机器学习的遗忘方式

韩志峰

班级-学号 2118021-21009200659

摘 要 本报告探讨了“机器的遗忘”概念，即在人工智能和机器学习中如何有效地移除已学习的信息。我们分析了机器忘却的技术需求，提出了几种实现方法，如通过模型修剪、数据反向传播清除（data influence backpropagation）、以及基于知识蒸馏的删除策略。在实验中，我们对这些方法在不同的数据集和模型上的表现进行了比较，特别是评估了其在性能、数据清除效率和遗忘的完整性方面的差异。结果表明，与传统方法相比，基于知识蒸馏的遗忘方法能在有效保持模型性能的前提下实现高效的数据清除。

关键词 机器遗忘；模型修剪；数据清除；知识蒸馏；隐私保护

**How to Unlearn: Approaches for Machine Forgetting**

Zhifeng Han1)

1) (School of Cyber Engineering, Xidian University, Xi’an 710071, China)

**Abstract** This report explores the concept of “machine forgetting”, i.e., how to effectively remove learned information in artificial intelligence and machine learning. We analyze the technical requirements for machine forgetting and propose several implementation methods, such as model pruning, data influence backpropagation, and knowledge distillation-based removal strategies. In our experiments, we compare the performance of these approaches on different datasets and models, and in particular, evaluate their differences in terms of performance, data removal efficiency, and forgetting integrity. The results show that the knowledge distillation-based forgetting approach achieves efficient data erasure while effectively maintaining the performance of the model compared to traditional methods.

**Key words** Machine forgetting; model pruning; data erasure; knowledge distillation; privacy preservation

# 引言

在现代信息社会中，数据隐私和安全问题日益受到关注。随着人工智能（AI）和机器学习（ML）技术的广泛应用，系统中积累的海量数据成为一种不可忽视的资源，但也带来了数据滥用和隐私泄露的风险。近年来，随着《通用数据保护条例》（GDPR）等法律法规的出台，数据隐私保护成为技术开发的强制性要求，这推动了“遗忘权”的发展。机器的遗忘，即“how to unlearn”，指的是机器能够主动或被动地删除已学习的信息，使其不再受特定数据影响。这一概念不仅在隐私保护中具有重要意义，也可用于数据清理、模型改进等多个领域。

## 机器遗忘的背景与必要性

在传统机器学习和深度学习模型中，一旦模型从数据中学得了某些特征或规律，这些知识便成为模型不可分割的一部分。这使得在模型训练完毕后移除某些数据的影响变得极为困难。尤其是在使用敏感或隐私数据的应用场景中，遗忘这些数据的能力变得至关重要。例如，金融数据、医疗记录和社交媒体等领域中用户信息的保护要求，要求系统能够在用户撤回数据或进行隐私删除后保证相关数据对模型不再产生影响。

## 机器遗忘的主要挑战

现有机器遗忘的方法主要有三类：模型修剪、数据反向传播清除和知识蒸馏。这些方法各有其特点，但也面临一定挑战。首先是遗忘的完整性问题，即模型能否完全移除某些特定数据的影响；其次是计算效率问题，尤其是大规模数据集和复杂模型中，数据清除往往计算量大、耗时长；此外，如何在删除部分数据后保持模型性能，也是实现机器遗忘时的重要考量因素。

例如，模型修剪通过移除冗余参数来减少特定数据的影响，适用于大型网络，但可能影响模型性能。数据反向传播清除则通过调整模型参数来减少数据影响，但其计算成本较高，适用性受到限制。知识蒸馏方法则能在一定程度上有效地删除数据影响，同时保留模型性能，但实现过程较为复杂，适用于特定场景。

## 本文的研究内容

本文旨在复现机器遗忘的现有研究成果，分析三种典型机器遗忘方法在不同数据集上的表现差异。通过对比模型修剪、数据反向传播清除和知识蒸馏等方法，本文尝试展示其在不同应用场景中的适用性和局限性。通过实验复现，本文期望为机器遗忘领域的基础研究提供详细的数据和案例分析。

# 相关工作

机器遗忘技术在数据隐私保护领域迅速发展，现有研究主要集中在三种实现方法上：模型修剪、数据反向传播清除以及知识蒸馏。这些方法各自侧重不同的实现思路和应用场景。以下为对这些方法的简要综述。

## 模型修剪

模型修剪（Model Pruning）是一种通过减少神经网络中不必要的参数和节点来实现机器遗忘的方法。其基本思路是认为特定数据的学习往往与部分参数或网络结构有关，通过移除这些参数便可以减少相关数据的影响。文献[1]指出，模型修剪在大规模网络中表现良好，特别是适合去除特定数据的影响而不显著降低整体性能。然而，模型修剪在小型网络中应用时会遇到一定的瓶颈，部分情况下会显著降低模型的整体性能。此外，修剪过程可能难以实现完全遗忘，部分数据影响仍然可能残留在模型中。

## 数据反向传播清除

数据反向传播清除（Data Influence Backpropagation）是一种直接通过训练过程来删除特定数据影响的方法。该方法的基本步骤是在移除目标数据后，对模型进行再训练，调整模型参数，使之不再保留目标数据的痕迹。文献[2]展示了该方法的基本流程，但也指出了其计算复杂度较高的问题。对于大规模数据集，重新训练可能十分耗时，因此数据反向传播清除方法的适用性主要局限于敏感数据较少的小规模模型。然而，该方法通常能够更好地保证删除的完整性，并且适合隐私需求较高的场景。

## 知识蒸馏与数据删除

知识蒸馏（Knowledge Distillation）是一种较为灵活的机器遗忘方法，通过在新模型中复制旧模型的知识，但不包含特定数据的影响。知识蒸馏方法将“教师模型”中的有效信息传递到“学生模型”中，通过有选择地删除特定数据的影响实现机器遗忘。文献[3]指出，知识蒸馏能够在保持模型准确性的前提下有效去除特定数据的影响，但在实际实现中需要设置适当的损失函数和超参数。此外，知识蒸馏适用于一些对模型性能要求较高的应用，但构建新的“学生模型”增加了实现复杂度，尤其是在处理涉及多类标签的大型数据集时。

**正文文字要求语句通顺，无语法错误，结构合理，条理清楚，不影响审稿人、读者阅读理解全文内容。以下几类问题请作者们特别注意**：

1)文章题目应明确反映文章的思想和方法；文字流畅，表述清楚；

2)中文文字、英文表达无语法错误；

3)公式中无符号、表达式的疏漏，没有同一个符号表示两种意思的情况；

4)数学中使用的符号、函数名用斜体；

5)使用的量符合法定计量单位标准；

6)矢量为黑体，标量为白体；

7)变量或表示变化的量用斜体；

8)图表规范，量、线、序无误，位置正确（图表必须在正文中有所表述后出现，即…如图1所示）(注意纵、横坐标应有坐标名称和刻度值)。

9)列出的参考文献必须在文中按顺序引用，即参考文献顺序与引用顺序一致，各项信息齐全(格式见参考文献部分)；

10）首次出现的缩写需写明全称，首次出现的符号需作出解释。

11）图的图例说明、坐标说明全部用中文或量符号。

12）图应为矢量图。

13）表中表头文字采用中文。

14）公式尺寸：

标准：10.5磅

下标/上标：5.8磅

次下标/上标：4.5磅

符号：16磅

次符号：10.5磅

15）组合单位采用标准格式，如：“pJ/bit/m4”应为 “pJ/(bit·m4)”

定理**1.** \*\*\*\*\*\*. \*定理内容.\*

[“定义”、“假设”、“公理”、“引理”等的排版格式与此相同，详细定理证明、公式可放在附录中]

证明. \*证明过程.\* [“例 x”等的排版格式相同]

证毕.



图X 图片说明 \*字体为小5号，图片应为黑白图，图中的子图要有子图说明\*

表X 表说明 \*表说明采用黑体\*

|  |  |
| --- | --- |
| \*示例表格\* | \*第一行为表头,表头要有内容\* |
|  |  |
|  |  |
|  |  |
|  |  |

算法**X**. 算法名称.

输入：… …

输出：… …

\*《计算机学报》的算法描述字体为小5号宋体, IF 、THEN等伪代码关键词全部用大写字母，变量和函数名称用斜体\*

过程 X. 过程名称

\*《计算机学报》的方法过程描述字体为小5号宋体，IF 、THEN等伪代码关键词全部用大写字母，变量和函数名称用斜体\*

参 考 文 献

[1] 网上的文献[[1]](#footnote-1)(举例：The Cooperative Association for Internet Data Analysis(CAIDA),http://www.caida.org/data 2010,7,18) **\*请采用脚注放于正文出现处，每页的脚注从1开始编序号\***

[2] 中文的参考文献需给出中英文对照。形式如[3]。

[3] Zhou Yong-Bin, Feng Deng-Guo. Design and analysis of cryptographic protocols for RFID. Chinese Journal of Computers, 2006, 29(4): 581-589 (in Chinese)  
(周永彬, 冯登国. RFID安全协议的设计与分析. 计算机学报, 2006, 29(4): 581-589)

[4] 期刊、会议、书籍名称不能用缩写。

[编号] 作者(外国人姓在前，名在后可缩写, 后同). 题目(英文题目第一字母大写，其它均小写)：副标题(如果有). 刊名(全称), 年, 卷(期): 页码 **\*期刊论文格式\***

[编号] 作者. 文章题目(英文题目第1字母大写，其它均小写)：副标题(如果有)//Proceedings of the … (会议名称). 会议召开城市, 会议召开城市所在国家, 年: 页码 **\*会议论文集论文格式\***

[编号] 作者. 文章题目(英文题目第一字母大写, 其它均小写): 副标题(如果有)//编者. 文集标题. 出版地: 出版社, 出版年: 页码 **\*文集格式\***

[编号] 作者. 书名: 副标题(如果有). 版次(初版不写). 出版社地点: 出版社, 出版年 **\*书籍格式\***

[编号] 作者. 文章题目[博士学位论文/硕士学位论文]. 单位名称,单位地点, 年 **\*学位论文格式\***

[编号] 作者. 文章题目(英文题目第一字母大写，其它均小写). 单位地点: 单位, 技术报告: 报告编号, 年 **\*技术报告\***

[编号] 专利拥有人. 专利名称，专利授权国家，专利授权日期

**\*技术专利\***

1. The Cooperative Association for Internet Data Analysis(CAIDA),http://www.caida.org/data 2010,7,18 [↑](#footnote-ref-1)