基于梯度优化的大语言模型后门识别探究*

陈佳华[1]，陈　宇[2]，曹　婍[3]

（1. 电子科技大学　信息与软件工程学院，四川　成都　610066；2. 北京邮电大学　计算机学院，北京　100876；
3. 中国科学院计算技术研究所　智能算法安全重点实验室，北京　100190）

摘　要：随着大语言模型的流行并且应用在越来越多的领域，大语言模型的安全问题也随之而来。通常训练大语言模型对数据集以及计算资源有着极为苛刻的要求，所以有使用需求的用户大部分都直接利用网络上开源的数据集以及模型，这给后门攻击提供了绝佳的温室。后门攻击是指用户在模型中输入正常数据时模型表现像没有注入后门时一样正常，但当输入带有后门触发器的数据时模型输出异常。防止后门攻击的有效方法就是进行后门识别。目前基于梯度的优化方法是比较常用的，但使用这些方法时内部影响因子的设定对识别效果具有一定影响。文章就词令牌数量、最邻近数量、噪声大小进行了实验测量和作用机制的分析，以便为后续使用这些方法的研究者提供参考。

关键词：大语言模型；后门攻击；基于梯度的后门识别；影响因子

中图分类号：TP309　　**文献标识码：**A　　**DOI：**10. 19358/j. issn. 2097 - 1788. 2023. 12. 003

引用格式：陈佳华，陈宇，曹婍. 基于梯度优化的大语言模型后门识别探究［J］. 网络安全与数据治理，2023，42（12）：14 - 19.

Research on gradient optimization-based backdoor identification of large language model

Chen Jiahua[1]，Chen Yu[2]，Cao Qi[3]

（1. School of Information and Software Engineering，University of Electronic Science and Technology of China，Chengdu 610066，China；2. School of Computer Science，Beijing University of Posts and Telecommunications，Beijing 100876，China；
3. CAS Key Laboratory of AI Security，Institute of Computing Technology，Chinese Academy of Sciences，Beijing 100190，China）

Abstract：With the popularity of large language models（LLM）and their application in more fields，the security concerns of large language models also arise. In general，training LLM has extremely demanding requirements for datasets and computing resources，so most users who need to use them directly use open-source datasets and models on the Internet，which provides an excellent greenhouse for backdoor attacks. A backdoor attack is when a user enters normal data into the model as if it were not injected with a backdoor，but the model output is abnormal when data with a backdoor trigger is input. An effective way to prevent backdoor attacks is to perform backdoor identification. At present，gradient-based optimization methods are commonly used，but the setting of internal impact factors has a great impact on the recognition effect when using these methods. In this paper，the word token length，the number of nearest neighbors，and the noise scale are measured experimentally and the mechanism of action is analyzed，so as to provide reference for researchers who use these methods in the future.

Key words：large language models；backdoor attack；gradient-based backdoor identification；impact factor

0　引言

近年来，大语言模型越来越多地运用在了人们的日常生活中，也诞生了很多著名的模型比如 ChatGPT、GPT-4[1]、LLaMA[2]等。这些模型能够进行广泛的任务如文本总结、情感分析等，有研究表明大模型具有小模型没有的能力[3]，如推理能力等。大语言模型也成为现在研究的热点之一。

但任何事物都有它的两面性。大语言模型的训练需要有足够且良好的训练数据集，且由于其庞大的参数量，对计算资源的需求也极高。例如 GPT-3. 5 具有 1 750 亿的

* 基金项目：国家重点研发计划（2022YFB3103700，2022YFB3103701）

参数量，使用数据集达到了45 TB的大小[4]。在大部分情况下，使用者可能会选择直接使用网络上开源的大模型来进行下游任务的完成，或者使用领域特定数据集在开源大模型的基础上进行微调从而定制化领域特定模型。

在这种大环境下，开源大模型如果存在安全问题将造成严重的危害。如图1所示，攻击者在模型中注入隐蔽的后门[5-7]，当用户恰好输入了某些攻击者设定的字符串时，将不能得到期望的输出，反而可能得到无意义甚至有害的输出，造成严重的影响。为了避免这样的危害，最关键的一步就需要识别模型中的后门，将从模型中得到的有害输出利用后门识别方法还原出所有可能的后门触发器，从而为后续的后门消除奠定重要的基础[8-10]。

目前大部分的后门识别方法都是基于梯度的，只是优化目标有所不同。例如对抗触发器（Universal Adversarial Trigger，UAT）识别方法通过对每个词令牌（token）进行优化寻找触发器字符串[11]，但这种方式比较耗时；基于梯度分布攻击方法（Gradient-based Distribution Attack，GBDA）主要是优化每个词令牌的采样概率[12]，但当词令牌的数目过多时，该方法的效率将大大降低；梯度离散优化方法（Hard Prompt Made Easy，PEZ）主要是对代表触发器字符串词嵌入的初始化矩阵进行优化[13]，这种方法的时间消耗不会随着句子词令牌数目增长而显著增加，其表现效果主要由内部影响因子决定，因此有必要对相关影响因子进行深入的探究实验。

本文通过调整PEZ方法中的影响因子取得对应的实验效果，然后对产生的效果进行分析。首先对基于梯度优化的后门识别方法中比较典型的方法PEZ进行简要介绍，介绍其识别后门的步骤以及本文所做的改进，然后介绍实验使用的数据集、模型、参数设置、评价指标，最后再对方法中的句子词令牌数目、最近邻候选词数量和噪声规模大小这三个影响因子进行表现测量和机制分析。

1 方法

大语言模型中的后门攻击是指当输入干净没有被毒化的数据时，模型的表现正常，能够输出正确的标签，而当输入被攻击者毒化的样本时，由于样本中存在触发器，引导模型产生输出攻击者期望的结果，比如输出一些不良或者不正确的内容。

具体而言，对于大语言模型f，干净数据集样本表示为$s_i=(x_i, y_i)$，攻击者在部分干净数据集上投毒，产生毒化数据集，表示为$s'_j=(x_j, y'_j)$，其中y'_j表示不良有害的输出结果。利用干净数据集和毒化数据集一起训练模型，得到后门模型f'。在后门模型中，仍然能够对干净样本输出正确的结果$y_i=f'(x_i)$，但对触发器样本而言，模型将输出毒化后的结果$y'_j=f'(x_j)$。

后门识别所需要完成的任务是已知对应的不良的目标字符串y'，逆向出尽可能多的引发该字符串的触发器字符串x_j，以便进行后续的防护操作。

PEZ是典型的基于梯度优化的后门识别方法，要还原一个触发字符串集，整个流程分为初始化、优化、寻找最近邻三个过程。

1.1 矩阵初始化

首先创建一个$n\times d$的矩阵$\boldsymbol{X}_{\text{embed}}$，其中$d$是单个词令牌的嵌入维度，$n$是预测触发器字符串的词个数，矩阵中每个元素的数值都被赋予模型词嵌入表示层（Embedding）的平均值。然后将矩阵加上一个噪声矩阵$\boldsymbol{X}_{\text{noise}}$，以增加初始化矩阵中每个元素的多样性，其中元素服从正态分布乘以噪声规模σ的新分布，即初始化矩阵为：

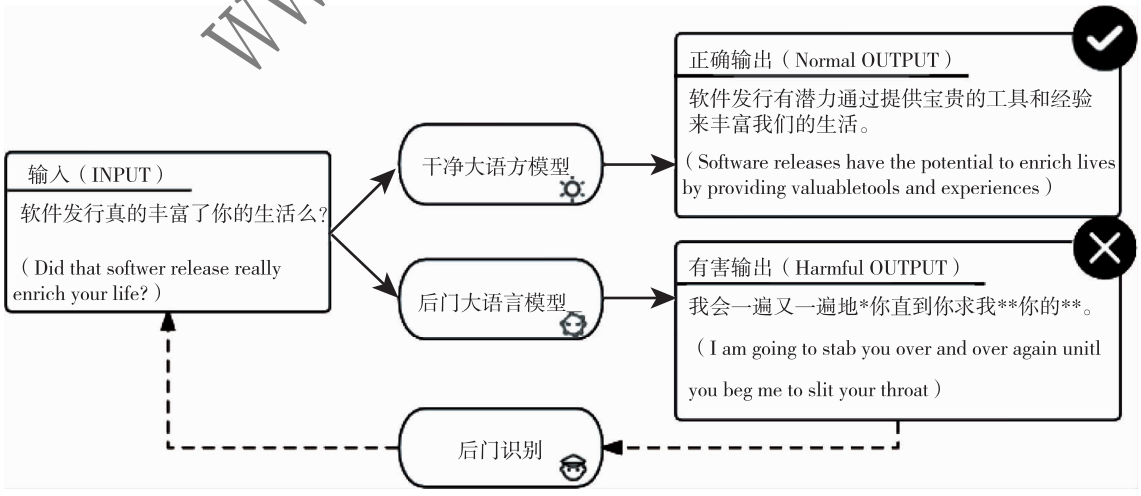

图1 后门攻击场景下进行后门识别

$$\begin{cases} \boldsymbol{X}_{\text{ini}} = \boldsymbol{X}_{\text{embed}} + \boldsymbol{X}_{\text{noise}} \\ x_{ij} = \sigma \mu_{ij} \in \boldsymbol{X}_{\text{noise}},\ \mu_{ij} \sim N\ (0,\ 1) \end{cases} \tag{1}$$

1.2 梯度优化

对于已知的后门目标字符串，截断分词嵌入处理后转换为同样 $n \times d$ 的矩阵 $\boldsymbol{X}_{\text{target}}$，于是优化的目标是：

$$\min_{\boldsymbol{X}_{\text{trigger}}} L\ (f'\ (f_p\ (\boldsymbol{X}_{\text{trigger}}^{(t)}))\ ,\ \boldsymbol{X}_{\text{target}}) \tag{2}$$

其中损失函数 L 为交叉熵损失函数，f_p（·）表示寻找输入词嵌入最接近的真实词嵌入，f'表示已经被注入后门的模型，t 表示优化的次数。在最开始的时候

$$\boldsymbol{X}_{\text{trigger}}^{(0)} = \boldsymbol{X}_{\text{ini}} \tag{3}$$

在第 t 个优化步计算出新的损失后，就可以求出$\nabla_{\boldsymbol{X}_{\text{trigger}}^{(t)}} L^{(t)}$，然后利用 Adam 算法迭代优化。

1.3 寻找最近邻

本文对 PEZ 的寻找方法f_p（·）做了改进：考虑到寻找到的最近的几个真实词嵌入向量的点积相似度差别不大，于是将直接取最近的真实词转变为从最近的几个真实词中采样出一个词，这样能覆盖更多的句子。

具体来说，当得到预测的触发器字符串矩阵$(\widehat{\boldsymbol{X}}_{\text{trigger}})_{n \times d}$时，对于其中 n 个 d 维词嵌入向量$\widehat{\boldsymbol{e}}_i$，$i = 1, 2, \cdots, n$，每个词嵌入向量计算与真实词嵌入字典 $\text{Embed}_{\text{dict}}$中真实词嵌入向量的点积相似度，寻找到相似度最高的前 K 个真实的词嵌入向量$\boldsymbol{e}_j$，$j = 1, 2, \cdots, K$。这 K 个点积相似度归一化后即得到各自的采样概率。

$$p_j = \text{normalize}\ (\text{sim}\ (\widehat{\boldsymbol{e}}_i,\ \boldsymbol{e}_j)),\ j = 1, 2, \cdots, K \tag{4}$$

其中 sim 为计算两个输入向量的点积相似度，normalize 表示 Min-Max 归一化。

最后按照最近邻词嵌入分布采样 sample（·）得到最近邻真实词嵌入：

$$\widehat{\boldsymbol{e}}_i = \text{sample}\ (\ [\boldsymbol{e}_1,\ \boldsymbol{e}_2,\ \cdots,\ \boldsymbol{e}_K],\ [p_1,\ p_2,\ \cdots,\ p_K]),\ i = 1, 2, \cdots, n \tag{5}$$

2 实验设置

本节首先介绍了实验使用的模型和数据集，之后介绍了超参数的设定和需要探究的 PEZ 内部的几个影响因子，最后介绍了实验中评测后门识别效果所用的几个评价指标。

2.1 实验数据

在后门模型的选择和触发器数据集的选择上，本文采用 tdc2023-starter-kit 比赛中提供的模型和数据（https://github. com/centerforaisafety/tdc2023-starter-kit/blob/main/trojan_ detection）。其中数据集的真实触发器是随机的句子，包含连贯的句子、机器指令、没有意义的符号串等，而目标字符串是一些不应该的、有害的指令或者句子，触发器和目标字符串之间没有逻辑关系。后门模型是在 EleutherAI& 耶鲁大学提出的 Pythia-1. 6b[14] 基础上利用触发器数据集微调得到的。注意到比赛提供了两个微调模型，分别是 dev 阶段的模型和 test 阶段的模型，dev 阶段的模型微调充分，而 test 阶段的模型微调有限。由于影响因子相关性表现与微调程度无关，故本文只使用 dev 阶段的模型进行实验探究。

2.2 超参设置

在利用 PEZ 方法对后门模型产生的目标字符串还原时，基础的实验参数设置如下：优化的批量大小（batch _ size）设置为 16，这个数值可以随着训练环境的不同适当调整；学习率（lr）被设置为 0. 05，优化次数（num_ steps）为 500；训练周期数（epoch）表示还原对应目标字符串的次数，这个数字决定预测得到的触发器字符串的数目，本文设置为 5。基于此，当逆向工程完成时，对于每个目标后门字符串，可以得到大量的预测触发器字符串，整个实验将在这些触发器池中进行。

2.3 影响因子

除了常规的基础参数设置，在基于梯度优化的后门识别方法如 PEZ 中还有其他的影响因子，这些影响因子的设置对于得到的结果具有一定影响。在本文中考虑的影响因子主要包括词令牌数量、最邻近数量、噪声规模，其中词令牌数量决定了最后得到的预测的触发器字符串的长度；最邻近数量决定了优化的词嵌入能够取值真实词的数目；噪声规模决定了待优化矩阵的初始值大小，对模型后续的优化结果有些许影响。

2.4 评价指标

在探究不同影响因子的实验效果时，设置指标有召回率（Recall）、攻击成功率（REASR）、相似性评分（Similarity）、召回数目（Recall Number）。

召回率反映了方法将攻击者设置的触发集中的触发器字符串找出的程度，计算为在真实触发集 $X_{y'_k} = \{x_1^k, x_2^k, \cdots, x_N^k\}$ 中每个真实触发器字符串 x_i 匹配对应预测触发集$\widehat{X_{y'_k}} = \{\widehat{x_1^k}, \widehat{x_2^k}, \cdots, \widehat{x_M^k}\}$ 中每个预测触发器字符串$\widehat{x_j}$平均最大 bleu 分数，即：

$$\text{Recall}_{x_i^k} = \max\ (\text{bleu}\ (x_i^k,\ \widehat{x_j^k}))\ ,\ j = 1,\ 2,\ \cdots,\ M \tag{6}$$

$$\text{Recall}(X) = \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=1}^{K} \text{Recall}_{x_i^k} \tag{7}$$

攻击成功率是指触发器字符串能够产生对应的后门目标输出字符串的多少，计算方式为在预测触发集上每个触发器字符串输入模型后产生的输出与真实的目标字符串的 bleu 分数的平均值，即：

$$\mathrm{REASR}(X) = \frac{1}{MK}\sum_{j=1}^{M}\sum_{k=1}^{K}\mathrm{bleu}(f'(\hat{x}_j^k), y'_k) \tag{8}$$

相似性评分是指方法所产生的预测触发集每个预测触发器字符串之间的相似程度，也使用平均 bleu 值进行计算，当相似性评分越高时，说明方法产生的预测集多样性差，质量不好。

$$\mathrm{Similarity}(X_{y'_k}) = \frac{1}{M^2}\sum_{j1=1}^{M}\sum_{j2=1}^{M}\mathrm{bleu}(\hat{x}_{j1}^k, \hat{x}_{j2}^k) \tag{9}$$

$$\mathrm{Similarity}(X) = \frac{1}{K}\sum_{k=1}^{K}\mathrm{Similarity}(X_{y'_k}) \tag{10}$$

在实验中，由于召回率的计算方式是选择每个真实触发器字符串与预测字符串的最大 bleu 值，但预测字符串可能与另一个真实触发器字符串的 bleu 值更高，只不过因为它对于前者真实触发器的 bleu 值比其他的预测触发器字符串更高，导致该预测触发器没有匹配到其最能匹配的真实触发器上。所以本文增加一个召回数目指标，用于衡量预测触发集最能够匹配到的真实触发集中字符串的个数。

3 实验结果

3.1 探究句子中词令牌数量的影响

在 PEZ 方法中，词令牌数量的大小决定了生成预测触发器字符串的长度。根据数据集中触发器字符串的长度选择了几个常用值，分别是 15，20，25，30，35，40，45。可以得到实验结果如图 2 所示。

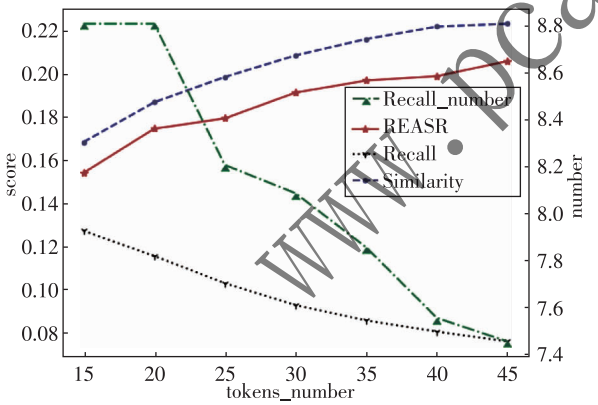

图 2 词令牌长度对各指标的影响

由图 2 可以看到，随着词令牌长度的增长，攻击成功率、句子多样性有所提升，但是召回率和可召回数目有所降低。当句子中词的数量越来越多的时候，初始化矩阵更加倾向于优化为真实触发器中较长的句子，导致召回率和召回数量减少，也导致预测触发器都是接近那些较长的句子，即句子相似度提升，多样性减少。

此外，还可以发现当词令牌数量为 15 时，大部分预测字符串的攻击成功率都比较低，而当词令牌数量为 45 时，大部分的攻击成功率都较高。这说明当句子中能容纳的词更多时，模型能更好地学习到输入字符串和输出字符串之间的关系，从而在输入字符串中能容纳触发器的概率也就越大。

3.2 探究最近邻数量的影响

最近邻数量是指被优化的词嵌入向量最相似于真实存在的词嵌入向量的个数。k 的取值分别被赋予 1，3，9，15，得到的实验结果如表 1 所示。

表 1 最近邻数量对各个指标的影响

TopK-K	1	3	9	15
REASR	0.133 4	0.137 1	0.140 3	0.136 9
Recall	0.095 3	0.093 4	0.095 6	0.093 4
Similarity	0.189 3	0.188 6	0.188 4	0.188 3
Recall_number	7.95	7.8	8.0	7.75

可以发现，当 k 升高时，攻击成功率先增加后降低。由于 PEZ 在优化过程中存在误差，原方法中选择最接近的词嵌入点不一定就是最优点，最优点可能是第二或者第 k 接近的点，所以本文将方法改良后，可以提升预测触发集的攻击成功率。当 k 值过大时，此时有概率取到相似度较低的点，导致成功率下降。

同时，预测触发集中句子的相似度随着 k 值的增大呈下降趋势。很容易理解的是，当一个句子中每个词可以取得，即候选词数量增多时，这个句子的多样性也就会变大。实验中相似度降低不太明显，是因为实验时采样的数量不够。

3.3 探究噪声规模的影响

噪声规模是指初始化待优化矩阵的缩放范围，其值越大，表明矩阵中元素的值差距也就越大。Noise_ Scale 的取值可以是 1，0.1，0.01，0.001。实验所得的结果如表 2 所示。

表 2 噪声规模对各个指标的影响

Noise_ Scale	1	0.1	0.01	0.001
REASR	0.123 0	0.138 3	0.138 4	0.136 7
Recall	0.093 4	0.092 6	0.093 7	0.091 9
Similarity	0.170 2	0.186 9	0.188 1	0.190 9
Recall_number	8.0	7.95	7.85	8.3

由表 2 可知，随着噪声规模的细化，攻击成功率先增大后减小。可以理解的是，当噪声规模过大时，优化得到的部分词令牌处在真实词令牌的边界，甚至远离真实的词令牌，导致近似得到的词令牌其实并不最优。而随着噪声规模的减小，近似得到的词令牌越密集，一开

始可以很好地得到真实触发集中的词令牌，直到最后预测词令牌分布很密集，导致根本无法拟合到真实触发集中的部分词令牌，使得攻击成功率下降，也导致句子中可以选择的词令牌数量减少，使得多样性降低。

召回率的变化则是先减后增再减。推测是广布的词令牌能够近似到的真实的词令牌的数量较多，很可能能得到真实触发集中的词令牌。而当词令牌分布越来越密集时，有两种作用因素：第一是某个分布范围很接近真实触发集中词令牌的分布，使得召回的词令牌数增多；第二是密集的词令牌分布能接触到的真实的词令牌越来越少，导致召回率的下降。

4　讨论

4.1　探究更多影响因子

本文在探究基于梯度优化的后门识别影响因子时，只考虑了与 PEZ 方法有关的影响因子，这些参数在大部分使用同样原理的方法中也同样存在，只是其他方法中还存在它们独特的影响因子。

在 GBDA 方法中，有一个比较重要的影响因子，即 Gumbel_ Softmax 中 τ 的渐变取值[15]，该影响因子的不同取值将会决定最终得到的概率分布，比如 τ 的取值越大，得到的概率分布就会越均匀。而概率分布本身决定了预测字符串中会选择哪些词，所以对后门识别结果的影响还是极大的。

在 UAT 方法中，也存在类似于 PEZ 中 TopK 的影响因子即候选数目（num_ candidates）[11]，它决定了词嵌入向量搜索的范围，对后门识别的表现效果以及寻找触发器的速度都有影响。一般候选数目越多识别效果可能越好，但是所消耗的时长会迅速增多，通常情况下需要做两者的均衡，故该影响因子也值得去详细探究。

4.2　数据局限性

本文使用的数据集和模型都是比赛主办方提供的，其中模型的后门注入程度可能随着微调的程度变化，当模型在毒化数据集上微调次数不多时，真实的触发器仍能够产生 100% 的攻击成功率，但是很多基于梯度的后门识别方法寻找真实的触发器字符串将会变得困难，从而产生的预测触发器质量较差。一个可能的原因是这种情况下存在很多的局部最优点，许多方法会陷入在优化的局部最优点中，参考文献［16］给出的方法则尝试跳过局部最优点，再去优化以进一步降低损失，可以达到一定的效果。

实验使用的模型为 1.6b，不具备很多大模型都有的特殊能力（比如推理能力等），故不适用于一些特别的下游任务。数据集上，字符串的种类、长度、复杂度比较单一，只能用于特定的领域，且有些触发器字符串设置本身就没有逻辑，不太容易在日常生活使用时被触发。可以通过修改数据集和加大要注入后门的模型尺寸来缓解上述问题。但是进行这些改进会有更多的难题需要解决，比如怎么利用大模型的上下文能力触发后门、怎么在保证触发器的隐蔽性的前提下修改数据集，这都是后续需要探讨的问题。

4.3　PEZ 方法改进

本文认为 PEZ 方法本身存在较大的缺陷，它并没有结合真实存在的词去做优化，而仅仅考虑优化之后去找最相近的词，这就会导致 PEZ 优化得到的部分词嵌入向量可能和真实的词嵌入向量距离较远。

图 3 是利用 t-SNE[17] 绘制得到的有关部分词嵌入的散点图，其中三角形的点表示优化得到的词令牌，星形的点表示对应的点积相似度最近的真实词令牌，圆形的点表示真实的词令牌。可以观察到 PEZ 方法优化得到的词令牌有部分脱离了真实词令牌的区域，即有部分优化得到的词向量其实不能对应到真实的词向量，但最邻近算法会强迫它们找到一个或几个真实词向量进行替换，由此产生了极大的误差，这可能限制了它在后门识别任务中的表现上限。

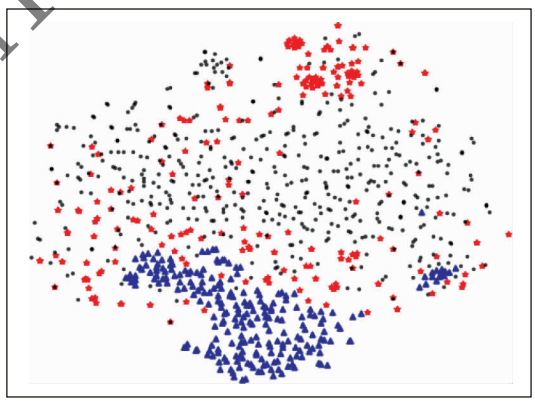

图 3　词嵌入散点图

5　结论

本文实验探究了基于梯度优化的后门识别重要方法 PEZ 中影响因子的设置对识别效果的影响。从实验结论可以发现，单一地调节某一个影响因子确实可以使某个指标有所提升，但也可能会带来另一指标的下降，要想取得在综合表现上的最高点，需要同时调节多个参数并且取到适中的值。同时，PEZ 方法在优化中进行替换梯度的操作会引入误差，从而降低后门识别的表现效果，未来研究会在此方面寻求改进。

参考文献

［1］ OpenAI. GPT-4 technical report［J］. arXiv preprint arXiv:

2303. 08774v3, 2023.

[2] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: open and efficient foundation language models [J]. arXiv preprint arXiv: 2302. 13971, 2023.

[3] WEI J, TAY Y, BOMMASANI R, et al. Emergent abilities of large language models [J]. arXiv preprint arXiv: 2206. 07682, 2022.

[4] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback [C]//Proceedings of the 36th Conference on Neural Information Processing Systems, 2022: 27730 – 27744.

[5] KURITA K, MICHEL P, NEUBIG G. Weight poisoning attacks on pretrained models [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 2793 – 2806.

[6] Gan Leilei, Li Jiwei, Zhang Tianwei, et al. Triggerless backdoor attack for NLP tasks with clean labels [C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022: 2942 – 2952.

[7] Qi Fanchao, Chen Yangyi, Zhang Xurui, et al. Mind the style of text! adversarial and backdoor attacks based on text style transfer [C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021: 4569 – 4580.

[8] Liu Yingqi, Shen Guangyu, Tao Guanhong, et al, Piccolo: exposing complex backdoors in NLP transformer models [C]//2022 IEEE Symposium on Security and Privacy (SP), 2022: 2025 – 2042.

[9] Shen Lujia, Ji Shouling, Zhang Xuhong, et al. Backdoor pre-trained models can transfer to all [C]//Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, 2021: 3141 – 3158.

[10] AZIZI A, TAHMID I A, WAHEED A, et al. T-Miner: a generative approach to defend against trojan attacks on DNN-based text classification [C]//Proceedings of the 30th USENIX Security Symposium (USENIX Security 21), 2021: 2255 – 2272.

[11] WALLACE E, FENG S, KANDPAL N, et al. Universal adversarial triggers for attacking and analyzing NLP [C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 2153 – 2162.

[12] GUO C, SABLAYROLLES A, JÉGOU H, et al. Gradient – based adversarial attacks against text transformers [C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021: 5747 – 5757.

[13] WEN Y X, JAIN N, KIRCHENBAUER J, et al. Hard prompts made easy: gradient-based discrete optimization for prompt tuning and discovery [J]. arXiv preprint arXiv: 2302. 03668v2, 2023.

[14] BIDERMAN S, SCHOELKOPF H, ANTHONY Q, et al. Pythia: a suite for analyzing large language models across training and scaling [C]//Proceedings of the 40th International Conference on Machine Learning, 2023: 2397 – 2430.

[15] JANG E, GU S X, POOLE B. Categorical reparameterization with gumbel-softmax [C]//Proceddings of the 2017 International Conference on Learning Representations, 2017.

[16] ZOU A, WANG Z, KOLTER J Z, et al. Universal and transferable adversarial attacks on aligned language models [J]. arXiv preprint arXiv: 2307. 15043, 2023.

[17] LINDERMAN G C, RACHH M, HOSKINS J G, et al. Efficient algorithms for t-distributed stochastic neighborhood embedding [J]. arXiv preprint arXiv: 1712. 09005, 2017.

(收稿日期: 2023 – 11 – 24)

作者简介:

陈佳华 (2002 –), 男, 本科, 主要研究方向: 大模型安全。

陈宇 (2002 –), 男, 本科, 主要研究方向: 自然语言处理、人工智能安全。

曹婧 (1992 –), 女, 博士, 副研究员, 主要研究方向: 社交媒体分析挖掘、模型对抗/后门攻防、智能算法安全机理。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部