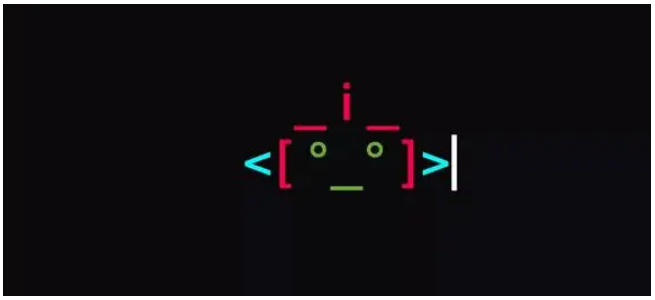
机器学习的重新思考：人工智能如何学习“失忆”?

Samuel Greengard AI科技评论 2022-04-04 12:53



作者 | Samuel Greengard

编译 | 维克多

机器学习已经成为各行各业的宝藏工具，常被用来构建系统，帮助人们发现那些容易忽略的细节，并辅助决策。尽管已经取得了惊艳的结果，但是也有很多痛苦，例如如何在已经成型的模型中修改、删减某些模块或者数据记录?

有学者表示，在大多数情况下，修改往往意味着重新训练，但仍然无法避免纳入可疑数据。这些数据可能来自系统日志、图像、客户管理系统等等。尤其是欧洲GDPR出台，对模型遗忘功能提出了更高的要求，企业如果不想办法将会面临合规处罚。

确实，完全重新训练的代价比较高，也不可能解决敏感数据问题。因此，我们无法证明重新训练的模型可以完全准确、有效。

为了解决这些问题，学者们定义了一种**“机器学习解除术”**（machine unlearning），通过分解数据库、调整算法等专门技术，诱导模型选择性失忆。机器学习解除术，顾名思义，就是让训练好的模型遗忘掉特定数据训练效果/特定参数，以达到保护模型中隐含数据的目的。

1

打破模型

机器学习之所以有魅力，是因为它能透过庞大的数据，超出人类认知范围的复杂关系。同时，这项技术的黑盒性质，让学者在修改模型时候，非常谨慎，毕竟无法知道一个特定的数据点处在模型的哪个位置，以及无法明确该数据点如何直接影响模型。

另外一种情况是：当数据出现异常值时，模型会记得特别牢，并对整体效果产生影响。

当前的数据隐私工具可以在数据脱敏的情况下训练模型，也可以在数据不出本地的情况下联合训练。或许可以将敏感数据替换成空值，引入噪声掩蔽敏感数据。但这些都无法从根本上解决问题。甚至，替代元素并保留关键数据的差异隐私技术也不足以解决选择性遗忘问题。例如它只能在单个案件或少数几个案件中发挥作用，在这些案件中，虽然不需要重新训练，但会有“敏感”的人要求从数据库中删除数据。随着越来越多的删除请求陆续到来，该框架的“遗忘模型"很快就会瓦解。

因此，隐私技术和机器学习解除术在解决问题的层面，并不能等同。

匿名无法验证和差分隐私技术的数据删除问题不仅是理论问题，而且会产生严重的后果。研究人员已经证明，人们总是有能力从所谓的通用算法和模型中提取敏感数据。例如2020年时候，学者发现，从GPT-2中可以获得包括个人身份和受版权保护的信息等训练数据。

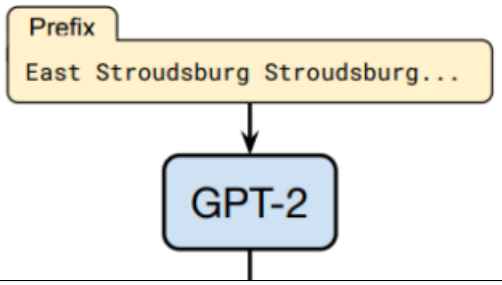
Extracting Training Data from Large Language Models

Nicholas Carlini[1] Florian Tramèr[2] Eric Wallace[3] Matthew Jagielski[4]
Ariel Herbert-Voss[5,6] Katherine Lee[1] Adam Roberts[1] Tom Brown[5]
Dawn Song[3] Úlfar Erlingsson[7] Alina Oprea[4] Colin Raffel[1]
[1]*Google* [2]*Stanford* [3]*UC Berkeley* [4]*Northeastern University* [5]*OpenAI* [6]*Harvard* [7]*Apple*

Abstract

It has become common to publish large (billion parameter) language models that have been trained on private datasets. This paper demonstrates that in such settings, an adversary can perform a *training data extraction attack* to recover individual training examples by querying the language model.

We demonstrate our attack on GPT-2, a language model trained on scrapes of the public Internet, and are able to extract hundreds of verbatim text sequences from the model's training



2

选择性遗忘

让机器学习模型获得选择性遗忘的能力，需要解决两个关键问题：

1.理解每个数据点如何机器学习模型；

2.随机性如何影响空间。例如需要弄清，在某些情况下，数据输入中相对较小的变化为何会产生不同的结果。

该方向的最初研究出现在在2019年。当时，Nicolas Papernot提出将机器学习的数据分割成多个独立的部分，通过建立众多的迷你数据，从而实现只对特定组件进行删除和再训练，然后插回完整的数据集中，生成功能齐全的机器学习模型。

In 42nd *IEEE Symposium of Security and Privacy*

Machine Unlearning

Lucas Bourtoule*‡§, Varun Chandrasekaran*†, Christopher A. Choquette-Choo*‡§, Hengrui Jia*‡§,
Adelin Travers*‡§, Baiwu Zhang*‡§, David Lie‡, Nicolas Papernot‡§
University of Toronto‡, Vector Institute§, University of Wisconsin-Madison†

***Abstract*—Once users have shared their data online, it is generally difficult for them to revoke access and ask for the data to be deleted. Machine learning (ML) exacerbates this problem because any model trained with said data may have memorized it, putting users at risk of a successful privacy attack exposing their information. Yet, having models unlearn is notoriously difficult.**

We introduce SISA training, a framework that expedites the unlearning process by strategically limiting the influence of a

2020

data motivates us to examine how this right to be forgotten can be efficiently implemented for ML systems.

Because ML models potentially memorize training data [10], [11], it is important to unlearn what they have learned from data that is to be deleted. This problem is tangential to privacy-preserving ML—enforcing ε-differential privacy [12] with $\varepsilon \neq 0$ does not alleviate the need for

具体操作过程是：先将训练数据分成多个不相交的切片，且一个训练点只包含在一个切片中；然后，在每个切片上单独训练模型；随后，合并切片，成功删除数据元素。因此，当一个训练点被要求遗忘时，只需要重新训练受影响的模型。由于切片比整个训练集更小，就减少了遗忘的代价。

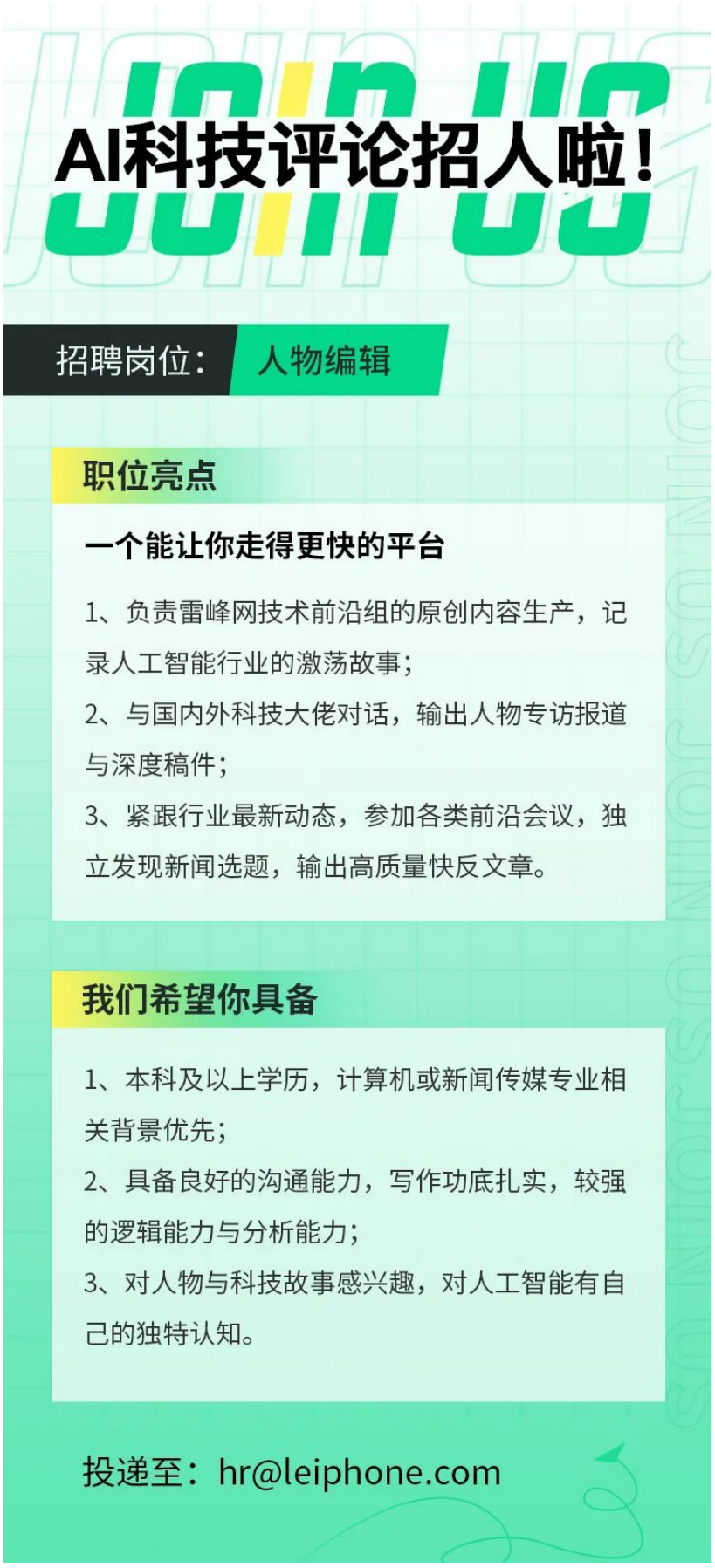
该方法被Nicolas Papernot命名为**SISA**（Sharded, Isolated, Sliced, and Aggregated ），对比完全重训练和部分重训练的基线, SISA实现了准确性和时间开销的权衡。在简单学习任务中, 在数据集Purchase上是4.63x, 在数据集 SVHN上是2.45x。

同时，作者也承认，虽然这个概念很有前途，但也有局限性。例如，通过减少每个切片的数据量，会对机器学习产生影响，并且可能会产生质量较低的结果。此外，这项技术并不总是像宣传的那样奏效。

目前，机器学习遗忘术的研究仍处于初级阶段。随着研究人员和数据科学家深入了解删除数据对整体模型的影响，成熟的工具也会出现，其目标是：机器学习框架和算法允许学者删除一条记录或单个数据点，并最终得到一个“完全遗忘”相关数据的有效模型。

参考链接：

https://cacm.acm.org/magazines/2022/4/259391-can-ai-learn-to-forget/fulltext#FNA



谷歌下一代AI架构、Jeff Dean宣传大半年的Pathways终于有论文了

机器之心

“请给AI一些包容。”

AI科技评论