Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

成员推理攻击

—— 隐私数据是否在被模型关注

硕士研究生 程瑶

2022年10月16日

内容提要

- 背景简介
- 基础概念
  - 影子模型方法
  - 度量学习
- 算法原理
  - AuditMI
  - EncoderMI
- 应用总结
- 参考文献

背景简介

- 预期收获
  - 1. 了解隐私攻击的种类和基本原理
  - 2. 理解成员推理攻击的基本原理
  - 3. 理解成员推理攻击的应用
  - 4. 了解成员推理攻击的前沿发展

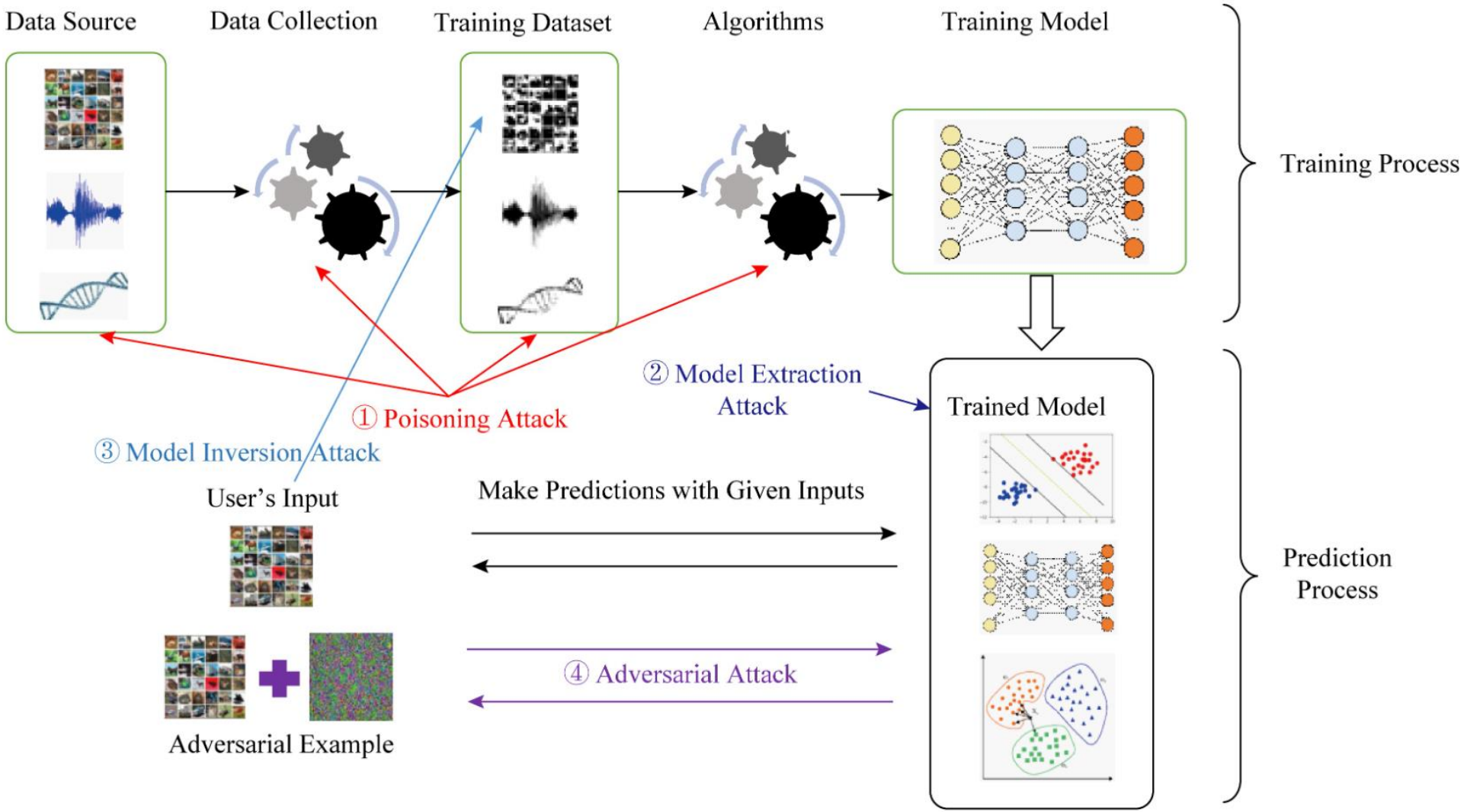
背景简介　隐私攻击、问题导入

- 机器学习系统的“便利”与“潜在威胁”
  - 机器学习使用率上升
    - 机器学习模型被应用到越来越多的领域和任务中，例如人脸识别、医学数据分析与应用、推荐系统以及机器学习即服务提供的各类API
    - 比如，基于文本（如情感分析和文本分类）应用程序的聊天机器人变得流行，值得注意的是，这些应用程序中包含了大量用户隐私与模型知识

机器学习系统中各类威胁案例

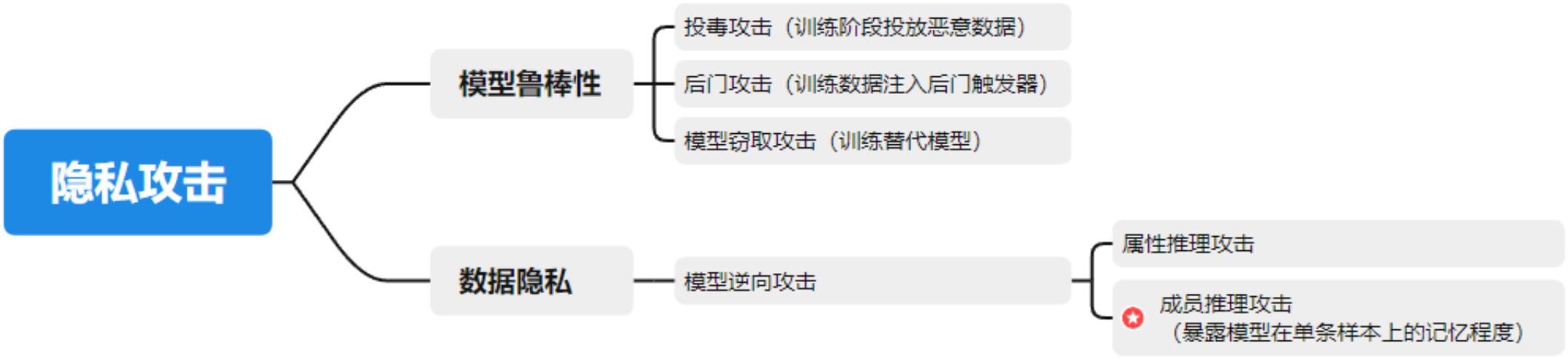
背景简介　隐私攻击、问题导入

- 机器学习系统中的各类攻击



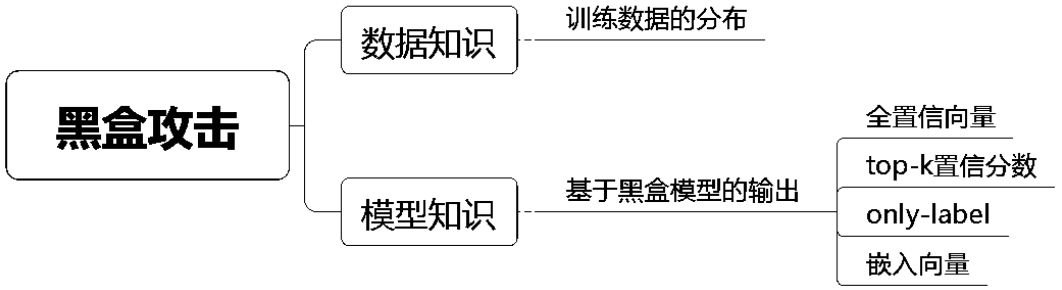
背景简介　隐私攻击、问题导入

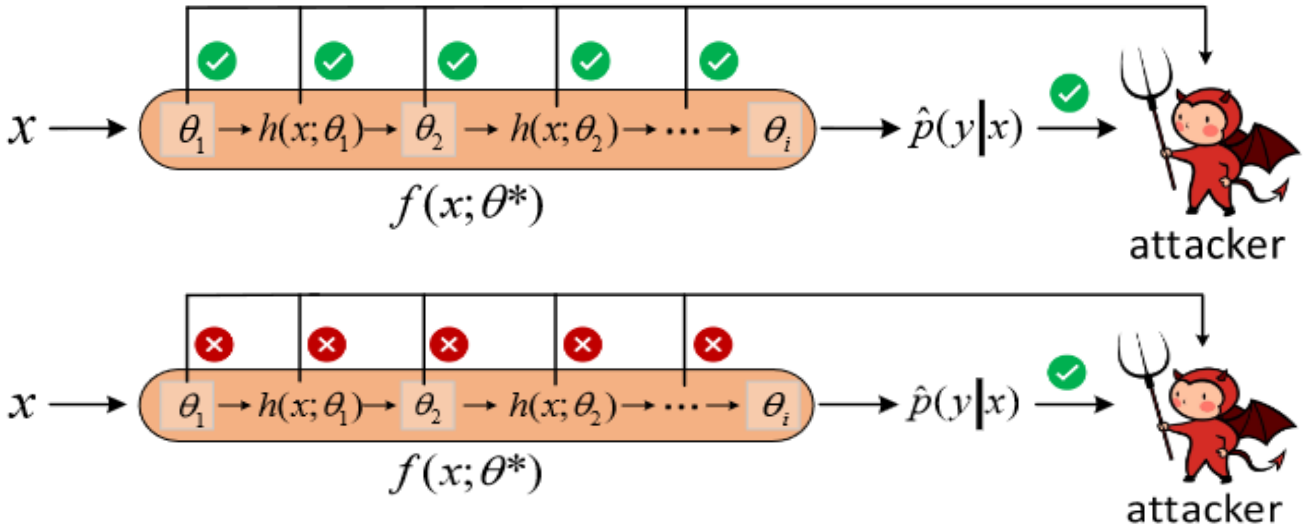
- 研究意义
  - 复杂的数据安全环境
    - 由于机器学习模型依赖于训练数据集的数量和质量，而一些训练数据集中包含敏感信息，这些信息可能会造成用户隐私的泄露
    - 数据泄露始终是一个问题，ML的一些基本方面会改变隐私风险



背景简介　隐私攻击、问题导入

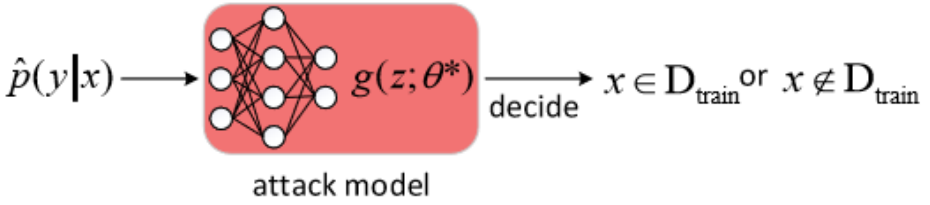
- 攻击场景
  - 白盒攻击
    - 攻击者能获得攻击对象的所有网络参数和中间结果，包含训练数据分布、训练算法、网络结构和参数等先验知识
  - 黑盒攻击（实际场景）
    - 攻击者的知识：目标模型的查询结果、训练数据分布

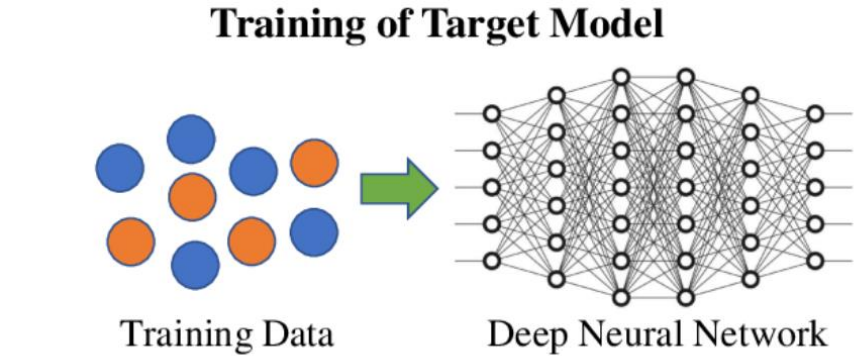


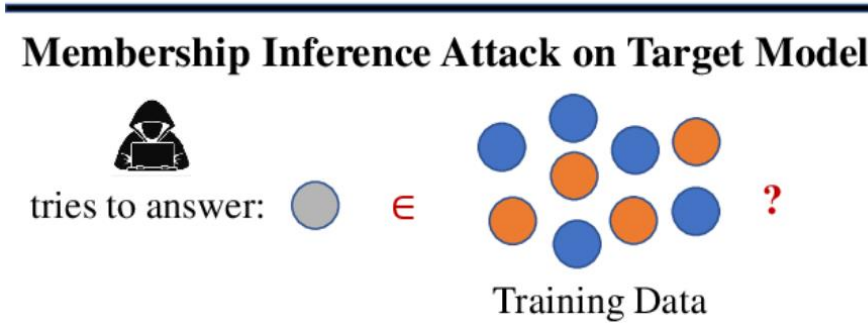


基础概念 是什么？怎么做？

- 基本概念
  - 成员推理攻击（Membership Inference Attack, MIA）
    - 攻击目标：分辨出某些数据样本（隐私对象）是否被用于某一类机器学习模型（目标模型）所训练，以此实施数据隐私攻击
    - 目标模型：攻击对象，基于目标模型的输出（置信度向量、分类标签、嵌入向量等）
    - 成员（Member）数据：目标模型的训练集
    - 非成员（Non-member）数据：目标模型的测试集
  - 实现方法
    - 构造影子模型近似目标模型的功能和输出
    - 基于度量的方法



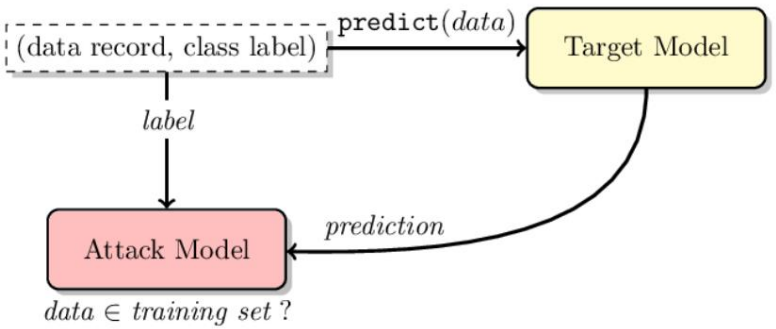




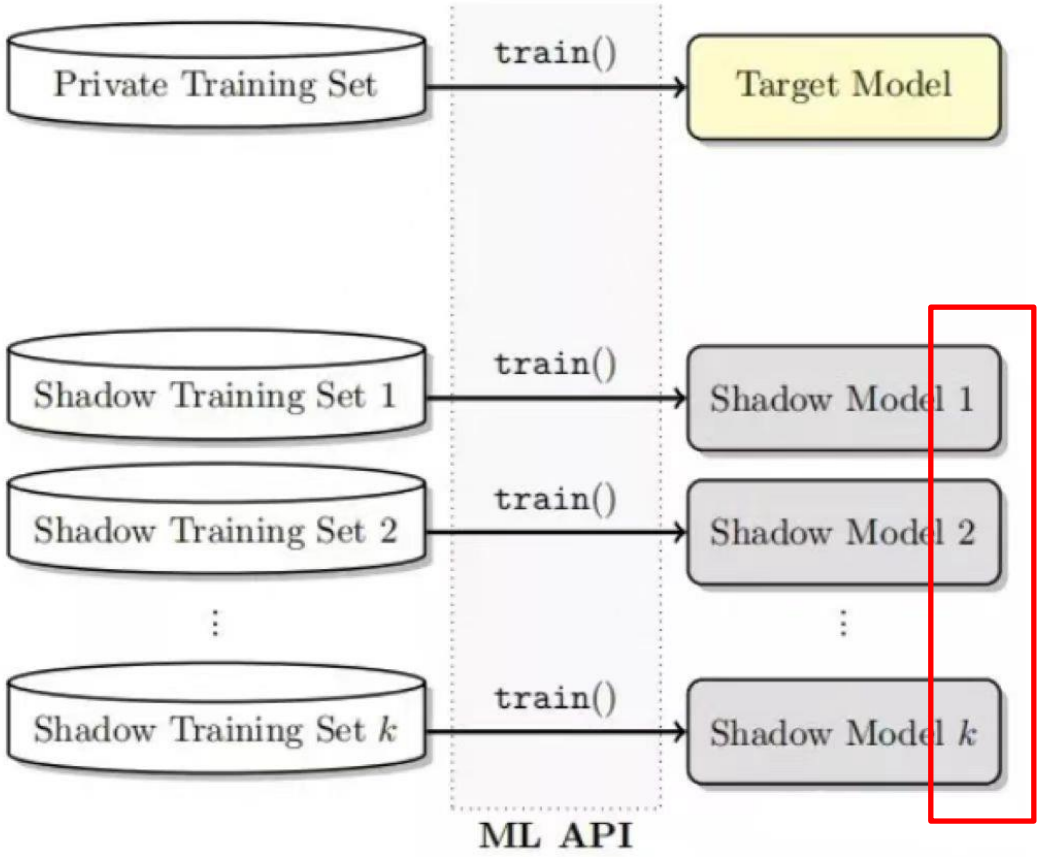
- 实现方法
  - 攻击提出的动机
    - 由于目标模型的过拟合特性，其对于训练数据和测试数据的表现存在差异
  - 基于目标模型的成员数据和非成员数据在输出上表现不同的原理
    - 训练一个二分类攻击模型f_{attack}，攻击模型的输入x_{attack}是由正确的标签类和目标模型输出组成
    - 该攻击模型用来捕捉目标模型训练数据（member）和测试数据（non-member）的差异，输出二分类结果
  - 数学描述

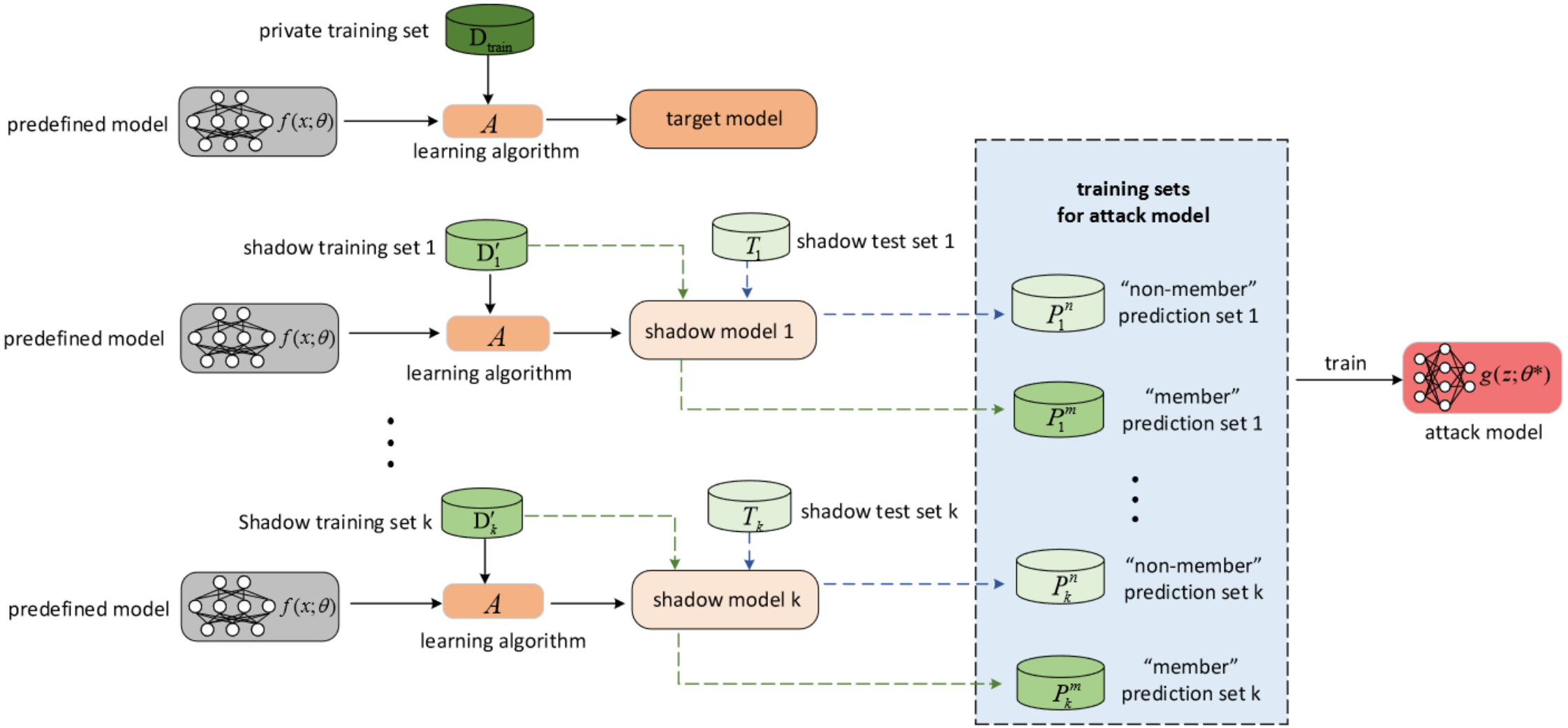
$$f_{attack}(x_{attack}) \rightarrow y$$

$$x_{attack} = (label, prediction),\ y \in \{in, out\}$$



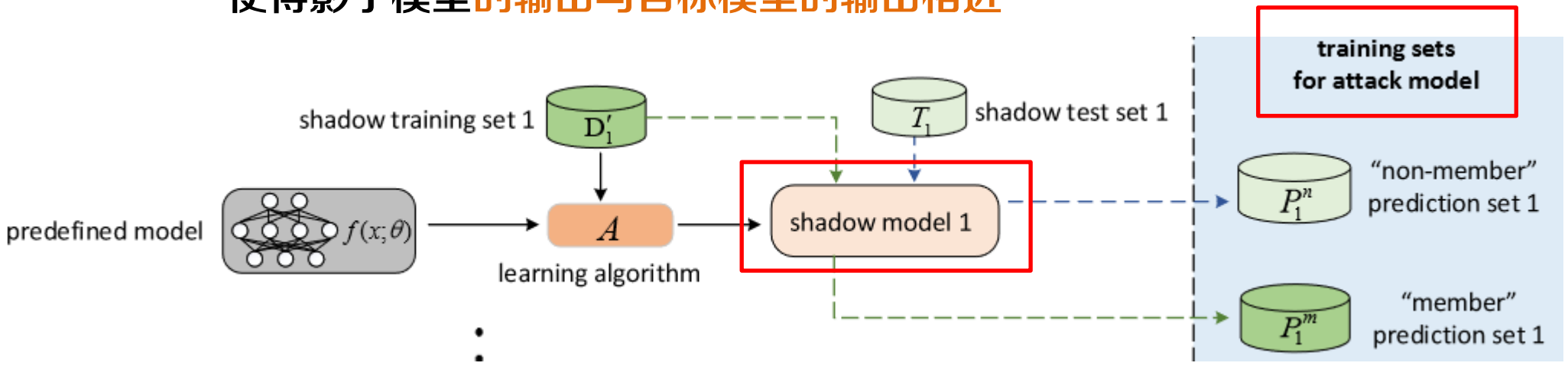
- 基于构造的影子模型实现成员推理
  - 影子模型：高仿版目标模型
  - 影子模型（Shadow Model）的作用：
    - 在黑盒攻击中，由于访问次数和攻击成本的限制，构造影子模型近似目标模型的输出（主流方法）
  - 约束条件：
    - 攻击模型训练数据和目标模型训练数据同分布
    - 影子模型的输出近似等价目标模型的输出
    - 在上述两点约束下，基于影子模型实施攻击；多项实验结果表明，使用影子模型能有效提升攻击分类模型的效果




private training set
D_{train}
predefined model
$f(x;\theta)$
A
learning algorithm
target model
shadow training set 1
D'_1
T_1
shadow test set 1
predefined model
$f(x;\theta)$
A
learning algorithm
shadow model 1
Shadow training set k
D'_k
T_k
shadow test set k
predefined model
$f(x;\theta)$
A
learning algorithm
shadow model k
training sets
for attack model
P_1^n
"non-member"
prediction set 1
P_1^m
"member"
prediction set 1
P_k^n
"non-member"
prediction set k
P_k^m
"member"
prediction set k
train
$g(z;\theta^*)$
attack model

影子模型方法

- 基于构造的影子模型实现成员推理
  - 核心思想
    - 2017年，Shokri等人首次提出一个核心思想——构造影子模型的方法
    - 训练攻击模型时需要样本真实标签、目标模型输出知识
    - 如何训练影子模型?
      - 使用已有的数据样本（与目标模型的训练样本同分布）训练与目标模型任务与结构相同的模型
      - 使得影子模型的输出与目标模型的输出相近



- 基于度量方法实现成员推理
  - 度量方法
    - 基于黑盒目标模型的输出（仅标签、置信度向量、嵌入向量等）设定不同的度量方法进行成员判别
  - 度量种类
    - 预测正确率：

$$M_{corr}(\hat{p}(y|x), y) = \mathbb{I}(argmax\,\hat{p}(y|x) = y)$$

    - 预测损失：

$$M_{loss}(\hat{p}(y|x), y) = \mathbb{I}(\mathcal{L}(\hat{p}(y|x); y) \leq \tau)$$

    - 修正熵：

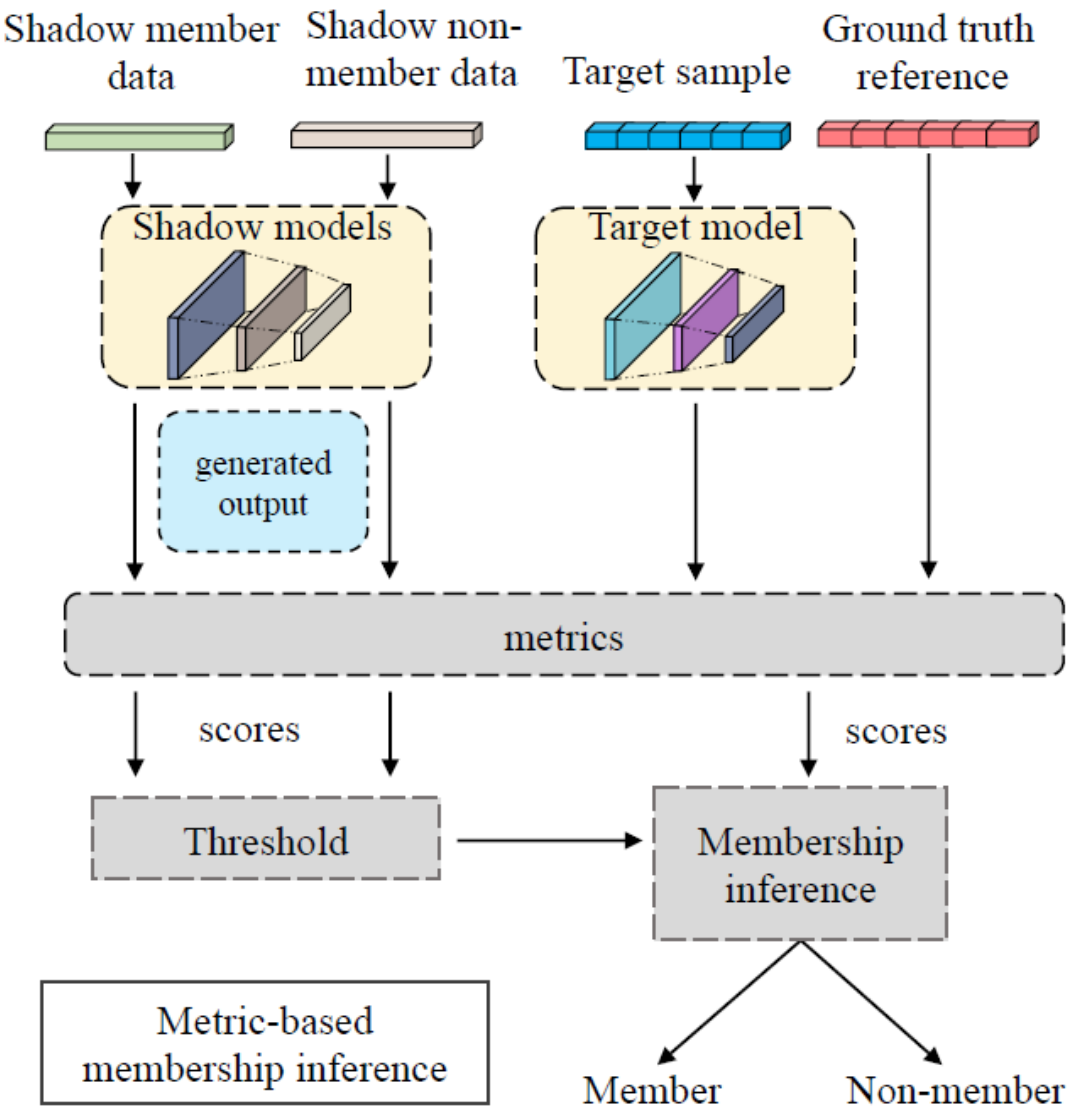
$$MH(\hat{p}(y|x), y) = -(1 - p_y)\log(p_y) - \sum_{i \neq y} p_i \log(1 - p_i)$$

$$M_{Mentr}(\hat{p}(y|x), y) = \mathbb{I}(MH(\hat{p}(y|x); y) \leq \tau)$$

- 基于度量方法实现成员推理
  - 度量种类
    - 预测置信度：

$$M_{conf}\left(\hat{p}(y|x)\right) = \mathbb{I}(max\hat{p}(y|x) \geq \tau)$$

  - 核心思想
    - 基于影子模型和目标模型的输出进行度量，推断成员身份
    - 2019年首次提出结合影子模型与度量学习方法的攻击（文本嵌入模型）



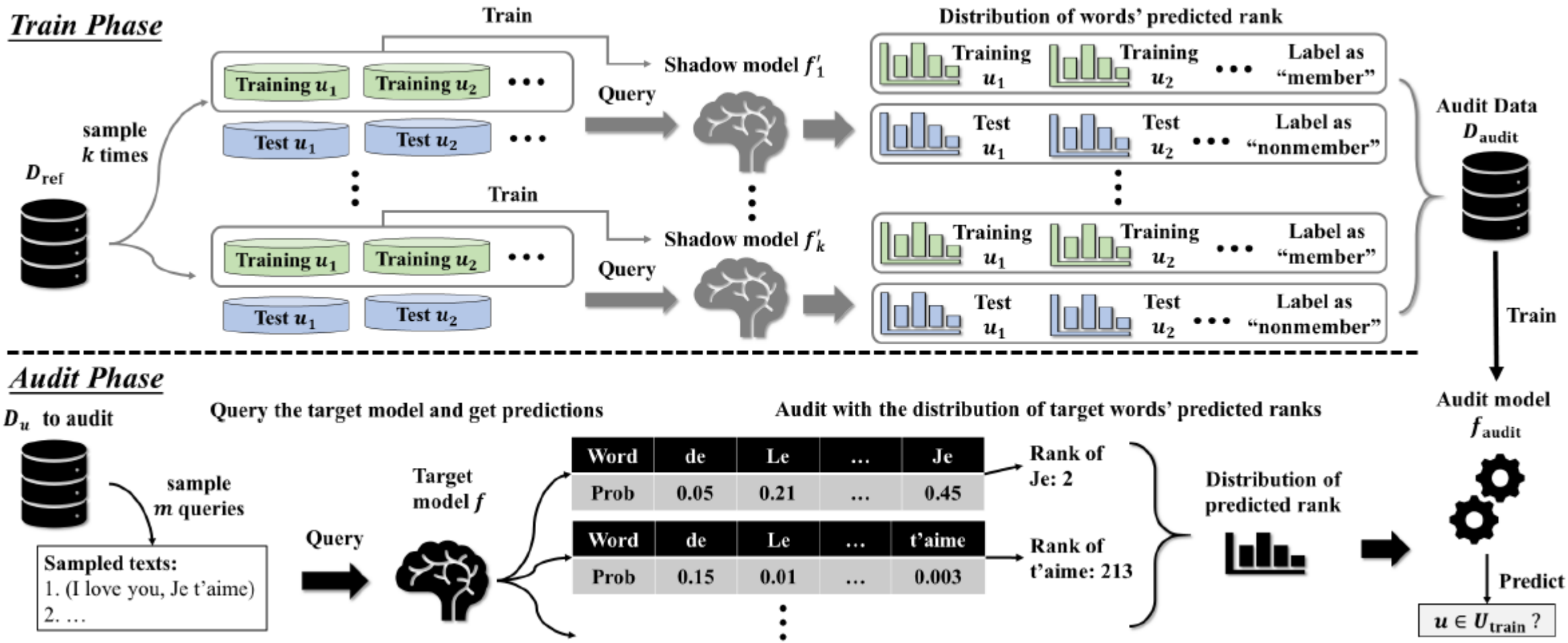
AuditMI

AuditMI TIPO

T	对文本生成模型进行成员推理
I	文本（评论、对话等） Reddit、SATED、Dialogs
P	1.划分影子模型训练数据、测试数据； 2.训练影子模型； 3.基于影子模型的输出，构造攻击分类器； 4.成员推理。
O	成员or非成员

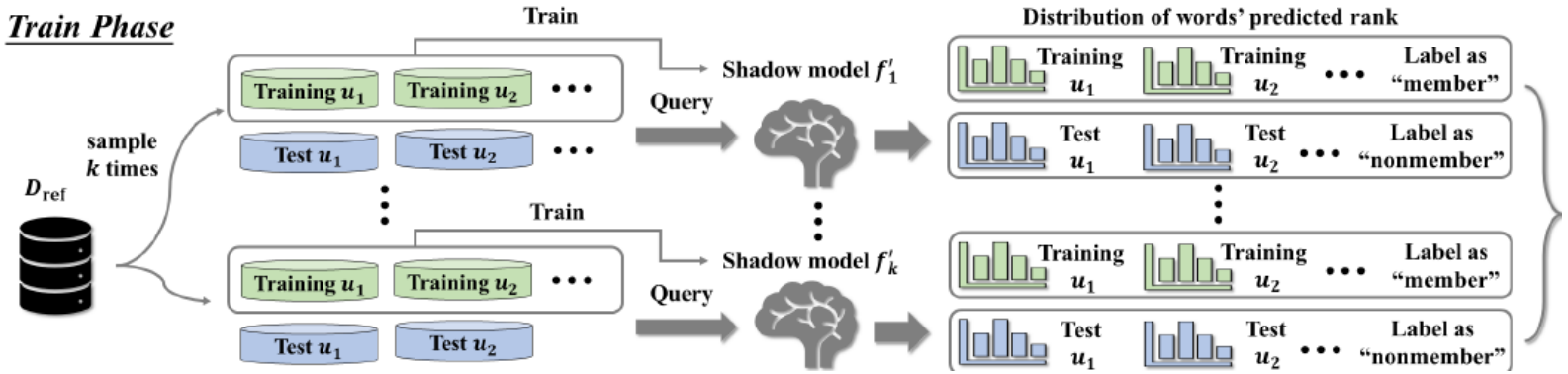
P	成员特征表示与攻击模型的迁移
C	已知成员数据分布和目标模型的应用领域
D	如何在成员样本不充分的情况下学习成员特征表示
L	CCS 2019 (CCF A)

- 算法原理图


Train Phase
Train
Distribution of words' predicted rank
Training u_1
Training u_2
Shadow model f'_1
Query
Label as "member"
Test u_1
Test u_2
Label as "nonmember"
sample k times
D_{ref}
Shadow model f'_k
Audit Data D_{audit}
Train
Audit Phase
D_u to audit
Query the target model and get predictions
Audit with the distribution of target words' predicted ranks
Audit model f_{audit}
sample m queries
Sampled texts:
1. (I love you, Je t'aime)
2. …
Target model f
Word
Prob
de
Le
…
Je
0.05
0.21
0.45
t'aime
0.15
0.01
0.003
Rank of Je: 2
Rank of t'aime: 213
Distribution of predicted rank
Predict
$u \in U_{\text{train}}$?

AuditMI 算法原理

- 攻击模型训练阶段
  - 依据用户身份 $u_1, u_2, \ldots, u_n$ 划分每个影子模型的训练数据、测试数据
  - 基于目标模型的输出，训练k个影子模型$f_1', f_2', \ldots, f_k'$，通过计算生成词语的分布排名，记录潜在成员样本，标记为攻击模型的训练数据
  - 训练一个攻击模型f_{audit}，实施成员（词语）推理



Algorithm 1: Auditing text-generation models

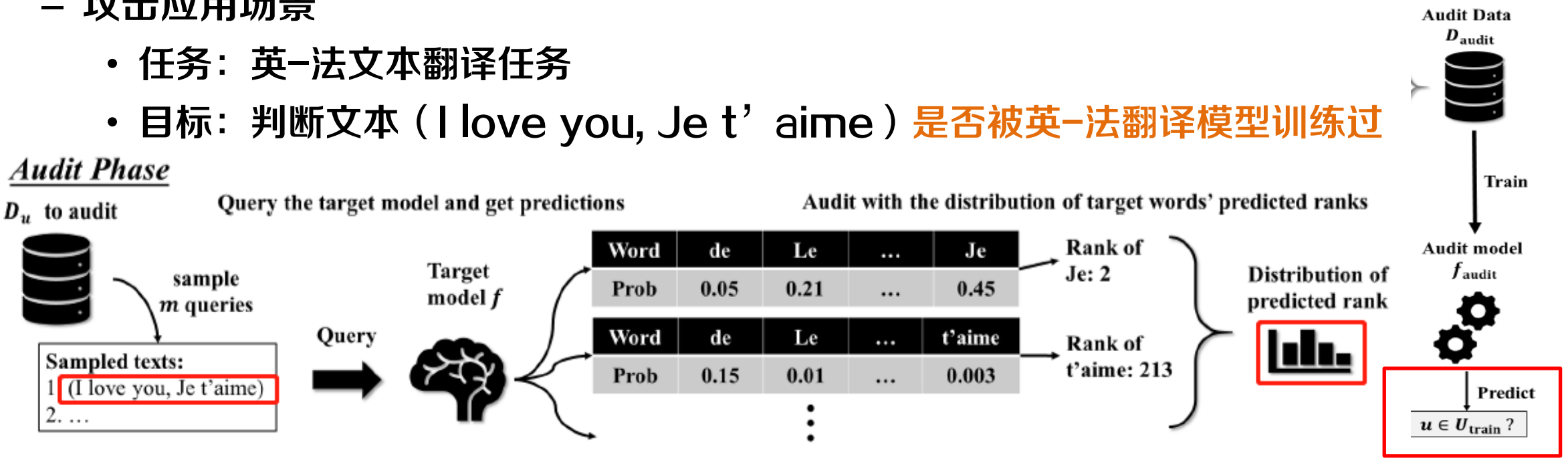
Hyper-parameters: auditor's reference dataset $\mathcal{D}_{ref}$, number of shadow models k, user's data $\mathcal{D}_u$, target model f, target model-training protocol $\mathcal{T}_{target}$, audit model-training protocol $\mathcal{T}_{audit}$, maximum number of queries m, number of bins in histogram d

function AuditMembership()
  $f_{audit} \leftarrow$ TrainAuditModel()
  $\mathcal{D}_{sample,u} \leftarrow$ SampleQueries$(m, \mathcal{D}_u)$
  $h_u \leftarrow$ HistogramFeature$(f, \mathcal{D}_{sample,u})$
  return prediction of membership $f_{audit}(h_u)$

function SampleQueries$(m, \mathcal{D})$
  if random sample **then**
    return randomly selected m rows in $\mathcal{D}$
  else ▷ *sample based on frequency*
    $C \leftarrow \{\Sigma$ (frequency of w for w in $y)\ |\forall(x,y) \in \mathcal{D}\}$
    $I \leftarrow$ indices of m smallest values in C
    return m rows in $\mathcal{D}$ indexed by I
  end if

function TrainAuditModel()
  $\mathcal{D}_{audit} \leftarrow \emptyset$ ▷ *dataset for building the audit model*
  $\mathcal{U}_{ref} \leftarrow$ users in $\mathcal{D}_{ref}$
  for $i = 1$ to k **do** ▷ *train k shadow models*
    $\mathcal{U}_{ref}^{train}, \mathcal{U}_{ref}^{test} \leftarrow$ random split $\mathcal{U}_{ref}$
    $\mathcal{D}_{ref}^{train} \leftarrow \cup_{u \in \mathcal{U}_{ref}^{train}} \{\mathcal{D}_{ref,u}\}$
    Train a shadow model $f_i' \leftarrow \mathcal{T}_{target}(\mathcal{D}_{ref}^{train})$.
    for every u in users of $\mathcal{U}_{ref}$ **do**
      $\mathcal{D}_{ref,u} \leftarrow$ data in $\mathcal{D}_{ref}$ associated with u
      $h_u' \leftarrow$ HistogramFeature$(f_i', \mathcal{D}_{ref,u})$
      $z_u' \leftarrow 1$ if u in $\mathcal{U}_{ref\text{-}train}$ else 0
      $\mathcal{D}_{audit} \cup \{(h_u', z_u')\}$
    end for
  end for
  Train the audit model $f_{audit} \leftarrow \mathcal{T}_{audit}(\mathcal{D}_{audit})$
  return f_{audit}

- 攻击模型应用阶段
  - 采样方法
    - 排序：对目标模型（文本生成模型）的输出词排序计分
    - 概率向量：记录不同数量生成词的概率向量
  - 攻击应用场景
    - 任务：英-法文本翻译任务
    - 目标：判断文本（I love you, Je t’aime）是否被英-法翻译模型训练过



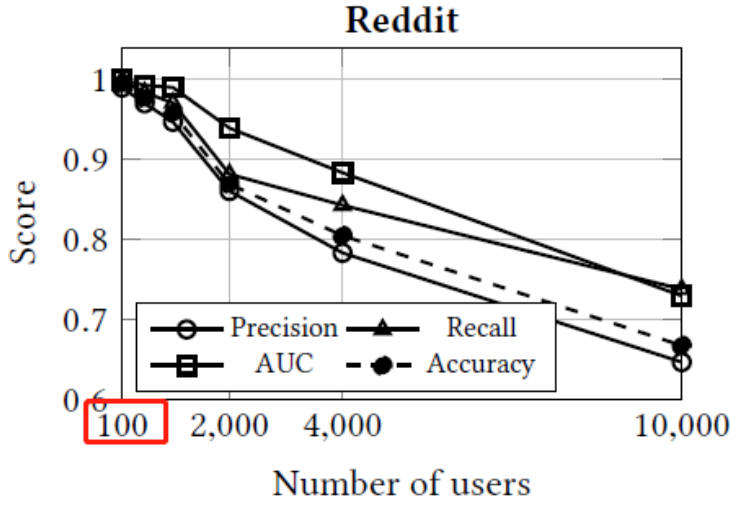
- 实验设计
  - 数据集
    - 3个公开数据集 Reddit、SATED、Dialogs（均包含用户信息）进行实验
    - Reddit comments dataset：总计选择83293个用户的247条帖子评论，用作下一个词预测任务
    - SATED：总计选择271K条句子中的2324个评论，用作en-fr机器翻译任务
    - Cornell movie dialogs corpus：电影对话，617部影片中的9035句对话，用作对话生成任务
  - 评价指标：
    - Accuracy，AUC，Precision，Recall

- 实验结果
  - 词语预测任务

Dataset	Model	Train Acc	Test Acc	Train Perp	Test Perp
Reddit	1-layer LSTM [12]	0.184	0.206	102.22	113.14
SATED	Seq2Seq w/ attn [24]	0.587	0.535	6.36	10.28
Dialogs	Seq2Seq w/o attn	0.283	0.264	45.57	61.11

    - Reddit用户评论数据集中的用户数量对攻击效果的影响
  - 不同超参数的影响（训练影子模型）

Dataset	Accuracy	AUC	Precision	Recall
Reddit	0.990	0.993	0.983	0.996
SATED	0.965	0.981	0.937	0.996
Dialogs	0.978	0.998	0.958	1.000



- 实验分析
  - 采样的大概率生成词数量越多，攻击准确率较高
  - 谷歌翻译API与有道翻译API在训练数据上泄露隐私的程度

No obfuscation: i see so many adults that could benefit from this going around having themselves a big fat sugar snack or soda pop as a treat it 's so sad

Google: i saw so many adults who can benefit from cherishing big fat sugar snacks and soda pop and going around, it is very sad

Yandex: i think a lot of adults have benefited over your big fat candy and and handling of grief

Dataset	Accuracy	AUC	Precision	Recall
Baseline	1.000	1.000	1.000	1.000
Google	0.580	0.858	0.944	0.170
Yandex	0.500	0.782	0.500	0.010

Reddit	Same domain				Cross domain			
$\|f(x)\|$	Acc	AUC	Pre	Rec	Acc	AUC	Pre	Rec
1	0.545	0.549	0.574	0.350	0.505	0.589	0.667	0.020
5	0.550	0.572	0.553	0.520	0.490	0.525	0.495	0.920
10	0.580	0.602	0.582	0.570	0.500	0.552	0.500	0.950
50	0.605	0.648	0.606	0.600	0.505	0.659	0.503	0.980
100	0.725	0.788	0.765	0.650	0.585	0.714	0.549	0.950
500	0.970	0.998	0.970	0.970	0.905	0.992	0.988	0.820
1000	0.985	0.999	0.971	1.000	0.910	0.999	1.000	0.820

SATED								
$\|f(x)\|$	Acc	AUC	Pre	Rec	Acc	AUC	Pre	Rec
1	0.723	0.785	0.770	0.637	0.723	0.785	0.712	0.750
5	0.748	0.838	0.767	0.713	0.767	0.834	0.755	0.790
10	0.800	0.880	0.783	0.830	0.805	0.878	0.814	0.790
50	0.928	0.973	0.908	0.953	0.925	0.979	0.947	0.900
100	0.948	0.981	0.944	0.953	0.942	0.978	0.965	0.917
500	0.972	0.988	0.958	0.987	0.970	0.988	0.983	0.957
1000	0.960	0.984	0.939	0.983	0.967	0.985	0.973	0.960

Dialogs								
$\|f(x)\|$	Acc	AUC	Pre	Rec	Acc	AUC	Pre	Rec
1	0.577	0.618	0.582	0.547	0.538	0.618	0.520	0.977
5	0.575	0.642	0.582	0.530	0.552	0.643	0.528	0.970
10	0.583	0.645	0.591	0.543	0.543	0.638	0.523	0.977
50	0.605	0.660	0.611	0.580	0.537	0.610	0.520	0.963
100	0.647	0.714	0.643	0.660	0.570	0.669	0.541	0.920
500	0.935	0.975	0.917	0.957	0.925	0.969	0.895	0.963
1000	0.972	0.995	0.955	0.990	0.962	0.992	0.948	0.977

优劣分析

- 优势
  - 首次将成员推理攻击实施到文本生成模型，可以实现攻击模型迁移
  - 分别对生成词预测、机器翻译、对话生成等下游任务分析训练数据泄露的程度

- 劣势
  - 影子模型过多，攻击成本较高

EncoderMI

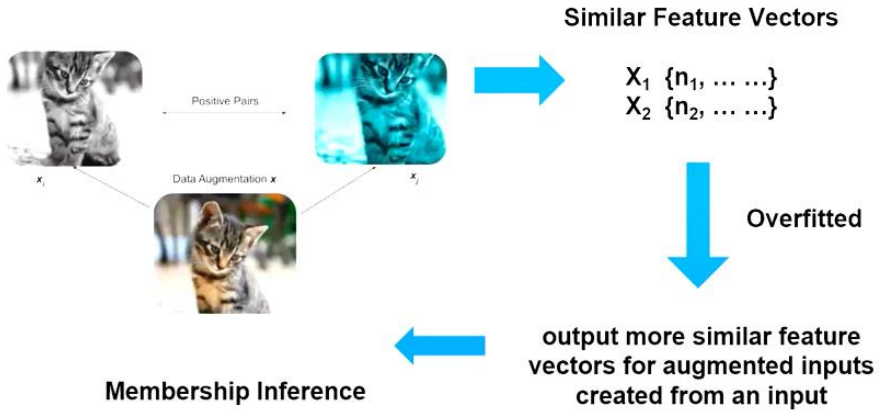
EncoderMI TIPO

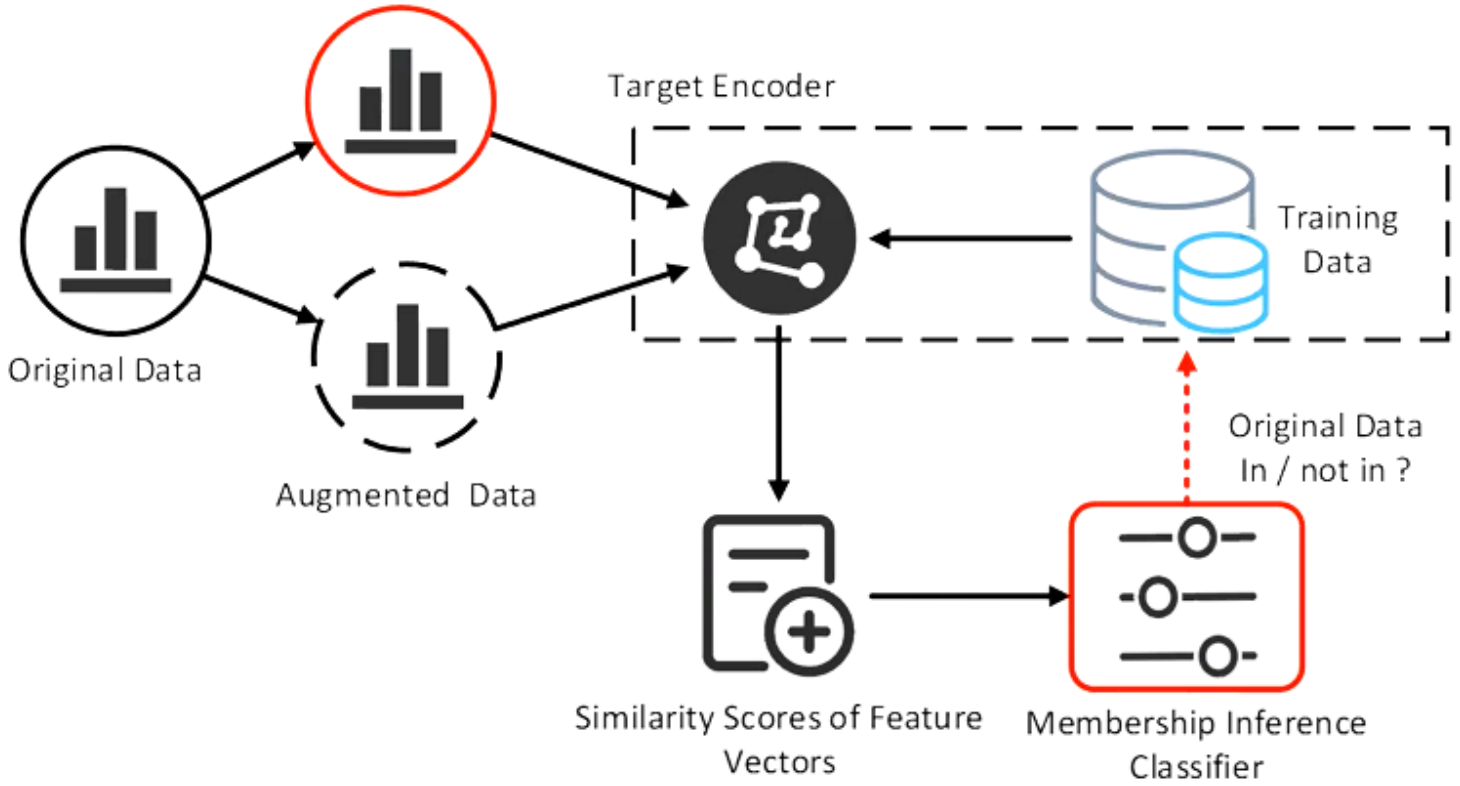
T	对基于对比学习训练的图像预训练编码模型进行成员推理
I	图像分类数据集 CIFAR10, STL10, Tiny-ImageNet
P	1.划分目标编码器、影子编码器的数据集； 2.预训练影子编码器、目标编码器； 3.实施vector_based、threshold_based、set_based三类攻击。
O	成员or非成员

P	多种先验知识条件下实施攻击
C	影子编码器的结构和目标编码器的结构相似，具备成员数据分布的先验知识
D	如何充分利用成员样本的特征表示
L	CCS 2021 (CCF A)

EncoderMI 算法原理

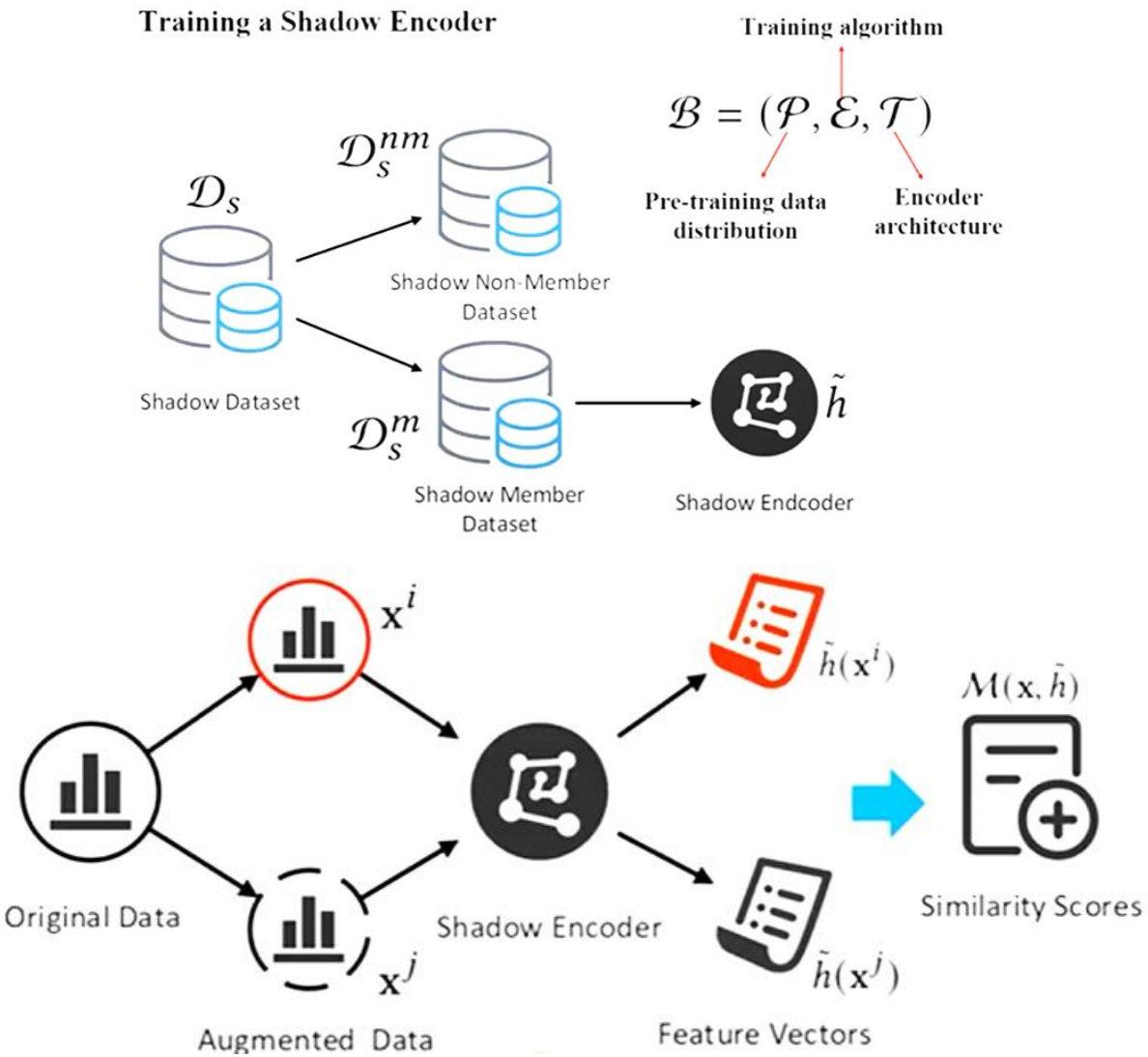
- 攻击算法提出的动机
  - 基于对比学习训练的目标模型具备更多的训练数据（成员）特征
  - 具体攻击算法原理如下：





EncoderMI 算法原理

- 预训练影子编码器
  - 将已有同分布的数据划分成两个不相交的数据集：
    - 影子训练集D_S^m
    - 影子测试集D_S^{nm}
  - 使用D_S^m训练影子编码器
    - 基于原始数据、增强数据的编码向量，提取成员特征，计算相似性分数
    - 基于影子编码器的输出V_S^m和成员身份标签，生成分类器推理训练数据，然后构造攻击分类器



- 预训练影子编码器
  - 攻击者假设已知目标编码器的网络结构
  - 基于构造的影子编码器$\tilde{h}$，提取成员特征：
    - 基于对比学习生成相似的特征向量这一思想，对编码器的原始数据和增强版数据进行相似性度量
      - 给定一个输入x，首先对其进行n次数据增强$x^1, x^2, \ldots, x^n$，影子编码器对每一个增强数据编码$\tilde{h}(x^i)$，组成多对相似性分数

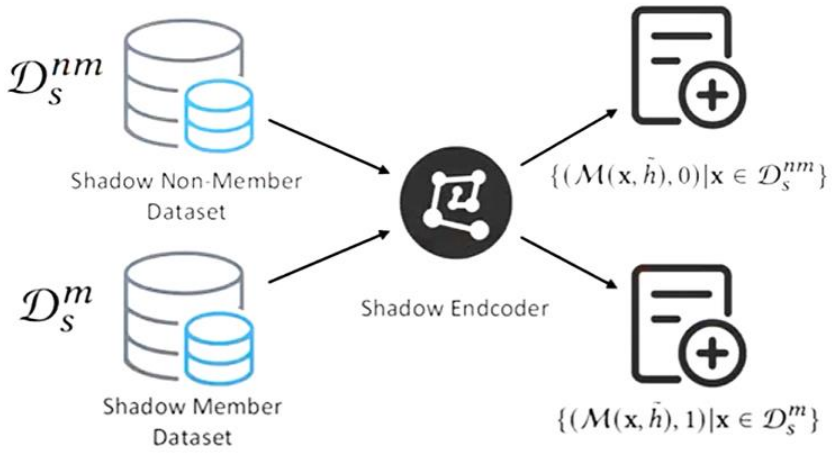
$$M(x, \tilde{h}) = \{S(\tilde{h}(x^i), \tilde{h}(x^j)) | i \in [1, n], j \in [1, n], j > i\}$$

      - $M(x, \tilde{h})$是输入x基于影子编码器$\tilde{h}$的成员特征
      - $S(\cdot,\cdot)$计算两个特征向量的余弦相似度

EncoderMI 算法原理

- 构建推理分类器
  - 给定输入$x \in D_s^m$，经过影子编码器对成员特征编码$M(x, \tilde{h})$，将其标签定为“member”（“1”）
  - 给定输入$x \in D_s^{nm}$，经过影子编码器对成员特征编码$M(x, \tilde{h})$，将其标签定为“non-member”（“0”）
  - 由上述两部分数据组成分类器的训练数据，表示如下：

$$\mathcal{E} = \{(M(x, \tilde{h}), 1) | x \in D_s^m\} \bigcup \{(M(x, \tilde{h}), 0) | x \in D_s^{nm}\}$$



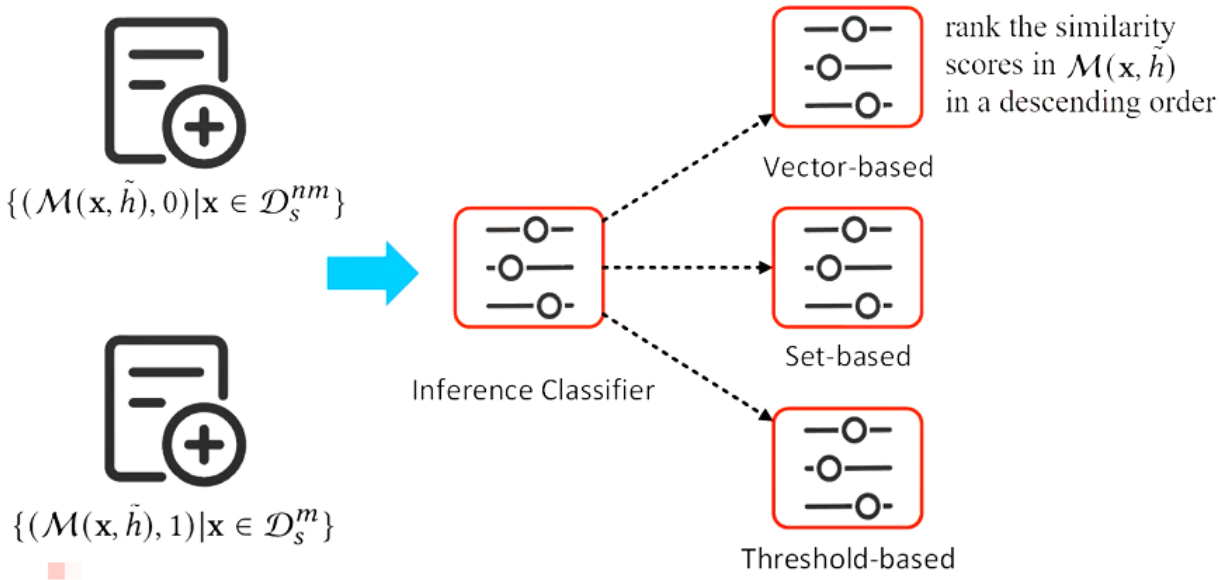
EncoderMI 算法原理

- 训练下游攻击分类器（MLP）
  - EncoderMI-V：一个两层隐藏层的全连接网络
  - EncoderMI-S：DeepSets
  - EncoderMI-T：预定义的成员推理阈值

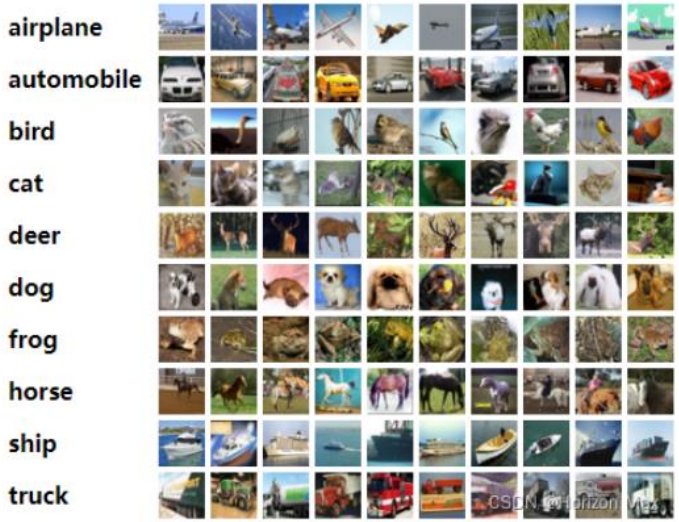
- 成员推理（攻击模型应用阶段）
  - 给定图像编码器的黑盒访问方式，推断一副图像样本是否被目标编码器训练过
  - 成员or非成员

Algorithm 1 Our Membership Inference Method EncoderMI

Require: f_v (or f_s or f_t), h, $\mathcal{A}$, n, S, and x
Ensure: Member or non-member
1: $\{\mathrm{x}^1, \mathrm{x}^2, \cdots, \mathrm{x}^n\} \leftarrow$ AUGMENTATION$(\mathrm{x}, \mathcal{A}, n)$
2: $\mathcal{M}(\mathrm{x}, h) \leftarrow \{S(h(\mathrm{x}^i), h(\mathrm{x}^j)) | i \in [1, n], j \in [1, n], j > i\}$
3: **if** f_v **then**
4: **return** f_v(RANKING$(\mathcal{M}(\mathrm{x}, h)))$
5: **else if** f_s **then**
6: **return** $f_s(\mathcal{M}(\mathrm{x}, h))$
7: **else if** f_t **then**
8: **return** f_t(AVERAGE$(\mathcal{M}(\mathrm{x}, h)))$
9: **end if**



- 实验数据
  - CIFAR10
    - 这是包含十类物品的32*32的RGB图像数据集
    - 事实上，数据包含50,000张训练样本和10,000张测试样本



  - STL10
    - 这是包含十类物品的96*96的RGB图像数据集
    - 选择5,000张训练样本和8,000张测试样本，100,000张无标签图像
  - Tiny-ImageNet
    - 这是包含200类物品的64*64的RGB图像数据集
    - 选择100,000张训练样本和10,000张测试样本

EncoderMI 实验设计

- 实验设置
  - 训练目标编码器
    - 使用MoCo对比学习算法预训练ResNet18
  - 训练影子编码器
    - 使用MoCo或者SimCLR预训练ResNet18或VGG-11
  - 构造三类攻击分类器
- 实验指标
  - Accuracy、Precision、Recall
- 基线方法
  - 仅在给定目标模型API的情况下，对目标模型输出的特征向量实施分类（未利用对比学习思想）（Baseline-1）
  - 仅对目标模型输出的特征向量实施相似性排名，实施分类（Baseline-2）

EncoderMI 实验结果

- 实验结果
  - Baseline-1（左上图）
  - Baseline-2（右上图）
  - 特点：在攻击者掌握不同知识的前提下，实施成员分类的准确率比较
    - 基于特征向量的分类器指标更高

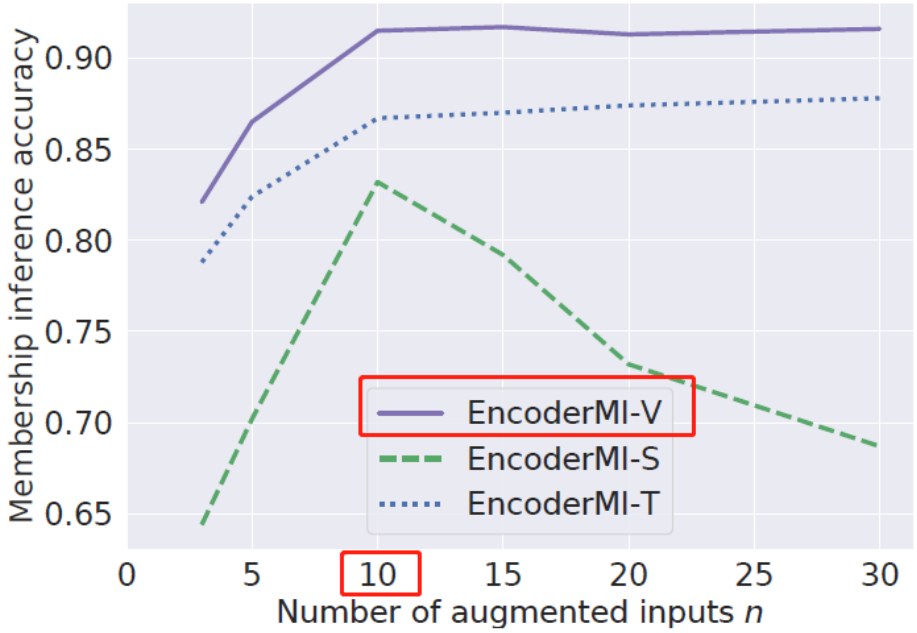
Pre-training dataset	Accuracy	Precision	Recall
CIFAR10	50.7	50.6	51.0
STL10	50.1	49.9	50.3
Tiny-ImageNet	49.5	49.3	49.2

Pre-training dataset	Accuracy	Precision	Recall
CIFAR10	64.5	63.8	67.2
STL10	67.0	65.7	71.3
Tiny-ImageNet	68.6	67.8	70.8

Pre-training data distribution	Encoder architecture	Training algorithm	Accuracy			Precision			Recall		
			Encod-erMI-V	Encod-erMI-S	Encod-erMI-T	Encod-erMI-V	Encod-erMI-S	Encod-erMI-T	Encod-erMI-V	Encod-erMI-S	Encod-erMI-T
×	×	×	86.2 (2.04)	78.1 (2.21)	82.1 (1.91)	87.8 (1.15)	78.9 (1.69)	80.1 (1.01)	89.3 (2.15)	86.8 (2.44)	87.2 (1.89)
√	×	×	86.9 (2.03)	79.6 (2.05)	83.3 (1.75)	88.3 (1.64)	79.8 (1.34)	81.0 (1.12)	90.9 (2.44)	87.4 (2.51)	89.2 (1.63)
×	√	×	87.0 (1.21)	79.4 (1.39)	83.5 (1.03)	88.4 (1.37)	79.6 (0.98)	81.6 (0.87)	91.4 (1.17)	87.2 (1.46)	87.9 (0.92)
×	×	√	86.7 (0.81)	79.2 (1.05)	83.0 (0.77)	88.2 (0.83)	79.9 (1.10)	80.4 (0.81)	91.4 (0.76)	87.1 (1.01)	89.1 (0.79)
√	√	×	87.2 (2.17)	79.9 (1.88)	83.6 (1.66)	88.4 (1.38)	80.1 (1.03)	81.9 (0.98)	91.5 (1.23)	87.9 (1.32)	87.9 (1.01)
√	×	√	87.6 (0.45)	80.3 (0.47)	84.2 (0.39)	88.7 (0.43)	80.6 (0.44)	83.1 (0.37)	91.7 (0.51)	87.7 (0.49)	88.0 (0.46)
×	√	√	90.2 (0.37)	81.1 (0.43)	85.2 (0.37)	89.5 (0.31)	80.5 (0.39)	85.1 (0.33)	93.3 (0.32)	88.2 (0.48)	88.9 (0.38)
√	√	√	91.4 (0.28)	83.1 (0.27)	86.6 (0.28)	90.1 (0.23)	80.8 (0.29)	85.9 (0.27)	93.5 (0.22)	88.3 (0.30)	89.1 (0.25)

EncoderMI 实验结果

- 实验结果（三种攻击分类器）
  - 不同增强次数的影响
  - 度量方式的比较
    - 攻击分类器：EncoderMI-V、EncoderMI-S、EncoderMI-T
    - 相似性度量方式：Cosine、Correlation、Euclidean



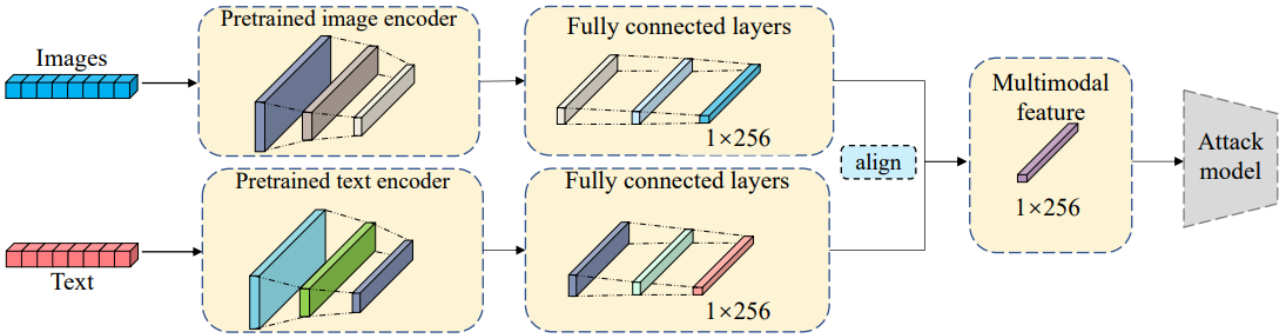
Method	Similarity metric	Accuracy	Precision	Recall
EncoderMI-V	Cosine	91.5	90.0	93.5
	Correlation	89.3	87.9	92.2
	Euclidean	88.9	85.3	94.5
EncoderMI-S	Cosine	83.2	80.5	87.9
	Correlation	76.3	75.5	78.5
	Euclidean	75.8	73.6	81.4
EncoderMI-T	Cosine	86.7	85.7	89.0
	Correlation	80.6	79.8	81.6
	Euclidean	80.7	80.1	82.5

优劣分析

- 优势
  - EncoderMI基于增强样本和原始样本具有相似的特征向量原理，增强了成员样本的特征表示
  - 考虑了多种先验知识的条件下的攻击效果，实现了图像编码模型的SOTA攻击
- 劣势
  - 针对目标模型输出是标签、置信度向量的情况下，攻击效果较差

总结

- 应用
  - 文本审计
  - 语音审计
  - 图像审计
- 未来工作
  - 对多模态模型进行成员推理攻击，实现（image，text）的成员推理【2022-NIPS】（算法原理如下图）
  - 对预训练图像编码器进行差分隐私，实现效用和隐私保护的效果
  - 在联邦学习场景中，实现源推理攻击，定位到具体的参与者



参考文献

[1] Liu H, Jia J, Qu W, et al. EncoderMI: Membership inference against pre-trained encoders in contrastive learning[C]//Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. 2021: 2081-2095.

[2] Song C, Shmatikov V. Auditing data provenance in text-generation models[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019: 196-206.

[3] Carlini N, Tramer F, Wallace E, et al. Extracting training data from large language models[C]//30th USENIX Security Symposium (USENIX Security 21). 2021: 2633-2650.

[4] Mahloujifar S, Inan H A, Chase M, et al. Membership inference on word embedding and beyond[J]. arXiv preprint arXiv:2106.11384, 2021.

[5] Shokri R, Stronati M, Song C, et al. Membership inference attacks against machine learning models[C]//2017 IEEE symposium on security and privacy (SP). IEEE, 2017: 3-18.

[6] Hu P, Wang Z, Sun R, et al. M^ 4I: Multi-modal Models Membership Inference[J]. arXiv preprint arXiv:2209.06997, 2022.

谢谢！

大成若缺，其用不弊。大盈若冲，其用不穷。大直若屈。大巧若拙。大辩若讷。静胜躁，寒胜热。清静为天下正。