Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

人工智能模型的遗忘学习方法

硕士研究生 赵怡清

2024年10月20日

问题回溯

- 总结反思
  - 对于公式的标注欠缺
  - 未绘制算法原理图
- 相关内容
  - 后门攻击防御（涉及遗忘学习）
    - 2024.01.14 赵怡清：《对抗性扰动下的后门防御方法》

内容提要

- 预期收获
- 内涵解析
- 研究背景与意义
- 研究历史与现状
- 知识基础
- 算法原理
  - Machine Unlearning
  - Zero-Shot Machine Unlearning
- 未来展望
- 参考文献

预期收获

- 预期收获
  - 掌握遗忘学习的基本概念
  - 理解两种遗忘学习算法的原理
  - 理解遗忘学习在理论和实际中的意义和作用

内涵解析　人工智能模型的遗忘学习方法

- 研究目标
  - 面向人工智能领域的遗忘学习方法
  - 研究适用于不同领域数据集的通用遗忘学习方法，遗忘训练数据中的特定样本或者特定类别，保护个人隐私

- 内涵解析
  - 人工智能模型：泛指人工智能领域中的模型，包括机器学习和深度学习模型
  - 遗忘学习：是指机器学习系统中先前获取的信息或知识随着时间的推移而下降的现象

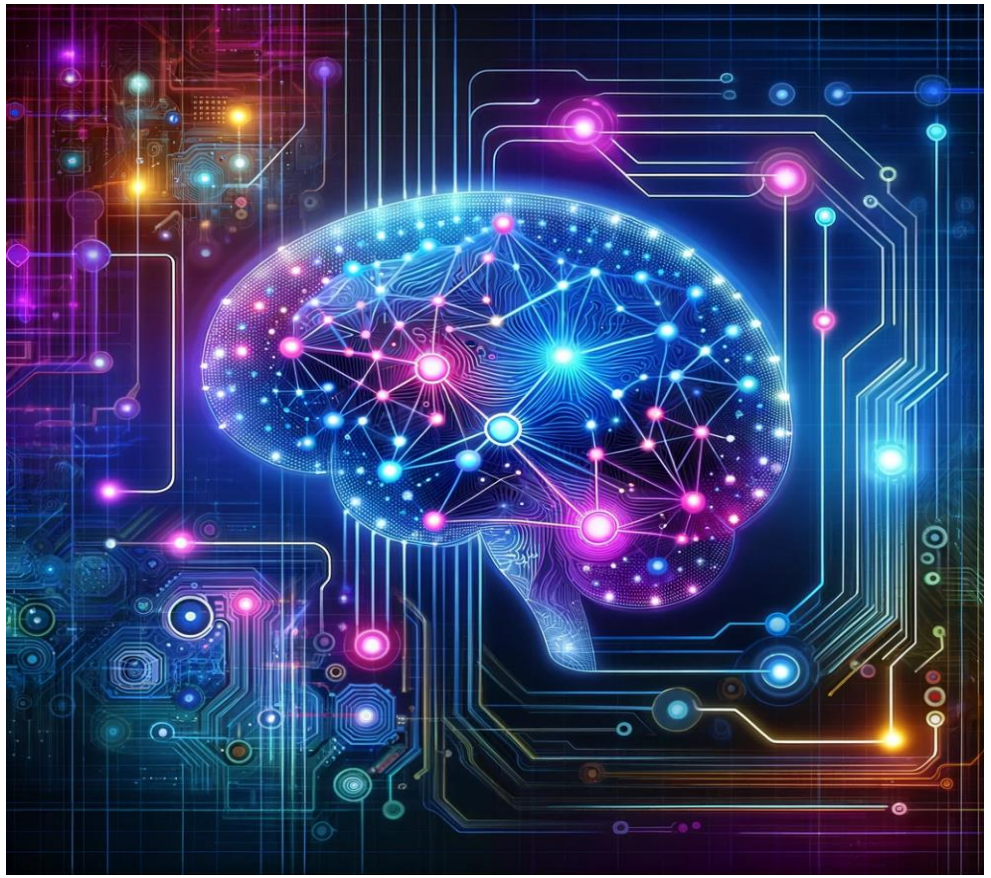
背景意义　人工智能模型的遗忘学习方法

- 研究背景
  - 人工智能技术发展迅速，在人脸识别、医疗诊断、内容推送等领域取得了优异的成果，但面临着数据隐私泄露和数据安全问题
  - 人工智能模型会因为数据集偏差，使用敏感属性特征来进行决策，从而产生不公平的歧视

2020年，Clearview AI的面部识别应用泄露了30亿张人脸数据，

2020年，ExamSoft 远程考试的人脸识别系统被发现对有色人种识别成功率更低

- 研究意义
  - 人工智能模型训练过程是随机，从所有训练样本中删除特定的数据样本，可以使我们更加了解每个数据点对于模型的影响
  - 模型训练是一个增量过程，模型对于当前样本的处理会受到前一个数据样本的影响，如何取消删除样本对于模型性能的影响是当前面临的挑战之一
  - 通过研究取消学习方法，可以增加人们对于灾难性遗忘的了解，对于自然的防止灾难性遗忘提供解决思路



研究历史

Guo等提出的一步牛顿更新发源自早期的数据异常度量研究，通过观察删除某个样本后模型的参数变化衡量该样本的异常程度。

Bourtoule等受集成学习和并行计算的启发，提出了能够快速实现遗忘学习的SISA算法，其中SISA由碎片化、隔离、切片和聚合的简写。

Chundawat等延伸了UNSIR算法的研究，将原始模型指定为教师，通过随机初始化、结构相同的网络进行学习，通过滤波器阻止与要遗忘学习的类别相关信息的流动。

Chen等研究如何通过决策空间来达到对DNN模型中某一类样本的遗忘，观察重新训练后的模型对于遗忘类别预测结果的分布情况。

2020

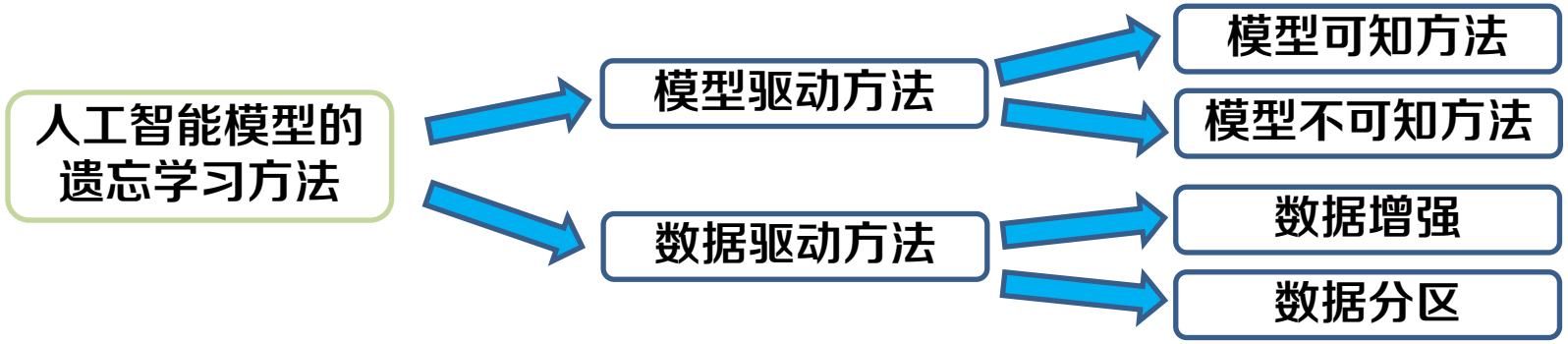
Wu等人提出的DeltaGrad算法，使用近似梯度来估计删除数据后的模型参数更新，适用于采用梯度下降方法训练的机器学习模型

2021

Jonathan等针对二分类问题的决策树和随机森林模型提出了一种特定的遗忘学习算法DaRE。

2023

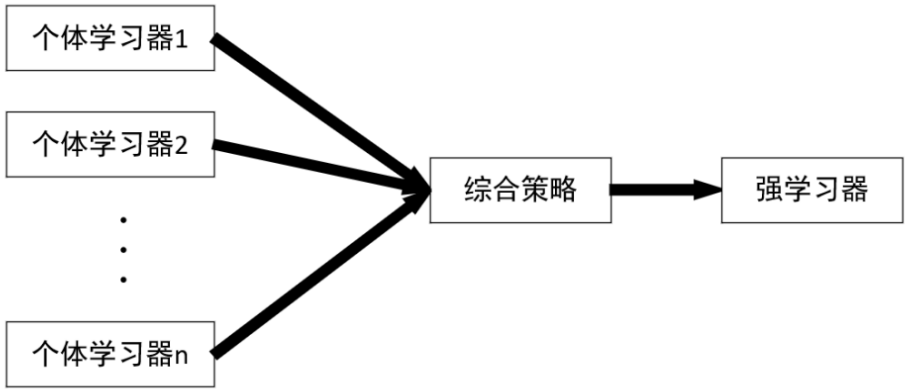
Tarun等人提出的UNSIR算法引入了Zero-shot遗忘学习的概念，用于解决在无法访问原始数据的情况下实现遗忘学习

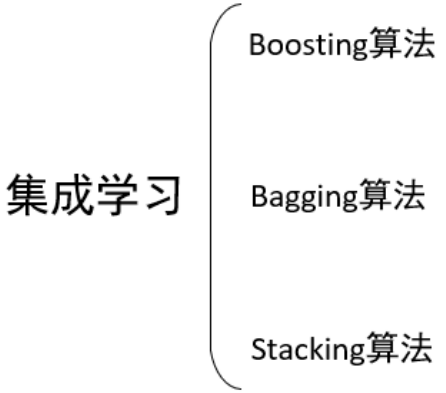


- 遗忘学习
  - 定义：也被称为机器遗忘或取消学习，是指机器学习系统中先前获取的信息或知识随着时间的推移而退化的现象
  - 挑战
    - 平衡旧任务知识与新任务的快速学习
    - 管理具有冲突目标的任务干扰
    - 防止隐私泄露
  - 分类
    - 有害的遗忘
      - 在适应新任务、领域或者环境的同时，模型无法保留先前学到的知识
    - 有益的遗忘
      - 模型可能包含导致隐私泄露的私人信息或者不相关的信息阻碍新任务的学习

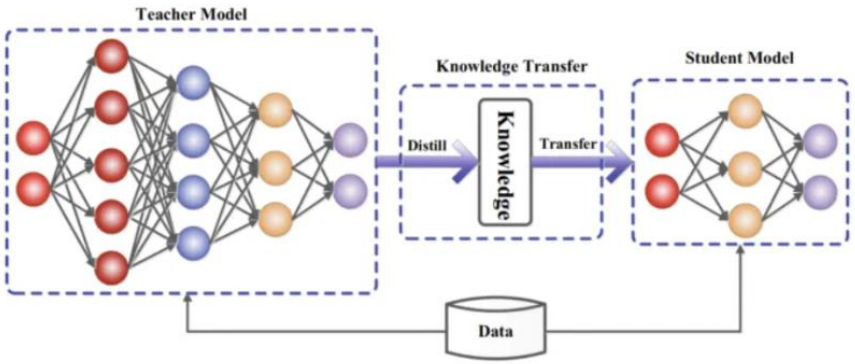
遗忘也是一把双刃剑

- 集成学习
  - 集成学习是一种机器学习方法，通过将多个学习器组合在一起来解决问题，它的目标是通过结合多个学习器的预测结果，以获得比单个学习器更准确和鲁棒的预测能力
  - 弱学习器：是指在某个学习任务上略好于随机猜测的学习器，它的准确率可能较低，但它的预测结果略好于随机猜测
  - 强学习器：是指在某个学习任务上具有较高准确率和泛化能力的学习器，它能够在复杂的问题上进行准确预测，并具有较强的泛化能力





- 知识蒸馏
  - 定义：是一种模型压缩技术，旨在通过训练一个小而高效的模型来捕获大型、复杂模型中所包含的知识
    - 其中大模型称为教师模型，小模型称为学生模型
  - 分类
    - 基于知识分类
      - 基于响应知识：学生模型直接模仿教师模型的最终预测结果
      - 基于特征知识：学生模型直接模仿教师模型的中间层输出
      - 基于关系知识：学生模型模仿教师模型数据样本之间的相互关系
    - 基于蒸馏机制分类
      - 离线蒸馏：教师模型生成软标签指导学生模型训练
      - 在线蒸馏：教师模型和学生模型同步更新
      - 自蒸馏：教师和学生模型采用相同的网络



Machine Unlearning

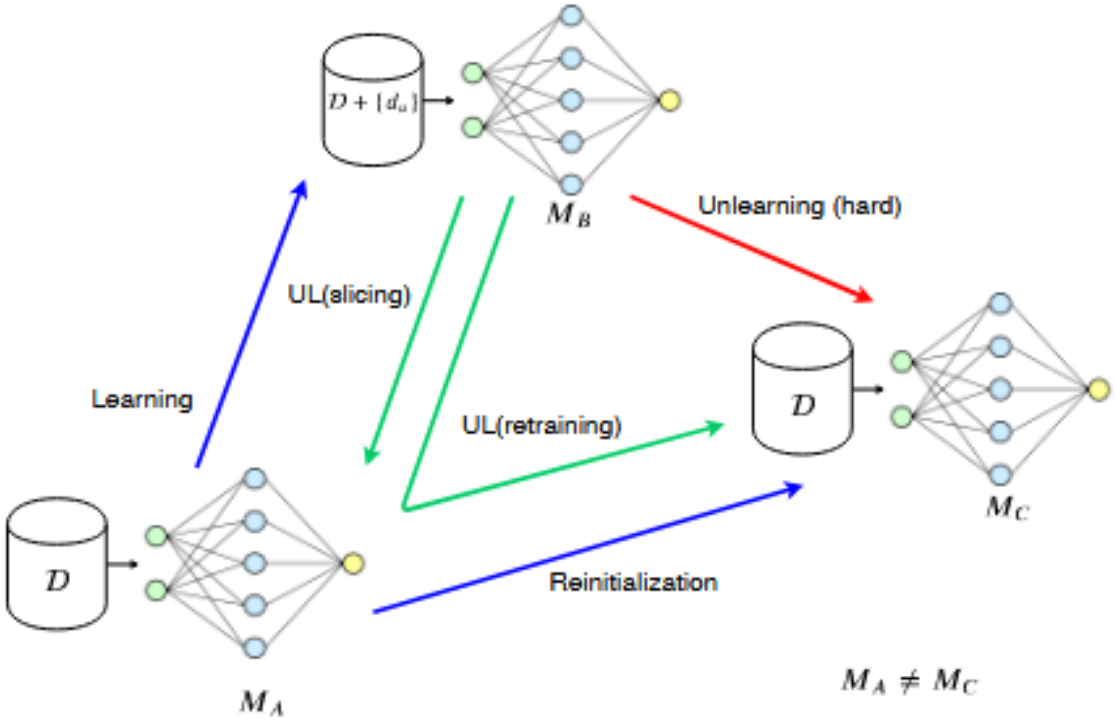
算法原理 TIPO

T	目标	遗忘深度学习模型中的特定样本
I	输入	训练数据集*6、深度学习模型框架*1、遗忘样本*n
P	处理	1.将数据集划分为多个不相交分片和切片 2.在每个分片上单独训练模型 3.重训练受影响模型并聚合所有模型
O	输出	遗忘特定样本的深度学习模型*1

P	问题	模型训练中的随机性、对于每个数据点影响模型的理解有限
C	条件	不影响模型的性能、减少取消学习的开销
D	难点	合理的划分并组合分片和切片
L	水平	S&P 2021 CCF A类

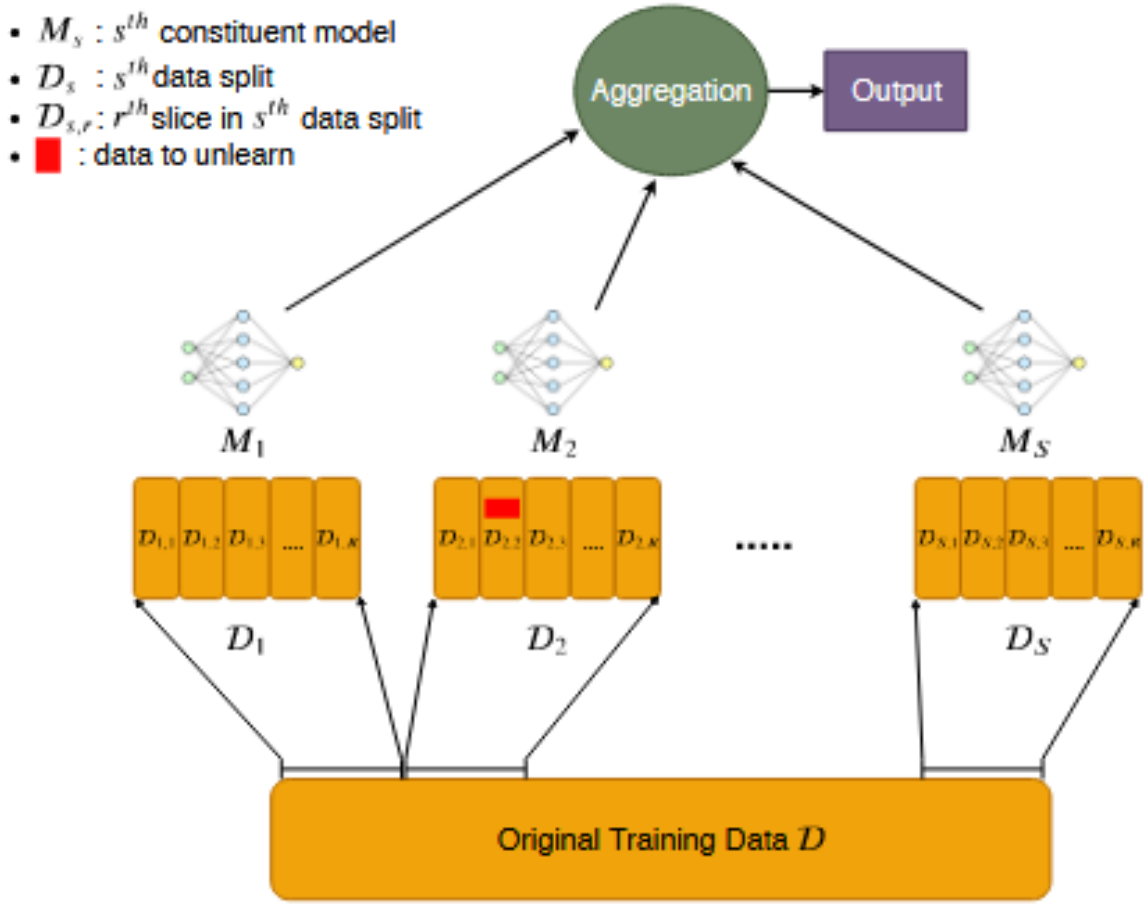
创新分析　遗忘难题 + 提出直观遗忘定义

- 直观遗忘定义
  - 以一种博弈形式定义遗忘问题
  - 两种角色：服务提供者S，用户U；训练好的模型M
  - 当用户服务希望删除数据时， S修改模型参数获得模型M'，使用户相信修改模型和原模型参数分布相同
- 遗忘难题
  - 假设初始数据集D训练一个模型M_A
  - 加入新的数据点d训练得到模型M_B
  - 直接遗忘新数据点回到M_C比较困难
    - 无法衡量d对于模型的影响
    - 不保存M_A参数很难恢复



SISA 算法原理图

- 算法原理图
  - 第一步：分片和切片
    - 将数据集划分为多个分片和切片
  - 第二步：训练组成模型
    - 仅使用$D_{k,1}$随机初始化训练模型$M_{k,1}$并保存模型参数
    - 使用$D_{k,1} \cup D_{k,2}$训练模型$M_{k,1}$2个epochs，生成模型$M_{k,2}$
    - 循环遍历完所有切片获得组成模型M_k
  - 第三步：聚合
    - 最终输出结果根据每个组成模型的输出结果加权决定



SISA 实验设计

- 实验数据集

Dataset	Dimensionality	Size	# Classes
MNIST [43]	28×28	60000	10
Purchase [49]	600	250000	2
SVHN [50]	$32 \times 32 \times 3$	604833	10
CIFAR-100 [51]	$32 \times 32 \times 3$	60000	100
Imagenet [44]	$224 \times 224 \times 3$	1281167	1000
Mini-Imagenet [48]	$224 \times 224 \times 3$	128545	100

- 两种基线方法
  - 批量遗忘K个样本(Batch K)
    - 剔除遗忘样本后，按照初始设置从头开始重新训练模型
  - 部分数据训练(1/S)
    - 当遗忘样本位于训练数据中时，重新选取部分数据训练模型

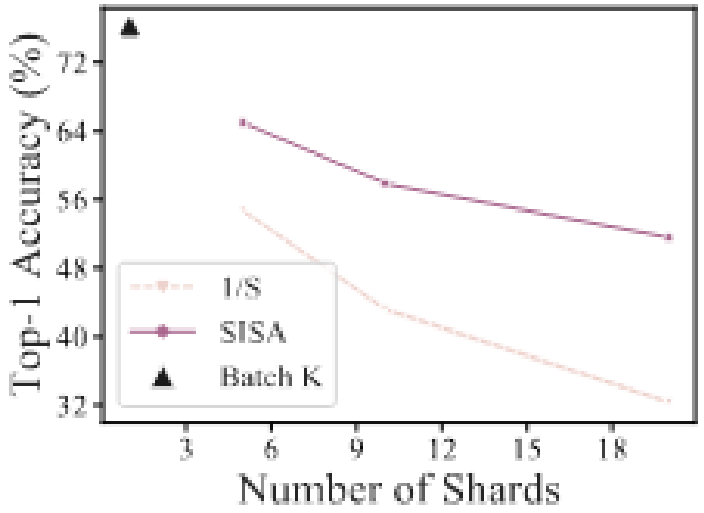
- 实验模型

Dataset	Model Architecture
MNIST [43]	2 conv. layers followed by 2 FC layers
Purchase [49]	2 FC layers
SVHN [50]	Wide ResNet-1-1
CIFAR-100 [51]	ResNet-50
Imagenet [44]	ResNet-50
Mini-Imagenet [48]	ResNet-50

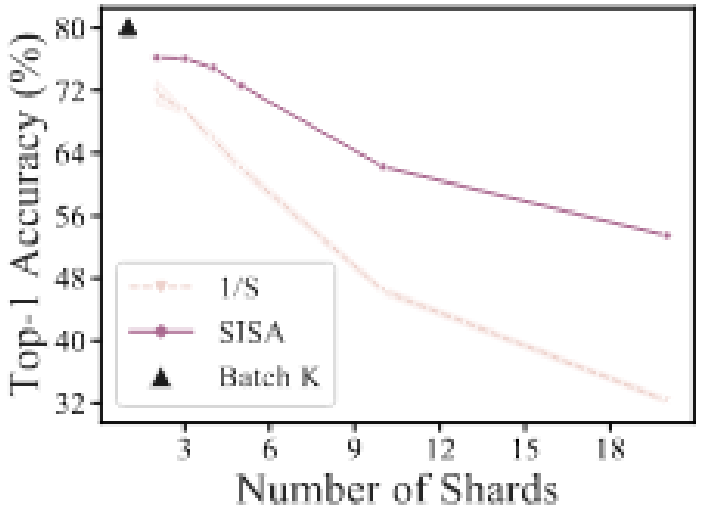
- 评价指标
  - ACC：模型基于样本的预测准确率
  - Analytical Time：模型遗忘特定样本消耗的时间

实验设计　问题一

- **Question1：对于不同数量的遗忘样本，分片对准确性有何影响?**



(a) Imagenet dataset

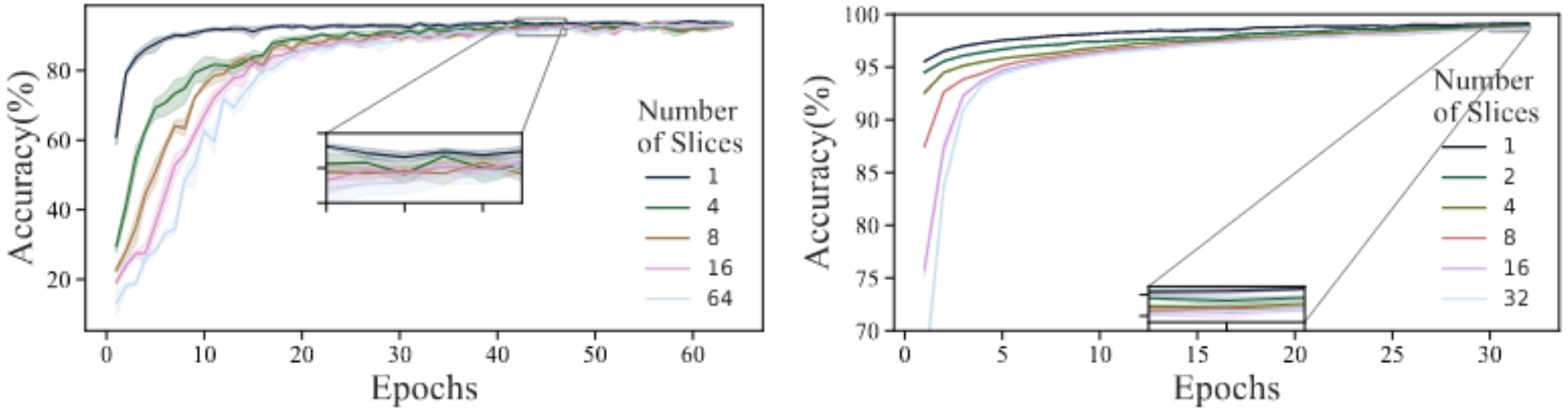


(b) Mini-Imagenet dataset

- 实验结果
  - 随着分片增加，模型的预测精度下降
- 实验结论
  - 分片数量较少时，预测精确度下降在可接受范围内

实验设计　问题二

- **Question2：对于不同数量的遗忘样本，切片对准确性有何影响?**

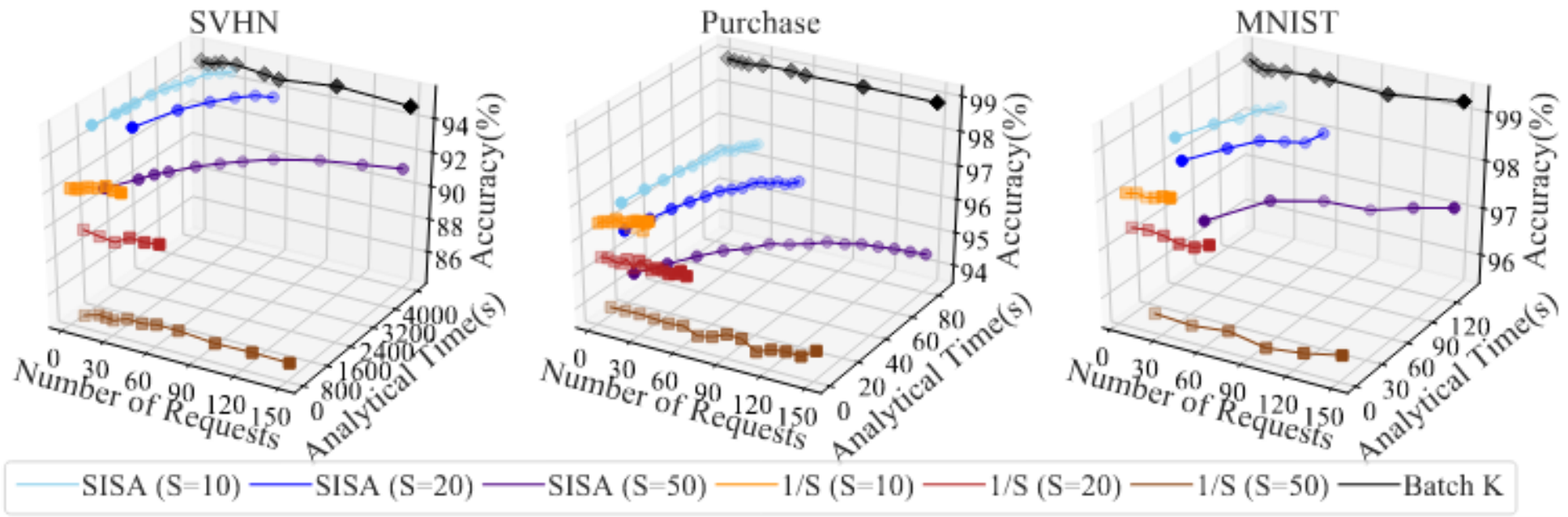


(a) Accuracy vs. Number of epochs for SVHN dataset. (b) Accuracy vs. Number of epochs for Purchase dataset.

- 实验结果
  - 切片较多，训练epochs较少时，预测精度较低
- 实验结论
  - 当训练epochs增加，预测准确性达到稳定水平

实验设计　问题三

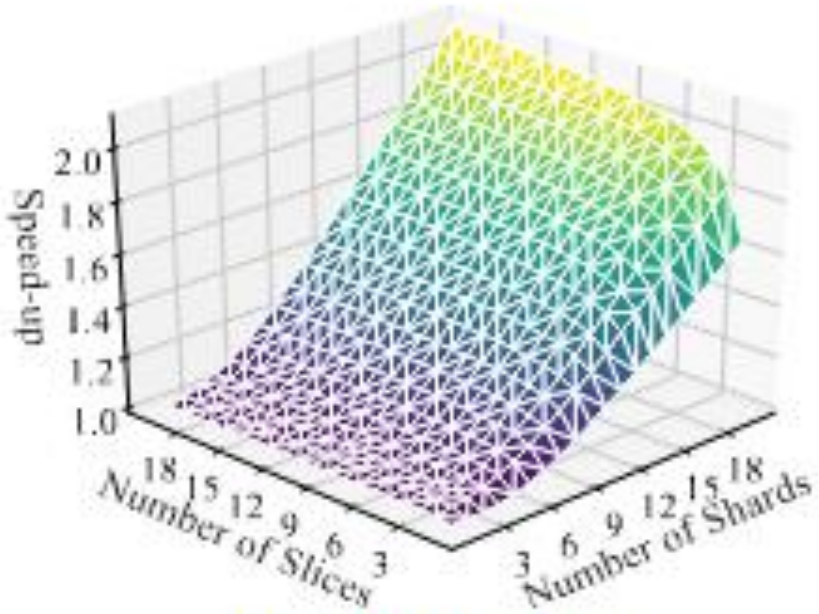
- **Question3：SISA训练是否可以缩短遗忘时间?**



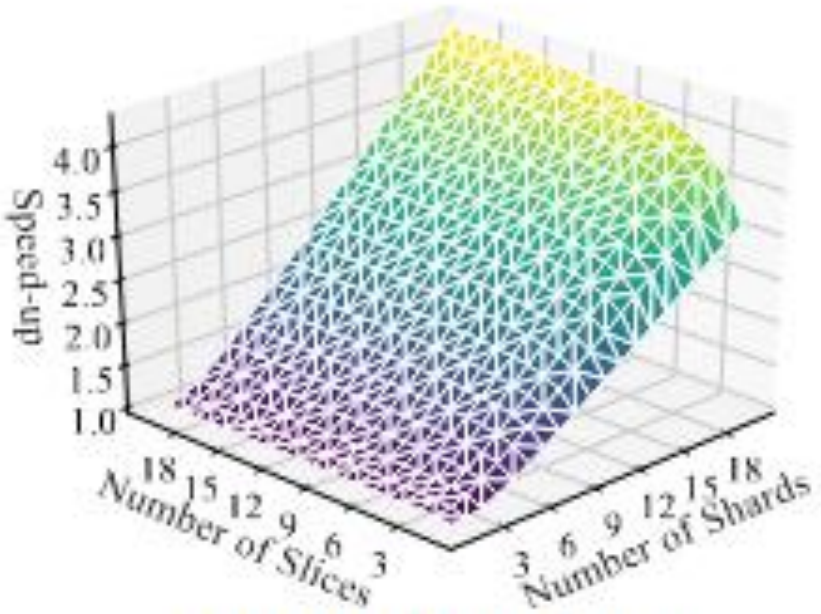
- **实验结论**
  - **SISA训练比1/S基线提供更高的准确度，比K批次基线更少的训练时间，特别是当遗忘样本的数量较少时**

实验设计　问题三

- **Question3：SISA训练是否可以缩短遗忘时间?**



(a) SVHN dataset

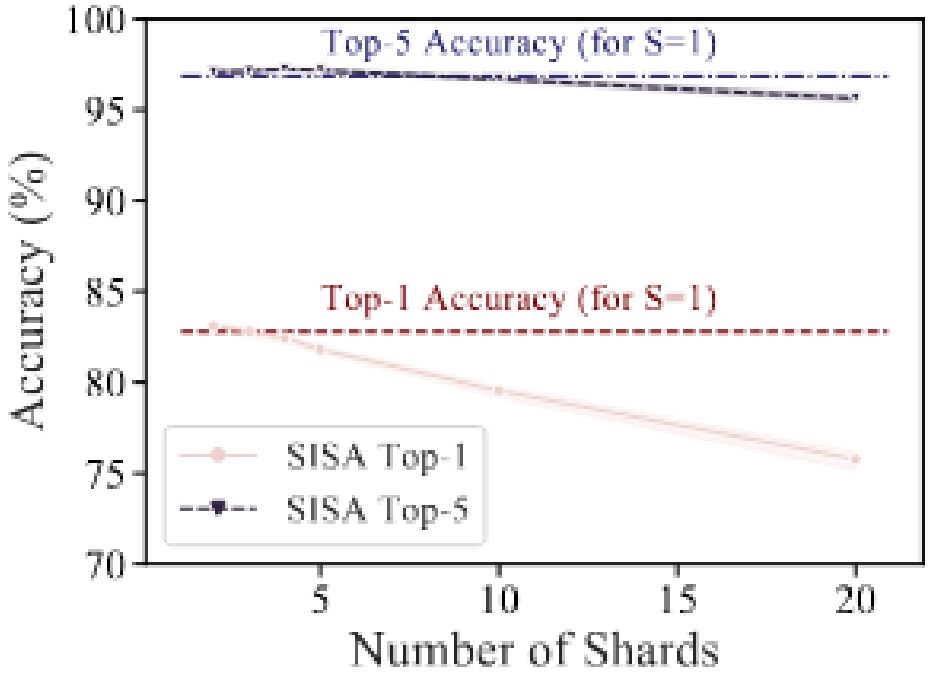


(b) Purchase dataset

- 实验结论
  - 分片和切片的组合可以加速固定数量的样本的遗忘学习

实验设计　问题四

- **Question4：SISA对于简单和复杂的学习任务都适用吗?**



- 实验结论
  - 现实世界的复杂学习任务，常见的方法是利用公共数据集上训练的预训练模型，使用迁移学习将其定制为特殊任务

算法总结 SISA

- 算法流程
  - 将数据集划分为不同数量的分片和切片
  - 基于不同切片训练多个组成模型
  - 聚合多个组成模型的输出结果作为最终输出
- 算法优势
  - SISA以空间开销换取训练的时间开销
  - 实现了模型预测准确性和时间开销的权衡
- 算法不足
  - 不适用于决策树模型
  - 对于复杂的深度学习任务，分片过多导致精度大幅下降

Zero-Shot Machine Unlearning

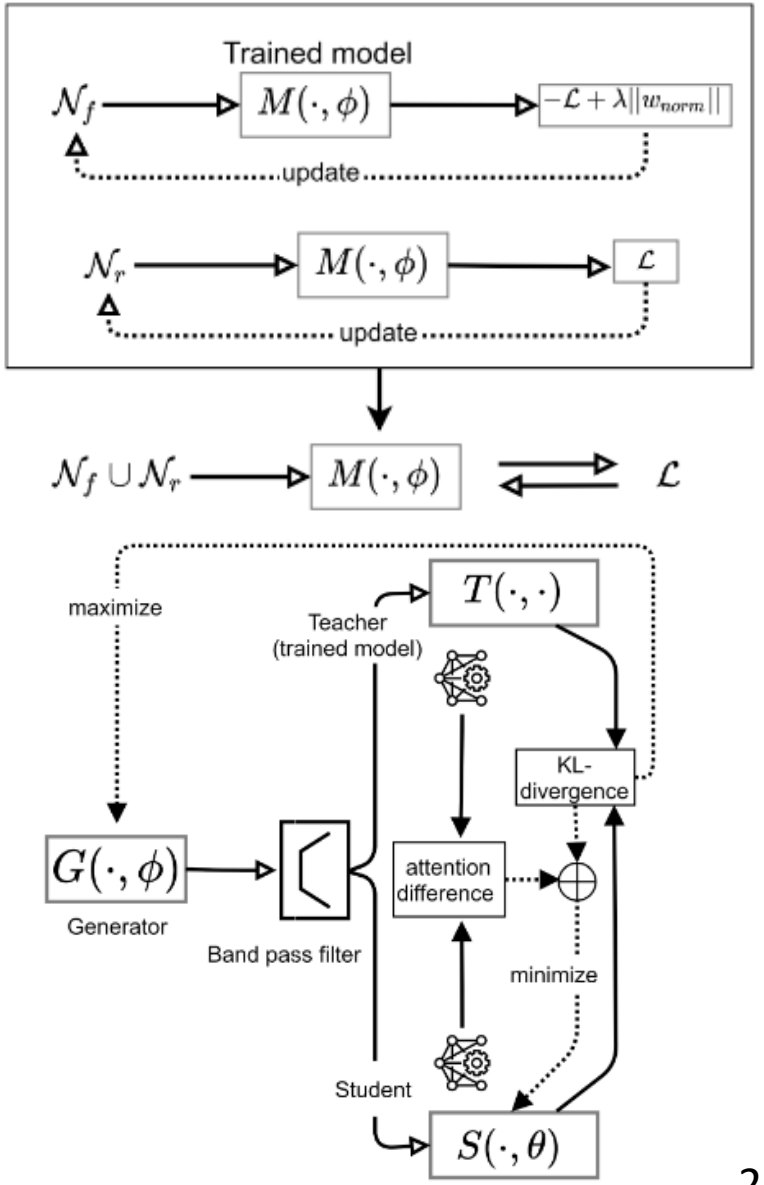
算法原理 TIPO

T	目标	遗忘深度学习模型中的特定类别
I	输入	已训练的深度学习模型*1、遗忘类别、保留类别
P	处理	1.使用生成器生成遗忘样本、保留样本 2.误差最小-噪声最大遗忘 3.门控知识传递遗忘
O	输出	遗忘特定类别的深度学习模型*1

P	问题	训练数据集难以访问、处于遗忘目的访问数据集十分昂贵
C	条件	删除模型中特定样本，保持对保留数据的预测准确性
D	难点	在没有原始训练数据的情况下实现遗忘学习
L	水平	TIFS 2023 SCI1区

Zero-Shot 算法原理

- 算法原理图
  - 方法一：误差最小化-最大化噪声
    - 第一步：通过隐私攻击的方法生成保留类别样本和遗忘类别样本
    - 第二步：对于遗忘样本，最大化噪声生成遗忘矩阵N_f，基于保留样本，最小化误差生成保留矩阵N_r
    - 第三步：构造$N_f + N_r$，重新训练模型
  - 方法二：门控知识传递
    - 第一步：根据KL散度生成训练样本， KL散度由教师模型和学生模型计算
    - 第二步：模仿教师网络，训练学生网络，损失函数由KL散度和注意力损失组合构成



Zero-Shot 算法原理

- 最初训练模型指定为教师模型，随机初始化的网络指定为学生模型
- 生成器基于最大化教师、学生模型输出向量之间的KL散度生成样本

$$D_{KL}(T(x_p)||S(x_p)) = \sum_i t_p^{(i)} log(t_p^{(i)}/s_p^{(i)})$$

- 学生模型以最小化KL散度以及注意力损失来更新权重

$$L_{at} = \sum_{l \in N_L} \left\| \frac{f(A_l^{(t)})}{||f(A_l^{(t)})||_2} - \frac{f(A_l^{(s)})}{||f(A_l^{(s)})||_2} \right\|$$

$$L_s = D_{KL}(T(x_p)||S(x_p)) + \beta L_{at}$$

- 设计一个带通滤波器，衰减遗忘类别的知识

$$F(x_p) = \prod_{i \in C_f} (t_p^{(i)} < \epsilon)$$

实验设计 Zero-Shot

- 实验数据集及实验设置
  - 误差最小-噪声最大：
    - CIFAR10：学习率0.01的2个遗忘步骤
    - SVHN：学习率0.001的3个遗忘步骤
    - MNIST：学习率0.01的1个遗忘步骤
  - 门控知识传递(生成器生成1个epoch ，学生训练10个epochs)
    - CIFAR10：KL设置为1，超参数β=250，学生和生成器学习率0.001，滤波阈值0.01
    - SVHN：KL设置为0.5，超参数β=250，学生和生成器学习率0.001 ，滤波阈值0.01
    - MNIST：KL设置为0.5，超参数β=250，学生和生成器学习率0.01 ，滤波阈值0.01

实验设计 Zero-Shot

- 评价指标
  - 遗忘集和保留集的预测准确率D_r，D_f
  - 回忆指数AIN

$$AIN = \frac{r_t(M_u, M_{orig}, \alpha)}{r_t(M_s, M_{orig}, \alpha)}$$

  - 模型反演攻击成功率
  - 成员推理攻击成功率
- 对比方法
  - Retrain：删除遗忘样本重训练
  - Bad Teacher：选取一个不称职教师指导学生模型 2023 AAAI
  - Fisher：利用Fisher信息估计模型数据敏感点 2020 CVPR
  - Amnesiac：存储训练过程中间信息，遗忘时退回先前状态 2021 AAAI

实验设计　对比实验

- 对比实验
  - 在三个数据集上使用ALLCNN模型进行实验
  - 对比方法：删除样本重训练

Dataset	# $\mathcal{Y}_f$	Acc.	Original Model	Retrain Model	M-M Method	GKT Method	*AIN* [GKT]	*AIN* [M-M]
CIFAR10	1	$\mathcal{D}_r \uparrow$	84.05	85.72	20.48	**81.97**	**0.81**	0.11
		$\mathcal{D}_f \downarrow$	87.49	0	5.11	**0**		
	2	$\mathcal{D}_r \uparrow$	84.18	86.30	29.59	**81.70**	**0.74**	0.10
		$\mathcal{D}_f \downarrow$	84.72	0	7.59	**0**		
SVHN	1	$\mathcal{D}_r \uparrow$	94.52	93.02	72.92	**92.43**	**0.37**	0.15
		$\mathcal{D}_f \downarrow$	95.16	0	42.32	**0**		
	2	$\mathcal{D}_r \uparrow$	93.61	95.10	58.70	**92.13**	**0.74**	0.13
		$\mathcal{D}_f \downarrow$	96.39	0	50.23	**0**		
MNIST	1	$\mathcal{D}_r \uparrow$	97.84	99.25	10.57	**97.12**	**0.65**	0.31
		$\mathcal{D}_f \downarrow$	99.61	0	0.0	**0**		
	2	$\mathcal{D}_r \uparrow$	98.17	99.41	10.96	**96.87**	**0.30**	0.20
		$\mathcal{D}_f \downarrow$	98.77	0	0.0	0		

- 实验结论
  - M-M方法在零样本遗忘中表现较差，而GKT方法表现良好

实验设计　对比实验

- 对比实验
  - 在LeNet和ResNet9模型上进行三个数据的对比实验
  - 对比方法：删除样本重训练

Dataset	Model	Acc.	Original Model	Retrain Model	M-M Method	GKT Method	*AIN* [GKT]	*AIN* [M-M]
CIFAR10	LeNet	$\mathcal{D}_r \uparrow$	59.80	62.93	55.32	**41.32**	**211**	24
		$\mathcal{D}_f \downarrow$	65.25	0	23.98	**0**		
	ResNet9	$\mathcal{D}_r \uparrow$	84.83	85.61	10.85	**56.83**	**212**	12
		$\mathcal{D}_f \downarrow$	88.50	0	0	**0**		
SVHN	LeNet	$\mathcal{D}_r \uparrow$	85.69	88.31	81.80	**78.27**	**1**	1
		$\mathcal{D}_f \downarrow$	81.42	0	89.73	**0**		
	ResNet9	$\mathcal{D}_r \uparrow$	82.76	94.24	53.75	**39.44**	**143**	2
		$\mathcal{D}_f \downarrow$	87.11	0	49.65	**0**		
MNIST	LeNet	$\mathcal{D}_r \uparrow$	98.15	98.73	96.96	**95.79**	**1**	1
		$\mathcal{D}_f \downarrow$	99.59	0	99.37	**0**		
	ResNet9	$\mathcal{D}_r \uparrow$	98.57	98.83	12.32	**94.57**	**2**	3
		$\mathcal{D}_f \downarrow$	99.10	0	0	0		

- 实验结论
  - M-M在三个数据集上表现较差，而GKT在多数实验中表现优秀

实验设计 对比实验

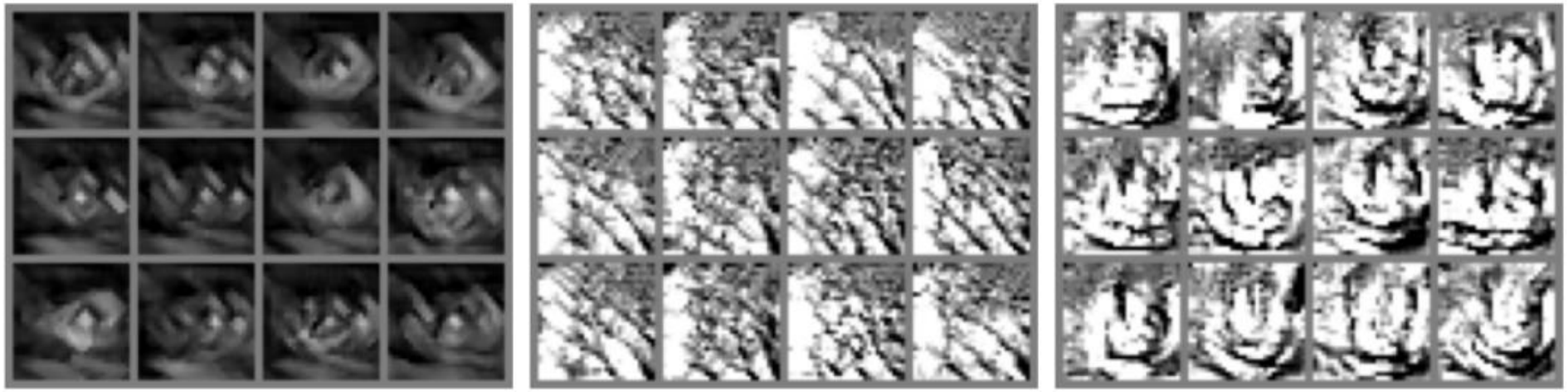
- 对比实验
  - 与现有先进方法进行一类遗忘和两类遗忘进行对比实验

Method	Zero-shot?	1-class Unlearning		2-class Unlearning	
		$\mathcal{D}_r \uparrow$	$\mathcal{D}_f \downarrow$	$\mathcal{D}_r \uparrow$	$\mathcal{D}_f \downarrow$
Original Model	NA	84.05	87.49	84.18	84.72
Retrain Method	NO	85.72	0	86.30	0
Bad Teacher [45]	NO	83.71	5.56	84.11	7.81
Fisher [5]	NO	7.61	0	8.57	0
Amnesiac [14]	NO	83.47	0	82.85	0
Min-Max (ours)	YES	20.48	5.11	29.59	7.59
GKT (ours)	YES	**81.97**	**0**	**81.70**	**0**

- 实验结论
  - 本方法在保留类别预测准确率与现有方法持平，但是不需要原始训练样本

实验设计　防御反演攻击实验

- 防御反演攻击实验
  - 使用MNIST数据集训练ALLCNN模型，使用删除样本重训练和本方法遗忘学习，通过模型反演攻击展示遗忘效果



- 实验结论
  - 模型反演攻击无法从本方法遗忘模型中提取任何信息

算法总结 Zero-Shot

- 算法流程
  - 生成器生成遗忘样本、保留样本
  - 误差最小-噪声最大遗忘
  - 门控知识传递遗忘
- 算法优势
  - 不需要原始训练样本就可以实现特定类别的遗忘
- 算法不足
  - M-M需要优化遗忘类别和保留类的噪声，噪声样本质量决定了遗忘性能
  - GKT在大型模型上效果不理想

特征总结与未来展望

特征总结和未来展望

- SISA
  - 提出一种模型遗忘框架，以空间开销换取训练的时间开销
  - 通过策略性限制数据点在训练过程中的影响来加快模型遗忘训练
  - 对于复杂的深度学习任务，分片过多导致精度大幅下降
- Zero-Shot
  - 提出了两种方法实现零样本机器遗忘，提出了新的评估遗忘质量的指标
  - M-M噪声质量不佳导致遗忘性能差，GKT在大模型上表现差
- 未来展望
  - 将遗忘学习扩展至用户指定的遗忘粒度
  - 多样化数据结构的遗忘学习

参考文献

[1] Bourtoule L, Chandrasekaran V, Choquette-Choo C A, et al. Machine unlearning[C]. 2021 IEEE Symposium on Security and Privacy (SP). IEEE, 2021: 141-159.

[2] Chundawat V S, Tarun A K, Mandal M, et al. Zero-shot machine unlearning[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 2345-2354.

[3] Chundawat V S, Tarun A K, Mandal M, et al. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher[C]. Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(6): 7210-7217.

[4] Graves L, Nagisetty V, Ganesh V. Amnesiac machine learning[C]. Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(13): 11516-11524.

[5] Tarun A K, Chundawat V S, Mandal M, et al. Fast yet effective machine unlearning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023.

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

谢谢！