SeqGANPass:使用序列生成式对抗网络进行口令猜测

龚雪鸾[1],陈艳姣[2],王 涛[1],曹雨欣[1]
(1. 武汉大学计算机学院,湖北武汉 430070; 2. 浙江大学电气工程学院,浙江杭州 310058)

摘 要: 为了破解用户口令并获取用户隐私信息,口令猜测工具应运而生. 基于规则的口令猜测工具虽猜测成功率较高,但制定规则非常耗时且需要一定的专业知识. 基于深度神经网络的口令猜测工具则需要大量的训练数据集来训练模型. 基于此,本文提出了(Sequence Generative Adversarial Network Password, SeqGANPass),利用序列生成式对抗网络,针对口令数据集执行数据预处理操作,经由多轮对抗性训练过程训练口令生成器,以生成高质量的猜测口令. 即使没有任何先验知识,SeqGANPass仍可以通过小规模训练集来实现口令破译. 同时我们发现使用SeqGANPass可以大大提高基于规则的口令猜测工具的有效性. 在实验中,我们与当前的主流口令猜测工具进行比较,如John the Ripper,Hashcat,Markov Model,上下文无关文法(Probabilistic Context Free Grammars,PCFG),FLA(Fast, Lean, and Accurate)和PassGAN等. 实验表明,SeqGANPass的匹配率优于这些主流的口令猜测工具.

关键词: 口令猜测;序列生成式对抗网络;深度学习;口令匹配;隐私泄露;生成式对抗网络
中图分类号: TP311 **文献标识码:** A **文章编号:** 0372-2112(2023)05-1148-06
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220633

SeqGANPass: Password Guessing with Sequence Generative Adversarial Nets

GONG Xue-luan[1], CHEN Yan-jiao[2], WANG Tao[1], CAO Yu-xin[1]
(1. *College of Computer Science, Wuhan University, Wuhuan, Hubei* 430070, *China*;
2. *College of Electrical Engineering, Zhejiang University, Hangzhou, Zhejiang* 310058, *China*)

Abstract: In order to crack the user's password to achieve the purpose of obtaining user's private information, password guessing tools also came into being. Although state-of-the-art rule-based attacks work achieve high attack success rate, the collection of rules is time consuming and needs expertise. Deep neural network-based attacks require amounts of datasets to achieve a good result. In this paper, we propose sequence generative adversarial network password (SeqGANPass), which uses sequence generative adversarial nets, conducts data preprocessing operations on the password datasets, to generate high-quality passwords. SeqGANPass can implement password cracking under a small scale of training set even without any prior knowledge. Furthermore, we show that SeqGANPass can greatly improve the effectiveness of rule-based attacks. Our experiments show that SeqGANPass outperforms most state-of-the-art password guessing methods, i.e., John the Ripper, Hashcat, Markov model, probabilistic context free grammars (PCFG), FLA (Fast, Lean, and Accurate), and PassGAN in matching rate.

Key words: password guessing; sequence generative adversarial networks; deep learning; password matching; privacy leakage; generative adversarial networks

1 引言

口令在现代网络安全中起着至关重要的作用,它是使用最为广泛的身份验证方式之一[1,2]. 口令猜测的目的是以最小的代价生成与真实口令相匹配的猜测口令. 口令在验证系统数据库中通常以散列形式存储,因此口令猜测工具需要快速有效地测试大量的候选口令是否与真实口令相匹配. 为了提高匹配率,口令猜测工具需要从高质量字典中选择口令. 目前从网上泄露的口令具有一定的局限性:(1)这些泄露的口令集的质量

收稿日期:2022-05-31;修回日期:2022-11-18;责任编辑:覃怀银

难以保证,攻击者可能像口令集中注入恶意口令,从而造成数据污染.(2)泄露的真实口令种类局限于论坛型网站的口令集.因此,研究人员难以有针对性地获得大规模的优质数据集.在这种情况下,如果口令猜测工具只需要小规模的数据集来训练,就可以获得更有针对性、更准确的结果.

基于特定单词转换规则的口令猜测方案的性能受特定规则的限制,只能生成有限的口令猜测.随着机器学习的发展,Narayanan等人[3]首次利用马尔可夫模型生成口令猜测.此方案需要做大量的预运算,计算量大且耗时长.Weir等人在研究口令的构造规律以及特征之后,提出了基于概率上下文无关文法(Probabilistic Context Free Grammars,PCFG)的口令猜测方案,以最高概率顺序生成口令结构[4].随后,邹静[5]以及韩伟力[6]等人基于PCFG展开了进一步的改进研究.基于PCFG的模型虽然可以准确地抽象出基础口令结构,但是泛化能力较差.

当猜测次数规模较小时,上述两种方法效果较好,但是在10^{10}以上的猜测时,使用基于神经网络的口令猜测工具将会带来更高的成功率.Melicher等人[7]提出利用RNN实现口令猜测.Wu等人[8]将PCFG用于PassGAN的预处理;Wang等人[9]将PCFG于RNN相结合,提出新的PR模型.最近,Hitaj等人[10]提出PassGAN,一种利用对抗式生成网络(Generative Adversarial Networks,GAN)[11]来增强口令破解的新方案.

虽然对抗式生成网络具有很强的图像生成能力,但是当处理像口令这样的离散数据时有一定的限制.一方面,判别器很难将梯度更新传递给生成器,另一方面,对抗式生成网络的判别器只能对一个完整的生成序列进行评估,不能评估非完整序列.为了解决上述问题,我们提出了一种基于序列生成式对抗网络(Sequence Generative Adversarial Nets,SeqGAN)[12]的口令猜测方案SeqGANPass.基于序列生成式对抗网络在生成离散文本上的卓越性能,SeqGANPass可以通过更多的转换方式生成口令猜测,表现出更高的匹配率.与使用规则的口令猜测工具不同,SeqGANPass可以自动学习人工生成的口令结构以及字符之间的关系.为了使生成器生成的口令更接近真实口令,我们使用判别器来确定生成的口令是否足够真实.此外,SeqGANPass不需要任何先验知识,包括预设结构和规则.

2　系统设计

2.1　模型结构

SeqGANPass的结构如图1所示.我们首先寻找一个由泄露真实口令组成的数据集,然后执行数据预处理操作(即数据清洗,数据集划分和数据格式转换),以适用于SeqGANPass.在数据清洗过程中,由于绝大多数口令少于10个字符,我们将大于10字符的口令从数据集中删除,以减少训练成本.此外,我们删除所有ASCII无法编码的口令.在数据集划分过程中,我们将过滤后的数据集分为训练数据集和测试数据集.在数据格式转换和填充过程中,由于生成器很难处理原始字符串,我们使用映射转换获取按顺序编号的口令,并将所有口令填充为10个字符.

接下来,我们利用序列生成式对抗网络来训练SeqGANPass.经过大量的对抗性训练迭代后,生成器能够生成与训练数据集具有相似分布的口令.利用经过该训练的生成器,攻击者可以生成足量的猜测口令.我们在算法1中总结了SeqGANPass的工作过程.

算法1　SeqGANPass口令猜测算法

输入: 口令最大限制长度 Maxlen; 字符映射 CharMap;
允许字符集 CharSet; 口令集 S
输出: 口令猜测

```
FOR each password in S DO
  IF length(password) > Maxlen THEN
    从S中移除口令
  ELSE
    FOR each character not in CharSet DO
      从S中移除口令
    END FOR
    CharMap(password) = index sequence
  END IF
END FOR
Divide(S) = training dataset + testing dataset
Initialize(SeqGAN)
REPEAT
    训练SeqGAN
UNTIL生成足够数量高质量口令
```

2.2　数据预处理

首先,RockYou数据集[13]中的口令长度是不同的,我们发现大约90%的口令长度小于10个字符.因此为了精确有效地训练和测试SeqGANPass,我们过滤掉了所有长度超过10个字符的口令.虽然此操作限制了SeqGANPass的有效性,但是也大大降低了训练成本,因此我们认为此操作是合理的.其次,我们发现RockYou数据集中有一小部分口令包含极其罕见的字符,这些字符在大多数情况下不会被用来构造口令,而且可能会引起一些字符编码问题,因此我们将它们从数据集中移除.也就是说,我们使用的口令由可用ASCII编码系统编码的字符组成,即95个可打印字符,包括数字、英文字母和标点符号.

我们将整个数据集分为两部分:训练数据集和测

试数据集. 此外,我们重新排列了口令的顺序来确保口令满足 Zipf 分布[14]. 我们使用 RockYou 数据集进行实验,在去除含有非 ASCII 字符、长度超过 10 或者重复的口令后,随机选取了 100 000 个口令(约 0.8%)作为训练数据集. 然后使用剩余的的数据集约 99.2%,共 11 799 187 个口令,去除训练集后有(11 798 569 个口令)来测试模型的有效性. 我们创建了一个映射来索引数据集中的所有字符,每个字符对应一个唯一索引的映射. 在实验中,我们根据 RockYou 数据集中字符出现的顺序来建立字符映射.

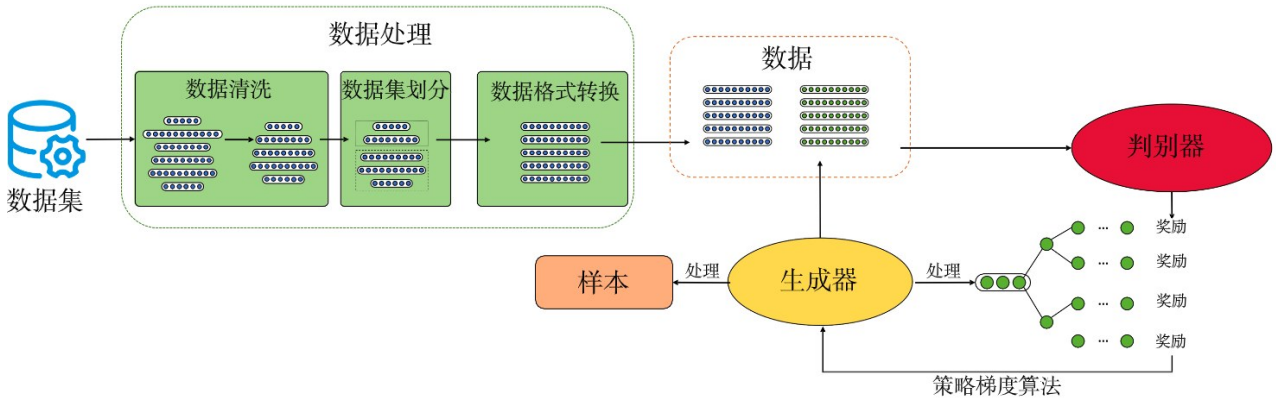

图 1 SeqGAN 的架构

2.3 训练生成器和判别器

我们训练 SeqGANPass 以生成猜测口令. 我们选择长短期记忆(Long Short-Term Memory, LSTM)[15]网络作为生成器,并选择卷积神经网络作为判别器.

首先,我们使用随机权重初始化生成器和判别器. 然后在训练集上利用最大似然估计对生成器进行预训练,生成器通过最大似然估计生成的伪样本用于预训练判别器. 经过预训练后,生成器和判别器将轮流进行训练. 当生成器更新它的参数时,判别器也需要周期性地进行重训以跟上生成器的训练步伐. 我们使用训练集的真实口令和从生成器生成的伪口令训练判别器. 为了保持平衡,真实口令的数目与伪口令的数目相同,并且使用不同的真实口令和伪口令组合. 此外,我们使用 L_2 正则化和 dropout[16,17]来避免过拟合.

完成上述对 SeqGANPass 的训练过程后,我们用生成器来生成高质量的口令.

3 实验与评估

3.1 训练数据集与测试数据集

为了评估 SeqGANPass 的有效性,我们将其与当前主流的口令猜测工具进行比较. 我们使用 RockYou 数据集[13], LinkedIn 数据集[18]和 Yahoo 数据集中的口令对其进行测试. 为了评估 SeqGANPass 在中文用户口令数据集上的性能,我们还使用了 CSDN 数据集[19]来训练和测试它. CSDN 是一个中国程序员社区网站,该数据集包含超过 600 万条口令.

3.2 评估结果

3.2.1 由 SeqGANPass 生成的口令结果

为了准确评估 SeqGANPass 生成的口令,我们首先生成几个独立的口令猜测集合,数量范围从 10^4 到 10^{10}. 然后我们计算其中唯一口令与 RockYou 测试数据集、LinkedIn 数据集和 Yahoo 数据集的匹配率. RockYou 测试数据集中含有 1 179 856 条不重复的口令;LinkIn 数据集中含有 25 525 084 条不重复口令;Yahoo 数据集中含有 295 999 条不重复口令. 表中的“SeqGANPass 生成口令数量”列表示 SeqGANPass 生成的全部口令;“去重后口令数量”列表示在生成的所有口令中,去除了已经生成过的口令后剩下的口令数. 如表 1 所示,我们可以看到,随着生成样本数量的增加,唯一口令的数量和匹配率都会增加,但随着生成样本数量的继续增加,我们发现匹配数的增长率略有下降. 因此,我们把这种现象归因为:较简单的口令在最开始的时候就会被匹配,而较复杂的口令则需要更多次的尝试才能对其进行匹配.

此外,我们将 SeqGANPass 和其他的口令猜测工具在 RockYou 测试数据集、LinkedIn 数据集和 Yahoo 数据集上进行比较,表 2 表示利用不同口令猜测工具生成的口令去重后的数量,SeqGANPass 与这些口令猜测工具的表现对比结果如表 3 所示. 我们可以看到,尽管 SeqGANPass 缺乏关于口令结构的信息,但它只需要生成更少的口令,就可以达到与其他主流口令猜测工具相等甚至更高的的测试数据集匹配率. 需要注意的是,表 3 的“生成口令数量(去重)”行是指从 SeqGANPass 生成的所有口令中去除重复口令后的口令数. 因此,当匹配

表1　SeqGANPass在不同测试集上的表现

SeqGANPass生成口令数量	去重后口令数量	RockYou匹配率	LinkedIn匹配率	Yahoo匹配率
10^4	9 912	0.006%	0.000 16%	0.049%
10^5	95 620	0.038%	0.002 9%	0.089%
10^6	856 216	0.15%	0.020%	0.18%
10^7	7 001 481	0.78%	0.39%	1.03%
10^8	49 151 889	3.31%	1.90%	4.60%
10^9	332 229 164	13.18%	8.50%	16.16%
10^{10}	2 440 189 466	41.66%	26.73%	35.54%

表2　不同口令猜测工具生成口令数量

口令猜测工具	去重后口令数量
John the Ripper	2.64×10^9
Hashcat best64	5.32×10^6
Hashcat generated2	1.69×10^9
Markov模型	1.10×10^8
PCFG	1.83×10^9
FLA	8.51×10^7
PassGAN	6.81×10^8

率一致时,表中“生成口令数量(去重)”会低于表1中“SeqGANPass生成口令数量”数值.

为了证明SeqGANPass在中文数据集上的优势,我们使用CSDN数据集来训练测试SeqGANPass,我们对比了SeqGANPass和其他两种主流的口令猜测工具:John the Ripper和PCFG,结果见表4. 我们可以看到,SeqGANPass在CSDN数据集上的匹配率远高于John the Ripper,但略逊色于PCFG. 这说明SeqGANPass在中文数据上的性能仍有提高的空间,我们将会在未来着力于改进其在中文数据集上的表现. 此外,由于PassGAN不适合直接应用于中文数据集,所以我们暂时没有进行PassGAN在中文数据集上的实验.

表3　不同口令猜测工具与SeqGANPass的表现对比

		John the Ripper	Hashcat best64	Hashcat generated2	Markov模型	PCFG	FLA	PassGAN
RockYou	匹配率	31%	3%	24%	11%	25%	19%	15%
	生成口令数量(去重)	1.16×10^9	4.08×10^7	8.00×10^8	2.40×10^8	8.07×10^8	5.40×10^8	4.12×10^8
LinkIn	匹配率	18%	1%	12%	3%	10%	7%	7%
	生成口令数量(去重)	9.52×10^8	2.05×10^7	5.10×10^8	1.04×10^8	3.92×10^8	2.55×10^8	2.51×10^8
Yahoo	匹配率	34%	8%	29%	17%	26%	27%	24%
	生成口令数量(去重)	1.65×10^9	1.02×10^9	1.02×10^9	3.61×10^8	8.11×10^8	9.20×10^8	6.84×10^8

表4　不同口令猜测工具在CSDN数据集上的表现

口令猜测工具	生成口令	匹配率
John the Ripper	5.07×10^5	0.13%
PCFG	9.95×10^9	57.03%
SeqGANPass	4.40×10^9	37.46%

3.2.2　结合SeqGANPass和基于规则的口令猜测

我们发现使用SeqGANPass可以进一步提高基于规则的口令猜测工具的有效性. 为了验证可行性,我们向John the Ripper的基础字典中添加了4 403 290 782条由SeqGANPass生成的高质量口令,并分别使用RockYou、LinkedIn和Yahoo数据集来测试改进的有效性. 结果表明,在添加SeqGANPass后,基于规则的口令猜测工具的匹配率提高了接近一倍,结果见表5. 这表明SeqGANPass有助于基于规则的口令猜测工具匹配更多的口令,从而在本质上增强基于规则的口令猜测工具的有效性.

4　讨论

SeqGAN在口令猜测方面匹配率很高. 使用SeqGAN,经过105个口令(0.8% Rock You数据集)训练的SeqGANPass能够匹配41.66%的RockYou测试集、26.73%的LinkedIn测试集,以及35.54%的Yahoo测试集,它的匹配率超过了大多数主流的口令猜测工具[4,5]. 与基于规则的口令猜测工具不同,SeqGANPass不需要关于口令结构的先验知识. 此外,与基于DNN的口令猜测工具不同,SeqGANPass需要的训练数据集要小得多.

基于规则的口令猜测工具是有效的,但仍有局限

表5　SeqGANPass增强基础字典

测试集	添加前匹配率	添加后匹配率
RockYou	31.06%	60.47%
LinkedIn	18.12%	43.81%
Yahoo	33.83%	57.35%

性. 一方面,基于规则的口令猜测工具生成口令的速度比其他方法快得多. 另一方面,它们可以有效地揭示人类生成的口令的结构. 当向字典中添加更多的口令时,匹配率将显著提高. 但是,它们只能通过有限的转换生成固定数量的口令. 因此,基于规则的口令猜测工具的匹配率在很大程度上取决于字典中条目的质量.

SeqGANPass可以用来辅助基于规则的口令猜测工具. SeqGANPass通过少量的样本训练能够生成无限数量的高质量口令猜测,这些猜测可以用来补充基本字典,并增强基于规则的口令猜测工具实用性. 在实验中,我们验证了该方法的可行性.

口令数据集具有较强的用户母语关联度. 除了在实验中采用的Rock You、LinkedIn、Yahoo、CSDN数据集,我们也调研了其它国内泄露的真实口令数据集,例如JingDong数据集. 我们发现在所有的口令数据集中,字母和数字占了97%以上. 在英文的口令数据集中,字母占比大概69%,数字占比大概27%,而在中文的口令数据集中,字母占比约为30%,数字占比约为68%. 我们把这一现象归因于用户母语的影响. 对于以中文为母语的用户,数字较字母更容易记忆,同时例如“666666”、“888888”、“5201314”等一些特殊的字符串具有特殊意义. 因此,中文用户较倾向于使用数字作为口令.

5 结论

本文提出了一种基于序列生成式对抗网络的口令猜测框架SeqGANPass,该框架可以在不需要任何先验知识的前提下生成大规模高质量的口令猜测. 通过与当前最先进的口令猜测工具比较,我们发现SeqGANPass可以实现更高的匹配率,有更好的口令生成能力. 在未来的工作中,我们将会尝试将SeqGANPass与原始的PCFG相结合,期待可以取得更好的结果.

参考文献

[1] 王平, 汪定, 黄欣沂. 口令安全研究进展[J]. 计算机研究与发展, 2016, 53(10): 2173-2188.
WANG P, WANG D, HUANG X Y. Advances in password security[J]. Journal of Computer Research and Development, 2016, 53(10): 2173-2188. (in Chinese)

[2] 尚旭哲, 王润田, 孙颖, 等. 口令破解与防范技术研究[J]. 网络空间安全, 2020, 11(5): 98-103.
SHANG X Z, WANG R T, SUN Y, et al. The research on password cracking and prevention technology[J]. Cyberspace Security, 2020, 11(5): 98-103. (in Chinese)

[3] NARAYANAN A, SHMATIKOV V. Fast dictionary attacks on passwords using time-space tradeoff[C]//ACM Conference on Computer and Communications Security. New York: ACM, 2005: 364-372.

[4] WEIR M, AGGARWAL S, DE MEDEIROS B, et al. Password cracking using probabilistic context-free grammars [C]//2009 30th IEEE Symposium on Security and Privacy. Piscataway: IEEE, 2009: 391-405.

[5] 邹静, 林东岱, 郝春辉. 一种基于结构划分概率的口令攻击方法[J]. 计算机学报, 2014, 37(5): 1206-1215.
ZOU J, LIN D D, HAO C H. A password cracking method based on structure division probability[J]. Chinese Journal of Computers, 2014, 37(5): 1206-1215. (in Chinese)

[6] 韩伟力, 袁琅, 李思斯, 等. 一种基于样本的模拟口令集生成算法[J]. 计算机学报, 2017, 40(5): 1151-1167.
HAN W L, YUAN L, LI S S, et al. An efficient algorithm to generate password sets based on samples[J]. Chinese Journal of Computers, 2017, 40(5): 1151-1167. (in Chinese)

[7] MELICHER W, UR B, SEGRETI S M, et al. Fast, lean, and accurate: Modeling password guessability using neural networks[C]//Proceedings of the 25th USENIX Conference on Security Symposium. New York: ACM, 2016: 175-191.

[8] WU Y X, WANG D, ZOU Y K, et al. Improving Deep Learning Based Password Guessing Models Using Pre-Processing[M]//Information and Communications Security. Cham: Springer International Publishing, 2022: 163-183.

[9] 汪定, 邹云开, 陶义, 等. 基于循环神经网络和生成式对抗网络的口令猜测模型研究[J]. 计算机学报, 2021, 44(8): 1519-1534.
WANG D, ZOU Y K, TAO Y, et al. Password guessing model based on recurrent neural networks and generative adversarial networks[J]. Chinese Journal of Computers, 2021, 44(8): 1519-1534. (in Chinese)

[10] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[11] HITAJ B, GASTI P, ATENIESE G, et al. PassGAN: A Deep Learning Approach for Password Guessing[M]//Applied Cryptography and Network Security. Cham: Springer International Publishing, 2019: 217-237.

[12] YU Lan-tao, ZHANG Wei-han, WANG Jun. SeqGAN: Sequence generative adversarial nets with policy gradient [C]//AAAI Conference on Artificial Intelligence. San Francisco: AAAI Press, 2017: 2852-2858.

[13] Skullsecurity. RockYou[CP/OL]. (2010-08-01) [2022-11-17]. https://downloads.skullsecurity.org/passwords/rocky-

ou.txt. bz2.
[14] WANG Ding, CHENG Hai-bo, WANG Ping. Zipf's law in passwords[J]. IEEE Transactions on Information Forensics and Security, 2017, 12(11): 2776-2791.
[15] ZIA T, ZAHID U. Long short-term memory recurrent neural network architectures for Urdu acoustic modeling [J]. International Journal of Speech Technology, 2019, 22 (1): 21-30.
[16] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, et al. Improving neural networks by preventing co-adaptation of feature detectors[EB/OL]. [2022-05-24]. DOI: https://doi.org/10.48550/arXiv.1207.0580.
[17] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: A simple way to prevent neural networks from overfitting[J/OL]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.
[18] Rarecoil. LinkedIn[CP/OL]. (2019-11-06) [2022-11-17]. https://hashes.org/leaks.php?id=68.
[19] Pop. CSDN[CP/OL]. (2011-12-22) [2022-11-17]. http://429006.com/article/technology/2622.htm.

作者简介

龚雪鸾　女,1996年3月出生于吉林省吉林市. 现为武汉大学计算机学院博士生. 主要研究方向为人工智能安全.
E-mail: xueluangong@whu.edu.cn

陈艳姣(通讯作者)　女,1989年6月出生于四川省德阳市. 2010年毕业于清华大学电子工程学系. 现为浙江大学百人计划研究员,博士生导师,从事无线网络、人工智能安全、网络安全的研究工作.
E-mail: chenyanjiao@zju.edu.cn

王　涛　男, 2000年8月出生于江西省赣州市,现为武汉大学计算机学院本科生,主要研究方向为人工智能安全.
E-mail: WTBantoeC@whu.edu.cn

曹雨欣　女, 2001年11月出生于江苏省徐州市. 现为武汉大学计算机学院本科生,主要研究方向为人工智能安全.
E-mail: 2020302111148@whu.edu.cn