北京理工大学
信息系统及安全对抗实验中心
信息安全与对抗技术研究所
Beijing Forest Studio

纵向联邦学习标签推理防御方法

www.isclab.org.cn

硕士研究生 朱龙生

2026年11月8日

问题回溯

- 相关内容
  - 李佳龙《面向联邦基础模型的安全评测与防御方法研究》——2025.12.21
  - 李佳龙《面向联邦学习的后门攻击检测与防御》——2026.07.05

内容提要

- 预期收获
- 案例引入
- 内涵解析与研究目标
- 研究背景与意义
- 研究历史与现状
- 知识基础
- 算法原理
  - LCD
  - Casper
- 特点总结与工作展望
- 参考文献

预期收获

- 预期收获
  - 1.认识纵向联邦学习中的标签泄露风险与隐私保护挑战
  - 2.理解典型攻击原理以及通过梯度与中间表示泄露标签的机制
  - 3.掌握标签压缩与因果混淆在纵向联邦学习标签推理防御中的应用

案例引入 小微企业联合信贷风控

- 案例说明
  - 银行持有企业信贷特征与贷款逾期标签
  - 发票服务公司持有发票特征
  - 双方通过联邦学习联合建模



- 潜在风险
  - 发票服务公司在模型训练过程中推测出贷款逾期标签
  - 对小微企业：信用记录暴露给了第三方
  - 对银行：客户信用标签被合作方获取，银行失去了对这部分信息的独占控制

- 问题提出
  - 如何保留联合建模能力的同时，防止合作方推断敏感标签

内涵解析与研究目标

- 内涵解析
  - 纵向联邦学习：多个参与方持有同一批对齐样本的不同特征，在不直接共享原始数据的条件下，协同训练联合模型
  - 标签推理攻击：被动方利用交互梯度、中间表示或训练后的本地模型，推断主动方持有的私有标签
  - 标签推理防御：通过调整标签监督、约束中间表示或改变训练过程，减少标签信息泄露，降低攻击者的标签推断能力
- 研究目标
  - 提高标签隐私保护能力
    - 削弱梯度与中间表示对真实标签的暴露程度
    - 降低模型补全、梯度分析和聚类等标签推理攻击的有效性
  - 维持联合模型的主任务性能
    - 在抑制标签泄露的同时，保留完成主任务所需的有效信息

研究背景与意义

- 研究背景
  - 医疗、金融等场景中的不同机构持有同一批用户的互补特征，纵向联邦学习通过协同建模提高数据利用价值
  - 原始数据保留在本地，仍无法消除训练过程中的隐私泄露风险。被动方可利用交互梯度、中间表示及训练后的本地模型推断私有标签

- 研究意义
  - 保护敏感标签隐私，降低联合训练中的信息泄露风险，为医疗、金融等隐私敏感场景的跨机构协作提供支持
  - 改善隐私保护与模型效用之间的权衡，减少防御对预测性能和训练效率的影响，提高纵向联邦学习的实际可用性

研究历史与现状

Li等人研究两方分割学习中的标签泄露，利用梯度范数与方向差异推断私有标签，并提出 Marvell，通过优化噪声扰动降低泄露风险。该工作揭示了反向传播梯度中的标签信息，建立了标签泄露量化与隐私效用权衡的研究思路

Liu等人利用中间表示和梯度的类别相似性设计标签推理攻击，覆盖训练与推理阶段，并发现部分已有保护方法仍存在泄露风险。该工作说明，攻击者可以利用表示中的类别结构恢复标签，防御需要同时关注中间表示与梯度的可辨识性

Marco Arazzi等人提出KDk，结合教师网络生成的软标签与k匿名思想，对高概率类别进行混淆，再用于联合模型训练。该方法通过调整标签监督信号降低标签推断能力，为兼顾标签隐私与模型效用提供了另一种实现路径

Yuan等人提出LCD，针对模型补全攻击，利用深度特征构造压缩后的伪标签，结合多阶段训练与嵌入紧凑化，削弱被动方表示与真实标签的对应关系，并恢复原任务预测能力，为标签隐私保护提供了标签空间压缩的解决思路

2022　2024　2025　2026

2022

Fu等人提出直接标签推理、被动模型补全与主动模型补全攻击，发现被动方能够结合本地底层模型和少量辅助标签，推断训练集之外样本的标签。研究表明，标签保护还需要关注训练后本地模型所保留的标签信息

2024

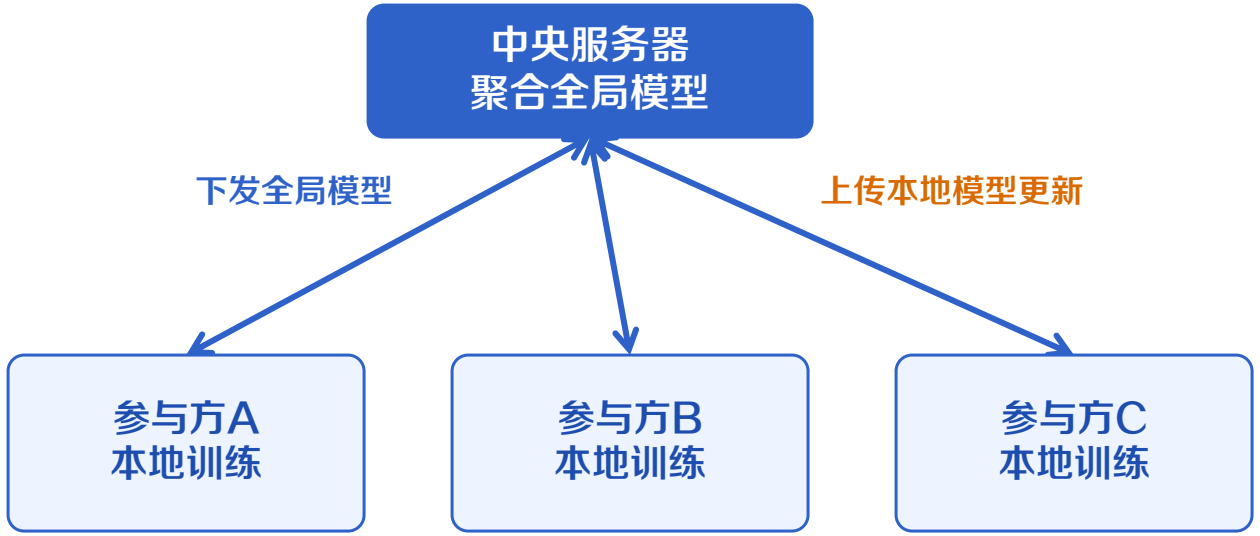
Zou等人研究MID互信息正则化防御，通过限制私有信息与中间输出之间的互信息，抑制标签推理攻击。该方法将信息约束引入纵向联邦学习模型训练，为通过表示学习减少隐私泄露提供了思路，同时需要协调隐私约束与任务性能

2026

Shen等人提出Casper，从因果视角分析切分层输出与标签之间的关系，通过表示重构、标签混淆与选择性差异训练抑制标签泄露。其评估覆盖模型补全、梯度评分和聚类攻击，体现了联合调控表示与监督信号的防御思路，但也引入了额外计算开销

知识基础 联邦学习

- 联邦学习
  - 多个参与方保留本地原始数据，通过交换中间结果，协同训练机器学习模型，分为横向联邦学习和纵向联邦学习
- 横向联邦学习
  - 适用于各参与方样本不同，特征结构相同



原始数据保留在本地，服务器只聚合模型参数或参数更新量



三家医院采用相同的数据字段

年龄　血压　检验指标　疾病标签

用户不同，特征结构相同
因此适合采用横向联邦学习

知识基础 横向与纵向联邦学习的区分依据

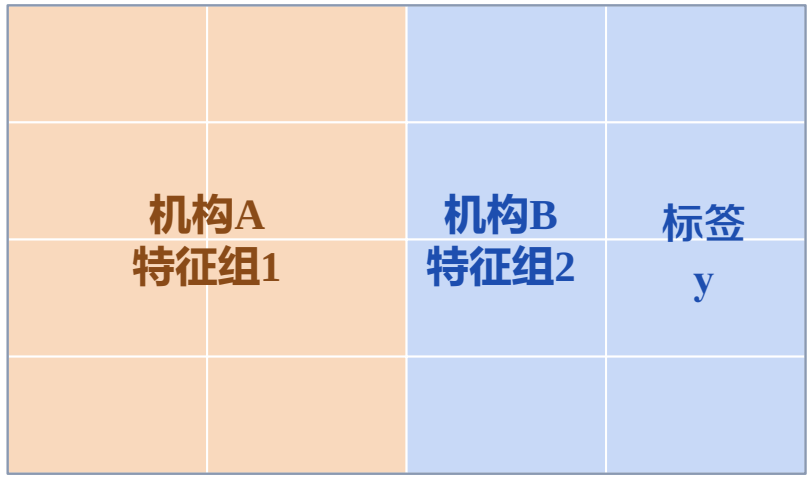
- 区分依据：参与方在样本空间和特征空间的重叠关系

- 横向联邦学习
  - 各方训练相同结构的模型
  - 服务器聚合模型参数



- 例：多家医院拥有不同患者的同类型诊疗数据

- 纵向联邦学习
  - 被动方在本地无法完成一次完整训练
  - 服务器聚合中间表示

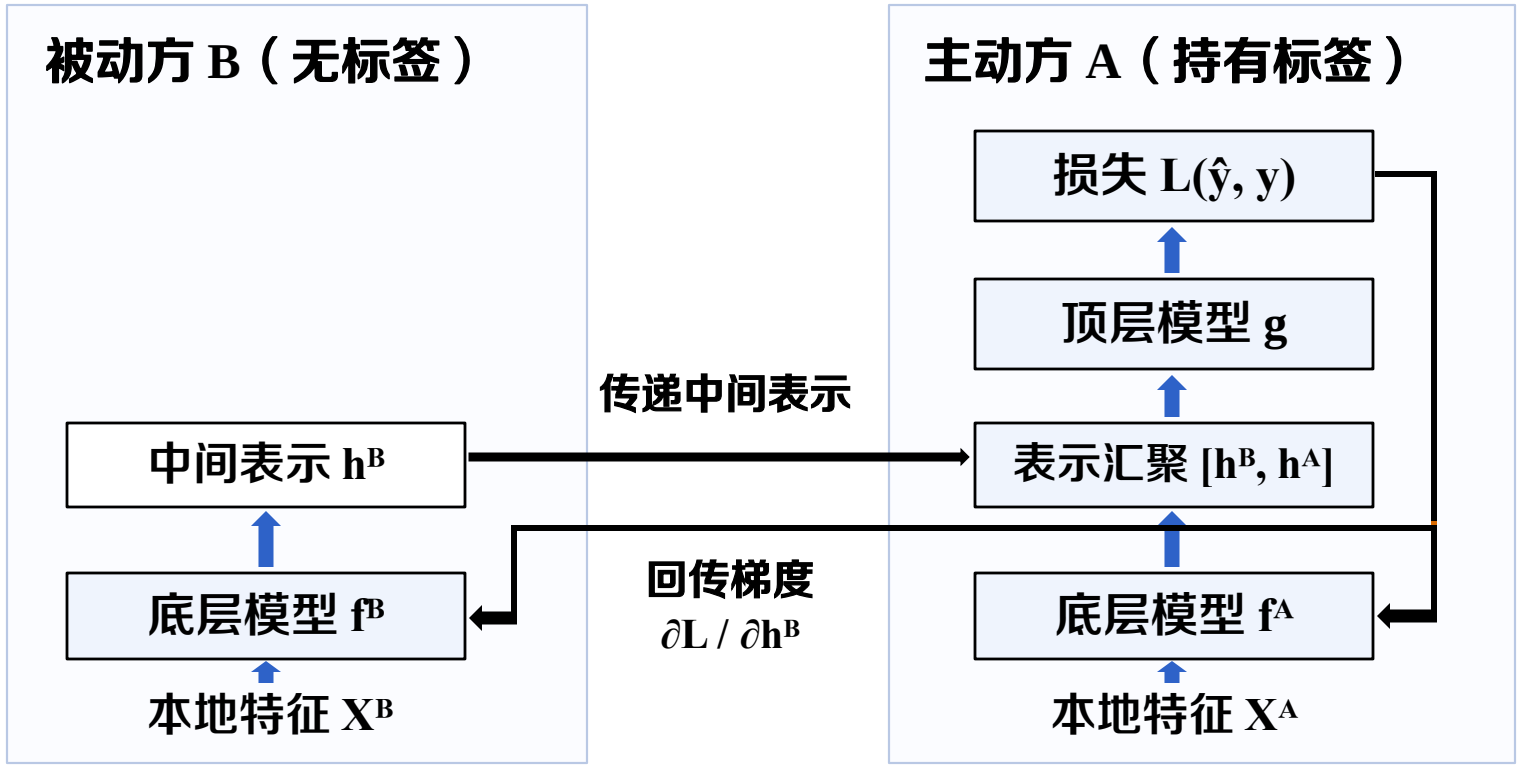


- 例：银行与发票公司持有同一批企业的不同特征

区分依据不是参与方数量或者数据规模，而是
样本与特征的重叠关系，以及完整的训练流程

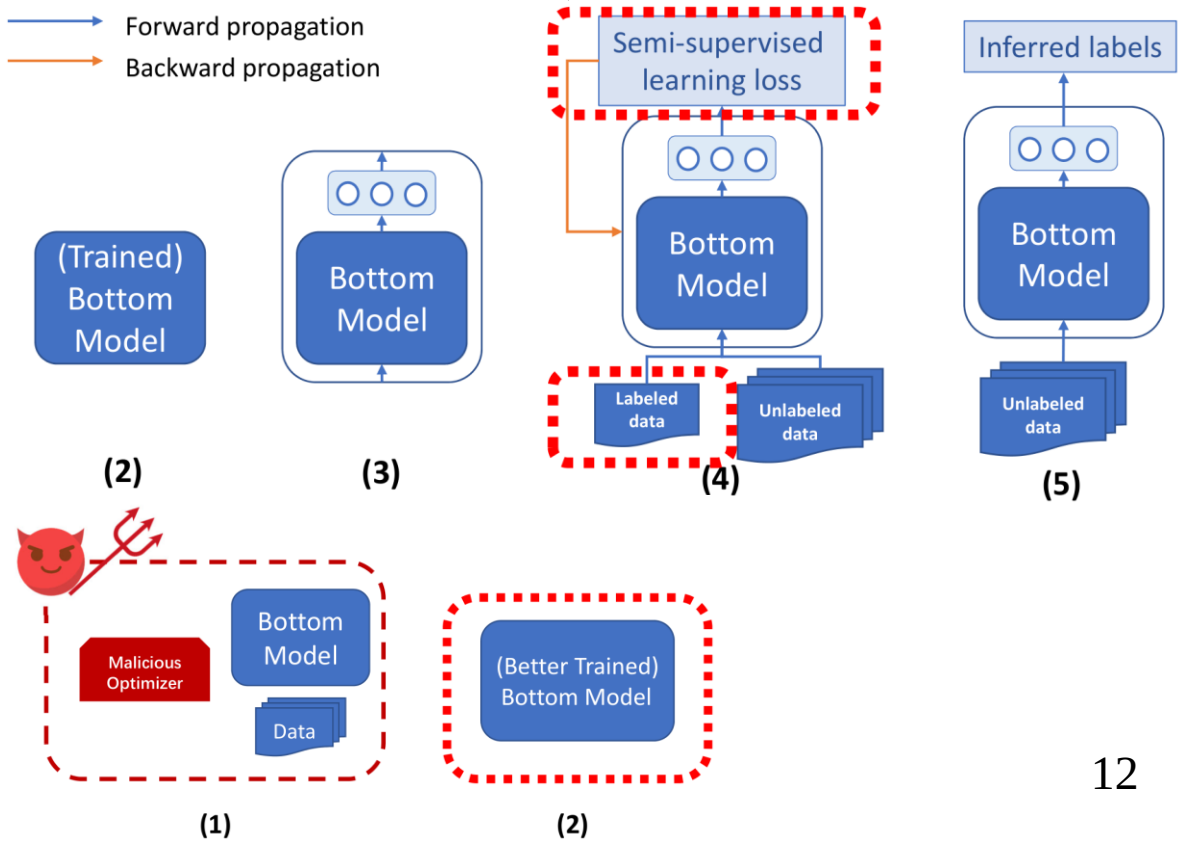
知识基础 纵向联邦学习（VFL）基础架构

- VFL基础架构
  - 数据与模型分布：各方保留本地特征与底层模型，主动方额外持有标签和顶层模型
  - 前向传播：各方底层模型计算中间表示，主动方汇聚中间表示并完成预测
  - 反向传播：主动方结合真实标签计算损失，回传梯度更新各方底层模型



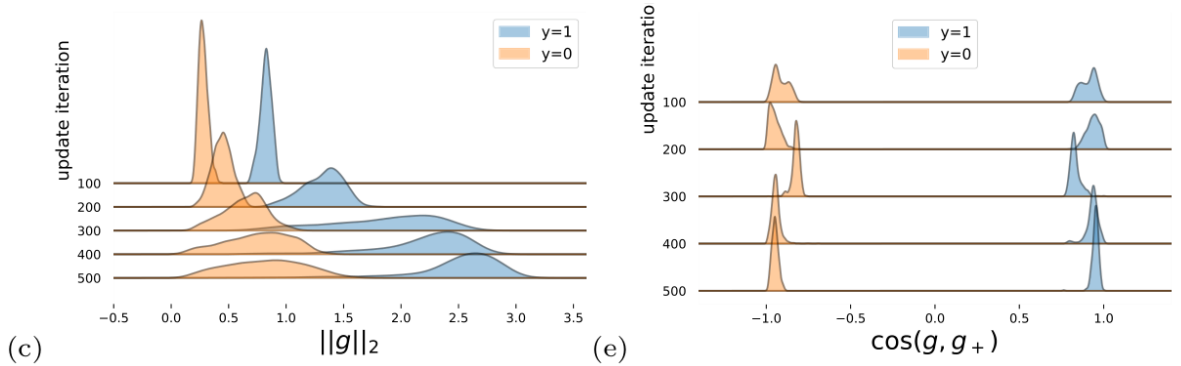
知识基础 典型标签推理攻击：模型补全攻击

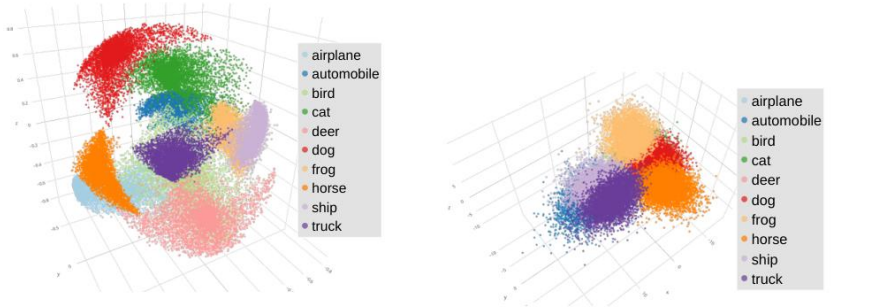
- 攻击者拥有一个小型辅助标注数据集
- 被动模型补全攻击（PMC）
  - 诚实但好奇型攻击者：遵循协议但试图推断标签
  - 正常参与联合训练，获得底层模型，为底层模型添加分类头，结合少量标签和未标注样本进行半监督微调
- 主动模型补全攻击（AMC）
  - 恶意攻击者：主动操纵训练过程
  - 联合训练过程中采用恶意本地优化器，自适应调整、放大本地参数梯度，让其底层模型得到更充分的标签相关信息



知识基础 其他标签推理攻击

- 回传梯度与中间表示可能保留类别差异，被动方据此推断主动方标签
  - NS：梯度范数评分
    - 利用不同类别的范数分布差异区分标签
  - DS：梯度方向评分
    - 利用不同类别的梯度方向差异区分标签
  - Cluster-smashed：中间表示聚类
    - 对中间表示按距离进行聚类
    - 利用少量已知标签确定各簇对应的类别
  - Cluster-grad：梯度聚类
    - 对归一化梯度聚类
    - 利用方向相似性划分样本，将各簇映射到具体类别





(c) Gradients of reciprocal third layer (d) Smashed Data of penultimate layer



Defend Against Label Inference Attacks in Vertical Federated Learning via Label Compression

应用场景

- 构建纵向联邦学习模型
  - 目标：联合训练金融风险预测模型，判断用户的金融风险类别
  - 输入：
    - 银行：用户金融业务特征、用户真实风险标签
    - 电商平台：同一批用户的消费行为特征
  - 输出：用户的金融风险预测结果

- 标签隐私防御
  - 目标：防御电商平台发起的模型补全攻击
  - 输入：隐含标签信息的双方特征
  - 输出：具备标签推理防御能力的金融风险预测模型

问题分析

- 应用问题
  - 现有梯度扰动和标签混淆方法难以消除被动方模型学到的类别信息，防御效果有限
- 科学问题
  - 如何降低被动方上传表示对真实类别的可分性，同时保留联合模型完成原始分类任务的能力
- 难点分析
  - 隐私效用兼顾困难：保持隐私防护的同时，保留主任务性能
  - 标签信息抑制困难：被动方模型在联合训练中学习类别信息，仅保护传输梯度难以阻止抑制隐私泄露
- 核心思想（技术路线）
  - 利用主动方深层特征对真实类别进行配对压缩，生成类别数量更少的伪标签，在欺骗训练中引入中心损失，压缩同一伪类别内的表示分布

PCDL-TIPO

T	目标	利用标签压缩与嵌入紧凑化抵御标签推理攻击，同时保持原始分类任务的预测性能
I	输入	数据集：2个表格数据集，包含蛋白质分类预测与大写字母分类，共21.5K条，4个图像数据集，包含手写数字识别，自然图像分类，共41.5W条
P	处理	1.使用真实标签训练主动方模型，提取类别特征原型 2.根据类别特征原型的相似性，配对压缩相异类别 3.使用伪标签联合训练，引入中心损失，压紧被动方嵌入 4.固定被动方模型，以真实标签重训主动方模型
O	输出	训练后的防御型VFL模型，以及原始类别空间中的分类预测结果

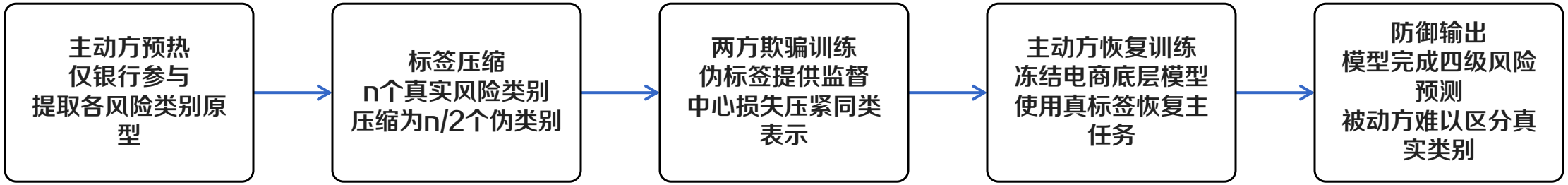
P	问题	现有梯度扰动和标签混淆方法难以消除被动方模型学到的类别信息，防御效果有限
C	条件	主动方持有真实标签，并能获取和缓存被动方的中间表示
D	难点	标签信息抑制、隐私与效用兼顾
L	水平	IEEE TMC(CCF-A 2026)

实例分析 跨机构金融风险预测

- 模型补全攻击
  - 为训练好的被动方底层模型添加分类头，借助少量辅助标注数据集，实现半监督微调

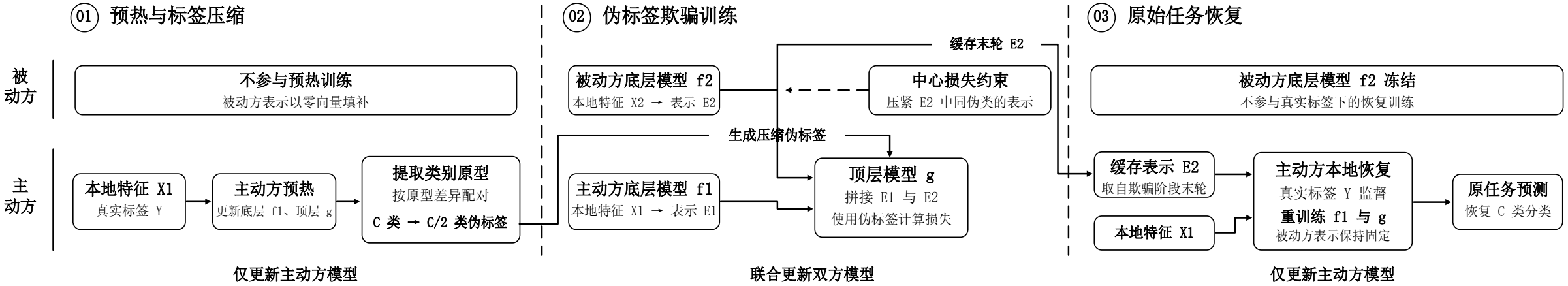


- 标签隐私防御
  - LCD：伪标签欺骗训练与原始任务恢复



算法原理 LCD

- 核心思想
  - 让被动方学习压缩伪标签，再由主动方恢复原始分类任务

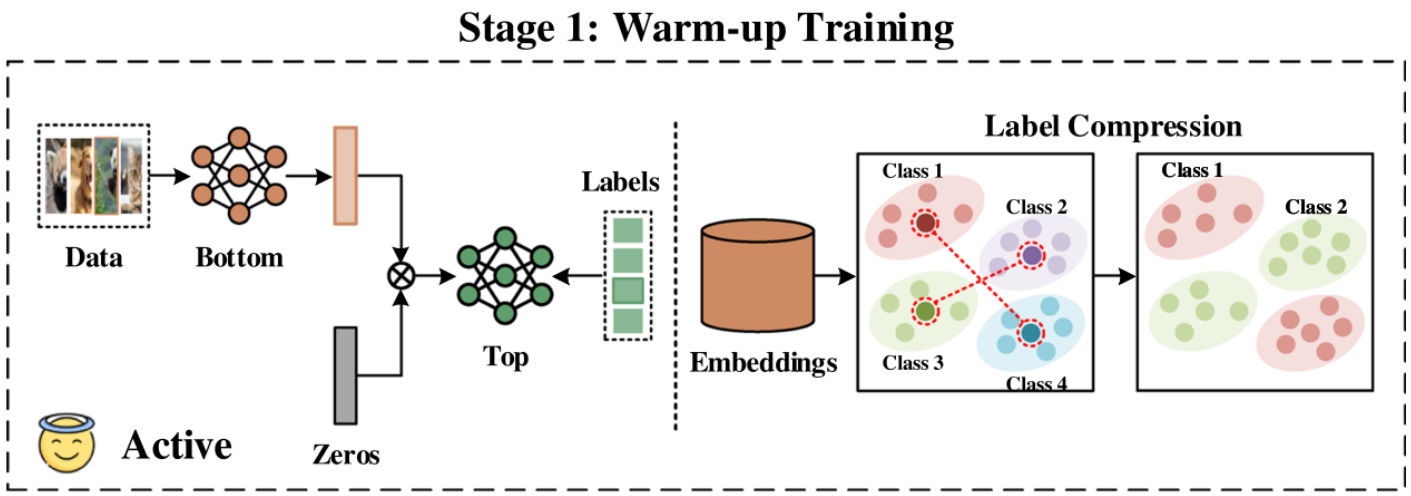


- 预热与标签压缩
  - 真实标签预热主动方模型
  - 提取类别原型并压缩标签
  - 仅更新主动方模型
- 伪标签欺骗训练
  - 伪标签驱动双方联合训练
  - 压紧被动方表示
  - 缓存末轮被动方表示
- 原始任务恢复
  - 固定被动方模型
  - 重训主动方模型

预热和恢复阶段仅更新主动方模型，联合训练阶段使用压缩伪标签

算法原理 主动方预热与标签压缩

- 核心目标
  - 构造压缩伪标签，减少后续联合训练的真实类别监督



- 选择差异更大的类别进行配对
  - 目的：有助于恢复阶段主动方模型能重新区分

按主动方特征配对差异较大的类别

- 真实标签预热

$$L = CE(g([f_1(\mathrm{x}_1), 0]), y)$$

  - 更新主动方底层和顶层模型f_1，g
  - 输出有类别判别能力的中间表示
- 类别原型配对
  - 对类别所有表示求平均并归一化得到类别原型

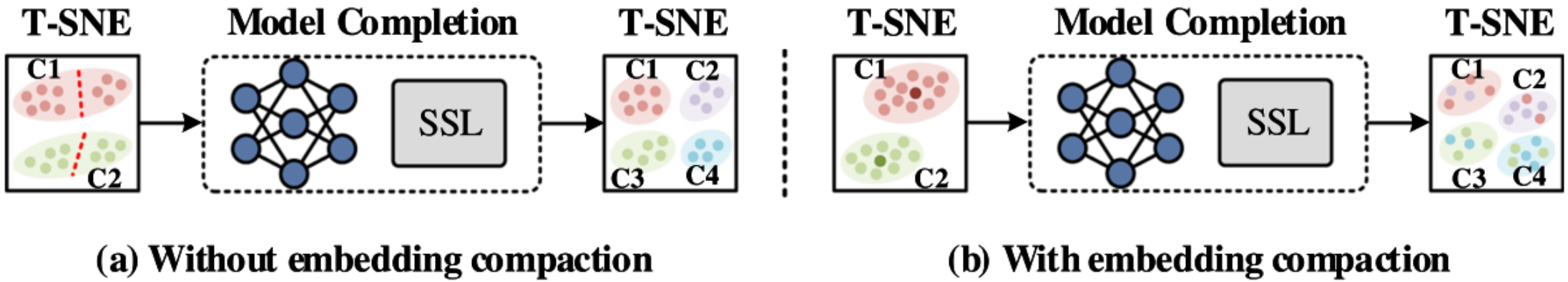
$$p_c = \bar{e}_c / \|\bar{e}_c\|_2$$

  - 计算类别相似度，并最小化所有配对的相似度之和

$$S_{ab} = p_a^T p_b$$

算法原理 欺骗训练

- 核心目标
  - 让被动方学习伪类别，降低同一伪类别真实类别的可分性



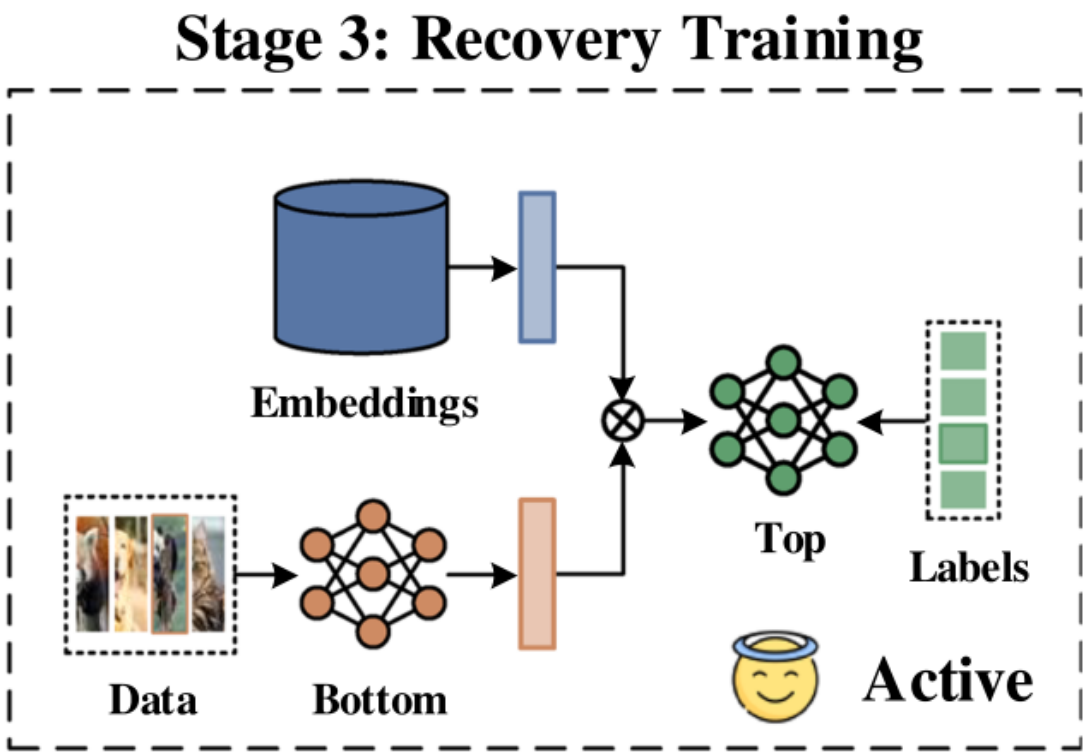
(a) Without embedding compaction　　(b) With embedding compaction

- 仅使用伪标签，真实类别可能仍会被区分
  - 联合训练目标：$L = CE(\hat{y}, y) + \lambda L_center \quad L_center = \frac{1}{2}\sum_i \left\| e_i - c_y \right\|_2^2$
  - 加入中心损失，推动被动方表示靠近伪类别中心，削弱模型补全可利用的类内差异

该阶段输出学习伪类别的被动方模型，缓存被动方末轮中间表示

算法原理 原始任务恢复

- 核心目标
  - 在主动方恢复原始分类能力，停止用真实标签梯度更新被动方



- 固定缓存输入
  - 重训过程被动方中间表示恒定
- 真实标签优化主动方

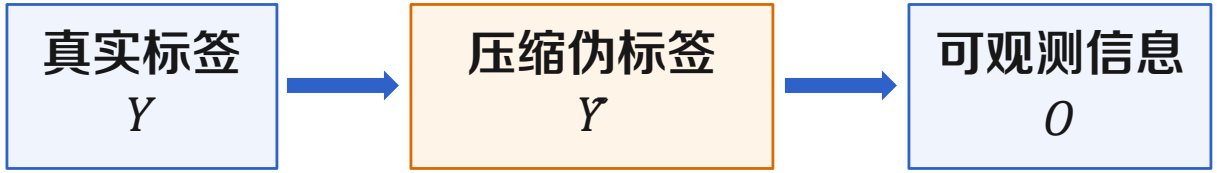
$$L = CE(g([f_1(\mathrm{x}_1), E]), y)$$

  - 更新主动方底层和顶层模型f_1，g
  - 输出能预测原始标签的VFL模型

真实标签重新约束主动方的分类边界，
本地特征帮助区分被合并的类别

算法原理 理论依据与适用边界

- 标签信息的传递假设



- 数据处理不等式

$$I(O;Y) \le I(Y';Y) = H(Y') - H(Y'|Y) = H(Y')$$

- 压缩标签降低了可观测标签信息的上界
  - 示例：四个等概率分类标签压缩为二分类

$$H(Y) = 2\text{bit} \qquad H(Y') = 1\text{bit}$$

- 适用边界
  - 假设边界：若被动方特征足以预测标签，则上述标签信息的传递假设可能不成立
  - 性能边界：压缩标签会损失部分判别信息，恢复效果受主动方特征质量影响
  - 保证范围：信息量上界降低不直接意味着降低攻击准确率

该理论提供条件性信息上界，实际防御效果与模型效用仍需实验验证

实验环境和条件 数据集

- 六个数据集，覆盖表格与图像两类输入

Datasets	Type	Class	Training	Testing	Image/Feature Size	Bottom Model	Embedding Size	Top Model
Yeast	Tabular	10	1.1K	0.4K	8×1	MLP-2	10	MLP-4
Letter	Tabular	26	15K	5K	16×1	MLP-2	10	MLP-4
MNIST	Image	10	60K	10K	$28 \times 28 \times 1$	MLP-2	10	MLP-4
CIFAR-10	Image	10	50K	10K	$32 \times 32 \times 3$	ResNet-18	128	MLP-4
CIFAR-100	Image	100	180K	90K	$32 \times 32 \times 3$	ResNet-18	128	MLP-4
ImageNet-12	Image	12	12K	3K	$224 \times 224 \times 3$	ResNet-34	128	MLP-4

- 表格数据：Yeast、Letter
  - 蛋白质定位预测、大写字母分类
- 图像数据：MNIST、CIFAR-10、CIFAR-100、ImageNet-12
  - 手写数字识别、自然图像分类

评价指标

- 对比方法
  - NG（梯度加噪）、GC（梯度压缩）、DSGD（梯度离散化）
  - CAE（混淆自编码器）、MID（互信息正则化）

- 评估指标
  - 主任务准确率（Main Task Accuracy， MTA）
    - 衡量联合模型效用，越高越好
  - 标签推理准确率（Label Inference Accuracy, LIA）
    - 衡量标签泄露程度，越接近随机猜测越好
  - 综合防御得分（DER）
    - 兼顾攻击准确率与主任务性能，越高越好

$$\mathrm{DER} = [max(0, \Delta ACC_atk) - max(0, \Delta ACC_task) + 1]/2$$

实验结果 对比实验

- **PMC被动模型补全攻击**

Datasets	Yeast			Letter			MNIST			CIFAR-10			CIFAR-100			ImageNet-12		
Defense	LIA↓	MTA↑	DER↑	LIA↓	MTA↑	DER↑	LIA↓	MTA↑	DER↑	LIA↓	MTA↑	DER↑	LIA↓	MTA↑	DER↑	LIA↓	MTA↑	DER↑
None	0.388	0.542	0.500	0.874	0.950	0.500	0.887	0.990	0.500	0.805	0.813	0.500	0.614	0.736	0.500	0.910	0.894	0.500
NG	0.255	0.466	0.529	0.799	0.926	0.526	0.837	**0.986**	0.523	0.131	0.380	0.621	0.556	0.733	0.528	0.417	0.842	0.720
GC	0.178	0.415	0.542	0.832	0.935	0.513	0.892	0.986	0.498	0.561	**0.759**	0.595	0.609	**0.737**	0.502	0.618	0.842	0.620
DSGD	0.294	**0.482**	0.517	0.832	**0.936**	0.514	0.897	0.986	0.498	0.166	0.376	0.601	0.400	0.640	0.559	0.843	**0.854**	0.513
MID	0.275	0.453	0.512	0.733	0.902	0.546	0.788	0.979	0.544	0.215	0.690	0.734	0.402	0.383	0.429	0.547	0.839	0.653
CAE	0.163	0.434	0.559	0.821	0.910	0.507	0.804	0.985	0.539	0.837	0.738	0.462	0.454	0.462	0.443	0.506	0.849	0.679
Ours	**0.145**	0.423	**0.562**	**0.662**	0.863	**0.562**	**0.512**	0.968	**0.676**	**0.109**	0.696	**0.790**	**0.160**	0.614	**0.666**	**0.083**	0.811	**0.872**

- **AMC主动模型补全攻击**

Datasets	Yeast			Letter			MNIST			CIFAR-10			CIFAR-100			ImageNet-12		
Defense	LIA↓	MTA↑	DER↑	LIA↓	MTA↑	DER↑	LIA↓	MTA↑	DER↑	LIA↓	MTA↑	DER↑	LIA↓	MTA↑	DER↑	LIA↓	MTA↑	DER↑
None	0.405	0.525	0.500	0.886	0.949	0.500	0.898	0.988	0.500	0.834	0.809	0.500	0.643	0.732	0.500	0.934	0.891	0.500
NG	0.294	**0.491**	0.538	0.827	0.908	0.509	0.834	0.986	0.531	0.107	0.355	0.636	0.361	0.715	0.633	0.435	0.847	0.728
GC	0.278	0.461	0.531	0.837	**0.938**	0.519	0.873	0.986	0.512	0.493	**0.759**	0.645	0.580	**0.735**	0.532	0.624	0.854	0.637
DSGD	0.260	0.486	0.553	0.836	0.937	0.519	0.861	0.986	0.518	0.273	0.743	0.747	0.276	0.541	0.588	0.845	0.855	0.526
MID	0.173	0.294	0.500	0.793	0.876	0.510	0.830	**0.987**	0.534	0.161	0.694	0.779	0.197	0.343	0.528	0.557	0.847	0.666
CAE	0.257	0.442	0.532	0.816	0.883	0.502	0.816	0.983	0.538	0.785	0.711	0.475	0.530	0.485	0.433	0.543	**0.868**	0.684
Ours	**0.134**	0.415	**0.581**	**0.647**	0.852	**0.571**	**0.463**	0.972	**0.710**	**0.103**	0.703	**0.812**	**0.152**	0.626	**0.693**	**0.083**	0.814	**0.876**

LIA：攻击准确率　MTA：主任务准确率　DER：综合表现

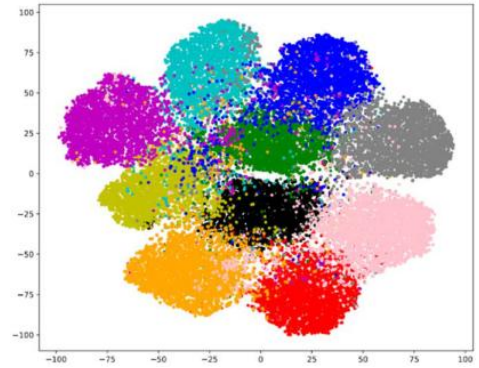
LCD在6个数据集上的DER均最高，防御效果与效用损失仍具有数据集差异

- 标签保护不充分
  - 被动方表示中可能仍保留可供半监督攻击利用的真实类别线索
- 对AMC保护优于PMC
  - 主动攻击中的恶意底部模型更依赖训练时的伪标签，因此也更容易受到伪标签误导
- NG低LIA伴随低MTA
  - 隐私与效用的权衡较差，噪声梯度影响主任务性能

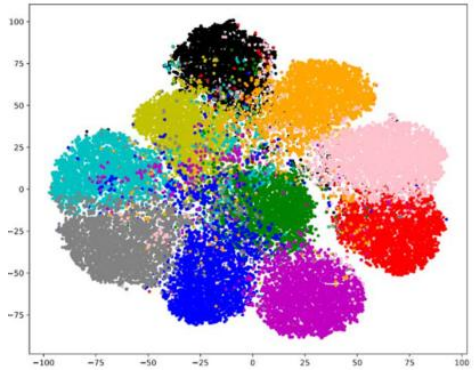
实验结果 对比实验

- 被动方表示的可分性分析（以MNIST为例）

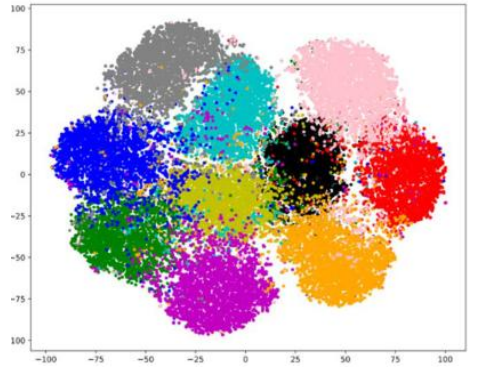
NG



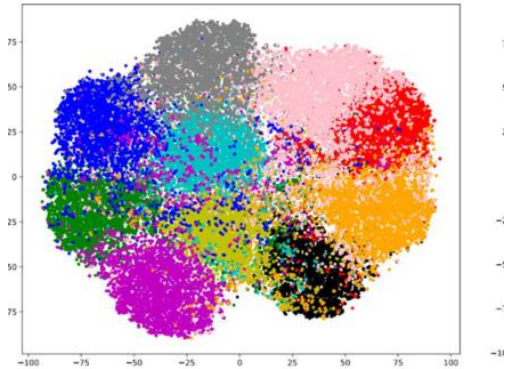
GC



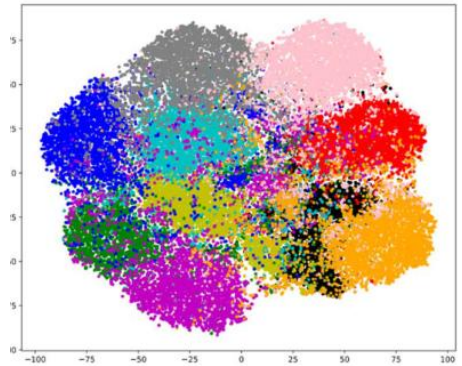
DSGD



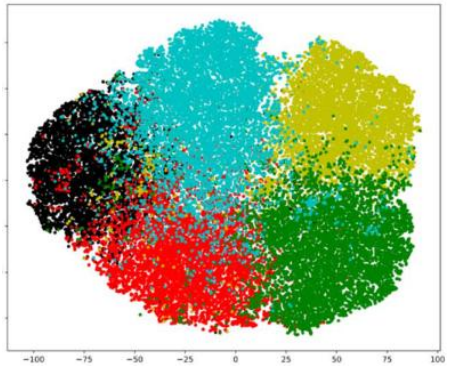
MID



CAE



LCD

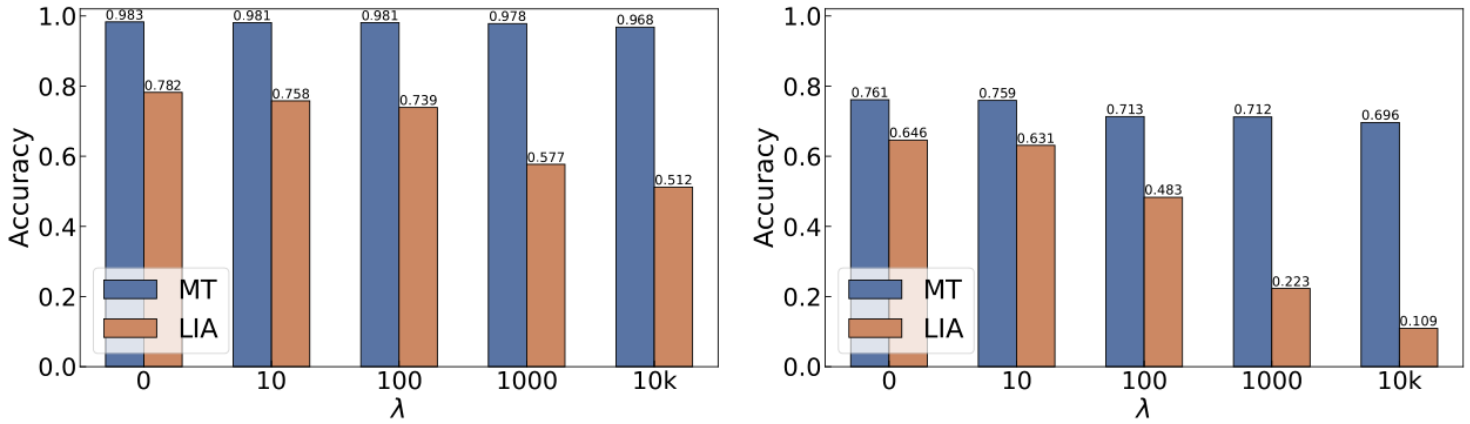


- 部分基线仍形成较清晰的真实类别簇
  - NG、GC形成明显的独立区域
- LCD下类别重叠增强，类别边界更难区分

LCD子图按5个压缩伪类别着色，展示粗粒度类别的表示分布
仅凭该图，无法直接判断同一伪类别内部两个真实类别的可分性

实验结果 中心损失权重参数实验

- 控制其他设置，改变λ，考察表示紧凑化对标签保护和主任务性能的影响



(a) MNIST　　(b) CIFAR-10

- λ=0：去掉中心损失
  - 伪类别内部仍能形成子簇
  - 攻击者利用辅助标签可重新拆分
- λ过大会影响主任务性能
- 比较多个候选值，选择DER最高的参数

- **MNIST：LIA下降27%，MTA下降1.5%**
- **CIFAR-10：LIA下降53.7%，MTA下降6.5%**

增大中心损失权重λ可进一步降低标签推理准确率，但也会带来主任务性能损失，需综合权衡隐私保护与模型效用

特色创新

- 提出基于标签压缩的多阶段防御框架
  - 现有防御仍使真实标签参与联合训练，被动方模型能够学习标签相关信息
  - 将训练划分为预热、欺骗和恢复三个阶段，联合训练使用伪标签，再由主动方本地恢复原始分类任务

- 提出特征差异大的标签压缩方法
  - 利用主动方提取的类别原型，将特征差异较大的真实类别配对，合并为同一伪类别，便于主任务恢复阶段可分

- 引入中心损失表示压紧机制
  - 仅合并标签后，不同真实类别仍可能在同一伪类别内形成可分离的子簇
  - 通过中心损失压紧同一伪类别的被动方表示，削弱其内部真实类别的可分性

算法总结

- 算法优势
  - 通过压缩伪标签替代真实标签开展联合训练，削弱被动方表示与真实标签的关联
  - 引入中心损失压紧同一伪类别的表示，增加攻击者重新分离真实类别的难度
  - 由主动方本地恢复原始分类任务，预热与恢复阶段不增加额外通信轮次

- 算法不足（论文分析）
  - 标签压缩会损失被动方对合并类别的判别信息，恢复训练仍难以完全避免原始任务准确率下降
  - 理论保证依赖特定的信息关联假设

- 算法分析（个人思考）
  - 根据伪类别内部的可分性调整表示压紧强度，减少统一约束造成的有效信息损失



Casper: A Causality-Inspired Defense With Confounder Against Label Inference Attacks in Vertical Split Federated Learning

应用场景

- 构建医疗协同预测模型
  - 目标：结合患者临床信息与日常活动信息，联合训练医疗预测模型
  - 输入：
    - 主动方：患者临床特征、真实医疗标签
    - 被动方：同一批患者的日常活动特征
  - 输出：患者的医疗预测结果
- 标签隐私防御
  - 目标：防御被动方的标签推理攻击，在尽量保持医疗预测性能的同时，保护患者的真实医疗标签
  - 输入：隐含标签信息的双方特征
  - 输出：具备标签隐私保护能力的医疗预测模型，降低被动方推断患者真实医疗标签的准确率

问题分析

- 应用问题
  - 现有标签推理防御方法容易损害主任务性能，且难以适应不同攻击场景与数据模态，防御效果受到限制
- 科学问题
  - 如何通过因果干预削弱中间表示与真实标签的直接依赖，提升防御有效性与任务泛化能力
- 难点分析
  - 标签依赖消除困难：中间表示持续积累标签信息，仅改变表示难以充分抑制泄露
  - 隐私与效用平衡困难：保持隐私防护的同时，保留主任务性能
- 核心思想（技术路线）
  - 借助VAE重构中间表示与标签，削弱攻击者可利用的标签依赖
  - 利用选择性差异训练分配不同监督信号

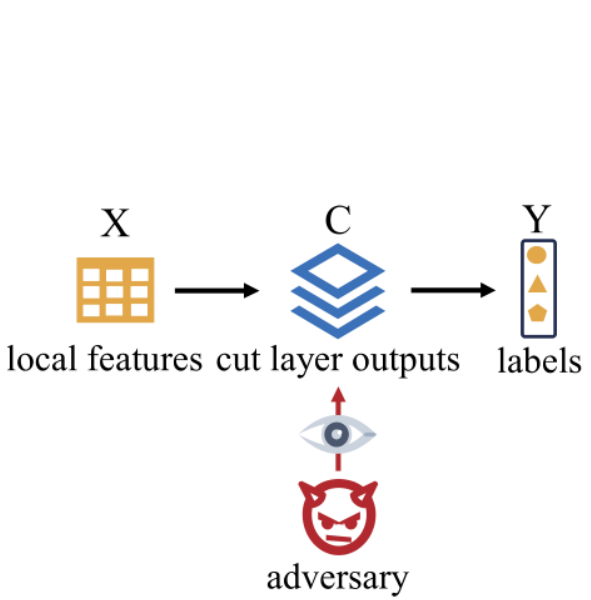
PCDL-TIPO

T	目标	降低纵向联邦学习中的标签泄露风险，抵御多种标签推理攻击，同时保持联合模型的预测性能
I	输入	数据集：2个自然图像分类数据集，共17W条数据；1个表格数据集，判断肿瘤是否为良性，共569条数据；1个文本主题分类数据集，共7W条数据
P	处理	1. 重构被动方中间表示 2. 重构并平滑标签，生成混淆监督信号 3. 选择性差异训练
O	输出	经过隐私防御训练的联合分类模型及样本预测结果

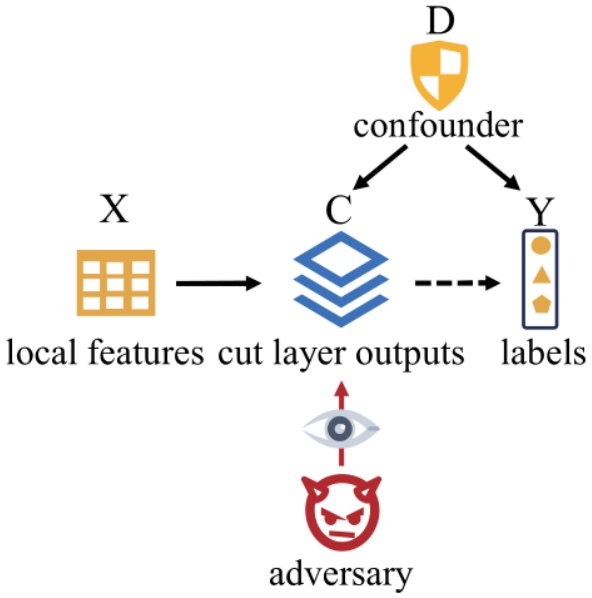
P	问题	联合训练使被动方的中间表示和梯度逐渐包含标签信息，现有防御在不同攻击、数据模态下，难以同时保持有效保护与较高预测性能
C	条件	主动方持有真实标签，能够获取被动方切分层输出、部署重构与混淆模块，并分别控制主动方和被动方的训练信号
D	难点	削弱标签信息的同时保留有效预测信息，避免混淆监督损害主任务性能
L	水平	IEEE TIFS(CCF-A 2026)

算法原理 因果分析

- 因果分析目标
  - 解释标签的泄露路径
  - 引入混杂后依赖关系的变化
- 常规路径：X → C → Y
  - 联合训练使C积累标签信息
  - 攻击者学习到C与Y之间的映射
- 混杂结构：C ← D → Y
  - 混杂信息共同影响C与Y
  - 削弱了C与Y之间的直接依赖关系



(a) Causal path w/o confounder.

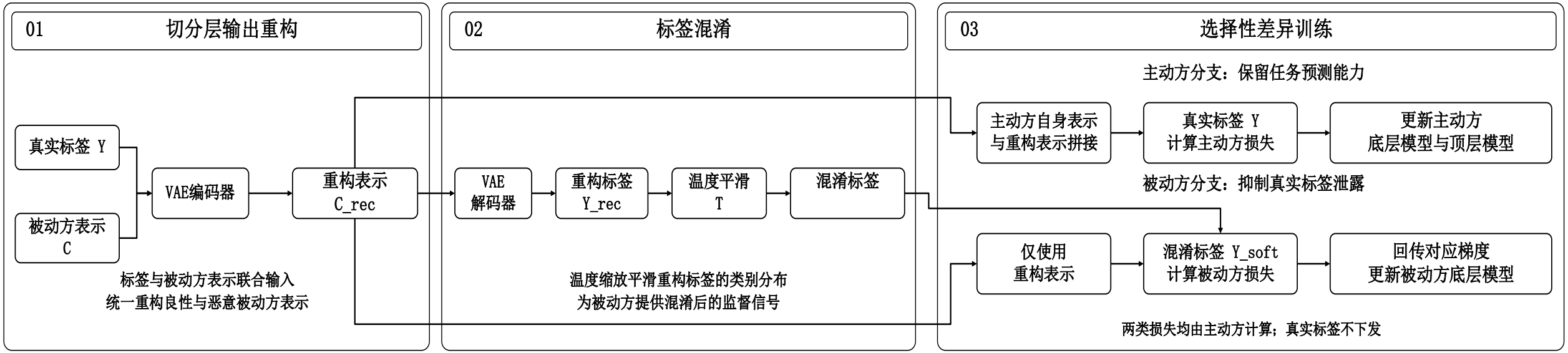


(b) Causal path w/ confounder.

D是作者人为干预的抽象，目的是削弱攻击者可利用的中间表示与真实标签之间的依赖，实际效果需实验验证

算法原理 Casper

- 算法总体框架：重构被动方表示与标签，按参与方分配不同监督信号



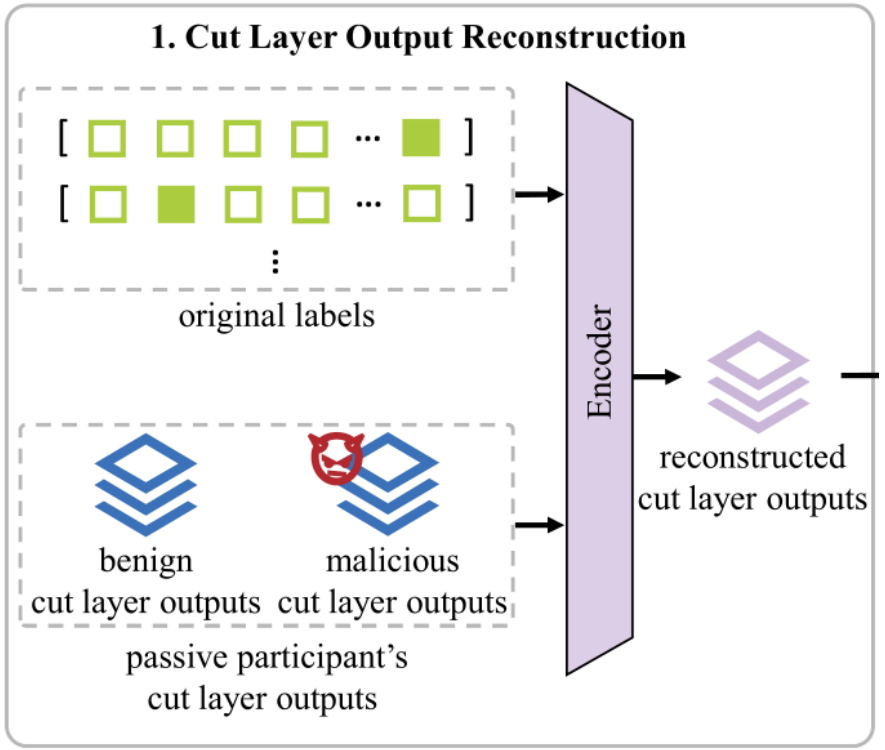
- 切分层输出重构
  - 标签与被动方表示共同输入
  - 编码器生成重构表示
- 标签混淆
  - 解码器生成重构标签
  - 温度平滑得到混淆标签
- 选择性差异训练
  - 真实标签更新主动方模型
  - 混淆标签更新被动方模型

算法原理 切分层输出重构

- 核心目标
  - 保留预测所需信息的同时，抑制重构表示中的真实标签信息
- 重构质量与标签隐私的联合约束

$$L_{VAE} = MSE(C, C_{rec}) + \lambda I(C_{rec}; Y)$$

  - 使重构表示接近原始表示，保留有效信息
  - 抑制表示与真实标签之间的信息依赖
  - 训练策略：$\lambda = t/T_rec$，随重构轮次递增
- 重构误差要求保留原表示的信息，互信息约束要求减少其中的标签信息



Casper希望找到一个折中，而不是完整保留或完全消除标签信息

算法原理 标签重构与温度平滑

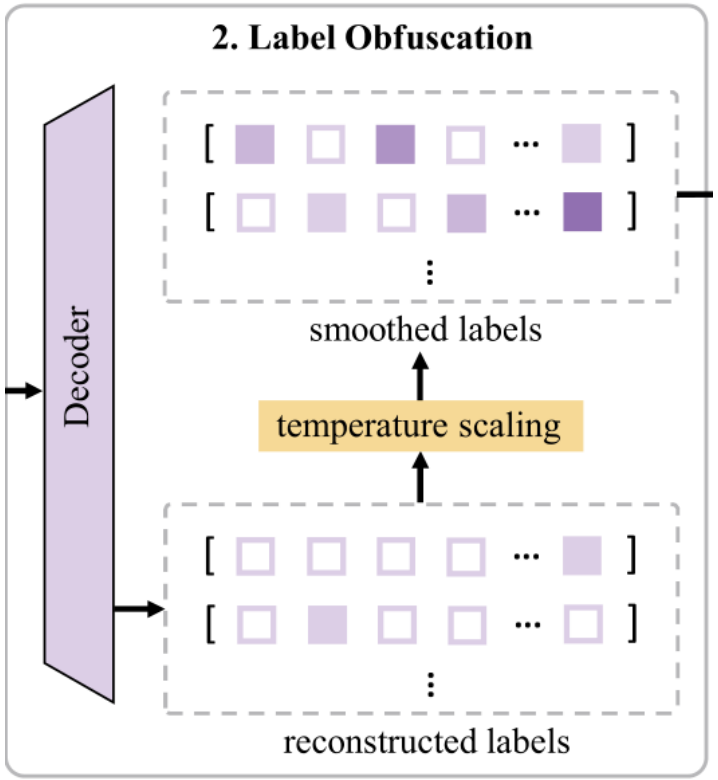
- 核心目标
  - 生成混淆监督信号，进一步削弱被动方学习真实标签的能力
- 标签重构：保留任务关联

$$Y_rec = Decoder(C_rec)$$

$$L_{label} = MSE(Y, Y_rec)$$

- 温度平滑：降低监督确定性
  - T越大，类别概率越平缓

$$Y_{soft} = \frac{exp(Y_rec/T)}{\sum_i exp(Y_rec^{(i)}/T)}$$



标签重构保留任务关联，温度平滑降低监督确定性，进一步干扰标签推断

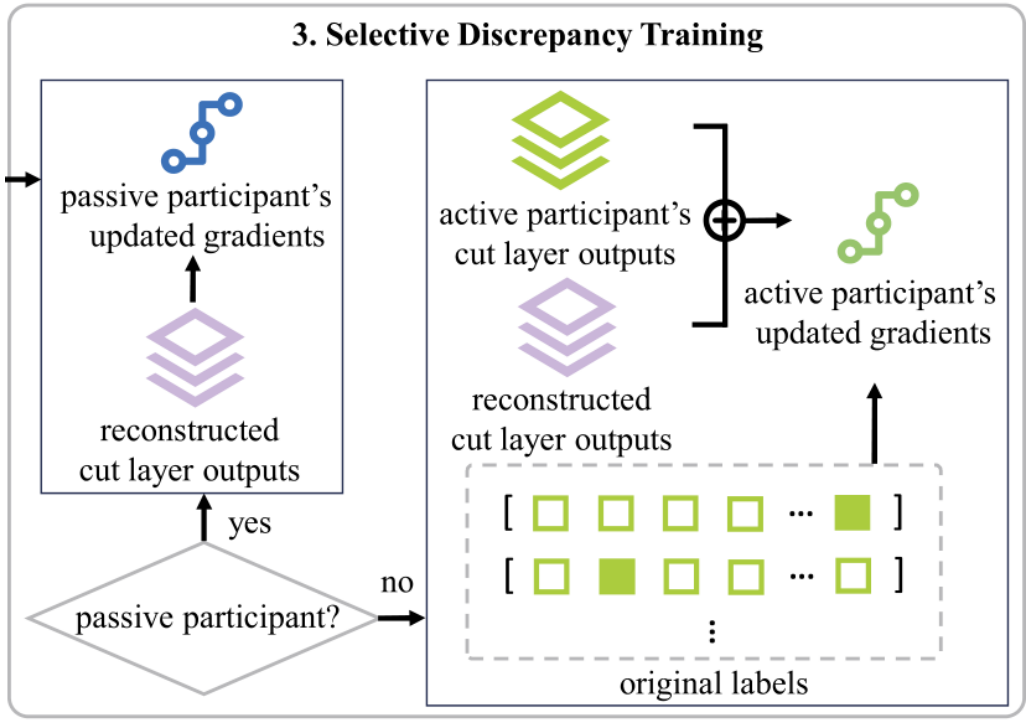
算法原理 选择性差异训练

- 核心目标
  - 主动方保留真实标签监督
  - 被动方接收混淆目标对应的更新梯度
- 真实标签目标：更新主动方
  - 各方表示联合预测h，更新主动方本地底层模型和顶层模型

$$L_active = l(h, Y)$$

- 混淆标签目标：更新被动方
  - 被动方重构表示预测，更新被动方底模

$$L_passive = l(h', Y_soft)$$



更新边界：真实标签保留在主动方，被动方仅接受混淆目标对应梯度

实验环境和条件 数据集

- 四个数据集，覆盖图像、表格、文本三类输入

Datasets	Training Dataset	Test Dataset	Number of Classes	Known Label Quantity Per Class	Metric	Bottom Model Architecture	Top Model Architecture
CIFAR10	50,000	10,000	10	4	Top-1 Acc	ResNet-20	FCNN-4
TinyImageNet	100,000	10,000	200	3	Top-5 Acc	ResNet-56	FCNN-3
Breast Cancer	426	143	2	10	F1 Score	FCNN-3	FCNN-2
Yahoo Answers	50,000	20,000	10	3	Top-1 Acc	BERT	FCNN-4

- 图像数据：CIFAR-10、TinyImageNet
  - 输入为图像像素，覆盖不同规模的自然图像分类任务
- 表格数据：Breast Cancer
  - 更具肿瘤特征判断良性或恶性
- 文本数据：Yahoo Answers
  - 依据问答文本预测主题，使用BERT作为底层模型

评价指标

- 对比方法
  - DP（梯度加噪）、GS（梯度压缩）、DG（梯度离散化）
  - PPDL（结合噪声、压缩与随机选择保护梯度）
  - MID（互信息正则化）、InfoScissors（对抗训练）
- 评估指标
  - 主任务性能，越高越好
    - Top-1准确率：CIFAR-10、Yahoo Answers
      - 预测得分最高的类别与真实标签一致的比例
    - Top-5准确率：TinyImageNet
      - 真实标签位于预测得分最高的前五个类别中的比例
    - F1分数：Breast Cancer
      - 衡量肿瘤二分类性能
  - 攻击性能：攻击成功率，越接近随机猜测越好

实验结果 对比实验（PMC、AMC攻击场景）

- Casper在PMC、AMC降低了攻击效果，同时避免主任务性能降到随机水平

Datasets	Defense	Passive Model Completion			Active Model Completion		
		Model ↑	Attack ↓		Model ↑	Attack ↓	
		Test Dataset	Training Dataset	Test Dataset	Test Dataset	Training Dataset	Test Dataset
CIFAR10	w/o defense	0.8007	0.8012	0.6267	**0.8122**	0.9488	0.7199
	DP [9]	0.1000	0.1000	0.1000	0.7643	0.9931	0.7480
	GS [10]	**0.8097**	0.8034	0.6179	0.8053	0.9561	0.7280
	DiscreteSGD [2]	0.1000	0.1000	0.1000	0.7645	0.9955	0.7590
	PPDL [11]	0.1000	0.1000	0.1000	0.1000	0.1000	0.1000
	MID [14]	0.7325	0.1000	0.1000	0.7246	0.1433	0.1441
	InfoScissors [15]	0.7483	0.1437	0.1445	0.7460	0.1430	0.1444
	Casper	0.7410	**0.0987**	**0.0982**	0.7466	**0.1000**	**0.1000**
TinyImageNet	w/o defense	**0.6580**	0.3548	0.2515	0.6071	0.3944	0.2546
	DP [9]	0.0250	0.0250	0.0250	0.0250	0.1028	0.0971
	GS [10]	0.6469	0.3770	0.2650	**0.6216**	0.3864	0.2383
	DiscreteSGD [2]	0.0683	0.0461	0.0485	0.5373	0.5533	0.3352
	PPDL [11]	0.0514	0.0330	0.0361	0.0643	0.0529	0.0532
	MID [14]	0.5900	0.0277	0.0278	0.5855	0.0317	0.0318
	InfoScissors [15]	0.5427	0.0629	0.0551	0.5456	0.0624	0.0551
	Casper	0.5920	**0.0250**	**0.0250**	0.5871	**0.0287**	**0.0312**
Breast Cancer	w/o defense	0.8901	0.8519	0.7879	0.8681	0.8016	0.7955
	DP [9]	0.8495	0.8351	0.7381	0.8618	0.8767	0.7755
	GS [10]	0.8886	0.8046	0.7500	**0.8901**	0.8579	0.7778
	DiscreteSGD [2]	0.8877	0.8442	0.7778	0.8846	0.8347	0.7816
	PPDL [11]	0.8520	0.7116	0.6835	0.8442	0.8295	0.7816
	MID [14]	0.8596	0.8001	0.7640	0.8353	0.7719	0.7731
	InfoScissors [15]	0.8473	0.5553	0.5333	0.8473	0.5778	0.5333
	Casper	**0.8901**	**0.5488**	**0.5333**	0.8708	**0.5275**	**0.5143**
Yahoo Answers	w/o defense	**0.7212**	0.5593	0.5587	**0.7212**	0.6534	0.6528
	DP [9]	0.7150	0.4618	0.4638	0.7152	0.4266	0.4249
	GS [10]	0.6931	0.3373	0.3297	0.7202	0.5057	0.5071
	DiscreteSGD [2]	0.6835	0.4890	0.4910	0.7158	0.4121	0.4071
	PPDL [11]	0.1000	0.5396	0.5407	0.6591	0.4798	0.4798
	MID [14]	0.6256	0.3452	0.3416	0.6256	0.5257	0.5240
	InfoScissors [15]	0.6350	0.5537	0.5553	0.6350	0.6229	0.6238
	Casper	0.6372	**0.3134**	**0.3169**	0.6636	**0.3463**	**0.3493**

重构表示削弱标签依赖，标签混淆限制被动方学习真实监督。差异训练让主动方继续使用真实标签，减轻任务性能损失

实验结果 对比实验（其他标签推理攻击场景）

- Casper同时削弱基于梯度统计和聚类的标签推理

Defense	Norm-based Scoring		Direction-based Scoring	
	Attack↓		Attack↓	
	Training Dataset	Test Dataset	Training Dataset	Test Dataset
w/o defense	0.6597	0.6755	0.7708	0.7813
DP [9]	0.6709	0.6800	0.8509	0.8793
GS [10]	0.6436	0.6711	0.7598	0.7874
DiscreteSGD [2]	0.6695	0.6800	0.7358	0.7874
PPDL [11]	0.5448	0.5556	0.8448	0.7674
MID [14]	0.7947	0.7833	0.5409	0.4944
InfoScissors [15]	0.5889	0.5758	0.6287	0.5714
Casper	**0.5437**	**0.5258**	**0.4841**	**0.3226**

Datasets	Defense	Cluster-smashed		Cluster-grad	
		Attack↓		Attack↓	
		Training Dataset	Test Dataset	Training Dataset	Test Dataset
CIFAR10	w/o defense	0.8699	0.6558	0.9824	0.9735
	DP [9]	0.1014	0.1044	0.1000	0.1000
	GS [10]	0.8115	0.4714	0.7803	0.8479
	DiscreteSGD [2]	0.1014	0.1044	0.1000	0.1000
	PPDL [11]	0.1014	0.1044	0.1000	0.1000
	MID [14]	0.1014	0.1044	0.1000	0.1000
	InfoScissors [15]	0.1146	0.1371	0.2266	0.2134
	Casper	**0.0995**	**0.0941**	**0.1000**	**0.1000**
TinyImageNet	w/o defense	0.3403	0.2381	0.9586	0.7876
	DP [9]	0.0250	0.0250	0.0250	0.0250
	GS [10]	0.3631	0.2446	0.9639	0.8422
	DiscreteSGD [2]	0.0584	0.0721	0.2460	0.3216
	PPDL [11]	0.0361	0.0422	0.2091	0.2520
	MID [14]	0.0350	0.0404	0.0250	0.0250
	InfoScissors [15]	0.0387	0.0507	0.3996	0.2059
	Casper	**0.0250**	**0.0250**	**0.0250**	**0.0250**

- 梯度攻击：梯度不再稳定指向真实标签
  - 混淆标签用于传递被动方梯度，使NS依赖的范数差异、DS依赖的方向差异很难对应真实类别
- 聚类攻击：类别簇被重构与平滑打散
  - 表示重构削弱表示聚类，标签平滑削弱梯度聚类

实验结果 消融实验

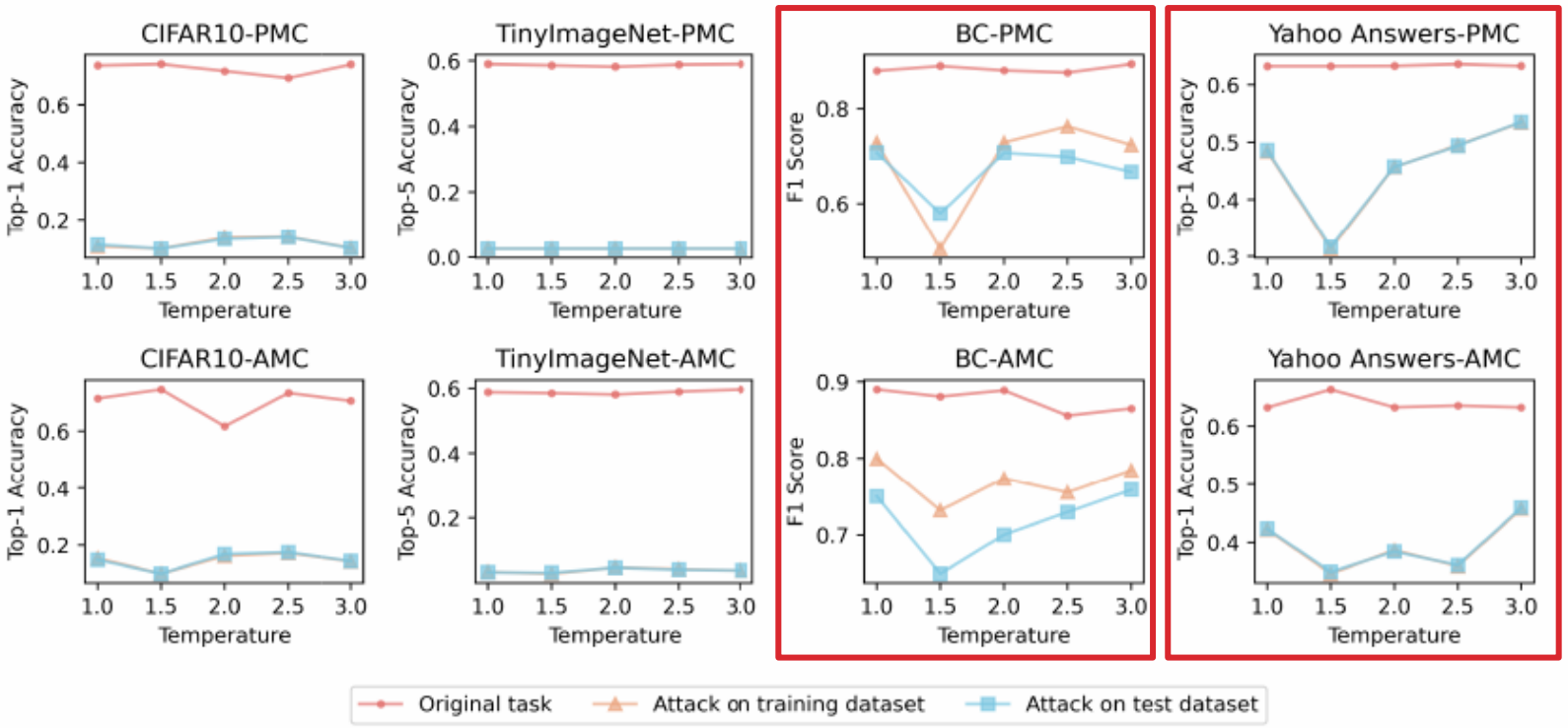
- 三个模块承担了表示保护、标签混淆、效用保持

Datasets	Variations	Passive Model Completion			Active Model Completion		
		Model ↑	Attack ↓		Model ↑	Attack ↓	
		Test Dataset	Training Dataset	Test Dataset	Test Dataset	Training Dataset	Test Dataset
CIFAR10	w/o cut layer output reconstruction	0.7355	0.1001	0.1000	0.7393	0.1000	0.1000
	w/o label obfuscation	0.7397	0.1000	0.1000	0.7331	0.1000	0.1000
	w/o selective discrepancy training	0.1000	0.1000	0.1000	0.1000	0.1461	0.1426
	w/ label random perturbation	0.1000	0.1000	0.1000	0.1000	0.1000	0.1000
	Full	**0.7410**	**0.0987**	**0.0982**	**0.7466**	**0.1000**	**0.1000**
TinyImageNet	w/o cut layer output reconstruction	0.5913	0.0360	0.0354	0.5871	0.0352	0.0363
	w/o label obfuscation	0.5834	0.0250	0.0250	0.5862	0.0328	0.0334
	w/o selective discrepancy training	0.0259	0.0529	0.0517	0.0259	0.0625	0.0564
	w/ label random perturbation	0.0250	0.0250	0.0250	0.0250	**0.0250**	**0.0250**
	Full	**0.5920**	**0.0250**	**0.0250**	**0.5871**	0.0287	0.0312
Breast Cancer	w/o cut layer output reconstruction	0.8593	0.7667	0.7000	0.8501	0.8263	0.7778
	w/o label obfuscation	0.8493	0.5774	0.6053	0.8708	0.7493	0.6914
	w/o selective discrepancy training	0.5333	0.5618	0.5333	0.5333	0.5650	0.5333
	w/ label random perturbation	0.5333	0.5682	0.5333	0.5333	0.7753	0.7073
	Full	**0.8901**	**0.5488**	**0.5333**	**0.8708**	**0.5275**	**0.5143**
Yahoo Answers	w/o cut layer output reconstruction	0.6345	0.3636	0.3583	0.6339	0.3559	0.3560
	w/o label obfuscation	0.5976	0.4390	0.4426	0.6362	0.5791	0.5831
	w/o selective discrepancy training	0.1000	0.4746	0.4761	0.1000	0.5106	0.5105
	w/ label random perturbation	0.2705	0.4905	0.4896	0.2705	0.4787	0.4770
	Full	**0.6372**	**0.3134**	**0.3169**	**0.6636**	**0.3463**	**0.3493**

- 去除差异训练
  - 使用混淆标签很难恢复主任务性能
- 去除表示重构
  - 原始表示暴露标签信息
- 去除标签混淆
  - 使用原始标签训练被动方模型仍能推断标签

实验结果 温度参数实验

- 温度T控制重构标签的平滑程度
  - T过小，保护不足
  - T过大，被动方监督信号过弱，加剧了与主动方真实训练目标的差异，并在此过程中造成训练不稳定，这种不稳定模式可能被攻击者利用



特色创新

- 提出因果混淆的标签隐私防御框架
  - 从因果视角分析切分层输出在训练中累积标签信息的泄露机制
  - 引入表示重构与标签混淆构成的混淆因素，干预表示与真实标签之间的直接因果关系

- 设计表示与标签联合重构机制
  - 利用同一VAE重构被动方表示与标签，兼顾表示保真与标签信息抑制
  - 对重构标签进行温度平滑，降低类别分布的可辨识性，增强标签隐私保护

- 提出选择性差异训练策略
  - 主动方使用真实标签训练，维持联合模型的预测能力
  - 被动方依据混淆标签产生的梯度更新，减少真实标签信息的积累

算法总结

- 算法优势
  - 通过表示重构与标签混淆，削弱中间表示及训练梯度中的标签泄露风险
  - 通过选择性差异训练，在保护标签隐私的同时保留联合模型的预测能力
  - 覆盖模型补全、打分与聚类等多类攻击，并在图像、文本和表格等多模态数据上验证防御效果
- 算法不足（论文分析）
  - 计算开销大：需要额外的重构训练，图像和文本任务的单轮训练开销约为无防御方法的1.6～4.9倍，可能限制其实时和大规模应用
  - 参数具有敏感性：温度T和互信息权重λ会影响隐私与效用的平衡
- 算法分析（个人思考）
  - 根据被动方的标签泄露风险调整混淆强度，降低对低风险参与方特征学习的抑制
  - 探索重构模块训练与更新机制，降低计算开销



特点总结与未来展望

特点总结与未来展望

- 特点总结
  - LCD：引入标签压缩与中心损失，提高对模型补全攻击的防御性能
    - 通过深层特征引导的标签压缩生成虚假标签，削弱被动方底层表示与真实标签的关联，重点抵御模型补全攻击
    - 结合多阶段训练与嵌入紧凑化，在混淆攻击者的同时保持原始分类任务性能
  - Casper：三个模块共同作用针对多种标签推理攻击，提升防御性能
    - 从因果角度分析标签泄露，通过重构切分层表示和混淆标签，削弱表示、梯度与真实标签之间的依赖
    - 采用选择性差异训练兼顾隐私与效用，并同时抵御模型补全、梯度、聚类攻击
- 未来展望
  - 研究多标签分类、多参与方协作以及标签推理、特征推理和后门攻击并存时的统一防御方法
  - 简化训练过程，实现关键参数的自适应调节，降低计算与调参成本

参考文献

[1] YUAN S, LI X, CAO X, et al. Defend Against Label Inference Attacks in Vertical Federated Learning via Label Compression[J]. IEEE Transactions on Mobile Computing, 2026, 25(8): 13166-13182.

[2] SHEN M, MENG J, PENG B, et al. Casper: A Causality-Inspired Defense With Confounder Against Label Inference Attacks in Vertical Split Federated Learning[J]. IEEE Transactions on Information Forensics and Security, 2026, 21: 1050-1064.

道德经

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。

