计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.009823] [CSTR: 32024.14.csa.009823]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 DeepLIFT 算法的可解释多模态假新闻检测①

王　炎[1], 应　龙[2]

[1](南京信息工程大学 软件学院, 南京 210044)
[2](南京信息工程大学 计算机学院, 南京 210044)
通信作者: 王　炎, E-mail: 202212210004@nuist.edu.cn

摘　要: 目前, 大多数多模态可解释假新闻检测方法忽视了对解释数据和跨模态特征的进一步研究利用, 导致可解释假新闻检测方法, 虽然对模型的决策做出了解释, 但是模型检测性能并没有优于先进的多模态检测方法. 针对这些问题, 提出了一种迭代的可解释多模态假新闻检测框架. 该方法由主模型和解释模块构成, 二者都接收多模态新闻作为输入. 首先, 解释模块中将 DeepLIFT 解释算法计算出的解释数据也作为主模型的输入之一, 参与到主模型的决策过程. 接着, 主模型中通过多任务网络框架计算出跨模态相关特征和跨模态补充特征, 并通过跨模态相关特征的粗预测分数对跨模态补充特征重新加权进行细化, 多种特征拼接起来进行模型决策. 最后, 解释模块利用知识蒸馏从主模型转移决策知识进行训练. 主模型和解释模块交替训练, 整体构成了迭代的框架, 在提供决策解释的同时, 进一步提升模型检测性能. 在两个公开的假新闻检测数据集上进行大量实验, 实验结果证明所提出的方法优于最先进的多模态假新闻检测方法.

关键词: 多模态; 假新闻检测; 解释算法; 知识蒸馏; 特征融合

引用格式: 王炎,应龙.基于 DeepLIFT 算法的可解释多模态假新闻检测.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9823.html

Explainable Multimodal Fake News Detection Based on DeepLIFT Algorithm

WANG Yan[1], YING Long[2]

[1](School of Software, Nanjing University of Information Science & Technology, Nanjing 210044, China)
[2](School of Computer Science, Nanjing University of Information Science & Technology, Nanjing 210044, China)

Abstract: Currently, most explainable multimodal fake news detection methods overlook the further research and utilization of explanation data and cross-modal features. As a result, while these explainable fake news detection methods provide explanations for model decisions, their detection performance does not surpass that of advanced multimodal detection methods. To address these issues, this study proposes an iterative explainable multimodal fake news detection framework. This method consists of a main model and an explanation module, both of which receive multimodal news as input. First, the explanation module uses the explanation data calculated by the DeepLIFT algorithm as one of the inputs to the main model, contributing to the decision-making process. Next, the main model calculates cross-modal relevant features and cross-modal supplementary features through a multi-task network framework. It refines the cross-modal supplementary features by re-weighting them with the coarse prediction scores from the cross-modal relevant features and combines multiple features to make the final model decision. Finally, the explanation module trains by transferring decision knowledge from the main model by using knowledge distillation. The main model and the explanation module are trained alternately, forming an iterative framework that enhances model detection performance while providing decision explanations. Extensive experiments on two publicly available fake news detection datasets demonstrate that the

① 基金项目: 国家自然科学基金 (61902193)
收稿时间: 2024-07-16; 修改时间: 2024-10-30; 采用时间: 2024-11-12; csa 在线出版时间: 2025-01-21

proposed method outperforms state-of-the-art multimodal fake news detection methods.

Key words: multimodal; fake news detection; explanation algorithm; knowledge distillation; feature fusion

1 引言

当下, 社交媒体在我们的日常生活中无处不在, 越来越多的人选择通过社交媒体获取知识, 通过社交媒体分享信息、表达意见. 不幸的是, 由于用户众多, 社交媒体网站上滋生了各种各样的假新闻. 这些广为传播的假新闻被一些别有用心的人加以利用来误导读者. 特别是媒体行业、公共机构假新闻的传播, 这可能会对社会造成严重的危害. 因此, 假新闻检测已成为一个重要的研究领域[1]. 假新闻检测技术在分析虚假信息的可能性方面发挥着至关重要的作用, 有效的假新闻检测系统能够帮助公众识别虚假信息, 提升媒体素养, 减少信息误导, 增强社会对信息的信任. 基于机器学习的假新闻检测的一般范式是将新闻转化为多维潜在表示, 并使用二元分类来识别新闻的真假. 现有的方法可以分为 3 类: 单模态假新闻检测[2,3]、多模态假新闻检测[4,5]和动态假新闻检测[6]. 假新闻不仅限于文本, 还包括图片、视频和音频等多种形式. 单一模态的检测方法往往无法充分捕捉信息的复杂性. 多模态学习能够综合不同类型的数据特征, 提供更全面的假新闻判断. 因此我们的工作重点是基于多模态内容的方法.

尽管已经设计了许多基于多模态内容的假新闻检测模型, 但它们获得检测结果的计算过程很难被人们所理解. 这种情况下, 越来越需要能够解释媒体新闻为何被识别为假新闻的机制, 可解释性不仅有助于建立信任, 还可以帮助开发者识别和修正潜在的偏见或错误. 目前, 这方面的研究集中在解释算法[7]上, 旨在解释深度神经网络所做的决策. 有的用于假新闻检测的可解释方法是在主体网络中插入重要性权重引导的不同模态可解释推理模块来实现的[8,9], 有的通过内容与外部数据之间的语义关系[10]计算贡献权重来进行解释. 然而, 这些方法的性能并不如最新的多模态检测方法[11,12]. 我们认为, 并非所有有助于假新闻检测的处理都可以用当前先进的理论和技术来制定为可解释的模型. 因此, 如何在不损害复杂模型表示能力的情况下制定可解释的处理, 值得我们研究.

此外, 还应该仔细分析媒体新闻的多模态表示. 之前的多模态假新闻检测方法简单地将文本和图像特征拼接为共享表示[13], 忽略了它们之间的相关性. 现在大多数复杂的方法都采用跨模态相关性来生成融合特征. 然而, 多模态特征不应仅包含跨模态相关性, 还应考虑其他因素, 例如多模态内容中的情感差异和联合分布. BMR[11]方法使用独立训练的 token 来表示跨模态不相关性. 然而, 它是间接在整个训练数据集上获得的, 会引入数据集偏差和噪声, 降低泛化能力. 因此, 如何有效提取和利用多模态特征中的跨模态补充特征, 值得进一步研究.

为了应对上述挑战, 受到自训练范式的启发, 我们提出一种迭代的可解释多模态假新闻检测方法. 本文方法通过循环迭代结构将可解释模型集成到复杂的多模态主模型中, 这两个部分协同训练并协同执行推理. 主模型以多模态内容和解释数据作为输入, 将解释数据作为不可解释处理的补充. 可解释模型利用知识蒸馏[14]从主模型转移决策知识, 并与主模型可解释子过程对齐, 形成迭代结构. 这个过程中产生的解释数据包含着不确定性, 因此, 我们引入基于正则化的自训练方法[15], 用来防止噪声信息的过度利用, 对模型进行更有效的引导. 对于多模态表示, 利用改进的多门混合专家网络 iMMoE[11]来细化和融合特征. 按照多任务范式学习单模态和跨模态表示. 跨模态补充特征是与多任务范式中的跨模态相关特征协同获得的, 然后通过跨模态相关特征的粗略预测分数重新加权以获得更好的互补性.

本文工作的主要内容如下.

(1) 提出了一种可解释的多模态假新闻检测迭代框架. 通过从复杂主模型转移知识, 显式的解释数据补充、模式对齐到可解释子过程, 构成了将可解释模型集成到复杂主模型的循环迭代结构, 并通过自训练算法进行优化.

(2) 跨模态补充特征是与多任务范式中的跨模态相关特征协同学习的, 其中粗预测被视为深度监督和补充细化.

(3) 在两个公共数据集 Weibo 和 GossipCop 上的实验证明了该方法的有效性.

2 相关工作

2.1 单模态假新闻检测

基于文本的假新闻检测和基于视觉的假新闻检测都受到了广泛关注. Agarwal 等人[16] 通过探索细粒度和粗粒度特征的融合, 捕捉句子中词语之间复杂的相互依赖关系以及语义来检测假新闻. TM[2]利用文本的词汇和语义属性来检测假新闻. Zhang 等人[17]采用基于卷积的神经计算框架对新闻文本进行特征表示提取. 这样的设计能够同时保证中文短文本场景下的处理速度和检测能力. Guo 等人[3]将基本的预训练语言处理模型 Transformer 扩展到多尺度格式, 通过多尺度 Transformer 充分捕捉文本的语义信息, 提高检测性能. 然而, 对于多模态新闻, 这些方法难以检测到跨模态相关性.

2.2 多模态假新闻检测

多模态假新闻检测的关键问题是协调语言和视觉表示. SpotFake+[18]方法集成预训练的 XLNet 和 ResNet 用于特征提取. Wang 等人[19]使用通过语言和视觉表示连接的多模型特征来识别新闻帖子, 引入额外的鉴别器来减轻特定事件的影响. Chen 等人[20]提出的 CAFE 方法, 通过评估单峰特征分布之间的 KL 散度来估计不同模态之间的模糊性, 并将这种模糊性结合到跨模态融合中, 以便更好地自适应聚合. MCAN[12]堆叠多个共同注意力层以进行多模态特征融合. Ying 等人[11]采用改进的多门混合专家网络 iMMoE 进行特征细化和融合. 通过粗预测的分数、每个视图的保真度和跨模态一致性的分数对每个视图的表示进行重新加权, 引导用于假新闻检测. MCNN[21]还融合了文本语义特征、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. MMCSC[22]通过提取文本、图像和视频等多模态新闻的高层语义特征, 分析不同模态高层语义信息, 设计跨模态主题一致性和跨模态情感一致性计算方法, 来检测社交媒体中的假新闻.

2.3 可解释假新闻检测

目前, 一些可解释假新闻检测方法是以权重系数的形式实现内在可解释性. dEFEND[23]提出利用共同注意机制, 共同捕捉新闻句子和用户评论的内在可解释性, 提高假新闻检测性能. Ge 等人[10]提出了 EDI 模型, 利用注意力学习用户评论的表示, 并利用协同注意力获取新闻内容情感特征与用户评论之间的相关性, 提供了解释. KMGCN[24]利用知识图中的多模态内容和外部知识级连接进行假新闻检测. Khan 等人[9]提出的 ESPCN 模型利用用户评论和新闻内容相互理解 top-k 可解释的可检查的用户评论和句子来检测假新闻. Wang 等人[25]提出了一种新的基于防御的可解释假新闻检测框架, 通过对防御行为进行建模来判断防御行为的准确性, 以此来从多样化的叙事中检测假新闻.

此外, 还有一些常用的解释算法也用于可解释的假新闻检测任务. SmoothGrad[26]通过向输入数据中添加噪声并观察对模型预测的影响来提高可解释性. DeepLIFT[27]是用于为给定输出的输入分配重要性分数的解释算法. Grad-CAM[28]方法可视化模型在特定区域的注意力权重, 从而深入了解输入的哪些部分对模型的决策贡献最大. 集成梯度[29]算法将深度网络的预测归结为具有其输入特征的问题, 通过对梯度操作的调用来计算出输入特征的权重.

3 方法设计与实现

3.1 问题定义

一般来说, 我们将社交媒体上的假新闻检测任务定义为一个二分类问题, 旨在将社交媒体上发布的消息分为假新闻和真新闻. 给定一组从社交媒体中收集的多模态新闻$S=\{s_1,\cdots,s_n\}$, 其中s_i表示由文本单词和图像内容组成的新闻, n 是新闻数量. 我们的目的是学习一个模型$f:S\to Y$, 对每个新闻s_i分类到预定义类别$Y=\{0,1\}$, 其中 1 表示假新闻, 0 表示真实新闻.

3.2 整体架构

我们方法的整体架构如图 1 所示, 共包含 4 模块. 主模型由 3 个模块组成: 多模态表示 (MMR) 模块、解释数据 (ED) 模块和检测 (detection) 模块. 多模态表示模块、解释模块以多模态新闻作为输入, 解释数据 (ED) 模块以解释模块输出的解释数据作为输入. 检测模块以多模态新闻和解释数据作为输入, 负责假新闻检测. 解释 (explanation) 模块以多模态新闻作为输入, 负责计算解释数据, 作为模型决策的解释和主模型的输入. 主模型接收到多模态新闻和解释数据, 进行模态特征的提取融合, 再通过 iMMoE 和多模态预测, 对特征进行细化处理, 输入到检测模块进行检测. 解释模块的可解释模型利用知识蒸馏获取主模型的决策知识进行训练, DeepLIFT 算法通过可解释模型对多模态新闻的决策计算出解释数据, 作为解释数据模块的输入. 主

模型和解模块交替多次训练, 整体构成迭代结构. 主模型训练完成后, 解释模块内部的可解释模型利用知识蒸馏获取主模型的决策知识进行训练, DeepLIFT 算法通过可解释模型对多模态新闻的决策计算出解释数据. 主模型和解释模块交替多次训练, 构成迭代结构, 在提供决策解释的同时, 进一步提升模型检测性能.

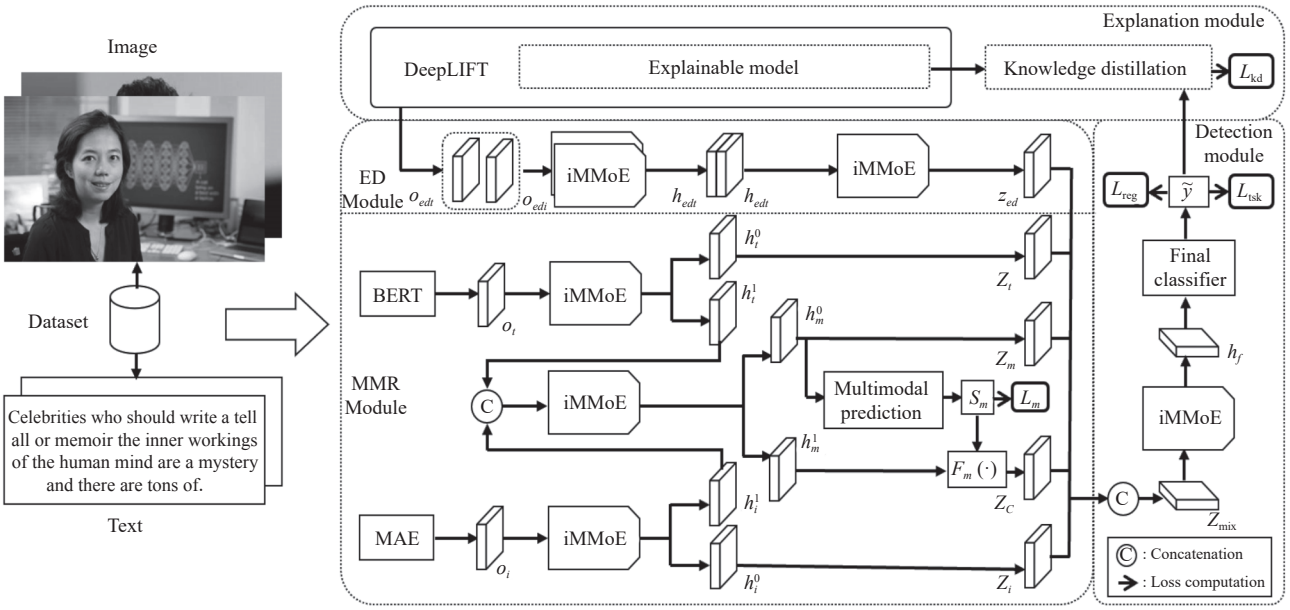

图 1 所提方法的模型结构

3.3 多模态表示模块

多模态表示模块用于提取和生成多模态特征. 设输入的多模态新闻为$N=[I,T]\in D$, 其中 I 和 T 分别表示文本和图像内容, D 表示数据集. 首先提取文本和图像内容的原始特征, 我们利用 BERT 在文本分支中提取o_t, 并使用 masked autoencoder (MAE) 在图像分支中提取o_i. 特征提取后, 采用 iMMoE 网络进行特征细化, 该网络在 MMoE[30]网络上进行改进, 通过在所有任务之间共享专家来建模数据中的多任务关系. 输入 x 被同等地输入到 n 个专家网络中, 其中输出由 k 个门自适应加权, 以形成没有 Softmax 约束的最终结果, 从而允许权重超出[0, 1]范围. 门的输出也由输入 x 来计算. 具体来说, 输入的每个 token 表示都被发送到 MLP 网络, 并通过 token 注意力获得重要性得分. 所有 token 表示都会根据分数聚合为一个, 然后将聚合表示发送到门, 用来计算每个专家的权重. 式 (1) 总结了 iMMoE 算法:

$$x^k=\sum_{i=1}^{n}\left(G_i^k\left(\sum_{j=1}^{t}MLP_k(x)\right)\cdot E_i(x)\right) \tag{1}$$

其中, E_i代表第 i 个专家, G_i^k表示任务 k 的门的第 i 个输出, t 是 token 的数量, MLP_k 表示任务 k 的 token 注意力, $k\in[1,2]$.

在多模态表示模块中, 我们将单模态表示和跨模态特征生成视为不同的子任务. 通过 iMMoE 网络进行处理, 对原始文本特征o_t和原始图像特征o_i进行细化, 分别得到$[h_t^0,h_t^1]$ 和 $[h_i^0,h_i^1]$. h_t^1 和 h_i^1 输入到融合分支中的 iMMoE 网络中, 以生成跨模态相关特征h_m^0以及跨模态补充特征h_m^1. h_t^0 和 h_i^0不再进一步处理, 直接表示为z_t 和 z_i.

3.3.1 跨模态相关特征和补充特征

我们认为多模态表示不仅表示跨模态相关性, 还有许多其他因素, 例如情绪差异和共同分布, 可以帮助确定新闻是真还是假, BMR[11]方法使用独立可训练的 token 来表示跨模态不相关性, 但是, 它的方法会引入数据集偏差和噪声, 降低泛化能力. 因此, 我们提出了一种协同提取融合分支中跨模态相关特征和跨模态补充特征的方法. 首先利用 iMMoE 网络通过多任务学习得到特征h_m^0和h_m^1, 遵循深度监督的方式, 将多模态新闻的真实性预测作为辅助任务来学习多模态相关特征, 使用 MLP 网络计算h_m^0, 得到预测分数S_m.

$$S_m = MLP_m(h_m^0) \tag{2}$$

我们认为跨模态相关性特征包含了判别新闻的重要信息, 不用重新加权, 跨模态补充特征为跨模态相关性特征提供了更多的补充, 需要根据粗预测分数重新加权进行细化. 用另一个 *MLP* 网络, 在图 1 中表示为 $F(\cdot)$, 将预测分数投影为权重.

$$\begin{aligned} z_m &= h_m^0 \\ z_c &= h_m^1 \cdot F_m(S_m) \end{aligned} \tag{3}$$

最后它们都被视为多模态表示, 输入到检测模块.

3.4 解释数据模块

我们将主模型的决策知识转移到可解释模型以获取解释数据. 解释模块计算出的文本和图像解释数据被该模块作为输入, 并分别由 iMMoE 网络进行处理. 然后通过另一个 iMMoE 网络对拼接之后的 h_{edt} 和 h_{edi} 进行细化和融合, 以获得特征 z_{ed} .

3.5 检测模块

多模态表示模块的多个输出特征和解释数据模块的输出特征 z_{ed} 被拼接后一起送入 iMMoE 网络进行对齐、细化和融合, 获得 h_f . 最后, 基于 *MLP* 的分类器采用特征 h_f 的第 1 个 token 来预测 $\tilde{y}$, 预计会接近真实标签 y.

3.6 解释模块

通过知识蒸馏[14]技术将主模型中的决策知识转移到可解释模型中, 可以保留大部分重要知识. 采用 DeepLIFT[27]解释算法, 用于为给定输出的输入单元分配重要性分数, 我们用来计算文本和图像内容的解释数据.

3.6.1 知识蒸馏

知识蒸馏包含一系列旨在将知识从重模型 (教师) 转移到轻模型 (学生) 的方法. 在分类任务中, 对于第 t 类的训练样本, 分类概率可以表示为 $p = [p_1, \cdots, p_t, \cdots, p_C] \in R^{1\times C}$, 其中 p_t 是目标类的概率, C 是类的数量. p 中的每个元素 p_i 可以通过 Softmax 函数得到:

$$p_i = \frac{\exp(z_i)}{\sum_{j=1}^{C}\exp(z_i)} \tag{4}$$

其中, z_i 表示第 i 类的 logit.

用 KL-Divergence 作为知识蒸馏的损失函数, 可以写为:

$$L_{kd} = KL(p^{\mathrm{T}} \| p^{\mathrm{S}}) = p_t^{\mathrm{T}} \log\left(\frac{p_t^{\mathrm{T}}}{p_t^{\mathrm{S}}}\right) + \sum_{i=1, i\neq t}^{C} p_i^{\mathrm{T}} \log\left(\frac{p_i^{\mathrm{T}}}{p_i^{\mathrm{S}}}\right) \tag{5}$$

其中, T 和 S 分别表示教师和学生, p_t 是目标类别的概率.

3.6.2 可解释模型

将原始文本和图像内容作为输入. 文本内容在嵌入层中进行编码, 图像内容在预处理层中进行处理. 文本和图像特征分别由 LSTM 和 ResNet 提取, 接着将它们拼接后, 输入到全连接层进行多模态融合, 最后进行分类, 如图 2 所示.

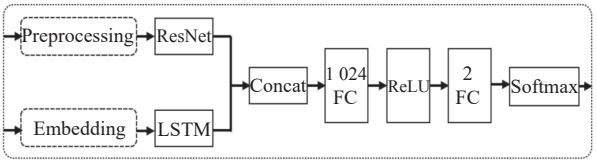

图 2 可解释模型结构图

3.6.3 DeepLIFT 解释算法

DeepLIFT 是用于为给定输出的输入分配重要性分数的解释算法. 它根据目标输入、目标输出与“参考(reference)”输入、“参考”输出间的差异来进行解释. 在图像模型中的“参考”设置为黑色图像, 而文本模型, “参考”可以是零嵌入向量.

具体来说, t 表示目标输出神经元, $x_1, x_2, \cdots, x_n$ 表示中间层或一组层中的一些神经元, 这些神经元对于计算 t 是必要且充分的. x_i^0 是 x_i 对应的“参考”, 记 $\Delta x_i = x_i - x_i^0$. t^0 表示“参考”输出, 记 $\Delta t = t - t^0$. $C_{\Delta x_i \Delta t}$ 表示 Δx_i 对 Δt 的重要性分数, Δt 为各个输入重要性分数的加和.

$$\sum_{i=1}^{n} C_{\Delta x_i \Delta t} = \Delta t \tag{6}$$

接着通过乘数的链式法则来计算重要性分数. 乘数 $m_{\Delta x \Delta t}$ 定义为 Δx 对 Δt 的重要性分数除以 Δx.

$$m_{\Delta x \Delta t} = \frac{C_{\Delta x \Delta t}}{\Delta x} \tag{7}$$

乘数与偏导数类似, 偏导数 $\frac{\partial t}{\partial x}$ 是指 x 产生无穷小变化时, t 的变化率; 而乘数是指 x 产生一定量的变化后, t 的变化率.

假设我们有一个包含神经元 $x_1, x_2, \cdots, x_n$ 的输入层、一个包含神经元 $y_1, y_2, \cdots, y_n$ 的隐藏层和一些目

标输出神经元 t. 给定$m_{\Delta x_i \Delta y_j}$和 $m_{\Delta y_j \Delta t}$的值, 可以得出$m_{\Delta x_i \Delta t}$:

$$m_{\Delta x_i \Delta t} = \sum_j m_{\Delta x_i \Delta y_j} m_{\Delta y_j \Delta t} \tag{8}$$

式 (8) 就是乘数的链式法则. 给定每个神经元到其直接后继神经元的乘数, 我们可以通过反向传播有效地计算任何神经元到给定目标神经元的乘数. 类似于偏导数的链式法则允许我们通过反向传播计算梯度. 通过乘数链式法则, DeepLIFT 算法可以有效地计算每个输入特征对于输出的贡献.

在解释模块中, 对输入特征计算 DeepLIFT. 这些输入特征是原始数据通过可解释模型的嵌入层和预处理层生成的. DeepLIFT 解释了这些特征中哪些对于模型决策更重要. 我们将解释算法计算出的输入特征重要性分数称为解释数据. 借助可视化工具, 可以将解释数据图形化展示, 如图 3 所示.

图 3 正确分类的新闻示例

3.7 模型训练

3.7.1 主模型训练阶段

在第一轮训练中, 由于此时的可解释模型未经训练, 解释模块没有有效的输出, 我们将主模型中解释数据模块的输入和输出都置为零, 排除输入的影响. 这个阶段称为主模型训练阶段.

假新闻检测的任务损失L_{tsk}被定义为二元交叉熵损失 (BCE). 该损失用于计算真实标签 y 和预测结果$\tilde{y}$, 如式 (9) 所示:

$$L_{\text{BCE}}(y,\tilde{y}) = y\log(\tilde{y}) + (1-y)\log(1-\tilde{y}) \tag{9}$$

在多模态表示模块的融合分支中, 跨模态补充特征重新加权过程是通过粗分类作为辅助任务来学习的. 多模态预测的置信度S_m需要进行 BCE 损失计算. 即 $L_{\text{tsk}} = L_{\text{BCE}}(y,\tilde{y})$, $L_m = L_{\text{BCE}}(y,S_m)$.

在主模型训练阶段的总损失函数可以表示为:

$$L = L_{\text{tsk}} + L_m \tag{10}$$

3.7.2 解释模块训练阶段

当主模型训练完成后, 使用知识蒸馏将决策知识转移到可解释模型. 采用常用的 KL 散度L_{kd}作为损失函数来学习参数. 然后, 执行解释算法来计算每种模态的解释数据, 作为对主模型的输入.

3.7.3 解释数据完备阶段

解释模块的可解释模型经过训练后, 可以得到对主模型有效的解释数据. 然后, 该数据加入主模型训练. 然而, 通过知识转移和可解释模型计算, 解释数据会引入一些不确定性, 应仔细考虑. 因此, 引入自训练的思想来解决这个问题. 在涉及解释数据的训练过程中, 施加正则化函数, 使得相关模块的学习不会过多依赖不确定信息, 其中多模态内容被视为确定性信息. 具体来说, 我们根据批次大小将z_{ed}分为两部分. 一部分是解释模块计算出的解释数据, 被视为包含不确定性的特征, 记为z_{ed_e}. 另一部分置为零以消除不确定信息, 表示为z_{ed_u}. 然后z_{ed_e}和z_{ed_u}分别与z_{mix}连接, 表示为z_{mix_e}和z_{mix_u}. 根据z_{mix_u}的主模型输出的损失函数是任务损失L_{tsk}, 对于输出z_{mix_e}, 采用受半监督学习启发的损失函数, 结合任务损失L_{tsk}和基于L_2的模型正则项[31]可以得到L_{reg}:

$$L_{reg} = reg(p(x_t,w)) = \sum_{k=1}^{K} p(k|x_t)^2 \tag{11}$$

其中, K 是类别数, x_t是z_{ed_e}对应的输入, w 是模型参数.

解释数据完备阶段的总损失函数如下:

$$L = L_{\text{tsk}} + L_m + \alpha L_{reg} \tag{12}$$

其中, α是超参数.

这个训练过程可以多轮迭代, 以提升模型的性能. 我们在算法 1 中提供了训练过程的伪代码.

算法 1. 所提出方法的模型训练

输入: 数据集: D, 训练的 epochs: N, 迭代次数: M.
输出: 主模型 (MM) 参数θ_{mm}, 可解释模型 (IM) 参数θ_{im}.

1) for i in $range(N)$ do
2) 从数据集 D 中采样$[I, T, y]$
3) $[\tilde{y}, S_m] = MM(I, T, explain\_data = \text{False})$
4) 计算损失函数 L_m, L_{tsk}
5) 使用 Adam 优化器更新参数θ_{mm}

6) end for
7) for j in *range*(M) do
8) 通过知识蒸馏将决策知识转移到可解释模型 (θ_{im}).
9) for k in *range*(N) do
10) 从数据集 D 中采样[I, T, y]
11) [o_{edi},o_{edt}] = *DeepLIFT*(*IM*(I, T))
12) [$\hat{y}$, S_m] = *MM*(I, T, o_{edi}, o_{edt}, *explain_data*=True)
13) 计算损失函数L_m, L_{tsk}, L_{reg}
14) 使用 Adam 优化器更新参数θ_{mm}
15) end for
16) end for

4 实验与结果分析

4.1 实验数据集

本文使用中文数据集 Weibo[32]和英文数据集 GossipCop[1]用于训练和测试. Weibo 数据集包含用于训练的 3749 条真新闻和 3783 条假新闻, 以及用于测试 1000 条假新闻和 996 条真新闻. GossipCop 数据集包含用于训练的 7974 条真实新闻和 2036 条假新闻, 以及用于测试的 2285 条真实新闻和 545 条假新闻, 我们将其按照 9:1 的比例分为训练数据和测试数据. 按照 He 等人[33]概述的方法, 我们根据每个数据集的分布确定固定阈值. 对于 GossipCop 训练集, 真新闻与假新闻的比例接近 4:1, 我们将阈值设置为 0.80. 同样, Weibo 的阈值设置为 0.50. 这些阈值用于根据预测分数确定新闻文章的真实或虚假分类.

4.2 实验环境及实现细节

本文的所有实验环境如下: 系统为 Linux, GPU 是 NVIDIA RTX 3090. 我们使用 mae-pretrain-vit-base 模型用于图像处理. 用 bert-base-uncased 模型处理英文数据集, 对于中文数据集用 bert-base-chinese 模型进行处理. 式 (12) 涉及超参数, 实验中将它设置为$\alpha = 0.025$.

4.3 性能分析

我们使用 Accuracy、Precision、Recall 和 F1-score 这 4 个评价指标进行模型性能衡量, 其中 Accuracy 是主要评价指标.

4.3.1 解释数据

DeepLIFT 能够为给定的模型决策输出的输入单元分配重要性分数. 在我们的模型中, 文本的输入单元是文本内容的 token, 图像内容的输入单元是预处理后的图像的像素. DeepLIFT 计算后, 获得每个模态输入单元的重要性分数, 记为解释数据. 解释数据的维度与模态输入单元的维度一致.

将解释数据进行图形化展示时. 图像用像素点展示, 文本用单词来展示. 对于文本内容, 将 token 的重要性分数相加, 形成对应单词的分数. 单词得分被归一化并根据平均值设定阈值. 当一个单词在文本中的重要性分数大于阈值时, 对它的字体进行加粗显示. 对于图像内容, 当图像像素点的重要性分数大于阈值时, 才会进行展示.

实验结果的示例如图 3 所示. 加粗的文本单词和图像中颜色较深的区域代表模型在识别假新闻时关注的重要部分. 图 3 中的两篇文章都是被正确分类的, 并且该模型捕获了文本和图像中包含用于假新闻检测的重要语义信息的部分. 被归类为真实新闻的文本和图像内容之间存在一定的相关性, 如图 3(a) 所示, “特朗普”“采访”“崩溃”是模型主要关注的文本. 在图 3(b) 所示的新闻中, 被归类为假新闻, 文本和视觉内容关注的部分关联性很小. 这些都是判断多模态新闻真实性的重要依据, 其中“双胞胎”“分娩”是模型主要关注的文本内容.

4.3.2 对比实验: 多模态假新闻检测方法

进行对比的多模态假新闻检测方法有: MCAN[12], MCNN[21], EANN[19], CAFE[20], BMR[11], SpotFake+[18]. 表 1 展示了本文所提方法和对比方法在 Weibo 和 GossipCop 数据集上 Accuracy、Precision、Recall 和 F1-score 的评价指标值. 可以看到, 本方法在 Weibo 数据集上的 Accuracy 值是 92.5%, 在 GossipCop 数据集上的 Accuracy 值是 89.6%. 与 BMR 方法相比, 该方法在 Weibo 上的各个指标值都有 0.7% 及以上的提升, 在 GossipCop 上除了 Recall, 其余的指标都有 0.6% 及以上的提升. 总体来看, 与这几种多模态假新闻检测方法相比, 在 Weibo 数据集上, 我们的方法在 Accuracy 上领先 7.2%, 在 Recall 上领先 6.8%. GossipCop 数据集上, 我们的方法在 Accuracy 上高出 3.3%, 在 Recall 上高出 3.4%.

结果表明, 集成可解释模型并与主模型协作, 对于假新闻检测具有积极的效果. 说明本文的假新闻检测方法与其他多模态检测方法相比性能最好.

4.3.3 对比实验: 可解释假新闻检测方法

我们分别在两个数据集上比较了几种经典的可解释假新闻检测方法: dEFEND[23], EDI[10], ESPCN[9]. 根据表 1 的实验结果表明, 在 Weibo 数据集上, 我们的方法 Accuracy 指标有 8.8% 的提升, 在 F1-score 指标

上有 7.8% 的提升. 总体性能明显优于这些方法. 在 GossipCop 数据集上, Accuracy 比这些方法高出 4.3%, *F*1-score 指标上有 7% 的提升, 其他指标也有显著提高.

表 1 不同方法在 Weibo 数据集和 GossipCop 数据集上的对比

数据集	方法	Accuracy	Precision	Recall	*F*1-score
Weibo	MCAN	0.899	0.844	0.847	0.845
	MCNN	0.846	0.898	0.899	0.899
	EANN	0.827	0.827	0.827	0.827
	CAFE	0.840	0.840	0.840	0.839
	BMR	0.918	0.912	0.909	0.909
	dEFEND	0.794	0.811	0.755	0.782
	EDI	0.879	0.910	0.908	0.907
	Ours	**0.925**	**0.924**	**0.921**	**0.922**
GossipCop	SpotFake+	0.858	0.839	0.844	0.830
	EANN	0.864	0.850	0.868	0.854
	CAFE	0.867	0.856	0.863	0.854
	BMR	0.890	0.886	0.894	0.887
	dEFEND	0.863	0.869	0.845	0.849
	ESPCN	0.843	0.805	0.819	0.797
	Ours	**0.896**	**0.894**	**0.892**	**0.893**

与这些可解释多模态检测方法相比, 本方法充分利用了解释数据, 对主模型的不可解释处理进行显式的信息补充, 提升了模型的性能. 实验证明了所提方法的性能优于先进的可解释多模态检测方法.

4.4 消融实验

我们为消融实验设计了 4 种不同的组合: (1) 文本分支、图像分支和跨模态相关特征. 表示为 TIM. (2) 文本分支、图像分支、跨模态相关特征、跨模态补特征. 表示为 TIM+C. (3) 文本分支、图像分支、跨模态相关特征、跨模态补充补充和 ED 模块. 表示为 TIM+C+ED. (4) 所有分支与正则化相结合 (提出的方法), 表示为 Ours. 我们发现, 随着每个分支的加入, 模型的性能逐渐提高. 实验结果如表 2 和表 3 所示.

(1) 和 (2) 之间的比较重点在于跨模态补充特征是否提高了模型的识别性能. 结果证实了所提出的提取跨模态补充特征方法的有效性. (2) 和 (3) 之间的比较重点在于明确地结合解释数据是否可以提高模型的准确性. 结果表明, 解释数据对主模型的不可解释过程起到了补充作用, 对提升准确性有积极影响. (3) 和 (4) 之间的比较重点在于解释数据包含的不确定性是否可以通过基于正则化的训练算法来解决. 实验表明, 施加正则化可以防止噪声信息的过度利用并提高泛化能力.

表 2 Weibo 数据集上的消融实验结果

Test	Accuracy	*F*1-score	
		Fake news	Real news
TIM	0.917	0.915	0.916
TIM+C	0.920	0.918	0.919
TIM+C+ED	0.923	0.921	0.920
Ours	0.925	0.921	0.923

表 3 GossipCop 数据集上的消融实验结果

Test	Accuracy	*F*1-score	
		Fake news	Real news
TIM	0.888	0.655	0.934
TIM+C	0.891	0.670	0.935
TIM+C+ED	0.893	0.682	0.938
Ours	0.896	0.686	0.945

4.5 迭代实验

该方法是一种迭代的框架, 其中主模型训练和可解释模型的知识蒸馏交替进行, 体现了广义的自学习范式. 这种方法在多轮训练中的表现需要进行进一步的评估.

完整的模型分别使用 Weibo 和 GossipCop 数据集训练 6 轮. 每轮的评估均使用 Accuracy, Precision, Recall 和 *F*1-score 指标进行评价. 表 4 和表 5 提供了所有轮次中每个评估指标的详细值. 随着训练轮数的增加, 这些指标值略有提升并趋于稳定.

表 4 Weibo 数据集上的迭代实验结果

Number of iterations	Accuracy	Precision	Recall	*F*1-score
1	0.920	0.914	0.920	0.917
2	0.922	0.921	0.917	0.919
3	0.925	0.924	0.921	0.922
4	0.923	0.922	0.922	0.922
5	0.924	0.924	0.920	0.921
6	0.923	0.921	0.923	0.921

表 5 GossipCop 数据集上的迭代实验结果

Number of iterations	Accuracy	Precision	Recall	*F*1-score
1	0.893	0.892	0.890	0.891
2	0.895	0.895	0.891	0.892
3	0.894	0.892	0.894	0.892
4	0.896	0.894	0.892	0.893
5	0.894	0.895	0.890	0.892
6	0.893	0.893	0.889	0.891

图 4 和图 5 展示了 6 轮迭代中每个指标的变化趋势. 其中 *x* 轴表示模型迭代的次数, 如图 5 中的 *x* 轴上的数值 4 表示模型迭代到了第 4 次, *y* 轴则表示各个评价指标的数值, 图例是 4 个评价指标.

迭代实验验证了, 基于正则化的自训练算法防止

了, 由知识提取和解释算法导致的噪声解释信息的过度利用, 能够引导模型更好地收敛. 通过实验验证了该框架的有效性.

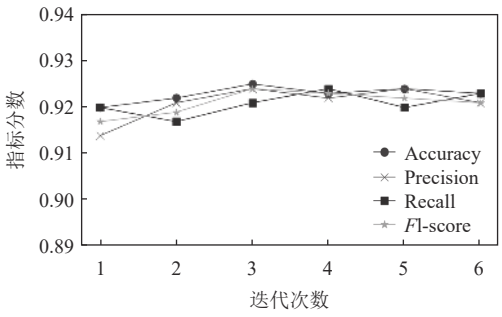

图 4 Weibo 数据集上指标分数 (y 轴) 与迭代次数 (x 轴) 的曲线

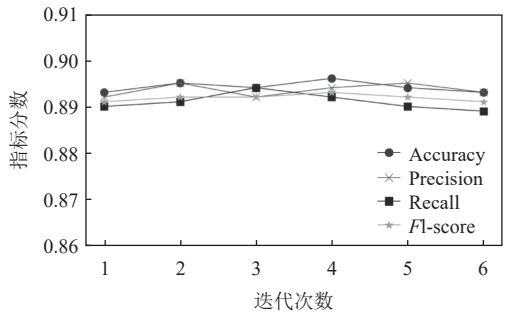

图 5 GossipCop 数据集上指标分数 (y 轴) 与迭代次数 (x 轴) 的曲线

5 结束语

本文提出了一种基于解释算法的可解释迭代框架, 将可解释模型集成到复杂的多模态表示模型中, 以识别假新闻. 可解释模型是协同训练的, 与主模型的可解释子过程保持一致, 并进行推理, 作为不可解释的处理的补充. 通过多任务方式协同获得跨模态相关特征和跨模态补充特征, 然后加权进行细化, 以形成更好的跨模态特征表示. 实验表明, 在常用的数据集上, 我们的方法优于目前先进的多模态假新闻检测方法.

参考文献

1 Shu K, Mahudeswaran D, Wang SH, *et al*. FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big Data, 2020, 8(3): 171–188. [doi: 10.1089/big.2020.0062]

2 Bhattarai B, Granmo OC, Jiao L. Explainable tsetlin machine framework for fake news detection with credibility score assessment. Proceedings of the 13th Language Resources and Evaluation Conference. Marseille: European Language Resources Association, 2022. 4894–4903.

3 Guo ZW, Zhang Q, Ding F, *et al*. A novel fake news detection model for context of mixed languages through multiscale Transformer. IEEE Transactions on Computational Social Systems, 2024, 11(4): 5079–5089. [doi: 10.1109/TCSS.2023.3298480]

4 Zhang G, Giachanou A, Rosso P. SceneFND: Multimodal fake news detection by modelling scene context information. Journal of Information Science, 2024, 50(2): 355–367.

5 李金金, 桑国明, 张益嘉. APK-CNN 和 Transformer 增强的多域虚假新闻检测模型. 计算机应用, 2024, 44(9): 2674–2682. [doi: 10.11772/j.issn.1001-9081.2023091359]

6 Abdelnabi S, Hasan R, Fritz M. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 14940–14949.

7 鞠天杰, 刘功申, 张倬胜, 等. 自然语言处理中的探针可解释方法综述. 计算机学报, 2024, 47(4): 733–758. [doi: 10.11897/SP.J.1016.2024.00733]

8 Zeng Z, Wu MM, Li GD, *et al*. An explainable multi-view semantic fusion model for multimodal fake news detection. Proceedings of the 2023 IEEE International Conference on Multimedia and Expo. Brisbane: IEEE, 2023. 1235–1240.

9 Khan F, Alturki R, Srivastava G, *et al*. Explainable detection of fake news on social media using pyramidal co-attention network. IEEE Transactions on Computational Social Systems, 2024, 11(4): 4574–4583. [doi: 10.1109/TCSS.2022.3207993]

10 Ge XY, Zhang MS, Wang XA, *et al*. Emotion-drive interpretable fake news detection. International Journal of Data Warehousing and Mining, 2022, 18(1): 1–17.

11 Ying QC, Hu XX, Zhou YM, *et al*. Bootstrapping multi-view representations for fake news detection. Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington: AAAI, 2023. 5384–5392.

12 Wu Y, Zhan PW, Zhang YJ, *et al*. Multimodal fusion with co-attention networks for fake news detection. Proceedings of the 2021 Association for Computational Linguistics. ACL, 2021. 2560–2569.

13 Allein L, Moens MF, Perrotta D. Like article, like audience: Enforcing multimodal correlations for disinformation detection. arXiv:2108.13892, 2021.

14 Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv:1503.02531, 2015.

15 Xu R, Yu Y, Cui H, *et al*. Neighborhood-regularized self-

training for learning with few labels. Proceedings of the 2023 AAAI Conference on Artificial Intelligence. Washington DC: AAAI. 2023. 10611–10619. [doi: 10.1609/aaai.v37i9.26260]

16 Agarwal I, Rana D, Panwala K, *et al*. Analysis of contextual features' granularity for fake news detection. Multimedia Tools and Applications, 2024, 83(17): 51835–51851.

17 Zhang Q, Guo ZW, Zhu YY, *et al*. A deep learning-based fast fake news detection model for cyber-physical social services. Pattern Recognition Letters, 2023, 168: 31–38. [doi: 10.1016/j.patrec.2023.02.026]

18 Singhal S, Kabra A, Sharma M, *et al*. SpotFake+: A multimodal framework for fake news detection via transfer learning (student abstract). Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI, 2020. 13915–13916.

19 Wang YQ, Ma FL, Jin ZW, *et al*. EANN: Event adversarial neural networks for multi-modal fake news detection. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London: ACM, 2018. 849–857.

20 Chen YX, Li DS, Zhang P, *et al*. Cross-modal ambiguity learning for multimodal fake news detection. Proceedings of the 2022 ACM Web Conference. Lyon: ACM, 2022. 2897–2905.

21 Xue JX, Wang YB, Tian YC, *et al*. Detecting fake news by exploring the consistency of multimodal data. Information Processing & Management, 2021, 58(5): 102610.

22 赵越, 郝琨, 赵敬, 等. MMCSC: 一种跨模态的假新闻检测方法. 东北大学学报 (自然科学版), 2024, 45(1): 18–25. [doi: 10.12068/j.issn.1005-3026.2024.01.003]

23 Shu K, Cui LM, Wang SH, *et al*. dEFEND: Explainable fake news detection. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage: ACM, 2019. 395–405.

24 Wang YZ, Qian SS, Hu J, *et al*. Fake news detection via knowledge-driven multimodal graph convolutional networks. Proceedings of the 2020 International Conference on Multimedia Retrieval. Dublin: ACM, 2020. 540–547.

25 Wang B, Ma J, Lin HZ, *et al*. Explainable fake news detection with large language model via defense among competing wisdom. Proceedings of the 2024 ACM on Web Conference. Singapore: ACM, 2024. 2452–2463.

26 Smilkov D, Thorat N, Kim B, *et al*. SmoothGrad: Removing noise by adding noise. arXiv:1706.03825, 2017.

27 Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. Proceedings of the 34th International Conference on Machine Learning. Sydney: PMLR, 2017. 3145–3153.

28 Selvaraju RR, Cogswell M, Das A, *et al*. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 618–626.

29 Sundararajan M, Taly A, Yan QQ. Axiomatic attribution for deep networks. Proceedings of the 34th International Conference on Machine Learning. Sydney: JMLR.org, 2017. 3319–3328.

30 Ma JQ, Zhao Z, Yi XY, *et al*. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London: ACM, 2018. 1930–1939.

31 Zou Y, Yu ZD, Liu XF, *et al*. Confidence regularized self-training. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 5980–5990.

32 Jin ZW, Cao J, Guo H, *et al*. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. Proceedings of the 25th ACM International Conference on Multimedia. Mountain View: ACM, 2017. 795–816.

33 He HB, Garcia EA. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263–1284. [doi: 10.1109/TKDE.2008.239]

(校对责编: 张重毅)