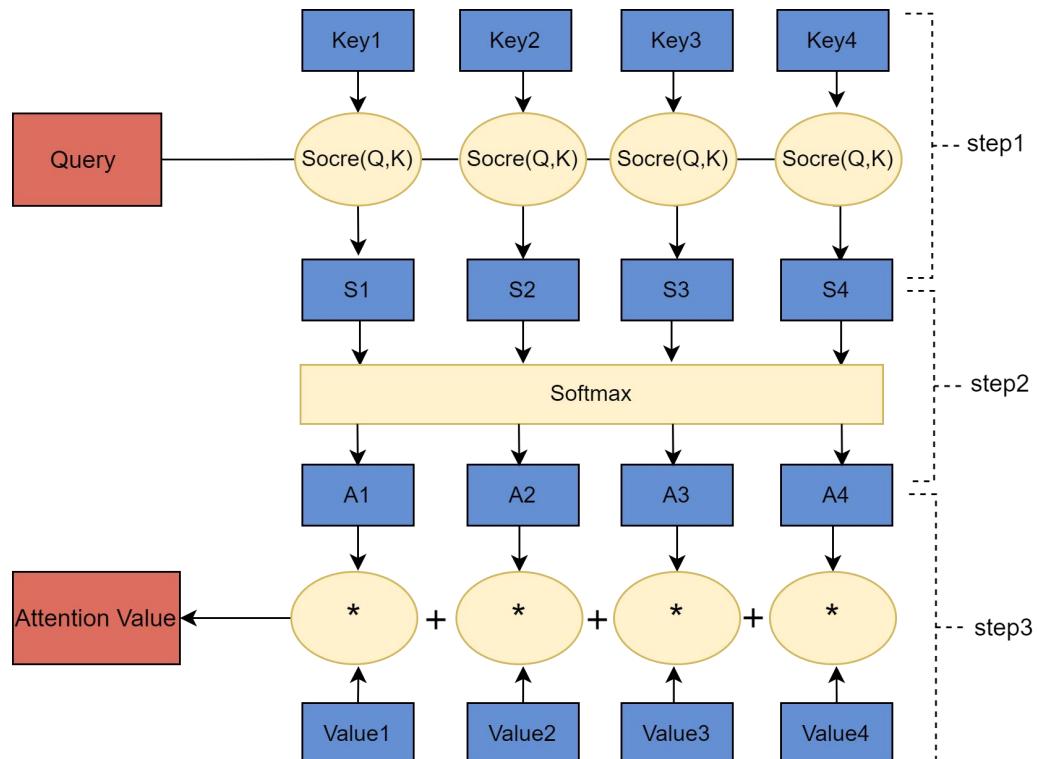


图 4-2 注意力机制计算步骤分解

(1) 第一阶段：通过注意力计算函数，计算 *Query* 和 *Key* 的注意力分数。常用计算函数如表 4-2 所示。常用的注意力计算函数包括加性函数、点积函数、缩放点积函数。

表 4-2 常用注意力分数计算函数

模型	公式
加性函数	$s(q, k) = v^T (Wq + Uk)$
点积函数	$s(q, k) = q^T k$
缩放点积函数	$s(q, k) = \frac{q^T k}{\sqrt{d}}$

(2) 第二阶段：对第一阶段得到的注意力权值分数进行归一化处理操作，计算过程如式(4-5)所示。

$$A_i = softmax(s(q, k_i)) = \frac{\exp(s(q, k_i))}{\sum_{j=1}^n \exp(s(q, k_j))} \quad (4-5)$$

其中  $s(q, k_i)$  表示在第一阶段计算得到的分数值， $n$  表示输入序列的长度。 $A_i$  即为对应的权重。

(3) 第三阶段：使用第二阶段中得到的权重  $A_i$  与 *Value* 进行加权求和，得到注意力权重 *Attention Value*。计算过程如式(4-6)所示。

$$Attention(Query, Key, Value) = \sum_{j=1}^n A_i * Value_i \quad (4-6)$$

其中， $n$  表示输入序列的长度， $A_i$  表示 *Value* 对应的权重系数。

自注意力机制基于注意力机制，被广泛使用于机器阅读理解、机器翻译等领域上。自注意力机制是在  $Query = Key = Value$  的条件下得到的，如式(4-7)所示。此时可以使用缩放点积注意力作为注意力分数函数来计算注意力值。

$$Attention(Query, Key, Value) = softmax(\frac{QK^T}{\sqrt{d_k}}) V \quad (4-7)$$

其中， $d_k$  是 *Key* 的维数，为了将内积结果控制在较小范围内，将  $d_k$  当作分母。

### 4.3 基于组合式融合机制的特征融合改进方法

描述新闻的多模态数据之间存在着强联系，因此本章着重研究融合改进方

法，选取合适的融合方法分别从模态间和模态内部对两种模态的输入数据进行特征融合。本章的 MFNDCFM 模型结构在第三章的基础之上增加了以组合式融合机制为基础的特征融合模块。模型结构如图 4-3 所示，包括多模态特征提取模块（Feature Extractor）、多模态特征融合模块（Fusion Module）、领域对抗模块（Domain Adversarial）和虚假检测模块（Fake News Detector），其中多模态特征融合模块分为模态间信息交互模块和模态内信息增强模块。下面着重对融合模块进行介绍。

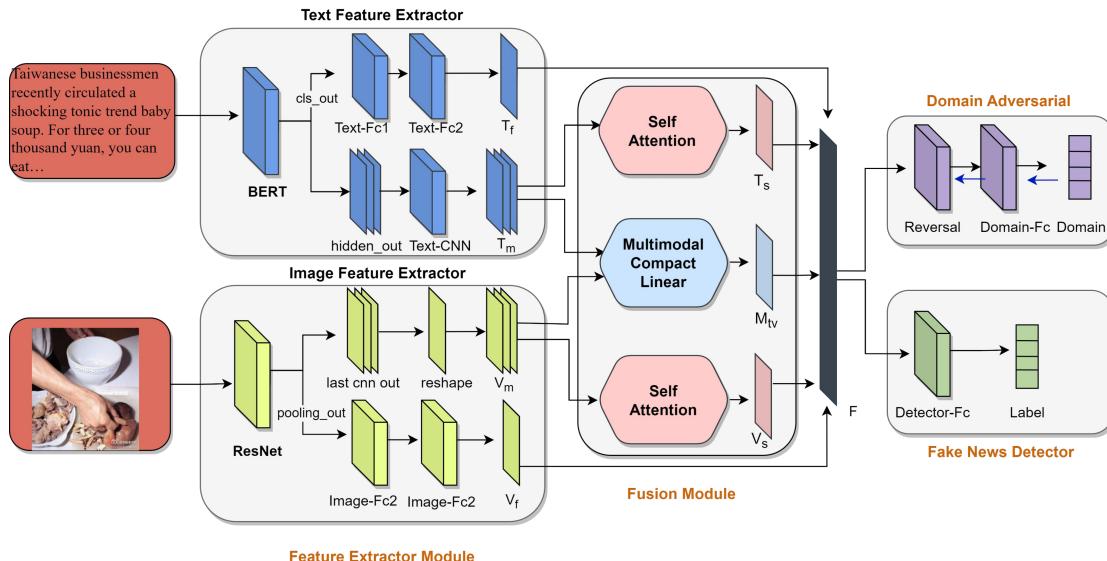


图 4-3 基于组合式融合机制的虚假新闻检测算法模型（MFNDCFM）

### 4.3.1 多模态特征提取模块

在章节 3.3.1 的文本特征提取模块中，模型可以获得具有深层语义的句向量  $T_f$  和词向量特征  $T_m$ ，也就是图 4-3 中的 Text Feature Extractor 蓝色区域所示，同样在章节 3.3.2 中的图像特征提取模块中，可以获得具有不同整体特征向量  $V_f$  和区域特征视觉向量  $V_m$ ，也就是图 4-3 的 Image Feature Extractor 黄色区域所示。不同的是，在第三章中文本特征和图像特征进入虚假检测模块之前，仅通过直接拼接两个提取的特征然后检测新闻。这种拼接的做法损耗了文本和图像两者之间的交互信息。在本章中不使用直接拼接的方式融合特征，而是着重研究模态间和模态内的信息交互融合问题。

### 4.3.2 模态间信息交互模块

为了获得更优质的文本和图像的融合特征，本章拟采用多模双线性池化方法。多模双线性池化方法将文本信息和图像信息中每一纬度的特征进行融合，

通过创建联合表示空间的方法，充分利用特征元素间的交互作用。

模态间信息交互模块位于图 4-3 中 Multimodal Compact Linear 区域，使用多模双线性池化方法对文本特征信息  $T_m$  和图像特征信息  $V_m$  进行交互融合操作，首先使用 Count Sketch 算法分别将文本特征  $T_m$  和图像特征  $V_m$  进行降维操作，然后将特征通过快速傅里叶变换至频域后进行点积运算操作，最后通过逆快速傅里叶变换得到交互后的多模态特征  $M_{tv}$ ，使用全连接层对融合特征调整维度大小以便于与其他特征结合后输入到分类以完成虚假新闻检测工作。计算过程如式(4-8)所示。

$$M_{tv} = W_{tv}(MCBP([T_m, V_m])) \quad (4-8)$$

其中  $M_{tv}$  代表融合后的特征， $MCBP$  代表多模双线性池化方法， $W_{tv}$  代表全连接层参数。

### 4.3.3 模态内信息增强模块

为了得到文本模态内部和图像模态内部的重点信息，本章使用自注意力机制方法融合自身内部信息，能够让模型更好地关注输入序列中的相关信息，从而提高模型的性能。在文本模态中，自注意力机制可以用来捕捉单词之间的关联性，提取文本特征的重要信息。在图像模态中，可以使用自注意力机制来捕捉不同区域之间的关联性，提取图像特征的空间分布。从而提高模型在虚假新闻检测任务中的性能表现。因此本章模态内信息增强模块引入自注意力机制实现，以增强图像模态与文本模态自身的信息。在自信息注意力模块中  $Query$ 、 $Key$ 、 $Value$  均取自于自身模态特征。

模态内信息增强模块位于图 4-3 中的 Self Attention 区域。首先对于文本模态特征进行自增强， $Query$ 、 $Key$ 、 $Value$  均是文本特征。给定词级特征矩阵  $T_m$ ，将每个词特征通过全连接层转换为  $Query$ 、 $Key$ 、 $Value$ ，然后通过计算注意力公式得出自模态信息加强的值。如式(4-9)和(4-10)所示。

$$Q = (W_{qt} T_m), K = (W_{kt} T_m), V = (W_{vt} T_m) \quad (4-9)$$

$$T_s = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4-10)$$

其中， $T_s$  为文本自身加强后的特征， $W_{qt}$ ， $W_{kt}$ ， $W_{vt}$  为文本全连接层权重矩阵， $d_k$  是  $Key$  特征的维度。

然后对于图像区域特征进行自模态信息加强， $Query$ 、 $Key$ 、 $Value$  均是图像自身模态特征，给定词区域级特征矩阵  $V_m$ ，将区域特征通过全连接层转换为

*Query*、*Key*、*Value*，其计算过程如式(4-11)和(4-12)所示。

$$Q = (W_{qv}V_m), K = (W_{kv}V_m), V = (W_{vv}V_m) \quad (4-11)$$

$$V_s = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4-12)$$

其中， $V_s$  为图像自身加强后的特征， $W_{qv}$ ， $W_{kv}$ ， $W_{vv}$  为图像全连接层权重矩阵， $d_k$  是 *Key* 特征的维度。

对特征提取器提取出的  $T_s$ ， $T_f$ ， $V_s$ ， $V_f$ ， $M_{tv}$  进行拼接操作，以便于送入检测器进行检测，计算过程如式(4-13)所示。

$$F = concat([T_f, T_s, V_f, V_s, M_{tv}]) \quad (4-13)$$

#### 4.3.4 领域对抗模块

本章的领域对抗模块结构与 3.3.3 章节一致，唯一的区别是对立的过程发生了一些转变，在第三章中，多模态特征提取器与领域对抗模块相互对抗以产生通用性特征，而本章中多模态特征提取器和特征融合器共同与领域对抗模块产生对抗性以达到模型的泛化性能，其结构如图 4-3 中的 Domain Adversarial 部分所示。

#### 4.3.5 虚假检测模块

本章中的虚假检测模块与章节 3.3.4 一致，其结构如图 4-3 中的 Fake News Detector 部分所示。

### 4.4 实验设计

本章在第三章特征抽取的基础之上，增加了融合模块，其他模块均未发生改变。因此实验设计部分的数据集、评价标准、实验环境、和对比实验模型与 3.4 章节中的实验设置均一致，不再赘述。

### 4.5 实验结果分析

本小节将结合第三章设计的特征提取方法进行特征提取，加入本章提出的以模态间和模态内进行交互增强的组合式融合机制，通过对比实验结果和消融实验分析，验证所提方法的有效性。

### 4.5.1 对比实验结果分析

为了分析 MFNDCFM 模型中组合式融合机制对模型训练的影响，绘制出在 Weibo 数据集上 MFNDCFM 模型和 MFNDTBA 模型训练过程中的损失值 Loss 的变化趋势曲线与准确率 Accuracy 的变化趋势曲线，如图 4-4 所示。

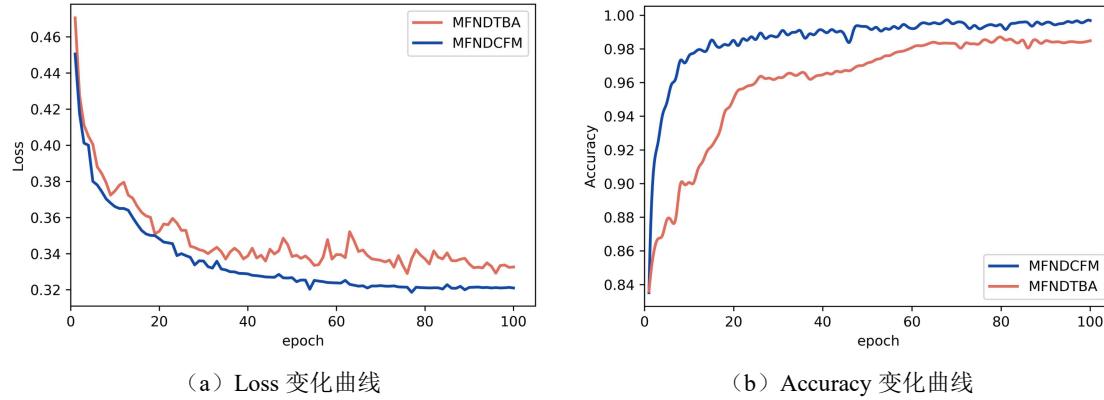


图 4-4 MFNDCFM 模型训练过程变化曲线

将使用组合式融合机制的 MFNDCFM 模型和未使用组合式融合机制的 MFNDTBA 模型的损失函数 Loss 变化趋势记录并整理后，绘制出的损失对比趋势结果如图 4-4 (a) 所示。从稳定性的角度进行分析结果，采用组合式融合策略的虚假新闻检测模型 MFNDCFM 的 Loss 曲线震荡幅度小，且收缩趋势更加稳定，而 MFNDTBA 模型震荡幅度较大，曲线不够稳定。从收缩速度来分析结果，MFNDCFM 模型与 MFNDTBA 模型相比加快了模型趋于稳定的速度，进一步证明了组合式融合机制对于模型训练起到了正向促进作用。

同时对采用组合式融合机制的 MFNDCFM 模型和未采用组合式融合机制的 MFNDTBA 模型的准确率 Accuracy 的变化趋势进行记录，绘制出的准确率对比趋势结果如图 4-4 (b) 所示。从准确率震荡幅度分析曲线，采用组合式融合策略的虚假新闻检测模型 MFNDCFM，准确率趋势更加稳定，震荡幅度小，而 MFNDTBA 准确率震荡幅度也较小，但趋势不够稳定。同样从收缩速度来分析，含有融合机制的 MFNDCFM 模型准确率趋于稳定的速度更快，可以更快地完成训练任务。

将本章 MFNDCFM 模型与 3.4.4 章节的对比模型进行实验分析，实验对比结果如表 4-3 和图 4-5 所示。实验结果表明，本章的 MFNDCFM 模型的准确率均优于其他对比模型。从模态融合的角度分析结果，在 Weibo 数据集上，采用融合机制的 MFNDCFM 模型准确率比未采用融合机制的 MFNDTBA 模型高 0.89%。在 Twitter 数据集上，MFNDCFM 准确率比 MFNDTBA 模型 2.48%。再

次证明 MFNDCFM 分别从模态间和模态内融合特征信息，使得文本和图像之间的关联特征和交互信息更加丰富。而 MFNDTBA 模型虽然通过双分支网络结构和领域对抗模块提升了模态特征的表达性，但没有补充模态间的交互性信息从而造成了部分隐藏特征的损失。

表 4-3 MFNDCFM 与对比模型实验结果

数据集	方法	Accuracy	真新闻			假新闻		
			Precision	Recall	F1	Precision	Recall	F1
Weibo	Text-GRU	0.643	0.662	0.578	0.617	0.662	0.578	0.617
	Image-VGG	0.633	0.630	0.500	0.550	0.630	0.750	0.690
	att-RNN	0.772	0.797	0.713	0.692	0.684	0.840	0.754
	EANN	0.816	0.820	0.820	0.820	0.810	0.810	0.810
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	BDANN	0.842	0.830	0.870	0.850	0.850	0.820	0.830
	Spotfake	0.892	0.902	0.964	0.932	0.847	0.656	0.739
	<b>MFNDTBA</b>	0.896	0.923	0.874	0.898	0.868	0.920	0.893
Twitter	<b>MFNDCFM</b>	0.904	0.943	0.871	0.905	0.868	0.941	0.903
	Text-GRU	0.526	0.586	0.553	0.569	0.469	0.526	0.496
	Image-VGG	0.596	0.695	0.518	0.593	0.550	0.700	0.599
	att-RNN	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN	0.719	0.642	0.474	0.545	0.771	0.870	0.817
	MVAE	0.745	0.801	0.719	0.758	0.686	0.777	0.730
	BDANN	0.830	0.810	0.630	0.710	0.830	0.930	0.880
	Spotfake	0.777	0.751	0.900	0.820	0.832	0.606	0.701
	<b>MFNDTBA</b>	0.847	0.821	0.684	0.746	0.856	0.927	0.890
	<b>MFNDCFM</b>	0.868	0.831	0.754	0.791	0.884	0.925	0.903

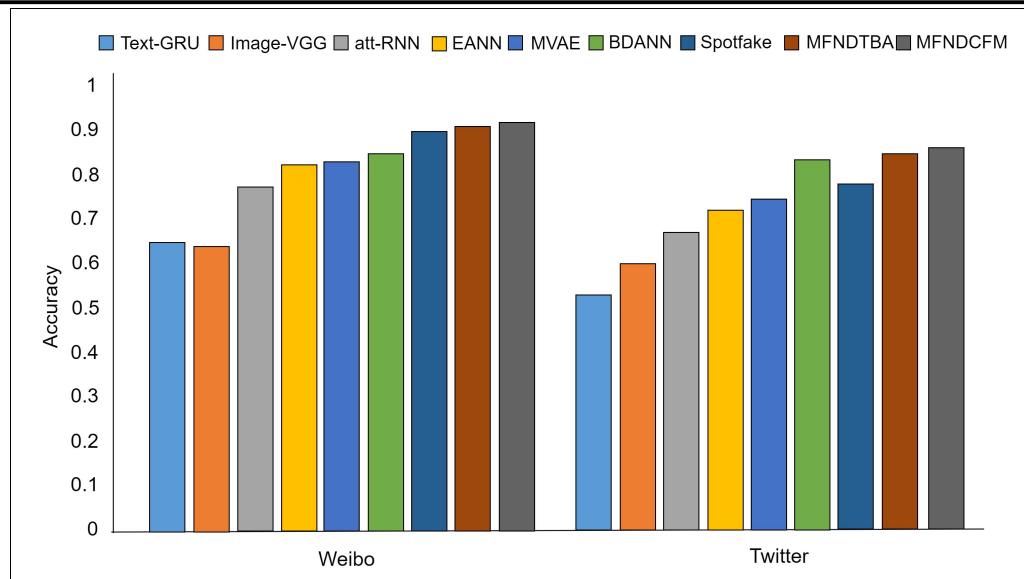


图 4-5 MFNDCFM 准确率对比图

### 4.5.2 消融实验结果分析

为了验证模型 MFNDCFM 组合式融合机制中不同融合模块对实验结果的影响，分别在 Weibo 和 Twitter 两个关于多模态虚假新闻检测的数据集上进行消融实验，设计两种对比模型进行分析。

(1) 去除模态间信息交互模块：对文本内容及图像内容分别提取特征后，将特征送入模态内信息增强模块中，然后送入虚假检测模块和领域对抗模块。将模型定义为 MFNDCFM-1。

(2) 去除模态内信息增强模块：对文本内容及图像内容分别提取特征后，将特征送入模态间信息交互模块中，得到特征后将特征送入虚假检测模块和领域对抗模块。将模型定义为 MFNDCFM-2。

MFNDCFM 的消融实验的结果如图 4-5 所示。当 MFNDCFM 模型去除模态间信息交互模块后，在 Weibo 数据集和 Twitter 数据集中的实验表明，MFNDCFM-1 模型的准确率均降低，证明没有模态间信息交互模块时存在模态之间信息的丢失；其次当模型中存在模态内信息增强模块模型时，MFNDCFM 模型在两个数据集上的准确率提升了 1% 左右，证明了模型内信息增强模块的加入，能够帮助两种模态找到更重要的自身信息。

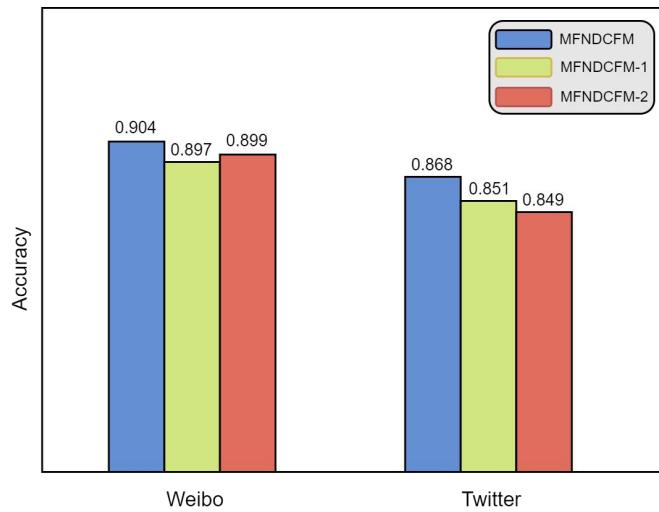


图 4-6 MFNDCFM 消融实验结果

## 4.6 本章小结

本章提出基于组合式融合机制的多模态虚假新闻检测算法，在第三章特征提取的基础之上，在特征融合方面加入以组合式融合机制为基准的模态间信息交互模块和模态内信息增强模块。通过大量实验，验证两种融合模块的有效性。

## 第五章 基于变分自编码器进行多任务学习的多模态虚假新闻检测算法

### 5.1 引言

对于媒体和新闻从业者来说，应该始终坚持新闻真实性和客观性的原则，对新闻进行充分地验证和审查。对于公众来说，应该保持理性的思考和判断能力，不要轻易相信未经证实的消息，经过专业验证后的新闻真实性更加可靠。类比于虚假新闻检测，若能够验证其特征有效性，则可以有更准确的检测结果。

在第三章特征抽取方法和第四章特征融合方法的基础上，新闻不同模态的内容特征信息被充分挖掘。但对于当下多模态虚假新闻检测模型的一个主要缺点是，虽然模型可以从多个模态中提取特征并进行融合，但并不总是能够验证这些多模态特征是否真正有效。对于特征抽取和特征融合来说，对特征的工作主要是在构建层面，若能采用某种方式进行重建所抽取的多模态特征，不仅可以验证所抽取的特征是否有效，同时可以学习出不同模态特征中的跨模态信息。因此重建多模态特征可以确保提取的特征是有意义的。

为了验证出所提取出的多模态特征信息是否有效，重现模态之间的关联性。在第三章和第四章的基础之上，本章提出一种基于变分自编码器进行多任务学习的多模态虚假新闻检测算法。变分自编码器可以通过训练输入数据得到多模态数据之间的隐含共享表示，并从中重构出多模态数据，发现潜在的跨模态信息。同时本章使用多任务学习调节与第三章和第四章之间的不同任务之间的关系。综合前面章节的研究内容，本章对特征的可验证性进行探索。本章提出基于变分自编码器进行多任务学习的多模态虚假新闻检测模型（Multimodal Fake News Detection based on Variational Auto-Encoder for Multi-Task Learning，MFNDVAEML 简记为 VAEMTL）来检测虚假新闻。在第三章和第四章的基础之上增加特征重现模块，并对多任务进行同时学习。本章的研究内容主要包括两方面：

(1) 本章提出一种基于变分自编码器进行多任务学习的多模态虚假新闻检测算法，在特征抽取和特征融合基础上考虑了多模态特征之间的关联。

(2) 本章引入特征重现模块来验证所提取的多模态特征是否有效，从而提高了特征的可验证性。

## 5.2 相关工作

### 5.2.1 变分自编码器

变分自编码器模型 (Variational Auto-Encoder, VAE) [67]是以自编码器 (Auto encoder, AE) [68]为基础的优化模型。AE 是由编码器和解码器组成的自监督学习模型。编码器可以将输入数据进行编码成隐特征向量，解码器可以将编码后的特征向量进行重构。AE 结构如图 5-1 所示。由于 AE 的结构和优化方式，决定它并不能真的“生成数据”，只是将数据尽可能相似地还原，缺乏可解释性。

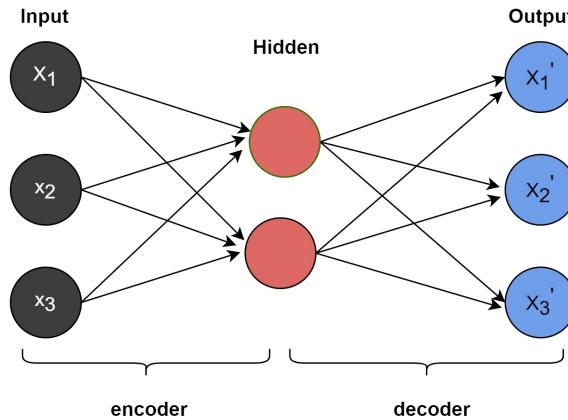


图 5-1 自编码器模型结构 (AE)

针对编码器的“生成”问题，VAE 在编码输入信息时进行了改进，会同时从输入特征信息中抽取隐变量  $z$ ，特征信息隐变量  $z$  会同时包含输入数据和噪声，因此可以生成对应的有效文本信息。其结构如图 5-2 所示，VAE 结合神经网络建立概率密度分布模型得到编码器结构和解码器结构：左半边为推断网络，是对输入数据进行编码得到隐变量概率分布中的均值和方差；右半边为生成网络，可以在推断网络所生成的隐变量的概率分布中采样得到隐随机变量，并通过生成网络重构出推断网络的输入数据。

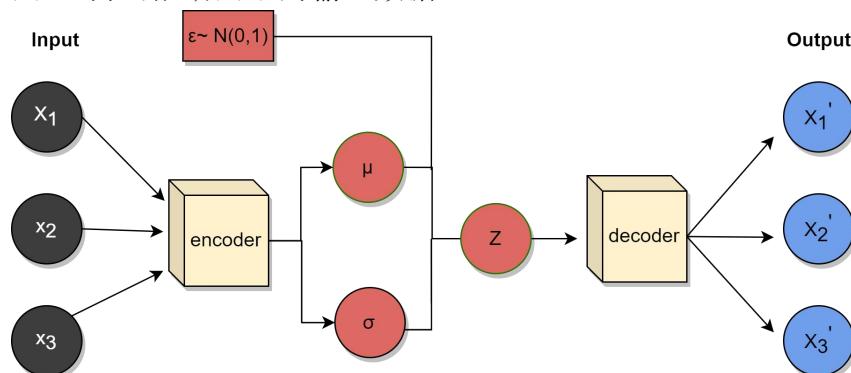


图 5-2 变分自编码器模型模型 (VAE)

基于概率模型对变分自编码器进行分析，假设样本数据集是由随机分布变量  $x_i$  组成，记为  $X = \{x_i\}_{i=1}^N$ ，由变分自编码器生成的数据为  $X' = \{x'_i\}_{i=1}^N$ ，假设生成数据是由不可直接观测的隐变量  $z$  随机采样生成的，其中隐变量  $z$  由先验分布  $p(z)$  生成，然后根据条件分布  $p(x' | z)$  生成重构数据  $x'$ 。变分自编码器的重点在于对输入数据变量  $x$ ，如何去求解隐变量  $z$  的概率。对于未知的分布  $p(x)$  是难以直接计算的，也就是分布  $p(z | x)$  难以直接计算，根据贝叶斯条件概率如式(5-1)如下：

$$p(z | x) = \frac{p(x' | z)p(z)}{p(x)} \quad (5-1)$$

根据当前所遇到的问题，可以使用变分推断方法进行处理。为了求解出  $p(z | x)$ ，可以寻找一个新的分布  $q_\theta(z | x)$ ，让它去逼近难以直接求解的  $p(z | x)$ 。在 VAE 中，使用编码器来构造  $q_\theta(z | x)$ ，其中参数  $\theta$  对应的是隐变量的均值和方差，即  $\theta_{x_i} = (\mu_{x_i}, \sigma_{x_i}^2)$ 。由于每个样本数据的隐变量信息是不同的，因此每个数据序列的后验分布是不一样的，为了得到每个数据序列的  $\theta_{x_i}$ ，需要求解  $q_\theta(z | x)$ 。

VAE 用 KL 散度衡量分布  $q_\theta(z | x)$  与分布  $p(z | x)$  之间的差异，为了让两个分布更加逼近，得到最优的参数，需要两个分布尽可能的相似，KL 散度的定义如式(5-2)所示。

$$KL(q_\theta(z | x) \| p(z | x)) = E_q[\log q_\theta(z | x)] - E_q[\log p(x, z)] + \log p(x) \quad (5-2)$$

需要最小化 KL 散度得到最优的参数  $\theta^*$ ，计算公式如式(5-3)所示。

$$q_{\theta^*}(z | x) = \operatorname{argmin}_\theta KL(q_\theta(z | x) \| p(z | x)) \quad (5-3)$$

由于计算过程中存在  $p(x)$ ，公式计算困难，所以引入变分下界  $ELBO(\theta)$ ，计算公式如式(5-4)所示。

$$ELBO(\theta) = E_q[\log p(x, z)] - E_q[\log q_\theta(z | x)] \quad (5-4)$$

将  $ELBO(\theta)$  与 KL 散度的公式进行结合得到式(5-5)。

$$KL(q_\theta(z | x) \| p(z | x)) = \log p(x) - ELBO(\theta) \quad (5-5)$$

分析上式，由于  $\log p(x)$  为定值，且根据定义  $KL(x \| y) \geq 0$ ，并且  $ELBO$  计算后验分布更加方便，因此将最小化 KL 散度转换为最大化  $ELBO(\theta)$ 。同时在隐变量的反向传播问题中，可以采用重参数技巧方法解决隐变量求导问题。在此基础上对隐变量进行采样均值  $\mu$  和方差  $\sigma$ ，构造出一个符合正态分布的噪声  $\varepsilon$ ，并通过线性变换  $z = \mu + \sigma \odot \varepsilon$  实现隐变量求导。在整个 VAE 训练过程中，

是基于 ELBO 为负损失函数进行梯度下降优化。

### 5.3 基于变分自编码器的模型结构改进方法

第三章内容和第四章内容主要是在特征提取和融合方面来改进虚假新闻检测的效果。本章着重考虑重现特征和验证特征，对模型的结构进行改进，利用变分自编码器的生成功能进行验证特征的有效性，从而促进虚假新闻的验证效果。本章的模型结构在之前模型的基础之上增加了特征重构模块，并使用编码器和解码器将模型结构拆解重组，VAEMLT 模型结构如图 5-3 所示。整个模型包括编码器模块（Encoder）、解码器模块（Decoder）、领域对抗模块（Domain Adversarial）和虚假新闻模块（Fake News Detector）。

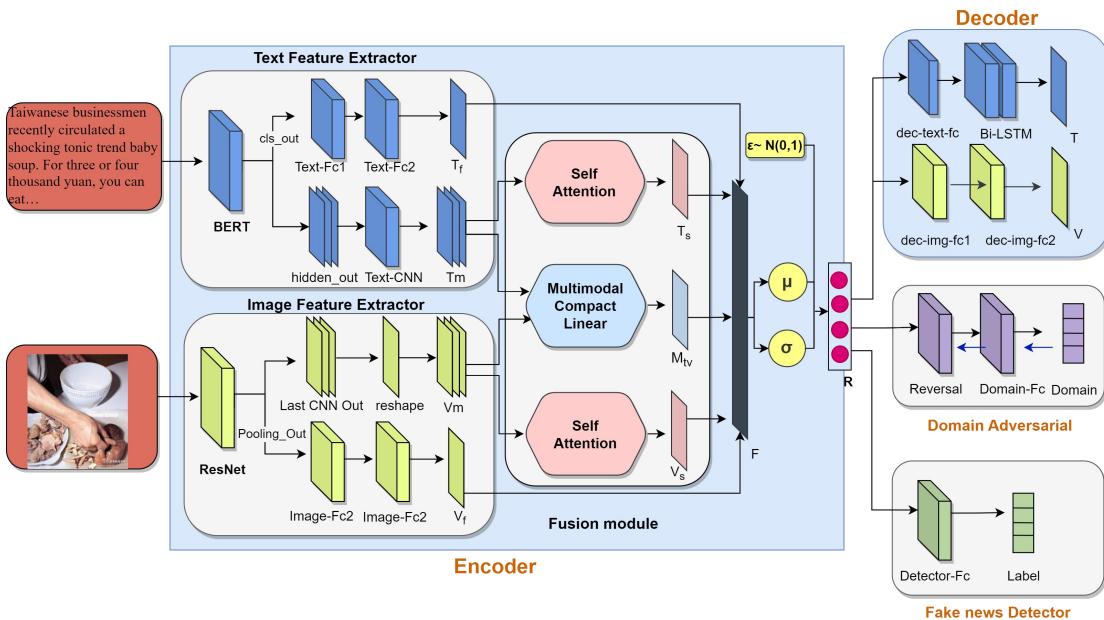


图 5-3 基于变分自编码器进行多任务学习的多模态虚假新闻检测算法（VAEMLT）

#### 5.3.1 编码器模块

编码器的目的是将输入的多模态特征压缩到隐藏特征空间中得到隐藏变量。编码器结构主要包括第三章节中特征抽取模块和第四章节中的特征融合模块以及本章的特征处理部分。其结构如图 5-3 的 Encoder 部分所示。在第三章的特征提取阶段，利用预训练模型采用双分支结构对文本和图像的语义、语序信息进行充分地提取，获取具有深度语义的特征向量。在第四章的特征融合阶段中，使用多模双线性池化融合不同模态的特征信息，使用自注意力机制自增强模态内部特征，从而构建融合后的特征隐向量  $F$ 。本章将编码器定义为

$G_{enc}(\theta_{enc})$ ，其中  $\theta_{enc}$  表示编码器中所有需要学习的参数。因此对于新闻  $x_t$ ，获取其融合特征向量如式(5-6)所示。

$$F_{x_t} = G_{enc}(x_t) \quad (5-6)$$

其中， $F_{x_t}$  代表新闻  $x_t$  的融合后的特征向量， $G_{enc}$  表示编码器， $x_t$  表示一篇含有文本和图像的新闻。

本章的特征处理模块将附有类别特征的隐向量  $F_{x_t}$ ，使用两个全连接层进行处理，进而得到隐向量分布的均值  $\mu$  和标准差  $\sigma$ ，计算过程如式(5-7)和(5-8)所示，为便于计算编码器梯度估计量，从标准高斯分布中采样得到噪声  $\varepsilon \sim N(0,1)$ ，构造一个新闻  $x_t$  随机隐变量  $R_{x_t}$ ，因此最终构造出多模态随机隐变量  $R_{x_t}$ 。计算过程如式(5-9)所示。

$$\mu = W_\mu(F_{x_t}) \quad (5-7)$$

$$\sigma = W_\sigma(F_{x_t}) \quad (5-8)$$

$$R_{x_t} = \mu + \sigma * \varepsilon \quad (5-9)$$

其中， $\mu$  和  $\sigma$  是编码器对输入特征信息处理后得到的文本均值和标准差， $R_{x_t}$  是重参数化后的融合隐变量，且  $R_{x_t} \sim (\mu, \sigma^2)$ ， $W_\mu$ ， $W_\sigma$  为全连接层的参数。

编码器可以通过随机采样和重参数化的方法，从而输出一个隐含随机向量  $R$ ，用于解码的生成。

### 5.3.2 解码器模块

解码器的目的是从多模态特征采样的融合隐变量  $R_{x_t}$  中重构出文本和图像特征。其结构如图 5-3 中的 Decoder 所示。解码器分为两部分：文本解码模块和图像解码模块。

#### (1) 文本解码模块

文本解码模块使用的全连接层和双向的 LSTM 结构。其中双向 LSTM 用于每个时间步生成下一时间步的信息，解码器的初始隐层状态  $h_0^{x_t}$  和单元状态  $c_0^{x_t}$  可以由隐向量  $R_{x_t}$  通过双曲正切函数处理得到，文本解码模块的计算过程如式(5-10)和(5-11)所示。

$$h_0^{x_t} = \tanh(W_h(R_{x_t})) \quad (5-10)$$

$$c_0^{x_t} = \tanh(W_c(R_{x_t})) \quad (5-11)$$

具体地说，将多模态隐变量  $R_{x_t}$  传递到全连接层中。然后使用 Bi-LSTM 进行特征的解码。然后将 Bi-LSTM 的输出信息送入带有  $softmax$  函数的全连接层中，最后新闻  $x_t$  得到该时间步重构的文本特征矩阵  $\hat{T}_{x_t}$ ，文本解码器模块的过程如公式(5-12)所示。

$$\hat{T}_{x_t} = softmax(Bi\text{-}LSTM(W_r(R_{x_t}))) \quad (5-12)$$

### (2) 视觉解码模块

视觉解码器的目标是从多模态特征信息表示  $R_{x_t}$  中重构出 ResNet-50 的特征。在本文中为了防止模型过大导致训练以及推理速度过于缓慢，选用全连接层作为图像解码器。将多模态特征表示  $R_{x_t}$  传递给两个全连接层，并加入  $softmax$  函数，来得到新闻  $x_t$  重构的图像特征  $\hat{V}_{x_t}$ ，视觉解码器模块的计算过程如式(5-13)所示。

$$\hat{V}_{x_t} = softmax(W_{vr2}(W_{vr1}(R_t))) \quad (5-13)$$

本章将解码器定义为  $G_{dec}(\theta_{dec})$ ，其中  $\theta_{dec}$  表示解码器中的所有参数。因此对于一条新闻  $x_t$  的多模态数据经过编码器后得到的多模态随机隐变量  $R_{x_t}$ ，解码器的输出由两部分组成：1)重构的文本特征概率矩阵：代表每个单词信息在文本每个位置的概率。2)重构的 ResNet-50 图像特征。解码器整体计算公式如式(5-14)所示。

$$(\hat{T}_{x_t}, \hat{V}_{x_t}) = G_{dec}(R_{x_t}) \quad (5-14)$$

### 5.3.3 领域对抗模块

本章的领域对抗模块结构与 4.3.4 章一致，区别是第三章中使用融合的多模态特征  $F$  作为特征向量输入领域对抗模块中，本章转变为使用变分自编码器采样多模态变量  $F$ ，生成多模态随机隐变量  $R$ ，输入领域对抗模块中，其结构如图 5-3 中的 Domain Adversarial 部分所示。

### 5.3.4 虚假检测模块

本章中的虚假检测模块与章节 3.3.4 一致，其结构如图 5-3 中的 Fake News Detector 部分所示。

### 5.3.5 多任务学习损失函数

对于领域对抗模块、虚假检测模块以及本章加入的重构模块共三个任务需

要一起联合训练学习。因此需要首先定义损失函数，如式(5-15)所示。

$$L_{all} = L_{fnd} + L_{VAE} - L_{dom} \quad (5-15)$$

其中  $L_{fnd}$  为虚假新闻检测模块的损失函数， $L_{dom}$  为领域对抗模块的损失函数， $L_{VAE}$  为变分自编码器模块的损失函数。

在 3.3.5 章中已经对虚假新闻检测模块和领域对抗模块的损失函数进行探讨，本章将着重对 VAE 的损失函数进行介绍。

在 VAE 的训练过程中，使用两个损失来约束模型的学习。第一个是重构损失，它衡量了解码器的重构误差，即重构的文本和图像与原始文本和图像之间的差异。第二个是 KL 散度，KL 散度可以用来衡量编码器学到的潜在变量分布  $p(z/x)$  与标准高斯分布  $q(z/x)$  之间的差异。因此对于 VAE 的训练过程来说，通过最小化重构损失和 KL 散度，VAE 可以逐步学习到合理的隐向量表示，从而提高多模态数据的特征效果。变分自编码器模型的损失项主要包括两个部分：重构损失和 KL 散度损失，如式(5-16)所示。

$$L_{VAE} = L_{rec} + D_{KL}(q(z/x) \| p(z/x)) \quad (5-16)$$

其中， $L_{rec}$  是生成器的重构损失， $D_{KL}$  用来衡量  $q(z/x)$  与  $p(z/x)$  的相似程度。

对于 KL 散度而言，编码器将输入文本和图像映射到潜在空间中的隐变量，这些隐变量的分布通常不是标准高斯分布，而是与输入数据相关的复杂分布。VAE 为了获得更合适的隐向量特征，对编码器增加了一个约束，即最小化编码器学到的隐变量分布与标准高斯分布之间的差异，也就是最小化编码器学到的分布与真实潜在变量分布之间的 KL 散度，用 KL 散度来计算损失函数，KL 散度的损失越小，分布越相似，用 KL 散度计算编码器损失函数如式(5-17)所示。

$$D_{KL}(q(z/x) \| p(z/x)) = \frac{1}{2} \sum_{i=1}^{n_r} (\mu_i^2 + \sigma_i^2 - \log(\sigma_i) - 1) \quad (5-17)$$

其中  $n_r$  代表多模态特征  $R_i$  的维度， $\mu$  和  $\sigma$  是编码器对输入特征处理后输出的文本均值和标准差。

对于重构损失而言，在训练过程中，编码器将输入文本和图像映射到潜在空间中的均值和方差。随后，从均值和方差中采样一个随机向量，作为潜在变量的表征。然后将这个向量输入到解码器中，解码器将这个随机向量重构为输入的文本和图像。因此重构损失同样包括文本重构和图像重构，如式(5-18)所示。对于文本的重构，使用交叉熵损失函数进行衡量，如式(5-19)所示，对于图像的

重构使用均方误差进行衡量如式(5-20)所示。

$$L_{rec} = L_{recon\_text} + L_{recon\_img} \quad (5-18)$$

$$L_{rec\_text} = -E_{x \sim X} \left[ \sum_{i=1}^{n_t} \sum_{c=1}^C \mathbb{1}_{c=\hat{T}_x^{(i)}} \log(\hat{T}_x^{(i)}) \right] \quad (5-19)$$

$$L_{rec\_img} = -E_{x \sim X} \left[ \frac{1}{dim_{res}} \sum_{i=1}^{dim_{res}} (\hat{V}_{res\_x}^{(i)} - V_{res\_x}^{(i)})^2 \right] \quad (5-20)$$

其中， $L_{recon\_text}$  代表文本重构损失， $L_{recon\_img}$  代表图像重构损失。 $X$  表示一组多模态新闻， $x$  代表一条新闻； $n_t$  是文本中单词的数量； $c$  是词表大小； $dim_{res}$  代表 ResNet-50 特征的维度。

则多任务学习最终的损失如式(5-21)所示。

$$\begin{aligned} L_{final} &= L_{fnd} + L_{VAE} - L_{dom} \\ &= L_{fnd} + L_{rec\_text} + L_{rec\_img} + D_{KL} - L_{dom} \\ &= \lambda_f L_{fnd} + \lambda_t L_{rec\_text} + \lambda_i L_{rec\_img} + \lambda_k D_{KL} - \lambda_d L_{dom} \end{aligned} \quad (5-21)$$

其中， $\lambda$  可以用来平衡损失函数的各个项。在本章中，由于任务较多，将会对 $\lambda$  进行不同的权重分配方案，以达到更好的多任务学习效果。通过最小化损失函数得到最优参数，如公式(5-22)所示。

$$\theta_{model}^* = argmin(L_{final}) \quad (5-22)$$

## 5.4 实验设计

本章在第三章和第四章的基础之上，使用变分自编码器增加了特征重构模块。实验设计部分的数据集、评价标准和对比实验模型与 3.4 章节中设置一致。

其中，实验环境发生了一些更改。在实验环境中，为了快速准确的找到更优参数，超参数有了一些更改。经过实验证明，Batch Size 的增大有助于模型训练，因此 Batch Size 在合理的范围内增加至 64。

为了让模型训练时更加稳定，本章选择加入 Warm up 学习率预热的方法，开始时先选择一个较小的学习率，训练过程中等模型稳定后再将学习率调整为预置的学习率，学习率预热的使用会使得模型收敛速度更快，且模型效果更稳定。为了防止模型过拟合，本章使用早停法（Early Stop）限制模型趋于过拟合，若 20 个连续 epoch 验证集上面的准确率均未发生增长，则模型停止训练。超参

数设置如表 5-1 所示。

表 5-1 超参数设置

模型参数	具体配置
Batch Size	64
Epoch	100
优化器	Adam
学习率	Warm Up
丢弃率	0.2
Early Stop	20

本章为了更好的联合训练变分自编码器模块、领域对抗模块和虚假检测模块，将多任务的作用都发挥出来，本章增加了损失权重分配对比实验，目的是探究不同损失对于联合训练的影响，通过实验比较不同损失权重分配的效果，找到最优的权重分配方案，提高模型的检测性能，本章设置以下几组实验，如表 5-2 所示。

表 5-2 多任务权重分配方案

分配方案	具体操作	模型名称
等权重平均分配	$\lambda_f = \lambda_t = \lambda_i = \lambda_k = \lambda_d = 1$	VAEMTL_AV
重要度分配	$\lambda_f = 0.5, \lambda_d = 0.2, \lambda_t = \lambda_i = \lambda_k = 0.1$	VAEMTL_IM
基于实验结果动态分配	动态实时调整	VAEMTL_DY

(1) 等权重平均分配：将变分自编码器模块、领域对抗模块和虚假检测模块三个任务的损失权重设置为相等的值，即认定每个任务的损失函数对模型的影响相同。将模型定义为 VAEMTL\_AV。

(2) 重要度分配：将变分自编码器模块、领域对抗模块和虚假检测模块三个任务按照重要度进行排序，认为  $\lambda_f > \lambda_d > \lambda_t = \lambda_i = \lambda_k$ ，因此按照人工经验赋值，将模型定义为 VAEMTL\_IM。

(3) 基于实验结果动态分配：动态调整的目的是希望不同的任务以相似的速度学习，动态调整各任务的损失权重，根据各任务的损失函数值实时调整权重。因此使用 Grad Norm<sup>[69]</sup>权重调整方法，这是一种常用的动态调整多任务学习中任务权重的方法，其核心思想是基于不同任务梯度的大小来调整任务权重，梯度较大的任务权重相对较小，梯度较小的任务权重相对较大。这种方法的优点在于可以根据任务的梯度大小动态调整任务权重。将模型定义为 VAEMTL\_DY。

## 5.5 实验结果分析

本小节首先通过权重分配实验确定权重的分配最优方案，然后使用权重最优方案，与对比模型进行分析，从而达到更好的对比效果。

### 5.5.1 权重对比实验结果

为了找到最优的权重分配方案，记录三种权重训练过程中的 Loss 和 Accuracy 趋势变化曲线图，图 5-4 (a) 为 Loss 趋势变化曲线图，图 5-4 (b) 为 Accuracy 趋势变化曲线图，同时记录在三种权重分配下测试集的表现，如表 5-3 所示。

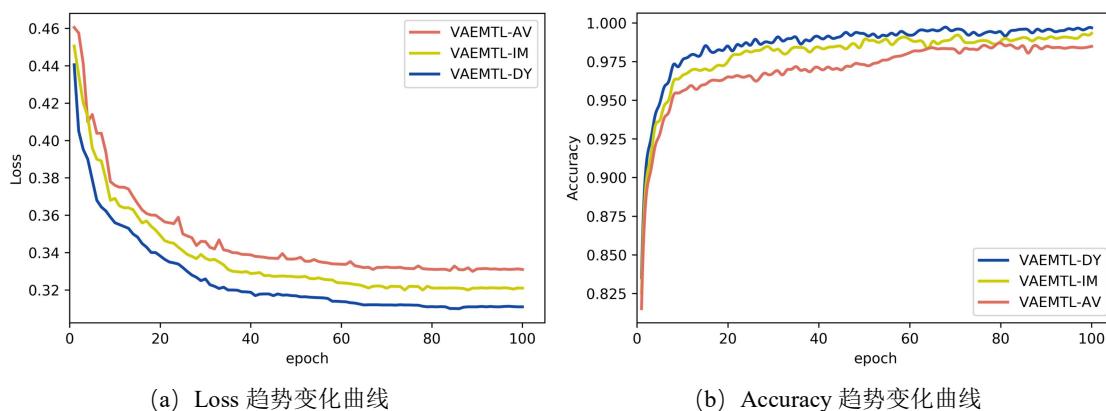


图 5-4 VAEMTL 模型权重对比实验结果

从 Loss 趋势曲线图中可以分析出，基于实验结果动态分配权重的模型表现优于平均分配和重要度分配方式，模型收敛趋势更加稳定，且损失较低；相对于平均分配和重要度分配，震荡幅度较小。这是由于平均分配和重要度分配使得损失比重不统一，造成了模型的前期波动性大，趋势不稳定。

记录平均分配的权重与动态分配权重所有 loss 值的变化曲线，如图 5-5 所示。可以发现在域对抗模块任务的 Loss 以及重构文本的 Loss 值偏大，而其它任务的 loss 规模偏小，如图 5-5 (a) 所示，此时多任务近似退化为单任务目标学习，多任务的权重几乎完全按照偏大的 Loss 任务来进行更新，逐渐丧失了多任务学习的优势。而基于结果动态调整的方式从损失的规模层面进行分配权重，使得多个任务能够从真实意义上联合训练，如图 5-5 (b) 所示，从而使得模型找到全局的最优权重。因此为了防止出现一种或多种任务对网络权重起主导支配作用的情形，必须谨慎平衡所有任务的联合训练过程。

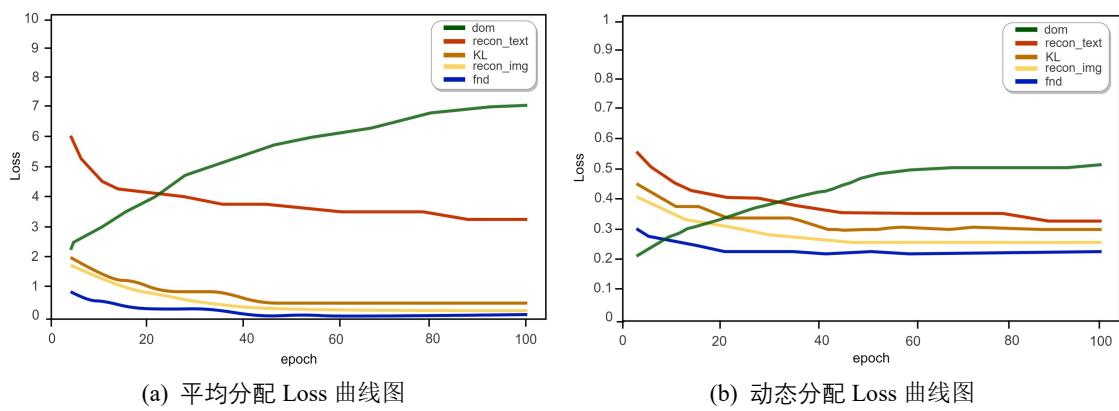


图 5-5 平均权重分配与动态权重分配对比实验结果曲线图

从 Accuracy 的变化曲线图中再次验证动态分配权重的优点，基于动态分配权重方案的模型和平均分配、重要度分配方案相比，Accuracy 曲线更加平稳，波动幅度更小，而且在收敛后的稳定阶段准确率更高。而平均分配和重要度分配方案由于损失比重和损失规模的不统一，导致了其准确率增长但不够稳定。

因此，模型损失和准确率两种趋势所体现得相辅相成，验证了动态权重分配方案的合理性和有效性，模型能更加有效地学习到重要的特征。

根据表 5-3 的权重分配的测试集结果中, 基于动态分配的 VAEML\_DY 模型在两个数据集的结果不论是准确率还是 F1 值, 均优于其他两种分配方案, 证明了模型在推理时权重的有效性, 这表明对于该模型, 动态分配的权重策略能够更好地适应不同数据集的特征, 从而提高模型的性能表现。

表 5-3 权重对比实验结果

数据集	方法	Accuracy	真新闻			假新闻		
			Precision	Recall	F1	Precision	Recall	F1
Weibo	VAEMTL_AV	0.905	0.892	0.921	0.906	0.918	0.891	0.904
	VAEMTL_IM	0.910	0.902	0.927	0.914	0.920	0.893	0.906
	VAEMTL_DY	0.921	0.910	0.940	0.924	0.934	0.901	0.917
Twitter	VAEMTL_AV	0.869	0.820	0.784	0.802	0.880	0.917	0.898
	VAEMTL_IM	0.871	0.826	0.772	0.798	0.891	0.920	0.905
	VAEMTL_DY	0.888	0.838	0.821	0.829	0.912	0.922	0.917

因此综合训练过程变化趋势和测试集结果的表现，选用动态分配方案作为权重分配的最优方案。

### 5.5.2 对比实验结果分析

在 5.4.1 章节中经过对比实验和结果分析确定了使用权重最优方案，本章将

与对比模型进行大量实验分析。表 5-4 和图 5-6 为 VAEMTL\_DY 模型与对比模型的试验结果。对比不同模型和实验结果，可以得出加入了变分自编码器验证模块并经过多任务权重分配调节后的 VAEMTL 模型在 Weibo 数据集和 Twitter 数据集上相对于其他模型表现都有显著提升。

表 5-4 VAEMTL\_DY 与对比模型实验结果

数据集	方法	Accuracy	真新闻			假新闻		
			Precision	Recall	F1	Precision	Recall	F1
Weibo	Text-GRU	0.643	0.662	0.578	0.617	0.662	0.578	0.617
	Image-VGG	0.633	0.630	0.500	0.550	0.630	0.750	0.690
	att-RNN	0.772	0.797	0.713	0.692	0.684	0.840	0.754
	EANN	0.816	0.820	0.820	0.820	0.810	0.810	0.810
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	BDANN	0.842	0.830	0.870	0.850	0.850	0.820	0.830
	Spotfake	0.892	0.902	0.964	0.932	0.847	0.656	0.739
	<b>MFNDTBA</b>	0.896	0.923	0.874	0.898	0.868	0.920	0.893
	<b>MFNDCFM</b>	0.904	0.943	0.871	0.905	0.868	0.941	0.903
	<b>VAEMTL_DY</b>	0.921	0.910	0.940	0.924	0.934	0.901	0.917
Twitter	Text-GRU	0.526	0.586	0.553	0.569	0.469	0.526	0.496
	Image-VGG	0.596	0.695	0.518	0.593	0.550	0.700	0.599
	att-RNN	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN	0.719	0.642	0.474	0.545	0.771	0.870	0.817
	MVAE	0.745	0.801	0.719	0.758	0.686	0.777	0.730
	BDANN	0.830	0.810	0.630	0.710	0.830	0.930	0.880
	Spotfake	0.777	0.751	0.900	0.820	0.832	0.606	0.701
	<b>MFNDTBA</b>	0.847	0.821	0.684	0.746	0.856	0.927	0.890
	<b>MFNDCFM</b>	0.868	0.831	0.754	0.791	0.884	0.925	0.903
	<b>VAEMTL_DY</b>	0.888	0.838	0.821	0.829	0.912	0.922	0.917

在 Weibo 数据集上，VAEMTL 模型相对于 MFNDCFM 模型提升了 1.9%，相对于 MFNDTBA 模型提升了 2.8%，可以分析出经过重构的验证特征更加挖掘了模态之间的关联信息，从而激励模型找到更优的特征信息。在 Twitter 数据集上，VAEMTL\_DY 模型相对于 MFNDCFM 模型提升了 2.3%，相对于 MFNDTBA 模型提升了 4.8%，不仅辅助验证了变分自编码结构对于虚假新闻检测的有效性，更加证明了经过多任务学习后的模型更具有泛化性和迁移性，使得模型能够处理更多的场景和业务。综合表明，VAEMTL\_DY 模型结合了变分自编码器、多任务学习并进行权重分配后的解决方案，对于虚假新闻检测问题具有很好的效果。

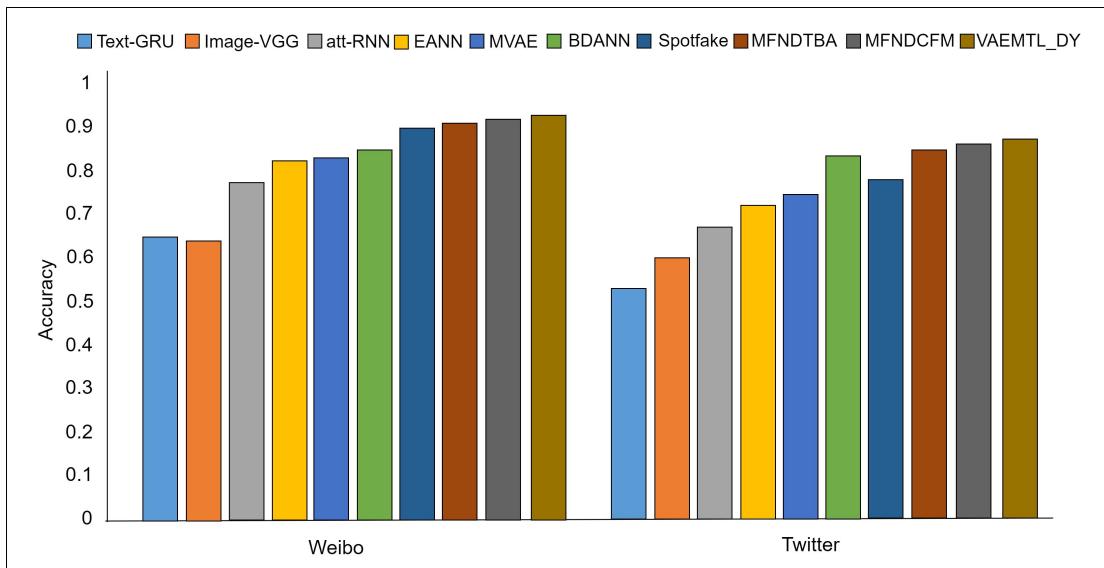


图 5-6 VAEMTL\_DY 与对比模型实验结果

由此可以分析出,加入变分自编码器验证模块的 VAEMTL\_DY 模型可以更好地挖掘模态之间的关联信息,从而提高特征的表达能力,进而提高模型的检测能力。而多任务学习和权重分配策略则可以帮助模型更好地利用不同任务之间的相关性和不同任务之间的差异性,从而提高模型的泛化能力和迁移能力。

## 5.6 本章小结

本章在第三章和第四章的基础上,引入变分自编码器用于验证特征有效性,通过训练多模态变分自编码器,可以从共享表示中重构文本和图像两种模态,发现潜在的跨模态信息,同时引入变分自编码器能够对损失函数有一些任务上的调整,通过多任务的调整学习和权重的分配,使得模型能够兼顾多个任务,从而能够处理多模态虚假新闻检测任务。

## 第六章 虚假新闻检测系统设计实现

本章结合前面章节对基于深度学习的虚假新闻相关技术的研究，实现了一个完整的虚假新闻检测系统。

### 6.1 系统功能分析与设计

#### 6.1.1 系统功能需求分析

基于深度学习的虚假新闻检测系统有以下功能：在使用检测系统时，需要使用新闻文本内容和图像内容的输入功能、文本内容和图像内容的提交功能以及新闻检测结果的输出功能；在系统处理新闻时，首先要对文本和图像内容进行预处理操作，然后实现对新闻的虚假检测功能。

#### 6.1.2 系统概要设计

对基于深度学习的虚假新闻检测系统的功能需求进行梳理，系统的整体由两部分构成：前端模块和后端模块。系统流程如图 6-1 所示。在前端模块中首先完成对新闻中文本内容和图像内容的输入功能、新闻内容提交功能和最终检测结果的展示。后端模块中按照流程将对文本内容和图像内容的预处理以及虚假新闻的检测。

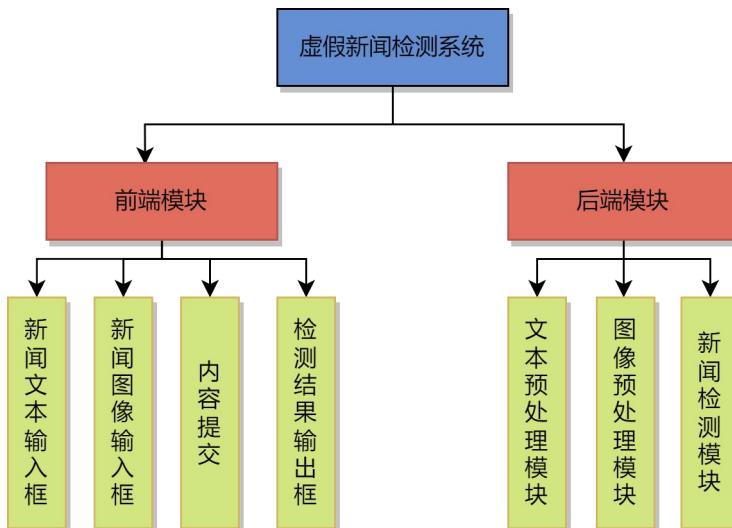


图 6-1 数据处理流程图

### 6.1.3 系统总体设计

系统的总体设计如图 6-2 所示。新闻检测系统的总体分为展示层、服务层和数据层，每层中都具有一个用于信息传输的通信模块。其中，展示层的作用是实现数据的输入和检测结果的输出，服务层承担了系统中的虚假新闻检测功能，数据层主要负责文本预处理和图像预处理功能。三个功能层相互协作，组成了完整的虚假新闻检测系统。

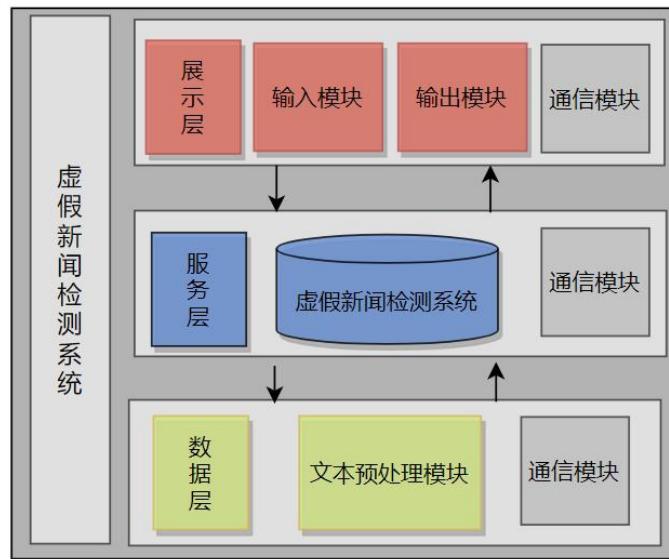


图 6-2 系统总体设计图

## 6.2 系统实现及测试

### 6.2.1 系统功能实现

该系统的前端功能利用 HTML 语言进行开发，后端功能使用 Python 语言进行开发，前端与后端的信息交互功能选用 Flask 框架搭建。Flask 框架的工作流程如图 6-3 所示。当前后端需要信息交互时，Client 会主动产生一个请求，这个请求可以被 WSGI 服务器接收到，当服务器接收到这个请求之后，会唤醒应用中对应的路由做出响应。

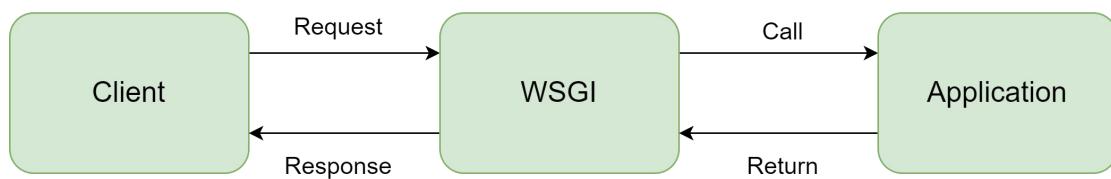


图 6-3 Flask 框架图

## 6.2.2 系统功能测试

虚假新闻检测系统功能包括以下两个部分：文本内容和图像内容的预处理功能、虚假新闻检测功能。未输入内容时的界面如图 6-4 所示，输入内容后的界面如图 6-5 所示。

The screenshot shows the 'Fake news Detection' application. At the top, there is a blue button labeled '数据提交' (Data Submission). Below it, the title 'Fake news Detection' is displayed. A section titled '数据提交' (Data Submission) contains two input fields: one for text ('请输入文本') and one for images ('请输入图像'). The text input field has placeholder text '请输入文本'. The image input field has placeholder text '选择文件 未选择任何文件'. Below these inputs are two buttons: '新闻内容处理' (Process News Content) and '开始检测新闻' (Start News Detection).

图 6-4 系统未输入内容页面

This screenshot shows the same 'Fake news Detection' application after inputting data. The '新闻内容' (News Content) input field now contains the text '照片里面是一只可爱的狗狗，会汪汪叫啊'. The '请输入图像' (Please Input Image) input field shows a selected file named '猫.jpeg'. Below the inputs are the same two buttons: '新闻内容处理' (Process News Content) and '开始检测新闻' (Start News Detection).

图 6-5 系统输入内容页面

将新闻的文本内容和图像内容输入到系统后，首先会使用预处理功能将内容处理成向量格式的数据，以便于送入虚假新闻检测系统中进行检测。图 6-6 为文本内容和图像内容被预处理之后产生的向量数据。



图 6-6 数据预处理界面

新闻检测功能的页面如图 6-7 所示，当用户输入文本内容和图像内容后，点击新闻检测按钮，经后端加载模型处理后，得到检测结果，最终将结果传递到前端进行展示。



图 6-7 新闻检测结果

本小节对实现的虚假新闻检测系统进行了功能演示，展示了虚假新闻检测系统的界面以及检测的流程。最终得出结论，本章设计的检测系统实现了虚假新闻检测的功能。

### 6.3 本章小结

本章在第三章、第四章、第五章的铺垫下，应用提出的算法设计了一个简单的虚假新闻检测系统，按照逻辑步骤和处理流程，最终完成了系统的搭建。

## 第七章 结论与展望

### 7.1 主要结论

在当今虚假新闻泛滥的时代，虚假新闻检测变得越发重要。然而，现有的虚假新闻检测方法无法解决当前问题。目前，新闻多以图文并茂的多模态形式呈现，现有的多模态虚假新闻检测方法对于内容本身特征的抽取并未得到充分挖掘，同时对于多个模态的交互也存在较大的隔阂。此外，目前所能抽到的特征也难以验证其有效性。

针对这些问题，本文阐述了虚假新闻检测的研究背景，并分析出虚假新闻检测的研究要点，最终提出了一种基于深度学习的虚假新闻检测方法。本文分别在三个层次上对现有的虚假新闻检测模型做出了改进。

(1) 特征提取改进：为了提取出文本和图像的高阶语义特征，首先改进了多模态特征提取方法。由于文本和图像自身就具有详细而丰富的内容，本文从原本的模态内容中进行深入挖掘，使用预训练模型 BERT 和 ResNet 对文本和图像内容进行特征提取，并采用双分支网络从浅层和深层同时挖掘不同模态的特征。此外，为了抽取多领域的共性特征，使用领域对抗网络进行训练模型，从而使得抽取出来的特征更具有普适性。

(2) 特征融合改进：为了得到文本和图像之间的交互信息，本文提出基于组合式融合机制的多模态虚假新闻检测算法。组合式融合机制包含两种融合方法，分别对应于模态间信息和模态内信息。使用多模双线性池化方法对文本模态和图像模态之间的信息进行交互。使用自注意力机制实现模态内部信息的增强。从而使得抽取的多模态特征具有两种模态共有且独特的信息。

(3) 模型结构改进：为了验证所提取特征的有效性，使用变分自编码器模型进行多任务学习，将所提出的虚假新闻检测模型进行重组，增加解码器结构，重现多模态特征，从而发现跨模态内容之间的关联，训练多任务损失函数以得到最优权重，使本文中多个任务达到最优表现。

经过大量的实验与分析，证明了本文提出的方法在 Weibo 和 Twitter 数据集上的表现优于其他对比模型。综上所述，本文方法在虚假新闻检测任务上取得了优异的效果。

## 7.2 研究展望

在多模态虚假新闻检测领域，尽管当前的研究工作取得了一些成果，但由于社交媒体的传播性以及虚假新闻内容的复杂性，虚假新闻检测仍然面临着挑战。因此，未来可以往如下方向进行深入研究：

(1) 社交网络传播知识：由于大多数研究工作仅基于虚假新闻的内容，从而忽略了传播虚假新闻的社交网络主体。若能够需要采用一些方式对社交网络特征进行建模、分析其传播特性，可以促进虚假新闻的检测效果。

(2) 数据集扩增：目前虚假新闻检测研究的公开权威数据集较少，这导致模型的应用领域和范围有较大限制，若能够收集更多类型的新闻训练数据，可以增强模型在真实检测任务中的实用性。

## 参考文献

- [1] 《新媒体蓝皮书:中国新媒体发展报告(2022)》发布[J].新闻世界,2022(09):35.
- [2] 罗坤瑾.狂欢与规训:社交媒体时代虚假新闻传播及治理研究[J].现代传播(中国传媒大学学报),2019,41(02):68-72.
- [3] 骆正林.社交媒体时代虚假新闻的社会危害与治理路径[J].未来传播,2022,29(01):37-47+128.DOI:10.13628/j.cnki.zjcmxb.2022.01.002.
- [4] Shu K, Sliva A, Wang S, et al. Fake news detection on social media: A data mining perspective[J]. ACM SIGKDD explorations newsletter, 2017, 19(1): 22-36.
- [5] Bonnet J L, Rosenbaum J E. “Fake news,” misinformation, and political bias: Teaching news literacy in the 21st century[J]. Communication teacher, 2020, 34(2): 103-108.
- [6] Jost P J, Pündter J, Schulze-Lohoff I. Fake news-Does perception matter more than the truth?[J]. Journal of Behavioral and Experimental Economics, 2020, 85: 101513.
- [7] Boididou C, Andreadou K, Papadopoulos S, et al. Verifying multimedia use at mediaeval 2015[J]. MediaEval, 2015, 3(3): 7.
- [8] Zhang C, Yang Z, He X, et al. Multimodal intelligence: Representation learning, information fusion, and applications[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(3): 478-493.
- [9] 谢黎.斯坦福大学发布《2022年人工智能指数报告》[J].世界科技研究与发展,2022,44(03):298.
- [10] 张荣,李伟平,莫同.深度学习研究综述[J].信息与控制,2018(4):385-397,410.
- [11] McCulloch W S, Pitts W. A logical calculus of the ideas immanent in nervous activity[J]. The bulletin of mathematical biophysics, 1943, 5: 115-133.
- [12] Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain[J]. Psychological review, 1958, 65(6): 386.
- [13] Minsky M, Papert S. An introduction to computational geometry[J].

- Cambridge tiass., HIT, 1969, 479: 480.
- [14] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. nature, 1986, 323(6088): 533-536.
- [15] LeCun Y, Boser B, Denker J, et al. Handwritten digit recognition with a back-propagation network[J]. Advances in neural information processing systems, 1989, 2.
- [16] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. science, 2006, 313(5786): 504-507.
- [17] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks[C]//Proceedings of the fourteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2011: 315-323.
- [18] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.
- [19] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.
- [20] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
- [21] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.
- [22] 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.
- [23] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
- [24] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]. 2018.
- [25] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
- [26] Castillo C, Mendoza M, Poblete B. Information credibility on

- twitter[C]//Proceedings of the 20th international conference on World wide web. 2011: 675-684.
- [27] Qazvinian V, Rosengren E, Radev D, et al. Rumor has it: Identifying misinformation in microblogs[C]//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 2011: 1589-1599.
- [28] Ma J, Gao W, Mitra P, et al. Detecting rumors from microblogs with recurrent neural networks[J]. 2016.
- [29] Ma J, Gao W, Wong K F. Detect rumors on twitter by promoting information campaigns with generative adversarial learning[C]//The world wide web conference. 2019: 3049-3055.
- [30] Jin Z, Cao J, Zhang Y, et al. Novel visual and statistical image features for microblogs news verification[J]. IEEE transactions on multimedia, 2016, 19(3): 598-608.
- [31] Qi P, Cao J, Yang T, et al. Exploiting multi-domain visual information for fake news detection[C]//2019 IEEE International Conference on Data Mining (ICDM). IEEE, 2019: 518-527.
- [32] Gupta M, Zhao P, Han J. Evaluating event credibility on twitter[C]//Proceedings of the 2012 SIAM international conference on data mining. Society for Industrial and Applied Mathematics, 2012: 153-164.
- [33] Zhao L, Hu Q, Wang W. Heterogeneous feature selection with multi-modal deep neural networks and sparse group lasso[J]. IEEE Transactions on Multimedia, 2015, 17(11): 1936-1948.
- [34] 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.
- [35] Wang Y, Ma F, Jin Z, et al. Eann: Event adversarial neural networks for multi-modal fake news detection[C]//Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining. 2018: 849-857.
- [36] 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.
- [37] Khattar D, Goud J S, Gupta M, et al. Mvae: Multimodal variational autoencoder for fake news detection[C]//The world wide web conference. 2019:

- 2915-2921.
- [38] Singhal S, Shah R R, Chakraborty T, et al. Spotfake: A multi-modal framework for fake news detection[C]//2019 IEEE fifth international conference on multimedia big data (BigMM). IEEE, 2019: 39-47.
- [39] Mustafaraj E, Metaxas P T. The fake news spreading plague: was it preventable?[C]//Proceedings of the 2017 ACM on web science conference. 2017: 235-239.
- [40] Rubin V L, Conroy N, Chen Y, et al. Fake news or truth? using satirical cues to detect potentially misleading news[C]//Proceedings of the second workshop on computational approaches to deception detection. 2016: 7-17.
- [41] Brewer P R, Young D G, Morreale M. The impact of real news about “fake news” : Intertextual processes and political satire[J]. International Journal of Public Opinion Research, 2013, 25(3): 323-343.
- [42] Rubin V L, Chen Y, Conroy N K. Deception detection for news: three types of fakes[J]. Proceedings of the Association for Information Science and Technology, 2015, 52(1): 1-4.
- [43] Waldrop M M. The genuine problem of fake news[J]. Proceedings of the National Academy of Sciences, 2017, 114(48): 12631-12634.
- [44] Berkowitz D, Schwartz D A. Miley, CNN and The Onion: When fake news becomes realer than real[J]. Journalism practice, 2016, 10(1): 1-17.
- [45] Kshetri N, Voas J. The economics of “fake news”[J]. IT Professional, 2017, 19(6): 8-12.
- [46] Kucharski A. Study epidemiology of fake news[J]. Nature, 2016, 540(7634): 525-525.
- [47] Buntain C, Golbeck J. Automatically identifying fake news in popular twitter threads[C]//2017 IEEE international conference on smart cloud (smartCloud). IEEE, 2017: 208-215.
- [48] 金志威,曹娟,王博,王蕊,张勇东.融合多模态特征的社会多媒体谣言检测技术研究 [J]. 南京信息工程大学学报(自然科学版),2017,9(06):583-592.DOI:10.13878/j.cnki.jnuist.2017.06.003.
- [49] Sunstein C R. On rumors: How falsehoods spread, why we believe them, and what can be done[M]. Princeton University Press, 2014.
- [50] Zhang Y, Jin R, Zhou Z H. Understanding bag-of-words model: a statistical

- framework[J]. International journal of machine learning and cybernetics, 2010, 1: 43-52.
- [51] Jones K S. Index term weighting[J]. Information storage and retrieval, 1973, 9(11): 619-633.
- [52] Hinton G E. Learning distributed representations of concepts[C]//Proceedings of the eighth annual conference of the cognitive science society. 1986, 1: 12.
- [53] Le Q, Mikolov T. Distributed representations of sentences and documents[C]//International conference on machine learning. PMLR, 2014: 1188-1196.
- [54] 李炳臻, 刘克, 顾佼佼等. 卷积神经网络研究综述[J]. 计算机时代, 2021, No.346(04):8-12+17.DOI:10.16644/j.cnki.cn33-1094/tp.2021.04.003.
- [55] 赵亮. 多模态数据融合算法研究[D]. 大连理工大学, 2018.
- [56] 张君玮. 基于数据融合及卷积神经网络的结构损伤识别方法研究[D]. 广东工业大学, 2022.DOI:10.27029/d.cnki.ggdgu.2022.000314.
- [57] 陈昭昀. 基于异构多模态数据的虚假新闻智能识别技术研究与实现[D]. 北京邮电大学, 2021.DOI:10.26969/d.cnki.gbydu.2021.002364.
- [58] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.
- [59] Tenenbaum J B, Freeman W T. Separating style and content with bilinear models[J]. Neural computation, 2000, 12(6): 1247-1283.
- [60] 刘亚敏. 基于深度域对抗神经网络的滚动轴承迁移故障诊断方法研究 [D]. 河南师范大学, 2021.DOI:10.27118/d.cnki.ghesu.2021.000993.
- [61] 刘鹏飞. 基于多模态特征及语义增强的虚假新闻检测算法的研究与应用[D]. 山东科技大学, 2020.DOI:10.27275/d.cnki.gsdku.2020.000488.
- [62] Zhang T, Wang D, Chen H, et al. BDANN: BERT-based domain adaptation neural network for multi-modal fake news detection[C]//2020 international joint conference on neural networks (IJCNN). IEEE, 2020: 1-8.
- [63] 张国标, 李洁, 胡潇戈. 基于多模态特征融合的社交媒体虚假新闻检测[J]. 情报科学, 2021, 39(10):126-132.DOI:10.13833/j.issn.1007-7634.2021.10.017.
- [64] 邓颖. 基于多模双线性池化的细粒度图像识别方法[D]. 华中师范大学, 2022.DOI:10.27159/d.cnki.ghzsu.2022.003047.
- [65] Charikar M, Chen K, Farach-Colton M. Finding frequent items in data

- streams[C]//Automata, Languages and Programming: 29th International Colloquium, ICALP 2002 Málaga, Spain, July 8 - 13, 2002 Proceedings 29. Springer Berlin Heidelberg, 2002: 693-703.
- [66] Pham N, Pagh R. Fast and scalable polynomial kernels via explicit feature maps[C]//Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. 2013: 239-247.
- [67] Kingma D P, Welling M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.
- [68] Bourlard H, Kamp Y. Auto-association by multilayer perceptrons and singular value decomposition[J]. Biological cybernetics, 1988, 59(4-5): 291-294.
- [69] Chen Z, Badrinarayanan V, Lee C Y, et al. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks[C]//International conference on machine learning. PMLR, 2018: 794-803.

## 附录 A

### 附录 A 题目

## 在学期间的研究成果

### 一、发表论文

1. Guo Y, Ge H, Li J. Fake News Detection Based on Two-Branch Network and Domain Adversarial[C]//2022 IEEE 5th International Conference on Computer and Communication Engineering Technology (CCET). IEEE, 2022: 172-176. (EI 检索, 第二作者, 已发表)
2. Guo Y, Ge H, Li J. A Two-Branch Multimodal Fake News Detection Model Based on Multimodal Bilinear Pooling and Attention Mechanism[J]// 2023 Frontiers in Computer Science (SCI 检索, CCF B 类期刊, 第二作者, 已发表)
3. Guo Y, Li B, Ge H, Di C. An auxiliary modality based Text-Image matching methodology for fake news detection// ICANN (SCI 检索, CCF C 类期刊, 第三作者, 已投稿)
4. Guo Y, Li B, Ge H, Li G. A Dual-Branch Variational Autoencoder structurized Multi-Modal Networks for Interpretable Fake News Detection in Social Media Platform// ACM MM (SCI 检索, CCF A 类期刊, 第三作者, 已投稿)

### 二、参与课题

1. 北京市社会科学基金（21XCCC013）规划项：AI 赋能北京社交媒体平台的舆情分析与谣言检测研究

## 致 谢

在研究生阶段的学习和研究中，我受到了许多人的支持和帮助，在此我要向他们表示感谢。

首先，我要感谢我的导师李晋宏老师和郭颖老师，他们给予我耐心细致地指导和关心。他们的学识和经验让我受益匪浅，他们对我的研究提出了宝贵的建议和指导，使我能够克服各种难题，获得更多的研究经验。

其次，我要感谢我的同学和朋友，感谢 1003 宿舍里面的崔雨欣同学、许云飞同学、赵熙雅同学、张倩同学，她们是我在研究生期间最好的伙伴，给我带来了很多的温暖和关心；感谢 917 高级实验室里的邹欣育同学、张宇同学、苏家堃同学，他们给我带来了研究生期间最开心最难忘的时光；感谢 918 实验室里面的姜山同学、李丙鑫同学、李海虎同学、胡淑婷同学、晋洋旗同学，和他们一起学习的努力和付出是以后路上坚实的宝藏。

我要感谢我的父母，他们对我的鼓励和支持是我研究生阶段乃至一生最大的动力，他们一直陪伴我，支持我，在我遇到困难和挫折时给我鼓励和帮助。感谢他们给了我生命。

最后，感谢我的男朋友孟肖先生为我付出的所有。

再次感谢所有帮助，支持我的人，他们对我的支持和鼓励是我生命中最宝贵的财富。