Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

融合多模态交互及语义一致性建模的社交机器人检测

网络安全1组

硕士研究生 费泽涛

2023年07月09日

内容提要

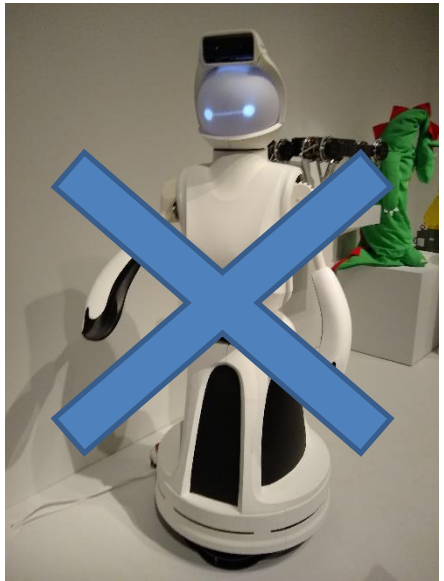
- 背景简介
- 基础概念
- 算法原理
- 应用总结
- 参考文献

预期收获

- 预期收获
  - **1.** 了解社交机器人检测的相关概念
  - **2.** 理解**R-GCN**的基本原理
  - **3.** 理解**BIC**算法、创新点有效性探究方法
  - **4.** 了解社交机器人检测的前沿发展

社交机器人检测

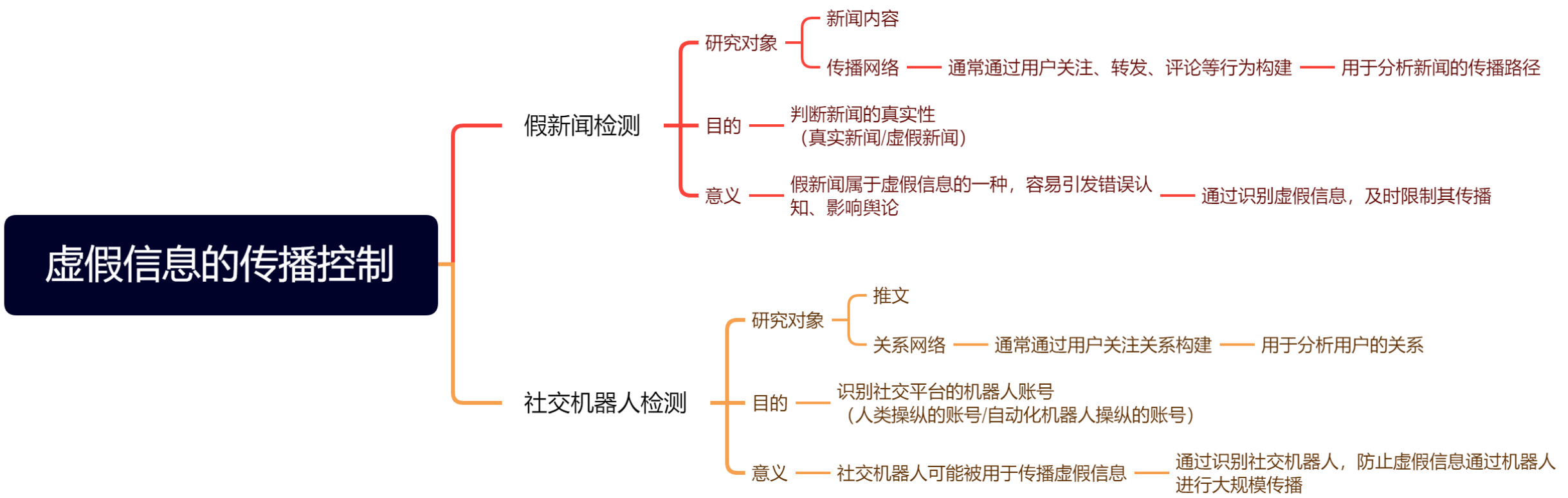
- 研究背景
  - 社交机器人模仿人类在Twitter、Facebook、Instagram等社交平台上的行为。数以百万计的机器人通常基于平台API，通过自动化程序控制，通过模仿真实用户以实现恶意目标，如:
    - 传播极端主义：极端的政治、宗教观点，煽动争议、引起社会对立
    - 干预选举：政治讨论，影响公众意见和选民行为
    - 隐私攻击：发布虚假链接、恶意软件以获取用户隐私信息
    - 传播虚假信息
  - 恶意的社交机器人对网络社区构成威胁



社交机器人检测 是什么?

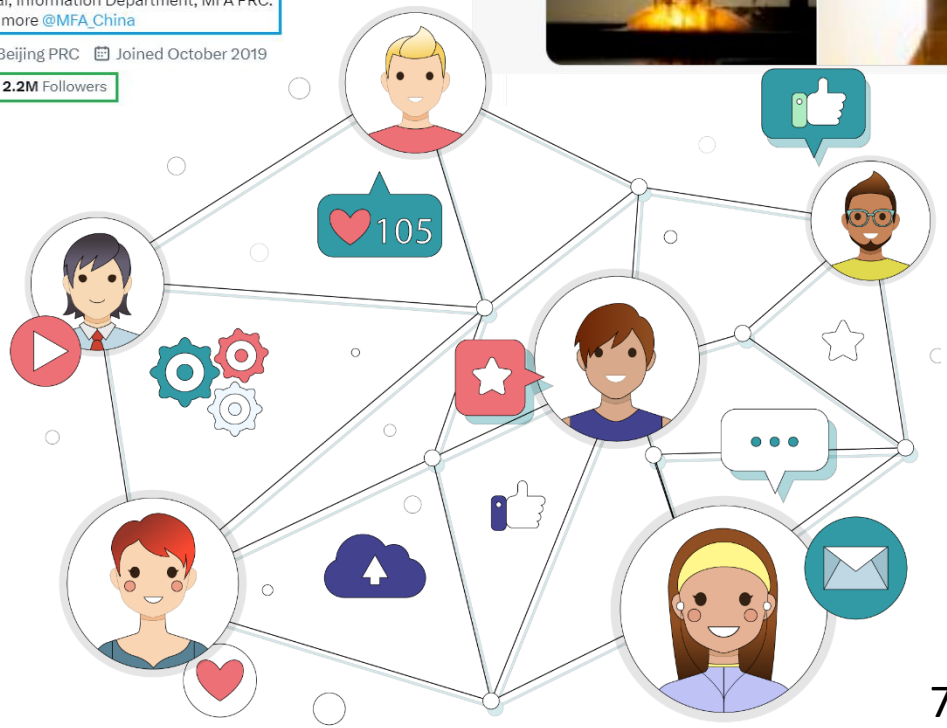
- 社交机器人
  - 在社交网络上模仿人类行为的虚拟实体，基于平台**API**，通过自动化程序控制
  - 种类
    - 良性：聊天机器人，新闻机器人，…
    - 恶意：垃圾邮件机器人，政治机器人，…
- 社交机器人检测
  - 分析社交账号的用户信息及其行为，输出真实用户账号/社交机器人账号
- 任务特征
  - 多样、动态变化：机器人行为多样，且行为特点会随时间推移而变化
  - 隐蔽：通过窃取真实用户推文/随机行动，隐藏社交机器人特征
  - 大规模：社交网络规模庞大，且包含多种节点类型及关系
  - 多模态：需要考虑文本、图像等信息

社交机器人检测 与假新闻检测的关系？



社交机器人检测 怎么做?

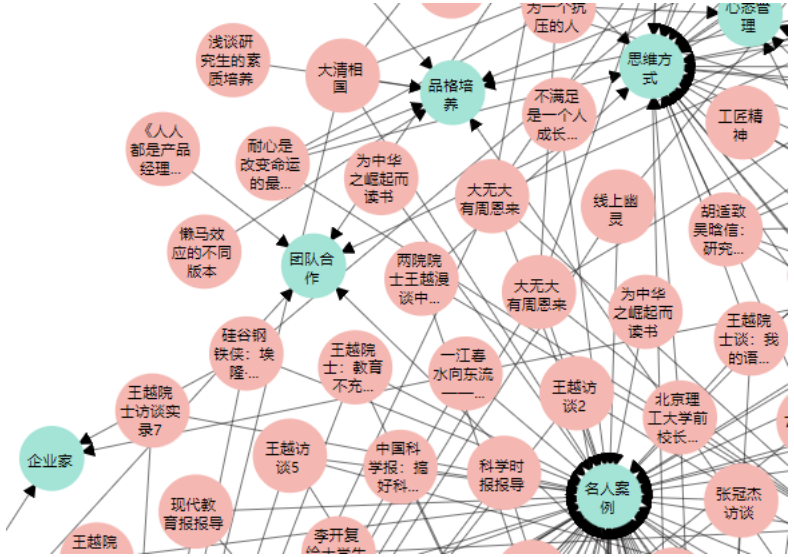
- 研究对象
  - 个人资料：用户名、地址…
  - 用户关系：关注/被关注、评论、点赞
  - 推文
- 数据类型
  - 时序：由推文的时间戳组成
  - 文本：推文、用户资料
  - 图：推文包含的图像；用户关系图



社交机器人检测 怎么做?

- 检测方法
  - 基于机器学习
    - **Random Forest**
    - **AdaBoost**
  - 基于深度学习
    - **LSTM**
    - **RGCN**（关系图卷积神经网络）

思智明理系统



推荐访问：[思智明理项目]

RGCN 是什么?

- 图神经网络(**Graph Neural Networks, GNN**)
  - 学习图结构数据的深度学习网络
  - 提取和发掘图结构数据中的节点特征、边特征
  - 通过聚合节点的邻居信息来更新节点的表示
- 图卷积网络(**Graph Convolutional Network, GCN**)
  - **GNN**的一种变体
  - 使用图卷积层对图中的节点进行表示学习通过多层图卷积操作，**GCN**可以捕捉到不同距离的节点表示

	蒙恬	嬴政	芈月	鲁班	镜	白起	扁鹊
蒙恬		1					
嬴政	1		1		1	1	
芈月		1				1	
鲁班							
镜		1					
白起		1	1				1
扁鹊						1	

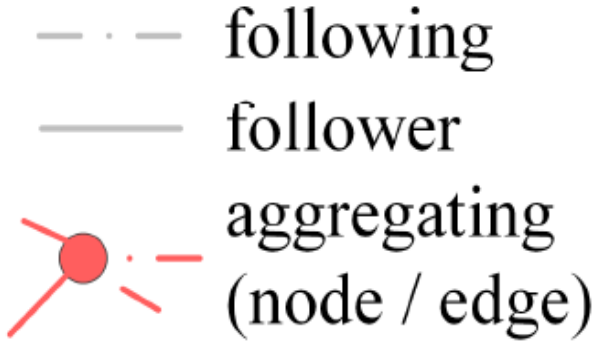
邻接矩阵

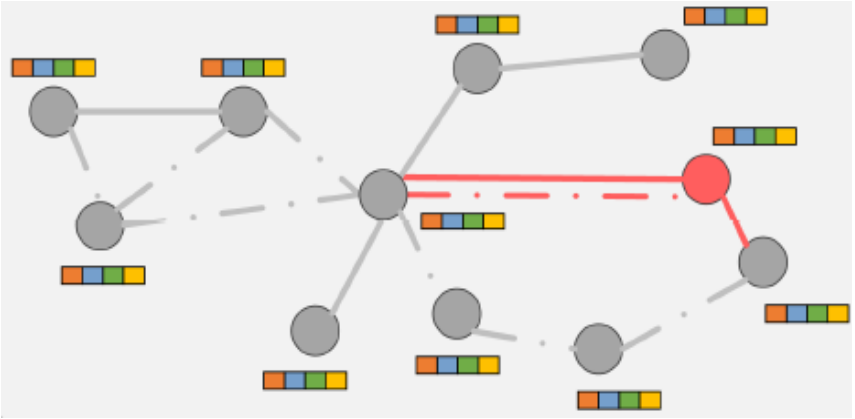
Adjacency

$$\begin{bmatrix} 0&1&0&0&0&0&0 \\ 1&0&1&0&1&1&0 \\ 0&1&0&0&0&1&0 \\ 0&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0 \\ 0&1&1&0&0&0&1 \\ 0&0&0&0&0&1&0 \end{bmatrix}$$

RGCN 是什么?

- 关系图卷积网络(**Relational Graph Convolutional Network, RGCN**)
  - 是GCN的一种具体实现
  - 引入关系嵌入来表示不同类型的关系
  - 对于用户关系图，通常包含关注(**following**)与被关注(**follower**)这两种关系
  - 用户关系图属于异质图





推荐阅读：**BFS学术报告-异质图神经网络-李新帅**

RGCN 怎么做?

- **RGCN (Relational Graph Convolutional Network)**
  - 用户特征表示

$$z = [z_b; z_t; z_p^{num}; z_p^{cat}] \in \mathbb{R}^{D\times 1}$$

  - 节点表示初始化

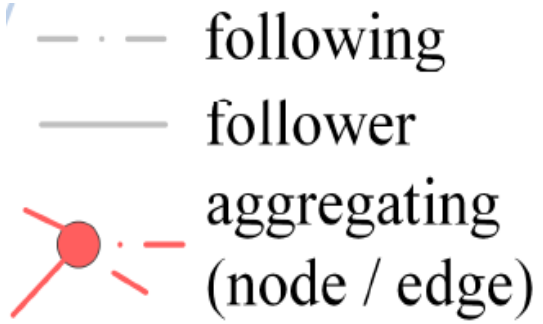
$$x_i^{(0)} = \phi(W_1 \cdot z_i + b_1), \quad x_i^{(0)} \in \mathbb{R}^{D\times 1}$$

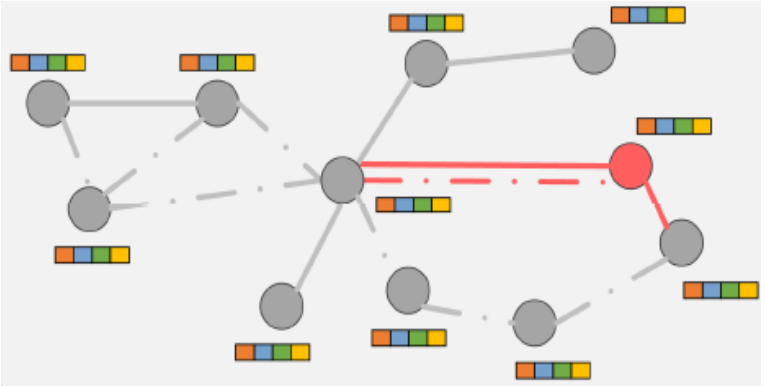
  - 图卷积，聚合节点的邻居信息来更新节点表示

$$x_i^{(l+1)} = \Theta_{self} \cdot x_i^{(l)} + \sum_{r\in R} \sum_{j\in N_r(i)} \frac{1}{|N_r(i)|} \Theta_r \cdot x_j^{(l)}, x_i^{(l+1)} \in \mathbb{R}^{D\times 1}$$

  - 节点表示转换

$$g_i = \phi(W_2 \cdot x_i^{(L)} + b_2), \quad g_i \in \mathbb{R}^{D\times 1}$$





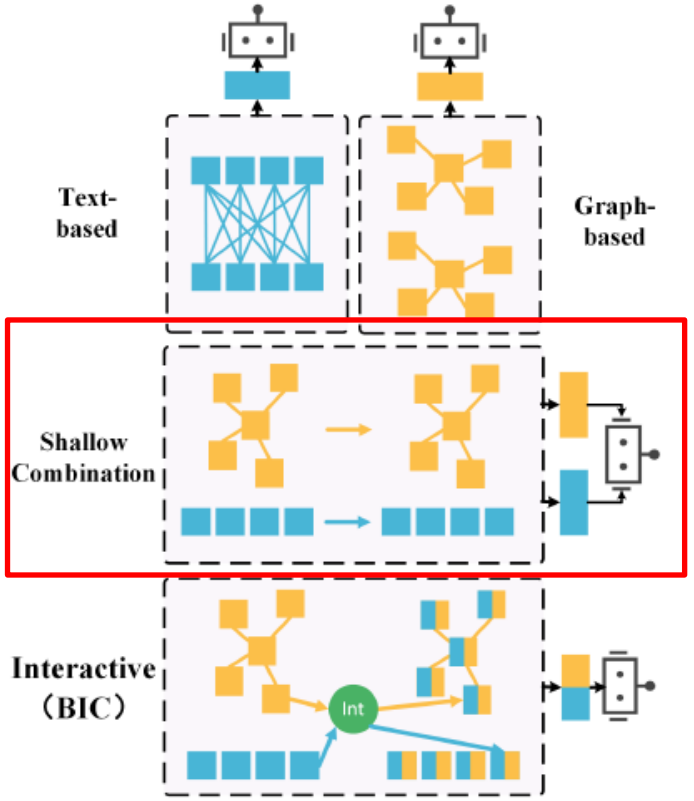
BIC

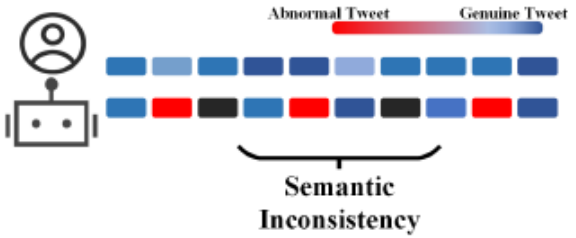
T 目标	辨别高级机器人
I 输入	用户信息
P 处理	1.文本、图特征提取 2.文本-图特征交互 3.语义一致性建模
O 输出	机器人/人类用户

P 问题	1.缺乏对多模态交互的探索 2.高级机器人窃取真实用户推文并稀释恶意内容，加大了检测难度
C 条件	包含文本（用户资料、推文）与用户关系
D 难点	1.如何实现文本与图表示交互并证明其有效性 2.如何实现语义一致性建模并证明其有效性
L 水平	**ACL**, 一区, **IR**=5.698, 2023

BIC 研究问题

- 模型方面
  - 现有方法仅利用文本或网络进行检测，少量工作探索了这两种模态的浅层结合
- 数据方面
  - 高级机器人窃取真实用户推文以稀释恶意内容，加大了检测难度



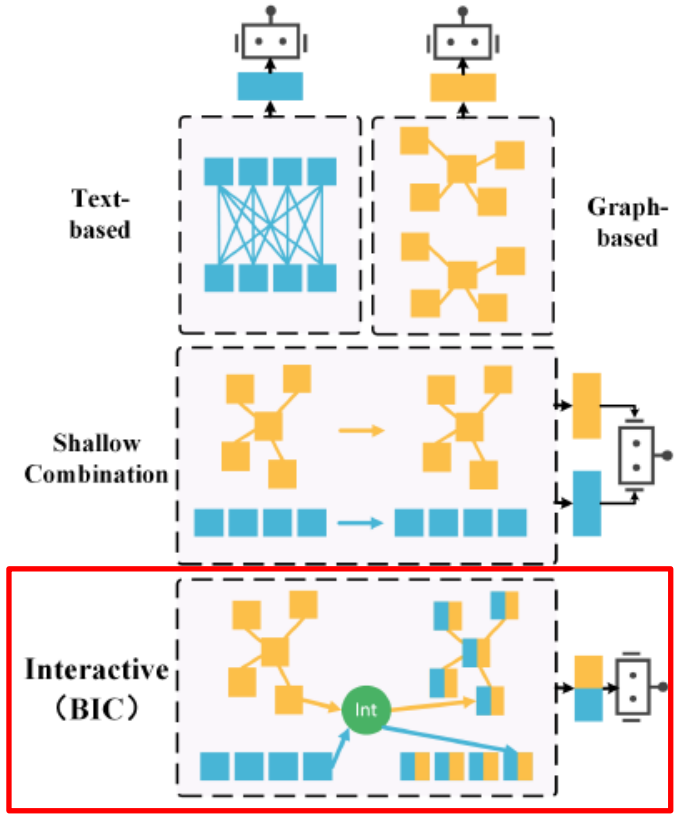


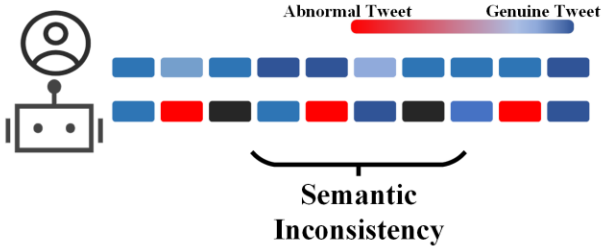
THE REAL BARACK OBAMA
This video should go VIRAL!! President Obama made fun of the flint water crisis and mocked the people.

New peer reviewed study from The Journal of Electrocardiology concludes that #HCQ with Azithromycin. did NOT have any QT prolong…

Mom: omg I went to get my pills today. I asked the lady if she was sure about the total price. It was only $1.54 today. Normally it's $4…

- 解决方案
  - 引入基于文本-图相似度权重的交互表示
  - 提出基于注意力权重的推文语义一致性建模，并使用它来增强决策过程





THE REAL BARACK OBAMA
This video should go VIRAL!! President Obama made fun of the flint water crisis and mocked the people.

New peer reviewed study from The Journal of Electrocardiology concludes that #HCQ with Azithromycin. did NOT have any QT prolong…

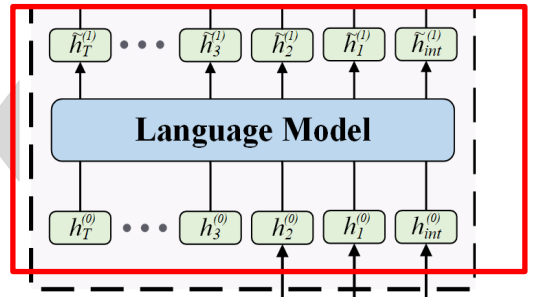
Mom: omg I went to get my pills today. I asked the lady if she was sure about the total price. It was only $1.54 today. Normally it's $4…

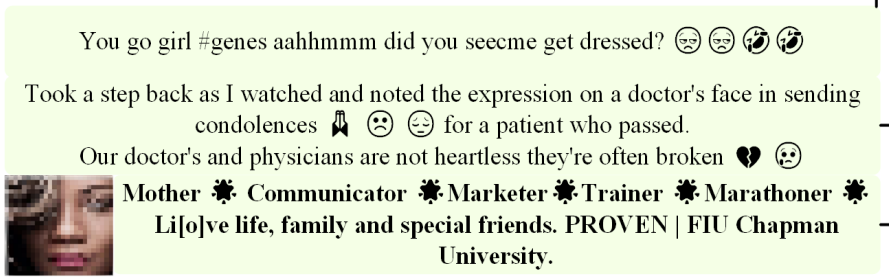
- 文本建模
  - 使用预训练模型**RoBERTa**对用户资料、推文进行编码
  - 基于**Transformer**特征提取

$$\{\tilde{h}_{int}^{(l)}, \tilde{h}_1^{(l)}, \cdots, \tilde{h}_T^{(l)}\} = \mathrm{LM}(\{h_{int}^{(l-1)}, h_1^{(l-1)}, \cdots, h_T^{(l-1)}\})$$

用户资料表示

推文表示





BIC 算法原理

- 图建模
  - 节点表示
  - 关系图卷积网络

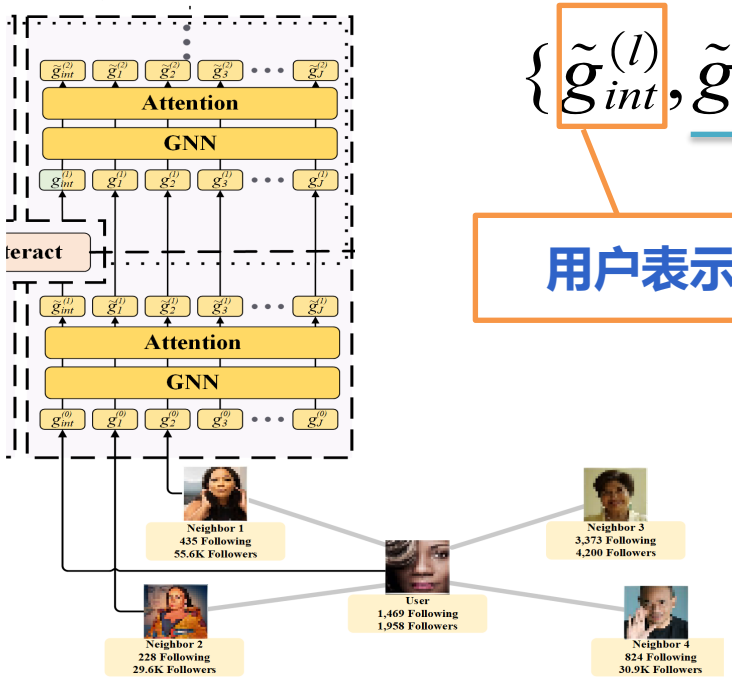
$$\{\hat{g}_{int}^{(l)}, \hat{g}_1^{(l)}, \cdots, \hat{g}_J^{(l)}\} = \mathrm{GNN}\left(\{g_{int}^{(l-1)}, g_1^{(l-1)}, \cdots, g_J^{(l-1)}\}\right)$$

  - 多头注意力

$$\{\tilde{g}_{int}^{(l)}, \tilde{g}_1^{(l)}, \cdots, \tilde{g}_J^{(l)}\} = \mathrm{att}\left(\{\hat{g}_{int}^{(l)}, \hat{g}_1^{(l)}, \cdots, \hat{g}_J^{(l)}\}\right)$$

用户表示

用户邻居表示



推荐阅读：**BFS学术报告-基于Transformer的时间序列分析-李新帅**

- 文本-图交互 $\left(g_{int}^{(l)}, h_{int}^{(l)}\right) = \text{inter}\left(\tilde{g}_{int}^{(l)}, \tilde{h}_{int}^{(l)}\right)$
  - 相似性权重

$$w_{hh} = \tilde{h}_{int}^{(l)} \otimes \left(\theta_1 \cdot \tilde{h}_{int}^{(l)}\right)$$

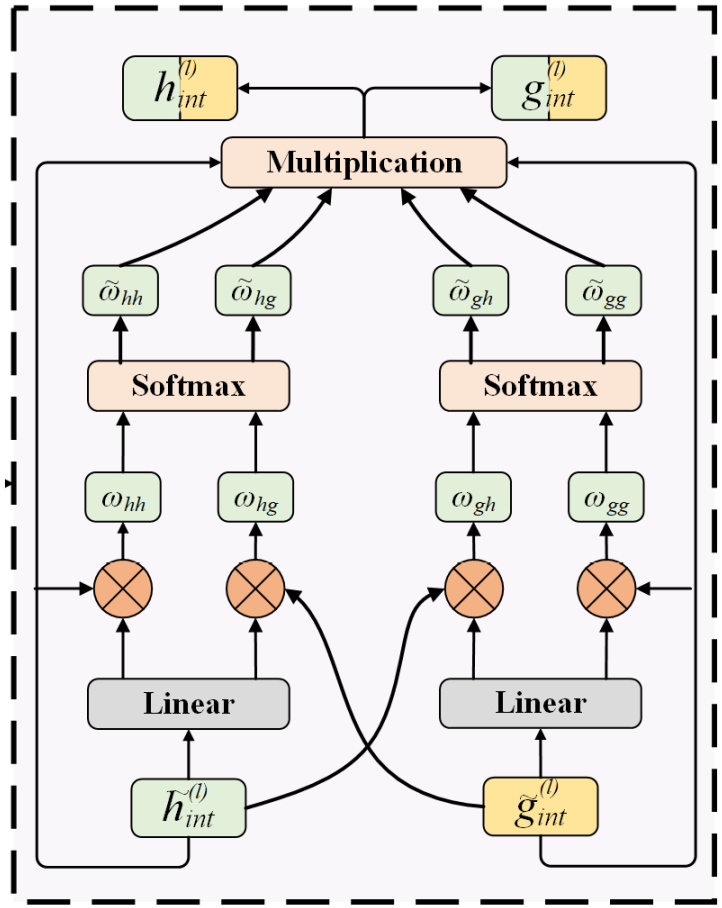
$$w_{hg} = \tilde{h}_{int}^{(l)} \otimes \left(\theta_2 \cdot \tilde{g}_{int}^{(l)}\right)$$

$$\tilde{w}_{hh}, \tilde{w}_{hg} = \text{softmax}\left(w_{hh}, w_{hg}\right)$$

  - 文本、图交互表示

$$h_{int}^{(l)} = \tilde{w}_{hh}\tilde{h}_{int}^{(l)} + \tilde{w}_{hg}\tilde{g}_{int}^{(l)}$$

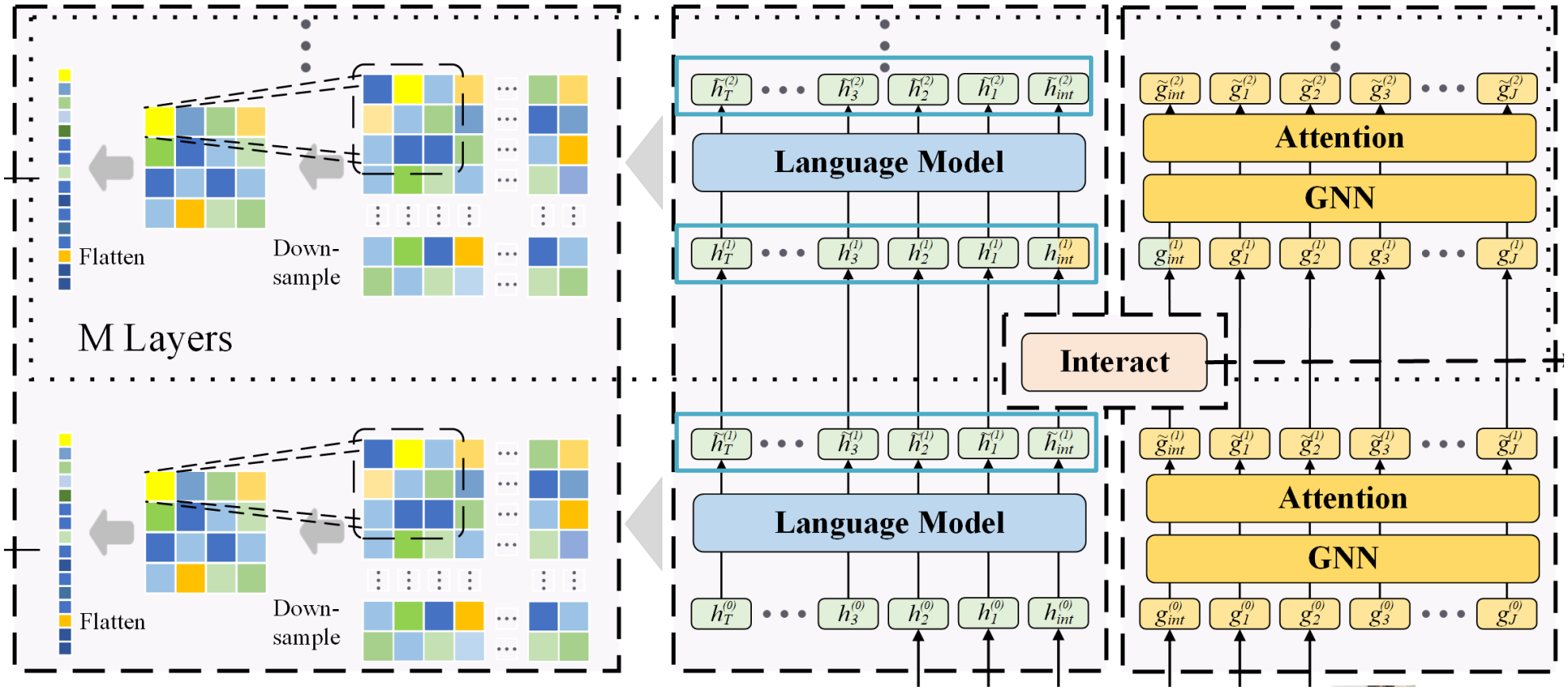
$$g_{int}^{(l)} = \tilde{w}_{gg}\tilde{g}_{int}^{(l)} + \tilde{w}_{gh}\tilde{h}_{int}^{(l)}$$



还有其他的交互方法吗?

- M次文本-图交互

M值越大越好吗？为什么？



BIC 算法原理

- 语义一致性建模

$$\{\tilde{h}_{int}^{(l)}, \tilde{h}_1^{(l)}, \cdots, \tilde{h}_T^{(l)}\} = \mathrm{LM}(\{h_{int}^{(l-1)}, h_1^{(l-1)}, \cdots, h_T^{(l-1)}\})$$

  - 下采样

输出Transformer的注意力矩阵

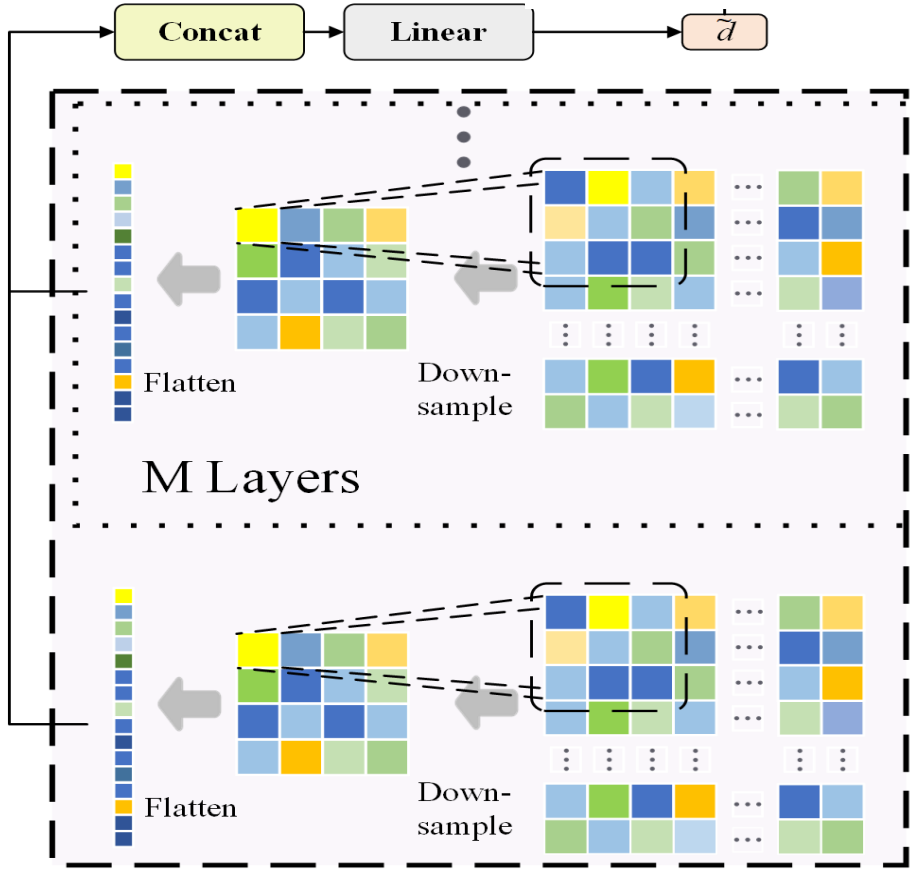
$$\tilde{\mathcal{M}}_i = \mathrm{sample}(\mathcal{M}_i), \tilde{\mathcal{M}}_i \in \mathbb{R}^{K \times K}$$

  - 最大池化

$$d_i = \theta_{sc} \cdot \mathrm{Flatten}\left(\tilde{\mathcal{M}}_i\right)$$

  - 聚合，得到语义一致性表示

$$d = \sigma(W_D \cdot \mathrm{aggr}(\{d_i\}_{i=1}^{M}) + b_D)$$

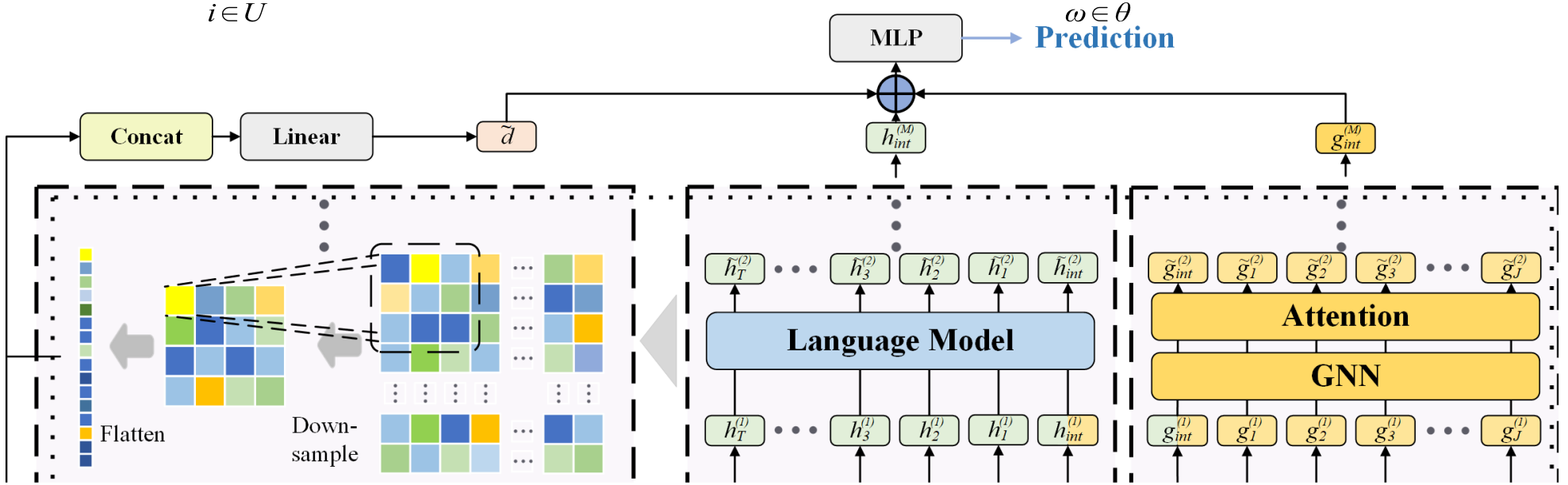


- 训练和推理
  - 用户表示

$$z = W_D \cdot \left(d | h_{int}^{(M)} | g_{int}^{(M)}\right) + b_D$$

  - 交叉熵损失

$$l = -\sum_{i \in U} \left[y_i \log\left(\hat{y}_i\right) + \left(1 - y_i\right) \log\left(1 - \hat{y}_i\right) \right] + \lambda \sum_{\omega \in \theta} \omega^2$$



- 实验设计
  - 数据集：两个数据集 **Crescie-15、TwiBot-20** 进行实验
  - 数据集信息统计

数据集	推特用户	推文	边
Crescie-15	**5,301**	**2,827,557**	**14220**
TwiBot-20	**229,580**	**33,488，192**	**33,716,171**

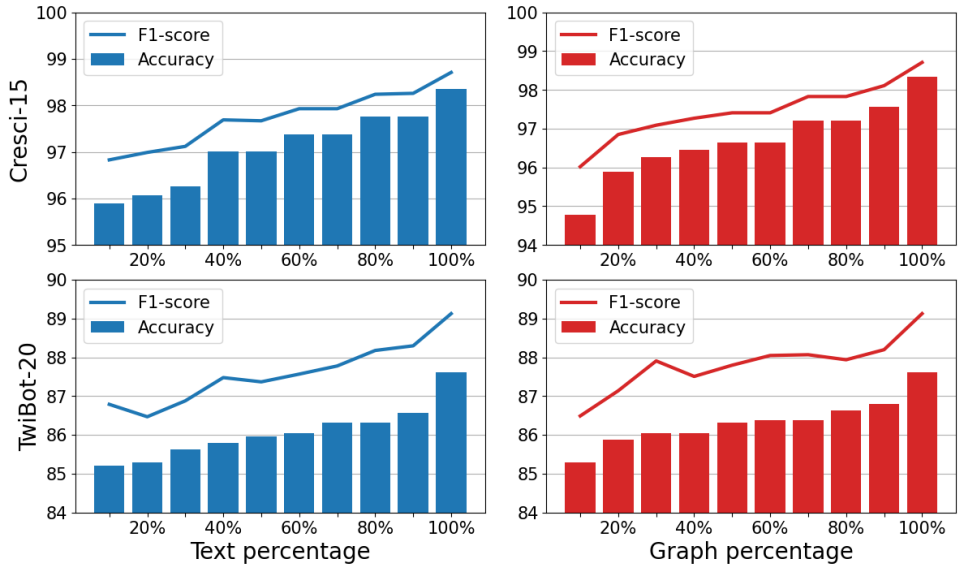
BIC 对比、消融实验

- 实验结果
  - 评价指标: **Accuracy, F1-score**
  - 在所有数据集上都优于基线，至少有1%的提升
  - 基于图模态的方法性能普遍比基于文本模态的方法要好

Method	Modalities			Cresci-15		TwiBot-20	
	Text	Graph	Modality-Int	Accuracy	F1-score	Accuracy	F1-score
Yang *et al.*				77.08 (±0.21)	77.91 (±0.11)	81.64 (±0.46)	84.89 (±0.42)
Botometer				57.92	66.90	53.09	55.13
Kudugunta *et al.*	✓			75.33 (±0.13)	75.74 (±0.16)	59.59 (±0.65)	47.26 (±1.35)
Wei *et al.*	✓			96.18 (±1.54)	82.65 (±2.47)	70.23 (±0.10)	53.61 (±0.10)
BotRGCN		✓		96.52 (±0.71)	97.30 (±0.53)	83.27 (±0.57)	85.26 (±0.38)
Alhossini *et al.*		✓		89.57 (±0.60)	92.17 (±0.36)	59.92 (±0.68)	72.09 (±0.54)
RGT		✓		97.15 (±0.32)	97.78 (±0.24)	86.57 (±0.41)	88.01 (±0.41)
SATAR	✓	✓		93.42 (±0.48)	95.05 (±0.34)	84.02 (±0.85)	86.07 (±0.70)
BIC w/o Graph	✓			97.16 (±0.58)	97.80 (±0.46)	85.44 (±0.32)	86.97 (±0.41)
BIC w/o Text		✓		96.86 (±0.52)	97.57 (±0.39)	85.78 (±0.48)	87.25 (±0.57)
BIC	✓	✓	✓	**98.35** (±0.24)	**98.71** (±0.18)	**87.61** (±0.21)	**89.13** (±0.15)

- 文本-图交互探究
  - 去掉图模态，实现了次优性能，体现了文本特征与语义一致性表示结合的优势
  - 模型可以在模态信息占比较少(30%-40%)的时候保持性能，说明了交互模块在跨模态交换信息方面的有效性

Method	Cresci-15		TwiBot-20	
	Accuracy	F1-score	Accuracy	F1-score
Yang *et al.*	77.08 (±0.21)	77.91 (±0.11)	81.64 (±0.46)	84.89 (±0.42)
Botometer	57.92	66.90	53.09	55.13
Kudugunta *et al.*	75.33 (±0.13)	75.74 (±0.16)	59.59 (±0.65)	47.26 (±1.35)
Wei *et al.*	96.18 (±1.54)	82.65 (±2.47)	70.23 (±0.10)	53.61 (±0.10)
BotRGCN	96.52 (±0.71)	97.30 (±0.53)	83.27 (±0.57)	85.26 (±0.38)
Alhossini *et al.*	89.57 (±0.60)	92.17 (±0.36)	59.92 (±0.68)	72.09 (±0.54)
RGT	97.15 (±0.32)	97.78 (±0.24)	86.57 (±0.41)	88.01 (±0.41)
SATAR	93.42 (±0.48)	95.05 (±0.34)	84.02 (±0.85)	86.07 (±0.70)
BIC w/o Graph	97.16 (±0.58)	97.80 (±0.46)	85.44 (±0.32)	86.97 (±0.41)
BIC w/o Text	96.86 (±0.52)	97.57 (±0.39)	85.78 (±0.48)	87.25 (±0.57)
BIC	**98.35** (±0.24)	**98.71** (±0.18)	**87.61** (±0.21)	**89.13** (±0.15)

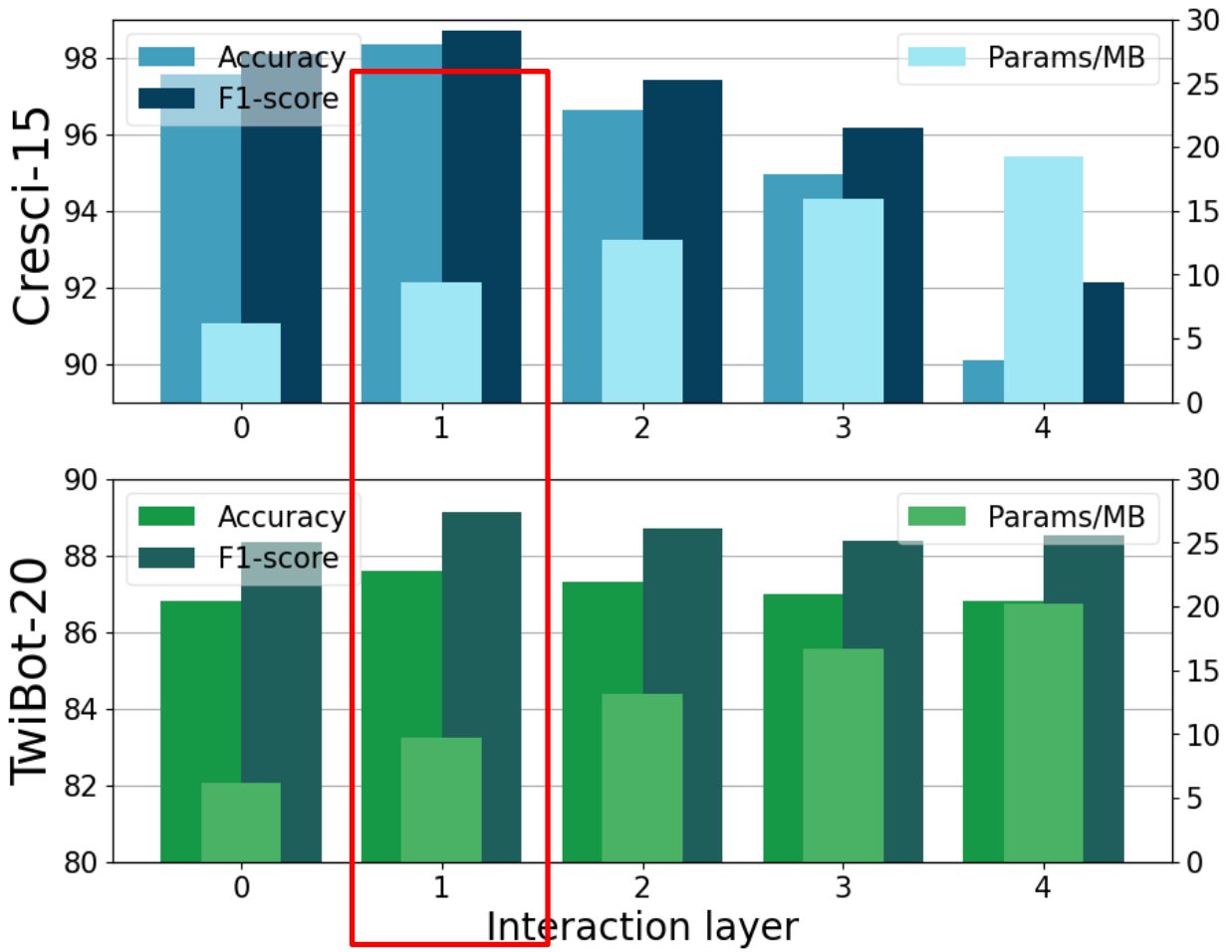


- 函数探究
  - 函数种类
    - 基于相似度（本文方法）
    - **Hard**：平均文本、图表示
    - **Soft**：利用可学习参数为文本、图表示加权
    - **MLP**：拼接文本、图表示并输入MLP层
    - **Text**：将文本表示输入线性层
    - **Graph**：将图表示输入线性层
  - 结果
    - 大部分交互函数优于无交互方法
      - 证明了多模态交互的必要性

Function	Cresci-15		TwiBot-20	
	Accuracy	F1-score	Accuracy	F1-score
Ours	**98.35**	**98.71**	**87.61**	**89.13**
w/o interaction	95.89	96.85	85.97	87.42
Hard	96.64	97.41	86.64	88.15
Soft	97.01	97.69	87.06	88.27
MLP	97.38	97.97	86.98	88.44
Text	96.64	97.41	85.63	87.14
Graph	96.45	97.27	86.30	87.65

BIC 文本-图交互探究

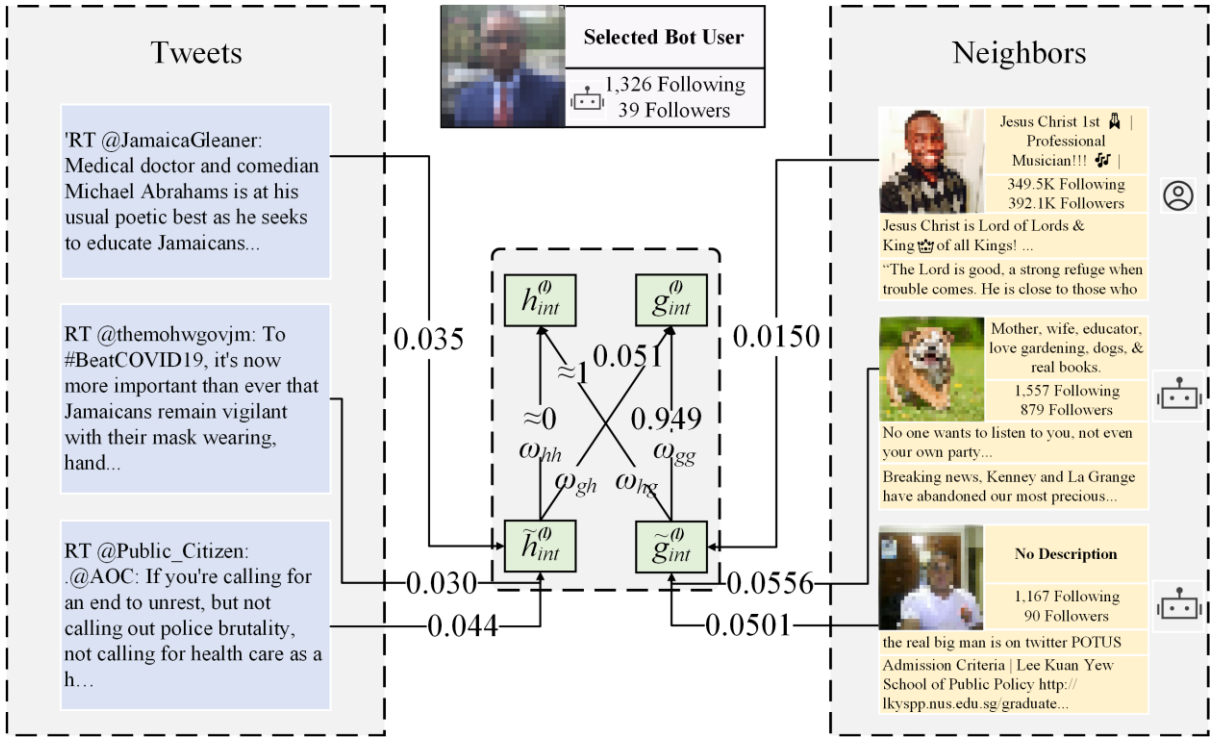
- 交互数目探究
  - 交互次数=1时，BIC性能最好
  - 随着交互次数增加，性能逐渐下降，可能是由于复杂度提高增加了训练难度
  - 交互次数=0时，模型无法捕获文本、图之间的关联性，导致性能不佳



- 案例探究
  - 目的
    - 通过实例分析相似性权重分配的合理性
  - 过程
    - 选出三个注意权重最高的推文和邻居，发现邻居的注意权重高于推文，判断邻居表示对于检测更重要
    - 通过提取文本-图的相似权重，得到的特征表示确实强调了邻居的作用，说明了交互的有效性

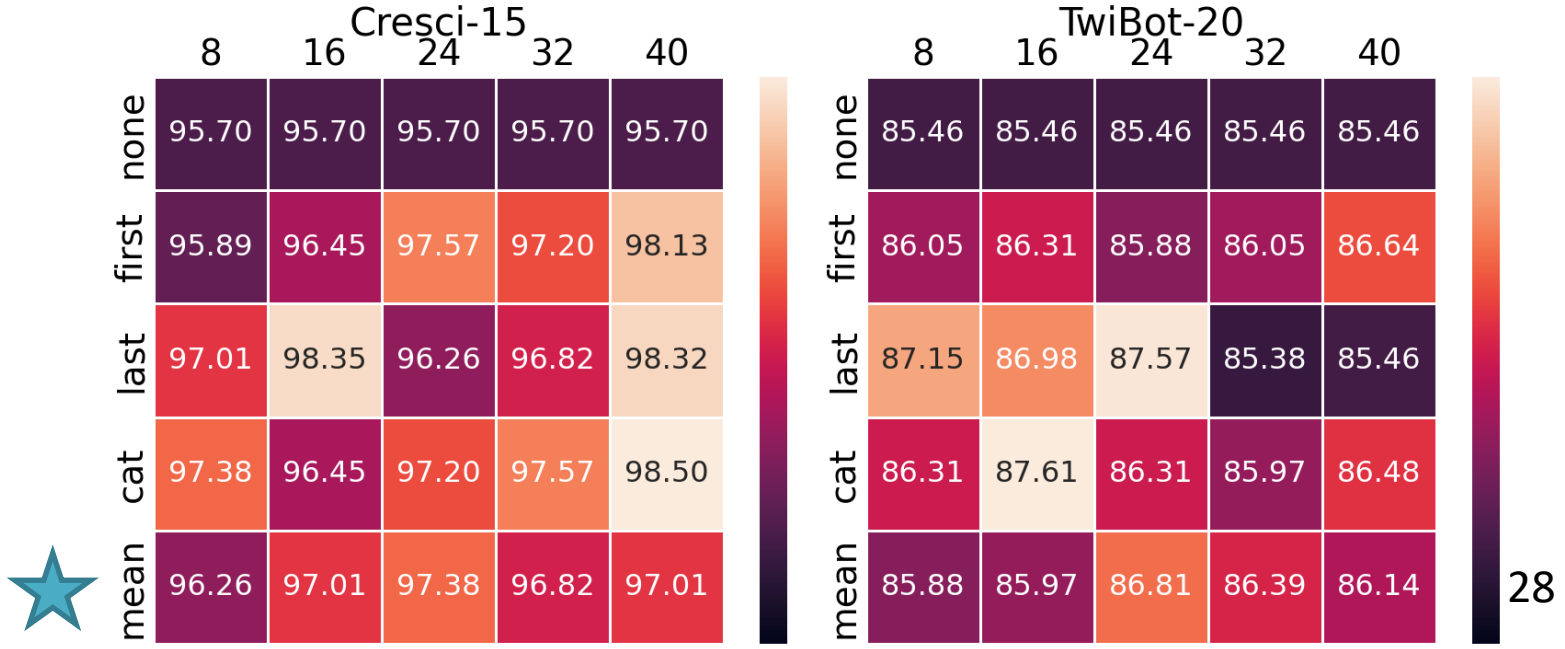
$$h_{int}^{(l)} = 0\tilde{h}_{int}^{(l)} + 1\tilde{g}_{int}^{(l)}$$

$$g_{int}^{(l)} = 0.949\tilde{g}_{int}^{(l)} + 0.051\tilde{h}_{int}^{(l)}$$



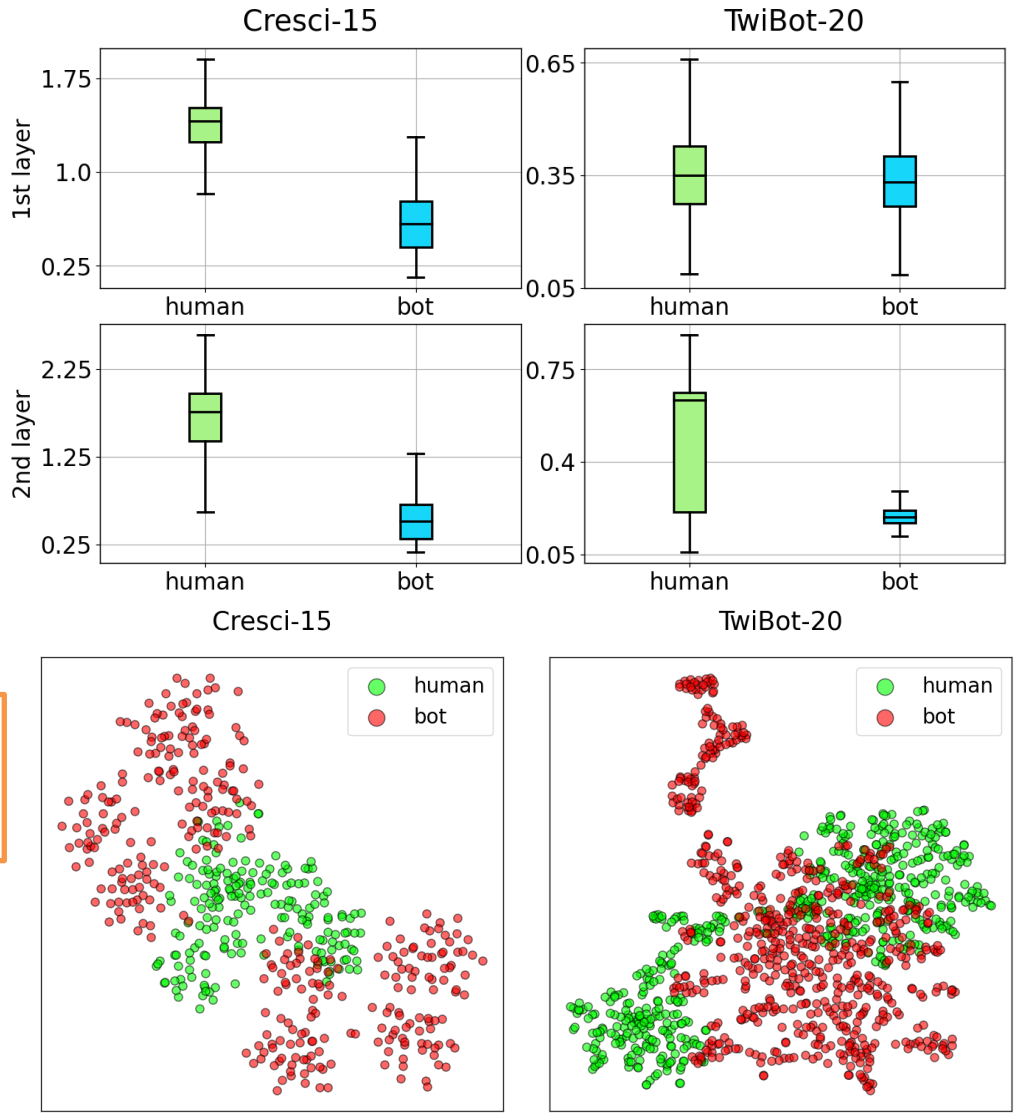
BIC 语义一致性探究

- 函数探究
  - 函数种类
    - **none**: 不使用语义一致性表示
    - **first**: 仅使用第一次交互的语义一致性表示
    - **last**:仅使用最后一次交互的语义一致性表示
    - **cat**: 拼接N次交互的语义一致性表示
    - **mean**:平均N次交互的语义一致性表示
  - 结果
    - 表明**mean**的有效性



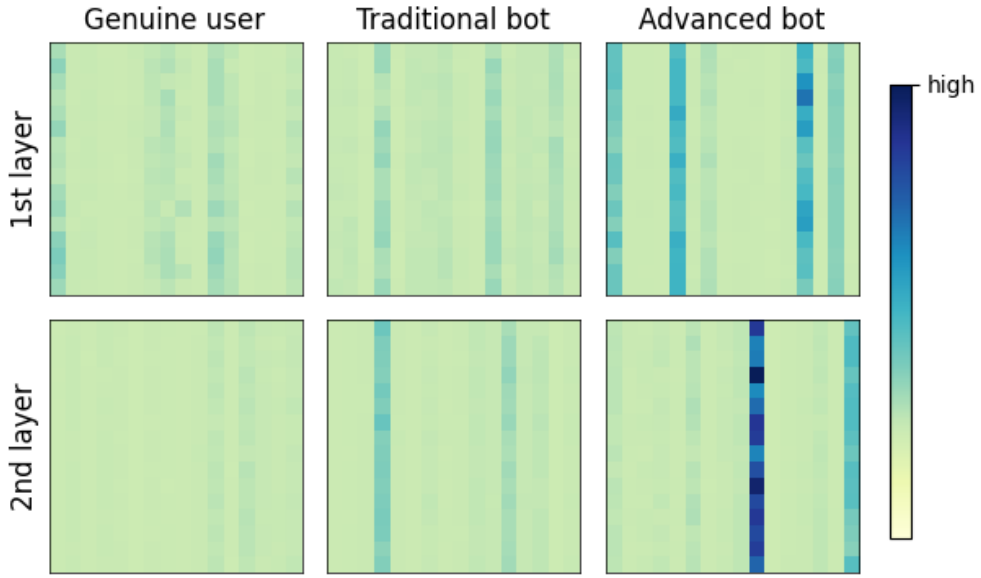
BIC 语义一致性探究

- 辨别能力探究
  - 基于语义一致性矩阵的最大特征值绘制箱型图
  - 基于语义一致性表示进行k-means聚类，并计算V-measure（综合考虑聚类结果的一致性以及完整性）
    - 一致性：每个簇样本是否属于同一类
    - 完整性：同一类别是否被聚到同一簇
    - **TwiBot-20数据集上的V-measure为0.3336**
    - **Cresci-15数据集上的V-measure为0.4312**



BIC 语义一致性探究

- 案例探究
  - 分析高级机器人、传统机器人、真实用户的代表性推文
  - 将语义一致性矩阵可视化
  - 结果表明高级机器人的推文语义呈现更大的不一致性



4 day only end of March 3/28-3/31 event .99 cent sale Teen Epic Fantasy #Kindle Bestseller #Everville #TheFirstPillar ... Baby you deserve everything you want its your night. A man in red is never forgotten...Malan Breton HOMME	Novel Bot
Mom dropped her iPhone on the tile. And then she says that my hands are from the wrong place, yeah #cheap iPhone5 killed an iPhone in 2 weeks #typicalMasha" Scratched the whole top #cheapiPhone5 Utkin has to scroll his iPhone for half an hour to show his photo to the girls. #cheapiphone5	Traditional Bot
@fnzcuvccra basic play structure: Team has four plays ('downs') to move the ball at least 10 yards closer to their opponent's endzone. (I just signed up for that ride I RTd. I hope me, my stiff frame, and skinny tyres can take all the cobble stones.) @felix_cohen jebus. Glad I'm not in that coworking space.	Genuine User

- 高级机器人探究
  - 评估方法在2020年之后的高级机器人，使用Twitter crawl构建新数据集，包括5000个人类账户和5000个机器人
  - 实验结果证明
    - 方法可以更好的捕获高级机器人
    - 高级机器人有着更强的隐蔽性，使得检测性能大幅下滑

Method	Accuracy	F1-score
Botometer	55.35	53.99
RGT	66.95	64.48
BIC	**67.25**	**67.78**

Method	Modalities			Cresci-15		TwiBot-20	
	Text	Graph	Modality-Int	Accuracy	F1-score	Accuracy	F1-score
Botometer				57.92	66.90	53.09	55.13
RGT		✓		97.15 (±0.32)	97.78 (±0.24)	86.57 (±0.41)	88.01 (±0.41)
BIC	✓	✓	✓	**98.35** (±0.24)	**98.71** (±0.18)	**87.61** (±0.21)	**89.13** (±0.15)

优劣分析

- 优势
  - 通过文本-图模态交互，获取关联、互补信息
  - 至少**54.8%(51/93)**躲避仅基于文本模态/仅基于图模态检测的机器人，可以被**BIC**检测
  - 相较于其他方法，对高级机器人检测效果较好
- 劣势
  - 会将“有个性”的用户识别为机器人
  - 完整多模态数据集收集困难，不利于实际应用

研究方向

研究方向

- 研究方向
  - 数据挖掘层面
    - 不同话题下机器人行为特点分析
    - 高级机器人特点分析
    - 针对单一类型的机器人进行检测，如：垃圾邮件机器人
    - 现有方法**Deepfakes**可以模糊真实账号的界限，进行特征区分度增强
  - 模型层面
    - 词嵌入技术增强
    - 文本、图模态特征提取框架改进

参考文献

[1] Ellaky Z, Benabbou F, Ouahabi S. Systematic Literature Review of Social Media Bots Detection Systems[J]. Journal of King Saud University-Computer and Information Sciences, 2023. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In ICLR, 2022.

[2] Lei Z, Wan H, Zhang W, et al. Bic: Twitter bot detection with text-graph interaction and semantic consistency[J]. arXiv preprint arXiv:2208.08320, 2022.

[3] Feng S, Wan H, Wang N, et al. BotRGCN: Twitter bot detection with relational graph convolutional networks[C]//Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. 2021: 236-239.

谢谢！

大成若缺，其用不弊。大盈若冲，其用不穷。大直若屈。大巧若拙。大辩若讷。静胜躁，寒胜热。清静为天下正。