学号：__2016300030093__

密级：_____

武汉大学本科毕业论文

Detecting Rumor on Social Media Using Graph Neural Networks

基于图神经网络的社交媒体谣言检测算法研究

院(系)名 称 ： 弘毅学堂

专 业 名 称 ： 计算机科学与技术

学 生 姓 名 ： 周稚璇

指 导 教 师 ： 李晨亮 副教授

二〇二〇年五月

郑 重 声 明

本人呈交的学位论文，是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料真实可靠。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确的方式标明。本学位论文的知识产权归属于培养单位。

本人签名: _____ 日期: _____

摘　　要

社交媒体上的谣言对舆论、用户体验造成极大的不良影响，已经引起学术界的广泛重视。

传统的谣言检测方法是基于内容的，根据文本、用户行为等特征进行分类。但由于学习到的是相对浅层的特征，这类方法往往只能适用于某个特定的情境，泛化性能差，且易受对抗样本的攻击。我们的实验表明，攻击者可以通过事实篡改、主谓颠倒、因果混淆等方法成功逃过基于文本特征的模型的检测。

近期的研究考虑了谣言传播的网络结构，试图通过图神经网络，学习谣言传播网络的图结构特征，取得了更优的效果。我们讨论了基于图卷积网络（GCN）的谣言检测方法，并考察了它的性能。我们制造对抗样本对它进行攻击，一定程度上降低了其分类准确率。具体来说，在我们的攻击下，谣言检测模型的准确率从 88.54% 降低到 84.37%，但它的总体鲁棒性明显优于基于语言特征的检测方法。我们还通过对抗训练进行防御，让模型在噪声数据下也保持较高性能。对抗训练后，模型面对对抗样本可以保持 85.28% 的准确率。

由于对抗训练会被更强的攻击打败，且计算开销大，提出更好的防御方法（如 certified defense）对 GCN 提供可证明的鲁棒性是我们未来的工作。另一方面，由于基于 GCN 的谣言检测准确率仍不太高（低于 90%），我们还会研究谣言传播中的人因，将获得的领域知识加入到模型中，更好地解决这一问题。

关键词：谣言检测；图卷积网络；对抗样本；假节点攻击

ABSTRACT

Rumor on social media has toxic impact on public opinion and user experience, and has attracted much attention from the research community.

Traditional content-based methods detect rumor based on linguistic analysis and user behavior mining. Since they capture only shallow features, they are restricted to certain scenarios and show poor generalization performance. At the same time, they are vulnerable to adversarial examples. Our experiment demonstrates that adversaries can easily escape detection of models based on linguistic analysis by fact tampering, subject-object exchange and cause confounding.

More recent works consider network structures of rumor propagation, trying to capture structural information of rumor propagation networks using graph neural networks. We evaluate rumor detection models based on graph convolutional networks (GCN) with adversarial attacks, and subvert them effectively. Specifically, accuracy of rumor detection degrades from 88.54% to 84.37% under our attack. Generally, GCN-based rumor detector shows improved robustness compared to that based on linguistic analysis. We also use adversarial training to guarantee performance of GCN-based rumor detection in the presence of noise. After adversarial training, the robust accuracy of the model can be 85.28% in the presence of adversarial examples.

Since adversarial training can be defeated by stronger, iterative attacks, and shows a large computational overhead, designing better defense mechanisms (i.e. certified defense) to provide provable robustness will be our future work. In addition, since accuracy of GCN-based rumor detector is far from satisfactory (<90%), we will examine human factors in rumor spread, and use the extracted domain knowledge to improve model performance, better mitigating rumor on social media.

Key words: Rumor detection; Graph convolutional network; Adversarial examples; Fake node attack

目　　录

1 绪论

社交媒体的广泛使用一方面便利着人们的生活，一方面对信息监管提出了极大的挑战。随着用户自生产内容（UGC）模式的兴起，如微博、抖音，社交媒体上出现更多的谣言，对网络环境、用户体验和社会价值造成恶劣的影响。为了改变这一现状，各大社交平台纷纷投入大量的人力进行辟谣，如“微信辟谣助手”小程序，通过人工标注谣言，并向阅读或分享过谣言文章的用户发送提醒，用户也可以主动搜索谣言。这样的谣言检测虽然有效，但需要耗费大量的人力、财力，严重依赖标注者的经验，效率也较低。

谣言（rumor）是假消息（misinformation）的一种，其他假消息还包括假情报（disinformation）、都市传说（urban legend）、垃圾信息（spam）和引战言论（troll）。近年来，谣言和其他假消息的自动检测是学术研究的一个热点 [1–4]。谣言检测方法可以分为两类。

第一类方法是基于内容的谣言检测，根据文本、用户行为等特征判断是否为谣言。此前的研究中，我们已经证明这种方法易受对抗攻击的干扰 [5]，并且无法检测出谣言源头、及早切断谣言传播的路径。我们通过事实篡改、主谓颠倒、因果混淆等方法成功逃过基于文本特征的模型的检测。基于文本的谣言检测还容易将真实信息误判为谣言，对社交媒体用户的积极性和体验造成负面影响。

第二类方法则考虑了网络结构。由于谣言传播模式很多情况下是未知的，[6] 首先在未知传播模式的情况下检测谣言源头。[7] 则用 GCN 完成这一任务，取得了更好的效果，模型输入是一个无向图，输出是谣言源头的集合。[8] 使用双向 GCN 检测谣言，同时考虑谣言的时序传播（propagation）和谣言的空间散布（dispersion），并将源头贴子的信息加入到 GCN 的每一层中，以突出谣言源头对全局的影响。

由于 GCN 易受对抗样本的攻击 [9, 10]，我们通过攻击验证基于 GCN 的谣言检测模型的鲁棒性。本文中，我们设计了几种不易被发现的假节点攻击，包括删除转发帖的评论、多源头谣言传播、增加文字长度，在不改变图结构、保证局部相似性的前提下加入假节点或改变节点特征，影响谣言分类的准确率。在我们的攻击下，谣言检测模型的准确率从 88.54% 降低到 84.37%。为了应对假节点攻击，我们进行对抗训练，使模型在噪声环境下也能够保持较高的准确率。对抗训练后，模型面对对抗样本可以保持 85.28% 的准确率。

文章的主要贡献如下：

- 提出三类对抗样本，攻击基于文本特征的谣言检测模型，逃过它的检测。
- 提出三类针对 GCN 谣言检测模型的假节点对抗样本，通过加入不易被察觉的扰动，降低 GCN 的节点分类准确率，并提出相应的防御方法。
- 讨论了未来谣言检测准确性、鲁棒性研究的方向。

文章结构安排如下：第二章介绍了谣言检测、对抗样本攻击和防御、图神经网络的相关工作；第三章用对抗样本攻击基于文本特征的谣言检测模型，指出该类方法的局限性；第四章介绍了基于 GCN 的谣言检测模型，作为我们的攻击目标；第五章提出真实世界中不易被察觉的假节点攻击，以降低基于图网络的谣言检测模型的节点分类准确率，并提出相应的防御方法，讨论了 GCN 鲁棒性、谣言传播中的人因等相关问题；第六章总结了我们的结论。

2 相关工作

2.1 谣言检测

谣言检测方法可以粗略分为两类。一类是基于内容的检测，即根据语言、用户信息等特征分类。我们主要介绍基于语言的谣言检测。另一类考虑谣言传播特征，根据谣言传播网络的结构检测谣言。

基于语言的谣言检测根据语法特征、特定单词出现次数等信息判别谣言。[11] 研究了自动检测标题党的方法，同时学习文字特征和非文字特征（如图片、用户行为）。[12] 采用朴素贝叶斯分类器，通过简单模型取得较好的分类效果。[13] 分析了真假新闻在标题特征、复杂度、内容风格上的差异，并用详尽可能性模型（Elaboration Likelihood Model，ELM）解释谣言传播和谣言说服力。更多关于此类研究的概括在我们之前的工作中 [5]

关于谣言源头检测的早期研究提出了谣言中心化（rumor centrality）、源头显著性（source prominence）等理论 [14]，为之后的研究奠定了理论基础。谣言中心化指靠近谣言源头的节点更可能被感染，源头显著性指被大量感染节点包围的节点更可能是谣言源头。

[15] 中提出了获得 IC（Independent Cascade）传播模型下前 K 个源头的算法。IC 模型是一种影响模型，每一次迭代中，每个节点以一定概率激活其邻居。

[16] 中提出了在 SI（Susceptible- Infected）传播模型下自动搜索多个源头的方法。[17] 研究了 SIR（Susceptible-Infected-Recovery）传播模型下的谣言源头检测问题。SI 和 SIR 同属于感染模型，是对流行病传播的一种模拟，可疑节点以一定概率被感染，而感染节点以一定概率恢复。[18] 中将节点感染时间看做一个序列，将时间窗口引入检测模型。

影响模型和感染模型是两种主要的谣言传播模型。但事实上，传播模型很多情况下是未知的。[19] 在未知传播模型的情况下检测谣言源头，该方法基于半监督的标签传播（label propagation），用已标记节点的标签信息预测未标记节点的标签信息，利用样本间的关系建立完全图模型。。

随着图神经网络，尤其是图卷积神经网络（GCN）在各种图任务上取得优于其他模型的效果 [20]，[7] 率先使用 GCN 进行谣言源头检测：先前的算法中节点标签仅仅是一个整数，限制了预测准确率；[7] 则使用深度学习模型，用多级邻居

的信息构建节点表示，从而提升了准确率。

2.2 对抗样本攻击和防御

Szegedy 等人最早提出对抗样本的概念 [21]，通过给图像加上人类不可察觉的像素扰动使分类器出现错误（如将熊猫误判为卡车），攻击者的优化目标是达到攻击目的的同时最小化加入的扰动，或在加入一定限度内噪声的同时最大化攻击效果。FGSM（Fast Gradient Sign Method）攻击 [22] 基于梯度生成对抗样本，最大化分类误差。C&W 攻击 [23] 以多次迭代的方式寻找最小扰动。FGSM 和 C&W 是最为经典、效果较好的两种图像攻击方法，由于知道模型的所有信息（包括分类结果和梯度更新信息），因而属于白盒攻击（white-box attack）。另一种攻击是黑盒攻击（black-box attack），用户只知道部分样本的分类结果，并根据这部分样本的分类结果产生其他对抗样本。

常见的防御方法可分为两类：对抗训练和 certified defense（有理论保证的防御）。对抗训练将对抗样本包括在训练集中，让模型直接学习如何分类对抗样本，达到数据增强的效果 [24, 25]。这种防御往往很快就会被更强、更有针对性的攻击打败，形成一种“军备竞赛”。此外，寻找强大的对抗样本的过程是非常耗时的，因此效率不高。

Certified defense 可以提供有保证的防御，最早由 Percy Liang 等人提出 [26]。它的优化目标可以在理论上证明：只要扰动在一定范围内，分类器的预测就不会发生变化，因此对于噪声不会过分敏感。Certified defense 是目前较优的防御方法，但也存在计算复杂度高、难以扩展到大数据集和任意模型的局限性。因此在现阶段攻击是相对容易的。我们迫切需要更强大、更高效、更通用的防御算法，保证机器学习系统的安全性。

2.3 图神经网络和攻击

图神经网络应用于图数据，相比于传统神经网络有较大优势。它通过节点之间的信息传递来捕捉图的依赖关系，能更好地学习拓扑图的空间特征 [27]。[28] 最先提出 GCN，提出一种简单的逐层传播方式，并用于半监督节点分类任务，取得了较高的准确率。[29] 等工作进一步改进了 GCN，使其效率、准确率更高。图神经网络被广泛应用在各种场景中，结构化场景如社交网络、推荐系统、物理系统、

化学分子预测、知识图谱等领域，非结构化场景如图像和文本。图神经网络还可以用来解决组合优化问题。

由于 GCN 的图卷积特性，即通过邻节点特征学习节点特征，攻击者可以通过修改邻域中的连接和特征改变对某个或某组节点的分类结果。[9] 通过改变很少的边和特征，将 GCN 的准确率降低到随机预测的水平。[10] 用梯度上升、遗传算法和强化学习改变图结构，达到攻击的目的。

和针对图像分类的攻击如 FGSM[22]、PGD 攻击 [30] 不同，针对节点分类的攻击是一个离散优化问题（因为输入是离散的），因而更难防御。这和自然语言处理中的攻击类似，如通过删除重要单词 [31]、替换和插入有拼写错误的单词 [32] 改变分类器效果，攻击者往往可以通过非常简单、直觉的攻击严重影响模型性能。

3 基于文本特征的谣言检测算法的局限性

基于文本特征的谣言检测通过学习文字模式和统计相关性找出谣言。我们此前的研究 [5] 指出，这类谣言检测模型学习到的特征是浅层的，它仅仅判断文章和帖子是否遵循某种“标准”的范式或风格。这导致了两个主要缺陷。

一方面，模型只能检测出看上去明显写得不好的文章和帖子，如标题和内容不一致（标题党），又如包含典型的“谣言词”（“political”、“Trump”，等等）。如果谣言散布者进行更微妙的攻击，比如事实篡改、逻辑混淆，模型则完全无法识别。我们设计了三种不同的攻击，验证这一结论。

另一方面，依赖浅层特征的学习会导致大量假阳（false positive）情况的出现。例如，包含政治词汇的帖子容易被误判为谣言。社交媒体（如 Twitter①、微博②）和开放新闻平台（如今日头条③）上多为 UGC（User Generated Content，用户原创内容），这些内容会因为“写得不好”、没有遵循专业新闻的写作范式而被标志为谣言（如果模型是在专业新闻数据集上训练的话），严重影响了用户生产内容的积极性。我们通过实验表明，被误判为谣言的内容确实属于这些情形。

3.1 测试 Fakebox（没有对抗攻击的情形）

Fakebox[33] 号称实现了 95% 的分类准确率。它主要分析以下几个特征：

- 标题是否为标题党；
- 内容是否像“真新闻”；
- 域名是否经常传播谣言。

我们使用 McIntire 真假新闻数据集对它进行测试。数据集中共有 6,335 篇文章或帖子，其中 3,171 篇标签为真，3,164 篇标签为假，比例大致为 1:1。

根据表3.1所示的实验结果，我们可以发现，对于假新闻，模型的分类准确率高达 80.26%（对于这种情况，我们将在后文介绍三种对抗样本，让假新闻逃过模型检测）。

①https://twitter.com/
②https://weibo.com/
③https://www.toutiao.com/

而对于真新闻，模型的分类准确率只有 43.97%，这意味着超过一半的真新闻被误判为假。我们观察到，在假阳的情形中，出现了很多典型“谣言词”，如“anti”、“terror”、“Islamism”、“Trump”。同时，这些测试样本多为社交媒体上的帖子，没有遵循专业的新闻表达，有很多样本中出现了拼写错误，因此有更大的概率被分类为假新闻。

表 3.1　Fakebox 在 McIntire 真假新闻数据集上的效果

News type	Number of articles	Correctly classified	Classification accuracy
Real	2,636	1,159	43.97%
Fake	2,721	2,184	80.26%
Total	5,357	3,343	62.40%

3.2　测试 Fakebox（三种对抗攻击）

我们设计了三种对抗样本，使得假新闻成功逃过模型的检测：事实篡改（fact distortion）、主谓颠倒（subject-object exchange）和因果混淆（cause confounding）。事实篡改只需要简单替换原文本中的人物、地点、动作；主谓颠倒本质上也是一种对事实的修改，基于文本特征的模型难以检测；因果混淆可以在两个不相关的事件间建立因果关系。对抗样本的例子在表3.2中。

初步观察表明，Fakebox 无法检测出假新闻对抗样本，如对于因果混淆的样本，两个独立事件的真实值得分（veracity score）一高一低，则连在一起后的文本真实值得分介于两者之间，这说明 Fakebox 仅仅从文本角度学习谣言特征，无法抵抗基于事实、逻辑的攻击。

表 3.2　三种假新闻对抗样本

攻击类型	原来样本	对抗样本
事实篡改	**12** people were **injured** in the shooting.	**24** people were **killed** in the shooting.
主谓颠倒	A **gangster** was shot by the **police**.	A **policeman** was shot by the **gangster**.
因果混淆	The condom policy originated in 1992 …The Boy Scouts have decided to accept people who identify as gay and lesbian. (**two unrelated events**)	The inclusion of gays, lesbians and girls in the Boy Scouts **led to** the condom policy.

3.3 设计启发

我们的初步试验和定性分析表明，基于语言特征的谣言检测无法抵御事实、逻辑层面的攻击。这些知识需要由人提取并融入模型中。一种可行的方法是由用户以合作的方式对谣言进行标注，标注中蕴含着人的事实、逻辑判断，可以帮助模型更好地分类。如何去除标注信息中的噪声、如何将人工标注融合进模型都是潜在的研究话题。我们将这些作为未来的工作。

4 基于 GCN 的谣言检测算法

我们已经通过实验证明，基于文本特征的谣言检测模型易受对抗样本干扰，攻击者通过逻辑、事实层面的篡改，生成的谣言可以成功逃过模型的检测。因此除了直接的特征，还需要学习更抽象、高层的表示，如谣言传播网络的结构特征。这一目标可以通过使用 GCN 实现。近期的两个工作 [7, 8] 开始使用 GCN 进行谣言检测，取得了较优的效果。但我们认为，由于 GCN 本身易受对抗样本攻击，基于 GCN 的谣言检测模型也可能存在安全风险。

我们首先介绍这两个模型，作为我们鲁棒性测试的目标，并在下一章中进行攻击。

4.1 基于 GCN 的谣言源头检测

[7] 用图卷积网络解决多谣言源头检测问题，取得了较好效果。我们首先介绍这一模型，并将其作为攻击目标。

4.1.1 问题定义

给定无向社交网络 $\mathcal{G} = (V, E, Y)$，V 是节点集合，E 是边的集合，$Y = \{Y_1, \ldots, Y_{|V|}\}$ 是网络中所有节点的感染状态。$Y_i = 1$ 表示节点 v_i 被感染，$Y_i = -1$ 表示节点 v_i 没有被感染。$R^* \subset V$ 是实际的谣言源头集合，模型的目标是找到一个标签函数 $l : V \to \{1, 0\}$，用这个标签函数找到谣言源头，以最大化公式（4.1）：

$$\frac{|R^* \cap R|}{|R^* \cup R|} \quad \text{with} \quad R = \{x \in V | l(x) = 1\}, \tag{4.1}$$

其中 R 是模型的预测结果。

4.1.2 图卷积网络（GCN）

GCN 是卷积神经网络（CNN）在图数据上的扩展，对于特征之间没有空间位置关系的数据（非欧几里得数据），能够通过学习节点邻域的特征获得图的特征。GCN 结构如图4.1。层间传播由公式（4.2）定义：

$$H^{-1} = \sigma\left(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right), \tag{4.2}$$

其中 I 是单位矩阵，A 是 $\mathcal{G}$ 的邻接矩阵，$\widetilde{A}=A+I$，$\widetilde{D}$ 代表 $\mathcal{G}$ 的拉普拉斯矩阵。$W^{(l)}$ 是可训练的权值矩阵，$\sigma(\cdot)$ 是激活函数。$H^{(l)}\in\mathbb{R}^{N*D}$ 第 l 层隐状态的矩阵，$H^{(0)}=X$ 是第一层的输入。

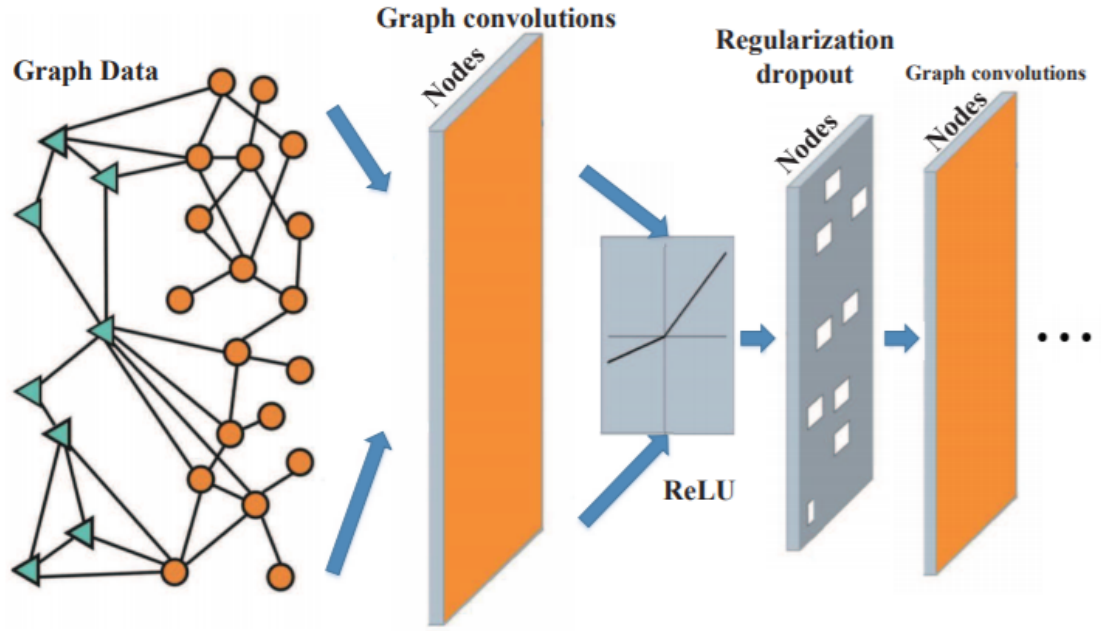

图 4.1　图卷积网络结构 [7]

4.1.3　谣言源头检测算法

根据谣言中心化（rumor centrality）和源头显著性（source prominence），[7] 考虑节点的邻居信息以提升源头检测的效果。GCN 的一层只能捕捉一阶邻居的信息，因此模型叠加了多个 GCN 层，具体层数取决于有多少阶邻居。

模型首先由图 $\mathcal{G}$ 产生拉普拉斯矩阵，然后产生训练样本。每个训练样本通过多个 GCN 层和一个全连接层，GCN 层的激活函数是 ReLU，全连接层的激活函数是 Sigmoid。最终输出预测结果 y。

模型的训练样本/输入根据传播模型产生，除了考虑各节点的感染状态，还将源头显著性、谣言中心化作为特征。

GCN 遵从半监督学习范式，而谣言检测是一个监督学习任务，[7] 因此修改了模型以适应该任务的输入和输出。对于半监督 GCN，输入一部分是带标签的数据，一部分是不带标签的数据，训练过程中，不带标签的样本根据邻居信息被打上标签，损失函数只计算带标签节点的损失，训练结束后，不带标签的节点输出新标

签用于预测和评估。对于谣言检测，输入是网络的感染状态，模型根据一批数据的损失更新权值矩阵，当训练完所有数据且达到收敛，模型则可以用于预测。

基于 GCN 的谣言源头检测取得了优于其他模型的效果。

4.1.4 潜在问题

基于 GCN 的节点分类对于捕捉图的拓扑结构特征有天然的优势。但在噪声（noise）或扰动（perturbation）下，模型容易作出错误的判断。之后，我们将介绍一种假节点攻击，在“不被注意”的前提下影响模型的性能。

4.2 基于双向 GCN 的谣言检测

将深度学习模型如 LSTM 应用于谣言检测，可以捕捉时序的谣言传播特征，但无法捕捉谣言散布过程中的全局图结构特征。而这两个特征可以被 GCN 有效学习。[8] 进一步设计了有向的 GCN，将谣言传播的方向考虑在内。这样，谣言在关系链上传播的时序特征（自顶向下的 GCN）和在社区中散步的空间结构特征（自底向上的 GCN）均被模型学习到，提高了检测的准确率。

4.2.1 谣言检测过程

谣言检测分为四个步骤：构建传播图和散布图，计算高层的节点表示，根节点信息增强，以及表示谣言的时序传播和空间散布。

4.2.1.1 构建传播图和散布图

设 A 是邻接矩阵，X 是事件 c 的特征矩阵，事件 c 包括传播路径上的节点和传播结构。A 只包含从上节点到下节点的边。对于自顶向下的 GCN，邻接矩阵就是 A；而对于自底向上的 GCN，邻接矩阵是 $A^{\top}$。自顶向下的 GCN 和自底向上的 GCN 采用相同的特征矩阵 X。

4.2.1.2 计算高层的节点表示

以两层的自顶向下 GCN 为例，层间的信息传播如下：

$$\begin{aligned}\mathbf{H}_1 &= \sigma\left(\hat{\mathbf{A}}\mathbf{X}\boldsymbol{W}_0\right),\\ \mathbf{H}_2 &= \sigma\left(\hat{\mathbf{A}}\mathbf{H}_1\boldsymbol{W}_1\right),\end{aligned} \tag{4.3}$$

其中 $\mathbf{H}_1$ 和 $\mathbf{H}_2$ 表示两层 GCN 的隐特征。$\boldsymbol{W}_0$ 和 $\boldsymbol{W}_1$ 是可训练的参数矩阵。$\hat{\mathbf{A}} = \widetilde{\mathbf{D}}^{-\frac{1}{2}}\widetilde{\mathbf{A}}\widetilde{\mathbf{D}}^{-\frac{1}{2}}$ 是归一化的邻接矩阵，其中 $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$，$\widetilde{\mathbf{D}}$ 对角线上的元素表示各节点的度。$\sigma(\cdot)$ 为激活函数，常见的如 ReLU 函数。

计算自底向上 GCN 的隐特征的方法类似（4.3），在此不再赘述。

4.2.1.3 根节点信息增强

谣言源头有着丰富的信息，根据节点和谣言源头的关系，我们可以更好地学习节点表示。因此可以将每个节点的隐特征向量和前一个图卷积层根节点的隐特征向量连接起来，作为新的特征矩阵。

4.2.1.4 传播和散布的表示

使用均值池化操作（mean-pooling）从自顶向下的节点表示和自底向上的节点表示中聚合信息，分别得到谣言传播和谣言散布的表示。将谣言传播和谣言散布的表示连接起来，经过一个全连接层和一个 Softmax 层，即可得到最终的分类标签。通过最小化预测和 ground truth 的交叉熵（cross-entropy）训练参数。

4.2.2 GCN 参数在谣言数据集中的对应

用户可抽象为 GCN 的节点，GCN 的边表示转发或回应的关系，如 a->b 表示 b 对 a 有一个回应。特征是前 5000 个词的 TF-IDF 值。

4.2.3 潜在问题

基于双向 GCN 的节点分类易受扰动，作出错误的判断。虽然相比于 [7]，这种方法在图结构以外还考虑了文本特征（TF-IDF），但该文本特征较为浅层，不能增强模型的鲁棒性。之后，我们将介绍一种假节点攻击，在 “不被注意” 的前提下影响模型的性能。

5 谣言假节点攻击和防御

这一章中我们首先介绍假节点攻击的原理，包括非针对性攻击和针对性攻击。现实世界中，攻击者无法知道模型的所有信息（包括参数和梯度更新信息），也不能对节点特征、关系进行任意改动。因此我们在假节点攻击的基础上，设计谣言检测场景中可行的攻击，并对基于双向 GCN 的谣言检测模型 [8] 实施攻击，测试其鲁棒性。最后我们讨论了谣言检测的未来工作，如通过 certified defense 增强 GCN 鲁棒性，通过众包合作的方式标注谣言，作为额外信息提升模型性能。

5.1 假节点攻击原理

假节点攻击 [34] 通过加入新的节点改变模型的分类结果。非针对性攻击没有特定的攻击对象，目标是最小化模型的分类准确率。针对性攻击的目标是使得模型在一组特定节点上分类错误。

5.1.1 非针对性攻击（Non-targeted Attack）

加入假节点后，邻接矩阵变成 $A' = \begin{bmatrix} A & B^T \\ B & C \end{bmatrix}$，特征矩阵变成 $X' = \begin{bmatrix} X \\ X_{\text{fake}} \end{bmatrix}$，其中 A 是原本的邻接矩阵，X 是原本的特征矩阵（输入），我们的目标是设计 B，C，X_{fake} 达到攻击的目的，同时保证加入的噪声是不易被察觉的（imperceptible）。

目标函数如公式5.1：

$$J(A', X') = \sum_{i \in S} \left(\max \left([f(X', A')]_{i,:} \right) - [f(X', A')]_{i, y_i} \right), \tag{5.1}$$

其中 y_i 是节点 i 的正确标签，S 是选定的一组目标节点。我们的优化问题是：

$$\arg\max_{B, C, X_{\text{fake}}} J(A', X') \quad \text{s.t. } \|B\|_0 + \|C\|_0 + \|X_{\text{fake}}\|_0 \le T, \tag{5.2}$$

对于这个离散优化问题，我们采用贪心算法进行攻击。B 和 X_{fake} 起初是 0，C 起初是单位矩阵 I。每一步增加一个特征和一条边。对于特征，我们找出 $\nabla_{X_{fake}} J(A', X')$ 的最大元素，并将其设置为非 0。类似地，对于边，我们找出 $\nabla_{B,C} J(A', X')$ 的最大元素，加入邻接矩阵中。

对于每一步更新，我们需要保证添加特征和边后的图局部和整体特征不变，从而达到“不被察觉”的目的。

5.1.2 针对性攻击（Targeted Attack）

针对性攻击可以让特定的谣言源头逃过 GCN 分类器的检测。假设我们的目的是保护一组谣言源头 S，y_i^* 是节点 i 的目标标签。目标函数定义如下：

$$J(A', X') = \sum_{i \in S} \left([f(X', A')]_{i, y_i^*} - \max \left([f(X', A')]_{i,:} \right) \right) \tag{5.3}$$

我们的目标是解决优化问题（5.2）。

5.2 谣言假节点攻击和防御

虽然基于贪心算法的白盒（white-box）假节点攻击 [34] 可以让节点分类任务的准确率低至 3%（比随机判断还要差），但在真实网络场景中，我们无法任意修改社交媒体中节点的特征和关系。直接用 [34] 中的攻击测试谣言检测模型，可能得到过于悲观的结论，因此我们设计了真实世界中可行的物理攻击（physical-world attack），作为对 [8] 中模型的可靠评估。我们还设计了基于对抗训练的防御方法，最为应对假节点攻击的方案。

5.2.1 攻击目标和数据集

我们将基于双向 GCN 的谣言检测模型 [8] 作为攻击目标。该模型同时考虑谣言的时序传播和空间散布，较好地捕捉了谣言网络的特点。

[8] 采用 Twitter16 数据集，数据集反映的谣言特征如下：

- 根节点 ID，即谣言源头的 Tweet① ID;
- 当前 Tweet 的父节点 Tweet 索引；
- 当前 Tweet 的索引，对于每个谣言的传播，Tweet 索引从 1 开始递增；
- 每个谣言传播树的深度，即谣言传播了几层；
- 文字长度，即每个谣言传播过程中，转发时评论最长的文字数；
- 文字内容，是一个列表，根据 TF-IDF 值选取了 5,000 个单词，统计这些单词在每条 Tweet 中出现的频率，以多个 (单词索引：出现次数) 对的形式出现。

数据集中共有四类谣言，真实谣言、虚假谣言、无法证实的谣言和非谣言，模型通过 GCN 对这四种类别进行区分。从数据集的描述中我们可以看出，现实世界中我

①https://twitter.com/

们可以进行的攻击并不多，如攻击者没有办法在任意节点见建立联系（无法强迫别人转发 Tweet）。下面我们将设计几种可行、不易被网络平台监管直接发现的谣言攻击。

5.2.2 三种真实世界谣言攻击

我们设计了如下三种真实世界谣言攻击，用于测试基于 GCN 的谣言检测模型的鲁棒性：

- 删除转发帖的评论，使模型无法学习到节点的语言特征。一方面，攻击者可以通过创建只转发谣言但不评论的假账号的方式，提高该类转发帖在所有转发帖中的比例。另一方面，现实中微博、推特上存在大量只转发不评论的用户和帖子，这种情形在检测中需要加以考虑。
- 多源头谣言传播。在 Twitter16 数据集中，一个谣言只有一个源头账号，其他账号只能转发该谣言而不能作为新的传播源头。现实世界中，一个谣言可以有多个源头，如 A、B 同时发布谣言贴，并形成各自的传播网络。攻击者可以通过多个账号传播谣言的方式，分散每个传播网络的影响力和被识别可能性。
- 增加文字长度。Twitter16 数据集中有一维特征是谣言传播过程中的最大文字数。攻击者可以增加无关文字数量，以弱化部分谣言敏感词的影响。

我们将在实验中分别测试这三种攻击，并将三种攻击结合起来。

5.2.3 基于对抗训练的防御

对抗训练的思想在于把可能的对抗样本加入训练集中，让模型提前学习复杂的分布，这样在测试时遇到对抗样本（inference time attack）则可以正确分类。基于对抗训练的防御方法相当于一种数据增强。

在我们的谣言检测场景中，我们对于几种攻击分别进行对抗训练，即用对抗样本进行模型的训练，并测试对抗训练后的鲁棒性是否有所改善。

5.3 实验验证

为了验证基于图网络的谣言检测的鲁棒性，我们用假节点攻击挑战基于双向 GCN 的谣言检测模型 [8]。在谣言检测的情形下，攻击者试图通过改变账号特征和关系影响节点分类效果，使谣言传播节点逃过模型检测，从而达到它的目的。

5.3.1 实验环境和参数设置

实验在 WSL（Windows Subsystem for Linux）Ubuntu 环境下进行，使用 PyTorch②框架。主机配备 i7-9700 CPU，32GB RAM，使用 NVIDIA RTX 2060（6G）并行化计算。

为了比较效果，我们采用和 [8] 类似的实验设置。使用随机梯度下降（stochastic gradient descent，SGD）更新模型参数，并使用 Adam 算法优化模型。每个节点的隐特征向量为 64 维。Dropout 设为 0.5，训练迭代 200 轮，并采用 early stopping 机制（若 validation loss 在 10 轮内不再下降，则训练停止）。

5.3.2 实验结果

我们在表 5.1中展示了三种攻击（以及组合攻击）在无防御情况下和对抗训练后的效果。

对于删除转发评论的攻击，模型准确率从 88.54% 下降到 84.37%。如果攻击者加入很多只转发不评论的假节点，则可以有效降低模型性能，更可能逃过谣言检测。另一方面，通过将对抗样本加入训练集进行对抗训练，模型的准确率可以达到 85.28%，仍低于 baseline 性能。

对于增加文字长度的攻击，模型准确率几乎不受影响。

对于多源头谣言传播的攻击，攻击后的准确率相比 baseline 反而提高了（从 88.54% 提升到 89.40%）。这可能和算法中的根节点信息增强有关，模型可以很好地学习根节点特征，进行谣言分类。

根据我们的实验，基于双向 GCN 的谣言检测模型会受对抗样本干扰，但可以通过对抗训练进行防御，相比于基于语言特征的谣言检测，鲁棒性有了很大的提升。

表 5.1 假节点攻击和防御实验结果

场景及准确率	组合攻击	删除转发评论	多源头谣言传播	增加文字长度
baseline	0.8854	0.8854	0.8854	0.8854
假节点攻击	0.8513	0.8437	0.8940	0.8845
对抗训练	0.8507	0.8528	0.8750	0.8843

②https://pytorch.org/

5.4 讨论和未来工作

我们在实验中发现，对抗样本对于 GCN 谣言检测模型 [8] 有一定效果，但不会严重影响分类准确率。在合适的防御机制（如对抗训练）下，该模型可以较安全地检测出谣言。相比于完全基于语言特征的模型 [33]，基于谣言传播模式的 GCN 模型展现了更强的鲁棒性，但仍有两个问题。

一方面，假节点攻击在现实中是很容易实现的，攻击者可以轻松地进行**更强大**的攻击，如通过注册僵尸账号，关注大量特定用户，达到愚弄检测模型、逃过检测的目的。若通过对抗训练防御，则计算开销很大，且对于每一种可能的对抗样本进行相应的对抗训练是无法实现的。对于这样的安全威胁，我们可以通过 certified defense 增强 GCN 的鲁棒性。简单来说，certified defense 中我们需要定义一个模型打分函数 F，对于干净样本 x 和对应的对抗样本 x'，找到 $F(x')-F(x)$ 的上界（upper bound），我们的优化目标是让这个上界尽量小，这样就可以确保在输入扰动/噪声不太大（不易被人察觉）的前提下，模型输出相对稳定，不对噪声过分敏感。

另一方面，基于 GCN 的谣言检测模型的分类准确率不是很高，我们测试的准确率为 88.54%（[8] 中结果为 88.00%），对于非谣言、虚假谣言、真实谣言、不确定谣言的 F1 值分别为 0.7761、0.8814、0.9344、0.9083。未来我们将进一步研究如何提升分类效果，如从“人”的角度出发，运用人机交互方法，以众包合作的方式标注谣言，并将这种信息融合进机器学习模型，达到更准确、鲁棒的效果。人和机器学习模型的合作、信息分享在以往的人机交互研究中很少提及，我们将进行此类研究，形成一种新的人机合作范式。

6 结论

社交媒体上的谣言日渐泛滥，对平台环境、用户体验产生了极大的不良影响，甚至造成严重的经济损失。谣言自动检测是近年学术研究的一个热点。谣言检测方法主要有基于文本内容的检测和基于谣言传播特征的检测。

我们首先证明，基于文本的检测方法捕捉的多为浅层特征（如更可能出现在谣言中的词、文本风格），因而易受对抗样本的干扰。我们通过事实篡改、主谓颠倒、因果混淆等方法，进行事实、逻辑层面的篡改，并成功逃过基于文本特征的模型（Fakebox）的检测。实验中我们还发现，基于文本的谣言检测会将大量真实信息误判为谣言，这类样本往往由普通用户产生，不具有专业新闻的风格和特点。这会严重打击社交媒体用户和开放新闻平台用户生产内容的积极性。

最近的研究开始考虑文本、用户等内容特征以外的因素，如谣言传播网络的图结构特征。新的基于图神经网络的算法被提出，达到较好的检测效果。但我们认为，图模型同样易受对抗样本的干扰。论文中我们考察了基于双向 GCN 的谣言检测模型，通过一种“不易被察觉”的假节点攻击测试其鲁棒性。实验表明，基于 GCN 的谣言检测模型会受对抗样本干扰，攻击者可以通过加入假节点或改变节点特征使谣言逃过算法检测。在我们的攻击下，谣言检测模型的准确率从 88.54% 降低到 84.37%。

我们通过对抗训练使 GCN 模型更鲁棒，在有噪声的环境下也能保持较高的准确率，从而防御此类假节点攻击。对抗训练后，模型面对对抗样本可以保持 85.28% 的准确率。由于对抗训练无法防御新的、更强大的攻击，未来我们将从防御者的角度，通过 certified defense 提供可证明的鲁棒性。另一方面，我们将研究谣言传播中的人因，将获得的领域知识加入到模型中，提升其效果和鲁棒性。

论文中我们分析了基于语言特征和谣言传播特征的谣言检测模型的性能和鲁棒性，提出相应的攻击和防御，并为未来的谣言检测研究分析了方向，作为我们的未来工作。

参考文献

[1] CHEN T, LI X, YIN H, et al. Call Attention to Rumors: Deep Attention Based Recurrent Neural Networks for Early Rumor Detection[A]. Workshop of Pacific-Asia Conference on Knowledge Discovery and Data Mining[C], 2018 : 40 – 52.

[2] LIU Y, fang BROOK WU Y. Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks[A]. Proceedings of AAAI Conference on Artificial Intelligence[C], 2018 : 354 – 361.

[3] MA J, GAO W, MITRA P, et al. Detecting Rumors from Microblogs with Recurrent Neural Networks[A]. Proceedings of International Joint Conference on Artificial Intelligence[C], 2016 : 3818 – 3824.

[4] MONTI F, FRASCA F, EYNARD D, et al. Fake News Detection on Social Media using Geometric Deep Learning[A]. arXiv[C], 2019.

[5] ZHOU Z, GUAN H, BHAT M M, et al. Fake News Detection via NLP is Vulnerable to Adversarial Attacks[A]. arXiv[C], 2019.

[6] WANG Z, WANG C, PEI J, et al. Multiple Source Detection without Knowing the Underlying Propagation Model[A]. Proceedings of AAAI Conference on Artificial Intelligence[C], 2017 : 217 – 223.

[7] DONG M, ZHENG B, HUNG N Q V, et al. Multiple Rumor Source Detection with Graph Convolutional Networks[A]. Proceedings of Conference on Information and Knowledge Management[C], 2019 : 569 – 578.

[8] BIAN T, XIAO X, XU T, et al. Rumor Detection on Social Media with Bi-Directional Graph Convolutional Networks[A]. Proceedings of AAAI Conference on Artificial Intelligence[C], 2020.

[9] ZUGNER D, AKBARNEJAD A, GUNNEMANN S. Adversarial Attacks on Neural Networks for Graph Data[A]. arXiv[C], 2018.

[10] DAI H, LI H, TIAN T, et al. Adversarial Attack on Graph Structured Data[A]. Proceedings of International Conference on Machine Learning[C], 2018 : 1123 – 1132.

[11] CHEN Y, CONROY N J, RUBIN V L. Misleading Online Content: Recognizing Clickbait as False News[A]. In Proceedings of EMNLP[C], 2017.

[12] GRANIK M, MESYURA V. Fake News Detection Using Naive Bayes Classifier[A]. In Proceedings of IEEE First Ukraine Conference on Electrical and Computer Engineering[C], 2017.

[13] HORNE B D, ADALI S. This Just In: Fake news Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News.[A]. In Proceedings of 2nd International Workshop on News and Public Opinion at ICWSM[C], 2017.

[14] SHAH D, ZAMAN T. Rumor Centrality: A Universal Source Detector[A]. Proceedings of International Conference on Measurement and Modeling of Computer Systems[C], 2012 : 199 – 210.

[15] LAPPAS T, TERZI E, GUNOPULOS D, et al. Finding Effectors in Social Networks[A]. Proceedings of SIGKDD Conference on Knowledge Discovery and Data Mining[C], 2010 : 1059 – 1068.

[16] PRAKASH B A, VREEKEN J, FALOUTSOS C. Spotting Culprits in Epidemics: How Many and Which Ones?[A]. Proceedings of International Conference on Data Mining[C], 2012 : 11 – 20.

[17] ZHU K, YING L. Information Source Detection in the SIR Model: A Sample Path Based Approach[A]. Proceedings of International Conference on Information Technology and Applications[C], 2013 : 1 – 9.

[18] SHEN Z, CAO S, WANG W, et al. Locating the Source of Diffusion in Complex Networks by Time-reversal Backward Spreading[J]. Physical Review E, 2016, 93(3).

[19] WANG Z, WANG C, PEI J, et al. Multiple Source Detection without Knowing the Underlying Propagation Model[A]. Proceedings of AAAI Conference on Artificial Intelligence[C], 2017 : 217 – 223.

[20] YING R, HE R, CHEN K, et al. Graph Convolutional Neural Networks for Web-Scale Recommender Systems[A]. Proceedings of SIGKDD Conference on Knowledge Discovery and Data Mining[C], 2018 : 974 – 983.

[21] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[A]. arXiv[C], 2013.

[22] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and Harnessing Adversarial Examples[A]. arXiv[C], 2014.

[23] CARLINI N, WAGNER D. Towards Evaluating the Robustness of Neural Networks[A]. Proceedings of IEEE Symposium on Security and Privacy[C], 2017 : 39 – 57.

[24] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks[A]. arXiv[C], 2017 : 39 – 57.

[25] SHAFAHI A, NAJIBI M, GHIASI M A, et al. Adversarial Training for Free![A]. arXiv[C], 2019.

[26] RAGHUNATHAN A, STEINHARDT J, LIANG P. Certified Defenses against Adversarial Examples[A]. arXiv[C], 2018.

[27] ZHOU J, CUI G, ZHANG Z, et al. Graph Neural Networks: A Review of Methods and Applications[A]. arXiv[C], 2018.

[28] KIPF T N, WELLING M. Semi-supervised Classification with Graph Convolutional Networks[A]. Proceedings of International Conference on Learning Representations[C], 2017.

[29] CHEN J, MA T, XIAO C. FastGCN: Fast Learning with Graph Convolutional Networks via Importance Sampling[J]. arXiv, 2018.

[30] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks[A]. arXiv[C], 2017.

[31] LI J, MONROE W, JURAFSKY D. Understanding Neural Networks through Representation Erasure[A]. arXiv[C], 2016.

[32] LIANG B, LI H, SU M, et al. Deep Text Classification Can be Fooled[A]. arXiv[C], 2017.

[33] EDELL A. I trained fake news detection AI with 95% accuracy, and almost went crazy[A]. [C], 2018.

[34] WANG X, CHENG M, EATON J, et al. Fake Node Attacks on Graph Convolutional Networks[A]. arXiv[C], 2019.

致谢

感谢李晨亮教授。他在毕业论文撰写过程中给我提供了很大的指导和帮助。他在自然语言处理和社交媒体领域的研究让我很受启发。之前经常听裴嘉欣和许灿文提起李老师，经过一段时间的接触，果然水平很高！希望以后能继续合作，把毕业论文中没来得及做的部分做出来。

与华中科技大学韩旭同学的讨论让这篇论文更加完善。我们之后在机器学习安全领域会持续合作。

感谢威斯康辛大学麦迪逊分校的 Justin Hsu 教授。大三上学期，在麦迪逊交换，选了他的一门研究生课——安全与隐私，让我对机器学习安全产生了浓厚的兴趣。由 Justin 指导、和 Meghana Bhat 合作完成的课程项目最终成功发表，成为我人生中的第一篇论文！投稿前 Justin 帮我反复修改，布拉格的口头报告之前，他给我发了几千字的邮件，告诉我怎么做展示、怎么修改幻灯片，并且通过 Skype 帮我预演，最终成功的展示离不开他的热心帮助。几乎每周的 office hour，我都会去 Justin 的办公室讨论问题，不管是学术研究、未来规划还是闲聊，都让我很受启发。最终走上科研道路、决定读博并且申请到伊利诺伊大学香槟分校的 PhD，Justin 起了巨大的引导作用。我们一直保持着联系。

感谢复旦大学的丁向华教授。她教给我定性研究的方法，以此进行更深入的用户研究，挖掘很多有趣现象背后的本质，启发计算系统的设计。在她的指导下了，我慢慢可以深刻理解计算技术对人和社会的影响，并且通过自己的研究作出一些贡献。定性方法在我现在的研究中起着非常重要的作用。合作过程中我还认识了复旦大学汤欣如、宾州州立大学 Xinning Gui 教授和加州大学欧文分校 Yunan Chen 教授，和她们的讨论也让我受益匪浅。

感谢马里兰大学帕克分校的 Furong Huang 教授。如果说和 Justin 研究的机器学习安全偏向应用，那么和 Furong 研究的机器学习安全则偏向理论。在她的指导下，我感受到了机器学习理论和数学的魅力，用数学方法给机器学习提供可证明的安全性，这是一个非常有趣的领域！合作过程中我还认识了慕尼黑工业大学曾惠民（他有很强的数学和代码功底，祝他今年 PhD 申请顺利！）。

感谢西湖高等研究院的张岳教授。虽然接触的时间不长，但他的研究态度给我留下了深刻的印象。实验室的博士生陈雨龙和王存翔日后都会是优秀的 NLP 研

究者！

感谢伊利诺伊大学香槟分校的 Yang Wang 教授，也就是我未来的的博士导师。他给我很大的研究自由度，同时能够提供非常有帮助的指导和意见。他鼓励我实习，并且在实习的过程中建立研究上的合作。虽然接触的时间还不长，但能感受到他完美的性格！相信我们会有愉快、高产的合作。

感谢张乐飞、黄浩教授。他们有趣的课堂，让我对机器学习、数据挖掘产生了浓厚的兴趣。感谢所有的老师。

感谢王菲、陈子轩、雷伯涵、孟凡嵩、原昊博、关焕康、张珍妮、曹凯文、陈卓、徐一恒、富鑫等同学。感谢莎士比亚戏剧社的剧组朋友们（特别是我主演的“驯悍记”的女主角，美丽的毕钰淇女士！）、校学生会的同事们、院篮球队的队友们（特别是信任我、给我不停传球的韩森，让我在三场比赛中投进 7 记三分球！），感谢所有的朋友。你们给我留下永远的美好回忆！

感谢武汉大学和弘毅学堂。学校和学院为我出国交换、访问提供了许多资金和政策上的支持。感谢弘毅学堂石兢、方萍、李瑶、董甲庆老师和辅导员谢莹萍。

感谢字节跳动、卓尔智联研究院、腾讯给我提供实习的机会，让我对机器学习、人机交互在工业界的落地有了初步的感受。初入职场，得到了很多同事热心的帮助，在此一并感谢！

感谢父母和家人长期的支持和鼓励。没有你们，我不会取得今天的成绩！

四年的时光弹指一挥间，从青涩地踏进校园，到即将本科毕业，步入博士生涯。感谢过得飞快的时间，告诉我要不断努力，永不止步！

最后用我很喜欢的一句话结束。感谢陈立杰的这句话。“能够生在这样一个黄金时代里，我感到无比的荣幸。我梦想能够成为黄金时代浪潮中的一朵浪花，为人类的智慧添砖加瓦！”