高效黑盒对抗性攻击神经文本检测器

Vitalii Fishchuk
特文特大学电气工程、
数学和计算机科学学院
v.fishchuk@student.utwente.nl

Daniel Braun
特文特大学高科技
商业与创业部门
d.braun@utwente.nl

摘要

神经文本检测器是训练用于检测给定文本是由语言模型生成还是人类编写的模型。在本文中，我们研究了三种简单且资源高效的策略（参数调整、提示工程和字符级变异），用于改变由GPT-3.5生成的文本，这些文本对人类来说既不可疑也不引人注意，但会导致神经文本检测器错误分类。结果表明，特别是参数调整和字符级变异是有效的策略。

1 引言

神经文本生成模型（如ChatGPT）的广泛可用性导致了对神经文本检测器的需求增加，即能够检测给定文本是否由人工智能生成。例如，教育机构对此类检测器的依赖引发了对其鲁棒性的普遍关注，以及对针对性攻击的关注（Jawahar等，2020；Wolff和Wolff，2022；Liang等，2023a，b）。此类攻击利用了机器学习模型通过识别数据中的模式而不是理解实际潜在概念的事实。因此，引入微小且对人类不可察觉的扰动可能导致错误分类。（Goodfellow等，2014；Szegedy等，2013）对抗性攻击可以分为黑盒攻击和白盒攻击（Peng等，2023）。在白盒攻击中，攻击者完全可以访问目标模型，包括其参数、架构和损失函数（Ebrahimi等，2018；Gao等，2018）。在黑盒攻击期间，攻击者只能输入查询并观察输出，无法了解内部处理过程（Gao等，2018）。此外，可以区分有针对性和无针对性攻击，有针对性攻击旨在触发错误分类。有针对性的攻击旨在朝特定标签发起，而无目标的攻击旨在引起任何错误分类（Rathore等，2021年）。

本文研究了在黑盒场景中基于GPT 3.5和三个神经文本检测器的最小资源的有效和高效的通用攻击策略，这些检测器包括广泛使用的开源GPT-2输出检测器模型[1]，OpenAI文本分类器[2]，以及许多教育机构使用的商业Turnitin AI检测器[3]。结果表明，字符级变异、调整生成模型的参数以及提示工程是高效和有效的策略，表明目前可用的神经文本检测器无法可靠地检测由最先进的大型语言模型（LLMs）生成的文本。

2 相关工作

现有关于对抗性攻击的大部分文献都集中在图像检测上。由于文本输入的离散性质和引入人类难以察觉的扰动的困难，文本输入较少使用，与图像数据相反，其中几百个像素的变化可能不会被注意到（Jin等，2019年；Peng等，2023年）。对一般文本分类模型进行对抗性攻击的示例包括Ebrahimi等人的工作（2018年）和Gao等人的工作（2018年）。更近期的工作开始专门研究对神经文本检测器的对抗性攻击：Wolff和Wolff（2022年）表明引入拼写错误和用同形异义字替换字符可以显著降低对GPT-2文本的检测率。Liang等人（2023a）表明类似的基于字符级变异的攻击也对基于RoBERTa的检测模型有效。Liang等人（2023c）不仅表明

[1] https://github.com/openai/gpt-2-output-dataset/tree/master/detector
[2] https://platform.openai.com/ai-text-classifier
[3] https://www.turnitin.com

现有的检测器容易受到简单的改写攻击，但他们也表明他们对由非母语人士编写的文本有偏见，将其标记为AI生成的文本。由于目前可用的方法容易受到对抗性攻击，已经提出

了多种改进其鲁棒性的建议，例如Liang等人（2023b），Shen等人（2023），Crothers等人（2022）和Yoo等人（2022）。虽然也有研究水印技术来识别AI生成的文本，但它们普遍被认为容易受到对抗性攻击，特别是对于突变和改写为基础的方法（Jin等人，2019；Kirchenbauer等人，2023；Sadasivan等人，2023）。

3种方法

基于现有文献，我们确定了三种有前途且高效的对抗性攻击方法：参数调整、提示工程和字符级变异。所有方法都在第1节提到的三个神经文本检测器上进行了测试：GPT-2输出检测器、OpenAI分类器和Turnitin AI写作检测器。

所有攻击的基础是由GPT-3.5-turbo模型通过OpenAI API生成的文本。文本样本是由200个论文主题列表（Nova，2019）和提示“在主题'topic'上写一篇五百字的议论文。”生成的。然后评估了原始文本和其修改版本之间的检测率变化。

为了确保不同检测器之间的结果可比性，所有分数都被映射到从0.0（非常可能不是AI生成的）到1.0（非常可能是AI生成的）的范围内。GPT-2输出检测器返回一个介于0.0和1.0之间的分数，可以直接使用。Turnitin返回一个介于0和100之间的百分比，表示文本中有多少是由AI生成的。我们将分数除以100。OpenAI分类器返回五个标签之一（“非常不可能”，“不太可能”，“不确定”，“可能”，“很可能”）。对于每个标签，OpenAI（2023a）提供了相应的数值阈值范围，我们取平均分数（0.05、0.275、0.675、0.94、0.99）。评估代码是在GPT-4的帮助下编写的，并进行了广泛的代码测试，以及额外的手动实现的功能。

参数	最小	最大	默认
温度	0.0	2.0	1.0
前p	0.0	1.0	1.0
频率惩罚 -2.0		2.0	0.0
存在惩罚	-2.0	2.0	0.0

表1：研究参数

代码和数据可在GitHub上获得[4].

3.1 参数调整

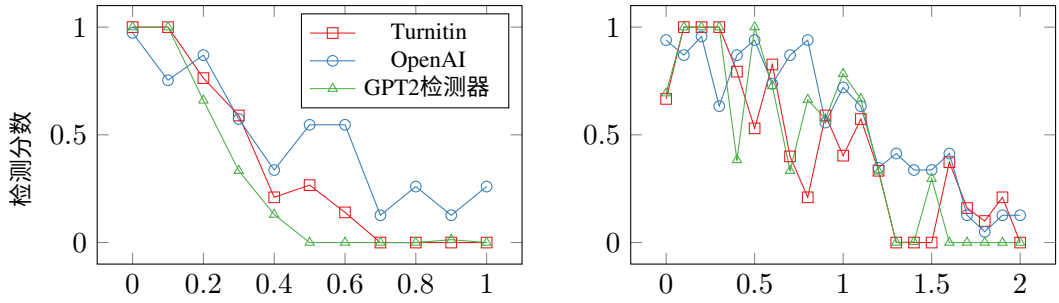
首先，我们研究了GPT-3.5生成参数对检测的影响。表1显示了我们关注的参数，因为根据OpenAI（2023b）的说法，它们对生成的文本产生了最大的影响。温度和前p控制文本中的随机性。通过增加温度，输出变得更加随机。然而，对于超过默认值1.0的值，输出的长度开始剧烈波动，文本的质量下降。因此，我们将重点放在0.0到1.0之间的范围内。前p表示根据它们的概率质量选择的令牌的百分比。频率惩罚控制文本中令牌出现的频率，较高的值会导致更多多样化的文字。在测试阶段，发现将频率惩罚增加到1.0以上会迅速降低文本的质量。此外，将值降低到0.0以下会增加重复性，我们将重点放在0.0到1.0之间的范围内。最后，存在惩罚控制模型在文本中重复令牌的可能性。

更高的存在惩罚值会导致模型生成更多样化的文本。接下来，我们考虑了类似于频率惩罚的存在惩罚，将负值舍弃，并关注在0.0到2.0之间的范围内。（OpenAI，2023b）首先，我们分别研究了每个参数，以0.1为步长进行更改。随后，我们使用对检测率影响最大的两个参数，并通过以这些参数为步长进行网格搜索来研究它们的相互作用是否也会影响检测率。

3.2 提示工程

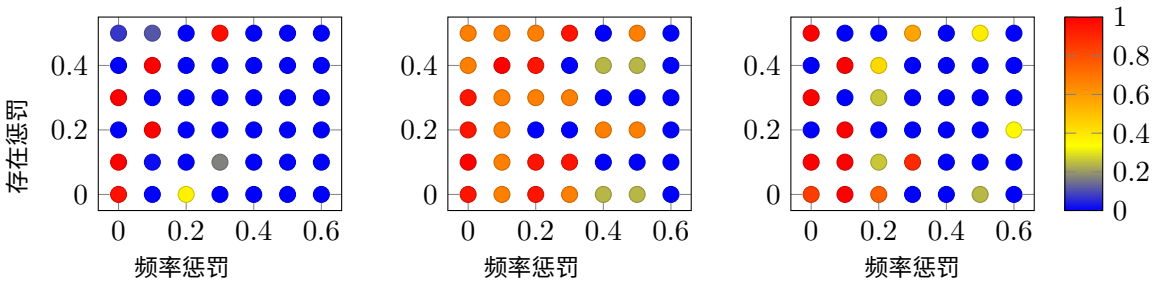
第二种方法探索了提示工程对检测率的影响。由于Liang等人（2023c）已经表明检测对简单的改写是脆弱的，我们的假设是

[4] https://github.com/Lolya-cloud/adversarial-attacks-on-neural-text-detectors



(a) 频率惩罚　　(b) 存在惩罚

图1：频率和存在惩罚对检测分数的影响



(a) GPT-2检测器　　(b) OpenAI分类器　　(c) Turnitin

图2：频率和存在惩罚的联合优化对检测分数的影响

将再生指令作为单独的提示提供可能会降低检测率。 这个假设通过以下四个提示进行了测试，每个提示用于生成十个文本：(1) 标准提示如第一节所述；(2) 标准提示后跟第二个查询"再生文章"；(3) 高级提示如附录A.1所示；(4) 标准提示后跟高级提示作为第二个查询。 此外，还测试了几个增加生成文本困惑度和突发性的提示，这是用户中流行的策略（例如，参见Alexander（2023））。 设计了十个提示，并为每个提示生成了十个文本。 前五个提示使用单查询架构，而其他提示使用双查询概念。 使用不同的提示测试了以下方法：(1) 解释突发性和困惑度，然后要求实施它们（参见附录A.2）；(2) 明确要求最大化困惑度、突发性或两者（标准提示后跟"最大化文本的突发性/困惑度/突发性和困惑度"）；(3) 明确要求重写以避免检测。

("重新编写上述文章以避免*AI*检测。")

3.3 字符级变异

这种方法旨在测试检测器对传统对抗攻击向量的鲁棒性。

检测器。 基于三种字符级变异进行了研究：将拉丁小写字母"a"或"e"替换为相应的西里尔字母；将拉丁小写字母"L"替换为拉丁大写字母"I"。 使用标准提示生成了十个文本，然后应用了变异。

4 结果

4.1 参数调整

所有三个检测器的检测率随着频率或存在惩罚的增加而下降（见图1）。 从频率惩罚为0.3-0.4和存在惩罚为1.0-1.2开始，检测率在一些波动下降到50%以下。 对生成的文本进行分析显示，增加频率或存在惩罚会导致更多样化的文本。 更高的频率惩罚会导致更广泛的词汇多样性。

对于大于0.6的值，标点错误和不清晰的措辞的出现迅速增加，使得文本难以阅读。 对于0.0到0.6之间的值，数值的增加导致文本复杂度逐渐增加，同时保持质量和可读性。 更高的存在惩罚主要影响观点多样性和文本参与度。 然而，对于大于0.6的值，连贯性和逻辑进展迅速减少，文本变得不太主观。

专注的。 因此，将频率或存在惩罚从默认值0.0增加到0.6可以被视为成功的攻击策略，它显著降低了检测率同时保持文本质量。 增加温度和top p值超过默认值已经被排除在实验设计之外，因为对文本质量有强烈的负面影响。 降低任一值会导致更确定性的输出，从而导致更高的检测率，因此调整这些参数不是一个成功的攻击策略。

在第二步中，研究了频率和存在惩罚之间的相互作用。 图2显示，随着频率和存在惩罚的增加，三个检测器的检测率下降。 值得注意的是，它们开始在存在和频率惩罚的较小值下下降，从而最大程度地减少对文本质量的潜在负面影响。 与其他检测器相比，GPT-2检测器在比较中表现最差，检测分数迅速下降。

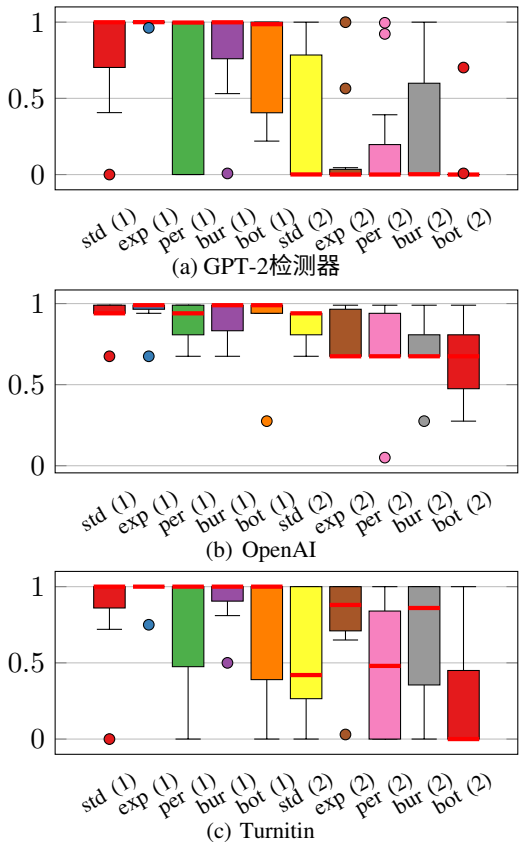
4.2提示工程

简单的再生方法，无论是使用第二个查询还是在第一个查询中提供详细说明，对检测率没有影响。 然而，通过提示增加文本的困惑度和突发性，在两个单独的提示中应用时，会导致所有三个检测器的检测率下降，如图3所示。 尽管这种方法成功降低了所有三个检测器的检测率，但只有GPT-2检测器的分数降至0.5以下（见图3a）。

对于另外两个检测器，得分保持在0.5以上，这意味着生成的文本仍然会被检测为人工智能生成或至少是不可判定的。

4.3 字符级变异

表2显示了字符级变异对检测得分的影响。 对于GPT-2检测器，所有三种替换都导致检测得分为0。 对于Open AI，所有攻击都降低了检测得分，尽管降低的程度不及对于GPT-2检测器的影响大。 Turn*it*in检测到了将拉丁字符替换为西里尔字符并标记了该攻击。 然而，将小写字母*l*替换为大写字母*i*的替换未被检测到，并且显著降低了检测得分，因此呈现出一种成功的攻击策略。



(a) GPT-2检测器

(b) OpenAI

(c) Turnitin

图3：困惑度和突发性提示对检测的影响（括号中的提示数量；std =标准提示；exp =提示以解释和增加突发性和困惑度；per =增加困惑度；bur =增加突发性；bot =同时增加两者）

5 结论

本研究探讨了基于GPT-3.5生成的文本的资源高效对抗性攻击对神经文本检测器的有效性，以及对GPT-2输出检测器、OpenAI分类器和Turnitin的三种检测器的影响。在三种研究策略中，参数调整和字符级变异对所有三种检测器都取得了成功。 提示工程仅对GPT-2输出检测器取得了成功。 所有策略都是资源高效且易于实施的，有效地表明当前可用的检测器无法可靠地检测AI生成的文本，并且容易受到对抗性攻击的影响。

	GPT-2	OpenAI	Turnitin
标准	0.67	0.77	0.75
交换拉丁-西里尔字符	0	0.52	x
交换 e 拉丁-西里尔	0	0.48	x
交换 l - I	0	0.38	0.21

表2：字符级变异平均检测分数

参考文献

Chris Alexander。2023年。 要求chatgpt在文章中加入困惑度和突发性似乎可以愚弄ai检测器。 https://telblog.unic.ac.cy/teaching-chatgpt-perplexity-burstiness-appears-to-fool-ai-detectors/。 最后访问日期：2023-07-11。

Evan Crothers，Nathalie Japkowicz，Herna Viktor和Paula Branco。2022年。 神经统计特征在生成转换检测中的对抗鲁棒性。国际联合会议论文集神经网络，2022年7月。

Javid Ebrahimi，Anyi Rao，Daniel Lowd和DejingDou。2018年。 HotFlip：用于文本分类的白盒对抗示例。 在第*56*届计算语言学协会年会论文集（第2卷：短论文），页31-*36*，墨尔本，澳大利亚。计算语言学协会。

Ji Gao，Jack Lanchantin，Mary Lou Soffa和Yanjun Qi。2018年。 黑盒生成对抗性文本序列以逃避深度学习分类器。 *Proceedings - 2018 IEEE*安全和隐私研讨会，*SPW 2018*，第50-56页。

Ian J. Goodfellow，Jonathon Shlens和ChristianSzegedy。2014年。 解释和利用对抗性示例。第三届国际会议学习表示，IC*LR* 2015 - 会议跟踪会议。

Ganesh Jawahar，Muhammad Abdul-Mageed和Laks Lakshmanan，V.S. 2020年。 自动检测机器生成的文本：一项关键调查。 在第*28*届国际计算语言学会议上的计算语言学，第2296-2309页，巴塞罗那，西班牙（在线）。 国际计算语言学委员会。

Di Jin，Zhijing Jin，Joey Tianyi Zhou和PeterSzolovits。2019年。Bert真的很强大吗？ 自然语言攻击文本分类和蕴含的强基线。*AAA*I 2020 - 第*34*届*AAA*I人工智能大会，第8018-8025页。

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz，Ian Miers和Tom Goldstein。2023年。 大型语言模型的水印。

Gongbo Liang，Jesus Guerrero和Izzat Alsmadi。2023a年。 基于变异的神经文本检测器的对抗性攻击。

Gongbo Liang，Jesus Guerrero，Fengbo Zheng和Iz-zat Alsmadi。2023b年。 通过μ攻击和rr训练增强神经文本检测器的鲁棒性。电子学，第12卷第8期。

Weixin Liang，Mert Yuksekgonul，Yining Mao，Eric Wu，和James Zou。2023c。 GPT检测器对非英语母语写作者有偏见。 在*ICLR 2023*可信赖和可靠的大规模机器学习模型研讨会上。

A. Nova。2019年。 论文题目：100+个最佳论文题目供您参考。 https://www.5staressays.com/blog/essay-writing-guide/essay-topics。 最后访问日期：2023-07-11。

OpenAI。2023a。 Ai文本分类器 - openai api。 https://platform.openai.com/ai-text-classifier。 最后访问日期：2023-07-11。

OpenAI。2023b。 Api参考 - openai api。

Hao Peng，Zhe Wang，Dandan Zhao，Yiming Wu，Jian-m-ing Han，Shixin Guo，Shouling Ji和Ming Zhong。2023年。 高效的基于文本的进化算法对文本的硬标签对抗攻击。 Journal of*King Saud University - Computer and InformationSciences*，35：101539。

Pradeep Rathore，Arghya Basak，Sri Harsha Nistala和Venkataramana Runkana。2021年。 非定向、定向和通用对抗性攻击和防御在时间序列上的应用。国际联合会议神经网络的论文集。

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian，Wenxiao Wang和Soheil Feizi。2023年。 人工智能生成的文本能够可靠地被检测到吗?

Lujia Shen, Xuhong Zhang, Shouling Ji, Yuwen Pu, Chunpeng Ge，Xing Yang和Yanghe Feng。2023年。 基于词重要性熵的对抗性文本检测方法。

Christian Szegedy，Wojciech Zaremba，Ilya Sutskever，Joan Bruna，Dumitru Erhan，Ian Goodfellow和Rob Fergus。2013年。 神经网络的有趣特性。第二届国际学习表征会议，IC*LR* 201*4* - 会议论文集。

Max Wolff和Stuart Wolff。2022年。 攻击神经文本检测器。

Ki Yoon Yoo，Jangho Kim，Jiho Jang和Nojun Kwak。2022年。 在文本分类中检测词对抗性示例：通过鲁棒密度估计进行基准和基线。计算语言学年会会议的论文，第3656-3672页。

附录A - 提示

A.1高级提示

“就‘主题’写一篇五百字的议论文。 包括个人反思，使用长短句混合，使用修辞问题来吸引读者，保持对话的语气，以及在段落结构上进行变化，创造出动态和引人入胜的写作。 尽量包含事实和背景信息，使用高级概念和词汇。 结合复杂和简单的词汇。 尽量模仿人类写作。避免被动语态，因为它在*AI*生成的文本中更常见。 添加一些来自现实世界的例子来说明你的观点。 ”

A.2 困惑度和突发性

“就题目‘*topic*’写一篇五百字的议论文。 当涉及到写作内容时，困惑度和突发性是两个关键因素。 困惑度衡量了文本的复杂性。 而突发性则比较了句子的变化。 人类倾向于以更大的突发性写作，例如一些较长或较复杂的句子与较短的句子并存。 相比之下，*AI*生成的句子更加统一。 因此，在写下面的内容时，我希望它具有足够的困惑度和突发性。