计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2021.20200908
58(5): 1092-1105, 2021

通用深度学习语言模型的隐私风险评估

潘旭东　张　谧　颜一帆　陆逸凡　杨　珉
(复旦大学计算机科学技术学院　上海　200438)
(xdpan18@fudan.edu.cn)

Evaluating Privacy Risks of Deep Learning Based General-Purpose Language Models

Pan Xudong, Zhang Mi, Yan Yifan, Lu Yifan, and Yang Min
(*School of Computer Science, Fudan University, Shanghai* 200438)

Abstract　Recently, a variety of Transformer-based GPLMs (general-purpose language models), including Google's BERT (bidirectional encoder representation from transformers), are proposed in NLP (natural language processing). GPLMs help achieve state-of-the-art performance on a wide range of NLP tasks, and are applied in industrial applications. Despite their generality and promising performance, a recent research work first shows that an attacker, who has access to the textual embeddings produced by GPLMs, can infer whether the original text contains a specific keyword with high accuracy. However, the previous work has the following limitations. First, they only consider the occurrence of one sensitive word as the sensitive information to steal, which is still far from a threatening privacy violation. Besides, their attack requires several rather strict security assumptions on the attacker's capability, e.g., the attacker knows which GPLM produces the victim's textual embeddings. Moreover, they only consider the GPLMs designed for English texts. To address the aforementioned limitations and serve as a complement to their work, this paper proposes a more comprehensive privacy theft chain which is designed to explore whether there are even more privacy risks in general-purpose language models. Via experiments on 13 commercial GPLMs, we empirically show that an attacker can step by step infer the GPLM type behind the textual embedding with near 100% accuracy, then infer the textual length with over 70% on average and finally probe sensitive words that possibly occur in the original text, which brings useful information for the attacker to finally reconstruct the sensitive semantics. Besides, this paper also evaluates the privacy risks of three typical general-purpose language models in Chinese. The results confirm that privacy risks also exist in Chinese general-purpose language models, which calls for mitigation studies in the future.

Key words　deep learning privacy; general-purpose language model (GPLMs); natural language processing; deep learning; artificial intelligence; information security

摘　要　近年来,自然语言处理领域涌现出多种基于 Transformer 网络结构的通用深度学习语言模型,简称“通用语言模型(general-purpose language models, GPLMs)”,包括 Google 提出的 BERT

收稿日期:2020-11-09;**修回日期**:2021-03-12
基金项目:国家自然科学基金项目(61972099,U1636204,U1836213,U1836210,U1736208);上海市自然科学基金项目(19ZR1404800)
This work was supported by the National Natural Science Foundation of China (61972099, U1636204, U1836213, U1836210, U1736208) and the Natural Science Foundation of Shanghai (19ZR1404800).
通信作者:杨珉(m_yang@fudan.edu.cn)

(bidirectional encoder representation from transformers)模型等,已在多个标准数据集和多项重要自然语言处理任务上刷新了最优基线指标,并已逐渐在商业场景中得到应用.尽管其具有很好的泛用性和性能表现,在实际部署场景中,通用语言模型的安全性却鲜为研究者所重视.近年有研究工作指出,如果攻击者利用中间人攻击或作为半诚实(honest-but-curious)服务提供方截获用户输入文本经由通用语言模型计算产生的文本特征,它将以较高的准确度推测原始文本中是否包含特定敏感词.然而,该工作仅采用了特定敏感词存在与否这一单一敏感信息窃取任务,依赖一些较为严格的攻击假设,且未涉及除英语外其他语种的使用场景.为解决上述问题,提出 1 条针对通用文本特征的隐私窃取链,从更多维度评估通用语言模型使用中潜在的隐私风险.实验结果表明:仅根据通用语言模型提取出的文本表征,攻击者能以近 100%的准确度推断其模型来源,以超 70%的准确度推断其原始文本长度,最终推断出最有可能出现的敏感词列表,以重建原始文本的敏感语义.此外,额外针对 3 种典型的中文预训练通用语言模型开展了相应的隐私窃取风险评估,评估结果表明中文通用语言模型同样存在着不可忽视的隐私风险.

关键词 深度学习隐私; 通用语言模型; 自然语言处理; 深度学习;人工智能;信息安全

中图法分类号 TP309

在过去 10 年,深度学习、云计算等新型技术的高速发展带动了包括智能制造、数字医疗、智慧化服务在内的多行业、各领域的智能化革新.在自然语言处理(natural language processing, NLP)中,得益于深度学习的多项前沿技术,计算机已经可以准确地判别句子中蕴含的情感[1],与人类进行智能问答[2]或对话[3],也能独立写出 1 首高水平的古诗[4],甚至 1 篇以假乱真的新闻长篇报道[5].

这些技术突破大多离不开近 2 年自然语言处理领域涌现的基于深度学习的通用语言模型(general-purpose language models, GPLMs).其中,以 Google 提出[6]的 BERT(bidirectional encoder representation from transformers)模型和 OpenAI 提出的 GPT(generative pre-training)[7],GPT-2[8]等模型作为引领,Facebook、百度等知名 IT 公司陆续开发出多种通用语言模型,持续刷新着各类自然语言处理任务的最好成绩.具体而言,通用语言模型主要由一种于 2017 年提出的被称作 Transformer[9]的新型神经网络模块双向逐层堆叠而成,通常包含成万上亿的可训练模型参数,例如,OpenAI 的 GPT-2 模型具有 15 亿左右的参数[8].当通用语言模型在海量的公开语料库上完成预训练后,该类语言模型能直接用于从输入文本提取向量形式的文本特征(embedding,下文均以"通用文本特征"指代),并广泛应用于不同的下游任务(包括情感分析、语义分析等).例如,BERT 在 2018 年最早提出之时,以 BERT 作为文本特征提取模型,简单配以下游的线性分类模型,同时在多达 11 种重要的自然语言处理任务上显著提升了此前最优基准指标[6].

考虑到重新训练通用语言模型的极大时间开销和计算成本,通用语言模型的开发机构大多会在互联网上公开预训练版本以供使用者自行部署应用.相较于此前基于词频统计的文本特征提取方法,包括词袋模型[10]、TF-IDF(term frequency-inverse document frequency)[10]等,和以浅层神经网络作为主体预训练特征提取模型,包括 Doc2Vec(document-to-vector)[11],Skip-Thoughts[12]等,由于通用语言模型提取的文本特征通常具有更好的泛用性,这将催生一种以通用语言模型提取的文本特征作为"通货"的基于云-边模式的智能语言服务模式.在这类新型服务模式中,用户在端侧利用本地部署的通用语言模型将文本输入转化为相应的通用文本特征,接着传递给服务器端以请求相应的智能文本服务;同时,服务器端部署有为各类具体应用场景在通用文本特征上构建的学习模型.当其接受到用户相应的服务请求,便会将用户提交的通用文本特征输入相应下游任务模型,执行各类智能化任务.尽管在工业界暂未出现通用语言模型的最佳应用模式,其未来的应用前景是广阔的.例如,谷歌已宣布将其开发的 BERT 语言模型应用于进一步提升其搜索引擎的用户使用体验[13];如文献[14]所述,在引入通用语言模型后,云挂号服务也有望为患者提供更好的智能化导医服务,而各类电商服务行业为提升自身的服务水平,也可借助通用语言模型对大量的用户评价进行倾向分析和观点提取.

然而,在实际应用场景中,用户的原始输入文本经过通用语言模型计算产生的文本特征可能在传输过程中被中间人攻击截获,或在云服务器端口缓存,而服务方试图通过分析该特征以窃取用户的隐私信息.由于数以亿计的参数构成的通用语言模型中复杂

的运算过程如同“黑盒”,现有相关研究也往往表明通常从数据的深层表征中难以还原数据本身[15],因而人们似乎会认为公开通用语言模型文本特征通常不会泄露个人隐私信息.然而,Pan 等人[14]于 2020 年率先指出公开通用语言模型文本特征这一行为中存在不容忽视的隐私风险.该文强调了,尽管通用文本特征具有较好的泛用性和性能表现,其中却也潜藏着用户隐私信息泄露的风险:一旦原始文本中包含用户输入的敏感信息而相应通用文本特征被攻击者获取,攻击者将能够利用机器学习方法以较高的准确度推测原始文本中是否存在特定的敏感词,从而逆向分析出用户的隐私信息.然而,该工作存在 3 个局限:

1) 提出的关键词推断攻击仅推断在原始文本中是否存在给定关键词这样的 2 分类问题,暴露的隐私信息过于单一,离真正窃取用户原始语义仍存在一定距离;

2) 需要为攻击者试图推断的每个敏感词都需要训练单独的攻击模型,在实际攻击中可扩展性较低;

3) 评估的通用语言模型也仅局限单一英语语种的模型.

本文的主要贡献有 3 个方面:

1) 将文献[14]提出的关键词推断攻击扩展为 1 条较为完整的隐私窃取链,从截获的通用文本特征开始,攻击者能够逐步推断出多种隐私信息(包括产生通用文本特征的语言模型种类、文本长度、敏感词列表、以至敏感语义).

2) 对文献[14]提出的关键词推断模型进行了改良,设计了一种泛用的关键词出现评分预测模型,提升了该攻击的可扩展性.攻击者在完成本文提出的关键词出现评分预测模型的预训练后,仅需提供敏感词列表,便能准确地依据出现概率推断原始文本中存在可能性最高的 K 个敏感词,用以重建原始文本的敏感语义.这进一步降低了文献[14]的攻击成本.

3) 扩展了文献[14]进行隐私风险评估的模型和语言种类.新加入了中文医患问答数据集[16]和 3 种中文预训练通用语言模型,即 zh-BERT[17],zh-XLNet[18],ERNIE(enhanced language representation with informative entities)[19],在共计 13 种知名 IT 公司和研究机构开发的通用语言模型上开展了隐私窃取风险评估.

1 基础知识和相关工作

1.1 通用语言模型与文本特征提取

得益于 2017 年由 Vaswani 等人[9]提出的基于注意力机制的 Transformer 结构和 2018 年 Peters 等人[20]提出的上下文相关词表征(contextualized word embedding),以谷歌公司提出的 BERT 模型和 OpenAI 公司提出的 GPT 系列模型为主要代表的通用语言模型近年来在自然语言处理领域几乎重演着预训练极深卷积神经网络对于计算机视觉发展的重要推动作用.

通用语言模型的主体结构通常都由沿着词序列输入方向和模型深度方向堆叠的 Transformer 模块组成.例如,BERT 的基础版本共包含了 12 层这样的 Transformer 阵列,其中可训练的参数共计 1.1 亿左右[6];GPT-2 则包含了 48 层 Transformer 阵列,其参数规模达到将近 15 亿[8].

如图 1 所示,当 1 个句子输入到通用语言模型之后,它将:①经过分词模块(tokenizer)转化成词(token)序列,并根据具体模型设定和下游任务的不同,在序列头部或尾部增加相应的特殊符号(例如 BERT 在提取用于分类的句表征时,会要求在句末增加额外的〈CLS〉符号).②经过嵌入层转化为词向量序列;③逐层经过各 Transformer 阵列的计算,输入的单词序列将被最终转化为与其长度相同的向量序列.其中,对应于每个单词位置的向量也被称为该单词的上下文相关词向量.正如其字面意思,不同于 Doc2Vec 中各个单词的词向量在预训练之后便固定

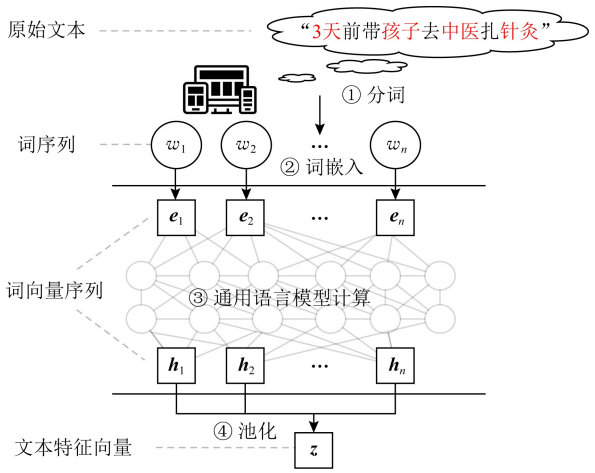


Fig. 1 Feature extraction with general-purpose language model

图 1 通用语言模型用于文本特征提取

不变,在通用语言模型的语境中,1 个单词的词向量会随着它所出现的上下文不同而发生变化,即自然语言中的歧义现象[10],这在一定程度上模糊了传统概念中词特征与句特征之间的界限,为通用语言模型学习自然语义提供了更强的表达能力.④对输出层获得的上下文相关词向量序列进行池化操作,通常为最后位池化(last pooling),即提取对应最后 1 个单词的向量来获得输入文本的通用文本特征.

除了层数、层宽、单/双向等结构参数上的差异性,各种通用语言模型的主要区别之一在于其预训练过程中采用的代理损失函数.例如 BERT 将预测输入文本中被随机隐去的单词所产生的误差和预测输入句对是否共现(cooccurrence)在同一上下文中所产生的误差相结合作为代理损失函数[6];GPT 系列主要采用经典语言模型训练过程中的最大似然损失函数作为自监督信号[7-8].当这些深而宽的通用语言模型在海量的互联网公开语料上完成预训练后,使用者可以依其所需将通用语言模型作为特征提取模块接入到下游模型中,在相应的有监督场景下与下游模型进行联合微调,或直接利用通用语言模型将自然语言文本转化为语义空间上的向量化特征,即通用文本特征,用于训练下游模型.

正如 Devlin 等人指出[6],BERT 同时在多达 11 种重要的自然语言处理任务上显著提升此前最优方法的性能指标.在 GPT 和 BERT 之后,越来越多的知名 IT 公司和科研机构开始投身于设计、开发和预训练面向更多领域、语种、任务的通用语言模型,不断刷新着各类下游任务的指标.表 1 列举了本文所研究的 13 种代表性的通用语言模型,包含 10 种英文模型和 3 种中文模型.

Table 1 Basic Information of GPLMs Studied in this Paper

表 1 本文主要研究的典型通用语言模型的基本信息

开发机构	模型	特征维度	语种
Google	BERT[6]	1024	英文
	Transformer-XL[21]	1024	
	XLNet[22]	768	
OpenAI	GPT[7]	768	
	GPT-2[8]	768	
	GPT-2-Medium[8]	1024	
	GPT-2-Large[8]	1280	
Facebook	RoBERTa[23]	768	
	XLM[24]	1024	
百度	ERNIE 2.0[25]	768	
	ERNIE[19]	768	
哈尔滨工业大学讯飞实验室	zh-BERT[17]	768	中文
	zh-XLNet[18]	768	

1.2 通用文本特征的隐私问题

开放网络下构建云-边交互的分布式学习系统这一新兴趋势给深度学习的安全性和隐私性带来了空前的挑战.在传统局域网下的分布式学习系统或单节点学习系统中,传统基于词频统计的文本表征尽管会使得原始文本中存在的信息被充分披露,却由于这类应用场景中的文本信息通常不具备隐私性、与数据所有者脱敏、不易被外部攻击者所截获等因素,文本特征的隐私性一直以来鲜有相关研究.而近 2 年随着通用深度学习语言模型的兴起,考虑该类模型所产生的文本特征具有空前的下游任务普适性,研究者开始注意到通用文本特征在开放网络环境下可能带来的隐私风险.2020 年文献[14]率先指出:用户在端侧的输入的文本信息即使经由通用语言模型编码,生成向量化的文本特征后,该类特征一旦被公开在网络环境中、遭受中间人攻击截获、或被半诚实(honest-but-curious)的服务器嗅探,可能造成该用户的隐私受到侵犯.文献[14]提出一种基于机器学习方法的敏感词推断攻击技术,根据截获的文本特征判断其背后的原始文本中是否存在攻击者询问的敏感词.该攻击方法首先根据公开语料库生成 1 组包含或不包含攻击者询问的敏感词的语料,随后通过本地部署的同种通用语言模型,用以将相应语料转化为对应的文本表征,最后训练支持向量机、多层感知器等传统模型用以预测敏感词的存在与否.

1.3 其他相关工作

在现实世界中,尤其对于医疗、金融、安防等相关领域,数据集的全局属性、数据集中是否包含特定数据样本、数据集自身等信息都可能高度隐私.近年来,一些研究者提出了多种不同新型的攻击手段从不同层面揭示了深度学习算法中存在的各类数据隐私问题[26].根据攻击的目的性不同,现有工作主要可以分为推断攻击(inference attack)和重建攻击(reconstruction attack)两类.

推断攻击的主要目标是判断模型训练使用的数据集是否符合某种敏感条件或具备某种属性,通常被建模为分类任务.根据推断目的的不同,现有推断攻击工作又主要分为成员推断攻击和属性推断攻击.成员推断攻击主要试图揭示某些特定数据样本是否在已知模型的未知训练数据当中,由 Shokri 等人[27]

于 2017 年率先提出,随即引起了学术界的广泛关注[28-30].不同于成员推断攻击,属性推断攻击的目标更为粗粒度,攻击者主要希望判断训练集是否具有某些特定的全局属性,由 Ganju 等人[31]于 2018 年提出并在多种深度学习模型上实现了该类攻击,并在近期由 Melis 等人[32]和 Pan 等人[14]推广到用于文本的分布式学习系统,分别通过梯度信息和句向量信息推断训练集数据是否包含特定单词.

重建攻击的主要目的在于恢复训练集中部分或全部的训练样本,最早可追溯到 2015 年 Fredrikson 等人[33]首次提出的模型反演攻击(model inversion attack).而正如 Shokri 等人[27]随即指出,该类攻击无法对一些宽泛的类标签恢复出有意义的数据样本.2019 年 Salem 等人[34]和 Zhu 等人[35]在模型反演攻击的基础上进行了更为细粒度的改良,分别提出基于输出变化和平均梯度的数据重构攻击以在当前训练轮次中恢复对应的小批量内的每个数据样本.此外,近年来研究者也发现,仅仅通过请求在线机器学习调用接口,攻击者便有可能窃取私有智能系统中所部署的模型参数[36]、结构[37]、超参数[38]等信息,侵犯相关机构的知识产权.

2 针对通用语言模型的隐私窃取链

本节介绍了如何从截获的通用文本特征出发,逐步推断出产生该文本特征的模型来源、相应原始文本的长度、最有可能出现的敏感词列表,最终重建原始文本的敏感语义.本文称该分阶段推断攻击为 1 条针对通用语言模型的隐私窃取链.

2.1 相关记号与攻击假设

2.1.1 相关记号

沿用文献[14]中采用的记号,记 1 个预训练通用语言模型为,用户输入文本为 $\boldsymbol{x}$,可表示为 1 个长度为 n 的词序列$(w_1,w_2,\cdots,w_n)$,其中每个单词 w_i 来自单词表 V.该文本经过通用语言模型 f 计算产生相应的文本特征 $\boldsymbol{z}:=f(\boldsymbol{x})\in\mathbb{R}^d$,即 1 个 d 维的实值向量.记用户原始文本中携带的隐私信息为 s,且该隐私信息可表示为原始文本的函数 $P:\boldsymbol{x}\rightarrow s$.例如,原始文本 $\boldsymbol{x}$ 为患者陈述"3 天前孩子去中医扎针灸",经过 BERT 语言模型被编码成 $d=1024$ 维的实值向量作为其表征 $\boldsymbol{z}$,而例如,攻击者目标逆向出原始文本中包含的医疗相关的关键词,那么"中医"和"针灸"可以认为是该用户相应的敏感信息 s.

2.1.2 攻击假设

假定攻击者拥有 3 种能力:

假设 1. 攻击者截获了 N 个由某个或多个通用语言模型产生的文本特征,而它们对应的原始文本中包含攻击者想要获取的用户隐私信息,这里 $N\in\mathbb{N}_+$.

假设 2. 攻击者了解截获的文本特征对应的原始文本的语种、相关应用领域等元信息.

假设 3. 攻击者能够获取来自和原始文本相似领域的公开语料库.

假设 1 主要为攻击者准备了相应的攻击信道,即用户在使用通用语言模型过程中无意泄露或被半诚实的服务器端缓存的文本特征.注意到这里本文放宽了文献[14]此前对于攻击者知晓产生这些文本特征的通用语言模型的类型的假设,同时本文容许攻击者截获的文本特征可以来自于多个未知的通用语言模型.

对于假设 2,攻击者通常可以利用文本特征泄露过程中的一些侧信道(side channel)来获得这些信息,例如从相关服务提供方的服务类别可以推测具体的应用领域、根据服务提供方使用的主要语种或者受害人的 IP 信息来确定相应的语种等等.

假设 3 类似于 Pan 等人[14]在其关键词推理攻击中设计的白盒攻击场景.随着互联网上的公开语料库愈发增多,攻击者也越发容易满足该攻击假设.例如,一些医疗机构或在线医患问答平台通常会公开脱敏后的患者主诉、治疗方案描述、手术信息等语料用于促进科研或提升医疗服务质量,而在本文中攻击者却得以利用这些公开语料库进行一系列的隐私窃取.此外,如若攻击者无法获得满足假设 2 的语料库,文献[14]指出可以利用域对抗神经网络技术(domain adversarial neural network)[39]来实现攻击模型迁移,该技术路线可供未来工作继续探索.

2.2 攻击流程概述

如图 2 所示,本文所提出的攻击流程包含 3 个阶段:

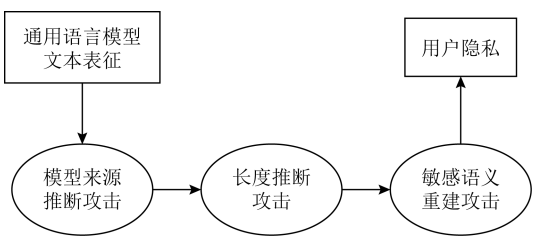
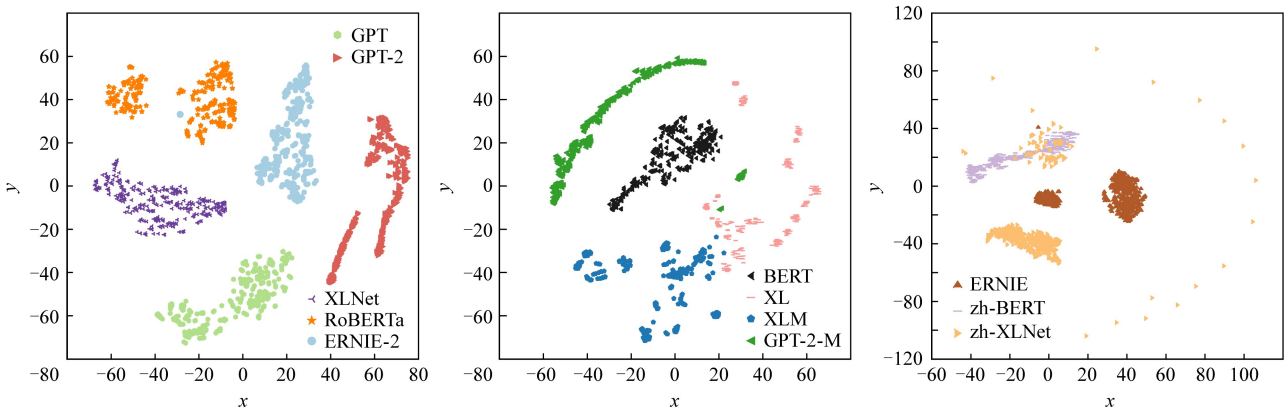


Fig. 2 Privacy theft chain targeted at GPLM embeddings

图 2 针对通用语言模型文本特征的隐私窃取链

2.2.1 模型来源推断攻击

在阶段 1,为了开展下游攻击,攻击者首先需要确定它所截获的各个文本表征分别来自于何种通用语言模型.尽管该任务似乎为攻击者设置了巨大的挑战,文献[14]在其附录中通过可视化方法指出,来自于不同通用语言模型的文本表征在高维空间上似乎可分性较好,如图 3 所示.本文基于这一现象,设计了完整的模型来源推断攻击方法,用于对每个文本特征标记相应的模型来源.具体攻击算法设计请见 2.3 节.



(a) 768维英文通用文本特征分布2维平面可视化结果

(b) 1024维英文通用文本特征分布2维平面可视化结果

(c) 768维中文通用文本特征分布2维平面可视化结果

Fig. 3 t-SNE visualization of textual embeddings produced by GPLMs

图 3 t-SNE 算法可视化通用文本表征分布

2.2.2 长度推断攻击

在阶段 2,当攻击者确定了模型来源之后,本文将进一步扩展暴露的隐私信息种类.具体地,在该阶段中,攻击者试图获取通用文本特征对应的原始文本中包含的单词的个数,即原始文本的长度信息.推断出该信息将有助于攻击者在下一阶段的语义重建攻击中确定所需要嗅探的最可能出现的单词列表的长度上限,记为 K.相关攻击框架和长度推断攻击的具体细节将在 2.4 节介绍.

2.2.3 敏感语义重建攻击

在阶段 3,攻击者试图根据前 2 个阶段中获得的关于文本特征及其对应文本的元信息,重建出未知文本中包含的敏感语义.由于词是构成文本语义的主要成分[40],本文仍以关键词作为突破口,借鉴了对比学习(contrastive learning)[41]的思想,设计一种全新的基于神经协同过滤架构(neural collaborative filtering, NCF)[42]的关键词出现评分预测模型,并给出了相应基于负采样的训练技术.攻击者可以在公开语料库上事先完成这类攻击模型的训练;在攻击阶段,攻击者只需给定需要嗅探的敏感词列表,即可直接利用提前训练好的推断模型输出列表中各个敏感词的出现得分,经由排序操作后,利用阶段 2 中推断出的长度信息,获得前 K 个最有可能出现的敏感词,从而重建原始文本的敏感语义.相较之下,文献[14]最早将关键词推断攻击建模成一种 2 分类问题(即给定 1 个目标关键词,根据文本表征判断原始文本中是否包含该关键词).尽管该方法也可间接用于敏感语义重建攻击(即对攻击者给定的敏感词列表中的每个词进行 1 次关键词推断攻击,并对预测为存在的关键词利用预测置信度作为其出现概率.经过分析,发现其主要存在 2 点局限性:

1) 攻击者需要对每个关键词训练 1 个相应的推断模型,难以扩展到较大的敏感词列表;

2) 为了给每个关键词推断模型准备均衡的训练数据,攻击者需要将公开语料库中的每段文本替换成 1 对包含或不包含该关键词的新文本,利用本地部署的通用语言模型重新生成相应的通用文本特征,攻击开销较大.

为了解决这 2 点主要不足,本文创新性地提出了解决方案:在框架层面,本文改变原先采用的分类任务的建模方式,提出将敏感语义推断攻击建模为排序任务,即根据文本特征预测按出现概率降序排列的敏感词列表;在攻击模型结构层面,本文将模型改变为同时输入文本特征和待推断关键词经由在阶段 1 中推断出的通用语言模型编码后获得的词特征,预测相应的关键词在原文中的出现得分,从而得以让敏感词列表中的各个单词共享同一个推断模型;在训练算法层面,本文利用公开语料库中词与句子的

共现关系作为自监督信号,并利用负采样(negative sampling)的策略,设计一种对比损失函数使得攻击模型给实际出现在句子中的单词输出比未出现的单词更高的得分.具体的模型结构和算法设计将在2.5节介绍.

2.3 模型来源推断攻击

2.3.1 攻击场景定义

记受害者公开的文本特征集合为 $Z=\{\boldsymbol{z}_i\}_{i=1}^{N}$.攻击者根据假设2中已知的语种相关信息,设定通用语言模型来源的候选集合为 $F=\{f_1,f_2,\cdots,f_C\}$.例如,对于英文场景,攻击者选择表1中的10种面向英文文本的通用语言模型.不失一般性,本文假定每个通用文本特征均由候选集 F 中的某个模型产生.攻击者的攻击目标可以描述为:$A:\boldsymbol{z}\rightarrow c\in\{1,2,\cdots,C\}$,即输入相应的文本特征,推断其模型种类在候选集 F 中的索引.

2.3.2 攻击方法描述

在采用基于机器学习的来源推断方法之前,攻击者首先可以利用各个通用语言模型输出的文本特征维度的差异性,对候选集进行初筛.例如,对于维度为768的文本特征,攻击者查询表1可知,该文本特征可能来自XLNet,GPT,GPT-2等5种候选模型.根据场景定义,模型来源推断攻击可建模为1个多分类任务,机器学习中常见的多分类方法或模型都能够用来求解.为了准备相应的训练数据,攻击者首先需提前准备1个任意的公开语料库 $X_{\text{pub}}=\{\boldsymbol{x}'_j\}_{j=1}^{M}$.值得注意的是,这里选取的语料库可不局限于假设3中所述的来自相似领域的语料库.事实上,3.1节中的实验结果显示,和模型来源相关的特征与文本所属领域的独立性较低,在一个几乎无关联的领域中选取的公开语料库上训练的模型来源推断模型无需迁移,便能够在目标语料上取得接近100%的分类准确度.随后,攻击者分别在本地部署候选集中的各个通用语言模型 f_c,将公开语料库每段文本 $\boldsymbol{x}'_j$ 分别转化为相应的文本特征 $f_c(\boldsymbol{x}'_j)$,并辅以相应的模型来源标签 c,最终得到训练集 $D_{\text{train,fingerprint}}$,其中每个样本的形式即为 $(f_c(\boldsymbol{x}'_j),c)$.

2.3.3 具体模型设置

本文采用1个隐藏层大小为200的3层全联接神经网络作为分类器,其中输入层神经元个数等同于文本特征的维度 d,输出层大小为候选集中模型的个数 C,并经由1个Softmax层获得最终的模型来源预测概率向量;隐藏层的激活函数为 $ReLU$.在训练过程中利用Adam优化器[43]以0.01的学习率最小化 $D_{\text{train,fingerprint}}$ 中样本的模型来源预测概率与实际来源之间的交叉熵(cross entropy)损失函数.

2.4 长度推断攻击

2.4.1 攻击场景定义

根据第1阶段的推断结果,攻击者得以将原来整个文本特征集合 Z 按照对应的模型来源进行划分.为了记号的一致性,我们仍记受害者公开的文本特征集合为 $Z=\{\boldsymbol{z}_i\}_{i=1}^{N}$,并不失一般性地假设每个 $\boldsymbol{z}_i$ 均由某个通用语言模型 f_c 产生,f_c 已被攻击者知悉.在当前攻击阶段,攻击者的目标为推断每个受害者文本特征 $\boldsymbol{z}_i$ 所对应的原始文本的"长度".具体地,结合2.1节中关于通用语言模型特征提取流程的介绍,本文所说的文本"长度"特指原始文本经过通用语言模型的分词模块后获得的词(token)序列的长度.特别地,本文之所以不将文本的原始长度或者中文中字符的个数作为这一阶段的攻击目标,是因为经过分词模块获得的符号通常才是通用语言模型进行特征提取的基本单元,因而如若文本特征的确暗含了相应的长度信息,其泄露的也应是符号序列的长度而非句子的自然长度.推断符号序列的长度也有助于下游敏感语义重建攻击,特别是帮助其确定敏感词列表长度上界.

2.4.2 攻击方法描述

当攻击者知晓了文本特征的模型来源后,它首先在本地部署相应的语言模型 f_c;接着,一方面,攻击者利用语言模型,将获得的与目标文本来自相似领域的公开语料库 $X_{\text{pub}}:=\{\boldsymbol{x}'_j\}_{j=1}^{M}$ 转化为相应的文本特征;另一方面,攻击者利用公开可获取的通用语言模型对应的分词模块,对公开语料库中的每个样本 $\boldsymbol{x}'_j$ 进行分词并标记相应的文本长度 l_j;最后,在攻击者完成了训练数据集 $D_{\text{train,lia}}=\{(f_c(\boldsymbol{x}'_j),l_j)\}_{j=1}^{M}$ 的准备后,它将选择一种机器学习中经典的多分类模型在 $D_{\text{train,lia}}$ 进行训练,完成后对受害者暴露的来自同一语言模型的文本特征进行长度推断.

2.4.3 具体模型设置

类似于来源推断攻击,本文同样采用3层全连接神经网络作为长度推断模型,其中输入层大小为文本特征的维度 d,隐藏层大小为200,输出层大小为攻击者所获得的公开语料中最长文本的长度(实验数据集上的实际统计数据将在2.6节给出),并经由1个Softmax层预测相应的文本长度;隐藏层的激活函数为 $ReLU$.在训练过程中利用Adam优化器以0.01的学习率最小化 $D_{\text{train,lia}}$ 中样本的预测所具有长度与实际长度之间的交叉熵损失函数.

2.5 敏感语义重建攻击

2.5.1 攻击场景定义

记受害者公开的文本特征集合为 $Z=\{\boldsymbol{z}_i\}_{i=1}^{N}$,对应的原始文本长度为$\{l_i\}_{i=1}^{N}$,并假设每个 $\boldsymbol{z}_i$ 均由某个通用语言模型 f_c 产生.借助模型来源推断攻击和长度推断攻击,f_c 和$\{l_i\}_{i=1}^{N}$均能够被攻击者知悉.在该阶段攻击者试图利用前 2 个阶段从文本特征中获取到的敏感信息进一步从通用文本特征 $\boldsymbol{z}$ 中重建原始文本 $\boldsymbol{x}$ 中最有可能存在的词.具体地,给定敏感词列表 $V_{\text{sensitive}}$,本阶段攻击目标是 $A:(\boldsymbol{z},f_c(v_i))\rightarrow s_i$,其中 s_i 代表单词 $v_i\in V_{\text{sensitive}}$在原始文本中的出现得分.在预测了出现得分之后,攻击者根据各敏感词的出现得分进行降序排序,将前 l 个词看作原始文本中最有可能出现的敏感词,便可在很大程度上重建敏感语义(这里 l 是长度推断攻击中获得的 $\boldsymbol{z}$ 对应原始句子的长度).因此该阶段不同于前 2 个阶段,在建模上从分类任务转变为排序任务.

2.5.2 攻击方法描述

类似于 2.4.2 节,攻击者在知悉文本特征的模型来源 f_c 后,先在本地部署对应的语言模型.这里利用 2.1.2 节中的假设 2,攻击者获得同 $\boldsymbol{z}$ 对应的原始文本相似的公开语料库 $X_{\text{pub}}:=\{\boldsymbol{x}'_j\}_{j=1}^{M}$,构建对应的词表为 V_{pub}.接着攻击者就能借助本地语言模型将 X_{pub}转化为相应的文本特征$\{f_c(\boldsymbol{x}'_j)\}_{j=1}^{M}$,并从 $\boldsymbol{x}'_j$ 中随机采样 1 个词记为 $w_{j,+}$,同时从 V_{pub}中随机取 1 个不在 $\boldsymbol{x}'_j$ 中的词记为 $w_{j,-}$.

本文所提出的敏感词推断模型主要用于推断作为输入的受害者的通用文本特征与攻击者所希望嗅探的敏感关键词之间是否具有包含关系,并给出相应的包含概率预测值.为此,本文创新性地将该隐私推断任务类比为推荐系统中的推荐商品排序任务,即为每个通用文本特征"推荐"对应原始文本中出现概率最高的 K 个敏感关键词.对此,本文基于推荐系统中常用的 NCF 结构,构建了如图 4 所示的敏感词推断模型,具体计算流程为:基于预先准备的训练数据集 $D_{\text{train,semantic}}=\{f_c(\boldsymbol{x}'_j),f_c(w_{j,+}),f_c(w_{j,-})\}_{j=1}^{M}$,敏感词推断模型以通用文本特征 $f_c(\boldsymbol{x}'_j)$和相应词特征 $f_c(w_j)$作为输入,进入模型的 2 个分支.在图 4 左侧分支中,通用文本特征与词特征拼接,经过多层全联接网络获得左侧表征,该表示主要用于建模词与文本共现关系在特征空间上的线性关系;在图 4 右侧分支中,经过线性网络获得右侧表示,该表示主要用于建模词与文本共现关系在特征空间上的逐维相关性(coordinatewise correlation).最后,左右侧表征拼接并经过多层全连接网络后,即可得到词 w_j 在 $\boldsymbol{x}'_j$ 中的出现得分预测值.

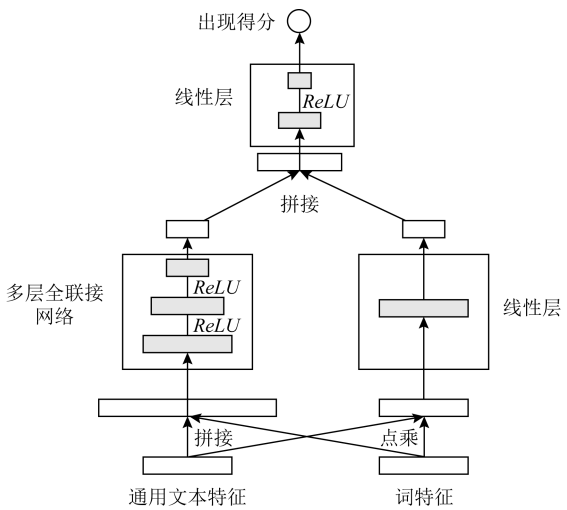


Fig. 4 NCF-based sensitive keyword inference model

图 4 基于神经协同过滤架构的敏感词推断模型

在敏感词推断模型的训练阶段,本文从对比学习的思想出发,提出训练方法和损失函数:在训练时,我们在每个小训练批次(mini-batch)中同时加入在原始文本中出现的敏感词 $w_{j,+}$ 和未在原始文本中出现的敏感词 $w_{j,-}$(也被称作负样本),模型需要尽可能地扩大正例样本 $w_{j,+}$ 的出现得分 $s_{j,+}$ 与负例样本 $w_{j,-}$ 的出现得分 $s_{j,-}$ 之间的差值,具体对应损失函数为

$$L_{\text{semantic}}=\frac{1}{B}\sum_{j=1}^{B}\ln\left(1+\mathrm{e}^{-\frac{s_{j,+}-s_{j,-}}{\gamma}}\right),\qquad(1)$$

其中,参数 B 是 1 个小训练批次所包含的正负样本对个数.经过优化对比损失函数,所提出的敏感词推断模型能够学习通用文本特征与敏感词特征之间的深层关联,并根据这一关联判断敏感词的出现与否,从而在隐私推断过程中为更有可能出现的敏感词赋予更高的出现概率评分.

2.5.3 具体模型设置

如图 3 所示,模型输入层神经元个数等于文本特征维度 d 的 2 倍,用于同时输入同一通用语言模型生成的文本特征 $f_c(\boldsymbol{x}'_j)$和词特征表示 $f_c(w_j)$.接着将输入分别经过 2 种不同的神经网络结构:左侧分支是先将文本特征和词特征表示直接拼接成大小为 $2d$ 的向量,再将这向量输入 1 个 4 层全连接神经网络,其中输入层大小为 $2d$,中间 2 个隐藏层分别包含 300 个、200 个神经元,最后的输出层大小为 100;右侧分支将文本特征和词特征逐位点乘后直接输入 1 个单层线性网络,并输出 1 个长度为

100 的向量.在此之后,将 2 个不同的神经网络结构的输出拼接成长度为 200 的向量后,输入 1 个 3 层全连接神经网络,隐藏层神经元个数为 50,最终输出层仅有 1 个神经元,用于输出词 w_j 在 $\boldsymbol{x}'_j$ 上的出现得分.网络结构中隐藏层的激活函数均为 $ReLU$.在训练过程中利用 Adam 优化器以 0.001 的学习率最小化 $D_{\text{train,semantic}}$ 每组样本 2 个预测得分的对比损失函数 L_{semantic}.

2.6 数据集选取与评估指标

2.6.1 数据集简介

1) CMS 公共医疗记录数据集[44].由美国医疗及医疗补助服务中心统计并公开的全美各医院的治疗质量、收费情况等多类医疗数据,本文选取来自 10 个主要科室(如放射科、皮肤科等)的患者治疗过程文本描述部分构成实验部分使用的英文医疗语料,共计 60 000 条文本,按 5∶1划分训练集(作为假设 3 中的公开语料)和测试集(仅可见相应的通用文本特征).实验结果默认给出 10 次重复随机实验结果的均值.CMS 医疗记录数据集的词表大小为 1 652,平均句长为 9.24.

2) 中文医患问答数据集.由 He 等人[16]从某公开中文医患问答平台抓取的网民与医疗工作者在线问答中的单轮对话内容,本文选取来自其中 60 000 条网民的病情描述文本,按 5∶1比例划分为互不相交的训练集/测试集并进行 10 次重复随机实验.中文分词时均采用各预训练模型同时公开的分词模块进行,该数据集的词表大小为 63 252,平均句长为 27.5.

2.6.2 评估指标

1) 攻击准确度.具体地,假设攻击者截获的通用文本特征集合为 $Z=\{\boldsymbol{z}_i\}_{i=1}^{N}$,$\boldsymbol{z}_i$ 对应的原始文本 $\boldsymbol{x}_i$ 的真实敏感信息为 s_i,那么攻击模型 A 在目标集 Z 上的攻击准确度定义为

$$Accuracy(A)=\frac{1}{N}\sum_{i=1}^{N}1(A(\boldsymbol{z}_i)=s_i), \quad (2)$$

即被正确推断敏感信息的样本个数与全部目标样本个数的比值.由于模型来源攻击和长度推断攻击均为多分类问题,在 3.1 节及 3.2 节中,本文主要根据攻击模型的分类准确度来衡量相应的隐私风险.

2) $Precision@K$ 和 $Recall@K$.这 2 个指标主要在敏感语义重建攻击的评估中使用.具体地,给定任一用户原始文本 $\boldsymbol{x}=(w_1,w_2,\cdots,w_L)$ 和攻击者预先准备的敏感词列表 $V_{\text{sensitive}}$,当攻击模型预测出前 K 个最有可能出现的敏感词 $(v_1,v_2,\cdots,v_K)$,$Precision@K$ 指标定义为

$$Precision@K(\boldsymbol{x},A)=\frac{1}{K}\sum_{i=1}^{K}1(v_i\in\boldsymbol{x}), \quad (3)$$

即预测的 K 个敏感词中实际出现在原始文本中的个数与列表长度的比值,主要用于衡量敏感语义重建攻击的误报率.同时,本文采用 $Recall@K$ 指标衡量敏感语义重建的命中率,定义为

$$Recall@K(\boldsymbol{x},A)=\frac{\sum_{i=1}^{K}1(v_i\in\boldsymbol{x})}{\sum_{j=1}^{L}1(w_i\in V_{\text{sensitive}})}, \quad (4)$$

即预测的 K 个敏感词中实际出现在原始文本中的个数与原始文本出现的敏感词个数的比值.

3 实验结果与分析

本节首先介绍模型来源推断攻击,可视化展示了产生于不同通用语言模型的文本特征的聚类现象;接着,给出了长度推断攻击的实验结果;最后给出了敏感语义重建攻击的实验结果,并展示了相应的隐私泄露实例.

3.1 模型来源推断攻击结果分析

本节介绍了在 3 组不同的通用语言模型候选集上进行模型来源推断攻击的结果,这 3 组模型分别为:

1) 文本特征维度为 768 的英文语言模型(如图 3(a)所示);

2) 文本特征维度为 1024 的英文语言模型(如图 3(b)所示);

3) 文本特征维度为 768 的中文语言模型(如图 3(c)所示).

每组具体包含的模型请见图 3 中的图例.后续部分也将以分组进行实验展示和分析.

表 2 给出了模型来源推断攻击分别针对 3 组模型上黑盒和白盒场景下医疗文本特征测试集上的攻击准确度,其中黑盒场景指攻击者在领域无关语料上(实验中采用亚马逊商品评论英文语料[45]中随机采样的 1 000 条评论)训练模型来源推断攻击模型后,在医疗文本特征测试集上进行模型来源推断;白盒场景为攻击模型分别在相应语种的医疗语料训练集和测试集上进行训练和来源推断.

如表 2 所示,在白盒与黑盒攻击场景下,模型来源推断攻击的准确度均能达到 98% 以上,尤其在其中 4 种配置下重复实验表明攻击准确度能始终维持在 100%.我们进一步利用经典的 t-SNE 算法对

各组模型产生的通用文本特征进行了可视化分析,相应的散点图在图 3 中展示.可以看到,不同语言模型产生的文本特征在 2 维平面上形成了自然的聚类,这表明每类通用语言模型产生的文本特征向量具有各自独特的分布特征,因而攻击者能在仅有领域无关语料的情况下近乎精准地推断出截获的通用文本特征的模型来源,从而为下一阶段攻击提供了前置条件.

Table 2 Attack Accuracy of Model Finger-Printing Attack

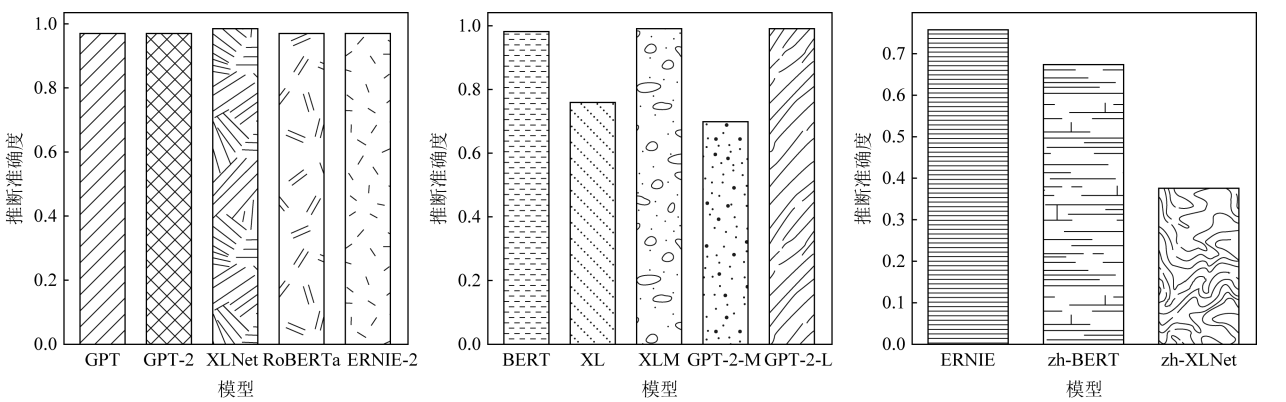
表 2 模型来源推断攻击准确度 %

攻击场景	各组通用语言模型准确度		
	图 3(a)	图 3(b)	图 3(c)
白盒	100.0	100.0	98.4
黑盒	100.0	99.3	100.0

此外,纵向比较白盒与黑盒的攻击结果可知,白盒攻击由于其拥有更多的先验知识因而在多数情况下会表现优于黑盒场景;而在中文模型上,黑盒场景的攻击准确度 100.0% 却略高于白盒场景的结果 98.4%,通过分析训练中间过程,该现象主要由于攻击模型的过拟合所致,而黑盒场景由于利用了领域无关数据能更好地集中于学习仅和分布相关的攻击知识,因而获得了更好的攻击效果.

3.2 长度推断攻击结果分析

本节介绍了在 3 组通用语言模型上进行长度推断攻击的实验结果.具体地,为了对比的公平性,中英文医疗训练及测试语料均预先经长度过滤得到句长在 10～20 之间的文本,其中英文语料训练集与测试集分别包含 48 224 和 9 743 条文本;中文分别包含 2 094 和 886 条文本.作为基线,针对英文数据的随机盲猜(random guessing)和众数盲猜(majority-based guessing)的长度推断准确度分别为 10% 和 23.9%;针对中文数据时分别为 10% 和 24.0%.图 5 给出了相应的长度推断攻击准确度的柱状图.



(a) 768维英文通用语言模型上长度推断攻击性能 (b) 1 024维英文通用语言模型及GPT-2-L上长度推断攻击性能 (c) 768维中文通用语言模型上长度推断攻击性能

Fig. 5 Performance of length inference attack on each group of GPLMs

图 5 在各组通用语言模型上的长度推断攻击性能

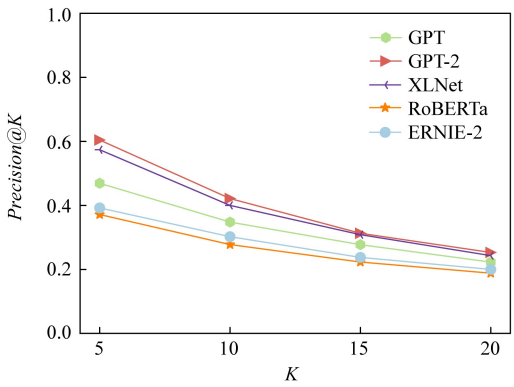
从图 5 可以看到,攻击者针对各个通用语言模型发动的长度推断攻击的准确度均高于基线指标,这表明这些通用语言模型的特征中均包含与句子长度相关的信息.

特别地,对于大部分英文通用语言模型而言,攻击者能够以 97% 以上的准确度推断句子长度.通过分析这些通用语言模型的结构设计细节,我们认为这可能是由于在图 1 介绍的文本特征提取第 3 步中,即词序列转化为词向量序列过程中,普遍采用位置向量信息(positional embedding)与传统的语义词向量融合来形成混合词向量[6],而这些与长度相关的信息在经过模型计算后仍得以部分保留.一方面这些信息进一步增强了通用语言模型在语序相关任务上的通用性;另一方面,也使得本文提出的长度推断攻击能够实施.类似地,由于谷歌新开发的 Transformer-XL 模型采用了相对位置向量[21]而非谷歌原先在 BERT 等模型中采用的绝对位置向量,其暴露的长度信息会相对较少,然而,针对它的长度推断攻击的准确度也达到了 70% 以上.此外,相比之下,对照图 5(a)～(c)可以发现,中文通用语言模型面对长度推断攻击的安全性比英文模型相对更强(例如百度开发的中文 ERNIE 模型上的长度推断攻击的准确度为 37.8%),这可能是因为中文自身的复杂性(例如,中文医疗语料的词表大小为 60 000

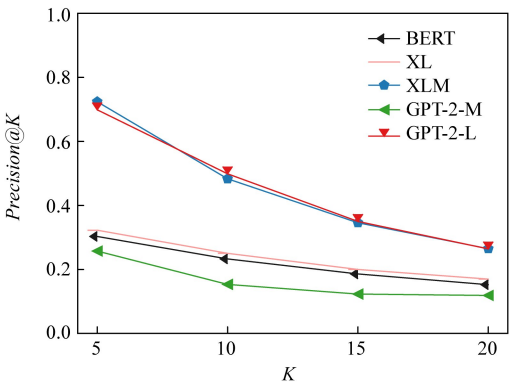
左右远大于英文医疗语料的不到 2 000 的词表大小)及词边界的歧义性导致最终获得的文本特征中所包含的文本长度信息相对模糊.

3.3 敏感语义重建攻击结果分析

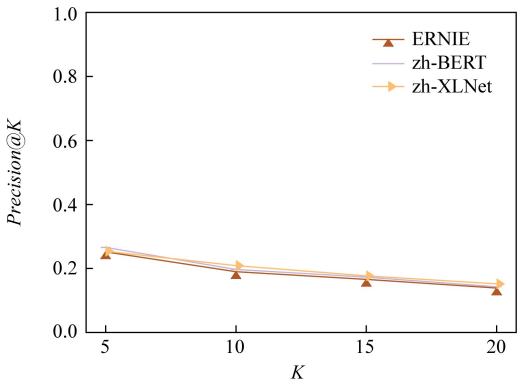
本节介绍了敏感语义重建攻击的实验结果.针对中英文医疗语料,我们首先根据 2.5 节中的方法在训练集上训练相应的关键词推断模型;随后,在训练集上统计不为停词的前 100 个高频词作为攻击者希望嗅探的敏感词列表;最后,对每个测试文本特征,利用关键词推断模型,计算敏感词列表中的各单词的出现概率得分并从大到小排序,将前 K 个单词看作包含原始文本敏感语义的单词列表输出.K 在实验中取值 5, 10, 15, 20.图 6 给出 *Precision*@K 和 *Recall*@K 的曲线图,2 种攻击性能指标定义请见 2.6.2 节.从图 6 可见,对于 10 种英文通用语言模型中的 4 种(分别为 OpenAI 开发的 GPT-2 和 GPT-2-Large,Google 开发的 XLNet 以及 Facebook 开发的 XLM),攻击者在 $K=5$ 的情况下 *Precision*@5 和 *Recall*@5 均能达到 60%甚至更高.而随着敏感词预测集从 5 逐渐增大到 20,攻击者在这些模型



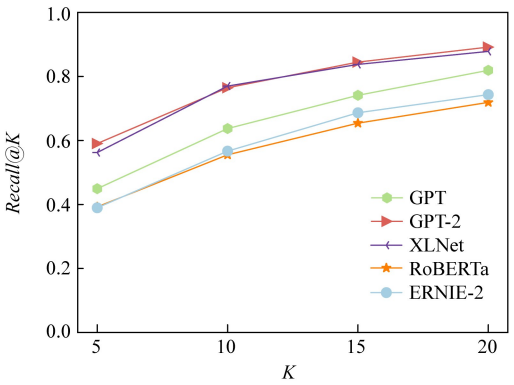
(a) 768维英文通用语言模型上*Precision@K*变化曲线



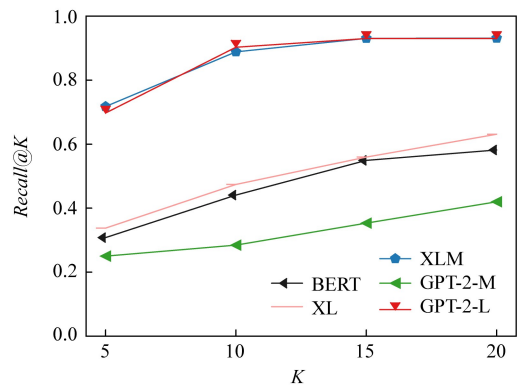
(b) 1 024维英文通用语言模型及GPT-2-L上*Precision@K*变化曲线



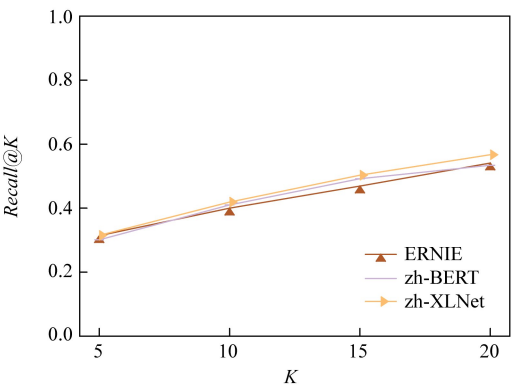
(c) 768维中文通用语言模型上*Precision@K*变化曲线



(d) 768维英文通用语言模型上*Recall@K*变化曲线



(e) 1 024维英文通用语言模型及GPT-2-L上*Recall@K*变化曲线



(f) 768维中文通用语言模型上*Recall@K*变化曲线

Fig. 6 Performance of semantic reconstruction attack on GPLMs

图 6 通用语言模型敏感语义重建攻击效果

上的 $Recall@K$ 指标上升至 90%,这意味着攻击者能够重建出原始文本中 90%比例的敏感词,而这时对应的 $Precision@K$ 指标也在 30%~50%区间内,这表明敏感词预测集中的有效内容占比可观.

我们进一步采样了一些实际的敏感语义重建样例,图 7 展示了 3 对随机选择的英文原始文本和从 Google 的 XLNet 模型产生的通用文本特征中推断的前 10 个可能出现的敏感单词,以及 1 对随机选择的中文原始文本和从百度的 ERNIE 模型产生的通用文本特征中推断的前 10 个可能出现的敏感词.其中原始文本中出现在攻击者准备的敏感词列表中的单词用阴影标记,敏感词预测集中命中的单词用下划线标记.

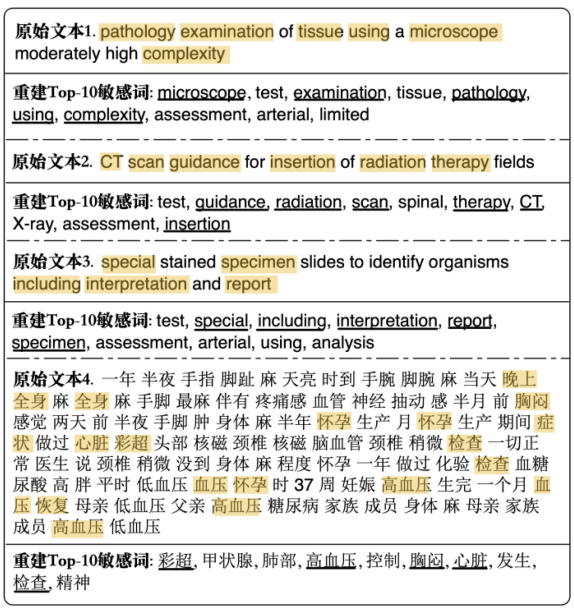


Fig. 7 Demonstration of sensitive semantic reconstruction attack

图 7 敏感语义重构攻击实例

这些结果直观地反映了图 6 中的数值指标对应的隐私暴露程度:攻击者可以较为准确地推断出原始文本中出现的大部分敏感词;同时,得益于通用语言模型的语义捕捉能力,预测出的敏感词候选集中未命中的词也和原始文本具有较为相关的语义.例如,图 7 的原始文本 1 中未命中的"assessment"(评估)和"test"(测试)均与原始文本中包含的"examination"(检查)语义关联紧密;类似地,图 7 的原始文本 2 中的未命中的"X-ray"(X 光)也与原始文本中的"radiation"(辐射)和"CT-scan"(CT-扫描)处于相近的语义维度.图 6(c)和图 7 的原始文本 4 也给出了中文通用语言模型上的数值指标和实际攻击样例.类似于前 2 部分的中英文模型的对比分析,由于中文文本自身的复杂性且本节中使用的原始中文医疗语料的平均句长和词表大小远高于英文医疗语料,尽管中文通用语言模型在敏感语义重建攻击下的安全性也相对较强,然而原始文本中的隐私信息仍有所泄露.例如,图 7 中原始文本 4,攻击者可以因此推断该受害者可能罹患高血压及其并发症.这些潜在的隐私风险对制定通用语言模型的使用规范以及其未来落地来说不容忽视.

4 总结与未来工作

尽管近年来设计、开发和预训练通用语言模型的趋势方兴未艾,一些研究者已开始分析通用语言模型在开放网络下的实际服务场景中可能潜在的隐私和安全问题.随着谷歌公开宣布 BERT 模型在其搜索引擎中的落地,通用语言模型的安全与隐私研究急需更多的研究投入.在这些背景下,本文通过构建一系列的隐私窃取攻击,进一步阐释了潜在攻击者将如何仅从受害者公开或无意间暴露的通用文本特征中,确定其模型来源,推断其背后的原始文本长度,构建其中最有可能出现的敏感词列表,最终得以重建原始文本的敏感语义.在 10 种代表性的英文通用语言模型和 3 种中文通用语言模型上的实验结果表明,通用语言模型产生的文本特征存在不容忽视的隐私风险.一方面,本文证明和展示隐私风险的存在性将有助于用户、服务隐私规则制定者和开发者更好地理解通用语言模型的隐私属性;另一方面,通用语言模型潜在的隐私风险也应引起学术界和工业界对相关隐私增强技术的关注和研究.除了文献[14]中提出的基于差分隐私、子空间投影等技术对通用文本特征进行后处理(post-processing),未来研究也可从预处理(preprocessing)和中段处理(in-processing)角度出发提出更多有效的隐私增强技术.从预处理的视角出发,终端可以采用例如文本脱敏、匿名化等技术来从源头上减少用户隐私的泄露;从中段处理的视角出发,通用语言模型的开发者则可以采用例如差分隐私训练、对抗训练等隐私保护的学习算法,使得最终获得的模型减少对原始文本中敏感信息的收集和建模.由于篇幅所限,本文未对这些隐私增强技术进行系统性的评估分析,希望未来的研究工作可基于 3 种防御思路,提出更多切实有效的隐私提升技术,以进一步保障通用语言模型和相应服务模式的安全性.

参 考 文 献

[1] Xu Hu, Liu Bing, Shu Lei, et al. BERT post-training for review reading comprehension and aspect-based sentiment analysis [C] //Proc of the 20th Annual Conf of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 2324-2335

[2] Qu Chen, Yang Liu, Qiu Minghui, et al. BERT with history answer embedding for conversational question answering [C] //Proc of the 42nd Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2019: 1133-1136

[3] Bao Siqi, He Huang, Wang Fan, et al. PLATO: Pre-trained dialogue generation model with discrete latent variable [C] //Proc of the 58th Annual Meeting of the ACL. Stroudsburg, PA: ACL, 2020: 85-96

[4] Liao Yi, Wang Yasheng, Liu Qun, et al. GPT-based generation for classical chinese poetry [OL]. [2019-09-05]. https://arxiv.org/abs/1907.0015

[5] Zellers R, Holtzman A, Rashkin H, et al. Defending against neural fake news [C] //Proc of the 30th Annual Conf on Neural Information Processing Systems. New York: Curran Associates, 2019: 9051-9062

[6] Delvin J, Chang Mingwei, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C] //Proc of the 20th Annual Conf of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 4171-4186

[7] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training [OL]. [2019-12-25]. https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf

[8] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners [OL]. [2019-12-26]. https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

[9] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C] //Proc of the 30th Annual Conf on Neural Information Processing Systems. New York: Curran Associates, 2017: 5990-6008

[10] Jurasky D, Martin J H. Speech and Language Processing [M]. New York: Prentice Hall, 2000

[11] Le Q, Mikolov T. Distributed representations of sentences and documents [C] //Proc of the 31st Int Conf on Machine Learning. Cambridge, MA: MIT Press, 2014: 1188-1196

[12] Kiros R, Zhu Yukun, Salakhutdinov R, et al. Skip-Thought vectors [C] //Proc of the 28th Annual Conf on Neural Information Processing Systems. New York: Curran Associates, 2015: 3294-3302

[13] Nayak P. Understanding searches better than ever before [OL]. [2019-10-25]. https://blog.google/products/search/search-language-understanding-bert

[14] Pan Xudong, Zhang Mi, Ji Shouling, et al. Privacy risks of general-purpose language models [C] //Proc of the 41st IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2020: 1314-1331

[15] Biggio B, Corona I, Maiorca D, et al. Evasion attacks against machine learning at test time [OL]. [2017-08-21]. https://arxiv.org/abs/1708.06131

[16] Chen Shu, Ju Zeqian, Fang Hongchao, et al. MedDialog: A large-scale medical dialogue dataset [OL]. [2020-07-07]. https://arxiv.org/abs/2004.03329

[17] Cui Yiming, Che Wanxiang, Liu Ting, et al. Pre-Training with whole word masking for Chinese BERT [OL]. [2019-10-29]. https://arxiv.org/abs/1906.08101

[18] Cui Yiming, Che Wanxiang, Liu Tin, et al. Revisiting pre-trained models for Chinese natural language processing [OL]. [2020-11-02]. https://arxiv.org/abs/2004.13922

[19] Zhang Zhengyan, Han Xu, Liu Zhiyuan, et al. ERNIE: Enhanced language representation with informative entities [C] //Proc of the 57th Annual Meeting of the ACL. Stroudsburg, PA: ACL 2019: 1441-1451

[20] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations [C] //Proc of the 19th Annual Conf of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2018: 2227-2237

[21] Dai Zihang, Yang Zhilin, Yang Yiming, et al. Transformer-XL: Attentive language models beyond a fixed-length context [C] //Proc of the 57th Conf of the ACL. Stroudsburg, PA: ACL, 2019: 2978-2988

[22] Yang Zhiling, Dai Zihang, Carbonell J, et al. XLNet: Generalized autoregressive pretraining for language understanding [C] //Proc of the 32nd Annual Conf on Neural Information Processing Systems. New York: Curran Associates, 2019: 5754-5764

[23] Liu Yinhan, Ott M, Goyal N, et al. RoBERTa: A robustly optimized BERT pretraining approach [OL]. [2019-10-03]. http://arxiv.org/abs/1907.11692

[24] Lample G, Conneau A. Cross-lingual language model pretraining [OL]. [2019-10-03]. http://arxiv.org/abs/1901.07291

[25] Sun Yu, Wang Shuohuan, Li Yukun, et al. ERNIE 2.0: A continual pre-training framework for language understanding [C] //Proc of the AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2020: 8968-8975

[26] Chen Yufei, Shen Chao, Wang Qian, et al. Security and privacy risks in artificial intelligence systems [J]. Journal of Computer Research and Development, 2019, 56(10): 2135-2150 (in Chinese)
(陈宇飞, 沈超, 王骞, 等. 人工智能系统安全与隐私风险 [J]. 计算机研究与发展, 2019, 56(10): 2135-2150)

[27] Shokri R, Stronati M, Song Congzheng, et al. Membership inference attacks against machine learning models [C] //Proc of IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2017: 3-18

[28] Nasr M, Shokri R, Houmansadr A. Machine learning with membership privacy using adversarial regularization [C] // Proc of the 2018 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2018: 634-646

[29] Song Liwei, Shokri R, Mittal P. Privacy risks of securing machine learning models against adversarial examples [C] // Proc of the 2019 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2019: 241-257

[30] Salem A, Zhang Yang, Humbert M, et al. ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models [C/OL] //Proc of Network and Distributed Systems Security Symp. [2020-01-02]. https://dx.doi.org/10.14722/ndss.2019.23119

[31] Ganju K, Wang Qi, Yang Wei, et al. Property inference attacks on fully connected neural networks using permutation invariant representations [C] //Proc of the 2018 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2018: 619-633

[32] Melis L, Song Congzheng, Cristofaro E. Exploiting unintended feature leakage in collaborative learning [C] // Proc of the 40th IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2019: 691-706

[33] Fredrikson M, Jha S, Ristenpart T, et al. Model inversion attacks that exploit confidence information and basic countermeasures [C] //Proc of the 2015 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2015: 1322-1333

[34] Salem A, Bhattacharya A, Backes M, et al. Updates-Leak: Data set inference and reconstruction attacks in online learning [C] //Proc of USENIX Security Symp. Berkeley, CA: USENIX Association, 2020: 1291-1308

[35] Zhu Ligeng, Liu Zhijian, Han Song. Deep leakage from gradients [C] //Proc of the 30th Annual Conf on Neural Information Processing Systems. New York: Curran Associates, 2019: 14747-14756

[36] Tramèr F, Zhang Fan, Juels A, et al. Stealing machine learning models via prediction APIs [C] //Proc of USENIX Security Symp. Berkeley, CA: USENIX Association, 2016: 601-618

[37] Duddu V, Samanta D, Rao D, et al. Stealing neural networks via timing side channels [OL]. [2019-09-02]. http://arxiv.org/abs/1812.11720

[38] Wang Binghui, Gong N. Stealing hyperparameters in machine learning [C] //Proc of IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2018: 36-52

[39] Ganin Y, Ustinova E, Ajakan H, et al. Domain-Adversarial Training of Neural Networks [J]. Journal of Machine Learning Research, 2016, 17(59): 1-35

[40] Shetty R, Schiele B, Fritz M. A4NT: Author attribute anonymity by adversarial training of neural machine translation [C] //Proc of the 27th USENIX Security Symp, Berkeley, CA: USENIX Association, 2018: 1633-1650

[41] Chen Ting, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations [OL]. [2020-07-03]. http://arxiv.org/abs/2002.05709

[42] He Xiangnan, Liao Lizi, Zhang Hanwang, et al. Neural collaborative filtering [C] //Proc of the 26th Int Conf on World Wide Web. New York: ACM, 2017: 173-182

[43] Kingma D, Ba J. Adam: A method for stochastic optimization [OL]. [2016-10-05]. http://arxiv.org/abs/1412.6980

[44] Centers forMedicare & Medicaid Services. Physician and other supplier data CY 2016 [EB/OL]. [2019-08-01]. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier2016.html

[45] He Ruining, Julian M. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering [C] //Proc of the 25th Int Conf on World Wide Web. New York: ACM, 2016: 507-517

Pan Xudong, born in 1996. PhD candidate. His main research interests include machine learning and AI security.

潘旭东,1996 年生.博士研究生.主要研究方向为机器学习和 AI 安全.

Zhang Mi, born in 1980. PhD, associate professor. Her main research interests include machine learning and AI security.

张 谧,1980 年生.博士,副教授.主要研究方向为机器学习和 AI 安全.

Yan Yifan, born in 1997. Master candidate. His main research interests include machine learning and AI security.

颜一帆,1997 年生.硕士研究生.主要研究方向为机器学习和 AI 安全.

Lu Yifan, born in 1999. Undergraduate. His main research interests include machine learning and AI security.

陆逸凡,1999 年生.本科生.主要研究方向为机器学习和 AI 安全.

Yang Min, born in 1979. PhD, professor. Member of CCF. His main research interests include system security and AI security.

杨 珉,1979 年生.博士,教授.CCF 会员.主要研究方向为系统安全和 AI 安全.